Classification of the Web: Challenges and Inquiries

Jens-Erik Mai

The Information School, University of Washington, Seattle
E-mail: jemai@u.washington.edu

Jens-Erik Mai is Assistant Professor at the Information School of the University of Washington. He teaches classes on indexing, classification, thesaurus construction, and the theoretical foundation of information science. His research is focused on usage-centered approaches to organization and representation of knowledge and he is particularly interested in understanding how people use information in their workplaces and how information should be organized and represented to support their work. He approaches this issue from theoretical as well as empirical perspectives, drawing on semiotics, philosophy of language, domain analysis, and Cognitive Work Analysis. Jens-Erik Mai holds a Ph.D. in Library and Information Science from the University of Texas at Austin and a Master's and a Bachelor degree from the Royal School of Library and Information Science, Denmark.

Mai, Jens-Erik. (2004). **Classification of the Web: Challenges and Inquiries.** *Knowledge Organization, 31(2).* 92-97. 33 refs.

ABSTRACT: This paper discusses the challenges faced by investigations into the classification of the Web and outlines inquiries that are needed to use principles for bibliographic classification to construct classifications of the Web. This paper suggests that the classification of the Web meets challenges that call for inquiries into the theoretical foundation of bibliographic classification theory.

Introduction

Increasingly more information is becoming available on the World Wide Web. However, searching for information on the Web can be wearisome and frustration can arise from finding too much information, too little information, or not being sure that the right information is found. This frustration arises partly from the type and quality of the current search services available on the Web, which is dominated by search engines based on free keyword searching. Some of the frustration might, however, be eased by using classificatory structures to search for information.

A user searching for documents that contain useful information on the Web can utilize two different strategies to identify those documents:

1) The user can utilize a search engine to search for specific words that the documents contain or have been assigned (e.g. free keywords), or
2) The user can employ a classified directory where the documents are organized systematically to facilitate navigation.

The initial indexing of documents determines the type of search strategy available to the user; however, any retrieval system can be set up to facilitate either one or both search strategies.

Each of the two types of strategies has its strengths and weaknesses. Keywords are very effective when searching for information about specific people, places, or things that have unique names. When searching for information about concepts that can be expressed in multiple ways it is more effective to use classified Web directories. A directory controls for synonyms and homonyms and provides context for the index terms by placing them in a hierarchical structure. It has often been suggested in the indexing literature that the two ways of indexing documents complement each other and should be used in combination (cf. e.g. Svenonius 1986; Fidel 1991; Rowley 1994).

The role and value of a classified approach to the retrieval of Web documents has not been crystallized yet. Since most Web search services stress the free keyword search strategy it would seem as if the major Web directories[1] are still struggling to find their role and being recognized as valuable by the users and the Web search services. The bibliographic classification research community can offer insights into how Web directories could be constructed more successfully and gain the status and role that bibliographic classifications have in the LIS community. I will, however, argue that the principles and techniques developed by the bibliographic classification community over the past century face major challenges in being applied in the Web environment. The challenges arise because the principles for constructing bibliographic classification systems are closely connected to the paper environment and to the scientific community and its discourse. In this paper I will discuss and outline fundamental challenges that will be encountered when attempting to classify the Web and basic inquiries that need to be pursued in order to contribute to and advance the classification of the Web.

Classificatory Principles

The development of principles for the construction of systems for the organization of information has a long history within bibliographic classification research and there are today a number of theoretical and practical principles and techniques for the construction of classification systems and retrieval thesauri.

The principles for organizing library material have been discussed in the literature at least since the middle of the nineteenth century (cf. Panizzi 1985/1848) and some of the basic principles that are still valid today were formulated in the late nineteenth century (e.g. Cutter 1876). The foundation for classification theory was laid in the first part of the twentieth century with the work of Sayers (1915), Bliss (1929), and Richardson (1935), among others. In the middle of the twentieth century a new bibliographic classification theory and new principles were introduced by Ranganathan (1962; 1967) that further added new principles to the organization of information. The British Classification Research Group further developed Ranganathan's ideas (cf. e.g. Vickery 1960; Classification Research Group 1957) and added to them. The principles for the construction of bibliographic classification schemes laid down by these authors are often used and referred to as the foundation for bibliographic classification theory.

