Stanford Project Proposal

Stanford is proposing two projects

- Tracer Bullets
  - Overview
  - Objectives
    - Pathway 1 - Traditional Vendor-supplied Cataloging
    - Pathway 2 - Traditional Original Cataloging
    - Pathway 3 - Self-Deposit to the Digital Repository
    - Pathway 4 - Ingestion of a Collection into the Digital Repository
- Performed Music Ontology (PMO)
  - Objectives
  - Deliverables

Tracer Bullets

According to the Agile Dictionary, a tracer bullet is “a set of work where interfaces are developed from beginning to end of a process. These interfaces may be very simplified or may just pass through. The purpose of the tracer bullet is to examine how an end-to- end process will work and examine feasibility.” Key to any successful transition of technical services functions to linked data will be the transition of its traditional workflows as these workflows account for the dominant part of a department’s throughput. Stanford will be applying the tracer bullet principles to its fundamental workflows.

Overview

As part of Stanford’s participation in the first Linked Data for Libraries project, much effort was put into the conversion of its MARC data to linked data, in this case, BIBFRAME. The predominance of a library’s metadata is captured in MARC and its conversion to linked data will be key to any successful shift in technology. This conversion process revealed a number of problem areas, however. MARC was intended not only to capture metadata about a resource, but also metadata about the metadata (for instance, when the metadata was created). These different types of metadata are difficult to sort out in the conversion to linked data. Likewise, many relationships, such as the link between a subject heading and a specific work, or the link between a performer and the particular work they performed, is not captured in MARC. It is assumed that the person viewing the data at a computer screen will make those connections. Because of this lack of internal connection within a MARC record, these relationships, and many others, will not be expressed in linked data after its conversion. All of these must be added by hand post conversion. It became clear, in order to avoid this extra labor, libraries would need to begin creating their metadata as linked data directly. This realization drove Stanford to choose the conversion of its workflows as its choice for an LD4P Institutional project. Although Stanford’s production workflows will be unique to that institution, the elements in the production chain can be generalized and shared with other members of LD4P. In addition, agreements for support with common vendors such as Casalini Libri or Backstage will support all.

The individual elements in this new production workflow are tantalizingly familiar. Stanford has had a many-year history with the conversion of MARC data, both bibliographic and authority, to linked data. They have had experience with a variety of triple-stores and the ingest of data. They have in-depth knowledge of RDF and extensive experience with identifiers. They have evaluated the Library of Congress’ linked-data production tools. And the Library of Congress tools are sophisticated enough at this point that LC will be training 40 catalogers to use them at that institution. But even given the experience with the elements of this new workflow, concerted effort will be needed to bind them together into a workflow that will function in Stanford’s environment.

Building upon the roadmap developed by BIBFLOW and its analysis of needs in a linked data environment, Stanford will focus these requirements on four key production pathways. Each pathway will be examined, from acquisition to discovery, as a tracer bullet. All key elements in those workflows will be converted to a process rooted in linked data but in a basic way. Emphasis will be on the completeness of the pathway. The workflows themselves will be expanded in future to account for additional complexities once the initial pathway has been established.

Stanford will be creating a parallel LOD processing stream to accomplish these goals. Resources flowing through these pathways first will be processed in the traditional way with traditional MARC or MODS metadata. This traditional metadata will be used for discovery purposes in the university’s discovery environment and for contributions to cooperative cataloging programs such as the Program for Cooperative Cataloging. A parallel, linked-data workflow will be created, however, for LD4P and duplicative metadata created. This metadata will feed into a parallel discovery environment as well so that we can mimic the entire processing stream. This parallel metadata can also be sent to various library vendors and programs so that they can begin to adjust their businesses to incorporate linked data. Although this proposed solution will require duplicative effort, it will allow Stanford to experiment with an alternative pathway without being dependent on the results for discovery. It also has the benefit of testing the new pathway with actual library resources and staff so that a true measure of effort and cost to implement the new paradigm can be evaluated.

In preparing for this grant proposal, Stanford held a series of meetings involving its Acquisitions Department, Metadata Department and Digital Library Systems and Services (DLSS). Key to the analysis was the mapping of its MARC workflows, MODS workflows, and original linked data workflows (including data from vendors, self-created linked data using LC’s BIBFRAME Editor, etc.). The elements of Stanford’s workflows and how they interact with these data types, data stores (Symphony (ILS), Digital Repository (SDR3)), and the triple store (Store Cache) were carefully mapped out. An analysis was done of what tools and environmental changes were needed in order to allow the individual workflows to proceed as a tracer bullet and the amount of effort needed to finish the work.

