<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper published in *The international information & library review*. This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

Citation for the original published paper (version of record):

Golub, K. (2016)

Potential and Challenges of Subject Access in Libraries Today on the Example of Swedish Libraries.

The international information & library review, 48(3): 204-210

<http://dx.doi.org/10.1080/10572317.2016.1205406>

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-57557>

Column Title: Advances in Library Data and Access

Column Editor: Heather Moulaison Sandy, Assistant Professor, iSchool at the University of Missouri, Columbia, MO USA moulaisonhe@missouri.edu

The "Advances in Library Data and Access" column examines technological advances internal and external to libraries. The focus is on how library data is created and used. The strength of the column is its broad, international focus and contributors are encouraged to explore issues and recent advances in information technology relevant to their geographical region, as well as the larger, global audience. Interested authors are invited to submit proposals and articles to the column editor at moulaisonhe@missouri.edu. Please include "ILR Submission" in the subject line of the email.

Potential and challenges of subject access in libraries today on the example of Swedish libraries

AUTHOR: KORALJKA GOLUB

Associate Professor

Department of Library and Information Science, School of Cultural Sciences, Linnaeus

University, Sweden

koraljka.golub@lnu.se

Abstract: Ensuring quality subject access in information services is one of the major tasks in libraries and related information institutions. An exploratory study of Swedish library catalogs indicates that subject access is not addressed systematically, that in new digital collections knowledge organization systems (KOS) are applied to a limited degree, and in integrated library and commercial databases the mappings between the different knowledge organization systems do not exist. Possibilities are suggested to help alleviate these issues, such as social tagging and automated subject indexing; how to best implement them calls for further research.

Keywords: subject access points, library catalogs, social tagging, knowledge organization systems (KOS), Swedish libraries

Introduction

Subject searching (searching by topic or theme) is the most common and at the same time the most challenging type of searching in library catalogs, compared to, for example, a known-title or known-author search. Subject index terms taken from standardized knowledge organization systems (KOS), like classification systems and subject headings systems, provide numerous benefits compared to free-text indexing of commercial search engines: consistency through uniformity in term format and the assignment of terms, provision of semantic relationships among terms, support of browsing by provision of consistent and clear hierarchies (for a detailed overview, see, for example, Lancaster, 2003).

However, controlled subject index terms are expensive to produce manually and there is a huge challenge facing library catalogs and digital collections of various types as to how to provide high quality subject metadata for increasing numbers of digital information at reasonable costs, and in interfaces that bring together a number of different collections and databases, each using a different KOS. (Semi)-automated mapping and subject indexing and end-user indexing represent some potential solutions to retain the established objectives of library information systems in the light of the exponentially increasing numbers of digital documents and integration databases utilizing various KOS.

The paper discusses the challenges of subject access today and the potential of social tags and semi-automated solutions to improve subject access points in library catalogs. Examples are taken from Swedish academic and public libraries. It is structured as follows: section 2 (Desirable features to support subject access) presents what options for subject access points library catalogs should

support; section 3 (Ways to provide subject access) provides an overview of KOS, subject access points through end user tags and (semi-)automatically derived index terms; section 4 (Subject access in Swedish library catalogs) presents the status of subject access in online catalogs of Swedish academic and public libraries; section 5 (Concluding remarks) wraps up the problems and suggests potential solutions as well as guidelines for further research.

Desirable Features to Support Subject Access

Already early research on online catalog use shows that most end users conduct subject searching compared to, for example, author or title searching (see, e.g., Olson & Boll, 2001, p. 2). While the latter types of searching were more common in paper catalogs, they were superseded in frequency by subject searching in online catalogs, implying the need to offer a variety of functions to support the subject searching. Moreover, the challenges with subject searching are bigger than with known-item searching; for example, concepts take various names (synonyms) and same names can refer to different concepts (homonyms); names take various forms (e.g., plural or singular); it is more difficult to conceptualize an information need and transform it into a search query, compared to a known-item search.