A related organizational and retrieval technique was also introduced in the middle of the twentieth century: the thesaurus. The thesaurus offers the advantage of searching by words instead of notations and is a widely used technique to organize specialized collections and bibliographic databases on particular topics. A number of textbooks on the construction of thesauri (cf. e.g. Soergel 1974; Lancaster 1986; Aitchison, Gilchrist, and Bawden 2000) and standards for the construction of thesauri (International Organization for Standardization 1986; National Information Standards Organization 1994) describe these principles and techniques. The theoretical and practical development of the principles and techniques behind thesaurus construction took place mainly in the 1950s and 1960s and the principles and techniques have changed very little since. In fact, the first editions of three aforementioned textbooks all appeared in the 1970s.

Although the principles and techniques for the creation of classificatory structures for the organization of bibliographic material have been explored in great detail, the basis for these principles and techniques is extant classifications. It is generally assumed that the classificatory structure should reflect some order that already exists; the reality and existence of these orders are often assumed to be a priori (cf. Mai 1999). The goal of bibliographic classification systems is to express somehow the already existing orders. These orders can be expressed through an ideology that the classification system serves, a particular social purpose that the classification system serves, and the order of the sciences or disciplines (cf. e.g., Langridge 1976; Miksa 1998).

Organization of the Web

Most bibliographic knowledge organization research concerned with improving access to Web documents has focused on enhancing the representation of individual documents and that mainly from a descriptive point of view, i.e., capturing the title, author, publisher, type, language, etc. of the document. The problems of representing the subject matter of Web documents have not been given the same amount of attention: Weinberg (1999) calls it "the stepchild of the metadata literature."

Most experimentation with bibliographic knowledge organization on the Web has been with classifi-

cation schemes[2] used in libraries. Saeed and Chaudry (2001) found that there are many possibilities for using the Dewey Decimal Classification system to organize the Web but concluded that more research in the area was needed. Schwartz (2001) has similarly found that there are many advantages to using known library classification schemes to organize the Web, for instance, the "institutionalization (an ongoing agency is responsible for maintenance and updating), familiarity to users and staff, and availability in machine-readable form" (Schwartz 2001, p. 76). Dodd (1996) and Vizine-Goetz (1996) have compared Yahoo!'s directory to traditional library classification schemes. Both found that the structures could be mapped to a certain degree but the terminology and arrangement differed.

Most Web directories, however, have not adopted a library classification scheme as their organizational structure, but have instead created their own structures: there is an immense interest in creating directories on the Web. Sherman (2000) reports that all major Web search services now offer directories and Lin and Chan (1999) noted that, "in the short history of Web searching, there is an increasing use of classification structures to impose order on the vast store of information. Increasingly, popular search services and directory services are employing hierarchical structures to organize Web resources" (Lin & Chan 1999, p. 155-156).

Despite the wide use of Web directories, their construction has not been guided by a set of principles such as those found in the bibliographic knowledge organization community. Williamson noted that, "It is essential that the principles of classification and classificatory structure be further explored and applied in conjunction with the Internet if it is to survive as an effective source of information" (Williamson 1997, p. 26). Williamson followed up on this paper in 2000 by taking a closer look at whether the guidelines for the construction of thesauri were adequate for the needs in the digital environment. Her answer was: "It doesn't appear so. Unequivocally, the answer is 'no'" (Williamson 2000, p. 272). Williamson further argues that more work is needed on how the guidelines for the construction of thesauri can be adapted to the digital environment. She finds that, "fundamentally to the work of revision, is the consideration of all relevant theoretical, intellectual, and technological issues. […] In particular, it is essential to address the problems of ambiguity, the effects of interdisciplinary, and the need for poly-hierarchical relationships" (Williamson 2000, p. 272). Hert, Jacob, and Dawson (2000) reached a similar conclusion when they investigated whether the traditional indexing theories and practices can be transferred to the digital environment. They found that the attempts to revise the guidelines "in light of the challenging information environment have not been successful" (Hert, Jacob, & Dawson 2000, p. 972).

It is clear that the community of bibliographic knowledge organization theory and practice has not been able to make its knowledge available to the Web directory community. It is also clear, however, that the principles developed in the bibliographic knowledge organization community need to be adapted in order to be useful in the digital environment.

Discussion

Much of the current practice in constructing search engines for the Web is based on research and development that have taken place in the information retrieval field over the past 50 years. However, there has not been the same interaction between the practices of constructing Web directories and the experience gained by the bibliographic classification community. The development of Web directories has taken place independently of knowledge organization research and practice possibly because the principles used in bibliographic classification research and practice face some challenges in being applied to the construction of Web directories.

There are two main challenges in applying bibliographic classification principles to the classification of the Web: 1) the principles are tied to the paper-based environment and 2) the principles assume that the material to be classified is scientific in nature.

