A DDC Visual Interface for Metadata Exploration

Xia Lin, Michael Khoo, Jae-wook Ahn
Drexel University, USA
{xlin, mjk326, ja626}@drexel.edu

Ceri Binding, Douglas Tudhope
University of South Wales, UK
{douglas.tudhope, ceri.bindling}@southwales.ac.uk

Hilary Jones, Diana Massam
MIMAS, University of Manchester, UK
{Hilary.Jones, Diana.Massam}@manchester.ac.uk

Abstract

This paper presents a visualization interface for DDC-enriched metadata collections. Three sets of metadata from three different digital libraries were aggregated and re-indexed. Automatic analysis was performed to assign one or more DDC classes to each individual metadata record. A comprehensive search and exploratory interface was designed and implemented to include dashboard views, localized views, and universe views of DDC and the metadata collections. Finally, an experiment was conducted to test and compare how subjects interacted with different views for metadata search, exploratory and resource discovery.

Keywords: visualization interface; metadata exploration; digging into metadata; Dewy Decimal Classification; DDC; automatic classification; interface testing and evaluation.

1. The Digging Into Metadata Project

As one of the "Digging into Data" projects (Digging into Data Challenge, n.d.), the *Digging Into Metadata* project investigated innovative methods for metadata enhancement and reuse. The project was conducted among our three research groups in the last three years (January 2012 to December 2014). It addresses three crucial needs of enhancing metadata for finding, retrieving, and sharing digital resources: the need to aggregate metadata records in multiple digital libraries, the need to perform automatic analysis of metadata collections and use the results to enhance individual metadata records, and the need to create new interfaces to access digital resources through the enhanced metadata.

An assumption being tested in this project is that some knowledge organization systems (KOS) can be mapped automatically to a collection of metadata to enhance semantic connections among the metadata records. The test bed for the project is the mapping of Dewey Classification System (DDC) numbers to an aggregated set of metadata from three digital libraries: the National Science Digital Library (U.S.A.: http://nsdl.org/) (also including the Digital Library for Earth Systems Education, DLESE: http://www.dlese.org/); the Internet Public Library (USA: http://www.ipl.org/) (also including the Librarians' Internet Index (LII); and Intute (U.K.: http://www.intute.ac.uk/). The IPL was founded in 1995 in the U.S. as an online reference service, and then began developing digital collections (Janes, 1998). In 2008, the IPL merged with the Librarians' Internet Index (LII), and the IPL and LII metadata was crosswalked to Dublin Core and added to a Fedora database (Khoo & Hall, 2010). The NSDL is an NSF-funded federated multi-disciplinary STEM library, with a central Dublin Core metadata repository currently housed at https://nsdl.oercommons.org/. The library includes metadata from a number of individual domain-specific portals, or 'Pathways' (e.g. Zia, 2004; Bikson et al., 2011). As the NSDL Pathways were independent entities, the same resource could be cataloged in different ways by different Pathways. Finally, Intute was developed in the U.K. by a grass-root community dedicated to online educational resource discovery (Joyce, 2008; Williams, 2006). Much Intute metadata was inherited from a series of previous partners and educational consortiums in the U.K., and as a result, each Intute resource has both a Dublin Core record, and can also have

additional subject classification metadata stored in separate SQL tables, partly a legacy of previous specific subject catalogs to suit the needs of users of particular collections. Each of these digital libraries therefore had histories dating back to at least the early 2000s. Further, the metadata for each collection included the standard Dublin Core elements (*title*, *description*, *subject*, *identifier*, etc.), although due to the large number of contingencies in the histories of each library, each catalog also contained a range of qualified elements. The harvesting and normalization processes were therefore quite complicated (Khoo et al., forthcoming).

For this purpose, we have designed and tested a workflow pipeline that includes: (1) harvesting metadata records; (2) extracting metadata from designated fields in each record; (3) analyzing this metadata and generating weighted key terms that represent 'aboutness'; (4) using these weighted key terms to generate one or more DDC numbers that can then be added back to the records concerned; and (5) using the new DDC numbers in each records to build tools that allow users to search and browse multiple DDC classes at the same time. The design and discussion of this pipeline has been reported in (Binding, et al., 2013, Khoo, et al., forthcoming).

The mapping of DDC to metadata records has also been reported in (Khoo, et al., 2012). For this process, two major components were developed: MASH (Metadata Analysis, Sharing, & Harvesting), and DISTIL (Document Indexing and Semantic Tagging Interface for Libraries). The MASH component includes the process of (1) cleaning metadata records harvested from multiple digital libraries, (2) extracting nouns from selected metadata elements of each record, (3) calculating Term Frequency (TF) scores after applying known language processing procedures such as tokenization, stop word removal, and stemming, and (4) ranking the noun list using a TF-score based weight schema. As a result, MASH produces a ranked list of nouns that can be sent to DISTIL for bulk analysis.