The resultant analysis led Stanford to choose four key workflows for conversion to linked data, two for traditional materials and two for digital: copy cataloging through the Acquisitions Department, original cataloging, deposit of a single item into the Digital Repository, and the deposit of a collection of resources into the Digital Repository. The tracer bullet will follow the life cycle of a resource, from its acquisition to discovery. Each process along the way will be converted to a linked data strategy. These processes simply need to be good enough to support an experimental workflow. Requirements for a full production workflow will be gathered iteratively and passed on through regular meetings to LD4L-Labs for development. The development of this skeletal architecture, however, will still demand a sizeable amount of effort. Through workflow analysis, we have determined eight key areas for initial development: implementation and enhancement of the LC MARC2BIBFRAME converter, functional installation of the LC BIBFRAME Editor, development of storage and caching mechanisms, development of a BIBFRAME bridge to the ILS, a BIBFRAME to Solr mapping for discovery in Blacklight, the publishing of the linked data output to the web, integration of the BIBFRAME Editor to the Digital Repository, and development of the systems architecture to answer such questions as database of record for resource metadata.

Objectives

Pathway 1 - Traditional Vendor-supplied Cataloging

Currently, 80% of Stanford's monographs come with some form of MARC copy. The material is received by lower level paraprofessional staff in Acquisitions and cataloged on receipt. Because of the large volume of materials, throughput must remain high. As the transition of the entire library ecosystem to linked data will be slow, having at least one flow that begins with MARC data will be inevitable. And because of its volume, the transition of this copy-cataloging workflow will be essential. The effort needed to convert this workflow has been analyzed. Of equal importance to the conversion of the workflow itself will be the broader questions that the Project Co-Manager can explore based on data gathered as resources are processed. The answers to these questions are not necessary for the successful completion of this phase of LD4P, but will prove invaluable for the planning of the next iteration of LD4P after these first two years are completed. For example:

- what is the best configuration of the MARC to BIBFRAME converter for this level of staff?
- is it possible to automate the creation of identifiers for controlled access points?
- what reconciliation processes are available during the automated process?
- what linked-data elements are most desirable post-conversion for highly functioning/integrable linked data
- how much additional cost does this workflow require?

Pathway 2 - Traditional Original Cataloging

Stanford produces original cataloging in all formats (books, serials, sound recordings, etc.) according to national standards for its own internal needs and to share with the broader library community. Although needing to make use of many of the same tools as Pathway 1 (e.g., a BIBFRAME Editor), as professional staff, the approach will be much different. These staff will be creating new linked data directly, not converting MARC data and enhancing it. As in Pathway 1, of equal importance to the conversion of the workflow itself, will be the broader questions that the Project Co-Manager can explore based on data gathered as resources are processed. For example:

- what is the best configuration of the BIBFRAME editor for the creation of new data
- how best to create new authorities and identifiers to support new controlled data
- how best to integrate these identifiers with the local identity management
- how best to expand the BIBFRAME ontology to cover multiple formats

Pathway 3 - Self-Deposit to the Digital Repository

Another major flow of metadata is for born digital objects deposited into Stanford's digital repository and from there into the discovery environment. This metadata currently is stored and maintained in the MODS format. Pathway 3 explores the self-deposit of a single digital resource into the digital repository. Issues to be explored:

- conversion of MODS metadata to BIBFRAME
- automated assignment of identifiers for controlled headings
- storage of linked data within the digital repository

Pathway 4 - Ingestion of a Collection into the Digital Repository

Stanford's digital repository also hosts an increasing number of large collections of digital objects. Metadata is often received in the form of a spreadsheet and is converted for deposit and typically enriched or remediated afterwards. All processes must be automated as the collections are large. Issues to be explored:

- conversion of a large collection of metadata to linked data
- automated remediation of the metadata either before or after processing

The Tracer Bullet projects will make use of common tools and environments as they proceed. A number of these tools, such as the BIBFRAME Editor, have already been developed and tested by the Library of Congress and will be used by them in their LD4P projects. Other elements such as the triple store or MARC conversion flow have been developed by LD4L and are already in place. The Tracer Bullet Technologist will be responsible for integrating these tools into a single flow within the local Stanford environment. As the tools are used, optional enhancements will be uncovered that would improve their functionality outside of the simple, tracer-bullet workflows.