The first challenge is that the bibliographic classification principles have been developed in the paper-based era and are to a large degree still tied to the paper-based environment (Clarke 2001). This is evident from at least three differences:

1) The principles for the construction of bibliographic thesauri and classification systems prescribe that the user first looks up the sought concept in a list of controlled terms and then goes to the documents. On the Web a user typically browses through a directory and accesses the documents directly by clicking the appropriate link.
2) The principles for the construction of bibliographic thesauri and classification systems often

advise that a notation is created to connect the different parts of the thesaurus or classification system. A notation is superfluous on the Web since the access mechanism and the documents are part of the same system.

3) The display issues in a paper environment are very different from those in the digital environment. The paper environment is constrained by the relative size of the book(s) in which the classification scheme is printed and by the fact that the user will have to leaf through a number of pages to see a particular topic area; however, the advantage is that the user has a sense of the relative size of the topic and the classification scheme. In the digital environment the user can navigate and jump between different topic areas relatively easy; however, the digital environment can give the sense of being lost and does not easily reveal the relative size of the topic and classification.

The challenge is to develop principles for the classification of documents in digital environments. These principles need to take into account the possibilities and limitations imposed by the digital medium.

The second challenge is that bibliographic classification theory and research has been primarily concerned with the organization of scientific or scholarly material. One of the guiding principles has been that any bibliographic thesaurus or classification system should reflect the order of the sciences as closely as possible because that will generate the most useful system. The relationship to the order of the sciences therefore has played a major role in bibliographic classification theory throughout the past century (cf. e.g. Miksa 1998). Web documents often represent ideas and information that are not part of the established order of the sciences. The sciences, therefore, cannot be used as a framework for the construction of Web directories. The material accessible via the Web deals with all kinds of topics and represents a greater portion of the universe of knowledge than simply the sciences.

The challenge is to develop principles for the construction of classification systems in environments that lack a clear already existing structure.

Bibliographic classification research has roots that go at least a century back and the classificatory principles have evolved over that time. However, for the principles to adapt to digital and non-scientific environments there need to be inquiries into the very foundation of classification theory and research. Bibliographic classification theory and practice has much to offer to the classification of the Web and it is also clear that there are some major challenges that need to be addressed for the principles to be applied in the Web environment.

Conclusion

For bibliographic classification research to contribute to the classification of the Web there need to be inquiries into the principles for classification in light of constraints and possibilities imposed by the Web. I have presented two major challenges that need to be addressed: the move from paper based environment to the digital environment and to construct classifications without an already existing structure. Both of these challenges imply inquiries that are much broader in scope and much more fundamental than simply classifying the Web. They illustrate a need to address the theoretical foundation of classification research and move the focus towards classifications based on needs and uses of the information rather than focusing simply on topicality (cf. Mai, 2004). This move, however, needs to develop and take shape as classification research moves forwards and expands into areas beyond traditional libraries.

Notes

1 Yahoo! (www.yahoo.com), The Open Directory Project (www.dmoz.org), and Zeal (www.zeal.org).

2 Cf. e.g. CyberStacks at www.public.iastate.edu/~cyberstacks/homepage.html; CyberDewey, at www.anthus.com/CyberDewey/CyberDewey.html; and CORC at www.oclc.org/ news/oclc/corc/index.html.

References

Aitchison, Jean, Alan Gilchrist & David Bawden. 2000. *Thesaurus Construction and Use: a Practical Manual*, 4th edition. Chicago: Fitzroy Dearborn

Bliss, Henry Evelyn. 1929. *The Organization of Knowledge and the System of the Sciences*. New York: Henry Holt

Clarke, Stella Dextra. 2001. Review of "Thesaurus Construction and Use: a Practical Manual" Jean Aitchison, Alan Gilchrist & David Bawden. *Knowledge Organization*. 28 (2): 100-102

Classification Research Group. 1957. The Need for a Facetted Classification as the Basis of all Meth-

ods of Information Retrieval. In: *Proceedings of the International Study Conference on Classification for Information Retrieval.* London: ASLIB.

Cutter, Charles A. 1876. *Rules for a Printed Dictionary Catalog*. Washington DC: US Government Printing Office

Devadason, Francis J., Neelawat Intaraksa, Pornprapa Patamawongjariya, and Kavita Desai. 2002. Faceted Indexing Based System for Organizing and Accessing Internet Resources. *Knowledge Organization*. 29 (2): 65-77.

