"What Came First – *Wellbeing* or *Sustainability*?" A Systematic Analysis of the Multi-dimensional Literature Using Advanced Topic Modelling Methods

Mubashir Qasim[1] and Les Oxley[1]

[1] *mq21@students.waikato.ac.nz,* [1]*loxley@waikato.ac.nz*
Waikato Management School, University of Waikato, Hamilton, New Zealand

Introduction

Both sustainability and well-being (SaW) are inter-dependent, inter-disciplinary, multi-dimensional, and international subject areas. However, people tend to interpret the subjects significantly differently based on their professional affiliation, academic background, geographical location etc., (Brunn, 2014; Roberts et al., 2013). A search of the SaW literature, using any scholarly search engine, generates results ranging from the thousands to millions creating a challenge for the researcher in picking the right papers; constructing a reasonable structure and synthesizing the vast material in order to conduct a comprehensive review of the literature. The work presented here relates to the use of a sophisticated method to exploit the explanatory power of metadata, attached to the results of a search query, to identify hidden patterns in the universe of given articles. The methods and metadata used to conduct the systematic analysis are briefly discussed under following headings.

Components of systematic literature analysis

Acquisition of data

Our quest begins with the analysis of key characteristics of metadata obtained from JSTOR *Data for Research* (DFR), which enables exploration of >9.2 million articles. We collected and analysed the metadata for a sample of 68,817 papers from DFR which related to SaW for this exercise. Metadata were generated against four queries with different sets of keywords as listed in Table 1. Analysis of the metadata was conducted in three steps: Step 1., analysis of keywords, subject and subject groups, disciplines and discipline groups, journals, authors and trends of publications (as presented in a recent study by (Brunn, 2014) but with slightly different approach). In Step 2., we applied the Latent Dirichlet Allocation (LDA) to study language differentiation between SaW themes. The main aim of this exercise was to identify complex hidden patterns in the data and present them in easily understandable ways. In Step 3., we used a reference manager software package called *Qiqqa* to identify key themes in the personal library and to identify seminal and frontier studies within each theme using cross references in the collection.

Table 1: Detail of search queries.

Query	Results	Search keywords	Search in
A	4,903	wellbeing OR well-being	Abstract
B	57,681	sustainability OR sustainable development	Title
C	5,472	sustainability; sustainable development; wellbeing; well-being	Any
D	761	sustainability OR sustainable development; well-being OR wellbeing	Abstract

Analysis of keyterms

We sampled 300 top keywords appearing in the corpus of each query to represent the frequently used language patterns in the subjects of SaW. The results are presented in the form of word-clouds in which the terms with high frequencies of occurrence are represented by the larger size of the word. Each word in the cloud indicates a dimension or issue in a subject (Jaewoo & Woonsun, 2014). Broadly discussed dimensions in the well-being literature include income, health, relationships, family, child, psychology etc., are correctly identified in our word-clouds.

Type of journals and subject group

Inter-relatedness of the SaW literature is established by confirming the large number of journals shared by SaW papers as suggested by (Mimno, 2012). Here, we extracted the names of the top 20 journals by number of articles in each query. Our analysis validates the assumption that many journals include papers on both aspects of the SaW literature. The interdisciplinary nature of the SaW literature is further established by similar categorization of SaW papers with respect to different subject groups.

Trends in publications

Many modern databases are devoted to tracking publications e.g., as Google Scholar, ISI Web of Science, JSTOR, SCOPUS, etc., and enable

scholars to perform quick and broad browsing of the literature (Hood & Wilson, 2003). Their expansions or contractions over time can indicate the interest of scholars in an area and the evolution of novel approaches (Adam, 2002; Casagrandi & Guariso, 2009).

