Augmenting Dublin Core Digital Library Metadata with Dewey Decimal Classification

Journal:	*Journal of Documentation*
Manuscript ID:	JD-07-2014-0103.R1
Manuscript Type:	Article
Keywords:	Cataloguing, Classification, Classification schemes, Digital libraries, Online catalogues, Online databases

Introduction

This paper addresses a well-known yet difficult question for digital libraries: How may a user search across multiple unrelated digital libraries with a single query? Depending on their information needs, a user may find it preferable to query multiple digital libraries at the same time, and have the results from each library gathered and combined into a single list. However, while individual digital libraries can provide access to a wealth of information from multiple domains and disciplines, there is often little integration between different libraries. Digital libraries often exist as stand-alone projects and institutions, with individual resources, catalogs, metadata, and discovery tools, and there is often little support or opportunity for querying multiple digital libraries from one location.

The question is not a new one, and a number of approaches have been proposed (Greenberg, Spurgin, & Crystal, 2006). These approaches can roughly be divided into two categories: (1) dedicated approaches that build interoperable metadata from the ground up, and (2) post-hoc approaches that augment metadata after its original creation. (Figure 1 provides an overview of this problem space, and the methodological choices made by this project.)

[INSERT FIGURE 1 ABOUT HERE]

Dedicated approaches aim to build metadata interoperability into digital libraries at the time of development, with project partners describing their resources by implementing a standard metadata format in similar ways (Woodley, 2008). One issue here concerns the choice of a standard. While widely adopted metadata standards have yet to emerge for digital libraries, some standards do appear to be 'more standard than others,' one example being Dublin Core metadata. Together with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), these provide a technical platform for federated discovery. One advantage of Dublin Core is that it allows for the relatively low-barrier construction of repositories; however, at the same time, there is also "no strict standard for consistent subject indexing" (Waltinger, Mehler, Lösch, & Hortsmann, 2011, p. 29). This may lead to heterogeneous implementation at the element level, with the result that "when it comes time to build services on [an] aggregated collection, the system architect finds that the lack of a uniform semantic basis is a major impediment to functionality" (Krown & Halbert, 2005, p. 46). "Normalizing the heterogeneous subject indexing of OAI records from different repositories" is therefore "central to the debate of an enhanced search experience within the digital library domain" (Waltinger et al., 2011, p. 30).

In contrast to dedicated approaches, post-hoc approaches to metadata interoperability seek to establish interoperability *after* metadata development has occurred. This may be necessary even if a standard such as Dublin Core has previously been adopted. For instance, if there is no prior history of collaboration between potential digital library partners, then element level differences in formatting, choice (or lack of) controlled vocabularies, etc., may be present, and the available metadata may not be fully interoperable. Post-hoc solutions may involve manual interventions, such as re-cataloging each resource from each library, but these are often not practical; manual classification is resource-intensive and time-consuming, and the number of collections and repositories that would benefit from additional metadata is growing more rapidly than the trained experts available to classify them (Greenberg, 2004; Greenberg, Spurgin, & Crystal, 2006; Wilson, 2007). Post-hoc solutions involving harvesting and/or crosswalking each library's existing metadata to a standard format can also require significant manual work to design mappings, normalize metadata schemas and elements across multiple collections, and evaluate crosswalk outcomes (Khoo & Hall, 2013).

An alternate group of post-hoc approaches involves automated metadata generation and augmentation, and the creation of one or more new elements to add to the original metadata records. In this group of approaches, it is advantageous to adopt an existing classification scheme as the target vocabulary, as such schemes represent significant previous intellectual effort by large numbers of people (Yi, 2007). One such existing scheme is the Dewey Decimal Classification (OCLC, n.d.), which is a widely established and implemented knowledge organization system (Sweeney, 1983), and thus is the one implemented in the research described below. The specific approach adopted involves generating new DDC classes for existing metadata records (in this case Dublin Core records from three digital libraries), adding these classes back to the individual records, and then using the augmented DDC metadata to support federated search and browse across these three different collections.

In general, this is not an easy problem to solve. In 1997, for example, OCLC reported on experiments in the Scorpion project to automatically classify DDC's own concept definitions with DDC using SMART (Thomson, Shafer, & Vizine-Goetz, 1997). One key finding here was that the meaning of a concept (class) required consideration of its hierarchy in addition to the text of its captions; and thus all captions of parents and immediate children were added to the text representing a given concept. The matching was based only on captions and only a single pass matching algorithm was employed rather than the two stage process also incorporating relative index terms described in this paper. In summary, therefore, creating good quality interoperable metadata that can be used by patrons to search across multiple digital libraries remains an ongoing challenge. Integrating metadata from multiple sources is a difficult task that, even when accomplished, does not necessarily fully provide the rich functionality expected from federated repositories.

The rest of this paper focuses on a description and evaluation of a post-hoc metadata augmentation strategy based on the automated generation of Dewey Decimal Class numbers from existing Dublin Core metadata.. Section 2 describes a range of existing approaches to metadata augmentation, focusing particularly on the approach adopted in this paper, automated document classification. Section 3 describes the project workflow, including the metadata harvest and processing, and the evaluation of that processing. Section 4 provides a discussion of the evaluation results, while Section 5 gives conclusions and possible directions for future work.

Approaches to metadata augmentation

Post-hoc methods for metadata augmentation generally rely on machine analyses of the content of a document (an academic paper, a web page, etc.), and/or the metadata (including keywords, title metadata, abstract metadata, etc.) that describes that document, in order to create additional subject metadata (such as DDC classes). Approaches to automated subject classification vary by analytical methods, size and type of corpus analyzed, target controlled vocabularies (domain-specific vocabularies, DDC, etc.), and other dimensions. This paper follows the approach of Golub (2006b) in characterizing post hoc approaches to metadata augmentation in terms of:

- text categorization/supervised machine learning
- document clustering/unsupervised machine learning
- document classification

This research follows a document classification approach. While it is therefore not a machine learning approach, to situate our approach and methodology, we first present a brief overview of approaches to automated metadata generation.

Machine Learning Approaches

Text categorization and document clustering approaches are built on supervised and unsupervised machine learning approaches respectively. They involve either (a) training an engine to recognize statistically examples of particular categories, by manual categorization of an initial set of documents, with the extracted characteristics then being used to categorize new documents; or (b) automatically generating categories *ab initio* through document comparison techniques, and subsequently assigning unclassified documents to these categories.

Waltinger et al. (2011) classified scientific documents to the first three levels of DDC by analyzing OAI metadata obtained from the Bielefeld Academic Search Engine (BASE: http://www.base-search.net/about/en/ BASE). They found an 'asymmetric distribution of documents across the hierarchical structure of the DDC taxonomy and issues of data sparseness” (p. 29) leading to a lack of interoperability that is a “severe problem” (p. 30). In related work, Lösch et al. (2011) describe the building of a DDC-annotated bilingual corpus to support experiments in text categorization. After manually constructing cross-concordances, they automatically mapped between 52,905 English and 37,228 German full text articles drawn from BASE, and DDC. They again note the uneven distribution of classified documents amongst DDC classes. Wang (2009) argues that DDC’s deep and detailed hierarchies can lead to data sparseness and thus skewed distribution in supervised machine learning approaches and proposes a method for creating a balanced DDC structure in machine learning classification.

