DiVA
http://www.diva-portal.org

This is the published version of a paper published in *Journal of Documentation*.

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

Golub, K., Szostak, R. (2025)
Information retrieval of humanities resources: subject searching from a user perspective
Journal of Documentation, 81(7): 376-398
https://doi.org/10.1108/jd-05-2025-0129

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-141916

Received 12 May 2025
Revised 12 August 2025
Accepted 15 August 2025

Information retrieval of humanities resources: subject searching from a user perspective

Koraljka Golub
iInstitute, Linnaeus University, Växjo, Sweden, and
Rick Szostak
Department of Economics, University of Alberta, Edmonton, Canada

Abstract

Purpose – This paper explores the longstanding disconnect between Knowledge Organization (KO) and Information Retrieval (IR), advocating for their integration to improve subject access in humanities and cultural heritage (CH) collections, including newer types of collections, such as those of research data.

Design/methodology/approach – Through a critical synthesis of literature, standards and recent advances in both KO and IR, the paper identifies key advantages and challenges and proposes a collaborative research agenda to address them.

Findings – While KO Systems (KOS) provide semantic depth and contextual accuracy, and IR systems offer scalability, their independent development has limited the effectiveness of search systems for humanities and CH collections. Today's operational search systems lack the capacity to support nuanced, exploratory search due to this disconnect. In addition, both KO and IR fields come with challenges which might be addressed via a complementary approach. The purposeful integration of KO and IR is necessary to address challenges such as opaque IR algorithms, underused or outdated KOS, and the need for context-aware, transparent and inclusive discovery environments.

Practical implications – Integrated KO-IR systems can support more accurate and inclusive discovery interfaces for libraries, museums and archives, as well as any search system, enhancing the visibility and usability of their resources.

Originality/value – The paper brings together perspectives from traditionally separate communities and calls for a de-siloed approach to designing subject access systems. It introduces key research questions and strategies for aligning KOS with advanced IR techniques.

Keywords Information retrieval, Knowledge organization, Controlled languages, Automated subject indexing, Social tagging, Knowledge organization systems

Paper type Conceptual paper

Introduction

The disciplines of Information Retrieval (IR) and Knowledge Organization (KO) are widely recognized as inherently complementary. However, the strengths and limitations of their respective approaches to organising and accessing information resources have been the focus of sustained debate over several decades. Rowley (1994) reviews the debate by first examining the dominant IR evaluation paradigm established through the Cranfield tests. These tests have been criticized for their artificial, laboratory-style conditions, reliance on binary and static relevance judgments and failure to account for user interaction or the evolving nature of information needs. More recently, Hider (2018) explains that content-based retrieval systems are effective at finding and obtaining digital resources but struggle with supporting user tasks like selection and exploration. This is because computers lack the broad real-world understanding that humans use to organize information meaningfully. Core KO textbooks

Journal of Documentation
Vol. 81 No. 7, 2025
pp. 376-398
Emerald Publishing Limited
e-ISSN: 1758-7379
p-ISSN: 0022-0418
DOI 10.1108/JD-05-2025-0129

© Koraljka Golub and Rick Szostak. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at Link to the terms of the CC BY 4.0 licence.

Downloaded from http://www.emerald.com/jd/article-pdf/81/7/376/10329593/jd-05-2025-0129en.pdf by Linneuniversitetet user on 13 October 2025

such as "The Organization of Information" by Joudrey and Taylor (4th ed., 2018) remain essential components of Library and Information Science (LIS) curricula worldwide as they comprehensively address new information organization standards, ensuring that IR teachings are complemented by robust KO principles. At the same time, many scholars contend that the automated methods characteristic of IR are adequate, particularly given the high costs associated with manual subject indexing through Knowledge Organization Systems (KOS), such as retrieval thesauri and subject heading schemes – even in the face of critiques regarding the decontextualized nature of IR evaluation (see e.g. Saracevic, 2008). In addition, patterns in research funding over recent decades reflect a pronounced bias towards computer science – the primary disciplinary home of IR – while LIS, traditionally associated with KO, has received comparatively limited support.

These developments have contributed to a growing divide between the two fields, evident not only in the separation of academic publication venues – such as distinct conferences and journals – but also in the limitations of contemporary search systems. These systems frequently rely on opaque, algorithmically driven ranking methods that offer users limited transparency or control over the search process. While increasing attention is being paid to enhancing the interpretability and transparency of AI-powered search systems, the prevailing reliance on so-called "black box" approaches remains a significant concern. This issue is particularly pronounced in collections of primary source materials used by humanities researchers, which are often unique and housed in archives and museums (Kamal and Golub, 2025). Furthermore, the challenges are compounded when dealing with multimedia or non-textual resources, which resist purely automated methods of retrieval. Even textual resources, such as literary works, present difficulties due to their frequent use of metaphor and indirect reference, which obscure explicit concept identification (see e.g. Golub, 2025 on LGBTQ + fiction).

However, the field of KO also faces a number of challenges. KOS are not always accurate or comprehensive representations of the world, and they are often criticized for being outdated, insufficiently maintained and non-neutral (Olson, 2002). Issues of consistency in indexing practices have also been widely documented (Leonard, 1977). Yet perhaps the most significant factor limiting the perceived value of KOS in end-user search is their underutilization within search interfaces (ISKO STAC Working Group on Subject Access Metadata, 2024; Golub *et al*., 2024). Even in systems where KOS are embedded in metadata records, their potential remains largely untapped at the user interaction level. As a result, search systems often fail to support critical functions such as disambiguating search terms, suggesting more specific or broader terms to refine or expand result sets and enabling users to navigate collections through topical structures – an experience analogous to browsing by subject in a well-organized public library.

To facilitate the development of more effective search services for humanities resources by leveraging the complementary strengths of IR and KO approaches, this paper seeks to identify key challenges associated with both domains, drawing on related literature and illustrative examples. It explores the potential contributions of KO to enhancing IR-based search systems, examines the requirements KO must meet to more effectively support IR and considers how IR methodologies can, in turn, be harnessed to strengthen KOS-based search functionalities within digital IR environments.

The remainder of the paper is structured as follows: the Background section provides an overview of users and their needs for subject access based on related research as well as the role of KOS and automated full-text retrieval; the Challenges section identifies problems of search and retrieval in both the KO and IR approaches; the Towards a Solution section outlines a possible path towards addressing concerns through collaboration between IR and KO; in the Concluding Remarks section, the main points are discussed and the need to collaborate with all the stakeholders emphasized, with a suggested future research agenda.

Downloaded from http://www.emerald.com/jd/article-pdf/81/7/376/10329593/jd-05-2025-0129en.pdf by Linneuniversitetet user on 13 October 2025

Background

Users

Subject-based searching remains one of the most common search behaviours among end users, despite being among the most challenging due to the inherent ambiguities of natural language. Evidence of this is found across a variety of digital search environments, including library catalogues (Hider and Liu, 2013; Hunter, 1991; Villén-Rueda *et al*., 2007), online museum platforms (Baca, 2004; Liew, 2004), bibliographic databases (Siegfried *et al*., 1993), institutional repositories (Heery *et al*., 2006), discovery services (Meadow and Meadow, 2012) and other related digital information systems (Patel *et al*., 2005).

In the context of the humanities, subject-based access to both primary sources (e.g. museum artefacts, archival materials such as diaries or letters and research datasets) and secondary resources (e.g. academic publications) is essential for a diverse range of user groups. These include academic researchers, students, educators (e.g. in preparing lesson plans or excursions) and members of the general public – including self-directed experts such as genealogy enthusiasts – as well as information professionals working in cultural heritage (CH) institutions. These professionals may, for instance, seek to increase the engagement and visibility of collections through augmented reality applications or digital storytelling (Frommholz *et al*., 2014).

The transition to digital platforms has significantly expanded the potential audience for cultural and scholarly content. As noted by The Getty Foundation (2012), while traditional print catalogues are typically targeted at scholarly audiences, the Internet affords institutions the opportunity to engage a much broader user base simultaneously. In their comprehensive literature review of user engagement with online archives, libraries and museums, Walsh *et al*. (2016) categorize users along three dimensions: (1) level of subject expertise, distinguishing between experts, semi-experts (including hobbyists) and non-experts; (2) type of information need, ranging from general visitors (seeking logistical information such as opening hours) to educational visitors (requiring content for lesson planning or project development) and specialist users (with a focused interest in specific collections); and (3) motivation and user role, identifying distinct categories such as explorers (motivated by curiosity), facilitators (supporting shared experiences), experience seekers (interested in intellectually or culturally significant content), professionals and hobbyists (with specific subject-related needs) and rechargers (seeking emotional or intellectual stimulation through art).

Additional research further highlights the breadth of user motivations for engaging with online CH resources. These include seeking inspiration, deriving enjoyment, accessing art-related news (Villaespesa *et al*., 2015), casual browsing (Fantoni *et al*., 2012) or simply passing time (Skov and Ingwersen, 2014; Walsh *et al*., 2018).

Cataloguing for subject access

The provision of subject access in the information systems of libraries, archives and museums is deeply rooted in longstanding cataloguing traditions. One of the earliest and most influential articulations of the objectives for subject access can be found in Charles Ammi Cutter's seminal work, *Rules for a Dictionary Catalog* (1876), in which he outlined what he termed the "objects" of the catalogue. Their purpose is to (1) enable finding an item of which the subject is known; (2) show what the library has on a given subject; and (3) assist in the choice of a book based on its topical character (Cutter, 1876, p. 5). Cutter's framework has had a profound and enduring impact, forming the basis for successive cataloguing codes and continuing to influence contemporary approaches to bibliographic control.

