2009 International Conference on Computational Science and Engineering

A Decision Support System for Institutional Research Management in Higher Education

Data Mining to determine Research Focus, Intensity and Synergy

Kobus Ehlers, Malan Joubert, Johann Kinghorn, Arnold van Zyl
Centre for Knowledge Dynamics and Decision Making
Stellenbosch University, South Africa
kobuse@sun.ac.za

Research orientated universities face a conundrum. On the one hand research areas are increasing and expanding; on the other research resources are – in relative terms – diminishing. As a consequence a more active approach to shaping the institutional research profile is called for. This presupposes a clear view of research activities at the institution, but this is notoriously lacking, universities being what they are. This paper reports on a DSS which was developed for research management at Stellenbosch University, South Africa. Within the context of a rather unique national management system for higher education, a data mining approach was developed and tested. A modified ontology model was built to bridge the inherent ambiguity when clustering outputs across different academic disciplines. The results enabled research managers to correlate (in a meaningful way) strategic goals with actual output. The outcome points to a number of further possibilities for the application of this system.

decision support system; data mining; university research; research management; ontology; folksonomy; taxonomy;

I. INTRODUCTION

THE advent of the modern university has meant that even the most idealistic of organisations has had to reconcile the entrenched principles of academic freedom with a focus on managed strategic planning [1], [2]. This of course meant that researchers no longer had the prerogative to pursue only their own whimsical interests, but rather that the research conducted by a specific organisation is loosely grouped around certain specified areas.

Concurrently research spending has become progressively more limited, leading to a much greater focus on the effective spending of funds. The era of large scale institutional or government research has also largely ended, with smaller, focused research initiatives becoming the mode *du jour*. This model has ushered in an era of university-government-industry co-operation on a level never seen before (the so-called *triple helix*) [3].

Modern institutions also face specific challenges with regards to corporate governance and financial accountability regarding their research projects and goals. Institutions now need to declare their spending and rationale for allocating these scarce funds. This is especially true with the growing trend of seeing universities as national assets belonging to all citizens – at least within the South African higher education sphere. (For an interesting discussion of the problematic, read Boulton & Lucas [4]).

The popularity of scientific management, combined with the abovementioned factors, has created a very definite need for Enterprise Research Management Systems (ERMS) or Research Information Management Systems (RIMS) to monitor the different projects initiated by researchers within a specific organisation.

One goal of these tools is to provide managers with an institutional level view of completed, ongoing and planned research. Another is to correlate these with an overview of the research funding and deliverables.

These issues are specifically relevant within the South African context where, being a developing economy, resources are relatively scarce. It is against this background that a project was undertaken at SU to construct a DSS with the aim of improving the efficiency of the use of available resources.

This paper will look at the use of data mining techniques for research management by evaluating a first phase pilot project executed at Stellenbosch University, South Africa, during July - November 2008. Specific issues with regards to data mining of this type of data are discussed and complemented by the definition of a conceptual model to address some of the difficulties encountered.

II. SOURCING THE DATA

Significant challenges faced the researchers in trying to identify and source relevant data to complete this analysis. Several potential sources existed, but novel approaches had to be developed to integrate these disparate sources into a coherent dataset and to construct a suitable model to integrate research outputs from several different disciplines or 'research areas'.

A. Data Mining vs. Data Elicitation

Ideally institutions would apply an enterprise level (or even national) RIMS to manage the complete research process from inception right through to publication and reporting. However, the invasive nature and substantial costs associated with implementing such a RIMS has meant that a large portion of universities and other research institutions currently have no such system in place. Current internal systems provided no method for mapping or analysing reported research outputs or assisting decision making in this regard[1].

[1] A national RIMS (*InfoEd*) is currently being established nationally. This study is being used to assist with scoping.

