# Improving Search Quality by Enhancing Access to Metadata

David Haynes[1], Koraljka Golub[2], Claudio Gnoli[3], Athena Salaba[4], Ali Shiri[5], Aida Slavic[6]

[1] *Edinburgh Napier University, United Kingdom*
[2] *Linnaeus University, Växjö, Sweden*
[3] *University of Pavia, Italy*
[4] *Kent State University, Ohio, USA*
[5] *University of Alberta, Edmonton, Canada*
[6] *UDC Consortium, The Hague, The Netherlands*

**Abstract**

Has something been lost in the move towards federated searching and relevance ranking in academic library search interfaces? The richness of metadata and other knowledge organisation systems (KOSs) such as classification schemes and controlled vocabularies is often not featured in library management system (LMS) procurement. The ISKO Scientific and Technical Advisory Committee (STAC) has set up a working group to develop a set of guidelines for academic and research libraries that are planning to procure an LMS. The purpose of the guidelines is to enable users to benefit from the intellectual effort that has been invested in assigning metadata to electronic resources. This paper reports on progress of the working group in its review of the literature and the plans for consultations with stakeholders.

**Keywords** [1]

Metadata, resource discovery, knowledge organizing systems, library management systems.

## 1. Introduction

Online library catalogues do not typically make effective use of knowledge organisation systems (KOS) elements to support search and discovery of content by subjects or topics. Content indexing represents considerable intellectual effort by cataloguers but is often overlooked during the procurement process (and sometimes in system design). This paper is based on a review article arising from a concern that library and similar search systems have been 'dumbed down' and that there is limited scope for structured subject searching of resources. Most systems offer a simplified search interface and a 'black box' approach which automatically modifies searches to retrieve large data sets. Much of the effort is then invested in some kind of ranking system using different algorithms based on a combination of statistical analysis and machine learning to present the results. These solutions often do not make use of the existing subject metadata created to increase precision and relevance.

In order to help address these challenges, the International Society for Knowledge Organization (ISKO) has set up an international working group to develop a set of metadata guidelines for procurement of library catalogues (ISKO STAC working group on Subject Access Metadata, https://www.isko.org/stac/metadata). The purpose of the guidelines is to ensure that metadata-based search systems such as those used in libraries enables users to get maximum value from subject

ISKO UK conference, July 24–25, 2023, Glasgow, Scotland

EMAIL: d.haynes@napier.ac.uk (David Haynes); koraljka.golub@lnu.se (Koraljka Golub); claudio.gnoli@unipv.it (Claudio Gnoli); asalaba@kent.edu (Athena Salaba); ashiri@ualberta.ca (Ali Shiri); aida.slavic@udcc.org (Aida Slavic)
ORCID: 0000-0001-9191-9247 (David Haynes); 0000-0003-4169-4777 (Koraljka Golub); 0000-0002-4721-7448 (Claudio Gnoli); 0000-0002-1694-4052 (Athena Salaba); 0000-0002-6635-2282 (Ali Shiri); 0000-0002-1934-6824 (Aida Slavic)

CEUR Workshop Proceedings (CEUR-WS.org)

metadata comprising classifications and controlled vocabularies. Although in its initial scope this review of metadata use in searching is focused on academic libraries and related discovery systems, it is hoped that many aspects of the guidelines will be applicable to other digital library search interfaces/solutions, institutional repositories as well as information systems of cultural heritage institutions like archives or museums.

The paper outlines the problem of lack of access to subject metadata and other KOSs in library discovery systems. It highlights some of the issues that need to be addressed and explains the approach being used by the working group to develop a set of guidelines for the procurement of library management systems. It draws on a previous review by Golub [1].

## 2. Challenges of subject searching

While support for subject searching has been traditionally advocated for in library catalogues, notably since Cutter's [2] objectives for library catalogues, research shows that since the library automation from the 1980s onwards, subject access in online library catalogues has not been satisfactory [3], [4]. Library catalogues are part of library management systems (LMS), which can take different forms, including Web-based discovery services which serve as one-stop-for-all resources to which the library has access. More recent developments and adoptions of such systems try to match users' expectations by implementing Google-like single search box interfaces. However, as it is not possible to apply efficient web search engine ranking mechanisms, retrieval failures are common. Exploitation of intellectual effort that has been invested into subject indexing and classification is missing from these services.

Subject searching is an important requirement in online search systems such as library catalogues [5], [6], [7], [8], bibliographic databases [9], repositories [10], discovery services [11], online museums [12], [13], and related digital search services [14].

