Archival Resource Keys for Collaborative Historical Ontology Publication

Mat Kelly[1,*], *Christopher B* Rauch[1], *John* Kunze[2], *Sam* Grabus[1], *Joan* Boone[1], *Peter M* Logan[3], and *Jane* Greenberg[1]

[1]College of Computing and Informatics, Drexel University, Philadelphia, USA
[2]California Digital Library, University of California, Oakland, USA
[3]Temple Libraries, Temple University, Philadelphia, USA

Abstract. This case study builds on previous research transforming the 1910 *Library of Congress Subject Headings* (1910 LCSH) into a computation-ready dataset, compliant with linked data standards. The Archival Resource Key (ARK) system offers persistent identifiers for vocabularies and is becoming more widely adopted. We provide background and motivation for our work with the Nineteenth-Century Knowledge Project (NCKP), review our data sources, and detail the implementation of ARKs. Finally, we review the Helping Interdisciplinary Vocabulary Engineering (HIVE) system and preset the unified workflow, which involved contributions from information science and humanities scholars, as well as computer science and information systems researchers. The conclusion provides insight into planned future work.

1 Introduction

The development of linked data has inspired large-scale digital archival services to publish their controlled vocabularies online. Many governmental research agencies, foundations, and institutions provide access to this metadata in the form of ontologies according to standards, such as the Simple Knowledge Organization System (SKOS) and Owl Web Ontology Language (OWL), with Resource Description Framework (RDF) at the base. This process permits the constituent terms to be represented by unique identifiers. While the availability of current vocabularies in this format is ubiquitous, historical versions seem to be primarily available in digital form as PDFs. There is an opportunity to make vocabularies that temporally align with resources more available as linked data [1]. This has been a goal of our current work with the LCSH 1910, historical editions of the *Encyclopedia Britannica* (EB), and ARKs. The sections that follow provide background information on the NCKP, which motivated this work and discusses the transformation required to support the assignment of persistent identifiers to EB entries representing 1910 LCSH concepts.

2 Background

The NCKP is an initiative supporting the study of the structure of nineteenth-century knowledge and its transformation. The project draws on historical editions of the EB as a model

*e-mail: mrk335@drexel.edu

for the representation of what constituted official knowledge throughout the English-speaking world, as well as ontological resources that provide a semantic representation of the knowledge of the time.

Initial work involved the creation of an accurate Text Encoded Initiative (TEI) dataset for the corpus. Preparation of the dataset is a multistep process that begins with the transformation of various editions of the EB into electronic form, character recognition, and automatic generation of metadata for each of the 113,000 entries. The entries are then tagged with terms drawn from a subject catalog closer in time to the source material (the LCSH 1910) that has been transformed through the processing of an automatic metadata extraction tool, HIVE. These terms are assigned to the Text Encoding Initiative (TEI) record, along with ARK persistent identifiers, preparing the data for computational manipulation and various forms of textual analysis.

3 Data Sources

The data sources of our initiative consist of two digitized resources: 1) selected historical editions of the EB and 2) 1910 LCSH. These two data sources are explained here, followed by an overview of the HIVE technology and our application of ARK persistent identifiers.

3.1 Encyclopedia Brittanica

Four historical editions of the EB (3rd- 1797, 7th- 1842, 9th- 1889, and 11th- 1911) are being digitally scanned for comparison. Scanning errors do arise, and these are generally resolved by a combination of data cleansing and human intervention. Early phase testing of subject metadata assignment, drawing from a contemporary vocabulary, revealed anachronistic results. To address these problems, domain experts worked together with technical staff to create an electronic representation of an ontology that made sense both in its original context and in electronic form. We selected the 1910 LCSH as an initial test.

Further processing resulted in a collection of text documents representing entries in the Encyclopedia. These were prepared in batch form and encoded according to the guidelines of the TEI (Text Encoding Initiative). Generally, TEI guidelines recommend a set of features for textual resources that facilitate machine processing of electronic representations of text [2]. Specifically, tags or semantic markers are applied to organize (or markup) the text of the entries so that it is more easily programmatically processed. Tags are rendered in Extensible Markup Language (XML) to surround and describe the properties of various selections of text. Here, XML serves as a mechanism to impose constraints on the storage layout and logical structure of the text [3].

