Stellenbosch University

Department of Industrial Engineering

A Framework for Exploiting Electronic Documentation in Support of Innovation Processes

J.W. Uys

Dissertation presented for the degree of Doctor of Philosophy at Stellenbosch University.

Promoter: Prof. Niek du Preez

Co-promoter: Assoc. Prof Eric Lutters

Date: February 2010

Declaration

Declaration

By submitting this dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the owner of the copyright thereof (unless to the extent explicitly otherwise stated) and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: 10 February 2010

Verklaring

Deur hierdie tesis elektronies in te lewer, verklaar ek dat die geheel van die werk hierin vervat, my eie, oorspronklike werk is, dat ek die outeursregeienaar daarvan is (behalwe tot die mate uitdruklik anders aangedui) en dat ek dit nie vantevore, in die geheel of gedeeltelik, ter verkryging van enige kwalifikasie aangebied het nie.

Datum: 10 Februarie 2010

Opsomming

Die belangrikheid van innovasie vir die daarstel van ‘n volhoubare mededingende voordeel word tans wyd erken in baie sektore van die bedryf. Ook die belangrikheid van die toeganklikmaking van relevante inligting aan werknemers op die geskikte tyd, word vandag terdeë besef. Die afhanklikheid van effektiewe, doeltreffende innovasieprosesse op die beskikbaarheid van inligting word deurlopend beklemtoon in die navorsingsliteratuur.

‘n Groot uitdaging tans is om die oorsake en impak van die inligtingsoorvloedverskynsel in ondernemings te bestry ten einde werknemers in staat te stel om inligting te vind wat voldoen aan hul behoeftes sonder om in die proses deur oormatige groot hoeveelhede inligting te sif. Die aanvanklike stappe van die innovasieproses, gekenmerk deur vrye assosiasie, semi-formele aktiwiteite, konseptualisering en eksperimentasie, is reeds geïdentifiseer as sleutelareas vir die verbetering van die effektiwiteit van die innovasieproses in sy geheel. Die afhanklikheid van hierdie deel van die innovasieproses op inligting is besonder hoog.

Om op ‘n doeltreffende en optimale wyse te innoveer, benodig elke onderneming ‘n strategie vir innovasie sowel as ‘n aantal goed gedefinieerde, ontplooide prosesse en metingskriteria om die innovasieaktiwiteite van die onderneming te dryf. Bykomend benodig ondernemings sekere innovasie-ondersteuningsmeganismes wat bepaalde sleutelaanlegde, -tegnologieë en kennis insluit. Kern tot hierdie navorsing, benodig organisasies ook ondersteuningsmeganismes om hul in staat te stel om meer doeltreffend innovasie-verwante inligting te bestuur en te gebruik. Inligting, gehuisves beide binne en buite die grense van die onderneming, word benodig om die innovasieproses te voer. Die bronne van sulke inligting is veeltallig en hierdie inligting mag gestruktureerd of ongestruktureerd van aard wees. ‘n Toenemende persentasie van innovasieverwante inligting is egter van die ongestruktureerde tipe, byvoorbeeld die inligting vervat in die tekstuele inhoud van verslae, boeke, e-posboodskappe en webbladsye. In hierdie navorsing word die innovasielandskap asook tipiese bronne van innovasie-verwante inligting verken. Verder word die landskap van teksanalitiese benaderings en -tegnieke ondersoek ten einde maniere te vind om meer doeltreffend en optimaal met ongestruktureerde, tekstuele inligting om te gaan. ‘n Raamwerk wat aangewend kan word om ‘n verenigde, dinamiese voorstelling van ‘n onderneming se innovasieverwante inligting, beide gestruktureerd en ongestruktureerd, te skep word voorgestel. Na afloop van implementasie sal hierdie raamwerk die innovasieverwante inligting van die onderneming organiseer en meer toeganklik maak vir die deelnemers van die innovasieproses. Daar word verslag gelewer oor die aanwending van twee nuwerwetse, komplementêre teksanalitiese tegnieke tot aanvulling van die raamwerk. Voorts word die potensiële waarde van hierdie tegnieke as deel van die inligtingstelsels wat die raamwerk realiseer, verder uitgewys en geïllustreer.

Synopsis

The crucial role of innovation in creating sustainable competitive advantage is widely recognised in industry today. Likewise, the importance of having the required information accessible to the right employees at the right time is well-appreciated. More specifically, the dependency of effective, efficient innovation processes on the availability of information has been pointed out in literature.

A great challenge is countering the effects of the information overload phenomenon in organisations in order for employees to find the information appropriate to their needs without having to wade through excessively large quantities of information to do so. The initial stages of the innovation process, which are characterised by free association, semi-formal activities, conceptualisation, and experimentation, have already been identified as a key focus area for improving the effectiveness of the entire innovation process. The dependency on information during these early stages of the innovation process is especially high.

Any organisation requires a strategy for innovation, a number of well-defined, implemented processes and measures to be able to innovate in an effective and efficient manner and to drive its innovation endeavours. In addition, the organisation requires certain enablers to support its innovation efforts which include certain core competencies, technologies and knowledge. Most importantly for this research, enablers are required to more effectively manage and utilise innovation-related information. Information residing inside and outside the boundaries of the organisation is required to feed the innovation process. The specific sources of such information are numerous. Such information may further be structured or unstructured in nature. However, an ever-increasing ratio of available innovation-related information is of the unstructured type. Examples include the textual content of reports, books, e-mail messages and web pages. This research explores the innovation landscape and typical sources of innovation-related information. In addition, it explores the landscape of text analytical approaches and techniques in search of ways to more effectively and efficiently deal with unstructured, textual information.

A framework that can be used to provide a unified, dynamic view of an organisation's innovation-related information, both structured and unstructured, is presented. Once implemented, this framework will constitute an innovation-focused knowledge base that will organise and make accessible such innovation-related information to the stakeholders of the innovation process. Two novel, complementary text analytical techniques, Latent Dirichlet Allocation and the Concept-Topic Model, were identified for application with the framework. The potential value of these techniques as part of the information systems that would embody the framework is illustrated. The resulting knowledge base would cause a quantum leap in the accessibility of information and may significantly improve the way innovation is done and managed in the target organisation.

Acknowledgements

I would like to start by thanking Prof. Niek du Preez, for his continued guidance and support from the conception to the completion of this research. I further want to extend thanks to Assoc. Prof Eric Lutters for the detailed input and guidance he provided at several occasions over the duration of this research.

I would like to thank Indutech (Pty) Ltd for providing me with the opportunity to perform such a project. I would like to thank the Enterprise Engineering Group of the GCC, their international research partners, and in particular Dr. Louis Louw, Dr. Heinz Essmann and Mr. Corné Schutte for their insightful contributions.

I would like to thank my wife, Lisa, for her support throughout this project, my parents whom provided me with the opportunity to engage in tertiary study, and my family (of which my parents in-law deserve special mention) and friends for always showing an interest in my work.

Finally, I thank our Heavenly Father for blessing me with the opportunity, health and abilities to complete this research.

Table of Contents

List of Figures

List of Tables

Glossary

Browsing	Refers to the process of finding information by guided selection. With browsing the information seeker is presented with some kind of structure representing an overview about what information can be found. The seeker can subsequently select a category to arrive at relevant information content or more categories. Browsing is better suited in scenarios where the information seeker is not too sure of exactly what he is looking for.
Classification	Classification is the categorisation of items where the categories being used are known beforehand.
Clustering	Clustering is the categorisation of items where the categories being used are determined as part of the process.
Concept-Topic Model	A type of statistical topic modelling technique that uses human-defined concepts to guide the process of constructing a generative model for the documents analysed.
Document	A “document” is a digital, demarcated source of human readable textual information stored in a format suitable for information extraction. “Document” may also mean a computer file. See also “electronic document”. In this research report the word document shall mostly mean “electronic document”.
Electronic Document	An “electronic document” is a digital, demarcated source of human readable textual information stored in a format suitable for information extraction.
Entity	The term “entity” shall refer to an actual instance of a given entity type. For example, “USA” is an entity with corresponding entity type “Country”.
Entity Type	The term “entity type” shall refer to the class of an entity or thing (i.e. the kind of thing of a given entity). For example, “Banana” is an entity and the corresponding entity type of this entity is “Fruit”.
Explicit Knowledge	Explicit knowledge is consciously understood and can be articulated implying that the ‘knower’ is aware of such knowledge and can further converse about it.
Information Profile	A profile for an individual, which is automatically constructed based on the information the individual interacts with, which estimates and portrays the topics the given individual is dealing with. Can be considered to be broader than a “Research Interest Profile” that deals with research-focused information.
Latent Dirichlet Allocation	A type of non-hierarchical (or flat) statistical topic modelling technique that constructs a generative model for a given document collection and uses such a model to infer the topics underlying the analysed document collection without any human guidance.
LDA	See “Latent Dirichlet Analysis”
Noise Topic	A topic, generated by a statistical topic model, which seems illogical and cannot be associated with any theme by the human evaluator.

Probabilistic Topic Model	See “topic model”.
Research Interest Profile	A profile for an individual, which is automatically constructed based on the research-focused information the individual interacts with, which estimates and portrays the research areas or topics the given individual is dealing with. Can be considered as a special case of an “Information Profile”.
Searching	Refers to the process of looking for information, usually be means of querying. When searching for information, the information seeker mostly has a definite idea about what he wants to find information. He translates his search need into a search query, executes the search, and subsequently scans through the returned results.
Software Agent	A “software agent” is a computer program that intelligently executes certain tasks without human interaction.
Statistical Topic Model	See “topic model”. Also known as “probabilistic topic model”.
Stopword	A stopword is a word that has little semantic value and that is usually excluded in text analysis exercises. Examples include words like “and”, “or”, “about”, “the”, “a”, “actually”, “later”, etc. Stopwords are also known as “noise words”.
Stoplist	A stoplist is a specified set of stopwords that should be excluded from a given text analysis.
Structured Data	See “structured information”.
Structured Information	“Structured information” is information whose format and structure is predictable and ordered. For example, the information in a database table. Also known as “structured data”.
Tacit Knowledge	Tacit knowledge is the type of knowledge that the ‘knower’ is not necessarily aware of possessing. The capture of such knowledge, if possible at all, can only be achieved with difficulty using special interviewing and observation techniques.
Topic	In general language a “topic” is synonymous with “subject” or “theme”. In statistical topic modelling, a “topic” is a facet of a topic model and represents a computer-generated theme that was based on the content of the analysed input data. The combination of such generated topics aims to explain the essence of the observed data.
Topic Model	A generative model generated by a given statistical topic modelling technique by training it on a set of input documents with the goal to explain the words encountered in the individual input documents in terms of mixtures of different deduced topics.
Unstructured Data	See “unstructured information”.
Unstructured Information	“Unstructured information” is information whose format and structure is not predictable. Unstructured information content is therefore not ordered or is ordered inconsistently. For example, the text of a newspaper article. Also known as “unstructured data”.

1. Introduction

"Innovation happens, within and between organisations, at the intersection of diverse information flows and knowledge exchanges." Krebs (2008)

The topic of innovation is receiving due attention in literature and business publications. Innovation is often regarded as a "soft" science and it is hard to define and measure (Hering & Phillips, 2005). Innovation can be described as the introduction of a value-adding new idea - in an organisation or market - in the shape of a new product or service, an improvement to a process or to the organisation itself. For organisations to differentiate themselves from their competitors and to truly innovate, they must be more efficient in leveraging the information residing outside the setting of their existing information systems (IBM, 2006). The same argument applies at departmental level for information residing outside the scope of a given department. Managing innovation is complex due to many reasons; some which include the fast rate at which new value offerings have to be realised, the large number of knowledge areas involved in a complete innovation, the potential communication barriers that may exist in multidisciplinary teams, and last but surely not the least, the resistance to change engrained in human nature. The availability of appropriate information to support decision-making throughout an organisation's innovation process as well as a shared understanding of the organisation's innovation concepts supporting effective communication may considerably facilitate the execution and management of innovation.

This chapter presents the backdrop of how information, knowledge and innovation are interrelated. The specific objectives of this study are addressed in Chapter 2, while the innovation landscape is presented in detail in Chapter 3.

1.1 The Importance of Information

The importance of information for economic growth and establishing a competitive advantage is increasingly recognised. Information is further increasingly recognised as a key ingredient to business innovation. To support this crucial role of information, the management of information across the organisation is becoming indispensible. Although most organisations have the information they require, they are ineffective in using it constructively due to the lack of understanding the organisation's information as a complete whole. The spread of information across heterogeneous sources in the organisation complicates the process of accessing information and further worsens the problem of putting information to valuable use (IBM, 2006). The one extreme is that the innovation workers (i.e. the employees contributing to innovation projects and activities) of an organisation are information-starved, whereas the other extreme is that innovation stakeholders are suffering from information overload; none of these extremes are conducive to an environment that promotes effective and efficient innovation. The desired situation is providing innovation stakeholders with access to low-volume, high-quality, contextualised information to facilitate their innovation efforts.

1.2 Some Challenges of Innovation

Although information is seen as one of the enablers of innovation, the availability of information alone does not lead to innovation (Dewett, 2001). Another important consideration of successful innovation is the way in which an organisation deals with integration (cf. Souder & Moenaert, 1992). Other factors such as the availability of resources and expertise, organisational culture, etc., further determine the rate of successful innovation in organisations. A common problem in large organisations is their inability to realise new business opportunities by combining resources from disparate parts of the organisation (cf. Kleinbaum, 2006). The fact that few organisations have a common understanding of what their innovation-related resources are, and what the capabilities of such resources are, adds to the difficulty of effectively exploiting innovation opportunities and managing the innovation process.

If organisations had an effective way of having an up-to-date description of all its innovation-information resources (including their inter-relationships) as well as knowledge profiles for all its key innovation workers (including affinities between the different profiles of such individuals), the process of combining, managing and growing such innovation resources would be much easier.

1.3 Intra-organisational Innovation Networks

Innovation networks mostly imply that different organisations collaborate to realise a complete innovation feat, actually referring to inter-organisational innovation networks. Innovation networks may also exist within a given organisation, referred to as intra-organisational innovation networks. In order to offer complete innovations, rather than just products, successful innovators join up to form innovation networks giving them access to competencies in a flexible manner. A trend to move away from the traditional linear model of research and development as the foundation of innovation towards a model where partners integrate complementary competencies in the form of a network currently exists.

It is not only inter-organisation networks that can promote successful innovation – intra-organisation innovations networks could also realise many of the benefits of inter-organisation innovation networks mentioned earlier. Therefore, organisations must adopt a more proactive, distributed approach to innovation within the organisation. This can be done by creating an organisation-wide, internal innovation network comprised of innovation workers. Moreover, the innovation processes in such an internal (or intra-organisation) innovation network should be integrated in such a manner that new innovation opportunities can be identified early and planning of the required innovation activities in the network can be done effectively (cf. Bullinger, Auernhammer & Gomeringer, 2004). To enable cross-divisional innovation (in the context of intra-organisational innovation networks) both informal structure - to alleviate the flow of information across internal organisational boundaries to uncover opportunities to collaborate - and formal structure - needed to officially acknowledge and implement cross-divisional innovation - are required (Kleinbaum, 2006).

1.4 Information Overload and Innovation

Information is as diverse as the doings and interests of mankind. Information has traditionally been seen as a sought after item; the more (relevant) information one can lay your hands on the better as it will provide you with additional or more complete insights which may distinguish you from others or improve the quality or speed of decisions. Therefore, there it would seem that there are no downsides to having or attaining information (Kotze, 2008).

The Information Revolution has equipped mankind with the technology and know-how to almost effortlessly communicate information on a global scale (Kotze, 2008). With the advent of the Internet a new era in information sharing has begun. New (software) tools were further developed as part of the Web 2.0 paradigm to empower users to generate and publish their own content on the Internet (e.g. blogs, wikis, social networking sites, etc.), a privilege previously reserved to a selected group of authors and publishers. Where in the past being starved for information, we now find ourselves overloaded with information (Borchers, Herlocker, Konstan & Riedl, 1998).

Due to the perceived overabundance of information in the world (for example modern scientists can search digital libraries spanning hundreds of years according to Blei and Lafferty (2009)), it is relatively easy to find information about almost any topic. The usefulness or appropriateness of a given piece of information is defined by the context of the user’s specific information needs, the quality of the retrieved information, and recentness of the retrieved information, to name a few characteristics. In summary, the retrieved information must thus conform to a set of contextual requirements to satisfy the information seeker’s needs (Kotze, 2008).

In the scenario where information is abundant, the knowledge about what information is useful and valuable matters most (Borchers et al., 1998) since it provides the key to the efficiency of the information retrieval process. Kotze (2008) puts forward that this knowledge generally eludes information seekers complicating the task of finding appropriate information in a time efficient way. Regarding information available on the Internet, this problem has partially been addressed by the advent of Web 2.0 technologies that make available the ‘wisdom of the crowds’, in the form of (quality or usefulness) ratings and descriptive tags, as filtering mechanism to information seekers.

The problem arising due to the condition of readily having access to vast amounts of information manifests in the adverse effect on the information retrieval effort. This condition is termed ‘information overload’ (Kotze, 2008). Information overload may be defined from an information processing view of the organisation as the situation where the information processing requirements exceeds the information processing capabilities (Eppler & Mengis, 2004). It is not only the quantity of information that leads to information overload however; the characteristics of information or the qualitative dimension of information may also contribute to information overload (Eppler et al., 2004). At present, where ongoing innovation is crucial to ensuring an organisation's sustained competitive advantage, the availability of time efficient ways of accessing high-quality information to support day-to-day business decisions is more important than ever. Organisations need to exploit available information (e.g. information pertaining to new markets, customer requirements, new technology, competitor activities, new legislation, etc.) and adapt to new opportunities in increasingly short time periods necessitating new methods to quickly arrive at relevant information. Information overload may therefore be an indirect threat to efficient innovation and the competitive advantage of the organisation.

The ease with which information can be captured and shared within an organisation presently, has led to the situation where internal information may also suffer from poor accessibility and quality problems. The organisation's internal information may also contribute to the information overload phenomenon (Kotze, 2008) in terms of sheer quantity as well as its characteristics. Kotze (2008) further discusses a survey on various information management issues (developed by VNU Exhibitions Europe with the support of the Information Management Professional Group). A total of 648 respondents completed the survey in 2006 and the main information management challenges of the participating organisations were concluded to be the following:

- Information overload and improving the quality of internal information (35%)
- Poor or no document management (25%)
- Sourcing the right external information (24%)

In spite of the efforts of organisations to capture, store and refine information only a small fraction of organisational information is actually ever used. Carlson (2003) reports that as little as 20% of all organisational information filed is ever used. This may be due to the following reasons:

- Mostly irrelevant information is gathered by the organisation
- Little awareness is created of what information the organisation actually possesses
- Poor access to information prevents it to be used since it cannot be found.

Moreover, Carlson states that 71% of organisational workers maintain that their main task is finding information. One may speculate that a lot of resources are actually wasted in finding inappropriate information. It may be concluded that much potential exists in improving awareness of what information the organisation has, improving access to information within the organisation, and lastly implementing more efficient and effective ways for employees to find information.

The information overflow phenomenon may also adversely affect the activities forming part of the fuzzy front end of the innovation process (refer to section 3.4 for a detailed discussion of the fuzzy front end innovation activities) where ideas and opportunities are contextualised and innovative concepts are developed and refined. The efficiency of the fuzzy front end innovation activities mainly depends on the knowledge, abilities, commitment and creativity of people. Another large factor influencing the efficiency of fuzzy front end innovation activities is the appropriateness, quality and availability of supporting information. The focus of this research is on developing a framework that would facilitate the supply of appropriate information to the organisation's innovation workers without inducing or aggravating information overload.

1.5 Knowledge Management and Innovation

The primary objective of knowledge management (KM) is to support decision making based on information and experience that exist in the organisation (Rath, 2003). Moreover, KM should fast-track the adoption of organisational learning and should grow the organisation's collective intelligence. On a high level, KM facilitates the knowledge cycle consisting of knowledge recognition, capture, distribution, utilisation and ultimately starting once more with the recapturing of knowledge. The rate at which organisations tap the knowledge cycle is progressively seen as a benchmark of the organisation's "fitness level". In KM, organisation's have a powerful instrument to combat the information avalanche by managing and accessing the endlessly growing amount of diverse, highly interlinked, and complex information.

Knowledge management initiatives over the past years can be classified into three generations (Walsham, 2001). The first generation of KM initiatives concentrated on the conception of knowledge repositories. KM solutions associated with this generation often turned out to be unsuccessful because much of the knowledge in such repositories was considered to be irrelevant to the personal context of the knowledge user. The second generation of KM initiatives concentrated on delivering more individualised knowledge to groups of knowledge users (e.g. push solutions). KM solutions of this generation were only successful where the target groups' requirements had been predicted successfully. The third (and current) generation of KM initiatives has the focus of sustaining communication between individuals, e.g. communities of practice (as an example of a knowledge network). Knowledge networks can be seen as being part of the third generation knowledge management initiatives (Schönström, 2005). Tacit knowledge is usually accessed by means of knowledge networks. It is therefore reasonable to speculate that informal knowledge networks in organisations can greatly support the activities of internal innovation networks. An effective knowledge network is comprised of a combination of individuals knowing how to do things and who knows how to do which things (Kogut & Zander, 1992; Carley, 1999; Monge & Contractor, 2002; Rogers, 1995).

It is widely accepted fact that the (tacit) knowledge and expertise of employees is the main asset of knowledge-based organisations. Expertise specifies an organisation's distinctive capabilities and core competencies (Finley, 2001; Holloway, 2000; Olson & Shaffer, 2002), while tacit knowledge is further regarded as a crucial component of innovation and product development processes (Grant, 1996; Hall, 1993; Nonaka & Takeuchi, 1995). Since tacit knowledge cannot readily be codified, it is not (yet) possible to diffuse it by means of information and communication technology (Bullinger et al., 2004). Therefore, to share tacit knowledge the spatial proximity of knowledge and innovation workers is necessary as it can only be transferred by means of richer communication media, such as face-to-face communication (Bullinger, et al., 2004; cf. Kleinbaum, 2006). The best way for an organisation to exploit the knowledge and expertise of its employees is by enhancing communication between employees in order for them to share their knowledge and expertise. Finding individuals with specific expertise, a process known as 'expertise matching', is receiving increasing attention because of its ability to connect people to each other (Liu & Dew, 2004). Given that the cutting edge is constantly changing, the real purpose of information systems in organisations should be to support the connection of people to people, so that they can share the expertise and knowledge they possess (cf. Stewart, 1997). Connecting people to people is an important enabler of successful intra-organisational innovation. Therefore, the role of technology in the challenge of unlocking knowledge in organisations is not only to facilitate access to explicit (documented) information and knowledge, but more importantly, to facilitate the process of accessing the tacit knowledge held by individuals. Technology can therefore fulfil a key role in the processes associated with intra-organisational innovation by catalysing collaboration in knowledge sharing among individuals (cf. Yimam-Seid & Kobsa, 2003). To make this possible, the visibility and traceability of knowledge in an organisation, tacit and explicit, needs to be enhanced. The need for internal expert location facilities as well as other effective information-, knowledge and communication infrastructures grows with the increase in the size of an organisation, an increase in the degree of geographical distribution, and an increase in the degree of heterogeneity in the composition of its members (Yimam-Seid et al., 2003). Organisations should endeavour to improve their information and communication infrastructures so that they are able to deal with the semantic descriptions of integrated, semi-structured data and offer advanced knowledge services on top of them (Maier, Hädrich & Peinl, 2005).

To move beyond the current state of the art in managing the knowledge assets of an organisation, a framework representing the information- and knowledge landscape of an organisation should be specified, created and maintained. This framework should be used to organise and make accessible the structured and unstructured information content of the target organisation to support and expedite the information-focused activities of employees. Such a framework needs to address all types of entities in the organisation that forms part of the organisation's collective intelligence. Such a framework would be instrumental in facilitating:

(i) Early detection of new information and knowledge,

(ii) Connecting knowledge and innovation workers to the most suitable explicit knowledge and information within the organisational boundaries,

(iii) Connecting knowledge and innovation workers to other experts in the organisation to trigger tacit knowledge transfer, cross-learning and the expansion of the informal knowledge network, and

(iv) Relieving the symptoms of information overload in the organisation.

Section 3.11 will elaborate further on the relationship between knowledge management and innovation.

1.6 Plan of Development

The structure of this dissertation is depicted in Figure 1 below.

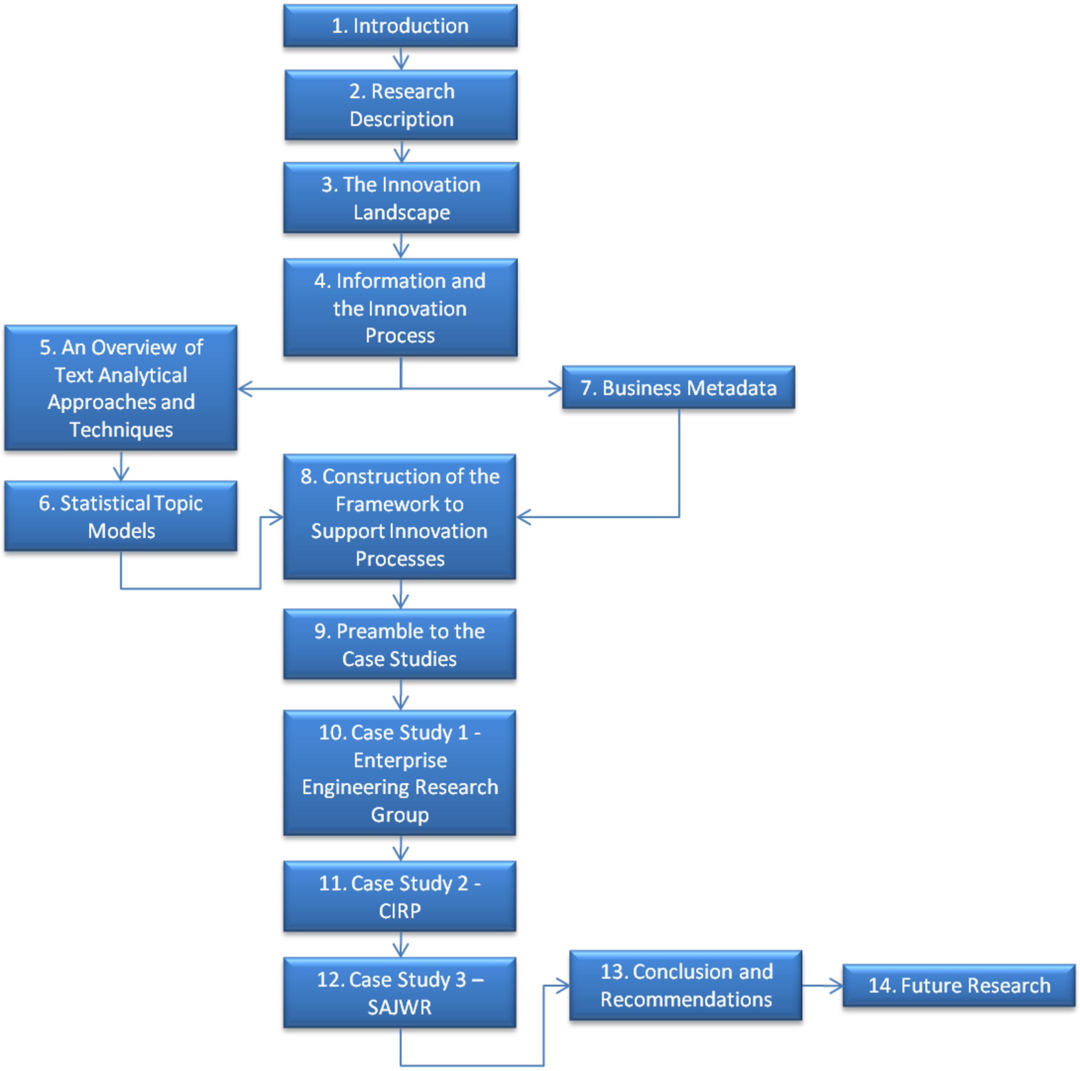

1. Introduction
2. Research Description
3. The Innovation Landscape
4. Information and the Innovation Process
5. An Overview of Text Analytical Approaches and Techniques
7. Business Metadata
6. Statistical Topic Models
8. Construction of the Framework to Support Innovation Processes
9. Preamble to the Case Studies
10. Case Study 1 - Enterprise Engineering Research Group
11. Case Study 2 - CIRP
12. Case Study 3 – SAJWR
13. Conclusion and Recommendations
14. Future Research

Figure 1: Outline of the development of this dissertation

2. Research Description

This chapter will define the research that will be completed in this study.

This chapter addresses the following:

- **State the objectives of this research study**
- **Explain the derivation of the high-level research objective of the study**
- **Specify the scope and focus of the research**
- **Describe the deliverables and outcomes of the study**
- **Discuss the contribution made by the research**
- **Describe the validation strategy**

Although organisations are normally complex, dynamic networks consisting of a myriad of different entities and processes, most modern organisations have at least two things in common. Firstly, they are made up of people and secondly, they generate and interact with textual documents on a daily basis. Documents are sources of information and explicit knowledge about a wide range of organisational topics. Documents are seldom shared throughout the organisation and often reside in departmental silos increasing the difficulty of finding and accessing such documents. Documents further have explicit and implicit relationships to one another, e.g. whether they address the same topics, have been created by the same author, are of the same type, have been created on the same day, etc. Other useful information to the organisation's endeavours resides outside the organisation in the form of scholarly articles website pages, magazines, brochures, etc. The organisation gathers such information as and when required to support specific activities. These retrieved documents are however not always stored in an accessible organisation-wide repository and shared with the appropriate people within the broader organisation.

The employees of the organisation, on the other hand, are bearers of tacit knowledge or expertise about the organisation and other matters. The subset of employees that busy themselves with information and knowledge work on a regular basis are called information workers and knowledge workers respectively. Employees also have explicit and implicit connections to one another, e.g. when they have the same background, similar business roles, works on similar projects, share the same ideas, read the same documents, etc. Employees interact with documents on a daily basis by generating, reading, editing and distributing such documents. Therefore, explicit and implicit relationships also exist between employees and documents in an organisational setting.

Every organisation further has available to them a number of potential innovation opportunities pertaining to its products, services, structure and processes. From this set of innovation opportunities, an organisation can formulate certain innovation goals to be achieved in the medium to longer term. The exploration of such opportunities and ideas normally happens in the Fuzzy Front End (FFE) of the innovation process, which is mostly characterised by unstructured thinking, conceptualising, free-association and exploration activities. A great potential exists to increase the throughput and the effectiveness of the innovation process and the early stages of the innovation process in particular, as it would ensure that resources are spent on identifying and maturing opportunities with high probability of success as opposed to trying to focus on everything and achieving little at the end.

By connecting the stakeholders of the innovation process with information appropriate to their needs, and with other persons in the organisation sharing similar ideas or having the knowledge or expertise to further the relevant innovation idea along the innovation lifecycle, the effectiveness of the innovation process may be increased substantially. The information needs of the various innovation workers may depend on their respective innovation role (refer to section 3.5 for a discussion of the various roles in innovation).

2.1 Definition of the Proposed Research Objectives

In the light of the backdrop sketched in the preceding paragraphs, the following research question would apply:

> *How can electronic, textual information be used to better support the innovation processes of an organisation?*

This leads to the following high-level aim of this research:

> *To develop a framework and working methods to exploit information contained in electronic textual documents to better support the innovation processes of an organisation.*

In relation to the research question above, the developed framework will be the vehicle required to organise and capture distillations of such textual information, while text analytical techniques will serve as the mechanisms required to distil such information from the multitude of sources of unstructured information. More specifically, such a framework and associated working methods will facilitate the compilation of a compact overview of all innovation-related information in an organisation. In addition, it will facilitate connecting individuals (e.g. knowledge workers, experts, etc.) to relevant information and to other individuals in context of a given domain topic. Even though this research is focused on supporting the innovation processes of an organisation, it has much wider applicability since it may be useful to any knowledge-driven organisation interacting with electronic documents.

Figure 2 summarises the derivation of the research high-level objective of this study.

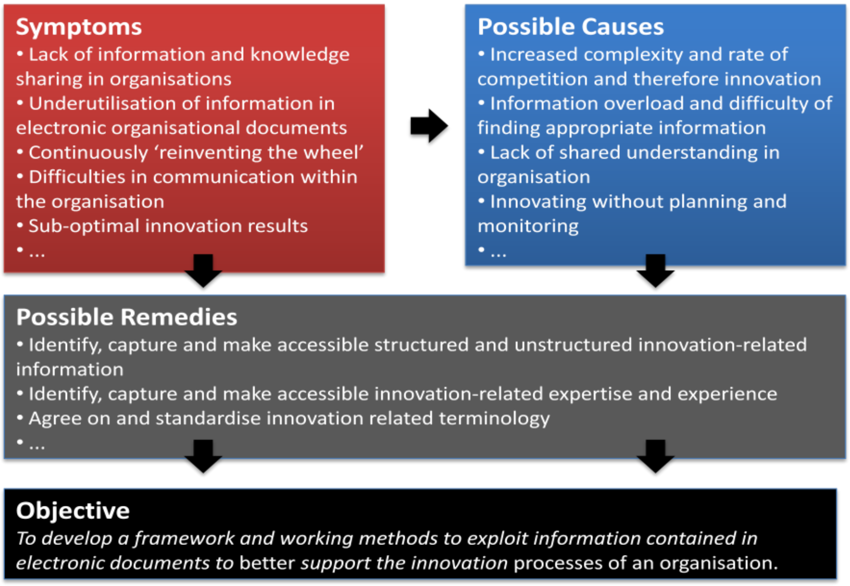

Symptoms
• Lack of information and knowledge sharing in organisations
• Underutilisation of information in electronic organisational documents
• Continuously 'reinventing the wheel'
• Difficulties in communication within the organisation
• Sub-optimal innovation results
• ...
Possible Causes
• Increased complexity and rate of competition and therefore innovation
• Information overload and difficulty of finding appropriate information
• Lack of shared understanding in organisation
• Innovating without planning and monitoring
• ...
Possible Remedies
• Identify, capture and make accessible structured and unstructured innovation-related information
• Identify, capture and make accessible innovation-related expertise and experience
• Agree on and standardise innovation related terminology
• ...
Objective
To develop a framework and working methods to exploit information contained in electronic documents to better support the innovation processes of an organisation.

Figure 2: Derivation of the research objective of this study

The more specific research objectives of this study are:

1. Develop a way to promote the dissemination of information in organisational documents
2. Develop a way to capture and make accessible indications of the expertise of individuals in the innovating organisation to promote communication and knowledge transfer between employees
3. Develop a way to realise and make accessible a 'corporate memory' with regard to innovation with the aim to unify structured and unstructured innovation-related information

2.2 Scope and Focus of the Study

This study will focus on transforming large quantities of unstructured, textual information into more concentrated, organised and useable forms. The specific focus would be on electronic, textual documents containing information about innovation-related matters. To achieve this transformation of textual information, suitable techniques will be investigated. The ways in which such textual information is gathered and stored will not form part of this study as it is assumed that the target organisation possesses such innovation-related, electronic, textual documents or the means to obtain these. This research will further focus on the design of an information framework, which will cater for the most important innovation-related information entities and the relationships between such entities, to be used to contextualise and make accessible information extracted from textual documents. Along with the framework appropriate working methods will be developed to guide the interaction with the framework to access the organisation's innovation-related information. In terms of the innovation process, this research will focus more, but not exclusively, on supporting the initial stages of the innovation process, namely the Fuzzy Front End, due to its unstructured nature and large dependency on unstructured information.

The developed framework and working methods should be applicable to both product and service organisations dealing with innovation. The usefulness of such a framework and accompanying working methods will not so much be determined by the size of the organisation, but more by its degree of geographical distribution and maturity in terms of innovation practices and capabilities. Lastly, the outputs of this research would support incremental innovation, but will possibly be even more useful to support radical innovation efforts.

2.3 Deliverables and Outcomes of the Study

The following deliverables are applicable to this research project:

- A framework and working methods that can be used to distil and make accessible the unstructured information in electronic documents
- A list of functionalities desired for the text analytical systems associated with the framework
- A basis for a strategy for implementing such a framework in an organisation

Moreover, the following outcomes are obtained:

- An overview of different text analytical techniques
- A description of the most prominent sources of innovation-related information
- Demonstrated added value of applying text analytical techniques to collections of electronic documents
- A consolidated portfolio of future research around the construction, implementation and refinement of the developed framework and working methods.

2.4 Contribution of the Research

This research expands the Innovation Management body of knowledge by exploring the information dimension of innovation. More specifically, it presents ways in which unstructured, textual information can be processed automatically and contextualised with regard to an organisation's key innovation-related information entities which are identified and characterised as part of the framework developed in this research. In addition, this research also contributes to the fields of Knowledge Management and Information Management by reflecting on possible ways to use such a framework and accompanying techniques to improve the way in which knowledge and information are used in an organisation.

Lastly, this research also made a contribution to the fields of Machine Learning and Text Analytics in that it explored ways to manipulate the results returned by the standard topic model to infer relationships between other entity types such as persons, departments, etc.

2.5 Validation Strategy

Possible ways of transforming the information contained in electronic documents and possible applications of such transformed information in an innovation setting were explored and validated by means of experimentation in three case studies. Although these case studies were not executed in a commercial innovation environment, each of them contains a great deal of elements that can be extrapolated to an actual innovation-driven organisation. In all three cases, the document collections to be analysed were selected so there would be persons available that understands the relevant fields and that can assist with interpretation and evaluation where required. Where human evaluators did not give sufficient evaluation feedback, ways to evaluate the accuracy of the transformed information based on the metadata of the analysed documents were formulated (e.g. using the combination of all keywords linked to documents associated with a given computer-generated topic, to evaluate the computer-generated words characterising the specific topic).

As an additional way of validating the working methods and techniques recommended by this research, the author extensively used topic modelling to organise his own research material throughout this study. This not only automatically identified the topics underlying the 400 odd electronic research documents collected during this study, but also helped the author to identify documents pertinent to a given topic of interest as well as to find documents closely related to a specific document. This sped up the process of interacting with research material and finding appropriate references tremendously.

This concludes the discussion around the description of this research. The next chapter will investigate the specific information sources that may contain information to support the innovation processes of an organisation.

3. The Innovation Landscape

"An innovation is an investment into future profits that will secure survival of the company in the market by maintaining or extending its market share." Pleschak & Sabisch (1996)

There are many different models and views on the innovation process ranging from simple linear processes to more complex, iterative, networked processes. All organisations have restrictions on available resources and allocating these resources over a large number of projects often leads to limited progress made on numerous fronts, but few actual projects reach the end goal of commercial success due to the too diverse focus (Pretium, 2005). It is therefore important to collect and characterise innovation ideas and concepts early in the innovation lifecycle to decrease uncertainty and increase the ease of making informed decisions about which ideas and concepts to pursue further to minimise risk and maximising possible future competitive advantage. Effective innovation requires a delicate mixture of the absence of structure – to foster creativity and the flow of information within the organisation – and formal structure – to drive the development of innovation ideas and concepts to commercial successes by allocating resources, budgets and milestones. The former requires an organisational environment conducive to innovation, while the latter requires a formal, institutionalised innovation process.

This chapter addresses the following:

- **Present an overview of different types of innovation**
- **Discuss the evolution of innovation models**
- **Present an overview of the innovation process according to the Fugle^{TM} model**
- **Introduce the Fuzzy Front End (FFE) of innovation**
- **Discuss the various stakeholder roles in the innovation process**
- **Elaborate on possible sources of innovation ideas**
- **Discuss different information realms pertinent to the information needs of the FFE**
- **Investigate the interaction between the information realms and the different stages of the FFE**
- **Investigate the involvement of information realms in terms of the different stakeholder roles**
- **Elaborate on the relationship between innovation and knowledge management**

The following definitions make explicit the intended meaning of some of the key innovation terminology used in this chapter.

An Opportunity: In the context of innovation an opportunity is a business or technology gap, recognised by an organisation or an individual, which exists between the current situation and an envisioned future, where the addressing of such a gap would result in capturing competitive advantage, responding to a threat, or solving or improving a problem. (Koen, Ajamian, Boyce, Clamen, Fisher, Fountoulakis, Johnson, Puri & Seibert, 2002)

An Idea: An idea is the most embryonic form of a new product or service that often consists of a high-level view of the solution envisioned for the problem identified by the opportunity. (Koen et al., 2002)

A Concept: A concept has a well-defined form, in terms of both written and visual descriptions, which includes its primary features and customer benefits combined with a broad understanding of the technology required. (Koen et al., 2002)

3.1 Types of Innovation

The most basic distinction between different types of innovation distinguishes between **radical innovation** and **incremental innovation**. Radical innovation creates dramatic change in technology, processes, products, and/or services that considerably transforms existing markets and industries, or even gives rise to new ones. Incremental innovation, on the other hand, is grounded on extending existing technologies, focusing on reduction of cost or the improvement of the features of existing products, services, or processes (Miller, Miller & Dismukes, 2006).

Moore (2005), in his book *Dealing with Darwin*, discusses fourteen types of innovation and where they fit in the innovation lifecycle. These different types of innovation become important in different stages of the innovation lifecycle. Moore distinguishes four innovation stages or zones:

1. Product leadership zone
2. Customer intimacy zone
3. Operational excellence zone
4. Category renewal zone

Firstly, **disruptive innovation** correlates with the early market phase of the technology adoption lifecycle (TALC) where a product or service offering makes its first appearance in the market in any form. With **application innovation** the product or service offering is gaining acceptance by covering a single, specific application need. In the case of **product innovation** the product or service is gaining widespread adoption for several applications due to increasing attractiveness in terms of price or performance. When the product or service offer achieves a state where it is an enabling component of new types of offerings it is known as **platform innovation**. These four types of innovation types are applicable to growth markets and forms part of the product leadership innovation zone.

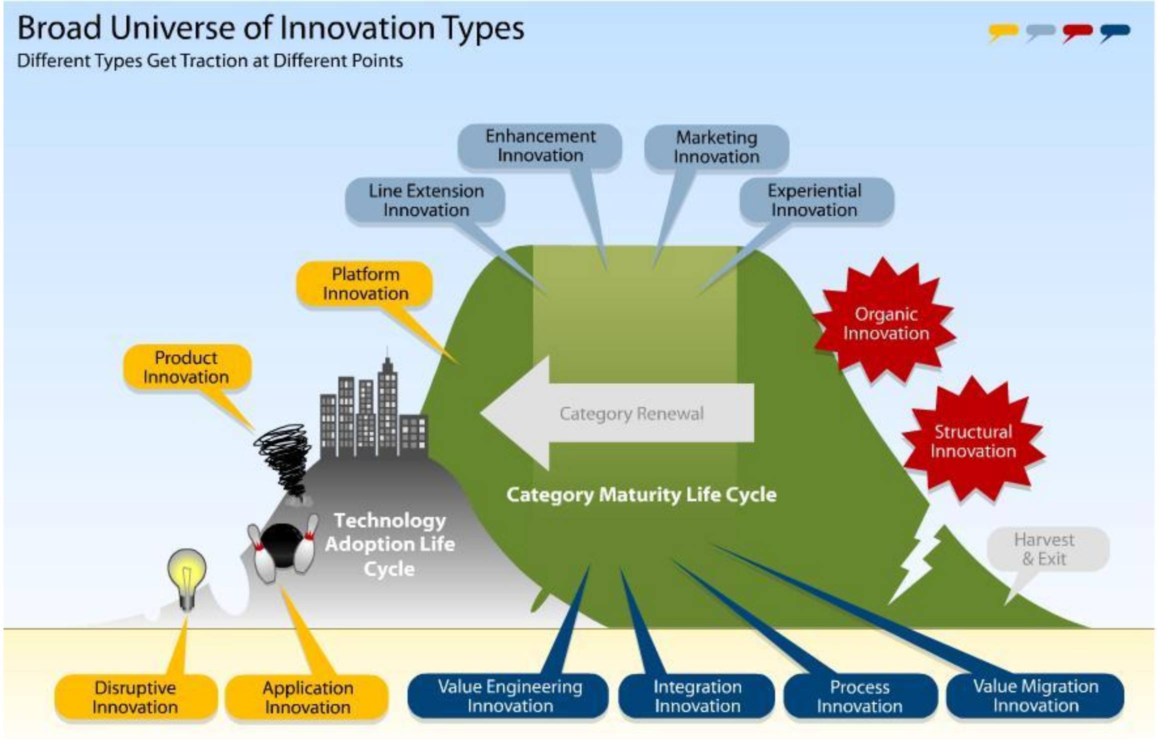


Figure 3: Moore's fourteen types of innovation[1]

For mature markets four innovation types, all having an optimisation feel, exists. **Line-extension innovation** entails structural alterations to an existing offer to establish a unique sub-offering in order to expand the market. **Enhancement innovation** expands on line-extension innovation by making further, finer adjustments to an existing offering to rekindle customer interest. **Marketing innovation** focuses on distinguishing the interaction with prospective customers during the purchase process to outsell competitors. **Experiential innovation** is the pinnacle of customer intimacy and is centred on distinguishing in terms of the experience of the offering. These four types of innovation form part of the customer intimacy innovation zone which focuses on the demand side of the market.

The operational excellence zone complements the customer intimacy zone, but focuses more on the supply side differentiation. Four innovation types are part of the operational excellence zone. **Value engineering innovation** reduces the cost of the realisation of an existing offering without changing its appearance. **Integration innovation** reduces the customer's cost of maintaining a complex process by the integration of heterogeneous components into a unified centrally managed system. **Process innovation** eliminates waste from the processes enabling the offer to increase profit margins. **Value migration innovation** entails redirecting the business model away from a saturated part of the market value chain toward a part characterised by more rewarding

[1] Source: Moore (2005)

margins. The category renewal zone deals with declining markets and two types of innovation pertain to this zone. **Organic innovation** (also known as organic renewal) occurs when an organisation repositions itself into growth category by using its internal resources and signifies a return to product innovation while remaining in the same product line. **Structural Innovation** (also known as acquisition renewal) addresses the category renewal challenge by means of merger with or acquisition by another company. The innovating company may be the acquiring party or the acquired party. (Moore, 2005)

These different innovation types, along with an understanding of where in the market lifecycle they apply, provides a framework for examining the market forces influencing the competitive advantage strategy of an organisation.

3.2 The Evolution of Innovation Models

The different phases of the innovation lifecycle – from idea to commercialised product or service – and the management of the innovation process are well described in literature (Rothwell, 1995; (Tidd, Bessant & Pavitt, 1998; Trott, 2005). The majority of the innovation process representations found in literature involves the following high-level phases:

- Generation and identification of innovation ideas
- Development of innovation concepts
- Evaluation and selection of innovation concepts
- Development of the actual new product, process or technology
- Implementing the developed product, process or technology

Seven generations of innovation process models have evolved over time (Rothwell, 1992; Du Preez, 2009) ranging from simple linear models to more complex interactive models.

Model Characteristic	Generation	Characteristic
Technology Push	1st	Simple linear sequential process, emphasis on R&D and science
Market Pull	2nd	Simple linear sequential process, emphasis on marketing, the market is the source of new ideas for R&D
Coupling Model	3rd	Recognizing interaction between different elements and feedback loops between them emphasis on integrating R&D and marketing
Interactive Model	4th	Combinations of push and pull models, integration within enterprise, emphasis on external linkages
Network Model	5th	Emphasis on knowledge accumulation and external linkages, systems integration and extensive networking
Open Innovation	6th	Internal and external ideas as well as internal and external paths to market can be combined to advance the development of new technologies
Extended Innovation Network	7th	Combining network models and open innovation into formally structured innovation and exploitation frameworks such as the Fugle™.

Table 1: Seven generations of innovation process models

The first and second generation innovation models can be typified as linear models – involving little or no interactive steps – that define innovation as either resulting from market needs or technology and scientific discoveries.

The third generation innovation model is a coupling model in that it takes into account the effects of technological capabilities of the innovating organisation as well as the applicable market needs. The coupling model contains feedback loops between its comprising steps, but is fundamentally a linear model due to limited functional integration. A well-known example of a linear innovation process model is the Stage-Gate model (Cooper, 1990) that promotes improved quality throughout the innovation process. The gates of the Stage-Gate model serve to ensure comprehensiveness in the innovation process by guarding against the omission of critical activities. This model may be criticised as being too stringent in terms of its gates that may impair freedom in the first part of the innovation lifecycle, namely idea and concept generation. Another critique voiced concerning this model is the linearity of the approach that may not be suitable for more radical innovations characterised by higher levels of uncertainty and requiring more lenient, learning-based approaches. Radical innovation processes calls for more iterative loops between the idea generation and concept definition stages to incorporate learning through experimentation and modelling.

The fourth generation innovation process model improves on the weak presence of functional integration in the linear models by catering for more interaction between the various stages of the innovation process (Rothwell, 1995). In this case the innovation process is represented as parallel activities across the functional boundaries of the organisation. An early example of a fourth generation model is the Minnesota Innovation Research Program (MIRP) model that was conceived during the 1980's (Hildrum, 2007). This model defines a number of fundamental innovation process characteristics that are involved as the innovation idea progresses through the innovation lifecycle. It further distinguishes between three successive periods in the innovation cycle, namely the initiation, development and implementation periods. The largest shortfall of this model is that it stops the innovation process before the actual implementation. In reality, the innovation process continues throughout the implementation period as continuous improvement and adjustments are often required.

The fifth generation innovation models can be described as network models. These models were conceived in the 1990's with the aim to represent the true complexity of the innovation process. Network models incorporate the effects of the external environment on innovation and cater for communication between internal and external innovation stakeholders. The Creative Factory Concept (Galanakis, 2006) can be considered as an example of a networked fifth generation innovation processes model and is based on a system thinking approach. In this model, the innovation process is comprised of three lower-level processes:

- Knowledge creation - exploiting information from public domain or industrial research

- New product development - transforming knowledge into new products
- Product success in market - determined by the functional competencies of the product and the competencies of the organisation to realise the product at a reasonable price and quality and adequately introducing it into the market.

Fifth generation models are mostly closed networks of innovation as innovation ideas are developed in secrecy within the boundaries of the organisation.

Open innovation models can be described as the sixth generation innovation model (Du Preez & Louw, 2008). Innovation models of this generation are also networked innovation models, but combine internal and external ideas and paths to the market to fast track the development of new products and technologies. The concept of open innovation was coined by Chesbrough (2003) and is considered beneficial due to the increased amount of ideas and technologies that are made available to the organisation to foster internal growth and innovation. Another benefit of open innovation is that it provides a mechanism to explore new growth opportunities at a lower risk (Docherty, 2006) and less resources. Organisations should develop integrated knowledge networks, spanning outside the traditional organisational boundaries, to fully exploit the concept of open innovation and support the innovation knowledge supply chain (Du Preez et al., 2008).

Lastly, the seventh generation model, according to Du Preez (2009), is the Extended Innovation Network that combines open innovation and network innovation models.

A synthesis from literature about innovation process models indicates the following key points (Du Preez et al., 2008):

- The majority of innovation process models comprises the following stages
  - Idea generation and idea identification
  - Concept development
  - Concept evaluation and selection
  - Actual new product, process or technology development
  - Implementation of the developed product, process or technology
- Innovation may result from a market pull, a technology push or a combination of these
- The integration of the various functions of the innovation process is of utmost importance to the success of innovation projects
- The latest innovation process models involves a network approach which extends the traditional internal focus to include external innovation partners and ideas
- Many innovation process models exclude the exploitation of the new innovation in the market. Such exploitation is central to increased competitiveness and should therefore be included in the innovation process.

A one-size-fits-all model for innovation is not possible due to the multitude of variables impacting the innovation and design process. A comprehensive innovation model is therefore required that can be adapted to the innovation needs of a given organisation. The next section describes the Fugle^{TM} innovation model to give the reader more insight into the activities of each of the stages of the innovation lifecycle.

3.3 The Fugle^{TM} Innovation Model

The Fugle^{TM} innovation model, developed by *Indutech (Pty) Ltd*, aims to assist organisations to identify, evaluate, develop, implement and exploit new products and services more efficiently and effectively. The model represents the generic innovation process and comprises two distinct parts (Du Preez et al., 2008):

- The innovation front end or funnel part that addresses the convergent processes of idea identification, concept formulation, evaluation and refinement
- The bugle part focussing on the divergent deployment and exploitation of the innovation

The innovation process described in the Fugle^{TM} model functions within the organisation, but all stages are linked to the external environment incorporating the innovation network and open innovation concepts. All stages could be influenced by external factors and may even be outsourced if required. In this model, the following factors guide and support the innovation process:

- Organisational strategies
- Organisational culture and people
- Organisational structure and processes
- Information and knowledge

The Fugle^{TM} model comprises the following stages and activities.

- Opportunity Identification and Idea Generation Stage
  - Collect, Categorise and Present Information
  - Generate and Collect Ideas
  - Capture Ideas
  - Idea Filter
- Concept Definition Stage
  - Develop Concepts
  - Incubate and Refine Concepts
  - Concept Filter
- Concept Feasibility and Refinement Stage
  - Determine Concept Feasibility
  - Develop Models and Prototypes
  - Refine Concepts
  - Funding Gate
- Portfolio Stage
  - Deployment Stage
  - Launch Gate
  - Plan Project
  - Detailed Design and Testing
  - Implementation
- Refinement and Formulisation Stage
  - Operate
  - Refine
  - Formalise
  - Exploitation Gate
- Exploitation Stage
  - Exploit business model

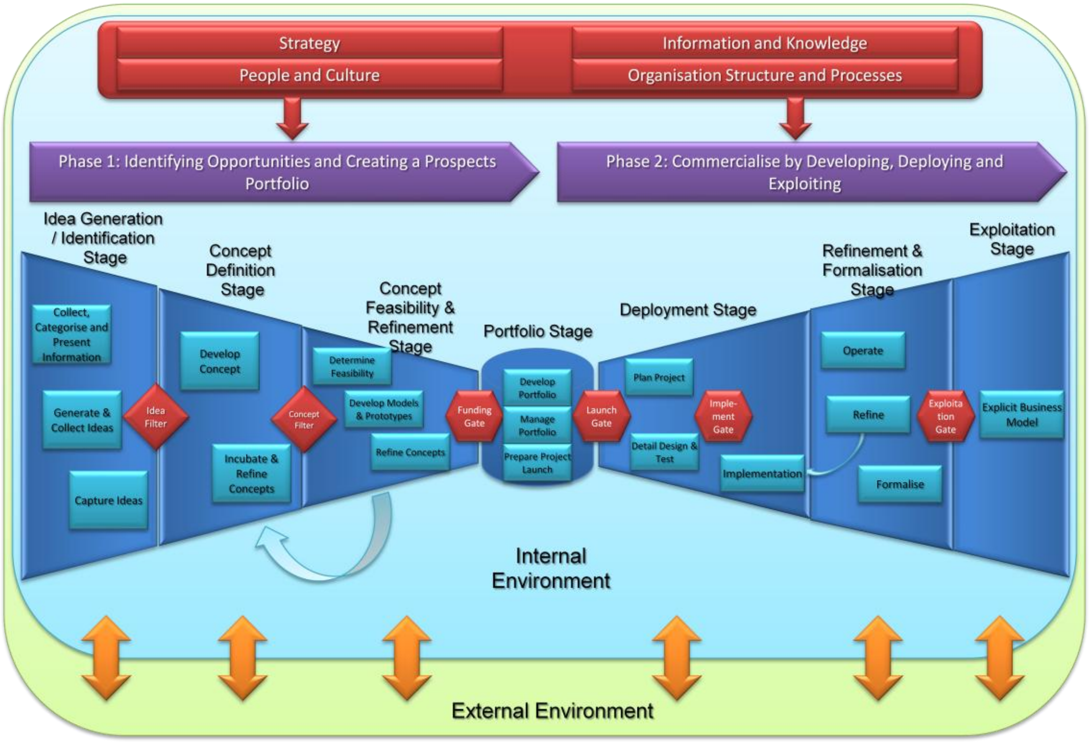


Figure 4: The Fugle^{TM} model for innovation

The model further incorporates gates and filters as decision points between certain activities and stages. Filters are used to sieve the attractive from the less attractive ideas and concepts in the first three stages of the model. It is important to note that even though not all ideas and concept will reach the deployment or even the portfolio stage, such less attractive ideas and concepts still need to be captured and stored for potential future revisit and evaluation. (Du Preez et al., 2008) The first three stages of the Fugle^{TM} model collectively are known as the 'funnel', while the last three stages are known as the 'bugle'.

3.3.1 Opportunity Identification and Idea Generation Stage

This stage embodies the bulk of the creative activities in the innovation process where new ideas are generated and opportunities are identified. New ideas can arise from internal sources (e.g. employees) or from external sources (e.g. customers, suppliers or competitors). Ideas may further be conceived as result of unplanned events (e.g. a sudden insights gained by an individual) or may result from more focused efforts (e.g. brainstorming sessions).

Collect, Categorise and Present Information

Information is an important enabler for the initial stage of the innovation process since it provides the context for innovative ideas and helps to decrease the high levels of uncertainty associated with this stage. Information about the following aspects need to be identified, captured, categorised and presented to act as important stimuli for the identification of new opportunities and the generation of new ideas:

- Current problems and problem areas in the organisation
- Competitors and their activities in the marketplace
- Customers and the relevant markets
- Current technologies, available technologies and future technologies
- Organisational strategies and objectives
- Laws and government policies

Generate and Collect Ideas

Even though ideas may result from unplanned events, ideas may also be generated as result of planned events such as brainstorming sessions or the application of other creativity tools. Providing the right individuals with the right information can act as catalysts to the generation of innovative ideas. A formalised knowledge supply chain may significantly improve the quality and quantity of innovation ideas.

Capture Ideas

All ideas generated should be formally captured and sufficiently detailed by the initiator so that it may be shared with other stakeholders and transformed into a more complete innovation concept (Gaynor, 2002). A repository of historic ideas further need to be constructed since not all ideas will immediately be considered feasible, but may become more relevant at a future point in time. Ideas should be captured in context of the innovation lifecycle, the relevant individuals involved, and the applicable external considerations in order to be easily interpretable in future. Once captured, ideas should be managed through their entire lifecycle.

Idea Filter

Not all ideas can be pursued further due to the restricted resources of the organisation. An intelligent, transparent evaluation process is required to filter out the most lucrative ideas for further development. The strategies and objectives of the organisation are useful guidelines for evaluating ideas since ideas that are obviously not aligned with the organisation’s strategy may be isolated in this manner. A set of evaluation criteria is required to formally evaluate ideas for potential promotion to the Concept Definition Stage. Ideas that are considered incomplete may be redirected to the initiators for rework and possible resubmission. Rejected ideas should be preserved in the idea repository together with the reasons for rejection. It is further crucial to keep initiators and other stakeholders informed about the status of their ideas to ensure future participation.

3.3.2 Concept Definition Stage

The purpose of the Concept Definition Stage is to grow selected ideas into more tangible concepts. Another filtering process is positioned after the concept definition stage to identify those concepts that seems the most promising and that should proceed to be evaluated further to determine its feasibility.

Develop Concepts

A concept may be created from a single idea or may crystallise from the combination of different complementing ideas. Concepts need to be thoroughly documented and sufficiently detailed to promote understanding and contributions from other stakeholders.

Incubate and Refine Concepts

Once concepts are sufficiently developed, time should be taken to share such concepts with stakeholders from different functional areas with the goal to further grow and refine such concepts.

Concept Filter

As for ideas, not all concepts can be pursued further due to resource restrictions. An evaluation filter, analogue to the Idea Filter, is required to select the most lucrative concepts for further evaluation and refinement. A more rigorous set of evaluation criteria is required to formally evaluate concepts for potential promotion to the concept feasibility and refinement stage. Concepts that are considered incomplete may be redirected to the concept development stage for rework and possible resubmission. Rejected concepts should be preserved in the concept repository together with the reasons for rejection. Once more, it is important to keep the stakeholders involved with the development of the concept informed about the status of the concept to ensure future participation.

3.3.3 Concept Feasibility and Refinement Stage

Activities in the Concept Feasibility Stage are focused on collecting more information about the concept in question as well as to create models and prototypes to determine the feasibility of the concept with higher levels of certainty. Throughout this stage concepts may further be refined. A formal funding gate succeeds this stage in order to identify concepts suitable for take up in the organisation’s innovation portfolio. The deliverable of this stage is therefore a list of prospective innovation projects.

Determine Concept Feasibility

To be able to determine the detailed feasibility of a concept, more information needs to be gather around all aspects of such an concept (e.g. technology implications, market perceptions, required partners, distribution methods, etc.). This activity therefore strives to decrease the uncertainty associated with the concept in question.

Develop Models and Prototypes

In order to experiment with a new concept, the concept may be instantiated by means of detailed models (conceptual or tangible) or prototypes that may be used to further test the concept in near to real life scenarios. External stakeholders (e.g. customers and partners) may further experiment with such models and prototypes and valuable inputs may be sourced in a relatively inexpensive manner. These activities may be considered as crucial learning experiences around the future success of the concept in question. By experimenting with the developed concepts in such a practical way, detrimental flaws in the concept may be detected early in the innovation lifecycle allowing the organisation to “fail fast and smart” (Wycoff, 2003), since it is more beneficial and cost effective for the concept to fail at this stage rather than during later stages.

Refine Concepts

Non-detrimental shortcomings identified during the earlier activities of this phase may be addressed in order to increase the robustness and feasibility of the concept in question. When the concept has been sufficiently refined and its feasibility determined to satisfaction, the concept may be submitted to the Funding Gate for potential inclusion in the organisation’s innovation portfolio.

Funding Gate

The Funding Gate is a formal selection process where concepts are considered, selected and prioritised, based on feasibility and potential gain in profit or competitive advantage, for inclusion in the organisation’s innovation portfolio.

3.3.4 Portfolio Stage

This stage builds upon the outputs of the Concept Feasibility and Refinement Stage to prioritise, schedule and align innovation project initiatives in the organisation’s holistic innovation portfolio. The following activities occur in the Portfolio Stage to ensure that the strategic objectives of the organisation are realised:

- Allocation of resources to projects
- Assignment of responsibility
- Continuous monitoring of innovation initiatives
- Understanding the aggregate effect of innovation portfolio projects
- Deciding when to launch a given innovation project

3.3.5 Deployment Stage

The Deployment Stage entails the design, implementation, testing and management of the specific innovation solution and builds upon the conceptualisations, models and specifications developed during the previous stages.

Launch Gate

The Launch Gate controls which projects in the organisation's innovation portfolio should proceed to the Development Stage at what point in time.

Plan Project

This activity addresses the detailed planning of the relevant project in terms of the required project steps, milestones, resources, deliverables and timelines to develop and deploy the innovation in question.

Detailed Design and Testing

During this activity the detailed specifications of the innovation product or process are determined, tested and refined.

Implementation Gate

The Implementation Gate acts as a final design review before implementation and use the results of the project planning and detailed design activities to determine whether the innovation product or process is fit for implementation.

Implementation

This activity involves the actual development of the design and the associated launch of the new innovation product or process.

3.3.6 Refinement and Formalisation Stage

Once implemented, the new innovation product or process is out in the market or in operation. Despite all the measures taken during the earlier stages, chances are that the innovation will initially not function optimally. The Refinement and Formalisation Stage deals with ensuring that the innovation product or process functions are according to specification and expectation and further entails monitoring, measuring, evaluating and refining the innovation solution.

Operate

This activity is about monitoring and measuring the innovation product or process as it functions in the intended real-world scenario. Deviations from specification and expectation are carefully noted for improvement.

Refine

This activity deals with adjusting the innovation product or process in order to address the undesirable characteristics detected in the previous activity.

Formalise

Once the innovation solution is performing to satisfaction the specifications, implementation procedures, and other operational documentation should be formalised.

Exploitation Gate

Before an innovation product or process may enter the Exploitation Stage, the Exploitation Gate evaluates which innovation solutions are candidates for further exploitation and determines how such solutions may be further exploited and at which point in time.

3.3.7 Exploitation Stage

Subsequent to the formalisation of the innovation solution, this final stage further exploits the innovation solution through new business models and markets to generate as much value from it as possible. The Fugle^{TM} model presented in the preceding sections may seem to be a linear, staged process, but many iterative loops and overlaps between the activities described are present. Some of the activities described may further occur in parallel (e.g. idea generation and opportunity identification) in some scenarios. Other activities such as portfolio management and information collection and management may occur throughout the innovation process.

The following section explains what is meant by the Fuzzy Front End of Innovation.

3.4 The Fuzzy Front End of Innovation

The innovation process is traditionally divided into the following three areas (Koen et al., 2002):

- Fuzzy front end
- New product development (NPD) process
- Commercialisation

The fuzzy front end (FFE) is located where the innovation process is initiated. Generally, more thinking than doing is involved here (Graham, 2005). According to Khurana and Rosenthal (1998) the FFE is inherently fuzzy in nature since it is an intersection where complex information processing, a wide spectrum of tacit knowledge, conflicting organisational demands (e.g. cross-functional inputs and resource tradeoffs), substantial uncertainty and high stakes have to converge.

The core objective of work done in the FFE is to create as many original ideas as possible, usually in a limited period of time. These ideas should not only be original, but should further be relevant and actionable to meet the needs of the business (Ishmael & Callahan, 2006). The second objective of the FFE is to identify those ideas having a high probability of success (Pretium, 2005).

3.4.1 New Concept Development Construct

Koen et al. (2002) describes a New Concept Development (NCD) construct shown in Figure 5. The NCD construct was developed to improve understanding of the FFE by describing it in unambiguous terminology. The two arrows pointing to the centre of the figure represents starting points of the innovation process and indicates that the process may either start with the generation of an idea or alternatively, by the identification of an opportunity. The arrow pointing away from the centre represents the point where concepts exit the new concept development process and the new product development (NPD) or technology stage gate (TSG) process starts.

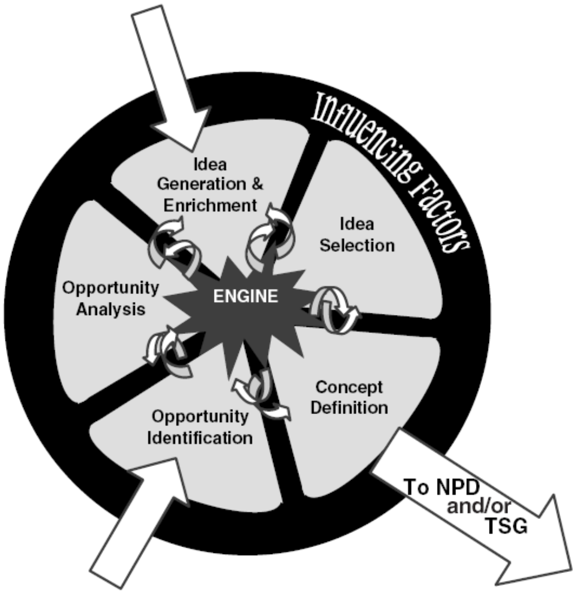


Figure 5: New Concept Development (NCD) construct[2]

The centre of the NCD construct (the "engine") consists of leadership, culture and business strategy of the organisation. It drives the five key controllable FFE elements or activities represented by the areas between the inner spokes in Figure 5, namely:

1. Opportunity identification
2. Opportunity analysis
3. Idea generation and enrichment
4. Idea selection
5. Concept definition

These activities may be performed concurrently and any activity may be iterated numerous times. These activities strongly correspond to the activities of the first two phases of the Fugle^{TM} model as Table 2 illustrates.

[2] Source: Koen et al. (2002)

Activities of the First Two Stages of Fugle^{TM} Innovation Model	FFE Activities according to the NCD Construct
Opportunity Identification and Idea Generation Stage	
Identify Opportunities	Opportunity Identification
Generate and Collect Ideas	
Capture and Manage Ideas	Opportunity Analysis
Idea Filter	
	Idea Generation and Enrichment
Concept Definition Stage	
Develop Concepts	Idea Selection
Incubate and Refine Concepts	
Concept Filter	Concept Definition

Table 2: Comparison between the activities of the first two stages of the Fugle^{TM} Model and the FFE activities according to the NCD construct

The influencing factors as per the NCD construct are as follows:

- Organisational capabilities of the innovating company
- Outside World
  - Distribution Channels
  - Law and Government Policy
  - Customers
  - Competitors
  - Political and Economic Climate
- Enabling Sciences and Technology (internal and external).

These factors are largely uncontrollable by the organisation and affect the full innovation process from opportunity identification through to commercialisation. The organisational capabilities factor determines the extent and manner in which opportunities are identified and analysed, how ideas are chosen and developed, and how concepts are created. These capabilities largely determine the organisation's ability to deal with the other influencing factors. Table 3 compares the influencing factors of the Fugle^{TM} model to that of the NCD construct. It can be seen that the influencing factors of the Fugle^{TM} model and the NCD construct agree to a large extent.

Influencing Factors of the Fugle^{TM} Model	Influencing Factors of the NCD Construct
Internal Environment	*Engine*
Strategy	Leadership
People and Culture	Organisational Culture
Information and Knowledge	Business Strategy
Organisation Structure and Processes	*Organisational Capabilities*
External Environment	*Outside World*
Customer Needs	Distribution Channels
Technological Advancement	Law and Government Policy
Socioeconomic Environment	Customers
Legal Environment	Competitors
Competition	Political and Economic Climate
	Enabling Sciences (internal and external)

Table 3: Influencing factors of the Fugle^{TM} model and the NCD construct

3.4.2 The Importance of the Fuzzy Front End of Innovation

The front end activities of the innovation process, which precede the more formal and structured processes, are increasingly receiving attention to increase the number, value and probability of success of the concepts entering the product development and commercialisation stages of the innovation process. Addressing the lack of in-depth research into the best practices in the FFE represents one of the most promising opportunities to improve the overall innovation process (Koen et al., 2002). In the past, efforts have been focused on increasing the efficiency (in other words, doing things right) of the innovation process with the greater focus on the later stages of the process (Bullinger, 2008). More recently, the focus started to shift toward improving the effectiveness (in other words, doing the right things) of the innovation process. To have the maximum impact, such efforts need to focus on the initial stages of the innovation process in order to better allocate resources and reduce risk, thereby reducing cost and increasing value (Bullinger, 2008). The decisions taken during the early stages of the innovation process, more specifically the first 20% of the process, typically determine up to 80% of costs of the process (Ehrlenspiel, 2007; Bullinger, 2008). Therefore, the early stages of the innovation process, specifically the evaluation and selection of ideas, are of critical importance to an organisation (Bullinger, 2008). The FFE is characterised by inter-departmental collaboration; often between individuals from different backgrounds. The differences in the business language and understanding of such individuals and the resulting potential for misunderstandings characterises the FFE. The fact that some knowledge is of tacit nature, making it difficult to share in a collaborative environment, aggravates this problem (Bullinger, 2008).

This study will predominantly focus on ways to support the innovation processes of a target organisation, which include the fuzzy front end innovation activities, with the information contained in organisational documents.

3.5 Stakeholder Roles in the Innovation Process

Innovation should be treated as a sustainable, repeatable business process that can be measured and managed to maximise commercial success. A collaborative, holistic approach to innovation is required to exploit the insights and skills of different people. Innovation cannot happen without people and therefore it is crucial to understand how people will implement the innovation process in the organisation. In a sense, innovation is an outgrowth of the people and the culture of the organisation. In the end, if the individuals and teams of the organisation do not participate and support innovation, the organisation will fail to be innovative regardless of the quality of ideas and rigour of the organisation’s processes. The key to successful innovation therefore is largely based on the organisation’s people and the roles that they play in the innovation process. (Hering et al., 2005)

When analysing the contribution of people to the innovation process the focus is not on the organisational title of the individuals (e.g. procurement manager), but rather on the role they fulfil in the innovation process. People are important to the innovation initiatives of the organisation partly because when it comes to innovation, metrics, measurements and practices are not very concrete – especially when it comes to the Fuzzy Front End of innovation. It is up to the people, culture and the processes to address these voids and uncertainties. It is important to expose ideas and concepts to a broad audience since it is the different skill sets and viewpoints of people in the organisation that can improve the probability of success of such ideas.

3.5.1 Description of the Roles of People in the Innovation Process

This section presents eight possible roles people can play in the innovation process according to Hering et al. (2005) and further motivates the importance of each of these roles.

The following eight important roles need to be filled for an organisation to be successful in innovation (Hering et al., 2005):

- Connectors
- Librarians
- Framers
- Judges
- Prototypers
- Metric Monitors
- Storytellers
- Scouts

Taylor (2007) defines the following seven innovation roles:

- Change Agent
- Value Creator
- Customer Satisfier
- Leader
- Strategy / Strategist
- Opportunity Generator
- Wealth Creator

Moreover, Kelly and Littman (2006) propose ten innovation roles divided into three areas:

- Learning Personas
    - The Anthropologist
    - The Experimenter
    - The Cross-Pollinator
- Organising Personas
    - The Hurdler
    - The Collaborator
    - The Director
- Building Personas
    - The Experience Architect
    - The Set Designer
    - The Caregiver
    - The Storyteller

Lastly, IBM (2004) discusses 27 innovation roles in terms of eight areas, namely:

- Idea & Insight
    - Explorer
    - Judge
- Research
    - Inventor
    - Advocate
    - Judge
    - Scrounger
- Development
    - Coordinator
    - Builder
    - Advocate
- Management & Strategy
    - Leader
    - Judge
    - Financier
    - Planner
    - Advocate
- Manufacturing & Distribution
    - Interpreter
    - Planner
    - Builder
    - Judge
- Marketing & Sales
    - Promoter
    - Salesperson
    - Analyst
- Customer
    - Buyer
    - Judge
    - Promoter
- Communities of Interest
    - Host
    - Subject Matter Expert
    - Connector

The eight innovation roles according to Hering et al. (2005) will be explained next in order to get a feel for what such roles involves.

Connectors

A Connector is an individual who is well connected to many people within and outside the organisation in question. The Connector builds proverbial bridges between other people and technologies within the organisation and connects the organisation to customers and business partners that may assist the organisation in its innovation endeavours. Having a Connector in the innovation team improves the probability of making valuable connections and drastically reduces the time required to find appropriate connections.

Librarians

The Librarian has the responsibility of collecting ideas and rendering organised access to other individuals who can assist in building a library of innovation-related information and who can further understand and disseminate the current collection of innovation-related information. The Librarian facilitates the process of checking in new ideas, solutions, problems, technologies and requirements and examining and expanding existing ones. The Librarian assists in the process of capturing ideas and providing metadata and tags to facilitate future information retrieval.

Framers

A Framer creates appropriate evaluation frameworks and criteria, in collaboration with various business functions and management, which teams should use to evaluate ideas in a transparent, fair and consistent manner.

Judges

The Judge applies the evaluation frameworks and criteria created by Framers to evaluate ideas. There may be various Judges for a given idea, each corresponding to the relevant business functions or silos involved by the idea in question. It is further the responsibility of the Judge to ensure that all appropriate business functions have contributed to the evaluation of a given idea, establish a verdict about the idea, promote the idea for further development, refer the idea back for further refinement or freeze the idea for future re-evaluation. Lastly, the Judge needs to ensure that the rationale for the relevant decision is documented in such a manner that others can understand it presently and in future.

Prototypers

The Prototyper is responsible for rapidly creating representations of ideas to make ideas more tangible to support the iterative process of testing, evaluating and refining. They create “wire model” versions of the envisioned product or service to provide users and customers with something more concrete to experiment with to be able to provide crucial feedback.

Metric Monitors

The Metric Monitor defines and monitors quantitative and qualitative metrics suitable to measure progress in each phase of the innovation process. These metrics should apply to each individual, team, division and ultimately the entire organisation. The Metric Monitor further need to expand and refine existing metrics based on the actual measured results of the metrics and the maturity of the organisation. Lastly, the Metric Monitor analyses successes and failures in the innovation process and communicate the findings to others to ensure that pitfalls can be avoided and successful practices promoted in future.

Storytellers

It is the responsibility of the storyteller to collect, preserve and recount stories about the organisation as people respond better to stories than other methods of communication. In this manner, the Storyteller assists in creating a culture supportive to innovation, particularly in the area of allowing people to fail. The Storyteller reminds people of what is important and continually reinforces the corporate culture.

Scouts

A Scout explores and analyses new and likely trends that may impact the business in future. Scouts gain insights into the market and future trends by actively meeting with customers, partners, vendors and influential stakeholders in the relevant industry. They synthesise the information gathered and present it to the innovation team for further examination. Scouts help the innovation team to understand what trends are significant and should be included in the innovation process.

3.5.2 Allocating Individuals to Innovation Roles

Many employees of the organisation will be suitable for multiple innovation roles. The goal is not necessarily to find an individual to fulfil each of these roles, but rather to ensure that the members of the innovation team spans each of the roles identified. Moreover, a given person often plays multiple roles - particularly in smaller organisations. According to Kelly et al. (2006) the innovation roles are not inherent personality traits or "types" that are permanently tied to one (and only one) person in the team as innovation roles are available to nearly anyone in the innovation team, and individuals can switch roles in reflection of their multifaceted capabilities.

Hering et al. (2005) provides the following guidelines for filling the various innovation roles in the organisation. Firstly, it is important to identify individuals that have an intellectual and emotional commitment to the innovation opportunities and ideas and who are willing to contribute their time and expertise to the innovation endeavour. The innovation opportunities and ideas should make sense to them, these opportunities and ideas should be something that they are passionate about and that they are able to execute and promote in their respective role. Secondly, individuals from every business function having a stake in the delivery of the envisioned product or service should be included in the innovation team. This is important to have as many viewpoints as possible to be able to uncover opportunities for products and services that might have been overlooked previously, as well as challenges and issues that may not be obvious. Thirdly, including individuals higher up in the management chain will increase the chances of success because of greater buy-in from management. On the other hand, it is possible to use individuals with less managerial influence and still attract the attention of management by showing quick successes in the innovation process. Fourthly, since failure will occur and is viewed as a crucial part of innovation, everybody should be comfortable with taking risks. Finally, keep in mind that tools to support innovation are useless and potentially dangerous without a guiding methodology and knowledgeable people to use them.

Not all of the these roles need to be fulfilled by actual employees of the organisation. For instance, the Scout role may be fulfilled by using trend monitoring organisations. The Judge role may also be outsourced to a consultant or third party not engaged in the day to day decision making processes. The Prototyper role may further be outsourced to other organisations specialising in creating such prototypes and the results integrated into the innovation process. The Connector, Librarian and Framer roles however, are too important and integrated in the innovation process to be filled by external parties.

3.6 Involvement of Innovation Roles in the Fuzzy Front End of Innovation

As different innovation workers (or even a single worker in some cases) assume different roles when participating in the innovation process a non-uniform involvement of the different roles in the different stages of the innovation process can be expected. Table 4 illustrates the likely degree of involvement of different innovation roles in the various stages of the FFE part of the innovation process.

	Opportunity Identification and Idea Generation Stage				Concept Definition Stage		
Innovation Roles	**Collect, Categorise and Present Information**	**Generate and Collect Ideas**	**Capture Ideas**	**Idea Filter**	**Develop Concepts**	**Incubate and Refine Concepts**	**Concept Filter**
Connector	2	1	0	0	2	1	0
Librarian	2	2	2	0	1	1	0
Framer	0	0	0	1	0	0	1
Judge	0	0	0	2	0	0	2
Prototyper	0	0	0	0	2	2	1
Metric Monitor	1	1	1	1	1	1	1
Storyteller	1	1	0	0	1	0	0
Scout	2	2	1	0	2	1	0

KEY: 0 = no involvement, 1 = passive involvement, 2 = active involvement

Table 4: Involvement of innovation roles in the FFE stages of the innovation process

The next section examines some possible sources of innovation ideas as well as the internal and external participants in the early stages of the innovation process.

3.7 Sources of Innovation Ideas

Innovation ideas, ranging from incremental to radical, may come from a variety of sources internal or external from the organisation. Bullinger (2008) presents the following sources of innovation ideas.

Idea Source Type	Idea Source	Idea Source Location	Examples
Research	Joint Projects	Internal & External	
	Literature	External	Books, Academic Journals, Management Journals, etc.
	Lectures	External	Universities Lectures
Analysis of Environment	Trend Reports	Internal & External	
	Research on Patents	External	
	Research on Market	External	
	Research on Technologies	External	
	Research on Competition	External	Benchmarking, Competitor Publications
Human Contact & Communication	Shareholders	Internal	
	Customers	External	Retailers, (End) Consumers
	Partners	Internal & External	Suppliers, Knowledge Brokers, Investors, (External) Consultants, Shareholders
	Universities	External	
	Competitors	External	
	Conferences	External	
	Team Talks	Internal	
	Innovative Culture	Internal	
	Social Activities	Internal & External	
Internal Analysis	Controlling	Internal	Sales Figures, Cost of R&D
	Consumer Complaints	External	
	Quality Reports	Internal & External	
	Information of Sales Representatives	Internal	
	Staff Surveys	Internal	
Spontaneous Ideas	Product Suggestion	Internal & External	
	Process Suggestion	Internal & External	
	Idea for Improvement	Internal & External	
(Systematic) Idea Generation	Workshops	Internal	
	Quality Circles	Internal	
	Training Programs	Internal & External	
	Communities of Practice	Internal & External	
	Continuous Improvement	Internal	

Table 5: Internal and external sources of innovation ideas

In addition, Bullinger (2008) identifies and describes the following participants which are active during the early phases of the innovation process. These participants were identified by means of empirical findings at four German SMEs involved in technology-based incremental product innovation.

Internal Participants (Members of Staff)	External Participants
(Financial) Controller	Competitor
Customer Centre	Component Supplier
Developer	Customer
Department Head (Middle Management)	End Consumer
Division Head (Upper Management)	External Consultant
Internal Consultant	Industrial Designer
Internal Patent Office	Lead User
Lower Management Member	Original Equipment Manufacturer (OEM)
Manufacturing Employee	University
Executive Board Member (Top Management)	
Advisory Board Member	
Product Manager	
Project Leader	
Quality (Assurance) Staff Member	
Sales Representative	

Table 6: Internal and external participants in early stages of innovation process

The next section presents the relationships between the internal and external information realms of the organisation and the fuzzy front end stages of the innovation process.

3.8 Information Realms and their Relationships to the FFE Stages of the Innovation Process

The innovation activities of an organisation may be impacted and supported by a vast range of information within and outside the boundaries of the organisation. This section identifies some examples of such information 'realms' internal and external to the organisation. An information realm can be described as a grouping of information based on the subject (i.e. what the information is about) of the information content.

The following are examples of **internal** information realms may typically impact and support innovation activities:

- Business Model Structure
- Capabilities/Competences/Strengths/Weaknesses
- Customer Information and Feedback
- Equipment and Assets
- Information Systems/IT
- Innovation Objectives
- Innovation Opportunities/Ideas/Concepts
- Lessons Learnt
- Organisational Strategies
- Organisational Structure/Human Resources
- Organisational Procedures
- Projects/Programmes
- Revenue and Cost Structures
- Value Chain/Business Processes

- Value Offering (Products/Services/Methodologies)

The following **external** information realms may typically be important for innovation.

- Competitor Activities
- Customer Needs
- Distribution Channels
- Enabling Sciences and Technology
- Law and Government Policies
- Market Trends
- Partners
- Political and Economic Climate
- Suppliers

If one can establish which information realms are pertinent to which stages of the innovation process, it may be possible to accelerate the task of finding information during the different innovation stages by designing and implementing mechanisms to aid information retrieval by primarily focusing on the information realms and sources relevant to the applicable stage of the innovation process. The following two tables attempt such a mapping on a generic level. Such mappings may need to be fine-tuned for a specific organisation over time based on its innovation experiences to be more accurate. Firstly, the likely relationships between the **internal information realms** and the FFE stages of the innovation process are presented in Table 7.

	Opportunity Identification and Idea Generation Stage				**Concept Definition Stage**		
Internal Innovation-Related Information Realms	**Collect, Categorise and Present Information**	**Generate and Collect Ideas**	**Capture Ideas**	**Idea Filter**	**Develop Concepts**	**Incubate and Refine Concepts**	**Concept Filter**
Business Model	1	1	0	2	1	2	2
Capabilities/Competences/Strengths/Weaknesses	2	2	2	2	2	1	2
Customer Information & Feedback	2	2	0	2	2	2	2
Equipment and Assets	1	1	0	2	2	1	2
Information Systems / IT	1	1	0	2	2	1	2
Innovation Objectives	1	2	1	2	2	2	2
Innovation Opportunities / Ideas / Concepts	2	2	2	2	2	2	2
Lessons Learnt	2	2	1	2	1	2	2
Organisational Strategies	1	1	1	2	1	1	2
Organisational Structure / Human Resources	1	1	1	2	2	2	2
Organisational Procedures	1	1	2	2	1	1	2
Projects / Programmes	1	1	1	2	2	2	2
Revenue and Cost Structures	1	1	0	2	2	2	2
Value Chain / Business Processes	1	1	1	2	2	2	2
Value Offering (Products/Services/Methodologies)	1	1	1	2	2	2	2
KEY: 0 = no relation, 1 = average relation, 2 = strong relation							

Table 7: Internal information realms and their relationships to the different FFE activities

Table 8 represents the likely relationships between the **external information realms** and the FFE stages of the innovation process.

	Opportunity Identification and Idea Generation Stage				Concept Definition Stage		
External Innovation-Related Information Realms	**Collect, Categorise and Present Information**	**Generate and Collect Ideas**	**Capture Ideas**	**Idea Filter**	**Develop Concepts**	**Incubate and Refine Concepts**	**Concept Filter**
Competitor Activities	2	2	1	2	1	2	2
Customer Needs	2	2	2	2	2	2	2
Distribution Channels	2	2	2	1	2	2	2
Enabling Sciences and Technology	2	2	2	2	2	2	2
Law and Government Policies	2	2	1	2	0	1	2
Market Trends	2	2	1	2	1	2	2
Partners	2	2	1	1	1	1	2
Political and Economic Climate	2	2	1	2	0	1	1
Suppliers	1	1	1	1	0	1	2
KEY: 0 = no relation, 1 = average relation, 2 = strong relation							

Table 8: External information realms and their relationships to the different FFE activities

Although the relationships illustrated in the two tables above are of a generic nature, it seems logical that different information realms will indeed have different levels of applicability to different stages of the innovation process.

3.9 Relationships between Innovation Roles and Information Realms

Since innovation cannot happen without people and people participate in the innovation process in terms of different innovation roles, the relationships between the various innovation roles and the different information realms are important in the light of the effectiveness and efficiency of the innovation process. More specifically, if one can establish the information needs of the different innovation roles it would be possible to facilitate the information distribution and retrieval activities of the different roles leading to the increased efficiency and possibly effectiveness of the innovation process.

Firstly, the likely relationships between the **internal information realms** and the different innovation roles are presented in Table 9.

	Innovation Roles							
Internal Innovation-Related Information Realms	Connector	Librarian	Framer	Judge	Prototyper	Metric Monitor	Storyteller	Scout
Business Model	1	1	2	0	0	1	0	0
Capabilities / Competences / Strengths / Weaknesses	1	1	2	1	1	2	1	1
Customer Information & Feedback	1	1	1	0	1	2	0	1
Equipment and Assets	1	1	2	0	1	1	0	1
Information Systems / IT	1	1	2	0	1	1	0	1
Innovation Objectives	2	1	2	1	1	1	2	1
Innovation Opportunities / Ideas / Concepts	2	2	2	2	2	2	2	2
Lessons Learnt	0	1	1	0	1	2	2	0
Organisational Strategies	1	1	1	1	0	1	1	1
Organisational Structure / Human Resources	2	1	1	1	1	2	2	1
Organisational Procedures	1	2	2	2	1	2	0	0
Projects / Programmes	2	2	2	1	1	2	1	2
Revenue and Cost Structures	1	1	2	0	0	1	0	1
Value Chain / Business Processes	1	1	1	0	0	2	0	1
Value Offering (Products / Services / Methodologies)	1	1	2	0	1	2	1	1
KEY: 0 = no involvement, 1 = passive involvement, 2 = active involvement								

Table 9: Internal information realms and their relevancy to different innovation roles

Table 10 shows the likely relationships between the **external information realms** and the different innovation roles.

	Innovation Roles							
External Innovation-Related Information Realms	Connector	Librarian	Framer	Judge	Prototyper	Metric Monitor	Story-teller	Scout
Competitor Activities	1	1	1	0	0	1	0	2
Customer Needs	2	1	1	0	1	1	1	2
Distribution Channels	2	1	1	0	1	1	0	2
Enabling Sciences and Technology	2	1	1	0	2	1	0	2
Law and Government Policies	1	1	1	0	0	0	0	2
Market Trends	1	1	1	0	0	0	1	2
Partners	2	1	1	0	1	1	0	2
Political and Economic Climate	1	1	1	0	0	0	0	2
Suppliers	2	1	1	0	1	1	0	2
KEY: 0 = no involvement, 1 = passive involvement, 2 = active involvement								

Table 10: External information realms and their relevancy to different innovation roles

These two tables illustrate that the respective innovation roles have different information requirements in terms of the internal and external information realms of the organisation pertaining to innovation. In order to better support these innovation roles in finding and sharing innovation-related information, a large subset of information realms need to be considered and a mechanism to facilitate improved access to their associated information need to be implemented.

Thus far this chapter dealt with the different types of innovation, the innovation process and its activities, the stakeholder roles in the innovation process, the sources of innovation ideas and different information realms in terms of innovation-related information. The next section introduces the concept of measuring innovation capability.

3.10 Measuring Innovation Capability

Maturity models arose in the field of software engineering (e.g. the Software Engineering Institute's SW-CMM®) to measure the maturity, capabilities and practices of organisations in developing software code. Organisations having higher levels of maturity are more likely to be successful in effectively and consistently creating high quality software. Fairly recently, the prospect of measuring the capabilities of an organisation relating to the execution of innovation endeavours was researched. More specifically, Essmann (2009) developed an Innovation Capability Assessment mechanism to assess the state of an organisation's capability to innovate sustainably. More specifically, he developed a framework that models and evaluates the innovation capability maturity of a target organisation. This framework consists of three dimensions, namely the Innovation Capability Construct, the Organisational Construct and the Capability Maturity Construct. The Capability Maturity Construct is comprised of five levels of maturity that is used to measure the maturity of the organisation's innovation capability. The capability of the organisation is determined by means of a questionnaire that is completed by people fulfilling different roles in the organisation and which evaluates 42 fundamental requirements of organisational innovation capability. The outcome of the assessment is a list of strengths of the organisation in terms of innovation capability as well as a list of weaknesses. The identified areas of weakness may then be subsequently addressed to improve the organisation's innovation capability. The processes of measuring and improving an organisation's capability to innovate should ideally be repeated at regular intervals to assist the organisation to continuously work towards higher levels of innovation maturity.

3.11 Knowledge Management and Innovation

"*Effective knowledge management not only forms the basis of successful innovation processes, it also greatly enhances an organisation's ability to innovate*." Sammer (2003)

Since innovation is founded on the development and application of new knowledge, strong links exists between knowledge and innovation. According to Sammer (2003) the knowledge activities of an organisation are mostly centred on the continued evolution of the main knowledge domains of the organisation. Usually, various knowledge holders are involved in such activities which may even include external knowledge domains and expertise from collaborative partners. The application and fusion of organisational knowledge yields returns which manifest in the creation of core competencies, core products, and finally, end products. Effective innovation management guides the knowledge development process in order to realise and commercially exploit the resulting knowledge. The impulsion for innovation projects can originate from two distinct sources, namely unplanned innovation and planned innovation. Unplanned innovation happens as part of the day-to-day business activities and is frequently the additive result of a client project. In contrast, planned innovation actively uses both the internal and external knowledge resources of the organisation in executing new value generating projects.

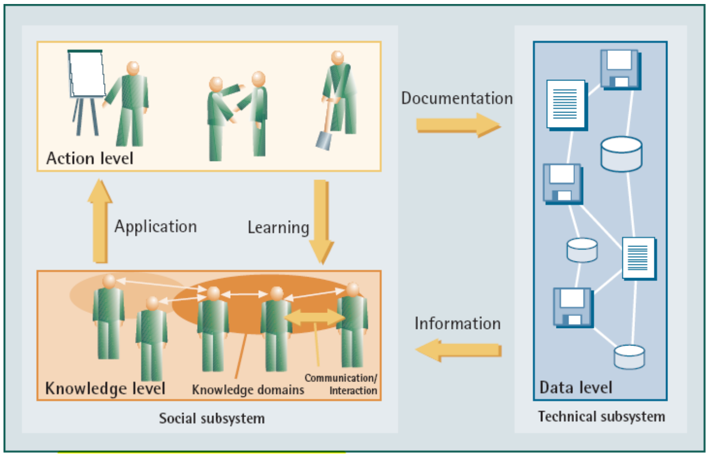


Figure 6: The Basic Model of Knowledge Management[3]

Sammer (2003) states that three main aspects exist with regard to knowledge, namely **individual knowledge**, **action** and **data**. The first aspect, individual knowledge, as the sum of a person's individual capabilities and experience, regulates the possible actions available to a person and therefore influences the contributions he is able to make to a specific project or task. The second aspect, action, covers both physical (e.g. operating a fax machine) and mental actions (e.g. problem solving). The actions necessitated to complete a given task frequently give rise to large

[3] Source: Sammer (2003)

amounts of data. Data, the third aspect, covers both internal data (e.g. data from related projects) and external data (e.g. data from Internet sources or libraries). These three aspects give rise to the three different operational levels of the basic model of Knowledge Management (KM) presented by Sammer (2003) and illustrated in Figure 6.

