JLIS.it

Library of Congress Classification as linked data

Kevin Ford

What is linked data

The Library of Congress has published a select number of classes from the Library of Congress Classification (LCC) system as linked data as a new offering of its Linked Data Service,[1] commonly known as id.loc.gov. The offering, while still considered a beta project, provides URIs for resources that represent a simplified version of the underlying data found in the source MARC Classification records. The beta service also furnishes URIs for classification number resources that either derive directly from the underlying data or are the result of a synthesis between a schedule resource and a table resource. Although the data are presented in MADS/RDF[2] and SKOS[3] where appropriate, LCC as linked data is accompanied by a small LCC ontology to more accurately describe the types of classification resources and the relationships between them, especially where MADS/RDF and SKOS Class and Property definitions were seen as insufficient. This paper explores the publication of LCC as linked data and the accompanying ontology by contextualizing them with

[1] http://id.loc.gov.
[2] http://www.loc.gov/mads/rdf.
[3] http://www.w3.org/2004/02/skos.

JLIS.it. Vol. 4, n. 1 (Gennaio/January 2013).
DOI: 10.4403/jlis.it-5465

respect to prior efforts representing LCC as linked data, representing Dewey as linked data, and the appropriateness of SKOS for library classification data, especially given the historical need for a distinct MARC format for Classification.

The Library of Congress classification system has existed since the late nineteenth century "to organize and arrange the book collections of the Library of Congress" (*Library of Congress Classification*). The system is organized into twenty-one classes, most of which are further divided into subclasses. Each class represents a field of knowledge, such as Art, Law, or History. Each subclass is further divided into more specific topics that basically adhere to a hierarchical representation of the field of knowledge. Like most classification systems, LCC is subject-based. The resulting "number", therefore, represents a distinct topic within the field of knowledge. For decades LCC has been printed, bound, and distributed (at cost, basically) and still is today. One may acquire, for a price, the entire 41-volume set or one may choose individual classes or schedules. LCC is also accessible via ClassificationWeb,[4] which is a sophisticated web application designed to assist catalogers with the assignment and creation of LCC classification numbers. It is offered as a subscription service for which LC charges a fee. Also for cost (basically), the Library of Congress Classification is available in MARC21 format and is made available as a bulk download, with periodic updates, from the Library's Cataloging and Distribution Service. Notably, the raw data, though available, requires purchase and is not presented in accordance with linked data methods and principles.

The Library of Congress Classification as linked data does have a history, albeit a short and little known one. Karen Coyle laboriously scraped the first four levels (more or less) of all LC Classification classes from PDF documents hosted on the LC website to a plain

[4] https://classificationweb.net.

text file (that is, something far more accessible for machines) and uploaded the resulting text file to archive.org.[5] This work dates to, and therefore the data predates, September 2007.[6] The PDF documents, which are still available (though perhaps updated since), present a detailed outline of LCC. Ed Summers then took the text file, generated a basic SKOS RDF representation from it, and developed a very simple website where he published the SKOS data.[7] This work was little publicized, but it is still active and accessible. Summers's code is on GitHub.[8]

Coyle's text file simply lists the classes (A, B, C, and so on) and the first three levels, if appropriate, of each subclass (AC, AE, AG, and so on). The concept's label at any given level is matched with the class number. Because only the first few levels of LCC are outlined, most classification numbers represent a range of more specific topics. Missing – nearly universally – from the detailed outline are language-specific divisions within topics, temporal divisions within topics, and form divisions within topics, in addition to simply greater granularity and specificity, such as the distinction between "General works" and "Special topics." From Coyle's text file, Summers generated a skos:Concept Resource for each classification number and associated label. He took each classification number and appended it to a base HTTP URI (in a namespace he controls) to create an unique identifier for the resource and he made the lexical label for the topic (and class number) the skos:prefLabel. He generated skos:broader and skos:narrower relationships between classification topics when the classification number represented an encompassing range or a more specific range respectively. Summers created something akin

[5] http://ia600304.us.archive.org/0/items/LcClassificationA-z/lc_class.txt.

[6] http://ia600304.us.archive.org/0/items/LcClassificationA-z/LcClassificationA-z_meta.xml.

[7] http://inkdroid.org/lcco.