The machine-readable description provided by XML coding of Encyclopedia entries provides a structure to which automatically generated metadata can be added. Additionally, the TEI body text for each entry can be automatically indexed for entities, such as topic and geographic location, which can then be mapped to the appropriate ontologies.

3.2 1910 Library of Congress Subject Headings

In 1901, the Library of Congress began distributing its cataloging cards to libraries throughout the United States. The distribution of these cards, which had been created according to international rules agreed upon by the American British Library Associations, enabled a cooperative approach to cataloging. By 1910, the cataloging rules represented by the common subject headings had become a national standard, published between 1910 and 1914 as *Subject Headings Used in the Dictionary Catalogues of the Library of Congress* [4]. This makes

the collected subject headings a historically appropriate vocabulary from which to generate metadata that better models the structure of knowledge as it was perceived during the time [5].

When preparing a historical ontology as a conceptual model to apply to the analysis of a body of work, the ontology itself must be examined for consistency and quality in its original context; otherwise, the insight that it might bring to analyzing a contemporaneous work could be distorted in some way. Even ontologies that are faithfully reproduced in electronic form from their print originals are subject to revision over time as perhaps secondary sources surface or inconsistency is identified. Ideally, this work involves both humanities scholars and information communication technologists.

Previously, the original 1910 Library of Congress Subject Headings publication was only available digitally through PDF scans in HathiTrust and Internet Archive. The transformation to a machine-readable digital ontological representation was an initiative led by the Metadata Research Center (MRC), Drexel University, as part of their participation in the NCKP. Character recognition errors and inconsistencies had to be corrected and resolved, as well as other mistakes or misattributions inherent to the complex transformation process. Additional complications included cross-reference and subheading inconsistencies by the Library of Congress cataloger in the original publication. These complexities were assessed on a case-by-case basis to ensure compatibility with electronic ontological use.

Employing the LCSH 1910 as a vocabulary in the automation of metadata generation for the EB is also facilitated by a markup convention, SKOS, and made accessible via the Helping Interdisciplinary Vocabulary Engineering (HIVE) application further reviewed below. SKOS is a World Wide Web Consortium (W3C) specification that provides a standard way to represent knowledge organization in RDF [6]. Such encoding is useful in representing the relationship between two entities- in this case, the terms of the 1910 LCSH, and by extension, the EB entries labeled with those terms.

Figure 1. Text to SKOS transformation of 1910 LCSH

The labeling process is represented in figure 1. The SKOS data model is used to represent subject headings as RDF triples. Subject headings are defined as concepts with a preferred label. See also headings are defined as related concepts. See headings are defined as alternative labels. Descriptive information is represented as scope notes. Each RDF triple consists of a subject, predicate, and object. For example, concept A is related to concept B.

4 Persistent Identifier Assignment (Archival Resource Key)

As the research community moves increasingly toward an open access publication model, research outputs become more widely available but are often distributed across multiple disparate platforms and archives [7]. Persistent identifiers (PIDs) can enable the consolidation and linking of these outputs. For PIDs, we chose ARKs (Archival Resource Keys) because of their low cost and flexibility. Any institution can create an unlimited number of ARKs and assign them to any research object using its own metadata schema. This follows from the structure of the identifier, which generally takes the form: *ark:/NAAN/Name*.

The Name is locally assigned, and the NAAN (Name Assigning Authority Number) usually represents a memory organization, scholarly society, or other type of cultural heritage institution that maintains archives. Each institution receives a numerical NAAN that identifies it uniquely. Because a NAAN is not a widely recognizable institutional brand, the stability of a collection of ARKs is not affected when collection objects and metadata are transferred to a successor institution, as when institutions merge, split, or are renamed.

Authoritativeness tends to follow ownership. For example, the Musée du Louvre recently launched a website making its entire collection available online [8]. The following ARK represents the Mona Lisa by Leonardo da Vinci in the Louvre's collection: *https://collections.louvre.fr/ark:/53355/cl010062370*.

Note the hostname in front of an ARK does not participate in ARK comparisons since hosting arrangements are no more permanent than institutions themselves. An ARK is globally unique without the hostname. While there are many representations of this painting and associated scholarship, the one from the museum that owns it is likely to be the most authoritative.

