

Please cite this article as:

Seth van Hooland, Ruben Verborgh, Max De Wilde, Johannes Hercher, Erik Mannens, and Rik Van de Walle (2011). *Free your metadata: Integrating cultural heritage collections through Google Refine reconciliation*. Pre-submission paper available on <http://freeyourmetadata.org/publications/freeyourmetadata.pdf>.

# Free your metadata: Integrating cultural heritage collections through Google Refine reconciliation

Seth van Hooland<sup>†\*</sup>, Ruben Verborgh<sup>§</sup>, Max De Wilde<sup>†</sup>,  
Johannes Hercher<sup>◇</sup>, Erik Mannens<sup>§</sup>, and Rik Van de Walle<sup>§</sup>

<sup>†</sup>Université Libre de Bruxelles  
Information and Communication Science Department  
Avenue F. D. Roosevelt, 50 CP 123  
B-1050 Brussels, Belgium  
{svhoolan, madewild}@ulb.ac.be  
Phone: +32 2 650 4765

<sup>§</sup>Ghent University – IBBT, ELIS – Multimedia Lab  
Gaston Crommenlaan 8 bus 201  
B-9050 Ledeborg-Ghent, Belgium  
{ruben.verborgh, erik.mannens, rik.vandewalle}@ugent.be  
Phone: +32 9 33 14959

<sup>◇</sup>Hasso-Plattner-Institute, University of Potsdam  
Prof.-Dr.-Helmert-Straße 2–3  
D-14482 Potsdam, Germany  
johannes.hercher@hpi.uni-potsdam.de  
Phone: +49 331 5509 547

October 2011

## Abstract

A varying level of courtship has always characterized the relation between the field of Library and Information Science and the Semantic Web community. The ambition to bring metadata into the Linked Data cloud draws the two communities closer together than they ever have been before, but metadata practitioners still lack a straightforward methodology and the tools to experiment with Linked Data. This paper gives a pragmatic overview of how a locally developed vocabulary can be successfully reconciled with the Library of Congress Subject Headings (LCSH) with the help of Google Refine. The different steps towards reconciliation are performed through freely available metadata and tools, making the process repeatable and understandable for metadata practitioners.

## 1 Introduction

### 1.1 General context and purpose of the paper

Courtship defines the period during which two persons, who sense a mutual attraction, get to know each other in order to decide whether a durable relationship can be established. The field of Library and Information Science and the Semantic Web community have been turning circles around one another ever since structured search for information on the Web became a major field of research in the second half of the 1990s. Both Library and Information Science and the Semantic Web community have been evolving significantly ever since. From the end of the nineties, the two

---

\*Corresponding author

communities invested considerably in the standard making process of metadata schemas and ontologies. However, the practice of gathering domain and technology experts to debate over a period of years to develop and fine-tune metadata schemas and ontologies has in both worlds lost a lot of its institutional support. The eContentplus funding program of the European Commission<sup>1</sup>, for example, explicitly did not fund the development of metadata schemas and the creation of metadata itself (van Hooland et al., 2010). The early-to-mid 2000s economic downturn in the US and Europe forced both fields to adopt a more pragmatic stance and to deliver short-term results towards grant providers. It is precisely in this context that the concept of Linked and Open Data (LOD) has gained momentum.

We believe that the integration of heterogeneous collections can be managed by using subject vocabulary for cross linking between collections, since major classifications and thesauri (e.g., LCSH, DDC, RAMEAU) have been made available following Linked Data Principles. Re-using these established terms for indexing cultural heritage resources represents a big potential for Libraries, Archives and Museums (LAM).

Therefore, the paper aims to examine the feasibility of using subject vocabularies as linking hub to the Semantic Web in advance of such effort. Namely, we will examine and answer the following key questions:

- What are currently the possibilities to reconcile metadata with controlled vocabularies already linked to the Linked Data cloud in a semi-automated, assisted manner with the help of non-expert tools?
- What are the characteristics of the reconciled metadata, and more specifically, do they offer a sufficient discriminatory value for search and retrieval?

## 1.2 Introducing Linked Data and SKOS

Tim Berners-Lee's TED 2009 mantra "More raw data now!"<sup>2</sup> resonated in the administrative and political domain.<sup>3</sup> The re-use of existing metadata and the attempt to gain more value out of them by offering publicly available metadata exports or a direct access over APIs is little by little unlocking some parts of the silos of metadata built up over decades. The creation and maintenance of highly structured metadata and controlled vocabularies has been the core business of cultural heritage institutions since their professionalization at the end of the 19th century. Accordingly, LAM have gained considerable attention as a valuable source for structured metadata and a growing number of projects, such as Isaac et al. (2008); Mäkelä et al. (2007); Neubert (2009) and Haslhofer and Isaac (2011), are making use of Semantic Web technologies to publish resources from the LAM domain.

The term Linked Data is referenced as a set of best practices to publish and connect entities rather than only documents. The ambition is to create a global data space of networked resources that can be queried with generic tools, rather than putting publishers and agents in charge to understand custom APIs. Using URIs to represent entities as well as their attributes and relations enables applications to process interlinked resources in an automated manner. Basically this is accomplished by using a set of web technologies along with generic data formats that extend XML. In this context, the Resource Description Framework (RDF, Brickley and Guha (2004)) and the Ontology Web Language (OWL, Hitzler et al. (2009)) make it possible to express meaning, i.e., constraints that define the semantics between resources. Additionally, one of the most beneficial characteristics is that the Linked Data approach provides a uniform model to access distributed and heterogeneous metadata. Datasets that are encoded in RDF/OWL can be queried using the SPARQL Protocol and RDF Query Language (SPARQL, (Prud'hommeaux and Seaborne, 2008)) as a generic query language independently of their structure. In order to allow interlinkage between resources, the metadata are often provided through so-called SPARQL endpoints or triple stores. We refer to Heath and Bizer (2011) for a state-of-the-art overview of Semantic Web technology and best practices.

In the context of making authority files and subject vocabularies part of the *Linked Data cloud*<sup>4</sup>, the Simple Knowledge Organizing System (SKOS, Miles and Bechhofer (2009)) has gained considerable popularity. SKOS was originally developed as a RDF Schema to represent thesauri in the Semantic Web (Alistair et al., 2005, p.17). It is basically compatible with relevant standards, such as ISO 2788:1985/1986 and ANSI/NISO Z39.19 and also has been preferred to other formats, such as Topic Maps<sup>5</sup> or zThes<sup>6</sup> because of its flexible structure and standardization by the W3C (Pastor-Sanchez et al., 2009).

Since SKOS provides straightforward relation types fit to describe hierarchy, synonymy and association relationships, it is well suited to express concepts of controlled vocabularies. Elements such as `skos:prefLabel` and

<sup>1</sup>[http://ec.europa.eu/information\\_society/activities/econtentplus/closedcalls/econtentplus/index\\_en.htm](http://ec.europa.eu/information_society/activities/econtentplus/closedcalls/econtentplus/index_en.htm)

<sup>2</sup>Tim Berners-Lee on the next Web, TED Talks, Feb. 2009.[http://www.ted.com/talks/tim\\_berniers\\_lee\\_on\\_the\\_next\\_web.html](http://www.ted.com/talks/tim_berniers_lee_on_the_next_web.html) minute 10:45

<sup>3</sup>See the following projects to illustrate the governmental support regarding the Open and Linked Data movement: <http://www.data.gov/>, <http://www.data.gov.au/>, <http://opendata-network.org/>, <http://openbelgium.be/>, <http://publicdata.eu/>.

<sup>4</sup>cf. Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch at <http://lod-cloud.net/>

<sup>5</sup><http://www.isotopicmaps.org/>

<sup>6</sup><http://zthes.z3950.org/>

`skos:altLabel` are used to describe the meaning of a concept. The relations between two `prefLabels` within one thesaurus can then be described using properties as `skos:narrower`, `skos:broader`, and `skos:related`. Additionally, SKOS is designed to match descriptors of different vocabularies by assigning `skos:closeMatch`, `skos:exactMatch` or `skos:narrowMatch` as properties between an object and a subject (Isaac et al., 2008).