Examples of unsupervised learning approaches include Krowne and Halbert (2005), who used a text-clustering approach to analyze the title, description and subject fields from the ‘americansouth.org’ digital library, and Newman et al. (2007), and Hagedorn, Chapman, and Newman (2007), who used a statistical topic model to enrich subject metadata in 7.5 million records in the OAIster Digital Library. Recently, Tuarob, Pouchard, and Giles (2013) described a method for generating tags from a domain-specific controlled vocabulary to augment metadata for resources from four different environmental data repositories associated with the DataONE program. They compared term frequency-inverse document frequency (TF- IDF) with a topic modeling approach (based on Latent Dirichlet Allocation) to metadata generation. The additional metadata ‘tags’ were matched against an existing controlled vocabulary of DataONE subject terms. The repositories sometimes contained sparse metadata and performance was influenced by the richness of the metadata and the frequency of tag utilisation.

Document classification approaches

In contrast, document classification approaches proceed by matching text in the documents to be classified against controlled vocabulary terms (Golub, 2006a). The preprocessing involved in document classification is similar in some ways to that involved in text categorization and document clustering approaches, e.g. initial text extraction, cleaning, stemming, weighting, and other types of preparation.

However, no learning, supervised or unsupervised, is subsequently involved. Instead, relevant terms are extracted from the text of the document and/or document metadata, and compared with terms in a controlled vocabulary. The approach described in this paper focuses on matching between Dublin Core metadata and DDC 23. Golub's study involved automated classification of engineering-related web pages against the Engineering Information thesaurus and classification scheme [Ei]. In other studies, the 'Enhanced Tagging for Discovery' project investigated the use of DDC suggestions for social tagging in an educational context using the Intute digital library, comparing a baseline social tagging system with an augmented version employing social tagging in combination with suggestions from DDC (Golub et al., 2014). Wartena and Sommer (2012) employed an automated text classification approach, based on using the German Subject Heading Authority File (mapped to the DDC) to classify the content of 3,826 documents and related abstracts from 7 different German universities, and they conclude that an automated document classification approach can compare favorably with the output of a supervised learning approach to the same corpus.

A general theme that emerges across these approaches is that of a trade-off between machine learning approaches and document classification. While machine learning can be applied to new datasets once trained, it can require large corpora and manual and/or automated training. With document classification approaches, once the pipeline itself has been identified and implemented, it does not require training and can be employed with knowledge organization systems with uneven hierarchies or sparse distribution across a given collection.

This paper describes and evaluates a document classification approach to metadata augmentation. The Digging Project (Digging Into Metadata, 2014) has been developing ways to provide federated discovery across three unrelated digital libraries - the Internet Public Library (IPL; http://www.ipl.org/); Intute (http://www.intute.ac.uk); and the National Science Digital Library (NSDL; http://nsdl.org/) – by adding to each Dublin Core metadata record in each library one or more DDC classes, based on the content of that particular metadata record. A document classification approach is used that extracts and weights key terms and noun phrases from each metadata record in each digital library. Note that the unit-of-analysis employed in this study is that of Dublin Core metadata records that describe an online resource in a digital library; that is, it is not the online resource itself that is analyzed, but the content of the Dublin Core record describing that resource. The broad goals of the project are as follows:

- To understand the effectiveness of a document classification approach in automated subject classification of large numbers of Web resource metadata records from heterogeneous digital libraries.
- To understand the general practical issues that can affect the construction of document classification pipelines in this context.

Project Workflow

The project workflow is as follows:

1) metadata records are harvested from each digital library;
2) for each metadata record, the content of the title, description and subject (including topic or keyword) fields is extracted, cleaned, and stored;
3) a text analysis of the extracted metadata is carried out that identifies and weights key terms and noun phrases;
4) the weighted key terms and noun phrases are used to generate one or more DDC classes for that record;
5) the DDC classes are added back to the original metadata record, to support the building of visualization tools for federated discovery across the collections.

[INSERT FIGURE 2 ABOUT HERE]

Figure 2 shows the three main processes (following a metaphor of distilling output metadata via a pipeline of refinement stages).

- MASH (Metadata Aggregation, Storage, and Handling)
- DISTIL processing (Document Indexing & Semantic Tagging Interface for Libraries)
- DRAMs (Dynamic Representations of Annotated Metadata).

This paper describes the first four stages of the pipeline, involving harvesting, processing, and DDC metadata generation.

Metadata harvesting

Both IPL and Intute provided database dumps of their catalogs. For NSDL, OAI-PMH was used to harvest the metadata. The harvest was affected by a number of legacy issues. While each digital library had adopted Dublin Core as a standard, there were differences in the ways in which it had been implemented to address the needs of different audiences. Metadata could also be stored in a variety of databases. In some instances, the metadata displayed in web views of the catalog differed from the metadata that could be found in various databases. These issues required further work in order to locate, understand, and then (if possible) address, including ongoing communication with each of the libraries in the project. These factors combined to make the harvest a significant manual exercise. After the issues were resolved, a total of 263,550 records were harvested: 40,973 from the IPL, 98,507 from the NSDL, and 124,070 from Intute.

Further post-harvest issues arose in the form of duplicate records within and between the digital libraries. There was no way to calculate precisely the extent of such duplication. Titles were good sources that represented the contents of a resource, although different resources could have the same title, especially when titles were shorter and consisted of common terms. URLs had less chance to be duplicate across different resources but care needed to taken with incorrect or insufficient information within URL strings (e.g. with typos, or when provided only with root URLs). Overall, there were 25,318 duplicate titles (9.6%), and 19,475 duplicate URLs (7.4%). Exact duplicates were relatively easy to identify and remove. However, non-identical duplicate records, such as different descriptions of the same resource, were more difficult to judge. This is not in itself a disadvantage. Given that metadata records are human-generated descriptions of documents that often take particular audiences into consideration, it suggested that different catalogers had decided that a particular resource could satisfactorily be described with at least two different sets of subject terms for different audiences, emphasizing different aspects of the resource. For instance, the official web site of the Chateau de Versailles has been cataloged by the IPL, by the Librarians' Internet Index (which merged with the IPL), and by Intute, in various ways. A comparison of the subject and description fields is given in Table 1. There is a wide variety in the descriptions supplied by each digital library, which is in turn reflected in the different DDC classes suggested by DISTIL for the different records.

[INSERT TABLE 1 ABOUT HERE]

In another example, five NSDL partners cataloged the web site for the National Science Teachers' Association (NSTA: http://www.nsta.org), in different audience-appropriate ways. One partner (ComPADRE) used five subject terms (*professional association*, *teaching tools*, *best practices*, *general* *physics*, *physics*), while another (the DLESE Community Collection) included twenty-three subject terms (*educational theory and practice*, *environmental science*, *policy issues*, *space science*, *science*, *earth science*, *physical sciences*, *chemistry*, *biology*, *education (general)*, *physics*, *astronomy*, *space sciences*, *education*, *ecology*, *forestry and agriculture*, *geoscience*, *social sciences*, *history/policy/law*, *space science*, *chemistry*, *physics*, *life science*, and *technology*).

These variations support Waltinger et al. (2011) regarding the '*lack of a uniform semantic basis*' in Dublin Core metadata. There is no reason to doubt that this may be a common occurrence amongst digital libraries with no prior record of collaboration. It suggests that the original catalogers for these libraries were often interested to provide audience specific points of entry to the resource.

Metadata cleaning and storage

After harvesting, the metadata from the title, description, and various subject and topic fields, was extracted from each catalog record. XML markup was removed and the cleaned metadata was stored in the MASH database in tuples that described the originating library, the original (harvested) record ID number, the harvested field, the type (a normalized field, for instance mapping topic and other similar fields to subject), and a value (in this case the text of the particular metadata field). The final MASH database contained approximately 4.89 million rows, each one representing a relevant metadata field from a record obtained through the harvest.