Sometimes the name assigning authority is indicated not by a simple NAAN but by a "shoulder," which is a NAAN extended with additional characters. For example, the ARK **ark:/99152/b4**2r3nz3s represents the term "hunting," and the shoulder, "ark:/99152/b4", indicates the 1910 LCSH ontology as we have prepared it. This can be useful for special classes of objects where the NAAN (the first part of the shoulder) signals certain immutable properties relevant to archival science. The NAAN 99152 represents controlled vocabulary and ontology terms, such as metadata element names and pick-list values [9]. The shoulder "/b4" is the extension we reserved for our version of the 1910 LCSH ontology.

The natural language string "hunting" may appear in many different ontologies, each with its own ARK, but ark:/99152/b42r3nz3s identifies the English language term as it was used by the Library of Congress in 1910 to represent a 20th-century concept within its subject headings. In addition, the ARK represents the efforts of researchers at MRC, Drexel University, who rendered these historical terms into a machine-readable format (SKOS) for semantic web use.

Another advantage of ARKs is that inflections and qualifiers can be added to modify a request to retrieve different forms of the data. They can be used in any common web browser. In principle, content negotiation achieves the same end but requires intermediation with a software agent that exchanges special HTTP protocol headers that both client and server must be configured to understand. Inflections make the entire request transparent by allowing users to modify the ending of the ARK in the web browser address bar. By publishing an endpoint that will always return the latest version of the ontology and exposing the services it can provide, we can return data in a variety of formats. For example, the University of North Texas Digital Libraries uses a qualifier to indicate a specific metadata scheme [10] such as dc.xml, dc.json, etc., or another type of domain-specific transformation- returning the data with, for example, instances of the em dash (—) and sharp s (ß) normalized or preserved.

From a practical perspective, having persistent but flexible identifiers assigned from the outset allows collaboration to begin sooner and on more solid ground. When designing algorithms to apply to a body of work using a particular ontology, having a well-established naming scheme for templating improves the availability of data for reuse. Since all identifiers would be resolved in the manner of our choosing, we could not only reference the ontology itself using the ARK but also deploy services against it.

5 Helping Interdisciplinary Vocabulary Engineering (HIVE) and Automatic Metadata Extraction

Because of the extensive size of the data sources, manual metadata assignment was not possible. HIVE is a linked data vocabulary server application that supports a workflow for automatically generating metadata for textual data, drawing terms from multiple controlled vocabularies [11].

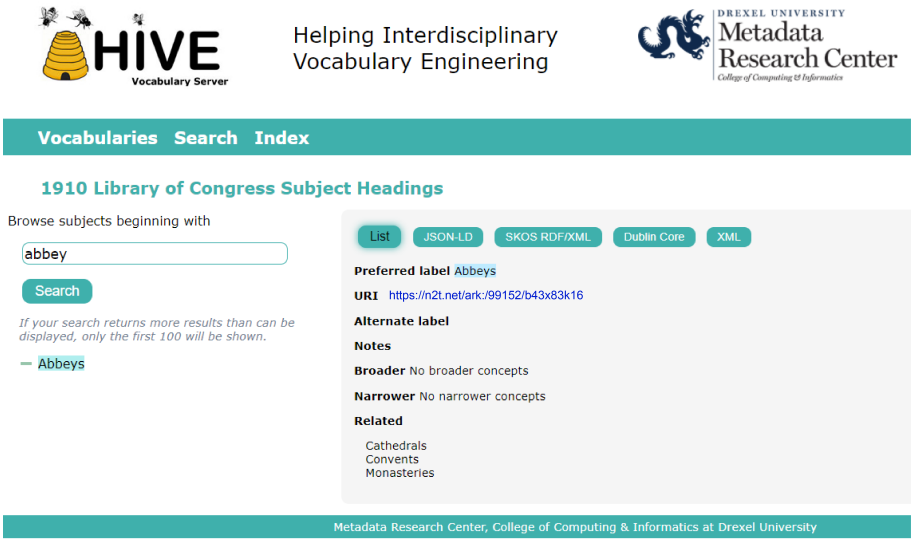

Figure 2. Fig 2: A 1910 LCSH entry after it has been processed by HIVE for automated metadata extraction.