### 1.3 Related work regarding the reconciliation of metadata and vocabularies

Isaac et al. (2008) identified four general approaches towards vocabulary reconciliation or alignment: 1) lexical alignment techniques, 2) structural alignment, 3) extensional alignment, and 4) alignment using background knowledge. We decided to focus on lexical alignment technologies as most of the terms can be reconciled by taking care of lemmatization, harnessing preferred labels or viewing string similarity (cf. 4.2.3). Lexical alignment can be performed easily by vocabulary managers using *Google Refine* for the reconciliation process, as will be demonstrated throughout the paper.

The reconciliation of a subject vocabulary is a crucial task for a wide range of applications in LAM institutions and has been tackled in various projects before (Tudhope et al., 2011; van der Meij et al., 2010; van Erp et al., 2011). Van Erp et al. (2011) demonstrate how to achieve a mapping between name-authority-thesauri using sophisticated NLP technology. Wikipedia articles of historic events are examined for significant phrases that indicate relationships between an entity (person, location, time) and an event. Entities, which occur in the same context and in different authorities at the same time, are used to determine a mapping between two thesauri classes. This method is only feasible for named entities (persons, locations, cities etc.) and did not work out for the subject vocabulary we intended to use. Furthermore, our approach differs because we chose to explore *non-expert* software, which can be learned and handled by vocabulary managers that do not necessarily have advanced skills in computer science or language technology.

In the scope of the CATCH/STITCH project, van der Meij et al. (2010) developed a service<sup>7</sup> to perform semiautomatic vocabulary alignment between large vocabularies, as e.g., RAMEAU, ICONCLASS and LCSH. Several vocabularies from the cultural heritage domain are imported, matched, and published using SKOS matching relations. Main attention is paid to the evaluation of machine created mappings between interlingual vocabularies derived from the MACS Project (Landry, 2009), leading to a hosting and alignment service for libraries which do not have the resources or the competencies to provide such infrastructure. However, performance and usability issues are largely neglected. Another drawback of van der Meij et al. (2010) is that vocabularies need to be transformed into RDF/SKOS first, whereas we can export our reconciliation results directly into RDF/SKOS format, which consequently can be exported into arbitrary triple stores for re-use.

Within the STELLAR<sup>8</sup> project, the CIDOC Conceptual Reference Model<sup>9</sup> was populated with mappings to glossaries and thesauri to support indexing and search within archeological resources. Tudhope *et al.* also provide a conversion service to transform CSV files into RDF/SKOS, but the scope of the STELLAR project is mostly limited to the archeological domain and its tools are not generalizable and as straightforward to use as *Google Refine*.

Crowdsourcing approaches are taken to encourage patrons in producing user-generated-metadata (van Hooland et al., 2011). The linkage between a folksonomy, captured from Delicious, and the LCSH<sup>10</sup> has been studied by Yi and Chan (2009). The authors selected three sets of 100 tags, based on their frequency (high, mid and low frequency) and analyzed the possibilities of automatically linking the tags with LCSH. Each user tag was compared on the basis of “complete-word matching”, meaning that a tag is linked with a subject heading when the tag “appears in the heading as a complete word” (Yi and Chan, 2009, p. 887). As the authors mention, this means that the tag “computer” results in a match with the subject heading “computer networks” but the tag “network” is not considered to be a complete-word match with the same heading as the tag is only a portion (“network” for “networks”) of the word in the subject heading, and “not as a complete word” (Yi and Chan, 2009, p. 887). This approach resulted in 60.9 percent of matches. Our reconciliation process differs significantly from Yi and Chan’s approach as we perform the reconciliation 1) on the entire corpus and not on statistically insignificant samples, 2) on the basis of a complete match between a locally developed vocabulary and the LCSH.

To conclude, we delineate from previous work because we provide a practical methodology and the tools for metadata practitioners to bring metadata into the Linked Data cloud. By providing a step by step guide that is reproducible with a low technical level, we aim to fertilize the interlinkage of smaller and locally developed vocabularies to larger authority files that are already interlinked within the Linked Data cloud. Vocabulary managers of small LAM prefer to learn a freely available tool that can be used in a more flexible manner rather than to rely on a specialized service/application. To our best knowledge, no attention has been paid to straightforward and handy

---

<sup>7</sup><http://www.cs.vu.nl/STITCH/repository/>

<sup>8</sup><http://hypermedia.research.glam.ac.uk/kos/stellar/>

<sup>9</sup><http://cidoc-crm.org/>

<sup>10</sup><http://id.loc.gov>

tools for vocabulary alignment/reconciliation. We believe our paper is a helpful contribution to catalyze large-scale experimentation with integrated and topic oriented navigation among heterogeneous collections.

## 1.4 Structure of the paper

The remainder of the paper is structured as following. Section 2 presents our methodology which is centered around a hands-on approach based on a case-study. The used metadata, the controlled vocabulary which already is connected to the Linked Data cloud and the reconciliation tool, all freely available, are described in detail. Before the reconciliation process can take place, we need to identify potential metadata quality issues which could lower the success of the reconciliation. The profiling and cleaning of the metadata is therefore described in Section 3. The reconciliation process itself is then explained in Section 4, which presents the results of the out-of-the-box reconciliation with the RDF extension for Google Refine but also shows how the results can be significantly augmented by a set of simple and automated manipulations. Section 5 proposes an in-depth analysis of the results, by studying the formal and semantic characteristics of the reconciled subject headings. We then get back to the two initial research questions in the conclusions regarding the insights gained throughout the paper and the added-value of Linked Data for the cultural heritage sector.

## 2 Methodology

The main impetus for this paper is a discomfort regarding the absence of real-life experimentation with Linked Data principles in small to mid-sized LAM. Recent efforts such as the W3C Library Linked Data Incubator Group<sup>11</sup> and the International Linked Open Data in Libraries, Archives, and Museums Summit (LOD-LAM)<sup>12</sup> are very encouraging and the Linked Data Pilot of Europeana<sup>13</sup> holds the potential of large groundings for interlinkage of LAM since the metadata of more than 17,000,000 digitized cultural heritage objects have been published following the Linked Data principles (Haslhofer and Isaac, 2011). These initiatives are essential, but we want to catalyze the uptake of Linked Data among collection holders who do not have the resources of large-scale international projects such as Europeana or ResearchSpace.<sup>14</sup>

We also clearly want to avoid the “black box problem”, as mentioned by Geoffrey Rockwell in the context of computational text analysis, but which is applicable to any humanities-related research project using computational techniques. When presenting research, “[...] *either the technique is encapsulated inside the black box of magical technology or it is unfolded in tedious detail obscuring the interpretation – tedious detail which ends up being a black box of tedium anyway*” (Rockwell, 2011). In order to side-step this problem, we provide all needed elements for practitioners and researchers to repeat the processes described in the paper, and only make use of freely available data and non-expert tools. It is outside the scope of the article to provide a fully detailed step-by-step description of the profiling, cleansing and reconciliation process, but a “Free your metadata”<sup>15</sup> project website has been created which offers in-depth documentation and screencasts on how to repeat the processes described in the paper.

The next sections will present the different elements of the case-study.

### 2.1 Metadata: export from the Powerhouse museum

We decided to make use of the freely available metadata export that the Powerhouse museum in Sydney provides on its website.<sup>16</sup> The museum is one of the largest science and technology museums worldwide, providing access to almost 90,000 objects which range from steam engines to fine glassware, from haute couture to computer chips. The Powerhouse has been very active disclosing its collection online and making most of its data freely available. From the museum website a tab-separated text file can be downloaded. The unzipped file (58MB) contains basic metadata (17 fields) for 75,823 objects, which are released under a Creative Commons Attribution Share Alike (CCASA) license.

The reconciliation process specifically focusses on the *Categories* field, which is populated with terms from the Powerhouse museum Object Names Thesaurus (PONT).<sup>17</sup> The thesaurus was created by the museum and first published in 1995. This controlled vocabulary is currently available as a downloadable pdf form<sup>18</sup> but an online thesaurus browser will be published from the museum’s website very shortly. PONT recognizes Australian usage

---

<sup>11</sup><http://www.w3.org/2005/Incubator/1ld/>

<sup>12</sup><http://lod-lam.net/>

<sup>13</sup><http://data.europeana.eu/>

<sup>14</sup><http://www.researchspace.org/>.

