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

Postprint

This is the accepted version of a paper published in *Journal of Documentation*. This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

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

Golub, K., Tyrkkö, J., Hansson, J., Ahlström, I. (2020)

Subject indexing in humanities: A comparison between a local university repository and an international bibliographic service

Journal of Documentation, 76(6): 1193-1214

<https://doi.org/10.1108/JD-12-2019-0231>

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-93446>

1 Introduction

While support for subject searching has traditionally been advocated for in library catalogues, particularly since Cutter defined objectives for library catalogues (1876), research shows that subject access in online library catalogues, repositories, and commercial services like bibliographic databases and discovery services has been less than optimal and often fails to meet established objectives for bibliographic systems (see, e.g., Markey, 2007; Golub, 2018). While the services try to match users' expectations by implementing Google-like single search box interfaces, it seems that efficient mechanisms such as ranking algorithms used by commercial search engines like Google, efficient exploitation of subject indexing, or even quality-controlled subject indexing per se, are still missing from these services, which leads to frequent retrieval failures.

The specific challenges of indexing humanities research output have been a cause for concern for researchers for several decades (Langridge, 1976; Tibbo, 1994). As part of the general development of digital scholarship, many disciplines and research areas within the humanities have developed new structures both within themselves and in relation to other disciplines within the humanities and beyond them (Borgman, 2007, pp. 212-224). For example, in the rapidly growing interdisciplinary field of digital humanities, it has become increasingly important to provide quality subject access to the vast variety of heterogeneous information objects catalogued by digital services. This includes both primary (see, e.g., Choi and Syn, 2016, on the use of tags in archival collections) and secondary sources, the latter being the focus of the present study. Although secondary sources in the humanities include monographs and book chapters much more than in the sciences, we will focus on journal articles because they tend to be more completely represented and reviewed in the bibliographic sense.

This exploratory study aims to paint a representative picture of the current state of affairs when it comes to the use of subject index terms in humanities journal articles, with particular reference to the well-established subject access needs of humanities researchers. The objective is to identify the needs for improvement in this specific context. The sample used comprises 649 peer-reviewed journal articles from across the humanities. The articles included are a part of the research output from an Arts and Humanities Faculty of a medium-sized Swedish university between the years 2010 and 2018. A comparison of subject metadata was carried out comparing DiVA, the university repository, and Scopus. The university repository follows local and national policies and Scopus claims to be the most comprehensive international abstract and citation database of research output (Elsevier, 2017), although it may be argued that the Web of Science is certainly comparable.

The paper is structured as follows. In the Background section, the stage is set by providing context on the objectives that subject access should meet in contemporary online search services, with particular focus given to the subject access needs of humanities researchers. A section on Methodology comes next, followed by Results. In the Conclusions section, a summary of the results is given, with suggested implications for future research and development.

2 Background

Subject searching is a common type of query in library catalogues (Hunter, 1991; Villén-Rueda et al., 2007), bibliographic databases (Siegfried et al., 1993), repositories (Heery et al., 2006), discovery services (Meadow and Meadow, 2012) and related digital search services (Patel et al.,

2005). In comparison to known-item searching (e.g., queries for information objects whose title, author, etc. are known beforehand) searching by subject is much more challenging. This is due to the difficulties of formulating queries with insufficient knowledge of the subject matter, as well as for other related reasons to do with information searching, semantic ambiguities inherent to natural language, and so on. In order to alleviate these problems, online search services (could) make available subject terms from controlled vocabularies such as subject heading systems, thesauri and classification systems, which can help the user select a more specific concept in order to increase *precision*, a broader concept or related concepts to increase *recall*, to disambiguate between homonyms, or to discover which term is most appropriate for a specific concept. In addition, hierarchical browsing of classification schemes and other controlled vocabularies could help the user improve their understanding of their information needs, which could consequently aid the user to formulate their queries more accurately.

Guidelines for providing subject access in bibliographic systems have their origins in the cataloguing standards used in libraries and related information services. The objectives of library catalogues for subject access are originally anchored in Charles Ammi Cutter's 'objects', as he called them, which are 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 as to its topical character (Cutter, 1876, p. 5). These objectives have been an integral part of cataloguing codes for nearly 150 years and continue to be so in contemporary FRBR (Functional Requirements for Bibliographic Records) family of conceptual models for catalogue functionality, which were in 2017 consolidated into the IFLA Library Reference Model (IFLA LRM, International Federation of Library Associations 2017). In the context of subject access, IFLA LRM and FRSAD (Functional Requirements for Subject Authority Data) tasks of finding, identifying, selecting, obtaining, and exploring, could be defined as follows:

- Find: to find resources embodying works that are described by a given subject label, for example, search using a nomen that is used in a subject headings system or a classification scheme;
- Identify: to clearly understand the nature of the resources found and to distinguish between similar resources, e.g., 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 virus from a zoological perspective versus virus from a medical perspective);
- Select: to determine the suitability of the resources found and to choose (by accepting or by rejecting) specific resources that seem the most relevant, e.g., due to certain aspects, facets or approach to the subject described;
Obtain: to access the content of the resource;
- Explore: to use the subject relationships between one resource and another to place them in a context, e.g., to browse around related topics such as through using related terms in a thesaurus, or to see narrower and broader terms or classes, in order to understand the relationships between various nomens for an entity such as: 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, e.g., finding a thesaurus descriptor which corresponds to a classification number.

Although publications in the humanities are also of interest to students and the general public, research so far has mainly focused on the information needs of humanities scholars. Professional

researchers in general require information with a high level of granularity when looking for information in bibliographic databases (De Andrade and Baptista, 2014), and this can be facilitated by applying indexing at a deep level of specificity. A major study of online database searching by humanities scholars was conducted over a two-year period by Bates and colleagues at Getty in 1990s combined search-log analysis and interviews (final report is given in Bates, 1996). The study showed that most searches were subject searches – 91% of natural language statements indicated a subject of some kind, strongly supporting the need for bibliographic databases to support subject searching; the interview study showed that their library's online catalogue would work better for known-item searching.

Furthermore, a comparative analysis of the Getty humanities scholars' queries against sciences scholars' queries, collected in an earlier study, revealed dramatic differences in the type of subjects the scholars searched for. Whilst scientific queries typically comprise common terms and only rarely other types of items, the Getty study showed that only 57% of the humanities queries contained any common terms at all. Bates (1996, p. 5250) defines common terms as uncapitalized terms which are none of the following categories: works or publications as subject, individuals as subject, geographical term, chronological term, disciplinary term, other proper term (capitalized). Instead, these other types of terms were often present (terms denoting named individuals, geographical terms, chronological terms, and disciplinary terms denoting academic disciplines, such as "history"). Similar findings have also been reported by Wiberley (1983, 1988), who observed that humanities subject terms are often highly precise proper names, and by Tibbo (1994), who summarised earlier studies observing that single proper terms like authors' names and the titles of works are common in certain disciplines such as literary studies, while common terms are more characteristic of field such as religion and philosophy. Likewise, Yi et al. (2006) found that most search terms of two history databases referred to specific instances of historical events, people, and regions, in contrast to the search terms used in a psychology database which were mostly common terms matching those of the dedicated classification scheme. All these findings suggest that there is a need for a faceted approach to controlled vocabularies, such as the Arts and Architecture Thesaurus for visual arts, rather than for pre-coordinated ones like the Library of Congress Subject Headings. Faceted vocabularies are more suitable because they support high specificity and can account for the different facets that are important to humanities scholars, such as geographical, chronological, and disciplinary terms (see Bates, 1996; Tibbo, 1994). Furthermore, facet selection and the query expansion based on such controlled vocabularies also needs to be implemented into the search interfaces, which is a feature currently limited to experimental interfaces (see, e.g., Alani et al., 2000; Tudhope et al., 2006) rather than being applied into practice across bibliographic systems.