978-0-7695-3823-5/09 $26.00 © 2009 IEEE
DOI 10.1109/CSE.2009.222

This effectively means that research is conducted in a decentralised fashion by individual researchers or departments. The traditional notion of an academic institution enjoying absolute academic and research freedom has also meant that the actual research focus (versus the declared institutional focus) at universities is largely unknown, even by the institutions themselves – perhaps with the exception of some high profile projects.

The obvious way to address this lack of information would be to proceed on an institution-wide audit of research activities at the university. One can envisage that such an audit would include interviews and comprehensive surveys of all relevant researchers.

At the institution in this study (Stellenbosch University, South Africa) this type of audit had already been completed in 2006 as part of a national quality audit (HEQC Audit [5]). That study also expanded on a research output specific audit conducted by CREST (Centre for Research in Science and Technology) [6] at the same institution.

These audits applied a combination of manual data analysis, interviews and surveys to gain some understanding of the research currently being conducted at the institution. However, it became clear that continuous monitoring, rather than explicit data elicitation, was needed for three reasons:

i.) Since funding models are built around subsidies determined by published research outputs, the focus should be on completed, reported research, rather than on ongoing projects or projects not producing accredited output.

ii.) Researchers already have to report published research for audit and funding purposes, so asking them to repeat this administrative action will provide little additional benefit in addition to that which can be gained from the available data.

iii.) Surveys and interviews are disruptive and invasive and generally not well received by researchers – especially if they have reported the same data to the institution previously.

These observations compelled the university to evaluate their existing data sources and try and implement a real time analysis system based on the data available, rather than commissioning continuous snap-shot audits.

One should also keep in mind that commissioned human audits are relatively expensive and take considerable time to execute, thus rendering them inappropriate for day-to-day operational decision making. That being said, these audits can certainly provide valuable insight when employed in a mixed quantitative/qualitative mode.

B. Available data sources

Fortunately a vast amount of information is already captured by South African High Education Management Information System (HEMIS) as part of the normal reporting process. This system is utilised to aggregate country-wide information on students, undergraduate courses and all subsidy-earning research outputs at accredited South African universities.

HEMIS is a rather unique system, integrating several different institutional parameters into one centralised, consolidated database. This type of approach makes sense specifically within the South African educational environment where universities are public, rather than private, institutions.

For the purposes of this project, the HEMIS reporting system provided a very accurate, objective record of all published research conducted at the institution. This data proved invaluable in mapping the research activities of the university and was utilised as a baseline for this project. The dataset was expanded by adding additional rich sources.

TABLE I. IDENTIFIED DATA SOURCES

Name	Description	Integrated
HEMIS Reporting	Department of Education reporting system for research	Yes
E-thesis	Electronic repository of all masters and doctoral dissertations	Yes
NRF/SARchI chairs	Elite research chairs (and accompanying projects)	Yes
OSP Flagships	Explicitly identified institutional flagship research projects (special status)	Yes
HR Data	Human Resources Employment Data	Yes
Budget proposals	Departmental research budget proposals	No
Research reports	Departmental annual reports on completed and ongoing research	No
Journal citations	Cited works by employed/associated researchers	No
Internet crawling	Other works (not accredited) by employed/associated researchers.	No

Additional research hotspots were identified by looking at the content of research projects specifically highlighted as high profile centres of excellence. These include research chairs awarded by the Department of Science and Technology and the National Research Foundation.

As part of the institutional Overarching Strategic Plan (OSP) several research areas were identified and specifically funded to act as flagship projects and concentrate resources and build capacity. These were considered in the mapping of the research landscape.

Another source for completed research is the recently developed digital repository of all accepted masters and doctoral dissertations (the majority of which is freely accessible through the World Wide Web). By combining these dissertations and information about the supervisors concerned, interesting comparisons between graduate research and post-graduate projects can also be made.

Human Resources (HR) data was used to provide a mapping between staff members and their associated academic departments in cases where this information was not specifically reported. An accurate list of current and past researchers is also needed when mining external data sources (such as journal citations and the internet).