The goal of the DISTIL component is to generate automatically one or more DDC class numbers for each metadata record, which can then be used to support searching and browsing. DISTIL follows a document classification approach with two main phases. The first phase attempts to match a weighted combination of the key terms in a metadata record against the entry vocabulary of DDC. This results in many matches both across different DDC hierarchies and at different levels within a given hierarchy. The second phase takes account of matches within hierarchies, aggregating lower level matches to broader parents. Depending on the configuration, outliers without any ancestor or descendant matches can be discarded. Essentially, DISTIL determines an overall degree of match between two sets of records: a metadata record, and DDC class headings, including DDC Relative Index headings. It then generates an output for each metadata record with the top *N* DDC numbers assigned to that record.

This paper reports the third part of the project on designing and developing new interfaces that will take advantages of the enhanced metadata (the metadata records with DDC numbers assigned) for resource discovery across multiple heterogeneous digital collections.

2. The Interface for DDC-enriched Metadata

Having the metadata records with DDC numbers automatically assigned has the potential to facilitate searching, browsing, and resource discovery. To fulfill the potential, specialized user interfaces need to be created (Slavic, 2006). When discussing desirable interface functions to support the use of classification systems for searching and browsing, Slavic emphasized that the advantages of using a hierarchical/facet classification for browsing and retrieval depend on the strength of the interface -- "The power of the interface is in supporting visualisation that will 'convert' what is potentially a user-unfriendly indexing language based on symbols, to a subject presentation that is easy to understand, search and navigate."

Creating visual exploratory interfaces that integrate classifications, subject terms, and search results is therefore a major goal of *the Digging into Metadata* project. In particular, we envision that:

- The interface should take full advantage of DDC classification structures to create "views" to guide the user. DDC has well-built hierarchical structures and associative class relationships. The structures may be used to create global views of the collections and localized views of queries and search results.
- The interface should utilize the new associations among metadata records as a result of assigning multiple DDC classes to metadata records. When metadata records are associated with DDC classes, new associative relationships are formed through the DDC-metadata relationships, which can also be converted into new metadata-metadata relationships. Both relationships might be used to guide user's searching and browsing activities, including querying and filtering.
- The interface should support user's interaction with both DDC class structures and search results. The use of DDC structures should help, rather than hinder, the user's interactions. The user does not need to be familiar with DDC structures, and the user should have choices of what they want to "see" and when to "see" or use the DDC knowledge structures.

These three requirements have become our design principles for creating the new interfaces. While they look simple and straight forward, the implementation has proved to be quite challenging.

2.1. The Interface and its Components

To build the interface, a solid indexing and searching platform was needed. We chose the open-source package, Solr (http://lucene.apache.org/solr), for this purpose. Three different data types are indexed in Solr for the project:

- Metadata: titles, descriptions, subject descriptors, and URLs of the digital resources
- DDC: DDC class, division, and section numbers and the class labels delivered from DISTIL
- Digging Statistics: context analysis results such as term frequencies and scores, DDC numbers co-occurrence frequencies, etc.

Currently, the Solr platform provides access to about 79,500 metadata records from IPL, Intute, and NSDL. These are the records that have been analyzed and for each of them one or more DDC classification numbers are assigned.

The front-end was built as a web application with html, JavaScript, and visualization tools such as D3.js (http://d3j3.org) and Sigma.js (http://sigmajs.org). Figure 1 shows a sample display of the interface. As shown in this example, the interface contains three major parts, as well as the top bar where the user may enter a search query. On the left is the area for DDC tree display; on the right, the bottom part is the display area for retrieval results and the top part is a tabbed area for three different visual widget displays, which are described next.