Knapp, Cohen, and Juedes (1998) established that the most effective way of searching databases in the humanities is to combine free-text searching with the use of controlled-vocabulary indexing; see also Rowley (1994) for a general overview of the debate of controlled versus natural indexing languages. Controlled vocabularies are particularly much needed in large databases covering many subjects (Markey, 2007; Tibbo, 1994) as well as in databases of primary sources (Bair and Carlson, 2008), which cannot be queried using full-text searches (alone). Tibbo (1994) makes the point that rather than improving information access, the exponentially increasing volume of information objects available online actually leads to information overload and entropy. Although full-text indexing works for some tasks, for others it creates information overload and prevents the searcher from gaining a comprehensive overview on a topic – if a query returns thousands of retrieved documents, few searchers will browse beyond the first dozen or two hits. It is also worth noting that humanities scholarship is linguistically very diverse: in addition to the considerably high volume of distinct concepts,

terminological synonymy and overlaps are also abundant, even within relatively well-defined subdisciplines. The terminological complexity complicates queries by placing the burden on the scholar, who would ideally need to include all possible synonyms in a query if a comprehensive set of results is desired. Homonymy is likewise also present, with the result that queries often end up producing false positives. Furthermore, terminological changes may also be a problem given that humanities scholars frequently work with materials and sources produced over long periods of time. Consequently, the index terms used may have different meanings when applied to the same term from different periods, or different terms in different periods may be used to refer to the same concept. When it comes to searching by titles, they may be metaphorical, contain allusions or intertextual references, or otherwise be less than descriptive, resulting in low recall as well as false positives. Similarly complex problems are also commonly encountered when it comes to primary sources. For example, in literary studies sexuality has not been investigated employing computational methods on large text corpora because the phenomenon studied is not manifestly present in the texts and calls for human subject indexing (Bergenmar and Golub, 2020). The lack of conceptualisation of terms extracted from text is common in humanities, for example concepts such as ‘dramatists’, or ‘persecution’ may not be articulated for a paper about particular playwrights or conflicts.

To counter high recall and low precision, specific subject indexing should be implemented, involving 1) indexing policies that promote a high level of specificity and 2) indexing languages that are deep and detailed for any given topic, especially in large bibliographic databases with tens of millions of records. The indexing language for humanities needs to be extensive in order to account for the fact that virtually any topic can appear in humanities writing and the same topics may be studied from a very wide variety of different perspectives. Furthermore, as a rather heterogeneous group, specific humanities disciplines will require their specific indexing languages, rather than a one-fit-all (Tibbo, 1994, p. 608-609); which then also requires a meta subject indexing language that brings them all together in order to support searching across disciplines in an interoperable manner.

Tibbo further notes that humanities scholars tend to use a ‘dense’ rather than ‘readable’ writing style, making it particularly challenging to create representative metadata (p. 609), the quality pending on information experts also being subject experts. An example of the latter also pertains to fiction: De La Tierra (2008) gives an example of LGBTQ literary work not being indexed as such, due to the lack of knowledge about this by the librarian, which is also related to the lack of conceptualisation mentioned above.

Tibbo (1994) highlights that the major challenges in providing subject indexing to support humanities scholarship are related to the need to devise and implement controlled vocabularies and indexing frameworks different to those of the sciences to reflect the different information seeking behaviours. She concludes that indexing for the humanities lags behind natural sciences because more funding is available in the sciences and because scientific literature is “simply easier to control”, particularly when it comes to “topical, lexical, and semantic heterogeneity and less structured nature of humanistic literature” (p. 616). Furthermore, the “interdisciplinary and unique research interests of humanistic scholars make convenient subject coverage with appropriate index terms challenging.”

According to East (2007), who studied subject retrieval from ten full-text databases in the humanities, the databases do not meet the needs of scholars. Only one of the databases studied assigned controlled subject index terms to each document, while two others assigned Library of Congress Subject Headings to whole books rather than to individual chapters and sections. East

concludes that in the “first phase of digitisation”, the excitement generated by the prospect of technology to make everything available at a few keystrokes, “has blinded many commentators to the distinct limitations”. He claims that the users will use these collections only “to look for documents that they have identified from other sources rather than performing subject searches to discover further resources”. East also expresses a wish for a second phase of digitization, “in which the ‘quick and dirty’ digital libraries of today will be enriched and enhanced to become resources that can effectively meet the information needs of scholars” (ibid., p. 239). The availability of funding, or lack thereof, is likely a root cause behind the differences in coverage and effectiveness of indexing between science and humanities, the latter having much less commercial relevance.

In spite of over 140 years having passed since the publication of Cutter’s Rules for a Dictionary Catalog, it has been said that bibliographic systems have never lived up to his original ideals. According to Markey (2007), there are ten reasons for why the desired functionalities have not been implemented online, including the failure of the research community to arrive at a consensus about the most pressing needs for online system improvement; decreasing funding; and, at the same time, the high cost of integrated library systems. As a result, by the time the World Wide Web became prevalent, the demand to implement functionalities similar to global search engines such as Google and other commercial services like Amazon, was increasing. These included the single search box, attractive web design, relevance ranking of results, recommendations, and access to a wide range of resources. Although seemingly attractive, these requirements come with problems related to searching based almost entirely on full-text indexing (cf. above), which means that each search would result in “millions of hits with no guarantee that the top-ranked ones will address your desired topic in depth or at your level of understanding” (Markey, 2007). Faceted navigation has become a standard feature in large bibliographic systems and subjects are often seen as only one of the facets (Chickering and Yang, 2014). However, studies show that this can be confusing to end users who frequently do not understand how facets work or know the types of terms included in them (Emanuel, 2011; Osborne and Cox, 2015).

The most common issues regarding subject searching today are inconsistency and incompleteness of metadata and the blending of controlled vocabularies, free keywords and full-text automatic indexing (Dempsey, 2012; Fagan, 2011; Golub, 2016). Challenges like this are particularly problematic for researchers conducting interdisciplinary searches common in the humanities and probably even more so in the emerging field of digital humanities. Commercial bibliographic systems claim to provide comprehensive coverage, yet their indexing policies are not standardised. Neither are there any international standards or common guidelines for implementing quality-controlled subject access in repositories; instead, a range of different vocabularies are applied across repositories and repository platforms (see, for example, Bundza, 2014). The National Library of Sweden (2019) provides guidelines for repositories that are made available in the Swedish national repositories service, SwePub (<http://www.swepub.se>), which includes the repository studied here. According to the guidelines, the National Subject Category, used mainly for statistical purposes by the Swedish statistics agency (Statistics Sweden, 2016), is the obligatory metadata value to choose from, while keywords do not have a pre-defined value set.

