Leveraging Semantic Analysis Technologies to Increase Effectiveness and Efficiency of Access to Information

Denise A. D. Bedford[1] and Karen F. Gracy[2]

[1]Information Architecture and Knowledge Management, Kent State University
[2]School of Library and Information Science, Kent State University

Abstract: This presentation will report on the results of a study of 245 publications and reports which were (1) classified to both LCSH for inclusion in a standard library catalog and classified and indexed to an economic development related topic classification scheme and institution specific thesaurus; (2) both manually classified and automatically classified using natural language processing (NLP) semantic technologies. The focus of the comparison is (1) a direct comparison of the quality and quantity of access points across schemes, and (2) the effect of variations on conceptual search (meaning versus word-based search systems). The research collection includes 245 digital publications and reports published by the World Bank which are accessible through OCLC's WorldCat catalog and through the World Bank's external web faceted Documents and Reports search system.

Keywords: Semantic analysis; natural language processing; automatic classification; Library of Congress Subject Headings; World Bank; OCLC WorldCat

Research Context

The inspiration for this research derives from lessons learned working with subject matter experts in the design of vocabularies to support effective search. Specifically, this and other in-progress research are being undertaken in the hopes of building a deeper understanding of the performance of source vocabularies in conceptual search. Conceptual search is designed to retrieve information that is "conceptually similar" to the information contained in the search query. A conceptual search looks for and tries to retrieve information that pertains to the ideas expressed in the search query. Conceptual search and term or word search are different approaches to an information retrieval challenge. Conceptual search focuses on the meaning expressed in the words of a query, rather than simple word or pattern matching. For this reason, the goodness and performance of conceptual search depends heavily on the underlying knowledge base. Conceptual search systems need deep knowledge of concepts, which are typically represented in knowledge organization systems (KOS) such as controlled vocabularies, thesauri and ontologies. Knowledge organization systems provide machine-level access to synonyms, hypernyms,

Received: 17 April 2012 / Accepted: 24 July 2012 ISSN 2241-1925

hyponyms, and variant or associative concepts for direct use in conceptual search.

Over the years, two primary approaches have dominated research into constructing the deep conceptual knowledge needed for concept search - those that follow the semantic analysis approach and those that follow the statistical approach. Statistical approaches to conceptual searching rely largely on word pattern detection and statistical alignment. Semantic analysis approaches to conceptual search leverage natural language processing, controlled vocabularies, thesauri, ontologies (sometimes referred to as Knowledge Organization Systems, or KOS), and more sophisticated query processing and query matching algorithms. Selective studies have demonstrated that concept search can be improved by the use of controlled vocabularies (Giunchiglia, Kharkevich, and Zaihrayeu, 2009; Miller, Beckwith, Fellbaum, Gross and Miller, 1990). The research reported in this paper focuses on the semantic approach but employs a hybrid approach between semantic and statistical.

Studies in the information science literature suggest that performance improvements in conceptual search can be achieved from the introduction of knowledge organization systems, specifically controlled vocabularies. However, these improvements have been qualified by the nature of the vocabularies and inherent level of human effort required to construct and maintain the controlled vocabularies. In particular, several studies have examined the nature and performance of Library of Congress Subject Headings (LCSH) (Fischer, 2005; Kreider, 2000; Marshall, 2003) and LCSH in comparison to other sources and approaches to vocabulary development, including author-assigned keywords, domain specific thesauri and controlled vocabularies (Dubois, 1984; Holley, 2007; Losee, 2009; McCutcheon, 2009; Olson, 2008; Schabas, 1982; Tonta, 1992). The studies have researched basic coverage and mapping of sources, the length and structure of terms, the level of generality-specificity of terms, and the use of the sources for indexing. However, these studies would tend to suggest that the maintaining these sources presents a significant constraint. The level of effort, required investment, and challenge of maintaining currency have generally precluded scaling this approach for conceptual search. A second constraint with this approach is the depth and extent of conceptual indexing required for individual documents and content objects. Over the years, these concerns have also led researchers to conclude that the statistical approach holds greater promise than the semantic.