<sup>15</sup><http://freeyourmetadata.org/>.

<sup>16</sup><http://www.powerhousemuseum.com/collection/database/download.php>

<sup>17</sup>[www.powerhousemuseum.com/collection/database/thesaurus.php](http://www.powerhousemuseum.com/collection/database/thesaurus.php)

<sup>18</sup><http://www.powerhousemuseum.com/pdf/publications/phm-thesaurus-sept09.pdf>

and spelling and reflects in a very direct manner the specificity of the collection. This results for example in a better representation of social history and decorative arts, whereas only a minimal number of object names exist for the domains of fine arts and natural history. According to Sebastian Chan, Head of Digital, Social and Emerging Technologies at the Powerhouse museum, the staff of the museum responsible for the *Categories* field are trained registrars, whereas curators provide the metadata regarding significance, history and provenance of the collection.

## **2.2 Controlled vocabulary: the LCSH**

To ease the linking of resources between collections, common vocabularies are needed in order to identify shared concepts. Some vocabularies enjoy a relatively broad and international uptake, such as the Arts and Architecture Thesaurus (AAT) which can be converted to SKOS with the help of the Annocultor tool.<sup>19</sup> However, the AAT is only available through a license and therefore unsuited for our approach. Secondly, one single gold standard vocabulary to which all cultural heritage institutions refer does not exist. The only feasible approach consists of aligning independent vocabularies, creating rich semantic networks which are extended every time a new vocabulary is aligned to one which already is connected to other thesauri or classification schemes.

In the absence of freely-available vocabularies specifically developed for object-centered collections, we decided to make use of one of the most popular controlled vocabularies from the cultural domain: the LCSH, being well aware that these have initially been developed for bibliographic descriptions. As O'Neill and Chan (2003) mention, the LCSH is the largest available controlled vocabulary in English, covering a large variety of subject areas. It is not a thesaurus in the strict sense of the word, but offers synonym and homograph control and cross references linking terms together. The SKOS-ification of the LCSH has been well-documented (Summers et al., 2008). The use of LCSH will also allow a future mapping to the Répertoire d'Autorité-Matière Encyclopédique et Alphabétique Unifié (RAMEAU) and SchlagWortnormDatei (SWD), the German indexing subject headings list, giving potential access to resources in French and German too. In 1998, four European national libraries established the Multilingual Access to Subjects (MACS) project in the context of the Conference of European National Librarians (CENL).

The manual alignment process between the RAMEAU, the LCSH and the SWD took ten years and is now implemented at the Swiss National Library and the Deutsche Nationalbibliothek (Landry, 2009). The STITCH project has re-used the outcomes of the MACS project and worked on vocabulary alignment by automatically detecting inter-vocabulary semantic mappings between vocabularies such as Iconclass, the Brinkman thesaurus and WordNet (van der Meij et al., 2010). The interlinking of these vocabularies in multiple languages has the potential to facilitate multilingual and cross-collection search and retrieval.<sup>20</sup>

## **2.3 Tool for the profiling, cleansing and reconciliation of metadata: Google Refine**

Previous research regarding data profiling of museum metadata relied on an open-source general-purpose data profiling tool.<sup>21</sup> Despite its merits, this profiler has some limitations: only five analyses are available and an XML profile specification file has to be manually set up.

Google Refine<sup>22</sup> (formerly Freebase Gridworks) is a tool designed to quickly and efficiently process, clean and enrich large amounts of data in a single interface. Made available in October 2010, it provides a number of analyses such as splitting or joining multi-valued cells, converting data into new forms, faceting textual or numerical values, detecting blank cells, trimming whitespace, etc. The tool also offers a powerful clustering functionality, based on the key collision and nearest neighbor algorithms, allowing to detect near-duplicates. Google Refine further allows to reconcile data with existent knowledge bases, creating a connection with the Linked Data community.

The DERI research group has developed an RDF extension for Google Refine, which can be freely downloaded.<sup>23</sup> The RDF extension allows users to add SPARQL endpoints to the reconciliation process. DBpedia is for example added, so that the content of the categories field can be matched to terms described as SKOS concepts in DBpedia. More specialized sources such as the DDC or the LCSH can also be used, as we will show later in this paper.

---

<sup>19</sup><http://annocultor.eu/converters.getty.html>

<sup>20</sup>For a concrete view on the scope of the project, consult the CATCH Vocabulary and alignment repository demonstrator available on <http://www.cs.vu.nl/STITCH/repository/>.

<sup>21</sup>The data profiler developed by Yves Bontemps is available at <http://sourceforge.net/projects/dataprofiler/>.

<sup>22</sup><http://code.google.com/p/google-refine/>

<sup>23</sup><http://lab.linkeddata.deri.ie/2010/grefine-rdf-extension/>

### **3 First step: profiling and cleansing of metadata**

#### **3.1 Context of data profiling and cleansing**

Data quality is obviously not a new issue for the Library and Information Science field. Especially in the library context, managers have been put under pressure since the 1980s to cut back on human resources attributed to cataloging, leading especially in the US to a vivid debate regarding what exactly quality cataloging beholds. The work of David Bade, such as Bade (2009), provides valuable insights into this discussion.

Metadata practitioners that worked on aggregation projects, harvesting metadata from different partners, must acknowledge that the quality of existing metadata is hardly questioned and only becomes visible once they are put to work and queried by a large number of users. After all, which collection holder wants to stand up in the middle of his or her peers and warn them about the low level of his or her metadata? This misplaced trust causes delays and failures when metadata do not live up to expectations (van Hooland et al., 2008). More importantly, we have to acknowledge that there are no established methodologies or tools for metadata quality evaluation, or to put it more bluntly in the words of Diane Hillmann: “There are no metadata police” (Hillmann and Phipps, 2007), even if the initiatives such as the Pedantic Web group<sup>24</sup> try to offer a typology of to be avoided errors when publishing metadata in the context of the semantic web. Conceptual frameworks on metadata quality criteria have been developed, but both practitioners and researchers hardly apply these theoretical frameworks to analyze their metadata.

In the absence of concrete methodologies and tools, metadata practitioners usually believe that producing information describing the quality of their metadata is too big a step to be taken. This paper will therefore present the notion of data profiling, defined by Olson (2003) as “*the use of analytical techniques to discover the true structure, content, and quality of a collection of data*”. Before asking the question how to link metadata from different sources, we need to develop strategies to check their initial quality and eventually solve issues which might disturb the reconciliation process amongst different resources. Only then will we be able to evaluate the real added-value of the Linked Data approach for the cultural heritage sector. We will illustrate now with the help of Google Refine how a quick overview of the metadata quality of a collection can be gained and which type of cleansing actions can be taken.

#### **3.2 Profiling and cleansing the Powerhouse Museum metadata**

Prior to presenting the results obtained from the profiling of the Powerhouse Museum metadata, we clearly want to state that we are not pointing fingers at issues from their particular collection, but want to give a honest appraisal of common issues, which can help other collection holders to act upon their metadata.

##### **3.2.1 Deduplication**

Before the metadata from the Powerhouse can be analyzed in Google Refine, they need to be converted to Unicode (UTF-8) as the initial text file is stored in the ISO-8859-1 (Latin 1) format, which results in character encoding problems. Once the metadata are loaded within the application, the first operation we need to perform is to detect and remove duplicates. This can easily be performed by sorting on *Record ID* and performing the *Blank down* command, detecting consecutive duplicated cells. In this manner, 86 records were identified and deleted from the metadata set. The sorting operation also allowed us to detect three records which did not contain a record ID or any other information, except for an invalid persistent link, which was automatically generated. These three records were also deleted.