Metadata analysis

A pilot manual pipeline was first constructed. A sample of fifty full metadata records was obtained (17 from both Intute and IPL, and 16 from the NSDL). Metadata from the title, description, and subject fields of each record was analyzed by term frequency, and noun phrase frequency. Noun phrases were identified through manual queries of NaCTeM's TERMINE (http://www.nactem.ac.uk/software/termine) term extraction system (Frantzi, Ananiadou, & Mima, 2000).

For each record, the individual terms and the terms in the noun phrases were stemmed, and stem frequencies per record were calculated. Stems were selected for further processing if they occurred over a specific threshold defined as follows, where TF = Term Frequency:

$$Threshold_{term} = mean(TF_{term}) + standarddeviation(TF_{term})$$

Following the manual pilot tests, the final version of the pipeline automatically extracts ranked/weighted key terms, and (for the evaluation) ranked/weighted noun phrases and applies preprocessing, including tokenization, stop-word removal, and Porter stemming. A total of 3,797,905 word stems were identified across the harvested records.

For most individual records, stems were extracted from the title and description fields, that were not extracted from the subject fields. That is to say, catalogers had used words in the title and description fields which they did not use in the subject fields. An average of 2.16 extra terms per record (an aggregate of 569,913 stems across the harvest) were located this way.

[INSERT TABLE 2 ABOUT HERE]

The stems were then annotated either with TF scores (or sum of TF scores for Phrases), as weights to be used by DISTIL metadata generation. The results were passed to the DRAMs database. The evaluation of the subsequent metadata generation compared the contribution of the various (stemmed) metadata elements processed by MASH to assist the analysis of the most appropriate strategy. Thus the original unweighted *Subject* metadata acts as a baseline for judging the contribution of the weighted *Subject* metadata, weighted *Terms* extracted by the pre-processing from Subjects, Title and Description, Termine derived Noun *Phrases* and various combinations of these elements. For example, would the additional metadata extracted from Title and description assist or hinder the steps in the DISTIL pipeline?

DDC metadata generation

DISTIL is a bespoke application for performing bulk processing of repository metadata records, producing a list of best match DDC classes to supplement the repository records. The generalised problem as illustrated in Table 3 is to determine an overall degree of match between two sets of typed and weighted metadata fields representing repository record subject fields and DDC class headings, including DDC Relative Index headings (OCLC, n.d.). Multiple fields of the same type may be present, and there are other possible field types not listed in this example. DISTIL attempts to find the main subject(s) for a repository item; DDC built (composite) numbers are outside current scope.

[INSERT TABLE 3 ABOUT HERE]

Obtaining reliable matches involves more than just textual comparison due to the nature of DDC class headings. Unlike a thesaurus, the same heading may appear multiple times at different positions in the hierarchy, the context of a particular heading being determined by hierarchical ancestry. E.g. "*scientific principles*" appears as a heading for a number of different DDC classes – e.g. under 200 (*Religion*), 401 (*Philosophy and Theory - languages*), 570 (*Life sciences; Biology*), 620 (*Engineering*), 630 (*Agriculture*) etc. It is therefore necessary to take account of the hierarchical context of candidate matches to determine the likelihood of relevance.

Broadly speaking, DISTIL follows a document classification approach with two main phases in a configurable pipeline. The first phase attempts to match a weighted combination of the metadata records against the entry vocabulary of the 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.

Input Data

A copy of DDC Version 23 was obtained from OCLC for use within the project. As this was provided in MARCXML format, a custom import routine was developed to read and parse the data, which was then used to populate an internal Apache Lucene index with the DDC class identifiers and associated labels.

The source and format of repository metadata to be used as input to the DISTIL process evolved throughout the course of the project. An initial implementation of DISTIL obtained repository metadata via online OAI-PMH interfaces. Following consolidation of the metadata records from the three separate repositories to a single MySQL database (MASH), the DISTIL application was revised to utilize a local copy of this database. The MASH database was subsequently used to populate an online Apache Solr repository (DRAMS), and at that point the DISTIL application was revised again to process metadata obtained via the DRAMS Solr API.

Data Processing

The DISTIL process uses repository metadata to search for suitable indexing, instead of the more usual case of using indexing to search for suitable repository records. The subject metadata of each repository record is used to build a Boolean query for retrieving a set of initial candidate DDC class matches from the internal Lucene index. A stop word list and Porter stemming provide some flexibility in matching. Queries can also use relative weightings to 'boost' scores for particular subjects. Phrases are treated as a group of words where all (stopped and stemmed) words must be present, though in any order. As an example, for the following set of weighted subjects:

```
Joint Diseases [3.000]
Medical Research [9.000]
Rheumatology [1.000]
Musculoskeletal Diseases [4.000]
Arthritis [8.000]
Charities [3.000]
Research Support [9.000]
Great Britain [2.000]
```

The following Boolean query is generated by DISTIL for use with Lucene. Note the application of word stemming and relative weightings:

```
((+label:joint +label:diseas)^3.0)
((+label:medic +label:research)^9.0)
label:rheumatolog
((+label:musculoskelet +label:diseas)^4.0)
label:arthriti^8.0
label:chariti^3.0
((+label:research +label:support)^9.0)
((+label:great +label:britain)^2.0)
```

This query retrieves an initial set of candidate DDC classes with associated scores, which is then refined via a series of successive filtering and aggregation stages to produce a shorter ranked list of the overall best matching classes. The process is repeated for each repository record, and then the consolidated results are exported to supplement the original repository records with their best matching DDC classes.

Pipeline

The filtering and aggregation stage of the process uses a pipeline architecture (Figure 3) comprising a series of sequential actions that may be enabled/disabled and reordered, allowing for experimentation with various configurations. There are general actions that would be applicable to any tabular result set, and more specialised actions relating specifically to the DDC.

[INSERT FIGURE 3 ABOUT HERE]

The pipeline actions are as follows:

- *Replace Values*: Replaces values in a specified column
- *Filter Rows*: Only allows rows matching the filter criteria e.g. “score > 0.5”
- *Sort Rows*: Sorts the results according to a column name and sort direction criteria e.g. “score DESC”
- *Limit Rows*: Returns a maximum number of results for each record; discard the rest
- *Normalize Values*: Applies normalisation to values in a specified column to obtain values in the range [0..1] using the following formula:

$$x_{new} = \frac{(x - x_{min})}{(x_{max} - x_{min})}$$

- *DDC Remove Outliers*: Removes DDC classes having a syntactic match but no other hierarchically related ancestors or descendants present in the results - this is an attempt to eliminate isolated single matches where the query terms had nothing else in common with the surrounding hierarchy, possibly indicating a homonym or a less relevant subject area.
- *DDC Remove Spans*: Removes any span classes from the results. These are organizational classes representing a fixed range of DDC numbers – e.g. “996.902-996.904”.
- *DDC Rule of Three*: Implements an aspect of the practice of manual indexers, the ‘Rule of Three,’ which states that any 3 or more matching classes with a common parent are replaced with that parent. The broader subject might not necessarily be present in the results at all, and so it is added and replaces the child classes. The sum scores of the replaced children are then added to the parent. (OCLC, n.d., page 8, section 5.7D: “Class a work on three or more subjects that are all subdivisions of a broader subject in the first higher number that includes them all.”)
- *DDC Summary Level Minimum*: This action mirrors another manual indexing procedure. The top 2 levels of the DDC are for hierarchical structure only – indexing should use as a minimum the 3rd level (3 digits). Any suggested classes having a notation of less than 3 digits are therefore removed from the results. (OCLC, n.d., page 37, section 13.3: “The classifier should never reduce the notation to less than the most specific three-digit number”.)
- *DDC Add Sum Descendant Score*: Performs upward score aggregation in which a class can inherit the aggregated sum of the scores of any hierarchical descendants present, effectively promoting it as a stronger match in the overall result list.