3 Methodology

3.1 Purpose and aims

In order to help improve quality subject access to humanities sources, the present study aims to establish the current state of affairs with particular reference to secondary sources in the humanities. Although such secondary sources include monographs and book chapters much more than the sciences, here it is peer-reviewed journal articles that are studied because they tend to be more completely bibliographically represented.

The research questions asked are the following:

- To what degree are established contemporary objectives for ensuring subject access to humanities journal articles supported by the metadata of a university repository (DiVA)?
- To what degree are established contemporary objectives for ensuring subject access to humanities journal articles supported by the metadata of a large, interdisciplinary bibliographic database (Scopus)?
- How do index terms used in the university repository and in Scopus compare to each other and against those found in the respective journal articles?

3.2 Data collection

Data collection is based on metadata of 649 journal articles extracted from the Linnaeus University repository of academic publications; of these, 321 were also found in Scopus. The repository software, DiVA (<http://www.diva-portal.org/>), conforms to local policies. Scopus was chosen because it claims to be the largest international abstract and citation database of peer-reviewed literature; although traditionally mostly science-oriented, it covers about 3,500 arts and humanities serials, meaning journals, book series and conference series (Elsevier, 2017).

The 649 articles represent the complete research output of peer-reviewed journal articles by the university's Arts and Humanities Faculty during the period 1 January 2010 to 31 December 2018. This period was chosen because the university was established in 2010 and at the same time, the university library started implementing quality control of repository submissions by authors. The university has about 33,000 students and, as of November 2019, 1,173 academic staff. It is structured into five faculties, of which Faculty of Arts and Humanities is one with 257 academic staff at the time of writing, most of whom are also researchers. The Faculty of Arts and Humanities comprises seven departments: Design, Film and Literature, Cultural Sciences, Media and Journalism, Music and Art, Languages, and Swedish. Each department has a number of disciplines and each discipline is listed as one value of the repository metadata element titled "research subject". The Faculty of Arts and Humanities comprises the following 31 subjects: archaeology; art science; comparative literature education; comparative literature; creative writing; cultural sociology; English education; English literature; English; Film studies; French linguistics; French literature German education; German literature; German history education; history; human geography; library and information science; linguistics; music education; music; musicology; Spanish literature; Spanish; study of religions; Swedish as a second language; Swedish didactics; Swedish; teaching and learning in French as a second language; visual culture.

Metadata from the repository were extracted on 19 March 2019. Metadata for the same journal articles were manually extracted from Scopus during the period between 20 March and 21 May 2019, when keywords from articles taken from journals were manually extracted.

Network analyses of the data were carried out using Cytoscape, an open source tool for visualizing complex networks. The pre-processing steps included transformation of multiword keywords into single-word items by replacing spaces with underscores, and the normalisation of apostrophes and dashes. Matches were not case sensitive.

4 Results

4.1 Repository

The DiVA repository user interface (<http://lnu.diva-portal.org>) offers simple and advanced search interface for research publications. The simple search interface supports

- 1) searching through a simple search box, whereby all metadata elements are searched apart from the research subject (explained below),
- 2) browsing by “subject”, where subject is the national subject categories (defined below). Advanced search interface allows searching by the following subject access points through selection from a drop-down list: 1) “category” (the national subject category), 2) “keywords” (which are keywords entered by author or by librarian from the published journal), and
- 3) “research subject” (local categories of subjects, explained below). Like in the simple search interface, one can also browse by “subject” (“subject” represents the national subject categories).

The repository uses three types of subject access index terms: uncontrolled keywords, national subject category, and university research subject. These are described and analysed in the three subsections that follow.

4.1.1 Keywords

Keywords are uncontrolled index terms either entered by author at the time of deposit to the repository (optional field), or subsequently added by the librarian. If left empty by the author, keywords are manually copied from keywords of the published article; if none are available, the field is left empty. In the repository data collection, a total of 514 articles (79.2%) had at least one keyword assigned.

As seen from the analysis below (section 4.4 Index terms on published journal articles), most authors did not assign keywords at the time of deposit since keywords were not an obligatory field. This is derived from the fact that the majority of “keywords” in the repository are the same as “author keywords” in Scopus, both of which seem to contain values taken from the published journal articles.

There are only five articles in the repository with keywords that are entirely different from journal keywords, i.e., those for which one can assume that the authors assigned them at the time of deposit. These are given below with those that are different are underlined:

- Repository: C-14 dating; Jordan; Middle Upper Paleolithic transition
versus
Scopus: ABOx-SC pretreatment; AMS; Biocultural evolution; Radiocarbon dating; Middle Upper Paleolithic transition

In the above example the author assigned one additional keyword referring to a geographical entity (“Jordan”). Both use different terms to describe the same method (the author uses “C-14 dating” referring to ¹⁴C-dating as its proper label and the journal (“radiocarbon dating”). Journal keywords are more exhaustive since they also cover specifics of the method (“ABOx-SC pretreatment”, Acid Base Oxidation-Stepped Combustion, a specific pretreatment protocol used to date particularly old samples and “AMS”, Accelerator Mass Spectrometry, the standard technique for ¹⁴C-dating generally). The article keywords are more exhaustive also in that they include another concept not included by the author (“biocultural evolution”).

- Repository: Higher education; language of instruction; lecturing; teaching; language choice

versus

ELF; medium of instruction; parallel-language education; teaching in English; university lecturing.

This example shows how the author and journal use near synonyms (“language of instruction” by author and “medium of instruction” by journal; “higher education” and “lecturing” assigned by author when used together are similar to “university lecturing”). The journal keywords are more specific (the author uses just “teaching”, the journal “teaching in English”). The author also used one concept not specified by the journal (“language choice”) and the journal used two keywords that are not covered by the author (“parallel-language education” and “ELF”).

- Linguistic landscape; English in Finland; International English

versus

English as a lingua franca; fieldwork; geosemiotics; globalisation and the English language; linguistic urban spaces vs. landscapes.

This example also shows a certain level of synonymy: “International English” by author and “English as a lingua franca” by journal; “linguistic landscape” by author versus “linguistic urban spaces vs. landscapes” whereby the journal provides a particular facet of comparison. The journal provides more concepts not mentioned by the author: “fieldwork”, “geosemiotics”, “globalisation and the English language”, while the author provides one more specific to a geographical facet: “English in Finland”.

- Mother tongue instruction; Sweden; teacher collaboration; minority languages

versus

edgelands; mono-cultural norms; mother tongue tuition; scheduling.