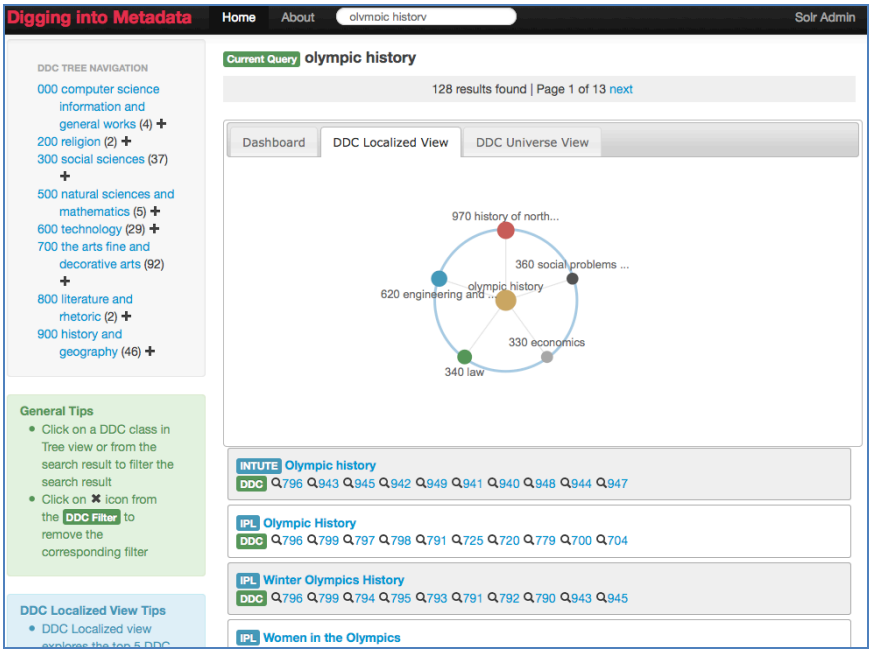


FIG. 1. A sample display of the digging user interface

2.2. Three Interactive Widget Displays

Using tabs is an effective way to provide alternative views of information within a limited display space. In figure 1, the tab selected is the "DDC localized view" which shows on a circle the 5 closest DDC classes related to the query. The circle format was chosen for its simplicity, interactivity, and easy repetition. In this example, the query "Olympic history" is most closely related to {DDC970, History of North America; 620, Engineering & allied operations; 340, Law; 330, Economics; & 360, Social problems & social services}. This list serves as the most succinct interpretation of issues related to the query "Olympic history." Displaying it on a circle is much easier to read and interact with. The circle effectively extends the query to a "query ring" with which the user can interact to refine his or her searches. For example, the user may click on a DDC number to bring up another "ring" with a new set of relevant DDC numbers, and he or she may choose a DDC number on the rings to add to the query to narrow down the search results. To a large extent, the usefulness of the DDC "query ring" will depend on the accuracy of the automatic DDC class assignment created by the DISTIL process. It will provide similar functions like those "synonym rings" described in Zeng (2006).

The other two tabs also provide unique functions for the user to interact with both search results and related DDC classes. Figure 2 shows the Dashboard display for the same search query "Olympic history." The display shows the top 10 DDC numbers occurring in the retrieved results, arranged by their occurrence frequency. The user can, for example, click on "725, public structures" to retrieve 55 documents which is the result of search query "Olympic history AND DDC:725."

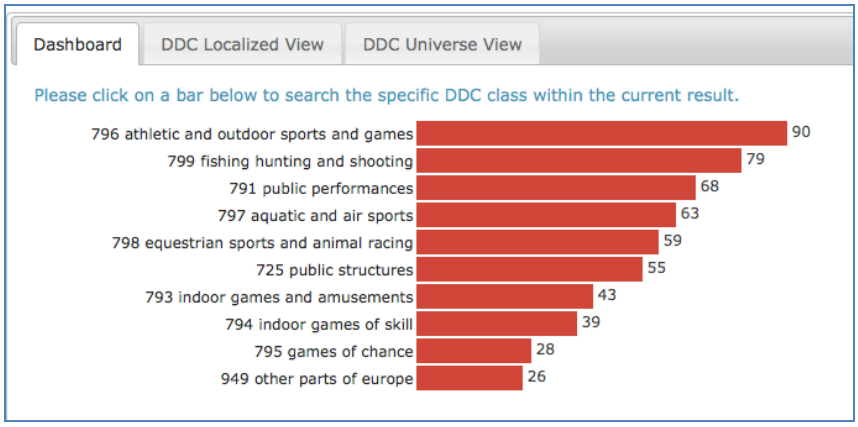


FIG. 2. The Dashboard display for the search query "Olympic history". It lists the top 10 DDC classes of the retrieved results by their occurrence frequencies.

The third tab, the "DDC Universe View", provides more elaborate functions to explore the network of metadata records (figure 3). While the DDC localized view is a bottom-up approach to exploring the collection – the user starts from a query and moves iteratively towards the target – the DDC Universe View is a top-down approach. It starts with showing the entire DDC universe of the underlying digital collections as a large graph in which DDC classes are represented as nodes and documents are depicted as links. Through the DDC Universe View, the user can zoom in and out dynamically, pan horizontally or vertically, or jump to a specific location of the network by a given DDC class. The user may (1) explore how a given DDC class is connected to other DDC classes in this collection, (2) learn what are the major clusters of the whole collection and what are the main classes within each cluster, and (3) locate seemingly unrelated but interesting new relationships by following the links. These activities help *serendipitous* discovery of new connections and benefit the exploratory nature of search. In this example, when the user zooms in to the first DDC class in the search result Dashboard display (DDC 796), he discovers that DDC 307.3 ("structure; Abandoned buildings") is one of the closest DDC nodes to DDC796. This indicates that the topic on "Abandon buildings" and "athletic and outdoor sports and games" is one of the common themes in this collection. The user may choose these two nodes to find the resources.