Based on research related to first three generations of online library catalogs including first experimental catalogs of the third generation, like Okapi and CHESHIRE, a literature-based analysis of desired features with focus on subject access was conducted and discussed by Golub (2003). The desired features include the following:

- Browsing by subject access points: subjects from KOSs (e.g., subject headings, captions from classifications systems, classmarks with captions), free keywords, and tags.
- Searching by individual words from various metadata elements and by words from subject access points and full-text.
- Searching and browsing by facets and individual concepts and facets from KOSs, such as individual terms from subject headings and individual notations from combined classmarks.
- Combining the search strategies above, including allowing unique searches by major and minor themes represented by KOSs if supported by the indexing policy.
- Presenting and browsing multiple classification hierarchies for each component concept of a classmark.
- Auto-suggesting of authorized KOS versions of entered search terms, presenting all the relationships and allowing further choice on browsing or searching the KOS.
- Suggesting auto-completion of search terms once the user begins typing.
- Suggesting corrected versions of mistypes.
- Combining subject searching with searching by other bibliographic fields (e.g., author, publication year, medium).
- Linking to other bibliographic records containing the same subject access points.
- Highlighting search terms in retrieved metadata and resources.
- Advanced searching by Boolean and proximity operators, truncation, and stemming.

- Suggesting changes or describing automatic steps such as stemming if no results are returned.
- Combining previous search formulations.

In addition, users should be made aware of which factors led to retrieving the resources – e.g., whether this was based on finding query words in one field or a combination of them. Therefore, the query expansion the system may suggest should also be interactive, allowing the end user to choose, for example, broadening or narrowing the search query by terms suggested by the system, rather than automatically performed by the system and simply providing the retrieval results without informing the user how the search was performed.

Ways to Provide Subject Access

Why Knowledge Organization Systems (KOS)?

Different people, end users, and authors often use different terms to refer to the same concept (synonyms); many words have many meanings (polysems), some of them completely unrelated (homonyms). These are major reasons why algorithms of search engines do not retrieve wanted information in response to a search query. In addition, too many hits, or hits that are not specific enough are other common problems.

To address these problems, KOS have been designed and implemented to ensure that a concept is represented by only one index term, but equivalent / related / narrower / broader terms are cross-referenced to it. This enables the user to search using his or her own words in the query, which are then translated through the system of cross references, allowing the user to choose

the most appropriate or additional query terms (e.g., accepted equivalent form, broader, narrower, related). For example, a search for “cats” in the user query could be disambiguated by the system that would offer different concepts under the homonym: domestic cat, a tea type, a whip type, a vehicle type, etc.; it could offer narrower terms like a specific breed of domestic cat; broader terms like other pets, etc.

What happens when the end user does not know which search term to use at all? In a library building, there are many users who browse the shelves where books are often arranged topically, following a classification system. In an online catalog, a hierarchical browsing tree linked to all resources on the topic is therefore desirable. In addition, classmarks are in many cases representations of several different concepts; in faceted classification systems like Universal Decimal Classification (UDC) classmarks are built through explicit combination of facets, which can therefore be made searchable individually and in any post-coordinate manner in the catalog, but using captions (class or facet names) rather than class numbers and notation – the end user will not know the latter.

However, as KOS and manual indexing are resource intensive, many argue for using only full-text indexing as common in Web search engines instead. Because of problems mentioned above, the search results will not suffice, especially when it comes to subject searching.

Library of Congress > LC Online Catalog > Browse > Browse Subjects

Print Subscribe Share/Save

LC Online Catalog LC Online Catalog Quick Search Search Browse Advanced Search Keyword Search

< Revise Your Search Search History Account Info Help LC Authorities

Search Results - Browse Subjects

Select a specific heading to display the titles for that heading or follow a cross-reference to get to a more appropriate heading.