Here one again sees synonyms: “mother tongue instruction” and “mother tongue tuition”. Other keywords are different; the author here like in two other examples adds a geographical facet (“Sweden”) as well as “teacher collaboration” and “minority languages”. The journal lists “edgelands”, “mono-cultural norms” and “scheduling”. Both seem to give a different idea on what the article is about and are probably best used together..

- School library; multilingual learning environments; literacy; intercultural school development

versus

diversity; intercultural education; multilingualism; reading for pleasure; school libraries.

Synonyms are present in this example, too: “school library” and “school libraries”. The foci of two related terms seem different: “intercultural school development” versus “intercultural education”. “Multilingual learning environments” by the author seem more specific than “multilingualism” by the journal. Both also add additional concepts not listed by the other: “literacy” was added by the author and “reading for pleasure” as well as “diversity” by the journal.

The journal keywords seem to be more exhaustive when it comes to indexing, but the authors occasionally provide additional concepts that are not provided by the journal, such as geographical facets that are commonly searched for by humanists (cf. Bates, 1996). In common with other repositories, authors are not trained or provided any indexing guidelines and this should be amended in the future.

4.1.2 National subject categories

Following repository indexing guidelines (National Library of Sweden, 2019), national subject categories are a required field to be filled at the time of deposit. One or more categories are to be selected by the author from a pre-set controlled list. The categories are quality-checked by the librarian after the deposit. The most common type of correction by the librarian is to assign a more specific category than what the author supplied, following repository indexing guidelines (ibid.).

National subject categories are broad disciplinary-level, controlled index terms from the Swedish standardised categorization of research areas (Statistics Sweden, 2016). There are 3 levels of hierarchical division: 6 top level categories, which are divided into a total of 42 categories at the second hierarchical level, followed by about 250 categories at the deepest level. The first two levels are the same as the international OECD’s classification of research topics (Organisation for Economic Co-operation and Development, 2007) while the third one is specific to Sweden. The main purpose of the categories is to facilitate the collecting of data for official national statistics on scientific publishing. However, it is also implemented in search interfaces of the national integrated repository titled SwePub and in the repository investigated here.

All 649 articles in the sample have at least 1 national category assigned. Most articles have only 1 category assigned (507 or 78.1%); 2 categories are assigned to 116 articles (17.9%), and 3 categories to 23 articles (3.5%). There are 2 articles with as many as 4 categories and 1 article with 5 categories. Most commonly assigned categories ($n > 10$) are: archaeology (assigned to 87 articles), history (78), specific languages (69), general literary studies (63), general language studies and linguistics (50), languages and literature (44), studies on film (39), specific literatures (37), religious studies (31), information studies (28), educational sciences (25), cultural studies (20), humanities (20), other humanities not elsewhere specified (17), history of religions (15), musicology (15), history and archaeology (14), didactics (12), and sociology (11). Figure 1 shows the number of rank order of national categories on the horizontal axis and the number of articles each has been assigned to in the DIVA repository.

Figure 1. Occurrences of national categories in the DiVA dataset

4.1.3 University research subject

University research subjects reflect actual division of research in departments at the university and are used for internal statistical purposes (see 3.2 Data collection). This is an optional field for the author at the time of deposit; one or more subjects can be chosen from a controlled list created by the university's Communications department. If necessary, the librarian reviewing the record adds additional subjects to match the national category or categories chosen by the author.

This list of categories has 3 hierarchical levels, numbering in total 136 subjects. At the top level there are 13 subjects. Of these, 10 have narrower subjects at the second level, making in total 86 subjects. Of these, 13 have subjects at the third level, with the total of 37 subjects at the deepest level.

In comparison to national subject categories, the humanities subjects in both lists contain almost the same amount of items: 31 national subject categories and 32 research subjects. Some categories from the national subject categories are not included in the research subjects simply because those subjects are not represented within university departments. Similarly, the research subjects list has some sections more expanded than the national subject categories list, such as languages: while the latter uses only 1 relevant category termed "specific languages", the research subject list has 14 subjects for subjects related to individual languages; for example, for English studies there are the subjects of "English Education", "English literature" and "English", the third item referring implicitly, and without guidance to the user, to English language or English linguistics. Other differences include different placements of some subjects; for example, library and information science in the list of research subjects is listed at level two under humanities, while in the national subject category list it is at level three under information studies (level 3) which in turn are under media and communications (level 2) that belong to social sciences (level 1).

As for the breakdown of LNU research subjects in the same 649 articles, 629 (96.9%) articles have at least 1 research subject assigned. Since this is not an obligatory field, it is easier also for librarians checking the records to miss it – hence the lack of 100%. Most have 1 subject

assigned (513 or 79.0%); 80 (12.3%) have 2; 26 (4.0%) have 3 categories; 6 (0.9%) have 4 categories; there are 4 articles each with 5, 6, 7 and 8 categories. Most commonly assigned ($n > 10$) are: humanities, archaeology (assigned to 93 articles); humanities, history (78); humanities, comparative literature (77); humanities (53); humanities, study of religions (42); humanities, film studies (34); humanities, English literature (32), humanities, English (31); humanities, library and information science (30); humanities, linguistics (29); humanities, Swedish (21); pedagogics and educational sciences (20); humanities, cultural sociology (15); humanities, art science (13); humanities, musicology (12); humanities, visual culture (11). Figure 2 shows the number of rank order of LNU categories on the horizontal axis and the number of articles each has been assigned to in the DiVA repository.

Figure 2. Occurrences of LNU research subjects in the DiVA Humanities dataset

Users are allowed to choose the top levels of the Research Subjects. It is also possible that an article registered into the system was not written when the author was doing research at LNU, but at a previous university. The subject of the article might not match any of the levels below the top level. The top level is then chosen instead. Librarians reviewing records in DiVA will choose a subject other than the top level if possible.

The network visualisation (Figure 3) illustrates the relationships between individual articles and both national categories and research subjects, and the object of interest is the amount of overlap between the nodes representing the categories and the research subjects. While some fields like history and archaeology show distinct clusters, most others are more dispersed. The figure shows the articles (purple rectangles), national categories (dark blue circles) and LNU categories (red circles). The sizes of the node markers are based on in-degree, or the number of articles that include that category; the layout algorithm used is an Edge-Weighted Spring-Embedded Layout. In many cases similar national categories and research subjects cluster well (=are spatially close together), but there are also differences. This visualisation shows how some fields like history and archaeology seem to be quite unambiguous and distinct, while others show a lot more fragmentation. For example, it seems counterintuitive that an article can be categorised under a specific language but not linguistics (or “languages and literature”).

Figure 3. Overlap between national subject categories and LNU research subjects.

The differences observed between research subjects and national subject categories can mostly be attributed to indexing inconsistency. If a topic is assigned by the researcher, the librarian will mostly focus on assigning the most specific hierarchical level of the national subject category, and complete the missing information for the research subject, but will not change the research subject chosen by the author. This is why, for example, there are 53 general humanities articles in research subjects, and 20 in the national subject category.