The three operational levels of the basic model of Knowledge Management are:

- Knowledge level
- Data level
- Action level

The **knowledge level** constitutes the knowledge of individual members of the organisation as well as the interaction with one another. The **data level** is made up of the collection of available documented knowledge (e.g. residing in organisational databases, printed and electronic documentation, etc.). The knowledge and data levels serve as inputs to the **action level**, representing the organisation's value creating processes where business processes are executed. These three operational levels coupled with the five core knowledge processes, namely **information**, **documentation**, **communication**, **application** and **learning**, constitute the basic model of KM.

The framework developed in this project is closely related to the aspects addressed in the abovementioned KM model. The prime focus of this research is to distil and make accessible the essence of innovation-related information stored in the electronic textual documentation of the organisation. This corresponds to a subset of the data level as well as the "information" process arrow linking the data and knowledge levels of the model. The framework will further attempt to facilitate interaction between the stakeholders of the innovation process by making explicit the organisation's innovation-related information entities[4] and the relationships among such entities. This corresponds to the "communication/interaction" process arrow in the knowledge level of the model. Information appropriate to the organisation's innovation-related information entities should therefore be made available to the innovation workers to support more effective and efficient actions on their part. The latter corresponds to the "application" process arrow linking the knowledge and action levels of the model. The framework should also serve as a mechanism to capture the outcomes of the action level and relate it to other innovation-related information entities of the organisation to possibly serve as future information. This corresponds to the documentation process arrow connecting the action and data levels. In summary, the intended framework would facilitate:

- Creating new knowledge from existing data or information
- Interactions between innovation stakeholders
- Putting existing knowledge into action

[4] The meaning of the term 'information entity' will become clear in Chapter 8 of this report.

- Capturing the outcomes of completed actions for possible future reuse

Although such a framework could have wider applicability in the organisation, for this research the focus will be on innovation-related electronic information[5], innovation workers[6] and innovation processes[7].

Availability of applicable knowledge is a key enabler of effective (operational and innovation) projects. Figure 7 illustrates the interactions between knowledge, documentation and projects.

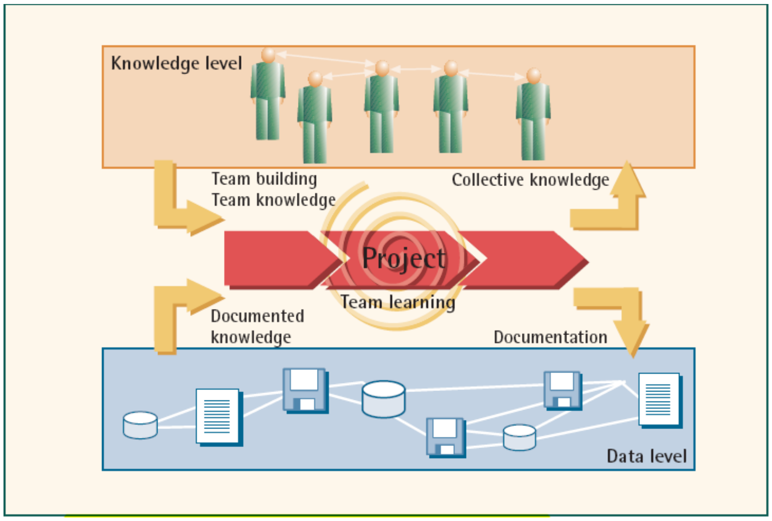


Figure 7: Projects as framework for knowledge creation and application[8]

Knowledge transfer in an organisation can occur via human networks (e.g. meetings or conversations around the coffee dispenser) or via information and communication networks (e.g. accessing stored data using an information system and turning it into knowledge or using video conferencing facilities for virtual meetings). Arguably, the most time consuming form of knowledge transfer, face-to-face meetings offer knowledge seekers the additional benefit of increasing their contextual knowledge as opposed to information- and document-based knowledge transfer where the knowledge seeker must have the relevant prior contextual knowledge to be able to understand the presented information. The developed framework will serve as a mechanism to provide its users with appropriate innovation-related information together with the contextual information (e.g. relationships of the given piece of information with other innovation-related entities) required for correctly interpreting the information in question.

[5] In terms of the data level of the basic model of KM
[6] In terms of the knowledge level of the basic model of KM
[7] In terms of the action level of the basic model of KM
[8] Source: Sammer (2003)

Sammer (2003) further posits that in order for KM to support effective and efficient innovation (management), there should be sufficient interaction between three different levels, namely the action, data and the knowledge levels of the basic model of KM.

The organisational experts, with their tacit knowledge and social skills comprise the knowledge level and they participate in the action level. The data and documents relevant to innovation are gathered at the data level and can be made available throughout the innovation process by means of information and communication tools. Sammer (2003) further stresses that good integration between the knowledge and data levels is an essential factor in successful innovation projects. The entities of the developed framework include items from the data level (e.g. specific documents or specific innovation ideas) as well as items from the knowledge level (e.g. specific innovation workers and knowledge areas) together with the relationships among such items. This could improve the integration required between these two levels.

3.12 Summary

This chapter introduced various aspects of innovation such as the different types of innovation, the evolution of innovation models, the Fugle^{TM} model for innovation, the fuzzy front end (FFE) of innovation and its importance, different roles in the innovation process, different sources of innovation ideas, and various information realms - both internal and external to the organisation - pertaining to innovation-related information. Of particular importance is the opportunity to increase the effectiveness and efficiency of the initial stages of the innovation process, corresponding to the fuzzy front end, which typically determine up to 80% of costs of the overall innovation process. This can be achieved by supporting decision-making, better allocating resources and reducing risk to reduce cost and increase value. The availability and utilisation of timely and accurate information in the early stages of the innovation process can facilitate these activities and therefore can be regarded as an important element in any initiative striving to optimise the innovation process and the FFE part in particular.

The fact that the FFE is characterised by inter-departmental collaboration, and communication between individuals from different backgrounds, a further opportunity exist to clarify and organise innovation-related information to reduce the potential for misunderstandings and increase knowledge sharing and reuse. The examination of the relationships among the activities in the FFE stages of innovation, the different innovation roles of stakeholders in the innovation process and the internal and external information realms of the organisation indicated the dynamic nature of the innovation process in term of the involvement of humans and information. Any mechanism aiming at supporting the innovation process and especially the FFE should therefore also be dynamic in order to be of practical value. To conclude the chapter, the relationship between Knowledge Management (KM) and innovation was discussed and it was concluded that KM constitutes the foundation of effective and efficient innovation and the management thereof.

4. Information and the Innovation Process

“Information and Information Technology will play a critical, lynchpin, role in accelerating radical innovation as knowledge workers provide most of the human infrastructure to innovation enterprises. In the future, rapidly accessible information will be the driving force of innovation, whether incremental, next-generation, or radical.” Miller et al. (2006)

The various generic internal and external information realms of an organisation possibly pertaining to innovation were presented in sections 3.8 and 3.9. Several sources of information may be used throughout the innovation process of an organisation. On a high level such information sources may be classified as sources internal and external to the organisation. Another possible classification may be structured versus unstructured information. This chapter will examine the information sources possibly containing innovation-related information with an emphasis on those containing information of unstructured nature as governed by the focus of this research. A mapping of such sources to the various innovation-related information realms defined in Chapter 3 will further be made.

This chapter addresses the following:

- **Explain the difference between structured, semi-structured and unstructured information**
- **Define what is meant by the term ‘electronic document’**
- **Present the internal, electronic sources of innovation-related information**
- **Map the internal information sources to the internal information realms**
- **Present the external, electronic sources of innovation-related information**
- **Map the external information sources to the external information realms**
- **Introduce the concept of a knowledge worker and its applicability to innovation**

4.1 Structured, Semi-structured and Unstructured Information

The way in which information is stored largely determines the ease of retrieving it. Traditionally, electronic information was stored in database tables in the form of records. Each record comprises a number of (text or numerical) values corresponding to the different fields defined for the relevant database table. Information stored in databases is a prime example of **structured information**[9]. With structured information it is usually clear what is stored where and how everything is related. This predictability makes it relatively easy to find information in a (well-designed) database. Today, databases are still largely employed to store electronic information.

[9] The term ‘information’ may be read as ‘information/data’.

For the last decade or so, the amount of information available electronically in free-form format increased dramatically, especially in the corporate environment, due to the advances in and adoption of digital authoring methods such as word processors, e-mail, web pages, wikis, blogs, to name a few. Information in this format is grouped under the umbrella term, **unstructured information**. With unstructured information anything can be said in any way and in any language. The format and structure is therefore not predictable and as result the unstructured information content are not ordered or is ordered inconsistently (Inmon, O'Neil and Fryman, 2008). This makes finding or processing information captured in unstructured format relatively difficult due to that fact that it has not been fielded as with structured information.

Some information usually regarded as unstructured actually is **semi-structured information**. For example, an e-mail message contains the fields sender, recipients, date, subjects, to name a few. Although these fields can be used to order e-mail messages or to find a specific e-mail message, e-mail messages are still not regarded as structured information due to the fact that arguably the most important part of the e-mail, the message body, contains free-form text that have to be processed using unstructured information processing methods.

According to Inmon et al. (2008), it is challenging to harvest and classify unstructured information. They further point out that a way is required to peer into the contents of unstructured information units (e.g. files) to discover what they contain (in other words, 'understand' the content electronically) in order to group (e.g. classify or cluster) them so that ultimately they can be found by businesspeople when required (e.g. innovation workers). Only when this has been achieved, can unstructured information be truly useful to the organisation. Chapters 5 and 6 will investigate different techniques to process and exploit unstructured information.

4.2 Defining Electronic Documents

The following are two of the definitions that WordNet (2006) provides for the term **document**:

"*Anything serving as a representation of a person's thinking by means of symbolic marks"*, and

"A computer file that contains text (and possibly formatting instructions) using seven-bit ASCII characters"

The focus of this research is on exploiting information contained in textual, electronic documentation to support an organisation's innovation processes. For the scope of this research the term **electronic document** shall refer to:

"A digital, demarcated source of human readable textual information stored in a format suitable for information extraction."

The term "digital" indicates that it should exist in a virtual, electronic format. The term "demarcated" implies that it should have a clearly defined boundary so that it forms a unit. The phrase "source of human readable textual information" implies that it should contain text understandable to humans, while the phrase “stored in a format suitable for information extraction” indicates that the text should be in a form that would allow for computer-based extraction. According to this definition an electronic audio file (e.g. *mp3* file), a picture file (e.g. *jpg* file) do not constitute an electronic document, since it do not contain text. Neither does a *pdf* file that represents text that has not been through an optical character recognition process as such text is not extractable.

4.3 Internal Innovation-related Information Sources

Internal innovation-related information sources may be described as information sources that are owned and managed by the organisation, as opposed to external information sources that typically reside outside the organisation and more importantly, are managed by external parties.

The following is an incomplete account of internal (electronic) information sources that may possibly contain information that could support the innovation processes of an organisation (listed in alphabetical order):

- Blogs
- Databases
- Electronic Files
- E-mails
- Enterprise Information Portals and Wikis
- Instant Messages
- Web Transaction History

These internal innovation-related information sources will next be described in greater detail.

4.3.1 Blogs

The term blog was derived from the term ‘web log’ and is a shared online journal where people can post daily entries (or posts) about their personal experiences and hobbies. Blog postings are usually in chronological order. Blogs are considered as a type of Web 2.0 technology. Other persons may make comments on such blog posts, thus providing a useful public platform for sharing information, having discourse on a certain topic, posing questions, etc. Recently, many organisations embraced the concept of blogs as an easy-to-use information sharing mechanism and implemented blogs within their respective organisations. For this reasons, blogs are mentioned here as an internal innovation-related information source. When used inside the boundaries of an organisation, blog posts are more focused on ideas, questions and issues in the respective organisation. Blogs entries are usually not screened, although the contributor’s name may appear as part of the respective post, providing employees with a large degree of freedom in making their contributions. A blog may be configured to have a number of categories to organise posts. The contributing party decides which category is the most applicable to their post as part of the process of creating the post.

A blog post is a form of unstructured information and satisfies the definition of an electronic document provided earlier.

4.3.2 Databases

Databases contain structured information about certain aspects of the organisation. Data in databases are contained in tables, each table containing a number of records. A database record is comprised of a number data entries corresponding to the fields of the relevant database table. Databases usually contain numbers, but often contain text as well. Databases are the storage mechanisms used by the information systems of the organisation and are therefore in a sense hidden from the typical information system user. The following are some examples of information systems that employ databases within the organisation:

- Client Relationship management systems
- Document management systems
- Financial administration systems
- Human resource management systems
- Idea management systems
- Inventory systems
- Issue tracking systems
- Production scheduling systems
- Project management systems
- Time and attendance systems
- Workflow systems

A database is a form of structured information and does not satisfy the criteria of an electronic document in most cases. Certain text fields contained in databases may however be viewed as electronic documents.

4.3.3 Electronic Files

Electronic files are required for and generated as result of executing day-to-day business activities and therefore are numerous in most organisations. The following is an incomplete list of the possible functional content of electronic files in an organisation (listed in alphabetical order):

- Asset registers
- Budgets
- Case studies
- Contracts
- Feasibility studies
- Invoices
- Meeting agendas
- Minutes of meetings
- Organisational procedures
- Organisational processes
- Performance appraisals
- Product designs
- Product documentation
- Quotations

- Request for proposals
- Research reports

The content of most electronic files is unstructured due to the absence of formal structure. This is in sharp contrast with the information for example contained in a relational database that is described by a table, field and row position as well as having specified data types. Some people consider electronic files as semi-structured due to the fact that there is not a total absence of structure. Electronic files usually have the following explicitly recognisable attributes:

- Filename
- Location
- Date/time modified
- Creator

Moreover, the following implicit attributes may be contained in electronic files:

- Title
- Abstract/summary
- Headings
- Author
- Keywords

The content of electronic files is comprised of words, numbers, punctuation marks, pictures, formatting characters, etc. More specifically, the following are examples of items that may occur in an electronic file containing textual information:

- Dates
- E-mail addresses
- Figures
- Monetary amounts
- Named entities (names of places, peoples, etc.)
- Numbers
- References
- Tables
- URLs

Electronic files may have different file formats. The following is an incomplete list of file formats of files containing textual information:

- Word processor files (e.g. *.rtf, *.doc, *.docx, *.tex, *.odt, etc.)
- Spreadsheets (e.g. *.xls, *.xlsx, *.xsc, etc.)
- Plain text files (e.g. *.txt, *.mw, etc.)
- Portable Document Format documents (*.pdf)
- PostScript files (e.g. *.ps)
- Presentations (e.g. *.ppt, *.pptx, *.pqf, *.sdp, etc.)
- Extensible Markup Language files (*.xml)
- Web pages (e.g. *.htm, *.html, etc.)

The electronic files of an organisation reside at one of the following locations in the organisation:

- The hard drive of an employee’s computer
- A shared drive or folder on the organisation’s network

- A document repository system
- An attachment in the employee's e-mail client

4.3.4 E-mail

In many organisations e-mail is the preferred mechanism for formal, internal communication due to its flexibility, traceability and ease of use. As result, e-mail message repositories are rich sources of organisational information. Although e-mail messages can be saved as separate electronic files, they are usually contained and managed in an e-mail system. The importance of e-mail as a source of organisational information warrants that it is discussed separate from electronic files or databases. The following are examples of explicit attributes that e-mail messages contain:

- Attachments in the form of electronic documents
- Date/time e-mail message was received
- Date/time e-mail message was sent
- E-mail message body text
- Recipients
- Sender
- Subject

E-mails may be stored in different formats:

- A collection of e-mail messages (e.g. *.pst,)
- A single e-mail message (e.g. *.msg, *.sdm, *.html, *.txt, *.eml, etc.)

Mining information from company e-mail messages may be subject to some regulations relating to employee privacy since it is not necessarily information that the employee elected to make public to a larger audience as with blog posts for instance. E-mails are a prime example of unstructured information and satisfy the criteria of electronic documents presented earlier.

4.3.5 Enterprise Information Portals and Wikis

"*Enterprise Information Portals (EIPs) are applications that enable companies to unlock internally and externally stored information, and provide users a single gateway to personalised information needed to make informed business decisions*." (Firestone, 2008). A wiki is a website or set of web pages that allows almost anyone to add and edit content. Although wikis became popular due to their availability on the Internet, many companies employ wikis internally as mechanisms to share information with employees. For this reason wikis are also presented as a possible internal source of innovation-related information. The reader is referred to section 4.4.15 for a further discussion around wikis. Wikis generally have less functionality than EIPs, but allows for employees to make contributions to the information content once they have signed in. Like blogs, wikis are considered as a type of Web 2.0 technology. The following are examples of information that may be available in the EIP or wiki of an organisation:

- Frequently asked questions and answers

- Organisational policies, procedures and best practices
- Information about products and services of the organisation
- Links to useful data sources of the organisation
- Information about employees and departments

The individual entries in a wiki may be regarded as electronic documents according to the definition provided earlier.

4.3.6 Instant Messages

In many organisations instant messaging has become a popular, cost effective, informal way for employees to communicate with other local or remote employees. It derives its usefulness from the ability to have interactive, text-based conversations with others. The history of such messages is stored on the individual’s computer. As with e-mail, mining instant messages may also be restricted by regulations pertaining to the privacy of the individual.

Another (Web 2.0) technology related to instant messages is Twitter, a free social networking and micro-blogging service that enables its users to send and receive updates, known as ‘tweets’, from one another. Tweets are text-based posts of up to 140 characters in size that are displayed on the author's profile page. Tweets are further delivered to the author’s followers, other users who have subscribed to follow the tweets of the given author. By default open access is allowed to all tweets, although authors can restrict delivery to followers in their circle of friends. Users can send and receive tweets using the Twitter website, Short Message Service (SMS) or external applications.

If instant message discussions and tweets are demarcated (e.g. different conversations are stored separately) and stored in a format suitable for information extraction such discussions may be considered as electronic documents.

4.3.7 Web Transaction History

Today the Internet is a crucial source of information to many organisations. Although the information on the Internet is actually an external information source, the information about what information employees have requested and retrieved from the Internet are stored within the organisation. The Internet browsing history are kept in either an Internet usage tracking system or at least in the logs of the web browsers (e.g. Firefox or Internet Explorer) located on the individual employees’ computers.

The following are some of the attributes of an Internet browsing transaction that may be retrieved from web browser logs:

- Address of web page visited
- Date of first visit
- Date of last visit
- Name of the web page
- Number of visits to the specific web page

Mining web browser transaction logs may also be restricted by regulations pertaining to the privacy of the individual as discussed earlier in the sections about e-mail messages and instant messages. Web transaction history alone does not represent truly useful electronic documents since it only contains the addresses of the actual web content that were browsed by employees.

Table 11 summarises the affinity between the internal innovation-related information realms and the internal information sources of the organisation.

	Internal Information Sources						
Internal Innovation-Related Information Realms	**Blogs**	**Databases**	**Electronic Files**	**E-mail**	**EIPs & Wikis**	**Instant Messages**	**Web Transaction History**
Business Model	0	0	2	0	1	0	0
Capabilities / Competences / Strengths / Weaknesses	1	0	1	0	0	1	0
Customer Information & Feedback	1	2	1	2	0	1	0
Equipment and Assets	0	1	1	2	1	2	0
Information Systems / IT	0	1	2	2	2	2	0
Innovation Objectives	1	0	2	0	2	0	0
Innovation Opportunities / Ideas / Concepts	2	0	2	2	1	2	0
Lessons Learnt	1	0	2	1	2	1	0
Organisational Strategies	0	0	2	0	2	0	0
Organisational Structure / Human Resources	0	1	2	1	2	0	0
Organisational Procedures	1	0	2	2	2	0	0
Projects / Programmes	1	2	2	2	2	2	0
Revenue and Cost Structures	0	2	2	1	0	0	0
Value Chain / Business Processes	0	0	2	1	0	0	0
Value Offering (Products / Services / Methodologies)	1	0	2	2	2	1	0
	KEY: 0 = no affinity, 1 = medium affinity, 2 = high affinity						

Table 11: Internal information sources versus internal information realms

Table 11 illustrates that the most important internal sources of innovation-related information are electronic files, e-mail messages, enterprise information portal and wiki entries as it can potentially provide information about a large variety of internal innovation-related realms.

4.4 External Innovation-related Information Sources

External information sources are those sources of information that is not owned and managed by the organisation in question, but that may be used by the employees of the organisation. The following sections will address some of most important external information sources potentially useful in an organisation's innovation endeavours.

The following is an incomplete list of (electronic) external innovation-related information sources (listed in alphabetical order):

- Book Reviews and Summaries
- Competitor Publications
- Conference Proceedings
- Consumer Websites
- Electronic Journals
- Foresight Studies and Services
- Market Research Services
- Electronic Newspapers
- Patent Databases
- Product Reviews
- Research or Publication Databases
- Scholarly Literature
- Technology Websites
- Technology Research
- Wikis
- Idea and Problem Repositories
- General World Wide Web and Social Bookmarking Services

Note that these external information sources are not necessarily mutually exclusive in terms of content.

4.4.1 Book Reviews and Summaries

Books are useful sources of information on a myriad of business-related topics. Obtaining an overview of books available on a certain topic and selecting books to acquire remains a challenge however. Many web-based services offer the tables of contents and review information of specific books (e.g. http://books.google.com and www.amazon.com). Other more specialised services offer book summaries for download (e.g. www.getabstract.com). Books can be ordered in hard copy, although the availability of electronic books (eBooks) is increasing. Book reviews and book summaries are typically unstructured information and when these are available as separate units (e.g. files) suitable for information extraction they may be regarded as electronic documents as defined at the start of this chapter.

4.4.2 Competitor Publications

Due to the increasing Internet presence of organisations a substantial amount of information is available about competitor products, services and technologies. Such information may be sourced and filtered for useful information in order for the target organisation to improve and maintain its competitive advantage. Competitor publications are typically unstructured information and when such publications are available as separate units (e.g. files) they may be regarded as electronic documents.

4.4.3 Conference Proceedings

Conferences organised by various disciplines address a wide range of topics and new techniques, new applications of existing techniques, new trends and findings are normally published and presented at conferences. The proceedings of conferences are valuable sources of information and may be acquired in electronic format at the start of or in some cases even after the conference event (e.g. the proceeding of various PICMET conferences may be ordered at https://www.picmet.org/new/db/pub_order.aspx). The various papers that form part of electronic conference proceedings can be regarded as electronic documents and are further a form of unstructured information.

4.4.4 Consumer Websites

Consumer websites can broadly be described as Internet-based mechanisms for consumers to voice their opinions, compare service quality and product features, read reviews from other consumers, and get detailed product information about different products and models. Some consumer websites only focus on customer service feedback from its users and give companies the opportunity to respond to consumer grievances or complements (e.g. www.hellopeter.com). Consumer websites may be a useful source of problems experienced by consumers that may be addressed in future products and services to gain competitive advantage. The individual pages of consumer websites may be regarded as electronic documents and is also a form of unstructured information.

4.4.5 Electronic Journals

Various electronic journals or magazines are available on the Internet (e.g. www.sciencemag.org) and are normally electronic versions of the actual printed magazine or journal, while others are only available online. Most electronic journals require users to subscribe to gain access to the content of its current and past issues. Another form of electronic journal service is electronic journal aggregators (e.g. http://ejournals.ebsco.com, www.jstor.org and www.ieee.org) which provide combined access to electronic journals from various publishers. Electronic journals may be a valuable source of innovation-related information. The individual articles in electronic journals are electronic documents and also unstructured information.

4.4.6 Foresight Studies and Services

Foresight studies and services (e.g. www.islandone.org/Foresight and www.foresight.gov.uk) aim to describe objective, likely future scenarios, developments and trends and are usually focused on specific themes (e.g. future medicine and treatment, future technology, future manufacturing, etc.). Future studies can be used to gain a view of future market needs, technology, and challenges which may aid the process of planning a longer-term innovation portfolio.

The findings and recommendations of such foresight studies and services are typically unstructured in nature. When these are available as separate units (e.g. files) suitable for information extraction they may be regarded as electronic documents as defined at the start of this chapter.

4.4.7 Market Research Services

Market research services (e.g. www.marketresearchworld.net) provide information about market trends, market research surveys and findings, and market opportunities for a wide variety of industries. Information obtained from market research services can assist organisations to more effectively position their innovation endeavours by better understanding the market they operate in as well as significant opportunities in such markets.

The findings and recommendations of such market research services are mostly unstructured in nature and may be regarded as electronic documents when available as separate units (e.g. files) suitable for information extraction.

4.4.8 Electronic Newspapers

Online newspapers publish material containing news, information, and advertising on the Internet and may be the electronic version of an actual printed newspaper (e.g. www.nytimes.com and www.dieburger.com). Most content can be accessed without having to subscribe although some content may only be accessible to subscribed readers. Online newspapers generally feature articles on political events, crime, business, art, entertainment, society and sports. Newsletter articles may contain some useful information with regard to the organisation’s innovation requirements (e.g. news about planned changes in legislation or the assessment of the political environment of a country). The textual content of newspapers is unstructured in nature and the individual electronic newsletter articles may represent electronic documents.

4.4.9 Patent Databases

Some websites provide the capability to search online patent databases for a wide variety of application areas (e.g. computers, software, chemical formulas, telecommunication, health, etc.). Examples of such websites include www.google.com/patents, www.freepatentsonline.com and www.micropat.com/static/index.htm. Patents can provide the organisation with an overview of competing solutions as well as useful components that may be sourced to integrate into their innovation projects speeding up the innovation lifecycle. Although such databases contain structured information the documents that describe individual patents are unstructured in nature and may be regarded as electronic documents.

4.4.10 Product Reviews

Websites providing product review services (www.consumersearch.com) may for example offer the following services to its users:

- Reviewed and ranked product reviews
- Identification of experts in a certain product field and their opinions about certain products
- Listing of the top-rated products, according to experts
- Prices and links to retailers that offer the recommended products

Product reviews can be used to gain an understanding of the shortcomings of an organisation's or a competitor's products as well as to identify opportunities for new products. Individual product reviews can be regarded as electronic documents and is unstructured in nature.

4.4.11 Research or Publication Databases

Research or publication database services (e.g. www.ebscohost.com) are online reference systems which provide access to a variety of proprietary full-text databases and popular databases from leading information providers. Publication databases represent early stage research by entities with a publishing culture (Tibbetts, 1997). For example, several universities provide online access to student dissertations (e.g. https://etd.sun.ac.za/jspui). New technological terms have the tendency to first appear in technical publication or research databases some years before actual patents and tens of years before it appears in business periodicals (Courseault, 2004). As a further example, Inspec (www.theiet.org/publishing/inspec) covers traditional and cutting-edge publications in the fields of physics, electrical and electronic engineering, communications, computer science, control engineering, information technology, manufacturing and mechanical engineering.

Research and publication databases include general reference collections and subject-specific databases for consumption by academics, medical practitioners, corporate users and scholars. Novel, applicable information may be obtained from such research databases that may aid the execution of an innovation project. The individual articles in research and publication databases are electronic documents and also unstructured in nature.

4.4.12 Scholarly Literature

This external information source is closely related to the previous one as with both sources the content have a strong research focus. Some websites (e.g. http://scholar.google.co.za, http://thomsonreuters.com/products_services/science/science_products/scholarly_research_analysis/research_discovery/web_of_science and www.scholarlyexchange.org) provide access to scholarly literature across variety of disciplines and sources, including theses, books, abstracts and articles. Such websites may be used to find scholarly research papers pertaining to a specific issue of an innovation concept. As for research and publication databases, the individual articles are electronic documents and unstructured in nature.

4.4.13 Technology Websites

Several websites and blogs provide news snippets on new technological breakthroughs (e.g. online (e.g. www.technologyreview.com and http://technology.newscientist.com) and hi-tech products (e.g. www.intechout.com/category/ngadget). Information sourced from such websites may aid an organisation’s innovation endeavours by broadening the understanding of new technologies that may form part of future innovations or open up new innovation opportunities.

If the individual technology snippets and reviews are available as separate units in a format that supports information extraction, these can be regarded as electronic documents and are also typically of an unstructured nature.

4.4.14 Technology Research

Some websites specialise in providing access to high-quality technology research and advice, compiled by leading analysts in the relevant field, to its subscribers (e.g. www.forrester.com). Technology research and advice can be used to benchmark innovation concepts as well as to align the innovation endeavours of an organisation with global technology best practices.

The findings and recommendations of such technology research services are mostly unstructured in nature and may be regarded as electronic documents when available as separate units (e.g. files) suitable for information extraction.

4.4.15 Wikis

Although already discussed in section 4.3.5 as a possible source of internal innovation-related information, wikis are also a source of external innovation-related information. To provide a different description as before, a wiki is a form of online encyclopaedia that is in most cases co-authored by a number of anonymous individuals wanting to share their knowledge on a given subject. An example is the well-known website www.wikipedia.org. In general wikis serve as good starting points to find information about a specific topic and usually contains references to other sources providing more in-depth information on the subject in question. The accuracy of information contained in wikis has been questioned since in many cases no explicit quality control is enforced in terms of its content.

The individual wiki entries are unstructured information and may be regarded as electronic documents.

4.4.16 Idea and Problem Repositories

The number of online, shared idea repositories available today is increasing steadily. Many such repositories are embodied in the form of a blog site where the author shares ideas, which he would like to see implemented but cannot accomplish himself, with the rest of the world without any cost or subscription (for example: http://ideas.menzieschen.com/). A slightly different, but closely related example is InnoCentive, an open innovation company (www.innocentive.com). InnoCentive gathers research and development problems in a wide range of domains, including engineering, computer science, math, chemistry, life sciences, physical sciences and business, and present them as "challenges". These challenges are open for solving to anyone who joined their “solver community”. InnoCentive rewards solvers who presented the best solutions according to the criteria of the challenge, with cash awards.

Online repositories gathering and presenting ideas and challenges can be used by innovation workers of an organisation to obtain new innovation ideas, to outsource some aspects of an innovation challenge to others or to establish if a given innovation idea or challenge was already addressed by someone else.

The individual idea or challenge repository entries are unstructured information and may be considered electronic documents.

4.4.17 General World Wide Web and Social Bookmarking

Another popular way to find information is of course searching the World Wide Web (WWW) using Internet search engines like Google (www.google.com), Yahoo (www.yahoo.com) and Clusty (www.clusty.com). The indexes built and maintained by Internet search engines keeps track of all information on the WWW not protected by access control systems and can therefore be used to find information on a myriad of topics. The shortcomings of Internet searches is the lack of quality control in terms of what is indexed and therefore can be retrieved, the lack of context of information in most cases as well as the absence of timestamps of information making it difficult to make relevancy judgements. Well-structured queries are required to find high-quality information due to the sheer amount of information available on the WWW.

Social bookmarking services, also known as collaborative tagging, social classification, social indexing, social tagging and folksonomy, are another form of Internet information retrieval technique that is rapidly growing in popularity. Social bookmarking are online services that aggregate the keywords (or tags) and quality ratings assigned to online resources (e.g. websites, books, articles, etc.) by numerous individuals subscribing to such social bookmarking systems, in a central repository. Social bookmarking services improve the way in which people discover, remember and share information on the Internet by enabling users to specify custom tags to describe web pages of interest in order to find such web pages once more at a later point in time. Users may also access the tags and bookmarks assigned by other users, therefore harnessing the collective opinions or ‘wisdom of the crowd’ in finding information published on the Internet.

Instead of having different bookmarks on different computers, social bookmarking sites allow users to have a single set of bookmarks online. Users can further send interesting bookmarks to other users of the social bookmarking system and identify and track the bookmarks of other users you find interesting. Lastly, users can discover the most pertinent bookmarks (in the opinion of the crowd) on specific topics by looking at popular bookmarks for a given tag. Social bookmarking systems may be used as an alternative to searching for information pertinent to a variety of innovation-related concepts published on the Internet. Examples of social bookmarking services include *Delicious* (previously known as del.icio.us; http://delicious.com), *Digg* (www.digg.com), and *Citeulike* (www.citulike.org). The collection of all user-specified tags and associated bookmarks is known as a folksonomy, short for 'folks taxonomy'. Critique against social bookmarking entails concerns for the quality of assigned tags (e.g. misspellings) as well as the ambiguity of assigned tags (e.g. a single tag may have numerous meanings).

In spite of the shortcomings mentioned, the WWW is still a good starting point to finding information about a wide range of topics pertaining to the innovation endeavours of an organisation. An organisation can use the Internet to track competitor activities, monitor customer perceptions, monitor government information, search for new ideas, collect customer feedback and gather international expertise by using the Internet (Teo, 2000).

The individual pages of websites available on the WWW may be regarded as electronic documents and are mostly unstructured in nature.

Table 12 presents the affinity between the **external innovation-related information realms** and the **external information sources** discussed above.

External Innovation-Related Information Sources	External Information Realms								
	Competitor Activities	Customer Needs	Distribution Channels	Enabling Sciences and Technology	Law and Government Policies	Market Trends	Partners	Political and Economic Climate	Suppliers
Book Reviews and Summaries	1	1	0	2	0	1	0	0	0
Competitor Publications	2	1	1	1	0	1	1	0	1
Conference Proceedings	1	1	1	2	1	2	1	1	1
Consumer Websites	2	2	0	0	0	1	1	0	1
Electronic Journals	1	2	1	2	1	2	2	1	2
Foresight Studies and Services	0	2	2	2	1	2	0	1	0
Market Research Services	2	2	1	1	2	2	1	2	1
Electronic Newspapers	1	0	0	1	2	1	1	2	1
Patent Databases	2	0	0	2	0	0	2	0	2
Product Reviews	2	2	0	2	0	1	1	0	2
Research or Publication Databases	0	0	1	2	1	2	1	2	1
Scholarly Literature	0	0	1	2	1	2	1	1	1
Technology Websites	1	2	0	2	0	1	2	0	2
Technology Research	2	2	1	2	0	2	2	0	2
Wikis	0	0	0	2	0	1	0	0	0
Idea and Problem Repositories	1	2	1	2	0	1	1	0	1
Worldwide Web and Social Bookmarking Services	2	1	1	2	1	1	2	1	2
KEY: 0 = no affinity, 1 = medium affinity, 2 = high affinity									

Table 12: External information sources versus external information realms

From Table 12 it can be seen that comparatively few electronic external information sources covers the information realms of distribution channels, law and government policies as well as political and economic climate. More importantly, no external innovation-related information source seems to stand out when it comes to finding information about a range of innovation realms external to the organisation. As results the organisation should select external information sources that satisfy their unique innovation-related information needs and source information from these sources on a regular basis.

The examples of external innovation-related information sources differ in the following dimensions:

- Who can access the associated information (e.g. unrestricted access or restricted access)
- Who can contribute to the associated information (e.g. read only, commenting allowed, editing allowed, deleting allowed, etc.)
- Size of associated information (e.g. sentences, paragraphs, articles, books, etc.)
- Quality of and extent of review involved with associated information (e.g. no review, informal and ad hoc review, governed by formal review process, etc.)
- Half-life of associated information in terms of future availability (e.g. available only for a brief period, available for years, etc.)
- Range of disciplines or subjects that they cover (e.g. narrow, general, etc.)

4.5 Knowledge Workers and Information

The term ‘knowledge worker’ was coined by Peter Drucker in the early 1960s. The most workers today participate in the information economy and produce abstract work outputs, primarily consisting of information, as opposed to tangible products such as bicycles and pens. Such workers are called knowledge workers. A knowledge worker is a person whose principal role is gathering and packaging information for the use by others. The knowledge worker’s repertoire of know-how does not necessarily include knowledge of all things, but more so knowing where to obtain information about a range of diverse topics. In the words of Samuel Johnson: “*We know a subject ourselves, or we know where we can find information upon it*”. Understanding widespread and heterogeneous pieces of data and information is part of the daily tasks of the knowledge worker. Feldman (2004) states that a typical knowledge worker spends anything between 15 and 35% of his time on searching for information and only 50% or less of such searches are successful. Miller et al. (2006) posit that knowledge workers will become highly valued as a large portion of the world’s service and manufacturing tasks become automated. Feldman (2004) identifies three costs that can be associated with searching for information:

1. Cost of employee time wasted on searching
2. Cost of (in many cases unknowingly) duplicating or reworking information
3. Opportunity cost of missed business opportunities

These costs can be contained by implementing mechanisms to improve the ease of finding information in the organisation. Many activities that the innovation workers engage in can be considered knowledge work. Thus, supporting the innovation workers in finding information can increase the efficiency of the innovation process and decrease the associated costs.

4.6 Summary

This chapter presented a large selection of (mostly unstructured) electronic information sources available to satisfy an organisation’s innovation-related information needs. Although the organisation’s internal information sources may contain a wide selection of innovation-related information, an even wider selection of innovation-related information resides outside of the organisation. An organisation should therefore include both internal and external information sources in their efforts to gather, organise and exploit innovation-related information.

Electronic files and e-mail messages seem to be the most important sources of internal innovation-related information, whereas no particular source of external innovation-related information could be identified on a generic level. To put to use information contained in electronic documents to further an organisation’s innovation processes, practical ways are required to understand and organise the content of such documents to support innovation workers in their attempts to find information. Chapters 5 and 6 will investigate and discuss techniques that can be used to exploit electronic, textual information in a semi-automatic fashion.

5. An Overview of Text Analytical Approaches and Techniques

The explosion in the amount of published text implies that it is impossible for even the most enthusiastic reader to keep abreast with all the reading in a given field, even disregarding adjacent fields. This phenomenon may cause new insights, ideas and knowledge to remain undiscovered in literature. Text mining may alleviate this problem by supplementing or replacing the human reader with automatic systems that can handle the effects of the text explosion. (Redfearn, 2006)

This chapter addresses the following:

- **Introduce natural language text as a medium for capturing information**
- **Discuss some popular approaches to analyse electronic document collections**
- **Present an overview of Natural Language Processing**
- **Discuss Text Mining and its related approaches**
- **Provide an overview of the Text Mining process**
- **Discuss the difference between classification and clustering**
- **Present a comparison between different text analytical techniques**
- **Present some applications of text analytical approaches**
- **Formulate the desirable criteria of the techniques to be applied in this research**

This chapter presents an overview of different approaches and techniques used to extract valuable information from electronic text. It is included as a precursor to the eventual framework and working methods aiming to support the innovation processes of a target organisation with the information contained in electronic documents. The objective of this chapter is not to present a technically or computationally exhaustive insight into the topic, but rather to achieve a shared understanding of popular text analytical approaches and techniques by providing a comprehensive overview.

The terms “Human Language Technology” (HLT), “Natural Language Processing” (NLP), “Text Mining” and “Text Analytics” are difficult to distinguish since they are closely related but do not have exact, unambiguous, widely accepted definitions. The discipline of Human Language Technology (HLT) is comprised of a number of focus areas such as Natural Language Processing (NLP), Speech Recognition, Machine Translation, Text Generation and Text Mining (Kao & Poteet, 2005).

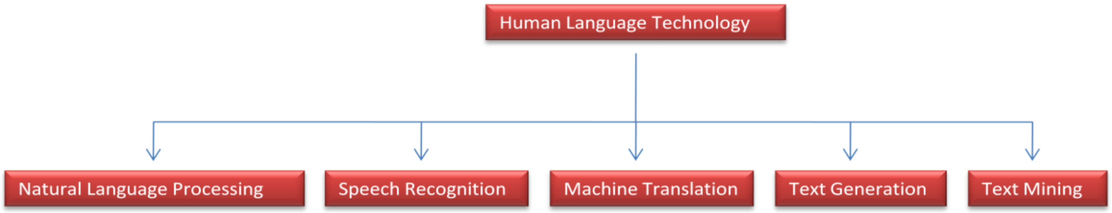


Figure 8: Overview of Human Language Technology

An overview of different text analytical approaches and associated fields such as Natural Language Processing and Text Mining, as well as other more detailed classification and clustering techniques, are presented in the sections that follow.

5.1 Natural Language Text as Unstructured Information

Although the difference between structured and unstructured information was already discussed in section 4.1, the distinction is reiterated here to some extent due to its importance for this project.

Unstructured information in the form of natural language text is abundant in various kinds of organisations (Cheung, Lee & Wang, 2005) and expands daily. To increase information sharing, organisational learning, decision-making and productivity, large amounts of unstructured text need to be read on a daily basis. Electronic (natural language) text is a convenient and common way to capture and store information about any topic thinkable. Natural language text is mostly classified as unstructured information due to the absence of explicit structure, found in relational databases and XML, for example. The effort associated with reading and understanding large collections of unstructured information – or natural language text – remains a challenge in spite of many technological advances in the field of information and communication technology (Nasukawa & Nagano, 2001). New tools are required for automatically organising, searching, indexing, and browsing the ever-growing, large collections of electronic documents (Blei & Lafferty, 2006).

Unstructured information is found in physical objects such as books, reports, academic dissertations, magazines, and newspapers, to name but a few. Unstructured information is also found in virtual objects such as text messages on cellular phones, web pages (including wikis, blogs, etc.), word processor and other computer files, e-mails, e-books, instant messaging messages, databases (e.g. issue tracking systems, customer relationship management systems, etc.) and many more. The time available for an individual to collect, read, interpret and act upon appropriate natural language text is limited both in corporate and research environments. Nasukawa et al. (2001) distinguishes between three (broad) types of technology that have been developed to address the problem of working with large collections of textual data typically found in electronic documents:

- Information retrieval technologies,
- Clustering/classification technologies, and
- Natural Language Processing (NLP), data mining, and visualisation technologies

Information retrieval technologies facilitate the process of searching for documents based on user supplied queries and return ranked sets of matching documents as output. Clustering/classification technologies, on the other hand, are mostly used to organise the individual documents of a document collection (corpus), by grouping such documents into representative categories based on the content of such documents. Some clustering/classification technologies further provide a characterisation of the concepts contained in the document corpus. This is achieved by supplementing each of the calculated clusters with a description, consisting of the key terms typifying the relevant cluster. Such descriptions provide a useful overview of the document corpus in terms of the topics addressed as well as the relative coverage of the topics in the corpus. In general, natural language processing (NLP) and data mining technologies have the aim to discover knowledge by extracting, uncovering and synthesising interesting information from document collections.

It is extremely difficult to provide a complete characterisation of well-formed linguistic utterances, since humans bend the rules to satisfy their communicative and creativity needs (Manning & Schütze, 1999). This complicates the task of automating understanding and interpreting linguistic, textual information. The innovation process, especially the FFE part, has a strong dependency on information and knowledge of which the information contained in electronic documents in the form of natural language text forms a substantial subset. Supporting the innovation process by providing improved awareness and access to the information in electronic documents require the automated understanding and interpreting of textual information. This chapter will therefore examine a range of approaches and techniques that may address this challenge.

5.2 Popular Approaches for Analysing Document Collections

Once a set of textual, electronic documents has been purposefully gathered, an individual may need to analyse such a document collection with a specific objective in mind (e.g. investigate the subjects addressed at a given conference and identify papers having potential relevance to a given innovation idea). The two most fundamental needs for analysing a document collection are to retrieve specific information and to gain an understanding of the contents of the document collection. Several traditional analysis approaches may be used to achieve the objective in question. One possible approach is to study the abstracts (where available) and section headings of each document in the collection to assess its potential usefulness given the objective. Subsequently, the individual may manually organise documents into representative categories according to the topics they address and their potential usefulness. Once all documents have been processed in this fashion, the analyst then decides which documents are the most relevant and proceeds to evaluate these documents in detail. The entire process relies on human intervention and judgement and therefore makes it time consuming, subjective but accurate.

Another approach is to use a summarisation tool (e.g. *Copernic Summarizer*[10]*)* or the summarisation functionality of a word processing tool to reduce the magnitude of the task. A concise summary of each document is generated, and subsequently scanned by the analyst, to

[10] http://www.copernic.com/en/products/summarizer/

assess the potential usefulness of each document. Finally, the analyst could again group the documents into categories. Although the second approach is less labour intensive, every document still has to be summarised and evaluated individually. Based on the author's experience, the affordable text summary tools available today are still largely ineffective in extracting the gist of the natural language text contained in lengthy documents.

Yet another popular approach used to analyse a large set of text documents, is constructing an index of the content of the document collection using software such as *dtSearch*[11], *Lucene*[12], or *Google Desktop*[13] . Once indexed, natural language queries similar to that of mainstream web search engines can be used to identify documents matching a specific query. The result set is usually ranked by the software according to the extent in which the individual documents match the user-supplied search query. The analyst can then inspect the different documents forming part of the result set. Occurrences of the supplied query terms are usually highlighted in the document text. The user can repeat this process by entering different queries and once more assessing the relevant documents returned in the result set. This approach generally involves less initial manual effort. It still relies heavily on the appropriateness and quality of the supplied queries. This is especially the case when the analyst is uncertain about what precisely to search for. This limitation is even more severe when the content of the document collection is largely new territory for the searcher (Nasukawa et al., 2001) as the case might often be when doing innovation-related research. The finding-a-needle-in-a-haystack approach of knowledge retrieval is not always optimal for answering all type of questions (Lieberman, 2007).

The popular or traditional approaches for analysing a collection of electronic documents discussed above have the following limitations:

- No overview of the concepts addressed in the document collection or in the individual documents is available. This makes it difficult to decide which documents to evaluate as well as where to start with the detailed evaluation process.
- The interdependence of the documents constituting the collection is usually only known after all documents have been analysed by the analyst, making it difficult to determine a 'reading path' through such a document collection.
- Documents need to be organised manually using, in most cases, nothing more than human judgement making it a labour intensive and time consuming process.

The following section presents the field of Natural Language Processing, a group of text analytical approaches, which deals with the automated analyses of language based information.

[11] http://www.dtsearch.com/
[12] http://lucene.apache.org/java/docs/
[13] http://desktop.google.com/features.html

5.3 Natural Language Processing

Natural Language Processing (NLP) now exists for several decades (Kao et al., 2005) and was previously known as 'Computational Linguistics'. The root of NLP research can be traced back to the late 1940s with the advent of computer-based Machine Translation. Apart from being a sub-discipline of HLT, NLP is also considered as a sub-discipline of Artificial Intelligence (AI). In fact, NLP is considered as one of the oldest and most challenging problems in the field of AI. NLP entails the analysis of human language with the ultimate goal being to enable computers to understand natural languages as humans do (Redfearn, 2006). In the last ten years of the previous century, the field of NLP started growing rapidly. This growth may be explained by the increased accessibility of large electronic text corpora, increased capabilities and availability of computers, and the coming of the Internet (Liddy, 2000). Although several definitions for NLP exist, a good definition is provided by (Liddy, 2003):

"*Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications*."

It is important to note that the texts referred to in the definition above may typically be oral or written as long as it is in a language used by humans to communicate with one another (Liddy, 2003).

NLP is a powerful enabling technology for a wide range of applications (Liddy, 2002), such as:

- Document Retrieval
- Question-Answering
- Information Extraction
- Text Mining
- Automatic Metadata Generation
- Cross-Language Retrieval
- Document Summarisation

NLP is not a text analytical technique per se but rather a family of techniques which provides the required mechanisms to enable Text Mining and the other applications mentioned above to automatically extract concentrated information or knowledge from text. A substantial amount of valuable information for mining exists in the form of natural language texts or in forms that can readily be converted to text (Liddy, 2000).

Several NLP techniques are inspired by linguistics. These techniques parse the syntax of the extracted text using a formal grammar and a lexicon. Subsequently, the results are then interpreted semantically and in some case meaning is inferred. Deep NLP entails parsing every element of the analysed text in an attempt to take into account the contribution to meaning of every element. On the other hand, shallow NLP identifies only the main grammatical elements (Redfearn, 2006) of the text and only performs limited semantic analysis (Koa et al., 2005).

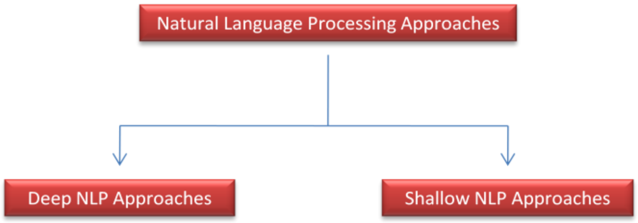


Figure 9: Different types of NLP approaches in terms of depth of semantic analysis

In some cases statistical techniques are used to disambiguate words with multiple meanings or the results of multiple parses of a specific sentence. The focus of NLP is normally single documents or text parts as opposed to collections of text parts or documents, since it is fairly computationally expensive. The following techniques are frequently used as part of an NLP system:

- **Stemming** – removing suffixes from words in an attempt to reduce the number of word tokens to consider
- **Lemmatisation** – replacing inflected words with their base forms
- **Multiword Phrase Grouping**
- **Synonym Normalisation**
- **Part-of-Speech Tagging** – Marking each word with its appropriate part-of-speech function (e.g. verb, noun, adjective, etc.)
- **Word-sense Disambiguation** – Determining the appropriate meaning of a word, having multiple meanings, in a sentence given its usage (Redfearn, 2006)
- **Anaphora Resolution** – Determining to which actual entity words like “they”, “the president”, etc. refers to
- **Role Determination** – Determining the subject and object of a sentence (Kao et al., 2005)

Over the years important contributions have been made to the NLP discipline and practice by the disciplines of Linguistics, Computer Science and Cognitive Psychology (Liddy, 2001) as illustrated in Figure 10.

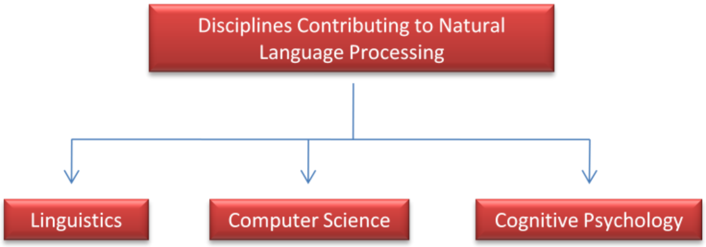


Figure 10: Disciplines contributing to NLP

The contribution made by Linguistics centres around the formal, structural models of language and finding of linguistic universals. Computer Science contributed to NLP by developing structures to represent data internally as well as the efficient processing of such structures. Cognitive Psychology views language usage as a window into the processes of human cognition and attempts to model the use of language in a psychologically credible manner (Liddy, 2001).

The field of NLP has two distinguishable foci, namely Language Processing and Language Generation (Liddy, 2001).

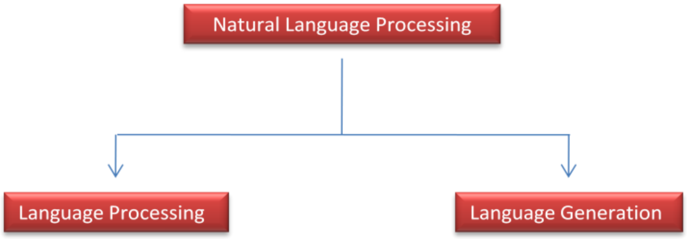


Figure 11: Different focus areas of NLP

The task of Language Processing may be considered analogue to the tasks of reading or listening, while the task of Language Generation may be considered as being analogue to the tasks of writing or speaking. Another traditional distinction in NLP is between Language Understanding - dealing with written language - and Speech Understanding, dealing with oral language and involving fields like acoustics and phonology.

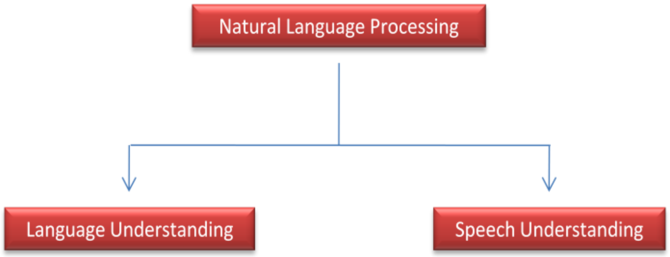


Figure 12: Distinctions in NLP

The Levels of Language, a synchronic model of language, best explains what actually takes place within a NLP system (Liddy, 2001). This model distinguishes the following seven levels of language:

1. *Phonology Level* - Involves the interpretation of speech sounds within and over words.
2. *Morphology Level* – Addresses the componential nature of words, which are composed of morphemes (the smallest units of meaning). An NLP system can gain and represent the meaning of a word by in turn recognising the meaning conveyed by each the word's constituent morphemes.
3. *Lexical Level* – Deals with interpreting the meaning of individual words at word-level. This activity often starts with assigning of a single part-of-speech tag to each word. Lexicons may further be used to provide semantic information for each word. Once the meanings of individual words are understood, meaning across words may be unified to produce complex interpretations.

4. *Syntactic Level* – Syntax carries meaning in most languages since order and dependency contribute to meaning. This level focuses on uncovering the grammatical structure of the sentence by analysing the words in a sentence. Both a grammar and a parser are required to achieve this. As output, a representation of the sentence is obtained which reveals the structural dependency relationships between the words in the sentence.
5. *Semantic Level* - Semantic processing establishes the potential meanings of a sentence by analysing the interactions among word-level meanings in the sentence and may include the semantic disambiguation of words with multiple meanings.
6. *Discourse Level* – This level considers the meaning conveyed by the properties of the text as a whole by making connections between constituent sentences. This may also include anaphora resolution – the process of replacing of words such as pronouns (e.g. “they”) that are semantically vacant with the appropriate denoted entity (e.g. “they” may actually denote “the soccer team” in a sentence).
7. *Pragmatic Level* – The purpose of this level is to account for how additional meaning may be interpreted in texts without actually being encoded in the texts. Context over and above the text content is used for understanding, which requires much world knowledge, including the understanding of intentions, plans, and goals (Liddy, 2001).

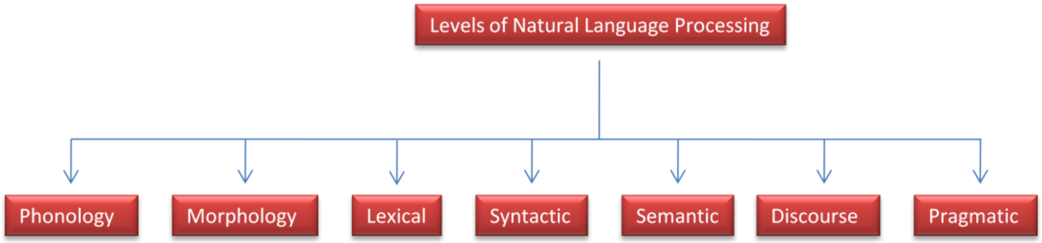


Figure 13: Seven levels of language processing

These seven levels are highly interactive during the process of understanding language. Each level of language carries meaning and humans use all these levels to understand language. The capability of an NLP system may be considered as directly proportional to the number of levels of language it utilises (Liddy, 2001).

NLP approaches may approximately be classified in the following categories:

- Symbolic Approaches
- Statistical Approaches
- Connectionist Approaches
- Hybrid Approaches

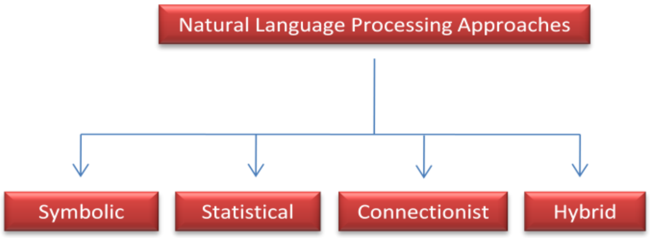


Figure 14: Different types of NLP approaches

Symbolic approaches carry out deep analysis of linguistic phenomena and are grounded on explicit representation of facts about language through well-understood knowledge representation schemes and associated algorithms (Manaris, 1998). Human developed rules and lexicons serve as the principal sources of evidence in symbolic systems. Examples of symbolic approaches are rule-based systems and semantic networks. Symbolic approaches have been used in the following applications (Liddy, 2001):

- Information extraction
- Text categorisation
- Ambiguity resolution
- Lexical acquisition

The following is a list of techniques associated with symbolic approaches (Liddy, 2001):

- Explanation-based learning
- Rule-based learning
- Inductive logic programming
- Decision trees
- Conceptual clustering
- K-nearest neighbour algorithms

Statistical approaches make use of assorted mathematical (or statistical) techniques and frequently also large text corpora to formulate estimated generalised models of linguistic phenomena based on actual examples of such phenomena in the analysed text. Such approaches usually do not require substantial linguistic or world knowledge in order to function. Contrasting with symbolic approaches, statistical approaches rather employ observable data as the principle source of evidence. Statistical approaches are typically used in the following applications (Liddy, 2001):

- Speech recognition
- Lexical acquisition
- Parsing
- Part-of-speech tagging
- Collocations
- Statistical machine translation
- Statistical grammar learning

The Hidden Markov Model (HMM), a statistical model, is often used in statistical NLP approaches.

A **connectionist model** is a network of interlinked simple processing units with knowledge stored in the weights of the connections between units. As for statistical approaches, connectionist approaches also formulate generalised models from examples of linguistic phenomena occurring in text corpora. The fact that connectionist models combine statistical learning with several theories of representation distinguishes it from normal statistical approaches. Therefore, connectionist representations allow transformation, inference, and manipulation of logic formulas.

In recent years, the popularity of **hybrid approaches**, incorporating the strengths of the formerly mentioned approaches, has increased due to their increased flexibility.

5.4 Information Retrieval

The term “Information Retrieval” (IR) is used to describe the task of recovering information from documents or retrieving relevant documents from a document corpus (the latter part is sometimes referred to as “Document Retrieval”). Information retrieval tasks usually starts out with the user expressing his information needs (e.g. the user wants to find information about “integrated manufacturing”) in the form of a query representation. In the background, documents are translated to document representations by the IR system. The IR system then uses the query representation and the document representations to determine which documents best satisfy the information needs. Boolean and vector space models of IR perform this matching done in a formal but semantically imprecise calculus of index terms. This impreciseness arises because given only the query, the IR system has an uncertain understanding of the information needs and therefore makes an uncertain guess as to which documents have content relevant to the specific information needs of the user (Manning, Raghavan & Schütze, 2008). IR approaches using probabilistic theories to reason under this uncertainty have been created to increase the quality of results and will be discussed in Chapter 6. Another relatively new approach to improve a computer system’s ability to understand a user’s information needs to provide more appropriate search results is the semantic web (Heim, 2006).

A large portion of document retrieval research is being performed to support World Wide Web search engines (e.g. Google). Document retrieval generally incorporate techniques that were initially deemed as post-retrieval techniques as part of the retrieval process (the PageRank link analysis technique coined by *Google* is a good example of this). Static and dynamic clustering techniques present documents to users in a manner that allows them to focus on areas of interest (Courseault, 2004). A further use of clustering is as a query expansion mechanism where documents not containing a specific search term are returned when these documents are clustered with the relevant search term.

Generally, various methods and interaction with the user are combined in information retrieval (Courseault, 2004). The purpose of information retrieval (IR) systems is to discover and return information or documents corresponding to a user’s search query from a corpus. Information retrieval systems can also be used to find physical items (e.g. physical books) using the electronic records representing such items (e.g. the electronic metadata of library books). The availability of digital versions of books, newspapers and journals are increasing, allowing IR systems to consider the content of physical information bearing sources (Redfearn, 2006).

5.5 Information Extraction

Information Extraction (IE) may be described as the process of automatically deriving structured data from unstructured natural language texts (Redfearn, 2006). This may involve defining general templates of the sought after information which are then used to facilitate the extraction process. The output of NLP systems is a crucial input to IE. The following are some examples of typical tasks that IE systems can perform (Redfearn, 2006):

- Term Analysis – Identifying the terms, which may be single (e.g. “project”) or multiple words (e.g. “project management”).
- Named Entity Recognition – Identifying the names occurring in a document. These may be names of people (e.g. “Bill Gates”), organisations (“Microsoft”), geographical places (“New York”) or natural attractions (e.g. “Table Mountain” or “Nile River”). Other patterns such as dates, monetary amounts, percentages, quantities, time expressions, etc., may also be recognised.
- Fact Extraction – Identifying and extracting complex facts from documents including relationships between certain entities as well as events.

The extracted information may be stored in appropriate structures that define the terms in the particular domain as well as the relationships these terms have to each other. Such structures may typically include dictionaries, thesauri, taxonomies, ontologies or semantic networks and these serve as inputs for subsequent analysis tasks such as data mining (Redfearn, 2006).

Information Extraction often forms part of a larger text analytical system where it is combined with complementary techniques to synthesise knowledge from large quantities of textual information.

5.6 Data Cleansing

Data Cleansing entails the algorithms and methods which determine the final information to be fed to subsequent analysis such as data mining, Text Mining, cluster analysis or link analysis. Moreover, it impacts the quality and structure of such information. An important part of the data cleansing process is selection – the manner in which words are identified as potential keywords for the analysis and calculating the significance of a specific word in a document.

The initial step in the selection process is defining what will be considered as a word (e.g. only single words versus single words in addition to multiword terms). The reduction in the number of identified words to be considered in the analysis is considered part of the selection process. A simple yet effective way to reduce the number of words for a given analysis is to only consider those words occurring in excess of a specified minimum frequency. Another way to achieve this reduction is to use compression techniques to group together variations of different words having the same meaning (e.g. “summarisation of documents” and “document summarisation”). Stemming[14] and lemmatisation algorithms may further be used to achieve compression. More advanced compression techniques consider words that are different but essentially have the same meaning (e.g. “Internet” and “World Wide Web”).

An additional issue closely related to data cleansing is determining the strength of association between words based on the relative location of words to each other (Courseault, 2004).

5.7 Data Mining

Data Mining (DM), also known as knowledge discovery, may be defined as the process of identifying patterns in large data sets with the objective of discovering previously unknown, useful knowledge (Redfearn, 2006). In the case where data mining is used as part of Text Mining, it is applied to the facts rendered as output of the information extraction phase with the aim to identify patterns in these facts. The results of the data mining process are typically stored in a separate database which may be queried by the end-user using an appropriate graphical interface or may be represented visually (Redfearn, 2006).

Generally, the activities of Link Analysis and Clustering are considered part of the Data Mining process. Data mining is often applied to quantitative data (e.g. sales figures of different products) contained in operational databases to uncover insightful trends and actionable information underlying such data. Link Analysis may be described as the process of linking information contained within documents (Courseault, 2004). In its most basic form it portrays networks of word relationships using some form of co-occurrence measure. More advanced forms of Link Analysis associate certain verb types with the appropriate actor and the objects of the relevant action. This may be used to identify links between entities occurring in texts as well as to identify major events in the analysed text (Courseault, 2004).

5.8 Summarisation

Text summarisation aims to identify and represent the essence of a given piece of natural language text as briefly as possible in order to give an overview of what the original text is about. Although, in most cases, summaries are created by humans many automated summary techniques and tools (e.g. *Copernic Summarizer* and *SUMMARIST*) are available. Automatic text summarisation is generally regarded as a branch of the field of NLP. For a more detailed discussion around automated text summarisation techniques the reader is referred to Mani

[14] Stemming is the process where inflectional (e.g. “seeing”) and sometimes derivationally related forms of a word (e.g. “saw” is reduced to a common base form (e.g. “see”). (Manning, Raghavan & Schütze, 2008)

(2001) or Mani and Maybury (1999). Automatic text summarisation forms part of many business intelligence-, Text Mining-, knowledge management- and document management systems.

5.9 Visualisation

An effective visual interface should permit the user to review, manipulate, search, explore, filter and understand large collections of data. The visualisation of results is the greatest challenge that impedes the effectiveness of text analytical techniques. An effective visualisation technique require merging the perceptual capabilities of the human mind with a computer, but has the restriction that communication must happen through a graphical display of some sort which hampers speedy, complex communication (Courseault, 2004).

Courseault (2004) further mentions that little research is done that concentrates on the theoretical principles of text data representation. Clustering is an aspect of text data mining that necessitates visualisation, even more so for large cluster maps. One of the challenges with the visualisation of cluster maps is providing the ability to navigate through a cluster map while focusing on the details of the cluster and maintaining context in terms of the entire cluster. Another clustering related visualisation challenge is finding particular details in a large cluster map. Kosara, Miksch and Hauser (2002) used existing focus and content methods (e.g. hyperbolic trees) to create a concept named Semantic Depth of Field where less important parts are displayed slightly blurred whereas important areas are displayed sharp (Courseault, 2004). Various visualisation techniques exist, some of which include (Courseault, 2004; Rauffet, 2007):

- One-dimensional representations
- Two-dimensional representations
- Three-dimensional representations
- Three-dimensional representations with time
- Perspective walls
- Network diagrams
- Three-dimensional geographical representations
- Animation
- Directed graphs with chronologically arranged vertices and nested maps
- Tag clouds
- Hyperbolic trees
- Graphs
- Trees
- Cluster Maps
- Fractal Maps
- Tree Maps

Visualisation is considered part of the challenge of retrieving high value information from large collections of electronic text and presenting it to humans in such a way that they can best understand and apply it. Careful thought has to be given to which visualisation techniques will be best suited to the needs of the users of a given text analytical system.

5.10 Text Mining

Text mining, also known as Text Data Mining, concerns a collection of methods that are applied to extract and identify meaningful patterns, relationships and rules from unstructured text data to create intelligence (Francis, 2006; Courseault, 2004). The lines between NLP and Text Mining are blurred making it hard to distinguish the differences between these two fields. In a way, Text Mining is a subset of NLP since it only focuses on textual data whereas NLP may also include speech data. Where NLP focuses more on understanding what a given speaker or writer has expressed, the focus of Text Mining is rather on extracting patterns across a large number of documents (Kao et al., 2005) and generating new information that has predictive value (Francis, 2006).

The patterns and relationships embedded in documents that can be extracted by Text Mining may be extremely difficult or even impossible to find otherwise (Redfearn, 2006). Most of the methods employed in Text Mining were developed in the fields of information retrieval, NLP, statistics, machine learning (Kao et al., 2005), information extraction and data mining (Redfearn, 2006). The various phases of the Text Mining process may be merged together in a single workflow (Redfearn, 2006). Though the interest in Text Mining is fairly recent, text analysis may be traced back to the late 1950s when the automatic abstraction of text was studied (Francis, 2006; Weiss, Indurkhya, Zhang & Damerau, 2005). Artificial intelligence researchers investigated natural language processing during the 1970s and 1980s, but the interest in text analysis declined since many of these early attempts did not yield commercially viable results. The interest in this field was rekindled in the 1990s however, due to new developments in Text Mining tools (Francis, 2006).

NLP techniques, such as part-of-speech tagging, sentence boundary detection and word-sense disambiguation, are used as part of the Text Mining process to provide the information extraction phase with the linguistic data required for meaningful extraction (Redfearn, 2006). While Text Mining and information retrieval often apply NLP techniques such as word stemming, the more advanced NLP techniques have rarely been used in Text Mining (Kao et al., 2005). Information retrieval techniques are used to reduce the number of documents to consider for a specific Text Mining analysis by identifying documents relevant to the particular focus of the analysis, thereby significantly speeding up the analysis process (Redfearn, 2006). Information retrieval may be regarded as the simplest type of Text Mining. However, Text Mining is normally much broader and consists of the following areas (Kao et al., 2005):

- Automatic text classification using a fixed set of categories
- Text clustering where the categories are determined as part of the categorisation
- Automatic text summarisation
- Topic extraction from texts
- Analysis of topic trends in text streams

The following are some of the typical applications of Text Mining (Kao et al., 2005):

- Information Extraction
- Classification
- Search and mining of search results

According to (Courseault, 2004), Text Mining approaches are mainly applied to the following domains:

- Web research publication databases (e.g. Medline and Engineering Index)
- Patent databases
- News source databases

Redfearn (2006) points out that Text Mining can be applied to analyse natural language documents about any subject. A significant new source of information could become available to organisations if information could automatically be extracted from unstructured data (Francis, 2006). Humans have the ability to connect seemingly unrelated facts to formulate new ideas or hypotheses. The explosion in the amount of published text however implies that it is impossible for even the most enthusiastic reader to keep abreast with all the reading in a given field, even disregarding adjacent fields. This phenomenon may cause new insights, ideas and knowledge to remain undiscovered in literature. Text mining may alleviate this problem by supplementing or replacing the human reader with automatic systems that can handle the effects of the text explosion (Redfearn, 2006). According to Francis (2006) Text Mining can be regarded as having two distinguishable phases, namely term extraction and feature creation.

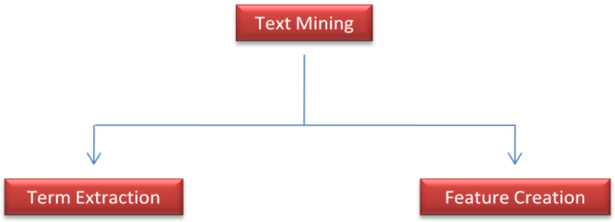


Figure 15: High-level phases in the Text Mining process

Francis (2006) further states that term extraction is the first step in inferring meaning or content from unstructured text. String manipulation functions are heavily used within the text extraction phase, but NLP techniques are also applied. Text is parsed into words during the term extraction phase while frequently occurring words that conveys little meaning, named stopwords (e.g. “a”, “the”, “and”), are further eliminated typically using a stoplist. By eliminating stopwords, the number of word tokens to be analysed may be reduced significantly. Also part of the term extraction phase is identifying words that constitute a unit (e.g. “Bill Branson” and “project management”). Subsequently, words having multiple versions and spellings must be addressed. Stemming is used to substitute all versions of a certain term (e.g. “photos”, “photography” and “photographer”) with the stem of that term (e.g. “photo”). As soon as parsing, stopword removal and stemming have been performed, the feature creation phase commences. In some cases other types of analysis (e.g. grammatical analysis) are performed before starting the feature creation phase.

In the feature creation phase (also known as feature selection) unsupervised learning techniques are used to reduce the potential number of features to a significantly smaller number of variables. This reduced number of features essentially represents the candidate dependent or predictor variables in the relevant analysis. Words and sequences of words are classified into groups carrying similar information as part of the feature creation phase using techniques such as cluster analysis and other dimensionality reduction techniques. Cluster analysis is an unsupervised learning technique that groups elements with similar values together and has the characteristic that it does not have a dependent variable. Clustering is applied in Text Mining to group records having similar words or words with similar meanings together (Francis, 2006). Many different techniques exist to perform feature selection; Manning et al. (2008) describes the following feature selection techniques:

- Mutual information
- χ^2 Feature Selection
- Frequency-based feature selection

Courseault (2004) considers Text Mining process to be constituted by the five following major technique categories:

- (Document) Retrieval
- (Data) Extraction
- (Data) Cleansing
- (Data) Mining
- Visualisation

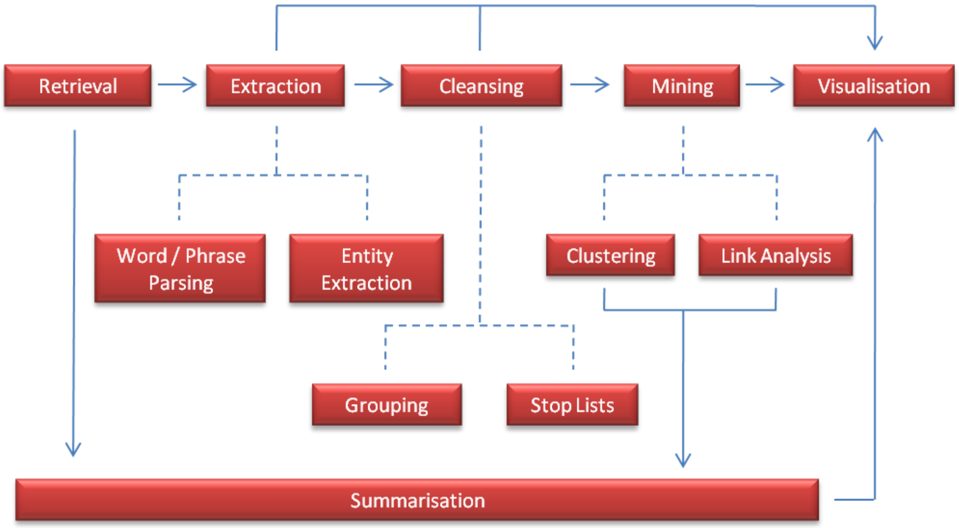


Figure 16: Major technique categories in the Text Mining process

A number of technique categories, such as Clustering, Link Analysis and Summarisation, may be considered as sub-categories of, or supplements to, the five major categories (Courseault, 2004). According to Courseault (2004) the prominence of Text Mining has increased to such an extent that data mining and statistical software tool vendors extended their offering to also include Text Mining products in the past years (e.g. SAS, SPSS, Inxight, LexisNexis, etc.). In spite of this, Text Mining is still a relatively new application and some consider Text Mining software to be relatively underdeveloped in comparison with other data mining applications (Francis, 2006) leaving ample room for future improvement. The cost of major Text Mining systems are currently too high to be economical viable to many smaller organisations.

Before discussing clustering in section 5.13, section 5.11 will first shed some light on the differences between clustering and classification, while section 5.12 will provides some insights into the difference between searching and browsing.

5.11 Clustering versus Classification

Although seemingly very close, clustering and classification are fundamentally different in the field of information retrieval. In both cases a set of elements (e.g. documents) are partitioned into groups, but with the difference that classification requires that the categories to be used be explicitly defined before classification commences, while clustering derives such categories as part of the clustering process. For this reason most classification techniques are **supervised learning techniques** as it also requires training data, a number of items that have been classified already, to train the classifier before it can be unleashed on the unclassified items. The learning algorithm is used to learn a classifier that can be used to map documents to the set of fixed, known classes (a.k.a. categories or labels). To train the classifier, the learning algorithm is provided with a training set consisting out of documents that have been labelled by a human. Once trained, the classifier can be applied to a test set consisting of one or more unseen documents. The classifier then assigns the relevant documents to the fixed classes thereby classifying the documents. The goal of text classification is achieving high accuracy in terms of classifying new data, which is quite different than achieving high accuracy on training data only. A frequently used restriction in text classification, though not very realistic, is that a document can only belong to exactly one class. Classification therefore tries to replicate the categorical differentiation that a human supervisor enforces on a data set. A class may be very specific (e.g. “multi-component injection moulding”) or more general (e.g. “manufacturing techniques”). The classification task may further be referred to as text classification, text categorisation, topic classification or topic spotting (Manning et al., 2008).

In the case of **unsupervised learning techniques**, and therefore clustering, there are no human inputs, and therefore no existing, explicit categories, available to guide the partitioning process. It is the distribution and composition of the data that determine cluster membership in clustering and other unsupervised learning techniques. Clustering is often preferred to classification in applications where little time is required for human input, where no existing categories or classes are available to serve as input, or where the information content is very dynamic in terms of the categories it pertains to, making it difficult to keep the categories up to date. In the light of this information clustering seems like a lucrative avenue to explore in terms of organising innovation-related information.

5.12 Searching versus Browsing

When confronted with the outputs of a text analytical or information retrieval system the user can in most cases either search for specific information or browse the result categories to identify the result elements that interest him. When searching for information, the information seeker mostly has a definite idea about what he wants to find information. He translates his search need into a search query, executes the search, and subsequently scans through the returned results. His information need may be satisfied, where he then aborts the searching process, or he may feel that his information need has not been satisfied, where he then refines his search query in an attempt to find more relevant or specific information. With searching, the information seeker is not presented with anything until such time as the search is executed. Therefore, searching is not as useful when it comes to discovering new or previously unknown subjects.

With information browsing on the other hand, the information seeker is presented with some kind of structure (e.g. a list of categories, a glossary of terms, a hierarchy of categories, etc.) which essentially groups underlying information elements. The information seeker is therefore presented with an overview about what information can be found and he can subsequently select a category to arrive relevant information content or more categories. Browsing is therefore better suited in scenarios where the information seeker is not too sure of exactly what he is looking for.

A combination of searching and browsing functionality is therefore required to enable the information seeker to explore and get an overview of a given field of interest as well as to find specific information by means of precise search queries. The mode selected by the user will depend on his knowledge about the relevant field, his ability to formulate his information need using natural language, the clarity of his information need, his personal information retrieval preferences and the availability, usefulness and accuracy of the browsing structures.

In the scenario where innovation-related information has to be presented to innovation workers in support of their various innovation-related activities the supporting mechanism must provide a blend of both searching and browsing functionality to cater for both modes of information seeking. The next section will examine clustering, a group of techniques that can be used to group information based solely on their content, that forms part of a substantial number of text analytical approaches.

5.13 Clustering

Clustering is a formal statistical approach used to group similar elements together, especially to perform dimensionality reduction along rows. When applied to text data it is often referred to as text clustering. Any type of text clustering is founded on the co-occurrence of words and has the objective to minimise associations between clusters while maximising the relationships within the respective clusters. However, different algorithms have different starting points although the results may be quite similar. The following are some of the differences between the various clustering techniques (Courseault, 2004):

- The type of element making up the clusters (e.g. terms or documents)
- Whether all elements have to be contained in at least one cluster or whether certain elements may be excluded from any cluster
- The number of clusters a certain element may be allocated to (disjoint clusters vs. overlapping clusters)
- The repeatability of the clusters calculated by the clustering algorithm over different runs (deterministic vs. non-deterministic algorithms)

Using clustering in information retrieval is grounded on the assumption that the cluster hypothesis applies:

"*Documents in the same cluster behave similarly with respect to relevance to information needs*."

In other words, given that a document from a specific cluster is relevant to a specific search request, it is probable that other documents from the same cluster would also be relevant to the given search request. This may be explained by the fact that clustering groups documents that have many terms in common. The cluster hypothesis is closely related to the contiguity hypothesis of vector space classification, stating that:

"*Documents in the same class form a contiguous region and regions of different classes do not overlap*." (Manning et al., 2008)

As mentioned before, clustering is applied in Text Mining to group records having similar words or words with similar meanings together (Francis, 2006). Several clustering techniques exist, which often renders the same results despite of the difference in starting point (Courseault, 2004). Most clustering techniques use a measure of dissimilarity to allocate similar elements to clusters. Dissimilarity measures used for numeric data are different to the dissimilarity measures used for categorical data. Text is generally regarded as categorical data, but when words are coded as binary indicator variables the dissimilarity measures used for numeric data may be applied. Since a given word may appear multiple times in text data, the frequency of occurrence of a word may be used instead of a binary indicator variable (Francis, 2006).

The following are some of the most popular measures of dissimilarity:

- **Euclidean distance** – Used for numeric variables and is based on the variable-specific squared deviation between the values of two variables representing two records.
- **Manhattan distance** – Uses absolute deviations rather than squared deviations.

- **Simple matching** – Compares the total number of non-matches to the total number of variables.
- **Rogers and Tanimoto** – More weight is given to dissimilarities than to similarities.

Some clustering techniques apply a measure of similarity rather than a measure of dissimilarity. A well-known measure of similarity is the cosine measure, which measures covariance, and is applied to records rather than variables. Instead of using binary indicator variables for the calculation of the cosine measure, it uses a value called the term frequency-inverse document frequency (tf-idf) statistic that is based on the frequency of a given term in a given record or document. The tf-idf statistic is normalised by dividing it by the total number of times the relevant term appears in all records or documents.

In most cases the aim of clustering text data is to estimate the value of a single new variable indicating to which cluster a given document or record has been assigned. Further useful information may be gained by examining the frequency of occurrence of each word of each cluster to establish which words are important in defining the respective cluster. It is also possible to automatically label the relevant cluster with the most frequently occurring words of that cluster.

5.13.1 Flat Clustering versus Hierarchical Clustering

Flat clustering is efficient, conceptually simple and non-deterministic. It requires the number of clusters to create as input and creates a one-dimensional cluster set, lacking any explicit structure in terms of inter-cluster relationships. Well known examples of flat clustering algorithms are K-means and Expectation Maximisation. (Manning et al., 2008)

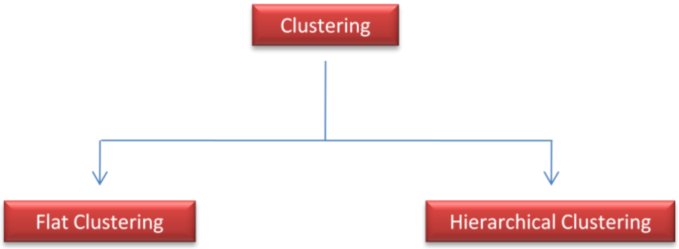


Figure 17: Types of clustering in terms of cluster structure

On the other hand, **hierarchical clustering** (a.k.a. hierarchic clustering) produces a hierarchy of clusters which is more informative than the flat clustering's unstructured cluster set. Hierarchical clustering methods group elements in a treelike structure (dendogram) as opposed to non-hierarchical methods which only divide the corpus into subsets. In the case of **partitional clustering**, a special case of hierarchical clustering, the resulting clusters are disjoint.

Other differences between hierarchical- and flat clustering are:

- Hierarchical clustering does not require the user to specify the number of clusters to create.
- Most hierarchical clustering algorithms are deterministic.

- Hierarchical clustering algorithms are less efficient (typically have quadratic time complexity) than flat clustering algorithms (typically have linear time complexity).

When efficiency is of the essence, flat clustering is normally preferred to hierarchical clustering whereas hierarchical clustering is preferred when the restrictions of flat clustering are problematic. There are researchers claiming that hierarchical clustering produces clusters having superior quality to that of flat clustering although there is not yet consensus over this issue.

5.13.2 Hard Clustering versus Soft Clustering

Clustering algorithms may further be distinguished as either hard or soft clustering algorithms. With **hard clustering** a document is either part of a given cluster or it is not – no partial membership is allowed. Conversely, with **soft clustering** a document's assignment to clusters is a distribution over all clusters allowing for partial cluster membership. Latent Semantic Indexing (LSI) and Expectation Maximisation are examples of soft clustering algorithms.

Some definitions for hard clustering allows a given document to be a full member of more than one cluster, whereas **partitional clustering** always adheres to the restriction of one cluster for any given document. Partitional clustering may therefore be regarded as a special case of hard clustering. (Manning et al., 2008)

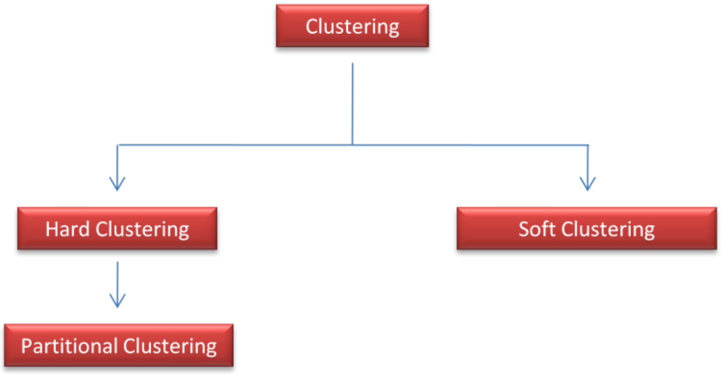


Figure 18: Types of clustering in terms of degree of cluster membership

5.13.3 Exhaustive Clustering versus Non-Exhaustive Clustering

Clustering algorithms may further be distinguished by the minimum number of clusters a given document may belong to. **Exhaustive clustering** requires that each document has to be assigned to a cluster whereas **non-exhaustive clustering** allows certain documents not to be assigned to any cluster. **Exclusive clustering**, a special case of non-exhaustive clustering, requires that each document is a member of either zero or one cluster. (Manning et al., 2008)

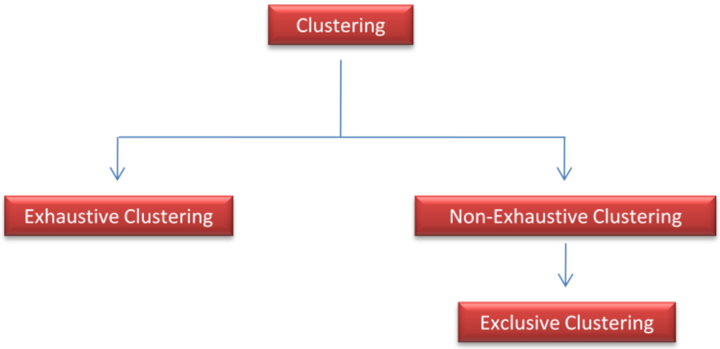


Figure 19: Types of clustering in terms of omission of elements

5.13.4 Cluster Cardinality

Cluster cardinality refers to the number of clusters that are created in a clustering exercise and is often represented by the letter “K”. Estimating the clustering cardinality required for optimal results is a well-known issue with parametric clustering algorithms (i.e. algorithms that require the user to specify K as part of the input) and often comes down to making a good guess based on domain knowledge and experience. Estimating the optimal number of clusters is even described as an art by some researchers. This however, is not a problem with non-parametric clustering approaches such as hierarchical clustering. (Manning et al., 2008)

Using too many clusters results in an over-parameterised model where noise is fitted together with pattern, a phenomenon known as **overfitting**. On the other hand, if too few clusters are used, the data is not modelled to satisfaction (Francis, 2006) as weakly related information is forced to share the same cluster.

Several approaches for determining the optimal number of clusters exist. A simple approach is manually investigating the centres of the clusters of a specific grouping to determine whether the clusters appear sensible. A quick and efficient approach is using forward stepwise regression, an automated procedure for selecting regression model variables, to calculate the best cluster size. It starts with a model having no predictors (i.e. a null model) and subsequently tests all possible candidate independent variables for a single-variable regression model. A significance level is selected and used as a threshold for selecting which variables to enter into the model. The variable that results in the best goodness of fit measure, typically the F-statistic, is then entered in step one. In the next step, all possible two-variable regressions are fit using the variable identified in the previous step. Once more, the variable which results in the largest goodness of fit measure improvement is entered into the model. This procedure continues up to the point where no significant improvement in fit is observed. This however will require that a number of different clustering runs are performed using different values of K.

A more formal way to determine the optimal number of clusters is using a statistical test like the *Swartz Bayesian Information Criterion* that can compare the likelihood functions of two models at a time (Francis, 2006).

5.13.5 Cluster Labelling

In many applications of flat and hierarchical clustering, particularly during cluster analysis and software user interfaces, users interact with clusters. In such settings, clusters should be labelled so that users can see what a cluster is about at a glance (Manning et al., 2008).

Differential cluster labelling selects cluster labels by comparing the distribution of terms in one cluster with that of other clusters using feature selection methods such as mutual information. **Cluster-internal labelling** computes a label that solely depends on the cluster itself, not on other clusters. Labelling a cluster with the title of the document closest to the centroid is one cluster-internal method. Titles are easier to read than a list of terms. A full title can also contain important context that did not make it into the top ten terms selected by feature selection techniques such as mutual information. On the web, anchor text can play a role similar to a title since the anchor text pointing to a webpage can serve as a concise summary of its contents. However, a single document is unlikely to be representative of all documents in a cluster.

It is also possible to use a list of terms with high weights in the centroid of the cluster as a label. Such highly weighted terms (or, even better, phrases, especially noun phrases) are often more representative of the cluster than a few titles can be, even if they are not filtered for distinctiveness as in the differential methods. However, a list of phrases takes more time for users to digest than a well crafted title. Cluster-internal methods are efficient, but they fail to distinguish terms that are frequent in the collection as a whole from those that are frequent only in the cluster.

For hierarchical clustering, additional complications arise in cluster labelling. Not only does the internal node need to be distinguished in the tree from its siblings, but also from its parent and its children. Documents in child nodes are by definition also members of their parent node, so it is not possible to use a naive differential method to find labels that distinguish the parent from its children.

5.13.6 Cluster Evaluation

Clusters may be evaluated in a number of ways. The following main categories of cluster evaluation techniques exist according to Courseault (2004):

- **Separateness** – measures the distinctiveness of every cluster and is based on the principle of minimising similarity between clusters.
- **Cohesion** – measures the density of clusters by taking into account the relationships between cluster elements of every cluster. May be interpreted as the maximum distance between any two elements of a cluster.

- **Precision** – indicates what fraction of the returned results is relevant to the information need (Manning et al., 2008). Can only be used as a clustering measure when all documents form part of at least one cluster, i.e. to evaluate clusters generated by exhaustive clustering techniques (Courseault, 2004).
- **Recall** - What fraction of the relevant documents in the collection were returned by the system (Manning et al., 2008). Similar to precision, recall may only be used as a clustering measure for exhaustive clustering techniques.

5.13.7 Applications of Clustering

Clustering has numerous applications. Some of the benefits that may be achieved by clustering text and documents include:

- Search result clustering (enhance user experience and quality of results)
- Enabling browsing of a document collection
- Improving search results
- Speeding up a search
- Novelty detection and near duplicate detection (Yang & Callan, 2006)
- Metadata discovery on the semantic web (Alonso, Banerjee, & Drake, 2006)

In Text Mining, clustering can be used in many ways including the following:

- Structuring a corpus by discovering topic hierarchies facilitating the organised exploration of the corpus (Larsen & Aone, 1999).
- Constructing a thesaurus, consisting of general and specific terms, based on statistical correlation information of terms. Such a thesaurus presents another way to explore a document corpus (Kaji, Yasutsugu, Toshiko & Noriyuki, 1999).
- As a visualisation technique grouping similar elements to aid information retrieval and navigation (Lowden & Robinson, 2002).
- As a way to understand concepts contained in documents (Courseault, 2004).

5.13.8 Types of Classification Methods

Although section 5.13 is primarily about clustering, this section sheds some light on the different types of classification approaches that are available to classify textual data. Manning, et al. (2008) describes the following types of text classification approaches:

- **Manual Classification** – typically performed without the aid of an computer (e.g. classifying books in a library using the Dewey classification scheme)
- **Hand-crafted Rules Classification** – rules created by humans (e.g. regular expressions) are used by the computer for classification
- **Machine Learning-based Text Classification** – rules are automatically learnt from training data
  - **Statistical Text Classification** – a subtype of machine learning-based text classification that uses statistical learning methods
  - **Vector Space Classification** - another subtype of machine learning-based text classification that uses vector space models for classification
  - **Support Vector Machines** – yet another subtype of machine learning-based text classification that uses support vector machines for classification

Classification techniques are especially useful when representative training data, labelled with the respective classes, are available and when the data to be classified are not too heterogeneous in nature.

5.14 Overview of Text Analytical-Related Techniques Encountered

Table 13 presents an overview of the specific text analytical techniques the author encountered during the process of reviewing the literature on this topic. For a more detailed discussion of these techniques, the user is referred to Appendix A of this report.

	Type	Supervised / Unsupervised	Hard / Soft	Grounds?	Flat / Hierarchical	Deep / Shallow	Exhaustive / Non-Exhaustive	Advantages	Shortcomings
Term Frequency Inverse Document Frequency Technique	Information retrieval	U	-	Algorithmic	-	S	E	Identifying discriminative for words for documents	Little reduction in document descriptions
K-means Clustering	Clustering	U	H	Algorithmic	F	S		Favourable time complexity	Need to specify number of clusters in advance
Hierarchical Clustering	Clustering	U	U	Algorithmic	H	S	E	Helps to find optimal number of clusters	Less efficient than k-Means Clustering
Bayesian Clustering	Clustering	U	S	Bayesian probability theory	H / F	S	E	?	Constitutes a firm basis for inferring the number of distinct components (i.e. clusters) in the model
Self-organising Maps	Dimensionality reduction & Clustering & Data visualisation	U	H	Neural Networks	H	S	E	Clustering is performed nonlinearly on the given input data sets.	Need to specify number of clusters in advance.
Factor Analysis	Clustering & Dimensionality reduction	U	H	Probabilistic (linear statistical approach)	F	S	E	Not extremely difficult to do, inexpensive, and accurate. Factor Analysis can be used to identify the hidden dimensions or constructs which may or may not be apparent from direct analysis.	Usefulness depends on the researchers' ability to develop a complete and accurate set of product attributes. Naming of the factors can be difficult - multiple attributes can be highly correlated with no apparent reason.
Principal Components Analysis	Clustering & Dimensionality reduction	U	H	Algorithmic	F	S	E	Relatively simple method for dimensionality reduction. Model parameters can be computed directly from the data.	Does not work well with high dimensional data or large numbers of data points. It is not obvious how to deal properly with a data set in which some data points are missing
Maximum Likelihood Estimation	Model parameter estimation	U	-	Probabilistic	-	-	-	Reasonably easy to use.	For small samples, the bias of maximum-likelihood estimators can be substantial.
Naïve Bayes Classification	Classification	S	H	Probabilistic	F	S	E	Simple, efficient and robust to noise features and concept drift	Makes low quality probability estimates.
Rocchio Classification	Classification	U	H	Vector Space	F	S	E	Simple and efficient	Inaccurate when spherical class assumption does not match data
K Nearest Neighbour Classification	Classification	U	H	Vector Space	F	S	E	Can handle non-spherical and complex classes better than Rocchio Classification if the training set is large	Less efficient that other classification methods
Expectation Maximisation	Model-based clustering	U	S	Probabilistic	F	S	E	More flexible than k-means and most hierarchical clustering methods	Bad seed documents may cause the model to get stuck at local optima
Support Vector Machines	Large-margin classifier	S	H	Vector Space	F	S	N	Works especially well with little training data. Performance of SVMs is at the state-of-the-art level.	Poor scaling, high computational cost
Latent Semantic Indexing	Clustering & Dimensionality reduction	U	S	Linear algebra	F	S	E	?	Poor scaling, high computational cost

Table 13: Comparison between different text analytical-related techniques

5.15 Applications of Text Analytical Techniques

In general text analytical techniques can be used to improve any application dealing with text. The following are some of the applications that frequently make use of NLP specifically (Liddy, 2001):

- Information Retrieval applications are ideal candidates for NLP techniques due to the significant presence of text in this application.
- Information Extraction applications recognise, tag and extract key elements of information from large text collections. These elements may then be used by other applications such as question-answering, visualisation and data mining applications.
- Question-Answering applications provide the user with either just the text of the answer to a specific question or passages that may such contain answers.
- Summarisation applications reduce a given piece of text into a more concise and rich narrative representation. Higher levels of NLP, e.g. the discourse level, may be used to achieve such summarisations.
- Machine Translation applications may utilise various levels of NLP.
- Dialogue systems currently focus on limited applications (e.g. a car's voice control system) and only utilise the phonetic and lexical levels of language. It is believed that using all levels of language may result in more powerful, useful dialogue systems.