As noted in table 1, several additional data sources have been identified, but were not data mined due to resource constraints in this pilot phase of the project. These additional sources all require more advanced content or semantic analysis. Some of the most promising sources included a cross-matching of journal citations with known researchers and their corresponding fields of expertise.

Internally most departments also produce an annual research report dealing with their specific areas of focus. Analysing these reports (and accompanying budget proposals) could provide important insights on active research projects that do not necessarily result in articles or papers published in accredited journals, but are still important in terms of establishing the research focus of the specific unit.

III. Defining the Institutional Brief

A. Immediate requirements

The proposed solution had to cope with several constraints. These included a very short development timeframe (six months), no direct data elicitation, a limited budget, and a requirement for a scalable model that should adapt, based on the available data sources.

The university also expressed specific requirements as to which data should be exposed through this study. For the first phase the focus was on a discovery of the conducted research and a comparison of this research with the stated research plan of the institution. Further phases would include more complicated analyses and incorporate budget and funding flows.

For this pilot phase the dimensions identified were

i.) Research focus – which topics of enquiry were currently being entertained at the institution

ii.) Research intensity – how many outputs relating to a specific research focus are being produced

iii.) Research synergy – how does the research landscape look with regards to the cooperation between researchers and departments

These different dimensions could then be plotted to create a 'topographical map' of the research space at Stellenbosch University. After such a map has been constructed several overlays could be utilised to determine the alignment between normative goals (e.g. stated institutional research focus, OSP, budget allocation, human resource deployment) and actual research outputs as identified by the mining process. No current system could provide this view.

The third dimension (research synergy) is especially important to identify 'orphan research' – idiosyncratic or novel project 'islands' that do not easily fit into the specific research environment. This dimension is also used to locate areas of cooperation between specific researchers or entire departments. Additionally researchers themselves may be able to use this mapping to find other researchers at the same institution who may have research interests or projects overlapping with their own.

B. Mode of presentation

The results of this project had to be available in two distinct modes to enable successful utilisation of the findings.

In the first case this 'topographical map' was required to be presented as a flat two-dimensional plot that enabled research managers to look at the visualised data and make certain high-level decision based on the findings. The focus here is in identifying general research patterns and macro trends with regards to research (some of these plots are shown in figures 2, 3 and 4).

This level of view is also suitable for making decisions on resource allocation and identifying key departments or units responsible for stimulating clusters of research.

The second case involves a more interactive mode. Considering the rich nature of the data available, it was required that the users of this system should be able to spatially manipulate and interrogate the constructed visual representation so as to enable them to identify individual researchers and research outputs and investigate their relationships. This view is essential to identify researchers who act as 'research catalysts' or 'research hubs' in bringing together different disciplines. The interactive mode also enables researchers themselves to browse the research space at the institution and look for potential opportunities for cooperation.

Quite often individual researchers are unaware that similar research is being conducted in a different department. With the increase in trans-disciplinary research this situation will only proliferate.

IV. Representing the Data

The starting point for this system was the construction of a modified ontology representing the landscape of research.

Unfortunately our research has shown that even formal systems for classifying research have proved relatively unsuccessful. One inherent problem is terminological confusion. Since the system will have to do textual semantic analysis on research outputs from various different disciplines of research, it becomes crucially important to avoid ambiguity when the same term (the specific word) is used with different meanings by different environments. Therefore, the construction of a 'scaffolding' or 'skeleton' becomes one of the most essential elements to building a useable system.

For the purposes of this project, we expanded on Gruber's understanding of ontologies within the IS context [7]: "*...an ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents.*"

Several different elements were mapped at once. To build a purposive ontology we had to consider the research outputs, research fields/disciplines, semantic context and researchers (agents). Of course we also had to keep track of the relationships between these different elements.

It is important to bear in mind that the levels of analysis are the research output, the researcher and the research unit. All of these levels need to be kept linked constantly.

