Skill-Building in Subject Representation: Assessing Learning Outcomes through Analysis of Student-Created Metadata

Vyacheslav Zavalin[a]

[a]Texas Woman's University, USA

vzavalin@twu.edu

ABSTRACT

Representation of topics, places, events and other entities that information resource is about (referred to as subject representation) is crucial in facilitating access to information. It is an integral part of one of the core competencies of information professionals defined by the American Library Association (ALA). Students in ALA-accredited professional degree programs develop knowledge and skills that prepare them for analyzing what information resource is about and representing this aboutness with a combination of free-text keywords/keyphrases and terms from controlled vocabularies that are widely used by practitioners in the field. This paper reports the results of the evaluation of professional competencies developed by future information professionals in the representation of an information resource's aboutness in a real-life practical setting. The dataset used in this evaluation consists of over 18 thousand metadata records -- created over 12 years, mostly by students -- in the major digital collection that provides access to historical patents.

ALISE RESEARCH TAXONOMY TERMS

Information organization and retrieval -- Metadata; Information practices -- Education; Education of information professionals -- Curriculum

AUTHOR KEYWORDS

Subject access; subject representation; STEM materials; practical competencies assessment; metadata evaluation

DOI: https://doi.org/10.21900/j.alise.2024.1665

INTRODUCTION

Subject access is one of the main kinds of access to information resources, especially in large and growing digital repositories (e.g., Bates, 2002). Subject access -- whether in digital repository of today or in analog databases of the past (e.g., print library catalogs) is enabled by providing subject metadata: the terms and phrases that represent what the resource is about, that was aptly termed "aboutness" by several influential researchers of information science (e.g., Fairthorne, 1969; Wilson, 1968). Subject access topics have been explored for a long time, especially in relation to information seeking and information retrieval (Hjørland, 1997) and have become fundamental concepts in the field of library and information science (e.g., Golub, 2014). Cochrane (1979) coined the operational definition of subject access as both the user exploration of the database by subject and the information-professional-performed subject cataloging. The latter includes systematic (e.g., classification system), topical (e.g., subject headings), and natural (e.g., keywords) representations of aboutness (Cochrane, 1979). Analysis and representation of information resource's aboutness is a complex task that includes multiple steps and is error-prone, especially for beginning metadata creators (e.g., Taylor & Joudrey, 2002; Zavalina & Burke, 2021). One of the most important criteria for evaluating the quality of metadata in general, and the quality of subject representation in particular is accuracy. Accurate metadata is free of factual and typographical errors, follows the guidelines and standards, does not misrepresent the information resource (e.g., Bruce & Hillmann, 2004).

Patents are important information resources for documenting inventions, tracing intellectual rights, learning the history of science, etc. Access to full-text of historical patents is provided online, including through the Texas Patents digital collection that includes over 18 thousand of official patents issued by the U.S. Patents Office in the 19th and early 20th century. The collection is well-known and is heavily used by those interested in the history of technology, including STEM researchers, educators, students, inventors, and the general public. For each item in this collection, in addition to the full-text searchable PDF document containing the full text of the patent and illustrations, descriptive metadata is provided to facilitate discovery of these materials that serve as important evidence of technological developments.

The Texas Patents collection's metadata records follow a locally-developed metadata scheme – used for all collections in this regional aggregation – based on the Dublin Core. Among the 21 metadata fields in this metadata scheme, 2 repeatable fields are intended for representation of the resource's aboutness. One of them, the Coverage field, is for representing the places (including the places of residence of inventors in patents metadata records) and the time periods that the resource covers. The Subjects metadata field is intended for representing other kinds of aboutness (topics, etc.) and can be used with terms from various controlled vocabularies such as Library of Congress Subject Headings, Thesaurus of Graphic Materials, etc., as well as with free-text keywords. Metadata managers of the aggregation that this collection is part of primarily rely on metadata students – and to much lesser extent, on volunteers, and other learners (e.g., new employees) – in

creating metadata for patents in this collection (Zavalina, Phillips, & Tarver, 2017). As part of patent upload to the digital repository, metadata manager creates a draft metadata record with very little information included and most applicable metadata fields left blank or having placeholder data values. This draft record remains hidden until a metadata editor (usually a metadata student) completes it in the beginner-friendly online form while consulting the collection-specific metadata guidelines, and makes it visible.