- *DDC Use Abridged ID*: Performs upward score aggregation from 'close' classification to 'broad' classification. For example the 'close' classification for a resource about *French cooking* would be 641.5944 (641.59 Cooking by place + 44 France), whereas the 'broad' class would be 641.5 (Cooking). The resource is "placed in a broad class by use of notation that has been logically abridged" (OCLC, n.d.). The broad class (a.k.a. abridged number) is not necessarily the direct parent class. The scores are aggregated to the associated broad class then the contributing results are removed.
- *DDC Use Summary ID*: Performs upward score aggregation to a consistent 3 digit DDC summary level. The process aggregates result scores up to the associated summary level ancestor then removes contributing results (see Figure 4).
- *DDC Add Dominant Summary Scores*: Boosts all scores to promote results originating from particularly strong subject areas. Scores are boosted by the overall sum of scores for each of the first 3 hierarchical levels. So in the example of Figure 4:
  - sum(level 1) is the sum of all scores for descendants of class "5"
  - sum(level 2) is the sum of all scores for descendants of class "55"
  - sum(level 3) is the sum of all scores for descendants of class "551"

  The new scores are then calculated using the following formula:

  $$new\ score = \frac{score\ +\ sum(level\ 1)\ +\ sum(level\ 2)\ +\ sum(level\ 3)}{overall\ sum\ of\ scores}$$

  Using this score manipulation technique the process effectively develops an overall 'opinion' on the most appropriate subject area(s) to use for classification and promotes results originating from those areas.

[INSERT FIGURE 4 ABOUT HERE]

Output Data

The DISTIL process outputs 2 result files. Firstly a comma delimited text file containing a list of repository resource identifiers and best matching candidate DDC class identifiers. This file can be used to supplement existing repository records. Secondly a text file including a record of the metadata used, the Lucene query generated and an explanation of the process applied to each resource. This information can be useful in subsequently determining the reasons behind any particular match.

Initial Testing

During initial testing we observed an encouraging initial overlap between the DISTIL output and manual indexing of a small subset of records. However a variation in the quality and quantity of the subject metadata was seen to be affecting the quality of some results – e.g. level of specificity, misleading or lacking metadata. Key subject elements were sometimes missing, or sometimes the DDC itself lacked sufficiently detailed subject coverage in some areas. In an effort to improve this situation a pre-processing phase (see Metadata Analysis above) supplemented the existing subject metadata with weighted subject keyword and phrase suggestions derived from titles and descriptions. The DISTIL process was subsequently run against a subset of 100,000 repository records.

Matching free text metadata against controlled terminology presented a number of issues:

- Subject phrases could sometimes be formatted in terms of a nested structure, using a local convention defined by punctuation, e.g. "*Arts & Humanities--History--History by Era--18th Century History*", "*History/Policy/Law*", "*Anatomy / physiology / morphology*".
- Variations in subject specificity were observed. Some general repository subject terms, for example, were not necessarily very useful e.g. "*General Resources*", "*People*", "*Places*", "*Projects*", "*Images*", "*Science*", "*Technology*".
- Repository subject terms occasionally held embedded encoded characters, stemming from their use within a web context e.g. "*Food &#38 Beverage*", "*Home & Housing*". This issue was resolved by adding these character-encoding sequences to the stop word list.
- Some subject metadata terms had little likelihood of matching DDC labels e.g.
  - Codes: "*artifact1200; artifact1137; artifact804;*", "*pi3731*"
  - Phrases and titles: "*Keystone Color Me Healthy*", "*Connecticut Butterfly Atlas Project*"
  - Spelling errors: "*muscoskeletal*", "*policytaxation*", "*intertial navigation*", "*filmsUKmarketing*"
- Misleading subject combinations, e.g."*SPACE*", "*training*", "*wireless networks*", "*mobile technology*" (the record actually referred to an Arts organisation called "*SPACE*")
- Variations in national language conventions. One of the repositories used in the project (INTUTE) originated in the UK, whilst the other two originated in the US. Although both nations use the English language, there are spelling differences between US and UK English for certain words. The DDC itself uses predominantly US English for class headings: e.g. "*color*", "*paleontology*", "*humor*", "*aluminum*", "*anemia*", resulting in no match on UK spellings of these words where they occurred in subject fields. The issue was resolved by adding a list of common

US/UK equivalents to the DDC23 index. So for example searching for the subject phrase "*movie theatre*" adds the following stemmed, nested Boolean query to the main Lucene query:

(+(label:movi label:film) +(label:theatr label:theater))

Initial observations of the relative accuracy of successive experimental runs of the DISTIL process were informal and subjective. Improving the process requires the ability to quantify the positive or negative effects of any changes. A more formal objective evaluation of DISTIL results was therefore required in order to better assess the quality of the DDC indexing being produced.

Evaluation: Comparison with intellectual DDC classification

In order to evaluate the DISTIL output, a trained librarian, affiliated with one of the project teams, intellectually indexed a sample of 50 records from the harvested metadata. The librarian selected 50 sample records, taken equally from across the three repositories (17 records from Intute, 17 records from IPL, and 16 records from NSDL), and covering numerous subject areas (one NSDL record was subsequently dropped, as it disappeared from the live repository during the project.). The librarian made a note of the title and description from the holding repository for each of the sample records, viewing 'more details,' where possible to capture any existing keywords (both controlled and uncontrolled). The librarian also looked up any existing subject classifications for any corresponding DDC number (using DDC23). Finally, the repository 'View Page Source' XHTML details were checked, to make sure that all the relevant metadata had been captured, in order to inform the intellectual indexing. The process was quite time consuming.

In the first phase of the intellectual indexing, the librarian assigned multiple DDC classes to each resource (an average of 4.5 classes per record). (This was motivated by current practice in assigning "multiple classifications to allow for the widest number of hits to be produced if people chose to browse by subject area."). This was modified in a subsequent second classification phase by the same librarian, where the task was to assign a single DDC classification of major subject when considered appropriate and multiple classes otherwise. Thus out of the 49 records, 2 classes were assigned in 19 cases, and 3 classes in 3 cases, in order to represent adequately the website represented by the record. The second phase classification was used as the basis for the evaluation of the automated DISTIL classification. Where the librarian assigned more than one class , a match by DISTIL against *any* of the (second phase) classifications was taken. Intellectual classification was given at the DDC level considered most appropriate and considered a match for DISTIL output identical or broader in the DDC hierarchy.

The evaluation exercise compared automated results from DISTIL with the second phase manual classification for the 49 records described above. DISTIL was configured to perform the following pipeline actions (see above for fuller descriptions of these actions):

1. DDC Summary Level Minimum
2. DDC Remove Spans
3. DDC Use Summary ID
4. DDC Add Dominant Summary Scores
5. Sort Rows (by descending score)
6. Limit Rows (maximum 10 results per record)

This process was run eight times, using as input various different combinations of pre-processed metadata fields (see 'Metadata analysis,' above). This produced a ranked list of DDC class suggestions for each repository resource. Only the top 10 ranked suggestions were considered (sometimes less than 10 suggestions were returned). The previously produced intellectual DDC classes were compared to those generated automatically by the DISTIL processing.