These are critical limitations that must be overcome if conceptual search is to become a practical search option in the future. An alternative hybrid approach has been proposed in the early 21st century, though – the use of semantic analysis methods to generate and maintain knowledge organization systems and to support deep automated conceptual indexing. The hybrid approach may be a viable alternative to either full manual or full machine-based approaches. Conceptual searching becomes more practical if the level of effort required to construct and maintain knowledge organization systems can be managed, and the productivity challenges of conceptual indexing can be met.

This research is a small step in exploring the feasibility of the hybrid approach. This research compares the potential value of two knowledge organization systems for conceptual search - specifically, (1) the comparative differences between the Library of Congress Subject Headings[i] (which are entirely manually constructed and designed for use in traditional library catalog "term-based" searching), and the World Bank Enterprise Topic Thesaurus[ii] (generated using a combination of semantic and manual methods). This research also explores the comparative differences in indexing practices achieved manually and using machine-based approaches. Observations on the implications for conceptual search are presented.

The Research Questions

The research explores two fundamental issues related to conceptual search. First, can a semantically generated knowledge organization system such as the World Bank Enterprise Topic Thesaurus perform as well or better than an established and universally accepted controlled vocabulary such as the Library of Congress Subject Headings? Second, can machine-based conceptual indexing methods perform as well or better than high quality manual subject indexing methods, i.e., can we make improvements in indexing productivity?

Research Methodology

The research team translated these issues into four specific research questions, including:

- Question 1: Given a controlled set of documents indexed with two distinct vocabularies, what is the rate of convergence of conceptual indexing terms and subject headings assigned, i.e., can we surpass the current performance of manually produced KOS using semantically generated KOS?
- Question 2: What is the nature of the convergence between indexing terms assigned from the two vocabularies (i.e., exact match, partial match, variants, etc.) and does the convergence have value for conceptual search?
- Question 3: Is there a variation in the construction of indexing terms across the two vocabularies, i.e., can we maintain the level of quality of indexing term construction using semantic approaches?
- Question 4: What is the comparative rate of assignment of manually generated indexing terms and machine assisted indexing terms, i.e., is the hybrid approach a feasible alternative for deep conceptual indexing?

The context for exploring these four questions was a set of controlled documents which had been (1) indexed to both Library of Congress Subject Headings, and the World Bank Thesaurus; and (2) both manually and machine-indexed. The World Bank's publicly available documents and publications were selected as the target collection for several reasons. First, the World Bank's documents and publications are widely respected and are likely to be found not only in the World Bank's repositories, but also in academic library collections. Where World Bank documents are found in academic library catalogs, they are likely to be manually indexed. In recent years, the World Bank implemented automated indexing methods for subject analysis.[iii] Second, World Bank's documents are likely to be indexed using both the World Bank Thesaurus indexing terms and the Library of Congress Subject Headings. The World Bank's machine assisted indexing methods leverage the World Bank Thesaurus as an embedded knowledge organization system. Academic libraries are likely to use Library of Congress Subject Headings. Third, there was a high probability that full metadata would be easily available on both the World Bank's external web search and through OCLC's WorldCat catalog.

Creating the Research Data Set

The first research task was to identify a broad set of documents and publications to construct the project data set. The World Bank's external website supports a faceted search for documents and publications.[iv] One of the most popular facets is searching by "Document Type." Document Types are organized into five high level categories – Board Documents, Country Focus, Economic & Sector Work, Project Documents, and Research and Publications. Of all five document types, Research and Publications were most likely to be found in academic library collections. The project team identified 337 World Bank research documents and publications, and downloaded their respective metadata as the control data set. Then, each document title was searched by author and title in WorldCat. Of the initial set of 337, a total of 245 (72.7%) were found in WorldCat. From WorldCat, the OCLC record numbers and LC Subject Headings were downloaded and added to the bibliographic data for each of the 245 titles.