As of 2017, over 300 library and information science students contributed metadata to Texas Patents collection through the hands-on real-life assignment that they completed in the introductory graduate metadata course in 11 semesters, with approximately 150 metadata records per semester (Zavalina, Phillips, & Tarver, 2017). This popular patent metadata assignment has continued to be used in the course and as of the time of writing this paper, it has been completed by students in a total of 31 semesters. Students work on this assignment after completing the first 4 course modules (out of the total of 8) that include introduction to metadata functions, uses and types, components of a metadata scheme, data content standards that guide metadata creation, data value standards (various controlled vocabularies) used in metadata creation, and syntax used for encoding metadata. In this practical assignment, students are expected to apply in patent metadata record creation everything they learned so far except the XML syntax that they apply in later assignments that focus on standard metadata schemes: Dublin Core, Metadata Object Description Schema, and Visual Resources Association Core 4.0.

The metadata management history of the Texas Patents collection makes it a highly suitable source of evidence-based data to evaluate the extent to which students develop the important professional competencies of describing information resources (specifically the STEM documentation). With no funding designated to support professional positions responsible for metadata editing, metadata managers rarely have a chance to make corrections in student-created patent metadata records, and limited evaluation of this metadata was done by the course teaching team during the first several years of implementing this assignment (Zavalina, Phillips, & Tarver, 2017). Since 2020, as part of course revisions efforts, evaluation of these metadata records has become much more involved and students are provided with detailed feedback (specifically related to subject representation). However, there is no requirement for students to improve their records using this feedback, so most of these records remain as they were when a student metadata editors completed them. I focused my analysis of student-created metadata in the Texas Patents collection on evaluation of professional competencies related to subject representation.

METHOD

The data was collected by the author in March 2024 through the Portal to Texas History metadata dashboard. It was exported, processed and analyzed with the help of spreadsheet software and Python Pandas. At that time, the Texas Patents collection included 22780 metadata records, including 18455 completed and visible to the public, which I analyzed with the focus on terms and

phrases used in the Subjects metadata field: one of the 2 metadata fields intended to represent aboutness in this digital repository. The visualizations presented in figures 2, 3, and 4 were created in Python.

This study sought to address the following research question: How is the use of free-text keywords and various controlled vocabularies distributed in the Subjects metadata fields in the dataset? To answer it, I examined the attribute-value pairs of the Subjects element that designate the way in which each instance of this field is used. For example, the "qualifier="KW"" indicates that this instance of the Subjects field relies on the free-text keywords, and the "qualifier="LCSH"" indicates that the term included in it is from the Library of Congress Subject Headings. I also examined the frequency distributions for various specific subject terms used by students in the Subjects field. I excluded from this analysis the generic placeholder subject terms added by the metadata manager in the process of uploading digitized patents: "Patents -- {StateName}" (n=12263) and "Inventions" (n=10387) in the Subject metadata field instances designated as using LCSH controlled vocabulary terms, and "Science and technology" (n=21433) in those designated as using the locally-developed hierarchical controlled vocabulary of topical terms.
In addition, my preliminary analyses explore the accuracy in subject representation, including:

- How accurately is the aboutness of a patent represented?
- How accurately are controlled vocabularies applied?