You Searched: SUBJECTS beginning with: macedonia

◀ Browse Subjects ▶

Macedonia. (34)	Library of Congress subject headings
Note: Here are entered works on the ancient country and kingdom of Macedonia, which later became a Roman province, as well as works on the region in the central Balkan Peninsula of somewhat indefinite boundaries, but including the Macedonia region of Greece, most of the middle Vardar Valley in the Macedonian Republic of the former Yugoslavia, and southwest Bulgaria west of the Mesta River. Works on the jurisdiction and the regions resulting from the division of Macedonia in 1912/1913 are entered under their own names.	
Macedonia--20th century--Biography. (1)	Library of Congress subject headings
Macedonia--Administrative and political divisions. (1)	Library of Congress subject headings
Macedonia--Administrative and political divisions--Maps. (1)	Library of Congress subject headings
Macedonia (Amazonas, Colombia)--Economic conditions. (1)	Library of Congress subject headings
Macedonia--Annexation. [from old catalog] (2)	Library of Congress subject headings
Macedonia--Antiquities. (38)	Library of Congress subject headings
Macedonia--Antiquities. (1)	not applicable
Macedonia--Antiquities--Bibliography. (1)	Library of Congress subject headings
Macedonia--Antiquities, Byzantine. (1)	Library of Congress subject headings
Macedonia--Antiquities--Catalogs. (2)	Library of Congress subject headings
Macedonia--Antiquities--Congresses. (2)	Library of Congress subject headings
Macedonia--Antiquities--Exhibitions. (5)	Library of Congress subject headings
Macedonia--Antiquities, Greek. (1)	Library of Congress subject headings
Macedonia--Antiquities--Periodicals. (2)	Library of Congress subject headings
Macedonia--Antiquities--Pictorial works. (1)	Library of Congress subject headings

Figure 1 shows an example of Library of Congress Subject Headings (LCSH) when browsing the Library of Congress catalog at <https://catalog.loc.gov/vwebv/searchBrowse> (selecting option “Subjects beginning with”, entering the query term “Macedonia”). The search retrieves a disambiguation note for the term stating that “Macedonia” is relevant for the Kingdom of Macedonia until 1912-1913 and acknowledges that it had shifting boundaries throughout this period, being located in what is now FYROM (Former Yugoslav Republic of Macedonia), Greece, and Bulgaria, and that for other periods different headings are used. This is then followed by an extensive list of specific subject headings, with see and narrower term references. For a further

discussion of issues with automatic full-text indexing and searching, see, e.g., Olson and Boll (2001, p. 48–51).

End User Index Terms

Social tagging, the term used to denote index terms added by end users, is yet another potential way to enhance subject access in digital collections. In addition to the fact that such subject access points are cheap to produce compared to resources needed for professional indexing, research has shown that tags offer several benefits. One often cited advantage would be that they are potentially more up-to-date, which is important, for example, for indexing documents about most recent scientific discoveries; another one is that they are expressed in end-user language, compared to KOS that for various reasons do not always include a sufficient amount of end-user terminology. Furthermore, they may provide multiple end-user perspectives, assign more specific topics than general subject headings would support (Lin et al., 2006), and allow for assignment of emotional tags that may in particular be popular for photographs (Choi, 2013).

However, disadvantages are also present. Firstly, the language control provided by a KOS is missing, causing problems with precision and recall. Therefore, it has been claimed how it may well be that savings made through low-quality indexing are “cancelled out by the high costs incurred by searchers who fail either to find everything that they want (low recall) or, often more frustratingly, to avoid everything that they do not want (low precision)” (Furner, 2010, p. 1861). Secondly, they are not assigned by trained indexers and therefore lack consistency, with some being only personally significant (e.g., “to_read”).

Finally, it has been shown that in existing services few end users actually do contribute (Ho & Horne-Popp, 2013).

One way to address the first issue is to provide taggers with a KOS as an initial suggestion for tags. A related study explored the use of Dewey Decimal Classification (DDC) with mappings to LCSH and showed that the KOS helped find focus for tagging, strengthened consistency and led to increase of access points in retrieval; 36% of additional resources could be found using end-user tags, and a bit more so when using index terms derived from the DDC, compared to the original manual indexing (Golub, Lykke, & Tudhope, 2014). Also, three times as many search terms were found in end-user index terms as in manually assigned controlled terms. In conclusion, both catalogers' index terms and end-users' index terms combined help improve retrieval.

Automatic Subject Indexing

Software vendors and experimental researchers speak of the high potential of automatic indexing tools. While some claim to entirely replace manual indexing in certain subject areas, others recognize the need for both manual (human) and computer-assisted indexing, each with its (dis)advantages. Reported examples of operational information systems include NASA's machine-aided indexing which was shown to increase production and improve indexing quality (Silvester, 1997); and the Medical Text Indexer at the US National Library of Medicine, which by 2008 was consulted by indexers in about 40% of indexing throughput (Ruiz, Aronson, & Hlava, 2008).