##### **3.2.2 Atomization**

Once the duplicate records have been eliminated, we can have a closer look at the individual fields. A quick glance at the values stored within fields such as *Title*, *Provenance (Production)*, and *Provenance (History)* illustrates one of the biggest hurdles for automated metadata analysis: field overloading. A title such as “*2001/32/1 Cricket ball and core, 'Platypus Gem', leather/ cork/ wool/ rubber, Platypus Sporting Goods (Dave Brown) Pty Ltd, Australia, 2000*”<sup>25</sup> regroups in one single unstructured field information which could be split out over more specific fields, such as registration number, material, manufacturer, place of production and date of production.

The grouping of these elements within one field obviously makes it difficult to compare and cluster values regarding exactly the same characteristic of an object. Google Refine offers the possibility to split the multi-valued content of fields out in extra cells or columns, on the basis of a delimiter (e.g., colon, semicolon, pipe...), field

---

<sup>24</sup><http://pedantic-web.org/>

<sup>25</sup>Persistent link to the object: <http://www.powerhousemuseum.com/collection/database/?irn=10019>

length or a regular expression. The success of this operation entirely depends on the consistent use of the same type of delimiter in all the fields of the whole metadata set.

Multi-valued fields with multiple instances of the same type of content also need to be atomized if we want to perform automated analyses such as faceting and clustering. Here we will focus on the content of the field *Categories*, which can contain within the initial export multiple values such as for example “*Abacus|Writing equipment|Calculating Instruments|Writing and Printing Equipment*”. On average, 2.25 categories are attributed per object. These values mostly go from the most specific to the most general heading. However, a lot of exceptions make it impossible to presume that we could automatically extract for example the most specific heading by taking the first value appearing within the field.

In the case of “*Specimens|Wool specimens|Animal Samples and Products*”, a general keyword precedes a second, more specific one. Other values do not necessarily have a broader/narrower term relationship but relate to different classes of keywords, as in the case of “*Advertising cards|Health and Medical Equipment*”. Outright doubles also occur, as in “*Photographs|Booklets|Documents|Photographs*”: 1668 records (about 2%) are concerned with double keyword entries. In order to analyze in detail the use of the keywords, we decided to split out the values of the *Categories* field out in individual cells on the basis of the pipe character (“|”), expanding the 75,823 records into 170,311 rows.

### 3.2.3 Blank values

Once the content of the different metadata fields has been properly atomized, filters, facets and clusters can be applied to give a quick and straightforward overview of classic formal metadata issues. By applying the custom facet “facet by blank”, one immediately gets an overview of the percentage of blank values in a field. Table 1 gives an overview of the fields for which the content remains blank in more than 50% percent of the cases.

Field name	Blank values
Marks	75%
Production date	74%
Provenance (production)	65%
Provenance (history)	86%
Height	60%
Width	52%
Depth	74%
Diameter	97%
Weight	99%

Table 1: In nine metadata fields, more than 50% of values are blank.

### 3.2.4 Formats and case inconsistencies

Facets give a quick overview of the distribution of the types of content within a metadata field, and to filter the entire collection on a specific value. This also allows in an indirect manner to analyze the different types of formats used to encode values. Applying a text facet on the field *Production date* shows us the different date formats used.

Regular expressions can also in this context be very helpful. On the first sight, the values contained within *Height*, *Width* and *Depth* are all described in mm but with the help of a regular expression (e.g., text filter on “cmlmm”) one can still detect some unity problems for a small number of objects. For example, all measures in the *Width* field are in millimeters except for 30 objects (8 in meters; 22 in centimeters).

### 3.2.5 Clustering

Applying a text facet on the *Categories* field allows us to have an overview of the totality of different headings used in the collection, in this case 4,895. The headings can be sorted alphabetically or by frequency, giving a list of the most-used terms to index the collection. The top three headings are “*Numismatics*” (8,012), “*Ceramics*” (7,389) and “*Clothing and dress*” (7,280).

After the application of a facet, Google Refine proposes to cluster facet choices together based on various similarity methods, such as nearest neighbor or key-collision. As Figure 1 illustrates, the clustering allows to solve issues regarding case inconsistencies, incoherent use of either the singular or plural form and simple spelling mistakes.

**Cluster & Edit column "Categories"**

Multiple cluster methods with multiple key/distance functions detect and solve several variations.

Method: key collision    Keying Function: ngram-fingerprint    Ngram Size: 3

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	9	<ul style="list-style-type: none"> <li>Air bricks (8 rows)</li> <li>Airbricks (1 rows)</li> </ul>	<input checked="" type="checkbox"/>	Air bricks
2	351	<ul style="list-style-type: none"> <li>Transport-Water (350 rows)</li> <li>Transport - Water (1 rows)</li> </ul>	<input checked="" type="checkbox"/>	Transport-Water
2	8	<ul style="list-style-type: none"> <li>Doorknobs (7 rows)</li> <li>Door knobs (1 rows)</li> </ul>	<input checked="" type="checkbox"/>	Doorknobs
2	2	<ul style="list-style-type: none"> <li>Band saws (1 rows)</li> <li>Bandsaws (1 rows)</li> </ul>	<input checked="" type="checkbox"/>	Band saws
2	12	<ul style="list-style-type: none"> <li>Bookmarks (11 rows)</li> <li>book marks (1 rows)</li> </ul>	<input checked="" type="checkbox"/>	Bookmarks
2	261	<ul style="list-style-type: none"> <li>Swatch books (207 rows)</li> <li>Swatchbooks (54 rows)</li> </ul>	<input checked="" type="checkbox"/>	Swatch books
2	11	<ul style="list-style-type: none"> <li>Mailbags (8 rows)</li> <li>Mail bags (3 rows)</li> </ul>	<input checked="" type="checkbox"/>	Mailbags
2	4	<ul style="list-style-type: none"> <li>Skullcaps (3 rows)</li> <li>Skull caps (1 rows)</li> </ul>	<input checked="" type="checkbox"/>	Skullcaps

Annotations: *spacing differences* (around Transport-Water), *capitalization differences* (around Bookmarks), *spelling differences* (around Mailbags).

11 clusters found

# Rows in Cluster: 0 — 680

Average Length of Choices: 8.5 — 16

Length Variance of Choices: 0.5 — 1

effect of the selected cluster method on the entire dataset

Select All    Deselect All    Merge Selected & Re-Cluster    Merge Selected & Close    Close

Figure 1: Clustering allows to detect equal values with different encodings.

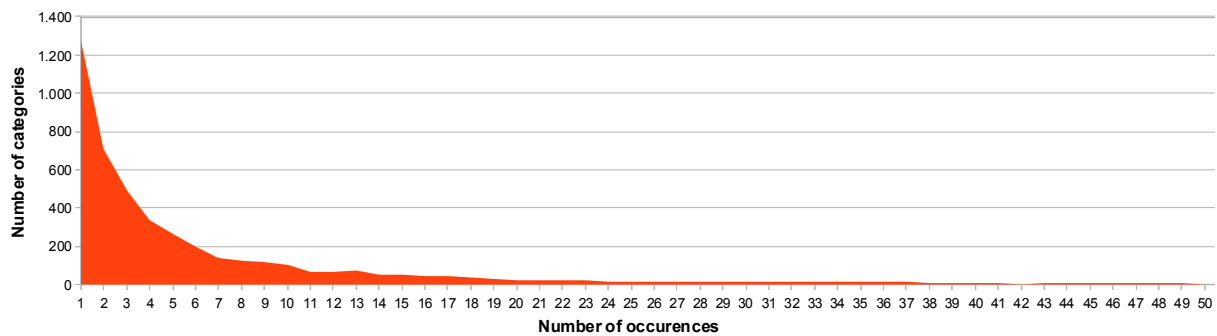


Figure 2: Most headings are only used a few times, resulting in a power-law distribution.

Google Refine presents the two related values and proposes a merger to the most recurrent value of the two.

Once the clustering process cleaned up the list of headings, we can easily export the list as tab-separated values to make, for example, visualizations such as a distribution graph, clearly illustrating a power-law distribution. The top 20 categories are used to describe 88% of the objects, and 1,272 categories (26%) are used only once in the collection; 3,654 categories (75%) are used less than 10 times.