These foundational principles have been carried forward into modern conceptual models of bibliographic and authority data. Notably, the *Functional Requirements for Bibliographic Records* (FRBR), alongside its complementary models – the *Functional Requirements for Authority Data* (FRAD) and the *Functional Requirements for Subject Authority Data* (FRSAD) – have collectively formalized and extended Cutter's objectives in light of

Downloaded from http://www.emerald.com/jd/article-pdf/81/7/376/10329593/jd-05-2025-0129en.pdf by Linneuniversitetet user on 13 October 2025

contemporary needs. These three models were eventually consolidated into the *IFLA Library Reference Model* (IFLA LRM) (Riva *et al*., 2017), which provides a unified framework for understanding and implementing user-centric catalogue functionality. The IFLA LRM articulates five high-level user tasks – *find*, *identify*, *select*, *obtain* and *explore* – which serve as the conceptual basis for structuring cataloguing practices across diverse information environments.

Within the specific context of subject access, these five user tasks are closely aligned with the objectives of KOS-supported IR. As described in both IFLA LRM and FRSAD (Zeng *et al*., 2011), the tasks can be operationalized as follows:

(1) Find: to locate resources that embody works described by a specific subject label, for example, search using a nomen that is used in a subject headings system or a classification scheme.

(2) Identify: to clearly understand the nature of the resources found and to distinguish between similar resources, for example, those that are indexed by homonyms or those with the same topic but from a different perspective (e.g. different branches of a classification system like a virus from a zoological perspective vs a medical perspective).

(3) Select: to determine the suitability of the resources found and to choose (by accepting or by rejecting) specific resources that seem the most relevant, for example, because of certain aspects, facets or approach of the subject described.

(4) Obtain: to access the content of the resource.

(5) Explore: to use the subject relationships between one resource and another to place them in a context, for example, to browse around related topics by using related terms in a thesaurus or similar; or to see narrower and broader terms or classes to understand the relationships between various nomens for an entity. Examples include the following: examine the variant names for a subject within a controlled vocabulary, survey the variant terms used in different contexts of use, which may include different languages; explore correlations between nomens for the same entity in different controlled vocabularies, for example, finding a thesaurus descriptor which corresponds to a classification number.

These functions underscore the critical role of well-structured subject access in supporting not only discovery and retrieval but also contextualization and intellectual exploration – functions that are particularly salient in the domain of the humanities, where conceptual nuance and thematic interconnections are central to scholarly inquiry.

The imperative to ensure robust online access to both primary and secondary resources in the humanities has been a longstanding focus of professional organizations such as the International Federation of Library Associations and Institutions (IFLA), the International Council of Museums (ICOM) and the International Council on Archives (ICA). Additionally, research communities such as the International Society for Knowledge Organization (ISKO) have contributed significantly to advancing theoretical and practical developments in the field. These collective efforts reflect a shared commitment to enhancing subject access as a means of supporting research, education and public engagement with cultural and scholarly resources in the digital age.

The role of KOS in online searching

The effectiveness of online IR is fundamentally dependent on the quality and sophistication of the underlying search systems. While known-item searching – where users seek a specific resource based on identifiable attributes such as title, author or publication date – can often be accomplished with relative ease, subject-based searching presents considerably greater

Downloaded from http://www.emerald.com/jd/article-pdf/81/7/376/10329593/jd-05-2025-0129en.pdf by Linneuniversitetet user on 13 October 2025

challenges. These difficulties stem from a range of factors that inhibit users' ability to formulate effective search queries and retrieve relevant results.

One major obstacle is the user's limited familiarity with either the subject domain or the structure and content of the information system or collection being queried. This lack of domain knowledge can hinder the selection of appropriate search terms or prevent the user from conceptualizing the information need in terms compatible with the indexing vocabulary of the system. In addition, insufficient knowledge of search techniques or IR strategies further complicates the task, as users may struggle to articulate queries that accurately reflect their information needs.

A third and particularly complex set of challenges arises from the semantic ambiguities inherent in natural language. These include *polysemy* (a single term having multiple meanings), *synonymy* (different terms referring to the same concept) and *homonymy* (identical terms with unrelated meanings). Polysemy can result in the retrieval of irrelevant documents, especially in large databases, where ambiguous terms may retrieve thousands of results that must be manually filtered by the user. Synonymy places a significant cognitive burden on the searcher, who must anticipate and include a comprehensive range of equivalent terms in their query to achieve recall completeness. Homonymy often leads to the inclusion of false positives in the result set, further diminishing retrieval precision.

Semantic ambiguity also manifests in the handling of multi-word expressions and complex conceptual structures. For instance, a query such as *"history of philosophy"* may retrieve documents pertaining both to the historical development of philosophical thought and to the *philosophy of history* – a distinct field concerned with the philosophical analysis of historical narratives. Similarly, the conceptual content of texts is not always explicitly articulated; authors may imply, rather than state, the subjects of their works. For example, research in the digital humanities may not employ the term *"digital humanities"* explicitly, perhaps because the author prefers a more specific disciplinary label such as *"digital archaeology,"* or because they consciously reject the broader terminology. Consequently, relevant resources may be omitted from search results unless the user accounts for such implicit variation in terminology.

These challenges are especially pronounced in humanities disciplines and in genres such as literary fiction, where metaphorical language, thematic complexity and porous conceptual boundaries are characteristic features. Subject indexing in these contexts becomes significantly more difficult. In the case of LGBTQ + fiction, for example, it has been reported that even trained information professionals may overlook relevant works unless thematic indicators are overtly present, such as through book covers or external reviews (De la Tierra, 2008; Golub *et al.*, 2022a). This highlights the limitations of relying solely on surface-level features or metadata, particularly when dealing with nuanced, marginalized or intersectional content.

Historical variation in language further complicates subject-based access. Texts from earlier periods often employ obsolete or regionally specific terminology, lexical forms and grammatical structures that differ markedly from contemporary usage. In addition to these linguistic shifts, historical texts may contain expressions now regarded as contested or problematic, introducing further challenges for retrieval systems based on modern controlled vocabularies. Moreover, digitized older texts often suffer from inaccuracies introduced by optical character recognition (OCR) errors, resulting in misspellings or misinterpretations that hinder discoverability. Such errors may prevent the retrieval of otherwise relevant documents or, conversely, introduce irrelevant material into result sets.

The complexities of subject access are further amplified in relation to non-textual media. As Svenonius (1994) observed, subject representation in non-verbal forms such as art, music and material culture presents fundamental challenges for IR systems. In the case of textual documents, subject analysis is inherently linguistic: analysts interpret natural language content, apply conceptual frameworks articulated in language and assign index terms from controlled vocabularies. By contrast, many museum and heritage objects lack inherent linguistic content. Such objects only rarely include text-based narratives that can be directly

Downloaded from http://www.emerald.com/jd/article-pdf/81/7/376/10329593/jd-05-2025-0129en.pdf by Linneuniversitetet user on 13 October 2025

interpreted or indexed. Even those that do, such as inscriptions or symbolic imagery, require interpretive translation into conceptual language suitable for information systems.

The representation of such artefacts depends heavily on the ability of cataloguers or subject indexers to accurately perceive, interpret and translate visual or material properties into abstract subject descriptors. This task poses significant difficulties, even for experienced professionals and is largely beyond the current capabilities of automated IR systems. Additionally, museums and archives often curate highly heterogeneous and unique items, which differ significantly from the mass-produced resources commonly found in library collections. These singular objects – often lacking duplicates or well-defined bibliographic identities – necessitate particularly careful and nuanced subject description to ensure meaningful access and discovery.

Similarly, ensuring subject access to humanities research datasets requires careful indexing by information professionals because these datasets often contain minimal textual content, making full-text retrieval insufficient on its own. Unlike traditional publications, humanities datasets – such as digitized images or coded field notes – frequently lack the narrative context or keyword density needed for effective discovery through automated search tools. Controlled vocabularies, when applied thoughtfully, provide a structured, consistent way to represent the themes, disciplines and concepts embedded in these sparse resources. By leveraging their domain expertise, information professionals can assign precise subject terms that enhance findability and ensure equitable access to knowledge, especially for interdisciplinary or underrepresented research areas.

In summary, the inherent complexity of subject-based searching – particularly in the humanities and in relation to non-textual resources, including intangible CH, as well as research data – demands sophisticated, semantically rich search systems that go beyond the capabilities of simple keyword-based or full-text retrieval. Effective subject access must address linguistic variability, conceptual ambiguity, historical and cultural specificity and media heterogeneity. These considerations underscore the continued importance of expert-driven knowledge organization practices, including the development and maintenance of robust controlled vocabularies and highlight the limitations of relying solely on automated approaches in contexts where nuanced interpretation is essential.

Full-text searching

Users often value full-text searching because of its simplicity and immediacy – it allows them to enter natural language queries without needing to understand or navigate controlled vocabularies. This ease of use lowers the barrier to entry, making IR accessible to a broader range of users. Full-text search can be particularly effective when users are looking for specific, well-defined pieces of information, such as a date, name or direct quote, where the exact term is likely to appear verbatim in the text. In such cases, the system can quickly locate relevant documents or passages, providing a fast and satisfactory result. While the breadth of results may sometimes be overwhelming due to the system's literal matching of search terms, this same comprehensiveness ensures that users are less likely to miss pertinent materials simply because they lacked the correct vocabulary. Thus, while not ideal for all research tasks, full-text IR can serve as a powerful and user-friendly tool for targeted queries and exploratory searches.

However, a substantial body of research has demonstrated that full-text searching, while valuable in many contexts, is often insufficient to address the complex information needs of users, particularly in the humanities. Humanities research frequently involves interpretive, contextual and thematic inquiries that cannot be fully captured through simple keyword matching. In this regard, Knapp *et al.* (1998) argue that the most effective approach to searching humanities databases involves a hybrid strategy: combining free-text queries with the structured use of KOS. This dual approach leverages the strengths of both retrieval paradigms – free-text search allows for flexibility and natural language input, while

Downloaded from http://www.emerald.com/jd/article-pdf/81/7/376/10329593/jd-05-2025-0129en.pdf by Linneuniversitetet user on 13 October 2025

KOS-based indexing introduces semantic control and conceptual precision into the search process. By anchoring queries in a controlled vocabulary, users can navigate content more systematically and avoid issues such as polysemy, synonymy and terminological inconsistency that frequently undermine full-text, keyword-based searching.