A. Envisioning the model

Several factors make the traditional understanding of ontologies problematic for application in this case. Objects of different orders/categories, terminology without clear definition or context and meshed methodologies are all complicating issues in this analysis.

The goal was *not* to build some kind of encompassing ontology of all knowledge present at the institution. The goal was

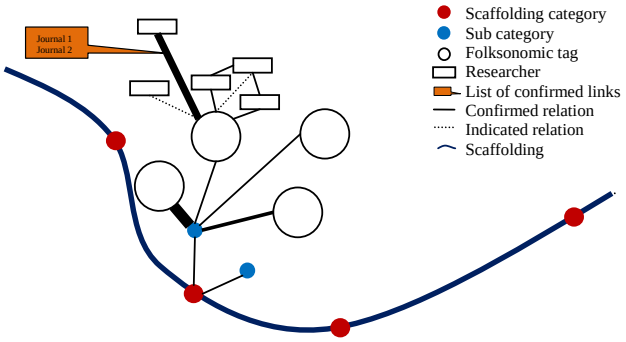

FIGURE 1 CONCEPTUAL MODEL

simply to try and construct the landscape of research activities that can be detected by looking at published research outputs.

In developing the eventual model for representing this data, it became quite clear that a taxonomy-based mapping of outputs would not be sufficient. Both through semantic analysis and user-tagging, data could be enriched and represented in a much more complex form than traditional hierarchical taxonomies would allow for.

As a consequence we attempted to mesh two approaches often juxtaposed or seen as dichotomous. We argued that the traditional taxonomy could be combined with a more contemporary folksonomy approach [8] [9]. This would allow users of the system to assign their own tags and categories to existing data[2].

The order and rigidity of formal hierarchies are needed to avoid semantic confusion across different disciplines and areas of study. These categories and 1-to-n relationships, however, are not sufficient to cope with the richness of the problem.

The proposed approach therefore constructed an ontology by linking relationships in a non-hierarchical fashion while scaffolding the data space with a more formal hierarchical taxonomy. In fact the taxonomy provides a type of skeleton on which these different keywords/tags can be 'hooked' (see a schematic representation in figure 1).

B. *Constructing the scaffolding*

Several options were evaluated for providing the overarching categories of possible study. Although many different systems exist, they all require significant translation from the source format to their coding system. In most cases one would have to manually assign a content category to each research output – this would be totally unfeasible.

The university already codes all (reported) research outputs in terms of the CESM (Classification of Education Subject Matter) categories devised by the South African Department of Education (DoE) (a full list of these categories can be obtained at [11])[3] saving significant amounts of labour. All research reported have to be divided into one of these twenty-two categories. Refined, second-level subcategories are also available for each top-level CESM-code.

Although the system is not particularly granular, it does provide enough structure to serve the role of scaffolding the landscape. Being a standardised taxonomy, it also provides data portability and can utilise existing processes for assigning a code to specific research or teaching. This proved to save a large amount of time and effort.

By contextualising every research output within its CESM-field it avoided possible terminological ambiguity and also assisted the system in learning the subject matter at hand.

This approach also saved a considerable amount of time when looking at the performance of departments or specific researchers, since both these units are also assigned a primary CESM code. This meant that accurate guesses about the content of an article could be made by simple cross-referencing the author and HR data – even if the article itself did not have an assigned CESM code.

A mapping of research intensity and synergy between specific CESM codes could also be compared with the OSP and declared institutional focus by simply assigning these goals specific CESM codes.

The specific weighting assigned to the different types of research output (conference proceedings, journal articles, books, etc) was determined by the subsidy model currently in use at the university. These weightings can, however, easily be modified by the users in accordance with their own expert judgements.

C. *Populating the model*

The first step was to create and populate a conceptual model (indicated in figure 1). This model bears some resemblance to the actual visualised system output, but being a conceptual model, it of course differs in some aspects.