### 3.3 Practical outcomes

We have demonstrated with the help of practical examples how Google Refine can be used to detect classic metadata quality issues such as duplicates, field overloading, blank values, format and case inconsistencies and spelling or typographic errors. The Google Refine interface allows the immediate manual editing of the values and, for some errors, the tool offers automated or semi-automated cleansing as it is the case with clustering. We have focused primarily on an analysis of the *Categories* field, as the reconciliation process described in the next section relies on the values contained within that field.

## 4 Second step: reconciliation of metadata

### 4.1 Context of reconciliation

For long, linguists and computer scientists have been trying to construct a comprehensive ontology of the world, enabling automated reasoning tasks on human-structured data. Vital to the success of such an ontology are the number and nature of the relationships between different concepts. In recognition of this importance, the Semantic Web community formed a Linked Data movement (Bizer et al., 2009), which strives to publish interlinked data in a structured format.

Berners-Lee (2006) has put forward a five-star scheme to score data quality against. The final, perfecting level can be reached when data is linked against other sources, implying the use of well-defined relationships such as equivalence, inclusion, inheritance, *etc.* As a result, machines are able to both broaden and deepen their understanding of data, since links provide the possibility to look up new data and to relate uninterpreted data to well-understood concepts.

Similarly, reconciliation is the process in which we map metadata concepts in a certain (often situation-specific) vocabulary to another (often more commonly used) vocabulary. In case the latter vocabulary forms part of the Semantic Web, this reconciliation act actually fulfills the fifth star in the Linked Data scheme, as it annexes the metadata to the Linked Data cloud. Subsequently, machines can now access and interpret these metadata, based on previously acquired knowledge. Reconciliation therefore plays a crucial role in the public availability and dissemination of metadata.

### 4.2 Reconciling the Powerhouse Museum metadata

#### 4.2.1 Library of Congress Subject Headings

The LCSH have been formalized in RDF using the SKOS vocabulary, which enables us to reconcile the Powerhouse Museum metadata against a Linked Data source. As of August 2011, they consist of 406,629 concept entities, all of which have a *preferred* label (`skos:prefLabel`) and some having one or multiple *alternate* labels (336,012 `skos:altLabels` in total). Different types of headings can share the same label, resulting in 403,894 unique labels.

The LCSH dataset can be queried through a SPARQL endpoint,<sup>26</sup> which unfortunately suffers from availability and performance problems. Alternatively, a serialized RDF version of the dataset is available at the Library of Congress website,<sup>27</sup> which we used to set up a private endpoint dedicated to the reconciliation tasks, eliminating performance issues.

#### 4.2.2 Initial reconciliation

When discussing reconciliation results, we will distinguish between the number of matched *rows* and the number of matched *records*. In total, 167,004 categories have been assigned to 75,275 objects, giving 2.2 categories per object on average. Google Refine presents each category assignment as a row and each object as a record. A successful automatic reconciliation of a category assignment is called a *matching row*, whereas we have a *matching record* when at least one of its assigned categories has been reconciled.

We first set up our LCSH endpoint in Google Refine by registering it as a SPARQL reconciliation service. Reconciliation is started on the *Categories* field, which has been cleansed, clustered, and split into individual values previously. Initially, we see that 20,185 rows (12.09%) or 19,056 records (25.32%) have been matched, if we only consider preferred labels. When we also take alternate labels into account, match scores go up to 29,710 rows (17.79%) or 26,471 records (35.17%). Since the LCSH has been originally developed to add subject headings to books, these initial numbers are surprisingly good by linking more than one third of the Powerhouse museum collection to the Linked Data cloud.

Google Refine provides us with the tools to investigate the unmatched rows, which roughly fall into two categories: those for which it was able to find one or several suggestions, and those that do not relate to any LCSH category. A suggestion consists of a category label and a percentual score indicating the correspondence between the label and the row category. Again, two cases occur: either only partial matches exist, or either multiple exact correspondences were found (since a single exact correspondence would cause a match).

The former case includes slight aberrations such as “*Ice-skates*” (PHM) versus “*Ice skates*” (LCSH) or serious semantic deformations such as “*Leaflets*” (PHM) versus “*Leaflets dropped from aircraft*” (LCSH). The latter case originates in the fact that some headings share labels. One prominent example is “*Numismatics*” (affecting

<sup>26</sup><http://api.talis.com/stores/lcsh-info/services/sparql>

<sup>27</sup><http://id.loc.gov/download/> choose “LCSH RDF/XML” or “LCSH N-Triples”

8,012 rows), which exists in LCSH both as a general subdivision and as a topical term. General subdivisions are used as a complement to another heading (e.g., “*Washington, George, 1732-1799–Numismatics*”), whereas topical terms are used directly in descriptions (Svenonius, 2000). Obviously, we want these rows to use the topical term, but Refine does not know how to reckon with these preferences.

### 4.2.3 Enhanced reconciliation

The analysis of the initial reconciliation results lead to an enhanced reconciliation method, which incorporates our preference of topical terms to subdivisions. To achieve this, we preprocessed the LCSH RDF dataset, retaining only one heading per unique preferred label. Alternate labels were added only to the extent that did not cause clashes with other labels. The resulting data set was fed into our SPARQL endpoint.

The success rate of the subsequent reconciliation attempt was astonishing: 70,270 rows (42.08%) or 55,379 records (73.57%) had been matched, and even 79,538 rows (47.63%) or 59,162 records (78.59%) when including alternate labels in the process.

Looking at the remaining mismatches, we also spotted a difference in grammatical number. In LCSH, concepts are in singular (e.g., “*Viscosity*”) and objects in plural (e.g., “*Dogs*”) as indicated by Broughton (2004).

The categories assigned to the Powerhouse Museum objects seem to always be plural. Therefore, we also performed an additional reconciliation on the singularized versions of the categories. Together, this resulted in 83,776 rows (50.16%) or 61,317 records (81.46%) with successful reconciliation.

Reformulating, we can state that *more than half* of category usages and *more than four-fifths* of collection objects have been reconciled automatically. It also implies that 81.46% of the Powerhouse Museum collection has been connected to the Linked Data cloud, with only minimal effort. The following section will analyze whether these surprisingly positive results have the potential to create meaningful ties between the Powerhouse objects and other resources.

## 4.3 Facilitating the reconciliation process

The Powerhouse Museum use case pinpoints some issues with the reconciliation process. Clearly, successful reconciliation demanded some advanced interventions that required more thorough technical knowledge. Fortunately, it is possible to abstract the technical inconveniences away into the framework. There are two main possibilities to achieve this:

**Using the preprocessed LCSH headings** The preprocessing steps detailed in 4.2.3 need to be carried out only once. Their result can be reused subsequently in other reconciliation processes. It is possible to make the preprocessed labels available using different triples, stored in a separate SPARQL endpoint. This is the easiest way, but it makes the metadata too dependent on the functionality of the Google Refine RDF extension.

**Enhancing the functionality of the Google Refine RDF extension** Alternatively, the preprocessing steps could become part of a future version of the RDF extension. By setting preferences in the interface, users could express how the RDF extension treats duplicate labels. For example, the user could give priority to preferred labels over alternate labels, and priority to topical terms over general subdivisions. This would help select the correct label in case of multiple choices. Additionally, a stemming mechanism could help with singular-plural issues.

## 5 Analysis of the reconciliation results

Three questions were asked regarding the reconciled headings: 1) what are their formal characteristics (the number of terms composing the headings and their phrasal construction, e.g., noun + adjective), 2) is there a semantic consistency between the matched headings from the local Powerhouse museum vocabulary and the LCSH (specific focus on potential issues regarding polysemy), and 3) do the reconciled headings provide a sufficient level of granularity to offer an added-value in the context of search and retrieval?

The first question was answered by automated analyses of the entire corpus of 167,004 rows. The second and the third question require the manual analysis of a sample population of matched headings. The success rate of the reconciliation process differs considerably on the basis of the number of terms composing a heading. Therefore, samples were taken from populations composed of either one (53,964 matches), two (8,532 matches) or three (7,797 matches) terms. The four term headings only contained 5 matches, consisting of the same heading (“*Single lens reflex camera*”). The sample populations were calculated on the basis of a level of confidence of 95% and an interval of confidence of 5, resulting in sample populations of 381 (one term heading), 368 (two term heading) and 366

(three term heading). To obtain statistical independence, every sample was selected by a pseudorandom number generator that allowed repetitions.