These comparable metadata for these 245 titles comprised the project data set. The data set included 81 publications (books), 94 journal articles, and 70 documents representing other types of research documents such as Annual Reports, Commodity Working Papers, Departmental Working Papers, Environmental Working Papers, Energy Sector Management Assistance Program (ESMAP) documents, Financial Flows papers, Global Development Finance papers, Global Environment Facility documents, Human Capital papers, Issues in Agriculture papers, research papers from Latin America and Caribbean Region, Managing Agricultural Development in Africa (MADIA), Poverty Assessments (PA), Poverty Research, Social Policy, Special Program for African Agricultural Research (SPAAR), United Nations Development

Programme (UNDP), and Water and Sanitation, and last, World Bank Institute reports.

Review and Evaluation Methodology

Excel spreadsheets were used to capture, count, analyze, and compare indexing terms and topics. The metadata and bibliographic data downloaded provided an exploratory base of 12,156 unique conceptual indexing terms from the World Bank Thesaurus, and a total of 614 unique LCSH subject headings. The research team reviewed the comparative metadata for the 245 titles item by item, and word by word. Across all documents in the project data set, the reviewers read and marked up 28,487 World Bank Thesaurus indexing terms and 835 LC Subject Headings assigned to the 245 titles.

Table 1.
Comparison of LCSH and World Bank Thesaurus Assigned and Unique Terms

Source	Total Indexing Terms Assigned	Total Unique Terms Used
LCSH Terms	835	614
Controlled Vocabulary/Machine Assigned	28,487	12,156

Research Results

The results for the four research questions are presented below. Observations and recommendations drawn are also included.

Question 1: Given a controlled set of documents indexed with two distinct vocabularies, what is the rate of convergence of conceptual indexing terms and subject headings assigned, i.e., can we surpass the current performance of manually produced KOS using semantically generated KOS?

The rate of matching is particularly important to conceptual search. Conceptual search systems need to have access to a broad knowledge base, including exactly matching concepts, partially matched concepts, variant terms. The research team compared conceptual indexing terms and subject headings using the well formed categories established by Strader in her study of Electronic Theses and Dissertations (2009). Strader identified six types of matches including: Exact Match, All Present, Partial Match, Needs 2 LCSH, Variant, and No Match. Of Strader's six match types, we found that four had value for this research. The match categories borrowed from Strader and used in this research are defined in Table 2.

Table 2.
Categories of Match

Type of Match	Description of Matching Criteria
Exact Match	Exact match of the words in the concept, including plurals (e.g., labor regulations and labour regulations), acceptable spellings (i.e., ecosystem and ecosystems), other language forms (i.e., -- include Spanish language forms here), or exact concept match including equivalent terms (e.g., biodiversity, species diversity, biological diversity)
Partial Match	Partial match of the concepts in the indexing term or subject heading such as a narrower or broader form of the concept (e.g., biodiversity and forest biodiversity), or a single indexing concept aligned with a compound LCSH (e.g., Roads & Highways vs. Roads, Highways)
Variant	Conceptual variant – covers the core concept but speaks to either a refinement or other associative relationship with the core concept (e.g., biodiversity and biodiversity management,
No Match	Terms were not present in any of the above forms. Concept was missed entirely by the other source or indexing method.

Strader also defined the criteria for matching as a "sequence of words." This research takes as a starting point the definition of a "concept match" – equivalent meaning rather than equivalent words. The idea of concept match derives from definitions of equivalent relationships in the ANSI/NISO Z39.19 standard governing Thesauri and Classification Systems. For example, in Strader's approach, Biodiversity, Biological Diversity, Species Diversity and Species Variation might be considered either Variants or Partial Matches. They would not be characterized as exact matches. Taking a conceptual match approach, these four concepts would be considered exact matches. Similarly, Labor and Labour would be a conceptual match, as would be Climate change and Climatic change.