A single populated model was used to represent all the available data. Different views and filters were constructed for specific outputs.

Each researcher and topic is represented as a node in the system. Spatial arrangement in the X and Y axis indicates the context and discipline on a curve ("backbone") anchored by the points specified in the taxonomy (CESM-categories).

Relationships between the different researchers and between researchers and the topics are defined in terms of research outputs, weighted according to expert judgement and specified in the ontology. In this graphical representation the thickness of links will indicate the extent of the relation between nodes. User-specified connections are also indicated as weak links.

These relationships are continually updated and dynamic. Nodes maintain relative positions but can be arranged and traversed as needed to visualize specific areas.

V. DATA ANALYSIS

Data were loaded from all the available sources and, for the pilot phase, only metadata analysis was performed. Unfortunately the HEMIS system does not currently expose the full text of the

[2] This approach is also argued and demonstrated in [10].

[3] The CESM system was devised in South Africa in the 1970's to form a set of 22 categories, building on research conducted in the USA in this same period.

articles in question, making a semantic analysis of the research outputs unfeasible.

In cases where research outputs were coded using the CESM system, those codes were used to sort and cluster the data. If no code was available, a deduction was made based on the researcher and department's coding in the HR database.

A. Analysis

The data were parsed using a custom J2SE application to preprocess the data and create the GDF output files for use in GUESS. Data visualization was performed using the GUESS visualization toolkit, an Exploratory Data Analysis (EDA) tool. [12][13]. To enable the clustering of results, a Fruchterman-Reingold Algorithm [14] was applied to the analyzed data.

B. Findings

Although the original intention of this study was to analyse the content of research outputs, the first discovery was that the current South African reporting standards simply do not require that the output itself be preserved as part of the reporting process. This creates the unfortunate situation where the text of articles, reports and dissertations cannot be easily mined to evaluate their content.

The initial study therefore had to focus on the available metadata. Regardless of this set-back it became quite clear that some very useful results could still be produced.

This analysis also enabled the effective visualisation of many variables. For example, figure 2 illustrates a random subset of data illustrating the years in which specific researchers published work, providing a user-friendly, graphic view to the underlying data.

The next step was to plot the various research outputs based on the CESM codes linked to the output (or the deduced CESM-code). This enabled us to create a broad research landscape for the entire institution.

The entire set of research outputs was analysed and linked based on mutual authors, resulting in an encompassing graph illustrating all collaborating researchers across the university. The resulting graph's links were superimposed on the list of departments, creating a new graph linking all departments that collaborate on research.

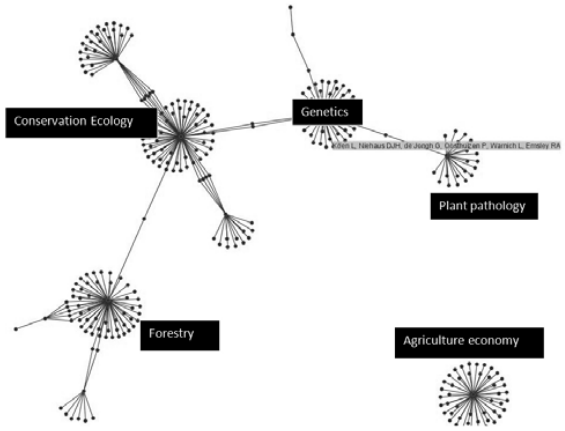

FIGURE 3 COLLABORATION BETWEEN DEPARTMENTS (PARTIAL BIOSCIENCE FOCUS)