The following are some of the applications of Text Mining:

- Biological Research – Automated Text Mining tools can be used for hypothesis generation. Redfearn (2006) provides an example where Text Mining was used to infer that the drug thalidomide could be used to treat several diseases it had not been connected to previously.
- Search engines – Clustering of search results (Manning et al., 2008)
- Identifying non-evident cross-relationships between topics in a body of researchers (Courseault, 2004).
- Gaining technical intelligence by rapidly monitoring key technical intelligence areas, cleansing and consolidating information into an understandable, concise representation of topics of interest (Courseault, 2004).
- Analysis of survey data – Text mining is used to automate the analyses of answers to open-ended questions in survey data (Francis, 2006).
- Spam filtering – Automatically analysing subject lines and content of e-mails to determine which e-mails are unwanted (Francis, 2006).
- Surveillance – Monitoring various types of communication (e.g. telephone, e-mail, etc.) to single out conversations that may be linked to crime or terrorism (Francis, 2006).
- Call centre routing – Automatically routing calls to help desks or technical support lines based on verbal answers to questions (Francis, 2006).
- Public health early warning – Monitoring news articles and other media from all over the world to provide early warning of possible public health threats (e.g. radioactivity dangers or disease epidemics) (Francis, 2006)
- Alias identification – Analysing the aliases of persons and organisations to identify cases of excessive billing and fraud (e.g. where a single claimant submit multiple insurance claims under different names) (Francis, 2006)
- Predicting important business outcomes (i.e. a dependent variable of interest to the analyst) based on new independent variables identified by the Text Mining process (Francis, 2006).

The applications of text analytical techniques will continue to grow as more types of textual information becomes available, the techniques become more efficient and sophisticated and the processing capabilities of computers expands even further.

5.16 Desirable Characteristics of Text Analytical Techniques for this Project

Through the investigation of the literature about text analytical approaches and their applications, the following characteristics were identified as being desirable for the candidate text analytical techniques to be applied to extract value from textual information applicable to the innovation endeavours of an organisation:

- As little as possible human intervention is desired to make the process as efficient as possible while still producing usable results
- The output should include concentrated, human interpretable abstractions of the input information to serve as a mechanism to gain an overview of the analysed information
- The technique should provide structure to the unstructured textual input data on the level of words as well as on the level of documents
- The technique should not solely be dependent on the availability of predefined classes to guide the output as innovation related information is often novel in nature making it difficult or insensible to predefine the desired output structure
- The technique should be scalable to be able to also handle large collections of electronic documents
- The technique should preferably be independent of language and style of the input textual information it operates on. In other words, it should be able to deliver sensible results when analysing everyday newspaper articles written in more ‘general’ language as well as biochemistry articles dealing with focused terminology and language. This implies that the technique should rather be linguistically shallow than deep (cf. Figure 9) in order not to depend too much on the rules, style and syntax of language.
- As part of the outputs of the candidate technique, quantifiable relationships between the various elements of the outputs should be deducible or included in the output (e.g. the similarity relationships between different documents).

A family of statistical text analytical techniques, known as statistical (or probabilistic) topic models, was identified as suitable in terms of these characteristics and will be presented in detail in the following chapter.

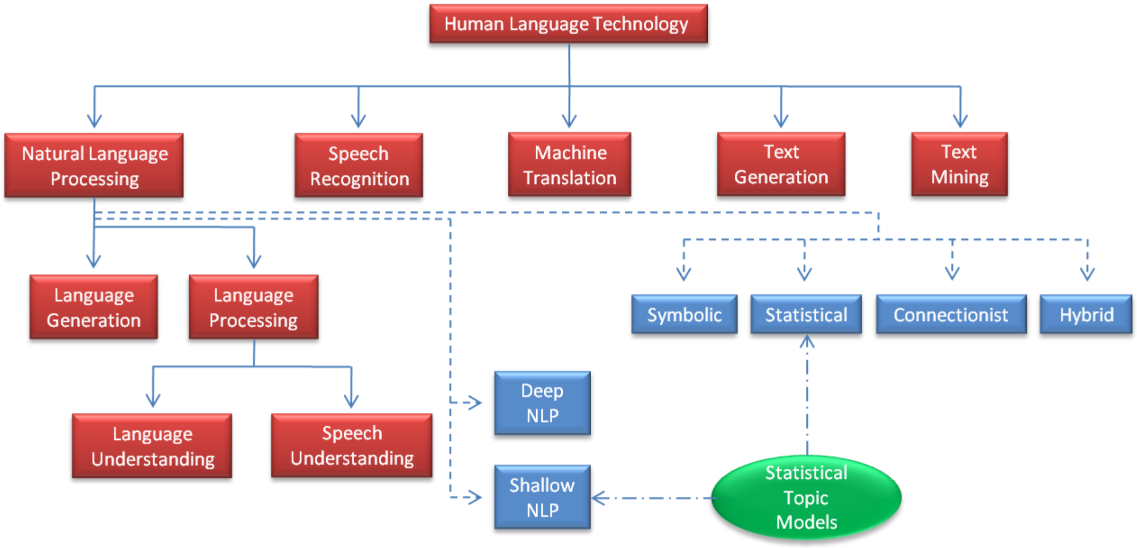


Figure 20: Positioning statistical topic models in relation to HLT and NLP

Figure 20 positions statistical topic models relative to Human Language Technology and Natural Language Processing using various aspects discussed in the introduction of this chapter as well as in section 5.3.

5.17 A Note on Text Analytical Software Encountered during this Project

The following are some of the most promising text analytical software that the author came across during the duration of this project:

- **Carrot^2** – Open source search results clustering engine (http://project.carrot2.org)
- **Digimimir** - Tool for rapid situation analysis of helpdesk and support e-mail (www.digimimir.org)
- **Inxight ThingFinder** - Advanced text analysis technology that automatically identifies and extracts key entities or other "things" from any text data source, in multiple languages (www.inxightfedsys.com/products/sdks/tf/default.asp)
- **Leximancer** - Analytics technology for qualitative data (https://www.leximancer.com)
- **LingPipe** - A suite of Java libraries for the linguistic analysis of human language (http://alias-i.com/lingpipe)
- **OntoGen** - A semi-automatic and data-driven ontology editor focusing on editing of topic ontologies combining text-mining techniques with an efficient user interface (http://ontogen.ijs.si/)
- **VantagePoint** – Commercial text mining software for technology management (www.thevantagepoint.com)

5.18 Summary

This chapter presented an overview of text analytical approaches, with special focus on Natural Language Processing, Text Mining and clustering, to portray the landscape of candidate techniques for extracting value from the textual content of documents containing innovation-related information. It was deduced that classification alone is insufficient to unlock the value of textual information pertaining to innovation as innovation often focuses on novel and relatively unknown concepts not present in the organisation's existing classification structures (cf. Rosen-Zvi, Griffiths, Steyvers & Smyth, 2004). Classification may however be useful in scenarios where known concepts (e.g. the different competitors of the target organisation) are monitored and corresponding information has to be identified.

Clustering, a mechanism that do not depend on existing classification structures to group information, may be useful in organising potentially diverse and unknown information making it better accessible to humans. A combination of clustering and classification may be ideal to best organise and make accessible innovation-related information.

As culmination to the chapter, several desirable characteristics of candidate techniques for further investigation and potential application in this research were identified. A concise list of text analytical software warranting further investigation was presented. The next chapter will investigate statistical topic models, a kind of linguistically shallow, statistical, natural language processing technique that clusters information in the form of interpretable topics while also presenting the relationships of input documents to formulated topics and vice versa.

6. Statistical Topic Modelling

“*A scientist, suddenly faced with access to millions of articles in her field, is not satisfied with simple search. Effectively using such collections requires interacting with them in a more structured way: finding articles similar to those of interest, and exploring the collection through the underlying topics that run through it. The central problem is that this structure—the index of ideas contained in the articles and which other articles are about the same kinds of ideas—is not readily available in most modern collections, and the size and growth rate of these collections preclude us from building it by hand. To develop the necessary tools for exploring and browsing modern digital libraries, we require automated methods of organizing, managing, and delivering their contents.”* Blei et al. (2009)

This chapter addresses the following:

- **Introduce the concept of statistical topic models and some of its applications**
- **Provide an overview of the different kinds of statistical topic modelling techniques**
- **Introduce the terminology used by topic modelling techniques**
- **Compare different statistical topic modelling techniques**
- **Explain the topic modelling process**
- **Discuss the need for linking the results of different topic models**
- **Discuss the need for extending the results of topic models**

This chapter builds on the foundation provided by the previous chapter and introduces statistical topic models, essentially a special kind of (linguistically) shallow (cf. Figure 9), statistical (cf. Figure 14) NLP technique. Many of the text analytical approaches discussed in the previous chapter have the shortcomings that they do not provide an overview of the underlying concepts in the document collection and further do little to capture and exploit inter- and intra document relationships, thereby limiting the inference capabilities and as result the usefulness of the output of the approach.

Topic models was explored, firstly because they provide a good fit in terms of the criteria presented in section 5.15. Secondly, due to the fact that the author had access, over a period of more than three years, to a number of prototypes where topic modelling techniques were implemented in custom developed software at *Indutech (Pty) Ltd* where he was employed during this period. The author could further, due to the nature of his role at *Indutech*, influence the functionalities of such prototypes to a large extent and perform evaluations using such prototypes. This provided the ideal environment for experimenting with the various facets of such topic modelling techniques. The author investigated the application of topic models to numerous collections of textual data as a potential means to semi-automatically extract concentrated, high-value information from the content of electronic document collections.

As with the previous chapter, the goal of this chapter is not to provide a technically or computationally exhaustive insight into the subject of statistical topic models, but rather to present a comprehensive overview to promote a shared understanding.

6.1 Introduction to Statistical Topic Models

Today, the demands for reading or processing textual information are substantial in many industries and organisations. Full text searching (for example using the popular tf-idf technique discussed in section 15.1 of Appendix A), employed by many Internet and personal computer search engines, alone is not sufficient as a first approach to help users understand what a collection of electronic documents is about, since it does not provide the user with an overview of the underlying concepts in the document collection and further does little to capture and exploit inter-document relationships. Many of the text analytical approaches presented in the previous chapter share these shortcomings. Statistical topic models address these shortcomings and were evaluated for their applicability to generate high-value abstractions of the information contained in electronic textual documents. The availability of a dynamic mechanism representing the essence of the document collection in terms of the subjects addressed, the association of individual documents to subjects and vice versa, as well as the associations of documents to other documents in the collection may provide significant assistance in exploring the document collection in question. Statistical topic modelling techniques are well-suited to construct such dynamic mechanisms (i.e. topic models) based on the content of the individual documents constituting the document collection. When such a mechanism is supplemented with another mechanism for finding precise information (e.g. using full text indexing and user-supplied queries) the resulting combined mechanism will be even more useful as it will be able to cater for both focused searching and exploratory browsing.

Blei and Lafferty (2007) states that topic models are useful mechanisms for identifying and characterising various concepts embedded in a document collection allowing the user to navigate the collection in a topic-guided manner. Topics, made up of significant words and terms, provide the user with an overview of the content of the document collection. Each document is represented as a mixture of the automatically constructed topics. The user may use these topics to select documents related to a specific topic of interest and vice versa. Similarities between documents may be found by looking at which documents are assigned to a specific topic enabling the user to find other documents related to a given document.

Topic models enable users to digest a larger number of documents, assisting them in spending more of their time in actually reading than finding relevant information. Topics models could be applied to assist knowledge workers in digesting large collections of textual documents by identifying various concepts embedded in a document collection, thereby allowing the user to navigate the collection in a topic-guided manner. The application of topic models to represent documents has recently received considerable attention in the field of machine learning (Wei & Croft, 2006). Topic models generate interpretable, semantically consistent topics, which can be represented by listing the most probable words describing each topic.

A topic model can be defined as a generative model (refer to section 6.2 for an explanation of generative models) for documents as it specifies a simple probabilistic procedure by which documents can be generated. More specifically, the following statistical inference problem may be formulated: given a set of words, infer the latent or underlying structure from which it was generated. In order to solve this problem the probabilistic process by which the set of words were generated need to be specified. This probabilistic process is known as a **generative model** in the fields of statistics and machine learning (Griffiths, Steyvers & Tenenbaum, 2007). Generative models are applied in machine learning for either directly modelling data, or as an intermediate step to creating a conditional probability density function. A conditional distribution can be created from a generative model through the application of Bayes' rule.

Using methods derived from Bayesian statistics, a set of topics can be learned automatically from a document corpus. According to Griffiths et al. (2007), this process may be considered as a computational parallel of how humans might formulate semantic representations through their linguistic experience. The following analogy can be used as a layman’s explanation of statistical topic modelling: When writing a piece of text, the author selects a distribution over topics (e.g. Topic 1 = “nutritional value of tropical fruit” 50%, Topic 2 = “nutritional requirements of children” 30% and Topic 3 = “salad recipes” 20%) and subsequently (and unknowingly) selects each written word (excluding words having little semantic value such as “a”, “the”, “them”, etc.) according to one of these topics and the specified topic distribution. During the training of topic models statistical techniques are applied to infer a set of likely topics responsible for generating a collection of documents, thus reversing the modelled authoring process (Steyvers & Griffiths, 2006).

Statistical topic models have been successfully used to analyse large quantities of textual information and other forms of discrete count data and are useful for a variety of tasks such as (Blei, Ng & Jordan, 2003; Blei & Laffertly, 2006 [1]; Li, & McCallum, 2006; De Waal, Venter & Barnard, 2007; Fei-Fei & Perona, 2005; Sivic, Rusell, Efros, Zisserman & Freeman, 2005; Pritchard, Stephens & Donnelly, 2000; Erosheva, 2002):

- Information organisation
- Document and text classification
- Document summarisation
- Collaborative filtering
- Language modelling
- Document summarisation
- Data mining
- Information retrieval
- Image analysis
- Biological data analysis

- Survey data analysis
- Forensic investigations
- Modelling of user profiles

Topic models further generate interpretable, semantically coherent topics, which can be examined by enumerating the most likely words for each topic (Mimno & McCallum, 2007 [1]). Topic models are well suited to cater for **synonymy** (multiple words with similar meanings) and **polysemy** (words with multiple meanings), since they assign words to topics based on the context of the document (Mimno & McCallum, 2007[2]). Apart from calculating the topics covered in a collection of documents, topic models also produce, for each document in the collection, a set of individual probabilities that the given document addresses the respective calculated topics. This allows one to learn which documents are significant in terms of which topics. A trained topic model calculates an estimate of the probability of a word given a topic, $P(w|t)$ and the probability of a topic given a document, $P(t|d)$ for all topics calculated and all documents analysed (Mimno et al., 2007 [1]). Document modelling, corresponding to estimating probability of a word given a document $P(w|d)$, is crucial to information retrieval. Under the Latent Dirichlet Allocation topic model, this probability is not explicitly given, but can be derived from the collection of $P(w|t)$ and $P(t|d)$ values. Topic models are a form of **unsupervised learning** since no form of human input or classification is required to learn the latent topics from the document corpus. The reader will recall that with unsupervised learning methods (e.g. clustering and topic models) only the distribution and composition of data influence cluster or topic membership. Supervised learning on the other hand, requires human defined classes or labels for training documents as input to the learning process. Semi-supervised learning involves a mixture of labelled and unlabelled documents (Manning et al., 2008). The reader is further referred to section 5.11 for a related discussion of clustering versus classification. Rosen-Zvi et al. (2004) summarises the major advantages of topics models over other document modelling approaches:

1. Topics are distilled in an entirely unsupervised fashion, requiring no predefined document labels and no special initialisation;
2. Each topic is individually interpretable, rendering a representation that the user can understand;
3. Each document can address multiple topics, therefore capturing the topic combinations that are manifested in textual documents.

Griffiths et al., (2007) state that topic models offer a starting point for further investigations of novel forms of semantic representation. The representation of words by means of topics intuitively corresponds to feature-based models of similarity. Words associated with a given topic in a topic model with high probability tend to be highly predictive of one another corresponding to stimuli that share numerous features that will be highly similar. It should be stressed that although numerous investigations into the application of topic models on textual data have been made, the application of topic models is not limited to text data but can be applied to any type of discrete count data.

6.2 Overview of Different Statistical Topic Modelling Approaches

Several topic modelling related approaches exists. The following are some of the most popular topic modelling related approaches addressed in information retrieval and machine learning literature (Steyvers et al., 2006).

- Unigram Model (Blei et al., 2003)
- Mixture of Unigrams Model, (Nigam, McCallum, Thrun & Mitchell, 2000)
- Probabilistic Latent Semantic Indexing (pLSI) (Hofmann, 1999)
- Latent Dirichlet Allocation (LDA) (Blei et al., 2003)
- Author-Topic Model (Rosen-Zvi et al., 2004)
- Hierarchical Latent Dirichlet Allocation (hLDA) (Blei, Griffiths, Jordan & Tenenbaum 2004)
- Dynamic Topic Model (DTM) (Blei et al., 2006 [2])
- Correlated Topic Model (Blei et al., 2006 [1])
- Pachinko Allocation Model (PAM) (Li et al., 2006)
- Non-parametric Pachinko Allocation Model (Li, Blei & McCallum, 2007)
- Hierarchical Pachinko Allocation Model (hPAM) (Mimno, Li & McCullum, 2007)
- Concept-Topic Model (Chemudugunta, Holloway, Smyth & Steyvers, 2008)
- Hierarchical Concept-Topic Model (hCTM) (Chemudugunta, Smyth & Steyvers, 2008)

Most of these approaches use some kind of dimensionality reduction technique to represent documents using fewer words. A further characteristic common to most of these approaches is that generative probabilistic models of language are employed to represent documents. A generative model is used to randomly generate observed data, usually including some hidden parameters. Another common characteristic of most topic modelling approaches is that they are based on the assumption that the order of words in a document can be neglected – the so-called **bag-of-words assumption**. In terms of probability theory, this assumption corresponds to an assumption of exchangeability for words in a document (Blei et al., 2003). The last communality of most topic modelling approaches is the assumption that the specific ordering of documents in a corpus can be ignored as well (Blei et al., 2003). The dynamic topic model (DTM), as presented in Blei et al. (2006 [2]) is an exception to this rule since the sequence of documents is indeed factored into this model.

The various topic models presented above are discussed in greater detail in Appendix B.

6.3 Topic Model Terminology

In order to formally introduce and compare different topic models, the following terms need to be defined (cf. Blei et al., 2003):

- A **word** is the basic unit of discrete data, and is defined to be an item from a vocabulary indexed by $\{1, \dots, V\}$. Words are represented by means of unit-basis vectors having a single component with a value of 1 and all other components having zero values.

Therefore, the v_{th}word in the vocabulary is denoted by a V-vector w having $w^v = 1$ and $w^u = 0$ for $u \neq v$.

- A **document** is defined as a sequence of N words represented by $\boldsymbol{w} = (w_1; w_2; \dots; w_N)$ where w_n is the n^th word in the sequence.
- Lastly, a **corpus** is a collection of M documents denoted by $D = \{\boldsymbol{w}_1; \boldsymbol{w}_2; \dots, \boldsymbol{w}_M\}$.

For a more detailed overview of the mechanisms behind the different topic models presented in the previous section, the reader is referred to Appendix B.

6.4 Comparing Different Topic Models

Table 14 presents a comparison of some of the key characteristics of the different topic modelling approaches listed in section 6.2 and discussed in more detail in Appendix B. Firstly, here follows a brief explanation of the characteristics presented in Table 14:

- **Unsupervised**: Does the technique require training data supplemented with classification labels (i.e. supervised technique) or not (i.e. unsupervised technique)?
- **Non-parametric**: Does the technique require that one specifies in advance the number of topics desired in the output (i.e. parametric technique) or not (i.e. non-parametric technique)?
- **Structure**: What is the structure of the resulting topics, e.g. flat (i.e. no vertical or horizontal relationships between topics), hierarchical (i.e. only vertical relationships), series (i.e. only horizontal relationships), or graph (i.e. horizontal and vertical relationships)?
- **Affinity between topics**: Are there some form of explicit affinity between topics included in the output?
- **Inputs**: What inputs are required for the respective techniques?
- **Cater for topic evolution**: Does the technique explicitly model the evolution of topics over time or is the time dimension ignored?
- **Explicitly models authors**: Does the technique explicitly include author-topic profiles as part of the output?
- **Probabilistic basis**: Are the respective techniques founded on a statistical foundation?
- **Automatic topic label generation**: Are topic labels included in the output?
- **Multiple topics per document**: Can a document be associated with more than one topic?
- **Word independence assumption**: Does the technique assume that words in the analysed documents are statistically independent?
- **Document independence assumption**: Does the technique assume that documents in the analysed document collection are statistically independent?
- **Distributed inference technique available**: Are there one or more inference techniques available for the respective techniques that will allow for multi-processor processing of the inference problem?

Criteria	LSI	Unigrams Model	Mixture of Unigrams Model	pLSI	LDA	Author-Topic Model	hLDA	DTM	CTM	PAM	Non-parametric PAM	Hierarchical PAM	Concept-Topic Model	HCTM
Unsupervised	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Non-parametric	N	N	N	N	N	N	Y	N	Y	N	Y	Y	N	Y
Structure	Flat	Flat	Flat	Flat	Flat	Flat	Hierarchy	Series	Graph	Graph	Graph	Hierarchy	Flat	Hierarchy
Affinity between topics	N	N	N	N	N	N	Vertical	Horizontal	Y	Y	Y	Y	N	Vertical
Inputs	Term- Document Matrix	Term- Document Matrix	Term- Document Matrix	Term- Document Matrix	Term- Document Matrix	Term- Document Matrix	Term- Document Matrix + Authors per document	Term- Document Matrix + Publication date per document	Term- Document Matrix	Term- Document Matrix	Term- Document Matrix	Term- Document Matrix	Term- Document Matrix + Concepts	Term- Document Matrix + Concept Hierarchy
Cater for topic evolution	N	N	N	N	N	N	N	**Y**	N	N	N	N	N	N
Explicitly models authors	N	N	N	N	N	**Y**	N	N	N	N	N	N	N	N
Probabilistic basis	**N**	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Automatic topic label generation	N	N	N	N	N	N	N	N	N	N	N	N	N	N
Multiple topics per document	Y	**N**	**N**	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Word independence assumption	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Document independence assumption	Y	Y	Y	Y	Y	Y	Y	**N**	Y	Y	Y	Y	Y	Y
Distributed inference technique available	N/A	N	N	N	**Y**	N	N	N	N	N	N	N	N	N

Table 14: Comparison between different topic modelling techniques

The Latent Dirichlet Allocation (LDA) topic modelling approach has been used in the majority of the investigations presented later in this research even though LDA is not the latest topic modelling technique available[15]. When viewed as a kind of clustering approach, LDA can be regarded as a **soft**, **flat**, **exhaustive** clustering technique.[16] LDA was used for the following reasons:

- The relative simplicity of programmatically implementing LDA compared to other topic modelling approaches (e.g. PAM or CTM)
- The fact that LDA is widely experimented with and well-addressed in academic literature makes it a fairly safe choice compared to other less published topic modelling approaches (e.g. DTM or the Concept-Topic Model)
- The availability of a distributed inference mechanism specially tailored to LDA allowing for the parallelisation of the inference problem over different processor cores of a multi-core processor computer. This can reduce the duration of the analysis by as much as 40% compared to the scenario where no parallelisation is used.

The author was involved in the specification, testing and refinement of several software prototypes, developed by Indutech, having LDA at its core. These prototypes however included a range of other complementary functionalities like information extraction from various electronic document formats, pattern extraction using regular expressions, post-processing LDA results to infer relationships not part of the standard output as well as to optimise the speed of interaction with the results database, constructing a vocabulary based on the significant words and terms of the analysed documents, presenting the topic model results in an interactive, visual environment enabling the exploration of the model, integrated searching to find topics, documents, text snippets and words associated with a given search query, integrated Internet look-up of unknown words, capability to rate the usefulness of individual documents and topics, to name a few.

The Concept-Topic Model was further identified for further investigation due to its unique ability to incorporate prior human knowledge and judgement into the process of constructing the associated generative model for the document collection analysed.

6.5 Topic Modelling Process

When considering the topic modelling process from start to end, two kinds of activities can be distinguished, namely analyst activities and activities part of computer-based processes. These two activity types are both important in formulating the desired end result – an interpretable topic model representative of the essence of the input document collection. Figure 21 summarises the analyst- and computer-based processes.

[15] In addition to LDA, the author has experience with LSI and the Concept-Topic Model.

[16] See section 5.13 for a discussion around clustering.

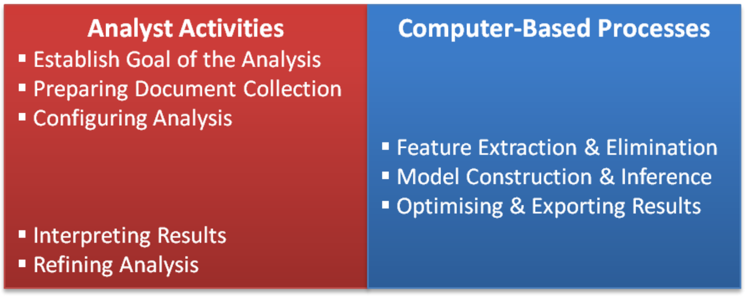

Analyst Activities
Establish Goal of the Analysis
Preparing Document Collection
Configuring Analysis
Interpreting Results
Refining Analysis
Computer-Based Processes
Feature Extraction & Elimination
Model Construction & Inference
Optimising & Exporting Results

Figure 21: Analyst- and computer-based processes involved in topic modelling

The majority of computer-based processes occur between the Configuration and Interpretation stages and essentially represent extracting text, constructing the topic model and retrieving the latent topics by means of inference techniques. The different phases in the topic modelling process will subsequently be discussed. Note that these phases may include some elements that are specific to the software, CAT[17], the author used for doing topic modelling analyses.

6.5.1 Preparation: Preparing the Document Collection

The preparation activities essentially entail the identification of the documents to analyse and ensuring that the target documents are in a format suitable for text extraction. For instance, documents should not be password protected and scanned-in documents should undergo an optical character recognition process to ensure that document text is stored as actual symbols and not just as images. Lastly, target documents should also be stored at a location accessible by the topic modelling software.

6.5.2 Configuration: Configuring the Analysis

Once the analyst completed the preparation activities, the next step is to configure the planned analysis by providing values and adjusting various parameters in the topic modelling software. Firstly, the analyst can supply a descriptive name for the analysis. The analyst then specifies the locations of the documents to be analysed by specifying one or more folder paths. Each specified folder may contain any number of sub-folders which may automatically be included in the analysis by setting the “include sub-folders” flag. For parametric topic modelling techniques like LDA, the analyst further has to specify the number of topics to be formulated based on the content of the analysed documents. Typically, less topics result in shorter analysis durations and more general topics. Conversely, more topics normally result in longer analysis durations and more specific topics. The analyst can further specify the minimum number of times a given word has to occur in the analysed documents before being included in the analysis (this corresponds to

[17] For more information about CAT, visit www.analyzecontent.com

the minimum frequency parameter). For example, if a minimum frequency value of “3” is specified, words and terms occurring only once or twice in all documents analysed will not be included in the analysis. This mechanism can be used to reduce the number of features (i.e. words) that enter the topic model construction phase without dramatically reducing the quality of the resulting topic model. The higher the minimum frequency value, the less features are included in the model and the shorter the analysis duration.

The analyst can also compose, edit and subsequently select one or more stoplists to eliminate specific words having little meaning, known as stopwords, from the analysis. The following parts of speech types are usually included in a typical stoplist:

- Articles (e.g. *a*, *the*, *an*, *his*, *that*, etc.).
- Prepositions (e.g. *of*, *with*, *about*, *from*, etc.).
- Conjunctions (e.g. *and*, *or*, *when*, *because*, etc.).
- Adverbs (e.g. *unfortunately*, *quickly*, *furthermore*, etc.)
- Pronouns (e.g. *he*, *she*, *they*, *them*, *that*, etc.)

Adjectives, e.g. *hot*, *small*, *old*, *new*, etc., may also be included in stoplists, but in the experience of the author they often form part of important terminology, often in the form of noun phrases like “hot forging” and “new product development”. For this reason it is wise not to include too many adjectives as part of the stopword lists. Stopwords are language specific[18] (e.g. “*die*” is a stopword in Afrikaans, but not in English), domain specific (e.g. the term “*sp*” occurs with high frequency in many biological and ecological fields and may be considered a stopword in such fields) as well as information source specific (e.g. in a given journal, the name of the journal will appear on every page, but carries no real meaning in this case). Stopwords are specified, firstly to reduce the number of features included in the model and therefore to reduce the duration of the analysis and secondly, to eliminate words and terms that may potentially be included in almost all topics therefore contributing little to the interpretation of such topics.

The analyst may further elect whether to include numbers in the analysis or only words and terms. The analyst can also specify whether to use **unigrams** (single words, e.g. “project”), **bigrams** (two-word terms, e.g. “Cape Town”), **trigrams** (three-word terms, e.g. “Latent Dirichlet Allocation”), or combinations thereof to during features extraction. Once the various parameters have been specified the analyst initiates the analysis.

With the **Concept-Topic Model** technique, the analyst has to additionally specify the individual concepts (i.e. a predefined topic), using characterising words, as part of the input to the analysis. The technique then uses the specified concepts to guide the generation of the topic model.

[18] The author has compiled and refined stoplists for English, Afrikaans, French, German, and Dutch during the course of this study.

6.5.3 Feature Extraction and Elimination

The first activity occurring after analysis initiation is extracting text from the individual documents of the specified document collection. Features correspond to the individual elements that will be used to construct the topic model. Features may of different types, e.g. unigrams, bigrams, trigrams, dates, named entities, years, etc. The topic modelling software determines which feature types can be extracted from document text. Naturally, the more feature types are elected for extraction the longer the extraction process will take, the more features will be created and the richer the resulting topic will be. Punctuation marks usually do not form part of features except maybe for words containing hyphens (e.g. "sub-populations"). Along with all extracted features, information about the specific documents in which the respective features occurred in, as well as the frequency of occurrence per document, is kept. Feature extraction is also known as tokenisation.

Feature elimination on the other hand entails winnowing the extracted features. All features contained in the stoplists specified for the analysis are eliminated as well as all features occurring less than the specified minimum frequency value. Numbers may also be eliminated if specified as such in the analysis configuration. The collection of all remaining features is known as the **corpus vocabulary**. These features are then used for constructing the topic model.

6.5.4 Model Construction and Inference

The input to the topic modelling process is a **word-document co-occurrence matrix**, where each row represents a word, each column represents a document, and the entries indicate the frequency with which the specific word occurred in the specific document. The word-document co-occurrence matrix is the output of the Feature Extraction and Elimination phase. The word-document co-occurrence matrix is used to infer the latent structures (topics) given the observed words in documents. The Latent Dirichlet Allocation (LDA) model assumes that words (which were observed in document text) are generated by a mixture of topics (i.e. latent multinomial variables) and that these topics are infinitely exchangeable within a document. Moreover, documents are represented as random mixtures over topics where each topic is defined or characterised by a distribution of (corpus vocabulary) words.

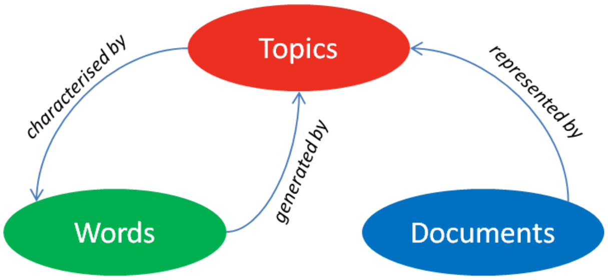


Figure 22: Conceptual framework illustrating the assumptions of the LDA topic model

All statistical topic modelling approaches specify a model that is assumed to explain a set of observed data. This model actually specifies a process for generating a document. LDA specifies the following iterative process steps for generating a document:

- For each **document** in the document collection, select a distribution of topics from a Dirichlet distribution.
- For each **word** in the given document, select a single topic from the distribution of topics that was selected in the previous step.
- For each **word** in the given document, select a word from the distribution corresponding to the topic selected for the specific word in the previous step.

The model specified by a statistical topic modelling approach has parameters that are fine-tuned in the training of the model to best explain the observed data, namely the documents and the words they comprise. The models that are used by the different statistical topic modelling approaches differ in its details but not in its principles. These models are multidimensional due to the extremely large parameter space associated with the problem at hand. The aim is to find the probability *p(observed data | model and its parameters)*. This probability is known as the likelihood function (of the model) and represents the likelihood of the parameters assigned to the model. Once we know the *Likelihood of the model* (or rather the likelihood function of the model), we can recover the latent variables of the model, which in the case of LDA are the multinomial variables representing the actual topics. Bayes' theorem states that: $p(x|y) = \frac{p(y|x)p(x)}{p(y)}$, meaning:

$$Posterior\ probability\ of\ parameters = \frac{Likelihood\ of\ parameters\ x\ Prior\ probability\ of\ parameters}{Evidence}$$

In this equation the *Evidence* term is known as it is represented by the words and their frequencies encountered in the different documents analysed. Also, the *Prior probability of the parameters* can be obtained by assuming a certain distribution (e.g. a Dirichlet distribution in the case of LDA). To get the *Likelihood of the parameters*, inverse probability can be used. The only unknown term is therefore the *Posterior probability of the parameters*. Because of the multidimensional nature of the model, exact inference of the posterior is impossible and therefore it has to be approximated. Here we need an intelligent mechanism to guide a random walk through the parameter space of the model to numerically estimate the *Posterior probability of the parameters*. Many Monte Carlo-like techniques, all based on random sampling, can be used for inferring the *Posterior probability of the parameters*. The act of solving the *Posterior probability of the parameters* is known as inference in Bayesian statistics. Statistical inference algorithms such as Expectation Propagation, Expectation-Maximisation, Markov Chain Monte Carlo, Variational Bayes, or Variational Expectation-Maximisation can be used to infer the posterior probability of the LDA parameters. Once the *Posterior probability of the parameters* has been approximated, the *Likelihood of the parameters* can be calculated and the latent topics can then be retrieved from this likelihood.

The inference results in a set of topics, including the words associated with each topic, together with the strength of association for each word-topic combination. For a more detailed explanation of the mechanics of LDA, the reader is referred to section 16.4 of Appendix B.

The standard outputs of a topic model include the topic-word matrix capturing the topic-word affinities and the document-topic matrix capturing the document-topic affinities. The relationships explicitly given by these matrices are:

- Topic-Word Affinities
- Word-Topic Affinities
- Document-Topic Affinities
- Topic-Document Affinities

Using the information in these two matrices the following additional, implicit relationships can be calculated:

- Topic-Topic Affinities
- Word-Word Affinities
- Document-Document Affinities
- Document-Word Affinities
- Word-Document Affinities

Figure 23 presents a conceptual framework of the different aspects of a standard topic model. Dotted lines represent implicit relationships inferable from data in the standard topic model.

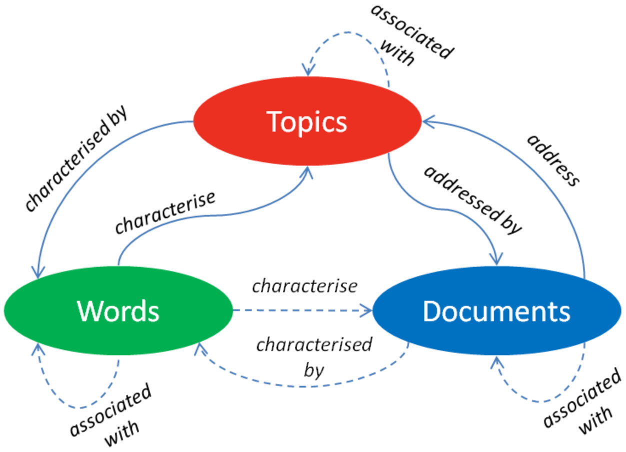


Figure 23: Conceptual framework illustrating the different aspects of a topic model

A note on randomness: Due to the necessity of using random-based inference techniques to find the topics underlying a given document collection in a realistic amount of time, consecutive runs on the same document collection using the same configuration are not guaranteed to have exactly the same output. This is especially the case for relatively small datasets. Often new topics appear in subsequent runs, although the majority of topics (about 70%) are found across two subsequent runs. In the view of the author, the advantages gained by using statistical topic modelling, including its associated random-based inference techniques, by far outweigh the associated disadvantages.

6.5.5 Optimising and Exporting Results

In this phase, the topic model results are optimised for exporting and viewing. When considering the complete set of (nine) possible relationships between topics, documents and words, the output can be quite substantial in size creating the need for optimisation before storage. The analysis results and statistics are typically exported to one or more formats, including (relational) databases, spreadsheets, XML or CSV files, for subsequent evaluation by the analyst and later for usage by other persons.

6.5.6 Interpretation

During the interpretation phase the analyst evaluates the formulated topics and supplement topics with descriptive labels. Figure 24 shows a conceptual framework explaining the interaction between humans and topics.

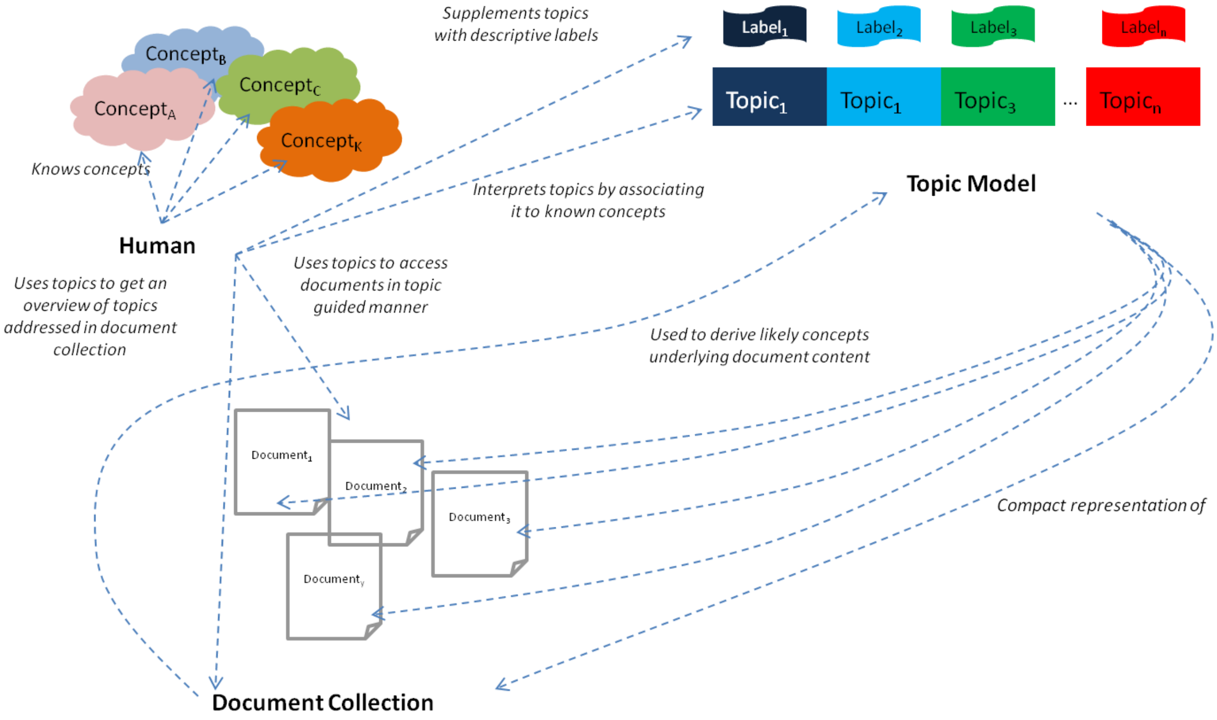
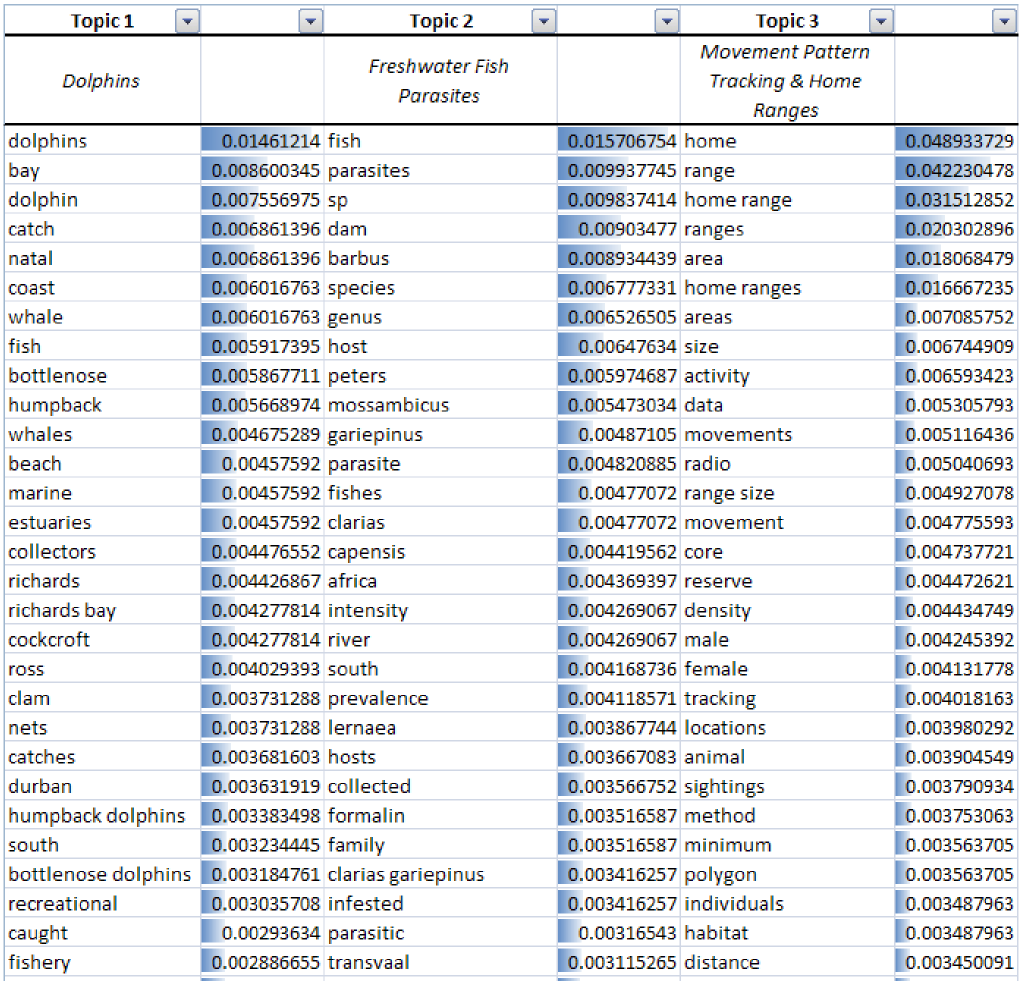


Figure 24: Conceptual framework of topic model interpretation and use

Although not strictly required, labels help humans to interpret topics as it captures the essence of the words constituting the topic. In the experience of the author, the following process can be followed to compose a label for a topic:

- Examine the words associated with the topics, as well as at their strength of association (i.e. the respective probabilities $P(word|topic)$ given by the topic-word matrix), and see if any concepts comes to mind
- Identify the documents having the strongest association to the topic (using the document-topic matrix), open two or three such documents and scan their filenames, titles, abstracts, paragraph headings or selected content to see if any concept comes to mind.
- Optionally, look at other topics indicated as being strongly related to the topic in question in order to put the topic in context. The availability of topic affinity scores depends on the topic modelling software used.

Table 15 shows an extract of three topics from the **topic-word matrix**, one of the main outputs of a topic model, of a 100-topic analysis on a medium-sized document collection. This collection contained 878 documents (research articles), published over a period of 38 years in a popular South African wildlife research journal, addressing a range of wildlife research related subjects. The objective of this topic modelling exercise was to obtain an overview of the types of wildlife information that were addressed in publications over the relevant 38-year period. For this analysis the corpus vocabulary contained a total of 44 011 unique words.



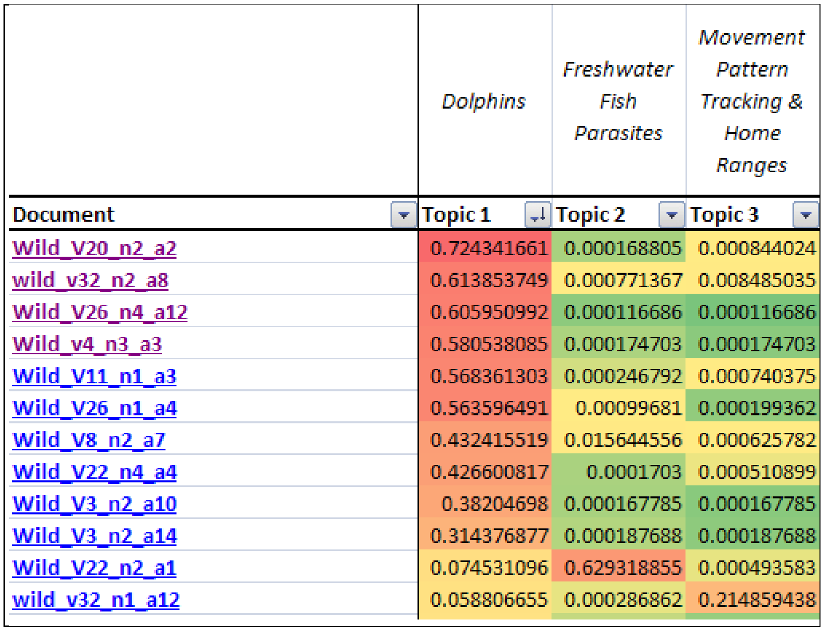
Topic 1		Topic 2		Topic 3	
Dolphins		*Freshwater Fish Parasites*		*Movement Pattern Tracking & Home Ranges*	
dolphins	0.01461214	fish	0.015706754	home	0.048933729
bay	0.008600345	parasites	0.009937745	range	0.042230478
dolphin	0.007556975	sp	0.009837414	home range	0.031512852
catch	0.006861396	dam	0.00903477	ranges	0.020302896
natal	0.006861396	barbus	0.008934439	area	0.018068479
coast	0.006016763	species	0.006777331	home ranges	0.016667235
whale	0.006016763	genus	0.006526505	areas	0.007085752
fish	0.005917395	host	0.00647634	size	0.006744909
bottlenose	0.005867711	peters	0.005974687	activity	0.006593423
humpback	0.005668974	mossambicus	0.005473034	data	0.005305793
whales	0.004675289	gariepinus	0.00487105	movements	0.005116436
beach	0.00457592	parasite	0.004820885	radio	0.005040693
marine	0.00457592	fishes	0.00477072	range size	0.004927078
estuaries	0.00457592	clarias	0.00477072	movement	0.004775593
collectors	0.004476552	capensis	0.004419562	core	0.004737721
richards	0.004426867	africa	0.004369397	reserve	0.004472621
richards bay	0.004277814	intensity	0.004269067	density	0.004434749
cockcroft	0.004277814	river	0.004269067	male	0.004245392
ross	0.004029393	south	0.004168736	female	0.004131778
clam	0.003731288	prevalence	0.004118571	tracking	0.004018163
nets	0.003731288	lernaea	0.003867744	locations	0.003980292
catches	0.003681603	hosts	0.003667083	animal	0.003904549
durban	0.003631919	collected	0.003566752	sightings	0.003790934
humpback dolphins	0.003383498	formalin	0.003516587	method	0.003753063
south	0.003234445	family	0.003516587	minimum	0.003563705
bottlenose dolphins	0.003184761	clarias gariepinus	0.003416257	polygon	0.003563705
recreational	0.003035708	infested	0.003416257	individuals	0.003487963
caught	0.00293634	parasitic	0.00316543	habitat	0.003487963
fishery	0.002886655	transvaal	0.003115265	distance	0.003450091

Table 15: Example of a partial topic-word matrix

Table 15 shows the 30 most significant words (unigrams or bigrams more specifically) characterising the topic. The topic modelling software used returns 100 characterising words per topic, but due to size restrictions the remaining 70 words are not illustrated in this table. For each topic, the words are listed in order of decreasing significance to the relevant topic determined by the calculated (conditional) probabilities of the respective words given the topic. Note that for each topic the topic model calculates a probability $P(word|topic)$ for all words in the corpus vocabulary. For any given topic, the sum of all probabilities $P(word|topic)$ always amounts to 1. This explains the relative small probabilities shown in Table 15.

Although Table 15 only presents a partial topic-word matrix, it illustrates that the calculated topics give a good overview of the subjects underlying the analysed document collection. For example, it is easy to distinguish Topic 1 from Topic 2 although both seem to deal with creatures living in water. Note that topics are generally not mutually exclusive with respect to the words that constitute them (e.g. the word "fish" occurs in both Topic 1 and 2). However, two topics can never share all of the same characterising words with identical probabilities.

Table 16 (below) shows an extract of the **document-topic matrix**, another output of the topic model, that lists the probability that a specific topic is described by a given document. Higher probabilities indicate stronger associations between topics and documents whereas lower probabilities indicate the opposite. Note that a topic can have one or many associated documents and that a document can have one or many associated topics.



	Dolphins	*Freshwater Fish Parasites*	*Movement Pattern Tracking & Home Ranges*
Document	**Topic 1**	**Topic 2**	**Topic 3**
Wild_V20_n2_a2	0.724341661	0.000168805	0.000844024
wild_v32_n2_a8	0.613853749	0.000771367	0.008485035
Wild_V26_n4_a12	0.605950992	0.000116686	0.000116686
Wild_v4_n3_a3	0.580538085	0.000174703	0.000174703
Wild_V11_n1_a3	0.568361303	0.000246792	0.000740375
Wild_V26_n1_a4	0.563596491	0.00099681	0.000199362
Wild_V8_n2_a7	0.432415519	0.015644556	0.000625782
Wild_V22_n4_a4	0.426600817	0.0001703	0.000510899
Wild_V3_n2_a10	0.38204698	0.000167785	0.000167785
Wild_V3_n2_a14	0.314376877	0.000187688	0.000187688
Wild_V22_n2_a1	0.074531096	0.629318855	0.000493583
wild_v32_n1_a12	0.058806655	0.000286862	0.214859438

Table 16: Example of a partial document-topic matrix

Each row in the document-topic matrix represents the respective allocations of a document to the respective topics. The individual document-topic affinity values are called **mixture ratios**. For each document, the sum of all its mixture ratios over all topics amounts to 1. An indication of topic coverage, an estimation indicating how well the given topic is represented in the analysed documents, can be obtained by calculating the sum of all mixture ratios of all documents for each respective topic (i.e. calculating the sum per column in the document-topic matrix). When normalising these sums a relative indication can be obtained of how well a given topic is covered in the document collection compared to other topics. For example, in the 100-topic analysis discussed previously, Topic 1 had a coverage score of 0.72%, Topic 2 a coverage score of 1.05% and Topic 3 a coverage score of 1.07% indicating that according to the topic model Topics 2 and 3 were slightly better represented in the analysed articles than Topic 1.

The document-topic matrix in conjunction with the topic-word matrix provides a useful mechanism to do the high-level exploration of a document corpus. The topic-word matrix can be analysed first to get a feel for the range of topics the document corpus contains, without necessarily reading any documents. Next, one can assess which topics are well represented in the document corpus and which are not, by looking at the normalised sum of mixing ratios of each topic. A topic of interest can be selected from the topic-word matrix and documents significantly describing this topic can be identified by inspecting the document-topic matrix. A prioritised list of documents on a certain topic can thus easily be obtained, and the process may then be repeated for other topics of interest. The topic model also organises the document collection by associating documents and topics – a process that can take a substantial amount of time when doing it manually (Uys, Du Preez & Uys, 2008).

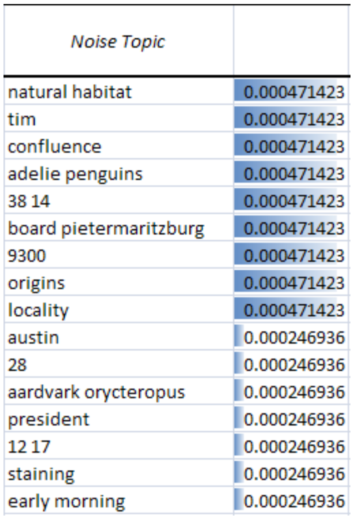
6.5.7 Refinement

The last analyst activity in the topic modelling process entails considering possible changes in the analysis configuration to further improve the results of subsequent analysis using the same document collection. Such possible changes include:

- **Identifying additional stopwords**. When the analyst observes candidate stopwords as part of the words characterising topics he may add such words to one or more stoplist for future elimination. Conversely, if the analyst finds that certain core words are missing in that they are not found in any topic, he should scan through the stoplists used for the specific analysis and eliminate such words from the stoplists where applicable.
- **Changing the minimum frequency value**. The analyst can experiment with different values of the minimum frequency parameter. Increasing this value will speed up the analysis and reducing it will have the opposite effect. When analysing very brief documents (e.g. abstracts) the analyst can make this value very low (e.g. 1 or 2).
- **Changing the number of topics**. When topics are too general or broad, the analyst can increase the number of topics. Conversely, if topics are too specific or narrow, the analyst can decrease the number of topics desired. Decreasing the number of topics will also speed up the analysis.
- **Altering the input document collection**. The topic model can help the analyst in identifying undesired documents, e.g. multiple copies of the same document or documents written in foreign languages. If desired, the analyst may subsequently remove such documents from the collection and repeat the analysis to achieve even better results.

- **Ignoring numbers**. When the words characterising topics are dominated by numbers, the analyst can repeat the analysis with the 'ignore numbers' flag set.

The quality of the analysis results can be judged (in terms of topic interpretability) by the number of topics where the respective characterising words more or less describe an apparent, unambiguous theme. In addition, the document-topic assignments should be logical for the majority of documents that the analyst samples. In most analyses it is not uncommon to find one or more 'noise topics', i.e. topics that seem illogical and cannot be associated with a theme. Where the probability distributions of words associated with normal topics portrays a fairly smooth curve, the probability distributions of noise topics are often abrupt or discontinuous like illustrated in Figure 25.



Noise Topic	
natural habitat	0.000471423
tim	0.000471423
confluence	0.000471423
adelie penguins	0.000471423
38 14	0.000471423
board pietermaritzburg	0.000471423
9300	0.000471423
origins	0.000471423
locality	0.000471423
austin	0.000246936
28	0.000246936
aardvark orycteropus	0.000246936
president	0.000246936
12 17	0.000246936
staining	0.000246936
early morning	0.000246936

Figure 25: Example of a noise topic

A note on adding documents to an existing topic model: New documents can be added to an existing topic model using a technique called "document fold-in". In short, the various topics of the topic model are then used as categories and the new documents are then associated with a mixture of the topics based on the words such documents contain. The shortcoming of this approach however, is that words encountered in the new documents that are not part of the corpus vocabulary of the existing topic model, will be ignored. Also, new possible topics that may be contained in the new documents will not be formulated as result. Thus the existing topic model remains static apart from the new topic-document relationships that are added to the model to cater for the newly added documents. The author have not experimented with the document fold-in technique, but believe that it can be used for the more frequent updates of a topic model so long as it is not used as the only way to update the topic model when many new documents are added to the associated corpus.

This concludes the discussion around the various phases of the topic modelling process. The next section sheds some light on the subject of linking the results of different analyses.

6.6 Linking the Results of Different Topic Modelling Analyses

The author needed to identify similarities between different topics in the same analysis to find for any given topic, other topics that deal with closely related subjects. It was also required to find similarities between topics of different analyses for the following reasons:

- To build a topic hierarchy by linking related topics constructed from the same document collection, but in different analyses in terms of the number of topics formulated. For example, when performing a 10-topic, 50-topic and 100-topic analysis on the same document set, the 10-topic analysis tends to have more general topics when compared to the 50- and 100-topic analyses. A topic hierarchy can be constructed by identifying for each topic in the 10-topic analysis results, fairly similar topics in the 50-topic analysis results, and subsequently, for each topic in the 50-topic analysis results finding fairly similar topics in the 100-topic analysis results. The resulting network actually represents an automatically created taxonomy with descriptive words and associated documents for each class.
- To build a topic network by finding similarities between topics constructed in different analyses using different document collections. Since topic modelling is a fairly computationally expensive technique, it is not always feasible to analyse very large document collections (e.g. more than 10 000 research articles) in one analysis. It would be more practical to analyse average-sized document collections (e.g. a 1000 research articles) and after the fact finding similarities between topics in different analyses in effect creating a topic network. Such a topic network may continuously be grown by associating topics formulated by the analysis of new documentation to the topics in the existing topic network.

Ways to create topic hierarchies and topic networks were developed as part of the CIRP case study and is discussed in section 11.8.

6.7 Extending the Topic Model Results

The aspects of a standard topic model (e.g. LDA) include **topics**, **documents** and **words** (refer to section 6.5.4). It is however desirable to have more aspects in the topic model to be able to associate topics and documents to people, organisational units, time periods, to name a few. Although it is possible to extend the structure of the standard topic model (in terms of its latent variables) to explicitly contain more dimensions (e.g. authors, time, etc.) this requires a deep understanding of the mechanics of the model as well as the implementation thereof not to mention the adaption of the inference techniques to recover the additional latent variables. Such an explicitly extended topic model will in all likelihood be more specialised in its application compared to the standard topic model that has wide applicability due to its relative simplicity. Specialised topic models, derived of the standard topic model, exist but do not have the same level of general applicability of the standard topic model. The Author-Topic Model (refer to Appendix B, section 16.5) and the Dynamic Topic Model (refer to Appendix B, section 16.7) are examples of such models. For these reasons the author investigated expanding the topic model results with different aspects after the fact. The standard topic model calculates document-topic affinities and associated strengths. Documents have associated metadata, for example:

- Title

- Author(s)
- Reader(s)
- Publication Date
- Associated Organisational Unit(s)

It should be possible to link other fields (e.g. authors) to the aspects present in the standard topic model results to arrive at extended topic model results. For example, the author of a specific document may be stored in a document management system. If such a document was included in the document collection used to construct a topic model, the following associations are explicitly stored in the topic model results or can be derived from the results:

- Associations between the document and **topics**
- Associations between the document and other **documents**
- Associations between the document and **words**

Therefore, using the document as a link, topics and words associated with the author may be found based on the fact that he authored the document in question. This can be repeated for all documents he authored and the results may be aggregated or combined to create an estimated author knowledge profile. Other documents that may be relevant for the given author may also be calculated using the associations between documents he authored and other documents associated with these documents according to the topic model. This can form the foundation for a mechanism to provide the author with other potentially useful documents. Figure 26 illustrates the scenario where the standard topic model results have been expanded with **Person**, **Organisational Unit** and **Publication Date** dimensions. Dotted lines represent the implicit relationships that can be inferred from the data in the standard topic model and newly included metadata fields.

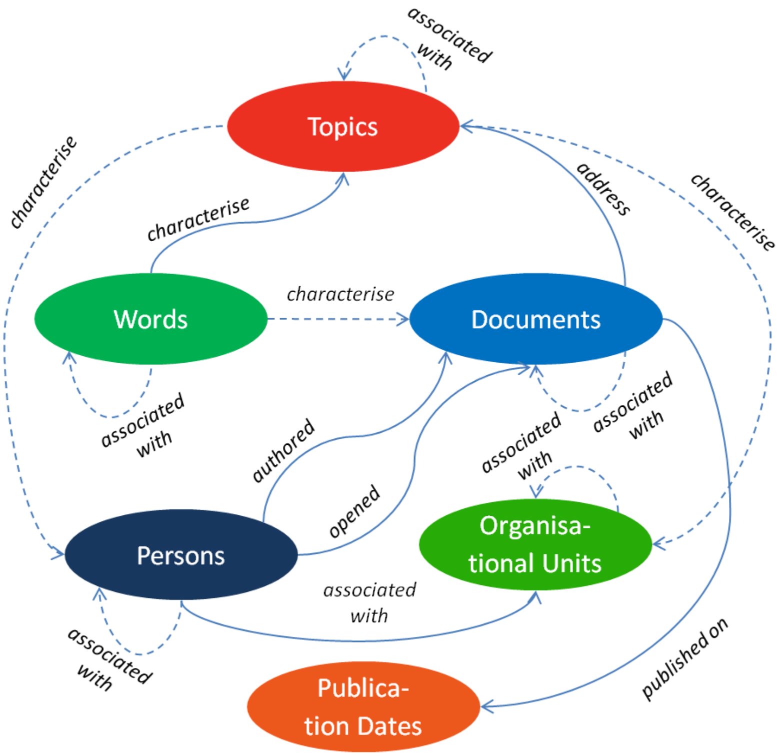


Figure 26: Conceptual framework illustrating possible aspects of an extended topic model

Similarly, when the creation dates of documents are known such information may be used to derive time trends for topics based on the documents strongly associated with them. Topic time trends may be used to estimate the currency of topics or the focus of a given organisational unit for a given time period.

The results of an investigation around the expansion of topic model results to include other dimensions are presented in the CIRP case study as presented in section 11.7.

6.8 Summary

In this chapter a family of methods, known as statistical topic models, was introduced for dealing with unstructured, textual data typically occurring in electronic documents. Topic models have the ability to derive interpretable structures underlying text without the need for training. Such structures, called topics, are driven by document content and can be used to obtain an overview of the information contained in a collection of documents as well as to group related documents together. Most topic models are comparable to (sophisticated) soft clustering techniques as documents may be associated with more than one topic (i.e. cluster).

Different topic models were compared and **Latent Dirichlet Allocation** (LDA) was selected as the preferred topic modelling technique to fulfil the clustering need for this research due to its efficiency and relative simplicity. The **Concept-Topic Model** was further identified to fulfil the classification need due to its unique capability to accept prior human knowledge and judgement captured in predefined concepts as input to guide model generation and therefore serve as an advanced classification mechanism. The topic modelling process, in the experience of the author, was further discussed from start to finish.

Information plays a key role in the innovation process, especially in the fuzzy front end. Organising information, with the goal to make it more accessible, using only manual methods may be accurate, but is too time-consuming to be practical and to make economic sense. Topic models may greatly assist innovation workers in understanding, organising and retrieving such information in a more effective way. Topic models are recommended as the means to bridge the gap between the organisational documents pertaining to innovation (discussed in Chapter 4) and the intended framework that will integrate, contextualise and make accessible such information to innovation workers (discussed in Chapter 8).

The topic model results may further be expanded or contextualised by other entities, their attributes and relationships that may exist in the organisation, but not necessarily included in the contents of electronic documents.

The next chapter introduces the concept of Business Metadata which deals with the meaning and context of information in the organisation.

7. Business Metadata

"The proper exploitation of business metadata, by both technologists and businesspeople, can revolutionize the use of technology, making it the facilitator of enterprise knowledge that it was always meant to be. Technology can lend assistance to how business metadata can enable business to be conducted on many levels, from facilitating communication and fostering true understanding to providing background information so that appropriate decisions can be made."
Inmon et al., (2008)

This chapter addresses the following:

- **Define the concept of business metadata**
- **Discuss the importance of business metadata**
- **Discuss how business metadata can assist in retrieving information**
- **Present the relationship between business metadata and semantics**
- **Present various mechanisms to capture business metadata**
- **Explain the applicability of business metadata to the innovation process**
- **Discuss the relationship between business metadata and knowledge management**

Metadata is generally defined as data about data. For example, the title, authors, publication date and keywords of an academic article may be considered as metadata of the article. Without metadata, data is fairly meaningless. For instance, one cannot interpret the data item "1/2" without metadata since it can be a date (day and month or month and day?), a cricket score (one run for two wickets or one wicket for two runs?), simply a fraction (one-half), or an imperial size in fractions of an inch equal to 13mm. Metadata therefore embodies meaning and enables understanding by providing context to data or information.

7.1 Types of Metadata in an Organisation

Metadata in an organisation can typically be classified as either **technical metadata** or **business metadata**. These two categories are actually the extremes of a gradual continuum of the focus of metadata in the organisation.

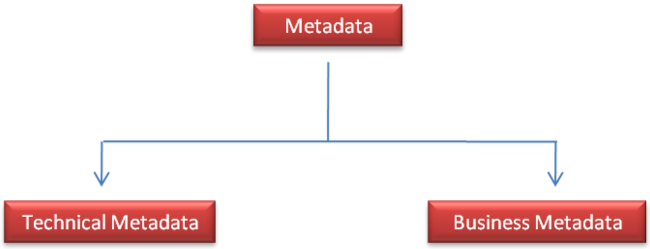


Figure 27: Types of metadata

An example of technical metadata is the data used by software developers or database administrators to characterise a database, such as:

- Table names (e.g. “Employees”, “Departments”, etc.)
- Table field names (e.g. “empl_name”, “empl_dateofbirth”, etc.)
- Field types (e.g. string, integer, date, etc.)
- Inter-table relationships (e.g. each employee has one and only one department, each department has one or many employees, etc.)

Business metadata, on the other hand, is metadata that is used by businesspersons to execute the day-to-day business activities. It provides the businessperson with the context and meaning of the data represented by the computer or contained in documents and enables the accurate use of data in the organisation. It is therefore imperative that business metadata be in the language of the businessperson. Business metadata is very valuable to the organisation because it facilitates understanding the data of the organisation. Business metadata can be found throughout an organisation, for example in:

- Spreadsheets
- Information system screens
- Reports
- Contracts
- Proposals
- Bank statements
- Manuals

Converse to technical metadata, business metadata is rarely explicitly documented and therefore are mostly scattered in unstructured data and specifically text throughout the organisation. Metadata may further be classified as either structured- or unstructured metadata. With **structured metadata** one can predict to a large extent where metadata will occur. For example, the various fields of an application form are examples of structured metadata. **Unstructured metadata** on the other hand is found at more unpredictable places and often appears as descriptive text. For example, unstructured metadata can be found scattered throughout the text of a contract document. When comparing a number of contracts, there would be little or no uniformity in the business metadata found the various contracts. (Inmon et al., 2008)

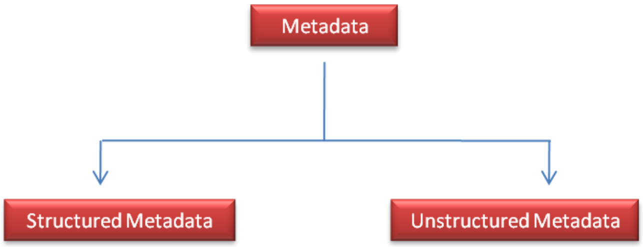


Figure 28: More types of metadata

Only a small fraction of business metadata is in the form of structured metadata; a much larger portion of metadata is of the unstructured type that is generally found in for example:

- E-mails
- Verbal communications between employees
- Organisational documents
- The minds of employees
- Workflow processes
- Business rules

Only when business data is abstracted to a broader organisational context, it becomes business metadata. Suppose “Peter Harper” works for a certain organisation. “Peter Harper” is not business metadata per se. However, when “Peter Harper” is referred to as an “employee”, then employees like “Peter Harper” may be considered as business metadata since “employee” gives context to data like “Peter Harper” and other employee names. Therefore, business metadata of the unstructured kind resides in text and exists in the form of abstractions. However, not all abstractions encountered in text are considered business metadata. A selective abstraction process is required to filter business metadata from unstructured information. Glossaries, taxonomies and ontology may further be developed to facilitate the identification of important abstractions from a collection of candidates retrieved from text. Statistical topic modelling, discussed in Chapter 6, is another alternative for creating abstractions of the content of a collection of unstructured text documents by extracting underlying themes and rendering candidate abstractions which may serve as business metadata items.

7.2 Typical Problems with Metadata in an Organisation

Inmon et al. (2008) presents two general problems with metadata in an organisation. Firstly, the fact that metadata occurs throughout the organisation and little coordination occurs between silos of metadata, may give rise to inconsistencies in metadata in different areas of the organisation. Therefore, a single metadata item may have different meanings in different organisational areas. Conversely, the same thing may be known by different names in different parts of the organisation. This problem is aggravated by the nature of language itself where phenomena like homonyms, homophones and synonyms create opportunities for ambiguity and misinterpretation.

The second problem with regards to metadata is gradual changes in the meaning of a given metadata item. Say for instance that people working on a contract basis for a given organisation was seen as “employees” (just as any other person working for the organisation) in year t and in year $t+1$ they were regarded as “contractors”; this will clearly befuddle any calculations regarding employees or contractors when the change in metadata definition and the exact point in time when the change occurred is not known and taken into account. Even worse, one part of the organisation may still apply the “employee” definition while another part may be using the new “contractor” definition. Inconsistencies in metadata may therefore lead to misunderstood data leading to errors in judgement made unknowingly. This especially applies to an innovation environment where the issues being investigated, implemented and refined are numerous and in constant flux in terms of meaning and applicability. Capturing and sharing innovation-related metadata in the organisation may therefore increase the effectiveness (i.e. doing the right things) and efficiency (i.e. doing things right) of the innovation process and will further facilitate communication between innovation workers from different backgrounds or parts of the organisation.

7.3 Consolidating Metadata in the Organisation

Just as organisations require data warehouses to create uniformity in data and to bring together data from various departments and divisions of the organisation for centralised analysis, the same requirement exists for metadata (Inmon et al., 2008). By gathering, integrating and localising metadata, it can more readily be rationalised across the organisation. With the absence of a metadata directory, the organisation tends to create everything from new each time a new information request is issued. Having a shared metadata directory improves the interpretability and reusability of information which can greatly improve the efficiency of gathering and using information. Business metadata of the organisation need to be consolidated for the following reasons:

- To capture the metadata from the organisation's employees[19]
- To aid employees in understanding an enterprise business process and the IT implementation of the process
- To treat metadata and information uniformly across the enterprise with the aim to increase consistency and quality
- To rationally plan the consolidation of enterprise systems
- To understand the various dimensions of the cross-functional organisation and the technical impacts of changes to be made to an organisational process or system

Most of these aspects can directly be applied to the innovation process and innovation-related information of an organisation. Not only does this imply that innovation-related metadata must be captured, but also that it has to be made accessible to the employees of the organisation once captured. A mechanism is therefore required to capture innovation-related metadata of the organisation and make it accessible to its innovation workers. The framework developed to address this issue is presented in Chapter 8.

[19] This corresponds to the Externalisation phase of the SECI model of knowledge creation where tacit knowledge is converted to explicit knowledge. Refer to section 19.2 in Appendix E for a more detailed discussion of the SECI model.

7.4 Characteristics of Well-formed Definitions

Business metadata essentially exists to add context to data. The meaning of business terms, acronyms, abbreviations and data elements need to be made explicit by capturing definitions (as part of business metadata) for such items in shared structures such as business glossaries, corporate dictionaries or corporate vocabularies. A standard for vocabulary management, ISO 11179, states basic guidelines of a well-formed definition whereas Inmon et al. (2008) present fourteen components of a well-formed definition. They further maintain that the text of a well-formed definition must make explicit at least two of the following three dimensions:

- **Broader term**: What is the class that the item being defined belongs to? This can be obtained by answering the question: *Item*$_x$ is a (kind of) what? For example, a "banana" is a kind of sub-tropical fruit.
- **Distinguishing characteristics**: How is the item being defined distinguished from other members of the class? This can be obtained by answering the question: *Item*$_x$ has a what? For example, a "banana" has a crescent shape**,** yellow colour, and soft, sweet flesh.
- **Function**: What is the item being defined used for? This is determined by the answer to the question: What is *item*$_x$ used for? For example, a "banana" is a high energy, natural snack eaten by humans and is further processed to a pulp used in several other foodstuffs.

Putting all three elements from these examples together into a single definition: "A banana is a kind of sub-tropical fruit, with a crescent shape, yellow colour and soft, sweet flesh and is a high energy snack eaten by humans and is further processed to a pulp used in several other foodstuffs." When the item being defined has more than one meaning (e.g. a homonym) the definition should state all possible meanings. It is further good practice to define all (non-trivial) terms used in definition text in the business glossary, dictionary or vocabulary. In the example above, the terms "sub-tropical fruit", "crescent", "pulp" and "foodstuffs" are candidates for being defined separately. Lastly, it is good practice not to define an item by itself (i.e. tautology). For instance, Inmon et al., (2008) warns against the following kind of definition: "Pulp is created when a substance is made into a pulp".

Innovation often deals with new technologies, techniques, materials or concepts and frequently involves stakeholders from different backgrounds, organisations or even different countries. Providing definitions for the terminology used when communicating about innovation-related issues in the organisation will facilitate a shared understanding and more effective communication between such stakeholders.

7.5 Business Metadata and Searching

When searching for information two kinds of business metadata are generally involved:

- The metadata structure common to all electronic documents (e.g. title, author, creation date, etc.)
- Tags, categories, topics, etc.

According to Inmon et al. (2008), one of the most fundamental ways to improve the likelihood of finding information located on the Internet, the corporate intranet and on personal computers is to improve the interpretability of filenames and web page names. For example, the filename “52-1-2003-9.pdf” says virtually nothing about the content of the file when compared to the filename “Assembly reliability evaluation method (2003).pdf”. Such meaningless filenames make files all but invisible on the corporate intranet. Document titles and even individual words in titles are valuable business metadata components. The title of web pages and the filenames of documents may be considered as the most important business metadata since it is in many cases the primary way information is found. Closely second in importance to finding information is how documents are classified. In other words, what is the topic of the file in question? And also, what topics are related to it? Having a classification system for documents would facilitate the process of retrieving not only a given document, but also other documents related to it.

One way of classifying organisational documents is by using a taxonomy, which may be described as a (hierarchical) scheme for organising items by means of their classification starting with the broadest or most general term at its root and having more narrow or specific terms as its branches. Taxonomy originated in the field of biology but is widely used in information retrieval applications to allow users to navigate the general to specific term structure to locate desired items and associated documents. A real strict or formal taxonomy (such as the taxonomies used in biology) only allows an item (e.g. a document) to be associated with a single taxonomy category. When dealing with the classification of documents this restriction is not realistic since a single document may be associated with more than one topic. For example, the document “Assembly reliability evaluation method (2003).pdf” may be associated with categories “Assembly Systems”, “Reliability” and perhaps “Evaluation Methods”. Associating more than one category to a document in many cases would increase the changes of more accurately describing the contents of the document and therefore the likelihood of finding the specific document using one or more of its associated categories. Therefore a clear distinction can be made between formal taxonomies (having strict rules) and business taxonomies which tend to be a bit more flexible. The purpose of a business taxonomy is to assist the user to navigate a website, document collection or corporate knowledge base to facilitate the ease of finding information by enabling them to discover information through browsing the taxonomy and viewing information in an intuitive and consistent manner (Wahl, 2006).

For effective business taxonomies it is important to use simple terminology and avoiding jargon to limit the risk of confusing potential users. The designers of such a taxonomy should identify the “lowest common denominator” of user types and construct the taxonomy using terms and topics that they can identify with.

Using taxonomies to classify documents and information however has the prerequisite that the different kinds of possible topics (corresponding to the taxonomy categories) have to be known and part of the taxonomy before the user can use it to classify a document. This implies the need that the business taxonomy must be kept up to date on a daily basis. Even then the scenario might arise where the categories to describe the contents of a document are not included in the business taxonomy at the time when the user wants to classify a document. Although taxonomies and classification in general are effective ways to organise information with the goal to make it more accessible, it should not be used as the only means to achieve this. Other techniques like clustering[20] where the categories are derived automatically from the content of the information to be organised, and specifically statistical topic modelling presented in Chapter 6, may be used in addition to classification to more dynamically organise information in environments characterised by diverse and rapidly changing information content.

7.6 Business Metadata and Semantics

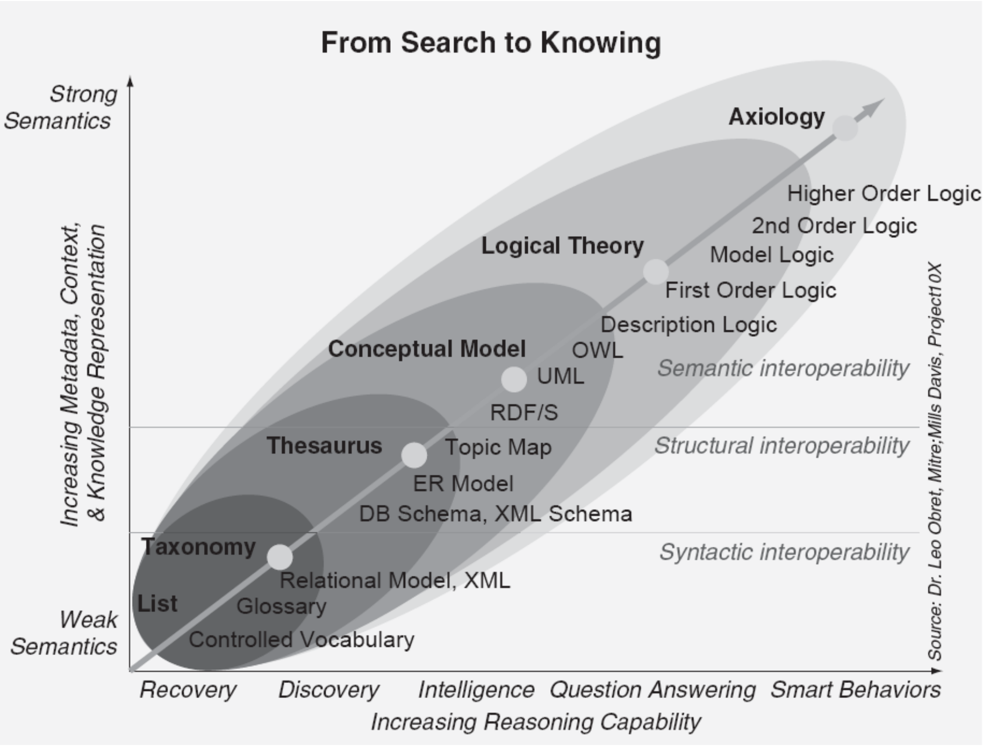
The word “semantics” was derived from ancient Greek philosophy and is concerned with the study of meaning (i.e. the meaning of words, phrases, sentences, or text in general). The business metadata (e.g. definition, assumptions, rules, examples, related concepts, etc.) describing a given business concept signifies the semantics of the concept. According to Inmon et al., (2008), two elements that embody the semantics of a typical organisation are the definitions of its business terms and its business rules. Semantics is closely tied to context - the meaning of a given term can change depending on which persons are using the term.

To achieve the vision of the semantic web, where the computer will ‘understand’ the meaning of a user’s query and ‘know’ all tasks associated to such a query, mechanisms for accurately capturing semantics and intelligent agents using such semantic representations for reasoning and executing appropriate actions are required. The requirement for searching for documents in the envisaged semantic web environment is that the computer must autonomously ‘understand’ the contents of documents (in addition to understanding the user’s search query) in order to provide the user with the best documents in response to his query. This implies that the computer must be able to decipher the indented sense of words (i.e. the intended meaning of a given word from the different possible meanings of the word) encountered in document text and users’ search queries. Word indices, document titles and associated tags are used to find documents. Word indices are automatically built by search engines[21], whereas tags are associated to documents by persons after the fact using social bookmarking systems such as *Delicious*[22].

Each information system in the organisation has an associated semantic set indicating what different data elements mean in the context of the specific system. Users just assume that what is

[20] See section 5.11 for the explanation of the differences between clustering and classification.
[21] See section 15.1 of Appendix A for more information on the tf-idf search scheme.
[22] Also see section 4.4.16 for more about social bookmarking.

called a ‘customer’ in one system signifies exactly the same thing in other systems since it has the same name – an understandable but potentially dangerous mistake. Human-computer interaction further complicates the issue of context – how can we be certain that the context of a user is the same as the context embedded in a given system? The discipline of (business) semantics is focused on capturing and sharing such definitions, assumptions and interdependencies and subsequently creating awareness about the contextual nature of data in general. Semantics is concerned with meaning and business metadata is concerned with supplementing data with meaning. Hence, any method that captures the semantics of data and that is able to show such meaning to a businessperson to clarify data, is providing users with business metadata. Different knowledge representation mechanisms or technologies have different associated levels of expressive power that is linked to their ability to capture semantics as indicated in Figure 29.



Source: Dr. Leo Obret, Mitre;Mills Davis, Project10X

Figure 29: Semantic ability and intelligence capabilities of different knowledge representation technologies[23]

As indicated in Figure 29, the following are popular knowledge representation mechanisms[24] which also involve capturing semantics (listed in order of increasing expressive power):

[23] Source: Davis (2006)

[24] See Appendix D for a more detailed discussion of these knowledge representation technologies.

- Controlled Vocabularies
- Dictionaries and Glossaries
- Taxonomy
- Thesauri
- Entity-Relationship Models
- Conceptual Models
- RDF and OWL
- Topic Maps and Concept Maps
- UML
- Ontology

If semantics are captured and represented accurately, meaning can be captured and used for the ultimate purpose of creating intelligence by means of reasoning capability. Unfortunately, the more powerful a technology is in terms of its semantic capability the less intuitive it becomes for humans due to complex representation languages and rules (Sammer, 2003). Similarly, in her discussion about the degree of formality of ontologies, Bullinger (2008) points out that the degree of formality (of an ontology) is positively correlated to the degree of machine processability and negatively correlated with the usefulness for (human-related) communication. This makes natural language ontology more suitable to facilitate communication in instances where automation is less important. She further states that an ontology intended for automated processing is more difficult to develop due to the high level of formality required to support computability.

According to Inmon et al., (2008) a shortcoming of semantic modelling technologies currently available is their lack of visual modelling capability. A combination approach, using both a dictionary and a concept map or an ER model, can be used to express business metadata in an organisation since the dictionary can cater for rigorous definitions, while the concept map caters for the graphical depiction of relationships. Once the challenge of representing OWL[25] in a form understandable to businesspersons is addressed, it will become an important part of business metadata expression of an organisation due to its capability of harnessing the reasoning capabilities of computers that are the building blocks for automatic reasoning and intelligent searches.

7.6.1 The Importance of Relationships

Definitions alone are not sufficient to capture and represent business metadata and semantics (McComb, 2004). In addition to definitions, relationships between entities are crucial to understanding and are therefore also part of business metadata. There are numerous ways to represent relationships. Taxonomy is well suited for representing vertical relationships between things, but hierarchies are not flexible enough to capture all relationships that exist in an organisation. More flexible mechanisms, catering for both vertical and horizontal relationships of

[25] “Ontology Web Language”; refer to section 18.7 in Appendix D for more information about OWL.

varying strengths are required to effectively capture and represent business metadata. The mechanisms located towards the higher end of the spectrum depicted in Figure 29 provide more richness of semantic expression. However, the current tools available to communicate the concepts embedded in such mechanisms to humans, and more so, to businesspeople are largely inadequate. Therefore, a non-technical individual normally finds it hard to understand the tools and languages that provide semantic richness.

The framework developed in this research should serve as a mechanism to capture and contextualise innovation-related information entities for use by businesspeople, more specifically, the innovation workers of the target organisation. The framework should describe such entities by means of their characterising attributes and relationships, both vertical and horizontal.

7.7 Knowledge Management and Business Metadata

Knowledge Management (KM) denotes a variety of practices applied by organisations to identify, create, represent and disseminate knowledge for the purposes of promoting knowledge reuse, creating awareness, and fostering learning throughout the organisation (Inmon et al., 2008; Sammar, 2003). KM initiatives are typically linked to organisational objectives and are meant to support the achievement of specific outcomes like increased performance, shared intelligence, higher levels of innovation and ultimately, competitive advantage. KM may be distinguished from organisational learning by its increased focus on the management of selected knowledge assets and the creation and maintenance of the ‘pipelines’ through which knowledge flows. Knowledge transfer, one facet of KM, traditionally manifested in organisations in the form of formal apprenticeship, peer discussions, professional training, mentoring programmes and corporate libraries, to name a few. With the dawn of the 21st century, technology was increasingly applied to the task of knowledge transfer in the form of knowledge bases, expert systems, and knowledge repositories.

Since business metadata is directly used and created by businesspersons, knowledge management can be regarded as a fundamental part of business metadata. Business metadata and KM does not merely cross, but are interlaced by definition as both have the central intention of providing value to the organisation by the identification, creation, representation and distribution of knowledge for the purposes of reusing, awareness creation, and learning. Moreover, business metadata makes out a portion of the organisational knowledge that KM is required to manage. KM can be distinguished from business metadata by the fact that KM focuses on the processes behind the creation and maintenance of knowledge, while business metadata captures the resulting artefacts from these processes. Therefore, KM is more process-orientated and business metadata is more data-oriented resulting in an interactive, symbiotic relationship. In partnership with business metadata, KM initiatives cultivate knowledge artefacts (a form of explicit knowledge), which is in turn stored as business metadata and made accessible to organisational users where and when it may be useful. It should be pointed out however, that not all business metadata originates from KM processes and conversely, KM does more than just creating business metadata.

As business economy is increasingly becoming centred on information, an organisation's success today greatly depends on its ability to manage information and apply corporate knowledge to attain the greatest benefit. Dixon (2000) points out that organisations are in need of ways to refrain from continually reinventing the wheel in the organisation. It is therefore important to take stock of what knowledge assets the organisation has, disseminate this knowledge and make it accessible to employees.

The framework developed in this research would essentially be a mechanism for capturing, relating and making accessible innovation-related business metadata of the target organisation.

7.7.1 Building the Corporate Knowledge Base

The Corporate Knowledge Base may be described as all things, pertaining to the business of an organisation, the organisation collectively knows. More specifically, according to Adelman and O'Neil (2007) it includes the following:

- Organisational knowledge
- Industry knowledge
- The ability to apply skills to complex situations
- Cognitive knowledge acquired from training and experience
- The system of understanding cause and effects
- The understanding of how the organisation functions
- Knowing how to prevent certain problems
- The knowledge of how to find information (knowing who knows what and where to look for what kinds of information)

The term "knowledge base" also has a more concrete meaning where it refers to a particular type of database for knowledge management used for the computer-aided collection, organisation and retrieval of knowledge. Only when the knowledge base of an organisation is transferred from the abstract form (i.e. where knowledge only exists in the minds of employees) to the concrete from (e.g. articulated knowledge artefacts stored in a database) it turns into business metadata. Business metadata is necessary to populate the corporate knowledge base, make it accessible and organising it so that it can be of use to employees.

According to Inmon et al. (2008) the following principles apply to building a Corporate Knowledge Base:

- A plan must exist for encouraging employees to share what they know
- Technologies that fits into the existing workflow of employees and that is easy to use must be incorporated into the knowledge base
- Technologies that make data accessible and reusable from other organisational software applications must be integrated into the knowledge base (e.g. accessing the corporate knowledge base from any application by pressing a keyboard shortcut).

When the Corporate Knowledge Base is represented in the form of a database or similar technology and becomes accessible, it becomes business metadata. Therefore business metadata and KM goes hand in hand in enabling effective growth and application of knowledge in an organisation. The reader is referred to section 8.2 for further discussion about the Corporate Knowledge Base. The framework developed in this research would in effect be a kind of innovation focused knowledge base.

7.7.2 Using Business Metadata to Support Knowledge Management

Business metadata essentially deals with the knowledge socialisation phenomenon[26] which enables it to be captured in a knowledge base as explicit knowledge, organised and made accessible for possible future application.

Business metadata can support KM in the following ways (Inmon et al., 2008):

- Alleviating knowledge capture by means of technologies (e.g. wikis, groupware, collaboration systems, etc.)
- Facilitating knowledge diffusion by means of technologies that enable relevant information to be accessed when and where it is needed
- Enabling the organisation of business metadata to improve the ease of finding information by means of categorisation schemes (e.g. controlled vocabularies, taxonomies, ontology, etc.).

The combination of business metadata and KM can be of great help to an organisation in capturing intellectual capital and institutional memory before it leaves the organisation along with ex-employees.

7.8 Business Metadata and Innovation

Business metadata provide context to the data and information of the organisation helping businesspeople to better understand, organise, retrieve, share and apply such data and information. Stakeholders and participants in the innovation process interact with data and information as well as generate data and information as part of their innovation-related activities (e.g. reviewing and compiling a shortlist of suitable technologies for a given product concept). Context to information is especially important to participants of the innovation process as innovation typically deals with a broad spectrum of fast changing concepts and issues from a number of different disciplines thereby increasing the difficulty for an individual to interpret, understand and communicate innovation-related data and information. The availability of quality business metadata will assist businesspeople (more specifically innovation workers) to better interact with information; more specifically innovation-related information.

[26] The knowledge socialisation phenomenon largely corresponds to the SECI model of knowledge creation presented in section 19.2 of Appendix E.

When combining the concentrated information generated by statistical topic models[27] when applied to unstructured information, with innovation-related business metadata (as a form of structured metadata), a knowledge base catering for both structured information and unstructured information can be realised.

7.9 Summary

This chapter presented the concept of business metadata and its importance to the organisation in the light of providing context to the organisation's data and information resources to facilitate the improved utilisation of such data and information. The relationship between business metadata and semantics was further discussed and the importance of well-formed definitions and relationships between business metadata entities were stressed in the light of providing context to data and information. The symbiotic relationship between Knowledge Management and business metadata was also addressed. The notion of a Corporate Knowledge Base and its tie to business metadata was further presented. Finally, this chapter discussed the applicability of business metadata to innovation and the innovation process.

In summary, the framework developed in this research will essentially be an innovated-focused, corporate knowledge base that will (once implemented in one or more computer system) facilitate the capturing, organising and making accessible data, information and business metadata pertaining to the organisation's innovation-related information entities with the goal to support the activities of its innovation workers. Lastly, this framework will serve as a mechanism to provide context to the information mined from the organisation's innovation-related documents and will therefore be a source of business metadata.

[27] Statistical topic models were discussed in the previous chapter.

8. Construction of the Framework to Support Innovation Processes

"...the worlds of structured data and unstructured data seem to operate in parallel universes that do not intersect. With few if any exceptions, there is no bridge or interface between the two worlds. However, if a bridge between the two worlds is finally created, the construction of entirely new kinds of systems will be possible." Inmon et al. (2008)

This chapter integrates various concepts introduced earlier into a framework to better organise and exploit information contained in electronic documents to support the innovation activities of an organisation, especially the front end innovation activities. At this stage it is important to point out that the developed framework is not an ICT system in itself, but rather represents a generic model or 'blueprint' that may be implemented in one or a combination of several (existing or custom developed) systems. Different organisations may implement this framework in different ways as influenced by their unique needs, existing processes and ICT systems. The intended framework is best described as a foundation for an innovation-focused information system that encapsulates information about key innovation-related information entities and their respective inter-relationships. In light of the previous chapter, which dealt with business metadata, the framework may also be seen as a mechanism for capturing and disseminating innovation-related business metadata, and information, sourced from various existing information systems and the employees of the organisation in addition to information and metadata mined from organisational documents. The framework should therefore host abstract (i.e. types of entities along with descriptions and inter-relationships; e.g. "project" and "competitor") and specific (i.e. instances of entity types; e.g. "Buro 2010" and "Unilever") innovation-related information.

This chapter addresses the following:

- **The need for integrating structured and unstructured organisational information**
- **Introducing the Corporate Knowledge Base**
- **Positioning the framework in context of the organisation**
- **Presenting typical use cases of the framework in the organisation**
- **Discuss some requirements for using electronic documents and the framework in the target organisation**
- **Define the information entities of the framework**
- **Explain the framework in relation to typical organisational information systems**
- **Discuss various mechanisms associated with the framework**
- **Discuss different working methods involving the use of the framework**
- **Demonstrate how statistical topic modelling techniques may benefit the framework**
- **Present some functionalities desired in the associated text analytics system**

8.1 Integrating Unstructured and Structured Information

As indicated in the quote at the start of this chapter, Inmon et al. (2008) states that structured data (information) and unstructured data (information) tend to function separately in most organisations as there is mostly no interface between these two realms, hampering integration. When this shortcoming is adequately addressed a completely new type of information system would be possible. Inmon et al. (2008) further posits that in order for unstructured data (information) to be used in the structured environment, it has to undergo an integration process that entails a systematic determination of the semantic and linguistic substance of the text. In this research, statistical topic models (refer to Chapter 6) are utilised to facilitate this integration process. Once text has been distilled and analysed, this concentrated output has to be harmonised with structured information in order to create an integrated information system solution. The framework developed serves as an encapsulation mechanism of such structured information, while textual documents will be the source of unstructured information. More specifically, the framework contextualises the distilled content of textual documents to better address the information requirements of the target organisation’s innovation processes and innovation workers.