## 5.1 Formal characteristics of the reconciled headings

Table 2 gives an overview of the success rate of the reconciliation process in relation to the number of terms composing a heading.

# terms	% keywords	% success	Examples
One term	49%	66%	Flatirons
Two terms	32%	16%	Chocolate molds
Three terms	14%	34%	Clothing and dress
Four terms	5%	0%	Single lens reflex camera
Five terms	0.10%	0%	Automata and Mechanical Musical Instruments

Table 2: Reconciliation is more successful with a low number of terms.

The single-term headings obviously score the best with a staggering 66% of headings matched to the LCSH. The percentage immediately drops to 16% with headings consisting of two terms. Here we really see the beneficial impact of the use of the alternate labels attached to the topical term. The `skos:altLabel` demonstrates its utility in the sample by respectively linking non-preferred terms used by the Powerhouse museum such as “*Hand loom*”, “*Chocolate moulds*”, and “*Personal effects*” to the topical terms from the LCSH “*Handlooms*”, “*Chocolate molds*”, and “*Personal belongings*”.

Intuitively, one would expect the percentage of reconciled headings to be negatively correlated with the number of terms but an impressive 34% of three term headings could be matched with the LCSH. This can be explained through the high percentage of three term headings composed of two general nouns, combined with a conjunction, as in the case of “*Clothing and dress*”. Throughout our entire corpus, only one value composed of four terms (“*single-lens reflex camera*”) provided a match with the LCSH. No matches were found for keywords consisting of more than four keywords.

## 5.2 Semantic consistency between the reconciled keywords and the LCSH

We expected to find some cases of polysemy during the manual analyses of the samples, particularly in the sample with single-term keywords. Examples mentioned by Yi and Chan (2009) regarding the matching of a folksonomy with the LCSH, such as “*ajax*” and “*hacks*” point out to the potential impact of polysemy. These terms relate in the LCSH respectively to Greek mythology and for example tie hacks, whereas taggers have probably used them to indicate technology-related concepts. We expected to find similar examples in our corpus. The Powerhouse museum could for example use the heading “*Wood*” to describe a photograph representing an area with many trees, which would then automatically be matched with “*Wood*”, the topical term `sh85147783` used to indicate a type of construction material or a composite of a tree. However, no example of polysemy was found among our samples of single-term or multi-term keywords. This is due to the fact that terms prone to polysemy, such as “*Cups*”, which could refer to an acronym, a teacup, a trophy or a contest, are avoided as subject headings.

## 5.3 Level of granularity and minimum level of depth

Apart from the semantic consistency between the reconciled keywords and the LCSH, we also need to investigate whether the reconciled keywords possess a sufficient level of granularity to offer a real added-value towards collection holders and end-users. If a majority of the matches with the LCSH would only concern very broad concepts, such as for example “*Photographs*”, the added-value of reconciliation remains rather limited. But how can we analyze the level of granularity?

There does not exist a formal method to categorize keywords in an deterministic manner into different levels of granularity, as this characteristic is subject to human experience and therefore context-dependent. We should thus underline the distinction between deterministic and empirical data, and hence the type of assumptions we can draw from them. As Isabelle Boydens clearly points out, deterministic data are “characterized by the fact that there is, at any moment, a theory which makes it possible to decide whether a value (v) is correct. This is the case with algebraic data: in as much as the rules of algebra do not change over time, we can know at any time whether the result of a sum is correct. But for empirical data, which are subject to human experience, theory changes over time along with the interpretation of the values that it has made possible to determine” (Boydens, 2011, p. 113).

Boydens mentions, for example, the medical domain, where theory evolves with the accumulation of experience, as witnessed, for instance, in the current research into influenza A(H1N1). Applied to the issue of keyword granularity, we could think of terms used as stop words in most domains such as “the” and “who” which could be discriminatory in the music domain when querying for “The Who”.

Despite the absence of a deterministic framework to define the level of granularity, we experimented with statistical analyses to get a general understanding of the level of specificity of the reconciled keywords. Intuitively, one would expect a relation between the number of terms composing a keyword or a subject heading and their level of specificity. During the manual analysis of the sample to detect semantic inconsistencies, we made the following observations. A minority of the single-term keywords relate to very broad and general types of objects, such as “*Photographs*”, “*Tools*” and “*Specimens*”. However, the majority of the single-term keywords deliver sufficient discriminatory value to perform interesting queries over large, heterogeneous metadata sets. Keywords such as “*Flatirons*”, “*Carburetors*” or “*Comptometers*” identify highly specific object types. The two-term keywords deliver a very precise description of the object, illustrated by reconciled keywords such as “*Babby rattles*”, “*Lawn bowls*”, “*Snuff bottles*”, “*Mustard pots*” and “*X-ray tubes*”. However, a big portion of the sample contained recurrent two-term keywords such as “*Botanical specimens*” and “*Personal Effects*”. The three-term keywords have the potential to describe very specific objects. But 80% of these values represent “*Clothing and dress*”, which remain fairly general. Mostly two nouns are combined, as in the case “*Clocks and watches*”. The only non-composite keyword matched to the LCSH is “*Telephone answering machines*”. The only matched four-term keyword is “*Single-lens reflex camera*”.

However, the attribution of headings to categories such as specific or generic remains inherently subjective. In order to have a more objective measure of the level of specificity, we calculated the level of depth of the reconciled headings. One heading can have several broader terms, which can have in their turn several broader terms. One heading can therefore have different path lengths, depending on what broader term is chosen to calculate the path length. For reasons of consistency and clarity, we decided to calculate the level of depth based upon the shortest path in the LCSH. We define the minimum level of depth  $\Lambda_{min}$  of a topic heading  $t \in T$  as follows:

$$\forall t \in T : broader(t) = \emptyset \wedge narrower(t) = \emptyset \Rightarrow \Lambda_{min}(t) = 0 \quad (0)$$

$$\forall t \in T : broader(t) = \emptyset \wedge narrower(t) \neq \emptyset \Rightarrow \Lambda_{min}(t) = 1 \quad (1)$$

$$\forall t \in T : \min_{\lambda} (\exists b \in T : b \in broader(t) \wedge \Lambda_{min}(b) = \lambda) \Rightarrow \Lambda_{min}(t) = \lambda + 1 \quad (2)$$

First, all headings that do not have any broader or narrower headings, and thus are not part of any hierarchy, are trivially assigned level 0 (0). Headings without broader, but with narrower headings, are assigned level 1 (1), and all other headings are assigned one level deeper than their highest direct broader heading (2).

Table 3 and Figure 3 give an overview of the results. Almost half of the LCSH (45.55%) are positioned on level 0, and therefore do not have any broader or narrower headings, but only 9.92% of the reconciled headings belong to this group. These are components of complex subject types, such as for example “*Specimen*” which is a component of the complex subject “*Printing–Specimens*”. As these terms do not express precise concepts, it is important to know that less than 10% of the reconciled keywords belong to this level. On the other hand, Figure 3 illustrates that there is a clear match of the presence of level of depth 1, 2 and 3 values between the LCSH and the reconciled PHM keywords. A surprising high amount of reconciled keywords are positioned between level 3 (9.74%), level 4 (23.76%), level 5 (4.90%), level 6 (10.60%), level 7 (6.00%) and level 8 (9.95%). As we previously stated, we cannot claim to determine the level of granularity in an absolute, deterministic manner but Table 3 and Figure 3 provide at least indicators that the reconciled keywords are not limited to general and broad concepts with no sufficient discriminatory value for search and retrieval.

## 6 Conclusions

The two following research questions were asked at the beginning of the paper:

- What are currently the possibilities to reconcile metadata with controlled vocabularies in a completely automated manner with the help of non-expert tools?
- What are the characteristics of the reconciled metadata and, more specifically, do they offer a sufficient discriminatory value for search and retrieval?