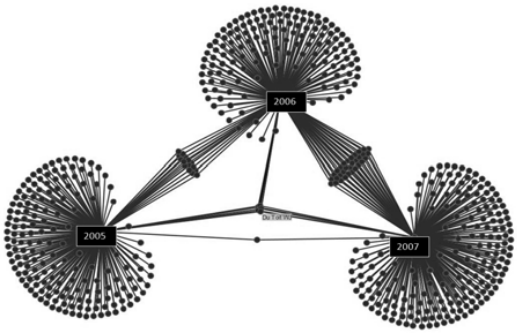

FIGURE 2 TOP PUBLISHING ACADEMICS (OUTPUTS PER YEAR)

Subsequently a Fruchterman-Reingold algorithm was applied, allowing us to cluster the outputs and present a view of these outputs indicating where subject areas link up (see figure 4. Each node indicates a research output). This algorithm attempts to evenly distribute vertices and minimize edge crossing. The resultant layout makes it reasonably easy to distinguish the automatically inferred "folksonomic" relationships[4].

By overlaying existing organisational structures on the visualised map, we have been able to ascertain where there is interdepartmental or inter-disciplinary cooperation and potential for research synergies within the institution (as illustrated in figure 3 for a small section of the biosciences).

These plots are also manually augmented with information such as the research chairs and strategic research projects, allowing the manager to see the web of research activities surrounding such a hub.

Examining the links between researchers, departments and topics has also enabled the institution to locate 'hot spots' of research activity not previously identified, and flagged these areas for possible institutional support. Orphaned outputs without any links and far away from the cluster of research activity was identified (and tracked over time) to investigate research allocation and alignment with the strategic research plan.

VI. CONCLUSIONS & FUTURE RESEARCH

We can conclude that data mining can be effectively utilised in a DSS analysing research outputs at universities and enable the identification of research focus, intensity and synergy. These reports were identified as particularly useful by institutional research management staff. Data mining also has the advantage of superior reliability and minimally invasive procedures. This allows for relatively inexpensive and quick results.

The obvious need for better central archiving of the full text of research outputs has been demonstrated and reporting systems are currently being modified to enable the future semantic analysis of outputs. This being said, an astounding amount of

[4] Although these inferred non-hierarchical relationships can be justified as analogous to explicit folksonomic links, one should note that the explicit user/expert defined links should have more weight. Both types of links can be shown on the resulting graph, but the system needs to allow the user to decide whether inferred links should be given equal weight.

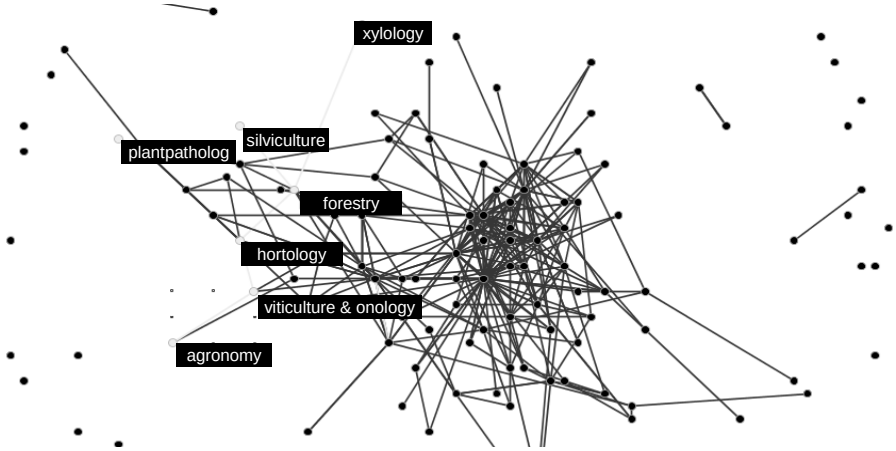

FIGURE 4 DATA WITH FRUCHTERMAN-REINGOLD LAYOUT APPLIED

information could be deduced by simply looking at the available reporting metadata.

Important progress was made in developing a suitable conceptual model to address the issue of disciplinary confusion within the higher education research sphere. Although this model requires more rigorous testing, it may find application in several areas where this ambiguity has traditionally proved challenging.