8.2 The Corporate Knowledge Base

The term “Corporate Knowledge Base”, as discussed in section 7.7.1, refers to virtually everything the organisation collectively knows about its business and related issues. Such knowledge may be contained in databases, but mostly it resides in the minds of employees and the content of organisational documents. According to Sammer (2003), a (collective) knowledge base has two important parts, namely the individual knowledge of the members of the organisation and the framework that connects them. In addition to this, interaction and communication structures further play a critical role. The bulk of explicit corporate knowledge is normally embedded in unstructured information in the form of electronic documents[28] (e.g. e-mails, word processing documents, Acrobat files, etc.). However, since a substantial portion of corporate knowledge is not made explicit (e.g. documented or recorded) it only exists in tacit form in the minds of the employees of the organisation.

Employees often have ideas about how to do their job better, but mostly such ideas are lost since there is no mechanism in place to easily capture them. Inmon et al. (2008) posit that individual knowledge capture has received limited attention to date in traditional knowledge management. The lack of a formal mechanism to capture ideas and insights on an individual level is aggravated by the fact that people generally are poor documenters because the benefits of such additional efforts are not always visible or understood. This task of documenting ideas and insights need to be done by the person that had the idea or insight in order to capture its true essence and context. The same

[28] Bergman (2005) states that as much as 90% of corporate memory exists in documents.

applies to business metadata – it must be directly captured from businesspeople and in the language of business people. Since the corporate knowledge base is built from contributions from individual members of the organisation a wide range of subjects will naturally result. Sammer (2003) recommends that the knowledge base should be organised in different knowledge domains to accommodate such diversity.

The concepts 'team memory', 'project memory', 'community memory', 'corporate memory' and 'organisational memory' are closely related to the concept of a Corporate Knowledge Base and are used widely to refer to the memories (and implicitly the expertise of) the organisation's individual employees plus technological extensions in the form of databases and knowledge bases (Shum, 2006). The purpose of a corporate memory, apart from being an information system, is that it should assist in converting information into action. According to Bullinger (2008) ontologies (see section 7.6) represent a possible means of constructing corporate memories.

In recent years, many new ways to socialise knowledge were introduced by advances in technology (phenomena like wikis, blogs, social bookmarking, social networking, etc., did not exist ten years ago for instance). In most organisations knowledge socialisation happens in an informal, uncontrolled way and the results of such socialisation are usually not captured for further expansion or future reuse (Inmon et al., 2008). For example, when two employees discuss a problem and its possible solutions during a coffee break, the resulting knowledge and troubleshooting process is not captured and therefore cannot be reused when the same problem occurs in a year's time. A more effective approach would be to 'socialise' the problem by creating an entry on the organisation's wiki (i.e. a form of a Corporate Knowledge Base) to allow future trouble-shooters to retrace the steps taken to solve the issue and understand the rationale behind each step. This would not only preserve and share the solution to the issue (in other words making it accessible to a larger audience) but will also provide the context of the issue and its solution.

8.3 Preliminary Requirements of the Framework

In order to make the envisaged framework more concrete, this section presents some initial requirements of the framework:

- The framework should be able to provide structure to information it embodies to facilitate searching and browsing for appropriate information.
- The framework should include and define the major categories of things important to innovation for association to different information entities. This is required to provide the different views required to facilitate the finding of pertinent information.
- It should be able to capture abstract and specific innovation-related information.
- It should not be purely hierarchical in its structure. It should rather be able to cater for horizontal and vertical relationships between the different information entities and entity types it embodies.
- Closely related to the previous requirement, it should be able to capture the relationships between different information entities to represent the true dynamic nature of innovation-related information.

- It should allow for capturing the innovation-related terminology of the target organisation and associated definitions as part of the innovation-related information it would embody.
- It should allow for only one entity type to be associated with a given information entity to aid inference.
- It should be able to accommodate links to information residing in other information stores of the organisation. In this case the relevant framework information entity instance should contain the reference to the appropriate information in the hosting system.
- It should allow for the addition, manipulation and interpretation of the information it embodies by software systems.
- It should also allow for capturing human input, with regards to the innovation-related information entities of the organisation and their inter-relationships, using suitable software system interfaces.
- The framework should be extendible to include new entity types and defining additional characteristics of existing entity types as the need for such extensions arise.
- The technology ultimately selected to embody the framework should be scalable to accommodate the core innovation-related information of small, medium and large organisations.
- The technology ultimately selected to embody the framework should allow for automatic inference to generate new information from the content of the framework.
- Relationships between different entity instances in the framework should not only be associated with binary relationships. It should rather be possible to capture the relative strengths of such relationships.
- Instances of information existing in the framework must be uniquely identifiable.
- No duplication should be allowed in terms of identical instances of information in the framework.
- The content of the framework should be stored in suitable formats to facilitate exporting such content to the mainstream formats (e.g. XML, RDF, OWL).

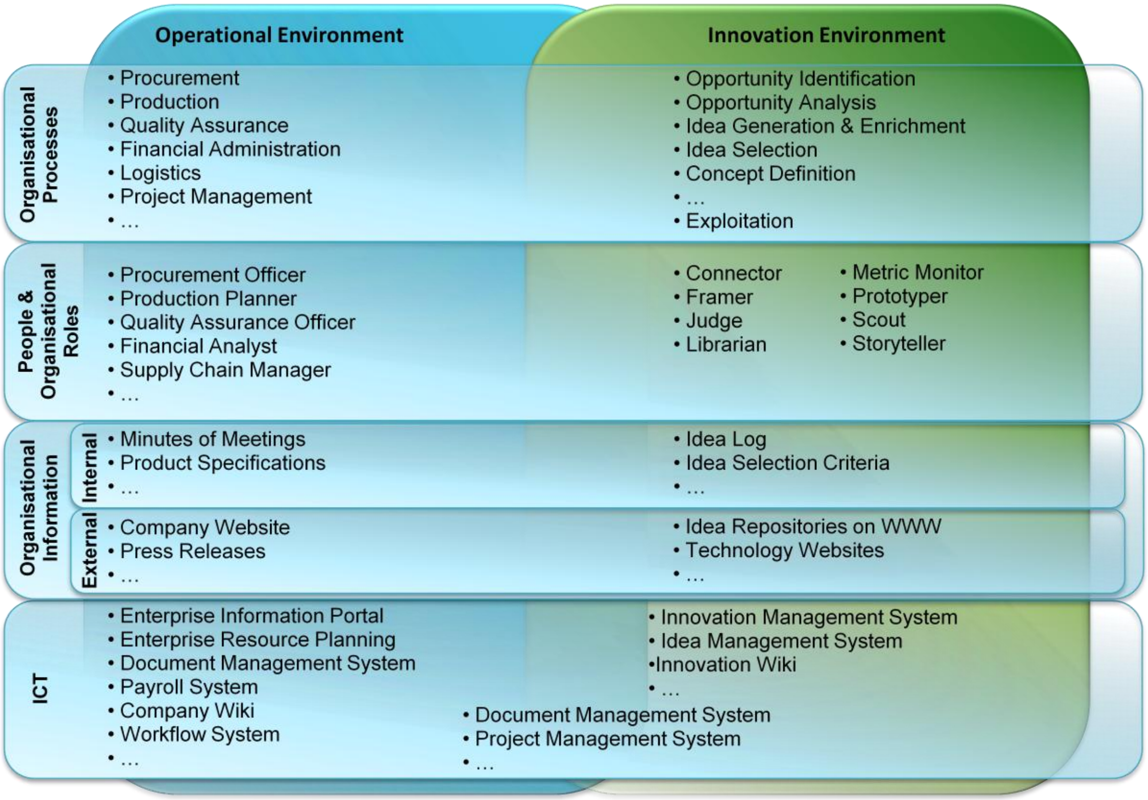
The list of requirements above is by no means exhaustive. As the framework and associated mechanisms mature in future, new requirements would necessarily arise and be supplemented to this list.

8.4 Positioning of the Framework in the Organisation

Few organisations exist with the sole purpose to innovate and no organisation can continue its operations in a sustainable way without innovating; a blend between operational and innovation activities therefore exists in most organisations. The framework touches on different dimensions of the organisation, but focuses on those dimensions pertaining to innovation. Some dimensions of an organisation that are important to the framework, also illustrated in Figure 30, are:

- Organisational processes
- People and their organisational roles
- Organisational information
- Information and communication technologies (ICT)

In any organisation a number **processes** are executed as part of the normal functioning of the organisation. Some processes are formal or explicit (e.g. the procurement process) while others are more informal or implicit (e.g. capturing the lessons learnt after the completion of a project). The framework will focus more on the innovation processes of the organisation, with a strong focus on the fuzzy front end innovation processes discussed in section 3.4. Although some processes can be automated, most organisational processes involve people at one or more stage of execution. The way in which **people** take part in a given process is determined by the role of the individual. Roles therefore may be regarded as part of the people dimension. The innovation front end processes in particular are very conceptual in nature and therefore cannot happen without humans as it requires the tacit knowledge and creativity of humans to create the desired outputs. The people dimension, with special focus on persons and roles involved in innovation, therefore have to be included as part of the framework.



		Operational Environment		Innovation Environment
Organisational Processes		• Procurement • Production • Quality Assurance • Financial Administration • Logistics • Project Management • …		• Opportunity Identification • Opportunity Analysis • Idea Generation & Enrichment • Idea Selection • Concept Definition • … • Exploitation
People & Organisational Roles		• Procurement Officer • Production Planner • Quality Assurance Officer • Financial Analyst • Supply Chain Manager • …		• Connector • Framer • Judge • Librarian • Metric Monitor • Prototyper • Scout • Storyteller
Organisational Information	Internal	• Minutes of Meetings • Product Specifications • …		• Idea Log • Idea Selection Criteria • …
	External	• Company Website • Press Releases • …		• Idea Repositories on WWW • Technology Websites • …
ICT		• Enterprise Information Portal • Enterprise Resource Planning • Document Management System • Payroll System • Company Wiki • Workflow System • …	• Document Management System • Project Management System • …	• Innovation Management System • Idea Management System •Innovation Wiki • …

Figure 30: Operational and innovation aspects of an organisation

One of the primary purposes of the framework is to more effectively disseminate organisational information in order to promote innovation activities; therefore it goes without saying that the **organisational information** dimension have to be included as part of the framework. The specific focus for this research however is on information contained in electronic organisational documents. Another important dimension of the organisation that concerns the framework is the organisation's **information and communication technology** (ICT). An organisation's ICT infrastructure enables its operations, but also innovation activities to a lesser extent. Information systems, which are part of the organisation's ICT infrastructure, hosts important information about innovation-related entities and ultimately need to be linked to the framework to ensure that the framework relays accurate information. One or more information systems are further required to implement the framework.

8.5 Use Cases of the Framework

Knowledge of the theory behind the framework alone will not contribute to the increased effectiveness of the innovation processes of an organisation; the framework has to be adapted, implemented and used in an organisation to bear fruit. The following are some possible high-level use cases of the intended framework in a target organisation:

- Serve as a corporate knowledge base of the target organisation with regard to actual entities related to innovation and their inter-relationships. Humans can access and exploit this knowledge base to gain contextualised information to support their innovation-related activities by interacting with the framework by means of one or more information system.
- Provide the means to structure unstructured information contained in electronic documents making it more accessible in context of existing innovation-related information entities.
- Aggregate the essence of innovation-related information (structured and unstructured) contained in disparate information sources across the organisation.
- Support communication about innovation-related issues between different individuals, possibly from different business units or functional areas, and between the organisation and external collaborators or innovation stakeholders (e.g. lead users, collaborating suppliers, etc.)
- Assist with decision-making concerning innovation-related matters by providing contextualised information about affected innovation-related entities.
- Serve as a tool to facilitate organisational understanding pertaining to innovation-related issues and entities.

8.6 Requirements for Using Electronic Documents and the Framework

The following is a partial list of requirements for implementing the intended framework in an organisation to better exploit innovation-related information:

- Availability of addressable, electronic document repositories accessible to text analytics tools.
- Continuous gathering and storing of information pertaining to the innovation-related information realms discussed in section 3.8.
- Descriptive filenames for all electronic documentation gathered to provide an indication of the file’s contents at a glance are preferred.
- Electronic documents should be in formats suitable for automated text extraction.
- Programmatic access to one or more databases containing employee information (e.g. name, contact details, business unit, etc.).
- Associations of people (e.g. employees, associates, authors, clients, etc.) to documents need to be captured (e.g. as part of the metadata of documents in a document management system) and programmatically accessible.
- Associations of organisational units (e.g. departments, teams, etc.) to documents need to be captured and programmatically accessible.
- The date of publication or last update of electronic documents needs to be programmatically accessible.
- A business vocabulary containing definitions of terminology and acronyms is required.
- An innovation management system facilitating the capture and management of opportunities, ideas and concepts and their respective inter-relationships as well as their respective links to people, documents and business units should be available and programmatically accessible.
- One or more text analytics systems are required to harvest and distil information from text.
- One or more information systems suitable to host and maintain innovation entities of the organisation and their respective relationships are required. This system should be customisable to accommodate new entities as well as to alter existing entities. These

systems should further be able to accommodate named relationships between any entities defined in the framework. As part of this system, a mechanism is further required to store and integrate information received from the text analytics system and other relevant information systems of the organisation. Such information systems should also have suitable interfaces to source data from and supply information to other systems.

In most organisations only some of these requirements will be met, however, the more requirements are satisfied, the better the context of organisational information captured and the more complete the support the framework can offer to innovation activities of the organisation.

8.7 Information Entities of the Framework

As mentioned earlier in this chapter, the framework can be described as the foundation of an innovation-focused information system that hosts and manages information about key innovation-related entities and their respective inter-relationships. This section will describe some typical innovation-related information entity types[29] that may be used as a starting point when implementing the framework in an organisation. Some may question the need for distinguishing between different types of innovation-related information. It is necessary to clearly classify innovation-information according to the respective entity types they deal with for the following reasons:

- To enable the grouping of related entities according to their entity type.
- To be able to compile very specific queries with the aim to find the answer to a very specific question, e.g. “Which documents are associated with this specific innovation idea?” or “Which persons in the organisation are associated with this specific technology?” This would only be possible if a distinction is made between the different types of entities.
- The types of information that one would like to store will not be the same for all types of entities.
- The types of operations possible given a specific entity type may differ from entity type to entity type.

Each entity type in the framework should be characterised by attributes and relationships to other entity types in the framework. As part of this research the following innovation-related information entity types were identified (listed in alphabetical order):

- Client/Customer/Consumer
- Collaborator/Distributor
- Competitor
- Concept
- Distribution Channel
- Document
- Idea
- Knowledge Area
- Market

[29] A note on the terminology used in this chapter from here and on: the term “entity type” shall refer to the class of an entity or thing (i.e. the kind of thing of a given entity). The term “entity” shall refer to the actual thing. For example, “Pat Harper” is an entity and the corresponding entity type of this entity is “Employee”.

- Need
- Objective
- Opportunity
- Organisational Unit
- Person
- Process/Activity
- Product/Service
- Project/Programme
- Skill/Competency
- Supplier
- Technology/Method/Tool/Equipment

In order to incorporate the distilled information of electronic documents generated by statistical topic models, the following additional entity types are included in the framework:

- Topic[30]
- Word[31]

To capture and make explicit the formal terminology of the organisation the following entity type is included in the framework:

- Business Term

The intention of this framework is to be generic to any organisation executing innovation in any industry. The framework can be implemented in a given organisation by adapting it to the specific innovation environment of the organisation in question. Specific entities of non-core entity types (i.e. entity types not explicitly included in the framework) may however be added to the framework as entities corresponding to the Business Terms entity type and may be associated with other entities by means of manually created relationships. A description of the different framework entity types follows. Note that even though not explicitly mentioned in the definition of all entity types, the relationships between each entity type and entity types Topic and Document are fundamentally part of the definition of all entity types.

Firstly, the **Business Term** entity type will be used to capture the meanings of important terminology used in an organisation to improve consistency in communication, facilitate a common understanding throughout the organisation and expedite organisational understanding. For each Business Term entity, the following attributes are captured:

- Unique business term identifier
- Possible meanings of the business term
- For each meaning:
    - Date of definition

[30] The “Topic” entity type shall represent entities corresponding to themes generated by statistical topic models.

[31] The “Word” entity type shall represent entities corresponding to the words that describe topics in statistical topic models. This entity type is not limited to entities represented by unigrams (e.g. “laser”), but may accommodate entities represented by n-grams as well (e.g. “selective laser sintering”).

    - Synonyms
    - Part of relationships
    - Example sentence
- Possible links to:
    - Other Business Terms
    - Other entity types

The collection of captured Business Term entities actually forms a corporate vocabulary or business glossary that will assist in promoting shared understanding and unambiguous communication in the target organisation.

The **Client/Customer/Consumer** entity type represents the direct and indirect users of the products and services of the organisation in question and may be internal or external to the organisation. Instances of this entity type may be associated with zero, one or more markets. The following attributes characterise the Client/Customer/Consumer entity type:

- Unique Client/Customer/Consumer identifier
- Client/Customer/Consumer name
- Client/Customer/Consumer description
- Client/Customer/Consumer location
- Client/Customer/Consumer keywords
- Client/Customer/Consumer type (individual, organisation, etc.)
- Status of relationship with Client/Customer/Consumer (current, potential, previous, etc.)
- Client/Customer/Consumer adopter status (innovator, early adopter, early majority, late majority, or laggard)
- Date of first association of Client/Customer/Consumer with organisation
- Attractiveness of Client/Customer/Consumer (e.g. low, medium, high)
- Level of risk associated with Client/Customer/Consumer (e.g. low, medium, high)
- Association of Client/Customer/Consumer with own products/services
- Association of Client/Customer/Consumer with competitors and associated products/services
- Needs associated with Client/Customer/Consumer
- Opportunities associated with Client/Customer/Consumer
- Ideas associated with Client/Customer/Consumer
- Concepts associated with Client/Customer/Consumer
- Distribution channels associated with Client/Customer/Consumer (e.g. mail order, internet, retailers)
- Possible links to:
    - Other Clients/Customers/Consumers
    - Other entity types

The **Collaborator/Distributor** entity type captures external persons or organisations working with the target organisation to achieve certain mutual goals (e.g. refining an innovation concept, building a product prototype or selling the target organisation's products and services). With regard to innovation initiatives, collaborators and potential distributors may be engaged as early as the idea generation or opportunity identification stages of the innovation lifecycle or later in the concept refinement and prototyping stages. The Collaborator/Distributor entity type is characterised by the following attributes:

- Unique Collaborator/Distributor identifier

- Collaborator/Distributor name
- Collaborator/Distributor description
- Collaborator/Distributor location
- Collaborator/Distributor keywords
- Collaborator/Distributor type (individual, organisation, etc.)
- Collaborator/Distributor maturity (e.g. low, medium, high, etc.)
- Collaborator/Distributor engagement date
- Type of relationship with Collaborator/Distributor (current, potential, previous, etc.)
- Knowledge Areas associated with Collaborator/Distributor
- Skills/Competencies associated with Collaborator/Distributor
- Technologies/Methods/Tools/Equipment associated with Collaborator/Distributor
- Ease of replacing the Collaborator/Distributor (e.g. low, medium, high)
- Own products/services affected by Collaborator/Distributor
- Needs fulfilled by Collaborator/Distributor
- Opportunities affected by Collaborator/Distributor
- Ideas affected by Collaborator/Distributor
- Concepts affected by Collaborator/Distributor
- Possible links to:
  - Other Collaborators/Distributors
  - Other entity types

The **Competitor** entity type will refer to an organisation that offers reasonably similar products or services to the same market sector as the target organisation. Competitors can be the direct source of ideas or may inspire new ideas to increase one’s competitive advantage in relation to one’s competitors. The following attributes describe the Competitor entity type:

- Unique Competitor identifier
- Competitor name
- Competitor description
- Competitor location
- Competitor keywords
- Competitor maturity (e.g. low, medium, high, etc.)
- Competitor inception date
- Status of relationship with Competitor (current, potential, previous, etc.)
- Level of threat associated with Competitor (e.g. low, medium, high)
- Competing products/services and link to own products/services
- Knowledge areas associated with Competitor
- Technologies/Methods/Tools/Equipment associated with Competitor
- Markets associated with Competitor
- Possible links to:
  - Other Competitors
  - Other entity types

The **Concept** entity type shall describe innovation concepts created from one or more idea that passed the idea filter stage. Concepts are evaluated in terms of financial feasibility, market acceptance, and practicality to name a few. The following attributes define Concept entities:

- Unique Concept identifier
- Concept name
- Concept description
- Concept owner
- Concept contributors

- Concept keywords
- Concept status (e.g. promoted to project, rejected, under investigation, on hold, assigned, incomplete, ready for evaluation, etc.)
- Concept creation date
- Concept evaluation date
- Concept evaluation remarks
- Level of risk associated with Concept (e.g. low, medium, high)
- Concept complexity level (e.g. low, medium, high)
- Level of potential benefit of Concept (e.g. low, medium, high)
- Concept priority (e.g. low, medium, high, urgent)
- Market(s) associated with the Concept
- Technologies/Methods/Tools/Equipment associated with Concept
- Knowledge Areas associated with Concept
- Ideas associated with Concept
- Projects/Programmes associated with Concept
- Possible links to:
  - Other Concepts
  - Other entity types

The **Distribution Channel** entity type represents entities corresponding to ways or methods of selling an organisation's product products or services. Examples include direct selling, third parties (e.g. retailers), Internet commerce websites, direct mail, customer self-service, etc. The following attributes characterise the Distribution Channel entities:

- Unique Distribution Channel identifier
- Distribution Channel name
- Distribution Channel description
- Distribution Channel keywords
- Status of Distribution Channel (e.g. existing, planned, existed previously, etc.)
- Distribution Channel lifecycle maturity level (e.g. immature, maturing, mature, decline, phase out, etc.)
- Attractiveness of Distribution Channel (e.g. low, medium, high)
- Level of risk associated with Distribution Channel (e.g. low, medium, high)
- Products/services association with Distribution Channel
- Competitors associated with Distribution Channel
- Needs associated with Distribution Channel
- Opportunities associated with Distribution Channel
- Ideas associated with Distribution Channel
- Concepts associated with Distribution Channel
- Collaborators/Distributors associated with Distribution Channel
- Markets associated with Distribution Channel
- Possible links to:
  - Other Distribution Channels
  - Other entity types

At minimum an entity corresponding to the **Document** entity type is a piece of text (i.e. a grouping of words) created as a single logical unit (e.g. a report, an article in Wikipedia, an electronic presentation, a book chapter or an entire book, etc.). Documents contain information about one or more topic and may describe one or more entities. The following attributes define the Document entity type:

- Unique Document identifier
- Document filename
- Document file type (e.g. word processor file, spreadsheet, presentation, design drawing, process diagram, etc.)
- Document title
- Document authors
- Document creation date
- Date on which Document was last edited
- Date on which Document was last opened
- Document abstract
- Document keywords
- Document status (in process, draft, completed, stale, etc.)
- Document rating (e.g. poor, average, high, brilliant)
- Topics associated with given document
- Possible links to:
    - Other Documents
    - Other entity types

An **Idea** entity type will represent innovation ideas conceived by a person. One or more idea may be expanded to a concept. The basic set of attributes required to characterise an entity of entity type Idea is as follows:

- Unique Idea identifier
- Idea name
- Idea description
- Idea origin (internal or external)
- Idea creator
- Idea contributors
- Idea keywords
- Idea status (e.g. promoted to concept, rejected, under investigation, on hold, assigned, incomplete, ready for evaluation, etc.)
- Idea creation date
- Idea evaluation date
- Idea evaluation remarks
- Level of risk associated with Idea (e.g. low, medium, high)
- Idea complexity level (e.g. low, medium, high)
- Level of potential benefit of Idea (e.g. low, medium, high)
- Idea priority (e.g. low, medium, high, urgent)
- Needs associated with Idea
- Opportunities associated with Idea
- Market(s) associated with the Idea
- Knowledge Areas associated with Idea
- Links to:
    - Other Ideas
    - Other entity types

The **Knowledge Area** entity type represents the (loosely demarcated) domains of knowledge that are required for the successful execution of innovation in the target organisation (e.g. rapid prototyping, design for reusability, project management, etc.). The Knowledge Area entity type describe “the categories of what a person knows something about” while Skill/Competency entity type deals more with the intangible tools an individual has to acquired over time to apply such knowledge. The following attributes describe the Knowledge Area entity type:

- Unique identifier for Knowledge Area
- Name of Knowledge Area
- Description of Knowledge Area
- Parent Knowledge Area
- Organisational unit(s) associated with Knowledge Area
- Keywords associated with Knowledge Area
- Innovation roles (e.g. connector, scout, prototyper, librarian, etc.) associated with Knowledge Area
- Persons associated with Knowledge Area
- Qualifications associated with Knowledge Area (e.g. M.Eng. in Industrial Engineering, B.Comm. in Marketing, B.Sc. in Computer Science, National Diploma in Graphical Design, etc.)
- Possible links to:
  - Other Knowledge Areas
  - Other entity types

The **Market** entity type denotes a specified category of potential buyers for the existing or envisaged products or services of a given organisation. One or more markets may be associated with an innovation initiative as early as the opportunity identification or idea generation stage, but is a crucial component of an innovation concept. The following attributes characterise the Market entity type:

- Unique Market identifier
- Market name
- Market description
- Market location
- Market keywords
- Market lifecycle maturity level (e.g. immature, maturing, mature, decline, phase out, etc.)
- Market inception date
- Level of engagement of target organisation with this Market (e.g. low, medium, high)
- Attractiveness of Market (e.g. low, medium, high)
- Level of risk associated with Market (e.g. low, medium, high)
- Association of Market with own products/services
- Association of Market with competitors and associated products/services
- Needs associated with Market
- Opportunities associated with Market
- Ideas associated with Market
- Concepts associated with Market
- Collaborators/Distributors associated with Market (e.g. lead users)
- Distribution channels associated with Market (e.g. mail order, internet, retailers)
- Possible links to:
  - Other Markets
  - Other entity types

The **Need** entity type serves to capture and structure anything pertaining to innovation in the organisation that is necessary but lacking. Objectives may be formulated to address certain needs, needs may be evaluated and formulated for a specific market. In essence, needs drive innovation. The Need entity type is characterised by the following attributes:

- Unique identifier for the Need
- Name of the Need
- Keywords associated with the Need
- Need identification date
- Need status (e.g. foreseen, fulfilled, unfulfilled, partially fulfilled, etc.)
- Persons associated with the Need
- Markets associated with the Need
- Products/Services associated with the Need
- Organisational Unit(s) associated with the Need
- Objectives associated with the Need
- Opportunities associated with the Need
- Ideas associated with the Need
- Concepts associated with the Need
- Projects associated with the Need
- Technology/Method/Tool/Equipment associated with the Need
- Possible links to:
  - Other Needs
  - Other entity types

The entity type **Objective** serves as a placeholder in the framework for (formally) defined goals describing an intended outcome pertaining to the organisation in question. The scope of such objectives may be a given project, department, or the organisation in its entirety. As a specific form of an objective, strategic objectives are specific goals at the level of a specific organisational unit (e.g. increase production uptime with 20% in next six months) or at the level of the entire organisation (e.g. grow target market with 10% during next financial year). Objectives may give rise to innovation opportunities and may trigger the conception of ideas. The Objective entity type is described by the following attributes:

- Unique Objective identifier
- Objective name
- Objective description
- Objective level (e.g. organisational unit level or organisation-wide level)
- Organisational units associated with Objective
- Objective keywords
- Objective type (e.g. cost reduction, income growth, increased quality, etc.)
- Objective status (e.g. planned, in force, achieved, not achieved, etc.)
- Parent of Objective
- Objective introduction date
- Objective validity period (e.g. start and end date)
- Key performance indicators associated with given Objective (e.g. lead time, production cost, market penetration, etc.)
- Level of risk associated with Objective (e.g. low, medium, high)
- Objective complexity level (e.g. low, medium, high)
- Level of potential benefit associated with Objective (e.g. low, medium, high)
- Needs associated with Objective
- Opportunities associated with Objective
- Ideas associated with Objective
- Concepts associated with Objective
- Possible links to:
  - Other Objectives

  - Other entity types

Entities corresponding to the **Opportunity** entity type can be defined as a possibility due to a favourable combination of circumstances or a change in the organisation's internal or external environment that may result in some kind of benefit. Opportunity identification concerns the searching, exploration, and problem identification activities that a organisation uses to identify new opportunities that are attractive to the organisation, thinking about what it might mean for existing and future products, services, processes, and business models. The opportunity identification process is a precursor to doing market research or creating a business case. It is a process that identifies needs and desires, possibilities, problems, limitations, bottlenecks or constraints that are the source of new concepts and ideas for new business development activities. Opportunity identification is one of the earliest innovation front end activities typically undertaken once the organisation's innovation strategy has been decided. Opportunities are typically identified by a person (internal or external) or an organisational unit (e.g. team or department). Opportunities can precede ideas or can be uncovered once an idea is identified. An opportunity may also be linked to zero, one or more needs; readily exploitable opportunities should address one or more needs. The following attributes characterise the Opportunity entity type:

- Unique Opportunity identifier
- Opportunity name
- Opportunity description
- Internally or externally identified Opportunity?
- Opportunity creator/owner
- Opportunity contributors
- Opportunity keywords
- Opportunity status (e.g. fulfilled, rejected, under investigation, on hold, assigned, to be reviewed, incomplete, etc.)
- Opportunity creation date
- Opportunity priority (e.g. low, medium, high, urgent)
- Needs associated with Opportunity
- Market(s) associated with the Opportunity
- Ideas associated with the Opportunity
- Concepts associated with the Opportunity
- Possible links to:
  - Other Opportunities
  - Other entity types

Entities corresponding to the **Organisational Unit** entity type are groupings of people that can be defined functionally, regionally, or according to their relation to products or services. Therefore, the Organisational Unit entity type can represent the organisation in its entirety, a division or sub-division of the organisation or even a single project team. Different organisational units of an organisation have different inputs into the innovation process and possess different skill sets. To achieve the best results a multi-functional approach to innovation should be taken from the earliest possible stage in the innovation lifecycle of a given idea. The Organisational Unit entity type is characterised by the following attributes:

- Unique Organisational Unit identifier
- Organisational Unit name
- Organisational Unit description
- Organisational Unit head
- Organisational Unit members
- Organisational Unit's parent unit
- Organisational Unit divisions
- Organisational Unit keywords
- Organisational Unit initiation date
- Organisational Unit status (e.g. planned, in existence, dissolved, etc.)
- Functions associated with Organisational Unit (e.g. financial control, product design, project management, competitive intelligence, quality assurance, etc.)
- Knowledge Areas associated with Organisational Unit
- Needs associated with Organisational Unit
- Skills/Competencies associated with Organisational Unit
- Technologies/Methods/Tools/Equipment associated with Organisational Unit
- Products/Services associated with Organisational Unit
- Projects associated with Organisational Unit
- Opportunities associated with Organisational Unit
- Ideas associated with Organisational Unit
- Concepts associated with Organisational Unit
- Possible links to:
  - Other Organisational Units
  - Other entity types

The **Person** entity type will represent entities corresponding to individuals internal to the organisation (e.g. an employee in general or a member of an innovation team) or external to the organisation (e.g. an external consultant). Innovation is driven by people and therefore this entity type is a mandatory element of the framework. The following attributes describe the Person entity type:

- Unique identifier for Person
- Name of Person
- Person's contact details
- Internal or external Person?
- Organisational Unit(s) associated with Person
- Keywords associated with Person
- Innovation roles associated with Person (e.g. connector, scout, prototyper, librarian, etc.)
- Skills/Competencies associated with Person (e.g. project management, C# programming, CAD modelling, etc.)
- Knowledge Areas associated with person
- Person's qualifications (e.g. M.Eng. in Industrial Engineering, B.Comm. in Marketing, B.Sc. in Computer Science, National Diploma in Graphical Design, etc.)
- Possible links to:
  - Other Persons
  - Other entity types

Entities represented by the **Process/Activity** entity type involve any single formal action (i.e. an Activity) or a structured set of activities (i.e. a Process) that is executed on a regular basis. Entities of this type are characterised by the following attributes:

- Unique Process/Activity identifier

- Process/Activity name
- Process/Activity description
- Process/Activity owner
- Persons associated with Process/Activity
- Process/Activity keywords
- Status of Process/Activity (e.g. planned, in existence, phased out, etc)
- Process/Activity initiation date
- Level of risk associated with Process/Activity (e.g. low, medium, high)
- Process/Activity complexity level (e.g. low, medium, high)
- Level of potential benefit of Process/Activity (e.g. low, medium, high)
- Level of Process/Activity cost (e.g. low, medium, high)
- Process/Activity priority (e.g. low, medium, high, urgent)
- Opportunities associated with Process/Activity
- Needs associated with Process/Activity
- Objectives associated with Process/Activity
- Knowledge Areas associated with Process/Activity
- Technologies/Methods/Tools/Equipment associated with Process/Activity
- Products/Services associated with Process/Activity
- Parent Process/Activity
- Sibling Processes/Activities
- Possible links to:
    - Other Processes/Activities
    - Other entity types

The **Product/Service** entity type represents the existing or planned products and services an organisation offers or intends to offer to its target market. This entity type may also describe products and services provided by other organisations and competitors. Innovation opportunities, ideas and concepts may pertain to products or services of the target organisation or the products and services of its competitors. The Product/Service entity type is described by the following attributes:

- Unique Product/Service identifier
- Product/Service name
- Product/Service description
- Product/Service owner (e.g. person, organisational unit, competitor)
- Product/Service keywords
- Product/Service status (e.g. planned, in development, in force, retired, etc.)
- Product/Service introduction date
- Level of associated risk for Product/Service (e.g. low, medium, high)
- Product/Service complexity level (e.g. low, medium, high)
- Profitability level of Product/Service (e.g. low, medium, high)
- Lifecycle stage associated with Product/Service priority (e.g. novel, maturing, mature, declining, etc.)
- Distribution Channels associated with Product/Service
- Needs associated with Product/Service
- Opportunities associated with Product/Service
- Ideas associated with Product/Service
- Concepts associated with Product/Service
- Knowledge Areas associated with Product/Service
- Possible links to:
    - Other Products/Services

  - Other entity types

Entities corresponding to the **Project/Programme** entity type are groupings of unique, formal activities performed by a person or group with predefined goals, budget and time limits. In terms of innovation, projects can be created to explore certain markets, certain needs or opportunities and further to explore and refine ideas, concepts or a combination thereof. Once a concept has passed the innovation funding gate, one or more projects may be defined to realise the relevant concept in the form of a product or service. An operational project may also involve high levels of innovativeness and are therefore not excluded from the scope of this entity type. The Project/Programme entity type is described by the following attributes:

- Unique Project/Programme identifier
- Project/Programme name
- Project/Programme description
- Project/Programme owner
- Project/Programme members
- Project/Programme keywords
- Project/Programme status (e.g. completed, in process, on hold, cancelled, etc.)
- Project/Programme innovation type (e.g. incremental, radical, operational, etc.)
- Project/Programme initiation date
- Project/Programme completion date
- Project/Programme location (e.g. external, internal, etc.)
- Level of risk associated with Project/Programme (e.g. low, medium, high)
- Project/Programme complexity level (e.g. low, medium, high)
- Level of potential benefit of Project/Programme (e.g. low, medium, high)
- Level of Project/Programme cost (e.g. low, medium, high)
- Project/Programme priority (e.g. low, medium, high, urgent)
- Objectives associated with Project/Programme
- Knowledge Areas associated with Project/Programme
- Skills/Competencies associated with Project/Programme
- Technologies/Methods/Tools/Equipment associated with Project/Programme
- Parent Project/Programme
- Sibling Project/Programme
- Possible links to:
  - Other Project/Programme
  - Other entity types

The **Skill/Competency** entity type attempts to represent the abilities, experience and tacit knowledge of individuals in the light of innovation. A skill can be defined as an ability acquired by training (e.g. spray painting), while competency (e.g. the quality of being able to communicate professionally or managing a project) is defined as the quality of being adequately or well qualified physically and intellectually (WordNet, 2006). The following attributes describe the Skill/Competency entity type:

- Unique identifier for Skill/Competency
- Name of Skill/Competency
- Description of Skill/Competency
- Keywords associated with Skill/Competency
- Organisational Units associated with Skill/Competency

- Knowledge Areas associated with Skill/Competency
- Innovation roles (e.g. connector, scout, prototyper, librarian, etc.) associated with Skill/Competency
- Persons associated with Skill/Competency
- Qualifications associated with Skill/Competency (e.g. M.Eng. in Industrial Engineering, B.Comm. in Marketing, B.Sc. in Computer Science, National Diploma in Graphical Design, etc.)
- Processes/Activities associated with Skill/Competency
- Possible links to:
    - Other Skills/Competencies
    - Other entity types

The **Supplier** entity type shall represent entities corresponding to organisations that provide products (including raw materials, equipment, etc.) or services (e.g. hosting a website) to the target organisation. Suppliers are instrumental in providing components or services required to realise existing or intended products or services of the target organisation. The Supplier entity type is characterised by the following attributes:

- Unique Supplier identifier
- Supplier name
- Supplier description
- Supplier location
- Supplier keywords
- Supplier maturity (e.g. low, medium, high, etc.)
- Status of relationship with Supplier (e.g. planned, current, historic, etc.)
- Supplier engagement date
- Replaceability of Supplier (e.g. low, medium, high)
- Skills/Competencies associated with Supplier
- Technologies/Methods/Tools/Equipment associated with Supplier
- Products/Services supplied by Supplier and link to own Products/Services
- Knowledge Areas associated with Supplier
- Possible links to:
    - Other Suppliers
    - Other entity types

The **Technology/Method/Tool/Equipment** entity type serves to capture the possible technologies, methods, tools and/or equipment that may be used to realise an improvement to or the creation of a product or service associated with a given opportunity, idea or concept. This entity type essentially describes various means to different innovation-related ends. In addition this entity type can be used to associate existing instance of technologies, methods, tools and equipment to other framework entity types other than Opportunities, Ideas or Concepts (e.g. Persons, Organisational Units, etc.). The Technology/Method/Tool/Equipment entity type is described by the following attributes:

- Unique Technology/Method/Tool/Equipment identifier
- Technology/Method/Tool/Equipment name
- Technology/Method/Tool/Equipment description
- Technology/Method/Tool/Equipment owner
- Technology/Method/Tool/Equipment keywords

- Technology/Method/Tool/Equipment maturity (e.g. experimental, maturing, mature, phase out, etc.)
- Technology/Method/Tool/Equipment target availability date
- Level of risk associated with Technology/Method/Tool/Equipment (e.g. low, medium, high)
- Technology/Method/Tool/Equipment complexity level (e.g. low, medium, high)
- Level of potential benefit of Technology/Method/Tool/Equipment (e.g. low, medium, high)
- Level of cost associated with Technology/Method/Tool/Equipment (e.g. low, medium, high)
- Persons associated with Technology/Method/Tool/Equipment
- Knowledge Areas associated with relevant Technology/Method/Tool/Equipment
- Skills/Competencies associated with relevant Technology/Method/Tool/Equipment
- Ideas associated with relevant Technology/Method/Tool/Equipment
- Concepts associated with relevant Technology/Method/Tool/Equipment
- Products/Services associated with relevant Technology/Method/Tool/Equipment
- Needs associated with relevant Technology/Method/Tool/Equipment
- Possible links to:
    - Other Technologies/Methods/Tools/Equipment
    - Other entity types

Entities corresponding to the **Topic** entity type are generated by statistical topic models and represent the themes embedded in a set of analysed documents. Topics consist of words with each word having a probability indicating the estimated strength of association between the word and the topic in question. A topic captures meaning and may be given a descriptive label by a human. Documents can be represented as a mixture of different topics. The following attributes are required to represent the Topic entity type:

- Unique Topic identifier
- Topic name (i.e. Topic label)
- Words associated with Topic with individual association strengths
- Documents associated with Topic with individual association strengths
- Other Topics associated with given Topic with individual affinity values
- Topic creation date
- Associated topic model
- Persons who interpreted the Topic
- Topic rating in terms of interpretability (e.g. poor, average, high, excellent)
- Possible links to:
    - Other entity types

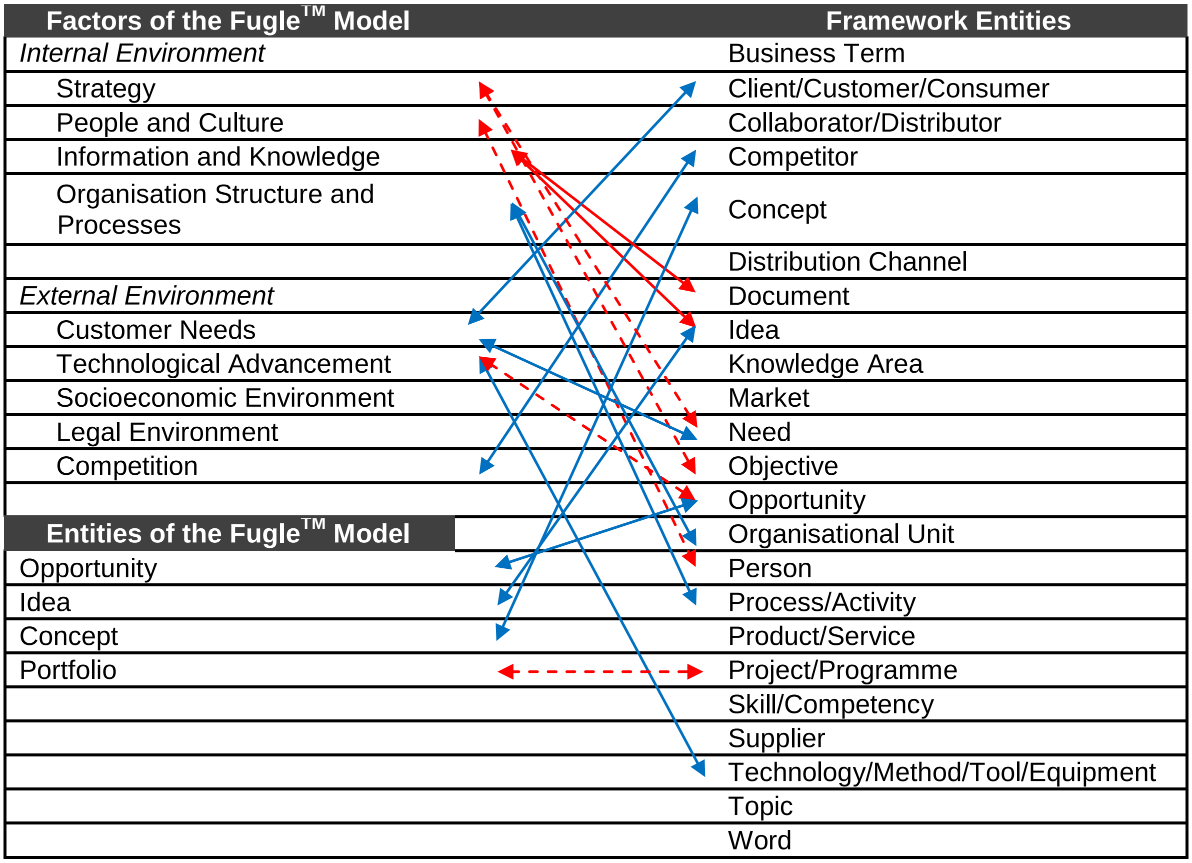
The **Word** entity type shall represent entities corresponding to the most basic units of meaning and the building blocks of sentences and document text. More specifically, entities of this entity type correspond to the words characterising topics in statistical topic models. This entity type may contain single words (e.g. ‘steel’, ‘jig’, etc.) or terms consisting of multiple words (e.g. ‘project management’, ‘financial needs assessment’, etc.). A single word may have one or more senses (i.e. meanings). Business vocabularies, collections of business terms, formally define words and terminology used in an organisation. Some entities may therefore correspond to both Word and Business Term entity types. The Word entity type is defined by the following set of attributes:

- Unique Word identifier
- Representing Word(s)
- Corresponding Topics along with individual strengths of association

- Corresponding Documents along with individual strengths of association
- Corresponding Words along with individual strengths of association
- Date of first observation
- Corresponding Business Terms
- Possible links to:
    - Other entity types

By populating the framework with actual entities corresponding to the entity types described in this section, innovation-related information and business metadata contained in disparate information sources across the organisation is integrated, structured and can be made available through a single mechanism. This populated framework can subsequently be used to gain a comprehensive overview of an organisation's innovation resources (e.g. persons, collaborators, knowledge areas, technologies, equipment, etc.), innovation potential (e.g. needs, opportunities, ideas, concepts, etc.) and innovation capabilities (e.g. skills/competencies and knowledge areas).

In an attempt to evaluate the comprehensiveness of the different framework entity types defined, Table 17 maps these entity types to the influencing factors and entities of Fugle^{TM} model for innovation (see section 3.3). In this table a dashed line represents an indirect match whereas a solid line represents a more direct match.



Factors of the Fugle^{TM} Model	Framework Entities
Internal Environment	Business Term
Strategy	Client/Customer/Consumer
People and Culture	Collaborator/Distributor
Information and Knowledge	Competitor
Organisation Structure and Processes	Concept
	Distribution Channel
External Environment	Document
Customer Needs	Idea
Technological Advancement	Knowledge Area
Socioeconomic Environment	Market
Legal Environment	Need
Competition	Objective
	Opportunity
Entities of the Fugle^{TM} Model	Organisational Unit
Opportunity	Person
Idea	Process/Activity
Concept	Product/Service
Portfolio	Project/Programme
	Skill/Competency
	Supplier
	Technology/Method/Tool/Equipment
	Topic
	Word

Table 17: Comparing the influencing factors and entities of the Fugle^{TM} model to the framework entity types

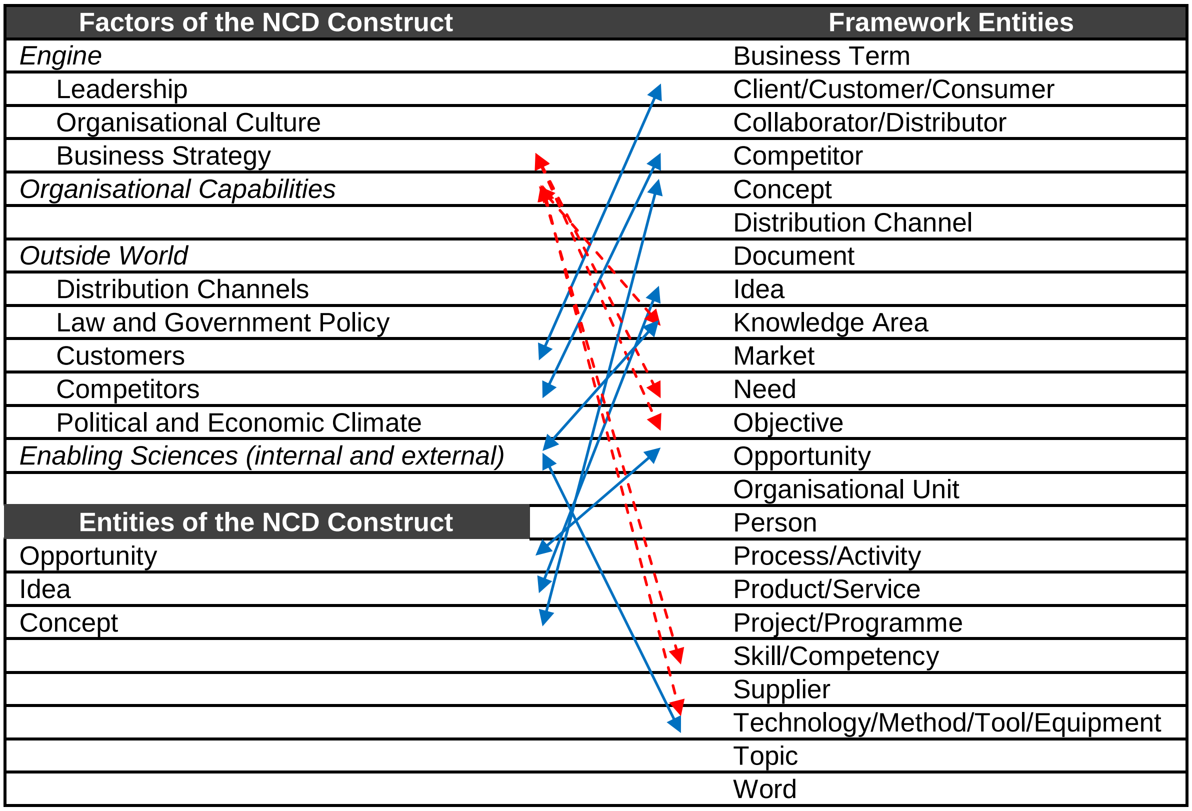
From Table 17 it can be seen that the following influencing factors of the Fugle^{TM} model are not represented as framework entity types:

- Socioeconomic Environment
- Legal Environment

Changes in the socioeconomic environment the target organisation operates in will be reflected as Opportunities, Needs and/or Markets which are included in the current framework definition. The same would apply for changes in the legal environment the target organisation operates in. The following influencing factors or entities of the Fugle^{TM} model are partially represented in the framework:

- Strategy
- People and Culture
- Portfolio

Strategy is partially covered by the Objective and Need entity types. People and Culture is partially covered by the Person entity type. The concept of organisational culture is too abstract to represent as an information entity and will therefore not be represented in the framework. The Portfolio factor is partially covered by the Project/Programme entity type as a project portfolio is just a logical, hierarchical grouping of different projects and their sub-projects.



Factors of the NCD Construct	Framework Entities
Engine	Business Term
Leadership	Client/Customer/Consumer
Organisational Culture	Collaborator/Distributor
Business Strategy	Competitor
Organisational Capabilities	Concept
	Distribution Channel
Outside World	Document
Distribution Channels	Idea
Law and Government Policy	Knowledge Area
Customers	Market
Competitors	Need
Political and Economic Climate	Objective
Enabling Sciences (internal and external)	Opportunity
	Organisational Unit
Entities of the NCD Construct	Person
Opportunity	Process/Activity
Idea	Product/Service
Concept	Project/Programme
	Skill/Competency
	Supplier
	Technology/Method/Tool/Equipment
	Topic
	Word

Table 18: Comparing the influencing factors and entities of the NCD construct to the framework entity types

Continuing in this vein, Table 18 maps the framework entity types to the factors and entities of the NCD construct (see section 3.4.1). As presented in this table, the following influencing factors of the NCD Construct are not represented as framework entity types:

- Leadership
- Organisational Culture
- Law and Government Policy
- Political and Economic Climate

Leadership is too abstract to represent as an explicit entity type in the framework. Leadership may however be indirectly reflected in the Need, Skill/Competency and Objective entity types. As discussed before, Organisational Culture is too abstract to represent as an explicit framework information entity. The Law and Government Policy factor and the Political and Economic Climate closely correspond to the Legal Environment factor and the Socioeconomic Environment factor of the Fugle^{TM} Model respectively. The reasoning maintaining that changes in these factors will be reflected as Opportunities, Needs and/or Markets therefore also applies here.

The following influencing factors or entities of the NCD Construct are partially represented in the framework:

- Business Strategy
- Organisational Capabilities

Business Strategy closely resembles the Strategy factor of the Fugle^{TM} model and is partially covered by the Objective and Need entity types. Lastly, the Organisational Capabilities factor is partially addressed by the Skill/Competency, Technology/Method/Tool/Equipment and Knowledge Area entity types.

The respective comparisons of the framework entity types to the influencing factors and entities of the Fugle^{TM} model and the NCD construct have not brought to light any major, concrete oversights in terms of the framework entity types. This, however, does not imply that the framework would be completely applicable, in its current form, to all industries and innovating organisations. It only indicates that the framework's information entities seem sufficient to capture and relate innovation-related information on a generic innovation process level.

The actual information entities that will be associated with the framework entity types can reside in many possible locations in the organisation including electronic documents, the databases of organisational information systems, etc. Moreover, other organisational information systems may consume the information embodied in the framework also implying another kind of information system framework dependency. The following section discusses which organisational information systems may be included in the implementation of the framework in an organisation.

8.8 The Relation of the Framework to Organisational Information Systems

The realistic application of the framework requires minimum human intervention to populate and maintain it without compromising information accuracy and applicability. Ideally, most of the actual information entities corresponding to the framework entity types would exist in one or more information system of the particular organisation. It is good practice that each information entity type be hosted primarily in one and only one 'source' system in order to promote information integrity, consistency and the ease of maintaining information about such entities. Instead of duplicating information, other systems requiring such information should source it from the relevant 'source' system. For example, if the HR system and the payroll system host the same kinds of information about employees separately, both systems need to be updated when an employee's information changes (e.g. an employee marries and her surname changes) to eliminate inaccurate or outdated information. If a given system must retrieve information about a specific information entity from another system, a certain level of inter-system integration is implied. Here the use of service-orientated architecture principles and web services may be useful. Ideally, a single system should be used to make addressable all business information; other systems should interface with this system to retrieve such information. It is therefore envisaged that the implemented framework would reference information in other systems, in cases where such systems are the primary host of the applicable information, without duplicating it if possible. Similarly, other systems should reference the information embodied in the framework in cases where the framework is the primary host of such information.

Most organisations have a range of information systems that are used to facilitate its day-to-day operations; each system carrying information about a fairly unique set of entity types and facilitating the execution of a unique set of business processes. No single information system will however host the full range of framework entity types and corresponding entities. Many information systems could be involved in the implementation of the framework in an organisation, some of which include (listed in alphabetical order):

- **Document Management System (DMS):** At minimum, a centralised document repository storing important electronic documents is required. In its simplest form this will be a shared network drive on the organisation's local area network, but ideally it would be a proper document management system that not only stores documents, but also captures all interactions with such documents (e.g. which employees interact with which documents, audit trail of document updates, etc.). Many systems having document management functionality are available and used in organisations including *Microsoft SharePoint Server*, *Documentum*, *Alfresco*, *EDEN* as well as custom-developed systems. In product focused organisations, systems aiding product design (e.g. CAD systems like *ProEngineer* or *Catia*) may host product designs in the form of electronic engineering drawings and designs. For the purpose of this research such systems are also regarded under the umbrella of document management systems. The DMSes of the target organisation would be the primary sources of textual information that will be analysed by the Text Analytics System that would in turn generate information to be used to populate the framework. DMSes may further provide information that relates different documents to other entity types of the framework (e.g. Organisational Unit, Person, Project/Programme, etc.)
- **Enterprise Search System (ESS):** The term 'Enterprise Search System' refers to any information system that facilitates searching for electronic, (typically unstructured) organisational information. An ESS may be part of an organisation's enterprise information portal[32] (EIP), it may be a system on its own or it may be a combination of different systems. The organisation's ESS and EIP will be an important component in delivering the information contained in the framework by means of structured browsing and intelligent searching using the innovation-related information entities and relationships between such entities that are stored in the framework.
- **Framework Management System (FMS):** The Framework Management System will host, integrate and maintain information corresponding to the various framework entities types. The actual information about some of the information entities corresponding to the framework entity types may primarily reside in other information systems as structured information (e.g. the information about the various employees of the organisation may reside in the database of the HR system). Other information corresponding to the framework entity types will only reside in unstructured text form. As part of the FMS, software agents[33] will continuously generate, store and enrich information about the various entities based on the content of electronic documents and the relationships of such

[32] EIPs are applications that enable organisations to unlock internally and externally stored information, and provide users with a single gateway to personalised information needed to make informed business decisions (Shilakes & Tylman, 1998).

[33] A "software agent" is a computer program that intelligently executes certain tasks without human interaction.

documents to other framework entities. Another set of software agents will be required to do the same, but only using the organisation's information systems as sources of such framework entity type information. The FMS can be regarded as the primary hosting platform of the framework and the entity types it comprises.

- **Innovation Management System (IMS):** An Innovation Management System facilitates the execution of an organisation's innovation processes in terms of capturing, sharing, managing, refining, evaluating and selecting ideas and concepts. The IMS may be one or more standalone systems, facilitating a specific part of the entire innovation process (e.g. an idea management system, a project management system, etc.) or a combination of different systems working in an integrated manner with the aim to facilitate the innovation process in its entirety (e.g. Microsoft's *Innovation Process Management* platform). When no formal IMS is implemented in an organisation, it is recommended that a workflow system be configured to facilitate the execution of the organisation's innovation processes. In the light of the framework, the IMS would be the source of innovation ideas and concepts as well as their relationships to other entity types such as Persons, Opportunities, Needs, etc.
- **Text Analytics System (TAS):** A Text Analytics System represents any system that analyses collections of electronic text and renders distilled information for use by humans or other information systems. With regard to the framework, one or more TAS is required to create concentrated abstractions of the individual documents stored in the target organisation's DMSes as well as to generate the relationships between such abstractions (e.g. topics) and other entities (e.g. specific documents, other topics, words, etc.). The resulting information is then used to populate the framework where it is linked to other innovation-related information entities.
- **Workflow Management System (WMS):** A Workflow Management System is a system that enables procedural automation of a business process by handling the sequence of work activities and by managing the required resources (e.g. people, data, applications, etc.) associated with the respective activity steps of the different business processes. Examples of WMSes include *AWD*[34] and *Eclipse*[35]. In the light of the framework, the WMS will serve as the proverbial glue between the different systems used in the implementation of the framework. More specifically, it may be used to automate the different processes involved in the population of and interaction with the framework. For example, if a new innovation idea is defined in the IMS, the WMS will trigger the creation of a new entity, of entity type Idea, in the FMS to represent this new idea in the framework. In addition, the WMS may serve to act upon certain predefined patterns of information in the framework. For example, a person may elect to be notified (e.g. via e-mail) when a new business term is defined or when an entity of type Technology/Method/Tool/Equipment is associated to a specific idea.

The goal of the list of systems presented above is to demonstrate the likely set of systems that may be required for implementing the framework in a generic organisation. The systems required may differ from industry to industry however. Figure 31 depicts a possible configuration of the various information systems that are typically required to implement the proposed framework.

[34] Automated Work Distributor by DST

[35] Refer to http://www.eclipse.org/jwt/ for more information about Eclipse.

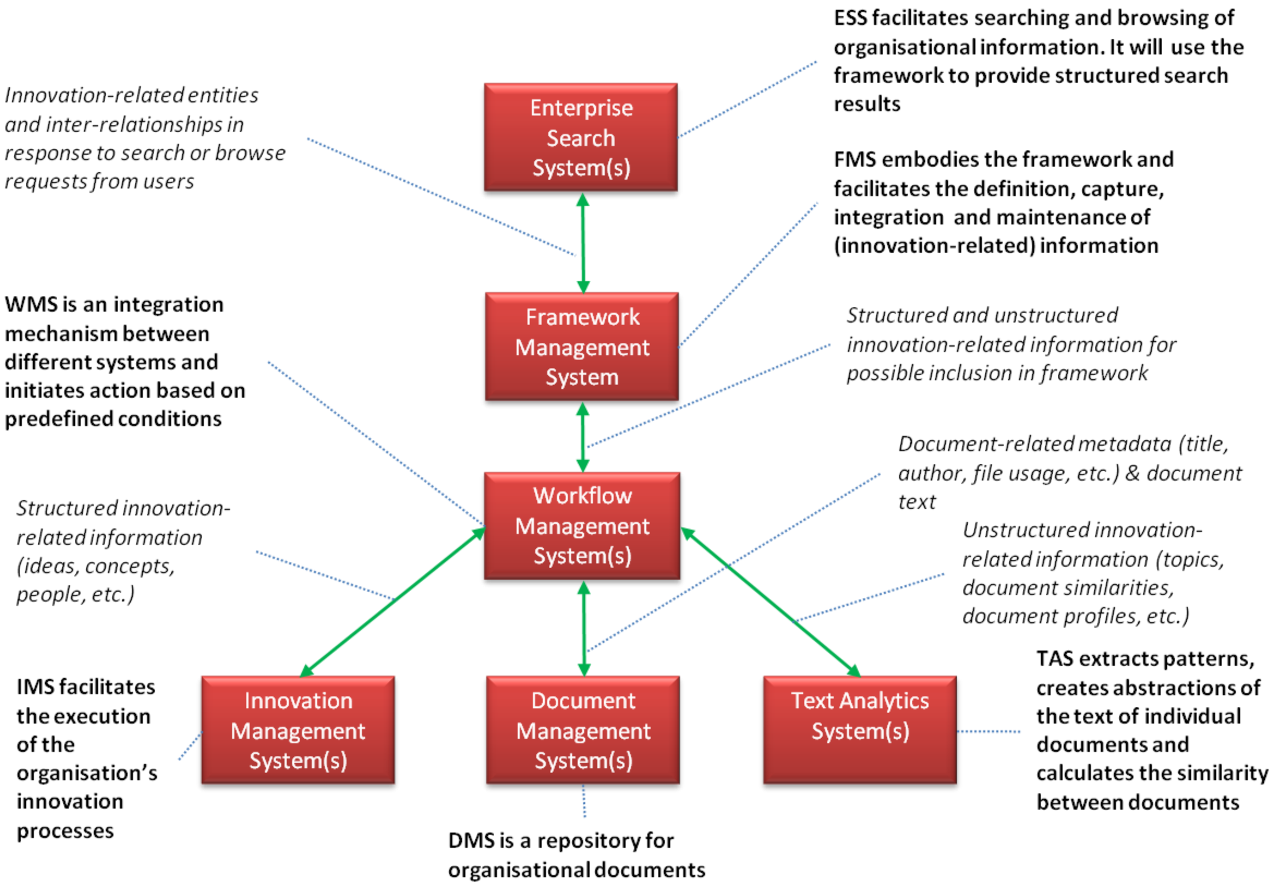

Enterprise Search System(s)
Framework Management System
Workflow Management System(s)
Innovation Management System(s)
Document Management System(s)
Text Analytics System(s)
ESS facilitates searching and browsing of organisational information. It will use the framework to provide structured search results
Innovation-related entities and inter-relationships in response to search or browse requests from users
FMS embodies the framework and facilitates the definition, capture, integration and maintenance of (innovation-related) information
WMS is an integration mechanism between different systems and initiates action based on predefined conditions
Structured and unstructured innovation-related information for possible inclusion in framework
Document-related metadata (title, author, file usage, etc.) & document text
Structured innovation-related information (ideas, concepts, people, etc.)
Unstructured innovation-related information (topics, document similarities, document profiles, etc.)
IMS facilitates the execution of the organisation's innovation processes
TAS extracts patterns, creates abstractions of the text of individual documents and calculates the similarity between documents
DMS is a repository for organisational documents

Figure 31: Different information systems and their relationships to the framework

This concludes the discussion about the likely ICT requirements for implementing the framework in an organisation. The following section addresses the working methods associated with the framework.

8.9 Associated Framework Mechanisms

Regardless of how the proposed framework is implemented in the target organisation, certain (manual and automatic) IT processes and mechanisms are required to populate, maintain the framework and to deliver value using the framework. Seven high-level processes can be distinguished in this regard:

1. Collecting information from the external environment
2. Instantiating and configuring the framework
3. Populating the framework
4. Inferring relationships between various entities
5. Distributing information using the framework
6. Retrieving information using the framework
7. Updating the framework

These processes will be discussed in more detail in the following sections.

8.9.1 Collecting Information from the External Environment

In section 4.4 some examples of typical external sources of (mainly unstructured) innovation-related information have been identified and discussed. The target organisation should identify, subscribe to (where required) and monitor the sources that best fit their innovation-related information needs. The organisation should further sanction employees to collect and store such information within the boundaries of the organisation. Alternatively or in addition, the organisation may choose to make use of external parties, services or software to collect and package such information on behalf of the target organisation. Where possible, such gathered external information should be in electronic format and should be stored in the organisation's document management system where interactions of people with such information can be recorded. The gathering and storing of innovation-related information from the external environment is an ongoing activity that needs to happen regardless of whether the framework is implemented in the target organisation.

8.9.2 Implementing and Configuring the Framework

In order to unify, organise and make addressable an organisation's (unstructured and structured) innovation-related information, the target organisation has to implement and configure the proposed framework. The following are some of the typical steps that have to be followed:

- Determine which organisational processes the framework would support once implemented.
- Identify individuals or groups that will use the framework once implemented.
- Determine the specific user needs with regard to innovation-related information.
- Evaluate the applicability of the framework entity types to target organisation.
- Modify the framework entity types according to the identified needs.
- Identify the set of desired functionalities that would be required to satisfy the identified user needs.
- Identify, using the desired functionality set, which of the organisation's existing information systems could be used as part of the framework implementation. Such systems may include, but are not limited to the systems discussed in section 8.8.
- Further identify, once more using the desired functionality set, which systems have to be acquired to implement the framework.
- Implement the framework using the required information systems and technologies.
- Configure the information systems as to support the identified organisational processes and satisfy the identified user needs.

The framework can be implemented in a gradual fashion and such implementation should be regarded and managed as a radical innovation project.

8.9.3 Populating the Framework

Once the implemented framework is ready to accept information, the following steps may be followed to start the population of the framework with the organisation's innovation-related information entities:

- Identify which of the organisation's information systems may possibly host information regarding the framework entity types determined to be relevant to the target organisation.

- Investigate the manifestation of such framework entity types in the identified systems. For example, some information about innovation ideas may be found in the organisation's wiki and is stored in html format. As another example, information about employees is stored in the HR system's database.
- Map the framework entity types and their characterising attributes to the data structures and elements of the appropriate systems. For example, the records of the Employees database table of the HR system correspond to instances of the Person framework entity type[36].
- Investigate the availability of suitable interfaces (e.g. web services or APIs) for interacting with the implicated systems.
- Design and implement the required integration or synchronisation mechanisms between the system hosting the framework, in other words the Framework Management System (FMS), and systems hosting (structured or semi-structured) information about framework entity types.
- Identify document repositories (e.g. document management systems or electronic storage locations) that may possibly host documents containing information entities corresponding to the defined framework entity types.
- Implement mechanisms to convert electronic textual documents to the formats supported by the implemented Text Analytics System.
- Map the entity types and outputs of the Text Analytics System (TAS) to the framework entity types. The TAS's output will possibly include entities corresponding to the Document, Topic and Word framework entity types.
- Configure the TAS to analyse the target document repositories at desired intervals.
- Implement one or more mechanisms to populate the framework with the information retrieved from document text by the TAS. For example, configure a workflow system to retrieve outputs of the TAS, modify such outputs to a format suitable for population in the FMS and actually populate such information in the FMS at regular intervals.
- Implement a mechanism to notify selected persons about the availability of new results for possible evaluation, refinement, and approval.

The population of the FMS with innovation-related information is an ongoing activity. It should be noted that the FMS does not necessarily carry the complete information instances of such innovation-related information entities; in some cases the FMS only carries partial entities that include references to the location where the actual corresponding information entities primarily resides. For example, the FMS may contain the entity "John Fourie", of type Person, which includes a reference to the database, table and record where the information entity "John Fourie" actually resides (e.g. the database of the HR system). Also, not all innovation-related information will reside in electronic documents or the information systems of the target organisation; a large portion of such information may only exist in the minds of employees. For example, the products of a given competitor and how such products relate to the products of the target organisation may not be documented, or may be documented but cannot be extracted in such a precise form. One or more employees of the target organisation may however know the sought information. The framework can be used to capture such knowledge from employees using suitable interfaces to the FMS so

[36] The reader is referred to Kotze (2008) for a detailed description of how to align, unify and make accessible the structured information of the organization to non-technical users.

that these individuals can capture the knowledge themselves, or more likely, using facilitated sessions to capture such knowledge in the FMS.

8.9.4 Inferring Relationships between Framework Entities

Once the framework has been populated with the actual entities corresponding to the respective innovation-related entity types sourced from electronic documents, organisational databases or even manually entered; the next step is to infer new information based on the entities and their attributes existing in the framework. It is unlikely that all attributes of a newly created entity of a given entity type be filled with the initial creation of such an entity in the FMS. Some attributes, especially those corresponding to relationships between different entities (e.g. the affinities between different persons based on the documents they interact with), have to be inferred from the initial information entered into the FMS for the corresponding entity. This requires a mechanism that can be tasked with the post processing of the information contained in the framework with the aim to deduce new information. For example, the discovery that "John Fourie" has the same innovation-related information interests as "Catherine du Toit"; therefore characterising such a shared interest and subsequently updating the FMS to reflect this newly discovered association. The reader is referred to sections 10.4, 10.5 and 11.7 for the discussion of experiments performed to investigate such possible inference by post processing the results generated by statistical topic models.

The FMS will therefore have to include software agents with the purpose of extending the entities existing in the FMS as well as their inter-relationships. The degree of inference possible depends on many factors, such as the inference capabilities of the technology (e.g. OWL, topic maps or a traditional relational database) used by the FMS to firstly define the framework in terms of its entity types and characterising attributes, and secondly to store the actual innovation-related information entities and their respective inter-relationships. Other factors that may influence the degree of inference possible include the completeness and accuracy of the information characterising each entity captured in the FMS.

8.9.5 Distributing Information using the Framework

The ultimate value of the framework is to provide stakeholders of the target organisation's innovation processes with actionable information that they can use to make better informed decisions, increase and share their knowledge, better allocate resources, to name a few. Information embodied in the FMS can be used proactively or reactively. **Proactive use** includes sending notifications about information, which the system deems applicable to specific individuals[37], to such individuals at regular intervals. An example of proactively using the FMS would be when an

[37] The approach where a system automatically filters a collection of information and sends 'personalised' information to specific individuals based on their information interest or preferences, is called 'collaborative filtering'. (cf. Snowdon & Grasso, 2002).

individual specifies, using the FMS, a standing query[38] or a trigger[39] that defines one or more conditions that should be met in order for the individual to be notified of such an event or associated information. For example, a user may instruct the FMS to notify him, via e-mail and RSS feed, when any document is entered into the framework that deals with a new[40] entity of the Technology/Method/Tool/Equipment entity type. A more complete set of conditions that may be used to trigger a notification to a user includes:

- When an entity is changed in the FMS (the user can select whether it should include changes to any entity, changes to one or more specific entities or changes to any entity of one or more specific entity types)
- When a new entity is added in the FMS (the user can select whether it should include the addition of any entity or an addition of an entity of one or more specific entity types)
- When a connection between one or more specified entities and another entity is added (the user may once more specify the specific target entities to act upon or their corresponding entity types)
- When a connection between one or more specified entities and another entity is changed (the user may once more specify the specific target entities to act upon or their corresponding entity types)

To expand the earlier example of “John” and “Catherine”, the FMS may proactively act on the newly discovered relationship[41] between “John” and “Catherine”, by determining whether there are some documents, corresponding to the interest area(s) shared by the two individuals, which are associated with one individual but not the other. This mechanism can then notify the individuals of such potentially interesting documents by sending an e-mail containing links to such identified documents to the individuals. Suppose that the mechanism detected that there is a document, say “Selective Laser Sintering.pdf”, read by “John”, corresponding to the identified interest area he shares with “Catherine”, but not yet read by “Catherine”[42]. The mechanism would then send an e-mail to “Catherine” containing the link to the document “Selective Laser Sintering.pdf” as well as the contact details of “John” and the interest area they have in common.

8.9.6 Retrieving Information using the Framework

Where the previous section discussed the proactive use of the information embodied in the FMS, this section deals more with the **reactive use** of such information. Reactive use of the information in the FMS includes scenarios where a person interrogates the FMS ad hoc to find information

[38] Basically, a ‘standing query’ is like any other query except that it is periodically executed on a document set to which new documents are incrementally added over time. (Manning, Raghavan & Schütze, 2008)

[39] ‘Triggers’ are mechanisms that determine when certain active objects evaluate their associated queries to recalculate their associated output or to take some other action. (Shipman & McCall, 1994)

[40] More specifically, ‘new’ in the sense that it was not previously embodied in the framework up to this point in time.

[41] Recall that this relationship was based on the automatic identification of one or more shared interest areas between the two individuals based on the documents they interact with.

[42] This may be based on the transactions with this document as retrieved from the document management system for example.

about a specific innovation-related issue. An example of the reactive use of the FMS would be to post a query to find all innovation ideas that deals with a specific need, say "Improving the shelf life of Product_X".

The following are considered part of the reactive use of the information in the FMS:

- Using an appropriate interface to browse through the information in the FMS using the various entity types to structure information[43]. For example, all framework entity types are displayed; the user the selects the Competitor entity type and all entities in the framework of type Competitor will be displayed.
- Using an appropriate interface to interactively browse through the information in the FMS using a specific entity as focus and using its relationships to other entities to change the focus to another entity of interest. This process may be repeated to explore the information in the framework in an interactive way. For example, selecting the 'XYZ Technologies' entity, corresponding to a competitor, and finding all their products and possibly the relationships between their products and that of the target organisation.
- Using free-form text queries to find entities matching the search request. For example, the user may enter the query "selective laser sintering" to find all matching entities. The search results may be organised according to the corresponding entity types of entities found.
- Using advanced or structured queries to specify which entity types should be considered in the search and possibly some attribute values of such entity types to reduce the number of hits. For example, the user may elect to only find entities of type Document and Person which are associated with "selective laser sintering" with documents not being older than 12 months.

A way of making accessible the information embodied in the FMS is to integrate the FMS with the target organisation's enterprise search system and/or information portal.

8.9.7 Updating the Framework

In order to reflect the latest view on the innovation-related information of the target organisation, the information embodied in the FMS have to be updated from time to time. A number of events may trigger the updating of the FMS information, including:

- An inaccuracy is found in one of the entities or its relationships to other entities. For example, an incorrect distribution channel, say 'Mail Order', was wrongly associated with one of the target organisation's competitor's products, say 'Product_z'.
- New information regarding one or more entities (including its relationships to other entities) becomes available. The current information about the entities in question may therefore be expanded. For example, competitor 'Product_z' is now distributed through an additional distribution channel, say 'Internet Commerce'.
- Information regarding one or more entities (including its relationships to other entities) changes. For example, competitor product, say 'Product_z', is rebranded as 'Product_m'.
- New innovation-related entities come into existence. For example, a new organisational unit, say 'New Business Development' is founded in the target organisation.
- One or more entities cease to exist. For example, competitor 'Product_z' is withdrawn from the market.

[43] This is also known as a 'faceted search', 'faceted navigation' or faceted browsing'.

Some of these framework update events may imply the creation of new entities or changes to existing entities. For instance, the fact that a competitor product is withdrawn from the market may trigger the creation of one or more entities of the type Idea or Opportunity.

The following requirements come to mind with regard to updating the information in the FMS:

- It should be possible for a user to mark a given attribute of an entity or a relationship between two given entities as suspect. Suspect entities or relationships should be investigated by one or more person, the suspect entities or relationships corrected and/or the suspect flag removed.
- The fact that a previously valid entity, entity attribute or relationship changes or cease to exist does not mean that it should completely disappear from the FMS as it embodies historic information which may still be potentially useful. The FMS should therefore have the capability to carry such history and clearly distinguish such history from the status quo.
- All transactions regarding entity types, entities and relationships in the FMS should be recorded including the date of such transactions, the persons or agents responsible for the respective transactions as well as the reasons for such transactions where applicable.
- Relationships between entities should not be limited to binary relationships that can only cater for the fact that a given relationship exists or do not exist at all. It should rather include weighted relationships that allow the strength of association between two entities to grow or decline as time progresses and more evidence is found confirming or refuting such relationships.

The FMS information will most likely be updated mainly by manual intervention. As technology continues to become more sophisticated, the FMS information may increasingly be updated automatically by intelligent software agents[44] [45].

A note on extending the framework: The data structure embodying the framework must be flexible so that new attributes can easily defined for a given entity type. Also, it should be possible to easily define new entity types and their relationships to existing entity types. When the various entity instances of a new entity type have to be populated automatically, the required text analytical systems also have to be updated to be able to identify and capture such instances autonomously. All users of the framework should be notified of the new framework entity types and attributes and the meaning and intention of such new elements should be carefully explained as it actually represents new business metadata.

The following section investigates how users may interact with the framework and associated FMS.

[44] The reader is referred to Miller et al.,(2006) for an in-depth discussion about the critical role of information and information technology in future (accelerated radical) innovation.

[45] Harrington and Clark (2007) discuss a prototype system that uses spreading activation-based algorithms to construct and maintain an automated semantic knowledge network. Such technologies may potentially be used as part of the software agents that are tasked with populating and updating the information in the FMS.

8.10 Working Methods Associated with the Framework

Although the framework is intended to act as an innovation-focused knowledge base within the target organisation, it may further support the knowledge lifecycle (Ribière & Román, 2006) of the organisation, illustrated in Figure 32, with specific focus on innovation-specific information and ultimately, knowledge. Similarly, the framework would also support the SECI knowledge creation cycle described in section 19.2 of Appendix E.

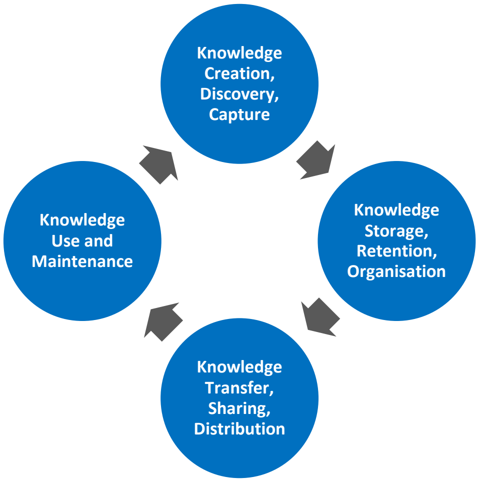

Knowledge Creation, Discovery, Capture
Knowledge Storage, Retention, Organisation
Knowledge Transfer, Sharing, Distribution
Knowledge Use and Maintenance

Figure 32: The Knowledge Lifecycle

To support the innovation processes of the target organisation, support has to be provided to the individuals involved in the execution of innovation. The foreseen working methods associated with the framework may therefore be best explained by discussing how the different roles in innovation, as presented in section 3.5.1, may use the framework in support of their innovation endeavours.

Table 19 illustrates the estimated level of interaction between the respective framework entity types and the different innovation roles. From this table one may theorise that the different innovation stakeholder roles will benefit in varying degrees by the implementation of the framework in the target organisation. It seems as if the Connector and Librarian innovation roles will interact with the largest number of framework entity types, while the Storyteller will probability interact with the least number of framework entity types. As a guesstimate, the framework, once implemented in the target organisation, would therefore likely be of the most benefit to the Connector and Librarian innovation roles in terms of supporting their innovation-related activities[46].

[46] Although the level of support the framework will provide to the various innovation roles cannot be judged alone on the number of entity types of the framework an innovation role would most likely interact with, it is certainly one of the major influencing factors when estimating the level of such support.

Framework Information Entity Type[47]	Connector	Framer	Judge	Librarian	Metric Monitor	Prototyper	Scout	Storyteller
Business Term	2	1	1	2	1	1	2	1
Client/Customer/Consumer	2	2	2	2	2	2	2	1
Collaborator/Distributor	2	1	1	2	2	2	2	1
Competitor	2	2	2	2	1	1	2	1
Concept	2	2	2	2	2	2	1	2
Distribution Channel	2	1	2	2	1	1	2	1
Document	2	1	1	2	1	1	1	1
Idea	2	2	2	2	2	2	2	2
Knowledge Area	2	2	2	2	2	1	1	1
Market	2	2	2	2	1	1	2	1
Need	2	2	2	2	2	2	2	2
Objective	2	2	2	2	2	2	2	2
Opportunity	2	2	2	2	2	1	2	2
Organisational Unit	2	1	1	2	2	1	1	2
Person	2	2	1	2	2	1	1	2
Process/Activity	2	1	1	2	2	1	1	1
Product/Service	2	2	2	2	1	2	2	1
Project/Programme	2	1	1	2	2	2	1	2
Skill/Competency	2	2	2	2	2	2	2	1
Supplier	2	1	1	2	1	1	2	1
Technology/Method/Tool/Equipment	2	2	2	2	2	2	2	1
KEY: 0 = no affinity, 1 = medium affinity, 2 = high affinity								

Table 19: Framework support to different innovation roles

The following sections illustrate how the framework and associated FMS may aid these respective innovation roles in their innovation-related activities. Moreover, these sections will outline the different working methods in dealing with the envisaged framework from the perspective of the different innovation roles.

8.10.1 Possible Interactions of the Connector with the Framework

In short, it is the function of the **Connector** role to build the proverbial bridges between innovation related entities (e.g. individual persons, technologies, ideas, collaborators, etc.) of the target organisation. The framework may aid the Connector in the following ways:

- Since the framework would embody most of the organisation's innovation related entities, it may help the Connector to identify new possible connections by examining unconnected entity pairs of different entity types (e.g. compare all entities of type Technology/Method/Tool/Equipment with all entities of type Idea that are not already connected by a relationship).
- The framework may further help the Connector to better comprehend the organisation's innovation landscape to be able to understand the current needs that would most benefit

[47] The Topic and Word entity types have been left out in this table since these represent intermediary or wildcard entity types that are required to combine the results of statistical topic models with the primary framework entity types.

from his actions (e.g. by examining the entities of type Need, Opportunity and Objective he can align his actions to address such items with higher priority).

- As a by-product, the framework can further serve to capture the network, consisting of the different connections, created by the Connector so that it can be more readily accessible to other innovation workers which may in turn trigger more ideas for possible connections.

8.10.2 Possible Interactions of the Librarian with the Framework

The **Librarian** is primarily responsible for collecting ideas, solutions, problems, technologies and needs. The duties of the Librarian further include enabling organised access to such collected items in order for other innovation workers to assist in building the organisation's innovation 'library' and making sense of the current contents of such a library. Of all the innovation roles describe here, the Librarian will arguably be best supported by the framework. The framework would support the Librarian in many ways, including the following:

- The Librarian largely determines what information has to be captured for each innovation entity type of the organisation. The framework described in this research does not only provide the Librarian with an initial list of entity types to consider for capturing; it also provides an initial set of attributes for consideration. The framework may therefore serve as a template that the Librarian may customise before implementation in the FMS.
- The framework would provide prime support to the Librarian in storing innovation-related information in an organised way.
- The way in which the framework can be used to retrieve innovation-related information would greatly facilitate the retrieval of such information by the Librarian in response to information requests by others.
- The framework and accompanying text analytics system will provide the Librarian with powerful tools to better organise and harness the information contained in electronic, textual documents.
- The framework may further aid the Librarian to identify individuals who have certain sought-after traits (e.g. having certain knowledge, skills, interests, etc.) as defined in an enquiry by another party. The Librarian may then put this party in contact with the identified individuals.
- The framework will go further by essentially being a self-help mechanism that would empower other innovation workers to retrieve such information without the constant assistance of the Librarian.

8.10.3 Possible Interactions of the Framer with the Framework

The **Framer** has the task of working with the managers of the target organisation to define appropriate evaluation schemes to be used to assess innovation ideas in a way that is fair, consistent and transparent. The framework may support the Framer in the following ways:

- The framework can assist the Framer in understanding the future direction of the organisation by providing him with a contextualised view of the organisation's strategic objectives, needs and opportunities as embodied in the framework. The Framer may then use these insights to better tailor the evaluation schemes to be aligned with the organisation's future direction.
- As the framework may contain useful details about the employees of the organisation, especially information about their skills, competencies, knowledge areas, and project experience, it can help the Framer to identify which persons should be involved in the evaluation of innovation ideas and concepts to make such evaluation as meaningful as possible.

- The attributes of the Idea and Concept framework entity types can further be expanded to include the evaluation criteria, as defined by the Framer, to make more transparent to the initiators and contributors of such ideas and concept how these would be evaluated. This may proactively increase the quality of ideas and concepts and may decrease the amount of rework required.

8.10.4 Possible Interactions of the Judge with the Framework

In short, the **Judge** uses the evaluation schemes created by the Framer to evaluate innovation ideas and concepts to determine which ideas should move forward, which ideas should be improved before possible re-evaluation, which ideas should be stored for possible consideration in future, and which ideas should be discarded altogether. The Judge may benefit from the framework as follows:

- The attributes that characterise the framework entities of type Idea and Concept, as well as the relationships of such entities with other entities captured in the framework, will provide the Judge with contextualised information assisting him make more informed and speedy decisions.
- The details of the evaluation (e.g. evaluation date, remarks and outcome) of an idea or concept can be captured as part of the respective entities representing such ideas or concepts. This will preserve such information for possible future consideration and further makes the evaluation process more transparent.

8.10.5 Possible Interactions of the Metric Monitor with the Framework

The **Metric Monitor** has the responsibility of defining measures to be used to determine the organisation’s progress and status in terms of innovation. In addition, the Metric Monitor has to monitor such metrics on a continuous basis and further has to suggest improvements or additions to such measures to influence the way in which innovation operations happen. The Metric Monitor examines successes and failures in the target organisation’s innovation processes to recognise patterns of what works and what to date has not worked in terms of innovation. The framework may support the Metric Monitor in the following ways:

- The framework may provide the information required by the Metric Monitor to monitor the respective innovation metrics (e.g. number of ideas captured in the last month, average number of ideas per employee, number of ideas successfully taken to market, etc.).
- The framework can further provide the Metric Monitor with the information required to analyse in order to identify patterns or trends in terms of innovation in the target organisation that can be used to further improve the organisation’s innovation process.
- Using the innovation-related information embodied in the framework, the Metric Monitor can help management to set appropriate and realistic goals with regard to innovation.
- The Metric Monitor can use the framework to determine which skills, competencies and knowledge areas are crucial to innovation. With this knowledge, employees having rare skills, competencies and knowledge can be better managed in an attempt to retain such skills, competencies and knowledge.

8.10.6 Possible Interactions of the Prototyper with the Framework

The function of a **Prototyper** is to quickly develop basic realistic, tangible representations of a product, service or process while embodying a sufficient amount of real life ‘feel’ so that people can physically work with the offering to in turn provide feedback to guide possible future developments. The framework may support the Prototyper in the following ways:

- The framework can provide the Prototyper with contextualised information around the idea or concept that necessitates a prototype. This would allow the Prototyper to better and quicker understand the gist of the idea or concept and may possibly improve the speed of development and quality of the rendered prototype.
- The framework entity types Technology/Method/Tool/Equipment, Skill/Competency and Knowledge Area may further assist the Prototyper to find suitable ways to create the desired prototype as well as to find people with the required skills, competencies or knowledge to aid in prototype development in such areas where the Prototyper lacks certain skills or knowledge.

8.10.7 Possible Interactions of the Scout with the Framework

The fundamental role of a **Scout** is to scan the future to perceive how the relevant industry is likely to change. In addition, the Scout aggregates and synthesises information that may include information about useful future technologies, future market trends, possible future scenarios, and potential threats from competitors. The Scout then brings such aggregated information to the target organisation's innovation team for further analysis. The framework may assist the Scout in many ways, including the following:

- The Scout can use the framework as a mechanism to capture and share the observed trends and findings in context of the organisation's innovation-related entities. This would not only make such information richer for the intended audience, but would also preserve it in an integrated manner for possible future use.
- The Scout may further use the framework to identify important focus areas in which to concentrate his exploration efforts (e.g. specific needs, objectives, competitors, technologies, etc.)
- The framework may further aid the Scout to find appropriate information (e.g. electronic documents) for analysis as a starting point for his exploration.
- Also, the framework can aid the Scout in identifying appropriate people to consult in terms of certain knowledge areas, competitors, needs, technologies, objectives, etc.

8.10.8 Possible Interactions of the Storyteller with the Framework

The **Storyteller**'s tasks are to collect, store, and recount stories about the organisation with the goal to establish a culture of innovation in the target organisation, particularly in the area of allowing people to fail and accepting failure as part of innovation and progress. The framework may assist the Storyteller in the following ways:

- The framework can help the Storyteller to identify which individuals are not taking part in idea generation and idea expansion in order to provide subtle encouragement to such persons.
- The framework can help the Storyteller to identify upcoming stars, in terms of innovation efforts spent and successes achieved, to use as examples to encourage others.
- The framework can provide the Storyteller with the information required to understand the status of innovation in the organisation to be able to accurately recount it to its employees in an effort to further the organisation's innovation culture.

8.10.9 Possible Support of the Framework to Innovation Workers in General

The previous eight sections outlined the support that the framework may provide to the various roles in innovation defined by Hering et al. (2005). This section will discuss how innovation workers in general may use the framework in support of their innovation-related activities. The framework may offer support to the innovation workers of the target organisation in several ways, including:

- Innovation workers can use the framework to log their innovation ideas and subsequently, to monitor the status of such ideas.
- Innovation workers can also use the framework to capture new innovation-related entities, other than ideas, and their inter-relationships with existing framework entities.
- They may further use the framework to view the ideas of others as well as their relationships to other innovation-related entities.
- They may further contribute to the innovation ideas of others using the framework.
- Innovation workers can use the framework and FMS to proactively find innovation-related information based on predefined information needs.[48]
- They can also use the framework reactively find information about certain innovation-related issues in an ad hoc way.[49]
- The framework and accompanying FMS can help innovation workers to uncover innovation-related commonalities that they respectively share with other individuals. Innovation workers may use this knowledge to approach such individuals about the topic in question to trigger face-to-face communication and ultimately the exchange of tacit knowledge.
- Innovation workers may further explore the information embodied in the framework, by browsing through the innovation-related entities and their inter-relationships, to gain a better understanding of the innovation landscape of the organisation.
- Innovation workers can look up the meanings of key business terms, which are stored as part of the framework and made accessible using an appropriate interface, in order to improve their understanding and the quality of their communication in a multi-disciplinary setting.
- Last but not the least, innovation workers may further request business terms or examples of the usage of such terms to be added to the business vocabulary when found absent when searching for the meaning of such business terms.

This concludes the discussion about the value the framework may offer to different types of innovation workers and the different associated working methods. The following section will report on how statistical topic modelling techniques may be used with the framework.

8.11 Using Statistical Topic Modelling to Associate Unstructured Information with the Framework

Statistical topic models were discussed in Chapter 6 as a way to distil and organise the content of textual documents. Two particular topic modelling techniques were identified for further investigation in this research. Firstly, Latent Dirichlet Allocation (LDA) was selected for its ability to formulate topics, as semantic representations of the contents of a set of documents, with little

[48] As discussed in section 8.9.5.
[49] As discussed in section 8.9.6.

guidance from the user. Another reason for choosing LDA was its relative simplicity. LDA can loosely be described as a comprehensive, intelligent clustering mechanism. Secondly, the Concept-Topic Model was further identified for further investigation due to its capability to accept detailed concepts as input to guide the generation of the topic model. The Concept-Topic Model may therefore serve as a classification mechanism. As previously discussed, a combination of clustering and classification is required to best address the discovery of new topics in a set of documents (i.e. using clustering) as well as to use existing knowledge to organise the information embodied in textual documents (i.e. using classification). This section will attempt to give some indication as how these two topic modelling techniques can be used in overlay with the framework to make accessible and contextualise the information embedded in textual documents.

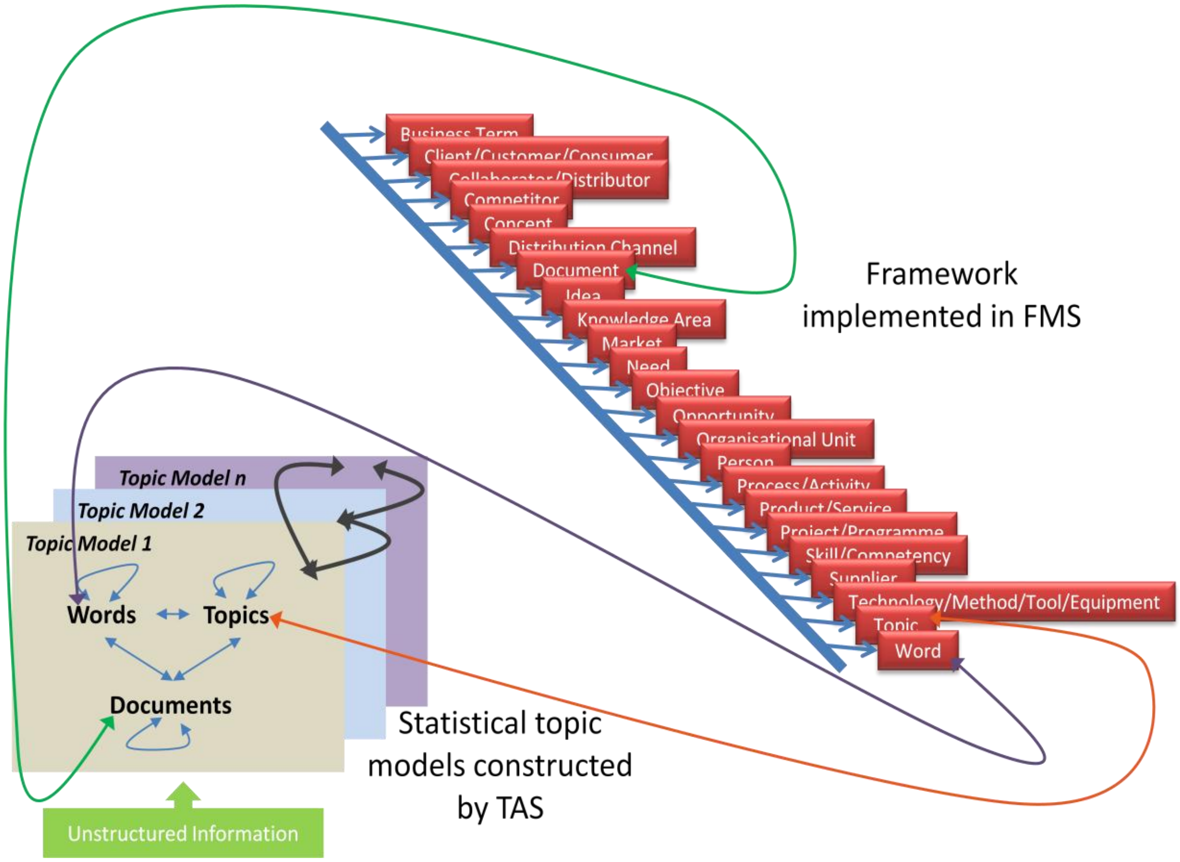


Figure 33: Using statistical topic models to link unstructured information to the framework

Figure 33 illustrates the concept of linking one or more topic models to the framework once it has been implemented in the FMS. A few questions may arise at this stage. Firstly, "Why do we need this framework or any framework?" In short, the framework is required to make explicit, capture and contextualise innovation-related entities in the organisation with the goal to make innovation-related information accessible to innovation workers in a structured way. Secondly, "Why do we bother with unstructured information?" Basically, because a large portion of organisational and innovation-related information is captured and stored in formats having little or no associated structure making it very difficult to intelligently and effectively find and use such information. Thirdly, "Why do we need topic models?" Topic models are used to derive likely and human interpretable structures from unstructured data. These structures are compact-sized abstractions representing the much larger quantity of textual data actually analysed. In this way the quantity of data to consider is significantly reduced without sacrificing too much of the intended meaning. Topic models further apply these structures to organise the unstructured information analysed. By linking the framework entities to the structures contained in topic models, one can use topic models as bridges to organise unstructured information in context of the key innovation-related entities of the organisation embodied in the framework. Lastly, "Why do we need more than one topic model?" Because of limitations in the current level of computer processing power, it is impractical to construct a single topic model using more than 10 000 documents[50] as input, unless one is prepared to wait longer than one day for the analysis to complete. Therefore, the author recommends using more than one topic model analysis to process large document collections and afterwards to derive relationships between the topics in the different topic models as illustrated in section 11.8.

