To make linked data work, the library community needs good data that is structured, unambiguous and published in a format that enables linking with data produced by other communities. Library data also needs to be more about the Things or the people, organizations, places and topics that users care about and that the library community has something to say about. These qualities are the keys to integrating libraries into the web, where users are now most likely to begin their quests for information.

Such conventions are already at work in Google’s production of Knowledge Cards, which integrate information mined from billions of web documents to produce simple and actionable displays about real-world Things or entities that underlie a search request issued in a particular language. For example, a search for “Chicago” returns the display shown in Figure 1. As a city, Chicago has a skyline and a map. It has points of interest such as parks, museums and universities. These entities have locations of their own. They may have hours of operation or host events for which tickets can be purchased. With
this display, the user can get stuff done that would be much more difficult to accomplish if the search had returned a list of documents instead.

Unfortunately, the library perspective is not well represented in Google’s Knowledge Cards, perhaps because much of our data is still confined to silos and is not comprehensible to the web at large. But Knowledge Cards offer both a glimpse of linked data’s promise and a warning that the library community has its work cut out to realize it.

In the rest of this article, we illustrate some milestones on the path from MARC to linked data by showing how translated works can be represented in a more deliberately defined entity-relationship model that can be expressed using persistent identifiers and revealed in MARC if certain practices are followed. We then discuss the efforts of the Program for Cooperative Cataloging (PCC) to convert recommendations by individual researchers for attaining greater machine understanding into best-practices conventions for the wider library community.

Defining Identifiers for Things

Researchers at OCLC and elsewhere are working to unlock the digital knowledge about creative works and their creators that libraries have been accumulating since the dawn of the computer age and make it available as linked data. Although libraries have been creating descriptions for decades using the MARC (or MAchine-Readable Cataloging) format, the results are intended primarily for human readers and are machine-processable only within library systems, not by third-party data consumers such as Google.

To move forward, data scientists need to make progress on three goals:

1. Define entities and relationships that are important to the library community. Many are commonplace and well understood, such as authors, subjects and publishers of creative works. But they are obscured in MARC and other library standards that rely too heavily on text and the tacit knowledge of human readers.

2. Use the best features of MARC while recognizing that the 51-year-old standard is on the path to obsolescence. Although a MARC description may consist primarily of text, it may also feature encodings that are machine-understandable and unambiguous. These encodings are linked-data gold.

3. Replace text with identifiers that may originate from librarianship but conform to linked data conventions. For example, a library authority file is a collection of records that define or establish names or subject headings and are typically associated with control numbers or record IDs that are usually understood only in local environments. Figure 2 shows a more web-friendly definition of identifier and shows that the Thing, or the real-world referent, is more important than the heading. For example, the internationally important entity known in English as the French National Library or the National Library of France is represented by many other text strings that only humans can read. But when associated with persistent identifiers instead, machine processes can assert that they all refer to the same Thing.

The need for these developments is also recognized by the Program for Cooperative Cataloging (PCC), which was founded in 1994 to promote greater uniformity and reduce cataloging costs by actively managing the evolution of MARC and other library standards. In the process, cataloging best practices are defined and promulgated throughout the library community. When necessary, the PCC also proposes incremental changes that respond to evolutionary pressures of librarianship. In the past two
years, these pressures have initiated defining better migration paths from MARC to linked data. As stated in the PCC 2015-2017 strategic directions document:

Existing methods of library authority control are based on constructing unique authorized access points as text strings (literals). This string-based approach works somewhat well in the closed environment of a traditional library catalog, but not in an open environment where data are shared and linked, and so require unique identifiers. The web presents both a challenge and an opportunity for libraries, which are now in a position to take advantage of data created outside of the library world, and also to contribute library authority data for use by other communities [2, p. 4].

The Virtual International Authority File (VIAF) aggregates identifiers for personal and organizational names, places and works from more than 40 sources around the world into a single, OCLC-hosted authority service. It relies on bibliographic information to determine whether two different text strings represent the same entity or not: if two sources for the same work each list an author, we can be confident that the author is the same even if the text strings differ. Figure 3 shows the various preferred forms of names in different sources linked to the same VIAF identifier 97450170 for a Japanese Nobel Prize winner for literature (romanized as Kawabata Yasunari).

Identifiers are published for entities by communities other than the library community and can be diffused across both domains and languages.