The problem of using both a classification scheme designed for statistical purposes and subjects assigned by authors to be used in subject searching by end-users can be illustrated by the following example:

Article title: “Anterior and posterior ERP rhyming effects in 3 to 5-year-old children”
 National categories: “languages and literature”; “psychology”; “learning”
 Research subject: “humanities, English”; “pedagogics and educational sciences, education”; “social sciences, psychology”

The national categories list the article as “languages and literature”, a very broad descriptor, while the keywords under research subjects focus specifically on English and say nothing about linguistics or languages in general. As it turns out, the actual article is not specifically about English at all, the only connection to English being that it is the language of the test subjects in the study reported by the article. However, this means that the article in question is not an English-studies article in the sense that it would say anything specific about the English language, but rather the article deals with psycholinguistics and more specifically with the neurocognitive development in children and their early literacy development.

Furthermore, the keyword “psychology” in the research subjects seems to be under “social sciences”, which is perhaps generally acceptable but in the case of this particular study, the specific field of psychology in question qualifies with some difficulty as a subfield of social science.

It seems obvious that these two lists of categories, primarily intended for statistical analysis, do not lend themselves well to information retrieval. With mere 30-ish categories across all of the humanities, they lead to information overload rather than ensuring the high level of granularity that would further group hundreds of resources into small sets of resources on actually same specific topics. In addition to sufficient granularity, these structures need to be created with the user in mind, rather than statistics, in order to support browsing in the sense of teaching the searcher on the structure of humanities disciplines, fields and subfields as well to support the searcher in query formulation.

4.2 Scopus index terms

Scopus user interface supports limiting search results by “keywords”; according to the Help file, this includes both “author” and “indexed keywords” (Elsevier, 2019a). Advanced search interface additionally enables the use of field codes whereby one can limit the selection to author keywords (“AUTHKEY”), indexed keywords (“INDEXTERMS”), or both (“KEY”); field code ‘SUBJAREA’ allows narrowing the search down based on broad fields, such as Arts and Humanities (ibid.).

Scopus’ indexing policy details that a “team of professional indexers” “manually add index terms for 80% of the titles included in Scopus” (x2017). However, Scopus’ comprehensive view of the world’s research output in all the fields is rather questionable, given that the controlled vocabularies listed by Scopus in the same document do not seem to cover arts and humanities, effectively preventing retrieval with high precision and recall and, therefore, access to information (cf. section 2.2 Subject access needs for humanities). The listed controlled vocabularies are:

- Ei Thesaurus (Engineering Information Thesaurus), for engineering, technology, physical sciences;
- Emtree medical terms (EMBASE thesaurus – Excerpta Medica dataBASE) and
- MeSH (Medical Subject Headings) for life sciences and health sciences;
- GEOBASE Subject Index for geology, geography, earth and environmental sciences;
- Regional Index for geology, geography, earth and environmental sciences;
- Species Index for biology and life sciences;
- FLX terms and WTA terms for fluid sciences and textile sciences (ibid.).

When it comes to the specificity and exhaustivity of indexing using these vocabularies, the following is stated: “There is no limit to the number of index terms that Scopus can add to records”. This information seems to be contradicted by information in another document where Elsevier states that controlled keywords are (2019a) “...chosen by content suppliers and are standardized based on publicly available vocabularies... the Indexed keywords take into account synonyms, various spellings, and plurals”. It is further stated that “Scopus has no influence over either Author or Indexed keywords because these are both determined by third parties” (ibid.).

Scopus classifies serial titles using the ASJC (All Science Journal Classification) scheme. Indexing is conducted at the time when Scopus decides to include the title in their database

(Elsevier, 2019b). The classification process is based on “the aims and scope of the title, and on the content it publishes” (ibid.). It is further stated that the classification scheme can be used to filter search results; however, it is not clear from the interface where this is possible to select. Advanced search interface offers limiting by “subject areas”; nowhere is it stated whether these are the same as ASJC classes.

While the search interface supports querying by author keywords, controlled keywords (called “indexed keywords” by Scopus), broad subject area and journal classification, retrieved records seem to comprise a different set of subject access terms: author keywords, controlled keywords and uncontrolled keywords. Uncontrolled keywords do not seem to be defined anywhere; broad subject area and journal classification are not listed with the retrieved record. The sections that follow analyse subject access points found in retrieved records of the sample: author keywords, controlled keywords, and uncontrolled keywords.

4.2.1 Author keywords

Elsevier (2019a) claims that author keywords are “chosen by the author(s) which, in their opinion, best reflect the contents of their document”. In the data collection, 234 articles (72.9%) have author keywords assigned in Scopus. No more details are available on when or how this is done, whether the author is given any guidelines, etc.; however, the analyses in section 4.4.1 shows that author keywords are essentially those taken from published journal articles whereby the author assigns keywords that normally form part of the abstract and associated text.

4.2.2 Controlled keywords

In the data collection, only 13.1% (42) of the articles have any controlled keywords assigned (called “indexed keywords” on Scopus interface). As explained above and written by Elsevier (2007), none of them have any controlled terms specific to the humanities; instead the following are used: Engineering Index (7 records); EMTREE (19 records); MeSH (17 records); GEOBASE (20 records). Of the 42 articles, 19 have terms from more than 1 controlled vocabulary assigned: most of them have both EMTREE and MeSH (15); 1 has EMTREE, MeSH and Engineering Index; 1 EMTREE and Engineering Index; 1 Engineering Index and GEOBASE; 1 GEOBASE, EMTREE and MeSH. The specificity and exhaustivity of indexing using these vocabularies does not seem to be known (cf. introduction to 4.2 Scopus index terms); on average there are 13 keywords assigned to each article, with the lowest number being 1 and largest 35.

4.2.3 Uncontrolled keywords

In the data collection, only 8 articles (2.5%) have assigned uncontrolled keywords. The Help file does not seem to specify how these are created or where they are derived from and in what way. They do seem to be exhaustive and specific in concept coverage – on average 9 terms are assigned. The uncontrolled keywords seem to be what Bates (1996) refers to as common terms. Many seem to be the same as keywords on published journal papers; for example, the uncontrolled keywords of one article are: Augmentative and alternative communication; Cerebral palsy; Children; Complex communication needs; Enjoyment; Participation; Peers; Play. In comparison to keywords on published article, six of them are identical; in addition, two are added as uncontrolled keywords (Robotics; Enjoyment).

4.3 Comparison of Scopus and repository index terms

While humanities articles in both databases generally have no controlled subject access with index terms from pertinent humanities controlled vocabularies, the humanities articles in

Scopus have author keywords on 72.9% of articles and the repository for 79.2%. In addition, national categories are assigned to all articles (100%) in repository, although the purpose is more for disciplinary-level browsing rather than specific topic or facet searching.

In order to compare index terms assigned to the same article in repository and in Scopus, the data subset was reduced to articles with metadata that had one or more keywords in both databases. This amounted to 203 articles; of these, only 29 articles had a controlled term from Scopus ('indexed keyword'), as well as Scopus author keywords, repository author keywords, repository national category and repository research subject; in addition, 5 articles also have Scopus uncontrolled keywords. The remainder of 174 articles had all of these, but lacked Scopus controlled and uncontrolled keywords.