[8] https://github.com/edsu/lcco.

to an LCSH-like pre-coordinated heading with the labels of narrower topics (i.e. those that fit contextually with broader topics): the skos:prefLabel of narrower topics contains the labels of its broader relations, the labels of which are separated by two hyphens. The data collected by Coyle, which may have been all that was reasonably possible to collect, were limited to a class number, label, and hierarchy. The first three levels of the Dewey Decimal Classification system – the Dewey Summaries – have been available as linked data since 2009.[9] OCLC published the full Dewey Decimal Classification as linked data in Summer 2012. As with Summers's design, each topic is a skos:Concept with broader or narrower relations to any given topic's hierarchical relatives. Published as it was by OCLC, the available data are richer, including information about provenance and licensing (no fewer than four statements for each Concept), creation and modification times, among a few others. Unlike Summers's design, OCLC reserved the skos:prefLabel exclusively for the lexical label of the given Concept – broader relations are not strung together with the topic's label to create the skos:prefLabel. OCLC's URI design patterns warrant special mention. Pains have been taken to embed some semantics into the URI pattern, reserving, essentially, one namespace each for "non-information resources (abstract or concrete real-world objects), generic resources, and their representations" (OCLC). Although some of the URI examples do not appear to function presently, the focus on URI composition and the need to represent a variety of different resource types bears on the representation of all aspects of publishing classification systems such as DDC and LCC as linked data.[10] A diverse number of resource types are also very relevant to LCC. In addition to the embedded semantics in the Dewey URIs, this issue received greater elucidation

[9] http://dewey.info.

[10] The actual service at http://dewey.info features diverse URI patterns, all of which appear to function, for all types of information resources.

by Panzer and Zeng in two related publications (Panzer and Zeng; Zeng, Panzer, and Salaba).

The authors explored how to model classification schemes (notably DDC) in SKOS. Among other findings, the authors discuss how classification systems include "assignable" and "non-assignable" concepts. In DDC, an example of a non-assignable concept is a centered entry, or a classification number range or span for which there are likely a number of more specific topics and, therefore, specific numbers. In LCC, this is referred to as a range. There is also the issue, as Panzer and Zeng note (2009), of synthesized concepts (a classification number and topic that are a result of combining two concepts in the classification system) and non-synthesized concepts. One risks some semantic incoherency when attempting to model all these types of things, and to establish appropriate relationships between them, purely in SKOS. Panzer and Zeng considered the need to create, minimally, an extension to the core SKOS vocabulary, but it was clear that an altogether separate attempt might be necessary, in a namespace entirely distinct from a SKOS one, to correctly capture the semantics and relationships. These same issues also materialized during the process of trying to represent LCC in SKOS.

SKOS – the Simple Knowledge Organization System – is designed "to support the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading lists and taxonomies within the framework of the Semantic Web".[11] SKOS has proven to be extremely versatile and effective at representing thesauri, subject heading lists, and taxonomies (though, in part as a result of being intentionally simple, there can be some loss of granularity with respect to library data). In fact, data represented using the MARC Format for Authority data, such as subject heading lists like LCSH, map effortlessly to SKOS. This is seen readily and simply

[11] http://www.w3.org/2004/02/skos.

when decomposing a MARC Authority record into MADS/RDF and SKOS. For MARC Authority, a valid (i.e. not deprecated) authority record is the Concept. The 1XX - the main heading - becomes the authoritative or preferred label. MADS/RDF provides a means to capture the type of concept, be it a Topic, Geographic, GenreForm, or Temporal notion, and a few others. MADS/RDF also provides support for better representation of pre-coordinated headings. MARC Authority 4XX fields are variant or alternate labels. 5XX fields represent various relationships between terms, of which broader and narrower relationships are the most popular. MADS/RDF added a few additional relationships, such as those needed to accurately record connections between earlier or later established concepts, and a new resource type to clearly denote deprecated resources. A number of note fields defined in MARC Authority also have one-to-one mappings to MADS/RDF and SKOS. But MADS/RDF and SKOS classes and properties have been far less amenable to classification data, or at least to library-specific classification systems such as DDC and LCC.[12] This is essentially the difficulty Panzer and Zeng encountered during their research and it is the same encountered when attempting to publish LCC as linked data. At least when it comes to library classification systems such as DDC and LCC, this is unsurprising.

The influential consideration here lies with the MARC21 format for Classification.[13] More specifically, its very existence. Formally but provisionally published in June 1990, the MARC21 Format for

[12]This probably has to do a lot to do with the relative complexity of classification systems, especially with respect to how classification numbers are constructed, when compared to thesauri or "subject heading lists;" the aggregate expertise of the SKOS designers and members of the working group with respect to classification systems; and, partly as a natural extension of the previous point, a certain amount of partiality and attention given to, and in favor of, thesauri and "subject heading lists" during the development of SKOS.