However, in comparison to known-item searching (e.g., queries for information objects whose title, author, etc. are known beforehand), searching by subject, even if all resources are available in digital form, often proves much more challenging. This may be due to:

- the difficulties of formulating queries with insufficient knowledge of the subject matter. One search box does not help the user see which KOSs are available beyond the search box;
- insufficient knowledge of the resources covered by the information system, resulting in an inability to use right search terms [16];
- insufficient knowledge of information searching (i.e. how to formulate a search query to reflect the information need);
- semantic ambiguities inherent to natural language: the same word can take on different meanings (polysemy) that could be completely unrelated (homonymy), while one concept can in turn be named using different words (synonymy);
- semantic ambiguities arising from multiple-word search terms;
- texts do not always explicitly name concepts that they write about. For example, searching for publications in the field of digital humanities will result in incomplete results because the term may not be used by an author who does not like that term or because the author works with digital archaeology and does not include the broader term digital humanities (or even the term digital archaeology as it may not been needed to mention;
- in many humanities disciplines and works of literary fiction, language is often metaphorical on purpose, related themes being intertwined with blurred boundaries between them;
- texts from different historical periods often use different terms due to lexical and grammatical changes for the same concepts than we do today, and these terms may also be expressed through contested historical language [17];
- older texts and manuscripts that have been digitised will also often have misspelt terms due to challenges with optical character recognition (OCR), resulting in not retrieving relevant documents or possibly false positives;

the problem is exacerbated with non-textual media that do not lend themselves to full text searching or do not have a narrative and are open to interpretation [18].

## 2.1. The role of KOSs

The general purpose of a KOS is to provide a means for organising information, through [15]:

- translation of the natural language of authors, indexers, and users into a vocabulary that can be used for indexing and retrieval;
- ensuring consistency through uniformity in term format and in the assignment of terms indicating semantic relationships among terms; and,
- supporting browsing by providing consistent and clear hierarchies in a navigation system supporting retrieval.

KOSs play a crucial role in resource retrieval and discovery. They improve the effectiveness of retrieval by helping to handle the sheer mass of available information. They also provide knowledge-based support for end users who access information without the help of an intermediary.

A well-structured KOS can be used as the knowledge base for an interface that can assist users with search topic clarification (e.g. through browsing well-structured hierarchies and guided facet analysis) and with finding good search terms (through query term mapping and query term expansion: synonyms and hierarchical inclusion).

Research has shown that KOSs are particularly needed in large databases covering broad areas of knowledge [16], [17] as well as in databases of primary sources [18] such as museum objects, which cannot be queried using full-text searches alone. Tibbo [17] makes the point that the exponentially increasing volume of information objects available online leads to information overload and entropy, rather than increasing benefit from access to information. Although full-text searching works for some tasks, for others it creates information overload and prevents the searcher from gaining a comprehensive overview of a topic: if a query returns thousands of retrieved documents, few searchers will browse beyond the first dozen or two hits.

Specific domains may require their own specialised indexing languages, rather than a one-size-fits-all approach [17] which then also requires a meta subject indexing language, usually called a ‘switching language’, that brings them all together in order to support searching across disciplines in an interoperable manner.

## 3. Library catalogues

Many researchers have addressed the problematic subject access to information in online library catalogues, pointing to continuing challenges for end users, for example, summarising a wide multi-year survey of Italian catalogues [19]. An overview, through a discussion of three generations of online library catalogues using the framework set by Hildreth [3], is given by Barton and Mak [20]. Key points are briefly presented below.

First generation online public library catalogues (OPACs) were developed with a focus on efficiency resulting from automation, rather than having service to end users in mind. Their functionalities were restricted to exact matching of known-item searches by author, title, or control number; effectively, this was a card catalogue in the online form. Second-generation online catalogues supported post-coordinate subject searching using Boolean operators, which, while an improvement in terms of functionalities, proved counterintuitive and hard to use. Third-generation catalogues were developed as experimental systems, e.g. Okapi and Cheshire, and research [5] concluded that their functionalities should include, among others, post-Boolean probabilistic searching, automatic spelling correction, term weighting, relevance feedback, output ranking, support for finding strategies.

Markey [4] provides ten reasons why these solutions were not applied to online library catalogues, among them: the failure of library systems' vendors to monitor shifts in information-retrieval technology and respond accordingly with system improvements; the failure of the research community to arrive at a consensus about the most pressing needs for online catalogue system improvement; decreasing funding and at the same time the high cost of integrated library systems.