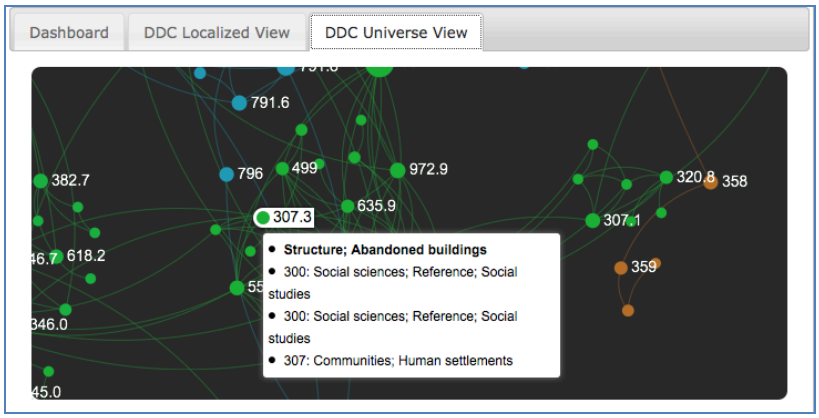


FIG. 3. DDC Universe View for the search query "Olympic history". When the user zooms in, he discovers that DDC 307.3 ("structure; Abandoned buildings") is closely related to DDC796 ("athletic and outdoor sports and games") in this metadata collection.

3. Testing and evaluating the interface

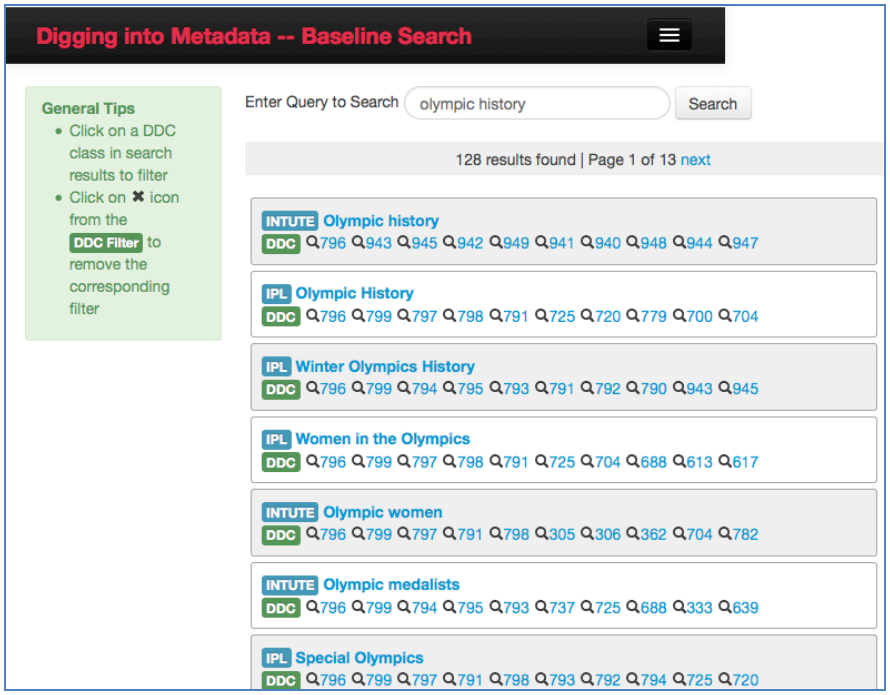
The interface was designed to serve several purposes. First, the interface will allow the user to access multiple metadata collections created with different standards or metadata structures, once these collections have been re-indexed. Second, the visual interactive functions of the interface will support a new way of exploring metadata collections through DDC distributions and their relationships to metadata records. Third, the interface will be a good testing platform to study how DDC may be applied to support searching, browsing and exploration of metadata collections. In this section, we reported our first effort in testing how users interact with the interface in an experimental setting.

To isolate the interactive functions we planned to test, we first separated the interface into three different implementations, each with a unique way of using the DDC classes for searching, browsing and exploration. In this experiment, the first interface is a simple search interface that returns search results with associated DDC numbers (the search interface; Figure 4(a)). The second adds a clickable DDC hierarchical tree to the search interface (the tree interface; Figure 4(b)). The third interface shows an interactive visual map of the search results and relevant DDC numbers (the visual interface – Figure 5).