The author recommends the following steps to integrate the results of different topic model analyses with the framework entities:

- One or more persons firstly identify the document collections to be analysed. Although not an absolute necessity, it would be of benefit if the persons who interacted with the respective documents and the dates and nature of such interactions (e.g. created, edited, read, etc.) are available electronically and programmatically accessible[51].
- Assign the document collections to one or more topic model analyses (multiple analyses are required if the sum of all documents to be analysed would exceed the capability of the applicable hardware and/or software if a single topic model analysis were to be used to analyse the entire collection).
- A person decides on the number of topics desired for each such analysis[52]. As a rule of thumb, the more topics one requests, the more specific such topics would become to the extreme were many of the calculated topics amounts to noise due to overfitting. The number of topics requested need not to be the same for different analyses. Also provide values for the other input parameters required by the software in question.
- A person decides if there are specific topics for which one wants to find associated documents for. Such topics have to be predefined by providing words that characterise such topics. The more characterising words provided, the better the quality of the resulting document associations. These predefined topics will serve as inputs to one or more Concept-Topic Model analysis. For each such predefined topic, a person should find a corresponding framework entity type. If not created previously, create an entity of the corresponding entity type. For example, say one of the predefined topics corresponded to one of the organisation's competitors. One would then create and populate an entity of type "Competitor" in the framework to represent this specific competitor. The words charactering

[50] Actually, the number of documents is not a good indicator of corpus size as documents varies in length. It is easier to grasp and measure, however, than the total number of words in such a corpus which may be several million in quantity.

[51] For example, as captured by a document management system.

[52] This is only required when using parametric topic modelling algorithms such as LDA for example.

this competitor, and therefore also the corresponding predefined topic, can be entered as attributes in the “Keywords” field of the corresponding entity.

- Initiate the associated topic model analyses using more than one computer in the case where a given document collection is analysed by more than one analysis. Such parallel analyses would reduce the total processing time of the document collection.
- If the software used provides the analyst with a list of documents that did not render any text directly after the extraction phase, a person can investigate and address such problems where possible.
- As an analysis is completed, a person investigates the formulated topics and document-topic associations provided as part of the analysis output. The person then decides if the results are to satisfaction or if changes should be made to the number of topics requested, the words characterising the predefined topics, etc. Rerun those analyses that did not render satisfactory results.
- For each topic formulated as part of those analyses deemed to be satisfactory, a person then provides a descriptive label. This is of course not necessary for predefined topics as they should already have descriptive labels.
- The FMS then creates entities of type Document for each document in each successful analysis taking care not to create duplicate entities. Using the topic model results, for each such entity representing a specific document in the framework, the FMS then adds as attributes which topics, words and other documents are associated with the document in question including the strengths of such intra-analysis associations.
- Also, the FMS creates entities of type Topic in the framework for each topic formulated in the respective successful analyses. Using the topic model results, the FMS then adds for each such entity representing a specific topic in the framework, attributes representing which documents, words and other topics are associated with the topic in question including the strengths of such intra-analysis associations.
- Likewise, the FMS creates entities of type Word in the framework for each word characterising a topic in the respective successful analyses, while again taking care to avoid the creation of duplicates. Using the topic model results, for each such entity representing a specific word in the framework, the FMS then adds as attributes which topics, documents and other words are associated with the word in question including the strengths of such intra-analysis associations.
- When all documents in all document collections earmarked for analysis have a corresponding entity in the framework, the FMS can calculate the affinity of documents, topics and words that were not part of the same analyses using the respective topic model results. The FMS updates the corresponding document, topic and word entities with the newly calculated inter-analyses affinities and the respective strengths of association.
- If available, the FMS may import references to all metadata and transactional data, including information about who contributed to the writing, initial storing and editing of each document as well as who opened such documents, from the relevant document management systems. Corresponding entities of type Person may subsequently be created, if not yet in existence, for each individual that interacted with a document. Information interest profiles may then be created for each person, based on the topics associated with the documents the person interacted with, and the corresponding Person entities may be updated with such information.
- The FMS then identifies similarities among the different information interest profiles of individuals and subsequently uses such similarities to identify possible relationships between individual persons. The corresponding Person entities are updated to reflect the newly calculated person-person affinities as well as the respective strengths of such associations.
- The FMS may be equipped with a mechanism to compare entities of type Word to identify words that do not have corresponding entities of type Business Term. The FMS may further filter the list of business term candidates (e.g. using number of topics and/or documents associated) to arrive at a list of candidate business terms. A person may then peruse this list and decide which of the candidates can be promoted to business terms. On such

promotion, an entity of type Business Term will be created by the FMS and the required attributes will be filled with data sourced from the person in question.

- Lastly, the FMS can compare each respective entity of type Word, to the attributes representing names, keywords and descriptions of entities having corresponding types other than Word, Document or Topic. The FMS may then suggest candidate relationships between such entities and the respective Word entities. A person may then confirm or refute such suggested relationships. The same process can be repeated to identify candidate relationships between entities of type Document and Topic and entities of types other than Word, Document or Topic.
- The FMS may further suggest additional characterising words for the different predefined topics. This can be done by using the words found to describe the machine-generated topics (e.g. generated by LDA) that were associated with documents that were in turn linked to the predefined topics by the classification technique used (e.g. the Concept-Topic Model). A human may then selected appropriate characterising words for each predefined topic from the list of suggested predefined topics.

These steps may be repeated at regular intervals to ensure that the framework remains recent. The resulting framework will have all the elements and inter-relationships to arrive from structured innovation-related concepts to representative documents, text, topics and vice versa.

The next section presents a preliminary list of functionalities desired for the text analytics system that will be used in overlay with the framework.

8.12 Text Analytics System Functionality Wish List

The following is an incomplete list of functionalities desired of the text analytics system to be used with the framework for exploiting the information contained in the textual documents and other textual information sources. The author discovered the need for these functionalities while dealing with prototype topic modelling software over a period of four years. Although some of these functionalities are somewhat specific to topic modelling based text analytical systems, most of these functionalities have wider applicability. The desired functionalities are discussed under the (slightly modified) main steps of the text mining process discussed in section 5.10.

8.12.1 Data Extraction Functionalities Desired

Data Extraction deals with the actions that are part of the process of extracting textual data from each individual document in the document collection and organising it in the format required as input to the actual analysis. The following functionalities are desired in this regard:

- The system should not by limited to extracting unigrams, but should also be able to extract bigrams, trigrams and even n-grams if desired.
- Automatically divide individual documents into their constituent chapters and sections while preserving the original filename as part of the new file names. This would enable a finer grained analysis as the text analytical tool can now readily find relationships between sections of the same or different documents.
- The text analytics system should not be limited to only processing text written in English.
- The system should be able to extract all dates (e.g. 2009, 11 May 2005, etc.), contained as part of document text, as separate logical units while also recording which date occurred in which document.

- The system should preferably be able to extract all proper nouns (e.g. “John Fourie”, “Table Mountain”, “Microsoft”, etc.) as separate logical units while also recording which proper noun occurred in which document.
- The system should further be able to extract all e-mail addresses contained in documents while also recording which e-mail address occurred in which document
- The system should be able to separately extract the title of each document where applicable while also recording which title occurred in which document.
- The system should be able to separately extract the abstract of each document where applicable while also recording which abstract occurred in which document.
- The system should be able to separately extract the author(s) of each document where applicable while also recording which authors occurred in which document.
- The system should be able to separately extract the headings of each document while also recording which heading occurred in which document.
- The system should automatically detect the language of each document.
- The system should extract and record the file type of each document.
- The system should extract and store the physical location of each document.
- The system should calculate and record the total number of words for each document.
- The system should offer the user the ability to specify one or more regular expressions[53] to be used to extract custom entities corresponding to certain patterns (e.g. telephone numbers) from document text as separate logical units while also recording which custom entity occurred in which document.
- The system should be able to optionally extract hyphenated words from document text.

8.12.2 Data Cleansing Functionalities Desired

Data Cleansing entails the operations required to remove unwanted units from the pool of data to be analysed as well as actions required to increase the quality of such data. The following data cleansing functionalities are desired:

- The system should be able to remove words with low semantic value (e.g. stopwords and common words), as specified by the analyst, from the data to be analysed.
- The system should indicate those documents from which no text could be extracted.
- The system should identify and indicate duplicate as well as near-duplicate documents, based on the content of such documents, with the option to remove such duplicates.
- The system should give the user the option to apply stemming to the input data.
- The system should offer the user the ability to specify one or more regular expressions to be used to eliminate custom entities corresponding to certain patterns from the input data.
- The system should give the user the ability to optionally exclude numbers from the input data.

8.12.3 Data Analysis Functionalities Desired

Data Analysis entails the actions that are part of the process of analysing the input data to compose distilled output information. The following functionalities are desired in this regard:

- The system should optionally be able to accept a list or hierarchy of seed topics, consisting of characterising words, and use these predefined topics to find associated documents as well as the strengths of such associations.

[53] A ‘regular expression’, or ‘regex’ for short, is a formal way of describing a template or pattern for a text string (e.g. telephones numbers, dates, e-mail addresses, etc.). A regular expression captures, in general terms, what features the text must have to conform to the template it specifies.

- The system should be able to automatically determine the optimal number of topics to use for a given analysis.
- For each word associated with the topics calculated by the system, the system should also give the strength of each such association (e.g. the probability calculated for each word-topic pair).
- Rendered topics should not only contain unigrams, but optionally also bigrams and trigrams to facilitate interpretability.
- The system should be able to give an indication of confidence for each of the topics calculated.
- For each topic calculated, the system should give the documents associated with the respective topic as well as the association strength for each document-topic pair.
- The system should indicate how well each calculated topic is represented in the source document set relative to other calculated topics.
- For each topic, the system should indicate which other topics are closely related and should include the estimated strength for each such topic-topic association.
- The system should calculate which set of words best describe the contents of each document in the collection.
- The system should automatically create a summary for each document in the collection.
- The system should cluster alike documents together.
- The system should have the ability to calculate an information interest profile per person based on the document set associated with the specific individual.
- The system should further have the ability to use the calculated interest profile of an individual to determine the following:
    - Additional documents strongly associated with the person including the strengths of such associations.
    - Topics strongly associated with the person including the strengths of such associations.
    - Words strongly associated with the person including the strengths of such associations.
    - Other persons strongly associated with the person including the strengths of such associations.
- The system should lastly be able to automatically suggest human interpretable labels for each topic calculated.
- The system should have the ability to update the results on a continuous basis as new information is added to the respective input document collection. This should include redoing the analysis using the same parameters but the updated document set in cases where the document set has changed dramatically. When only a few changes occurred in the document set, the system should be able to update the existing results by integrating the new or changed documents (a process known as “document fold-in”; refer to section 6.5.7 for more information about this process).

8.12.4 Visualisation Functionalities Desired

Visualisation in this setting deals with the way in which the outputs of the text analytics system are presented to the users. In this respect, the following functionalities are desired:

- On a general level, the system should enable users to review, manipulate, search, explore, filter, and understand large volumes of textual data.
- The system should be able to present an overview of the information contained in selected sources by showing what underlying topics or concepts are addressed in the set of input documents.
- The system should optionally be able to represent a hierarchy of concepts or topics extracted from the input document set with the most general topics at the top of the hierarchy and more specific topics at the bottom.

- When focusing on a topic, the system should display hyperlinks representing all words, documents and other topics associated with the specific topic to enable easy navigation.
- When focusing on a document, the system should show hyperlinks representing all the words, topics and other documents associated with that document to facilitate navigation.
- When focusing on a word, the system should display hyperlinks representing all topics, documents and other words associated with the specific word to facilitate navigation.
- The system should be able to show a list of all words (excluding stopwords), extracted from source information, which should include frequency of occurrence, associated topics, associated documents and some score indicating the estimated level of generality of each word compared to other words in the list.
- The system should be able to apply stemming to words characterising topics, documents and other words in a reversible way.
- Wherever lists of topics, words or documents are shown, the system should have a filtering mechanism to reduce such lists to the items matching the supplied filtering criterion.
- In addition to linear lists, the system should have alternative views using colour and spatial arrangements to display data (e.g. tag clouds).
- The system should be able to use the document inter-relationships calculated as part of the analysis to plot a document affinity graph where highly similar documents are positioned closely together and vice versa.
- Similarly, the system should be able to use the topic inter-relationships calculated as part of the analysis to plot a topic affinity graph where highly similar topics are positioned closely together and vice versa.
- Also, the system should be able to use the inter-relationships between the information interest profiles, calculated for different individuals as part of the analysis, to plot a person affinity graph where persons having highly similar information interest profiles are positioned closely together and vice versa.

8.12.5 Document- and Information Retrieval Functionalities Desired

Information Retrieval and Document Retrieval deals with the operations required to find and return information matching the user’s information requests. Functionalities envisaged here are:

- The system should have a unified search capability where strongly related topics, documents and other words are returned in response to a query entered by the user.
- The system should be able to retrieve and display the content of the respective source document wherever document filenames are displayed as part of the results.
- The system should offer the user the ability to search for topics, documents, or words across different analyses result sets.
- The system should further have the ability to look up the meanings of words from Internet-based sources (e.g. online dictionaries, gazetteers, encyclopaedias, etc.).
- The system should have the ability to identify documents that span a combination of topics specified by the user (i.e. “bridging documents”).
- The system should be able to identify documents that significantly address more than one topic.
- The system should allow the user to export all words and documents having custom entered descriptions.
- The system should be able to export the results of an analysis to a spreadsheet to enable the user to further manipulate such results.
- The system should offer functionality that would allow the user to export the custom developed thesauri to one or more spreadsheets or XML files.

8.12.6 Information Enrichment Functionalities Desired

Enrichment deals with the actions performed by the user with the goal to improve the interpretability or depth of the results rendered by the text analytics system. The following functionalities are desired to this end:

- For each topic calculated by the system, have the ability to manually supplement the topic with a human interpretable, descriptive label.
- The system should offer the user the ability to add a description to any document.
- The system should offer the user the ability to add a usefulness score to individual documents.
- The system should offer the user the ability to add the creation date of the document for individual documents.
- The system should offer the user the ability to add the authors of the document for individual documents.
- The system should offer the user the ability to add the source or bibliographical information for individual documents.
- The system should offer the user the ability to compare topics between different analysis result sets.
- The system should offer the user the ability to mark additional stopwords to be ignored in all subsequent analyses or alternatively specific subsequent analyses.
- The system should give the user the ability to mark documents that should be ignored in subsequent analyses.
- The system should give the user the ability to add a description to any word or term encountered in the results.
- The system should have functionality to assist the user in building one or more custom thesauri based on words selected from one or more analyses.

8.12.7 Information Sharing Functionalities Desired

Information Sharing deals with how the system can be used to proactively share relevant information with others. Here, the following functionalities are envisaged:

- The text analytics system should have the ability to use the calculated information interest profiles of different individuals to identify potentially useful, ‘unseen’ documents per individual for possible exchange between individuals.
- The system should allow users to adjust their information interest profiles when required.

As a general requirement, the desired text analytics system must have suitable programmatic interfaces available for other information systems to be able to programmatically address this system to realise seamless integration.

8.13 Summary

This chapter presented the culmination of this research by combining concepts of the disciplines of Knowledge Management, Innovation, Text Analytics and Business Metadata into a framework and accompanying processes.

The preliminary elements of the framework, in the form of entity types, intended to organise, contextualise and make accessible innovation-related information, were described. Note that these entity types are definitely not exhaustive and may be expanded as the framework is implemented in different organisations in future.

The author further shared his views on the information systems that would typically be involved in the implementation of the framework in the target organisation. High-level use cases of the framework were further discussed and some thoughts were presented on how the framework and accompanying framework management system might be used by persons fulfilling different innovation roles in the organisation. Although such innovation roles do not form part of the actual framework entity types currently defined, it proved to be a useful way of investigating the potential application of the framework in the target organisation from different viewpoints corresponding to the different innovation roles.

Also, some views on how statistical topic modelling techniques may be used in concert with the developed framework were conveyed.

Lastly, some important functionalities of the text analytics system to be used as part of the implementation of the proposed framework were presented.

The following chapters report on selected case studies and experiments using topic models. Although the framework was not implemented and evaluated in its entirety during this study, the case studies that follow shed some light on the first two research objectives stated in Chapter 2, namely:

1. Develop a way to promote the dissemination of information in organisational documents.
2. Develop a way to capture and make accessible indications of expertise of individuals in the innovating organisation to promote communication and knowledge transfer between employees.

At the end of each of the following three case studies, the relevance of the specific case study to the proposed framework will be discussed.

9. Preamble to the Case Studies

During the period 2005 to 2009, the author was engaged in numerous investigations and applications of text analytical techniques to unstructured information contained in electronic documents. Each of these experiences contributed to the author's understanding of the challenges and possible solutions in this field and therefore directly or indirectly contributed to the outcomes of this study.

Chapters 10, 11 and 12 present the details of three selected investigations dealing with some aspect of the automated analysis of electronic documents and further inference using the results of such analyses. These case studies were selected to shed some light on parts of the process supporting the framework as well as on some of the possible working methods when dealing with the framework once implemented. All investigations undertaken and reported about here were governed by the availability of information to analyse, the accessibility of experts familiar with the domains implied by information analysed, as well as by the business requirements expressed by Indutech (where the author has been employed for the duration of this study) where these investigations were performed.

The framework was not implemented and evaluated in its entirety in this study due to the following reasons:

- The unavailability of one or more participating organisations and access to their innovation workers and innovation-related information stores.
- Sensitivity issues around the innovation-related information of a participating organisation.
- The absence of access to a suitable technology platform to host the framework.
- Lack of resources to develop the required interfaces between the platform hosting the framework and the required text analytical system.
- The immaturity of the text analytical system used in terms of interfacing with other systems (e.g. CAT currently does not have an API to facilitate such integration).
- Lastly, implementing and evaluating the proposed framework in its entirety would have caused a substantial increase in the scope of the study.

The three case studies that follow therefore focus more on the mechanisms required to distil and organise information in such a way as to 'feed' the proposed framework. However, in each of these case studies the implications in terms of the greater framework will be discussed.

Table 20 provides an overview of the case studies that will be addressed in the next three chapters. Also, Appendix C provides details about the characteristics of the software used in the respective investigations.

Research Objectives	Case Study 1 EE Group	Case Study 2 CIRP	Case Study 3 Wildlife Research
Develop a means to promote the dissemination of information contained in organisational documents to the members of a specified group.	Documents were identified for exchange between participants using calculated research interest profiles.	The LDA technique was investigated as a possible means to automatically cluster documents based on their contents.	The Concept-Topic Model technique was investigated as a possible means to automatically classify documents based on their contents.
Develop a means to capture and make accessible indications of the expertise of individuals in the innovating organisation to promote communication and knowledge transfer between employees	Research interest profiles were calculated based on documents associated with individual participants.	The results of the LDA technique were extended to capture and represent the topic profiles of authors by using explicit author information.	Concepts, capturing the interests of one or more person, were defined and used to find appropriate documents.
Develop a means to realise and make accessible a 'corporate memory' with regard to innovation with the aim to unify structured and unstructured innovation-related information	Investigated the usefulness of the LDA technique as a means to update and populate the framework by organising unstructured information by detecting underlying topics.	Investigated the usefulness of the LDA technique as a means to update and populate the framework by organising unstructured information by detecting underlying topics.	The Concept-Topic Model technique was investigated as a possible means to populate and update the framework information associated with specific concepts.

Table 20: Overview of the three case studies

10. Case Study 1: The Enterprise Engineering Group

The first case study focused on the Enterprise Engineering Group (EE Group), a division of the Department of Industrial Engineering of Stellenbosch University. The high-level aims of the case study were firstly to establish whether it is possible to construct information profiles for different individuals based on a collection of personal documents, and secondly to establish whether it is possible to use such information profiles to identify candidate documents for exchange between different individuals.

Both Indutech and the EE Group are actively doing research on knowledge management, enterprise engineering and innovation-related issues and have substantial documentation on these knowledge areas. The possibility of identifying potentially useful documents for exchange between individuals based on the documents these individuals interact with was further investigated.

This case study explored the following:

- **Whether it is possible to automatically "calculate" a research interest profile for different persons, based on documents associated with the respective persons, using topic models.**
- **Establishing the quality of the resulting research interest profiles.**
- **Finding a mechanism to identify similarities between different research interest areas of the respective persons.**
- **Testing whether the calculated topics can be interpreted by humans.**
- **Establishing whether significantly similar research interest areas of different persons indeed indicate similarity in their research and possibly their tacit knowledge.**
- **Finding a mechanism to identify documents for exchange between persons having similar research interests.**
- **Evaluating the usefulness of the exchanged documents with the receiving parties.**
- **Establishing whether it is possible to calculate a participant keyword profile for individual persons using a topic model.**

At the time of the experiment four members of the EE group were engaged in PhD studies in the fields of innovation and knowledge management, while another member was in the process of completing his Master's degree in essentially the same fields. These five individuals participated in this experiment.

The first phase of this case study entailed the following steps:

1. Collect the research documentation from each of the participants while maintaining a reference of what documents were collected from whom.
2. Calculate the respective research interest areas addressed by the individual members as well as by the EE Group in general.
3. Identify the similarity between the various research interest areas calculated.
4. Identify significant similarities between participants based on shared research interest areas for potential tacit knowledge transfer between such individuals.

5. For each participant, identify documents of other participants that are potentially relevant to that specific individual.
6. Construct participant keyword profiles for each of the individual participants.
7. Evaluate the interpretability of the identified research interest areas with the assistance of the applicable participants. For each candidate research interest area the participant should give a descriptive label (where possible) and also indicate whether the relevant research area is part of his core research. Calculated research interest areas that cannot be given descriptive labels are further marked accordingly.
8. Identify the significant "unseen" documents per participant by taking into account the participant's core research areas as identified in step 7 as well as which documents he already have.
9. Validate the usefulness of the identified, potentially useful "unseen" documents with the relevant participants.

The following sections will discuss each of these steps in greater detail.

10.1 Collecting Research Documentation

Table 21 summarises the documentation collected from the various participants. In total 1942 files were considered for the analysis, while 1703 of the files rendered text for extraction and were successfully processed. The reasons why some documents could not be processed by the analysis software[54] includes:

- No text could be extracted from some files due to protection on the respective files (e.g. password protected files)
- No text could be extracted from some files because these files do not contain programmatically recognisable text (e.g. pdf files that contain pages that are actually images)

The total size of the documents submitted for analysis amounted to 1950MB. For each participant, all documents relating to his respective research project to date were collected.

Participant	# Files Processed	# Files Successfully Processed	Total Size (MB)
Participant 1	209	181	443
Participant 2	166	149	111
Participant 3	715	613	576
Participant 4	564	479	654
Participant 5	288	281	166
TOTAL	1942	1703	1950

Table 21: Characteristics of the documents collected from EE Group participants

10.2 Identifying Research Interest Areas

For each participant, a 20-topic LDA analysis was performed individually in order to estimate the knowledge areas underlying the participant's research documentation. The number of topics to use, twenty, was selected after doing a few exploratory runs and investigating the specificity of

[54] See Appendix C for the characteristics of the software used for these analyses.

the topics returned. A total of 100 topics (i.e. 20 topics for each of the five participants) were thus generated by doing the five individual LDA analyses. It was assumed that each topic, generated by the LDA topic modelling technique, represented a candidate research interest area. The same stoplists and parameters were used for all five analyses.

The software used for the analyses, an in-house developed prototype by Indutech, characterised each topic with 40 words (unigrams and bigrams more specifically). All results were returned in a *Microsoft Excel* spreadsheet. These results are comprised of the topic-word matrix and the document-topic matrix. For illustrative purposes, Table 22 shows the first topic from the 20 topics represented in the topic-word matrix calculated by the LDA topic modelling technique for each participant[55]. According to the calculated topic model, the words listed first in a given topic are more significant in terms of characterising the topic compared to those listed lower down. The version of the software used for this case study did not explicitly quantify the significance of the individual characterising words of the respective topics.

Topic 1 Participant 1	Topic 1 Participant 2	Topic 1 Participant 3	Topic 1 Participant 4	Topic 1 Participant 5
development	concept	methods	system	innovation
system	map	design	systems	network
innovation	maps	method	engineering	networks
linux	knowledge	process	requirements	research
social	concept maps	model	team	technology
internet	concepts	problem	design	development
community	immediate solution	surface	js ol	companies
technology	learning	selection	process	communication
network	students	product	data	innovations
practice	data	engineering	development	industry
time	information	time	objective	management
project	performance	system	program	technologies
communities	reporting	evaluation	technical	knowledge
computer	mapping	value	management	organization
source	problems	engineers	generation donnors	business
technological	novak	criteria	implementation	programme
knowledge	lane ibm	classification	analysis	work
systems	applications	paper	product	organizational
work	science	development	risk	information
...	...	...	...	...

Table 22: Examples of one out of twenty topics calculated for each participant

As part of the LDA results, the document-topic matrix indicates which documents are associated with which topics as well as the respective document-topic association weights for each

[55] The number of characterising words shown per topic in this table is limited to the first 20 of the possible 40 words returned by the software to conserve space.

document-topic combination. Note that a document can be associated with one or many topics and that a topic may have one or many associated documents.

For each document analysed, the topic model contains a vector comprised of an array of values (i.e. the document-topic weights or the so-called mixture ratios) equal to the number of topics specified for the analysis (i.e. 20 topics in this case). Each value (0 ≤ value ≤ 1) in this array represents the extent to which the document corresponds to each of the generated topics. A value of 1 indicates the strongest possible association between a document and a topic, a value of 0.5 indicates a relationship of average strength, while a value of 0 indicates that there are no association whatsoever between the document and the given topic[56]. Documents that were not processed successfully (due to reasons explained earlier) have equal weights for all topics. The sum of all array values for a given document always amounts to one. Table 23 shows an extract from the document-topic matrix of one of the five analyses.

Document	Topic 1	Topic 2	Topic 3	…	Topic 20
Faceted classification_WIKIPEDIA.pdf	0.000	0.014	0.008	…	0.000
Relationship Analytics_COGITO.pdf	0.300	0.344	0.030	…	0.111
Part-of-speech Tagging_WIKIPEDIA.pdf	0.000	0.003	0.017	…	0.004
Managing Collective Intelligence_ZARA.pdf	0.000	0.007	0.000	…	0.000
Euclidean Embedding of Co-occurence data_GLOBERSON.pdf	0.000	0.000	0.003	…	0.000
The Business of Information Sharing (1)_SIERRA_SYSTEMS.pdf	0.027	0.593	0.003	…	0.006
Ontologies and Folksonomies Can They Coexist_FAY.pdf	0.001	0.791	0.002	…	0.035
Probabilistic Latent Semantic Indexing_HOFMANN.pdf	0.000	0.000	0.854	…	0.000
…	**…**	**…**	**…**	**…**	**...**
Disruptive technology roadmaps_KOSTOFF.pdf	0.691	0.001	0.017	…	0.001
Normalised Sum of Topic-Documents Weights (%)	**4.72%**	**14.33%**	**8.77%**	…	**4.06%**

Table 23: Extract of the document-topic matrix generated as part of the LDA results

The normalised sum of the document-topic weights for a given topic indicates how well the specific topic is represented, relative to other topics, in the relevant document collection. For example, it can be seen in Table 23 that Topic 2 is better represented in the given document collection than Topic 20 (since 14.33% >> 4.06%).

10.3 Identifying Similarities between Research Interest Areas

As mentioned earlier, the 20 topics identified per participant are considered to be estimators of the research interest areas the relevant participant collected information about. It was decided to

[56] This corresponds to soft clustering as discussed in section 5.13.2.

use correlation to measure the level of similarity between the different topics of the different participants with the goal to identify shared research interest areas. All calculations for this case study were done in *Microsoft Excel* and were based on the results of the five topic models generated by the software in question.

Firstly, all topics generated for the five participants were consolidated in one spreadsheet resulting in a total of 100 topics. In order to use correlation, the individual topics first needed to be quantified in some way. To achieve this, all the words characterising each of the 100 topics were combined into a global list (or a corpus vocabulary) and duplicate terms were eliminated. Table 24 gives the total number of characterising words per participant, the total number of unique words per participant, the total number of words for all participants as well as the total number of unique words for all participants.

Participant	# Words	# Unique Words	Total # Words	Total # Unique Words
Participant 1	800[57]	340	4000	1294
Participant 2	800	439		
Participant 3	800	398		
Participant 4	800	411		
Participant 5	800	412		
	4000	**2000**		

Table 24: Number of words in the individual and combined LDA results

The average number of unique words per participant was calculated as 400 with a standard deviation of 36.7 words. Of the 2000 words that were unique per participant (the sum of the unique words for the five participants), only 1294 words were globally unique among all participants. These 1294 words were then used to create the global word list.

Next, for each word in the global word list the frequency of occurrence per participant was determined. The total frequency of occurrence for each word in the global word list was then calculated as the sum of the occurrence frequencies per participant. For example, in Table 25 the word “activity” occurred once in Participant 1’s 20 topics and once in Participant 4’s 20 topics resulting in a total frequency of two.

For each word in the global word list, it was further determined in which of the 100 topics the specific word occurred. In Table 25, it can be seen that the word “activity” specifically occurred in Topic 1 of Participant 1 (the fact that it also occurred in Topic 7 of Participant 4 is not shown in this table due to size restrictions).

[57] Twenty topics per participant with forty characterising words per topic results in 800 characterising words per participant (which are not necessarily all unique as a given word may occur in more than one topic of a given participant and also in more than one topic when considering the topics of all participants).

Term	Topic 1 P1	Topic 2 P1	…	Topic 19 P5	Topic 20 P5
_______________	0	0	…	0	0
µm	0	0	…	0	0
abstract	0	0	…	0	0
...	…	…	…	...	…
ceo	0	0	…	0	0
change	1	1	…	0	0
...	…	…	…	...	…
information	1	1	…	1	1
information overload	0	0	…	0	0
...	…	…	…	...	…
network	1	0	...	0	1
networks	1	0	…	0	1
…	…	…	…	...	…
ys	0	0	…		0
	40	**40**	**…**	**40**	**40**

Table 25: Extended global word list

The extended global word list, shown in Table 25, therefore represents a binary-valued vector for each of the 100 topics with each such vector having 1294 dimensions corresponding to the number of words in the global word list. Each of the 100 topics is made up of 40 words from this global word list.

Next, a symmetrical matrix was drawn up having a total number of rows and columns equal to the total number of topics resulting in a 100 x 100 matrix. For each cell in this matrix, the correlation between the topic represented by the given row and the topic represented by the given column was calculated using *Microsoft Excel*'s correlation function using the binary-valued vectors of the respective topics as input. The calculations were based on the makeup of the 100 topics in terms of the 1294 possible words as represented by the global word list discussed earlier. A part of this matrix is shown in Table 26.

	Topic 1 P1	Topic 2 P1	Topic 3	Topic 4	Topic 5 P1	…	Topic 20 P5
Topic 1 P1	1	0.1670778	0.4102298	0.2199413	0.324635941	…	0.27012056
Topic 2 P1	0.1670778	1	0.2370973	0.1540103	0.096888577	…	0.36309089
Topic 3 P1	0.4102298	0.2370973	1	0.3198167	0.265410601	…	0.14055562
Topic 4 P1	0.2199413	0.1540103	0.3198167	1	0.181125993	…	0.12609970
Topic 5 P1	0.3246359	0.0968886	0.2654106	0.181126	1	…	0.16371169
…	…	…	…	…	…	…	…
Topic 20 P5	0.27012056	0.36309089	0.14055562	0.12609970	0.16371169	…	1
Normalised Sum of Topic-Correlation Weights (%)	**1.234%**	**1.593%**	**1.492%**	**1.493%**	**1.126%**	**…**	**1.059%**

Table 26: Extract of the topic-topic correlation matrix indicating topic similarities

It can be seen that each topic has the strongest possible correlation (i.e. a value of 1) with itself since it consists of the exact same words. Subsequently, for each of the 100 topics, the three topics with the highest correlation score (excluding the topic in question itself) were determined. All topics pairs having a correlation score of more than 0.50 were viewed as having significant[58] overlaps. The 0.5 threshold value was determined empirically. All such significantly overlapping topics were then identified. Significantly overlapping topics involving topics of two participants signify possible shared research interest areas between such participants. Significantly overlapping topics associated with a single participant signify closely related research areas of that participant and may also indicate that these topics may be viewed as a single topic. The latter occurrence can possibly imply that 20 topics may have been too many topics given the variety of the participant's research documentation.

The sums of the correlation scores of each topic were further calculated and normalised. The five topics with the highest and lowest overall normalised correlation score sum were then identified. The sum of the correlation scores is a dimensionless figure that gives one an indication of how unique or general a given topic is compared to other topics.

	1st	2nd	3rd	4th	5th
Minimum	**Topic 9 P5**	**Topic 20 P4**	**Topic 16 P3**	**Topic 3 P2**	**Topic 13 P3**
	-0.8484898	0.69090191	0.93335021	1.91914045	2.73131102

Table 27: Topics with the lowest overall correlation score sums

The two topics with the lowest correlation score sums turned out to be meaningless topics after closer investigation. The topics with the third, fourth and fifth lowest sums on the other hand, dealt with very specific concepts (i.e. "Cutting - a machining process", "A tool for documenting human rights violations", and "Grinding – a machining process").

	1st	2nd	3rd	4th	5th
Maximum	**Topic 17 P1**	**Topic 14 P1**	**Topic 14 P4**	**Topic 5 P3**	**Topic 18 P1**
	0.75935599	0.69700317	0.66861449	0.64619468	0.62623824

Table 28: Topics with the highest overall correlation score sums

Topics having relatively high correlation score sums generally indicate that these topics overlap with many other topics in terms of characterising words. The topics with the five highest sums dealt with more general knowledge management and innovation related concepts (e.g. "Knowledge Management Tools", "Knowledge Generation/Learning", etc.).

[58] By no means does the term "significant' used here imply statistical significance.

10.4 Identify Significant Similarities between Participants

Using the topic correlation scores discussed in section 10.3, shared research interest areas between the various participants were determined by, for each topic, identifying topics of other participants having correlation scores of greater or equal to 0.50 with the topic in question. This was done for each topic of the 100 topics. The tables below represent the significant overlapping topics for each participant with the respective correlation scores[59].

Topic in Focus	Best Match	Score	2nd Best Match	Score
Topic 2 P1	**Topic 7 P5**	0.56316484	**Topic 19 P2 ***	0.55443699
Topic 3 P1 *	**Topic 18 P5 ***	0.66861449	-	-
Topic 4 P1	**Topic 8 P4 ***	0.50146628	-	-
Topic 5 P1 *	**Topic 10 P5 ***	0.50043210	-	-
Topic 14 P1	**Topic 14 P5**	0.59177282	-	-
Topic 15 P1 *	**Topic 14 P3**	0.58466176	-	-
Topic 18 P1	**Topic 14 P3**	0.53165509	**Topic 10 P5 ***	0.50043210

Table 29: Topics significantly similar to Participant 1's topics

Topic in Focus	Best Match	Score	2nd Best Match	Score
Topic 9 P2 *	**Topic 19 P5**	0.64619468	-	-
Topic 12 P2	**Topic 11 P3 ***	0.75935599	-	-
Topic 19 P2 *	**Topic 7 P5**	0.59632836	**Topic 2 P1**	0.55443699

Table 30: Topics significantly similar to Participant 2's topics

Topic in Focus	Best Match	Score	2nd Best Match	Score
Topic 6 P3	**Topic 11 P4**	0.62623824	-	-
Topic 11 P3 *	**Topic 12 P2**	0.75935599	-	-
Topic 14 P3	**Topic 15 P1 ***	0.58466176	**Topic 18 P1**	0.5316551

Table 31: Topics significantly similar to Participant 3's topics

Topic in Focus	Best Match	Score	2nd Best Match	Score
Topic 8 P4 *	**Topic 4 P1**	0.50146628	-	-
Topic 11 P4	**Topic 6 P3**	0.62623824	-	-

Table 32: Topics significantly similar to Participant 4's topics

[59] Topics marked with an "*" character corresponds to topics corresponding to key research interest areas of the different participants as identified by them later in this case study.

Topic in Focus	Best Match	Score	2nd Best Match	Score
Topic 7 P5	**Topic 19 P2 ***	0.59632836	**Topic 2 P1**	0.56316484
Topic 10 P5 *	**Topic 5 P1 ***	0.50043210	**Topic 18 P1**	0.50043210
Topic 14 P5	**Topic 14 P1**	0.59177282	-	-
Topic 18 P5 *	**Topic 3 P1 ***	0.66861449	-	-
Topic 19 P5	**Topic 9 P2 ***	0.64619468	-	-

Table 33: Topics significantly similar to Participant 5' topics

The average topic correlation score of all possible topic correlations was calculated as 0.126 with a standard deviation of 0.137. The topics of a given participant may further have significant correlations with other topics of the same participant. This occurrence indicates overlaps in the research interest areas of the specific individual. Table 34 shows some examples of this occurrence.

Topic in Focus	Best Match	Score	2nd Best Match	Score
Topic 9 P1	**Topic 10 P1**	0.69700317	-	
Topic 10 P1	**Topic 9 P1**	0.69700317	**Topic 19 P1**	0.50352327
Topic 19 P1	**Topic 10 P1**	0.50352327	-	
Topic 4 P4 *	**Topic 14 P4 ***	0.50488048	-	
Topic 14 P4 *	**Topic 4 P4 ***	0.50488048	-	

Table 34: Topics of a participant having significant similarities with other topics of the same participant

Table 35 summarises the significant similarities between the research interest areas of the five participants.

	Participant 1	Participant 2	Participant 3	Participant 4	Participant 5	# Other Participants Overlapping With
Participant 1	3	1	2	1	5	**4**
Participant 2	1	0	1	0	2	**3**
Participant 3	2	1	0	1	0	**3**
Participant 4	1	0	1	2	0	**2**
Participant 5	5	2	0	0	0	***2***
# Overlaps with other Participants' Research Interest Areas	**9**	**4**	**4**	**2**	**7**	

Table 35: Summary of significant topic overlaps between participants

Participant 1 has overlapping research interest areas with all other participants, but has large overlaps with Participant 5 (i.e. five out of twenty topics). Participant 2 has the largest overlap with Participant 5 (two out of twenty topics), while Participant 3 has the largest overlap with Participant 1 (two out of twenty topics). Participant 4 has limited overlaps with Participant 1 and

Participant 3 (one out of twenty topics). Participant 5 has a large overlap with the research interest areas of Participant 1 (five out of twenty topics) and to a lesser extent overlaps with Participant 2's research interest areas (two out of twenty topics). Figure 34 illustrates the affinity among the five participants based on the calculated number of significantly overlapping research areas among the five participants.

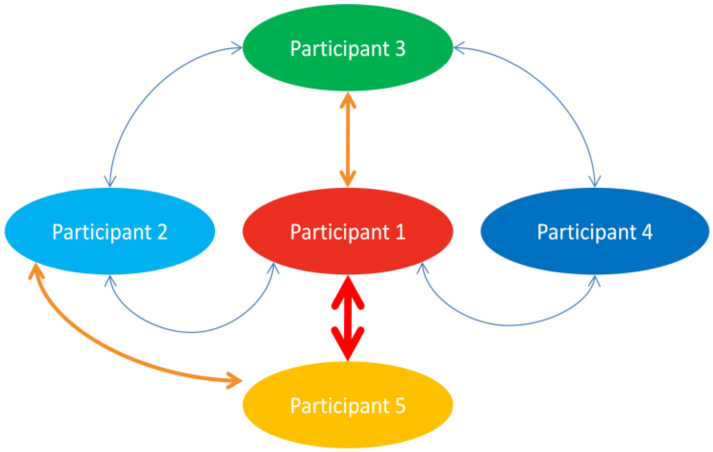


Figure 34: Affinities between participants' research interest areas

10.5 Identify Potentially Relevant "Unseen" Documents for each Participant

Using the significant overlaps between topics of different participants, as calculated in the previous section, the documents associated with these topics can now be determined. Once identified, these documents may then be distributed to the other participants that have been identified as having significant overlaps with the topic in question. This is done for all topics that have significant overlaps with topics of other participants. For each topic identified as having significant overlaps with the topics of other participants, documents having document-topic weights[60] (i.e. mixture ratios) greater than or equal to 0.90[61] were identified. As example, Table 36 shows the documents significantly[62] associated with Topic 2 of Participant 1. The table indicates that this topic has two significant overlaps with topics of other participants, namely Topic 7 of Participant 5 and Topic 19 of Participant 2. Therefore, the four documents identified in this table are candidates for sharing with Participants 2 and 5 under the following conditions:

- These documents do not already form part of the research corpus of the identified two participants.

[60] Refer to section 10.2 for an explanation of the relevance of document-topic weights.
[61] This threshold value was once more determined empirically.
[62] Once more, the term "significantly" does not imply statistical significance here.

- The respective matching topics (i.e. Topic 7 of Participant 5 and Topic 19 of Participant 2) form part of the two participants' core research interest areas as opposed to research areas of little importance to these individuals.

Topic 2 P1	Springer-Verlag Ontology Management Semantic Web, Semantic Web Services, and Business Applications.pdf		0.998
	knowledge-mining-proceedings-of-the-nemis-2004-final-conference.9783540250708.27387.pdf		0.996
	Management of dynamic knowledge.pdf		0.967
	Wiley Towards The Semantic Web Ontology-Driven Knowledge Management.pdf		0.993
Share With	**Participant 5**	**Participant 2**	
Matching Topic	**Topic 7 P5**	**Topic 19 P2**	

Table 36: Documents significantly associated with Topic 2 of Participant 1

In section 10.7 the topics corresponding to core research areas of the individual participants are identified. This was achieved by each individual participant evaluating each of the twenty calculated topics corresponding to his candidate research interest areas. Once it is known which topics of each individual participant correspond to core research interest areas, these topics can be examined to find the topics of other participants significantly overlapping with such topics. Subsequently, the documents significantly attributed to such significantly overlapping topics can be identified. Such documents will have possible relevance to the individual participants associated with these topics and are therefore candidates for exchange between the participants involved. Once such documents are identified, a further check may be done to ensure that the identified documents were not part of the target candidate's research corpus. The reader is referred to section 10.8 for the identification of such core, significant documents.

10.6 Constructing Participant Keyword Profiles

A participant keyword profile characterises the individual participant in terms of his research interest and consist out of a number of characterising words. The participant keyword profile of each individual participant was constructed as follows:

1. Compile a word list using the first three charactering words of each of the twenty topics calculated per participant.
2. Sort this list alphabetically.
3. Change all plural words to their singular form.
4. Remove all obvious meaningless words from the list.
5. Eliminate all duplicate words from the list and keep count of the frequency of occurrence of each duplicate word.

Table 37 shows the resulting participant keyword profiles for the five participants in terms of the keywords characterising each individual along with the occurrence frequency of such keywords.

#	Participant 1	Freq	Participant 2	Freq	Participant 3	Freq	Participant 4	Freq	Participant 5	Freq
1	knowledge	11	information	10	design	10	innovation	7	information	8
2	management	6	ontology	7	knowledge	5	business	6	knowledge	8
3	innovation	5	data	4	management	4	knowledge	3	document	4
4	information	4	knowledge	4	process	4	management	3	network	4
5	system	3	term	4	product	4	process	3	topic	3
6	creativity	2	map	3	information	3	development	2	data	2
7	firm	2	topic	3	system	3	engineering	2	innovation	2
8	research	2	concept	2	enterprise	2	market	2	management	2
9	business	1	document	2	method	2	product	2	ontology	2
10	capital	1	model	2	ontology	2	project	2	semantic	2
11	community	1	tag	2	power po	2	software	2	term	2
12	cost	1	class	1	project	2	technology	2	web	2
13	data	1	controlled	1	tool	2	appraisal	1	controlled	1
14	development	1	event	1	architecture	1	architecture	1	journals	1
15	enterprise	1	information overload	1	business	1	coach	1	map	1
16	group	1	management	1	cde	1	company	1	model	1
17	growth	1	manufacturing	1	cutting	1	competency	1	papers	1
18	justification	1	name	1	design process	1	countries	1	people	1
19	learning	1	overload	1	dr	1	enterprise	1	research	1
20	market	1	owl	1	family	1	indicator	1	services	1
21	network	1	person	1	grinding	1	information	1	similarity	1
22	ontology	1	product	1	manufacturing	1	involvement	1	tags	1
23	organization	1	query	1	model	1	model	1	technology	1
24	practice	1	ranking algorithm	1	problem	1	organization	1	user	1
25	product	1	set	1	qoc	1	performance	1	word	1
26	technology	1	warehouse	1	surface	1	project management	1	-	-
27	user	1	web	1	time	1	requirement	1	-	-
28	venture	1	xml	1	-	-	scenario	1	-	-
29	venture capital	1	-	-	-	-	system	1	-	-
30	work	1	-	-	-	-	tool	1	-	-
31	-		-	-	-	-	workforce	1	-	-
		57		**60**		**59**		**55**		**54**

Table 37: Participant keyword profiles of the individual participants

Table 38 summarises the sizes of the individual participant keyword profiles. For example, Participant 2's participant keyword profile consists of 28 unique words occurring with total frequency of 60.

	Participant 1	Participant 2	Participant 3	Participant 4	Participant 5
# Unique Words in Profile	30	28	27	31	25
Total Word Frequency	57	60	59	55	54

Table 38: Size of the individual participant keyword profiles

These participant keyword profiles represent easily understandable overviews of the participants' (estimated) research interests. The calculated participant keyword profiles were validated by die respective participants as discussed in section 10.7.

To calculate the overlaps between the various participant keyword profiles, the same process explained in section 10.3 was followed. Firstly, a global word list consisting of the unique words occurring in the various participants' keyword profiles, along with their respective frequencies of occurrence, was constructed. This list consisted of 91 words that cumulatively occurred 285 times in the respective participant keyword profiles.

Table 39 shows an extract of the global participant keyword profile word list.

#	Terms	P1	P2	P3	P4	P5	Total
1	appraisal	0	0	0	1	0	1
2	architecture	0	0	1	1	0	2
3	business	1	0	1	6	0	8
4	capital	1	0	0	0	0	1
5	cde	0	0	1	0	0	1
6	class	0	1	0	0	0	1
7	coach	0	0	0	1	0	1
…	..	…	…	…	…	…	…
31	indicator	0	0	0	1	0	1
32	information	4	10	3	1	8	26
33	information overload	0	1	0	0	0	1
34	innovation	5	0	0	7	2	14
35	involvement	0	0	0	1	0	1
…	...	…	…	…	…	…	…
88	word	0	0	0	0	1	1
89	work	1	0	0	0	0	1
90	workforce	0	0	0	1	0	1
91	xml	0	1	0	0	0	1
		57	**60**	**59**	**55**	**54**	**285**

Table 39: Extract of the global participant keyword profile word list

Table 40 shows the correlation scores of the research interest profiles of the five participants.

	P1	P2	P3	P4	P5
P1	1	0.311403	0.364064	0.463965	0.628738
P2	0.311403	1	0.18104	-0.04486	0.69262
P3	0.364064	0.18104	1	0.215267	0.227903
P4	0.463965	-0.04486	0.215267	1	0.149485
P5	0.628738	0.69262	0.227903	0.149485	1
Normalised Correlation Score Sums	24.33%	18.81%	17.47%	15.68%	23.72%

Table 40: Correlation scores of the individual semantic research profiles

Participant 1's profile significantly correlates with Participant 5's profile, while Participant 2's profile also significantly correlates Participant 5's profile. Participant 3's profile has no significant correlations with other participants' keyword profiles. Participant 4's profile significantly correlates with Participant 1's profile, while Participant 5's profile has fairly strong correlations with Participant 1's and Participant 2 keyword profiles. Looking at the normalised correlation score sums per participant it can be seen that Participant 1 has the most overlaps with other participants' keyword profiles, closely followed by Participant 5, while Participant 4 has the least overlaps with other participants' keyword profiles. This corresponds closely with the affinities between the individual participant's research interest areas shown in Figure 34.

10.7 Evaluate Identified Research Interest Areas/Profiles with Participants

The objectives of this step are as follows:

- Identify a label for each topic
- Identify meaningless topics
- Identify topics corresponding to the participant's core research areas
- Evaluate appropriateness of the calculated participant keyword profiles

These objectives were achieved by evaluating each of the twenty calculated topics, as well as the calculated participant keyword profile with the respective participants individually. The topics were evaluated by inspecting each topic individually in terms of its characterising words and the relevant order of such words. Words that the individual considers as key to identifying the topic's label were highlighted. In some cases the filenames of documents associated with a topic are further investigated for possible clues to the identity of the topic. For really elusive topics, a number of the documents associated with the topic were opened and the headings and overview of its content investigated. On average it took about five minutes to evaluate and label a topic.
Table 41 shows the topics of Participant 1 and the descriptive labels identified by evaluating the individual topics. Topics that involved much difficulty in assigning a descriptive label are marked using the "#" character.

Topic	Label	Core?
Topic 1 P1	Role of Internet Communities in Open Innovation	N
Topic 2 P1	Ontology and Semantic Web Technologies and Applications	N
Topic 3 P1	Innovation Networks	**Y**
Topic 4 P1	Product Lifecycle Management | Problem Solving Techniques	N
Topic 5 P1	Knowledge Networks and Communities of Practice	**Y**
Topic 6 P1	Resolving Knowledge Inconsistencies	N
Topic 7 P1	Acceptability of Knowledge	N
Topic 8 P1	Organisational Innovation and the Impact on Growth	N
Topic 9 P1	Innovation Advancement Policies on Institutional Level	N
Topic 10 P1	Innovation Advancement Policies on Institutional Level	N
Topic 11 P1	Knowledge Management in Organisations	**Y**
Topic 12 P1	Guidelines for Scientific Writing | Innovation and Growth [#]	N
Topic 13 P1	Sharing Expertise	N
Topic 14 P1	Knowledge Generation/Learning	N
Topic 15 P1	Setting up Knowledge Networks	**Y**
Topic 16 P1	The Role of Creativity in Innovation	N
Topic 17 P1	Knowledge Management Tools	N
Topic 18 P1	Knowledge Management	N
Topic 19 P1	Making Innovation Sustainable	N
Topic 20 P1	Investing in Innovation	N

Table 41: Evaluation results for Participant 1's research interest areas

Participant 1 identified four topics (i.e. Topics 3, 5, 11 and 15) as part of his core research focus for his PhD study. He also indicated that the word "open" was missing as a characterising word for Topics 1 and 3. On closer inspection the word "open" was found to be part of the stoplist used for the analyses with the implication that it was removed before the analysis started.

Two topics (i.e. Topics 4 and 12) contained characterising words associated with dual themes that were not trivially reconcilable even after investigating the significantly associated documents. These two topics therefore have dual labels to cater for this observation (i.e. composite topics). It was fairly difficult to find a label for Topic 12. The documents with the highest document-topic weights for almost all topics generated for Participant 1 were further found to be electronic books. Such books in most cases consist out of chapters with varied focus. Since a book is considered as a single document by the analysis software, the wide scope of such books influenced the nature of the formulated topics. Electronic articles were also associated with topics, but with lower document-topic weights. No truly meaningless topics were identified for Participant 1.

Table 42 shows the topics of Participant 2 and the descriptive labels identified by evaluating the individual topics.

Topic	Label	Core?
Topic 1 P2	Concept Maps | Concept Mapping (minor presence)	**Y**
Topic 2 P2	Modelling information based on relations	**Y**
Topic 3 P2	Events Standard Formats (outlier topic)	N
Topic 4 P2	Information Overload	**Y**
Topic 5 P2	Formal Ontologies	N
Topic 6 P2	Information Management and Decision Support Systems	**Y**
Topic 7 P2	Web Ontology Applications & Technologies	N
Topic 8 P2	Information Searching & Retrieval	**Y**
Topic 9 P2	Thesauri and Controlled Vocabularies	**Y**
Topic 10 P2	Database Design and Utilisation	N
Topic 11 P2	Topic Maps | Global Information Use	**Y**
Topic 12 P2	Information Management in Product Development (outlier topic)	N
Topic 13 P2	Information Search & Retrieval	**Y**
Topic 14 P2	Information Modelling	**Y**
Topic 15 P2	Topic Maps |Semantics	**Y**
Topic 16 P2	Multi-domain Ontology Implementation	**Y**
Topic 17 P2	Concepts and Linking between Concepts	**Y**
Topic 18 P2	Enterprise Data Warehousing | Participation in Virtual Communities of Practise	**Y**
Topic 19 P2	Ontology Applications / Implementation Methodologies	**Y**
Topic 20 P2	(Information Context,) Tagging / Folksonomies	**Y**

Table 42: Evaluation results for Participant 2's research interest areas

Out of the 20 topics, 15 topics were identified to be core topics in terms of Participant 2's research project. Four composite topics were further identified (Topics 1, 11, 15 and 18), while no meaningless topics were identified. Topics 3 and 12 were identified as outlier topics in terms of his research interest.

Table 43 shows the topics of Participant 3 and the descriptive labels identified by evaluating the individual topics.

Topic	Label	Core?
Topic 1 P3	Classification and Selection of Design Processes | Creativity?	**Y**
Topic 2 P3	Product Development Process/Models | Software Tools	**Y**
Topic 3 P3	Detailed Design Process	**Y**
Topic 4 P3	Problem Solving Analysis/Synthesis in Engineering Design	**Y**
Topic 5 P3	Knowledge and Information Management (in Engineering Design)	N
Topic 6 P3	Enterprise Design/Enterprise Reference Architectures	N
Topic 7 P3	Decision Making Process in Design	N
Topic 8 P3	Production Management and Control	N
Topic 9 P3	Information Handling by Designers	**Y**
Topic 10 P3	Product Family and Product Platform Design	N
Topic 11 P3	Information Handling in the Production Process	**Y**
Topic 12 P3	Business Aspects of Engineering Companies	**Y**
Topic 13 P3	Fabrication Processes	N
Topic 14 P3	Knowledge Management in Companies	N
Topic 15 P3	Design Tools	**Y**
Topic 16 P3	Material Removal Machining Processes	N
Topic 17 P3	Knowledge Structures/Ontologies | Problem Solving/Decision-making	N
Topic 18 P3	Design and Product Development Process	**Y**
Topic 19 P3	Knowledge Management in Companies	N
Topic 20 P3	Product Development Process/Models	**Y**

Table 43: Evaluation results for Participant 3's research interest areas

Out of the 20 topics generated, 10 topics were identified as core topics in terms of Participant 3's research. No meaningless topics and three composite topics (Topics 1, 2 and 17) were identified. Table 44 shows Participant 4's topics and the descriptive labels identified after evaluation.

Topic	Label	Core?
Topic 1 P4	Business Management and Leadership	N
Topic 2 P4	Noise Topic (contains elements of Corporate Communication/Knowledge Networks) $^{\#}$	N
Topic 3 P4	Management Maturity Models (PM3)	**Y**
Topic 4 P4	Frameworks for Innovation	**Y**
Topic 5 P4	Disruptive Innovation, Change and Planning for the Future	**Y**
Topic 6 P4	Business Development Maturity Model	N
Topic 7 P4	Software Engineering Best Practices	N
Topic 8 P4	Business Strategies/Business Models for Innovation $^{\#}$	**Y**
Topic 9 P4	Innovation Output Indicators/Metrics	**Y**
Topic 10 P4	System Engineering	N
Topic 11 P4	Design/Enterprise Architectures	N
Topic 12 P4	People Capability Maturity Model	N
Topic 13 P4	Software Capability Maturity Model	**Y**
Topic 14 P4	Innovation Frameworks	**Y**
Topic 15 P4	Capability Maturity Model Integration™	**Y**
Topic 16 P4	Knowledge Management Practices	**Y**
Topic 17 P4	Management and Innovation Management	N
Topic 18 P4	Stakeholder/People Involvement	**Y**
Topic 19 P4	e-Business Coaching	N
Topic 20 P4	Noise Topic (containing several Acronyms)	N

Table 44: Evaluation results for Participant 4's research interest areas

Out of the 20 topics generated, 10 topics were identified as core topics in terms of Participant 4's PhD research. Two meaningless topics (i.e. Topics 2 and 20) and no composite topics were identified. It proved to be very difficult to find descriptive labels for Topic 2 and 8.

Lastly, Table 45 shows Participant 5's topics and the descriptive labels identified after evaluation.

Topic	Label	Core?
Topic 1 P5	Innovation Networks	**N**
Topic 2 P5	Expertise Management/Matching/Location	**Y**
Topic 3 P5	Text Mining	**Y**
Topic 4 P5	Information Categorisation	**Y**
Topic 5 P5	Knowledge Technologies and Knowledge Representation	N
Topic 6 P5	Text Mining Using Database Tomography	N
Topic 7 P5	Ontology Management and Ontology Applications	N
Topic 8 P5	Topic Maps | Dublin Core	N
Topic 9 P5	Noise Topic	-
Topic 10 P5	Knowledge Networks and Knowledge Transfer	**Y**
Topic 11 P5	Topic Modelling Techniques	**Y**
Topic 12 P5	KM and Collective Intelligence	N
Topic 13 P5	Text mining and Knowledge Discovery	**Y**
Topic 14 P5	Infrastructure for Knowledge Management	N
Topic 15 P5	Information Extraction and Searching	**Y**
Topic 16 P5	Information Systems and Knowledge Sharing	**Y**
Topic 17 P5	Applications of Semantics and Natural Language Processing	N
Topic 18 P5	The Role of Intangible Networks in Innovation	**Y**
Topic 19 P5	Controlled Vocabularies | Concept Maps	N
Topic 20 P5	Natural Language Processing/Text mining | Tagging/Tag Clouds #	N

Table 45: Evaluation results for Participant 5' research interest areas

Out of the 20 topics generated, 9 topics were identified as core topics in terms of Participant 5's PhD research. One meaningless topic was identified (i.e. Topic 9). Three composite topics were further identified (i.e. Topics 8, 19 and 20). Lastly, it was challenging to find a descriptive label for Topic 20.

Table 46 shows the significant topic overlaps among participants where at least one of the overlapping topics corresponds to a topic representing a core research interest area. In the case where both overlapping topics represent core research interest areas a "2" is shown in Table 46 whereas a "1" represents occurrences where only one of the two overlapping topics corresponds to a core research interest area of a participant.

Source → Sink ↓	Participant 1	Participant 2	Participant 3	Participant 4	Participant 5	Sink for how many Participants?
Participant 1	-	1	0	1	2+2+1	**3**
Participant 2	0	-	1	0	0	**1**
Participant 3	1	0	-	0	0	**1**
Participant 4	0	0	0	-	0	**0**
Participant 5	2+2	1+1	0	0	-	**2**
Source for how many participants?	**2**	**2**	**1**	**1**	**1**	
# Overlaps with other Participants (core to core)	**2**	**0**	**0**	**0**	**2**	
# Overlaps with other Participants (core to non-core)	**1**	**3**	**1**	**1**	**1**	

Table 46: Summary of core topic overlaps between participants

Figure 35 is a refined version of Figure 34 as it only includes overlaps between topics where at least one of the topics represent a core research interest area. In this figure, the arrowhead points in the direction of the sink participant (i.e. the eventual, potential recipient of the identified documents). Lines with arrowheads on both sides represent the scenario where a core research interest area of one participant is matched with a core research interest area of another participant. This implies that these participants are both a source and a sink.

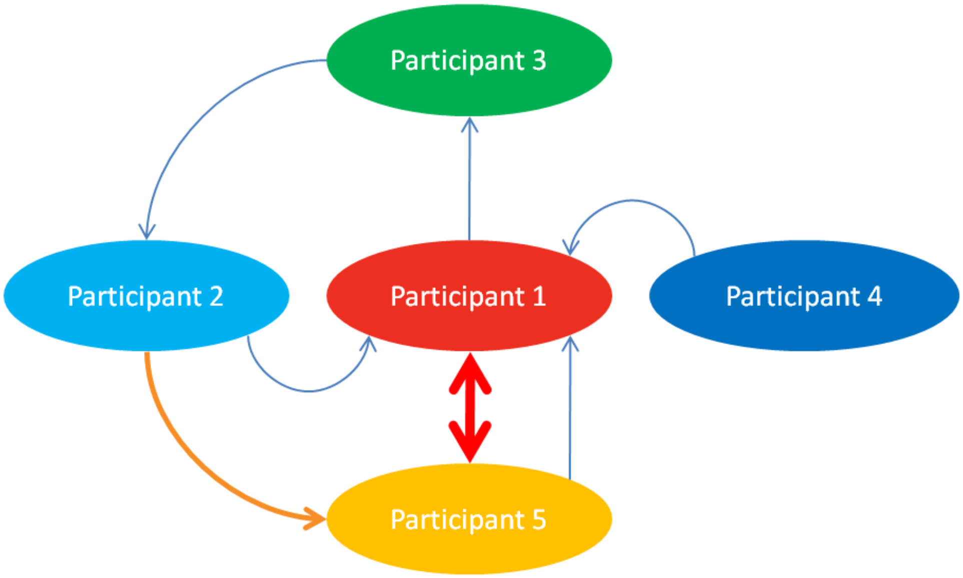


Figure 35: Affinities between participants' core research interest areas

Table 47 shows the participant keyword profiles of the five participants. Words in bold print represent those words that the individual participants indicated as being key to describing their respective research foci at that point in time.

#	Participant 1	Freq	Participant 2	Freq	Participant 3	Freq	Participant 4	Freq	Participant 5	Freq
1	**knowledge**	11	**information**	10	**design**	10	**innovation**	7	**information**	8
2	**management**	6	**ontology**	7	**knowledge**	5	**business**	6	**knowledge**	8
3	**innovation**	5	**data**	4	**management**	4	**knowledge**	3	**document**	4
4	**information**	4	**knowledge**	4	**process**	4	**management**	3	**network**	4
5	system	3	**term**	4	**product**	4	**process**	3	**topic**	3
6	**creativity**	2	map	3	**information**	3	development	2	data	2
7	firm	2	**topic**	3	system	3	engineering	2	**innovation**	2
8	**research**	2	**concept**	2	enterprise	2	**market**	2	**management**	2
9	business	1	document	2	**method**	2	**product**	2	ontology	2
10	capital	1	model	2	ontology	2	**project**	2	**semantic**	2
11	**community**	1	**tag**	2	power po	2	software	2	**term**	2
12	cost	1	class	1	**project**	2	**technology**	2	web	2
13	data	1	controlled	1	tool	2	**appraisal**	1	controlled	1
14	development	1	event	1	architecture	1	architecture	1	journal	1
15	enterprise	1	**information overload**	1	business	1	coach	1	**map**	1
16	**group**	1	**management**	1	cde	1	company	1	**model**	1
17	growth	1	manufacturing	1	cutting	1	**competency**	1	paper	1
18	justification	1	**name**	1	**design process**	1	countries	1	**people**	1
19	learning	1	**overload**	1	dr	1	**enterprise**	1	research	1
20	market	1	**owl**	1	family	1	**indicator**	1	service	1
21	**network**	1	person	1	grinding	1	**information**	1	**similarity**	1
22	ontology	1	product	1	manufacturing	1	**involvement**	1	tags	1
23	**organization**	1	**query**	1	**model**	1	model	1	technology	1
24	**practice**	1	**ranking algorithm**	1	problem	1	**organization**	1	user	1
25	product	1	set	1	qoc	1	**performance**	1	word	1
26	technology	1	warehouse	1	surface	1	**project management**	1	-	-
27	user	1	web	1	time	1	requirement	1	-	-
28	venture	1	**xml**	1	-	-	scenario	1	-	-
29	venture capital	1	-	-	-	-	system	1	-	-
30	work	1	-	-	-	-	tool	1	-	-
31	-		-	-	-	-	workforce	1	-	-
		57		**60**		**59**		**55**		**54**

Table 47: Results of the participants keyword profiles evaluation

Table 48 summarises the number of the candidate participant keyword profile words the individual participants regarded as being central in describing their research foci.

	Participant 1	Participant 2	Participant 3	Participant 4	Participant 5
# Unique Words in Profile	30	28	27	31	25
Core Words Identified by Participant	11 (36.7%)	16 (57.1%)	10 (37.0%)	18 (58.1%)	13 (52%)

Table 48: Summary of number of core words per participant keyword profile

On overall, 48% (i.e. 68 of 141) of the calculated semantic research profile words were deemed as core words in terms of the respective participant keyword profiles. It seems that a somewhat accurate keyword profile of an individual can be automatically compiled from the topics calculated using the documentation such an individual has interacted with.

10.8 Identifying the Significant "Unseen" Documents Matching the Participants' Core Research Interest Areas

This step entailed trying to identify possible documents for exchange between participants using the topics associated with their respective core research interest areas. The process of identifying such documents was briefly discussed in section 10.5, but is repeated here:

- For each topic identified as corresponding to a core research interest area, find topics[63] of other participants having a significant overlap with such topics.
- For each topic identified in this way, find the documents significantly associated with the respective topic by indentifying those documents related to the topic in question with a document-topic weight of 0.90 or larger.
- For each document identified for possible exchange, ensure that the intended recipient indeed did not already have the document in question. This was done manually by opening each document in question, copying an entire sentence from the text and performing a full-text search against the intended participant's research document corpus. When no matches are returned, the intended recipient indeed does not have the document in question. If one or more matches have been returned, the matching documents were investigated to verify whether they are indeed the same as the document in question.[64]

The following tables represent documents[65], corresponding to a non-core research area of one participant that corresponded to a core research area of another participant. Topics corresponding to core research interest areas are once more indicated by an "*".

Owner's topic label: "***Ontology and Semantic Web Technologies and Applications***"

Topic label of recipient's corresponding topic: "***Ontology Applications / Implementation Methodologies****"

	Filename	Weight	Seen?
Topic 2 P1	Springer-Verlag Ontology Management Semantic Web, Semantic Web Services, and Business Applications.pdf	0.998	N
	knowledge-mining-proceedings-of-the-nemis-2004-final-conference.9783540250708.27387.pdf	0.996	N
	Management of dynamic knowledge.pdf	0.967	N
	Wiley Towards The Semantic Web Ontology-Driven Knowledge Management.pdf	0.993	N
Share With	Participant 2		
Matching Topic	Topic 19 P2		

Table 49: Documents of Participant 1 to potentially share with Participant 2

Owner's topic label: "***Product Lifecycle Management | Problem Solving Techniques***"

[63] These may be core or non-core topics.

[64] The reason for not simply using the filenames of documents to establish uniqueness is that documents may have the same content, but different filenames.

[65] Note that filenames are presented exactly as sourced from the participants without correcting spelling.

Topic label of recipient's corresponding topic: "***Business Strategies/Business Models for Innovation****"

	Filename	Weight	Seen?
Topic 4 P1	The Logistic Innovation Approach and the theory of inventive problem solving.pdf	0.909	N
	indeng_v17_n2_a6.pdf	0.916	N
	IdeationBrainstorming.pdf	0.963	N
	Global_Product.pdf	0.999	N
Share With	Participant 4		
Matching Topic	Topic 8 P4		

Table 50: Documents of Participant 1 to potentially share with Participant 4

Owner's topic label: "***Knowledge Management***"

Topic label of recipient's corresponding topic: "***Knowledge Networks and Knowledge Transfer****"

	Filename	Weight	Seen?
Topic 18 P1	Schwartz D. Encyclopedia of Knowledge Management.pdf	0.888[66]	N
Share With	Participant 5		
Matching Topic	Topic 10 P5		

Table 51: Documents of Participant 1 to potentially share with Participant 5

Owner's topic label: "***Information Management in Product Development***"

Topic label of recipient's corresponding topic: "***Information Handling in the Production Process****"

	Filename	Weight	Seen?
Topic 12 P2	Thesis.pdf	1	Y
Share With	Participant 3		
Matching Topic	Topic 11 P3		

Table 52: Documents of Participant 2 to potentially share with Participant 3

Owner's topic label: "***Knowledge Management in Companies***"

Topic label of recipient's corresponding topic: "***Setting up Knowledge Networks****"

[66] Although 0.888 is lower than the threshold value of 0.90, no documents were associated with the relevant topic with a value of 0.90 or higher. The document with the highest affinity was selected in this case.

	Filename	Weight	Seen?
Topic 14 P3	Total Knowledge Management.pdf	0.974	N
	comprehensive KM.pdf	0.955	N
	Knowledge Management-A state of the art guide.pdf	0.999	N
	Frameworks for knowledge a contribution towards conceptual c.pdf	0.95	N
	Organizing_Knowledge.pdf	0.997	N
	KM Emerging Discipline Rooted in a Long History.pdf	0.992	N
	Ch 7 Competencies 06.08.02.pdf	0.942	N
	The Intelligent Enterprise and KM.pdf	0.993	N
Share With	Participant 1		
Matching Topic	Topic 15 P1		

Table 53: Documents of Participant 3 to potentially share with Participant 1

Owner’s topic label: “***Ontology Management and Ontology Applications***”

Topic label of recipient’s corresponding topic: “***Ontology Applications / Implementation Methodologies****”

	Filename	Weight	Seen?
Topic 7 P5	Managing Ontologies A Comparative Study of Ontology Servers_AHMAD.pdf	0.991	N
	Computational Ontologies and Information Systems Foundations_KISHORE.pdf	0.993	Y
	Computational Ontologies and Information Systems Formal Specification_SHARMAN.pdf	0.994	Y
	Ontology Management Semantic Web, Semantic Web Services, and Business Applications_HEPP.pdf	0.999	N
Share With	Participant 2		
Matching Topic	Topic 19 P2		

Table 54: Documents of Participant 5 to potentially share with Participant 2

Owner’s topic label: “***Controlled Vocabularies | Concept Maps***”

Topic label of recipient’s corresponding topic: “***Thesauri and Controlled Vocabularies****”

	Filename	Weight	Seen?
Topic 19 P5	Matching Knowledge Elements in Concept Maps using a Similarity Flooding Algorithm_MARSHALL.pdf	0.95	N
	Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies_NISO.pdf	1	Y
	Afrikaans-English cross-language information retrieval_COSIJN.pdf	0.991	N
Share With	Participant 2		
Matching Topic	Topic 9 P2		

Table 55: Documents of Participant 5 to potentially share with Participant 2

The following tables represent documents, corresponding to a core research area of one participant that corresponded to a core research area of another participant.

Owner’s topic label: “***Innovation Networks****”

Topic label of recipient’s corresponding topic: “***The Role of Intangible Networks in Innovation****”

	Filename	Weight	Seen?
Topic 3 P1	Never_Ending_Friending_April_2007.pdf	0.983	N
	ISTE Publishing Innovation Engineering The Power of Intangible Networks.pdf	0.939	Y
	GIO_2005_for-printing.pdf	0.965	N
	von Hippel Con Krogh - Free revealing and the private collective model.pdf	0.965	N
	Innovation Engineering.pdf	0.938	Y
	The MIT Press Democratizing Innovation.pdf	0.983	N
	GIO_2005.pdf	0.958	N
	Horizontal innovation networks.pdf	0.952	N
Share With	Participant 5		
Matching Topic	Topic 18 P5		

Table 56: Documents of Participant 1 to potentially share with Participant 5

The documents “GIO_2005_for-printing.pdf” and “GIO_2005.pdf” turned out to be the same document with different filenames. Only one of these was therefore recommended to the relevant participant for validation. The same applied to the documents “ISTE Publishing Innovation Engineering The Power of Intangible Networks.pdf” and “Innovation Engineering.pdf”.

Owner’s topic label: “***Knowledge Networks and Communities of Practice****”

Topic label of recipient’s corresponding topic: “***Knowledge Networks and Knowledge Transfer****”

	Filename	Weight	Seen?
Topic 5 P1	TeiglandthesisKnowledgeNetworking.pdf	0.997	N
	Communities of practice and organizational performance.pdf	0.906	N
	lesser.pdf	0.942	N
	Dualities, distributed communities of practice and knowledge management.pdf	0.949	N
Share With	Participant 5		
Matching Topic	Topic 10 P5		

Table 57: Documents of Participant 1 to potentially share with Participant 5

The documents “Communities of practice and organizational performance.pdf” and “Lesser.pdf” turned out to be the same document with different filenames. Only one of these was therefore recommended to the relevant participant for validation.

Owner’s topic label: “***Knowledge Networks and Knowledge Transfer****”

Topic label of recipient’s corresponding topic: “***Knowledge Networks and Communities of Practice****”

	Filename	Weight	Seen?
Topic 10 P5	Knowledge Networks Mapping and Measuring Knowledge Creation_KREBS.pdf	0.975	N
	Creating knowledge networks- lessons from practice_SCHONSTROM.pdf	0.979	Y
	Decision Making Challenges in 'Co-opetitive Learning and Knowledge Exchange Networks'_LOEBBECKE.pdf	0.986	N
	Information Processing Model of Information Technology Adaptation - An intra-organizational diffusion perspective_COOPER.pdf	0.996	N
	ICT and Information Flow Theory_YUDONG.pdf	0.98	N
	Mining Version Histories to Verify the Learning Process of Legitimate Peripheral Participants_HUANG.pdf	0.952	N
	Knowledge Transfer within Interorganisational Networks_PRIESTLEY.pdf	0.998	N
	Uses of information sources in an Internet-era Firm_QUAN-HAASE.pdf	0.985	N
	Using Bayesian Agents to Enable Distributed Network Knowledge A Critique_POTGIETER.pdf	0.981	N
	Information Sharing in Innovation Networks_PRIESTLEY.pdf	0.959	N
	Knowledge Networks Innovation through Communities of Practice_HILDRETH.pdf	0.991	Y
	Outcome Ambiguity in Inter-organizational Knowledge Transfer Do Various Network Forms Make a Difference_SAMMADAR.pdf	0.947	N
	Enabling Complex Adaptive Process through KM_RUGGLES.pdf	0.953	N
Share With	Participant 1		
Matching Topic	Topic 5 P1		

Table 58: Documents of Participant 5 to potentially share with Participant 1

Owner's topic label: "***The Role of Intangible Networks in Innovation****"

Topic label of recipient's corresponding topic: "***Innovation Networks****"

	Filename	Weight	Seen?
Topic 18 P5	Innovation Engineering The Power of Intangible Networks_CORSI.pdf	0.977	Y
	Inter- and intraorganisational Learning processes in the interaction between firms and patent offices_CHRISTENSEN.pdf	0.969	N
Share With	Participant 1		
Matching Topic	Topic 3 P1		

Table 59: Documents of Participant 5 to potentially share with Participant 1

Table 60 summarises, for each participant, the number of documents that participant has to share with other participants and also the number of documents that were matched to the specific participant for evaluation. Values in parentheses indicate the number of documents that the respective recipient already had in his possession. Participant 1 had the most documents identified for sharing with other participants (17 documents), closely followed by Participant 5 (16 documents). Participant 2 and Participant 4 had no documents identified for sharing.

Conversely, Participant 1 had the most documents to evaluate (20 documents), followed by Participant 5 (9 documents). Participant 3 had no documents to evaluate.

		Recipient					
		P1	P2	P3	P4	P5	Total docs to share
Source	P1	-	4 (0)	0 (0)	4 (0)	9 (1)	**17**
	P2	0 (0)	-	0 (1)	0 (0)	0 (0)	**0**
	P3	8 (0)	0 (0)	-	0 (0)	0 (0)	**8**
	P4	0 (0)	0 (0)	0 (0)	-	0 (0)	**0**
	P5	12 (3)	4 (3)	0 (0)	0 (0)	-	**16**
	Total docs to evaluate	**20**	**8**	**0**	**4**	**9**	

Table 60: Summary of documents shared and to be received by participant

The following section reports on the perceived usefulness of the individual documents exchanged among the participants.

10.9 Validating the Usefulness of the Exchanged Documents

The final step in this case study, was for the participants to evaluate the usefulness of the documents identified as matching their core research areas. An e-mail message containing the following information was sent to each recipient participant:

- The label of the source participant's topic that the documents were associated with. This label was included to provide context to the attached documents.
- The label of the recipient's topic that was determined to be significantly similar to source recipient's topic. Again this label was included to provide context to the documents.
- A table containing the filenames of the documents to be evaluated.
- As part of this table, columns to score the usefulness of each document as either "Irrelevant", "Interesting but not Useful" or "Useful".
- The actual files to be evaluated were added as attachments to the e-mail.

One e-mail message was sent to each recipient for each topic of another participant that significantly matched one of his core research interest areas. The name of the source participant of the relevant documents was not mentioned to the recipient in order to prevent any potential bias concerning the usefulness of the documents to be evaluated.

Table 61 contains Participant 1's feedback concerning the usefulness of the eight documents that matched one of his core topics, labelled "***Setting up Knowledge Networks****". The source participant's corresponding topic was labelled "***Knowledge Management in Companies***".

Filename	Irrelevant	Interesting, but not useful	Useful
Total Knowledge Management.pdf		X	
comprehensive KM.pdf		X	
Knowledge Management-A state of the art guide.pdf		X	
Frameworks for knowledge a contribution towards conceptual c.pdf	X		
Organizing_Knowledge.pdf		X	
KM Emerging Discipline Rooted in a Long History.pdf			X
Ch 7 Competencies 06.08.02.pdf		X	
The Intelligent Enterprise and KM.pdf			X

Table 61: Evaluation results of documents evaluated by Participant 1 (1 of 3)

Table 62 contains Participant 1's feedback concerning the usefulness of the eleven documents that matched one of his core topics, labelled "***Knowledge Networks and Communities of Practice****". The source participant's corresponding topic was labelled "***Knowledge Networks and Knowledge Transfer****".

Filename	Irrelevant	Interesting, but not useful	Useful
Knowledge Networks Mapping and Measuring Knowledge Creation_KREBS.pdf		X	
Decision Making Challenges in 'Co-opetitive Learning and Knowledge Exchange Networks'_LOEBBECKE.pdf			X
Information Processing Model of Information Technology Adaptation - An intra-organizational diffusion perspective_COOPER.pdf			X
ICT and Information Flow Theory_YUDONG.pdf		X	
Mining Version Histories to Verify the Learning Process of Legitimate Peripheral Participants_HUANG.pdf		X	
Knowledge Transfer within Interorganisational Networks_PRIESTLEY.pdf			X
Uses of information sources in an Internet-era Firm_QUAN-HAASE.pdf		X	
Using Bayesian Agents to Enable Distributed Network Knowledge A Critique_POTGIETER.pdf		X	
Information Sharing in Innovation Networks_PRIESTLEY.pdf			X
Outcome Ambiguity in Inter-organizational Knowledge Transfer Do Various Network Forms Make a Difference_SAMMADAR.pdf			X
Enabling Complex Adaptive Process through KM_RUGGLES.pdf		X	

Table 62: Evaluation results of documents evaluated by Participant 1 (2 of 3)

Table 63 contains Participant 1's feedback concerning the usefulness of a single document that matched one of his core topics, labelled "***Innovation Networks****". The source participant's corresponding topic was labelled "***The Role of Intangible Networks in Innovation****".

Filename	Irrelevant	Interesting, but not useful	Useful
Inter- and intraorganisational Learning processes in the interaction between firms and patent offices_CHRISTENSEN.pdf	X		

Table 63: Evaluation results of documents evaluated by Participant 1 (3 of 3*)*

Table 64 contains Participant 2's feedback concerning the usefulness of two documents that matched one of his core topics, labelled "***Thesauri and Controlled Vocabularies****". The source participant's corresponding topic was labelled "***Controlled Vocabularies | Concept Maps***".

Filename	Irrelevant	Interesting, but not useful	Useful
Matching Knowledge Elements in Concept Maps using a Similarity Flooding Algorithm_MARSHALL.pdf		X	
Afrikaans-English cross-language information retrieval_COSIJN.pdf		X	

Table 64: Evaluation results of documents evaluated by Participant 2 (1 of 3)

Table 65 contains Participant 2's feedback concerning the usefulness of two documents that matched one of his core topics, labelled "***Ontology Applications / Implementation Methodologies****". The source participant's corresponding topic was labelled "***Ontology Management and Ontology Applications***".

Filename	Irrelevant	Interesting, but not useful	Useful
Managing Ontologies A Comparative Study of Ontology Servers_AHMAD.pdf		X	
Ontology Management Semantic Web, Semantic Web Services, and Business Applications_HEPP.pdf			X

Table 65: Evaluation results of documents evaluated by Participant 2 (2 of 3)

Table 66 contains Participant 2's feedback concerning the usefulness of four documents that matched one of his core topics, labelled "***Ontology Applications / Implementation Methodologies****". The source participant's corresponding topic was labelled "***Ontology and Semantic Web Technologies and Applications***".

Filename	Irrelevant	Interesting, but not useful	Useful
Springer-Verlag Ontology Management Semantic Web, Semantic Web Services, and Business Applications.pdf			X
knowledge-mining-proceedings-of-the-nemis-2004-final-conference.9783540250708.27387.pdf		X	
Management of dynamic knowledge.pdf		X	
Wiley Towards The Semantic Web Ontology-Driven Knowledge Management.pdf			X

Table 66: Evaluation results of documents evaluated by Participant 2 (3 of 3)

Table 67 contains Participant 4's feedback concerning the usefulness of four documents that matched one of his core topics, labelled "***Business Strategies/Business Models for Innovation****". The source participant's corresponding topic was labelled "***Product Lifecycle Management | Problem Solving Techniques***".

Filename	Irrelevant	Interesting, but not useful	Useful
The Logistic Innovation Approach and the theory of inventive problem solving.pdf		X	
indeng_v17_n2_a6.pdf			X[67]
IdeationBrainstorming.pdf		X	
Global_Product.pdf		X	

Table 67: Evaluation results of documents evaluated by Participant 4 (1 of 1)

Table 68 contains Participant 5's feedback concerning the usefulness of three documents that matched one of his core topics, labelled "***Knowledge Networks and Knowledge Transfer****". The source participant's corresponding topic was labelled "***Knowledge Networks and Communities of Practice****".

Filename	Irrelevant	Interesting, but not useful	Useful
TeiglandthesisKnowledgeNetworking.pdf		X	
Communities of practice and organizational performance.pdf		X	
Dualities, distributed communities of practice and knowledge management.pdf		X	

Table 68: Evaluation results of documents evaluated by Participant 5 (1 of 3)

[67] Participant 4 remarked that the document "indeng_v17_n2_a6.pdf" was extremely relevant to his field of study.

Table 69 contains Participant 5's feedback concerning the usefulness of three documents that matched one of his core topics, labelled "***Role of Intangible Networks in Innovation****". The source participant's corresponding topic was labelled "***Innovation Networks****".

Filename	Irrelevant	Interesting, but not useful	Useful
Never_Ending_Friending_April_2007.pdf		X	
GIO_2005_for-printing.pdf		X	
von Hippel Con Krogh - Free revealing and the private collective model.pdf		X	
The MIT Press Democratizing Innovation.pdf		X	
Horizontal innovation networks.pdf		X	

Table 69: Evaluation results of documents evaluated by Participant 5 (2 of 3)

Lastly, Table 70 presents Participant 5's feedback concerning the usefulness of a document that matched one of his core topics, labelled "***Knowledge Networks and Knowledge Transfer****". The source participant's corresponding topic was labelled "***Knowledge Management***".

Filename	Irrelevant	Interesting, but not useful	Useful
Schwartz D. Encyclopedia of Knowledge Management.pdf			X

Table 70: Evaluation results of documents evaluated by Participant 5 (3 of 3)

Table 71 shows the breakdown of the usefulness of the documents evaluated by each participant.

Recipient	Irrelevant	Interesting, but not useful	Useful	Total
P1	2 (10%)	11 (55%)	7 (35%)	20
P2	0 (0%)	5 (62.5%)	3 (37.5%)	8
P3	0 (0%)	0 (0%)	0 (0%)	0
P4	0 (0%)	3 (75%)	1 (25%)	4
P5	0 (0%)	8 (88.9%)	1 (11.1%)	9
	2 (4.9%)	**27** (65.9%)	**12** (29.3%)	

Table 71: Breakdown of the usefulness of evaluated documents per recipient

Of the 41 potential useful documents sent to four of the five participants, 29% (12 documents) were deemed to be useful by the recipients, while only about 5% (2 documents) were deemed as being irrelevant to the participants' core research interests. Table 72 summarises whom the respective sources of the documents deemed useful by the respective recipients, were.

	Source					
Recipient	**P1**	**P2**	**P3**	**P4**	**P5**	**Total**
P1	0	0	2	0	5	7
P2	2	0	0	0	1	3
P3	0	0	0	0	0	0
P4	1	0	0	0	0	1
P5	1	0	0	0	0	1
# Useful Docs supplied by	**4**	**0**	**2**	**0**	**6**	**12**

Table 72: Sources and recipients of the useful evaluated documents

Out of the 12 useful documents, 50% (6 documents) originated from Participant 5, 33% (4 documents) from Participant 1 and about 17% (2 documents) from Participant 3. None of the documents of Participant 2 and Participant 4's, which were identified as potentially relevant to other participants, were found to be useful by the recipients.

Table 73 shows, for each participant, which percentage of the total documents identified as potentially relevant to other participants, were actually considered as useful by the recipient participants.

	Source				
	P1	**P2**	**P3**	**P4**	**P5**
# Useful Docs supplied by Participant	4	0	2	0	6
Total # Docs of Participant identified for evaluation	17	0	8	0	16
% Useful Docs of Total Shared Docs	**23.5%**	**0.0%**	**25.0%**	**0.0%**	**37.5%**

Table 73: Percentage useful documents supplied per participant

For this case study, the average ratio of the number of documents per participant identified as potentially useful to the number of documents actually deemed as useful was found to be about 29%.

The following section concludes the discussion of the case study.

10.10 Summary

The issues explored in this case study, conducted with five members of the Enterprise Engineering Research group of Stellenbosch University, are presented below.

This case study explored the following:

- **Whether it is possible to automatically "calculate" a research interest profile for different persons, based on documents associated with the respective persons, using topic models.**
- **Establishing the quality of the resulting research interest profiles.**
- **Finding a mechanism to identify similarities between different research interest areas of the respective persons.**
- **Testing whether the calculated topics can be interpreted by humans.**
- **Establishing whether significantly similar research interest areas of different persons indeed indicate similarity in their research and possibly their tacit knowledge.**
- **Finding a mechanism to identify documents for exchange between persons having similar research interests.**
- **Evaluating the usefulness of the exchanged documents with the receiving parties.**
- **Establishing whether it is possible to calculate a participant keyword profile for individual persons using a topic model.**

Table 74 shows some characteristics about the quality of the topics calculated as part of this case study and address the first two objectives of this case study.

	Topic Quality Characteristics					
Participant	**'Good' Topics**	**Topics with Composite Labels**	**Topics Difficult to Label**	**Noise Topics**	**Core Research Topics**	**Total**
P1	18	2	1	0	4	20
P2	16	4	0	0	15	20
P3	17	3	0	0	10	20
P4	17	0	2	2	10	20
P5	16	3	1	1	9	20
Total	**84** (84%)	**12** (12%)	**4** (4%)	**3** (3%)	**48** (48%)	**100** (100%)

Table 74: Overview of topic evaluation

A large number of topics (84 of 100 topics) were considered to be sensible, semantically consistent descriptions of research interest areas. A small number (12 of the 100 topics) of the calculated topics had dual labels indicating that these actually each addressed two research areas. It is likely that this is an indication that too little topics were used for participants having topics with dual labels (4 of the 5 participants). An even smaller number of the topics proved very difficult to label (4 of the 100 topics), while yet a smaller number (3 of the 100 topics) were judged as being totally meaningless (i.e. noise topics). In total, the amount of topics representing core research areas of participants amounted to 48 of the total of 100 topics identified.

Overall, these results are quite satisfactory as it indicates that the topics generated by statistical topic modelling techniques can be good estimations of research interest areas. Some further improvements may be made to calculate the optimal number of topics to reduce the number of topics with composite labels.

Correlation was used to calculate the degree of similarity between research interest areas (i.e. topics) and fair results were obtained judged by the percentage of useful documents of the total number of documents shared between participants that were identified using topic similarities. Correlation is further not necessarily the best method for determining topic similarities since a correlation score may vary between -1 and 1; negative correlation scores do not have any intuitive meaning in terms of similarities between topics. Other techniques (e.g. the information radius technique) may be better to measure the similarity between topics. The threshold value, which is used to determine what correlation values are considered to indicate significantly similar topics, was determined empirically in this study. Statistical methods may be used to determine this value with larger confidence and transparency.

The research area affinity among participants, calculated based on the number of research interest areas significantly similar among the different participants, was tested with each participant. They unanimously agreed that this is a good representation of the real world situation.

The degree of usefulness of the “unseen” documents identified and exchanged between participants was also to satisfaction. Of the 41 potential useful documents, 29% were deemed to be useful by the recipients in terms of their research, about 66% were considered to be interesting, but not useful, while only about 5% were deemed as being irrelevant to the participants’ core research interests. These results may be improved further by increasing the quality of the topics and finding more accurate ways to calculate similarities between topics.

On overall, 48% of the words calculated for the participants keyword profiles were deemed as describing the core of the individual participants’ research. This value may be improved considerably by only evaluating documents identified by the participants as core to their respective research or again by improving the quality of the topics calculated.

This case study firstly illustrated that topics, which are derived from the textual content of documents different persons have interacted with, may be used as semantic estimations of the interest areas of such persons. Secondly, it was shown that areas for the exchange of tacit knowledge (e.g. by engaging in conversation) or explicit knowledge (e.g. exchanging research articles) may be identified by finding significant similarities between the topics of different persons. The participants involved in this case study are well acquainted and come in contact with one another at least on a weekly basis. Despite of this fact, a number of documents significantly pertaining to the research areas of other members of the research group would most likely not have been shared with the respective participants if it was not for this case study. This indicates that such techniques have value, even in small, non-dispersed groups of people. One may argue that such techniques would even be of more value in larger and more dispersed groups or organisations.

The following are some positive findings in terms of the topic modelling technique (LDA), and the prototype software employed in this case study:

- Files that are actually identical except for having different filenames were grouped in the same topic with the same document-topic weights.
- Files that are very close to identical in terms of content (e.g. a file containing the contents of a web page and another file containing the printer friendly version of the web page content), and also having different filenames, were grouped in the same topic with the nearly the same document-topic weights. This finding together with the abovementioned finding confirms the validity of the calculated topics to some extent.
- Topics having the smallest topic correlation sums proved to be either meaningless topics or very specific topics. Intuitively, specific topics are expected to have little overlap with other topics and therefore this finding supports this speculation.
- It was found that documents of the same author are often grouped in the same topic, as expected, since a given author frequently publishes about a specific set of topics.

The following is a list of some limitations of the approach, techniques and software used in this case study:

- It is not trivial to decide on how many topics to use for a given analysis. Some mechanism for measuring the overall quality, in terms of interpretability, is required to help the analyst find the desired resolution in terms of number of topics.
- The quality of topics seems to be influenced negatively by lengthy documents (e.g. books dealing with a diversity of concepts) allocated to the specific topic. Better results may be achieved by dividing such documents into chapters.
- The fact that topics with composite labels were identified possibly indicates that the number of topics used were too few in some cases.
- For ten out of the 100 topics, the presence of certain documents were hard to explain in context of the other documents significantly associated with these topics. One can speculate that such documents do not have a corresponding topic due to the limited number of topics the model was instructed to formulate (i.e. 20 topics). By increasing the number of topics such documents may be associated to more suiting topics or to topics of their own.
- The fact that three levels of usefulness were used to gather feedback from participants in terms of the usefulness of the “unseen” documents sent to participants, may have biased the use of the middle level (i.e. “Interesting but not useful”) since it may be considered a neutral response.