A further variation on Strader's approach is the focus only on subject terms and headings. Library of Congress Subject Headings often contain subdivisions that pertain to countries, resources, time periods, or other non-topical terms; none of these subdivisions were included in the comparisons. The World Bank Thesaurus coverage focuses on topical subject domains. Names of countries, regions or types of resources are treated as separate and distinct metadata attributes and supported by distinct controlled reference sources. To ensure comparability, LC Subject Headings subdivisions that pertained to countries, regions, or types of resources were removed from consideration. In addition, due to common indexing practices in academic libraries, LCSH subdivisions often contain repetitive terms. For research purposes, repetitive LCSH terms were counted only once for the same bibliographic record. For example, *Economic development – Honduras – Case Studies* and *Economic development*

– Kenya – Manuals – generated only one LCSH for comparison: *Economic development*.

Table 3 indicates that 66.45% of the Library of Congress Subject Headings used were matched to World Bank Thesaurus conceptual indexing terms. This is a high rate of coverage for two distinct vocabularies, and is likely due to the conceptual foundation of the World Bank Thesaurus. 41.93% of the total matches were exact matches. 17.74% of all matches were Partial Matches and Variant Matches accounted for the remaining 40.32%.

Table 3.
Raw Counts of LCSH Matches

Category	**Numbers**	
LCSH Exactly Matched by Indexing Terms	234 *(41.93%)*	
LCSH Partially Matched by Indexing Terms	99 *(17.74%)*	
LCSH Matched by Variant Indexing Terms	225 *(40.32%)*	
Total LCSH Headings Matched by Indexing Terms	558	*66.45%*
LCSH Not Matched by Indexing Terms	280	*33.55%*
Total Headings	*835*	*100.00%*

In contrast, the Library of Congress Subject Headings missed 85% of the conceptual indexing terms (Table 4). The difference in coverage was due in part to the difference in number of terms assigned in traditional academic library indexing and conceptual indexing approaches. A total of 835 subject headings were assigned, compared to 28,487 conceptual indexing terms. Of the conceptual indexing terms assigned, Library of Congress matched only 14.99% - 85.01% were not matched in any form. Only 8% were exact matches, 20.13% were Partial Matches, and the remaining 71% were Variant Matches.

Table 4.
Raw Counts of Controlled Vocabulary Indexing Terms

Category	**Numbers**	**%**
WB Thesaurus Terms Exactly Matched by LCSH	245 *(8.22%)*	
WB Thesaurus Terms Partially Matched by LCSH	610 *(20.14%)*	
WB Thesaurus Terms Matched by Variant LCSH	2,170 *(71.64%)*	
WB Thesaurus Terms Matched by LCSH	3,029	*14.99*
WB Thesaurus Terms Not	24,216	*85.01%*

Matched by LCSH		
Total Matches	*28,487*	*100.00%*

Table 5 provides a side-by-side comparison of the convergence rates of Library of Congress Subject Headings and World Bank Thesaurus conceptual indexing terms.

Table 5.
Breakdown of Matches of LCSH and World Bank Thesaurus Indexing Terms

Category	**% LCSH Assigned Terms Matched by World Bank Thesaurus**	**% World Bank Thesaurus Assigned Terms Matched by LCSH**
Not Matched	*33.55*	*85.01%*
Total Matches	*66.45*	*14.99%*
Breakdown of Matching Terms		
Exact Match	41.94%	8.22%
Partial Match	17.74%	20.14%
Matched by Variant	40.32%	71.64%

Question 2: What is the nature of the convergence between indexing terms assigned from the two vocabularies (i.e., exact match, partial match, variants, etc.) and does the convergence have value for conceptual search?

The research team also examined the nature of each category of match. Tables 6 through 10 provide illustrative examples of Exact Match, Partial Match, Variant Match, and missed (Not Matched) headings and indexing terms. Strader's matching criteria also provide a framework for understanding how the semantically generated vocabulary can be leveraged in conceptual searching. Exactly matched terms provide insight into the potential use of "equivalent meaning" for automated expansion of query terms. Table 6 provides examples of exact matching terms found in both sources, including one example of exactly matching "meanings" beyond simple word matches.