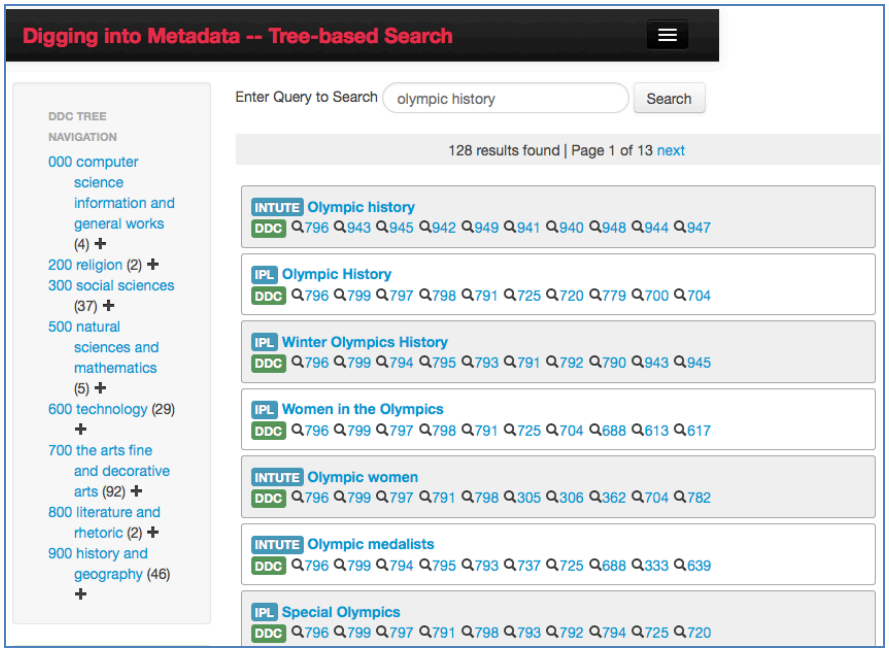
After getting the Institute Review Board (IRB) approval, we recruited 30 subjects, mostly undergraduate students (22 males and 8 females), for the experiment. They were paid $10 each for the experiment that lasted for about an hour. Each of them first completed a pre-questionnaire and watched a short video that introduced the three different styles of the interface. They were than asked to complete one search task with each of the three interfaces. The same three search tasks (see Appendix), and the three interfaces, were rotationally assigned to the subjects to avoid any order impact or bias. After completing a search task, each subject completed an interface-specific post-questionnaire that asks questions such as how easy to use the search interface, how useful the DDC specific functions, whether they have a positive experience with the interface, and how satisfied are they with the search results, etc.

As the first step of data analysis, we focused on comparing the three different interfaces and how the subjects interacted with DDC classes shown on the interfaces. Two main results are reported here.

The first concerns the general impression of the interfaces. The results indicate that the subjects understood how to use DDC to filter or narrow down search results on all three of the interfaces. As shown in Figure 6, the subjects favored the tree interface consistently across the four categories: Easy of use, Usefulness, Positive experience, and Satisfaction with the search results. The differences, however, are small and not significant. While the search interface is perceived more easy to use than the visual interface, the visual interface seems to have achieved more satisfied results and is perceived more useful than the search interface. In some of the verbal comments that the subjects made, they also confirmed that the visual interface both most interested and most confusing to them, but they managed to use it nevertheless.



(a) Experimental interface 1 -- The search interface



(b) Experimental interface 2 -- The tree interface.

FIG. 4. The experimental interface 1 and 2.