Of the 29 articles, several are discussed that illustrate differences, problems and advantages. The underlined terms are those that are uniquely mentioned in terms other than author keywords (titled "keywords" in repository).

Example A.

Title: "The need and potential for an archaeology orientated towards the present"

- Repository index terms
 - Keywords: public archaeology; archaeology in society; new heritage; applied archaeology
 - Categories: archaeology
 - Research subjects: humanities, archaeology
- Scopus index terms:
 - Author keywords: public archaeology; archaeology in society; new heritage; applied archaeology
 - Indexed keywords:
 - Engineering Index: archaeology; civil society; cultural heritage; policy making

Common to all the examples is that repository keywords are the same as Scopus author keywords, taken from the published journal article. This example also demonstrates how research subject adds the broader disciplinary term of 'humanities', and that controlled Scopus keywords add specificity beyond those given in author keywords. Controlled and author keywords combined provide a larger potential set of subject access points.

Example B.

Title: "At the threshold of retirement: From all-absorbing relations to self-actualization"

- Repository index terms
 - Keywords: cultural structures; emancipation; gender equality; I-we balance; retirement; Sweden
 - Categories: cultural studies; sociology
 - Research subjects: humanities, cultural sociology
- Scopus index terms
 - Author keywords: Cultural structures; emancipation; gender equality; I-we balance; retirement; Sweden
 - Indexed keywords:
 - EMTREE: gender; human; life; retirement; self actualization; social structure; Sweden; welfare; female; human relation; lifestyle; perception; psychology; retirement; satisfaction; self concept; Sweden;

- MeSH: female; humans; interpersonal relations; life style; personal satisfaction; retirement; self concept; social perception; Sweden; women, working

In addition to characteristics from previous examples, here we also see that many controlled terms enhance the number of potential access points in retrieval. Also, it is apparent that applied vocabularies have not been mapped to each other and there does not seem to be an indexing policy in place that combines them in a user-friendly way. Instead, the user is given conceptual duplicates like “human relation” by EMTREE and “interpersonal relations” by MeSH; or “female” by EMTREE but “women” by MeSH. One other interesting point is that the repository classifies this as a cultural sociology paper, while EMTREE designates it as a psychology paper. Lack of policies in terms of singular and plural is also visible (“human” versus “humans”).

Example C.

Title: “Freedom to divorce or protection of marriage?: The divorce laws in Denmark, Norway and Sweden in the early twentieth century”

- Repository index terms
 - Keywords: Scandinavia; Scandinavian legislative cooperation; divorce; marriage; laws; women’s status; gender equality; individuality
 - Categories: economic history
 - Research subjects: humanities, history
- Scopus index terms
 - Author keywords: Scandinavia; Scandinavian legislative cooperation; divorce; marriage; laws; women’s status; gender equality; individuality
 - Indexed keywords:
 - EMTREE: article; Denmark; divorce; economics; education; ethnology; female; gender identity; history; individuality; jurisprudence; legal aspect; marriage; Norway; psychological aspect; spouse; Sweden; women’s rights;
 - MeSH: Denmark; divorce; gender identity; history, 20th century; individuality; judicial role; jurisprudence; marriage; Norway; spouses; Sweden; women; women’s rights

Example C further illustrates the issues discussed above. It also shows conceptual duplicates resulting from lack of indexing policies aligning author keywords and controlled keywords: the author uses “Scandinavia”, and both EMTREE and MeSH apply “Denmark”, “Norway” and “Sweden” – again leaving the burden for the user to align instead of having a rich cross-reference system to lead the user to all documents indexed with any of the relevant Scandinavian country. Similarly, “laws” is used in author keywords and “legal aspects” in EMTREE. Also, a genre keyword “article” gives little meaning here and confuses the user further. This example appears to illustrate what Bates et al. (1993) saw that the humanities need – facets of subjects expressed as geographical terms and periods (“history, 20th century”).

Example D.

Title: “Participation and enjoyment in play with a robot between children with cerebral palsy who use AAC and their peers”

- Repository

- Keywords: augmentative and alternative communication; cerebral palsy; children; complex communication needs; peers; robotics; play; participation; enjoyment
- Category: general language studies and linguistics
- Research subjects: humanities, Swedish
- Scopus
 - Author keywords: augmentative and alternative communication; cerebral palsy; children; complex communication needs; peers; robotics; play; participation; enjoyment
 - Controlled keywords:
 - Engineering Index: diseases; gesture recognition; human rehabilitation engineering; robotics; robots; augmentative and alternative communication; cerebral palsy; children; complex communication needs; enjoyment; participation; peers; play
 - Emtree: adult; cerebral palsy; child; communication aid; female; human; male; middle aged; peer group; preschool child; recreation; robotics; social participation
 - MeSH: adult; cerebral palsy; child; child, preschool; communication aids for disabled; female; humans; male; middle aged; peer group; play and playthings; robotics; social participation
 - Uncontrolled keywords: Augmentative-and-alternative communication; Cerebral palsy; Children; Complex communication needs; Enjoyment; Participation; Peers; Play

Here we see that there is a large number of exact duplicates between Emtree and MeSH. Evidence of the lack of consistent indexing policy in Scopus is visible from using too broad terms like “human”, which may lead to high recall at the expense of precision; lack of policy is also seen from non-consistent usage of singular versus plural (e.g., “human” by Emtree and “humans” by MeSH).

4.4 Index terms on published journal articles

In the sample, 203 (63.2%) articles had author keywords in Scopus. Of these, a majority of 141 (69.5%) had the same keywords both in the repository and in Scopus, because they are taken from the published article. There are in total 1015 keywords in the repository and 1035 in Scopus, of which 136 keywords only appear in the repository and 156 only appear in Scopus.

Figure 4 below illustrates the differences in the remaining 62 articles. A closer qualitative analysis shows that the differences are often due to individual additional terms in either the repository or Scopus. Here the differences are often minor and the index terms are conceptually similar. In a number of cases, the differences are due to minor variations in spelling and punctuation (e.g., Nyström vs. Nystrom, far-right vs. far right). Others are due to the use of Swedish or, more rarely, other languages like French and Japanese, as repository keywords, instead of English. Near synonyms also exist: for example, in one case Scopus uses the term “relations” and the repository uses “relationships”, in another Scopus uses “secondary school” while the repository uses “high school students”.

Figure 4. Number of unique keywords per article (only showing Jaccard matches <1)

A closer analysis against journal article keywords revealed that Scopus collects the keywords from pdf files, while the repository collects them from the article websites. These differ sometimes in singular versus plural, in spelling and in using completely different words, and occasionally also other languages (e.g., the pdf has them in English, while the website has the same keywords in Swedish). A few errors in copying have been found both in Scopus. In one case, the keyword was “Keywords_heritage”, suggesting that Scopus imported also the heading “Keywords”; in another, Scopus only copied the first keyword; in a few instances Scopus divided a multiple-word keywords into three separate keywords; and in one case, “D140” and “G200” were assigned as keywords by the repository, but they are not actually author keywords at all but classification codes, which is made clear when looking at the pdf of the article, but not at the article website.