HIVE uses the Rapid Automatic Keyword Extraction algorithm (RAKE), an unsupervised keyword extraction method [12]. For this case study, the HIVE workflow applies the RAKE algorithm to the digitized versions of the EB entries to extract candidate keywords. These are normalized and mapped to controlled vocabularies. Keywords (or phrases) that match vocabulary terms are selected as metadata for each EB entry and are made available in a computation-ready dataset. This dataset enables the final processing step where metadata is added to the EB entry.

6 Conclusion and Future Work

In order to better understand the state of knowledge as it existed in past centuries, we have transformed the printed form of the 1910 Library of Congress Subject headings into more of

an ontological structure, using SKOS. This transformation supports automatic indexing and allows us to represent the structure of knowledge, including relationships, closer to the time of the creation of the resource. Through automatic metadata extraction and tag assignment, we have labeled the data to facilitate computational processing, text analysis, and linked data. The use of ARK persistent identifiers was important to our process. We were able to create our own persistent identifiers as we needed them to represent unique digital objects either temporarily or persistently. ARKs also allow us to stably describe relationships among data not only in institutional repositories but also open data on the web. In future work, we expect to assign ARKs to EB entries and provide case studies in linked data practices facilitated by persistent identifiers.

7 Acknowledgements

Research is supported in part from NEH-HAA-261228-18 and MRC/Drexel University research funds.

References

[1] Kelly, M., Greenberg, J., Rauch, C. B., Grabus, S., Boone, J. P., Kunze, J. A., & Logan, P. M. (2020). A Computational Approach to Historical Ontologies. *2020 IEEE International Conference on Big Data*, **1878–1883**. https://doi.org/10.1109/BigData50022.2020.9378268

[2] TEI Consortium. (2021). TEI P5: *Guidelines for Electronic Text Encoding and Interchange*. https://doi.org/10.5281/ZENODO.3413524

[3] *Extensible Markup Language (XML) 1.0 (Fifth Edition)*. (n.d.). Retrieved April 22, 2021, from https://www.w3.org/TR/xml/#sec-common-syn

[4] Yee, M. M. (2011). "Wholly Visionary." *Library Resources & Technical Services*, **53(2)**, 68–78. https://doi.org/10.5860/lrts.53n2.68

[5] Foskett, A.C. (1996). *The Subject Approach to Information*. Facet Publishing.

[6] *RDF - Semantic Web Standards*. (n.d.). Retrieved April 22, 2021, from https://www.w3.org/RDF/

[7] Persistent Identifiers Connect a Scholarly Record with Many Versions. (2021, February 18). *Association of Research Libraries*. https://www.arl.org/blog/persistent-identifiers-connect-a-scholarly-record-with-many-versions/

[8] The Musée du Louvre Launches Online Collection Database and New Website. (2021, March 25). *Espace Presse du Musée du Louvre*. https://presse.louvre.fr/le-musee-du-louvre-met-en-ligne-ses-collections-et-devoile-son-nouveau-site-internet-3/

[9] *ARK Alliance*. (n.d.). Retrieved April 23, 2021, from https://arks.org/

[10] Phillips, M. E. (2008, June 5). *Using Archival Resource Keys (ARKs) for Persistent Identification* [Presentation]. Texas Conference on Digital Libraries (TCDL), 2008, Austin, Texas, United States. https://digital.library.unt.edu/ark:/67531/metadc28359/

[11] Greenberg, J., Losee, R., Agüera, J. R. P., Scherle, R., White, H., & Willis, C. (2011). HIVE: Helping interdisciplinary vocabulary engineering. *Bulletin of the American Society for Information Science and Technology*, **37(4)**, 23–26. https://doi.org/10.1002/bult.2011.1720370407

[12] Huang, H., Wang, X., & Wang, H. (2020). NER-RAKE: An Improved Rapid Automatic Keyword Extraction Method for Scientific Literatures Based on Named Entity Recognition. *Proceedings of the Association for Information Science and Technology*, **57(1)**, e374. https://doi.org/10.1002/pra2.374