The following is a list of possible improvements to the overall approach and the techniques used for the application in future case studies:

- In this study, a collection of all research documentation of a given participant was gathered and analysed. A list of core research documents should rather be obtained from each participant to be able to calculate better research interest profiles.
- An objective mechanism to measure the level of interpretability of the topics is required to help the analyst to decide how many topics to use for an analysis of a specific corpus. It is important to point out that the number of topics used need not be the same for all participants. A possible solution may be to do several analyses per participant using a different number of topics for each such analysis, and then to combine the results of the

different analyses per participant before matching participants, to cater for possible differences in the required levels of resolution.

- A quantitative measure is required indicating how significant a given word is to a given topic. This would facilitate the identification of the core words per topic as well as to estimate the quality of individual topics.
- Correlation is not necessarily the best way to calculate similarities between topics. Other techniques should be researched and tested.
- One case was found were a characterising word was missing from a given topic (the word "open" was not part of the characterising words of a topic dealing with "open innovation"). On closer inspection this word was found to be part of a stoplist and therefore not considered for the analysis. Care should be taken to ensure that potentially useful words are not part of a stoplist before any analyses are begun.
- Better topics and document similarities may be obtained by dividing electronic books into their respective chapters. Then individual chapters may be allocated to topics and similarities between individual chapters can be found in the topic model output.
- A mechanism is required to identify duplicate and near-duplicate documents before the topic model calculation process commences. This would make it possible to only include one copy of each document in order not to skew the results by including multiple copies of a given document.
- Rather use four levels to judge the usefulness of shared documents. This will eliminate the presence of a perceived neutral level.

A note on the implication of the case study in terms of the proposed framework:

Even though this case study concentrated more on ways to identify possible relationships between different persons in an organisation, based on the documents they interact with, it has important implications in terms of the framework formulated in Chapter 8. It illustrated the potential value of knowing the relationships between individual persons and individual documents, the relationships between individual documents as well as the relationships between individual persons, which would be embodied in the envisaged framework once implemented.

In terms of the framework entity types described in Chapter 8, the following entity types explicitly exist in the Enterprise Engineering Group: Concept, Collaborator/Distributor, Document, Idea, Knowledge Area, Need, Objective, Opportunity, Person, Process/Activity, Product/Service, Project/Programme, Skill/Competency, Technology/Method/Tool/Equipment. The envisaged framework, once implemented, would definitely provide a useful knowledge base for the Enterprise Engineering Group and would play a key role in capturing and disseminating the knowledge of this group.

This concludes the discussion around the Enterprise Engineering Group case study. The next case study presents a detailed analysis of a collection of documents of a European production engineering research academy named CIRP.

11. Case Study 2: CIRP

This case study explores what value can be obtained by analysing a collection of electronic documents associated with a specific organisation using a statistical topic modelling technique.

This case study explores the following:

- **Automatically revealing which subjects were addressed in CIRP technical papers for the period 2001 to 2006 using a topic model**
- **Investigate the interpretability of the calculated topics and how to find an appropriate resolution in terms of number of topics**
- **Determine which documents are associated with which topics and vice versa**
- **Identify the documents that are significantly associated with more than one topic**
- **Establish the degree of similarity between topics to uncover topic affinities**
- **Establish how well each topic is represented in the input document set**
- **For each document, identify which other documents are relevant to the document in question**
- **Identify which words are strongly associated with a given word**
- **Identify which topics are associated with a given word**
- **Investigate ways of identifying the general and specific terms for the given field**
- **Investigate whether the topic model may be used as a starting point to construct a vocabulary for the CIRP community**
- **Identify which authors are associated with each topic and vice versa**
- **Identify which STCs are associated with each topic and vice versa**
- **Investigate ways to use the topic model results to estimate topic-time associations**
- **Establish if it is possible to construct a network or hierarchy of topics from the topic model resulting from different runs with varying levels of resolution (in terms of number of topics)**
- **Investigate ways of validating the topic model results**

11.1 Background to CIRP

According to their website[68], the International Academy for Production Engineering, better known as CIRP, was founded in 1951 for the purpose of scientifically addressing issues related to modern production science and technology by means of international cooperation. It currently includes more than 500 members from 46 countries. The name “CIRP” originated from the French acronym of “College International pour la Recherche en Productique”. The number of members is intentionally limited in order to better facilitate informal scientific information exchange and personal contact.

[68] www.cirp.net

Although CIRP is an academic organisation, it promotes the participation of industry in its activities. There are approximately 140 Corporate Members who participate in the research work of the academic members of CIRP, and who frequently contribute to the information exchange within CIRP by presenting their views on needs and perspectives of industry. CIRP has the following general objectives:

- Encouraging scientific research related to the following research areas:
    - Manufacturing processes,
    - Production equipment and automation,
    - Manufacturing systems and
    - Product design and manufacturing
- Encouraging cooperative research among its members and creating opportunities for informal contacts among CIRP members
- Encouraging the application of the fundamental research work in industry and at the same time receiving feedback from industry about their needs and their development
- Organising an annual General Assembly with keynote and paper sessions and meetings of the STCs, publishing papers, reports, annals and other technical information, as well as organising and sponsoring international conferences.

CIRP is composed of a number of Scientific Technical Committees (STCs) and Working Groups (WGs), covering numerous research areas. Each year CIRP organises a General Assembly, which lasts for a week, where over 140 technical paper presentations from different areas of manufacturing take place in addition to the presentation of a number of keynote papers at the opening of the conference, and the presentation of technical work within the different STCs.

The author decided to analyse the CIRP documentation for the following reasons:

- The author had access to their formal documentation for a period of six years.
- At the time there was a need to show some of their members what may be achieved by applying text analytical techniques to their unstructured information.
- The author knows five CIRP members who could provide possible inputs to the study.
- CIRP can be viewed as an actual organisation with the STCs being the departments of such an organisation.

The following section will provide more details about CIRP's respective Scientific Technical Committees (STCs).

11.2 CIRP's Research Focus

The Scientific and Technical Committees (STCs) are the groups responsible for managing the collaborative research project run by CIRP. The various STCs are described in the following sections.

Assembly

The Assembly STC (STC-A) deals with the techniques, processes, equipment and design for assembly and disassembly, recycling, handling, micro-assembly, product integration and interconnection, including human operations such as maintenance services, quality testing, and logistics. It further includes monitoring, diagnostics and control, information and communication systems, structure and organisation of assembly processes.

Cutting

The Cutting STC (STC-C) addresses the processes and techniques used to shape components by material removal (turning, milling etc.), including the processes of chip formation, the physical laws governing the wear of cutting tools and the factors influencing surface finish.

Design

The Design STC (STC-Dn) covers the conceptual and innovative processes in engineering design. More specifically, the following areas fall within the domain of this STC: Design for economic manufacture, coordination with manufacturing, computer-automated systems and the integration of technological and economic methods, interfacing of CAD/CAM systems and databases for CAD systems.

Dictionary

This STC has the responsibility for publication of CIRP dictionaries on Advanced Manufacturing Engineering. The dictionaries cover definitions and technology for manufacturing processes, machines, tooling, materials and system formulated by the other STCs.

Electro-physical and Chemical Processes

The Electro-physical and Chemical Processes STC (STC-E) deals with research into material removal processes of a physical, physicochemical or chemical nature, such as electro-discharge machining (EDM), electrochemical machining (ECM) and the use of high energy laser, electron and ion beams.

Forming

The Forming STC (STC-F) addresses processes in which components are shaped by plastic deformation, including pressure joining and separation techniques such as stamping, shearing, etc. It further covers the application of the theory of plasticity to industrial forming processes with reference to tribology and materials engineering aspects.

Abrasive Process

The Abrasive Process STC (STC-G) is involved in research into material removal processes using hard abrasive grains such as grinding and finishing. Attention is largely focus on the mechanics of grinding and the economics of abrasive processes.

Machines

The Machines STC (STC-M) covers the design, manufacture and use of manufacturing equipment, including the study of performance-related factors, such as the static and dynamic behaviour, efficiency and resistance to wear. The control of production processes and the application of new materials as well as the automation, interface and control systems are also covered.

Nanotechnology

The Nanotechnology STC (STC-N) was merged with the Precision Engineering and Metrology STC (STC-P) in 2003.

Optimisation of Manufacturing Systems

The Optimisation of Manufacturing Systems STC (STC-O) deals with the design for production, factory equipment selection and lay-out, numerical and adaptative control, application of computers to manufacturing, information technology and human factors in production engineering. This STC also has the role of advising the other STCs regarding the optimisation of manufacturing systems.

Precision Engineering and Metrology

The Precision Engineering and Metrology STC (STC-P) focuses on the development and application of measuring techniques to be used for quality control procedures, involving the measurement of size, shape and positional relationships in manufactured components and assemblies. Nanotechnology processes and equipment also falls within the domain of this STC.

Surfaces

Lastly, the Surfaces STC (STC-S) performs research into the geometrical, physical and chemical properties of the workpiece surface in relation to the production process concerned. This has involved the preparation of a CIRP standard for measuring roughness parameters and collaborative projects on measuring surface hardness, residual stresses and crack detection on workpiece surfaces.

11.3 Information about a Previous Investigation into CIRP Documentation

The author and some Indutech colleges already started with ad hoc investigations, about how text analytical techniques may be used to aid knowledge sharing within CIRP, in 2005. Initially, prototype software implementing the LSI[69] technique to cluster documents, extracting text patterns using regular expressions and collocations, among others, was developed by Indutech

[69] See section 15.14 in Appendix A for more information about LSI.

and was used for such investigations. LSI proved to have many shortcomings, but its large requirements on computer resources and poor scalability led us to investigate other techniques related to soft clustering. In 2006, the first implementation of LDA[70] in another software prototype was completed. This prototype was used as part of a formal investigation into what can be achieved in terms of knowledge sharing in CIRP by analysing the content of CIRP technical papers. A report summarising the investigation and its findings was compiled and submitted to one of their longstanding members. The investigation was based on 641 technical papers presented at the CIRP General Assemblies of 2002 to 2006. This investigation had the following objectives:

1. Automatically group papers into meaningful categories based on their content.
2. Determine descriptive terms for each category.
3. Determine the overlaps between various categories in terms of descriptive terms.
4. For each paper, determine which other papers are conceptually similar to the relevant paper with a certain level of significance.
5. Determine descriptive terms for each of the technical papers.
6. Determine descriptive terms for each of the STCs.
7. Extract metadata from technical papers.
8. Determine descriptive terms for each author.
9. Determine in which STC a given paper would fit best.
10. Determine outlier papers for a given STC.
11. Determine the most conforming papers for a given STC.

A very basic implementation of the LDA technique along with custom developed text extraction algorithms, all developed by Indutech, were used to perform the relevant analyses[71]. Although only the absolute essential details of this investigation will be presented in this section, it provides some background to the subsequent analysis of CIRP documents, which is the actual focus of this chapter.

The following is a list of the key conclusions made after the completion of the first investigation:

- **Automatically group papers into meaningful categories based on their content**: The classification of papers obtained through the dynamic classification process does not match the CIRP classification. Dynamic classification, as per LDA, detects the underlying topics of the given document collection. While a paper may superficially be designated as belonging to the 'Design' STC, or to the 'Optimisation' STC, the object of the paper might be to design a machine or to optimise one; thus the majority of the topic weight of the paper falls to the topic of 'Machining'. It is also important to note that LDA does not only assign one topic per document: every document has associated with it a 'topic vector', which is a vector with dimension equal to the number of topics, representing how strongly each topic is represented in the document. The method used to cluster the documents has simply been to take the topic with the largest weight and assign the document to that topic. A more sophisticated method would be to use some other form of clustering vectors (e.g. K-means clustering) to obtain more topically coherent clusters.

[70] See section 16.4 in Appendix B for more information about LDA.

[71] See Appendix C for the characteristics of the software used for this analysis.

- **Determine descriptive terms for each category:** The topics themselves, on the other hand, are much more like the original CIRP classification: all the STC's are well represented, except the Optimisation STC. This might be because optimisation does not so much use a vocabulary of its own than use primarily the vocabulary of the system it is aiming to optimise. For this reason, the optimisation topic 'dissolves' into the other categories, and optimisation keywords will be found scattered among the other topics. The problem of the LDA algorithm 'degenerating', i.e. not being able to find any topics when too many topics are specified, remains to be solved.

- **Determine the overlaps between various categories in terms of descriptive terms:** The topic overlaps reveal that the shared words between topics are mostly words signifying general concepts, like 'system', 'process', 'cutting', and so forth. This is understandable since the individual topics themselves provide the additional detail. Also, the number of times a word occurs in a topic overlap gives an indication as to its prevalence. For instance, for some years 'cutting' occur the most whereas in other years 'process' or 'system' occurs the most. This indicates the dominant concept for the papers in that year.

- **For each paper, determine which other papers are conceptually similar to the relevant paper with a certain level of significance:** The document similarity results are reasonably accurate. One can note that the documents ranked as most similar to any given document within a year collection usually falls outside the STC of the original document, whereas for the five year collection (all papers from 2002 to 2006) the most similar documents for any given document usually falls within the STC of the original document. This is because for both collections the same number of topics was used – hence for the one year collection dynamic classification is more discriminatory whereas for the five year collection it is less discriminatory, and more generalising.

- **Determine descriptive terms for each technical paper:** The descriptive terms per paper provide a very good characterisation of the content of the paper. Both ends of the spectrum are covered: the single-word key terms provide a more abstract, general view on the paper, whereas the two-word key terms give a more specific, detailed view on the paper, mostly consisting of the jargon of the subject at hand. Multi-word key terms may also be useful and may be included in future versions.

- **Determine descriptive terms for each collection:** The descriptive terms of a collection provide an excellent characterisation of the documents in the collection. The most highly ranked two-word key terms give an idea as to which specific techniques are mentioned most often in a given field of study.

- **Extract metadata from technical papers:** The metadata extraction works well in general – there are some documents, however, where the extraction fails. This can be attributed to the non-standard way in which most CIRP PDF documents are formatted and possible sensitivity of the extraction algorithm. The inconsistent formatting issue may be the result of using different PDF converters to convert the document to the PDF file format. This is an area where there is room for improvement – both in the extraction process itself but more importantly, in the amount of other metadata that can be extracted.

- **Determine descriptive terms for each CIRP member:** The results of this experiment are best verified by the authors themselves, or at least experts in the field, but a comparison of the content of the papers of an author and the author's assigned keywords agree fairly well.

- **Determine in which collection a given paper would fit best:** For many papers the calculated STC does not equal the assigned STC simply because the vocabulary of the calculated STC was used more often in the paper than the vocabulary of the assigned STC. The assigned STC does turn up in the top three most fitting STC's in the majority of cases. This demonstrates the power of dynamic classification to find the underlying topics of a collection of documents. Another possible explanation may be that the classification of technical papers is currently done manually by CIRP and since most papers pertains to more than one STC, it is difficult to determine the most fitting STC by inspection.

- **Determine outlier papers for a given collection:** The outlier papers in a given collection are determined by finding the paper that differs the most from other papers in the collection in terms of concepts addressed. Outlier papers may indicate new research in a given collection (when compared to the other papers in the collection). The outlier papers identified are best verified by an expert due to the specialised nature of the field.

- **Determine most conforming papers for a given collection:** On the other hand, the most conforming papers in a given collection are determined by finding the paper that is the most similar to other papers in the collection in terms of concepts addressed. The most conforming papers may indicate the most representative papers of a collection (when compared to other papers in the collection) and may further indicate a good starting point when reading about an unknown field. Once more, the most conforming documents identified are best verified by an expert due to the specialised nature of the field.

The results of this initial CIRP study have shown that useful information can be obtained by analysing an electronic document collection. Automation can be done to a large extend and may further be refined to obtain even better results. As part of the finding of this study, it was mentioned that in order to be of maximum use, the entities (i.e. papers, authors, key terms, abstracts, titles, STCs, etc.) and relations (most conceptual similar documents, most suitable STC per paper, etc.) should be stored in a flexible framework that can be used to readily explore the information at hand. Using such a framework, one can start at familiar entity and discover associations with other entities in this network. Such a framework would be of tremendous value for any virtual research organisation to further interactions between its members as well as to aid the retrieval of information.

In terms of future refinements, the greatest challenge is to find a way to estimate the maximum number of topics that the dynamic classification algorithm can determine in a collection without 'degenerating'. Topic degeneration is in reality a failure of the underlying inference model to converge. Using different techniques to perform this inference, or adjusting the parameters governing convergence is an obvious way to improve the performance.

The dynamic classification algorithm used can also be extended to work with n-grams – that is, groups of words – as it currently only caters for unigrams (i.e. one-word terms). This would be possible as the underlying model will stay the same. This would mean that topic qualifiers will no longer be restricted to one-word terms only.

Another area for future research is regular expressions: finding more effective and accurate ways of extracting word patterns, and finding novel ways using regular expressions to extract information like person names, place names, dates, monetary values, etc.

In conclusion, at the time of completing this specific investigation it was the opinion of author that satisfactory results were obtained for all experiments although not all results have been validated by experts.

The remainder of this chapter will discuss the second analysis of CIRP documentation, the focus of this case study.

11.4 Electronic Documents Analysed

All technical papers and keynote papers presented at the CIRP General Assemblies of 2001 to 2006[72] were processed as well as supporting files of the given years (e.g. lists of authors, tables of contents, cover page of proceedings, etc.) with the goal to investigate what can be achieved by analysing the documentation of a research driven organisation with special focus on the objectives listed at the beginning of this chapter. Table 75 shows some statistics about the documents analysed in this case study.

Publication Year	Number of Files Processed	Number of Files having Extraction Problems	Number of Duplicate Files	Size of Document Corpus
2001	95	3	2	60.3 MB
2002	136	1	0	121.0 MB
2003	128	0	2	89.2 MB
2004	117	1	2	64.8 MB
2005	133	1	1	85.7 MB
2006	143	0	0	129.0 MB
Total	**752**	**6**	**7**	**550.0 MB**

Table 75: Characteristics of the documents analysed for the CIRP investigation

The reason for including files representing lists of authors, cover pages and tables of contents of proceedings, in spite of the fact that such files do not carry much semantic value, was to decrease the need for human intervention in the analysis process and to test whether such files will negatively impact the overall quality of the resulting topic model.

11.5 Analysis Process

Papers were already grouped into folders, one for each year of publication, and each such folder containing sub-folders representing the various Scientific Technical Committees (STCs). A

[72] The author did not have access to the CIRP technical papers for the years 2007 and 2008 when this case study was in progress.

custom-developed English stoplist with 821 entries was used to eliminate the words with little semantic value. Analyses were performed on the same document set using 10, 30 and 100 topics respectively. A minimum frequency value of 5 was used for all three analyses. The results of the 100 topic analysis proved to be the easiest to interpret and were therefore used for further manual processing. Table 76 shows the run statistics for each of the three analyses performed. All three analyses were performed on a 2GHz, 8-core CPU computer, using an improved version[73] of Indutech's Content Analysis Toolkit (CAT)[74].

Topics	Files	Minimum Frequency	n-gram Level	Total Words	Unique Words	Average Words per File	Run Time (hh:mm:ss)
10	757	5	unigrams + bigrams	1644406	47738	2172	01:02:40
30	757	5	unigrams + bigrams	1644406	47738	2172	02:49:17
100	757	5	unigrams + bigrams	1644406	47738	2172	09:49:31

Table 76: Characteristics of the 10-, 30- and 100-topic analyses

The software exports the results to an Excel file that contains the following information:

- A spreadsheet containing the calculated topics. For each topic, the characterising words along with the associated probabilities that the individual words describe the individual topics are given. This corresponds to the topic-word matrix discussed earlier. For each topic it is possible to add a descriptive label that is automatically populated to other spreadsheets.
- A spreadsheet containing the document-topic matrix. This matrix has the individual documents as rows, the calculated topics as columns with the individual cells containing the applicable document-topic mixture ratios, indicating how strongly a given document is associated with a given topic.
- A spreadsheet indicating how well a given topic is represented in the input documents relative to other topics. The topic 'coverage' results are shown in a pie chart.
- A spreadsheet containing the degree of similarity for all topic pairs.
- A spreadsheet containing the two closest topics to any given topic.
- A spreadsheet containing the degree of similarity for all document pairs.
- A spreadsheet containing the two closest documents to any given document.
- A spreadsheet containing a list of all unique words characterising the respective topics along with the number of topics each word characterises and an indication of the specific topics it characterises. In addition, for each word in the list the associated probability is given for the respective topics it characterises. For each word in the list a score is given indicating the level of generality of the given word for the corpus analysed. General words (e.g. "cutting") have larger scores than specific words (e.g. "asynchronous").

In addition to the results exported in an Excel spreadsheet, CAT also saves the results in a proprietary database. Using this database, the software also has the ability to show such results in a visually interactive way by means of the *CAT Visualisation Centre*.

[73] See Appendix C for the characteristics of the software used for these analyses.
[74] For more information about CAT, visit www.analyzecontent.com.

11.6 Interpreting the Topic Model

The interpretation of a topic model was discussed in section 6.5.6. The first step in evaluating the results of the three analyses forming part of this case study is labelling the individual topics. The following tables present the labels, as identified by the author, for each of the three analyses.

Topic Number	Topic Label
Topic 1	Tools, Models & Systems
Topic 2	Forming
Topic 3	Surfaces & Abrasive Processes
Topic 4	Design & Optimisation of Manufacturing Systems
Topic 5	Abrasive Processes | Electro-physical & Chemical Processes
Topic 6	Machines
Topic 7	Precision Engineering & Metrology
Topic 8	Assembly
Topic 9	Optimisation of Manufacturing Systems
Topic 10	Cutting

Table 77: Topic labels assigned to the topics generated by the 10-topic analysis

The fact that the calculated ten topics were very broad in nature indicated that more topics are required.

Topic Number	Topic Label
Topic 1	Coated Cutting Tools & Wear
Topic 2	Electric Discharge Machining & Electrodes
Topic 3	Assembly & Product Lifecycle
Topic 4	Milling
Topic 5	Production Systems & Production Planning
Topic 6	Micro-products & Associated Processes
Topic 7	Manufacturing System Configuration & Control
Topic 8	Machine Tool Control
Topic 9	Chip Formation & Associated Factors
Topic 10	Surface Roughness & Associated Factors
Topic 11	Vibration Cutting
Topic 12	Applications of Modelling & Simulation
Topic 13	Grinding & Grinding Wheels
Topic 14	Virtual Models of Tools & Workpieces
Topic 15	Parallel Kinematic Machining
Topic 16	Design Complexity & Axiomatic Design
Topic 17	Monitoring, Measurement & Sensors
Topic 18	Polishing
Topic 19	Deep Drawing & Biological Simulation
Topic 20	Optical Measurement
Topic 21	Forming Processes, Stress & Strain
Topic 22	Spindle Bearings & Spindle Systems
Topic 23	Supply Chain Control, Flexible Automation & Production Planning
Topic 24	Product Disassembly & Recycling
Topic 25	Laser Sintering
Topic 26	Punching, Drawing & Extrusion & Associated Factors
Topic 27	Temperature & Manufacturing Processes
Topic 28	Measurement & Metrology
Topic 29	Gears & Gear Metrology
Topic 30	Knowledge as Support to Product Development & Design Processes

Table 78: Topic labels assigned to the topics generated by the 30-topic analysis

No topics with composite labels were found in the 30-topic analysis. In spite of this fact a 100 topic analysis was performed to investigate whether more specific topics would result.

Topic Number	Topic Label
Topic 1	Quality Controls & Process Chains
Topic 2	Forming Processes
Topic 3	Gears & Gear Metrology
Topic 4	Polishing & Spray Painting
Topic 5	Milling
Topic 6	Electrochemical Processes
Topic 7	Grippers & Part Handling
Topic 8	Dies & Ironing
Topic 9	Stepping, Piezoelectric Motors & Actuators
Topic 10	Production System Control
Topic 11	Grinding Processes
Topic 12	Simulating Molecular Dynamics
Topic 13	Robot Vision
Topic 14	Cutting & Chip Formation
Topic 15	Tool Path Generation
Topic 16	Wrinkling
Topic 17	Silicon Wafers & Substrates
Topic 18	Forming Processes
Topic 19	New Methodologies (Chatter Detection | Enterprise Knowledge Model| Components Reuse)
Topic 20	Vibration Cutting
Topic 21	Air Flow & Friction
Topic 22	Tool Wear & Tool Life
Topic 23	Micro Products & Micromachining
Topic 24	Virtual Reality & Haptic Systems
Topic 25	Injection Moulding & Temperature
Topic 26	Workpiece Surfaces & Tool Life
Topic 27	Dressing & Trueing
Topic 28	Glass Drilling & Cracks
Topic 29	Production & Supply Chain Management
Topic 30	Temperature & Heat Transfer

Table 79: Topic labels assigned to Topics 1 to 30 of the 100-topic analysis

Topic Number	Topic Label
Topic 31	Roll Forming
Topic 32	Authors List | Table of Contents | Specification of Direct Drive Motors
Topic 33	Optical Measurement Systems
Topic 34	Measurement Sensors
Topic 35	Laser Forming & Bending | Lifecycle Management & Recycling
Topic 36	Photochemical Machining | Ceramic Sol-Gel Coatings
Topic 37	Manufacturing System Configuration
Topic 38	Damping Ratios
Topic 39	Machine Tool Structure Models & Optimisation
Topic 40	Parallel Kinematic Machines
Topic 41	Lapping & Current Control
Topic 42	Product Lifecycles
Topic 43	Rapid Prototyping & Laser Sintering
Topic 44	Drilling & Turning Cooling Lubricants, Tool Coating Design | Surface Roughness in Hard Turning
Topic 45	Friction
Topic 46	Fe & Melting
Topic 47	Monitoring Grinding Processes & Acoustic Emission
Topic 48	Atomic Force Microscopy & Surface Measurement
Topic 49	Reconfigurable Enterprises & -Production Systems & Distributed Enterprises | Collaborative Design
Topic 50	Bone Cutting
Topic 51	Punching
Topic 52	Costing & Cost Models
Topic 53	Product Families, Product Design & Configuration
Topic 54	Product Development, Design Process & Knowledge
Topic 55	Bearings & Rotation
Topic 56	Cooling & Heat Transfer
Topic 57	Ultrasonic Cutting
Topic 58	Finite Element Analysis & FEM Simulation
Topic 59	Micro Objects & Micro Handling
Topic 60	Drilling & Bore Quality

Table 80: Topic labels assigned to Topics 31 to 60 of the 100-topic analysis

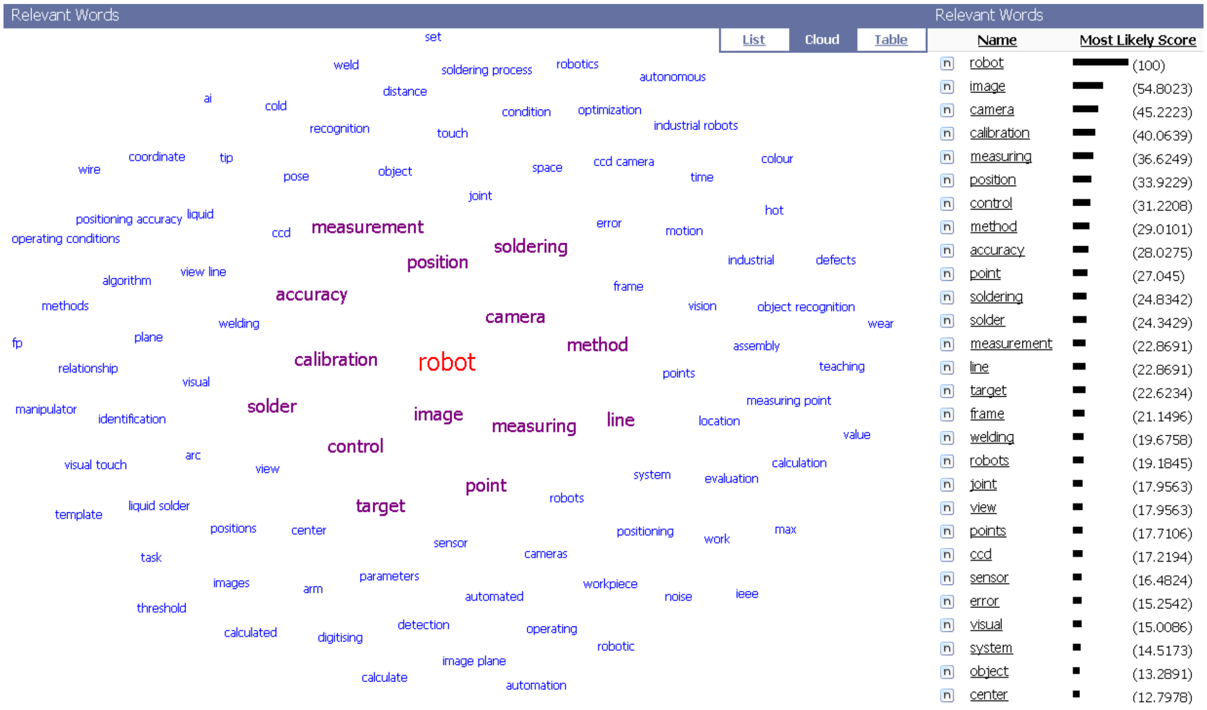
Topic Number	Topic Label
Topic 61	Grinding Wheels
Topic 62	Surface Topography & Processing
Topic 63	Axiomatic Design, Functional Requirements and Design Parameters
Topic 64	Burr Formation
Topic 65	Minimal Quantity Lubrication Cutting
Topic 66	Reverse Engineering & Surface Reconstruction
Topic 67	Product Disassembly, Remanufacturing & Recycling
Topic 68	Automatic Revolving Door Safety
Topic 69	Assembly Faults
Topic 70	Tools, Models & Systems for Optimising Design- & Production Processes
Topic 71	Feature Recognition
Topic 72	Titanium & Deep Drawing
Topic 73	Coating & Wear
Topic 74	Sheet Metal Forming Processes
Topic 75	Measurement Accuracy & Calibration
Topic 76	Fuzzy Decision-making Methods
Topic 77	Strength & Residual Stress
Topic 78	Extrusion
Topic 79	Product-, Process & Operational Complexity
Topic 80	Abrasive Material Removal
Topic 81	Reuse & Reusability
Topic 82	Semi-solid Metal & Alloy Processing
Topic 83	Manufacturing Paradigms
Topic 84	Production Planning & Resource Scheduling
Topic 85	Surface Roughness & Droplet Bouncing
Topic 86	Tungsten Clad Rods & Residual Stress
Topic 87	Drive System Control
Topic 88	Self-Organising Biological Manufacturing Systems
Topic 89	Machine Tool Control
Topic 90	Thread Tapping Tolerance
Topic 91	Noise Topic (only symbols)
Topic 92	Steel Sheet Strength
Topic 93	Polishing & Material Removal
Topic 94	Diamond Machining Processes & Surface Roughness
Topic 95	Centerless Grinding & Stability
Topic 96	Electric Discharge Machining & Electrodes
Topic 97	Optical Measurement & Metrology
Topic 98	Laser Machining
Topic 99	Multiresolution CSG Models | Carbon Nanotube Assembly
Topic 100	Assembly Processes & Systems

Table 81: Topic labels assigned to Topics 61 to 100 of the 100-topic analysis

By comparing the topics of the 100-topic analysis to that of the 10-topic and 30-topic analyses, it can be seen that topic specificity increases with the increase in number of topics as expected. The topics of the 100-topic analysis were generally easy to interpret and provided a fine-grained segmentation of the field of study.

11.6.1 Finding the Words that Characterises a Topic

Each of the topics listed in the preceding tables is characterised by 100 words each having an associated weight (a probability to be more specific) for a specific topic. A given word can be associated with more than one topic when the word can be used in more than one context. Figure 36 shows an example of the words characterising the topic labelled *Robot Vision* (Topic 13 of the 100-topic analysis) as presented in the software. In the table representation, at the right hand side of Figure 36, the words associated with the topic are shown in order of decreasing certainty for the specific topic (note that not all of the 100 words associated with this topic are displayed in this figure due to size restrictions). The tag cloud representation, at the left hand side of Figure 36, gives a spatial representation of the words and terms associated with the given topic. Here, the words in the centre correspond to the words that are highly likely to be associated with the given the topic.



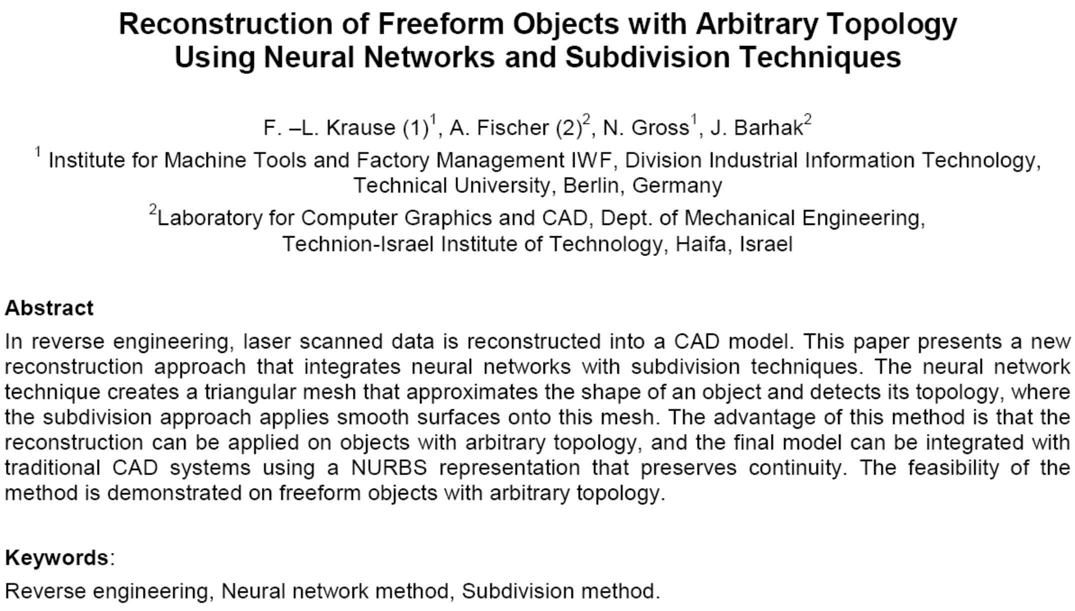
	Name	Most Likely Score
n	robot	(100)
n	image	(54.8023)
n	camera	(45.2223)
n	calibration	(40.0639)
n	measuring	(36.6249)
n	position	(33.9229)
n	control	(31.2208)
n	method	(29.0101)
n	accuracy	(28.0275)
n	point	(27.045)
n	soldering	(24.8342)
n	solder	(24.3429)
n	measurement	(22.8691)
n	line	(22.8691)
n	target	(22.6234)
n	frame	(21.1496)
n	welding	(19.6758)
n	robots	(19.1845)
n	joint	(17.9563)
n	view	(17.9563)
n	points	(17.7106)
n	ccd	(17.2194)
n	sensor	(16.4824)
n	error	(15.2542)
n	visual	(15.0086)
n	system	(14.5173)
n	object	(13.2891)
n	center	(12.7978)

Figure 36: Words characterising the topic "Robot Vision"

Overall, the quality of the topics seemed very good since no noise topics were identified for the 10- and 30-topic analysis and only one noise topic was identified for the 100-topic analysis. The results of the 100-topic analysis were used for further analysis due to its easily interpretable, very specific topics.

11.6.2 Finding Topics Relevant to a given Document

Using the 100-topic CIRP topic model it is possible to find the topic(s) pertaining to any given paper. The following paper (52-1-2003-125.pdf), published in the in 2003 under Design STC (STC-Dn), addressed a single topic that was labelled *Reverse Engineering & Surface Reconstruction* (Topic 66).



Reconstruction of Freeform Objects with Arbitrary Topology Using Neural Networks and Subdivision Techniques

F. –L. Krause (1)[1], A. Fischer (2)[2], N. Gross[1], J. Barhak[2]

[1] Institute for Machine Tools and Factory Management IWF, Division Industrial Information Technology, Technical University, Berlin, Germany

[2]Laboratory for Computer Graphics and CAD, Dept. of Mechanical Engineering, Technion-Israel Institute of Technology, Haifa, Israel

Abstract

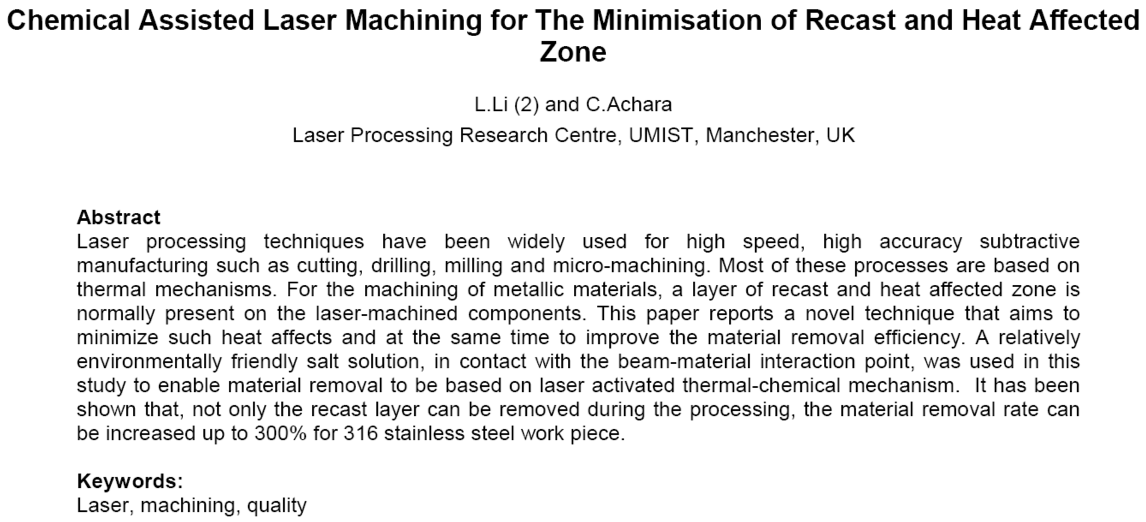
In reverse engineering, laser scanned data is reconstructed into a CAD model. This paper presents a new reconstruction approach that integrates neural networks with subdivision techniques. The neural network technique creates a triangular mesh that approximates the shape of an object and detects its topology, where the subdivision approach applies smooth surfaces onto this mesh. The advantage of this method is that the reconstruction can be applied on objects with arbitrary topology, and the final model can be integrated with traditional CAD systems using a NURBS representation that preserves continuity. The feasibility of the method is demonstrated on freeform objects with arbitrary topology.

Keywords:

Reverse engineering, Neural network method, Subdivision method.

Figure 37: Title and abstract of document linked to the topic "Reverse Engineering & Surface Reconstruction"

It is however possible for a paper to pertain to more than one topic. For example, the following paper (E02_Li.pdf) published in the 2004 under the Electro-physical and Chemical Processes STC (STC-E), addressed two topics according to the topic model, namely *Laser Machining* (Topic 98) and *Surface Topography & Processing* (Topic 62).



Chemical Assisted Laser Machining for The Minimisation of Recast and Heat Affected Zone

L.Li (2) and C.Achara
Laser Processing Research Centre, UMIST, Manchester, UK

Abstract
Laser processing techniques have been widely used for high speed, high accuracy subtractive manufacturing such as cutting, drilling, milling and micro-machining. Most of these processes are based on thermal mechanisms. For the machining of metallic materials, a layer of recast and heat affected zone is normally present on the laser-machined components. This paper reports a novel technique that aims to minimize such heat affects and at the same time to improve the material removal efficiency. A relatively environmentally friendly salt solution, in contact with the beam-material interaction point, was used in this study to enable material removal to be based on laser activated thermal-chemical mechanism. It has been shown that, not only the recast layer can be removed during the processing, the material removal rate can be increased up to 300% for 316 stainless steel work piece.

Keywords:
Laser, machining, quality

Figure 38: Title and abstract of a document linked to topics "Laser Machining" and "Surface Topography & Processing"

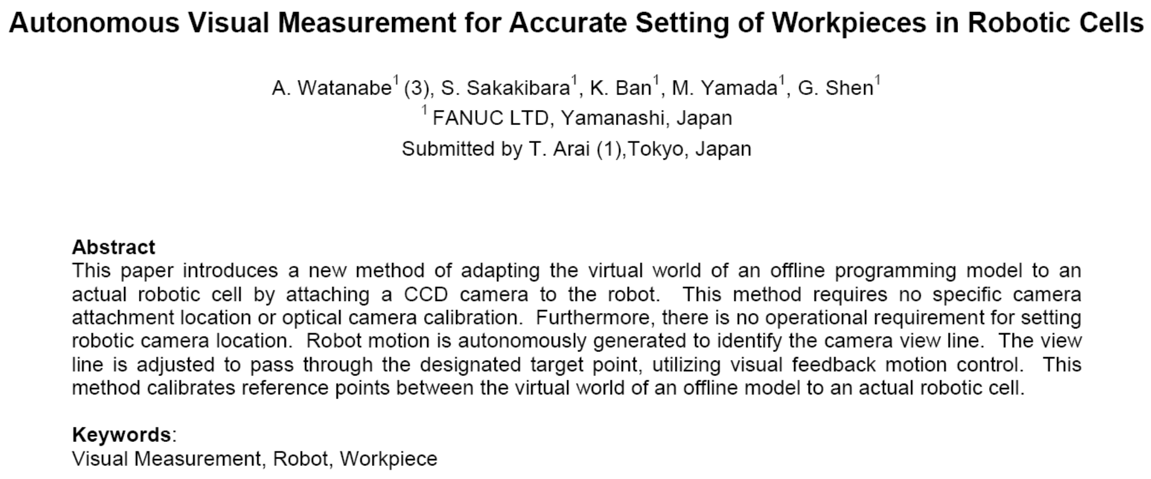
It can be seen from the titles and abstracts of these papers that they indeed correspond to the identified topics.

A mechanism was constructed in the relevant spreadsheet to identify topics that significantly address more than one topic. Such documents can be called 'bridging documents' since they essentially form a connection between two or more topics. This was done by finding all documents that have two or more document-topic weights between 0.30 and 0.70[75] (a document has a number of document-topic weights equal to the number of topics which is 100 in this case). For example, a document having two document-topic weights between 0.30 and 0.70 essentially "bridges" two topics, while a document having three document-topic weights between these values "bridges" three topics. A total of 105 documents of the just more than 750 documents analysed were identified as bridging documents.

11.6.3 Finding Documents Relevant to a given Topic

Conversely, one can use the topic model to find papers pertaining to a specific topic or combination of topics. For example, when interested in finding the articles most relevant to the topic *Robot Vision* (Topic 13) presented earlier, the following three articles (published in Assembly STC, STC-A, in 2005, 2006 and 2002 respectively) have the strongest association of the 10 papers associated with the given topic according to the topic model.

[75] The values of 0.30 and 0.70 were once more determined empirically.



Autonomous Visual Measurement for Accurate Setting of Workpieces in Robotic Cells

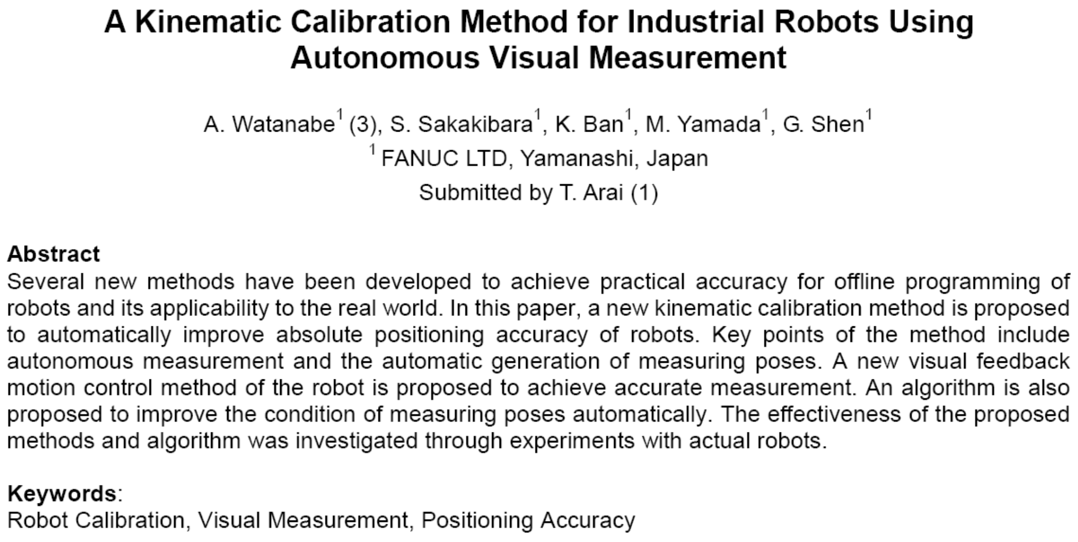
A. Watanabe[1] (3), S. Sakakibara[1], K. Ban[1], M. Yamada[1], G. Shen[1]
[1] FANUC LTD, Yamanashi, Japan
Submitted by T. Arai (1),Tokyo, Japan

Abstract

This paper introduces a new method of adapting the virtual world of an offline programming model to an actual robotic cell by attaching a CCD camera to the robot. This method requires no specific camera attachment location or optical camera calibration. Furthermore, there is no operational requirement for setting robotic camera location. Robot motion is autonomously generated to identify the camera view line. The view line is adjusted to pass through the designated target point, utilizing visual feedback motion control. This method calibrates reference points between the virtual world of an offline model to an actual robotic cell.

Keywords:
Visual Measurement, Robot, Workpiece

Figure 39: Title and abstract of a document with highest affinity with topic "Robot Vision"



A Kinematic Calibration Method for Industrial Robots Using Autonomous Visual Measurement

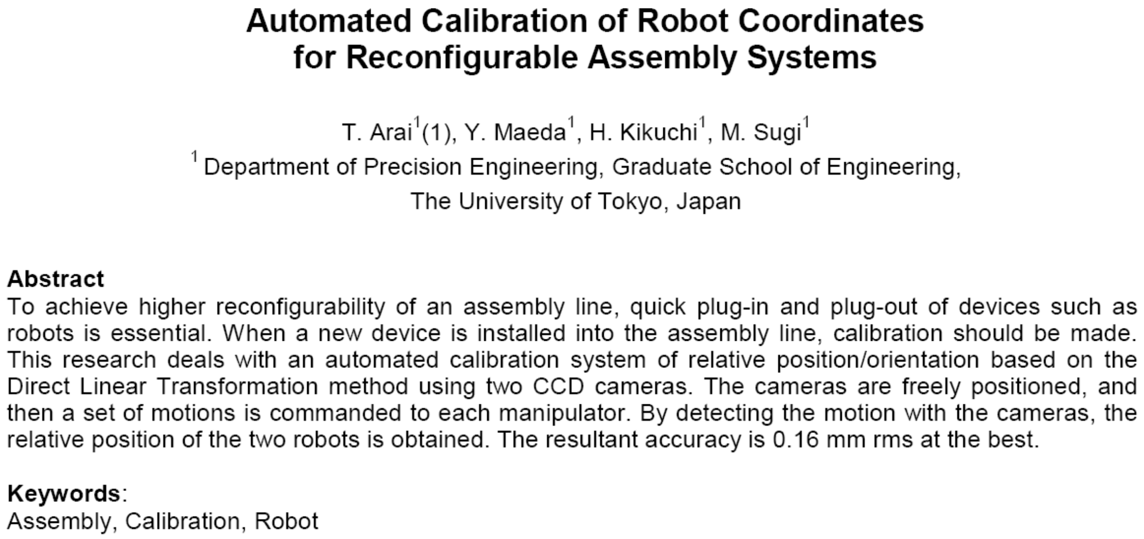
A. Watanabe[1] (3), S. Sakakibara[1], K. Ban[1], M. Yamada[1], G. Shen[1]
[1] FANUC LTD, Yamanashi, Japan
Submitted by T. Arai (1)

Abstract

Several new methods have been developed to achieve practical accuracy for offline programming of robots and its applicability to the real world. In this paper, a new kinematic calibration method is proposed to automatically improve absolute positioning accuracy of robots. Key points of the method include autonomous measurement and the automatic generation of measuring poses. A new visual feedback motion control method of the robot is proposed to achieve accurate measurement. An algorithm is also proposed to improve the condition of measuring poses automatically. The effectiveness of the proposed methods and algorithm was investigated through experiments with actual robots.

Keywords:
Robot Calibration, Visual Measurement, Positioning Accuracy

Figure 40: Title and abstract of a document with second highest affinity with topic "Robot Vision"



Automated Calibration of Robot Coordinates for Reconfigurable Assembly Systems

T. Arai[1](1), Y. Maeda[1], H. Kikuchi[1], M. Sugi[1]
[1] Department of Precision Engineering, Graduate School of Engineering,
The University of Tokyo, Japan

Abstract
To achieve higher reconfigurability of an assembly line, quick plug-in and plug-out of devices such as robots is essential. When a new device is installed into the assembly line, calibration should be made. This research deals with an automated calibration system of relative position/orientation based on the Direct Linear Transformation method using two CCD cameras. The cameras are freely positioned, and then a set of motions is commanded to each manipulator. By detecting the motion with the cameras, the relative position of the two robots is obtained. The resultant accuracy is 0.16 mm rms at the best.

Keywords:
Assembly, Calibration, Robot

Figure 41: Title and abstract of a document with third highest affinity with topic “Robot Vision”

By scanning the titles and abstracts of these papers it can be seen that they indeed correspond to the subject of “Robot Vision”.

11.6.4 Finding Topics Closest to a given Topic

The topic model, generated by CAT, also captures the calculated similarity among topics. For instance, the two topics closest to *Robot Vision* (Topic 13) are given as *Stepping, Piezoelectric Motors & Actuators* (Topic 9), *Drilling & Turning Cooling Lubricants, Tool Coating Design* **|** *Surface Roughness in Hard Turning* (Topic 44). Although Topic 9 may correspond to the control of robots, it is unclear (to a non-expert) how Topic 44 can relate to *Robot Vision*. As a second example, the topics found to be closest to *Bone Cutting* (Topic 50) were *Vibration Cutting* (Topic 20) and *Drilling & Turning Cooling Lubricants, Tool Coating Design* **|** *Surface Roughness in Hard Turning* (Topic 44). Again, the association of the second topic is hard to explain (for a non-expert) in the light of the topic *Bone Cutting*. This might be due to the fact that most topics in the 100-topic analysis are very specific and therefore have little in common with other topics. As a third and last example, the topics found to be closest to topic *Tool Wear & Tool Life* (Topic 22) were *Cutting & Chip Formation* (Topic 14) and *Coating & Wear* (Topic 73) which seem more satisfactory (to a non-expert) than the first two examples.

11.6.5 Finding Topics having the Best Coverage

Using the document-topic assignment scores of the topic model, it is possible to calculate the topics that are best represented in the analysed papers as explained in section 6.5.6. Table 82 shows the ten topics having the highest coverage scores, in term of the analysed papers, relative to other topics. The percentage given for each topic indicates the degree of representation of the topic in question relevant to other topics in the 100-topic analysis.

Topic	Topic Coverage
Topic 70: Tools, Models & Systems for Optimising Design- & Production Processes	11.21%
Topic 62: Surface Topography & Processing	11.04%
Topic 87: Drive System Control	2.54%
Topic 75: Measurement Accuracy & Calibration	2.47%
Topic 5: Milling	2.44%
Topic 61: Grinding Wheels	1.99%
Topic 54: Product Development, Design Process & Knowledge	1.89%
Topic 58: Finite Element Analysis & FEM Simulation	1.83%
Topic 22: Tool Wear & Tool Life	1.82%
Topic 14: Cutting & Chip Formation	1.81%

Table 82: Ten topics having the highest coverage in the analysed document collection

These topics may be interpreted as topics that have, according to the calculated topic model, good representation in the analysed documentation. These topics may be regarded as either topics central to the field of manufacturing and production processes or very general topics.

11.6.6 Finding Topics having the Poorest Coverage

Conversely, Table 83 presents the topics that have the weakest representation in the analysed articles according to the generated topic model.

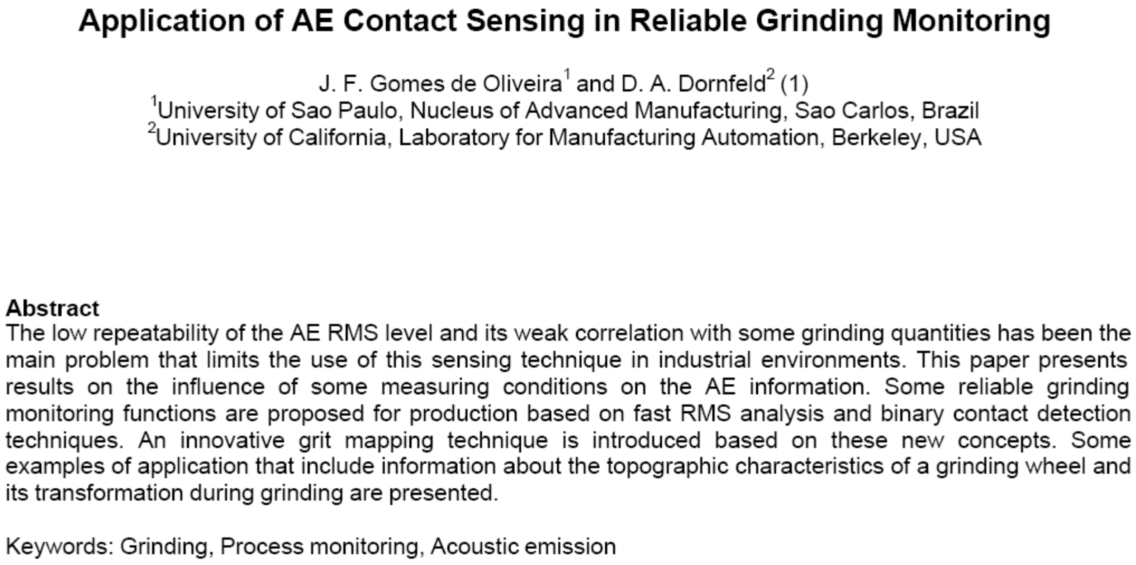
Topic	Topic Coverage
Topic 44: Drilling & Turning Cooling Lubricants, Tool Coating Design | Surface Roughness in Hard Turning	0.13%
Topic 28: Glass Drilling & Cracks	0.19%
Topic 19: New Methodologies (Chatter Detection | Enterprise Knowledge Model| Components Reuse)	0.21%
Topic 86: Tungsten Clad Rods & Residual Stress	0.22%
Topic 59: Micro Objects & Micro Handling	0.22%
Topic 83: Manufacturing Paradigms	0.25%
Topic 38: Damping Ratios	0.28%
Topic 9: Stepping, Piezoelectric Motors & Actuators	0.28%
Topic 68: Automatic Revolving Door Safety	0.30%
Topic 99: Multiresolution CSG Models | Carbon Nanotube Assembly	0.30%

Table 83: Ten topics having the poorest coverage in the analysed document collection

These topics may be interpreted as topics sitting on the boundary of the field or being very recent or very specific topics and therefore not having many associated papers. When investigating the labels of the weakly covered topics, they indeed seem to be more specific than the best covered topics.

11.6.7 Finding other Documents relevant to a given Document

The topic model, generated by CAT, also captures affinities between the different documents in the analysis. These affinities may be exploited to find other documents related to a given document. For example, the following paper (G04_Oliveira.pdf) was published in 2001 in the Abrasive Process STC (STC-G).



Application of AE Contact Sensing in Reliable Grinding Monitoring

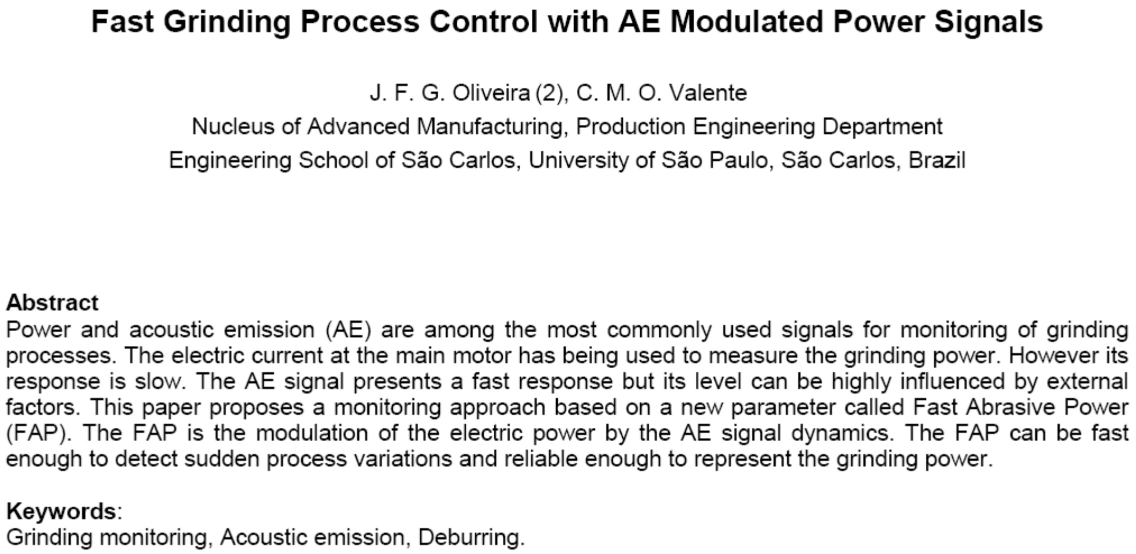
J. F. Gomes de Oliveira[1] and D. A. Dornfeld[2] (1)
[1]University of Sao Paulo, Nucleus of Advanced Manufacturing, Sao Carlos, Brazil
[2]University of California, Laboratory for Manufacturing Automation, Berkeley, USA

Abstract
The low repeatability of the AE RMS level and its weak correlation with some grinding quantities has been the main problem that limits the use of this sensing technique in industrial environments. This paper presents results on the influence of some measuring conditions on the AE information. Some reliable grinding monitoring functions are proposed for production based on fast RMS analysis and binary contact detection techniques. An innovative grit mapping technique is introduced based on these new concepts. Some examples of application that include information about the topographic characteristics of a grinding wheel and its transformation during grinding are presented.

Keywords: Grinding, Process monitoring, Acoustic emission

Figure 42: An example of a target document

The following two documents (G06_Oliveira.pdf and G09_Brinksmeier.pdf published in 2004 and 2005 respectively in STC G) were found to be the closest to this document



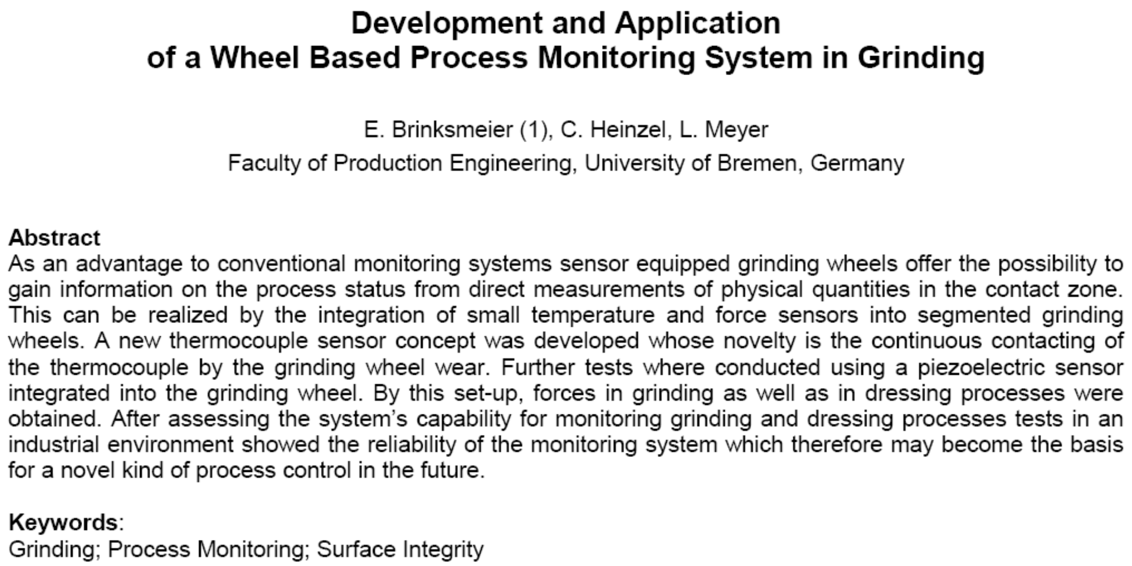
Fast Grinding Process Control with AE Modulated Power Signals

J. F. G. Oliveira (2), C. M. O. Valente
Nucleus of Advanced Manufacturing, Production Engineering Department
Engineering School of São Carlos, University of São Paulo, São Carlos, Brazil

Abstract
Power and acoustic emission (AE) are among the most commonly used signals for monitoring of grinding processes. The electric current at the main motor has being used to measure the grinding power. However its response is slow. The AE signal presents a fast response but its level can be highly influenced by external factors. This paper proposes a monitoring approach based on a new parameter called Fast Abrasive Power (FAP). The FAP is the modulation of the electric power by the AE signal dynamics. The FAP can be fast enough to detect sudden process variations and reliable enough to represent the grinding power.

Keywords:
Grinding monitoring, Acoustic emission, Deburring.

Figure 43: Document having the highest affinity with target document



Development and Application
of a Wheel Based Process Monitoring System in Grinding

E. Brinksmeier (1), C. Heinzel, L. Meyer
Faculty of Production Engineering, University of Bremen, Germany

Abstract
As an advantage to conventional monitoring systems sensor equipped grinding wheels offer the possibility to gain information on the process status from direct measurements of physical quantities in the contact zone. This can be realized by the integration of small temperature and force sensors into segmented grinding wheels. A new thermocouple sensor concept was developed whose novelty is the continuous contacting of the thermocouple by the grinding wheel wear. Further tests where conducted using a piezoelectric sensor integrated into the grinding wheel. By this set-up, forces in grinding as well as in dressing processes were obtained. After assessing the system's capability for monitoring grinding and dressing processes tests in an industrial environment showed the reliability of the monitoring system which therefore may become the basis for a novel kind of process control in the future.

Keywords:
Grinding; Process Monitoring; Surface Integrity

Figure 44: Document having the second highest affinity with target document

By investigating the keywords, titles and abstracts of the three papers, it can be concluded that they address the monitoring of the grinding process. What is further interesting is that the paper identified as the closest to the paper in focus shared the same main author (J.F.G. Oliveira). These results are therefore quite satisfactory. The document affinity relationships may be extremely useful in identifying more documents to read about the topics addressed in a given document. This capability may be especially helpful when one is exploring a large collection of documents to gather new insights as in the case of the initial stages of the innovation process.

11.6.8 Finding Words Associated with a given Word

In addition to the associations between words, topics and documents, the topic model can also be used to represent associations between individual words. A minimum frequency value of 5 was specified for this analysis implying that words occurring less than five times in total in the input document set will be discarded in the analysis.

Figure 45 represents the words associated with the term “laser sintering”[76] that formed part of the relevant topic model as shown in the *CAT Visualisation Centre*.

[76] ‘Selective laser sintering’ is an additive rapid manufacturing technique which employs a high power laser to fuse small particles of plastic, metal, ceramic, or glass powders into a solid representing a desired 3-dimensional object.

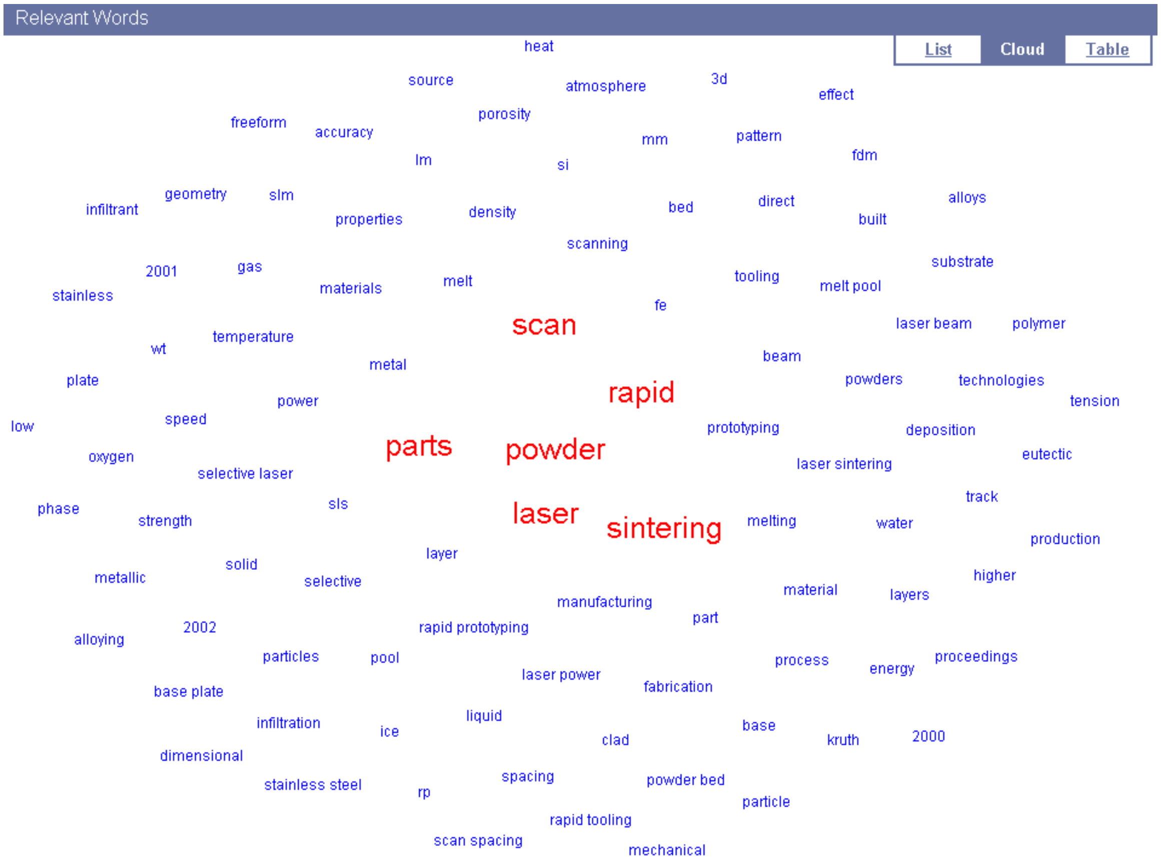


Figure 45: Tag cloud view of the words associated with topic "Laser Sintering"

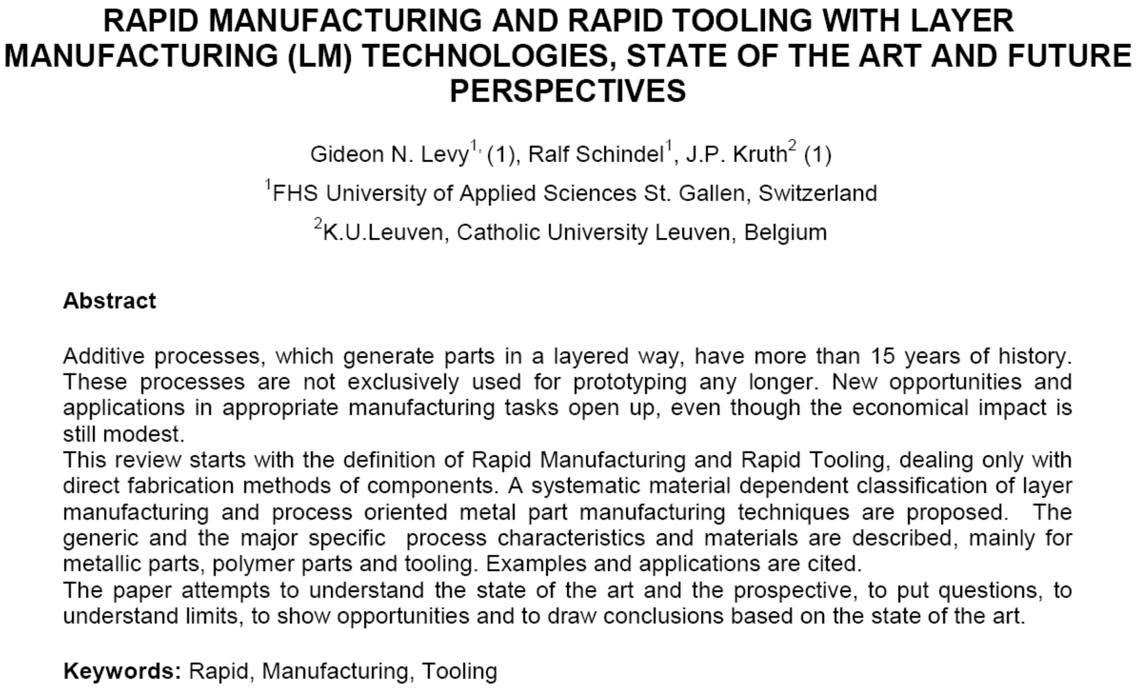
The words presented in Figure 45 give one a good indication of what the technique entails.

11.6.9 Finding Topics Associated with a given Word

When focusing on a given word, the topics associated with the word in focus may also be found. The topic *Rapid Prototyping & Laser Sintering* (Topic 43) has the strongest association with the term "laser sintering" which is a satisfactory result. As a second example, the following topics were found to be associated with the word "oil": *Minimal Quantity Lubrication Cutting* (Topic 65), *Friction* (Topic 45), *Bone Cutting* (Topic 50), *Punching* (Topic 51) and *Electric Discharge Machining & Electrodes* (Topic 96).

11.6.10 Finding Documents Associated with a given Word

The documents associated with a given word may also be found using the topic model. The following two papers (published in 2003 and 2006 respectively under Electro-physical and Chemical Processes STC, STC-E) have the strongest association with the term "laser sintering".



RAPID MANUFACTURING AND RAPID TOOLING WITH LAYER MANUFACTURING (LM) TECHNOLOGIES, STATE OF THE ART AND FUTURE PERSPECTIVES

Gideon N. Levy[1,] (1), Ralf Schindel[1], J.P. Kruth[2] (1)

[1]FHS University of Applied Sciences St. Gallen, Switzerland

[2]K.U.Leuven, Catholic University Leuven, Belgium

Abstract

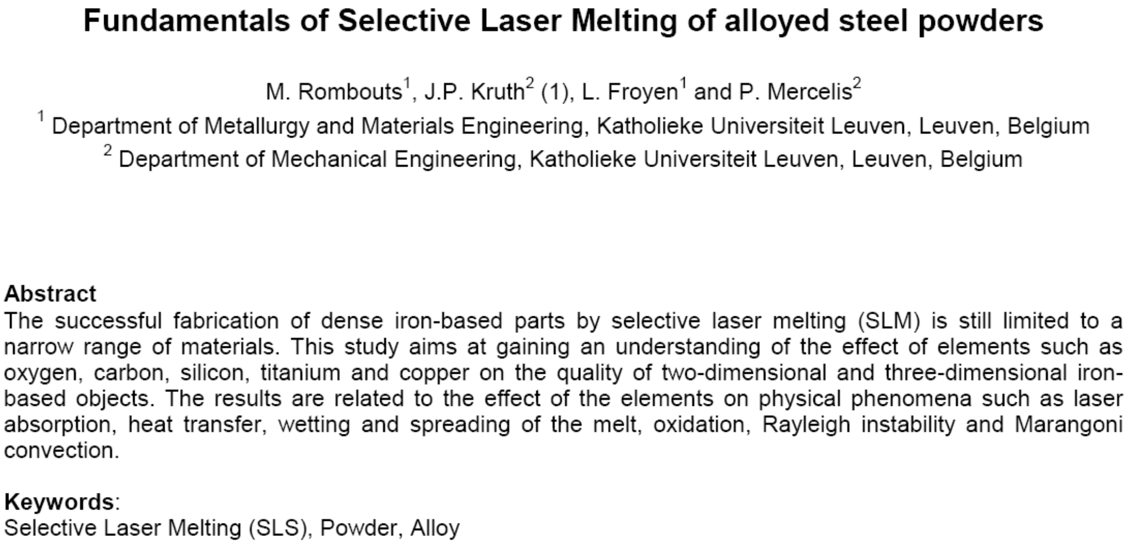
Additive processes, which generate parts in a layered way, have more than 15 years of history. These processes are not exclusively used for prototyping any longer. New opportunities and applications in appropriate manufacturing tasks open up, even though the economical impact is still modest.

This review starts with the definition of Rapid Manufacturing and Rapid Tooling, dealing only with direct fabrication methods of components. A systematic material dependent classification of layer manufacturing and process oriented metal part manufacturing techniques are proposed. The generic and the major specific process characteristics and materials are described, mainly for metallic parts, polymer parts and tooling. Examples and applications are cited.

The paper attempts to understand the state of the art and the prospective, to put questions, to understand limits, to show opportunities and to draw conclusions based on the state of the art.

Keywords: Rapid, Manufacturing, Tooling

Figure 46: Document having the highest affinity with term "Laser Sintering"



Fundamentals of Selective Laser Melting of alloyed steel powders

M. Rombouts[1], J.P. Kruth[2] (1), L. Froyen[1] and P. Mercelis[2]

[1] Department of Metallurgy and Materials Engineering, Katholieke Universiteit Leuven, Leuven, Belgium

[2] Department of Mechanical Engineering, Katholieke Universiteit Leuven, Leuven, Belgium

Abstract

The successful fabrication of dense iron-based parts by selective laser melting (SLM) is still limited to a narrow range of materials. This study aims at gaining an understanding of the effect of elements such as oxygen, carbon, silicon, titanium and copper on the quality of two-dimensional and three-dimensional iron-based objects. The results are related to the effect of the elements on physical phenomena such as laser absorption, heat transfer, wetting and spreading of the melt, oxidation, Rayleigh instability and Marangoni convection.

Keywords:

Selective Laser Melting (SLS), Powder, Alloy

Figure 47: Document having the second highest affinity with term "Laser Sintering"

By examining the titles, abstracts and keywords of these papers it seems as if these documents are indeed very applicable to the term "laser sintering".

11.6.11 Finding Author Profiles

One may further attempt to use the topic model to find a keyword profile for individual authors by exploiting the associations between words captured in the model. A profile of a given author can be established by focusing on the surname of the author and obtaining all words associated with the person in question. Figure 48 shows the author profile of a CIRP author with the surname "Uhlmann" as shown in the *CAT Visualisation Centre*.

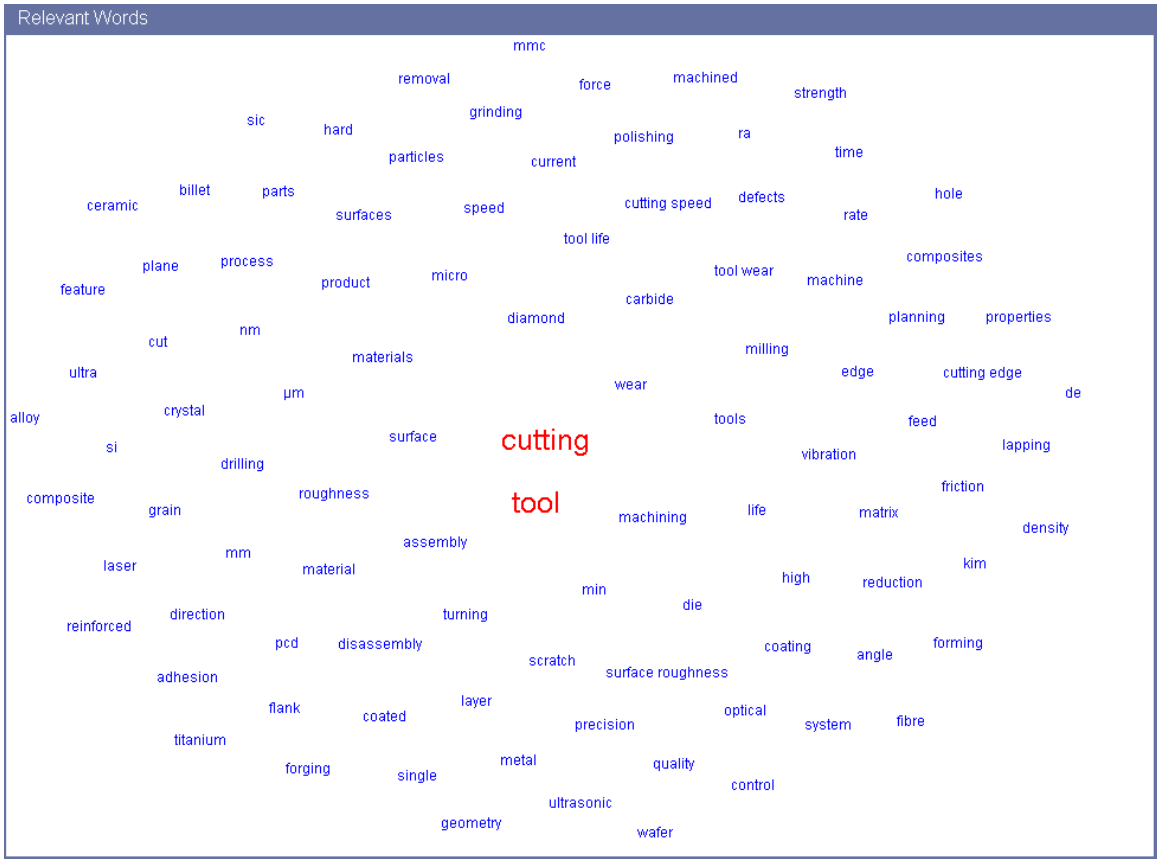


Figure 48: Words associated with the word "Uhlmann"

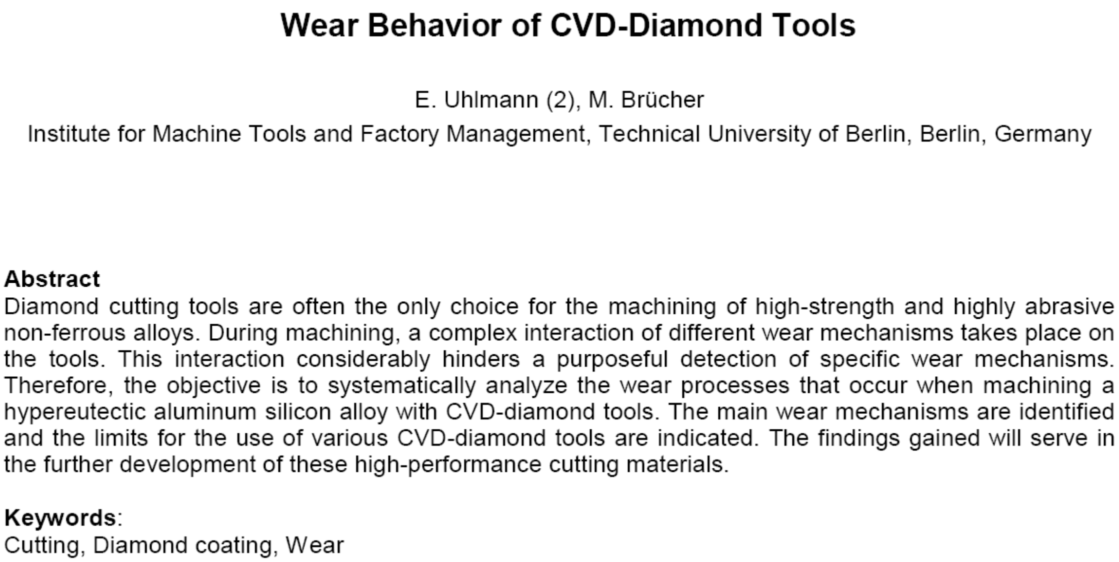
It must be mentioned that this approach will only give sensible results for well-published authors since authors names are only words in the topic model and not explicitly differentiated from other words in the model. Another problem with using author surnames to find author profiles is that a given surname may refer to more than one person which may result in a combined profile of all authors with the surname in question. The results of this approach may therefore not always be accurate. A more accurate technique for obtaining author profiles is presented in section 11.7.1 of this chapter.

The following documents were found to be associated with the word "Uhlmann" using the topic model:

- Auteurs 2005.pdf
- Auteurs 2003.pdf
- Auteurs 2004.pdf

These documents contain only the surnames and initials of authors that published in the specific year. On closer examination only the first document contained the surname “Uhlmann” at all. No other paper was significantly associated with this author even though the author actually authored 4 papers. The author-document association proved to be wanting in this case affirming the need for a more explicit method of relating words and documents to authors.

The model further indicated that Topics 22 (Tool Wear & Tool Life) and have the strongest associations with the author’s surname. On closer inspection it was found that two documents authored by “Uhlmann” were under the top 25 documents associated with this topic (positions 7 and 23 respectively). Figure 49 shows the title and the abstract of the paper authored by “Uhlmann” having the 7th highest association with Topic 22.



Wear Behavior of CVD-Diamond Tools

E. Uhlmann (2), M. Brücher
Institute for Machine Tools and Factory Management, Technical University of Berlin, Berlin, Germany

Abstract
Diamond cutting tools are often the only choice for the machining of high-strength and highly abrasive non-ferrous alloys. During machining, a complex interaction of different wear mechanisms takes place on the tools. This interaction considerably hinders a purposeful detection of specific wear mechanisms. Therefore, the objective is to systematically analyze the wear processes that occur when machining a hypereutectic aluminum silicon alloy with CVD-diamond tools. The main wear mechanisms are identified and the limits for the use of various CVD-diamond tools are indicated. The findings gained will serve in the further development of these high-performance cutting materials.

Keywords:
Cutting, Diamond coating, Wear

Figure 49: Example of a document associated with topic “Tool Wear & Tool Life”

Judging by the title and abstract of this paper, it seems as if the topic with the highest association to the author can reasonably be justified.

11.6.12 Finding General Terms for the given Field

As part of the analysis, the software automatically constructed a vocabulary based on the given input documents analysed. For each word in the vocabulary, the topics associated with the given word are given along with the strength of such associations. The associated topics can be seen as providing context to the words of the vocabulary. A generality score is given for each word indicating how general or specific the given word is relative to all other words in the vocabulary.

The words (ranked vertically in columns according to descending generality scores) in the left hand section of Table 84 are associated with many topics with high association scores and may therefore be considered as general terms in the light of CIRP. The words and terms (ranked vertically in columns according to ascending generality scores) in the right hand section of Table 84, on the other hand, are associated with only a single topic and may therefore be considered to be more specific.

Most General Terms	
cutting	machine
tool	control
surface	micro
product	assembly
design	forming
system	production
process	workpiece
machining	laser
temperature	material
grinding	die

Most Specific Terms	
fixed frame	chip thicknesses
lower resolution	coating systems
person	dispersion
piezo actuators	eliminated
position accuracy	energies
rotations	notching
university uk	polar components
advanced	solids
asynchronous	surface energies
austenitic steel	tool defects

Table 84: General and specific terms for the analysed document collection

The specific words that are found using the associated topic modelling software, CAT, in some cases represent actual, specific terminology (e.g. “notching” and “austenitic steel” in Table 84). The specific terms are however mixed with words that do not represent actual terminology (e.g. “advanced” and “person”).

A total of 4987 words formed part of the vocabulary calculated as part of the 100-topic analysis. These words may serve as a starting point to develop a taxonomy, thesaurus or glossary for CIRP or any other target organisation. CIRP has existing terminology lists for “Assembly Systems”, “Design”, and “Manufacturing Systems” as well as a unified keyword list which are all available on their website[77].

The generated vocabulary may be used to expand the CIRP terminology and keyword lists as new terms are introduced over time. If a mechanism can be found to identify the new, candidate terms introduced each year, it will greatly facilitate the maintenance of such terminology and keyword lists.

11.7 Extending the Topic Model Analysis Results

The standard topic model has three dimensions, namely topics, words and documents as previously shown in Figure 23. During the inference phase, the topics that best explain the occurrence of the relevant words in respective documents analysed, are calculated and

[77] www.cirp.net/index.php?option=com_content&task=view&id=27&Itemid=71

characterised. Also, each document is assigned to one or more topic with a given strength of association.

All words encountered in documents (more specifically, those words which are not eliminated because of stopword lists and minimum frequency filters) are merely words to the topic model in the sense that no distinction is made between say ordinary words (e.g. “sintering”) and author names (e.g. “uhlmann”), and publication year values (e.g. “2007”). In order to readily integrate the topic modelling results with the framework discussed in Chapter 8, it is required to reliably associate topics to the different entities of the “Person” entity type as well as the entities of the “Organisational Unit” entity type to be able to create information interest profiles for such individuals and groups.

There are ways of extending the topic model approach by formally and explicitly adding either an author dimension (e.g. Rosen-Zvi et al., 2004) or a time dimension (e.g. Blei et al., 2006 [2]). These techniques were however not investigated due to the difficulty of programmatic implementation and the fact that they cater for very specific cases whereas the general LDA topic model used thus far has good all round performance. Therefore, the author attempted to devise ways to explicitly incorporate time, author and STC dimensions into the standard topic model results data rather than attempting to change the LDA technique itself. This process is explained in the following sections.

11.7.1 Incorporating Explicit Author Information

The need to find author-topic affinities motivated the explicit integration of the author dimension into the topic model results. The advantages of having this information available will be illustrated throughout this section.

In order to incorporate author (closely related to the “Person” entity type of the framework) information into the topic model, each paper was opened individually and the authors were copied and pasted into the Excel output file of the 100-topic analysis generated by the software, to create a “Documents-Authors” matrix in a new spreadsheet. This matrix contained the following information[78]:

- Filenames of all files analysed
- STC associated with each respective file
- Year of publication of each file
- Authors of each file

Extreme care was taken to ensure that all author names were in the same format to minimise duplication when a unique author list will be compiled later in the process. Since most of the

[78] If the documents resided in a document management system and the authors, (readers,) publication (or modification) dates and STCs (departments) associated with the different documents were recorded as part of the metadata of the individual documents, this process can easily have been automated.

authors names were in the format *initial(s). surname*, cases where full first names of authors were listed were converted to the format above. Unnecessary spaces were further eliminated and all initials not followed by a full stop were corrected in order to promote consistency. All authors listed for a given paper were captured with the maximum number of authors encountered being thirteen. No distinction was made between the first author and other authors assuming that all listed authors are equally associated with a given paper. An example of the data captured for two papers, as example entries in the "Documents-Authors" matrix, are shown below.

Filename	STC	Year	Author	Author	Author	Author
52-1-2003-41	STC_C	2003	Y. Takeuchi	M. Murota	T. Kawai	K. Sawada
F01_Hirt	STC_F	2004	G. Hirt	J. Ames	M. Bambach	R. Kopp

Table 85: Extract from the "Documents-Authors" matrix

Once the "Documents-Authors" matrix has been fully populated, the next step was to find the authors associated with each of the discovered topics. This was achieved by identifying all documents associated with a given topic with a weight[79] of more than or equal to 0.20. The value of 0.20 was empirically determined by examining what the lowest document-topic weight is that returns papers relevant to a given topic. The filenames of the documents fulfilling this criterion were then copied and pasted next to the relevant topic number in a new spreadsheet. The "Topics-Documents-Authors" matrix was constructed in this way. This process was repeated for all 100 topics. Excel's *vlookup* function was subsequently used to retrieve the authors of each of the respective papers from the "Documents-Authors" matrix constructed in the previous step. Table 86 shows an example of the documents and authors associated with a single topic, namely "Forming Processes" (Topic 2).

Topic Number	Topic Label	Significant Documents	Authors					
Topic 2	*Forming Processes*	**F06_Yoon**	S.J. Yoon	D.Y. Yang	0	0	0	…
		52-1-2003-201	S.J. Yoon	D.Y. Yang	0	0	0	…
		G04_Yanagihara	T. Tateishi	Q. Gao	Y. Tani	K. Yanagihara	H. Sato	…
		Dn06_Guttman	G. Guttman	M. Shpitalni	0	0	0	…
		Dn14_Schuh	G. Schuh	0	0	0	0	…

Table 86: Extract from the "Topics-Documents-Authors" matrix

Once the "Topics-Documents-Authors" matrix was fully populated, another matrix, called the "Topics-Authors" matrix was created in a new spreadsheet by processing the entries in the "Topics-Documents-Authors" matrix to create a unique list of authors for each topic. The author

[79] Recall that document-topic weights (also known as document mixture ratios) are values between 0 and 1.

names were further split into initials and surname and the author list for each topic was sorted alphabetically using the authors' surnames. Table 87 shows an extract of this matrix.

	Topic 3		**Topic 4**			**Topic 81**		
…	***Gears & Gear Metrology***		***Polishing & Spray Painting***		…	***Reuse & Reusability***		…
…	Goch	G.	Arai	T.	…	Arai	T.	…
…	Guenther	A.	Arikan	M.A.S.	…	Ham	M.	…
…			Balkan	T.	…	Jeswiet	J.	…
…			Bley	H.	…	Kondoh	S.	…
…			Braun	P.	…	Lung	N.	…
…			Fischer	N.	…	Muller	A.	…
…			Kikuchi	H.	…	Portman	V.	…
…			Maeda	Y.	…	Rubenchik	Y.	…
…			Sugi	M.	…	Shimomura	Y.	…
…			Touge	M.	…	Shneor	Y.	…
…			Watanabe	J.	…	Shuster	V.	…
…					…	Sugino	T.	…
…					…	Suhner	M.C	…
…					…	Umeda	Y.	…
…					…	Véron	M.	…

Table 87: Extract from the "Topics-Authors" matrix

The "Topics--Authors" matrix gives one a good impression of which persons are knowledgeable on what topics. It is important to note that a given person may be an expert in more than one topic, as in the case of person "T. Arai" in Table 87.

Subsequently, the "Author-Topic" matrix was constructed by copying all authors from the "Topics-Documents-Authors" matrix and removing all duplicates in terms of author names to form the vertical axis of the matrix, and using all topics to form the horizontal axis of the matrix. Excel's *countif* function was then used to fill the cells of the matrix with numbers indicating how many times a given author is associated with a given topic in the "Topics-Documents-Authors" matrix. The matrix contained a total of 1481 authors resulting in a 100 by 1481 matrix. Duplicates in author names may still be present due to the following reasons:

1. Some authors having more than one initial gave a single initial in some papers and all initials in other papers (e.g. H.M. Wang and H. Wang).
2. Surnames containing special characters were not always spelled consistently in all papers (e.g. R. Züst and R. Zust).

These suspected duplicates were however not removed due to the danger of removing false positives. Table 88 shows an extract of the "Author-Topic" matrix.

				Quality Controls & Process Chains	Forming Processes	Gears & Gear Metrology	…
Author Number	Author Surname	Author Initials	Author Name	Topic 1	Topic 2	Topic 3	…
..	..	..	..	..	..	..	..
395	Glardon	R.	R. Glardon	0	0	0	..
396	Glass	R.	R. Glass	1	0	0	..
397	Glavonjic	M.	M. Glavonjic	0	0	0	..
398	Göbel	R.	R. Göbel	0	0	0	..
399	Goch	G.	G. Goch	1	0	1	..
400	Göhringer	J.	J. Göhringer	0	0	0	..
..	..	..	..	..	..	..	..
				18	11	2	

Table 88: Extract from the "Author-Topic" matrix

Using the "Author-Topic" matrix one can find the topics a given person specialises in as well as the experts for a given topic.

The following topics were found to be associated with the author "Uhlmann" using the "Author-Topic" matrix: "Milling" (Topic 5), "Tool Wear & Tool Life" (Topic 22), "Micro Products & Micromachining" (Topic 23), "Surface Topography & Processing" (Topic 62), "Coating & Wear" (Topic 73), "Drive System Control" (Topic 87). This technique of calculating author-topic associations are far more accurate than the technique presented in section 11.6.11.

A total of 280 unique persons were identified as possible research partners for "Uhlmann" by looking at the authors associated with either one of the six topics linked to "Uhlmann" in the "Topics Authors" spreadsheet. The associated authors were then copied to a list and duplicates were removed to create the "Associated Authors List" for the given author. Table 89 shows an extract from the "Associated Authors List" of "Uhlmann". Note that the topics these authors are associated with, which are not associated with "Uhlmann" also, are not shown in Table 89.

		Milling	*Tool Wear & Tool Life*	*Micro Products & Micromachining*	*Surface Topography & Processing*	*Coating & Wear*	*Drive System Control*
Author Number	**Author Name**	**Topic 5**	**Topic 22**	**Topic 23**	**Topic 62**	**Topic 73**	**Topic 87**
5	E. Abele	1	0	0	1	0	0
…	…	…	…	…	…	…	…
38	M. Arizmendi	0	1	0	0	0	0
…	…	…	…	…	…	…	…
446	A. Hansel	1	0	0	0	0	1
447	H.N. Hansen	1	0	1	1	0	0
448	M. Hao	1	0	0	0	0	0
…	…	…	…	…	…	…	…
550	I.S. Jawahir	0	0	0	0	1	0
…	…	…	…	…	…	…	…
1451	M. Zatarain	1	0	0	1	0	1
…	…	…	…	…	…	…	…

Table 89: Extract of the possible research partners identified for person “Uhlmann”

Using this approach, potential collaborators could be identified on one or more specific topics or a combination of specific topics.