The necessity of integrating KOS becomes even more pronounced in large-scale databases that span multiple disciplines or encompass a wide array of subjects. In such environments, the diversity and ambiguity of user queries, coupled with the heterogeneity of content, often exceed the capacity of full-text systems to return coherent and relevant results (Markey, 2007; Tibbo, 1994). Without the semantic scaffolding that KOS provide, users are left to rely on surface-level textual matches, which can lead to large, unfocused result sets. These issues are further compounded in digital collections comprising primary source materials, such as museum artefacts, archival documents, historical photographs or other non-textual objects. These resources are frequently accompanied by limited or non-standardized descriptive metadata and thus cannot be effectively queried using full-text techniques alone (Bair and Carlson, 2008). In such cases, subject indexing through KOS becomes essential for enabling access, discovery and contextualization.

Tibbo (1994) offers a critical perspective on the limitations of full-text searching by highlighting a paradox at the heart of digital IR: while the Internet and digital libraries have exponentially increased the volume of accessible information, this proliferation does not necessarily translate into increased informational benefit. On the contrary, the sheer abundance of unstructured content often results in information overload and conceptual entropy, making it more difficult for users to discern relevance, establish connections and gain a meaningful overview of a given topic. The lack of semantic organization leads to an excess of retrieved items that are only marginally relevant, reducing the efficiency of the search process and increasing the cognitive load on users.

Consequently, while full-text indexing may function adequately for narrowly defined, fact-based queries as mentioned above, it tends to falter when users seek broader, more abstract insights or wish to explore a conceptual domain holistically. For instance, thematic investigations into subjects like “colonial resistance,” “urban memory” or “aesthetic theory” often return voluminous result sets when queried via full-text search engines. Faced with thousands of documents, most users are unlikely to examine more than the first few pages of results, which are typically ranked according to opaque algorithmic criteria. As a consequence, highly relevant but less prominently ranked materials may remain undiscovered. This phenomenon not only diminishes the comprehensiveness of the user’s findings but also undermines the potential for serendipitous discovery, which is often a valued aspect of humanities scholarship.

By contrast, KOS – such as thesauri, classification systems and subject headings – can support users in navigating these challenges by providing a semantically structured entry point into the information space. They enable disambiguation of terms (e.g. distinguishing between “Jaguar” the animal and “Jaguar” the automobile), facilitate the identification of narrower or broader related concepts and support topical browsing that mirrors traditional practices of shelf browsing in physical libraries. In doing so, KOS not only enhance precision but also supports exploratory search strategies, which are particularly important in domains where users may not have well-formed queries at the outset. Thus, rather than replacing full-text searching, KOS should be seen as a necessary complement – particularly in information-rich environments where the goal is not merely retrieval but understanding, contextualization and discovery.

Challenges

Heterogeneous document types in museums and archives

The international ISO standard for subject indexing, first established in 1985 and reaffirmed in 2020 (International Organization for Standardization, 1985), outlines general principles and

Downloaded from http://www.emerald.com/jd/article-pdf/81/7/376/10329593/jd-05-2025-0129en.pdf by Linneuniversitetet user on 13 October 2025

techniques to be applied in subject indexing by "any agency in which human indexers analyse the subjects of documents and express these subjects in indexing terms" (International Organization for Standardization, 1985, p. 1). The standard defines a *document* as "any item amenable to cataloguing or indexing," explicitly including non-print media and three-dimensional objects or *realia*. According to the standard, manual subject indexing is a document-centric process comprising three sequential stages: (1) identifying the subject content of the document; (2) conducting a conceptual analysis to determine which aspects of the content should be represented; and (3) translating those concepts into terms from a KOS.

Nonetheless, varying types of information resources require additional contextual considerations. While the underlying principles of subject indexing for museum objects are analogous to those applied to printed publications (Will, 1993), museum professionals must also account for distinctive characteristics inherent to such objects. As mentioned above, many museum artifacts lack narrative content and are instead categorized based on material, form and function. Consequently, effective indexing necessitates capturing not only depicted motifs (*ofness*) and implied meanings (*aboutness*) but also the intrinsic nature of the object itself (*isness*) and its functional purpose (see Golub *et al.*, 2022b).

Ofness refers to what an object visually represents – elements discernible to a non-expert viewer, such as animals, plants or everyday objects – whereas *aboutness* pertains to the narrative, symbolic or thematic content of the work. These two dimensions reflect the dual nature of visual representation, as described by Jack (2001), wherein an image can simultaneously denote a tangible subject (e.g. an owl) and an abstract concept (e.g. wisdom). Both dimensions are vital for subject access: one user may seek visual representations of owls (*ofness*), while another may be interested in works conveying the theme of wisdom (*aboutness*). The interpretive complexity increases when considering culturally specific symbolism; for instance, while owls are traditionally linked to wisdom in Western iconography, they may symbolize death in many Native American traditions. Such nuances often elude automated IR systems and present challenges even for experienced human indexers.

Unique archival materials – such as unpublished manuscripts, personal letters, institutional records or ephemera – often resist standardized descriptive practices due to their singularity and contextual richness. At this level, collection-level description remains the dominant archival practice. Rather than item-level indexing, archives typically describe materials in aggregate, offering contextual information about the provenance, function and content of entire collections or sub-series. This is particularly true for large archival holdings, where individual description of every object would be prohibitively resource-intensive.

Consequently, controlled vocabularies or KOS are rarely applied at the item level for such materials. In many cases, even subject indexing at the collection level is minimal or absent, limiting discoverability through subject-based retrieval methods. The idiosyncratic nature of archival documents – often lacking titles, consistent formats or easily identifiable subjects – further complicates the use of traditional indexing tools.

High indexing specificity and faceted KOSs

Empirical research has underscored the importance of specificity in subject indexing, especially when aiming to mitigate issues of high recall and low precision in large-scale IR systems. Achieving this requires both (1) indexing policies that promote detailed, specific subject representation and (2) comprehensive, hierarchically deep indexing languages. The necessity of an extensive KOS stems from the diverse contexts in which any given topic might appear and the many disciplinary lenses through which subjects may be approached. Accordingly, while broad coverage is essential due to the diverse contexts in which any given topic might appear, as well as the myriad disciplinary lenses through which subjects may be approached, it should be informed and structured by domain-specific expertise (Tibbo, 1994).

Downloaded from http://www.emerald.com/jd/article-pdf/81/7/376/10329593/jd-05-2025-0129en.pdf by Linneuniversitetet user on 13 October 2025

Studies in bibliographic information systems further demonstrate that researchers across disciplines exhibit divergent requirements for subject access (Golub *et al*., 2020). In the humanities, for example, users of primary sources benefit from faceted vocabularies – such as the Art and Architecture Thesaurus – over pre-coordinated indexing systems like traditional subject headings. Faceted controlled vocabularies allow for greater specificity and enable the articulation of multidimensional facets relevant to humanities research, including geographical, chronological and disciplinary distinctions (Bates, 1996; Tibbo, 1994). However, the effective integration of such facets into search interfaces remains largely experimental, with limited practical implementation across mainstream digital platforms (e.g. Alani *et al*., 2000; Tudhope *et al*., 2006).

Biases in KOS and algorithms

Like most information structures, KOS are not neutral; they reflect the values, assumptions and worldviews of their creators. In the context of the arts and humanities – fields particularly attuned to issues of representation, voice and interpretation – these biases become especially consequential.

Numerous scholars have critiqued the inherent Western-centrism, coloniality and gender biases embedded in widely used systems such as the Library of Congress Subject Headings (LCSH) and Dewey Decimal Classification (DDC). For example, Olson (2002) argues that KOS often mirror dominant social ideologies, marginalizing or misrepresenting minority perspectives. In the LCSH, historically marginalized communities have been described using outdated or pejorative terms and the structures themselves often render non-Western knowledge systems invisible. A well-known example is the former use of the term "Illegal Aliens" in the LCSH (https://www.loc.gov/catdir/cpso/illegal-aliens-decision.pdf).

In the realm of art and CH, classification systems such as the Art and Architecture Thesaurus (AAT) and Iconclass have also been scrutinized for perpetuating Western-centric taxonomies. Iconclass, for instance, is based largely on Christian iconography and European visual culture, rendering it less effective – or even inappropriate – when applied to non-Western or Indigenous art traditions.

Moreover, KOS often operate under assumptions of objectivity and fixity, which conflict with the interpretive and evolving nature of humanities scholarship. Terms and categories used to describe artworks or historical events may shift over time as scholarly paradigms change. For example, the use of the term "primitive art" to describe African and Oceanic objects reflects a colonial perspective that has been increasingly rejected, yet traces of such terminology may persist in metadata.

In response to these issues, scholars and practitioners have advocated for more inclusive and participatory approaches to knowledge organization. Initiatives such as the Homosaurus (a linked data vocabulary for LGBTQ + terms) (https://homosaurus.org/) and community-driven projects like Local Contexts (which supports Indigenous metadata labels) (https://localcontexts.org/) aim to redress historical imbalances by centering marginalized voices in the development of vocabularies.

While some researchers have advocated for replacing the biases and inconsistencies of KOS and professional subject indexing with automated approaches (see e.g. Malmsten *et al*., 2024), these technologies are far from neutral. Automated systems, like their manual counterparts, are embedded with human decisions and socio-cultural assumptions that shape their outcomes. As Heider and Sundin (2019) argue, search engines – particularly widely used ones such as Google – are often mistakenly perceived as neutral tools. In reality, they are socio-technical systems influenced by various forms of bias: data bias, which reflects dominant cultural narratives; algorithmic bias, shaped by opaque ranking and personalization systems; and user bias, derived from prior knowledge, behaviours and search histories. These layers of bias collectively influence the visibility, credibility and reinforcement of information presented to users.