The clear mapping of the research space has also created a useful base on which further research projects can build. Many important questions were previously left unanswered due to the lack of such a generally acceptable, scientific understanding of the real research being conducted at the institution.

Although this pilot study only intended to construct this base and illustrate the viability of the approach, several potential areas of research were identified by the researchers.

The first obvious addition would be to complete the semantic analysis and try and identify linkages that are not present in the metadata. A content analysis would also greatly benefit the existing system enabling, amongst others, the automatic suggestion of relevant keyword based on output content.

A comparative study between the semantic analysis and metadata analysis may also enable research managers to evaluate the current structures of research and their efficiency in supporting the research.

Mapping the research space, at an institution, over time, may also allow the user to detect certain trends and adjust planning and strategy accordingly.

Perhaps the most exciting future development would be the full implementation of the interactive mode of interfacing with the DSS. This would entail enabling the user to explore the visualised data and also allow for the folksonomic tagging and user-specified relationship identification the conceptual model calls for. This could also be the first steps toward building a directory of expertise for the researchers at the institution.

Currently researchers have limited effective tools with which to practically manage their research portfolios. In most cases control and influence boils down to financial decision making and budgeting.

Utilising the model constructed here, one can now go about mapping the research funding flows at the institution in the same fashion. This will empower decision makers to evaluate the return on investment of allocated funds (as far as reported research is concerned) and also identify strategic areas that could benefit from the allocation of additional funding.

VII. REFERENCES

[1] P.G. Altbach, "Academic freedom: International realities and challenges", *Higher Education*, vol. 41, pp. 205–219, 2001.

[2] S. Slaughter and L.L. Leslie, "Academic Capitalism: Politics, Policies, and the Entrepreneurial University", Baltimore, MD: Johns Hopkins University Press, 1997.

[3] L. Leydesdorff, H. Etzkowitz, "Emergence of a Triple-Helix of university–industry–government relations", *Science and Public Policy*, vol. 23 (5), 1996.

[4] G. Boulton and C. Lucas, "What are universities for?", position paper, LERU, 2008.

[5] Higher Education Quality Comitee, "Audit Report on Stellenbosch University", Council on Higher Education Report, 2006.

[6] CREST, "Research at the University of Stellenbosch, A bibliometric Study", Centre for Research on Science and Technology, Report, December 2004.

[7] T. Gruber. "Toward principles for the design of ontologies used for knowledge sharing", *International Journal of Human-Computer Studies*, vol 46 (5-6), 907-928, 1995.

[8] T. Vander Wal. "Folksonomy". [online]. Available at http://vanderwal.net/folksonomy.html. [Accessed 17 March 2009].

[9] C. Shirky. "Ontology is overrated: Categories, links and tags.". [online]. Available at http://shirky.com/writings/ontology_overrated.html. [Accessed 21 March 2009].

[10] T. Gruber. "Ontology of Folksonomy: A Mash-Up of Apples and Oranges". *International Jounral on Semantic Web & Information Systems.* vol. 3 (1). 2007.

[11] Department of Education. "*First and Second Order Classification of Education Subject Matter (CESM) codes*". Internal document. Available at http://web.wits.ac.za/NR/rdonlyres/BEB43AE1-5327-4780-B16E-CBEC83939415/0/PUBSCESMS.doc. [Accessed 14 October 2008].

[12] E. Adar. "*The Graph Exploration System*". GUESS. [online]. Available at http://graphexploration.cond.org/. [Accessed 6 April 2009]

[13] E. Adar. "GUESS: A Language and Interface for Graph Exploration". University of Washington, Computer Science & Enginering. Available at http://graphexploration.cond.org/chi2006/guess-chi2006.pdf.

[14] T.M.J. Fruchterman & E.M. Reingold. "Graph Drawing by Force-Directed Placement". *Software: Practice and Experience.* vol 21 (11). 1991