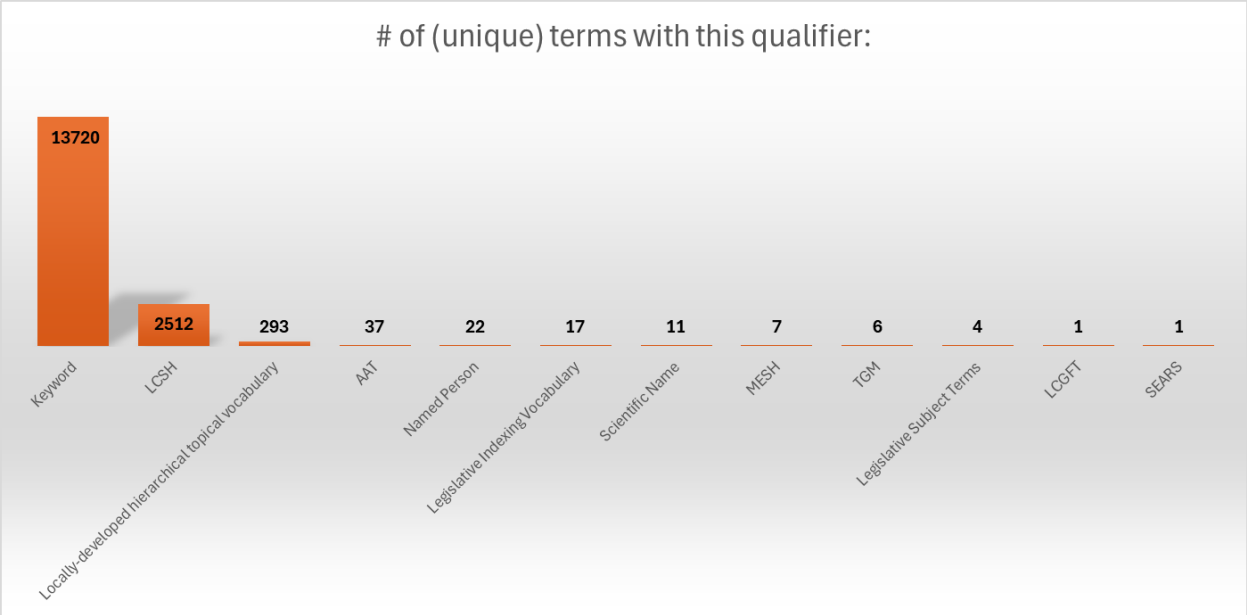
FINDINGS AND DISCUSSION

This paper reports a research project that used an evidence-based case-study approach to examine the effectiveness of integrating the subject representation professional competency in education to prepare students to real-life organizing information tasks. Results demonstrate three major uses of the Subjects field in my dataset of student-created patent metadata dataset (Figure 1). The majority of terms used were the free-text keywords (n=13720 unique terms or 82.5 % of all unique terms observed in the Subjects metadata fields). In addition, terms from two controlled vocabularies that collection-specific metadata guidelines strongly recommend using in Subjects fields were often added by students: Library of Congress Subject Headings LCSH (n=2512), and the locally-developed hierarchical controlled vocabulary of topical terms (n=293). Nine additional controlled vocabularies were found to be used at much lower levels: Art and Architecture Thesaurus AAT (n=37), locally-developed Named Person controlled vocabulary (n=22), Legislative Indexing Vocabulary LIV (n=17), local Scientific Name vocabulary (n=11), Medical Subject Headings MESH (n=7), Thesaurus for Graphic Materials TGM (n=6), Legislative Subject Terms (n=4), Library of Congress Genre/Form Terms LCGFT (n=1), and Sears list of subject terms (n=1). These additional controlled vocabularies are not highlighted in the collection-specific metadata guidelines, but some of them are covered in the metadata course before this practical assignment and in other related courses (e.g., cataloging and classification),

so student metadata creators might be including terms from them based on their pre-existing knowledge of these controlled vocabularies.

Figure 1

Controlled vocabularies and free-text subject terms used in the dataset



I observed the long-tail distribution of the free-text keywords added by student metadata editors, with 96.56% of terms appearing 10 or fewer times and 74.26% only once. The head of this distribution consists of 472 unique terms used between 10 and 243 times (Figure 2).