Downloaded from http://www.emerald.com/jd/article-pdf/81/7/376/10329593/jd-05-2025-0129en.pdf by Linneuniversitetet user on 13 October 2025

The challenges of neutrality in IR extend beyond commercial platforms and into domains such as CH. Here too, IR systems are shaped by the cultural frameworks and institutional priorities of those who design and maintain them. Relevance – the core concept in IR – is itself inherently subjective. As Saracevic (1975) notes, relevance varies across users depending on their intent, background and context. Algorithms tasked with determining and ranking relevance must therefore make value-laden decisions, inevitably privileging certain interpretations or uses of information over others. The training data that underpin IR systems further reinforce these biases. Collections often reflect institutional or disciplinary perspectives, embedding assumptions into the very structure of search and retrieval. Jo and Gebru (2020) and Noble (2018) have shown how such biases can reinforce dominant ideologies and marginalize alternative views, including those of underrepresented communities. Even in the absence of commercial imperatives, the design of ranking algorithms in non-commercial IR systems can introduce unintended biases. Decisions about which materials to prioritize – those with more metadata, from certain institutions or with higher usage – can lead to skewed representations. Collections from well-resourced institutions or with greater digital visibility may dominate search results, sidelining less visible but equally significant materials.

Ultimately, CH search engines are not neutral platforms. The design and implementation of such systems involve interpretive decisions that reflect broader cultural, institutional and epistemological values. Recognizing this lack of neutrality is not merely an academic concern; it is essential for building more inclusive and equitable information infrastructures.

Indexing consistency

The ISO 5963:1985 standard outlines manual subject indexing as a structured, three-step, document-oriented process. It begins with identifying the subject content of a document, followed by a conceptual analysis to determine which aspects of that content should be represented. Finally, these identified concepts are translated into a controlled vocabulary. This model emphasizes the internal characteristics of the document itself, prioritizing an objective representation of its topics.

In contrast, request-oriented indexing shifts the focus to the anticipated needs of users. Rather than indexing solely based on what a document *is about* – its aboutness – this approach also considers its potential *relevance*, that is, how the document might be used. Since aboutness is rarely explicit and often shaped by cognitive, contextual and cultural factors, the process of determining it is inherently subjective. Indexers must interpret meaning based on shared knowledge, user expectations and intended use cases, making the task deeply contextual.

The quality of subject indexing is closely tied to institutional policies, particularly in terms of exhaustivity, specificity and the intended target audience. Exhaustivity refers to the number of topics selected for indexing, where a high level includes many concepts and a low level includes only the most essential ones. Specificity concerns how precisely concepts are described, ranging from broad categories to more detailed, granular topics. Both factors must be aligned with the information needs of the intended user group – be it scholars, students or the general public.

Errors in manual indexing often arise from mismatches in the application of these policies or from inconsistencies in practice. Common issues include assigning too many or too few subjects (exhaustivity errors), using overly broad or narrow terms (specificity errors), omitting important topics or assigning subjects that are clearly incorrect. Consistency in indexing is a related challenge. Inter-indexer consistency – agreement between different indexers – and intra-indexer consistency – consistency in an individual indexer's work over time – vary significantly, with studies showing agreement rates ranging from 4% to 84%. Factors such as the number of subjects indexed and the size of the vocabulary available can decrease consistency. Importantly, high consistency does not guarantee high quality; indexers may be

Downloaded from http://www.emerald.com/jd/article-pdf/81/7/376/10329593/jd-05-2025-0129en.pdf by Linneuniversitetet user on 13 October 2025

consistently wrong and therefore consistency should not be taken as a sole indicator of indexing accuracy.

Automatic indexing attempts to replicate human judgment but encounters several limitations. Algorithms may miss relevant terms or assign them inappropriately, especially when nuances or context are involved. Evaluation of automated systems often relies on metrics like precision and recall, but these typically do not account for the indexing policies of specific collections. Inter-indexer consistency is sometimes used as a proxy for a gold standard in evaluation, yet this must be approached cautiously, given the different factors affecting consistency; also, while human indexers may be inconsistent, they only rarely assign erroneous subjects.

Interoperability

Interoperability of KOS is a critical concern in the development and functioning of cross-search databases. As digital collections become increasingly distributed across institutions, regions and disciplines, the ability to integrate and search across diverse metadata schemas and subject vocabularies becomes essential for supporting effective IR. Interoperability in this context refers to the capacity of different KOS to function cohesively, allowing for seamless cross-referencing and searchability of resources curated using different terminologies (Zeng and Chan, 2004).

Linked data frameworks, such as those based on the Semantic Web, have offered promising mechanisms for enabling such interoperability by representing KO structures in machine-readable formats and linking equivalent or related concepts across systems. Technologies such as the Simple Knowledge Organization System (SKOS) and OWL (Web Ontology Language) facilitate the expression of relationships between terms in different KOS, including equivalence, hierarchy and associative links. However, the implementation of these standards often requires substantial conceptual mapping work between vocabularies, which poses significant challenges due to differences in scope, granularity, cultural perspectives and linguistic expressions across KOS.

The process of mapping between KOS is inherently complex and resource-intensive. Conceptual mismatches – such as partial overlaps or non-equivalent terms – can lead to semantic inconsistencies or misinterpretations when users search across systems. Inaccurate or overly simplistic mappings may result in retrieval failures or misleading results, particularly when culturally specific concepts are forced into alignment with more generalized or dominant KOS structures (Vizine-Goetz *et al.*, 2004). Additionally, automated mapping approaches, while scalable, often lack the contextual sensitivity required to preserve the nuanced meanings embedded in certain terminologies, especially in fields such as CH.

For end users, the implications of poor or absent mapping are significant. When interoperability is not effectively established, as in discovery systems (Golub, 2018), users of cross-search systems may experience fragmented search results, redundant information or an inability to discover relevant resources located under alternative terminologies. This leads to frustration and reduced trust in the system's comprehensiveness or accuracy. Furthermore, users may be unaware of the underlying KOS structures or mappings, making it difficult for them to refine their queries or understand the reasons behind missing or irrelevant results. The opacity of these structures, combined with inconsistent application, can ultimately undermine the promise of integrated access to distributed collections.

Addressing these challenges requires sustained investment in the development and maintenance of high-quality mappings. It also necessitates a user-centered approach to interface design that makes the semantic relationships between terms transparent and navigable. As digital scholarship increasingly depends on cross-disciplinary and cross-institutional research infrastructures, the demand for interoperable KOS will only grow. Ensuring semantic coherence across systems is thus both a technical and epistemological

Downloaded from http://www.emerald.com/jd/article-pdf/81/7/376/10329593/jd-05-2025-0129en.pdf by Linneuniversitetet user on 13 October 2025

challenge, requiring collaboration across the KO and IR domains as well as the involved institutions and cultures.

Towards a solution

Automated subject indexing

The integration of automated subject indexing as a complement to professional human indexing has emerged as a promising strategy for improving the efficiency and scalability of subject access, particularly in large-scale and multidisciplinary digital collections, highlighting their ability to assist rather than replace human indexers in the subject metadata creation process (Golub, 2006; Golub *et al*., 2016). One of the key advantages of automated subject indexing lies in its ability to process large volumes of content with speed and consistency. Human indexing, while often more nuanced and context-aware, is time-consuming and resource-intensive, especially in environments where documents are rapidly accumulating, such as institutional repositories or digital archives. Automated tools can assist in the initial subject assignment, offering suggestions derived from training data or rule-based algorithms, which can then be verified or refined by professional indexers. This model – sometimes referred to as *semi-automatic indexing or machine-aided indexing* – can significantly improve indexing throughput while maintaining a high standard of quality when human expertise is applied in the final stage.

Additionally, automated subject indexing has demonstrated value in identifying terms that might be overlooked by human indexers. This capacity can be particularly useful in interdisciplinary or emergent research areas, where professional indexers may lack domain-specific familiarity or where controlled vocabularies are not fully developed. By a controlled integration of machine-generated suggestions, human indexers can broaden the scope of subject access and improve the discoverability of resources for a more diverse range of users.

Social tagging

The rise of participatory digital culture has brought along social tagging – or user-generated metadata – as a complementary approach to professional subject indexing. Social tagging allows users to assign descriptive keywords or "tags" to resources based on their own interpretations, needs and terminologies. This bottom-up form of KO offers a means of enriching and diversifying subject access in digital environments, particularly when used alongside established KOS.

Integrating social tagging into traditional subject indexing practices in libraries has shown promise, especially in areas where common controlled vocabularies may fall short. This is particularly relevant for the indexing of fiction, where themes often extend beyond conventional metadata categories such as genre, time period and geographical setting. Social tagging systems can capture more nuanced and reader-driven interpretations, enriching subject access by highlighting thematic, emotional or identity-based elements of works. As Johansson and Golub (2019) have demonstrated, user-generated tags contribute to a more granular and diverse thematic representation of fiction, offering perspectives that are often overlooked in professional indexing.

This complementary function of social tagging is also evident in the context of LGBTQ + literature. Adler's (2009) comparative study of twenty books on transgender topics in WorldCat and LibraryThing reveals significant discrepancies between professional and user-generated metadata. Terms commonly used within LGBTQ + communities – particularly those relating to specific gender identities and expressions – were frequently underrepresented or entirely absent from the Library of Congress Subject Headings used in WorldCat records. In contrast, LibraryThing's user-generated tags captured a broader and more contemporary vocabulary, offering more inclusive and representative subject access. This suggests that

Downloaded from http://www.emerald.com/jd/article-pdf/81/7/376/10329593/jd-05-2025-0129en.pdf by Linneuniversitetet user on 13 October 2025

folksonomies can fill important gaps in controlled vocabularies, particularly in domains where language is evolving rapidly or where mainstream systems lag.