As a result, by the time the World Wide Web became prevalent, OPACs were still second-generation catalogues, and the demand to implement functionalities of global search engines such as Google and other commercial services like Amazon, was increasing. These included a single search box, attractive web design, relevance ranking of results, recommendations, and access to a wide range of resources. However, Markey [4] argued that the new directions of developments towards simplification would not attract users back to the online catalogue. In integrated library catalogues each search would result in "millions of hits with no guarantee that the top-ranked ones will address your desired topic in depth or at your level of understanding" [4].

## 4. Metadata interoperability

Discovery services today predominantly operate on one integrated index of metadata from all resources involved. A single index provides faster retrieval compared to distributed searching which compiles information from different databases on the fly [23]. In order for this one central index to operate well, contributing metadata elements and its values need to be interoperable.

A move towards standardisation in order to bridge issues preventing unified search is NISO Open Discovery Initiative (ODI) [21], [22]. ODI creates a technical recommendation and model for data exchange, which allows libraries, as content providers, to work with discovery service vendors. Apart from simplifying the data exchange, ODI ensures that the vendors follow fair and unbiased indexing and linking practices.

Quality-controlled subject access in examined discovery services seems severely hindered. For an overview, see Golub [1]. This is in spite of the fact that huge resources have been allocated to adding subject index terms from indexing languages to library catalogue records. Little of this is adding value to existing interfaces. While imitating Google's black box approach, the task to retrieve relevant resources to a search query is addressed without making use of the existing index terms, relationships and structures of applied subject indexing languages.

In terms of the IFLA Library Reference Model (LRM) and the Functional Requirements for Subject Authority Data (FRSAD), the potential of controlled vocabularies has not been utilised to address the following user tasks:

- to find, as different resources are indexed using different controlled vocabularies, and also most probably following different indexing policies as they come from different collections of resources;
- to identify, as homonyms are not disambiguated, different perspectives are not disambiguated, at least not systematically by taking advantage of controlled vocabularies;
- to select, as aspects, facets or approach to the subject are not accounted for;
- to obtain, as useful resources are not located as a consequence;
- to explore, as it is not possible to, e.g., browse around related topics such as through using related terms in a thesaurus, or see narrower and broader terms or classes, in order to understand the relationships between various *nomens* for an entity.

## 5. Subject search functionalities on interface

In order to alleviate these problems, library catalogues and related information retrieval systems should employ a number of measures. They should include controlled subject terms from vocabularies such as subject headings systems, thesauri and classification systems, to help the user to, for example, choose a more specific concept to increase precision, a broader concept or related concepts to increase recall, to disambiguate homonyms, or to find which term is best used to name a concept. The ability of browsing classification schemes and other controlled vocabularies with hierarchical structures, would help the user further her understanding of the information needs and provide support to formulate the query more accurately. Interactive online help and instruction on information searching should teach users about search strategies, search techniques and query formulation. This would include thesaurus-based search term suggestions to support dynamic query reformulation and expansion and thesaurus navigation. To this end thesauri should be integrated into search engines as well as systems to support exact and partial matches to the query terms. They should also be used to support automatic and interactive search term selection for query formulation and expansion. The use of multilingual thesauri would support multilingual searching and browsing [23].

Golub et al [24] point to 18 functionalities common across cultural heritage institutions as well as three additional image-related ones. These include those related to KOS-based searching, browsing and indexing, as listed below in no particular order.

## 5.1. KOS-based searching

KOS-based searching functionalities include:

1. Searching using KOS concepts, including terms in the form of single or compound words, phrases, pre-coordinated headings, and class captions from classification systems.
2. Searching - using individual facets or concepts from KOS that compose a complex term (e.g., a class).
3. Searching using any combination of individual concepts and facets (as above);
4. Automatic alignment of user search terms into KOS terms (e.g., automatic synonym searching.
5. Disambiguation -- offering the user different concepts (e.g., are you looking for jaguar as an animal or Jaguar as a car?).
6. Linking any index term found in a metadata record to all other metadata records with the same index term.
7. Searching using major and minor themes represented in KOS.

## 5.2 KOS-based browsing

KOS-based browsing functionalities include:

8. Browsing by concepts from KOS, which is especially useful for those new to the document collection.
9. Browsing by facets, aspects and individual concepts from controlled vocabularies, such as individual terms from subject headings, as well as captions and notations representing individual concepts from synthesised classmarks (e.g., in Universal Decimal Classification).
10. Showing the narrower terms and broader terms to the search terms.
11. Displaying results in systematic order(s). As the function of classification notation is to control the sequential order of concepts [25], this should also be exploited to present results of a search in a meaningful way, making their examination and selection easier.