Figure 2

472 most frequently used free-text keywords

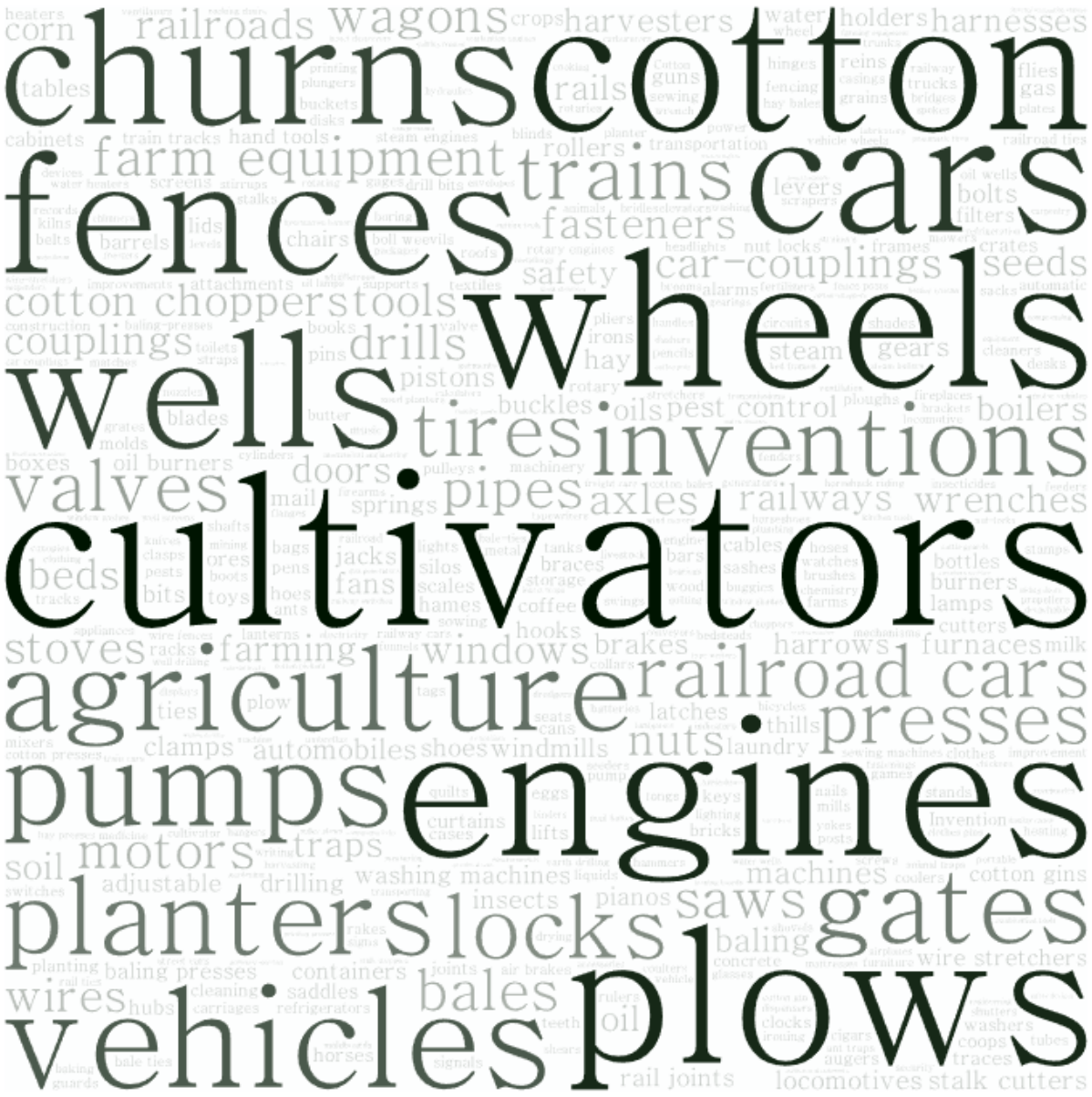


Table 1 shows the distribution of the top 10 most commonly used subject terms for each of the 3 most common uses of the Subjects metadata field by student metadata editors. It represents free-text keyword terms used 136 or more times, LCSH terms used 12 or more times, and locally-developed hierarchical topical controlled vocabulary terms used 313 or more times. Figures 3 and 4 show all terms from these two controlled vocabularies observed in the dataset. The terms have mostly agricultural and oil industry focus.