A wide variety of performance measures were available in principle. Our research question concerned automated classification rather than immediate retrieval from a set of queries. Since we had a Gold Standard available in the 49 intellectually classified records, the performance measure was per record. The data was too sparse to report on performance of DDC classes themselves. While it would be possible to treat the problem as a binary classification problem, DISTIL returns a ranked list of possible DDC classes and we wished to characterise the performance of the set of highest ranking results (not only the top result). This was partly due to the indexing consistency issues discussed below; there might be more than one reasonable answer. Thus we employed the widely used Mean Reciprocal Rank (MRR) as the main measure, which is bounded (0 – 1) and averages well (Voorhees 1999). An automated result that matches the Gold Standard with the first choice scores 1 but a lower ranking result that matches will gain some lesser degree of credit. MRR was also used by Wartena and Sommer (2012), the most closely related previous study, making a direct comparison possible. As they also observe, a motivating use case for this work is a recommendation system to assist human indexers, where a ranked list of results is helpful. The current state of play is likely to require a final human inspection element to validate correctness of the automated classification rather than a completely automated operational system. To complement MRR of the top 10 ranked results, we included a binary measure of whether the Gold Standard DDC class was found in the top 5 automated results. The two measures are defined as:

- *Mean Reciprocal Rank (MRR)* - The reciprocal rank (RR) is calculated as 1 divided by the ranked position of the first result relevant to the manual classification(s), in descending score order. The Mean Reciprocal Rank (MRR) is then the overall average of the RR scores across the entire result set.
- *Recall at 5 (Rec@5)* - measures whether or not the manual DDC classification appears within the first 5 DISTIL results in descending score order.

Table 4 shows an example, for manual DDC classification of 330 – Economics.

[INSERT TABLE 4 ABOUT HERE]

Results

The MRR and Rec@5 scores were calculated for each resource, overall averages of these scores at each of the first 3 hierarchical levels of the DDC were then calculated for the sample set. Table 5 shows (for both measures) that compared to the baseline original Subject metadata, TF pre-processing of Subjects or Terms (from Subjects, Title, Description) improved performance but Phrases (alone) did not. Any combination improved performance but the best results (highlighted) were obtained using a combination of Subjects, MASH Terms and TERMINE Phrases. Thus results clearly show a benefit (for this DISTIL pipeline configuration) to applying TF to Title and Description (with just a slight benefit from including Phrases). This was striking for some individual records with sparse original Subject metadata. As expected, performance declines with increased specificity of DDC level, with MRR approximately 0.7 for Level 2 and 0.5 for level 3.

[INSERT TABLE 5 ABOUT HERE]

[INSERT TABLE 6 ABOUT HERE]

[INSERT TABLE 7 ABOUT HERE]

Splitting the results by originating repository for this field combination only (Table 6), we see a variation in performance across the different libraries. The lower performance for NSDL is possibly due in part to differences in the subject metadata; several NSDL records in the sample had just a few, very general subject metadata terms, such as ‘Education’ or ‘Technology’, which poses more difficulties for DISTIL’s matching of the DDC entry vocabulary than metadata elements comprising several more specific terms. The effect will have been mitigated by the pre-processing of Title and Description fields but may have contributed to the difference in results observed.

Finally, a further experimental run of the DISTIL process was undertaken, this time using a slightly different pipeline configuration, to aggregate scores up to abridged DDC numbers:

1. DDC Summary Level Minimum
2. DDC Remove Spans
3. DDC Remove Outliers
4. DDC Use Abridged ID
5. DDC Add Dominant Summary Scores
6. Sort Rows (by descending score)
7. Limit Rows (maximum 10 results per record)

Table 7 shows a fairly linear degradation of MRR and Rec@5 scores through the 5 hierarchical DDC levels for the abridged. Overall scores at levels 1 to 3 are lower than the previous pipeline but results from abridged levels are made possible. Some abridged results are accurate but offset by less accurate results generally. Introducing ‘Rule of 3’ aggregation to these results may improve this although that might then tend to aggregate to DDC level 3.

Comparison with related work

One of the closest recent studies is Wartena and Sommer (2012), who also report results on automated DDC metadata generation. In this study, the input data consisted of subject keywords, title and abstract (similar to the present case). The information resources were a collection of German scientific papers (from 7 university repositories). The project matched against a thesaurus (the German Subject Heading Authority File), which in turn was mapped to DDC. Use of a thesaurus as an entry vocabulary resembles DISTIL's matching against the DDC Relative Index (plus Captions), although DISTIL directly engages with the DDC entry vocabulary. The results are reported at DDC Level 1 and 2 from the (OAI-PMH) repository of the Hochschule Hannover. Their best results at Level 2 use the combination of Title + Abstract + Keywords and yield MRR 0.61 and Rec@5 0.77. They report that the results are competitive with a state of the art machine-learning system ACT-DL (University of Bielefeld Automated Classification Toolbox for Digital Libraries).

In comparison, DISTIL's best results at Level 2 using Subjects + Terms + Phrases (Table 5) yield MRR 0.70 with Rec@5 0.76, comparing favourably on the generally more severe MRR measure. Level 1 results show better performance by the DISTIL approach (bearing in mind the caveats discussed earlier). Level 3 results are only returned by DISTI - while performance is lower than Level 2 (as expected) at MRR 0.5 and Rec@5 0.61, the results suggest that automated Level 3 DDC subject metadata could be appropriate for some use cases, for example semi-automated suggestion systems, recall enhancing configurations, or the visualisation discussed in future work.

Discussion and limitations

A method has been described for the lightweight automated augmentation of metadata from unrelated digital libraries. The method includes an integrated pipeline and set of tools for metadata harvesting and document classification. The pipeline generates DDC classes from metadata harvested from each digital library (in this case Dublin Core metadata). The modular nature of the pipeline means that it should be relatively easy to adapt and scale it to new collections of metadata. Evaluation results are generally encouraging, both for the harvesting and processing pipeline, and the automatically generated DDC. The following discussion falls into two parts: evaluation of the overall technical pipeline; evaluation of the results (which can also be seen as evaluation against an equivalent human pipeline).

In the overall pipeline, the project encountered a number of practical issues in the metadata harvest (Khoo et al., 2013). While they can be seen as 'normal' problems to be faced in any harvest, taken together they illustrate some of the more general issues that need to be addressed in harvesting workflows. Particularly, as each of the libraries in the project had a complex organizational history, this led to specific legacy metadata issues that had to be addressed on a case-by-case basis. These legacy issues were not immediately obvious, and often only came to light during the harvest itself, adding to the time, resources, and manual intervention required. This finding points to an ongoing need for tools to identify these issues. and support metadata analysis at the harvest stage.