## 5.3 Enhancing KOS-based indexing

Some functionalities reflect the need to complement controlled vocabularies with other ways of information retrieval:

12. Searching using words from various metadata elements as well as from full-text.
13. Combining searching using controlled subject with searching using other bibliographic fields.
14. Adding, searching and browsing end-user tags. This allows  to consider end-user perspectives and  include most current terms from the literature.
15. Linking concepts from one KOS to other relevant ones. This calls for mapping across KOSs in order to support searching across different databases, including multilingual searching.

## 5.4 Image-related functionalities

KOS-based image-related functionalities include:

16. Searching using image- characteristics (e.g. size, orientation -- portrait/landscape).
17. Searching using content-based image retrieval (CBIR) methods (e.g. query by example image).
18. Searching using features enabled by IIIF (e.g. deep Zoom viewing).

# 6 The working group

The Subject- Access Metadata working group was initially suggested by Patrick Lambe and Koraljka Golub following a talk given to ISKO's Singapore chapter [26]. A small working group on Subject Access Metadata was created under the auspices of the ISKO Scientific and Technical Advisory Council (STAC) in 2022. The work of the group falls within the remit of the STAC Working Group on KO and subject access in information retrieval. Its members are: David Haynes (Edinburgh Napier University, United Kingdom), Koraljka Golub (Linnaeus University, Sweden), Athena Salaba (Kent State University, USA), Ali Shiri (University of Alberta, Canada), Claudio Gnoli (University of Pavia, Italy) and Aida Slavic (UDCC, The Netherlands).

The working group is publishing a literature review In the Knowledge Organization journal, which identifies the issues of subject searching and the reduction in subject access that has occurred from the early days of library automation. There are undoubted benefits in having a single search interface across different repositories and data sets, as well as across an entire library collection. The problem has become more evident with the greater access to digital resources such as electronic journals and digital versions of (or, increasingly, digital only) books.

The review will consider the main findings of user studies of discovery systems, particularly in academic and research libraries. It will also consider the experience of other collections such as public libraries, museums and galleries. What are the retrieval requirements and behaviours of their users? What are the barriers to good retrieval?

The knowledge organization community, and academic librarians are being consulted by questionnaire to identify the key issues around subject access. Using ISKO as a platform will make good use of the international nature of ISKO and its chapters as well as other professional and academic groups.

This will be followed up by online focus groups to discuss some of the issues raised during the survey to look forward to potential solutions. It will also be an opportunity to consider the main barriers to good subject access in discovery systems used by academic and research libraries.
The outcome will be a discussion paper for wider circulation in the academic and research library sector and among vendors, standards bodies, professional groups and user communities.

The final stage will be a draft set of guidelines for consideration by ISKO and IFLA and for eventual endorsement and publication. The intention is that this will generate debate about the issue of subject access, and the development of similar principles for other sectors such as museums and galleries, where there are particular concerns about image retrieval.

This could eventually lead to some kind of quality stamp or award to encourage uptake by vendors of library management systems and digital discovery tools. In time, the guidelines could be incorporated into the development of national or international standards for subject searching and resource discovery.

## 7 Conclusion

A review of the literature has demonstrated there has been a concern about lack of searchability of controlled vocabularies and access to KOSs in discovery systems since the first OPACs appeared. There has been some progress towards the development of a set of functionalities that discovery systems should have to improve the quality of searches and specifically to improve subject access to digital resources. The knowledge organization community as represented by ISKO members has a role to play in raising awareness of the subject access deficit in current discovery systems. Members' knowledge and experience of subject retrieval and contact with user communities will help to inform the development of guidelines for subject access in discovery systems.

## 8 References

[1] K. Golub, Subject Access in Swedish Discovery Services, Knowledge Organization, 45.4 (2018), 297–309. doi: 10.5771/0943-7444-2018-4-297.

[2] C. A. Cutter, Rules for a Printed Dictionary Catalogue, Washington DC: United States Bureau of Education, 1876.

[3] C. R. Hildreth, Pursuing the ideal: generations of online catalogs, in: Proceedings of a Library and Technology Association Preconference Institute, June 23-24 1983, American library Association, Chicago, IL, USA,1984, pp. 31–56.

[4] K. Markey, The online library catalog: paradise lost and paradise regained?, D-Lib magazine, 13.1, ( 2007) 4. doi: 10.1045/january2007-markey.