Research related to automated subject indexing or classification is spread around three major areas: document clustering, text categorization and

controlled-vocabulary based string matching (Golub, 2006). In document clustering, both clusters (classes) into which documents are classified and, to a limited degree, relationships between them, are produced automatically. Labelling the clusters is a major research problem, with relationships between them, such as those of equivalence, related-term and hierarchical relationships, being even more difficult to automatically derive (Svenonius, 2000, p. 168). In addition, “[a]utomatically-derived structures often result in heterogeneous criteria for category membership and can be difficult to understand” (Chen & Dumais, 2000, p. 146). Also, clusters’ labels, and the relationships between them, change as new documents are added to the collection; unstable class names and relationships are user-unfriendly in information retrieval systems, especially when used for subject browsing. Related to this is keyword indexing whereby topics of a document are identified and represented by words taken from the document itself (also referred to as derived indexing). Text categorization (machine learning) is the most widespread approach to automated classification of text. Here characteristics of subject classes, into which documents are to be classified, are learnt from documents with manually assigned classes. However, manually classified documents are often unavailable in many subject areas, for different document types or for different user groups. Also, text categorization algorithms only perform well on new documents if they are similar enough to the training documents.

In controlled-vocabulary based string matching, matching is conducted between a controlled vocabulary and the text of documents to be classified. A major advantage of this approach is that it does not require training documents, while still maintaining a pre-defined structure of the controlled vocabulary at

hand. If using a well-developed classification scheme, it will also be suitable for subject browsing in information retrieval systems. Apart from improved information retrieval, another motivation to apply controlled vocabularies in automated classification is to re-use the intellectual effort that has gone into creating such a controlled vocabulary.

Hard evidence on the success of automatic indexing tools in operating information environments is scarce; research is usually conducted in laboratory conditions, excluding the complexities of real-life systems and situations. Having reviewed a large number of automated indexing studies, Lancaster concluded that the research comparing automated versus manual indexing is “seriously flawed” (2003, p. 334). For a suggested comprehensive framework for evaluation of automatic indexing, see Golub et al. (2016).

Integrating Subject Access Across Collections

A variety of digital collections, like repositories, archival and museum collections, and journal databases can today be integrated with library catalogs. While integration and mapping are successful to the degree that the underlying metadata standards can be technically aligned, the largest challenge lies in integrating subject access fields. If quality subject access is to be preserved in integrated databases, subject access fields need to be merged, mapped to one another, or transformed to match the domain, language, or granularity.

The costs involved may be high if the mappings of the KOS involved do not pre-exist. For example, when mapping between KOS in different languages, what needs translating is concepts rather than terms, and many have different contexts in different countries or cultures – think, for example, about the concept

of an education system across countries. When it comes to mapping KOS between different communities, each community develops KOS specific to their needs; at the same time, end users want to use a single search interface to find resources in multiple databases serving different domains and different KOSs, across which there might be no consensus regarding concepts, terminology, and knowledge organization.

Subject Access in Swedish Library Catalogs

An exploratory study of Swedish academic and public libraries was conducted to determine common issues and discuss potential solutions. In total 46 catalogs were examined, out of which 5 catalogs provided by the National Library of Sweden; 20 university libraries of the biggest Swedish universities (counted by the number of full-time students at undergraduate and graduate levels), out of the total of 50 [nationwide? Kora – please address]; and, 21 public libraries, each representing the main library in each of the 21 Swedish counties. The analysis was conducted by opening the catalogs and examining possible searching and browsing options, in the period between 1 March to 15 April 2016. The observations were categorized in a spreadsheet file.

While detailed analysis will be reported at a later stage, the general observations will be discussed here. Most libraries offer a simple search interface as the starting option, which comprises a simple search box; however, in the vast majority of cases this comes without any direct guidance or instructions of how to conduct a search or what is being done by default. The black box approach imitates commercial search engines like Google, but end users will not learn how to improve their search unless help is provided. The

Help document tends to be available only once the advanced search interface is reached in academic libraries, while there are only a few public libraries that offer search help at any point. Most libraries offer advanced search interface which offers better control of the fields selected, and is often accompanied by support of Boolean operators, truncation and stemming options.