Table 6.
Sample Exactly Matched LCSH Headings and Indexing Terms

LCSH Headings	**Controlled Vocabulary Indexing Terms**
AIDS (Disease)	AIDS
Carbon content	Carbon content
Climatic change	Climate change, climate sensitivity, climate variability, climate variation, climate variations, climate vulnerability, climatic change, climatic

	changes
Education, Secondary	Secondary education
Labor market	Labour market
Clothing trade	Clothing industry

Strader's categories for "partial match" will require some translation and interpretation before the true value for conceptual search might be assessed. The challenge is the varied nature of "partial match" in a word-based versus a conceptual search context. In some instances, "partial matches" might be narrower concepts in conceptual search, and in others they may be treated as close equivalent or exactly matching terms.

Table 7.
Sample Partially Matched LCSH Headings and Indexing Terms

LCSH Headings	Controlled Vocabulary Indexing Terms
Integrated water development	Water development
Occupational retraining	Retraining
Rural conditions	Rural areas
Teacher education	Training of teachers
Agricultural extension work	Agricultural extension services

Strader's categories for "variant terms" surface perhaps the clearest indication of potential value for conceptual searching of the hybrid approach, grounded in a semantically generated controlled vocabulary. Table 8 provides a view of the rich expansion of variant concepts from Library of Congress Subject Headings and the World Bank Thesaurus.

Table 8.
Sample Variant Matched LCSH Headings and Indexing Terms

LCSH Headings	Controlled Vocabulary Indexing Terms
Tariff	Tariff barriers, tariff concessions, tariff equivalent, tariff equivalents, tariff levels, tariff negotiations, tariff preferences, tariff protection, tariff rates, tariff reduction, tariff reform, tariff revenue, tariff revenue losses, tariff schedule, tariff setting, tariff structure, tariff structures
Transportation	Transport, transport agencies, transport agreements, transport capacity, transport corridors, transport cost, transport costs, transport equipment, transport facilitation, transport infrastructure, transport investment, transport investments, transport market, transport modes, transport network, transport

	operators, transport price, transport projects, transport regulation, transport research, transport sector, transport service
Rice	Rice areas, rice crop, rice cultivation, rice husks, rice milling, rice mills, rice prices, rice production, rice research, rice varieties, rice yields
Politics and political science	Political change, political channels, political control, political decisions, political economy, political environment, political influence, political instability, political interference, political leadership, political opposition, political participation, political parties, political power, political process, political rights

Table 9 provides examples of entire concept areas that were missed either as a result of lack of coverage in the Library of Congress Subject Headings or due to manual indexing practices and constraints.

Table 9.
Sample Clusters of Indexing Terms Lacking in LCSH

Consumers, consumer choice, consumer demand, consumer goods, consumer groups, consumer preferences, consumer price indexes, consumer prices, consumer products, consumer protection, consumer protection regulations, consumer surplus
Community-driven development, community access, community action groups, community benefits, community capacity, community colleges, community development, community education, community engagement, community facilities, community infrastructure, community involvement, community management, community organizations, community participation
Aquifers, groundwater

Taking the LC Subject Headings as a baseline for comparison, only one topic area was missing from the World Bank Thesaurus – designation of Communist countries and other terms that may reflect a political perspective. While there was extensive coverage of politics and political systems, the World Bank Thesaurus did not use the designation “Communist countries” to refer to subject matter content. All other areas identified at the concept level in LCSH were covered in the World Bank Thesaurus.

Question 3: Is there a variation in the construction of indexing terms across the two vocabularies, i.e., can we maintain the level of quality of indexing term construction using semantic approaches?

We expect the conceptual indexing terms to achieve a greater level of specificity and granularity. This can be represented simply in more atomic concepts or in a more extensive description of the concept, as represented by words used in the Heading or indexing term. While it was the case that more atomic level concepts were included in the World Bank Thesaurus, Table 10 indicates that there was no variation in the number of words per heading or indexing term. Both sources followed good practice guidelines for indexing term construction.

Table 10
Average Number of Terms

Source	**Ave. # Word Per Heading or Indexing Term**
Library of Congress Subject Headings	2.073
World Bank Thesaurus Indexing Terms	2.029

Question 4: What is the comparative rate of assignment of manually generated indexing terms and machine assisted indexing terms, i.e., is the hybrid approach a feasible alternative for deep conceptual indexing?