Table 1

Top 10 most commonly used terms from 3 groups

Locally-developed hierarchical topical controlled vocabulary terms	Count	LCSH terms	Count	Keywords terms	Count
Science and Technology - Tools	2199	Tires	19	cultivators	243
Agriculture - Farm Equipment	1515	Agricultural machinery	18	cotton	195
Agriculture - Farming - Cotton	801	Automobiles	18	plows	178
Business, Economics and Finance - Transportation - Automobiles	544	Cotton-picking machinery	16	wheels	164
Agriculture - Farm Equipment - Plows	531	Cotton-picking machinery -- Patents.	16	engines	163
Business, Economics and Finance - Transportation - Railroads - Trains	424	Oil well drilling	16	fences	153
Business, Economics and Finance - Transportation - Railroads	394	Wheels	14	agriculture	151
Agriculture - Processing and Storage	377	Agriculture	13	inventions	151
Architecture - Construction	371	Farm equipment	13	vehicles	141
Agriculture	313	Locks and keys	12	planters	136

Figure 3

Distribution of subject terms from LCSH

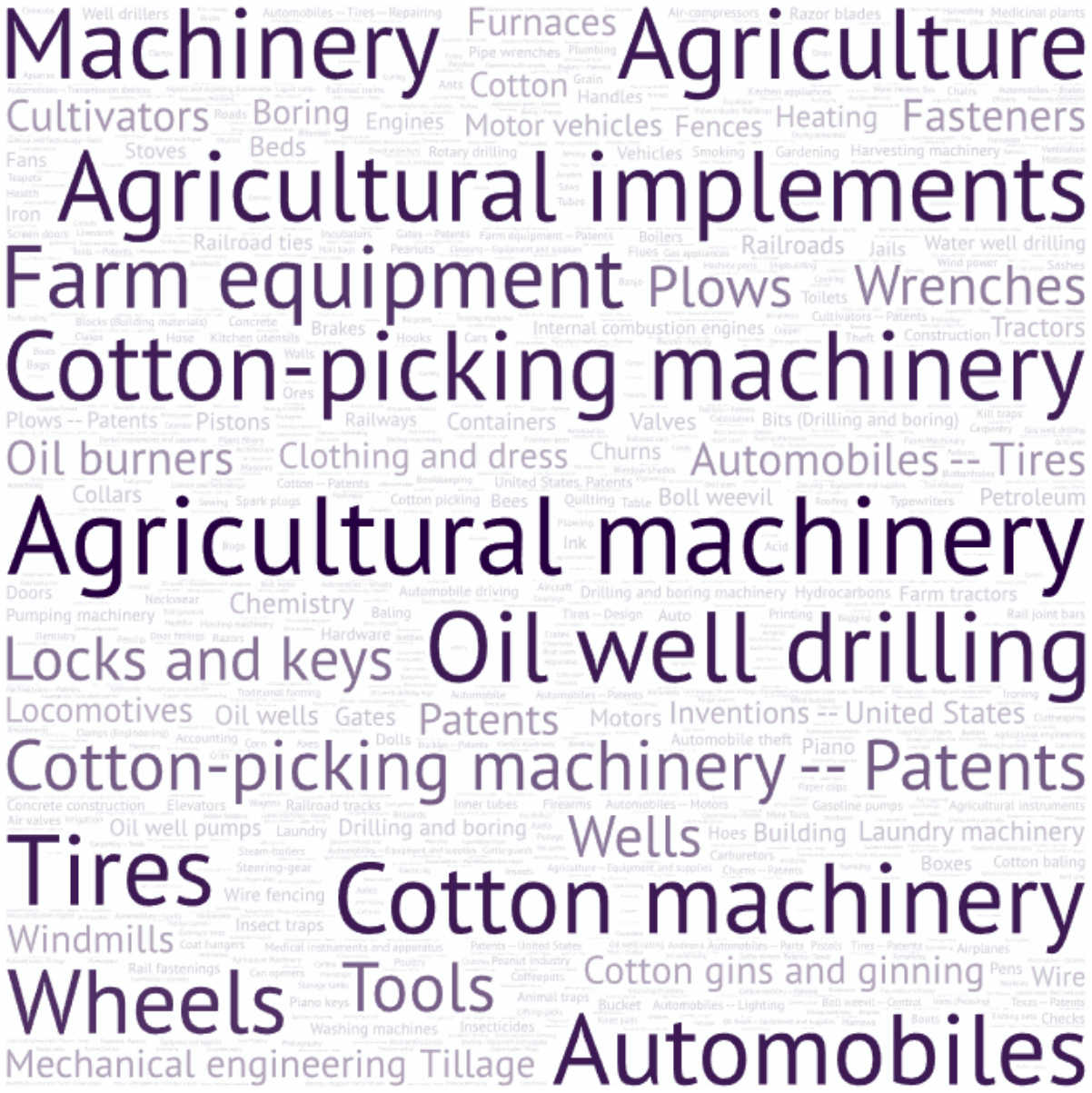


Figure 4

Distribution of subject terms from the locally-developed hierarchical controlled vocabulary

Preliminary analysis of the accuracy of the patient aboutness representation revealed a substantial number of cases of misplaced use of subject terms by student metadata editors. In 7.85% of the Subjects field instances following the locally-developed controlled vocabulary, geographical headings that are not part of this topical vocabulary were used (e.g., “Places - United States - New Jersey”, etc.). My analyses also examined the accuracy of the data entry of the authorized forms of the controlled vocabulary terms. After identifying and fixing data inconsistencies and duplicates and removing leading, trailing, multiple spaces, and other characters, the total number of unique LCSH terms in the dataset was reduced from 2512 to 2219, which means 11.5% of LCSH terms were entered by students inaccurately. While

punctuation and capitalization issues might not matter for the users who have already found and are examining the metadata record in a repository, they impede the discovery of metadata records in searching by subject by negatively affecting the collocation of search results (e.g., Beall & Kafadar, 2013).

CONCLUSIONS AND FUTURE RESEARCH

This study sought to address the following research question: How is the use of free-text keywords and various controlled vocabularies distributed in the Subjects metadata fields in the Texas Patents collection that heavily relies on student-created metadata? The preliminary findings reported in this paper demonstrate the prevalence of free-text keywords in representing the aboutness of patents by metadata students and the use of several controlled vocabularies, with varying degrees of accuracy (with accuracy issues assessed at the level of formatting errors, and the use of the Subjects field for categories of terms that belong to another metadata field). The preliminary analysis focused on only one of the two metadata fields intended to represent aboutness in this digital repository. My next step is to analyze geographical aboutness representation in the Coverage metadata field. I will also further evaluate accuracy of student-created descriptions in this dataset in terms of adequacy of selected topical and geographical terms to represent what the patent is about. I will compare the findings with those obtained in other existing studies of student-created metadata (e.g., Zavalina & Burke, 2021; Zavalin & Zavalina, 2023; Aljalahmah & Zavalina, 2023, etc.)