Of the libraries studied, the largest range of subject access functionalities is available at “Extended search” by LIBRIS, the union catalog by the National Library of Sweden (http://libris.kb.se/form_extended.jsp). Of subject access points, one can search on: keywords from all fields, words from the title, subjects and classification. The “subject” field refers to different KOS, including both the more traditionally used KOS in library catalogs, subject headings (e.g., SAO – Swedish Subject Headings, MeSH – Medical Subject Headings), but also thesauri like AGROVOC, a multilingual agricultural thesaurus maintained by the Food and Agriculture Organization of the United Nations. Other libraries in many cases use the term “Subject” but do not state what kind of KOS it is; in academic libraries the situation seems to be particularly complex because the catalog is often an integrated interface to access commercial databases as well as the library catalog, but in many cases it is not clear which KOS it is, and the mappings between them are largely missing.

In LIBRIS “Extended search” the “classification” field allows searching by classmarks of SAB (Swedish Library Classification System), DDC, and UDC; however, this requires knowing the symbolic notation which cannot be expected by end users – instead, it should be possible to search by class captions (actual terms and words from the natural language that are represented by classmarks) and relative index terms. LIBRIS is the only interface found in the sample that

supports subject browsing through SAB and DDC. The interface designs for browsing the two, however, are not identical. In addition to supporting browsing, both SAB and DDC support searching through their captions as well as their classmarks.

Neither in classification nor in the subject field are the different KOS mapped to one another, which prevents truly integrated cross-searching. This means that resources on a certain subject from one KOS that have been indexed using terms from another KOS, will not be retrieved in a query which the searcher only uses terms from the first. For example, searching by “LGBTQ” in Swedish Subject Headings will retrieve no results as the term does not exist, neither is it cross referenced from “HBTQ” which exists. If we search “LGBTQ” in KVINNSAM, we will get results; however, if we compare results retrieved using the latter query and “HBTQ” in KVINNSAM, we will get a different set of resources retrieved, both sets incomplete. Again, it cannot be expected on the end user to be acquainted with terms used by different KOS.

Furthermore, as seen from the example in Figure 2, there is an obvious loss of the specificity and granularity that KOS traditionally used by libraries have provided, for example in subject headings. Unlike when we search on Macedonia in LCSH (Figure 1), many academic library catalogs integrated with commercial database search do not provide any disambiguation (in our example how “Macedonia” is defined) or specific subtopics (examples from Figure 1: “20th century”, “Biography”, “Administrative and political divisions”, “Maps”).

Social tagging as an option is not supported by LIBRIS; there are a few public libraries in Sweden that allow this option. As to automatic indexing, some records in LIBRIS will have a DDC note stating “machine generated”, which refers to those records where mappings from SAB to DDC were produced automatically based on mapping tables created by human classifiers. However, the note is not explained to the end user. No other evidence of (semi-)automatic solutions were found in the sample.

Furthermore, the interfaces of newer digital collections have limited subject fields and it is not always obvious which KOS is supported, if any, for searching,

while subject browsing options are entirely absent. Examples of such newer collections that could benefit from improved subject access points include:

- SwePub, the Swedish national repository (<http://swepub.kb.se>);
- The catalog of SND (Swedish National Data Service) (<https://snd.gu.se/en/catalogue>);
- The Swedish Media Database (<http://smdb.kb.se>);
- The National Library of Sweden's collection of TV, radio, video, and movies presented in theaters, CDs, and multimedia; and,
- Sondera (<http://sondera.kb.se>), a cross-search service of LIBRIS, the Swedish Media Database and the Swedish National Archives.

Concluding Remarks

Ensuring quality subject access in information services is an important task of information-providing organizations. Today the seas of information resources make this task rather challenging; solutions like social tagging and (semi-)automated indexing support may help alleviate the problem and at the same time provide addition subject access points.