Bates and Rowley (2011) similarly found that LibraryThing's tagging system improved the discoverability and visibility of LGBTQ + materials. They emphasized, however, that social tagging is not without its limitations. While folksonomies can democratize subject representation, they also reflect the cultural and demographic biases of their contributors. In LibraryThing, the predominance of US-based users resulted in the marginalization of certain ethnic minority perspectives, highlighting the importance of critically evaluating whose voices are represented in tagging communities.

Despite these advantages, significant challenges remain in integrating social tagging with professional subject indexing. One key issue is semantic inconsistency: tags tend to be uncontrolled, non-hierarchical and highly variable, which can lead to problems of synonymy, polysemy, homonymy and spelling variation, leading to retrieval failures. Furthermore, not all users tag resources with the same level of insight or intention; some tags may be vague, idiosyncratic or irrelevant from a retrieval standpoint. Efforts to reconcile social tagging with KOS have proven fruitful in experimental contexts (Golub *et al*., 2014), but these methods remain unapplied in operational systems.

Phenomenon-based classification

Related to the need for faceted vocabularies in the humanities and CH mentioned above, a relevant solution to both subject indexing and subject searching is using phenomenon-based classification (for a recent literature overview of the topic, see Gnoli *et al*. (2024). Such classifications, grounded in the phenomena we study rather than in disciplines, have many advantages for the project of this paper. Most obviously, phenomenon-based classifications pursue facet analysis: subject headings are formed by synthetically combining terms from different schedules. Most classification schedules are of phenomena, but there are also schedules of verb-like relators and adjective/adverb-like properties. Szostak (2017) had noted that museums and art galleries have not found existing systems of library classification congenial as they seek to make their collections accessible. A phenomenon-based classification, precisely because it had lengthy schedules of "things," is much better suited to providing subject headings for items that a museum or gallery possesses. And a synthetic approach to subject headings allows a museum or gallery to readily capture what is special (or perhaps typical) about a particular object: a golden ceremonial axe, for example.

Note that such a synthetic subject heading can potentially grapple with *ofness* as well as *aboutness* (Szostak, 2015 described how authorial perspective could be captured synthetically). Indeed, an approach to classification that allows us to freely combine terms from schedules of things, relators and properties can allow us to potentially capture many aspects of a work of art in a subject heading: style, provenance, form, technique (Szostak, 2014a). Note here that many museum artefacts are both aesthetically appealing and useful: a subject heading may wish to treat elements of both. A classifier can decide which characteristics of an artefact to stress in a subject heading – but can only do so if they have access to a KOS that allows the formulation of complex synthetic subject headings. As noted elsewhere, there are legitimate concerns that classifiers may be biased in what and how many subject headings they choose to emphasize: We can aspire to identify unbiased principles to guide these choices. Yet we can note here that facilitating the choice of an expansive set of (individually complex) subject headings will reduce the risk that important subjects are ignored. Users may not find every artefact with a particular form or meaning or political aspect – if the classifiers of particular objects have not stressed these elements – but should be able to readily find a set of objects for which classifiers did stress the particular characteristic that a user seeks. If classifiers have the time and inclination to choose lengthy subject headings, users will be guided to even larger sets of relevant objects.

Downloaded from http://www.emerald.com/jd/article-pdf/81/7/376/10329593/jd-05-2025-0129en.pdf by Linneuniversitetet user on 13 October 2025

At present, museums and galleries often turn to art-specific sources of controlled vocabulary, such as the Getty Museum Art and Architecture Thesaurus. Yet we want classifiers to be readily able to capture important political, economic, technological or cultural (or other) characteristics of a work. It is thus invaluable to have schedules specific to art nested within a broader classification so that classifiers can capture how art relates to other aspects of human existence. This becomes even more important when we turn to humanities scholarship, for humanities scholars should and do investigate how art influences and is influenced by all aspects of the human condition. The central importance of art to the human experience is lost if we can not readily capture the myriad influences on and of art (Szostak, 2014a).

Existing enumerated library classifications have inherently illogical hierarchies precisely because enumerated subject headings are complex. Recycling is thus treated as a subset of garbage even though it is something we do to garbage rather than a type of garbage (Mazzocchi *et al.*, 2007). Within a phenomenon-based classification, things have separate schedules from relators and properties and each schedule is logically constructed. The logical nature of such classifications should facilitate their integration with IR techniques. In particular, a logical structure should facilitate the application of automatic subject indexing.

In the KO literature, a standard critique of post-coordinated approaches to classification (that is, one in which a subject heading is created synthetically by combining terms from different schedules rather than choosing a complex heading from one schedule as happens with the classifications used in almost all of the world's libraries) is precisely that a search for "history of philosophy" will receive many false hits for philosophy of history. But this argument assumes that we cannot instruct a search engine to prioritize the order in which search terms are entered. Renwick and Szostak (2020) explored the possibility of just such an interface. This is one area in which the expertise of IR can usefully be combined with KO: to create a search interface that can achieve precision in searches involving synthetic subject headings.

Szostak (2017) imagined a search interface that would allow a user to see how a particular search term was nested within a broader classification. That is, a user should be able to easily move towards broader or narrower terms – a task that will be much more useful when navigating a KOS composed of logically structured schedules. A user should be able to see alternatives for each term in a synthetic subject heading. If you have found a gold ceremonial axe, you may become curious about bronze ceremonial axes or battle axes or ceremonial spears. Ideally, you should be able to add terms to your subject heading, such as how to build a ceremonial axe or how to use it. Note here that we could work towards an interface that worked across databases, allowing the user to move from admiring a museum artefact towards an article in a library describing how it was made or used.

A logical KOS that is easy to navigate should also facilitate social tagging employing controlled vocabulary. As noted elsewhere, social tagging has been shown to be valuable in a variety of applications, but its value is limited by idiosyncratic use of terminology. Users could be encouraged to employ the terminology of a phenomenon-based classification.

A further advantage of a synthetic approach should be noted. An important strand of KO scholarship emphasizes the biases inherent in the enumerated classifications utilized in the world's libraries. The managers of those classification systems have been slowly removing some biases rooted in the 19th-century genesis of these classifications. Szostak (2014b) has noted that a synthetic approach is automatically far less biased. An enumerated classification may create a subclass of "male nurse" and thus treat male nurses far differently from female nurses. A phenomenon-based classification forms a subject heading of "male nurse" synthetically in precisely the same way as "female nurse" (or indeed a nurse of any gender), by combining terms from a gender schedule and an occupation schedule. Similarly, works of art from all countries or regions of the world can be classified with respect to a classification of countries and regions that does not privilege one region over another.

The point here is not just to advocate for developing KOSs grounded in facet analysis. The point here is that we need KO and IR expertise working together. As the "philosophy of

Downloaded from http://www.emerald.com/jd/article-pdf/81/7/376/10329593/jd-05-2025-0129en.pdf by Linneuniversitetet user on 13 October 2025

history" example shows, we can make misguided decisions about KO development if we do not design KOSs with user interfaces in mind. The same precise problem exists on the IR side. IR scholars should not imagine that they are limited in the design of search interfaces by existing imperfections in KOSs. Rather, we should work together in developing combinations of interface and KOS.

We argued in previous sections that users needed to be guided with precision to both works of art and humanities scholarship. And we recognized that users could have a very wide range of search queries. We thus need an interface/KOS combination that can deliver precision across the widest possible range of queries. Yet it is also crucial that a user then be able to seamlessly move on to identify related works (of art or scholarship). We hope in this section to have identified a path forward that can achieve these goals.

A further collaboration with museum and gallery curators and archivists is called for. What sort of user needs do they recognize and what sort of interface/KOS combinations appeal to them? Importantly, what sort of tagging can they provide for the items in their collections? We should design our interface/KOS with the capabilities and needs of museums and galleries in mind. (Note here that phenomenon-based classifications are well-suited to providing terminology for the Semantic Web, whose RDF Triples take a thing/relator or property/thing format. Museums and galleries in tagging items for an interface/KOS could simultaneously tag them for the Semantic Web).

Automated information retrieval

In addition to the need to complement IR with KOS for humanities resources as discussed above, it is important to point out that the field of IR has witnessed a significant transformation in recent years, particularly with the advent of deep learning and transformer-based architectures (Ferrando *et al*., 2024). Neural network-based retrieval models have emerged as the state of the art, demonstrating superior performance over traditional methods, especially in contexts that demand semantic matching beyond exact term overlap. One of the most prominent developments in this regard is the shift from sparse vector representations – where document and query vectors are aligned with vocabulary indices – to dense retrieval models such as Dense Passage Retrieval. These models encode documents and queries as low-dimensional continuous vectors, enabling efficient similarity computation and improved retrieval accuracy (Karpukhin *et al*., 2020).

Similarly, recent advancements in computer vision and deep neural networks have significantly enhanced capabilities in image retrieval, enabling more sophisticated and accurate retrieval methods across various application domains (Vo *et al*., 2019). Zhou *et al*. (2017) highlight the increasing importance of deep learning techniques within Content-Based Image Retrieval (CBIR). As a subset of machine learning (ML) under the broader umbrella of artificial intelligence, deep learning allows for the extraction of high-level, abstract features from images. These features often correspond more closely to human visual perception and cognitive processing than traditional hand-crafted features. In the context of CH and museum collections, this technological shift opens up new possibilities – not only for the automatic recognition of objects (*ofness*) and individuals (e.g. through facial recognition), but also for more sophisticated interpretations, such as detecting mood, emotional expression or atmospheric qualities. This high-level semantic understanding is particularly valuable in the humanities, where visual materials often convey complex, non-literal meanings that are not easily captured by conventional indexing methods.