[13]http://www.loc.gov/marc/classification.

Classification was specifically developed to facilitate the exchange and printing of classification data, most notably LCC and DDC (Guenther). Importantly, the new MARC format was, however, the result of an attempt to modify the MARC format for Authority data (this work started in 1987/1988). After identifying most of the changes that would be required of the MARC Authority format, a draft of the proposed changes was presented to the committee overseeing changes to the MARC formats (MARBI). Following this review, and the early development period generally, it was clear that "there was less overlap with the authority format than originally anticipated, and ... [the MARC Authority] codes and conventions were too constraining" (Guenther). The proposal for classification data was rewritten to be a separate format, which would become the MARC Format for Classification by 1991.

The MARC Format for Classification – and its development process – took into consideration the very same semantic difficulties encountered by Panzer and Zeng, and the present author, when faced with "skosifying" complex library classification data, and a difficulty that is compounded by the unsuitable nature of the RDF data element semantics. The MARC Format for Classification can represent class schedules and tables, neither of which is necessarily assignable as is. The format can represent ranges and hierarchy. Naturally, it has full support for notes and index terms. But SKOS semantics are not rich enough this type of information. That said, SKOS can reasonably represent (assignable) classification topics and even class number ranges. It is with this information in mind, and the background work by Panzer and Zeng, that it was decided to present LCC as linked data as much as possible in MADS/RDF and SKOS but to define a small vocabulary in OWL to faithfully represent LCC-specific data and data elements where MADS/RDF and SKOS fall short.[14]

[14] http://id.loc.gov/ontologies/lcc.

Although there are a few ontological constraints on the data, constraints do not presently extend to how the data are used. For example, while it could be possible to infer "assignable" versus "non-assignable" resources from the intersection of select Classes in the ontology, this type of modeling has not been undertaken. As such, it is an experimental offering that attempts to make no semantic restrictions on its use but which strives to represent the derived and underlying data accurately. The ontology is also specific to LCC; it makes no attempt to model data elements specific to other classification systems, such as DDC. Also, though it would be unwise to rule out OCLC developing an ontology for DDC, the explicit declaration of classes in the small LCC ontology transfers the semantics embedded in dewey.info URIs to the data itself. ("Smart" URIs and clear data semantics are not mutually exclusive and could, in fact, be complementary.) A select number of Library of Congress Classification classes are available from LC's linked data Service,[15] commonly known as id.loc.gov.[16] This offering - at the time of this publication - is very much a beta offering. During this stage, the data and its representation are subject to change, especially as more is learned about how the data is used and better ways for it to be represented are determined or developed. Nevertheless, it is an attempt not only to publish an RDF representation of the underlying data used to construct classification numbers but also to publish the classification numbers themselves. To this end, an effort has been made to apply the tables to schedules, thereby synthesizing a classification number, as appropriate.

In order not to become too mired in MADS/RDF [17] and SKOS [18] semantics and restrictions, everything is a MADS/RDF Authority

[15] http://id.loc.gov.
[16] http://id.loc.gov/ontologies/lcc.html.
[17] http://www.loc.gov/mads/rdf.
[18] http://www.w3.org/2004/02/skos.

and SKOS Concept, with the exception of Index Terms, which can be interpreted as variants. They are therefore instantiated as MADS/RDF Variants and SKOS/XL Alternate Labels. The authoritative label - the preferred label and the tightly controlled term - is reserved for the main caption or term. This is therefore similar to how OCLC created Dewey resources and a departure from how Summers presented the data. The full lexically represented hierarchy that one finds in the source MARC records is recorded simply as an rdfs:label so that it is still available for parsing and potentially for display purposes. The classes and properties in the LCC ontology, therefore, are the real carriers of distinction between Library of Congress Classification resources published at id.loc.gov.[19] The LCC ontology provides a way to describe the "underlying data," which is a reference to the data one would find in a MARC classification record. Data in the MARC classification record include information about classification-specific resource types such as tables and schedules, and data describe details about how to apply table numbers to base numbers to generate and assignable classification number. As such, the LCC ontology defines Classes and Properties sufficient enough to accurately represent LCC data in RDF and sufficient enough to synthesize class numbers from schedules when and however appropriate. The ontology is a significant simplification of the MARC Classification codes, data element definitions, and conventions. One such simplification touches on the identification of different types of ranges defined in MARC Classification. Because there appears to be no meaningful distinction between a MARC Summary Range and MARC Defined Range with respect to their representation in RDF, specifically for LCC, these types are simply an LCC Range. On the

