Tracer Bullet 4: Ingestion of a Collection into the Digital Repository

This workflow tests approaches to the automated conversion of metadata for a digital collection to linked data. The sample data used was a replayable spreadsheet with metadata for the Classics South Africa collection, which in current practice was converted to MODS for ingest to the repository.

SUL digital collections metadata context

- Currently no authority control for metadata in digital repository
- Entering URIs is recommended but time-consuming, especially for large collections
- Much digital collections content is rare or unique, with names that are unreconcilable or of local significance only
- Local process developed to derive string matches from a local copy of the LC authority file

LC MODS to BIBFRAME mapping

The Library of Congress has made available a mapping of MODS to BIBFRAME, which was consulted for the replayable spreadsheet mapping. As the mapping is still in draft form, there is not an XSLT or other conversion mechanism available yet for use in an application.

KARMA

KARMA is a tool for modeling tabular data and converting it to linked data. We attempted to model the metadata spreadsheet via KARMA, but encountered some problems.

1. The creation of blank nodes often failed, requiring all modeling info to be re-created, after which the action would often fail again. This proved an unsurmountable barrier, as BIBFRAME relies heavily upon blank nodes.

2. The force-directed graphical tree interface was challenging to use, and became more confusing when part of the tree was off-screen. Attribute bubbles sometimes overlapped each other such that it was impossible to edit one of them. Not being able to see the tree in a comprehensive, static way made it difficult to construct and to navigate.

3. The BIBFRAME conversion from MODS is sometimes highly dependent on the actual data (for example, the role associated with an agent determines whether that contribution is associated with the Work or the Instance). Representing those dependencies required complex Python scripting beyond the tiny snippets the application expected.

4. The likelihood of making mistakes in the modeling by clicking the wrong moving bubble negatively impacts the AI autosuggest feature.

Ruby

Given the blocks to making effective use of Karma for the time being, we developed a Ruby script that mapped spreadsheet data to BIBFRAME fragments. For the purposes of this work, the mapping coverage was limited to the data elements and values present in the sample set. The flexibility of the script and the limited scope made it possible to express all relevant data dependencies. The high amount of resources required for this process, however, suggest that they would be more usefully applied direclty to a shared standard rather than a local metadata template facilitating data creation according to that standard.

Bibcat

The Python-based application Bibcat focuses on source data in MARC, but does provide a MODS to BIBFRAME mapping, which we tested. However, the mapping covered only a small subset of available MODS data, and its use of RML (rather than the W3C standard R2RML) raised questions of sustainability.

Issues for further exploration

1. Refinement of MODS mapping to account for data not yet mapped and entity generation.
2. Incorporating reconciliation.
3. Evaluation of target RDF models other than BIBFRAME.