A prominent area of innovation within image retrieval is the development of cross-modal and multimodal retrieval systems, which allow for the querying of images through different modalities. Cross-modal retrieval typically involves the use of a textual input to retrieve visual content but also includes approaches such as sketch-based retrieval or query-by-example, where a sample image is used as the input query. In multimodal image retrieval, the query itself may combine multiple forms of input – most commonly an image paired with textual

Downloaded from http://www.emerald.com/jd/article-pdf/81/7/376/10329593/jd-05-2025-0129en.pdf by Linneuniversitetet user on 13 October 2025

descriptors. Such approaches have been employed to support complex retrieval tasks such as query completion or the refinement of search results in digital repositories containing scanned artefacts or CH images (Jaiswal *et al*., 2021). This multimodal alignment opens promising avenues for addressing challenges related to *aboutness*, especially in domains such as CH or symbolic imagery, where visual content often conveys abstract or culturally situated meanings. For example, an image of an owl might be associated with the concept of "wisdom" in Western cultural contexts. By embedding both the image and the associated textual concept in a shared representational space, multimodal retrieval systems can potentially support more conceptually rich queries – such as retrieving images that symbolically represent "wisdom" – even when those concepts are not explicitly depicted in a literal sense. This highlights the potential of multimodal techniques to bridge the gap between user intent and content representation, offering new opportunities for semantically nuanced image retrieval in complex or interpretive domains.

Applying dense retrieval approaches to domains such as the humanities, particularly within the CH online catalogues and collections, offers promising opportunities to derive meaning from unstructured and semantically rich content. In this context, dense retrieval can be seen as a bottom-up method for structuring knowledge, particularly where existing indexing is sparse or inconsistent. However, the success of these methods is contingent on the availability of large-scale, high-quality training data – something that is often lacking in CH collections. Humanities datasets frequently include rare or domain-specific terminology that is underrepresented in general corpora, which presents a major obstacle to the effective learning of word embeddings or document representations. This challenge is further exacerbated in the case of smaller institutions or repositories with limited digitized content, resulting in insufficient data to train robust models capable of capturing nuanced semantics.

Furthermore, implementing state-of-the-art IR systems in these contexts requires substantial expertise. Neural and transformer-based retrieval systems not only demand large-scale datasets but also involve complex architectures, model tuning and domain adaptation. These processes typically require interdisciplinary collaboration between computer scientists, linguists and domain experts – resources that may not be readily available in many CH institutions. The technical and infrastructural barriers are especially pronounced in smaller or underfunded organizations, where staffing and computational resources are often limited.

Concluding remarks

This paper addresses the growing disconnect between the fields of IR and KO, despite their widely acknowledged complementarity. It responds to the persistent challenges facing each field – such as the limited transparency and contextual sensitivity of automated IR systems and the underutilization and maintenance issues of KOS – by arguing that more effective search systems, particularly for humanities and CH collections, require a purposeful integration of both approaches.

Users of CH and humanities collections are diverse, ranging from academic researchers and educators to hobbyists and casual browsers. Subject-based searching remains a fundamental method for users to discover both primary and secondary resources. It enables users to find specific materials, explore topical content and contextualize information. Users differ widely in expertise, needs and motivations – ranging from lesson planning to curiosity-driven exploration – yet all benefit from reliable subject access that helps them navigate complex, often unique collections.

Subject-based searching remains one of the most complex yet essential modes of IR, especially in the humanities. Users often face significant challenges in articulating their information needs due to unfamiliarity with subject domains, limited knowledge of indexing vocabularies or a lack of search expertise. These difficulties are further compounded by the semantic ambiguities inherent in natural language, such as polysemy, synonymy and

Downloaded from http://www.emerald.com/jd/article-pdf/81/7/376/10329593/jd-05-2025-0129en.pdf by Linneuniversitetet user on 13 October 2025

homonymy, as well as by the implicitness of concepts in scholarly texts. These issues are especially pronounced in domains such as literary studies or LGBTQ + fiction, where meaning is often metaphorical or indirect, and in historical materials, where outdated or OCR-compromised language adds further barriers. Non-textual resources in museums and archives present additional challenges, as they require interpretive translation into text – an area where current automated systems fall short. In these contexts, well-developed KOS offer a crucial tool for enhancing subject access, enabling precise disambiguation, thematic exploration and conceptual navigation.

While full-text search has enabled flexible access to digital collections, it often lacks the semantic depth needed to support the interpretive and exploratory nature of humanities research. Simple keyword searches may retrieve large, unfocused result sets or miss key resources altogether due to terminological variation or implicit subject content. These limitations become especially problematic in large, interdisciplinary databases or collections of primary sources with sparse metadata. Without the structure provided by KOS, users may experience information overload and find it difficult to identify relevant materials. KOS offer a vital complement by introducing semantic control, supporting disambiguation and enabling users to browse and refine their queries systematically. This is particularly valuable in exploratory search contexts, where users may not begin with well-defined queries. By combining the strengths of full-text retrieval and structured subject access, integrated systems can offer more meaningful, nuanced and effective tools for scholarly discovery in the humanities and CH sectors.

While KOS play a critical role in enabling subject-based access, they face challenges. One of the main issues lies in the complexity and diversity of the resources they aim to describe – ranging from non-verbal artefacts and archival documents to contemporary fiction and multimedia. These materials often lack standard structures or explicit subject cues, requiring sophisticated, nuanced interpretation by indexers. Even with established guidelines like ISO 5963, the indexing process can be inconsistent due to subjectivity, cultural context or policy variations. Moreover, traditional KOS, such as LCSH or Iconclass, have been critiqued for inherent biases that marginalize non-Western perspectives and evolving terminologies. Newer approaches like phenomenon-based classification offer a more flexible and inclusive alternative by allowing synthetic subject headings that better reflect multidimensional attributes of cultural objects. Yet such systems demand thoughtful implementation and must be accompanied by user-friendly search and browse interfaces to unlock their full potential. Interoperability between vocabularies, another pressing challenge, remains difficult due to conceptual mismatches and the high resource cost of mapping, further complicating cross-institutional search and discovery.

Contemporary IR systems – particularly those powered by neural networks and dense retrieval models – offer advanced capabilities for semantic matching and improved retrieval. These technologies are well-suited for content-based retrieval and can support interpretive tasks in visual and textual domains. However, they are not without significant drawbacks. Dense retrieval approaches require massive, domain-specific training data, which is often unavailable or sparse in humanities and CH collections. Additionally, deep learning systems introduce new layers of opacity and potential bias, mirroring the socio-cultural assumptions of their training data and algorithmic design. This is especially problematic in domains where representation and context are crucial, as biased algorithms risk reinforcing dominant ideologies and marginalizing minority perspectives. Technical and resource constraints further hinder adoption in underfunded institutions and the implementation of these systems demands interdisciplinary collaboration that many organizations may be ill-equipped to sustain.

Given the respective limitations of both KO and IR approaches, a compelling case emerges for their integration in order to provide more effective, inclusive and user-responsive search environments for humanities and CH resources. A hybrid approach should include the development of search interfaces that can interpret synthetic subject headings, navigate hierarchically structured vocabularies and dynamically adapt to user queries. This, in turn,

Downloaded from http://www.emerald.com/jd/article-pdf/81/7/376/10329593/jd-05-2025-0129en.pdf by Linneuniversitetet user on 13 October 2025

requires the development of faceted KOSs with logical hierarchies, which should be developed to mesh with search interfaces. These coordinated innovations in IR and KO require close collaboration between IR and KO specialists, as well as active engagement with CH professionals and diverse user communities to design systems that reflect real-world needs and values. Social tagging and semi-automated indexing also offer valuable augmentation, particularly when used to complement expert knowledge. Ultimately, by uniting the strengths of both fields, we can create search systems that are not only functionally robust but also ethically and epistemologically attuned to the demands of digital scholarship and inclusive cultural memory.

To fully realize the potential of subject-based search in the digital age, there is an urgent need to dismantle the persistent silos that separate the communities of KO and IR. While KO has developed sophisticated, semantically rich frameworks over centuries – such as those codified in the IFLA Library Reference Model and supported by organizations like IFLA, ICOM, ICA and ISKO – these have often evolved separately from the rapid, data-driven advancements in IR, largely driven by the computer science community and disseminated through venues such as Special Interest Group on Information Retrieval (SIGIR), Text Retrieval Conference (TREC), European Conference on Information Retrieval (ECIR) or ACM International Conference on Information and Knowledge Management (CIKM). The result is a fragmented landscape where deep semantic structures and scalable search technologies exist in parallel but rarely intersect meaningfully.

To bridge this divide, a structured and sustained collaboration agenda must be initiated between the KO and IR communities. This agenda could begin with joint workshops and special issues across major venues – bringing together ISKO, IFLA, ICOM and ICA scholars and practitioners with researchers active in SIGIR, TREC and related initiatives. Collaborative research projects and funding applications should be explicitly designed to combine IR and KO strengths. A key goal should be the development of hybrid models that incorporate KOS into IR workflows, such as integrating subject hierarchies into retrieval models or leveraging faceted vocabularies in semantic query expansion. Equally important is the co-design of interfaces that draw from both traditions – providing users with both relevance-ranked results and structured pathways for exploratory browsing, contextual navigation and disambiguation. We have outlined many potential synergies between KO and IR in this paper, but we strongly suspect that further unimagined synergies will emerge if IR and KO experts collaborate in the development of search interfaces. In such a setting, scholars can share the practical difficulties they are facing, and colleagues can suggest constructive solutions.

Critical research questions include:

(1) How can KOSs be embedded in IR models to improve interpretability and conceptual precision?