For example, Wikidata creates machine-understandable descriptions for entities harvested from Wikipedia, assigns identifiers to them and associates other identifiers such as those defined in national authority files, VIAF, the International Standard Name Identifier (ISNI), Freebase and GeoNames. Wikipedia pages in multiple languages may embed one or more of these identifiers, thus enhancing library authority file descriptions with photographs, biographies or historical context and expanding the range of names or labels for an entity.

Figure 4 shows how the description of Kawabata is affected. The content of each Wikipedia page differs, but each refers to the same person and includes the same set of identifiers, including the VIAF identifier in Figure 3. The result increases the options for an international audience of readers to enter the linked-data cloud of trusted and authoritative information about a Nobel Prize-winning author.
A Use Case for Identifiers: Multilingual Description

The cream of the world’s cultural and knowledge heritage is shared by being translated. Great works are often translated multiple times. The most important works will more likely have at least one MARC record in the WorldCat database, which is more complete or comprehensive than other catalogs, with more than 350 million records representing holdings of the world’s libraries, of which more than half are in languages other than English. Aggregating all the translations of a given work allows us to leverage the value of the most complete records. For example, as long as one cataloger has supplied a uniform title, we can use it for the entire Work cluster, even if no other cataloger supplied one. We can take advantage of the Cyrillic-Russian records to display the Russian title in Cyrillic on the piece, rather than one of different romanization schemes few Russian speakers could anticipate. Machines can process the identifiers for a work, author and its translators and then present the labels in the preferred language and script of the user (for example, 川端康成 for a Japanese or Chinese reader and Kawabata Yasunari for an English reader).

The relationship of a work (with an author) and its associated translations (with their respective translators) is relatively straight-forward, with each translation linking back to the original work via the property isTranslationOf. A work can have any number of translations, and there can be multiple translations into the same language, which is why identifying the translator is so important. Since it’s been library practice not to include a controlled form or added entry of the translator’s name, we need to parse the statement of responsibility using a long table of “translated” or “translator” in different languages to extract the translator from the MARC record. But once extracted, this information (work, author, translator, language of translation) could be presented in the same kind of Google Knowledge Card presented for “Chicago” at the beginning of this article, with the option to borrow the book from a local library.

Although translations are often described in text, MARC also allows machine-understandable encodings of the most important relationships. The conversion to linked data could be automated if a MARC record for a translated work included the following descriptors:

- The language code of the original and the language code of the translation (and, if appropriate, the language of the intermediate translation)
- Uniform title, in the script of the original
- Added entry for translator
- Roles for each personal entity

A bibliographic description from a MARC record for an English translation of a Chinese work might look like this when mapped to Schema.org:

```html
# Original Work (in Chinese)
<http://worldcat.org/entity/work/id/1215997>
a schema:CreativeWork;
  schema:creator <http://viaf.org/viaf/102266649> ;
  # “Gao, Xingjian”
  schema:inLanguage “zh”;
  schema:name "靈山"@zh-hant.

# Translated Work (in English)
<http://worldcat.org/entity/work/id/145209748>
a schema:CreativeWork;
  schema:creator <http://viaf.org/viaf/102266649> ;
  # “Gao, Xingjian”
  schema:translator <http://viaf.org/viaf/81663420> ;
  # “Lee, Mabel”
  schema:inLanguage “en”;
  schema:name “Soul Mountain”@en ;
  schema:translationOfWork <http://worldcat.org/entity/work/id/1215997>
```

If the information curated by libraries and museums is formulated as linked data according to our recommendations, a machine process can customize the display of search results according to the user’s language preferences. But the underlying data is not only more accessible but even richer because library resource descriptions such as VIAF are already linked with third-party resources such as Wikidata. They can be discovered and mined more easily to support scholarly inquiry. For example, a researcher...
interested in gauging information sharing across cultures might ask these questions: Which authors are translated the most? Which works have been translated into the most languages? How many translations are from the original work and not from a translation of a translation? Which countries or regions are the focus of the greatest translation activity, and what are the most common source and target languages? Answers to such questions can be gleaned from searches on today’s library databases, but only with much manual intervention. Linked data implementations promise more precision and comprehensiveness with less effort. Experimental prototypes have demonstrated some impressive results.

From Innovation to Best Practice

The examples described so far illustrate what catalogers have probably always understood about MARC: the standard supports a range of options for description – from text strings intended for human readers to terse but precise encodings that are designed primarily for machine processes. Since linked data is about greater machine understandability, it makes sense to craft descriptions higher on the scale.