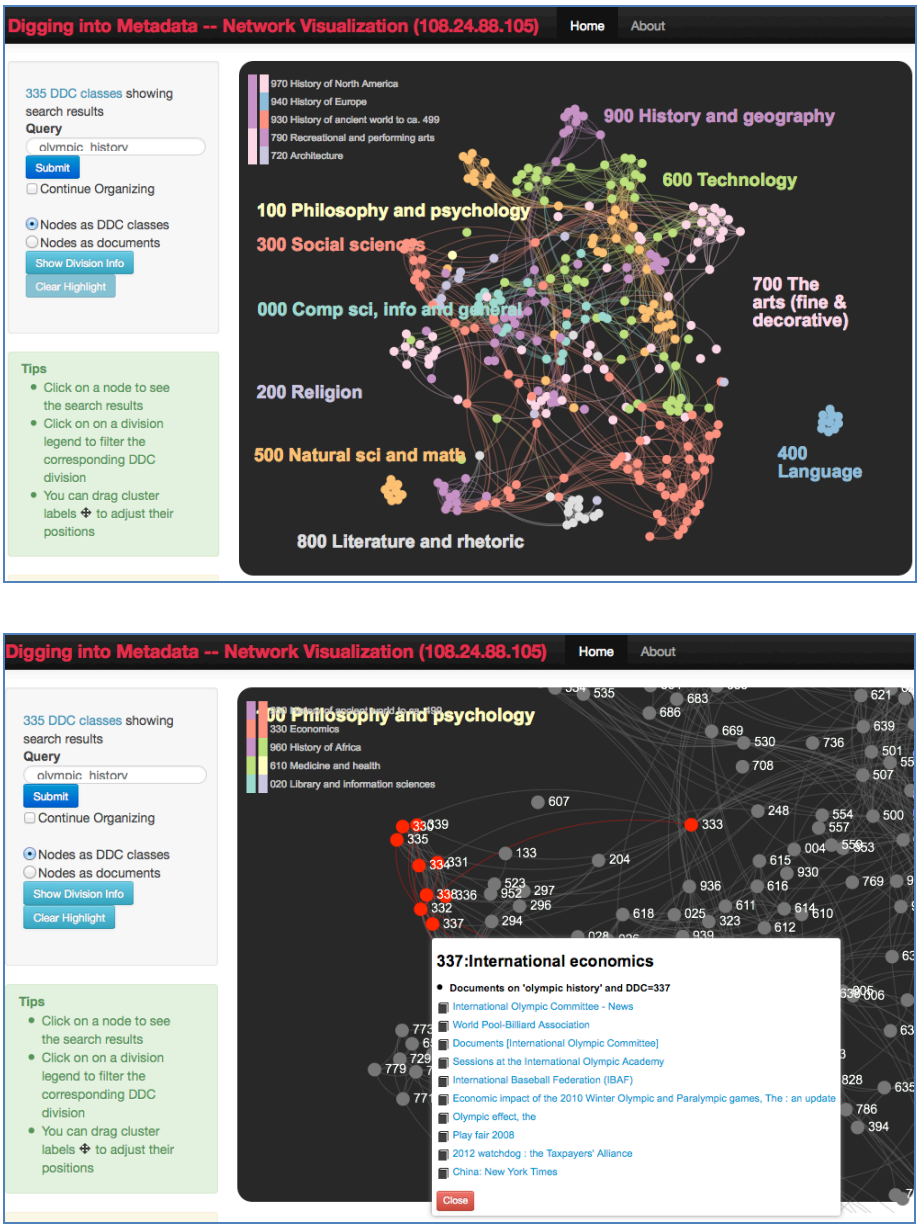


FIG 5. The experimental interface 3. The top one is the initial visual view of the interface 3 used in the experiment for the query "Olympic history". The bottom one is the zooming view when the user clicked on the DDC label "330 Economics" and then DDC node "337, International economics".

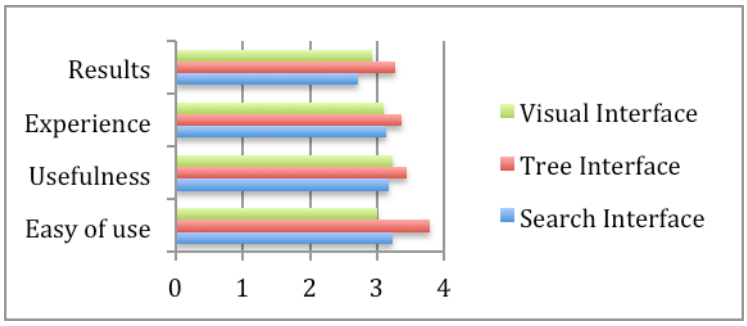


FIG. 6. Subjects' responses to post-task questionnaires. On a scale of 1 to 5, users were asked to rank each interface for its easy of use, usefulness, positive experience, and satisfaction of results. The tree interface is consistently better in all the four categories.

Table 1: DDC classes chosen for the three different interfaces

Search Tasks	Interface	DDC divisions chosen to explore (each subject could choose more than one DDC class)
Nuclear Testing	Search	3 (subjects) opted 620 (Engineering) 3 opted 621 (Applied physics) 3 opted 623 (Military & nautical engineering)
	Tree	5 opted 621 (Applied physics) 4 opted 628 (Sanitary engineering) 3 opted 623 (Military & nautical engineering)
	Visual	2 opted 572 (Biochemistry) (all other classes were selected by only one subject)
Water Cycles	Search	5 opted 551 (Geology, hydrology, meteorology) 4 opted 628 (Sanitary engineering) 3 opted 333 (Economics of land & energy)
	Tree	5 opted 550 (Earth sciences & geology) 5 opted 551 (Geology, hydrology, meteorology) 4 opted 628 (Sanitary engineering) 4 opted 333 (Economics of land & energy) 3 opted 577 (Ecology) 3 opted 621 (Applied physics)
	Visual	4 opted 628 (Sanitary engineering) 3 opted 620 (Engineering) 2 opted 631 (Specific techniques; apparatus, equipment, materials)
Hurricane Katrina	Search	5 opted 363 (Other social problems & services) 5 opted 551 (Geology, hydrology, meteorology) 4 opted 973 (United States) 3 opted 979 (Great Basin & Pacific Slope region of United States)
	Tree	2 opted 970 (History of North America) 2 opted 973 (United States) 2 opted 979 (Great Basin & Pacific Slope region of United States) 2 opted 620 (engineering)
	Visual	3 opted 970 (History of North America) 3 opted 973 (United States)