Indexing practices for both the traditional academic library context and the machine-generated metadata context are prescribed. The results of the exploratory research are not unexpected. In the academic library setting, a generally accepted practice is to assign three, but generally not more than five, subject headings (Studwell, 1990). The research data set was entirely representative of academic library indexing practices. The average number of Library of Congress Subject Headings assigned was 3.36. Some bibliographic records had no subject descriptors. Machine assisted indexing identified a significantly greater number of relevant conceptual indexing terms than were suggested through manual indexing. In the machine-generated metadata context, the number of conceptual indexing terms is determined by the density and coverage of the content. A general upward limit of 330 is placed on documents and publications available on the Bank's external web. The limit is set higher for internally accessible documents. On average, the same documents had 115.77 conceptual index terms.

Table 11.
Average, Median, Maximum and Total
Indexing Terms and LCSH per Title

Source	**Average**	**Minimum**	**Maximum**	**Median**
World Bank Thesaurus / Machine Assigned	115.77	3	245	113.5
LCSH/Manually Assigned	3.36	0	11	4

Conclusions

The research explored two fundamental issues related to conceptual search. First, we questioned whether a semantically generated knowledge organization system such as the World Bank Enterprise Topic Thesaurus could perform as well or better than an established and universally accepted controlled vocabulary such as Library of Congress Subject Headings. Second, we asked whether machine-based conceptual indexing methods could perform as well or better than high quality manual subject indexing methods (i.e., can we make improvements in indexing productivity?). The research results suggest that the semantically generated controlled vocabulary achieves a comparably high performance level. For conceptual search, the results suggest that the semantically generated controlled vocabulary provides greater value than manually assigned subject headings. The research results also suggest that hybrid machine-based indexing methods that are grounded in rich knowledge organization systems are a feasible and scalable alternative to either pure manual or pure statistical indexing methods.

Acknowledgements

The authors wish to thank their research assistants, Tom Burdick and Nicole Yoder, for their contributions to this paper.

References

Chan, L.M., (1998). Still Robust at 100: A Century of LC Subject Headings. *Library of Congress Information Bulletin*, Vol. 57, no. 8, 200-201. http://www.loc.gov/loc/lcib/9808/lcsh-100.html

Dubois, C., (1984). The Use of Thesauri in Online Retrieval. *Journal of Information Science*, Vol. 8, No. 2, 63-66.

Fischer, K., (2005). Critical Views of LCSH, 1990-2001: The Third Bibliographic Essay. *Cataloging & Classification Quarterly*, Vol. 41, no. 1, 63-109.

Giunchiglia, F., Kharkevich, U., and Zaihrayeu, I., (2009). Concept Search. Proceedings of the *6th European Semantic Web Conference* (ESWC 2009): The Semantic Web: Research and Applications. Heraklion, Crete, Greece, May 31–June 4, 2009. 429-444. http://www.ulakha.com/concept-search-eswc2009.html

Holley, R.P., (2007). Subject Access Tools in English for Canadian Topics: Canadian Extensions to U.S. Subject Access Tools. *Library Resources and Technical Services,* Vol. 52, No. 2, 29-43.

Kreider, L., (2000). LCSH Works! Subject Searching Effectiveness at the Cleveland Public Library and the Growth of Library of Congress Subject Headings Through Cooperation. *Cataloging & Classification Quarterly*, Vol. 29, nos. 1-2, 127-34.

Losee, R., (2004). A Performance Model of the Length and Number of Subject Headings and Index Phrases. *Knowledge Organization* 31 (2004), No. 4, 245-251.

Marshall, L., (2003). Specific and Generic Subject Headings: Increasing Subject Access to Library Materials," *Cataloging and Classification Quarterly* 36 (2003), no. 2, 59-87.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K., (1990). Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, Vol. 3, No. 4, 235-244.

McCutcheon, S. (2009). Keyword vs. Controlled Vocabulary Searching: The One with the Most Tools Wins. *The Indexer*, Vol. 27, no. 2, 62-65.