The evaluation, results are at least competitive with related work. Comparison is however complicated by differences in datasets, vocabularies, and evaluation methodologies. Some studies involve more homogeneous datasets, sometimes using domain specific subject vocabularies. The Digging Project involved what might be considered more heterogeneous source material and factors arising from this heterogeneity should be taken into account when considering the evaluation. First, the general problem space addressed by the Digging Project is relatively heterogeneous, for instance in terms of the resources described (web sites), the metadata harvested (various flavors of native and crosswalked Dublin Core), and the domains, disciplines, and audiences covered. Second, at the input stage of the pipeline, the text that is being analyzed is the resource metadata rather than the resource itself. Third, the resource metadata is a snapshot of a description of a web site at the point of harvest, and it is possible that while a web site (unlike a published conference paper) can change over time, the attached metadata itself might not be updated (Intute, for example, closed in July 2011, and the metadata has not been updated since). In addition, the 'live' repository web page for a resource may not necessarily display all the metadata that is held for a resource, or otherwise differ from the harvested via OAI-PMH - the manual indexing process carried out by the librarian occasionally used slightly different metadata to that available to the DISTIL process. Fourth, complications arise if the evaluation considers the whole pipeline (including metadata harvesting), as differences in the configuration of any stage of the pipeline can introduce one or more confounders into any comparison of methods.

As a rough check of manual subject indexing consistency, a subset of the records were independently classified by a second librarian with experience in DDC classification, from an institution external to the project. This exercise classified 14 of the records (6 IPL, 4 NSDL, 4 Intute). The second librarian was allowed to select more than one class if considered appropriate but elected to return a single result for the major classification except for one case where an alternate was given as equally valid. This was compared with the outcomes from the second classification phase by the original librarian for the same records. Where the original librarian returned more than one class, a match on any was taken as a positive match (as in the comparison with the automatically generated classes) and similarly for the second librarian single alternate. Out of the 14 records, 12 matched to the top 3 DDC levels (in fact 9 were complete matches) and one matched to 2 DDC levels. There was one complete non match, which illustrates some of the difficulties in arriving at a single class in a discipline-based classification (mathematical principles in computer science vs programming aspect of mathematics).

Thus the exercise showed perhaps a surprisingly high level of agreement in the intellectual subject indexing. One factor that possibly supported this level of agreement was that both librarians were not cataloging *ab initio*, but rather were working to assign the harvested metadata records to the same controlled vocabulary, i.e. DDC 23 (c.f. Mann, 1997, who observes that many studies cited as evidence of low inter-cataloger reliability are studies that allowed the catalogers to choose their own subject terms).

Additionally, the exercise was to generate a DDC classification rather than more detailed (thesaurus) subject indexing.

The methodology of constructing a 'Gold Standard' is a complex issue which affects direct comparison of the technical pipeline with a human version of the same pipeline. It is not clear that an automatically assigned DDC class that differs from that supplied by a human cataloguer is necessarily incorrect in comparison with human judgment. As we see in the non match example above, this is particularly the case with discipline-based classifications such as DDC, where a subject can occur in very different hierarchies, depending on the focus of the cataloguer. This issue was noted in a study by Golub & Lykke (2009), who combined a study of user hierarchical browsing behavior via automatically assigned classes by a document classification algorithm for a set of engineering web pages, with an investigation of the correctness of automatically assigned classes assigned as perceived by the users. They reported differences in the human judgments, and that some web pages posed particular issues for judgment of appropriate classes due to a lack of text. Wartena and Sommer (2012) make a similar point that *"in many cases there is more than one possible label that could be regarded as true and a more or less arbitrary choice had to be made by the annotators. In fact labels closely related to the ground truth could be considered as correct as well"* (p. 43). This is true of the current study, involving complex, multi-faceted resources such as websites, where single subject classification can be difficult. The (original) librarian's comments on one resource, assigning two classes (616.x and 362.x) illustrate this point: *"616.742 (Fibromyalgia) AND 616.0478 (Chronic Fatigue Syndrome (CMS)) AND 362.1960478 (services to patients with CFS) as website includes resources, coping techniques and equipment to aid sufferers not just about medical conditions"*. The librarian also makes the general point *"… I think it is best to show as many classes as are applicable to highlight all the relevant resources that may be found when browsing by subject area."*. Of course, this is related to the issue of the intended use case – what activity is the evaluation aiming to support?

In terms of future work, there is clearly a need to conduct research into a more objective and comprehensive evaluation methodology that can take account of the issues discussed above concerning differences in legitimate answers. This should encompass the intended use case to be supported by the evaluation and ecological validity, issues of consistency, the possibility of multiple valid classifications from different points of view and the notion of close matches. There is also scope to expand the application of the current configuration of the pipeline. For instance, the resulting DDC Summary numbers could be expressed as dewey.info Linked Data for LOD applications. Future plans include visualization and search interfaces for end-users, to help them navigate the aggregated metadata and develop understanding of possible connections between repository items.

Conclusion

An ongoing question in digital library research concerns how to support users to search across unrelated digital libraries with a single query. One useful approach involves the automated augmentation of metadata records from different libraries, in order to create a central repository that has one or more fields in common. This paper has demonstrated the functionality of a prototype pipeline to support such an approach, from metadata harvesting, through text analysis, to the generation of DDC classes for metadata records. The method does not require training data matched to the hierarchical structure of the DDC or indeed any training set. The evaluation results are encouraging, particularly for the complex harvesting and processing pipeline. While currently specific to the DDC, generalization of the pipeline to other knowledge organization systems would not be a large step. The DISTIL pipeline is understandable to humans and can be configured differently depending on the intended use case, for example whether recall or precision enhancing.

The approach is novel on various levels. It addresses the normalization problem as it relates to metadata descriptions of Web sites, which tend to be more heterogeneous documents than articles, dissertations, etc. The automated classification method matches a combination of weighted pre-processed metadata records against the entry vocabulary of the DDC, before a further phase takes account of matches within hierarchies, aggregating lower level matches to broader parents. From this point of view, the algorithm can be considered to resemble the practice of a human DDC cataloguer; first identifying candidate hierarchies via the relative index table and then selecting the most appropriate hierarchical context for the main subject. Results suggest that adding weighted terms extracted from Title and Description can improve performance. Long-term development options include scaling the harvest to include other DLs; extending general application to other domains and knowledge organization systems. Overall, the approach is applicable to other metadata repositories that seek to add value for their users, and a natural next step would be to apply the method to academic research abstracts.

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.

Frantzi, K., Ananiadou, S. and Mima, H. (2000). Automatic recognition of multi-word terms. *International Journal of Digital Libraries 3*(2), pp.117-132.

Golub, K. (2006a). Automated subject classification of textual Web pages, based on a controlled vocabulary: Challenges and recommendations. *The New Review of Hypermedia and Multimedia 12*(1): 11-27.

Golub, K. (2006b). Automatic subject classification of textual Web documents. *Journal of Documentation, 62*(3), 350-371.

Golub, K., & Lykke, M. (2009). Automatic classification of web pages in hierarchical browsing. *Journal of Documentation, 65*(6), 901-925.

Golub K, Lykke M, Tudhope D. (2014). Enhancing social tagging with automated keywords from the Dewey Decimal Classification. *Journal of Documentation,* 70(5), 801-828. Emerald.

Greenberg, J. (2004). Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications. *Journal of Internet Cataloging, 6*(4): 59-82.

Greenberg, J., Spurgin, K., Crystal, A. (2006). Functionalities for automatic metadata generation applications: a survey of metadata experts' opinions. *Int. J. Metadata, Semantics and Ontologies, 1*(1), 3-20.

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

Hiom, D. (2006a). Retrospective on the RDN. *Ariadne Issue 47*, http://www.ariadne.ac.uk/issue47/hiom/

Hiom, D. (2006b). RDN Timeline. *Ariadne Issue 47*, http://www.ariadne.ac.uk/issue47/hiom

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 M., Hall, C. (2013). Managing metadata: Networks of practice, technological frames, and technical work in a digital library. *Information and Organization, 23*, 81–106.

