What to expect when you're analyzing, transforming, and inputting: A linked data guide
MLA Annual Meeting 2018, Portland, OR

Summary written by Michelle Urberg, ExLibris (ProQuest)/University of Washington Libraries
Session Presenters: Richard P. Smiraglia (University of Wisconsin–Milwaukee), Nancy Lorimer (Stanford University), Tracey Snyder (Cornell University)

With the LC BIBFRAME Pilot Phase 2 webinar soon approaching (March 7, 2018 for those interested), the panel hosted by the Cataloging and Metadata Committee on linked data at MLA in Portland could not have been more timely. The panel brought together three speakers each addressing aspects of linked data work highlighted in the title: analyzing, transforming, and inputting data. Each of the talks focused on different challenges catalogers, metadata librarians, and those in other roles in the library will be facing, as well as important issues to consider in creating linked data environments. All three highlighted that their work included developing robust and well thought-out standards, that this work was undertaken with team-based approaches, and that the linked data world is highly complex, involving many different moving parts to link pieces of data together.

Richard Smiraglia started the panel with his discussion of the "Digging into the Knowledge Graph project "(DiKG) (see http://di4kg.org/), which is an international and interdisciplinary linked open data project (LOD) studying the process of analyzing and transforming economics and music metadata. Smiraglia's talk at MLA introduced the project, as well as presented the challenges and process by which his team is enhancing the knowledge presentation of single item artifacts. The case studies for this project are in music and economics, but Smiraglia's talk for MLA focused primarily on the process of creating and analyzing linked data sets for musical artifacts.

DiKG approaches creating LOD using the five principles Tim Berners-Lee has developed for evaluating robust linked metadata. If DiKG does its job effectively, at a full five-star rating according to the Berners-Lee scale, the data sets will use vocabulary that is available on the web, that is machine-readable, in a non-proprietary format, published using open W3C standards, and that is linked to other vocabularies. DiKG researchers are striving to meet all of these standards through careful work in the early stages of research and project planning. One key strategy is to find existing robust data sets of musical data, one of which is the Computerized Mensural Music Editing Project (The CMME Project), which already is committed to providing open source data and has a growing body of scores complemented by contextual information (see http://cmme.org/about, especially the section about edition projects and metadata). To CMME data, DiKG will supplement entities and attributes by adding LC genre and form terms (used in the 655 field in MARC 21 records), as well as concepts supplied by the Basic Concepts Classification schema (https://goo.gl/CmguVh) that fellow DiKG team member Rich Szostak developed. Smirgalia's talk gave a very brief overview of the project (more here: https://diggingintodata.org/awards/2016/project/digging-knowledge-graph) and ended on a positive note, by observing that the human interaction with analyzing and transforming data for musical artifacts will be significant and that music specialists will be required to make the leap from proof of concept to functional LOD product. More work for the metadata and catalogers among us!

The second presentation was given by Nancy Lorimer: "URIs in MARC: Enhancing Bibliographic Records for Conversion". The work described by Lorimer, like that of Smiraglia, is work that has been undertaken by a group of experts who are exploring how to transform flat data points into data that can be linked at a large scale. This project is a task force within the Program for Cooperative

Cataloging (PCC) and is integrated with the work of other PCC task forces, including the Work-Entity task force, the Identity Management task force, the BIBFRAME task group, the ISNI pilot, and the Linked Data Advisory Group. Each of these groups has its own interest in supplementing cataloging records with URIs. Future documents to be released by the URI task force will include documentation about how to formulate and obtain URIs as well as one on how to use URIs in OCLC cataloging.

Although no final conclusions about URIs have been made by the task force, Lorimer's presentation showed what types of data become URIs in MARC records and how they are an important step toward making data linked. First, challenges currently exist to adding URIs to MARC records because not all URIs will created "linked data." Common types of URIs used in MARC records are websites (as in the $u field of the 856, which takes a URL as a source of information for the item described in the record). These URLs are, however, not linked data. Second, linkages do exist with controlled author (100), subject (650), genre (655), and added name (700/710) fields, but there is currently no set practice for what types of URIs to use for these fields if the authority files will not be the primary source of URI links. Third, these URIs match on strings, not a persistent identifier associated with an authority record control number or a real-world object. The persistent identifier is a preferred choice for creating robust linked data. Currently, the task force is exploring the choice to use authority file records or real world identifiers as URIs. Lorimer concluded her talk by presenting some links to relevant documents, including one that outlines basic information about URIs (see: https://goo.gl/1qrVGz).

Tracey Snyder concluded the panel, with "LD4P & LPs: Development of a Linked Data Editor", a discussion of a small pilot linked data project undertaken at Cornell University. LD4P stands for Linked Data for Production, part of the Linked Data for Libraries grants project, which Cornell Library was granted to create an ontology and RDF editor, which in turn is used to make hip-hop LPs discoverable in a linked data environment, called VitroLib. The software for VitroLib is based on the software developed for VIVO, a platform designed for sharing scholarship of faculty members, librarians, and other scholars. The editor links out to existing LC subject headings and genre/form terms, as well as applies RDA relationship designators (e.g. performer, composer) and Work-Manifestation-Item-level data collection. When one enters a specific work in VitroLib, it is complemented by information about the specific manifestation that is kept at Cornell. The information compiled with each work, manifestation, and item then provides the data for a larger network of linked data accessible through the ux-design of VitroLib.

While the technical aspects of Snyder's talk revealed challenges of reworking a platform like VIVO for a new purpose in VitroLib, for me, the most important point she brought out was how this project took a team people with different skillsets to make even a test environment become a reality. Catalogers completed only a portion of the laobr done to make VitroLib work. The vision of having a functional linked data editor of any size or scope will only be realized when groups collaborate, demo the beta platforms, and adequately document the results. Entering data into this environment is undoubtedly among the most important later steps of the process of making VitroLib function for users, but a large portion of work is done by developers, ontologists, metadata librarians, and cataloging coordinators long before data is entered and users are involved. The linked data world proposed by the creation of VitroLib is one that defies silos that often exist in large library settings (and possibly reflects the future of libraries??).