The following section explains how STC information was incorporated into the topic model results.

11.7.2 Incorporating Explicit STC Information

The main reason for explicitly incorporating the STC dimension into the results of topic model is to find implicit relationships between STCs and the topics generated from the analysed documentation. This actually corresponds to finding which topics are associated with which organisational unit (e.g. team, department, etc.) of the innovating organisation and vice versa. This would facilitate knowledge sharing and the assignment of appropriate resources to innovation tasks to name but a few advantages. The process of incorporating a STC dimension into the topic model results will now be explained.

The “Topics-Documents-STC” matrix was constructed in a new spreadsheet by making a duplicate of the spreadsheet containing the “Topics-Documents-Authors” matrix and replacing the author columns with a STC column, indicating in which STC each paper was published. Table 90 shows an extract of this matrix.

Topic Number	Topic Label	Significant Documents	STC
Topic 2	*Forming Processes*	**F06_Yoon**	F
		52-1-2003-201	F
		G04_Yanagihara	G
		Dn06_Guttman	Dn
		Dn14_Schuh	Dn

Table 90: Extract from the "Topics-Documents-STC" matrix

Lastly, the "STC-Topic" matrix was created in a new spreadsheet with topics on the vertical axis and STCs on the horizontal axis resulting in an 11 by 100 (representing the different topics calculated in the 100-topic analysis) matrix. Microsoft Excel's *countif* function was once more used to fill the cells of the matrix with numbers indicating how many times a given topic is associated with a given STC in the "Topics-Documents-STC" matrix. Table 91 presents a part of this matrix.

Topic Label	Topic Number	*STC-A*	*STC-C*	*STC-Dn*	*STC-E*		# STCs
Quality Controls & Process Chains	**Topic 1**	0	0	0	0	…	**4**
Forming Processes	**Topic 2**	0	0	2	0	…	**3**
Gears & Gear Metrology	**Topic 3**	0	0	0	0	…	**1**
Polishing & Spray Painting	**Topic 4**	2	0	0	0	…	**3**
Milling	**Topic 5**	0	14	0	0	…	**3**
Electrochemical Processes	**Topic 6**	0	1	0	9	…	**3**
Grippers & Part Handling	**Topic 7**	2	0	0	0	…	**2**
Dies & Ironing	**Topic 8**	1	0	0	0	…	**3**
…	…	…	…	…	…	…	…
# Unique Topics Associated with STC		**25**	**30**	**24**	**20**		

Table 91: Extract from the "STC-Topic" matrix

The bottom row in Table 91 indicates how many different topics are associated with each STC, while the rightmost column indicates how many different STCs are associated with each topic. This matrix may be used to get an indication of which topics the respective STCs deal with and to what extent (e.g. from Table 91 it may be inferred that STC-C deals much more with milling than with electrochemical processes). Conversely, one could also find which STCs deals with a specific topic (e.g. it can be seen from the example data in Table 91 that if one is interested in the topic "Grippers & Part Handling", the research of STC-A may be appropriate). Therefore, the table can be used to decide which STCs to include in a multi-disciplinary team to do collaborative research on a given topic. On the CIRP website the following 'research tracks'[80] were found:

[80] In CIRP, 'research tracks' represent combinations of different STCs participating in collaborative meetings and research.

Research Track 1: Gathers STC-C, STC-E, STC-F and STC-G. Using the “STC-Topic” matrix the following topics were identified as common among these four STCs:

- Topic 62 (Surface Topography & Processing)
- Topic 77 (Strength & Residual Stress)

Research Track 2: Gathers STC-M, STC-P and STC-S. The following topics were identified as common among these three STCs:

- Topic 62 (Surface Topography & Processing)
- Topic 75 (Measurement Accuracy & Calibration)
- Topic 94 (Diamond Machining Processes & Surface Roughness)

Lastly, STC-A, STC-Dn and STC-O participate in **Research Track 3**. For this research track, the following topics were identified as common among these three STCs:

- Topic 24 (Virtual Reality & Haptic Systems)
- Topic 35 (Laser Forming & Bending | Lifecycle Management & Recycling)
- Topic 37 (Manufacturing System Configuration)
- Topic 42 (Product Lifecycles)
- Topic 66 (Reverse Engineering & Surface Reconstruction)
- Topic 70 (Tools, Models & Systems for Optimising Design- & Production Processes)
- Topic 76 (Fuzzy Decision-making Methods)
- Topic 90 (Thread Tapping Tolerance)

It can be seen that Research Track 3 deals with a much wider range of topics when compared to Research Track 1. Table 92 shows the number of distinct topics associated with each STC.

STC	# Topics
STC-A	25
STC-C	30
STC-Dn	24
STC-E	20
STC-F	27
STC-G	24
STC-M	36
STC-N	9
STC-O	31
STC-P	18
STC-S	19

Table 92: Number of distinct topics associated with the respective STCs

STC-N is associated with the least number of topics which can be justified since it was merged with STC P after 2002. STC-M is associated with the most topics and may be considered very broad. Conversely, STC-S and STC-P have the smallest number of associated topics after STC-N indicating that these STCs are narrower in focus when compared to say, STC-M.

The following topics were found to be associated with a single STC. These topics may be candidates for expansion by involving other STCs to gain new perspectives.

Topic Label	Topic Number	Associated STC
Gears & Gear Metrology	Topic 3	P
Production System Control	Topic 10	O
Cutting & Chip Formation	Topic 14	C
Forming Processes	Topic 18	F
Production & Supply Chain Management	Topic 29	O
Roll Forming	Topic 31	F
Measurement Sensors	Topic 34	P
Parallel Kinematic Machines	Topic 40	M
Drilling & Turning Cooling Lubricants, Tool Coating Design | Surface Roughness in Hard Turning	Topic 44	C
Cooling & Heat Transfer	Topic 56	F
Titanium & Deep Drawing	Topic 72	F
Sheet Metal Forming Processes	Topic 74	F
Extrusion	Topic 78	F
Production Planning & Resource Scheduling	Topic 84	O
Tungsten Clad Rods & Residual Stress	Topic 86	E
Self-Organising Biological Manufacturing Systems	Topic 88	O
Centerless Grinding & Stability	Topic 95	G
Assembly Processes & Systems	Topic 100	A

Table 93: Topics associated with a single STC

On the other hand, the following topics were associated with more than one STC.

Topic Label	Topic Number	Associated STCs
Injection Moulding & Temperature	Topic 25	C, Dn, E, M, P
Dressing & Trueing	Topic 27	A, Dn, E, G, M
Laser Forming & Bending | Lifecycle Management & Recycling	Topic 35	A, C, Dn, E, O
Surface Topography & Processing	Topic 62	C, Dn, E, F, G, M, N, P, S
Tools, Models & Systems for Optimising Design- & Production Processes	Topic 70	A, C, Dn, E, F, M, O
Measurement Accuracy & Calibration	Topic 75	Dn, M, O, P, S
Strength & Residual Stress	Topic 77	C, Dn, E, F, G, O
Thread Tapping Tolerance	Topic 90	A, C, Dn, O, S
Diamond Machining Processes & Surface Roughness	Topic 94	C, G, M, P, S

Table 94: Topics associated with more than one STC

One may speculate that these topics are already more multidisciplinary in nature since several STCs are involved in the associated research.

The next section outlines how the topic model results were extended to include a time dimension.

11.7.3 Incorporating Explicit Publication Date Information

The main reason for explicitly integrating publication date information into the topic model results was to incorporate a time dimension and more specifically to find topic-time trends. Such trends can be used to estimate the recentness and time span associated with the different topics. The remainder of this section will explain how this was achieved and will also illustrate some findings in this regard.

The calculated topic model results were manually extended with the year of publication of each paper analysed, enabling one to estimate a time trend for each topic. The "Topics-Documents-Years" matrix was constructed in a new spreadsheet by making a duplicate of the "Topics-Documents-STC" matrix and replacing the STC column with a Year column indicating in which year each paper was published. Analogues to the author and STC dimensions explained earlier, in the case of the time dimension the papers associated with a given topic with a weight of 0.20 or higher were captured in the last mentioned matrix. Table 95 shows a part of this matrix.

Topic Number	Topic Label	Significant Documents	Year
Topic 2	*Forming Processes*	F06_Yoon	2005
		52-1-2003-201	2003
		G04_Yanagihara	2006
		Dn06_Guttman	2006
		Dn14_Schuh	2004

Table 95: Extract from the "Topics-Documents-Years" matrix

For each of the 100 topics in this matrix, the mean publication year was calculated as well as the standard deviation of the publication years associated with the respective topics. An interval was then calculated with its lower limit being the mean publication year less the standard deviation and the upper limit being the sum of the mean publication year and the standard deviation. For each topic the minimum and maximum publication year were identified. This information was captured in a matrix in a new spreadsheet and this matrix was called the "Topics-Time" matrix. The minimum, lower limit, upper limit and maximum publication year data for each topic were then used as input for Microsoft Excel's *Stock Chart* graph. Figure 50 shows how 50 of the 100 topics[81] are positioned in the time over period 2001 to 2006. In Figure 50, the earliest published topics are shown first and the more recent published topics are shown last. It would be interesting to construct a similar graph showing topic time trends of a longer period, say 10 or 20 years, as

[81] A graph containing all 100 topics can unfortunately not be shown here in a practical way due to size restrictions. The 50 topics shown were filtered from the possible 100 topics by selecting those 50 topics having the best representation in the input documents according to the topic model.

little variation is visible in terms of the calculated topics over the six year window addressed in this study[82].

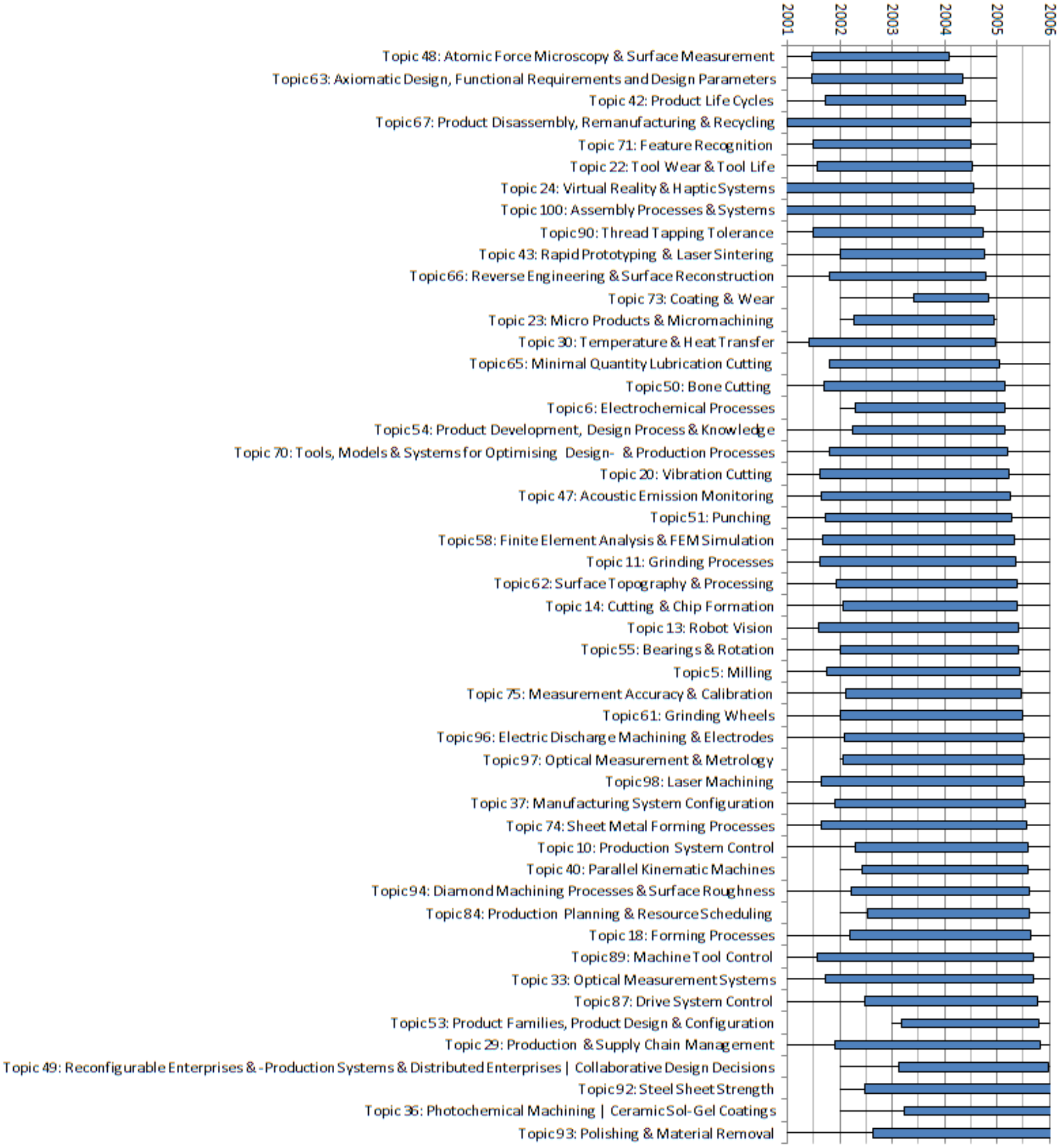


Figure 50: Topic time trends of fifty topics having highest coverage

Figure 51 shows how the 50 topics having the lowest representation in the analysed documents are positioned in the time during the period 2001 to 2006. The earliest published topics are once more shown first and the more recent published topics are shown last. One may speculate that

[82] The author performed two similar studies where the articles of two South African journals published over periods of 20 and 38 years respectively were analysed. The estimated topic time trends in these two cases have shown much more variation over the two periods studied respectively.

these 50 topics are more specific and therefore display more interesting variation over time when compared to the previous 50 which may be more general topics in the field and therefore receiving more constant attention in terms of publications.

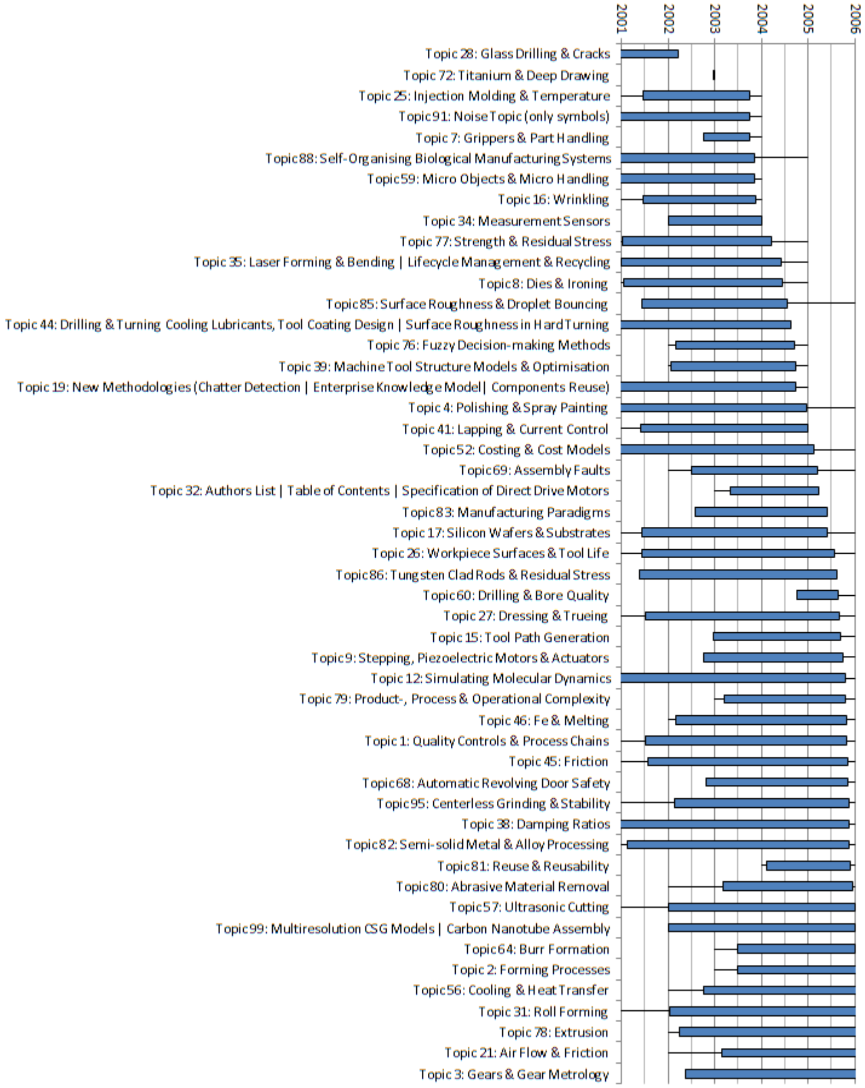


Figure 51: Topic time trends of fifty topics having lowest coverage

In two preceding figures the start and end of the thin black line of each topic bar respectively represents the year of publication of the earliest technical paper and the most recent technical paper associated with the specific topic. The middle of the blue rectangle of each topic bar represents the average year of all papers associated with the given topic and the width of the rectangle is determined by the standard deviation of all publication years of technical papers associated with the topic in question. Several views can be constructed using the time dimension of topics as explained in the following sections.

Longstanding (Persistent) Topics

An additional field, named “Max-Min Spread” was added to the “Topics-Time” matrix and associated values were calculated for each of the 100 topics by determining the difference between the maximum and minimum publication years associated with each topic. The following ten topics were identified as having the longest duration in years (in the period 2001 to 2006) by sorting according to descending “Max-Min Spread” values.

1.	Damping Ratios	6.	Roll Forming
2.	Simulating Molecular Dynamics	7.	Quality Controls & Process Chains
3.	Semi-solid Metal & Alloy Processing	8.	Friction
4.	Costing & Cost Models	9.	Dressing & Trueing
5.	Polishing & Spray Painting	10.	Workpiece Surfaces & Tool Life

Table 96: Ten topics having the longest calculated duration

Brief topics

Conversely, the following ten topics were identified as having the shortest duration in years (in the period 2001 to 2006) by sorting according to ascending “Max-Min Spread” values.

1.	Titanium & Deep Drawing	6.	Measurement Sensors
2.	Glass Drilling & Cracks	7.	Authors List | Table of Contents | Specification of Direct Drive Motors
3.	Grippers & Part Handling	8.	Reuse & Reusability
4.	Drilling & Bore Quality	9.	Gears & Gear Metrology
5.	Manufacturing Paradigms	10.	Tungsten Clad Rods & Residual Stress

Table 97: Ten topics having the shortest calculated duration

These topics were therefore only addressed in concentrated time periods whereas the longstanding topics were addressed over a longer period of time.

Recent topics

Moreover, the following topics were identified as recent (i.e. regardless of duration, topics that were addressed only in the period 2003 to 2006). These topics were found by sorting according to descending “Maximum Publication Year” values and identifying those topics also having “Minimum Publication Year” values greater or equal to 2003.

1.	Drilling & Bore Quality	6.	Product Families, Product Design & Configuration
2.	Reuse & Reusability	7.	Tool Path Generation
3.	Burr Formation	8.	Stepping, Piezoelectric Motors & Actuators
4.	Product-, Process & Operational Complexity	9.	Automatic Revolving Door Safety
5.	Forming Processes	10.	Gears & Gear Metrology

Table 98: Most recent topics

The recent topics view is interesting since it may be useful to detect newer developments in a given field. This may assist the Scout and Connector roles in the innovation process to identify potential topics to investigate further.

Outdated topics

Moreover, the following topics were identified as outdated (i.e. regardless of duration, topics that were not addressed after 2002). These topics were identified by sorting according to ascending "Maximum Publication Year" values and identifying those topics having "Maximum Publication Year" values less or equal to 2004.

1.	Glass Drilling & Cracks	5.	Drilling & Turning Cooling Lubricants, Tool Coating Design | Surface Roughness in Hard Turning
2.	Titanium & Deep Drawing	6.	Micro Objects & Micro Handling
3.	Grippers & Part Handling	7.	Wrinkling
4.	Measurement Sensors	8.	Injection Moulding & Temperature

Table 99: Most outdated topics

Although these topics were not significantly addressed in publications in the period 2005 to 2006, it does not mean that they are obsolete; only that they have received less attention in formal research in terms of CIRP publications in the recent years of the studied period.

Although not shown here, the standard deviation values of topics could also be used to identify compact topics (having small standard deviation values) and spread-out topics (having large standard deviation values). This corresponds to longstanding and brief topics to a large extent.

11.8 Calculating Topic Networks or Topic Hierarchies

The last investigation made in this case study aimed to establish whether it is possible to construct a network or hierarchy of topics by combining the results of different analyses (corresponding to different topic models) on the same input documents with different levels of granularity in terms of number topics. The availability of such a topic network or hierarchy may facilitate topic interpretation since the user can now find topics related to a given topic at different levels of aggregation.

The technique used to obtain the similarities between different topics of different analyses is very similar to the technique explained in section 10.3 dealing with finding similarities in the research interest profiles of different persons. The only difference is that in the case of research interest profile similarity calculation, the strength of association of a given word to a given topic was not part of the output of the software used at that time. Therefore, when constructing the word vectors, which are required to perform the correlation calculations, a given word was either part of a topic (denoted by a 1) or not (denoted by a 0). This resulted in binary word vectors for all topics. In the present case study, the much improved software used was able to indicate how strong a given word is associated with a given topic. The corresponding word vectors of topics, on which correlation calculations were based upon, could now be enhanced to indicate the exact strength of association of a given word in context of a given topic. Such vectors would enable more accurate similarity comparisons between different topics.

A 140 by 140 “Topic Correlations” matrix was compiled representing the inter-topic affinities of the topics respectively generated in the 10-, 30-, and 100-topic analyses. Table 100 illustrates all topics from the 10-, 30- and 100-topic analyses which are related to the first five topics of the 10-topic analysis with a correlation score of 0.40 or higher (ordered according to decreasing correlation scores). The numeric value in each cell in Table 100 is the relevant correlation score, indicating the strength of association between the relevant topic pairs.

Topic (10) 1: ***Tools, Models & Systems***	**Topic (10) 2:** ***Forming***	**Topic (10) 3:** ***Surfaces & Abrasive Processes***	**Topic (10) 4:** ***Design & Optimisation of Manufacturing Systems***	**Topic (10) 5:** ***Abrasive Processes & Electro-physical- & Chemical Processes***
Topic (30) 15: **Parallel Kinematic Machining** 0.761400555	Topic (30) 21: **Forming Processes, Stress & Strain** 0.844083181	Topic (30) 10: **Surface Roughness & Associated Factors** 0.850439998	Topic (30) 16: **Design Complexity & Axiomatic Design** 0.826356701	Topic (30) 13: **Grinding & Grinding Wheels** 0.848862547
Topic (100) 66: **Reverse Engineering & Surface Reconstruction** 0.623180229	Topic (100) 74: **Sheet Metal Forming Processes** 0.64456388	Topic (30) 18: **Polishing** 0.698731051	Topic (30) 30: **Knowledge as Support to Product Development & Design Processes** 0.807811722	Topic (100) 61: **Grinding Wheels** 0.778828379
Topic (30) 12: **Applications of Modelling & Simulation** 0.611092387	Topic (30) 26: **Punching, Drawing & Extrusion & Associated Factors** 0.63563363	Topic (100) 48: **Atomic Force Microscopy & Surface Measurement** 0.687277754	Topic (100) 63: **Axiomatic Design, Functional Requirements and Design Parameters** 0.772884404	Topic (100) 11: **Grinding Processes** 0.56776767
Topic (100) 70: **Systems supporting the Design Process** 0.464401469	Topic (100) 18: **Forming Processes** 0.609668419	Topic (100) 94: **Diamond Machining Processes & Surface Roughness** 0.579170112	Topic (100) 54: **Product Development, Design Process & Knowledge** 0.635035309	Topic (100) 95: **Centerless Grinding & Stability** 0.520063506
Topic (100) 15: **Tool Path Generation** 0.444839966	Topic (100) 58: **Finite Element Analysis & FEM Simulation** 0.544664078	Topic (100) 62: **Surface Topography & Processing** 0.525205573	Topic (30) 3: **Assembly & Product Lifecycle** 0.57759792	Topic (30) 2: **Electric Discharge Machining & Electrodes** 0.478280701
	Topic (100) 62: **Surface Topography & Processing** 0.49189002	Topic (100) 93: **Polishing & Material Removal** 0.518604489	Topic (100) 70: **Systems supporting the Design Process** 0.533448372	Topic (100) 62: **Surface Topography & Processing** 0.455448331
	Topic (30) 27: **Temperature & Manufacturing Processes** 0.441169425	Topic (100) 11: **Grinding Processes** 0.428251721	Topic (100) 53: **Product Families, Product Design & Configuration** 0.525787198	Topic (100) 47: **Acoustic Emission Monitoring** 0.453250039
	Topic (100) 92: **Steel Sheet Strength** 0.420057807		Topic (10) 8: **Assembly** 0.506875161	Topic (30) 25: **Laser Sintering** 0.422959985
			Topic (100) 42: **Product Lifecycles** 0.501425426	Topic (100) 96: **Electric Discharge Machining & Electrodes** 0.408067719
			Topic (100) 69: **Assembly Faults** 0.415732281	

Table 100: Examples of topics having affinities with topics in other topic models

Table 101 illustrates all topics from the 10-, 30- and 100-topic analyses which are related to the last five topics of the 10-topic analysis with a correlation score of 0.40 or higher (once more ordered according to decreasing correlation scores).

Topic (10) 6: Machines	**Topic (10) 7: Precision Engineering & Metrology**	**Topic (10) 8: Assembly**	**Topic (10) 9: Optimisation of Manufacturing Systems**	**Topic (10) 10: Cutting**
Topic (30) 8: **Machine Tool Control** 0.788612288	Topic (30) 28: **Measurement & Metrology** 0.848236086	Topic (30) 3: **Assembly & Product Lifecycle** 0.783335996	Topic (30) 5: **Production Systems & Production Planning** 0.814901902	Topic (30) 9: **Chip Formation & Associated Factors** 0.946551693
Topic (100) 87: **Drive System Control** 0.745551557	Topic (100) 97: **Optical Measurement & Metrology** 0.779484712	Topic (30) 30: **Knowledge as Support to Product Development & Design Processes** 0.702599592	Topic (30) 7: **Manufacturing System Configuration & Control** 0.810367985	Topic (100) 22: **Tool Wear & Tool Life** 0.885449194
Topic (30) 22: **Spindle Bearings & Spindle Systems** 0.521156632	Topic (100) 75: **Measurement Accuracy & Calibration** 0.77429305	Topic (100) 70: **Systems supporting the Design Process** 0.695365	Topic (100) 10: **Production System Control** 0.737570959	Topic (30) 1: **Coated Cutting Tools & Wear** 0.827172802
Topic (100) 89: **Machine Tool Control** 0.516450946	Topic (30) 20: **Optical Measurement** 0.652633255	Topic (30) 23: **Supply Chain Control, Flexible Automation & Production Planning** 0.6264931	Topic (30) 23: **Supply Chain Control, Flexible Automation & Production Planning** 0.656557052	Topic (100) 14: **Cutting & Chip Formation** 0.81377754
Topic (30) 4: **Milling** 0.503959009	Topic (100) 33: **Optical Measurement Systems** 0.504193857	Topic (100) 54: **Product Development, Design Process & Knowledge** 0.57126192	Topic (100) 70: **Systems supporting the Design Process** 0.639520731	Topic (30) 4: **Milling** 0.73958361
Topic (100) 5: **Milling** 0.453978364	Topic (30) 29: **Gears & Gear Metrology** 0.496515135	Topic (30) 6: **Micro-products & Associated Processes** 0.547032192	Topic (100) 37: **Manufacturing System Configuration** 0.621346501	Topic (100) 5: **Milling** 0.730559944
Topic (100) 55: **Bearings & Rotation** 0.427107	Topic (100) 3: **Gears & Gear Metrology** 0.411358031	Topic (30) 24: **Product Disassembly & Recycling** 0.536012994	Topic (100) 29: **Production & Supply Chain Management** 0.502803268	Topic (30) 11: **Vibration Cutting** 0.714122334
		Topic (100) 100: **Assembly Processes & Systems** 0.514527201	Topic (100) 84: **Production Planning & Resource Scheduling** 0.49517677	Topic (100) 50: **Bone Cutting** 0.650722833
		Topic (100) 42: **Product Lifecycles** 0.514388664		Topic (100) 73: **Coating & Wear** 0.554795155
		Topic (10) 4: **Surfaces & Abrasive Processes** 0.506875161		Topic (100) 65: **Minimal Quantity Lubrication Cutting** 0.430544654
		Topic (100) 53: **Product Families, Product Design & Configuration** 0.408338411		

Table 101: More examples of topics having affinities with topics in other topic models

Judging by the labels of the topics significantly associated with the respective topics of the 10-topic analysis, these identified inter-topic relationships seem to be very sensible. The data in the "Topic Correlations" matrix can be used to construct a topic network representing the closest topics to any of the 140 topics generated by the three topic models. Alternatively, the relevant data can be used to generate a topic hierarchy where specific topics (e.g. the topics of the 100 topic-analysis) are linked to more general topics represented by the level immediately above in terms of number of topics (e.g. the topics of the 30-topic analysis). The construction of such a topic graph or topic hierarchy can be automated fairly easily. The only manual work involved is for the analyst to provide descriptive labels for all topics in the individual analyses and to provide a threshold value indicating which inter-topic correlation values should be considered to represent significant inter-topic relationships.

The relationships between the 140 topics can be depicted as a graph where the thickness of the lines connecting the relevant topic nodes is proportional to the association strengths of the respective topic pairs.

For an innovating organisation, the value in having a system that allows users to navigate between topics of different levels of specificity may be the following:

- The ability to discover which newly generated topics which are associated with known topics can aid individuals to keep track with the fast pace of evolution in the knowledge areas associated with the organisation's innovation endeavours.
- The ability to find which specific topics are linked to a more general topic, and vice versa, will help an individual to better understand the field in question.
- Using the calculated affinities between topics, innovation workers may uncover previously undiscovered relationships between topics which may aid the organisation's innovation processes in many ways (e.g. generating new ideas based using such knowledge).
- The ability to find the specific documents associated with any topic in the system will increase the speed with which innovation workers can find applicable documentation.

This concludes the discussion around the calculation and application of topic networks and topic hierarchies. Lastly, the subsequent section will present some views regarding the validation of the CIRP topic modelling results.

11.9 Validating the Topic Model

George Edward Pelham Box is recognised as father of the following well-known statement: *"Remember that all models are wrong; the practical question is how wrong they have to be to not be useful."*

This quote certainly also apply to topic models. It is an almost impossible feat to validate all topics in terms of their relationships to documents and individual words, let alone the relationships of topics to individual authors, time periods and STCs. This case study aimed to investigate and report on different (potentially automatable) techniques envisaged by the author to aid in deriving more value from collections of electronic documents. The expanded topic model results were presented to a longstanding member of CIRP in order to get a feel for its level of accuracy. In the process some labels were refined to be more representative of the actual topics. The accuracy of the CIRP topic model was found to be satisfactory as no obvious flaws could be pointed out during this session.

The author subsequently devised a more objective way to determine the accuracy of the words characterising the individual topics. The majority of the CIRP technical papers analysed have, as part of their contents, keywords assigned by the authors and which were selected from a prescribed CIRP Unified Keyword List[83] that contains 411 keywords. Such keywords are meant to give one a (extremely) compact overview of the content of a given paper from the perspective of the authors. A similar process to the one described in section 11.7.1 was used to generate a "Keyword-Topic" matrix showing which author-specified keywords are associated with which topics and with which associated frequency. Conversely, the matrix also gives the topics associated with any of the author-specified keywords. The following matrix compares the 25 most probable characterising words of the 100 words characterising the topic "Robot Vision" (Topic 13) according to the topic model, to the set of 24 author-specified keywords associated with this topic as obtained using the "Keyword-Topic" matrix.

Words Characterising Topic	Topic-Word Probability	Author-Specified Keywords	Keyword Frequency
robot	0.022209735	calibration	2
image	0.012171437	robot	2
camera	0.010043754	object recognition	2
calibration	0.008898079	visual measurement	2
measuring	0.008134295	assembly	1
position	0.007534179	control	1
control	0.006934064	co-operative assembly	1
method	0.006443060	deterioration evaluation	1
accuracy	0.006224836	disassembly processes	1
point	0.006006612	disassembly tools	1
soldering	0.005515608	image analysis	1
solder	0.005406497	image processing	1
line	0.005079161	micromanipulator	1
measurement	0.005079161	online control	1
target	0.005024605	operation plan	1
frame	0.004697269	positioning accuracy	1
welding	0.004369933	product accompanying information systems	1
robots	0.004260821	reliability improvement	1
joint	0.003988041	robot calibration	1
view	0.003988041	robot safety system	1
points	0.003933485	uncertainty	1
ccd	0.003824373	visual control	1
sensor	0.003660706	welding	1
error	0.003387926	workpiece	1
visual	0.003333370		

Table 102: Comparison of words characterising the topic "Robot Vision" and the associated author-specified keywords

[83] The guidelines provided by the CIRP technical secretary states that authors have to select three keywords from the CIRP Unified Keyword List - one keyword to identify the general subject of the paper and two keywords to provide detail about particular aspects of the paper.

In Table 102, green cells represent cases where author-specified keywords and characterising words directly correspond, where yellow cells indicates partial matches between keywords and characterising words. Although not all of the author-specified keywords are found under the top 25 words characterising the topic as determined by the topic model, a fair number of matches were found indicating that the relevant topic to a large degree corresponds to a human interpretable concept.

The process of comparing words characterising topics to author-specified keywords associated with topics may be automated and topics having large discrepancies between the two word sets may be identified for manual inspection. Another level of validation may be to compare the author-specified keywords of a given paper to the most probable characterising words the topic model assigned to such a paper. Table 103 shows the three author-specified keywords for a paper “C07_Grzesik.pdf”, titled “An Investigation of the Thermal Effects in Orthogonal Cutting Associated with Multilayer Coatings”, published under STC-C in 2001. Table 103 further shows the ten most probable words of the 100 words characterising this document as determined by the topic model. The same colour coding applies here as did for Table 102.

Author-Specified Keywords	Most Probable Characterising Words
cutting	cutting
tool coating	tool
thermal analysis	coating
	contact
	bearing
	temperature
	surface
	chip
	thermal
	speed

Table 103: Comparison of words characterising the paper “C07_Grzesik.pdf” and the associated author-specified keywords

Once more, the words the topic model associated to this paper bear good resemblance to the author-specified keywords.

This concludes the main discussion around the CIRP case study. The essence of this case study will be presented in the following section.

11.10 Summary

The objectives of this case study, as stated at the beginning of this chapter, were the following.

This case study explored the following:

- **Automatically revealing which subjects were addressed in CIRP technical papers for the period 2001 to 2006 using a topic model**
- **Investigate the interpretability of the calculated topics and how to find an appropriate resolution in terms of number of topics**
- **Determine which documents are associated with which topics and vice versa**
- **Identify the documents that are significantly associated with more than one topic**
- **Establish the degree of similarity between topics to uncover topic affinities**
- **Establish how well each topic is represented in the input document set**
- **For each document, identify which other documents are relevant to the document in question**
- **Identify which words are strongly associated with a given word**
- **Identify which topics are associated with a given word**
- **Investigate ways of identifying the general and specific terms for the given field**
- **Investigate whether the topic model may be used as a starting point to construct a vocabulary for the CIRP community**
- **Identify which authors are associated with each topic and vice versa**
- **Identify which STCs are associated with each topic and vice versa**
- **Investigate ways to use the topic model results to estimate topic-time associations**
- **Establish if it is possible to construct a network or hierarchy of topics from the topic model resulting from different runs with varying levels of resolution (in terms of number of topics)**
- **Investigate ways of validating the topic model results**

In this case study CIRP was used as a model organisation and some documents generated by this organisation over a six year period were analysed to establish the possible benefits of using statistical topic models to analyse and organise the largely unstructured information of a given organisation. The different STCs comprising CIRP can be viewed as the departments of the organisation. A total of 140 topics were generated by three individual topic model analyses on some 757 documents. These topics generally exhibited high levels of human interpretability and presented a good overview of the subjects addressed by the organisation's research initiatives. It was further shown how the relationships between topics, documents and words, as embodied in the topic model, may be utilised to find specific information about CIRP-related topics, documents and terminology.

Ways of extending the topic model results by explicitly incorporating author (i.e. person), STC (i.e. organisational unit) and publication year (i.e. time) information, associated with individual documents, were further illustrated along with possible views that may be constructed using the calculated associations between topics, authors, STCs and publication years. Lastly, it was shown how the author-specified keywords may be used to validate the calculated topics as well as the words charactering individual documents in the topic model.

A note on the possible value of the developed mechanisms to CIRP:

The topic modelling mechanisms illustrated in this case study have several applications in CIRP. Firstly, a member of CIRP suggested that it may be used to identify technical papers that ‘significantly’ address the same two STCs (e.g. STC-Dn and STC-O). This way one or more common sessions can be organised at the annual General Assembly where such ‘overlapping’ papers are presented. This would lead to the exchange of knowledge between these STCs (i.e. the hypothetical departments of the organisation).

A second application envisaged by a CIRP member is using the proposed topic modelling mechanism to suggest suitable, previously published CIRP papers to the authors of new papers upon submittal of the draft versions. This would promote knowledge reuse, increase the quality of new papers and improve the citation index of the CIRP community.

A last suggested application is using the topic modelling mechanism to update the keyword lists of each STC as well as the CIRP dictionary in an objective way every second year. This would ensure that the vocabulary of the CIRP community remains up to date and that the appropriate keywords are available to the members of CIRP to describe their technical papers.

A note on the implication of the case study in terms of the proposed framework:

Although this case study focused more on the possible applications of statistical topic modelling techniques and possible extensions of the output of such techniques, its findings are of great importance to the proposed framework. It has shown that topic modelling can be used to distil and organise collections of highly technical, electronic document content – an important input to the eventual framework.

In terms of the framework entity types described in Chapter 8, the following entity types explicitly exist in CIRP in the opinion of the author: Concept, Document, Idea, Knowledge Area, Need, Objective, Opportunity, Organisational Unit, Person, Process/Activity, Product/Service, Skill/Competency, Technology/Method/Tool/Equipment. The proposed framework could therefore serve as a useful knowledge base for the CIRP community once implemented. New entity types may further be defined to cater for CIRP’s unique needs.

The next case study will investigate another topic modelling technique which allows the analyst to predefine the topics sought in the topic modelling analysis.

12. Case Study 3: South African Journal of Wildlife Research

The main purpose of this case study was to investigate the possibilities around using a statistical topic modelling technique, called the Concept-Topic Model[84], to classify a set of documents using a set of predefined concepts.

This case study explored the following:

- **Investigating whether the Concept-Topic Model statistical topic modelling technique can be used to classify documents using predefined concepts.**
- **Defining such concepts using appropriate characterising words.**
- **Evaluating the output generated by the model in terms of document-concept assignments.**
- **Calculating the relative coverage the individual concepts have received in publications over the analysis period.**
- **Calculating the time trends of the individual concepts over the analysis period.**

12.1 Electronic Documents Analysed

For this study, the articles published in the *South African Journal of Wildlife Research* (SAJWR) during the period 1970 to 2008 were selected for the following reasons:

- The author obtained access to the electronic versions of these articles as well as the consent of the journal to perform such an analysis.
- The author had good access to an ecologist being an expert in most of the knowledge areas covered by this journal. This would be essential for defining sensible concepts.
- The document set covered a substantial time period of 38 years.
- The fields of study included in the scope of this journal are favourable for the definition of sensible categories describing such fields.
- The editor of the relevant journal had a need to understand the relative coverage of different wildlife research related subjects in the past as part of an investigation into changing the name of the journal in the near future.

The main purpose of this analysis was to investigate the prevalence of the respective subjects associated with the specified concepts throughout the entire existence of the SAJWR journal in an objective manner. Although a LDA analysis was also performed on the document collection and interpretable topics were generated, LDA cannot report about the prevalence of specific topics. The Concept-Topic Model technique was selected to address this need. Prototype software implementing the Concept-Topic Model technique, developed by Indutech, was used for this analysis[85].

Table 104 presents some statistics about the document collection analysed in this case study.

[84] The Concept-Topic model is discussed in section 16.12 of this report.
[85] See Appendix C for the characteristics of the software used for this analysis.

Publication Years	Number of Files	File Format	Number of Files having Extraction Problems	Number of Duplicate Files	Document Collection Size	Total Number of Words	Number of Unique Words
1970 - 2008	878	Adobe PDF	1	0	449 MB	1 898 987	44011

Table 104: Characteristics of the documents analysed for the SAJWR analysis

All files representing the individual articles for the period 1970 to 2000 were scanned-in, OCRed versions of the original hard copies as no electronic copies were available for that period. Despite of this, surprisingly few extraction problems were encountered during the analysis of such files. Once obtained, all electronic articles were grouped in a folder based on the year of publication.

12.2 Formulating the Concepts

The Concept-Topic Model technique differs from LDA in that it requires the analyst to specify certain categories, called 'concepts' by the creators of this model, which serve as input to the analysis. These concepts are specified by creating the different labels for such concepts (e.g. "Cars") and for each concept specifying a number of characterising words associated with the concept in question (e.g. "engine", "steering", "wheels", "belt", "wipers", "windscreen", etc.). The Concept-Topic Model technique then uses the specified concepts to guide the construction of its associated generative model[86].

The concepts defined for this analysis, in collaboration with an expert in the field of wildlife management and ecology, are shown in Table 105.

[86] The prototype used for this analysis was able to accept either a non-hierarchical or hierarchical list of concepts. For the purpose of this study, the non-hierarchical option was used.

Concept Number	Concept Name	# Charac-terising Words
1	Reptiles	53
2	Amphibia	16
3	Fish	366
4	Birds	386
5	Fresh Water Mammals	9
6	Marine Mammals	16
7	Bats	10
8	Small Terrestrial Mammals	40
9	Large Carnivores	51
10	Large Herbivores	116
11	Mega Herbivores	12
12	Primates	15
13	Mollusks	43
14	Myriapoda	11
15	Arachnids	32
16	Parasitism	75
17	Insects	143
18	Wildlife Management Techniques	140
19	Evolutionary Biology (Anatomy, Evolution, Taxonomy & Genetics)	356
20	Population Dynamics	90
21	Physiology (Reproduction & Growth & Nutrition)	351
22	Behaviour (Reproductive & Nutrition)	208
23	Conservation, Utilisation, Sustainable Use & Biodiversity	129
24	Ecology	308
	Total number of characterising words defined	**2976**
	Total number of UNIQUE characterising words defined	**2761**

Table 105: Characteristics of the concepts defined for the SAJWR analysis

Of the 24 concepts defined, 16 concepts (i.e. Concepts 1 to 15 as well as Concept 17) represented specific taxonomic groups in the biological taxonomy *Animalia*. The remaining 8 concepts (i.e. Concepts 18 to 24 and Concept 16) dealt with more general wildlife research-related subjects. The number of characterising words specified varied dramatically and was influenced by the ease of identifying words that describe the given concept and preferably not any of the other concepts included in the scope of the study. Only unigrams were used to specify concepts as the software used for this analysis could not accommodate bigrams, trigrams, etc. at the time of the study. This of course increased the difficulty of identifying unambiguous characterising words for the 24 concepts.

The following process was followed to define the concepts in question:

- For each concept, use appropriate text books and online encyclopaedias to identify words (singular words where such words are nouns) charactering the concept.
- Where possible look the label of the concept (e.g. “Amphibia”) up in the *WordNet* electronic dictionary to find words that are related to the concept (e.g. frogs, toads, newts, salamanders, caecilians are listed as related words to the word “amphibia” in this electronic dictionary).

- For each characterising word identified, also include the plural form as well as possible derivations (e.g. “sloughing” is a derivation of the word “slough”) of the word where sensible and applicable.
- Use a spell checker to ensure the correct spelling of all words specified. For words not included in the spell checkers of the standard office document editors, use *WordNet* to ascertain that the specified words exist and are spelled correctly.

Table 106 provides an example of a concept and its characterising words.

Reptiles			
adder	crocodile	python	snakebite
adders	crocodiles	pythons	snakebites
agama	elapidae	reptile	snakes
agamas	gecko	reptiles	terrapin
alligator	geckos	reptilian	terrapins
alligators	hydropiidae	reptilians	tortoise
boomslang	iguana	rinkhals	tortoises
boomslange	iguanas	serpent	turtle
chameleon	lizard	serpents	turtles
chameleons	lizards	skink	venom
chelonia	mamba	skinks	viperidae
cobra	mambas	slough	
cobras	monitor	sloughing	
colubridae	monitors	snake	

Table 106: Words specified for concept “Reptiles”

Although it is a tedious process to accurately define such concepts, once defined they can be reused for future analyses. Information that may assist in the definition and the refinement of such concepts are increasingly becoming available on the Internet in the form of shared word lists, taxonomies, ontologies, etc.

12.3 Analysis Process

Once all 24 concepts were specified they were inputted into the format required by the prototype software used for the analysis. Except for the concepts, the software requires the following inputs to an analysis:

- The words that have to be excluded from the analysis. A custom-developed English stoplist with 821 entries was provided to eliminate the words with little semantic value.[87]
- The minimum number of times a word has to appear in the combination of all documents analysed before it is included in the analysis. A minimum frequency value of 5 was specified for this analysis.
- The path to the documents to be analysed.

After all the necessary inputs were specified, the analysis was initiated.

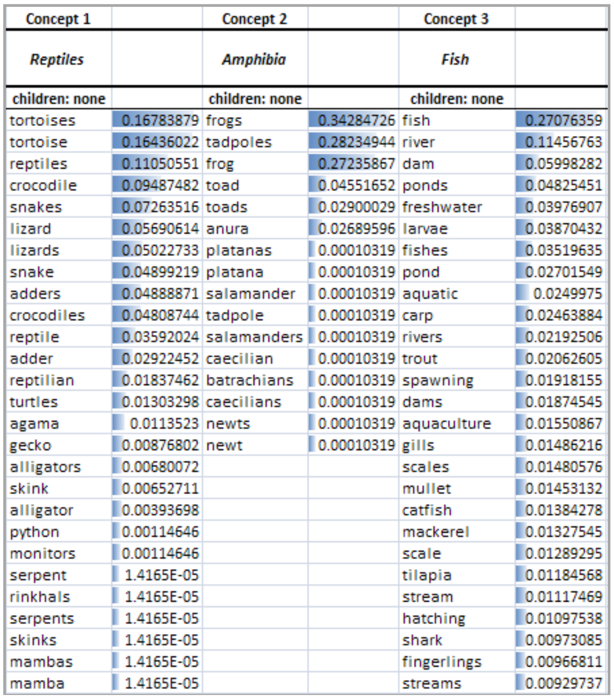
[87] This is the same stoplist that was used in the CIRP case study.

12.4 Evaluating and Refining the Analysis Results

A 2GHz, 8-core CPU computer was used for the analysis which took 12:06:53 to complete[88]. The software exports the results to a *Microsoft Excel* file containing the following information:

- A spreadsheet containing the concepts, along with their respective charactering words as specified by the analyst. For each characterising word a probability is given indicating the confidence the model has calculated, based on the contents of the specific documents analysed, that the given word forms part of the given concept. For each of the concepts the associated charactering words are arranged in order of descending probability.
- A second spreadsheet containing the calculated document-concept associations. This essentially corresponds to the document-topic matrix generated by the LDA technique that was discussed in sections 6.5.6 and 10.2. As with LDA, a document may have one or many associated concepts and a concept may have one or many associated documents. The individual assignment weights of a given document to the various concepts included in the analysis are called the document mixture ratios.

Figure 52 shows an extract of the first spreadsheet generated by the software, giving the probabilities calculated for the respective words charactering the concepts included in the analysis.

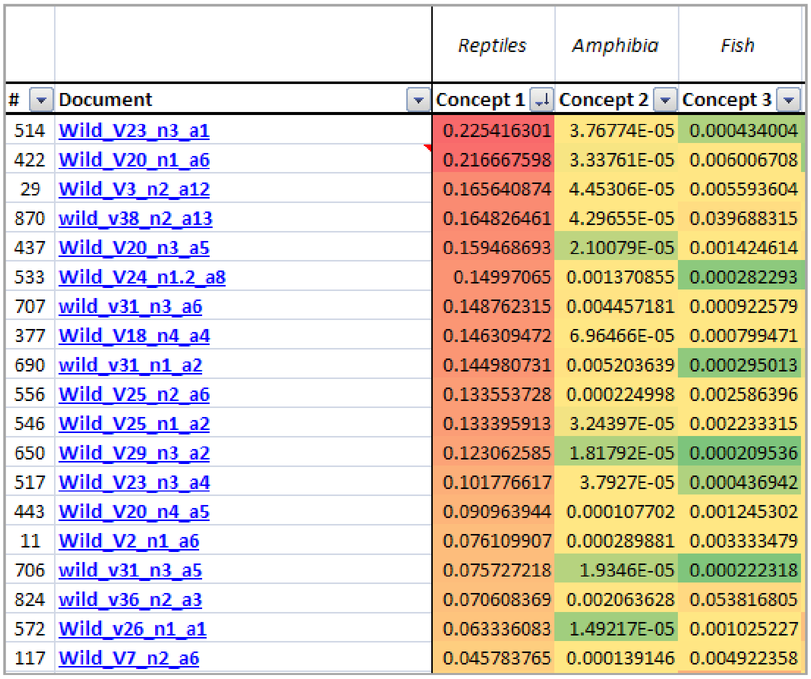


Concept 1		Concept 2		Concept 3	
Reptiles		*Amphibia*		*Fish*	
children: none		children: none		children: none	
tortoises	0.16783879	frogs	0.34284726	fish	0.27076359
tortoise	0.16436022	tadpoles	0.28234944	river	0.11456763
reptiles	0.11050551	frog	0.27235867	dam	0.05998282
crocodile	0.09487482	toad	0.04551652	ponds	0.04825451
snakes	0.07263516	toads	0.02900029	freshwater	0.03976907
lizard	0.05690614	anura	0.02689596	larvae	0.03870432
lizards	0.05022733	platanas	0.00010319	fishes	0.03519635
snake	0.04899219	platana	0.00010319	pond	0.02701549
adders	0.04888871	salamander	0.00010319	aquatic	0.0249975
crocodiles	0.04808744	tadpole	0.00010319	carp	0.02463884
reptile	0.03592024	salamanders	0.00010319	rivers	0.02192506
adder	0.02922452	caecilian	0.00010319	trout	0.02062605
reptilian	0.01837462	batrachians	0.00010319	spawning	0.01918155
turtles	0.01303298	caecilians	0.00010319	dams	0.01874545
agama	0.0113523	newts	0.00010319	aquaculture	0.01550867
gecko	0.00876802	newt	0.00010319	gills	0.01486216
alligators	0.00680072			scales	0.01480576
skink	0.00652711			mullet	0.01453132
alligator	0.00393698			catfish	0.01384278
python	0.00114646			mackerel	0.01327545
monitors	0.00114646			scale	0.01289295
serpent	1.4165E-05			tilapia	0.01184568
rinkhals	1.4165E-05			stream	0.01117469
serpents	1.4165E-05			hatching	0.01097538
skinks	1.4165E-05			shark	0.00973085
mambas	1.4165E-05			fingerlings	0.00966811
mamba	1.4165E-05			streams	0.00929737

Figure 52: Extract from the Concept-Word matrix

Figure 53 shows an extract from the second spreadsheet that contains the calculated document-concept assignments. In the figure, the documents having the strongest associations with the "Reptiles" concept are shown in order of decreasing strength of association.

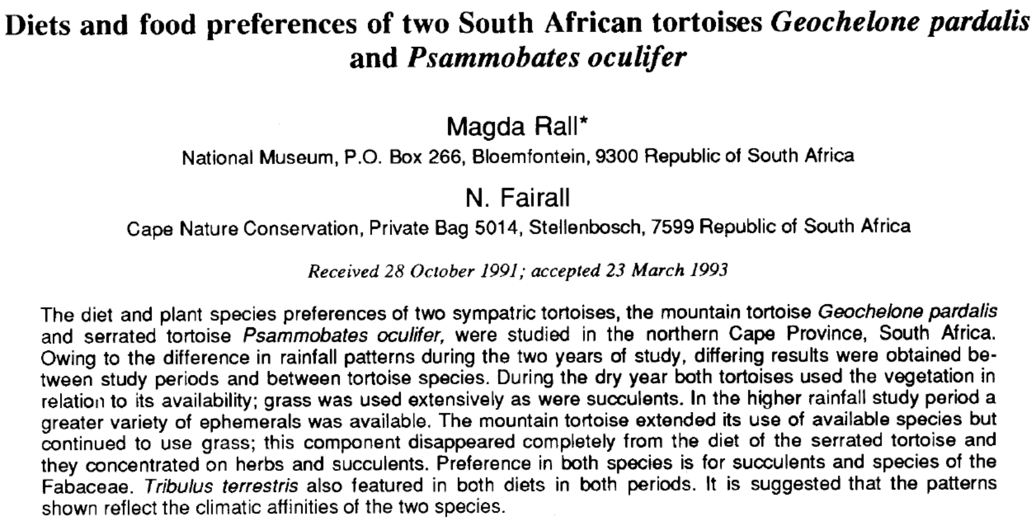
[88] This prototype could only utilise a single CPU core at a time during the inference phase which contributed to the long processing time.



		Reptiles	Amphibia	Fish
#	Document	Concept 1	Concept 2	Concept 3
514	Wild_V23_n3_a1	0.225416301	3.76774E-05	0.000434004
422	Wild_V20_n1_a6	0.216667598	3.33761E-05	0.006006708
29	Wild_V3_n2_a12	0.165640874	4.45306E-05	0.005593604
870	wild_v38_n2_a13	0.164826461	4.29655E-05	0.039688315
437	Wild_V20_n3_a5	0.159468693	2.10079E-05	0.001424614
533	Wild_V24_n1.2_a8	0.14997065	0.001370855	0.000282293
707	wild_v31_n3_a6	0.148762315	0.004457181	0.000922579
377	Wild_V18_n4_a4	0.146309472	6.96466E-05	0.000799471
690	wild_v31_n1_a2	0.144980731	0.005203639	0.000295013
556	Wild_V25_n2_a6	0.133553728	0.000224998	0.002586396
546	Wild_V25_n1_a2	0.133395913	3.24397E-05	0.002233315
650	Wild_V29_n3_a2	0.123062585	1.81792E-05	0.000209536
517	Wild_V23_n3_a4	0.101776617	3.7927E-05	0.000436942
443	Wild_V20_n4_a5	0.090963944	0.000107702	0.001245302
11	Wild_V2_n1_a6	0.076109907	0.000289881	0.003333479
706	wild_v31_n3_a5	0.075727218	1.9346E-05	0.000222318
824	wild_v36_n2_a3	0.070608369	0.002063628	0.053816805
572	Wild_v26_n1_a1	0.063336083	1.49217E-05	0.001025227
117	Wild_V7_n2_a6	0.045783765	0.000139146	0.004922358

Figure 53: Extract from the Document-Concept matrix

The document filenames contained in this spreadsheet are actually hyperlinks which allow the analyst to open any of these documents directly from the spreadsheet by simply clicking on the relevant document filename. Figure 54 shows the title, author information and abstract of the article having the strongest association with the "Reptiles" concept.



Diets and food preferences of two South African tortoises *Geochelone pardalis* and *Psammobates oculifer*

Magda Rall*
National Museum, P.O. Box 266, Bloemfontein, 9300 Republic of South Africa

N. Fairall
Cape Nature Conservation, Private Bag 5014, Stellenbosch, 7599 Republic of South Africa

Received 28 October 1991; accepted 23 March 1993

The diet and plant species preferences of two sympatric tortoises, the mountain tortoise *Geochelone pardalis* and serrated tortoise *Psammobates oculifer*, were studied in the northern Cape Province, South Africa. Owing to the difference in rainfall patterns during the two years of study, differing results were obtained between study periods and between tortoise species. During the dry year both tortoises used the vegetation in relation to its availability; grass was used extensively as were succulents. In the higher rainfall study period a greater variety of ephemerals was available. The mountain tortoise extended its use of available species but continued to use grass; this component disappeared completely from the diet of the serrated tortoise and they concentrated on herbs and succulents. Preference in both species is for succulents and species of the Fabaceae. *Tribulus terrestris* also featured in both diets in both periods. It is suggested that the patterns shown reflect the climatic affinities of the two species.

Figure 54: Example of an article having a strong association with the "Reptiles" concept

For each concept, a number of the associated documents were opened to evaluate the accuracy of the assignment of documents to concepts. A few cases were detected where documents were wrongly assigned to concepts. For these cases, the characterising words of the associated concepts were investigated in an attempt to identify the words responsible for the wrong assignment. Once found, the concept was updated to be less ambiguous. For example, the word “capensis” was removed from the “Fresh Water Mammals” concept because although it forms part of the biological name of a certain type of otter, it also forms part of the biological names of other animals not being fresh water mammals. Also, the words “herd” and “herds” was removed from the concept “Large Herbivores” since the term “herd” is also used to refer to a collection of seals which are marine mammals and not large herbivores. Updates were made to five of the 24 concepts and the analysis was subsequently initiated once more.

The document-concept matrix also contains a computer-generated concept, called “Root”, to which the portions of documents are assigned that cannot be explained by the concepts provided by the analyst. Finding the documents that have strong calculated associations with this concept corresponds to finding those documents that have weak associations with the concepts defined by the analyst. After more closely investigating a sample of such documents, it was found that mostly documents primarily dealing with vegetation, plants, trees and soil were grouped under this concept since the defined concepts did not include these subjects or their characterising words specifically. This to some extent confirms the accuracy of the classification.

12.5 Extending the Analysis Results

The objectives of this case study included the estimation of the relative coverage the individual concepts have received during the analysis period. Figure 55 shows the Concept Coverage Graph that the author constructed by calculating the sum of the mixtures ratios per concept and subsequently normalising such sums analogous to the process explained in section 6.5.6. These sums were then used as input to *Microsoft Excel*’s pie chart functionality to create the graph.

This graph shows that approximately 48% of the content of the observed documents can be explained (or strictly speaking, generated) by the concepts defined. For the defined concepts, the concept “Ecology” has the received the most attention in the articles published during the analysis period of 38 years, while the concept “Myriapoda” received the least. The relative coverage of the different concepts can also be read from this graph. For example, it can be seen that about twice as much was published about the concept “Large Herbivores” than about the concept “Large Carnivores”.

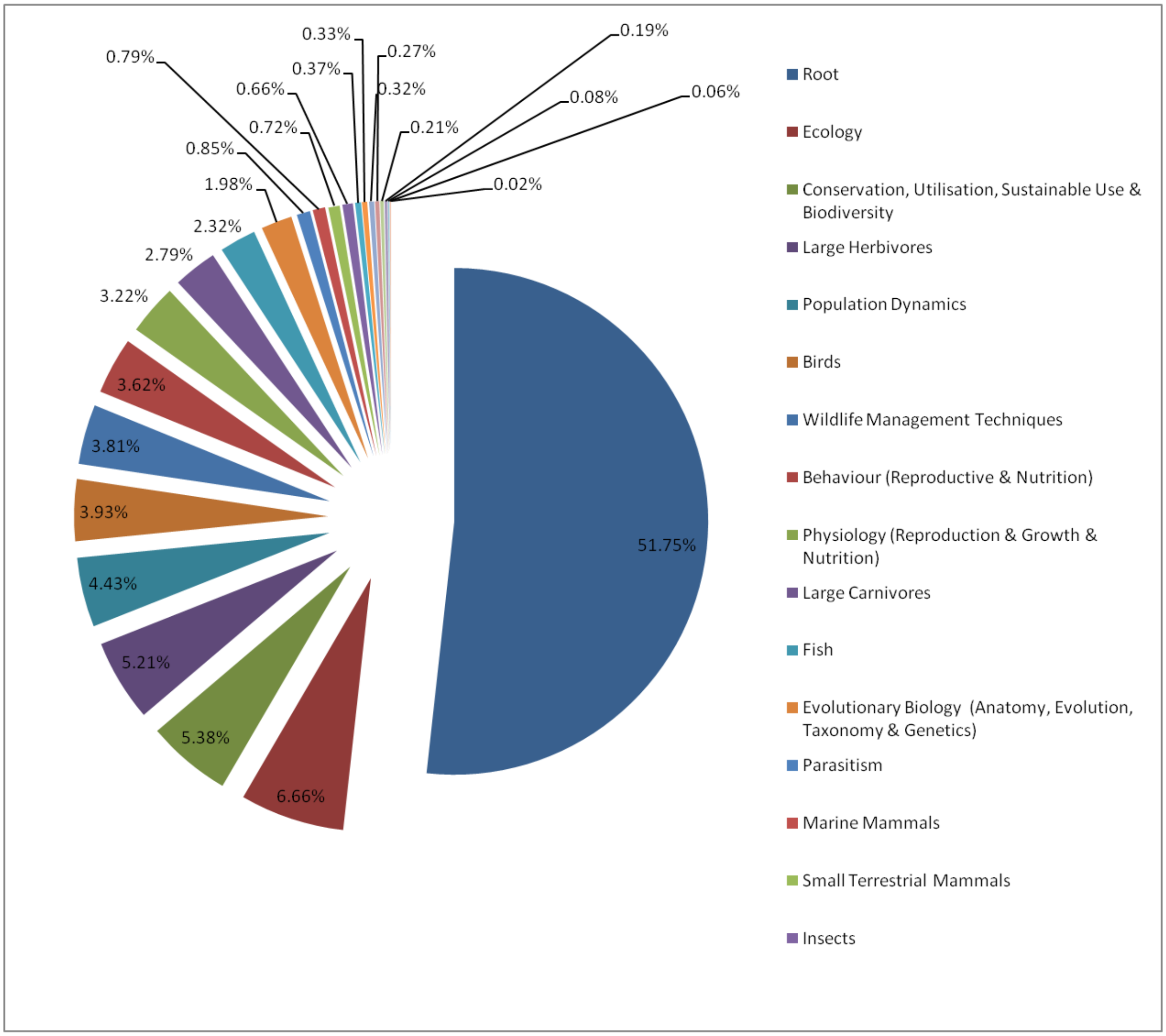


Figure 55: Graph indicating the relative coverage of the different concepts

Next, it was attempted to estimate the time trends of the individual concepts during the analysis period. Essentially the same process explained in section 11.7.3 was followed to achieve this with the exception that the threshold mixture ratio value used to determine whether a given document is adequately addressing a given concept (using the document mixture ratios in the concept-document matrix) was empirically determined to be 0.03. This means that documents having mixture ratios of 0.03 or greater for a given concept are considered to adequately address that concept. After calculating the mean publication year, standard deviation and minimum and maximum publication year values of the documents assigned to concepts in this way, the following graph (Figure 56) was constructed by using *Microsoft Excel*'s *Stock Chart* functionality as in the previous case study.

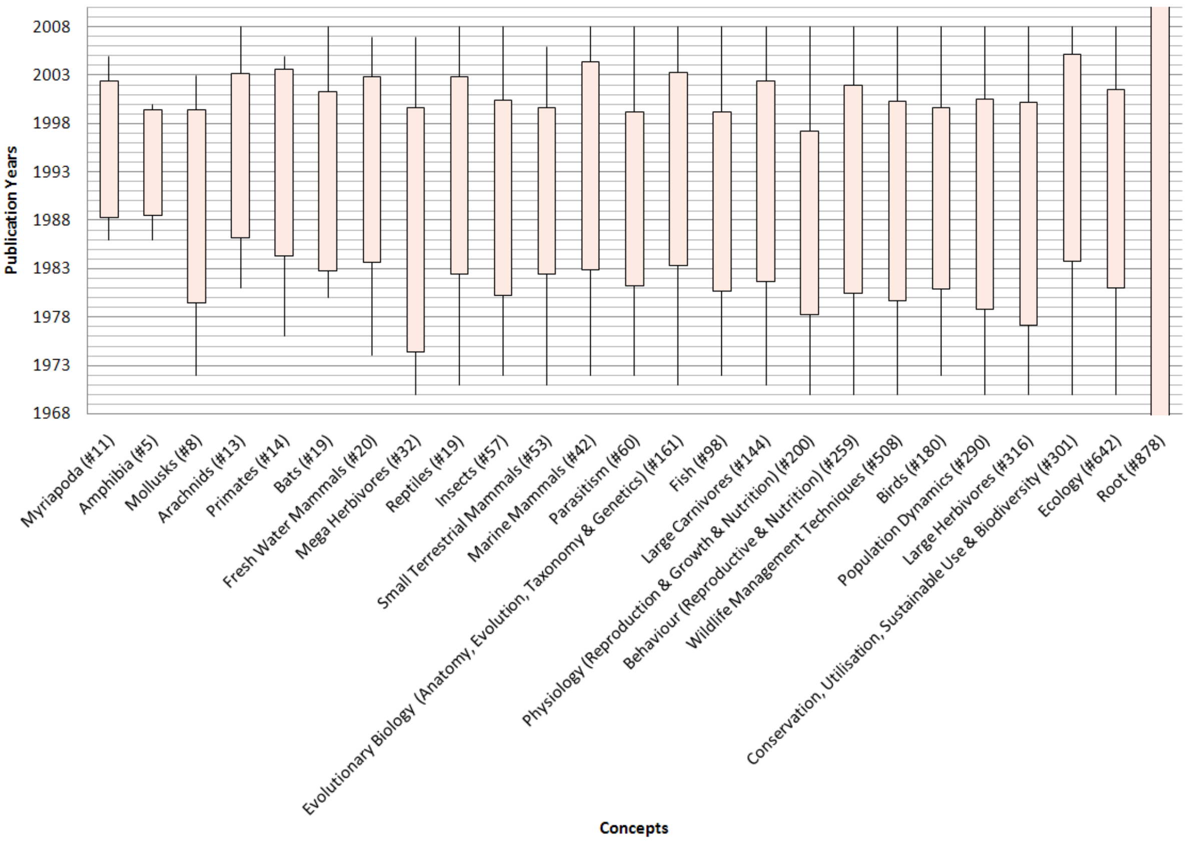


Figure 56: Time trends of the concepts defined for the SAJWR analysis

In Figure 56 the leftmost concept is the concept that received the least coverage (in terms of the number of associated publications) over the analysis period, while the concepts to the right are the concepts that received increasingly better coverage in the corresponding period. The numbers in parenthesis succeeding the individual concept labels indicate the number of papers that have document mixture ratios of 0.03 or more for the given concept.

This concludes the discussion about the third and last case study presented in this report. The next section summarises the case study.

12.6 Summary

Firstly, the objectives of this case study are restated.

This case study explored the following:

- **Investigating whether the Concept-Topic Model statistical topic modelling technique can be used to classify documents using predefined concepts.**
- **Defining such concepts using appropriate characterising words.**
- **Evaluating the output generated by the model in terms of document-concept assignments.**
- **Calculating the relative coverage the individual concepts have received in publications over the analysis period.**
- **Calculating the time trends of the individual concepts over the analysis period.**

The statistical topic modelling technique investigated in this case study, the Concept-Topic Model, proved to be a very useful way to classify documents using a number of predefined categories (concepts). What makes this technique especially useful is the fact that it allows for partial membership of a given document to a number of categories (concepts). This is a realistic way of classifying documents as most documents address multiple categories (concepts) in real life. Also, an additional category is automatically created by the technique to which all portions of documents that cannot be explained by the defined categories (concepts) are assigned to. By investigating the contents of the documents strongly assigned to this category the analyst can identify candidates for new concepts for future analyses (e.g. the need for the categories “Soil”, “Trees” and “Vegetation” was detected in this way in the present case study).

It was further found that it is indeed possible to define such concepts by identifying a number of words that characterise the individual concepts. The ability to use bigrams to define concepts will greatly improve the ease of defining such concepts and should be listed as a high priority improvement to the prototype software used.

The 24 concepts defined in this study rendered satisfactory results and refinements to the concepts were made, after investigating the results of the first analysis, to further improve the accuracy of the results. The results were further extended by estimating the relative coverage the individual 24 concepts have received in publications during the 38-year period analysed. As a second extension to the results, the 24 concepts were positioned in time by calculating and visualising the time trends of the individual concepts.

A note on the implication of the case study in terms of the proposed framework:

The Concept-Topic Model technique has great significance in terms of the framework defined earlier to organise and make accessible the innovation-related information of the target organisation. Firstly, this technique can be used to automatically identify information pertaining to specific entities that are of interest to a given person, team, or other organisational unit (e.g. entities corresponding to the competitors of the target organisation or entities corresponding to new technologies). Some thoughts in this regard were put forward in section 8.11. To recapitulate, this technique will play an important role in linking the actual instances of the different entity types of the framework to the content of electronic documents. This can be done by using such entity instances, including their characterising attributes, as the predefined concepts this technique requires as input. Once a Concept-Topic Model has been generated for a given document collection in this way, the sought-after relationships between the various instances of the different entity types of the framework and the different documents analysed will be quantified and would be available for incorporation into the framework.

Lastly, the Concept-Topic Model technique (as a classification technique) complements the LDA technique (as a clustering technique) to form a bouquet of text analytical techniques to aid the process of organising and making accessible unstructured information from two ends; from what the organisation already knows and has made explicit (classifying information using existing concepts) as well as from the unfamiliar and informal end (using clustering to find new topics or to aid in making explicit informal concepts).

This concludes the discussion of the three case studies included in this research report. The following section presents the conclusion to this research as well as some recommendations as to the possible application of the work presented here.

13. Conclusion and Recommendations

Today, electronic text is a convenient and in many cases the preferred way to capture and share information about almost any subject. Electronic text is generally regarded as unstructured data or information[89] due to its free-form nature as opposed to the very structured data contained in relational databases or XML. Unstructured information, in the form of natural language text, is abundant in various kinds of organisations and it expands daily. Examples include web pages (including wikis, blogs, etc.), word processor files, e-mails, e-books, instant messaging messages, the text fields of issue tracking system databases, and customer relationship management systems.

The effort associated with skimming through, reading and understanding large collections of unstructured textual information remains a challenge in spite of many technological advances in the field of information- and communication technology. The time available for collecting, reading, interpreting and acting upon appropriate textual information is limited in both corporate and research environments. In order for an organisation to continuously excel in innovation at the fast pace dictated by its competitors and customers, many elements should be in place including appropriate strategies, capable staff, sufficient resources, enabling technologies and accessible, current and accurate information. A combination of applicable structured and unstructured information, residing within and outside the boundaries of the organisation, is required to feed the target organisation's innovation processes. Three requirements arise in this regard.

1. Being able to (for a large part) automate the process of filtering, abstracting and organising large quantities of unstructured, textual information on an ongoing basis in order to arrive at distilled information that can potentially feed the organisation's innovation processes.
2. Being able to structure such (distilled) unstructured, textual information and identifying relationships between such information items and the organisation's core innovation-related information entities.
3. Being able to unify an organisation's structured and unstructured innovation-related information and subsequently making it accessible to the organisation's innovation workers in support of their daily activities.

This research has investigated the innovation landscape in terms of various types of innovation, different innovation models, different views on roles in the innovation process as well as various generic information 'realms' that may be applicable to innovation[90]. The opportunity, as identified in literature, to improve the efficiency, but especially the effectiveness of the innovation process was further expressed. Specific sources of innovation-related information, both internal and

[89] The reader is referred to section 4.1 for a detailed explanation about the differences between structured and unstructured information.

[90] The reader can read more about this topic in Chapter 3.

external to the organisation, were identified and its relationships to the generic innovation process were discussed[91]. Subsequently, the landscape of text analytical approaches was explored to get a grip on possible ways of analysing textual information to arrive at distilled information required to feed the innovation processes of the target organisation[92]. A specific set of text analytical techniques, namely statistical topic models, was subsequently identified and thoroughly researched[93]. Two specific topic modelling techniques, namely Latent Dirichlet Allocation (LDA) and the Concept-Topic Model, were selected for their potential usefulness and complementary nature. The concept of business metadata, dealing with the contextualisation of organisational data and information to facilitate information usage by non-technical employees, was further researched and reported on[94]. Lastly, a framework was developed to address the second and third requirement identified earlier in this section[95].

This framework – constituting an initial design for an innovation-focused knowledge base for the target organisation - is intended to organise, contextualise and make accessible innovation-related information to the innovation workers of the target organisation. The framework consists of 23 generic information entity types corresponding to the fundamental elements that the target organisation has to gather, capture and disseminate information about in the light of its innovation activities. The different possible information systems that would typically be involved in the implementation of the framework in the target organisation were further discussed as well as some high-level use cases of the framework. The working methods associated with the framework and the potential role of the two identified statistical topic modelling techniques in populating and updating the information in the framework were also discussed. Lastly, some significant functionalities of the text analytical system to be used as part of the implementation of the framework were presented.

Three case studies were further presented where the identified statistical topic modelling techniques were applied to three fundamentally different document collections in an attempt to address the first and second requirements mentioned earlier in this section. On a high level, these case studies investigated the potential support that text analytical techniques may offer in automatically distilling large quantities of textual information. Even thought the proposed framework was not implemented and evaluated in its entirety, the case studies investigated important aspects of the framework and its envisaged interaction with the text analytical techniques identified for application in this research. At the end of each such case study, the implications of the specific case study to the proposed framework are discussed.

[91] The reader is referred to Chapter 4 for a discussion about this subject.
[92] More details about this in Chapter 5.
[93] See Chapter 6 for a discourse on this topic.
[94] The reader is referred to Chapter 7 for more information on this topic.
[95] The details about this framework are presented in Chapter 8.

Two types of text analytical approaches were investigated in the case studies, namely clustering and classification[96]. LDA, essentially a sophisticated type of clustering technique, was investigated in two of the case studies. It was shown that it is capable of automatically generating human-interpretable topics that represent the different subjects underlying the content of electronic documents analysed. It was further shown how such topics can be used to construct information interest profiles for different individuals as well as how such profiles may be used to identify areas for possible tacit knowledge exchange (by means of face-to-face conversation) between such individuals or to identify documents for possible exchange. It was further illustrated how the relationships between topics, documents and words, embedded in the generated topic model, may be used to more effectively explore unstructured textual information. Ways of extending the topic model results, by adding author, time and organisational unit aspects, were further suggested and illustrated. Moreover, ways of automatically constructing topic hierarchies or topic networks by linking the results of different topic modelling analyses were presented. The major benefit of using LDA on innovation-related information is to detect new topics that may possibly be relevant to the organisation's innovation workers.

The Concept-Topic Model technique, the second text analytical technique explored as part of the case studies, is essentially a kind of advanced classification technique and was investigated as a possible means of automatically identifying documents dealing with a number of predefined concepts[97]. The main benefit of using this technique to analyse an organisation's innovation-related information is to find information about specific, existing innovation entities (e.g. a certain new technology) that form part of the implemented framework.

It can be concluded that text analytical techniques, and statistical topic modelling techniques in particular, may be extremely useful in automatically distilling large quantities of textual information to be more palatable to innovation workers. Moreover, when using such techniques in parallel with the framework proposed in this research, great potential exists to provide innovation workers with the ability to browse the innovation-related information in a structured way, compose intelligent search queries with all information shown in context of the organisation's innovation-related information entities. The framework is not limited to improving the accessibility of unstructured information. It should be used to make accessible structured information contained in disparate databases of the organisation's information systems bringing together the traditionally separated universes of structured and unstructured information. Association and context is driving innovation. The framework, once implemented, populated and made accessible, can help to improve the status quo of gathering information in many disparate locations only for it

[96] The reader is referred to section 5.11 for a discussion of the difference between clustering and classification.

[97] The reader is referred to Case Study 3 for more information in this regard.

to collect dust and clutter the digital storage media of the organisation. It can help to unlock the true value of one of the innovating organisation's key resources – the organisation's generated and meticulously gathered information – by putting it into the hands of its innovation workers in a structure that contextualises such information in terms of the fundamental subjects involved in the organisation's innovation endeavours. The application of such a framework is not limited to a single organisation. It may be very useful in structuring and making accessible innovation-related information and terminology in the scenario where a number of organisations are engaged in collaborative innovation.

Finally, the following **recommendations** should be noted around innovation, its link to information and any information-driven support to the innovation processes of an organisation:

- Innovating organisations need to constantly gather information about the key elements affecting their innovation endeavours to be able to make informed decisions about their innovation-related matters.
- Innovating organisations need to capture key information, especially innovation-related information, in a structure that facilitates question-answering by linking it to several information entities (e.g. the various entity types of the framework). This would allow more powerful search capability (reactive) as well as possibilities for automatic distribution of information (proactive).
- The framework and its associated processes has to be fine-tuned to support the various roles of individuals in the target organisation (e.g. the various innovation roles), its benefits should be made clear to such individuals, and it should further be made part of day-to-day innovation-related tasks of innovation workers to ensure its successful adoption and continued utilisation.
- The framework should be implemented in such a way that it gives its users the opportunity to provide inputs regarding the usefulness of different information entities embodied in the framework. Such contributions should be made available to other innovation workers to harness the principle of the 'wisdom of the crowd' and enable 'collective intelligence' with regard to innovation in the target organisation.
- Lastly, without the commitment, support and active participation of the organisation's management the implementation of the framework, or any other system to support the organisation's innovation endeavours, will ultimately be unsuccessful.

On a more philosophical note, this now concludes a part of the journey to the development of the framework and operational protocol for better managing innovation-related information. Others may continue the journey from this point onward in search of the potential benefits envisaged with the eventual deployment and operation of such a framework. As the nature of academic research demands, others may choose to retrace some steps along the road I followed to arrive here, exploring alternate routes of their own making to arrive here and maybe beyond, always advancing the point that we call *here*. But ultimately the words of *Stephen King* capture the truth: "*Only God gets it right the first time*".

The following section will shed some light on possible areas for future research in the light of this research.

14. Future Research

This research merely chiselled away the top layer of sediment in the hypothetical archaeological site of information-related support to innovation to uncover but the first signs of true deeper-lying value. In the opinion of the author the following avenues exist for future research in relation to the research presented here:

- Investigate and experiment with suitable software platforms and technologies for implementing the envisaged framework and the associated processes (i.e. building the first working prototypes).
- Design and apply a suitable and extendible data model for embodying the framework.
- Implement, operate, refine and extend the framework and its associated entity types and processes. This should be done in a variety of organisations as well as in different industries (i.e. building scalability and industry specific experience).
- Investigate mechanisms for automatically and continuously acquiring new external information related to the content of the framework to broaden the target organisation's information base.
- Explore the full potential of the framework in a business context:
    - Developing detailed use cases for the framework and its associated processes
    - Evaluating the usefulness of the framework in different industries
    - Investigating the relationship between the organisation's innovation maturity and the usefulness of the framework
    - Determining the influence of organisation size on the usefulness of the framework
    - Evaluate the value of the framework in a multi-organisation innovation setting
- Investigate the possibility of extending the framework to cater for the extended information needs of the target organisation.
- Investigate, apply and compare the usefulness of more text analytical techniques and systems with regard to the automated analysis of textual information.
- Elaborate on possible ways of updating the information content of the framework as well as ways of inferring new information based on the information currently in the framework.
- Investigate how to successfully motivate people to make effective contributions to the knowledge base embodied by the implemented framework.

With regards to statistical topic modelling techniques and specifically the LDA and Concept-Topic Model techniques, the following research topics are foreseen:

- Determine the difference in the results when analysing document abstracts versus analysing entire documents.
- Determine the difference in the results when analysing chapters of documents individually versus analysing such documents in their entirety.
- Determine a mechanism or heuristic to determine the optimal number of topics for a LDA analysis for a given document collection.
- Establish the efficiency and quality of the outputs of non-parametric topic modelling techniques when compared to LDA.
- Determine runtime equations for LDA and the Concept-Topic Model that would predict the duration of the analysis based on a number of parameters that can be measured beforehand (e.g. number of documents, number of words in corpus, number of topics or concepts, etc.)

- Further investigate ways of improving the visualisation of topic model results so that the user can explore a document collection in an interactive, topic-guided fashion as well as to support the user in obtaining information about specific questions he might have.
- Investigate the implementation, accuracy and usefulness of the document fold-in technique with LDA topic models with the goal to add new documents to existing topic models. This technique will reduce the frequency of creating topic models from new when minor changes occur to the document collections in question.
- Investigate ways of further speeding up the generation of topics models by means of distributed processing and possibly using graphical processing units (GPUs) of computers to do the required calculations.

By addressing these issues in future research and structured investigations, the research presented in this study would be taken to a next level of maturity bringing it ever closer to ultimate adoption in industry.

References

1. Adelman, S., O'Neil, B. (2007), *Capturing Intellectual Capital in Metadata*, DM Review, http://www.dmreview.com/editorial/dmreview/print_action.cfm?articleId=1075079
2. Alonso, O., Banerjee, S., Drake, M. (2006), *GIO: A semantic web application using the information grid framework*, In Proc. WWW, pp. 857-858. ACM Press.
3. Bergman, M.K. (2005), *Untapped Assets: The $3 Trillion Value of U.S. Enterprise Documents*, BrightPlanet Corporation.
4. Blei, D., Griffiths, T., Jordan, M., Tenenbaum, J. (2004), *Hierarchical Topic Models and the Nested Chinese Restaurant Process*, *NIPS*, 2004.
5. Blei, D., Lafferty, J. (2007), *A Correlated Topic Model of Science*, The Annals of Applied Statistics 2007, vol. 1, no. 1, pp. 17–35.
6. Blei, D., Lafferty, J. (2006), *Correlated Topic Models*, in Advances in neural information processing systems, vol. 18.
7. Blei, D., Lafferty, J. (2006), *Dynamic Topic Models*, in Proceedings of the 23rd International Conference on Machine Learning.
8. Blei, D., Lafferty, J. (2006), *Modeling Science*.
9. Blei, D., Lafferty, J. (2009), Topic *Models,* In A. Srivastava and M. Sahami, editors, Text Mining: Theory and Applications. Taylor and Francis, in press.
10. Blei, D., Ng, A., Jordan, M. (2003), *Latent Dirichlet Allocation*, Journal of Machine Learning Research, vol. 3, pp. 993 – 1022.
11. Borchers, A., Herlocker, J., Konstan J., Riedl, J. (1998), *Ganging up on Information Overload*, IEEE Computer Society Press, Computer, vol. 31, no. 4, pp. 106-10.
12. Bullinger, A.C. (2008), *Innovation and Ontologies: Structuring the Early Stages of Innovation Management*, Gabler, Germany.
13. Bullinger, J., Auernhammer, K., Gomeringer, A. (2004), *Managing innovation networks in the knowledge-driven economy*, International Journal of Production Research, Vol. 42, No. 17, pp. 3337 – 3353.
14. Carley, K.M. (1999), *Learning within and among organizations,* Advances in Strategic Management, Vol. 16, pp. 33-53.
15. Carlson, C.N. (2003), *Information overload, retrieval strategies and Internet user empowerment*, Proceedings of The Good, the Bad and the Irrelevant (COST 269), University of Art and Design, Helsinki, Finland, pp. 169-173.
16. Casey, R.M. (2005), *Bioinformatics Tools for Gene Sequence Analysis,* Business Intelligence Network.
17. Cheeseman, P., Stutz, J. (1996), *Bayesian classification (AutoClass): Theory and results*, In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., editors, Advances in Knowledge Discovery and Data Mining, pp. 153–180. MIT Press, Cambridge, MA.
18. Chemudugunta, C., Holloway, A., Smyth, P., Steyvers, M. (2008), *Modeling documents by combining semantic concepts with unsupervised statistical learning*, In International Semantic Web Conference 2008 (ISWC 2008), pp 229-244.
19. Chemudugunta, C., Smyth, P., Steyvers, M. (2008), *Combining Concept Hierarchies and Statistical Topic Models* (CIKM'08), Napa Valley, California, USA.
20. Chen, A., McLeod, D. (2005), *Collaborative Filtering for Information Recommendation Systems*, Encyclopaedia of Data Warehousing and Mining, Idea Group.
21. Chesbrough, H. (2003), *Open Innovation: The New Imperative for Creating and Profiting from Technology*, Harvard Business School Press.
22. Cheung, C., Lee, W., Wang, Y. (2005), "*A multi-facet taxonomy system with applications in unstructured knowledge management*, Journal of Knowledge Management, vol. 9, issue 5, pp. 76 – 91.

23. Cooper, R.G. (1990), *Stage-Gate systems: a new tool for managing new products - conceptual and operational model*, Business Horizons, May-June: pp 44-53.
24. Courseault, C.R. (2004), *A Text Mining Framework Linking Technical Intelligence from Publication Databases to Strategic Technology Decisions*, Georgia Institute of Technology.
25. Davis, M. (2006), *The Semantic Wave 2006, Part 1: Executive Guide to Billion Dollar Markets*, Project10X.
26. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R. (1990), *Indexing by latent semantic analysis*, Journal of the American Society for Information Science, vol. 41, issue 6, pp. 391-407.
27. Dempster, A.P., Laird, N.M., Rubin, D.B. (1977), *Maximum Likelihood from Incomplete Data via the EM Algorithm*, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1, pp.1-38.
28. De Waal, A., Barnard, E. (2008), *Evaluating topic models with stability*, Nineteenth Annual Symposium of the Pattern Recognition Association of South Africa (PRASA 2008), Cape Town, South Africa, 27-28 November 2008, pp 79-84.
29. De Waal A., Venter J.P., Barnard E. (2007), *Applying Topic Modelling on Forensic Data: A Case Study*, International Federation for Information Processing, Advances in Digital Forensics IV, eds. Shenoi, S., Vol 242, pp 303-315, Springer Boston.
30. Dewett, T. (2001), *The role of information technology in the organization: a review, model, and assessment*, Journal of Management, Vol. 27, No. 3, pp. 313-346.
31. Dixon, N. (2000), *Common Knowledge*, Cambridge, MA: Harvard Business School Press.
32. Docherty, M. (2006), *Primer on open innovation: Principles and practice*, PDMA Visions, Nr. 2, pp 13-17.
33. Du Preez, N. (2009), *Innovation Completes the Full Circle*, Indutech News. Vol. 8, issue 1, pp. 1-2.
34. Du Preez, N., Louw, L. (2008), *A Framework for Managing the Innovation Process*, PICMET 2008 Proceedings, 27-31 July, Cape Town, South Africa. PICMET. pp. 546-558.
35. Ehrlenspiel, K. (2007), *Integrierte Produktentwicklung: Denkablaeufe, Methodeneinsatz, Zusammenarbeit* (3rd rev. ed.), Muenchen: Carl Hanser Verlag.
36. Eppler, M.J., Mengis, J. (2004), *The Concept of Information Overload: A Review of Literature from Organisation Science, Accounting, Marketing, MIS and Related Disciplines*, The Information Society, vol. 20, pp. 325-344.
37. Erosheva, E. (2002), *Grade of membership and latent structure models with application to disability survey data*, PhD thesis, Carnegie Mellon University, Department of Statistics.
38. Essmann, H.E. (2009), *The Integration of Project Management Processes with a Methodology to Manage a Radical Innovation Project*, PhD Thesis in Industrial Engineering, Stellenbosch University, 2009.
39. Fei-Fei, L., Perona, P. (2005), *A Bayesian hierarchical model for learning natural scene categories*, IEEE Computer Vision and Pattern Recognition.
40. Feldman, S. (2004), *The High Cost of Not Finding Information*, KM World, March 1, 2004. http://www.kmworld.com/Articles/PrintArticle.aspx?ArticleID=9534
41. Finley, M. (2001). *Expertise Management 101*, In Future Shoes: February 16, 2001. http://www.computeruser.com/articles/daily/7,5,1,0216,01.html
42. Firestone, J.M. (2008), *Defining the Enterprise Information Portal*, DSs`tar.
43. Francis, L.A. (2006), *Taming Text: An Introduction to Text Mining*, Casualty Actuarial Society Forum.
44. Galanakis, K. (2006), *Innovation process: Make sense using systems thinking*, Technovation, 26(11), pp 1222-1232.
45. Gaynor, G.H. (2002), *Innovation by Design*, AMACOM, New York.
46. Graham, G. (2005), *Fuzzy Front End of Innovation". Broken Bulbs: Innovation*, http://orxilinasia.blogspot.com/2005/12/fuzzy-front-end.html.
47. Grant, R.M. (1996), *Prospering in dynamically-competitive environments: organizational capability as knowledge integration*, Organization Science, Vol. 7, No. 4, pp. 375-87.
48. Griffiths, T., Steyvers, M. (2002). A probabilistic approach to semantic representation. In Proceedings of the 24th Annual Conference of the Cognitive Science Society.
49. Griffiths, T., Steyvers, M., Tenenbaum, J. (2007), *Topics in Semantic Representation*, Psychological Review, 114(2): 211-244.

50. Gruber, T. R. (1993), *A Translation Approach to Portable Ontologies*, Knowledge Acquisition, 5(2):199-220.
51. Hall, R. (1993), *A framework for linking intangible resources and capabilities to sustainable competitive advantage*, Strategic Management Journal, Vol. 14, pp. 607-18.
52. Halkidi, M., Vazirgiannis, M. (2001), *Clustering Validity Assessment: Finding the Optimal Partitioning of a Data Set*, Athens University of Economics & Business, Department of Informatics.
53. Hall, R. (1993), *A framework for linking intangible resources and capabilities to sustainable competitive advantage*, Strategic Management Journal, Vol. 14, pp. 607-18.
54. Harrington, B., Clark, S. (2007), *ASKNet: Automated Semantic Knowledge Network*, Association for the Advancement of Artificial Intelligence.
55. Heim, D. (2006), *Semantic Wikis in Knowledge Management*, Master's thesis, Fachhochschule Kaiserslautern.
56. Hering, D., Phillips, J. (2005), *Innovation Roles - The People You Need for Successful Innovation*, NetCentrics Corporation.
57. Hildrum, J. (2007), *Does the emergence of distributed innovation call for new innovation process theories?*, Centre for Technology, Innovation and Culture, University of Oslo, http://www.cas.uio.no/research/0708innovation/CASworkshop_Hildrum.pdf
58. Hofmann, T. (1999), *Probabilistic latent semantic indexing*, 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA.
59. Holloway, P. (2000), *How to Capture and Deploy Tacit Knowledge in Your Organization*, In Braintrust: January, 2000. http://www.knowledgeharvesting.org/presentations/How%20to%20Capture,%20Package%20and%20Deploy%20Tacit%20Knowledge.pdf
60. IBM (2004), *Roles for innovation: The right people at the right times*, White Paper, IBM Business Consulting Services.
61. IBM (2006), *Driving information-led business innovation with IBM Information Server*, Delivering information you can trust, September 2006, IBM Corporation. ftp://ftp.software.ibm.com/software/bigplays/tbc/us/Driving_Info_Paper.pdf
62. Iles, P., Yolles, M., Altman, Y. (2001), *HRM and Knowledge Management: Responding to the Challenge*, Research and Practice in Human Resource Management, 9(1), 3-33.
63. Inmon, W.H., O'Neil, B.K., Fryman, L. (2008), *Business Metadata: Capturing Enterprise Knowledge,* Morgan Kaufmann, Burlington, USA.
64. Ishmael, G.S., Callahan, R.H. (2006), *Looking for Ideas in All the Wrong Places: An Argument for Staying in the Box*, Decision Analyst, Inc.
65. Jordan, M. editor. (1999), *Learning in Graphical Models*, MIT Press, Cambridge, MA.
66. Kaji, H., Yasutsugu, M., Toshiko, A., Noriyuki, Y. (1999), *Navigation in an Association Thesaurus Automatically Generated from a Corpus*, Proceedings of the Sixteenth Joint Conference on Artificial Intelligence-IJCAI99 Workshop IRF-3: Text Mining: Foundations, Techniques, and Applications.
67. Kao, A., Poteet, S. (2005), *Text Mining and Natural Language Processing – Introduction for the Special Issue*, SIGKDD Exploration Newsletter, ACM Press, 7(1), pp.1-2.
68. Kelly, T., Littman, J. (2006), *The Ten Faces of Innovation: IDEO's Strategies for Beating the Devil's Advocate & Driving Creativity Throughout Your Organization*, Profile Books.
69. Khurana, A., Rosenthal, S. (1998), *Towards Holistic 'Front Ends' in New Product Development*, Journal of Product Innovation Management, 15(1), pp. 57–74.
70. Kleinbaum, A.M. (2006), *Measuring mail: New analyses of e-mail data for the study of cross-divisional innovation*, Harvard Business School, http://www.hbs.edu/doctoral/pdf/KleinbaumAOMPaper.pdf
71. KnowledgeRush (2009), *Simplex*, February 26, 2009, http://knowledgerush.com/kr/encyclopedia/Simplex/
72. Koen, P.A., Ajamian, G.M. Boyce, S., Clamen, A., Fisher, E., Fountoulakis, S., Johnson, A., Puri, P., Seibert, R. (2002), *Fuzzy-Front End: Effective Methods, Tools and Techniques*, In Belliveau, P., Griffen, A. and Sorermeyer, S., eds. PDMA Toolbook for New Product Development. New York: John Wiley and Sons, pp. 2-35.
73. Kogut, B., Zander, U. (1992), *Knowledge of the firm, combinative capabilities, and the replication of technology*, Organization Science, Vol. 3, No. 3, pp. 383-397.