Khoo, M., Tudhope, D., Binding, C., Jones, H., Orrego, I. (2013). OAI-PMH and Metadata Aggregation From Heterogeneous Digital Libraries: Three Case Studies. iConference 2013, Fort Worth, TX, February 12-15, 497-501.
Available online at: https://www.ideals.illinois.edu/handle/2142/42563

Krowne, A., & Halbert, M.: An initial evaluation of automatic organization for digital library browsing. JCDL 2005, 246–255.

Lösch, M., Waltinger, U., Hortsmann, W., & Mehler, A. (2011). Building a DDC-annotated Corpus from OAI Metadata. *Journal of Digital Information, 12*(2).

Thomas Mann (1997) “Cataloging Must Change!” and Indexer Consistency Studies: Misreading the Evidence at Our Peril. *Cataloging & Classification Quarterly, 23*:(3-4), 3-45.

Newman, D., Hagedorn, C., Chemudugunta, C., & Smyth, P. (2007). Subject metadata enrichment using statistical topic models. JCDL 2007, 366-375.

Nichols, D., Chan, C-H., Bainbridge, D., McKay, D., & Twidale, M. (2008). A lightweight metadata quality tool. JCDL 2008, 385-388.

Online Computer Library Center (OCLC) (n.d.) *Introduction to the Dewey Decimal Classification*. http://oclc.org/content/dam/oclc/webdewey/help/introduction.pdf

Sweeney, R., (1983). The Development of the Dewey Decimal Classification. *Journal of Documentation, 39*(3), 192-205.

Thompson, R., Shafer, K., & Vizine-Goetz, D. (1997). Evaluating Dewey concepts as a knowledge base for automatic subject assignment. Second ACM Int. Conf. on Digital libraries (DL '97), pp. 37-46.

Tuarob, S., Pouchard, L., & Giles, C. L. (2013). Automatic Tag Recommendation for Metadata Annotation Using Probabilistic Topic Modeling. JCDL 2013, pp. 239-248.

Voorhees E. (1999). The TREC-8 Question Answering Track Report. Proceedings of the 8th Text Retrieval Conference, 77–82. NIST Special Publication 500-246. http://trec.nist.gov/pubs/trec8/t8_proceedings.html

Waltinger, U., Mehler, A., Lösch, M., & Horstmann, W. (2011). Hierarchical Classification of OAI Metadata Using the DDC Taxonomy. Advanced Language Technologies for Digital Libraries - Lecture Notes in Computer Science Volume 6699, 2011, 9-40.

Wang, J. (2009). An extensive study on automatic Dewey Decimal Classification. *Journal of the American Society for Information Science and Technology (JASIST), 60*(11), 2269–2286.

Wartena, C., & Sommer, M. (2012). Automatic classification of scientific records using the German Subject Heading Authority File (SWD). Proceedings of the 2nd International Workshop on Semantic Digital Archives (SDA 2012, Paphos, at TPDL 2012), 37-48.

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

Wilson, A. (2007). Toward Releasing the Metadata Bottleneck. *Library Resources and Technical Services, 51*(1), 16-28.

Woodley, Mary S. (2008). Crosswalks, Metadata Harvesting, Federated Searching, Metasearching: Using Metadata to Connect Users and Information. In Baca, M. (Ed.), Introduction to Metadata. Los Angeles, CA: The Getty Trust.

Yi, K. (2007). Automatic Text Classification Using Library Classification Schemes: Trends, Issues, and Challenges. *International Cataloguing and Bibliographic Control Journal*, *36*(4), 78-82.

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.

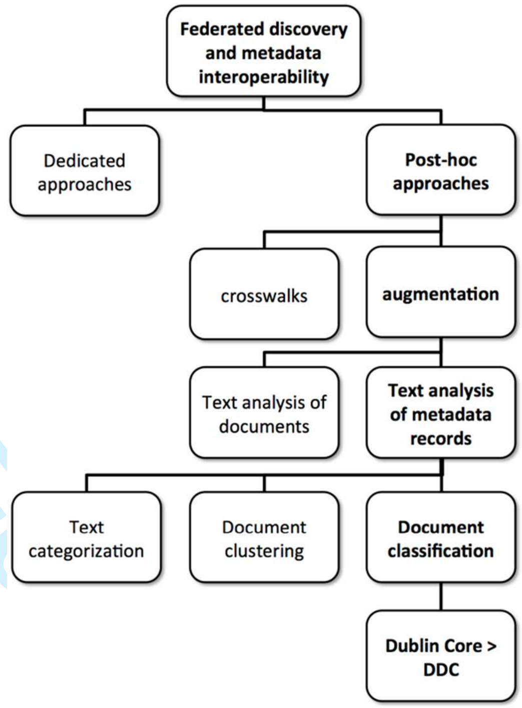

Figure 1: Problem space definition, showing general methodological choices used in the analysis

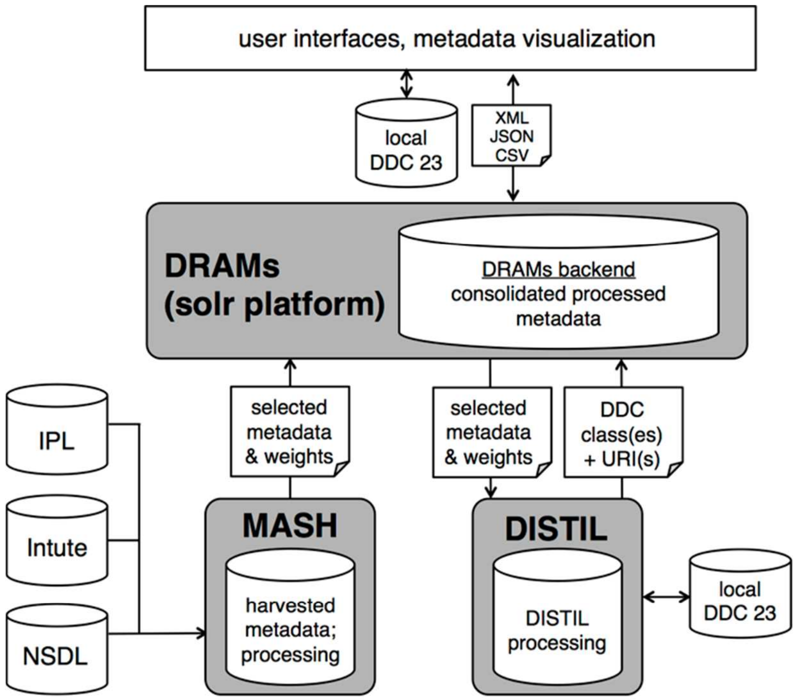

Figure 2. High-level architecture of the Digging Project.

