Bulk downloads of Works and Instances

Introduction

This bulk download is the first attempt to share the Library of Congress BIBFRAME file with the community. The file was developed from the Library of Congress MARC unit records converted to BIBFRAME RDF, creating linked BIBFRAME Work (over 19 million) and BIBFRAME Instance (around 24 million) descriptions. The MARC bibliographic record targets an RDA manifestation description, but also contains elements associated with the current RDA work and expression definitions, all combined in the same record. Data in the MARC unit record has been divided into BIBFRAME Work descriptions and BIBFRAME Instance descriptions. The Work descriptions generated by this process were then merged (deduped). The MARC Title Authority records, which are similar to BIBFRAME Work descriptions, were matched with the new Work descriptions from the bibliographic records and merged. This process sounds simple but it is difficult to do accurately because of the textual nature of much MARC data, the mixture of full and partial records in the file, the use of the MARC Authority format for titles, and the inconsistency of data and diversity of cataloging rules over time (more than 100 years).

Contents

This export of RDF for BIBFRAME Works and Instances uses document-based configurations for the data. We exported Works into one set of ntriples files, and Instances into another set, zipped up on the id.loc.gov download page.

Known Issues

We have found the following are some issues and will continue to address these and others in upcoming releases:

1) The URIs for Works and Instances in the resulting export files will not resolve outside the Library of Congress network because they point to a separate database that is not open to the public. We continue to examine issues related to making the URIs resolve, as well as how to represent the final node in the URI, whether it be with LCCNs, different prefixes for different sources, etc.

***Current numbering schemes**:*

- /resources/works/n1234: works from name-title or title authorities converted from MARC records at id.loc.gov
  (n1234 is authority LCCN)
- /resources/works/c1234: works from bib records converted from MARC (1234 is Voyager bib id)
- /resources/works/e1234: works created from scratch in BIBFRAME editor (1234 is LCCN from Instance)
- /resources/instances/c12340001: instances from bib records converted from MARC (1234 is Voyager bib id plus 4 digit one-up number, starting at 0001)

- /resources/instances/e12340001: instances created from scratch in BIBFRAME editor (1234 is LCCN, plus 4 digit one-up number, starting at 0001)

***Editing & Resource numbering Rules followed*:**

- If a new description created in the BIBFRAME editor lacks an LCCN, the e# will be an internal ID assigned in the editor, eg., e1521740260032422468226941184332968546 57
- If resources are edited and lack an LCCN, the original URI is retained. Once an LCCN is assigned in an instance, that Instance gets it's a new URI based on the LCCN.
- If a description is called up that already has an LCCN and a new LCCN is assigned, this is a copy function, and the description is stored under the new URI with LCCN.
- Generally RDF descriptions are not posted to the database until the LCCN is assigned, but saved in the temporary space while being edited.

2) Since the triples are stored within documents in our system, there is a lot of duplication on export. For example:

<http://id.loc.gov/vocabulary/languages/eng> <rdf:type> <http://id.loc.gov/ontologies/bibframe/Language> .

This makes the export file larger, but a triplestore will de-duplicate them.

3) We did not use URI reconciliation services while loading originally, so there are many unlinked names and subjects for this round.

4) Merge problems identified have been corrected by re-merging small sets of data. Re-merging after the fact may leave behind "orphan" works whose instance is now attached to another work. These are being fixed as found.

5) Works were created from the split of data in a MARC record. The resultant Works were then merged based on a combined name/title string. For example "Twain, Mark, 1835-1910. Adventures of Huckleberry Finn" .

The Work being merged from was discarded but its subjects and classification numbers were are copied and deduped onto the merged Work. Instances were re-linked to the merged Work. Some expression Works may not have correctly benefitted from this merging. Future software improvements are planned to retain a unique expression Work and to build an "expressionOf" link to the appropriate Work description.

6) When 7XX tags are converted from MARC records, brief Work descriptions are created. In the future, these brief Works will be merged with the existing Works.

7) Works created from name-title MARC authority records will be linked to each other in a future update.

Comments

Please send any comments to ndmso@loc.gov. If possible, please include the LCCN of any examples.