		3 opted 976 (South central United States) 3 opted 324 (The political process)

The second result relates to how the three interfaces make the subjects "see" different DDC classes for the same search. With each interface, the subjects started by entering their own queries for the search task. Based on what they saw on the interface, they could choose one or more DDC classes to add to the query, or select DDC classes to explore related resources. Table 1 shows the DDC classes chosen by the subjects for each search task (Only those classes chosen by multiple subjects were shown in the table). Clearly, most of the DDC classes are relevant to the topics. The interfaces did have significant impacts on what the subjects perceived as relevant DDC classes to the query. The classes identified by the search and tree interfaces are significantly overlapped; however, the visual interface seems to lead to unique and diverse classes. Different subjects tended to see different relevant DDC classes with the visual interface. The tree interface also helps the subjects focused on the same branch of the DDC hierarchy (such as 621, 623, 628, and 550, 551, etc.).

4. Discussions & Conclusions

As the overall goal of our project, we successfully integrated three sets of metadata from different digital libraries, created a set of tools and procedures to automatically assign one or more DDC classes to individual metadata records, and established a new indexing service to provide access to the enhanced metadata collection with richer semantic connections. We believe that a new interface is still needed in order to make the best use of the new metadata collection.

Building DDC-based interactive interfaces have been reported in a number of cases (Pollitt & Tinker, 2000; Chowdhury and Chowdhury, 2004). There are also various research projects on taxonomy-based interfaces where hierarchical structures and categories of terms or concepts are utilized for searching and browsing (Khoo, Wang & Chaudhry, 2012). Another example is the metadata interface for enhanced metadata records with additional terms generated by a Topic Modeling algorithm (Hagedorn, Chapman, and Newman, 2007). The authors in particular discussed the benefits and limitations of using automated classification techniques to enrich metadata for searching and browsing. Building on similar ideas, our design goal is to integrate multiple views of DDC hierarchical structures, query-based contextual structures, and classification-based semantic structures for the purpose of interactive searching, exploration and discovery.

We have built a prototype interface to demonstrate the feasibility of such integration (available for testing at: http://mcd.ischool.drexel.edu/ddcvisual). Testing and evaluation of such interfaces, however, remains a challenge. For an interface for metadata exploration, there are issues of metadata integration and indexing, content representation and organization, and interactions and usability, to name just a few. All these issues have significant impacts on how well the interfaces could be used by users for their intended purposes. In the experiment reported here, we attempted to isolate some of the issues and focus on how users perceive DDC classes presented on the interfaces and how they used DDC classes for searching and exploration. Initial findings indicate that the subjects understand the values of DDC and found it useful for searching and exploration. They liked to interact with the DDC classes, and use them to filter the search results. The results also show that how DDC classes are presented on the interfaces will make a major difference.

Each of the three interfaces used in the experiment has some advantages and disadvantages. The search interface can quickly lead the user to see DDC classes most relevant to the user's query. The classification codes provide additional semantic links that the user might be able to follow to find relevant items. But in general the classification codes may not increase the precision for searching, as commented by a subject, "Once I was in a detailed topic I found it hard to return to look through broad DDC codes" (subject 9). The DDC tree interface, as another subject remarked, "is easy to use but it did not help much for this topic." (subject 30). The data

showed that the visual graphs help the subjects see different DDC classes, but it is not clear that what the subjects saw was new insights that might not be seen in other interfaces. The visual interface "was most interesting but I feel like it was a bit hard to find the information that I wanted" (subject 21). Other comments indicate the visual interface was "very confusing" and with too much information, "the graph was easy to understand, but a lot of things were unrelated to the search" (subject 23).

While we are inspired by the subjects' favorable impressions of the interfaces, it is also clear that the interfaces have not yet optimized for making the best use of DDC structures in an interactive and visual setting. In the future, we plan to conduct more experiments to understand how subjects interact with the DDC classes. A new experiment will be run for more specific exploratory tasks. Additional experimental modules will be implemented to log user's interactions with the interfaces. We hope that the detailed log analysis will help us understand further how significant DDC plays in completion of the exploratory tasks and what additional benefits that automatic DDC classification will bring to the metadata collections.