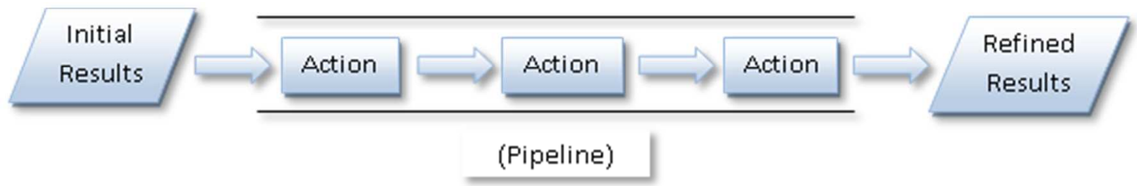

Figure 3 – DISTIL process pipeline

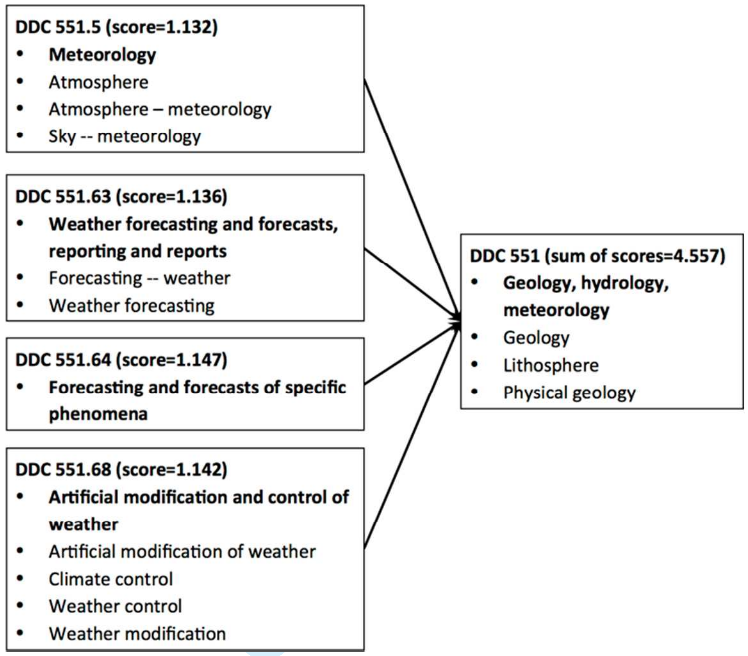

Figure 4. Upward score aggregation to summary level.

Table 1. Example of variations in duplicate records for the same resource.

<u>IPL – Chateau de Versailles</u>
Description
This museum is located near Paris and includes many masterpieces. This website describes the history of the chateau through the buildings, gardens and famous royalty that have lived there. Take a tour with the interactive map.
Subject
Chateau; Louis XIV; Marie-Antoinette; Marie-Antoinette's estate; Palace; french court; Grand Trianon; hall of mirrors; formal gardens;

<u>LII – Chateau de Versailles</u>
Description
This site contains an introduction to the palace at Versailles, France. Find history of its construction, images, and brief biographies of some of the historic figures in French history. Visiting information and events are provided. Available in English, French, and Japanese.
Subject
Architecture; Dragons, Dreams Daring Deeds; Castles Palaces; Palaces;

<u>Intute – Chateau de Versailles</u>
Description
This is the official website of the Château de Versailles. Dating back to the 17th Century, Versailles is most closely associated with Louis XIV and became, in 1682, the official residence of the Court of France. The site contains detailed information about the Château, including 360 degree panoramic views of rooms and a photographic history of the buildings and the landscaped grounds. There is also information about the notable figures associated with Versailles and some details about life as it would have been lived in the Château. Versailles is also the home of the Museum of French History and houses many works of art, some of which are detailed under the 'Masterpieces' section. The site is available in both French and English
Keywords – Controlled
Château de Versailles; French; landscape architecture; chateaux; fine arts; country houses; paintings; furniture; Baroque; Versailles--Ile-de-France--France; Louis XIV, King of France, 1638-1715;
Classification
Architecture and planning > Architectural history > Periods, styles and movements > 17th century > Baroque
Architecture and planning > Built environment > Buildings and structures > Residential buildings and structures
Architecture and planning > Landscape architecture > Garden design
Creative and performing arts > Visual arts > Art history > Museums and galleries > International

Table 2. Terms found in different elements

Term origin	Total	Average
terms in title and description not appearing in subject elements	569,913	2.16
terms from subject elements only	2,566,332	9.74
terms common to (title & description fields) and subject elements	661,661	2.51
total terms from all elements	3,797,905	14.41

Table 3. Matching between repository record and DDC class.

Resource [id=Intute:12345]		
Field type	**Field label**	**Weight**
subject	*Atmospheric science*	1.500
subject	*Climatology*	1.220
subject	*Geoscience*	0.865
subject	*Meteorology*	0.973
title	...	0.000
description	...	...

Match?

DDC Class [id=551.6]		
Field type	**Field label**	**Weight**
label	*Climatology and weather*	1.000
label	*Climate*	1.000
label	*Climatology*	1.000
label	*Weather*	1.000

Table 4. Example DDC classification.

Manual DDC classification for repository record: “330 – Economics”		
Top 10 DISTIL DDC results, based on repository record metadata		
Rank	**DDC class**	
1	336 - Public finance	
2	333 - Economics of land and energy	
3	338 - Production	
4	332 - Financial economics	
5	331 - Labor economics	
6	339 - Macroeconomics and related topics	
7	330 - Economics	
8	335 - Socialism and related systems	
9	337 - International economics	
10	334 - Cooperatives	
Level 1 RR:	**1.000**	DDC Level 1 “3” - matches “**3**36” at rank 1 (RR=1/1)
Level 2 RR:	**1.000**	DDC Level 2 “33” - matches “**33**6 at rank 1 (RR=1/1)
Level 3 RR:	**0.143**	DDC Level 3 “330” - matches “**330**” at rank 7 (RR=1/7)
Level 1 Rec@5:	**1**	DDC Level 1 = “3” - occurs within first 5 results
Level 2 Rec@5:	**1**	DDC Level 2 = “33” - occurs within first 5 results
Level 3 Rec@5:	**0**	DDC Level 3 = “330” - does not occur within first 5 results

Table 5. Mean Reciprocal Rank (MRR) and Recall at 5 (Rec@5) at first, second and third DDC levels.

Metadata fields	**DDC Level 1**		**DDC Level 2**		**DDC Level 3**	
	MRR	**Rec@5**	**MRR**	**Rec@5**	**MRR**	**Rec@5**
Subjects (no weighting)	0.651	0.673	0.453	0.531	0.294	0.449
Subjects (MASH weighting)	0.668	0.714	0.530	0.592	0.351	0.490
(MASH) Terms	0.713	0.755	0.575	0.633	0.393	0.449
(TERMINE) Phrases	0.447	0.531	0.303	0.388	0.191	0.265
Subjects + Terms	0.789	0.878	0.676	0.735	0.490	0.592
Subjects + Phrases	0.739	0.776	0.607	0.673	0.427	0.571
Terms + Phrases	0.711	0.796	0.608	0.694	0.420	0.551
Subjects + Terms + Phrases	**0.823**	**0.898**	**0.702**	**0.755**	**0.497**	**0.612**

Table 6. MRR & Rec@5 for subjects + MASH terms + TERMINE phrases, split by originating repository.

Repository	DDC Level 1		DDC Level 2		DDC Level 3	
	MRR	Rec@5	MRR	Rec@5	MRR	Rec@5
Intute	0.897	0.941	0.794	0.824	0.582	0.647
IPL	0.838	0.882	0.729	0.765	0.496	0.647
NSDL	0.722	0.867	0.567	0.667	0.400	0.533

Table 7. MRR & Rec@5 for Subjects, MASH terms, TERMINE phrases, aggregation to abridged vs. to summary level.

DDC Level	This pipeline - aggregation to abridged level		Previous pipeline - aggregation to summary level	
	MRR	Rec@5	MRR	Rec@5
1	0.737	0.755	0.823	0.898
2	0.594	0.612	0.702	0.755
3	0.390	0.408	0.497	0.612
4	0.235	0.245	n/a	n/a
5	0.046	0.082	n/a	n/a