Each journal will have different guidelines for keywords and few will have any controlled indexing guidelines. For example, the advisory entitled ‘Guide for Authors’ by Elsevier’s Journal of Academic Librarianship only states that manuscripts should “include keywords” (Elsevier, 2019d). Another of Elsevier’s journals, Journal of Archaeological Science: Reports, provides more detailed instructions: “provide 3-7 keywords, using British spelling and avoiding general and plural terms and multiple concepts (avoid, for example, ‘and’, ‘of’)” and “[b]e sparing with abbreviations: only abbreviations firmly established in the field may be eligible” (Elsevier, 2019e). These are just two examples but show that the instructions for authors can vary for journals with the same publisher.

5 Conclusion

As the humanities develop and expand to include more pronounced digital scholarship, it is important to provide quality subject access to the vast range of heterogeneous information

objects in digital services. This includes both primary sources and secondary ones, the latter being the focus of this study. The general objective of bibliographic systems for subject indexing should be that they allow the user to find anything and everything in the collection that is relevant to a certain topic, and this requires that controlled vocabularies need to be applied to ensure high precision and recall. In addition, humanities scholars would benefit from highly specific subject indexing, applications of controlled vocabularies that are faceted rather than pre-coordinated to cater for a range of possible topics discussed from different perspectives, and the inclusion in such vocabularies of named individuals as well as facets for space and time.

This study has shown that established bibliographic objectives to ensure subject access for humanities journal articles are not adequately supported in either the world's largest commercial abstracts and citation database Scopus, or the local repository of a public university in Sweden. The indexing policies in the two services do not seem to address the needs of humanities scholars for highly granular subject index terms with appropriate facets: indeed, no controlled vocabularies for any humanities discipline (e.g., Arts and Architecture Thesaurus) are used whatsoever. This is due to the fact that the repository categories primarily are intended for statistical analysis rather than retrieval, while Scopus simply prevents effective retrieval through omissions in its controlled vocabularies. This finding is well in line with East (2007), who established that ten individual databases in the humanities that he studied provided no controlled vocabularies for humanities resources. Only a minor portion of all articles in the study have any controlled vocabulary terms assigned in Scopus; those that do use index terms do so relying on controlled vocabularies from outside the humanities (EMTREE; MeSH, GEOBASE). The findings also demonstrate the lack of mapping between the vocabularies, which produces duplicates and renders the users unable to use terms from one vocabulary across all the resources.

Author keywords are complementary to controlled keywords; author keywords entered at the time of deposit also complement those found on journal articles. However, as usual in most repositories, authors are not trained in indexing nor are they provided with any indexing guidelines. To enhance and speed up the depositing process, training and guidelines should be provided to the authors. Social tagging could also be offered, with automatically produced tagging suggestions to choose from to enhance the uptake of this optional field. If the suggestions are derived from a suitable controlled vocabulary, retrievability of their articles would most likely contribute towards optimal precision and recall (cf. Golub et al., 2014).

In all, not much has changed since Tibbo (1994) wrote that indexing for the humanities lags behind the sciences because there is more funding available in the sciences, and because scientific literature is "simply easier to control" (p. 616). This seems to be the case in digital services such as repositories but also web archives and digitised cultural heritage collections, as well as cross-collection search services that cater for all scientific disciplines and fields, such as union catalogues and discovery services (see Golub, 2016; Golub, 2018). Notably, inconsistent and incomplete metadata and blending of controlled vocabularies, free keywords, and full-text automatic indexing create the biggest problems for subject searching (Dempsey, 2012; Fagan, 2011).

In addition to the action called for above, future research should include subject indexing and retrieval studies of monographs and book chapters that are common in the humanities, and probably less completely bibliographically represented and reviewed. Subject access in digital services for primary sources should also be studied in order to establish the current status and

to pinpoint the needs for improvement. Interfaces should be designed and tested to support query expansion, word-sense disambiguation, etc. based on specific user needs. All these should include user studies, analyses of real search sessions of humanities scholars and interdisciplinary scholars, as well as university students, cultural heritage professionals and the general public.

References

Alani H., C. Jones and D. Tudhope. (2000), "Associative and spatial relationships in thesaurus-based retrieval", in J. Borbinha and T. Baker (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, available at: https://doi.org/10.1007/3-540-45268-0_5

Bair, S. and Carlson, S. (2008), "Where keywords fail: using metadata to facilitate digital humanities scholarship", *Journal of Library Metadata*, Vol. 8 Iss. 3, pp. 249-262, available at: <https://doi.org/10.1080/19386380802398503>

Bates, M. J., Wilde, D. N. and Siegfried, S. (1993), "An analysis of search terminology used by humanities scholars: The Getty Online Searching Project Report Number 1", *Library Quarterly*, Vol. 63 No. 1, pp. 1-39, available at: <https://doi.org/10.1086/602526>

Bates, M.J. (1996). "The Getty end-user Online Searching Project in the humanities: report no. 6: Overview and conclusions", *College & Research Libraries*, vol. 57 no. 6, pp. 514–523, available at: https://doi.org/10.5860/crl_57_06_514

Bergenmar, J. and Golub, K. (2020), "Recovering a lost heritage. The challenge of LGBTQ literature for subject indexing", paper to be presented at Digital Humanities in the Nordic Countries Conference (DHN), 17-20 March, Riga, Latvia.

Borgman, C. L. (2007), *Scholarship in the digital age: information, infrastructure and the Internet*, MIT Press, Cambridge, MT.

Bundza, M. (2014), "The Choice Is Yours! Researchers Assign Subject Metadata to Their Own Materials in Institutional Repositories", *Cataloging & Classification Quarterly*, Vol. 52 Iss. 1, pp. 110-118, available at: <https://doi.org/10.1080/01639374.2013.852439>

Chickering, F. W. and Yang, S. Q. (2014), "Evaluation and Comparison of Discovery Tools: An Update", *Information Technology and Libraries*, Vol. 33 No. 2, pp. 5-30, available at: <https://doi.org/10.6017/ital.v33i2.3471>

Choi, Y. and Syn, S. Y. (2016), "Characteristics of Tagging Behavior in Digitized Humanities Online Collections". *Journal of the Association for Information Science and Technology*, Vol. 67, Iss. 5, pp. 1089-1104, available at: <https://doi.org/10.1002/asi.23472>

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

De Andrade, M. C. and Baptista, A. A. (2014), "Researchers' information needs in the bibliographic database: A literature review", *Information Services & Use*, Vol. 34 No. 3-4, pp. 241-248, available at: <https://doi.org/10.3233/ISU-140744>

De La Tierra, T. (2008), "Latina Lesbian Subject Headings: The Power of Naming", in Roberto, K. R. (Ed.), *Radical Cataloging: Essays at the Front*, McFarland, North Carolina, NC, pp. 94-102.