[5] P. Hider, Y.-H. Liu, The Use of RDA Elements in Support of FRBR User Tasks, Cataloging & Classification Quarterly, 51.8, (2013) 857–872. doi: 10.1080/01639374.2013.825827.

[6] R. N. Hunter, Successes and failures of patrons searching the online catalog at a large academic library: a transaction log analysis, RQ, vol. 30.3(1991) 395–402.

[7] L. Villen-Rueda, J. A. Senso, and F. de Moya-Anegón, The use of OPAC in a large academic library: a transactional log analysis study of subject searching, The Journal of Academic Librarianship, 33.3 (2007) 327–337. doi: 10.1016/j.acalib.2007.01.018.

[8] D. Wells, Online public access catalogues and library discovery systems, ISKO Enclyclopedia of Knowledge Organization. ISKO, 2020.

[9] S. Siegfried, M. J. Bates, and D. N. Wilde, A profile of end-user searching behavior by humanities scholars: The Getty Online Searching Project Report No. 2, Journal of the American Society for Information Science, 44.5(1993) 273–291. doi: 10.1002/(SICI)1097-4571(199306)44:5<273::AID-ASI3>3.0.CO;2-X.

[10] R. Heery, A. Powell, Digital repositories roadmap: Looking forward, Joint Information Systems Committee, Bath, 2006.

[11] K. Meadow, J. Meadow, Search Query Quality and Web-Scale Discovery: A Qualitative and Quantitative Analysis, College & Undergraduate Libraries, 19.2–4 (2012) 163–175. doi: 10.1080/10691316.2012.693434.

[12] M. Baca, Fear of Authority? Authority Control and Thesaurus Building for Art and Material Culture Information, Cataloging & Classification Quarterly, 38. 3–4 (2004) 143–151. doi: 10.1300/J104v38n03_13.

[13] C. Li Liew, Online cultural heritage exhibitions: a survey of information retrieval features, Program, 39.1 (2005) 4–24. doi: 10.1108/00330330510578778.

[14] M. Patel et al., Semantic interoperability in digital library systems, University of Bath, 2005.

[15] NISO, ANSI/NISO Z39.19-2005 (R2010). Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies, Baltimore, MD, 2010. doi: 10.3789/ansi.niso.z39.19-2005R2010.

[16] K. Markey, Twenty-five years of end-user searching, Part 1: Research findings, Journal of the American Society for Information Science and Technology, 58. 8 (2007) 1071–1081. doi: https://doi.org/10.1002/asi.20462.

[17] H. Tibbo, The Epic Struggle: Subject Retrieval from Large Bibliographic Databases, The American Archivist, 57. 2 (1994) 310–326. doi: 10.17723/aarc.57.2.f0650763x258t4p5.

[18] S. Bair, S. Carlson, Where Keywords Fail: Using Metadata to Facilitate Digital Humanities Scholarship, Journal of Library Metadata, 8.3 (2008) 249–262. doi: 10.1080/19386380802398503.

[19] E. Casson, A. Fabbrizzi, A. Slavic, Subject Search in Italian OPACs: an Opportunity in Waiting?, in: Subject Access: Preparing for the Future. IFLA Series on Bibliographic Control. Vol 42, P. Landry, L. Bultrini, E. O'Neill, and S. K. Roe, Eds. De Gruyter Saur, Berlin, 2011.

[20] J. Barton and L. Mak, Old hopes, new possibilities: Next-generation catalogues and the centralization of access, Library trends, 61.1, (2012) 83–106. doi: 10.1353/lib.2012.0030.

[21] J. Walker, The NISO Open Discovery Initiative: promoting transparency in discovery, Insights: the UKSG journal, 28.1 (2015) 85–90, 2015. doi: 10.1629/uksg.186.

[22] NISO, NISO RP-19-2020, Open Discovery Initiative: Promoting Transparency in Discovery, Baltimore MD, 2020. doi: 10.3789/niso-rp-19-2020.

[23] A. Shiri, Powering Search: the role of thesauri in new information environments, ASIST Mono. Medford, NJ: Information Today Inc, 2012. doi: doi.org/10.7939/R3J679046.

[24] K. Golub, P. M. Ziolkowski, G. Zlodi, Organizing subject access to cultural heritage in Swedish online museums, Journal of Documentation, 78.7 (2022) 211–247. doi: 10.1108/JD-05-2021-0094.

[25] C. Gnoli, Notation, ISKO Encyclopedia of Knowledge Organization, ISKO, 2018.

[26] D. Haynes, Doing Magic with Metadata, ISKO Singapore monthly meeting, Singapore, ISKO Singapore, 2021.