In our analysis, we find the first article related to Query A, appears in 1919 and the number of publications remains trivial until the 1970's. Thereafter, a huge influx of papers begins in the late 1970's with 30 papers per year, peaking at 311 papers in 2012. In contrast, papers related to sustainability in Query B started much earlier with the first paper published in 1800. This number reaches to 50 papers per year in the next 100 years and steadily increase thereafter for another 50 years to around 250 papers per year in 1950. Post-1950, the number of scholarly articles grew five fold over the next five decades and peaked in 2005 at 1304 papers per year. Articles related to both SaW in Query C emerge in the late 1970's and grow exponentially over the next 40 years. As Query D is a subset of Query C they exhibit similar trends. A comparison of these trends with the papers in the entire DRF corpus of 9.3 million articles indicates the level of interest of the scholars over different years.

Authors of publications and places

Another way to consider the SaW literature is to analyse the country of the main author(s) of an article in order to answer the key question "what countries are leading the SaW agenda?" We select the top 20 authors in each set of documents based on their number of publications. Their country is established from the place of their affiliation at the time of publication. Our results show 74 unique authors from 12 different countries wrote 1,869 SaW paper. Not unexpectedly, 9 of these countries are developed OECD countries with the United States the home of 61% of SaW authors and 29% of this literature is produced by people from Europe, Canada and South Africa and rest of them are from Australia, India and Botswana.

Differentiating language using LDA

Finally, we conducted probabilistic analysis of the SaW literature using Latent Dirichlet Allocation (LDA) in order to establish underlying topics within the corpus of documents in each query (a topic is a set of co-occurring words). Our analysis helps understanding what sort of language is used within and across disciplines; what clusters of words happen to occur together; and how the use of language changes overtime. Results are shown by java based interactive visuals made in the programing language R. Each topic provides a clear structure to build a paragraph in a literature review and the cluster of topics gives a clear indication of the categories/themes within each set of documents.

Identification of seminal and frontier studies

Most dominant papers in our set of documents are identified using in-bound references assuming that heavily cited and highly ranked articles are the key papers in each collection. Identification of these articles provides the best starting point to begin the traditional literature review with. We used network diagrams using a reference manager called Qiqqa to conduct this exercise.

Validation of results

The results are validated using the metadata from another widely used scholarly source called Web of Science. Most of our results exhibit the same characteristics as the results of DFR data.

References

Adam, D. (2002). Citation analysis: The counting house. *Nature*, *415*, 726–729. doi:10.1038/415726a

Brunn, S. D. (2014). Cyberspace Knowledge Gaps and Boundaries in Sustainability Science: Topics, Regions, Editorial Teams and Journals. *Sustainability*, *6*(10), 6576–6603. doi:10.3390/su6106576

Casagrandi, R. & Guariso, G. (2009). Impact of ICT in Environmental Sciences: A citation analysis 1990–2007. *Environmental Modelling & Software*, *24*(7), 865 – 871. doi:http://dx.doi.org/10.1016/j.envsoft.2008.11.013

Hood, W. & Wilson, C. (2003). Informetric studies using databases: Opportunities and challenges. *Scientometrics*, *58*(3), 587–608. Kluwer Academic Publishers. doi:10.1023/B:SCIE.0000006882.47115.c6

Mimno, D. (2012). Computational Historiography: Data Mining in a Century of Classics Journals. *J. Comput. Cult. Herit.*, *5*(1), 3:1–3:19. New York, NY, USA: ACM. doi:10.1145/2160165.2160168

Roberts, L., Brower, A., Kerr, G., Lambert, S., McWilliam, W., Moore, K., Quinn, J., et al. (2013). A Good Life: How nature's ecosystem services contribute to the wellbeing of New Zealand and New Zealanders. Department of Conservation.

Jaewoo, C. & Woonsun, K. (2014). Themes and Trends in Korean Educational Technology Research: A Social Network Analysis of Keywords . *Procedia - Social and Behavioral Sciences, 131*(0), 171 – 176. doi:http://dx.doi.org/10.1016/j.sbspro.2014.04.099