74. Kosara, R., Miksch, S., Hauser, H. (2002), *Focus + Content Taken Literally*, IEEE Computer Graphics and Applications, pp 22-39.
75. Kotze, D.J. (2008), *The development of an Implementation Methodology for a Conceptual Framework Tool used for the Improved Viewing and Utilisation of Organisational Information*, Stellenbosch University.
76. Krebs, V. (2008), *Knowledge Networks Mapping and Measuring Knowledge Creation*, March 19, 2008, http://www.knetmap.com/knowledge-networks-mapping.html
77. Larsen, B., Aone, C. (1999), *Fast and Effective Text Mining Using Linear-time Document Clustering*, In: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 1999).
78. Li, W., Blei, D., McCallum, A. (2007), *Nonparametric Bayes pachinko allocation*, in The 23rd Conference on Uncertainty in Artificial Intelligence (UAI).
79. Li, W., McCallum, A. (2006), *Pachinko allocation: DAG-structured mixture models of topic correlations*, International Conference on Machine Learning (ICML).
80. Liddy, E.D. (2000), *Text Mining*, Bulletin of the American Society for Information Science. Vol. 27(1).
81. Liddy, E.D. (2003), *Natural Language Processing*, In Encyclopedia of Library and Information Science, 2nd Ed. Marcel Decker, Inc.
82. Liddy, E.D. (2002), *A Breadth of NLP Applications*, ELSNEWS. Newsletter of the European Network in Human Language Technologies.
83. Lieberman, J. (2007), *From Metadata to Megadata: Deriving Broad Knowledge from Document Collections*, Collexis, Inc., Whitepaper.
84. Liu, P., Dew, P. (2004), *Using Semantic Web Technologies to Improve Expertise Matching within Academia*, Proceedings of I-KNOW ’04, Graz, Austria, June 30 - July 2, 2004.
85. Lowden, B.G.T., Robinson, J. (2002), *An Analysis of File Space Properties using Clustering*, Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics: International Conference on Information Systems, Analysis and Synthesis: Computer Science I. v(5).
86. Maier, R., Hädrich, T., Peinl, R. (2005), *Enterprise Knowledge Infrastructures*. Springer-Verlag Berlin Heidelberg.
87. Manaris, B.Z. (1998), *Natural Language Processing: A Human-Computer Interaction Perspective*, Advances in Computers 47: 2-68.
88. Mani, I., Maybury, M.T. (1999), *Advances in Automatic Text Summarization*, MIT Press.
89. Mani, I. (2001), *Automatic Summarization*, John Benjamins Publishing Company.
90. Manning, C., Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA. 1999.
91. Manning, C., Raghavan, P., Schütze, H. (2008), *Introduction to Information Retrieval*, Cambridge University Press, New York, USA.
92. McComb, D. (2004), *Semantics in Business Systems*, San Francisco: Morgan Kaufman.
93. McGuinness, D. L. (2001), *Ontologies Come of Age*, in The Semantic Web: Why, What, and How (D. Fensel, et al., eds.), MIT Press, 2001.
94. Miller, L., Miller, R., Dismukes, J. (2006), *The Critical Role of Information and Information Technology in Future Accelerated Radical Innovation*, Information Knowledge Systems Management, vol. 5 pp. 63–99, IOS Press.
95. Mimno, D., Li, W., McCallum, A. (2007), *Mixtures of Hierarchical Topics with Pachinko Allocation*, in Proceedings of the 24th International Conference on Machine Learning.
96. Mimno, D., McCallum, A. (2007), *Expertise Modeling for Matching Papers with Reviewers*, Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’07), San Jose, California, USA.
97. Mimno, D., McCallum, A. (2007), *Mining a Digital Library for Influential Authors*, Joint Conference on Digital Libraries (JCDL ‘07), Vancouver, British Columbia, Canada.
98. Minka, T.P., Lafferty, J. (2002), *Expectation-propagation for the generative aspect model*, In Uncertainty in Artificial Intelligence (UAI).
99. Monge, P. R., Contractor, N. (2002), *Theories of communication networks*, New York: Oxford University Press.

100. Moore, G.A. (2005), *Dealing with Darwin: how great companies innovate at every phase of their evolution*, Portfolio, USA.
101. Nasukawa, T., Nagano, T. (2001), *Text Analysis and Knowledge Mining System*, IBM Systems Journal, vol. 40, no. 4.
102. Nigam, K., McCallum, A., Thrun, S., Mitchell, T. (2000), *Text classification from labeled and unlabeled documents using EM*, Machine Learning, vol. 39, no. 2/3, pp. 103–134.
103. Nonaka, I., Takeuchi, H. (1995), *The Knowledge-Creating Company – How Japanese Companies Create the Dynamics of Innovation*, Oxford University Press, New York, NY.
104. Olson, L., Shaffer, R. (2002), "*Expertise Management – and Beyond*".
105. Pretium (2005), *Structured Innovation Empowered by TRIZ*, Pretium Consulting Services, LLC.
106. Pritchard, J., Stephens, M., Donnelly, P. (2000), *Inference of population structure using multilocus genotype data*, Genetics, 155:945–959.
107. Pleschak, F., Sabisch, H. (1996), *Innovationsmanagement*, Stuttgart: Schaeffer-Poeschel.
108. Ramoni, M. Sebastiani, P., Cohen, P. (2002), *Bayesian Clustering by Dynamics*, Machine Learning, vol. 47(1), pp. 91–121 (2002) 2001 Kluwer Academic Publishers, Boston.
109. Rath, H.H. (2003), *The Topic Maps Handbook*, Empolis, GmbH.
110. Rauber, A., Paralic, J., Pampalk, E. (2000), *Empirical Evaluation of Clustering Algorithms*, Vienna University of Technology, Department of Software Technology. Technical University of Kosice, Department of Cybernetics and Artificial Intelligence.
111. Rauffet, P. (2007), *A Methodology for the Application of an Automated and Interactive Reification Process in a Virtual Community of Practice*, Master's Thesis. l'École Centrale Nantes et l'Université de Nantes. France.
112. Redfearn, J. (2006), *Text Mining – Briefing Paper*, National Centre for Text Mining. JISC. UK.
113. Ribière, V.M., Román, J.A. (2006) *Knowledge Flow*, Chapter In: Encyclopaedia of Knowledge Management. Idea Group Inc.
114. Rogers, E. M. (1995), *Diffusion of innovations*, 4th ed., New York: The Free Press.
115. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P. (2004), *The author-topic model for authors and documents*, AUAI '04: Proceedings of the 20th conference on Uncertainty in artificial intelligence. AUAI Press. pp. 487-494.
116. Rothwell, R. (1992), *Successful industrial innovation: critical factors for the 1990s*, R&D Management. Vol 22:3. pp. 221 – 239.
117. Rothwell, R. (1995), *Industrial innovation: success, strategy, trends*, In M. Dodgson and R. Rothwell, (Eds). The Handbook of Industrial Innovation (pp.33–53). Aldershot: Edward Elgar, Hants.
118. Salton, G., McGill, M. (1983), *Introduction to Modern Information Retrieval*, McGraw-Hill.
119. Sammer, M. (ed.) (2003), *An Illustrated Guide to Knowledge Management*, Graz: Wissensmanagement Forum
120. Schönström, M. (2005), *Creating knowledge networks: lessons from practice*, Journal of Knowledge Management, Vol. 9, No. 6, pp. 17 - 29, Emerald Group Publishing Limited.
121. Shah, C. (2002), *Automatic Organization of Text Documents in Categories Using Self-Organizing Map (SOM)*, IEEE's Regional Student Paper Contest.
122. Shilakes, C., Tylman, J. (1998), *Enterprise Information Portals*, Merrill Lynch, Inc., New York, NY, November 16, 1998.
123. Shipman, F.M., McCall, R. (1994), *Supporting Knowledge-Base Evolution with Incremental Formalization*, In Proceedings of Conference on Human Factors in Computing Systems (CHI) '94, April 24-28, Boston, Mass., USA. pp. 285-291.
124. Shum, S.B. (2006), *Knowledge Technologies in Context*, The Open University, Milton Keynes, UK.
125. Sivic, J., Rusell, B., Efros, A., Zisserman, A., Freeman, W. (2005), *Discovering objects and their location in images*, In International Conference on Computer Vision (ICCV 2005).
126. Snowdon, D., Grasso, A. (2002), *Diffusing Information in Organizational Settings: Learning from Experience*, In Proceedings of Conference on Human Factors in Computing Systems (CHI) 2002, April 20-25, 2002, Minneapolis, Minnesota, USA.
127. Souder, W.E., Moenaert, R.K. (1992), *Integrating marketing and R & D project personnel within innovation projects: An information uncertainty model*, Journal of Management Studies, Vol. 29, No. 4, pp. 485-512.

128. Stewart, T. (1997). In interview “*Tom Stewart on Intellectual Capital*”. Knowledge Inc., May 1997, http://webcom.com/quantera/llstewart.html
129. Steyvers, M., Griffiths, T. (2006), *Probabilistic Topic Models*, in Latent Semantic Analysis: A Road to Meaning, Trends in Cognitive Science. vol. 10, issue 7, pp. 327 – 334.
130. Steyvers, M., Smyth, P., Rosen-Zvi, M., Griffiths, T. (2004), *Probabilistic Author-Topic Models for Information Discovery*, In: 10th ACM SigKDD conference knowledge discovery and data mining (Seattle, 2004).
131. Taylor, D. (2007), *The Seven Core Roles of Innovation*, The Business Blog at Intuitive.com, August 14, 2008, http://www.intuitive.com/blog/index.shtml
132. Teh, Y., Jordan, M., Beal, M., Blei, D. (2005), *Hierarchical Dirichlet processes*, Journal of the American Statistical Association.
133. Teo, T.S.H. (2000), *Using the Internet for Competitive Intelligence in Singapore*, Competitive Intelligence Review, Vol. 11(2), 61-70.
134. Tibbetts, J. (1997), *Technology Scouting. Keeping Abreast of Science and Technology: Technical Intelligence for Business* (W.B. Ashton, R.A. Klavans eds.), Battelle Press, Columbus, Ohio.
135. Tidd, J., Bessant, J., Pavitt, K. (1998), *Managing Innovation: Integrating Technological, Market and Organizational Change*, West Sussex, England: John Wiley & Sons Ltd.
136. Trott, P. (2005), *Innovation Management and New Product Development*, 3rd edition. Harlow, England: Pearson Education Limited.
137. Uys, J.W., Du Preez, N.D., Uys, E.W. (2008), *Leveraging Unstructured Information Using Topic Modelling*, PICMET 2008 Proceedings, 27-31 July, Cape Town, South Africa. PICMET. pp. 955-951.
138. Von Krogh, G., Ichijo, K., Nonaka, I. (2000), *Enabling Knowledge Creation*. New York: Oxford University Press.
139. Wahl, Z. (2006), *Masterclass: Business Taxonomy, Part I*, Inside Knowledge, October 31, 2006, http://www.ikmagazine.com/
140. Walsham, G. (2001), *Knowledge management: the benefits and limitations of computer systems*, European Management Journal, Vol. 19, No. 6, pp. 599-608.
141. Wei, W., Croft, W. (2006), *LDA-Based Document Models for Ad-hoc Retrieval*, in Proceeding of 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR ‘06), Seattle, WA, USA.
142. Weiss, S., Indurkhya, N., Zhang, T., Damerau, F. (2005), *Text Mining*, Springer.
143. Wikipedia (2009), *Maximum Likelihood*, August 5, 2009, http://en.wikipedia.org/wiki/Maximum_likelihood
144. WordNet Database (2006), Princeton University.
145. Wycoff, J. (2003), *Project Management vs. Managing Innovation Projects*, www.innovationtools.com
146. Yang, H., Callan, J. (2006), *Near-duplicate detection by instance-level constrained clustering*, In Proceedings of the Twenty Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle.
147. Yimam-Seid, D., Kobsa, A (2003), *Expert finding systems for organizations: Problem and domain analysis and the Demoir approach*, Journal of Organizational Computing and Electronic Commerce, Vol. 13, No. 1, pp. 1-24.

15. Appendix A - Specific Text Analytical Techniques

15.1 Term Frequency Inverse Document Frequency Technique

Term Frequency Inverse Document Frequency (tf-idf) is a basic methodology suggested by information retrieval researchers for exploiting information in text corpora (Blei et al., 2003). This popular methodology has been successfully deployed in many modern Internet and desktop search engines (i.e. search engines employed on individual computers). The tf-idf technique represents each document in the corpus as a real-value vector constituting ratios of counts (Salton & McGill, 1983). A basic vocabulary of words or terms is created and for every individual corpus document, a count is formulated based on the total number of occurrences of each of the words or terms in the given document (i.e. the term frequency counts). Only using raw term frequency counts to identify relevant documents has a critical disadvantage though: All terms are viewed as equally important when used to assess the relevancy on a query. Since certain terms have little or no discriminating power when it comes to determining document relevance (e.g. a corpus of documents on the higher education research is likely to have the term "education" in about every document). A mechanism for attenuating the effect of frequently occurring terms in a given collection is therefore required to sensibly perform relevance determination (Manning et al., 2008). To achieve this, the individual normalised (per document) term frequency counts are scaled by multiplying it with a new measure, called the inverse document frequency to obtain a more useful relevancy weight for each term. The inverse document frequency of a given term *t* (idf_t) is calculated as follows:

$$idf_t = log\frac{N}{df_t}$$

In the equation above, *N* represents the total number of documents in the corpus and df_t represents the document frequency of term *t*. The *tf-idf* weight of a given term and document combination may be calculated by the following equation:

$$tfidf_{t,d} = tf_{t,d} \times idf_t$$

The $tfidf_{t,d}$ weight assigns to term *t* a weight in document *d* and has the following characteristics:

1. This weight is the highest when the term *t* occurs frequently within a few documents (thereby bestowing high discriminating power to those documents)
2. This weight is lower in cases where the term occurs fewer times in a document, or occurs in many documents (therefore offering a less pronounced relevance indication);
3. This weight is the lowest when the term occurs in nearly all documents.

A term-by-document matrix is created having the individual vocabulary words/terms as rows and columns containing the tf-idf values for each respective document. The tf-idf technique reduces documents of arbitrary length to fixed-length lists of real numbers (Blei et al., 2003). Given a freeform query (i.e. a simple set of words without any connecting search operators), documents relevant to the individual query words may be identified by using the term-by-document matrix to find the respective documents having favourable tf-idf scores for the respective query words (Manning et al., 2008).

The tf-idf reduction technique is attractive for many reasons, one being its basic ability to identify a set of words that are discriminative for documents in a corpus. The shortcomings of this technique are the relatively small reduction in the descriptions of the documents constituting the corpus as well as that it exposes little in terms of the statistical inter- and intra-document structure (Blei et al., 2003).

15.2 K-means Clustering

K-means clustering, more commonly used than hierarchical clustering, is a clustering technique that accepts the number of clusters to form and uses a statistical dissimilarity measure between elements to group most similar elements and separate elements that are the most dissimilar (Francis, 2006). The objective of K-means Clustering is to minimise the average distance between documents and their centroids which corresponds to maximising the similarity between documents and their centroids. This technique tries to find cluster centroids that are good representatives of the data. The cluster centroids may be regarded as the model that generates the data. A document may be generated using this model by choosing a centroid at random and subsequently adding some noise. K-means Clustering is often regarded as the most significant flat clustering method and has favourable time complexity since it is linear in all relevant factors, namely:

- Iterations
- Number of clusters
- Number of vectors
- Dimensionality of the space

This makes K-means Clustering more efficient than hierarchical clustering algorithms. (Manning et al., 2008)

15.3 Hierarchical Clustering

Hierarchical clustering is a popular method used in text mining to cluster terms with the goal to discover content or to create features to be used for further analysis. Hierarchical clustering is a stepwise procedure of sequentially combining clusters in close proximity until some termination criteria are satisfied. It produces dendograms (i.e. treelike structures) of each of the various clustering steps assisting the user in determining the optimal number of clusters to select. Both text records and variables may be analysed by means of Hierarchical Clustering. Hierarchical Clustering can be used to gain insight into word combinations that tend to occur together in a set of data and that may possibly be associated with similar events or concepts (Francis, 2006).

15.4 Bayesian Clustering

Bayesian Clustering, a probabilistic approach, employs Bayesian probability theory to determine the probability that a certain element belongs in a certain group (Rauber, Paralic & Pampalk, 2000; Courseault, 2004). Cheeseman and Stutz (1996) pioneered Bayesian clustering methods for static datasets (i.e. assumption that the data are independent and identically distributed). However, a Bayesian approach is very suitable to clustering dynamic processes since it provides a principled way to combine prior and current evidence (Ramoni, Sebastiani & Cohen, 2002).

15.5 Self-organising Maps

Self-organising Maps (SOM), invented by Dr. Teuvo Kohonen, is an artificial intelligence based, data visualisation approach that uses unsupervised self-organising neural networks to reduce the dimensionality of data and display similarities and relationships in data (Casey, 2005; Courseault, 2004). The set of reduced data dimensions are plotted using a variety of colour schemes and two-dimensional or three-dimensional graphs to further ease of interpretation. One application of Self-organising Maps is clustering documents using similarities between the documents' semantic maps (Shah, 2002). Another popular application of SOMs is visualising gene expression profiles derived from DNA microarray data (Casey, 2005).

15.6 Factor Analysis

Factor Analysis is a fairly linear statistical clustering approach that uses similarity measures to partition documents (Halkidi & Vazirgiannis, 2001). As for Principal Components Analysis, Factor Analysis may be used to discover underlying factors or constructs which explain the correlations among a set of elements. It is further used to represent a large number of elements by a smaller number of derived items, named factors.

15.7 Principal Components Analysis

Principal Components Analysis (PCA) is a popular technique for reducing the dimensionality of data and is often used in biological statistics (e.g. analysis gene expression data). It tries to capture the variation in multidimensional datasets and map the variance into a number of principal components which can be used to explain variance in the observed data to a large extent. The complex dataset can then be represented by these simpler, more comprehensible dimensions or principal components (Casey, 2005).

15.8 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is a well-known statistical method for fitting a statistical model to a given set of data and offers a way of finding estimates for the parameters of the model to provide a good fit to the actual data. Given a data set and an underlying probability model, MLE selects those model parameter values that make the data "more likely" than any other model parameter values would make them (Wikipedia, 5 August 2009).

15.9 Naïve Bayes Classification

Naive Bayes (NB) Classification is a supervised learning algorithm based on probabilistic foundations. There are two ways to setup a Naïve Bayes Classifier, namely using a multinomial model (giving rise to multinomial NB model[98]) or a Bernoulli model (giving rise to a multivariate Bernoulli model[99]). In the former NB model, one term is generated from the vocabulary in each position of the document where a generative model is assumed. In the latter model, an indicator for each term in the vocabulary is generated where a one indicates the presence of the specific term given the document and a zero indicates its absence in the document. The multinomial NB model and the multivariate Bernoulli model have the same time complexity (Manning et al., 2008).

The multivariate Bernoulli model employs binary occurrence information when classifying a test document, ignoring the number of occurrences. On the other hand, the multinomial NB model considers multiple occurrences. The multivariate Bernoulli model typically makes many misclassifications when operating on long documents and is therefore more suited to the classification of short documents. The multinomial NB model does not have the same shortcomings when it comes to the classification of longer documents. Another difference between these two models is how they cater for non-occurring terms. The multinomial NB model can cater for more features than the multivariate Bernoulli model. Another difference between the two classifiers is that the multivariate Bernoulli model captures non-occurrences (of words in documents) explicitly where the multinomial NB model does not.

NB models are based on the conditional independence assumption stating that features (i.e. words or terms) are independent of each other given the class, which is hardly true for terms in documents. In most cases, exactly the opposite is true since terms are actually highly dependent on other terms. For example, when dealing with the class “United States of America”, the word “new” is highly dependent on the words “york” as well as “jersey”. The relatively unrealistic conditional independence assumption let on to the name “Naïve Bayes” as it is in truth a fairly naïve as a model of natural language. In addition to the conditional independence assumption,

[98] The multinomial NB model is formally identical to the Multinomial Unigram Language Model.
[99] The multivariate Bernoulli model is equivalent to the Binary Independence Model.

the multinomial NB model assumes positional independence. In other words, the multinomial NB model assumes that a word occurring in the heading of a given document is equally important to a word occurring in a footnote.

The multivariate Bernoulli model only takes into account the absence or presence of a term given the document and therefore ignores positions of words in documents completely. The so-called bag-of-word model, employed in the multivariate Bernoulli model and many other language models, disposes of all information that is conveyed by the order in which words occur in natural language sentences. Naïve Bayes is reckoned as a good text classifier despite of it excessive simplicity. Although NB models make low quality probability estimates, it often makes surprisingly good classification decisions. Further advantages of NB models are that it is reasonably robust to noise features (words and terms not contributing to a given class) and concept drift (the gradual change over time of the concept underlying a given class) as well as its efficiency being its main strength (training and classification is achieved with a single pass over the data). Therefore NB models are frequently used as a baseline in text classification research. NB models are preferred in the following scenarios:

- When slight improvement in accuracy is not worth the extra effort required by other text classification methods
- A large quantity of training data is available not suitable for more accurate text classification methods
- Robustness to concept drift is required

However, it has been proven that NB classifiers cannot compete with classifiers like SVMs in terms of average effectiveness when these classifiers are trained and tested on independent, identically distributed data. Given real world data, not being independent, identically distributed and displaying signs of concept drift, the difference in the consistency of classification quality between NB classifiers and other more accurate classifiers is often marginal. In conclusion, the quality of classification results depends on much more than just the machine learning algorithm used as the nature of the document collection, the experimental setup and the class definition significantly contributes to the end results (Manning et al., 2008).

15.10 Rocchio Classification

Rocchio Classification, a form of Rocchio Relevance Feedback, is based on a vector space which is divided into regions centred on centroids for each class. The centroid is effectively the calculated centre of mass of all documents of a given class. Rocchio Classification has the advantages that it is simple and efficient, but has the disadvantage that it is inaccurate when the approximation that classes are spheres having similar radii does not apply to the data. (Manning et al., 2008)

15.11 K-Nearest Neighbour Classification

k-Nearest Neighbour (kNN) classification operates on the principle that the majority class of k nearest neighbour classes is assigned to a test document. An unprocessed training document set can directly be used for classification as kNN does not require explicit training. It has the advantage that it can handle non-spherical and complex classes better than Rocchio Classification if the training set is large. kNN has the disadvantage that it is less efficient that other classification methods. (Manning et al., 2008)

15.12 Expectation Maximisation

Expectation Maximisation (EM), a generalisation of K-means Clustering, is a form of model-based clustering and was first introduced by Dempster, Laird and Rubin (1977). Model-based clustering accepts that the observed data was generated by a model and attempts to recover this model from the data. The recovered model defines the respective clusters and the assignments of the documents to these clusters. The maximum likelihood criterion is generally used for estimating the model parameters. Model-based clustering is more flexible than k-means and most hierarchical clustering methods, which makes somewhat rigid assumptions about the data (e.g. k-means clusters are assumed to be spheres), as the clustering model can be adjusted to incorporate what the user knows about the distribution underlying the data.

The Expectation-Maximisation (EM) algorithm, often used for model-based clustering and many other types of probabilistic modelling, is an iterative algorithm that maximises the objective function that measures the goodness of the clustering. It is further a soft clustering approach. In the EM model, a document is generated by selecting a cluster ω_k with probability α_k and subsequently generating the document terms according to the parameters q_{mk}. EM normally converges to a soft assignment. The importance of identifying good seed documents is even more critical for EM than for K-means as bad seed documents may cause the model to get stuck at local optima resulting in substandard clustering results. (Manning et al., 2008)

EM algorithms further represent a standard procedure to estimate the maximum likelihood of latent variable models (Chen & McLeod, 2005). Specific EM algorithms have been used to estimate different variants of the aspect model for collaborative filtering (Hoffman, 1999).

15.13 Support Vector Machines

The name “Support Vector Machine” (SVM) arose from neural network literature where learning algorithms were referred to as architectures or machines. The distinguishing element in support vector machines is that the selection of the decision boundary to use is entirely decided (or supported) by a few training data points known as the support vectors. An SVM is a type of large-margin classifier. More specifically, it is a vector space based machine learning method that has the objective of finding the decision boundary between two classes that is the maximum distance apart from any point in the training data. In the process some data points may be disregarded as outliers or noise. In the past 20 years intensive research has been done to improve the effectiveness of classifiers leading to a new paradigm of state-of-the-art classifiers such as:

- Support Vector Machines (SVMs)
- Boosted decision trees
- Regularised logistic regression
- Neural networks, and
- Random forests

Many of these methods have been applied in information retrieval problems and text classification especially. The performance of SVMs is at the state-of-the-art level and SVMs currently have considerable theoretical and empirical attractiveness. However, SVMs are not necessarily superior to other machine learning methods with the exception of cases where little training data is available. The following extensions to the SVM model have been developed:

- Soft margin classification – caters for the improved separation of the bulk of the data while ignoring a few uncharacteristic noise documents.
- Multiclass SVMs – caters for classes that are not just a set of independent, categorical labels, but may be arbitrary structured objects with relationships defined between them.
- Nonlinear SVMs – caters for data sets not linearly separable
- Transductive SVMs – performs semi-supervised instead of supervise learning

The training time of SVMs largely depends on the time required to solve the quadratic programming problem underlying most SVMs, and therefore varies depending on the method used to solve it. Typically the training time required for SVMs is the cubic of the size of the data set. It is therefore difficult, if not impossible, to use conventional SVM algorithms on really large training data sets due to their super-linear training time. A promising new training algorithm, based on cutting plane techniques, addresses the issue of the high training time requirements of SVMs, but still remains slower than the Naive Bayes model where simple term counting is involved instead of performing quadratic optimisation. (Manning et al., 2008)

15.14 Latent Semantic Indexing

Latent Semantic Indexing (LSI) uses a linear algebra technique, singular value decomposition (SVD), and the bag-of-words representation of text documents for extracting words with similar meanings (Hofmann, 1999; Deerwester, Dumais, Furnas, Landauer & Harshman, 1990). LSI is also known as Latent Semantic Analyses (LSA). A natural representation of a set of documents may be achieved by viewing the document set as a collection of vectors using a term-document matrix. A term-document matrix is an M x N matrix; its rows representing the M terms (corresponding to words found in the document set) and its columns representing each of the N documents. (Manning et al., 2008) Each document is therefore represented by indicating in the column, corresponding to the given document, which words or terms (the respective rows of the matrix) are found in the document in question. Cells corresponding to terms found in the given document are populated with ones, while cells corresponding to terms not found in the given document are filled with zeros. A term-document matrix is usually a sparse matrix since it contains many more zeros than ones. A more compact representation of each document in the collection may be obtained by finding a low rank approximation of the term-document matrix using a technique such as SVD to reduce the number of rows and columns representing the terms and documents. Apart from obtaining the low rank representation, LSI is also the process of casting queries into this low-rank representation providing the ability to calculate query-document similarity scores in this representation. When reducing the initial higher dimensional space to a more compact k-dimensional space, SVD merges terms with similar co-occurrences causing little reduction in retrieval quality. Some important steps in the LSI process are:

1. Using SVD to find a low rank approximation of the term-document matrix
2. Using the lower dimensional LSI representation to calculate similarities between document vectors
3. Using cosine similarities to calculate the similarity between a query and a document, between two documents, or between two terms

A mechanism is also provided to add (or “fold in”) a new document (a document that was not in the initial collection) to the LSI representation allowing the incremental addition of documents to an LSI representation. However, new terms only contained in these new documents cannot be included in the representation and such incremental addition further fails to capture the co-occurrences of the newly added documents. This causes the quality of the LSI representation to degrade as the number of incrementally added documents increases; eventually necessitating the recalculation of the LSI representation. SVD has a significant computational cost and poses the biggest obstruction to the widespread acceptance of LSI. To date no successful experiments using LSI have been done with over one million documents. LSI further has two drawbacks in common with vector space retrieval models:

1. No good way exists to express negations (e.g. finding documents containing the word “trees” but not “decision”)
2. No way of enforcing Boolean conditions

Except for the large time requirement of SVD computation causing poor scaling, inability of expressing negations and enforcing Boolean conditions, LSA has the additional drawback that it is a non-probabilistic model (Blei et al., 2003). LSI may be regarded as a soft clustering technique by viewing each dimension in the low rank representation as a cluster and the value corresponding to a given document in that dimension as its fractional membership in that cluster. The vector space representation is unable to cope with two classical problems associated with natural languages, namely synonymy and polysemy. LSI however, overcomes these shortcomings by using co-occurrences of terms to capture latent semantic associations (Manning et al., 2008; Blei et al., 2003). Lastly, LSI functions best in scenarios where little overlap between queries and documents exist (Manning et al., 2008).

16. Appendix B - Topic Modelling Techniques

16.1 Unigram Model

As many other statistical topic models, the unigram model operate in the space of distributions over words. Such a distribution can be regarded as a point on the $(V - 1)$-simplex, called the word simplex (recall that V denotes the number of words in the vocabulary of the corpus). In geometry, a simplex refers to an n-dimensional figure (in this case an n-simplex is implied), being the convex hull of a set of (n + 1) affinely independent points in a given Euclidean space[100]. A point therefore represents a 0-simplex, a line segment represents a 1-simplex, a triangle represents a 2-simplex and a tetrahedron represents a 3-simplex, where each of these geometric shapes includes an interior (KnowledgeRush, February 2009).

The Unigram Model identifies a single point on the word simplex and postulates that all words in the document corpus origins from the applicable distribution (a single "topic" for all documents is therefore implied). In the case of the unigram model, the words of each document in the analysed corpus are drawn independently from a single multinomial distribution. The following equation gives the probability of a document under the Unigram Model:

$$p(\boldsymbol{w}) = \prod_{n=1}^{N} p(w_n)$$

.

A useful way of depicting the differences between different latent topic models is by regarding the geometry of the associated latent spaces of each model, and examining how a document is represented in the respective geometries under each model (Blei et al., 2003). More specifically, topics models can be depicted graphically using directed graphs (De Waal & Barnard, 2008). In such a graph, variables are represented by nodes, dependencies between variables by edges, and replication by plates (where plates can also be nested within one another).

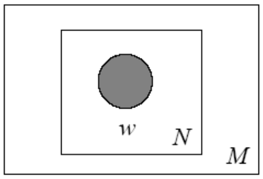


Figure 57: Plate notation of the Unigram Model

[100] A Euclidean Space is a set of points satisfying the condition that no m-plane contains more than (m + 1) of these points.

In the graphical model representation above, M is a variable that represents the total number of documents in corpus, N is a variable representing the number of words in a given document and w is a variable representing the individual words of the vocabulary. The boxes in this representation are “plates” representing replicates. The outer plate denotes documents, while the inner plate denotes the repeated choice of words within a document (Blei et al., 2003).

16.2 Mixture of Unigrams Model

The Mixture of Unigrams Model employs a probabilistic generative model for the data, but assumes that each document is characterised by exactly one topic (Blei et al., 2003). This assumption is regarded to be too simplistic to effectively model a large document collection (Wei et al., 2006). The mixture of unigrams model can be obtained by augmenting the Unigrams Model with a discrete random topic variable z. Under this model, each document is generated by first choosing a topic z and subsequently generating N words independently from the conditional multinomial $p(w|z)$. When the word distributions are estimated from a corpus, these distributions can be regarded as representations of topics under the assumption that each document presents a single topic only.

The following equation gives the probability of a document under the Mixture of Unigrams Model:

$$p(\boldsymbol{w}) = \sum_{Z} p(z) \prod_{n=1}^{N} p(w_n|z)$$

The different latent variable models (i.e. topic models) evaluates k points on the word simplex and create a sub-simplex based on these points, called the topic simplex. Any point on the topic simplex is also a point on the word simplex. It is this topic simplex that various latent variable models use in different ways to generate a document (Blei et al., 2003). Figure 58 illustrates the topic simplex for three topics embedded in the word simplex for three words. The three corners of the triangle representing the word simplex match the three distributions where each word has a probability of exactly one respectively. The three points of the triangle that represents the topic simplex correspond to three different distributions over words. The Mixture of Unigrams Model places each document at one of the corners of the topic simplex (since a document may only have a single associated topic under this model). More specifically, this model posits that for each document, one of the k points on the word simplex (this corresponds to one of the corners of the topic simplex) is chosen randomly and all the words of the document are drawn from the distribution corresponding to that point (Blei et al., 2003).

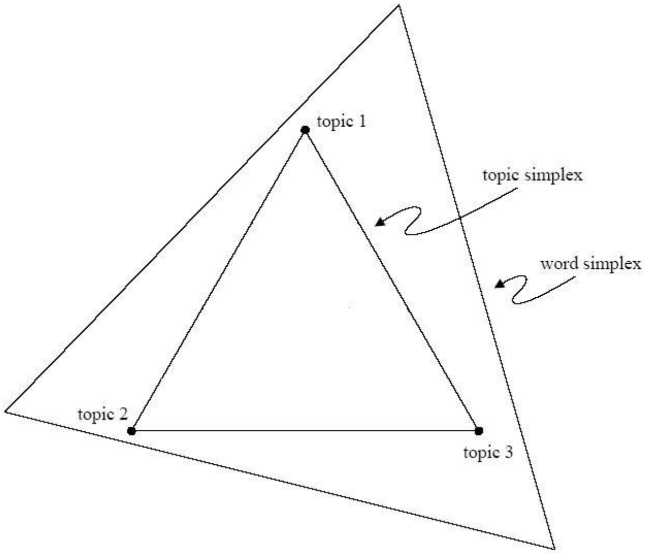


Figure 58: Topic simplex for three topics in the word simplex for three words under the Mixture of Unigrams Model[101]

Figure 59 represents the plate notation of the mixture of unigrams model.

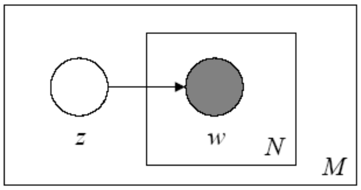


Figure 59: Plate notation of the Mixture of Unigrams Model

In the plate notation representation above, the outer plate denotes documents, while the inner plate denotes the repeated choice of words within a document. As before, M is a variable that represents the total number of documents in corpus, N is a variable representing the number of words in a given document, w is a variable representing the individual words of the vocabulary and z is a variable denoting a discrete random topic variable.

Lastly, it is worth mentioning that the Mixture of Unigrams Model corresponds to the supervised Naive Bayes Model (Blei et al., 2003).

[101] This figure was adapted from Blei et al. (2003)

16.3 Probabilistic Latent Semantic Indexing (pLSI)

Probabilistic Latent Semantic Indexing (pLSI), the probabilistic variant of LSI, has a statistical foundation and attempts to define a generative data model. pLSI is also known as the aspect model. A probabilistic model has the advantage that standard statistical techniques can be applied for questions like model fitting, model combination, and complexity control (Hofmann, 1999). The pLSI model views each word in a document as a sample from a mixture model comprised of multinomial random variable components. These random variables may be regarded as topic representations. Each word is thus generated from a single topic, while different words in a document may be generated from different topics. Each document is represented by a list of the mixing proportions for the respective (topic) mixture components (i.e. the multinomial random variables corresponding to topics) resulting in a reduced probability distribution on a set of topics. pLSI, however, does not provide a probabilistic model at the level of documents since there is not generative probabilistic model for the topic mixing proportions.

Blei et al. (2003) states that the pLSI model posits that a document label d and a word w_n are conditionally independent given an unobserved topic z:

$$p(d, w_n) = p(d)p(w_n|z)p(z|d)$$

In the pLSI model, a document may be generated from more than one topic – an improvement over the mixture of unigrams model – with $p(z|d)$ serving as the mixture weights of the topics for a given document d. Blei et al. (2003) reports that a major shortcoming of the pLSI model is that it learns the topic mixtures only for documents it has been trained on having the restriction that there is no way to assign topics to a previously unseen document. A second shortcoming of this model is the fact that the number of parameters that must be estimated increases in a linear fashion with the number of training documents causing the model to scale poorly with the size of the corpus and making it prone to overfitting.

In terms of the graphical representation of word and topics simplexes, illustrated in Figure 60, the pLSI model induces an empirical distribution on the topic simplex denoted by the respective x's.

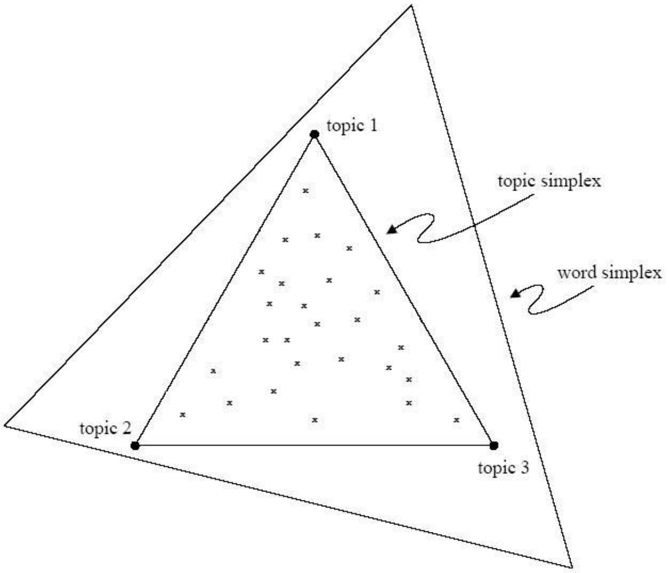


Figure 60: Topic simplex for 3 topics in a 3-word simplex under the pLSI Model[102]

Figure 61 represents the plate notation of the pLSI model.

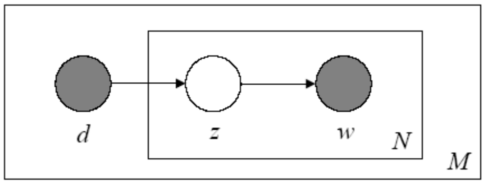


Figure 61: Plate notation of the pLSI Model

Under the pLSI model each word w of a training document d originates from a randomly chosen topic z. The topics are drawn from a document-specific distribution over topics corresponding to points on the topic simplex (i.e. the x's in Figure 60 above). The training document set defines an empirical distribution on the topic simplex since there is one document-specific distribution over topics for each document (Blei et al., 2003).

16.4 Latent Dirichlet Allocation (LDA)

The Latent Dirichlet Allocation (LDA) model assumes that words are generated by (a mixture of) topics and that such topics are infinitely exchangeable within a document (Blei et al., 2003; Blei et al., 2007). Moreover, documents are represented as random mixtures over latent (i.e. underlying) topics where each topic is defined or characterised by a distribution of (corpus vocabulary) words (Mimno & McCullum, 2007 [2]). The LDA model's latent multinomial variables are referred to as

[102] This figure was adapted from Blei et al. (2003).

topics. LDA is a true generative probabilistic model of a document corpus and potentially overcomes the drawbacks of earlier topic models (e.g. pLSI, Mixture of Unigrams, etc.) due to its generative properties and the possibility of assigning multiple topics per document (Uys et al., 2008). It further caters for synonymy (Griffiths et al., 2007) and polysemy (Steyvers et al., 2006). LDA offers a fresh, interesting approach to model documents (e.g. when compared to the standard query likelihood model) (Wei et al., 2006). It further represents a very useful model for deriving structure in otherwise unstructured data as well as generalising new data to fit into that structure (Blei & Lafferty, 2006 [3]).

On a high level, the generation of a document corpus in LDA is modelled as a three step process. The first step entails sampling a distribution over topics from a Dirichlet distribution for each document. Subsequently, a single topic is selected from this distribution for each word in the document. The last step involves sampling each word from a multinomial distribution over words corresponding to the sampled topic (Steyvers, Smyth, Rosen-Zvi & Griffiths, 2004).

More specifically, the following generative process is assumed for each document $\boldsymbol{w}$ in a corpus D (Blei et al., 2003)

1. Choose $N \sim Poisson(\xi)$
2. Choose $\theta \sim Dir(\alpha)$
3. For each of the N words w_n:
   a. Choose a topic $z_n \sim Multinonial(\theta)$
   b. Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

Blei et al. (2003) simplifies this process to arrive at the following:

1. Choose $\theta \sim Dir(\alpha)$
2. For each of the N words w_n:
   a. Choose a word w_n from $p(w_n|\theta, \beta)$

The simplified process above defines a marginal distribution[103] of a document $\boldsymbol{w}$ as the following continuous mixture distribution:

$$p(\boldsymbol{w}|\theta, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} p(w_n\,|\theta, \beta) \right) d\theta$$

In the equation above $p(w_n|\theta, \beta)$ represents the mixture components and $p(\theta|\alpha)$ the mixture weights. There are three levels to the LDA representation as depicted in the plate notation in Figure 62. The highest level, say the corpus level, involves corpus level parameters α and β which are assumed to be sampled once in the process of generating a corpus. The second level, say the document level, is linked to document-level variables θ_d which are sampled once per

[103] Given a joint distribution of two random variables a marginal distribution of one variable can be obtained by integrating out the other variable.

document. The third and most detailed level, the word-level, is represented by word-level variables z_{dn} and w_{dn} which are sampled once for each word in each document.

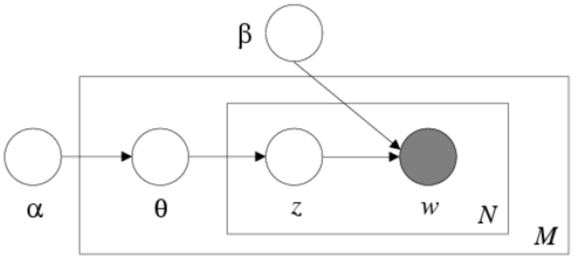


Figure 62: Plate notation of the LDA Model

In terms of the graphical representation of word and topic simplexes, illustrated in Figure 63, the LDA model places a smooth distribution on the topic simplex denoted by the contour lines in this figure. More specifically, LDA posits that each word of observed and unseen documents is generated by randomly selecting a topic from a distribution with a randomly chosen parameter which is sampled once per document from a smooth distribution on the topic simplex.

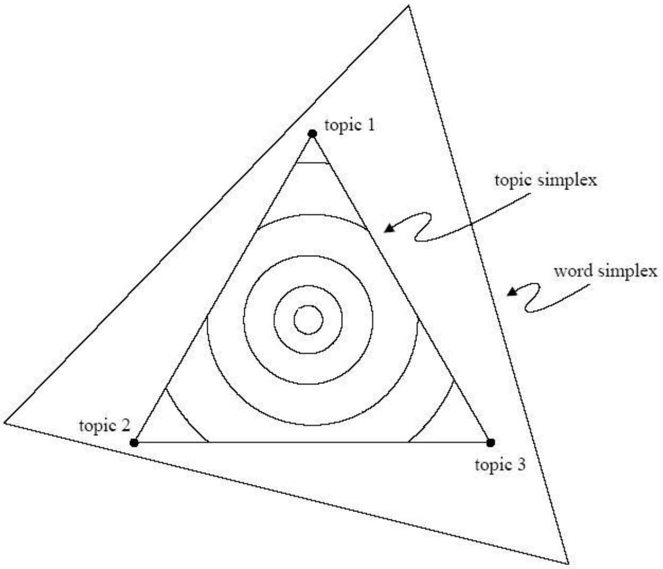


Figure 63: Topic simplex for 3 topics in a 3-word simplex under the LDA Model[104]

The major inferential problem to be solved in order to use LDA is that of calculating the posterior distribution[105] of the hidden variables given a document. Regrettably, this distribution is

[104] This figure was adapted from Blei et al. (2003).

[105] The posterior probability of a random event is the revised probability of past events based on new evidence. The posterior probability distribution of one random variable given the value of another can be calculated using Bayes' theorem.

intractable for calculating exact inference. However, an assortment of approximate inference algorithms, e.g. Laplace approximation, variational approximation, Markov chain Monte Carlo, expectation propagation and collapsed Gibbs sampling, can possibly be applied to LDA (Jordan, 1999; Minka & Lafferty, 2002; Griffiths & Steyvers, 2002; Blei et al., 2003).

A shortcoming of LDA, shared by most probabilistic topic models, stems from the bag-of-words assumption which allows words that should be generated by the same topic (e.g. “Heart and Stroke Foundation Eating Plan”) to be allocated to various deferent topics (Blei et al., 2003). LDA has wider applicability than only text since it may be applied to address problems associated with other data collections such as content-based image retrieval and bioinformatics (Blei et al., 2003).

16.5 Author-Topic Model

The Author-Topic Model (Rosen-Zvi et al., 2004) is a generative model for documents which extends the LDA model to include an author dimension. More specifically, the Author-Topic Model extends LDA by permitting the mixture weights of individual topics to be determined by the document’s authors. Where LDA is an informative model about the content of documents, it provides no explicit information about the interest of the authors of documents. A given author may produce several documents, each containing one or more topics. It will be extremely useful to be able to describe the interests of authors by means of the topics in the associated documents.

One of the fundamental problems raised by large document collections is modelling the interests of associated authors. A variety of important questions about the content of document collection may be answered by means of an appropriate model of author interests:

- Determining the subjects an author writes about
- Finding which other authors are likely to have documents similar to an given document
- Identifying authors who produce similar work

Rosen-Zvi et al. (2004) reports that research on author modelling is inclined to focus on the problem of determining who wrote which document - a problem that may be solved by relatively trivial discriminative models. In the Author-Topic Model, each author has an associated multinomial distribution over topics, while each topic in turn has an associated multinomial distribution over words. For a document with multiple authors, the document is modelled as a mixed distribution over topics comprised of the different topic distributions associated with the respective authors. A set of topics featuring in the analysed corpus including their relevance to individual documents as well as topic-author affinities may be obtained by learning the parameters of the model. Figure 64 shows a graphical model of the author-topic model.

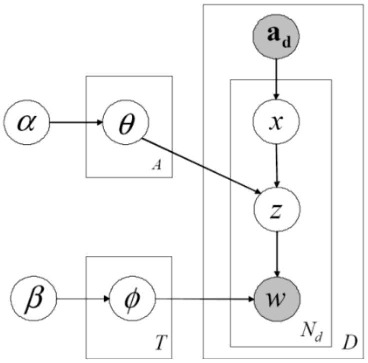


Figure 64: Plate notation for the Author-Topic Model

In this representation, d denotes a document (a vector of N_d words) and x denotes the specific author of the group of authors $\boldsymbol{a}_d$ responsible for a given word w. Each author is associated with a distribution of topics θ, which is drawn from a symmetric Dirichlet(α) prior. A topic z is selected by using the mixture weights associated with the given author and a word w is generated according to the distribution φ, associated with that topic, which is drawn from a symmetric Dirichlet(β) prior.

It is possible to obtain information about the set of topics individual authors usually write about by estimating the parameters φ and θ. In addition, a topic-based representation of the content of each document will also result with the estimation of these parameters. A range of methods have been applied in topic model parameter estimation. Rosen-Zvi et al., (2004) however prefers Gibbs sampling in the author-topic model setting due to its relative simplicity in obtaining parameter estimates under Dirichlet priors and the fact that it allows for the combination of estimates from various local maxima of the posterior distribution.

On a qualitative level, Rosen-Zvi et al., (2004) reports that rather specific topics were discovered by the author-topic model in the analysed test corpora. In addition, the resulting author lists were also fairly satisfactory. On a more quantitative level, they evaluated the perplexity[106] of the author-topic model to test its ability to predict words of unseen documents. They conclude that the Author-Topic Model provides a significant increase in predictive power when compared to the simplistic author model, where author interests are modelled directly by probability distributions over words. On the other hand, when comparing the perplexity of the author-topic model to that of LDA the Author-Topic Model has lower perplexity for a small amount of training words since it

[106] A quantitative measure to compare language models widely used to compare the predictive performance of topic models (Chemudugunta, Smyth and Steyvers, 2008). Although perplexity does not necessarily directly measure aspects of a model such as interpretability or coverage, it is nonetheless a useful general predictive metric for assessing the quality of a language model (Chemudugunta, Holloway, Smyth and Steyvers, 2008).

utilises its knowledge of the author to generate a better prior for the document content. However, as the amount of training words increases the predictive performance of LDA relative to that of the Author-Topic Model improves due the LDA's superior flexibility to adapt to the content of specific documents.

The ability of the Author-Topic Model to predict authors of documents is further examined in Rosen-Zvi et al. (2004). They show that the inclusion of authorship information significantly improves the accuracy of the model in terms of the expectation about the content of documents by specific authors. A related application of the Author-Topic Model may therefore be the identification of possible authors for novel documents.

In conclusion, the core value of the Author-Topic Model lies in its ability to explicitly include authors in the document models, which results in a general framework for answering queries and predicting at both author and document level.

16.6 Hierarchical Latent Dirichlet Allocation (hLDA)

An even more complex challenge than learning topics underlying unstructured text is learning a topic hierarchy from unstructured text. Given a collection of documents, each containing a set of words, the challenge in this case is not only to discover the underlying topics in the documents, but further to organise these topics into a hierarchy. An extension of the LDA model, named Hierarchical LDA (hLDA), has been developed for this purpose (Blei et al., 2004). hLDA specifies a generative probabilistic model for hierarchical structures and solves the problem of learning such structures from data using a Bayesian approach. In the hLDA model the random variables of the LDA model is extended to include another set of random variables representing the hierarchies. These new random variables are specified according to an algorithm which builds the hierarchy as data becomes available. The Chinese restaurant process, a distribution on partitions of integers, is used in hLDA as the probabilistic object to cater for a hierarchy of partitions. hLDA associates each hierarchy node with a topic, where a topic is a distribution over (corpus vocabulary) words. Under this model a document is generated by selecting a path from the root of the hierarchy to a leaf while repeatedly sampling topics along the traversed path as well as sampling words from the those topics. The sampler alternates between selecting a new path through the tree for each document and assigning each word, while learning the model from data.

The aim of organising topics into a hierarchy is to capture the context of the usage of topics throughout the corpus that reflects the embedded syntactic and semantic concepts of generality and specificity (Blei et al., 2004). With hLDA, the quality of the topic tree determines the quality of the distribution of topic mixtures (Mimno et al., 2007). A Gibbs sampling algorithm is used to approximate inference of the implied nested Chinese restaurant problem and the corresponding topics in the hLDA model, thereby learning the structure of the tree as well as the topics themselves. Future improvements to the hLDA model include:

(a) Allowing documents to be generated from a mixture of paths in the topic hierarchy

(b) Allowing different documents to have different path lengths in terms of the number of topics included in the path.

16.7 Dynamic Topic Model (DTM)

Most static topic models, like LDA, assume that documents are exchangeable or in other words, that the joint probability of corpus documents is invariant to permutation. In some cases this assumption is too restrictive, since documents about the same topic are not necessarily exchangeable as topics evolve over time. Document collections such as scholarly journals, e-mail, news articles, and search query logs all exhibit evolving content (Blei & Lafferty, 2006 [2]). In the Dynamic Topic Model (DTM) approach, the document corpus is divided into sequential segments (e.g., by year) and the document exchangeability assumption is made stricter by assuming that only the documents in each segment are exchangeable (Blei & Lafferty, 2006 [2]). DTM is then applied to the segmented corpus allowing topic distributions to evolve from segment to segment resulting in a hierarchical model of sequential document collections. DTM attempts to specify a statistical model of topic evolution. More specifically, Dynamic topic models provide a qualitative overview of the contents of a large document collection and also give quantitative, predictive models of a sequential document corpus (Blei & Lafferty, 2006 [2]). Assuming the data is divided by time slice, DTM models the documents of each slice using a K-component topic model, where the topics associated with time slice t evolve from the topics associated with time slice $t-1$.

In language modelling applications, Dirichlet distributions are typically used to model the uncertainty about the distributions over words. Dirichlet distributions are however not suitable for sequential modelling. In DTM the natural parameters of each topic k in time slice t are chained in a state space model which evolves with Gaussian noise. Therefore, in DTM the sequences of compositional random variables are modelled by chaining Gaussian distributions in a dynamic model and mapping the resulting values to the simplex. The sequential tying of a collection of topic models is thus achieved by chaining together topics and topic proportion distributions.

Whereas the document specific topic proportions θ are drawn from a Dirichlet distribution in the LDA model, a logistic normal with mean α is used to incorporate uncertainty over proportions in the DTM model. Figure 65 illustrates the graphical representation of a DTM with three time slices. $\beta_{t,k}$ represents each topic's natural parameters which evolve over time, along with the mean parameters α_t of the logistic normal distribution for the topic proportions.

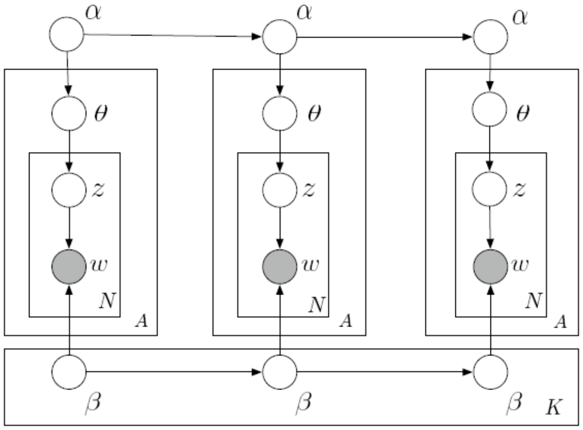


Figure 65: Plate notation of a Dynamic Topic Model for 3 time slices[107]

Utilising a time series over the natural parameters of topics makes possible the application of Gaussian models for representing the time dynamics. The non-conjugacy of the Gaussian and multinomial models makes posterior inference intractable. DTM circumvents this problem by using a deterministic variational method to approximate posterior inference. More specifically, Kalman filters and wavelet regression are used to achieve variational approximations (Blei & Lafferty, 2006[2]).

The most promising extension to DTM is to integrate a model of how new topics appear and disappear over time in the given corpus. This would be an improvement on the current model that assumes a fixed number of topics.

16.8 Correlated Topic Model (CTM)

Another limitation of static topic models, such as LDA, is its inability to model correlation between topics. In the case of LDA, this limitation stems from the independence assumptions implicit in the Dirichlet distribution that is used to model the variability among the topic proportions. The components of the topic proportions vector are nearly independent under a Dirichlet distribution, leading to a strong and unrealistic modelling assumption that the presence of one topic is not correlated with the presence of another (Blei & Lafferty, 2006 [1]).

The correlation between topics may be very useful when, for instance, a researcher in a narrow sub-discipline, is searching for a particular research article, and finds that this article is highly correlated with another topic that the researcher may not have been aware about and that is not explicitly included in the article (Uys et al., 2008). With the knowledge of the existence of this new related topic, the researcher could explore the document collection in a topic-guided manner to enquire connections to a body of work the researcher was previously unaware of. In most text

[107] Source: Blei & Lafferty (2006 [2])

document collections, it is realistic to anticipate that subsets of the underlying latent topics will be highly correlated. Another, more flexible, topic modelling approach, called the Correlated Topic Model (CTM), builds on LDA and employs an alternative, more flexible distribution for the topic proportions which provides for covariance structure among topics (Blei & Lafferty, 2006 [1]). This results in a more realistic model of the latent topic structure where the presence of one latent topic may be correlated with the presence of another. CTM employs a logistic normal distribution on the simplex which caters for a general pattern of variability between the individual components by transforming a multivariate normal random variable.

Figure 66 depicts example densities of the logistic normal distribution on a 2-simplex. The leftmost simplex shows diagonal covariance and nonzero-mean, the centre simplex shows negative correlation between components 1 and 2, while the rightmost simplex shows positive correlation between components 1 and 2.

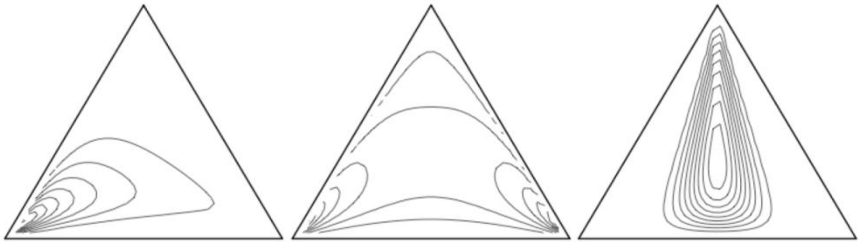


Figure 66: Example densities of the logistic normal distribution on a 2-simplex[108]

The process of generating a N-word document under CTM is identical to the generative process of LDA with the exception that topic proportions are drawn from a logistic normal distribution rather than a Dirichlet distribution. This removes the strong (topic) independence assumption implicit to the Dirichlet distribution. CTM is more expressive than LDA, since the covariance matrix of the logistic normal distribution employed in CTM caters for topic correlations. It therefore provides a natural mechanism for visualising and exploring unstructured data sets.

Posterior inference is the main challenge to using CTM (Blei & Lafferty, 2006 [1]). Variational Expectation-Maximization (EM) is used to maximise the bound on the log probability of a collection given by summing over the log probability all documents.

Blei and Lafferty (2006 [1]) compared LDA and CTM and reports that CTM provides a better model for the document collection used in the experiment. When compared to LDA, CTM better predicts the remaining words of a document after observing a portion of the document. This can be explained by the fact that CTM uses topic correlation to infer that words in a related topic may also be probable. On the other hand, LDA requires that a large portion of the document be

[108] Source: Blei & Lafferty (2007)

observed, in order for all the topics of the document to be represented, before being able to predict the remaining words of the document well.

CTM has the shortcomings that (a) only pair-wise correlations are modelled, and (b) the number of parameters in the covariance matrix increases as the square of the number of topics (Li et al., 2006).

16.9 Pachinko Allocation Model (PAM)

As mentioned before, the topics discovered by LDA capture correlations among words, but does not explicitly model correlations among topics due to sampling topic proportions of each document from a Dirichlet distribution. The fact that LDA neglects to take into account topic correlations hinders its ability to discover a large number of fine-grained, tightly-coherent topics (Li et al., 2006).

The Pachinko Allocation Model (PAM) (Li et al., 2006) presents a flexible alternative to CTM (Blei & Lafferty, 2006 [1]), which has the limitation that it only captures correlations between pairs of topics. PAM captures arbitrary, nested, and possibly sparse correlations between topics using an arbitrary, directed acyclic graph (Li et al., 2007). The leaves of the directed acyclic graph (DAG) symbolise individual vocabulary words, whereas the respective interior nodes corresponds to topics and capture the correlation among its children (which may in turn be words or other interior nodes). A traditional topic under the LDA model would be represented in the PAM model as an interior node having children that are all leaves (corresponding to vocabulary words). Under PAM, some interior nodes may also have some children that are also interior nodes (other topics) resulting in a mixture of topics. For many topics, PAM thus not only captures correlations among words (like LDA does) but in addition also captures correlation among topics. In other words, PAM extends the concept of topics to be distributions not only over words, but also over other topics.

With PAM the respective interior node distributions can be parameterised in an arbitrary fashion. One possibility is to use a Dirichlet distribution to represent the distribution of each interior node over its children. A PAM model therefore consists of a DAG with each interior node having a distribution over its children. Under PAM, a document may be generated by starting at the root of the DAG and subsequently sampling one of the children of this node according to its multinomial. This process is repeated for subsequent children of the selected interior nodes until a leaf of the DAG (representing a word) is reached.

The DAG structure in PAM is very flexible since it can accommodate hierarchies and arbitrary DAGs (e.g. DAGs with inter-connected edges or with edges skipping levels) and nodes may be sparsely or fully connected. LDA may be regarded as a special case of PAM since the LDA involves a three-level DAG with a root node on the first level, connected to all the topics on the second level, with all topics in turn connected to all vocabulary words on the third level.

The hidden variables to be learnt in PAM include the sampled multinomial distributions of each topic, the respective topic assignments as well as the parameters in the individual Dirichlet distributions. Gibbs sampling is subsequently applied for inference and parameter learning.

Li et al. (2006) compare PAM to LDA in terms of (a) topic clarity determined by human judgement, (b) likelihood of held-out test data, and (c) the accuracy of document classification. The human judges found PAM topics to be more semantic coherent and specific than LDA topics constructed from the same data. In terms of likelihood to held-out test data, PAM consistently has higher likelihood compared to LDA, especially for a large number of topics (e.g. 100 topics or more). They further report that CTM performs better than both LDA and PAM for a small number of topics (e.g. less than 50 topics). In terms of document classification the improvement in classification accuracy of PAM over LDA was determined to be statistically significant with p-value of less than 0.05.

Comparing PAM to other topic models, PAM is more flexible than hLDA since it samples a topic path for each word instead of each document. PAM is also more flexible than CTM because although both these models try to model topic correlations directly, PAM can capture n-ary and nested correlations where CTM is limited to pairwise correlations. Where with CTM parameters have to be estimated for each possible topic pair in the covariance matrix, with PAM there is no need to model every topic pair but only sparse mixtures of correlations regulated by the number of super-topics. On the other hand, it is more difficult to identify the appropriate topic structure for a given dataset when using PAM compared with some other simpler models like LDA (Li, et al., 2007).

16.10 Non-parametric PAM

PAM offers a powerful way to characterise inter-topic correlations and formulate large numbers of fine-grained topics from electronic text. It however shares a practical difficulty with many other topic models, namely determining the number of topics to use for the corpus at hand (Li et al., 2007). In the case of PAM, which can also accommodate hierarchies and arbitrary networks, the number of possible structures further complicates the task of selecting the most suitable one. Choosing the most suitable topic structure is therefore even harder with PAM when compared to simpler models such as LDA where structures are simpler (e.g. non-hierarchical) and yields fewer possibilities.

The most suitable number of topics can be learnt automatically using a non-parametric prior like the hierarchical Dirichlet process (HDP) described in Teh, Jordan, Beal and Blei (2005). HDPs are meant to model data collections having a pre-defined hierarchical structure, but have the limitation that it cannot automatically discover topic correlations from unstructured data. HDPs have been successfully applied to LDA for selecting the appropriate number of topics. A variant of the standard HDP mixture model, where HDPs are now built on dynamic groupings of data according to topic assignments, may be used as a non-parametric Bayesian prior for PAM (Li, et al., 2007) resulting in an extension to the fixed-structure PAM described in Li et al. (2006). More specifically, these authors describe how the non-parametric prior can be regarded as a variant of the Chinese restaurant process (CRP). Where Dirichlet processes have been used in many topic models as priors for mixture models, the HDP can be utilised for problems that require mixture components to be shared among multiple Dirichlet processes.

To generate a document under the non-parametric PAM model, multinomial distributions are firstly sampled over topics from the corresponding HDPs. Subsequently, a topic path is repeatedly sampled according to the multinomials for each word in the document. In order to generate a word in a document, a super-topic is first sampled according to the CRPs that sample the respective category. Given the super-topic, a sub-topic is sampled from the CRPs. Lastly, a word is sampled from the relevant sub-topic according to its multinomial distribution over words.

In addition to determining the suitable number of topics, non-parametric PAM also discovers a sparse connectivity between super-topics and sub-topics with the result that words may be grouped dynamically in terms of super-topic assignments.

As for inference, non-parametric PAM also employs Gibbs sampling for this task. Li et al. (2007) compares non-parametric PAM to HDP, PAM and hLDA in terms of likelihood of held-out test data. They found by means of paired t-tests that non-parametric PAM matches the best performance of PAM and performs significantly better than both HDP and hLDA.

16.11 Hierarchical Pachinko Allocation Model (hPAM)

The Hierarchical Pachinko Allocation Model (hPAM) addresses the shortcoming of PAM, namely that it cannot represent a nested hierarchy of topics. hPAM explicitly represents a topic hierarchy and is an enhancement of PAM. It combines the advantages of hLDA with the ability of PAM to mix different leaves of the topic hierarchy (Mimno et al., 2007). Topic models catering for a hierarchal topic structure are intuitively appealing and by taking topic hierarchy into account two advantages arise when compared to “flat” topic models such as LDA. Firstly, by explicitly modelling hierarchical topic co-occurrence patterns, more accurate models may be learnt having improved predictive capability. Secondly, the organisation of a corpus can be described more accurately by hierarchical topics models than their flat counterparts.

In contrast to hLDA, hPAM permits higher-level topics to share lower-level topics. Mimno et al. (2007) use a fixed number of topics, but mentions that non-parametric priors over the number of topics may be used to determine the number of topics automatically. Although PAM does not represent word distributions as parents of other distributions, it portrays some hierarchy-related characteristics. For example, trained PAM models frequently exhibit a single topic with high probability in all super-topic nodes, which has the appearance of a “background” topic (Mimno et al., 2007).

hPAM extends the basic PAM structure to represent hierarchical topics and may be defined as a PAM model where all nodes are associated with a distribution over the vocabulary (as opposed to PAM only allowing the nodes on the lowest level to have associated word distributions). This extension results in a highly flexible framework for hierarchical topic modelling. Mimno et al. (2007) further present two possible variations of hPAM. The first variation (i.e. hPAM1) contains a distribution over the levels of a path for every path through the DAG. These distributions are shared by all documents. The second variation (i.e. hPAM2) is similar to hPAM1 except for the absence of the distributions over path levels. In hPAM2 the Dirichlet distribution of each internal node has one additional dimension instead.

Both hPAM variations are trained by means of Gibbs sampling. Where hLDA learns a tree structure of topics, hPAM represents the hierarchical structure of topics by means of the Dirichlet-multinomial parameters of the internal node distributions. Training these parameters resembles a crucial part of hPAM. A good hierarchical topic model should have the ability to generalise to unobserved data. In terms of the empirical likelihood of held-out data that measure the ability of the model to predict unseen documents, hPAM produces better results that PAM, hLDA and LDA (Mimno et al., 2007). Ideally, a model should be good at predicting unseen documents with the maximum number of topics for finer granularity and improved interpretability. The likelihood measure for LDA decreases dramatically for 20 topics or more. PAM performs better in terms of the likelihood measure for larger numbers of topics and peaks at 40 sub-topics. hPAM portrays little decline for more than 60 topics and is more stable than at larger topic numbers compared to LDA and PAM and performs better than hLDA for most topic configurations. hPAM1 and PAM show better performance for more super-topics, while the performance of hPAM2 remains fairly stable for different numbers of super-topics.

Mimno et al. (2007) further measure the ability of different topic models (i.e. hLDA, PAM and hPAM) to discover the hierarchical structure of a corpus by focusing on the top-level branches of the hierarchy generated by each model and calculating the mutual information between these topics and human-generated categories corresponding to the journals in which corpus articles was published in. The characteristic of a model that effectively distinguishes hierarchical structure of a corpus is that it should at minimum be capable of dividing the corpus into its main topical elements. The aim is to measure how well each model can predict the journal corresponding to a word that is assigned to a given super-topic, and vice versa. They found that hPAM1 consistently performs better than the other models at predicting the journal to which a given word belongs to given its topic assignment. The performance of hPAM2 and hLDA are close, but hPAM2 having superior performance at its best topic configuration. Surprisingly, the best results of non-hierarchical LDA are close to the results of hPAM using small numbers of super-topics. The performance of PAM is extremely poor with the super-topic/journal mutual information measure being nearly statistically independent. In summary, Mimno et al. (2007) reports a 1.1%

improvement in the empirical log likelihood of hPAM over hLDA and a five-time improvement in super-topic/journal mutual information. In summary, they found that the super-topics generated by hPAM tend to be readily interpretable and that super-topic distributions over sub-topics are inclined to be very non-uniform, making them sparse and interpretable.

Compared to hLDA, hPAM is significantly faster. The hPAM model is also fairly robust to bad parameterisations. For example, in the case where hPAM is given too many topics it can simply refrain from assigning any words to one or more sub-topics thereby effectively reverting to a flat LDA model. Therefore, a badly parameterised hPAM model may not necessarily render a good topic hierarchy, but it will not completely fail in the task of creating a topic model.

16.12 Concept-Topic Model

Probabilistic topic models render a general data-driven framework for automated discovery of high-level knowledge from large corpora of text documents (Chemudugunta, Holloway, Smyth & Steyvers, 2008). Although topic models can discover a wide variety of topics in a data set, the interpretability of such topics are not always optimal. On the other hand, human-defined concepts are inclined to be semantically richer since the words defining such concepts are usually selected carefully, but such concepts usually do not exhaustively cover the themes present in the data set. Data-derived topics further have the advantage of being aligned to the themes of the specific corpus they are learnt from. Human-defined concepts are created for example during the construction of thesauri and ontologies where prior human knowledge and judgement are used to associate a small set of significant words with each concept. The definition of human-defined concepts may also include concept names and relations between different concepts. Thus a concept in its most basic form is represented by a set of words and a name (e.g. “Tropical Fruit”). Concepts are often ordered in a tree-like structure, e.g. in a taxonomy. Human-defined concepts and machine-learnt topics represent similar information in different ways.

Chemudugunta, Holloway, Smyth and Steyvers (2008) introduces the Concept-Topic Model that combines data-driven topics with human-defined semantic concepts to automatically annotate documents. The Concept-Topic Model is an extension of the standard topic model where C concepts are added to the topic model’s T topics resulting in an effective set of $T + C$ topics for each document d. Since human-defined concepts are essentially a set of words, such concepts only renders a membership function of words (i.e. whether a given word is a member of a given concept or not). Concepts can easily be integrated into the topic model by converting them to “topics” by representing them using probability distributions over their associated word sets. A concept c may therefore be represented by a multinomial distribution so that $p(w|c) = 1$ in the case where the word w is part of concept c’s word set and $p(w|c) = 0$ where w is not part of the respective concept’s word set. As result a document is now represented as a distribution over concepts and topics.

In the generative process for a document corpus under the Concept-Topic Model, the only known elements are the words in the corpus and the membership of words in the human-defined concepts. Given the words in the corpus, the inference problem includes estimating the respective topic word distributions, the respective concept word distributions as well as the distribution over topics and concepts. The Gibbs sampling method used for performing inference for topic models is adjusted to perform inference for the concept-topic model (Chemudugunta, Holloway, Smyth & Steyvers, 2008). Chemudugunta, Holloway, Smyth and Steyvers (2008) reports a significant improvement in the predictive performance of the Concept-Topic Model when compared to the standard topic model (e.g. LDA) in terms of perplexity.

16.13 Hierarchical Concept-Topic Model (HCTM)

Chemudugunta, Smyth & Steyvers (2008) extends the Concept-Topic Model framework to the Hierarchical Concept-Topic Model (HCTM) that combines a hierarchy of human-defined semantic concepts with probabilistic topics to merge the benefits of these two approaches.

Similar to the concept-topic model introduced earlier, HCTM also has T topics and C concepts resulting in a possible $T + C$ “topics” for each document. For each document d, a “switch” distribution $p(x|d)$ is used to determine whether a word should be generated via the topic route or concept route. A binary “switch variable” x is associated with each word in the corpus vocabulary. Where $x = 0$, the standard topic model mechanism is used to generate the word. More specifically, a topic t is selected from a document-specific mixture of topics $p(t|d)$ and subsequently a word is generated from topic t’s associated word distribution. Where $x = 1$, a word is generated from one of the C concepts in the concept tree. This is achieved by associating a document-specific multinomial distribution with dimensionality $N_c + 1$ with N_c represents the number of children of concept node c. Using this distribution, it is possible to traverse the concept tree and exit at any one of the tree’s C nodes. For instance, when located at concept node c, there are N_c child concepts to select from and an additional option to select an “exit” child to exit the concept tree. The concept tree is traversed starting at the root node by selecting a child node from the children of the root node. This process is repeated until an exit node is chosen and the word is generated from the parent of the corresponding exit node. There are exactly C different ways to select a path and exit the tree in a concept tree having C nodes.

In effect, HTCM portrays a document as a weighted combination of mixtures of T topics and C paths through the concept tree. HCTM is fairly flexible since it can accommodate any directed acyclic concept graph. The word generation mechanism via the concept route in HCTM is somewhat similar to that of the hPAM2 model described earlier.

Chemudugunta, Smyth & Steyvers (2008) uses two corpora and two concept sets (containing 2000 and 10 000 concepts respectively) to test the predictive performance of HCTM compared to the Concept-Topic Model and the standard topic model (e.g. LDA) by comparing the perplexity of unseen words in test documents. They conclude that the perplexity of the Concept-Topic Model and HCTM, which includes concepts and a concept-hierarchy, shows a substantial improvement in terms of perplexity compared to the other purely data-driven models. The difference becomes even more significant when the model is trained on one genre of documents and tested on documents of another genre. This suggests that models using concepts are robust and can handle noise. The standard topic model is completely data-driven as it does not include any human knowledge and portrays lower levels of robustness. When comparing CTM and HCTM where no words are generated via the topic route (i.e. where $T = 0$), the performance of HCTM is consistently better than that of the Concept-Topic Model. This effect may be explained by the inclusion of correlations between child concepts in the case of HCTM.

17. Appendix C - Characteristics of Software Used for Investigations

	CIRP 1	Enterprise Engineering Group	CIRP 2	SAJWR
Completion Date	Jan 2007	Aug 2008	Jan 2009	Aug 2009
Software Used	Custom Prototype	CAT 0.1.0.0.6	CAT 0.8.10.26	Custom Prototype
Topic Modelling Technique	LDA with Expectation Maximisation	LDA with Expectation Maximisation	LDA with Gibbs Sampling	Concept-Topic Model
Multi-core CPU Support?	N	N	Y	N
Supported File Types	pdf + txt	pdf + doc + ppt + txt	pdf + MS Office	pdf + MS Office
Output Formats	Text files	Excel only	Excel + database	Excel only
Regular Expression-based extraction?	Y	Y	N	N
Collocations Returned	Y	Y	N	N
Level of n-gram support	unigrams only	uni-, bi- and trigrams	uni-, bi- and trigrams	unigrams only
Automatic document **title** extraction?	Y	N	N	N
Automatic document **author** extraction?	Y	N	N	N
Automatic document **abstract** extraction?	Y	N	N	N
Topic-Word probabilities part of output?	N	N	Y	Y
Topic-Document probabilities part of output?	Y	Y	Y	Y
Document-Word probabilities part of output?	N	N	Y	N
Topic-Topic affinities automatically calculated?	N	Manually calculated using correlation	Y	N
Document-Document affinities automatically calculated?	N	N	Y	N
Word-Word affinities automatically calculated?	N	N	Y	N
Number of words shown per topic	20	40	100	N/A
Catering for topic hierarchy?	N	Manually calculated using correlation	Manually calculated using correlation	N/A
Suggestion for number of topics?	N	N	N	N/A
Technique used for determining topic similarity	not calculated	Correlation (done manually using Excel output)	iradius (Automatic)	N/A
Automatic Vocabulary List?	N	N	Y	N
Mechanism for capturing user feedback about results?	N	N	Y	N
Catering for interactive result visualisation?	N	N	Y	N

Table 107: Characteristics of the software used for the various case studies

18. Appendix D - Knowledge Representation Technologies

18.1 Controlled Vocabularies

A controlled vocabulary regulates the usage of terms in a given domain (e.g. an organisation) by restricting terms used to the set of terms specified in the controlled vocabulary. A controlled vocabulary may include 'preferred terms' which are linked to and should be used instead of other (non-preferred) terms. Controlled vocabularies can be used by both humans and machines and are very basic in terms of representing meaning. (Inmon et al., 2008; Manning et al., 2008)

18.2 Dictionaries and Glossaries

A fundamental component of well-captured semantics is a good dictionary containing definitions of business terms that are 'set complete' in order to make explicit what constitutes membership to the specific set (i.e. the specific term defined). It is a challenge for businesspeople to specify thorough, set complete definitions since they are generally not trained to be precise in spite of the ambiguous nature of language. Refer to section 7.4 for a discussion about the characteristics of a good definition. Dictionaries and glossaries are part of the core of any business metadata solution since they are mainly human orientated. (Inmon et al., 2008)

18.3 Taxonomy

As mentioned earlier in section 7.5, taxonomy is a hierarchical classification scheme that organises a set of terms by means of parent-child or broad term-narrow term relationships. In terms of semantic capability, taxonomy is more advanced than dictionaries or controlled vocabularies since it also adds relationships (in the form of hierarchy) in addition to a list of definitions. Taxonomies can be used to facilitate enterprise search and to classify things like products for instance. Taxonomies can be used by both humans and computers to organise information and is considered as a carrier of business metadata. (Inmon et al., 2008)

18.4 Thesauri

Whereas taxonomy is generally restricted to hierarchical (i.e. vertical) relationships between terms, thesauri are more focused on horizontal relationships between words (e.g. synonyms). Thesauri are well known to authors and it are used among other things to locate other terms having the same meaning as a given term to improve the quality of writing. Thesauri, more shaped for use by humans, can be used by computers in a limited way. (Inmon et al., 2008)

18.5 Entity-Relationship Models

The entity-relationship (ER) model is used to define entities (concepts) in terms of the various attributes (characteristics) that define such entities. In addition, it captures the relationships between entities. It is an excellent mechanism for representing relationship semantics due to its rich and robust means of semantic expression, which includes cardinality and optionality. In data modelling, three types of ER models exist, namely:

- Conceptual ER models – focus on the primary concepts required to do business
- Logical ER models
- Physical ER models – include physical implementation details of concepts

An ER diagram (ERD), the visualisation of an ER model, represents business rules by means of two implied unambiguous relationship sentences associated with a given inter-connected entity pair. Different notations, for example Barker notation, Crow's Foot notation and Bachman notation, can be used to create ER models.

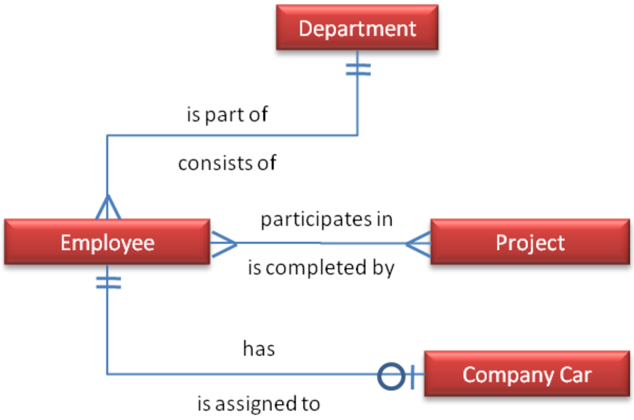


Figure 67: An example of an ER diagram (ERD)

In the example of an ERD shown in Figure 67, the following relationships are implied:

- An employee is part of one and only one department
- A department can consist of many employees
- An employee can participate in many projects
- A project can be completed by many employees
- An employee can have zero or one company cars
- A company car is assigned to one and only one employee

Well-defined ER models can facilitate communication between non-technical businesspeople and are considered to represent business metadata in the form of relationships between concepts in a graphical and linguistic way. ER models are also used to generate technical metadata (e.g. database schemas) and communicate technical metadata of data elements to technical people (e.g. software developers). Due to the capability of ER models to express richer relationships than taxonomies and thesauri, its semantic ability are considered superior to that of taxonomy and thesauri. (Inmon et al., 2008)

18.6 Conceptual Models

A concept differs from a business term as more than one term can exist for a given concept and vice versa implying that many-to-many relationships may exist between terms and concepts. A conceptual model represents individual concepts and allows data elements to be mapped to concepts to illustrate the system location of different business concepts. Conceptual models can therefore express individual concepts and the relationships between concepts independent of how such concepts are implemented in systems. A semantic model can be described as a conceptual model with a more rigorous manifestation of meanings for concepts to resolve possible ambiguities. ER models can be used to formally represent conceptual models.

Conceptual models therefore represent another mechanism to provide business metadata to businesspeople. The fact that a single concept may have more than one associated data element implies that a conceptual model differs from a data model. Also, data models are examples of technical metadata whereas conceptual models are mostly business metadata orientated. (Inmon et al., 2008)

18.7 RDF and OWL

Modelling languages like Resource Description Framework (RDF) and Web Ontology Language (OWL) express concepts and relationships in a format that enables processing by computers. Constructs in RDF are called “triples” because they consist of concept-relationship-concept combinations. These two languages in combination are considered by some to be the building blocks of the semantic web. These languages, although they can express rich relationships between entities, do not directly represent business metadata since they are not interpretable by businesspeople since they are expressed in XML syntax. (Inmon et al., 2008)

18.8 Topic Maps and Concept Maps

Topic maps constitute topics (i.e. concepts or entities), the associations (relationships) between topics and occurrences (information about a topic). The standard ISO/EIC 13250:2003 describes and regulates topic maps. A concept map can be considered as a variant of a topic map. Topic maps and concept maps can be used to represent RDF triples in a form that is understandable by humans. Since topics maps and concepts maps are tuned for human interpretability they are considered as representations of business metadata. (Inmon et al., 2008)

18.9 UML

Unified Modelling Language (UML) is a standard set of models developed by the Object Management Group that serves as a graphical notation system for object-oriented analysis and design using class hierarchies. UML's ability to capture semantics is limited to hierarchical design and therefore Inmon et al. (2008) argues that it should be lower in the spectrum depicted in Figure 29. UML models are intended for technical persons making it difficult for businesspeople to interpret them. The fact that UML models do not express semantics to businesspeople means that they are generally not considered to be carriers of business metadata. (Inmon et al., 2008)

18.10 Description Logics and Other Forms of Logic

Logic enables the ability to reason and thereby creating new knowledge from existing knowledge. Using logic, relationships can be derived in the corporate knowledge base. For instance, if two employees have the same surname, shares the same address and have different genders one could derive that they are possibly married. Description Logics (DL) represents a set of knowledge representation languages that use formal logic and that can be translated to first-order logic - a system of deductive reasoning. (Inmon et al., 2008)

18.11 Ontology

Ontology stems from philosophy and in that context is defined as the metaphysical study of the nature of being and existence. In the context of computer science and information management it can be defined as:

"*Ontology is a formal specification of concepts of the domain of interest."* (Gruber, 1993)

and

"*Ontology provides a reference domain model that both human and software can refer to for various purposes such as search, browsing, interoperability, integration, and configuration.*" (McGuinness, 2001)

In other words, an ontology is essentially a model representing a collection of concepts within a domain as well as the relationships between such concepts. Ontologies are used to reason about objects in a specific domain and have applications in artificial intelligence, the semantic web, software engineering and knowledge engineering. An ontology is a way of representing knowledge, which includes relationships and classifications, and can be represented as a graph or network diagram. Ontologies can also be represented in topic maps, ER models and concepts maps. Generally, ontologies allows for more detailed relationships to be expressed compared to ER models and UML (although an ER model is strictly a kind of ontology). OWL, introduced earlier, is a kind of ontology language having many inherent relationship types that contribute to its inferencing capability. (Inmon et al., 2008)

19. Appendix E - Preamble to Knowledge Management

19.1 Types of Knowledge

When structuring knowledge in terms of articulability it can be divided into two types, namely **explicit knowledge** and **tacit knowledge**. Explicit knowledge is consciously understood and can be articulated implying that the 'knower' is aware of such knowledge and can further converse about it. In the case of tacit knowledge, the 'knower' is not necessarily aware of possessing such knowledge and the capture of such knowledge, if possible at all, can only be achieved with difficulty using special interviewing and observation techniques. 'Know-how' or procedural knowledge - the ability to simply knowing how to accomplish a task or detect when something seems suspect - without being able to explain it, also forms part of tacit knowledge Tacit knowledge also includes experiential knowledge. Since tacit knowledge is difficult to articulate and therefore to transform to business metadata, the relationship between KM and business metadata becomes less pronounced when it comes to tacit knowledge. (Sammer, 2003)

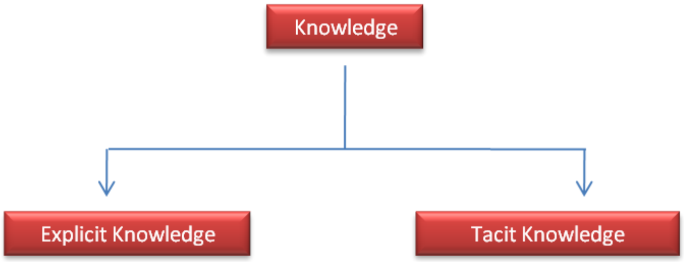


Figure 68: Types of knowledge

Another way of distinguishing different types of knowledge is in terms of the knowledge holder. Sammer (2003) distinguishes between **individual knowledge** and **collective knowledge**.

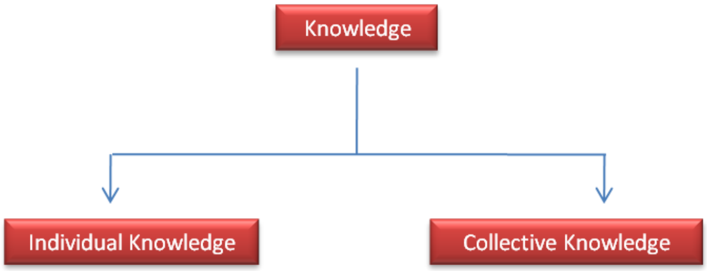


Figure 69: More types of knowledge

Individual knowledge is held by a person, is not dependent on a particular context and is controlled by the individual in question. Collective knowledge on the other hand is knowledge that is applicable in a particular environment (e.g. a discipline, organisation, hobby, etc.). It may be a combination of the individual knowledge of several persons resulting in synergy. It may also be knowledge common to all members of a group (e.g. everybody in an organisation knows who to approach when they are experiencing problems with their computer).

19.2 The Knowledge Creation Process

Nonaka et al. (1995) proposed the SECI model of knowledge creation. In this model a cycle of knowledge creation results from conversion processes between tacit and explicit knowledge. This model distinguishes the following four knowledge conversion processes:

- **Socialisation** – transferring tacit knowledge from the mind of the knower to the minds of other individuals creating sympathised knowledge as result
- **Externalisation** – transforms tacit knowledge to explicit knowledge and creates conceptual knowledge as result
- **Combination** – merging different instances of explicit knowledge to form new explicit knowledge
- **Internalisation** – transforms explicit knowledge to tacit knowledge and creating operational knowledge as result

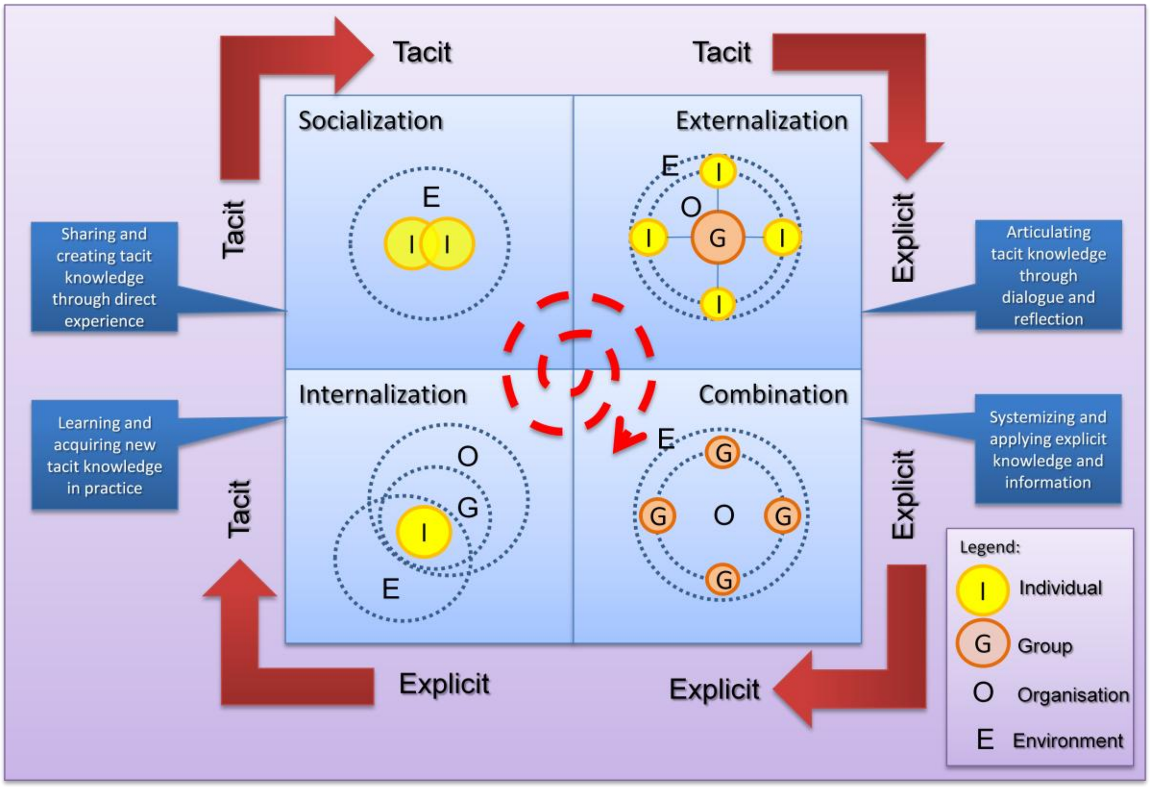

Tacit
Tacit
Socialization
Externalization
Tacit
Explicit
Sharing and creating tacit knowledge through direct experience
Articulating tacit knowledge through dialogue and reflection
E
I I
E
O
I G I
I
I
Internalization
Combination
Learning and acquiring new tacit knowledge in practice
Systemizing and applying explicit knowledge and information
O
G
I
E
E
G
G O G
G
Tacit
Explicit
Legend:
I Individual
G Group
O Organisation
E Environment
Explicit
Explicit

Figure 70: SECI model of Knowledge Creation[109]

[109] This figure was adapted from Nonaka et al. (1995) by C. Schutte.

The knowledge creation cycle initially starts with the knowledge socialisation process where tacit knowledge is synthesised by sharing experiences between individuals, developing shared mental models and technical skills. This process is mainly experiential and connects people through their tacit knowledge sets (Iles, Yolles & Altman, 2001). Subsequently, the externalisation process commences where conceptual knowledge is created by making tacit knowledge explicit. This happens by means of knowledge articulation involving a communication process that uses language in dialogue as well as collective reflection and discussion. Next, the combination process commences where explicit knowledge is integrated, expanded, combined and categorised to create new explicit knowledge. The internalisation process is the last process in the knowledge cycle and is essentially a learning process that involves behavioural development of operational knowledge from explicit knowledge (e.g. manuals, documented examples, etc.). The SECI cycle then starts once more with knowledge socialisation as the spiral in Figure 70 suggests.

19.3 Knowledge Management and Tacit Knowledge

Business metadata is created by formulating knowledge and storing it as data in a database management system (DBMS), wiki, thesaurus, dictionary or ER model for dissemination to business users (Inmon et al., 2008). One of the myths concerning knowledge sharing is that technology can substitute face-to-face interactions according to Dixon (2000). Tacit knowledge exists in the minds of people and therefore cannot be captured as business metadata except for when it is converted to explicit knowledge. Therefore the tacit knowledge aspect of KM cannot be codified as business metadata. Arguably one of the main challenges for KM and the endeavour of capturing business metadata is providing support (and incentives) to expert employees in expressing their tacit knowledge to capture it as business metadata where possible.

In some cases just the act of socialisation can shake loose tacit knowledge, make it concrete, and develop the required artefacts for addition to the corporate knowledge base. ICT systems like groupware (e.g. Lotus Notes and Microsoft SharePoint) have the ability to create an audit trial of the socialisation process to help one track the source of ideas and their subsequent development.

Although ICT systems can assist with transferring tacit knowledge, some tacit knowledge necessitates socialisation in the form of face-to-face interaction (e.g. dialogue, mentoring, etc.). An underutilised potential source of business metadata is the notes made by junior employees while undergoing mentoring. The problem with notes is that they are often haphazard, too cryptic, unintelligible and difficult to organise. One potential way of exploiting such notes is for the mentor to evaluate the trainee’s notes to identify and elaborate on important points to jointly develop knowledge artefacts for deposition in the corporate knowledge base. By elevating the skill of adequate note-taking may greatly facilitate business metadata management. It is crucial to make such artefacts easily accessible and creating awareness about the location of such artefacts to encourage knowledge reuse. New employees can for instance be introduced to the corporate knowledge base as part of their induction course. Another knowledge socialisation technique, called serial transfer, is introduced in Dixon (2000). Serial transfer is a process which shifts the unique knowledge (about a given activity, idea, issue, scenario, etc.) developed by individuals into a group setting with the goal of synthesising and integrating such knowledge in such a way that it is comprehensible to the entire team. Serial transfer typically occurs in the form of a debrief meeting which may happen face-to-face or online. It is more than just sharing knowledge – it is about individuals using what others have said to re-evaluate their own understanding of the situation.

Similarly, Sammer (2003) discusses the good practise of closing every project with a 'Lessons Learned Workshop' to discuss and transfer individual experience as well as to improve the insights of individuals and the group. This process of building collective knowledge iteratively based on the knowledge of individuals evokes the rethink of causes and effects that leads to the identification of discrepancies in what was perceived and subsequently creates new generalisations which may direct future actions. Knowledge socialisation – the way in which people share knowledge about a topic – therefore improves the quality and thus the value of knowledge itself (i.e. knowledge evolution) in addition to making the participants smarter.

The field of KM is centred around the environment of knowledge sharing and knowledge transfer methods with strong emphasis on face-to-face interactions. Innovation can result from understanding tacit knowledge transfer according to Von Krogh, Ichijo and Nonaka (2000). It may not be possible to articulate tacit knowledge itself, but the factors that enable its transfer may be captured and constitute business metadata (Inmon et al., 2008).