Subject representation is a complex process, and building subject representation skills requires practice. With the patent metadata project being the first practical assignment in which students of this graduate course create metadata, the level of practical experience in subject representation (and especially that with the use of controlled vocabularies) most students have at the time of working on this assignment is very limited. Future studies will track the evolution of the subject representation skills over time by comparing the subject representation effectiveness in this assignment completed earlier in the semester with three practical metadata creation assignments that students complete later in the course. With the emphasis on renewal and strengthening the connections between teaching, learning, and practice, I assume that more library and information science educators include in their courses real-life practical assignments. Comparative studies of the openly-accessible output of such practical projects (e.g., in the form of harvestable metadata) across different courses and programs are needed. Such studies will allow assessing the effectiveness of the professional degree curricula in developing the key skills in adequately representing various information resources to facilitate their discoverability through subject access. Their findings will help augment the curriculum design of these practical experiences to best prepare the students to meet the information representation needs and to thrive in various information contexts.

REFERENCES

Aljalahmah, S.H., & Zavalina, O. L. (2023). Dublin Core metadata created by Kuwaiti students: Exploration of quality in context. In, *Information for a Better World: Normality, Virtuality, Physicality, Inclusivity: 18th International Conference, iConference 2023, Proceedings* (pp. 544–551). https://doi.org/10.1007/978-3-031-28032-0_41

Bates, M. (2002). After the dot-bomb: Getting web information retrieval right this time. First Monday, 7(7). Retrieved from https://firstmonday.org/ojs/index.php/fm/article/view/971/892

Beall, J., & Kafadar, K. (2004). The effectiveness of copy cataloging at eliminating typographical errors in shared bibliographic records. *Library Resources & Technical Services, 48* (2), 92–101. https://doi.org/10.5860/lrts.48n2.92-101

Bruce, T. R., & Hillmann, D. I. (2004). The continuum of metadata quality: Defining, expressing, exploiting. In D. Hillman & L. Westbrook (Eds.), *Metadata in practice* (pp. 238–256). Chicago, IL: American Library Association. Retrieved from https://www.researchgate.net/publication/247818823_The_Continuum_of_Metadata_Quality_Defining_Expressing_Exploiting

Cochrane, P. (1979). Universal Subject Access (USA): Can anyone do it? In *Redesign of Catalogs and Indexes for Improved Online Subject Access*: Selected papers of Pauline A. Cochrane, Phoenix, Ariz.: Oryx Press, 1985, pp. 223–238.

Fairthorne, R.A. (1969). Content analysis, specification and control. *Annual Review of Information Science and Technology, 4,* 73–109.

Golub, K. (2014). Subject Access to Information: An Interdisciplinary Approach. Libraries Unlimited.

Hjørland, B. (1997). The concept of subject or subject matter and basic epistemological positions. In Information Seeking and Subject Representation: An Activity-Theoretical Approach to Information Science. Westport CT: Greenwood Press, 55-103.

Taylor, A. G., & Joudrey, D. N. (2002). On teaching subject cataloging. Cataloging & Classification Quarterly, 34(12), 221–230. https://doi.org/10.1300/j104v34n01_13

Wilson, P. (1968). Two kinds of power: An essay on bibliographic control. Berkeley: University of California Press.

Zavalin, V.I., & Zavalina, O.L. (2023). Exploration of accuracy, completeness, and consistency in metadata for physical objects in museum collections. In, *Information for a Better World: Normality, Virtuality, Physicality, Inclusivity: 18th International Conference, iConference 2023, Proceedings* (pp. 83-90). https://doi.org/10.1007/978-3-031-28032-0_7

Zavalina, O. L., & Burke, M. (2021). Assessing skill building in metadata instruction: Quality evaluation of Dublin Core metadata records created by graduate students. *Journal of Education for Library and Information Science, 62*(4), 423–442. https://doi.org/10.3138/jelis.62-4-2020-0083

Zavalina, O.L., Phillips, M., & Tarver, H. (2017). Quality assurance and evaluation of change for patent metadata. *Proceedings of the Association for Information Science and Technology, 54* (1), 842-843. https://doi.org/10.1002/pra2.2017.14505401180