(2) What methods best support the automatic or semi-automatic mapping of user queries to KOS terms in interdisciplinary, multimodal and multilingual contexts?

(3) How can we design user interfaces that enable transitions between structured browsing and unstructured search while preserving semantic coherence?

(4) How can we develop, implement and sustain the best KOS, and how can to best combine KOSs with social tagging and automated subject indexing?

(5) What strategies can support the development of bias-aware and culturally inclusive KOS–IR frameworks, particularly for underrepresented or non-Western knowledge systems?

(6) How can interoperability across KOS and IR systems be operationalized to support cross-institutional, cross-domain discovery environments?

Downloaded from http://www.emerald.com/jd/article-pdf/81/7/376/10329593/jd-05-2025-0129en.pdf by Linneuniversitetet user on 13 October 2025

We would note that these are largely empirical questions. The theoretical case for a useful collaboration between KO and IR is strong, but what is needed is clear empirical evidence that we can together address user needs better than we can separately.

Practical steps towards this integration should go beyond high-level calls for collaboration and focus on building tangible, reusable solutions. One priority is the joint design of KOSs that are optimized for machine processing as well as human interpretation, ensuring that vocabulary structures, hierarchies and relationships can be seamlessly incorporated into IR pipelines. We should actively encourage IR systems to employ KOSs as core components of their architecture, rather than as optional add-ons, so that semantic richness becomes an integral part of retrieval. This may involve developing interoperable, API-accessible vocabularies that can feed directly into ranking algorithms and query expansion. Pilot implementations could test faceted search interfaces enriched by both automated indexing and expert curation. Cross-domain mapping tools –combining machine learning with human oversight – could support discovery across institutions using different vocabularies.

A shared research infrastructure – including open datasets, annotated collections, KOS mappings and reusable evaluation frameworks – should be prioritized. Embedding these developments within open, shared infrastructures would ensure that innovations are replicable, scalable and accessible to institutions of varying sizes, making the KO–IR partnership a practical reality rather than an abstract ideal. Ultimately, the de-siloing of KO and IR is not only a technical imperative but also a cultural and epistemological one. It is essential for building retrieval systems that support not only efficiency and scalability but also equity and transparency.

Acknowledgments

We would like to thank Ingo Frommholz and Adam Jatowt for their valuable discussions on IR, particularly, in the context of the humanities and CH.

References

Adler, M. (2009), "Transcending library catalogs. A comparative study of controlled terms in library of congress subject headings and user-generated tags in LibraryThing for transgender books", *Journal of Web Librarianship*, Vol. 3 No. 4, pp. 309-331, 10.1080/19322900903341099.

Alani, H., Jones, C. and Tudhope, D. (2000), "Associative and spatial relationships in thesaurus-based retrieval", in Borbinha, J. and Baker, T. (Eds), *Proceedings (ECDL 2000) 4th European Conference on Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science*, Springer, Berlin, pp. 45-58, doi: 10.1007/3-540-45268-0_5.

Baca, M. (2004), "Fear of authority?: authority control and thesaurus building for art and material culture information", *Cataloging and Classification Quarterly*, Vol. 38 Nos 3-4, pp. 143-151, doi: 10.1300/J104v38n03_13.

Bair, S. and Carlson, S. (2008), "Where keywords fail: using metadata to facilitate digital humanities scholarship", *Journal of Library Metadata*, Vol. 8 No. 3, pp. 249-262, doi: 10.1080/19386380802398503.

Bates, M.J. (1996), "The Getty end-user online searching project in the humanities: report no. 6: overview and conclusions", *College and Research Libraries*, Vol. 57 No. 6, pp. 514-523, doi: 10.5860/crl_57_06_514.

Bates, J. and Rowley, J. (2011), "Social reproduction and exclusion in subject indexing: a comparison of public library OPACs and library thing folksonomy", *Journal of Documentation*, Vol. 67 No. 3, pp. 431-448, doi: 10.1108/00220411111124532.

Cutter, C.A. (1876), *Rules for a Printed Dictionary Catalogue*, Government Printing Office, Washington.

De la Tierra, T. (2008), "Latina Lesbian subject headings", in Roberto, K.R. (Ed.), *Radical Cataloging. Essays at the Front*, McFarland, Jefferson NC, pp. 94-102.

Downloaded from http://www.emerald.com/jd/article-pdf/81/7/376/10329593/jd-05-2025-0129en.pdf by Linneuniversitetet user on 13 October 2025

Fantoni, S.F., Stein, R. and Bowman, G. (2012), "Exploring the relationship between visitor motivation and engagement in online museum audiences", Museums and the Web, available at: https://www.museumsandtheweb.com/mw2012/papers/exploring_the_relationship_between_visitor_mot

Ferrando, J., Sarti, G., Bisazza, A. and Costa-jussà, M.R. (2024), "A primer on the inner workings of transformer-based language models", *arXiv: 2405.00208*

Frommholz, I., Graves, D., Liu, H., Kumar, A. and Brady, G. (2014), "Great war stories told by the people – crowdsourced cultural heritage in digital museums", *Proceedings Digital Libraries 2014 ACM/IEEE Joint Conference on Digital Libraries (JCDL 2014) International Conference on Theory and Practice of Digital Libraries (TPDL 2014)*, pp. 419-420, doi: 10.1109/jcdl.2014.6970204.

Gnoli, C., Smiraglia, R. and Szostak, R. (2024), "Phenomenon-based classification", *Annual Review of Information Science and Technology*, Published online January 9, 2024, doi: 10.1002/asi.24865.

Golub, K. (2006), "Automated subject classification of textual web pages", *Journal of Documentation*, Vol. 62 No. 3, pp. 332-349, doi: 10.1108/00220410610666564.

Golub, K. (2018), "Subject access in Swedish discovery services", *Knowledge Organization*, Vol. 45 No. 4, pp. 297-309, doi: 10.5771/0943-7444-2018-4-297.

Golub, K. (2025), "Challenges in AI: indexing LGBTQ+fiction", DHNB 2025, available at: https://dhnb.eu/wp-content/uploads/2025/03/Schedule_DHNB2025.pdf

Golub, K., Lykke, M. and Tudhope, D. (2014), "Enhancing social tagging with automated keywords from the Dewey Decimal Classification", *Journal of Documentation*, Vol. 70 No. 5, pp. 801-828, doi: 10.1108/JD-05-2013-0056.

Golub, K., Soergel, D., Buchanan, G., Tudhope, D., Lykke, M. and Hiom, D. (2016), "A framework for evaluating automatic indexing or classification in the context of retrieval", *Journal of the Association for Information Science and Technology*, Vol. 67 No. 1, pp. 3-16, doi: 10.1002/asi.23600.

Golub, K., Tyrkkö, J., Hansson, J. and Ahlström, I. (2020), "Subject indexing in humanities : a comparison between a local university repository and an international bibliographic service", *Journal of Documentation*, Vol. 76 No. 6, pp. 1193-1214, doi: 10.1108/JD-12-2019-0231.

Golub, K., Bergenmar, J. and Humlesjö, S. (2022a), "Searching for Swedish LGBTQI fiction : challenges and solutions", *Journal of Documentation*, Vol. 78 No. 7, pp. 464-484, doi: 10.1108/JD-06-2022-0138.

Golub, K., Ziolkowski, P.M. and Zlodi, G. (2022b), "Organizing subject access to cultural heritage in Swedish online museums", *Journal of Documentation*, Vol. 78 No. 7, pp. 211-247, doi: 10.1108/jd-05-2021-0094.

Golub, K., Gnoli, C., Haynes, D., Salaba, A., Shiri, A. and Slavic, A. (2024), "Library catalog's search interface: making the most of subject metadata", *KO Knowledge Organization*, Vol. 51 No. 3, pp. 169-186, doi: 10.5771/0943-7444-2024-3-169.

Heery, R., Lyon, L., Tsinaraki, C., Brody, T., Koch, T. and Doerr, M. (2006), "Report on digital repositories: an evaluation study on the development and implementation of community repositories to support research (and learning and teaching)", DELOS2 Network of Excellence on Digital Libraries, Deliverable 5.1.1, available at: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=19458833E5FA5CC45117776085BD0E87?doi=10.1.1.101.5976&rep=rep1&type=pdf

Heider, J. and Sundin, O. (2019), *Invisible Search and Online Search Engines: The Ubiquity of Search in Everyday Life*, Routledge, Milton Park, Abingdon, OX and New York.

Hider, P. (2018), *Information Resource Description : Creating and Managing Metadata*, 2nd ed., Facet, London.

Hider, P. and Liu, Y.H. (2013), "The use of RDA elements in support of FRBR user tasks", *Cataloging and Classification Quarterly*, Vol. 51 No. 8, pp. 857-872, doi: 10.1080/01639374.2013.825827.

Downloaded from http://www.emerald.com/jd/article-pdf/81/7/376/10329593/jd-05-2025-0129en.pdf by Linneuniversitetet user on 13 October 2025

Hunter, N.R. (1991), "Successes and failures of patrons searching the online catalog at a large academic library: a transaction log analysis", *RQ*, Vol. 30 No. 3, pp. 395-402, available at: https://www.jstor.org/stable/25828813

International Organization for Standardization, ISO (1985), Documentation – methods for examining documents, determining their subjects, and selecting indexing terms (5963:1985), available at: https.

ISKO STAC Working Group on Subject Access Metadata (2024), "aaa", available at: https://www.isko.org/stac/metadata

Jack, C. (2001), "State of the arts: current applications for indexing images", available at: https://web.archive.org/web/20010210080529/http://www.slis.ualberta.ca/599/cjack/599.htm

Jaiswal, A., Liu, X. and Frommholz, I. (2021), "Multimodal query completion for digital libraries", *Proceedings of the 21st ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 297-300, doi: 10.1109/JCDL52503.2021.00069.