Section 4 presented the reconciliation results in a two-fold process. By taking into account the alternate labels, 29,710 rows (17.79%) or 26,471 records (35.17%) were matched in the initial reconciliation. The analysis of these results lead to an enhanced reconciliation method, which incorporates our preference of topical terms to subdivisions.

<b>Level</b>	<b># reconciliations</b>	<b>% reconciliations</b>	<b>% headings</b>	<b>Examples</b>
0	7,894	9.92%	45.55%	textiles, specimens, models, negatives, coins, computers
1	2,933	3.69%	6.24%	personal effects, musical instruments, toy trucks
2	16,584	20.85%	21.89%	numismatics, toys, glass, medals, books, scientific instruments
3	7,746	9.74%	5.75%	electronics, badges, tools, furniture, measuring instruments
4	18,902	23.76%	5.71%	ceramics, clothing and dress, packaging, philately, engines
5	3,897	4.90%	3.99%	bottles, jackets, stockings, scarves, shawls, collars, coats
6	8,4292	10.60%	2.86%	vases, jewellery, lighting, jugs, dresses, shoes, hats, postcards
7	4,773	6.00%	1.75%	botanical specimens, sculpture, brooches, spoons, finger rings, earrings, bracelets
8	7,918	9.95%	1.55%	photographs, cups and saucers, tsuba, mugs, aeroplanes, drinking cups, prints, busts
9	318	0.40%	2.35%	building stones, lithographs, bricks, puppets, sheet music, groats, wood engravings
10	140	0.18%	1.98%	chromolitographs, tea, glove puppets
11	4	0.01%	0.33%	cigars, cigarettes
12	0	0.00%	0.04%	-
13	0	0.00%	0.00%	-

Table 3: Assigned levels of depth of the reconciled subject headings

The LCSH RDF dataset was therefore preprocessed, retaining only one heading per unique preferred label. Alternate labels were added only to the extent that did not cause clashes with other labels. These actions resulted, together with the inclusion of alternate labels and additional reconciliation on the singularized versions of the categories, in the successful matching of 83,776 rows (50.16%) or 61,317 records (81.46%) of the Powerhouse museum with the LCSH. We indicated how the proposed preprocessing could be incorporated into the Google Refine RDF extension. Such functionality would remove the burden for end-users, allowing them to obtain similar results without the additional technical effort.

We can therefore conclude that it is currently possible to reconcile in a satisfactory manner uncontrolled keywords by using a non-expert tool such as Google Refine. Even if there remains a huge potential for the reconciliation of existing metadata, collection holders will perhaps, over the next few years, steadily move over to the implementation of controlled vocabularies as web services, to be used within cataloging software offering an on-the-fly reconciliation during the encoding of metadata. The open-source collection management software CollectiveAccess has for example implemented the LCSH as a dynamic field which can be activated within the chosen metadata scheme. An on-the-fly check-up is being made within the LCSH when catalogers start typing in a keyword.<sup>28</sup>

The second research question was tackled in Section 5, in which we analyzed the characteristics of the reconciled headings and undertook the difficult process of evaluating their discriminatory value for search and retrieval. Firstly, we had a look at the correlation between the number of terms composing a heading and its reconciliation success. As expected, the large majority of the Powerhouse museum headings consist of one term, out of which 66% have been successfully reconciled. The success rate significantly drops with the increase of the number of terms, but two (16%) and three (34%) term headings still score reasonably well. Additionally, an evaluation of the semantic consistency between the reconciled headings and the LCSH was performed, based upon the manual analysis of a

<sup>28</sup>An online demo is freely accessible from the CollectiveAccess website: <http://demo.collectiveaccess.org/>.

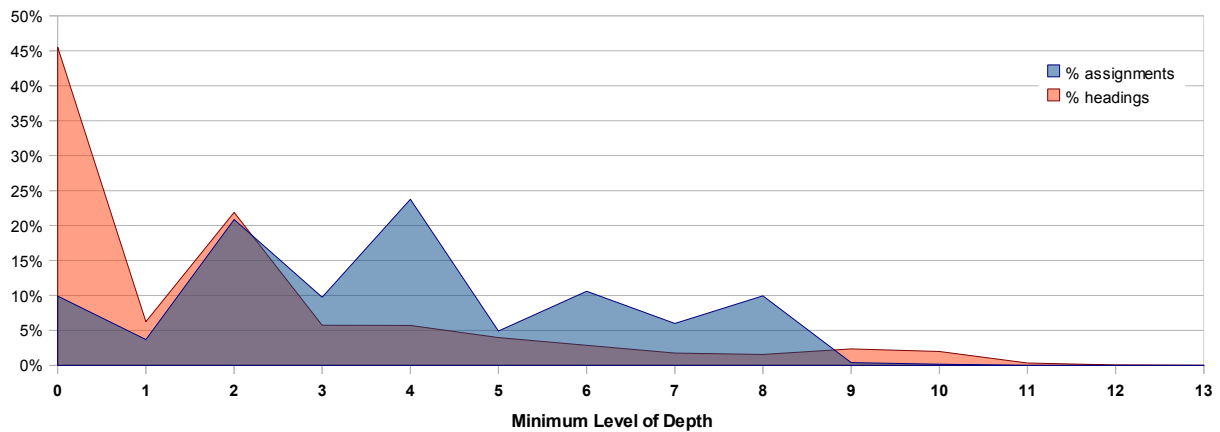


Figure 3: The level of depth assigned to the PHM collection items is generally higher than expected from the LCSH level of depth distribution.

random sample. Some cases of polysemy were expected and searched for, but not found. This can be attributed to the exclusion of semantically ambivalent concepts in the LCSH. The third analysis provided statistics regarding the level of depth of the reconciled headings. Table 3 and Figure 3 clearly demonstrate that the majority of reconciled headings have a minimum level of depth between 2 and 6. As the relevance and the discriminatory value of a heading is context-dependent, we can not propose a deterministic evaluation of the possible role of the reconciled heading for search and retrieval. However, the combination of the above mentioned three analyses, and especially the calculation of the minimum level of depth, make it clear that the reconciled headings are not limited to general concepts which describe an object type, such as for example “Photographs”, but also comprise highly specific concepts, ranging from single-term headings such as “Flatirons” to multi-term keywords such as “Chocolate molds”.

The positive outcomes regarding our two research questions immediately lead us to a next question: once the reconciliation process is mastered and understood, how can we exploit the automatically created interconnections between metadata and vocabularies and achieve clear benefits for end-users and collection holders? As objects from different datasets are interlinked, they can be recommended as additional resources. Future work will focus on how browser plug-ins can automatically display and recommend linked resources from other collections to end-users. Collection holders can also provide a higher recall in users searches thanks to the gathering of alt-labels, and leverage the visibility of their resources since crawlers rely on links. The application of the Linked Data principles also forces us as metadata practitioners to re-evaluate our metadata and vocabularies, which need to be modified in an iterative process to facilitate the reconciliation, thereby showing similarities with hermeneutical practices, in which a constant process of going back and forth allows a deeper understanding of the studied phenomenon.

However, we should be wary of taking the outcomes of the linking of vast sets of freely available data at face value. Lawrence Lessig demonstrated in his essay *“Against transparency: the perils of openness in government”* (Lessig, 2009) that a lack of contextual understanding of politics can lead to overhasty conclusions based upon a direct mapping between political spendings and the decision making process of US senators. We can also apply Lessig’s warning to the cultural heritage context: an exact match between two strings of characters does not necessarily provides a deeper understanding of the cultural resource for the end-user. We should be careful not to promote the atomization of the metadata of our cultural institutions too much at the expense of lengthy, unstructured and costly narratives but which are essential for an in-depth understanding of complex cultural resources. The tension between the database and the narrative as competing models to convey information has been eloquently described in the work of Lev Manovich: “As a cultural form, the database represents the world as a list of items, and it refuses to order this list. In contrast, a narrative creates a cause-and-effect trajectory of seemingly unordered items (events). Therefore, database and narratives are natural enemies. Competing for the same territory of human culture, each claims an exclusive right to make meaning out of the world” (Manovich, 2001, p. 225). We should therefore keep the possibilities but also the limitations of both models in mind when elaborating future metadata management strategies.