Adopting linked data as a replacement for MARC is a long-term goal for the library community, expressed in strategy documents published by the Library of Congress and many national libraries. Progress towards this goal is often achieved through the insights of individual researchers and catalogers. But innovation must be propelled into best practices and amplified into community-wide change through engagement with standards groups.

Since 2009, MARC standards committees have recommended the addition of URIs to bibliographic and authority records. According to a MARBI (Machine-Readable Bibliographic Information Committee) position paper published in 2009

The use of a URI instead of plain text is particularly applicable to situations where the value of the...element comes from a controlled vocabulary, which could be an authority list or formal thesaurus (e.g., a name from the LC Name Authority File or a topic for an LSCSH heading) or any other list of controlled codes or terms (e.g. the MARC Code List for Languages). [3]

Sample outcomes are a MARC subject field such as ‘650 #0 $a Courtship $v Fiction $0 (uri) http://id.loc.gov/authorities/subjects/sh2008100298’ or a slightly more complex MARC author field such as ‘110 2# $a University of Texas. Dept. of Anthropology. $0 http://lccn.loc.gov/n86041077 $4 spn http://id.loc.gov/vocabulary/relators/spn.’ In these examples, URIs identify authoritative web locations for the topical heading “Courtship—Fiction,” the corporate name heading “University of Texas Department of Anthropology” and the relationship code “spn” or “sponsor.”

These uses of URIs promote greater consistency and machine understanding, but it is not obvious that the real-world Things behind the headings are identified. The intent is clearer in an authority record for a living person. The example below is an excerpt from the Library of Congress Name Authority record for Donald Trump, with details pulled while he was a 2016 U.S. presidential candidate. Of interest here are the 024 fields, some listed below, which list URIs associated with the entity named by the heading “Donald Trump.”

024 7_ |a http://dbpedia.org/resource/DonaldTrump |2 uri
024 7_ |a http://vocab.getty.edu/ulan/500082105 |2 uri
024 7_ |a http://id.worldcat.org/fast/174117 |2 uri
024 7_ |a https://viaf.org/viaf/49272447 |2 uri
024 7_ |a http://www.imdb.com/name/nm0874339 |2 uri
024 7_ |a http://id.ndl.go.jp/auth/ndlna/00476339 |2 uri

... 001 1_ |a Trump, Donald, |d 1946-

Nevertheless, the resources accessible from the URIs illustrate a variety of options for human and machine consumption. Some are only human-readable English-language text, such as the Internet Movie Database, or IMDB. Some are modeled as linked data and exported as a human-readable view, such as the Japanese-language texts published by the National Diet Library (NDL). And some are primarily about controlled headings for Donald Trump, while others are about Donald Trump the person. Perhaps such noise or ambiguity is inevitable, but all are labeled as URIs in the same MARC field. As a result, a machine process can mine this information and
construct a composite view of Donald Trump as an author, a performer, a media celebrity and a political figure, which leverages the knowledge of many stakeholders and spans many languages and domains of interest. This example illustrates what might be interpreted as a full embrace of URIs as defined by the linked data conventions. URIs refer to a Thing, something real in the world. When URIs are dereferenced, a machine-understandable description links to other descriptions of the Thing, or to related Things, to echo Tim Berners-Lee, who first articulated the linked data vision over a decade ago. When the library community retires the MARC standard and adopts the linked-data paradigm, these examples are a hint of where they might end up – where it becomes possible to project the library’s knowledge onto the web and to assimilate knowledge contributed by others. After all, other communities may not understand library data, but all groups benefit when they come together around the task of merging what is known about famous people.

But the to-do list for realizing this vision from the starting point of existing library standards is still daunting. We have to manage the transition from the old to the new culture of description. We have to create MARC records now that can be easily converted to linked data later, even as the still-experimental research projects being conducted at OCLC and elsewhere are advancing to production. We need to identify the machine-understandable URIs more clearly so that the benefits of linked data can be realized even before the transition to linked data is complete. We have to recognize that data expressed in MARC is only part of the problem, because the same issues arise in all collections of legacy metadata and includes descriptions of all kinds of objects collected and managed by libraries, not just the published books we have highlighted. Translating the good ideas that emerge from this work into best-practices recommendations is the charge of the URI task force sponsored by the PCC, which has been working since late 2015 to identify and address policy issues surrounding the use of URIs in MARC records that should be resolved before implementation can proceed on a large scale. But once implemented, URIs empower every individual with expert knowledge to contribute a link, a fact or a simple association to a collective effort that has the potential to be transformative for libraries and the information-seeking public.

Resources Mentioned in the Article