However, subject access in new digital collections and integrated search interfaces seems to be rather limited. The exploratory study of Swedish online catalogs shows that subject access is not addressed systematically, that in new services KOS are applied to a limited degree, and in integrated ones the mappings between the different KOS do not exist. While reasons for this are probably attributed to lack of resources, the final product for the end user is deficient even in basic subject access possibilities that were supported by subject

headings systems. Searching and browsing options for improved subject access, promoted by first researchers of online library catalogs, are few.

Options suggested to help alleviate these issues, such as social tagging and automated subject indexing are largely underexplored and unrecognized. For example, research dealing with Swedish DDC in relation to end-user indexing or automated classification does not exist. In addition, the combined value and relation of catalog librarians, expert end users, inexperienced end-users and automated indexers has not been explored before in general.

Further research is needed to gain a scientifically sound understanding of the level to which it is possible to apply automated subject indexing in the library contexts, using assigned indexing with KOS, derived indexing of free keywords from the resource at hand itself, as well as to determine the value of those automatically assigned index terms, in combination and comparison with end-user assigned index terms as well as catalogers' assigned index terms in the process of information retrieval by end users.

Acknowledgements

Many thanks to Madeleine Lundman, a Master-level student in Library and Information Science at Linnaeus University, who helped collect the data on Swedish public libraries. Special thanks to Miriam Nauri of the Swedish National Library, for an interview on subject access practice in the Swedish Union Catalog

References

Chen, H., & Dumais, S. (2000). Bringing order to the Web: Automatically categorizing search results. In T. Turner et al. (eds.), *Proceedings of the ACM*

- International Conference on Human Factors in Computing Systems* (pp. 145-152). New York: ACM.
- Choi, Y. (2013). Social indexing: A solution to the challenges of current classification. *Journal of Documentation*, 70(5), 107-135.
- Furner, J. (2010). Folksonomies. In M. J. Bates (ed.), *Encyclopedia of library and information sciences*, 3rd ed. (pp. 1858- 1866). New York: Taylor and Francis.
- Golub, K., Soergel, D., Buchanan, G., Tudhope, D., Hiom, D., & Lykke M. (2016). A framework for evaluating automatic indexing or classification in the context of retrieval. *Journal of the Association for Information Science and Technology*, 67(1), 3-16.
- Golub, K. (2006). Automated subject classification of textual web documents. *Journal of Documentation*, 62(3), 350-371.
- Golub, K., Lykke, M., & Tudhope, D. (2014). Enhancing social tagging with automated keywords from the Dewey Decimal Classification. *Journal of Documentation*, 70(5), 801-828.
- Golub, K. (2003). Predmetno pretraživanje u knjižničnim katalozima s web-sučeljem. MA thesis, University of Zagreb. Accessed July 21, 2014. <http://koraljka.info/publ/Magisterij-hrv.pdf>.
- Ho, B., & Horne-Popp, L. (2013). VuFind: An OPAC 2.0?. In J-R. Park, L. C. Howarth (eds.), *New directions in information organization* (pp. 159-71). Bingley: Emerald Group Publishing Limited.
- Lancaster, F. W. (2003). *Indexing and abstracting in theory and practice*. 3rd ed. London: Facet.

- Lin, X., Beaudoin, J. E., Bui, Y., & Desai, K. (2006). Exploring characteristics of social classification. In J. Furner and J. Tennis (eds.), *Proceedings of the 17th Annual ASIS&T SIG/CR Classification Research Workshop*. Accessed June 1, 2016, <http://journals.lib.washington.edu/index.php/acro/article/viewFile/12491/10990>
- Olson, H. A., & Boll, J. J. (2001). *Subject analysis in online catalogs*. 2nd ed. Orlando, FL: Academic Press.
- Ruiz, M. E., Aronson, A. R., & Hlava, M. (2008). Adoption and evaluation issues of automatic and computer aided indexing systems. In *Proceedings of the American Society for Information Science and Technology*, 46(1), 1-4.
- Silvester, J. P. (1997). Computer supported indexing: A history and evaluation of NASA's MAI system. *Encyclopedia of Library and Information Services*, 61(24), 76-90.
- Svenonius, E. (2000). *The intellectual foundation of information organization*. Cambridge, Mass.: MIT Press.