Dempsey, L. (2012), "Thirteen Ways of Looking at Libraries, Discovery, and the Catalog: Scale, Workflow, Attention", *EDUCAUSE Review*, available at: <https://er.educause.edu/articles/2012/12/thirteen-ways-of-looking-at-libraries-discovery-and-the-catalog-scale-workflow-attention> (accessed 29 November 2019).

East, J. W. (2007), "Subject Retrieval from Full-Text Databases in the Humanities", *Portal: libraries and the academy*, Vol. 7 No. 2, pp. 227-241, available at: <https://doi.org/10.1353/pla.2007.0018>

Elsevier (2017), "Scopus: Content coverage guide", available at: https://www.elsevier.com/_data/assets/pdf_file/0007/69451/0597-Scopus-Content-Coverage-Guide-US-LETTER-v4-HI-singles-no-ticks.pdf (accessed 27 November 2019).

Elsevier (2019a). "How do Author / Indexed keywords work?", Scopus: Access and use Support Center, available at: https://service.elsevier.com/app/answers/detail/a_id/21730/supporthub/scopus/kw/keyword/ (accessed 27 November 2019).

Elsevier (2019b), "What are the most frequent Subject Area categories and classifications used in Scopus?", Scopus: Access and use Support Center, available at: https://service.elsevier.com/app/answers/detail/a_id/14882/supporthub/scopus/kw/subjarea%28arts%29/ (accessed 27 November 2019).

Elsevier (2019d), "Guide for Authors [The Journal of Academic Librarianship]", available at: <https://www.elsevier.com/journals/the-journal-of-academic-librarianship/0099-1333/guide-for-authors> (accessed 26 November 2019).

Elsevier (2019e), "Guide for Authors [Journal of Archaeological Science: Reports]", available at: <https://www.elsevier.com/journals/journal-of-archaeological-science-reports/2352-409x/guide-for-authors> (accessed 26 November 2019).

Emanuel, J. (2011), "Usability of the VuFind next generation Online Catalog", *Information Technology and Libraries*, Vol. 30 No. 1, pp. 44-52, available at: <https://doi.org/10.6017/ital.v30i1.3044>

Fagan, J. C. (2011), "Discovery Tools and Information Literacy", *Journal of Web Librarianship*, Vol. 5 Iss. 3, pp. 171-178, available at: <https://doi.org/10.1080/19322909.2011.598332>

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, available at: <https://doi.org/10.1108/JD-05-2013-0056>

Golub, K. (2016), "Potential and Challenges of Subject Access in Libraries Today on the Example of Swedish Libraries", *International Information & Library Review*, Vol. 48 Iss. 3, pp. 204-210, available at: <https://doi.org/10.1080/10572317.2016.1205406>

Golub, K. (2018), "Subject Access in Swedish Discovery Services", Knowledge Organization, Vol. 45 No. 4, pp. 297-309, available at: <https://doi.org/10.5771/0943-7444-2018-4-297>

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> (accessed 30 November 2019).

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> (accessed 30 November 2019).

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", *Library Quarterly*, Vol. 68 No. 4, pp. 406-30, available at: <https://doi.org/10.1086/603001>

Langridge, D. W. (1976), *Classification and indexing in the humanities*, Butterworths, London.

Markey, K. (2007), "The Online Library Catalogue: Paradise Lost and Paradise Regained?", *D-Lib Magazine*, Vol. 13 No. 1/2, available at: <https://doi.org/10.1045/january2007-markey>

Meadow, K. and Meadow, J. (2012), "Search Query Quality and Web-Scale Discovery: A Qualitative and Quantitative Analysis", *College & Undergraduate Libraries*, Vol. 19 Iss. 2-4, pp. 163-75, available at: <https://doi.org/10.1080/10691316.2012.693434>

National Library of Sweden (2019), "Nationella riktlinjer för beskrivning av forskningsoutput: Version 1.3", available at: https://www.kb.se/download/18.d0e4d5b16cd18f600e1ecd/1573725693925/NR_v1_3.pdf (accessed 30 November 2019).

Organisation for Economic Co-operation and Development (2007), "Revised field of science and technology (FOS) classification in the Frascati manual", available at: <http://www.oecd.org/science/inno/38235147.pdf> (accessed 30 November 2019).

Osborne, H. M. and Cox, A. (2015), "An Investigation into the Perceptions of Academic Librarians and Students towards Next-generation OPACs and Their Features", *Program: Electronic Library and Information Systems*, Vol. 49 No. 1, pp. 23-45, available at: <https://doi.org/10.1108/PROG-10-2013-0055>

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> (accessed 30 November 2019).

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. 1, pp. 108-118.

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 Iss. 5, pp. 273-291, available at: [https://doi.org/10.1002/\(SICI\)1097-4571\(199306\)44:5<273::AID-ASI3>3.0.CO;2-X](https://doi.org/10.1002/(SICI)1097-4571(199306)44:5<273::AID-ASI3>3.0.CO;2-X)

Statistics Sweden. (2016). "Standard för svensk indelning av forskningsämnen 2011: Uppdaterad augusti 2016", available at: <https://www.scb.se/dokumentation/klassifikationer-och-standarder/standard-for-svensk-indelning-av-forskningsamnen/> (accessed 30 November 2019).

Tibbo, H. R. (1994), "Indexing for the humanities", *Journal of the American Society for Information Science*, Vol. 45 Iss. 8, pp. 607-619, available at: [https://doi.org/10.1002/\(SICI\)1097-4571\(199409\)45:8<607::AID-ASI16>3.0.CO;2-X](https://doi.org/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, available at: <https://doi.org/10.1108/00220410610673873>

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", *Journal of Academic Librarianship*, Vol. 33 Iss. 3, pp. 327-37, available at: <https://doi.org/10.1016/j.acalib.2007.01.018>

Yi, K., Beheshti, J., Cole, C., Leide, J. E. and Large, A. (2006), "User search behavior of domain-specific information retrieval systems: An analysis of the query logs from PsycINFO and ABC-Clio's Historical Abstracts/America: History and Life", *Journal of the American Society for Information Science and Technology*, Vol. 57 Iss. 9, pp. 1208-1220, available at: <https://doi.org/10.1002/asi.20401>

Wiberley, Jr., S. E. (1983), "Subject access in the humanities and the precision of the humanist's vocabulary", *Library Quarterly*, Vol. 53 No. 4, pp. 420-433, available at: <https://doi.org/10.1086/601405>

Wiberley, Jr., S.E. (1988), "Names in space and time: The indexing vocabulary of the humanities", *Library Quarterly*, Vol. 58 No.1, pp. 1-28, available at: <https://doi.org/10.1086/601949>

Wiberley, Jr., S. E. (1991), "Habits of humanists: Scholarly behaviour and new information technologies", *Library Hi Tech*, Vol. 9 No. 1, pp. 17-21.

Wiberley, Jr., S. E., and Jones, W. G. (1989), "Patterns of information seeking in the humanities", *College & Research Libraries*, Vol. 50 No. 6, pp. 638-645, available at: https://doi.org/10.5860/crl_50_06_638