## Acknowledgements

The authors would like to thank the Powerhouse Museum for making their metadata freely available and therefore allowing us to perform this case-study on the basis of metadata which can be freely used under the CCASA license. Muriel Foulonneau (Knowledge Intensive Systems and Services, Henri Tudor Research Centre Luxembourg), Rick Block (Metadata librarian at Seattle University) and David Miller (Head of technical services, Levin Library, Curry College Massachusetts) provided useful feedback and comments.

The research activities as described in this paper were partly funded by Ghent University, the Interdisciplinary Institute for Broadband Technology (IBBT), the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research Flanders (FWO Flanders), and the European Union.

## References

- M. Alistair, B. Matthews, D. Beckett, D. Brickley, M. Wilson, and N. Rogers. SKOS: A language to describe simple knowledge structures for the web. 2005. URL <http://epubs.cclrc.ac.uk/bitstream/685/SKOS-XTech2005.pdf>.
- David Bade. The perfect bibliographic record: Platonic ideal, rhetorical strategy or nonsense? *Cataloging & Classification Quarterly*, 46(1):109–133, 2009.
- Tim Berners-Lee. Linked Data, 2006. URL <http://www.w3.org/DesignIssues/LinkedData.html>.
- Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data – the story so far. *International Journal On Semantic Web and Information Systems*, 5(3):1–22, 2009.
- Isabelle Boydens. *Practical Studies in E-Government : Best Practices from Around the World*, chapter Strategic Issues Relating to Data Quality for E-government: Learning from an Approach Adopted in Belgium, pages 113–130. Springer, 2011.
- Dan Brickley and Ramanathan V. Guha. Rdf vocabulary description language 1.0: Rdf schema. Technical report, 2004. URL <http://www.w3.org/TR/2004/REC-rdf-schema-20040210/>.
- Vanda Broughton. *Essential classification*. Facet Publishing, 2004.
- Bernhard Haslhofer and Antoine Isaac. data.europeana.eu – The Europeana Linked Open Data Pilot. In *Proc. of Int. Conf. on Dublin Core and Metadata Applications*, 2011.
- Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 1st edition, 2011. ISBN 9781608454303. URL <http://linkeddatabook.com/>.
- Diane Hillmann and Jon Phipps. Application profiles: Exposing and enforcing metadata quality. In *International Conference on Dublin Core and Metadata Applications*, pages 53–62, 2007.
- Pascal Hitzler, Markus Krötzsch, Bijan Parsia, Peter F. Patel-Schneider, and Sebastian Rudolph. OWL 2 Web Ontology Language Primer. W3C Recommendation, World Wide Web Consortium, October 2009. URL <http://www.w3.org/TR/owl2-primer/>.
- Antoine Isaac, Stefan Schlobach, Henk Matthezing, and Claus Zinn. Integrated access to cultural heritage resources through representation and alignment of controlled vocabularies. *Library Review*, 57(3):187 – 199, 2008. URL [www.emeraldinsight.com/10.1108/00242530810865475](http://www.emeraldinsight.com/10.1108/00242530810865475).
- Patrice Landry. Providing multilingual subject access through linking of subject heading languages: The macs approach. In Raffaella Bernardi and Sally Chamers, editors, *Proceedings of the workshop on advanced technologies for digital libraries*, pages 34–37, 2009.
- Lawrence Lessig. Against transparency : the perils of openness in government. *The New Republic*, October 2009, 2009.
- Eetu Mäkelä, Osmo Suominen, and Eero Hyvönen. Automatic exhibition generation based on semantic cultural content. In Lora Aroyo, Eero Hyvönen, and Jacco van Ossenbruggen, editors, *Cultural Heritage on the Semantic Web*, Bexco, Busan, Korea, November 2007. Workshop Proceedings of the 6th International Semantic Web Conference (ISWC) and 2nd Asian Semantic Web Conference (ASWC). 12. November 2007.

- Lev Manovich. *The language of new media*. MIT press, 2001.
- Alistair Miles and Sean Bechhofer. SKOS Simple Knowledge Organization System. W3C Recommendation, 2009. URL <http://www.w3.org/TR/skos-reference/>.
- Joachim Neubert. Bringing the “thesaurus for economics” on to the web of linked data. In *Proceedings of the Linked Data on the Web Workshop (LDOW2009), Madrid, Spain, April 20, 2009, CEUR Workshop Proceedings*, volume 538, 2009. URL [http://ceur-ws.org/Vol-538/ldow2009\\_paper7.pdf](http://ceur-ws.org/Vol-538/ldow2009_paper7.pdf). LDOW2009, April 20, 2009, Madrid, Spain.
- Jack Olson. *Data quality: the accuracy dimension*. Morgan Kaufmann, 2003.
- E.T. O’Neill and L.M. Chan. Fast (faceted application of subject terminology): a simplified vocabulary based on the library of congress subject headings. *IFLA Journal*, 29(4):336–442, 2003.
- Juan-Antonio Pastor-Sanchez, Francisco Javier Martínez Mendez, and José Vicente Rodríguez-Muñoz. Advantages of thesauri representation with the simple knowledge organization system (SKOS) compared with other proposed alternatives for the design of a web-based thesauri management system. *Information Research*, 14(4), 2009. ISSN 1368-1613. URL <http://informationr.net/ir/14-4/paper422.html>.
- Eric Prud’hommeaux and Andy Seaborne. SPARQL Query Language for RDF. W3C recommendation, W3C, 2008. Published online on January 15th, 2008 at <http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/>.
- Geoffrey Rockwell. There’s a toy in my essay: Problems with the rhetoric of text analysis. *Draft article*, 2011.
- Ed Summers, Antoine Isaac, Clay Redding, and Dan Krech. Lcsh, skos and linked data. In *Proceedings of the International Conference on Dublin Core and Metadata Applications (DC 2008)*, pages 25–33, 2008.
- Elaine Svenonius. LCSH: Semantics, syntax and specificity. *Cataloging & Classification Quarterly*, 29(1-2):17–30, 2000. doi: 10.1300/J104v29n01\_02.
- Douglas Tudhope, Ceri Binding, Stuard jeffrey, Keith May, and Andreas Vlachidis. A stellar role for knowledge organization systems in digital archaeology. *Bulletin of the American Society for Information Science and Technology*, 37(4):15–18, April/May 2011.
- Lourens van der Meij, Antoine Isaac, and Claus Zinn. A web-based repository service for vocabularies and alignments in the cultural heritage domain. In *Proceedings of the 7th European Semantic Web Conference (ESWC)*, volume 6088, pages 394–409, 2010.
- Marieke van Erp, Johan Oomen, Roxane Segers, Chiel van den Akker, Lora Aroyo, Geertje Jacobs, Susan Legêne, Lourens van der Meij, Jacco van Osssenbruggen, and Guus Schreiber. Automatic heritage metadata enrichment with historic events. In J. Trant and D. Bearman, editors, *Museums and the Web 2011: Proceedings*. Archives & Museum Informatics, Toronto, 2011.
- Seth van Hooland, Seth Kaufman, and Yves Bontemps. Answering the call for more accountability: applying data-profiling to museum metadata. In *International conference on Dublin Core and metadata applications, 22-26 september 2008*, pages 93–103, 2008.
- Seth van Hooland, Eva Mendez, and Françoise Vandooren. Opportunities and risks for libraries in applying for european funding. *The Electronic Library*, 29(1):90–104, 2010.
- Seth van Hooland, Eva Mendez, and Isabelle Boydens. Between commodification and sense-making. on the double-sided effect of user-generated metadata within the cultural heritage sector. *Library Trends*, 59(4):707–720, 2011.
- Kwan Yi and Lois Mai Chan. Linking folksonomy to library of congress subject headings: an exploratory study. *Journal of Documentation*, 65(6):872–900, 2009. ISSN 0022-0418. doi: 10.1108/00220410910998906. URL <http://www.emeraldinsight.com/journals.htm?articleid=1823651&show=html>.