Acknowledgements

The funding support from IMLS and JISC to the "*Digging into Metadata*" project was gratefully acknowledged. Thanks also go to OCLC for permitting the use of DDC for this research.

References

Bikson, T., Kalra, N., Galway, L., & Agnew, G. (2011). Steps Toward a Formative Evaluation of NSDL. RAND Technical Report. http://www.rand.org/content/dam/rand/pubs/technical_reports/2011/RAND_TR998.pdf.

Binding, C., Tudhope, D., Ahn, J-W., Khoo, M., Lin, X., Massam, D., & Jones, H. (2013). Digging Into Metadata. 12th European Networked Knowledge Organization Systems (NKOS) Workshop at the TPDL Conference, Valletta, Malta, Thursday 26th September 2013.

Chowdhury, S. and Chowdhury, G. (2004) Using DDC to create a visual knowledge map as an aid to online information retrieval. In: 8th International ISKO Conference: Knowledge organization and the Global Information Society, 2004-07-13 - 2004-07-16, London. Avail from: http://strathprints.strath.ac.uk/2624/1/strathprints002624.pdf

Digging into Data Challenge. (n.d.). Retrieved from http://diggingintodata.org/ .

Janes, J. (1998). The Internet Public Library: An Intellectual History. Library Hi Tech, 16(2), 55-68.

Joyce, A., Wickham, J., Cross, P., & Stephens, C. (2008). Intute integration. Ariadne 55. http://www.ariadne.ac.uk/issue55/joyce-et-al

Khoo, C.; Wang, Z.; & Chaudhry, A.S. (2012). Task-based navigation of a taxonomy interface to a digital repository. Information Research, 17(4), 2012. Available at: www.informationr.net/ir/17-4/paper547.html

Khoo, M., Ahn, J. W., Binding, C., Jones, H., Lin, X., Massam, D., & Tudhope, D. (forthcoming). Augmenting Dublin Core Digital Library Metadata with Dewey Decimal Classification. Paper accepted in The Journal of Documentation.

Khoo, M., & Hall, C. (2010). Merging Metadata: A Sociotechnical Study of Crosswalking and Interoperability. 10th ACM/IEEE Joint Conference on Digital Libraries, Brisbane, Australia, June 21-25, 2010, pp. 361-36.

Khoo, M., Tudhope, D., Binding, C., Abels, E., Lin, X., & Massam, D. (2012). Towards Digital Repository Interoperability: The Document Indexing and Semantic Tagging Interface for Libraries (DISTIL). Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2012, Vol.7489, pp.439-444.

Hagedorn, K., Chapman, S.; & Newman, D. (2007). Enhancing search and browse using automated clustering of subject metadata. D-Lib Magazine, 13(7/8). Available at: http://www.dlib.org/dlib/july07/hagedorn/07hagedorn.html

Pollitt, A. S.; Tinker, A. J. (2000), "Enhanced view-based searching through the decomposition of Dewey Decimal Classification Codes", Proceedings of the Sixth international conference of the International Society for Knowledge Organization, 10-13 July 2000, Toronto, Canada. Eds. C. Beghtol, L. Howarth and N. J. Williamson. Würzburg: Ergon Verlag, 2000. (Advances in Knowledge Organization 7). 288-294.

Slavic, A. (2006). Interface to classification: some objectives and options. (UDC paper) http://arizona.openrepository.com/arizona/handle/10150/105459 .

Williams, C. (2006). Intute: The New Best of the Web. Ariadne 48. http://www.ariadne.ac.uk/issue48/williams.

Zeng, M. L. (2008). Knowledge organization systems. Knowledge Organization, Vol. 35, No. 2-3: 160-182.

Zia, L. (2005). The NSF National Science, Technology, Engineering, and Mathematics Education Digital Library (NSDL) Program. D-Lib Magazine 11(3). http://www.dlib.org/dlib/march05/zia/03zia.html.

Appendix: The search tasks used in the experiment:

1. Please find best Web resources that a high school student should read when working on a paper for nuclear testing sites and its impact to the environments. What DDC classes would be useful for this topic?
2. You have been asked to prepare a class project on the water cycle, and to identify some of the current environmental, social, political, and other issues associated with different stages of the water cycle. Please identify relevant web resources and DDC classes.
3. Hurricane Katrina was one of the largest storms to make landfall in the United States, and the costliest in terms of damage to New Orleans and other places. Your project is to collect information for writing a timeline for Hurricane Katrina. The timeline should not just focus on the storm itself, but also look at such issues as the history of New Orleans, the social and political issues that were raised after the storm, the reconstruction, how the storm has been remembered, how the storm has affected peoples' lives today, and so on.