Metadata reuse (WP2)

Summary and background

In the first Linked Data for Production grant, Stanford’s Tracer Bullet 1 was defined as copy-cataloging making use of vendor-supplied metadata in the MARC formats. This first Tracer Bullet will be core to libraries’ new linked data workflows as the majority of traditional materials acquired will be received with some form of MARC-based copy. In this workflow, we identified metadata supplied by Casalini in our ILS database and converted it to BIBFRAME 2.0 making use of a MARC to BIBFRAME converter supplied by the Library of Congress. We then stored the data in a local triple store. As an enhancement to this process, Stanford developed a pipeline so that this process could be done at scale instead of breaking the data into small chunks for repetitive processing. LD4P2 will move this workflow and pipeline into an implementation phase. The pipeline will be available in the cloud environment and the data it produces fed into the data pool in the cloud.

There will be three key aspects to the next phase of work:

- The inclusion of identifiers within the MARC data supplied by vendors will greatly simplify the conversion process by eliminating the need for reconciliation of those entities. As vendors supply a sizeable portion of a library’s metadata, this agreement on a new standard for the MARC metadata supplied by them will be a key step forward. Casalini has already adopted this standard and will make this enhanced metadata available to all their customers if desired. In the next phase of LD4P, Stanford will request that Coutts/Proquest add identifiers to a selection of the MARC metadata they supply as has been developed with Casalini. As they supply cataloging for all of our US/UK imprints, they will be the single most important vendor with whom we must work.
- The conversion pipeline developed for the first phase of LD4P will need to be refined and enhanced. As we expand our flow to include MARC metadata of less high-quality than that supplied by Casalini, or MARC metadata lacking the addition of identifiers for known entities, processes such as reconciliation become important in the data conversion process. A second area of development is updating. The MARC metadata we have converted to linked data may receive metadata corrections and enhancements over time from vendors that still supply data in the traditional MARC formats. These additional metadata sources must be folded into the production pipeline and their updates applied to the metadata previously converted. As part of this project, Stanford will investigate and trial different technical approaches to external reconciliation and updates such as the complex reconciliation service offered by Casalini through SHARE-VDE and incorporate both into its MARC-to-BIBFRAME conversion pipeline.
- A final area to resolve will be the direct use of pre-existing RDF metadata for a resource as opposed to the conversion of matching MARC copy. The use of native RDF descriptions will be a fundamental shift for libraries, finally moving them away from a record-based ecosystem. In this phase of LD4P, partners will be able to make use of SHARE-VDE or the data included in the cloud environment to fully articulate our needs. As per agreement with the LD4P Cohort members, the cloud environment will include a current instantiation of all of their institutions' metadata in RDF. As such, it will become an ideal source for metadata reuse. Policies will need to be developed for working within a communal environment and the retention of provenance for individual statements. A key partner will be the Program for Cooperative Cataloging in the development of such policies. While this effort will be targeting integration with and use of pre-existing RDF metadata in the Casalini SHARE-VDE (largely because it is a readily accessible, high quality source of metadata with a high degree of interest to the participating PCC libraries), the mechanisms and policies established here should be extensible to the use of existing RDF metadata in other environments (commercial, consortial or institutional).

Enhancement of the current conversion pipeline and the connection of the cloud-based editing environment to Casalini SHARE-VDE will be supported by the developers at Stanford. The development of policies for these activities will be created through partnership with the PCC and LD4P2 and fostered through PCC leadership based at Harvard.