Olson, Tony, (2008). LCSH to MeSH, MeSH to LCSH. *Cataloging & Classification Quarterly*, Vol. 46, no. 4, 438-439.

Schabas, A. H., (1982). Postcoordinate Retrieval: A Comparison of Two Indexing Languages. *Journal of the American Society for Information Science*, Vol. 33, no. 1, 32-37.

Strader, C. R., (2009). "Author-Assigned Keywords versus Library of Congress Subject Headings: Implications for the Cataloging of Electronic Theses and Dissertations," *Library Resources and Technical Services*, Vol. 53, no. 4, 243-250.

Studwell, W., (1990). Subject Suggestions 6: Some Concerns Relating to Quantity of Subjects. *Cataloging & Classification Quarterly*, Vol. 10, no. 4, 99-104.

Tonta, Y., (1992). LCSH and PRECIS in Library and Information Science: A Comparative Study. *Occasional Papers*, No. 194.

[i] Library of Congress Subject Headings (LCSH) is a controlled vocabulary, originally developed in 1898 by the Library of Congress for its own internal cataloging needs, which has become an internationally adopted standard source for topical terms. It contains over 250,000 subject headings, and has been described as "the most comprehensive general controlled vocabulary" in the world (Chan, 1998). LCSH terms are arranged in alphabetic, or dictionary, order. They are not organized by facets or by topical term groupings, however, all terms are related to one another through the tracing of equivalent (synonymic), hierarchical (broader/narrower term), and associative (related term) relationships. A primary feature of LCSH is the extensibility of the headings through subdivision. Topics that have been chosen to describe the subject matter of a work may be further specified using topical, geographical, chronological, or genre subdivisions. An example, drawn from a record analyzed for this study, would be: Conservation of natural resources—Thailand—Periodicals. As noted above, however, for the purposes of this study the main entry point of the heading was usually chosen to determine whether or not there was a match between the LCSH heading manually assigned to the work and World Bank thesaurus term automatically assigned to the work. LCSH is a general vocabulary that may be used for cataloging all types of library materials. Its level of granularity makes it useful for providing subject access to larger library collections of monographs, periodicals, multimedia and audiovisual materials, and

other materials commonly found in libraries. It is less helpful for providing entry points for technical literature and articles within periodicals, due to its generalist nature and the limits to which it can reflect depth of coverage of topics in works (most works cataloged using LCSH are assigned no more than three or four headings).

[ii] The World Bank Thesaurus is an organizational thesaurus designed to facilitate conceptual searching of documents, publication and all other types of content related to any aspect of economic development. The Thesaurus is inclusive of all Bank interests and issues over the Bank's 70-year history. It is also intended to be sufficiently flexible and robust to manage those interests in the future. The Thesaurus covers approximately 28 high level topics (Agriculture, Communities and Human Settlements, Conflict and Development, Education, Energy, Environment, Finance, Gender, Governance, Health, Industry, Poverty Reduction, Rural Development, Urban Development, Transport, etc.). Each high level topic is subdivided in to up to 36 subtopics. Each subtopic has an elaborated controlled vocabulary of between 500 and 10,000 conceptual indexing terms. The Thesaurus is comprised of approximately 500,000 terms, which are designed to support conceptual indexing and searching.

[iii] The conceptual indexing terms assigned to the World Bank documents and publications are generated automatically using the SAS/Content Categorization Suite. The Categorization Suite includes grammar based concept extraction, rule-based concept extraction, dynamic categorization, rule based categorization, and summarization technologies. These technologies have underlying Natural Language Processing capabilities, leverage extensive internal organizational knowledge bases and support the construction of institution-specific rules for conceptual indexing. Automatically generated metadata is stored as persistent metadata in metadata records, which are managed in organizational data stores.

[iv] http://www-wds.worldbank.org/WBSITE/EXTERNAL/EXTWDS/0,,detailPagemenuPK:64187510~menuPK:64187513~pagePK:64187848~piPK:64187934~searchPagemenuPK:64187283~siteName:WDS~theSitePK:523679,00.html