Performed Music Ontology (PMO)

The Performed Music Ontology Project is a collaborative effort of Stanford University, the Music Library Association (MLA), the Association for Recorded Sound Collections (ARSC), the Library of Congress, and the PCC, with participation of LD4P partner institutions. The project aims to develop a BIBFRAME-based ontology for performed music in all formats, with a particular emphasis on clarifying and expanding on the modelling of works, events, and their contributors. Building on the work being completed by the Library of Congress in early 2016 and using BIBFRAME as a core ontology, the project members will collaborate to expand that core with domain-specific enhancements for use as a common standard by the library and archival communities, and establish a model by which these extensions can be created, endorsed, and maintained by the community in the future. Along with the development of the new ontology, the project team will explore ways of enhancing existing MARC records to make them more conversion-friendly to the newly developed ontology.

Linked data, including BIBFRAME, provides a major opportunity for describing performed music resources. The rich complex of associations in and among sound recording resources can be expressed through machine-linking of the data elements, and made available for further enhancement as linked open data on the Web. Without the restrictions imposed by the MARC format, more subtle relationships amongst performed music instances, the interrelationships of musical groups and musicians, the relationships between works, all can potentially be better expressed in linked open data. BIBFRAME, however, does not in itself provide a complete model for expressing all these relationships, its initial creation having been based strongly on the contents of a MARC record, and thus inheriting some of its inherent drawbacks for non-book resources. To best expose and exploit the relationships in a performed music resource, the BIBFRAME ontology needs to be extended, either through extending the ontology from within, or by using pre-existing ontologies. BIBFRAME would thus be conceived of as a core ontology, to which additional domain-specific ontologies may be added. This approach is highly appealing, not only to communities interested in performed music, but also to other specialist communities, in that they have the opportunity to better shape ontologies and vocabularies to respond to their particular needs and domain outlook. Ideally, using BIBFRAME as a core ontology can allow for data exchange and compatibility amongst different domains, while still being tailored to the needs of each.
The risk with this model for BIBFRAME development is that without a cooperative effort in development and maintenance, conflicting ontologies may be developed for the same domain and data exchange hindered. The library community maintains itself by creating metadata to very specific standards that can be exchange with little or no additional effort, the overall load of work being far too great for any one institution. To continue this successful collaboration in the linked data sphere, it is vital that communities come together to create enhancements and agree upon a common ontology, and establish a model by which these enhancements can be developed and maintained by the community going forward.

The project will emphasize the refinement and extension of the BIBFRAME 2.0 model as developed by the Library of Congress. Special emphasis will be placed on the modelling of events, technical characteristics, performers, and the integration of the new ontology for medium of performance. Project team members are drawn from the partner institutions, and serve as primary liaisons for the project; there will also be major contributions from, and testing by, catalogers and selected working groups within each organization. This collaboration of the primary stakeholders is vital in developing a successful, sharable standard.

Objectives

- Building on LC’s work, evaluate the BIBFRAME ontology for describing performed music, both for mainstream and archival performed music collections.
- Evaluate other available linked data ontologies and vocabularies for extending and/or expanding BIBFRAME to better accommodate performed music.
- Create and document a linked data ontology for performed music, using BIBFRAME and either BIBFRAME extensions or already existing ontologies.
- Develop an RDA profile to complement the ontology for use by the library community.
- Suggest preferred vocabularies for use by the library community.
- Catalog a representative selection of performed music resources using the new ontology, and make them publicly available.
- Determine any augmentations/manipulations of MARC records for performed music that would enhance results when converted to the new ontology.
- Evaluate results of project and share set of recommendations for further research and development.
- Inform the development of metadata production tools to ensure compatibility for describing performed music.

Deliverables

- Provide a written evaluation of the BIBFRAME ontology in describing performed music, including a set of use cases to justify extending and/or expanding the ontology.
- Define a linked data ontology with a BIBFRAME core with for describing performed music.
- Create an RDA profile to complement the ontology for shared use by the library community, including list of preferred vocabularies.
- Provide a representative selection of resource descriptions created using the Performed Music Ontology.
- Provide a strategy for augmenting and/or manipulating performed music MARC bibliographic records to improve conversion to the Performed Music ontology.
- Write an evaluation of the project findings with a set of recommendations for further research and development.
- Present project findings to appropriate library and linked data communities such as the Music Library Association, the Association of Recorded Sound Collections, the Library of Congress A-V Division, the Canadian Association of Music Libraries, and ALA.