[19] I have endeavored to capitalize the word "Class" (and Property) when referring to an OWL or RDF Class (or Property). Whenever referencing an entity associated directly with LCC - such as classification number, LCC class, class schedule, or class number - I have presented the word in all lowercase letters.

other hand, it was deemed necessary to define an additional Table type - a Guide Table - where the MARC Classification format made no clear distinction between the two. A Guide Table is hierarchically the broadest table concept and carries the Table Rule, which is the instruction needed to synthesize a classification number between an LCC Schedule and an LCC Table. The small LCC ontology includes Classes for a Schedule, Range, Table, Guide Table, and Table Rule, all of which are types of resources that are somewhat unique to classification schemes. Additionally, classification-specific properties have been defined that relate these classes to each other, such as one that relates a Table to its Guide Table or another that relates a Guide Table to one or more Schedules, to which the Guide Table may apply. At all other times, MADS/RDF, which is fully mapped to SKOS, is employed (all data are, of course, also outputted as SKOS). Naturally, these Table, Guide Table, and Schedule resources are "underlying data" and are generally considered to be "non-assignable," that is they are resources that should not be used to describe another resource, such as a bibliographic one. Because these resources often have a one-to-one relationship with an underlying MARC Classification record, the LCCN of the underlying record has been used as part of the URI scheme. An LCCN that begins with "CF" represents a schedule; one that begins "CT" represents a Guide Table or Table. However, when classification resources are described with the ClassNumber OWL Class, the resource could be described as assignable. The URIs for these resources end in a classification number or range.

A ClassNumber resource may be an LCC Range or a MADS/RDF Topic. The former - an LCC Range - generally represents a group of concepts hierarchically related to the broader concept represented by the range. Of course, ranges are not assignable when traditionally assigning classification numbers to physical bibliographic resources. MADS/RDF Topic was used when the resource represented a single,

`lcc:GuideTable/lcc:Table`	http://id.loc.gov/authorities/classification/ct96152584
`lcc:Schedule`	http://id.loc.gov/authorities/classification/cf94051344
`lcc:ClassNumber`	http://id.loc.gov/authorities/classification/ND1360-ND1360.6

Table 1: Table showing example URIs based on different LCC types. Note how LCCN is last token of URI in the first two examples versus the classification number range in the last example.

distinct concept.[20]

MADS/RDF and SKOS broader and narrower relationships were asserted between all concepts whether they represented non-assignable underlying data or assignable classification numbers and ranges. However, broader and narrower relationships are expressed between concepts based on whether they represent underlying data (schedules, tables, and guide tables) or classification numbers. Schedules link to tables, guide tables, or other schedules for example; classification numbers link to other classification numbers. For example, an LCC Schedule or LCC Table, both of which are considered non-assignable resources and represent underlying data, may record broader or narrower relationships to other LCC Schedules or LCC Tables respectively, but will not carry such a relationship to an LCC Class Number. That said, there are defined relationships in the LCC ontology created expressly to accurately capture the relationship between underlying data resources, such as an LCC Table, and an LCC Class Number. For example, lcc:isynthesizedFromTable and lcc:synthesizedFromSchedule records from which LCC Schedule or LCC Table the LCC Class Number derives.

The LCC ontology has helped considerably in maintaining a separation of concerns and avoiding the pitfalls of representing this information purely, or at least mainly, in SKOS. Additionally,

[20] http://id.loc.gov/authorities/classification/B4877.S4.html.

because the non-assignable or underlying data has also received representation in RDF, it is possible for others to experiment with this information. In fact, it is known beyond any doubt that the representation of LCC Table resources as tables, LCC Guide Tables as guide tables, LCC Schedules as schedules, and the inclusion of Table Rules in RDF is sufficient to derive and synthesize classification numbers from these resources. The creation of classification numbers and resources, as seen at id.loc.gov, is the result of programming manipulation of LCC Schedule and LCC Table numbers (when tables were required and as part of the process of applying the table rules) and smart querying of the LCC Table data in RDF loaded into a triplestore. Ultimately, focus to date has been almost entirely on the accurate generation of classification numbers from LCC Schedules and, when required, LCC Tables. The MARC Classification records contain numerous ways to link one classification schedule or range to another, often in a separate class altogether. No attempt has been made to extract this information and establish the relationship between the two concepts in the data. Where MADS/RDF or SKOS relationship properties are insufficient, it is anticipated that new LCC properties will be created. Additionally, the data, as presently available, represents a snapshot of any given class - no updates or changes to those classes have been taken into consideration. The types of changes classification numbers undergo, how those changes are recorded in the data, and how changes may or may not affect the RDF representation of LCC remain open, and as yet unexplored, issues. Given that the use of MADS/RDF provides a means to indicate the type of concept - here everything is a MADS/RDF Topic - one wonders whether it would be possible to identify the type of concept especially at the narrower hierarchical levels where the concept might be distinctly temporal in nature (18th century) or a form (General works or Cantos) or a specific language (Russian).