Jo, E.S. and Gebru, T. (2020), "Lessons from archives: strategies for collecting sociocultural data in machine learning", *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT20)**, pp. 306-316, doi: 10.1145/3351095.3372829.

Johansson, S. and Golub, K. (2019), "LibraryThing for libraries : how tag moderation and size limitations affect tag clouds", *Knowledge Organization*, Vol. 46 No. 4, pp. 245-259, doi: 10.5771/0943-7444-2019-4-245.

Joudrey, D.M. and Taylor, A. (2018), *The Organization of Information*, 4th ed., Libraries Unlimited, Santa Barbara, CA.

Kamal, A.M. and Golub, K. (2025), "Subject matters : metadata standards and subject access for library and museum catalogues", in *The Hermeneutics of Bibliographic Data and Cultural Metadata*, pp. 204-239, available at: https://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-136374

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D. and Yih, W.-t. (2020), "Dense passage retrieval for open-domain question answering", *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, pp. 6769-6781, doi: 10.18653/v1/2020.emnlp-main.550.

Knapp, S.D., Cohen, L.B. and Juedes, D.R. (1998), "A natural language thesaurus for the humanities: the need for a database search aid", *The Library Quarterly*, Vol. 68 No. 4, pp. 406-430, doi: 10.1086/603001.

Leonard, L.E. (1977), "Inter-indexer consistency studies, 1954-1975: a review of the literature and summary of study results", Occasional papers (University of Illinois at Urbana-Champaign. Graduate School of Library Science); no. 131.

Liew, C.L. (2004), "Online cultural heritage exhibitions: a survey of information retrieval features", *Program Electronic Library and Information Systems*, Vol. 39 No. 1, pp. 4-24, doi: 10.1108/00330330510578778.

Malmsten, M., Lundborg, V., Fano, E., Haffenden, C., Klingwall, F., Kurtz, R. and Börjeson, L. (2024), "Without heading? Automatic creation of a linked subject system".

Markey, K. (2007), "The online library catalogue: paradise lost and paradise regained?", *D-lib Magazine*, Vol. 13 Nos 1/2, doi: 10.1045/january2007-markey.

Meadow, K. and Meadow, J. (2012), "Search query quality and web-scale discovery: a qualitative and quantitative analysis", *College and Undergraduate Libraries*, Vol. 19 Nos 2-4, pp. 163-175, doi: 10.1080/10691316.2012.693434.

Mazzocchi, F., Tiberi, M., De Santis, B. and Plini, P. (2007), "Relational semantics in thesauri: some Remarks at theoretical and practical levels", *Knowledge Organization*, Vol. 34 No. 4, pp. 197-214, doi: 10.5771/0943-7444-2007-4-197.

Noble, S.U. (2018), *Algorithms of Oppression: How Search Engines Reinforce Racism*, NYU Press, New York.

Downloaded from http://www.emerald.com/jd/article-pdf/81/7/376/10329593/jd-05-2025-0129en.pdf by Linneuniversitetet user on 13 October 2025

Olson, H.A. (2002), *The Power to Name: Locating the Limits of Subject Representation in Libraries*, Kluwer Academic, Dordrecht.

Patel, M., Koch, T., Doerr, M. and Tsinaraki, C. (2005), "Semantic interoperability in digital library systems", DELOS2 Network of Excellence on Digital Libraries, Deliverable 5.3.1, available at: http://delos-wp5.ukoln.ac.uk/project-outcomes/SI-in-DLs/SI-in-DLs.pdf

Renwick, T. and Szostak, R. (2020), "A thesaural interface for the basic concepts classification", *Proceedings of the International Society for Knowledge Organization Conference, 2020*, retrieval test results: A historical perspective, *Library Trends*, Vol. 56 No. 4, pp. 763-783, doi: 10.1353/lib.0.0000.

Riva, P., Le Bœuf, P. and Žumer, M. (2017), "IFLA library reference model: a conceptual model for bibliographic information", *International Federation of Library Associations and Institutions*.

Rowley, J. (1994), "The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research", *Journal of Information Science*, Vol. 20 No. 2, pp. 108-118, doi: 10.1177/016555159402000204.

Saracevic, T. (1975), "Relevance: a review of the literature and a framework for thinking on the notion in information science", *Journal of the American Society for Information Science*, Vol. 26 No. 6, pp. 321-343, doi: 10.1002/asi.4630260604.

Saracevic, T. (2008), "Effects of inconsistent relevance judgments on information".

Siegfried, S., Bates, M.J. and Wilde, D.N. (1993), "A profile of end-user searching behavior by humanities scholars: the Getty online project report no. 2", *Journal of the American Society for Information Science*, Vol. 44 No. 5, pp. 273-291, doi: 10.1002/(SICI)1097-4571(199306)44:5<273::AID-ASI3>3.0.CO;2-X.

Skov, M. and Ingwersen, P. (2014), "Museum web search behaviour of special interest visitors", *Library and Information Science Research*, Vol. 36 No. 2, pp. 91-98, doi: 10.1016/j.lisr.2013.11.004.

Svenonius, E. (1994), "Access to nonbook materials: the limits of subject indexing for visual and aural languages", *Journal of the American Society for Information Science*, Vol. 45 No. 8, pp. 600-606, doi: 10.1002/(sici)1097-4571(199409)45:8<600::aid-asi15>3.0.co;2-6.

Szostak, R. (2014a), "Classifying the humanities", *Knowledge Organization*, Vol. 41 No. 4, pp. 263-275, doi: 10.5771/0943-7444-2014-4-263.

Szostak, R. (2014b), "Classifying for social diversity", *Knowledge Organization*, Vol. 41 No. 2, pp. 160-170, doi: 10.5771/0943-7444-2014-2-160.

Szostak, R. (2015), "Classifying authorial perspective", *Knowledge Organization*, Vol. 42 No. 7, pp. 499-507, doi: 10.5771/0943-7444-2015-7-499.

Szostak, R. (2017), "A grammatical approach to subject classification in museums", *Knowledge Organization*, Vol. 44 No. 7, pp. 494-505, 2017, doi: 10.5771/0943-7444-2017-7-494.

The Getty Foundation (2012), "Moving museum catalogues ONLINE, an interim report from the getty foundation", available at: https://www.getty.edu/foundation/pdfs/osci_interimreport_2012.pdf

Tibbo, H.R. (1994), "Indexing for the humanities", *Journal of the American Society for Information Science*, Vol. 45 No. 8, pp. 607-619, doi: 10.1002/(SICI)1097-4571(199409)45:8<607::AID-ASI16>3.0.CO;2-X.

Tudhope, D., Binding, C., Blocks, D. and Cunliffe, D. (2006), "Query expansion via conceptual distance in thesaurus indexed collections", *Journal of Documentation*, Vol. 62 No. 4, pp. 509-533, doi: 10.1108/00220410610673873.

Villaespesa, E., Tate, U. and Stack, J. (2015), "Finding the motivation behind a click: definition and implementation of a website audience segmentation", in *MW2015: Museums and the Web 2015*, available at: https://mw2015.museumsandtheweb.com/paper/finding-the-motivation-behind-a-click-definition-and-implementation-of-a-website-audience-segmentation/

Villén-Rueda, L., Senso, J.A. and De Moya-Anegón, F. (2007), "The use of OPAC in a large academic library: a transactional log analysis study of subject searching", *The Journal of Academic Librarianship*, Vol. 33 No. 3, pp. 327-337, doi: 10.1016/j.acalib.2007.01.018.

Downloaded from http://www.emerald.com/jd/article-pdf/81/7/376/10329593/jd-05-2025-0129en.pdf by Linneuniversitetet user on 13 October 2025

Vizine-Goetz, D., Hickey, T.B., Houghton, A. and Thompson, R. (2004), "Vocabulary mapping for terminology services", OCLC Research, available at: https://www.oclc.org/research/publications/library/2004/2004-03.pdf

Vo, N.N., Le, T.H., Hoang, V.T. and Nguyen, D.D. (2019), "A comprehensive survey on recent developments in image retrieval using deep learning", *Journal of Imaging*, Vol. 5 No. 7, pp. 1-26.

Walsh, D., Clough, P. and Foster, J. (2016), "User categories for digital cultural heritage", *First International Workshop on Accessing Cultural Heritage at Scale*, pp. 3-9, available at: https://www.researchgate.net/publication/304114334_User_Categories_for_Digital_Cultural_Heritage

Walsh, D., Hall, M.H., Clough, P. and Foster, J. (2018), "Characterising online museum users: a study of the National Museums Liverpool Museum website", *International Journal on Digital Libraries*, Vol. 21 No. 1, pp. 75-87, doi: 10.1007/s00799-018-0248-8.

Will, L. (1993), "The indexing of museum objects", *The Indexer*, Vol. 18 No. 3, pp. 157-160, doi: 10.3828/indexer.1993.18.3.6, available at: https://www.theindexer.org/files/18-3/18-3_157.pdf

Zeng, M.L. and Chan, L.M. (2004), "Trends and issues in establishing interoperability among knowledge organization systems", *Journal of the American Society for Information Science and Technology*, Vol. 55 No. 5, pp. 377-395, doi: 10.1002/asi.10384.

Zeng, M., Žumer, M. and Salaba, A. (2011), *Functional Requirements for Subject Authority Data (FRSAD): A Conceptual Model*, De Gruyter Saur, Berlin and New York.

Zhou, W., Li, H. and Tian, Q. (2017), "Recent advance in content-based image retrieval: a literature survey", [cs], available at: http://arxiv.org/abs/1706.06064

Corresponding author
Koraljka Golub can be contacted at: koraljka.golub@lnu.se

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: **permissions@emeraldinsight.com**

Downloaded from http://www.emerald.com/jd/article-pdf/81/7/376/10329593/jd-05-2025-0129en.pdf by Linneuniversitetet user on 13 October 2025