Dodd, David. 1996. Grass-roots Cataloging and Classification: Food for Thought from World Wide Web Subject-Oriented Hierarchical Lists. *Library Resources and Technical Services*. 40: 275-286

Fidel, Raya. 1991. Searchers' Selection of Search Keys: II. Controlled Vocabularies or Free-Text Searching. *Journal of the American Society for Information Science*. 42 (7): 501-514

Hert, Carol A., Elin K Jacob and Patrick Dawson. 2000. A Usability Assessment of Online Indexing Structures in the Networked Environment. *Journal of the American Society for Information Science*. 51 (11): 971-988

International Organization for Standardization. 1986. *Documentation – Guidelines for the Establishment and Development of Monolingual Thesauri*. 2nd ed. Geneva: International Organization for Standardization. ISO 2788-1986

Lancaster, F.W. 1986. *Vocabulary Control for Information Retrieval*, 2 ed. Arlington, VA: Information Resources Press

Langridge, Derek. W. 1976. *Classification and Indexing in the Humanities.* London: Butterworths.

Lin, Xia and Lois Mai Chan. 1999. Personalized Knowledge Organization and Access for the Web. *Library and Information Science Research*. 21 (2): 153-172.

Mai, Jens-Erik. 2004. Classification in Context: Relativity, Reality, and Representation. *Knowledge Organization.* 31 (1): 39-48.

Mai, Jens-Erik. 1999. A Postmodern Theory of Knowledge Organization. In *Knowledge: Creation, Organization and Use. Proceedings of the ASIS Annual Meeting*, 36: 547-556.

Miksa, Francis L. 1998. *The DDC, the Universe of Knowledge, and the Post-Modern Library*. Albany, NY: Forest Press

National Information Standards Organization. 1994. *Guidelines for the Construction, Format, and Management of Monolingual Thesauri.* Bethesda, MD: NISO. ANSI/NISO Z39.19-1993 R1998

Panizzi, Sir Antonio. 1985. Mr. Panizzi to the Right Hon. the Earl of Ellesmere.--British Museum, January 29, 1848. In: *Foundations of Cataloging: A Sourcebook*. Michael Carpenter and Elaine Svenonius, eds. Littleton, Colo.: Libraries Unlimited.

Ranganathan, S.R. 1962. *Elements of Library Classification*. Bombay, India: Asia Publishing House

Ranganathan, S.R. 1967. *Prolegomena to Library Classification*, 3rd ed. Bombay, India: Asia Publishing House.

Richardson, Ernest Cushing (1935). *Classification: Theoretical and Practical*. New York: H.W. Wilson Co.

Rowley, Jennifer. 1994. The Controlled Versus Natural Indexing Languages Debate Revisited: a Perspective on Information Retrieval Practice and Research. *Journal of Information Science*. 20 (2): 108-119.

Saeed, Hamid and Abdus Sattar Chaudry. 2001. Potential of Bibliographic Tools to Organize Knowledge on the Internet. *Knowledge Organization*. 28 (1): 17-26

Sayers, W.C. Berwick. 1915. *Canons of Classification*. London: Grafton & Co.

Schwartz, Candy. 2001. *Sorting out the Web: Approaches to Subject Access*. Westport, CT: Ablex Publ.

Sherman, Chris. 2000. Google Introduces Web Directory Using Netscape's Open Directory Project Data. *Information Today*. 17 (5): 14

Soergel, Dagobert. 1974. *Indexing Languages and Thesauri: Construction and Maintenance*. Los Angles: Melville Publ. Co.

Svenonius, Elaine. 1986. Unanswered Questions in the Design of Controlled Vocabularies. *Journal of the American Society for Information Science*. 37 (5): 331-340

Vickery, Brian C. 1960. *Faceted Classification: a Guide to Construction and Use of Special Schemes*. London: Aslib

Vizine-Goetz, Diane. 1996. *Using Library Classification Schemes for Internet Resources*. OCLC Internet Cataloging Colloquium Position Paper. www.oclc.org/oclc/man/colloq/v-g.htm

Weinberg, Bella Hass. 1999. Improved Internet Access: Guidance from Research on Indexing and Classification. *Bulletin of the American Society for Information Science*. 25 (2): 26-29.

Williamson, Nancy. 1997. Knowledge Structures and the Internet. In: *Knowledge Organization for Information Retrieval. Proceedings of the Sixth International Study Conference on Classification Research*. The Hague, Netherlands: International Federation for Information and Documentation

Williamson, Nancy. 2000. Thesauri in the Digital Age: Stability and Dynamism in their Development and Use. In: *Dynamism and Stability in Knowledge Organization. Proceedings of the Sixth International ISKO Conference*. Wurzburg, Germany: Ergon