Although a little time has been spent linking LCC resources with LCSH and LC Names resources, more work can be done here too. Naturally, linking LCC to Dewey would be a high priority endeavor. There has also been a long-standing desire to use the Library of Congress Classification as an entry point to the bibliographic catalog versus merely a means to locate a book on a shelf (Chan).
Considerable work remains, but it is hoped that this beta offering will energize developers and stimulate additional innovation. In particular, we look forward to learning of new use cases, especially ones that will explore new uses of the data. For our part, we will continue to make entire classes available as time and resources permit. And we will continue to augment the data and accompanying ontology to ensure that the data being offered is as rich as possible and necessary to accurately represent the data and promote new development.

References

Beall, Julianne and Joan S. Mitchell. "History of the Representation of the DDC in the MARC Classification Format". *Cataloguing & Classification* 48.1. (2010): 48–63.

Chan, Lois Mai. "Library of Congress classification as an online retrieval tool. Potentials and limitations". *Information Technology and Libraries* 5.3. (1986): 181–192. (Cit. on p. 173).

Guenther, Rebecca S. "The Development and Implementation of the USMARC Format for Classification Data". *Information Technology and Libraries* 11.2. (1992): 120–131. (Cit. on p. 167).

Library of Congress. *Library of Congress Classification*. http://www.loc.gov/catdir/cpso/lcc.html. (Cit. on p. 162).

OCLC. *Using the API*. 2010. http://oclc.org/developer/documentation/dewey-web-services/using-api. (Cit. on p. 164).

Panzer, Michael and Marcia Lei Zeng. "Modeling Classification Systems in SKOS: Some Challenges and Best-Practice Recommendations". *Semantic interoperability of linked data: Proceedings of the International Conference on Dublin Core and Metadata Applications, Seoul, South Korea*. Ed. S. Oh, S. Sugimoto, and S.A. Sutton. 2009. Seoul: Dublin Core Metadata Initiative and National Library of Korea, 2009. 3-14, http://dcpapers.dublincore.org/index.php/pubs/article/view/974/944. (Cit. on p. 165).

Zeng, Marcia Lei, Michael Panzer, and Athena Salaba. "Expressing Classification Schemes with OWL 2 Web Ontology Language". *Paradigms and conceptual systems in Knowledge Organisation: Proceedings of the Eleventh International ISKO Conference, University of Rome, Italy*. Ed. C. Gnoli and F. Mazzocchi. 2010. 356–362. (Cit. on p. 165).

KEVIN FORD, Library of Congress.
kefo@loc.gov

Ford, K. "Library of Congress Classification as linked data". *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #5465. DOI: 10.4403/jlis.it-5465. Web.

ABSTRACT: In 2009 and in 2011, the Library of Congress made two of its largest authority files –Subject Headings and Names - available as linked data via LC's linked data service, id.loc.gov. Both are offered in MADS/RDF and SKOS. It is LC's objective, in 2012, to publish another of its largest authority files as linked data: LC Classification. However, whereas the source records for Subject Headings and Names are encoded in the MARC Authority format, from which there is a relatively straightforward mapping to MADS/RDF and SKOS, LC Classification records rely on the MARC Classification format. Mapping from LC Classification to MADS/RDF or SKOS has been a little more challenging. For example, records that represent classification ranges, which are not Concepts intended to be assigned, are not easily accommodated in SKOS. This presents additional problems when needing to accurately represent the relationships in RDF for LC Classification. With comparison to the publication of LCSH and Names at id.loc.gov, this paper will examine issues encountered – and how those challenges were addressed – during the conversion of LC Classification to MADS/RDF and SKOS for release as linked data at id.loc.gov.

KEYWORDS: Library linked data; Library of Congress Classification; Ontology; SKOS; MARC21; Authority control

Submitted: 2012-04-25
Accepted: 2012-08-31
Published: 2013-01-15