Building a BIBFRAME Catalog

Bibliographic Records

BIBFRAME descriptions

nametitles, titles

id.loc.gov

BIBFRAME database
Initial Works File

- Extract nametitle/title Authorities from ID.loc.gov
- Transform to BIBFRAME (see github)
- Ingest to database
Bibliographic Conversion

- **ILS Export**
  - MARC2bibframe2 transform (see github)
  - Match to existing bf:Works with same nametitle

- **BIBFRAME database**
  - Found bf:Work?
    - Yes:
      - Merge, Dedup Subjects, Classifications
      - Store in Found Work
      - Adjust uris to found Work, Store new Instances, Items
    - No:
      - Store as new bf:Work
      - Store new Instances, Items
  - Store in Found Work
  - Adjust uris to found Work
  - Store new Instances, Items
BIBFRAME Descriptions

• Create new bf:Work, Instances, Items
• Ingest (what is the uri?)

• Create Instance, Items(s)
• Look up a bf:Work in BIBFRAME database
• Ingest with link to the Work
Infrastructure

- MarkLogic NoSQL Server (3 node cluster) for ID
  - Storage, search/display, RDF triplestore

- MarkLogic 3 node cluster
  - for BIBFRAME and ID ingest, processing, testing

- Apache/Varnish Web Cache
  - (2 VMs for load balancing)

- Xquery, SPARQL code base for ingest, search/display

- Javascript codebase for BIBFRAME editor

- XSL for MARCXML, ONIX data transformations
Infrastructure Updates

• Added new node to MarkLogic production cluster for ID
• Added 1 varnish web cache server
• Added 2 new nodes for BIBFRAME processing MarkLogic cluster
• Upgraded from MarkLogic version 5 to version 8
• MarkLogic Semantics replaces 4store triplestore
  o Document-based triples for ease of updates
• New BIBFRAME database added to id database
  o Still not public
• HTTPS support just added (not mandated)
Software updates I

• New MARC Conversion in xsl instead of xquery

• Installation of conversion in Metaproxy, yaz

• New Authorities transform for nametitles

• Comparison program online to show MARC and BIBFRAME side by side in rdfxml and ttl serializations.

• Merge/ingest programs (nametitles and bibliographic records) updated for BIBFRAME2 vocabulary

• New search/display interface
Software updates II

• Use SPARQL to show links to parent Work/Instance, sibling Instances, Item titles

• New templates for BIBFRAME2 vocabulary in Editor, new lookups for controlled vocabularies

• Editor now has lookups to BIBFRAME database for attaching Instances to Works

• Storing “published” BIBFRAME descriptions in database

• Daily nametitle and bib ingests from ILS to database to simulate the real catalog
Some Numbers

ID.loc.gov: 10.5M Names, Subjects, vocabularies
  - 300M triples
  - subjects: 21M
  - predicates: 768
  - objects: 25M

BIBFRAME Database: 65M Works, Instances, Items
  - 4 Billion Triples
  - subjects: 500M
  - predicates: 14,615
  - objects: 800M
Merge/Match Specs

- Based on 130/240 uniform titles indexed as “nametitle”

- New bf:Works stored with “nametitle” index and so become match point for future records

- For each new work from MARC, concatenate primary contributor and title (not from MARC 880)
  
  <bflc:name00MatchKey>Twain, Mark, 1835-1910.</bflc:name00MatchKey>
  
  <bflc:title00MatchKey>Adventures of Huckleberry Finn</bflc:title00MatchKey>

- (strip trailing slash)

- Match to existing database index entries.

- Suppressing “Untitled”, null etc., going forward
Merge Stats

- 1.2M nametitles/titles as Works
- 17M Bibliographic descriptions
- 1.2M Works have merged instances
- 1.4M Instances merged altogether (onto nametitles/titles or other bibs)
- 530K Instances merged onto nametitle/title works
  - (still verifying these results)
Runaway mittens
Instance from Bib

Instance
resources/instances/c00147719200001
Title Runaway mittens
LCCN 87012024
ISBN 0688070531
ISBN 068807054X (lib. bdg.)
Physical details col. ill.
Creative responsibility statement by Jean Rogers ; pictures by Rie Muñoz
Mode of issuance http://id.loc.gov/vocabulary/issuance/mono

Publication
http://id.loc.gov/ontologies/bibframe/Publication
Date [PDT]1988
Place http://id.loc.gov/vocabulary/countries/nyu

Publication
http://id.loc.gov/ontologies/bibframe/Publication
Date c1988
Place New York
Agent Greenwillow Books

Dimensions 26 cm.
Has Item resources/items/c00147719200001

Instance Of
› Runaway mittens /(c001477192)
Sibling(s)
› Runaway mittens / c00147719200001
› Runaway mittens /(c01184554400001)
Has Items(s)
› c00147719200001

Other Views for This Description
› BIBFRAME RDF
› BIBFRAME Turtle
› BIBFRAME JSON
› BIBFRAME JSON-LD
› METS
Title authority collocating mechanism, probably not a pure bf:Work. But results from cataloging decisions.
SPARQL Use 1

Display Instance parent, sibling title info using SPARQL
SPARQL Use I

Display Instance parent, sibling title info using SPARQL
SPARQL Use I

Display Instance parent, sibling title info using SPARQL
SPARQL Use II
Display Item title, parent info from other docs using SPARQL

```xml
- <rdf:RDF>
  - <bf:Item rdf:about="/resources/items/c0014771920001">
    - <bf:shelfMark>
      - <bf:ShelfMarkLcc>
        <rdfs:label>PZ7.R6355 Ru 1988</rdfs:label>
      </bf:ShelfMarkLcc>
    </bf:shelfMark>
    <bflc:derivedFrom rdf:resource="/resources/bibs/1477192.marcxml.xml"/>
    <bflc:itemOf rdf:resource="/resources/instances/c0014771920001"/>
  </bf:Item>
</rdf:RDF>
```
SPARQL Use II

Display Item title, parent info from other docs using SPARQL
SPARQL Use II

Display Item title, parent info from other docs using SPARQL

The Library of Congress > BIBFRAME Database

Runaway mittens
Item

LCC call number PZ7.R6355 Ru 1988
Source DLC

Source metadata /resources/bibs/1477192.marcxml.xml

Item Of Runaway mittens /(c0014771920001)

Other Views for This Description
- BIBFRAME RDF
- BIBFRAME Turtle
- BIBFRAME JSON
- BIBFRAME JSON-LD
- METS

Bookmark This Description
Display Item title, parent info from other docs using SPARQL

The Library of Congress > BIBFRAME Database > Detail

Runaway mittens

Source metadata

/loc/lcctpражения/c0014771920001

Other Views for This Description
- BIBFRAME RDF
- BIBFRAME Turtle
- BIBFRAME JSON
- BIBFRAME JSON-LD
- METS

Bookmark This Description
Issues Already Encountered

• Serializations are an ongoing issue:
  o `<rdf:Description><rdf:type rdf:resource="bf:Work"/></rdf:Description> == <bf:Work/>`

• Huge number of triples: how to limit, dedup on the way in, cache labels, etc.

• Merge: MARC 130s are problematic for title authorities; too many “Untitled” etc.
  o eg., photographs

• Merge: Record load sequence affects matching on initial build and reload.
  (Daily records okay)

• BIBFRAME conversion spec changes affect existing descriptions: need update mechanisms that don’t affect merges

• Plenty of interesting examples of merging, conversion, or inadequate data in so many descriptions from varying cataloging rules over the years.
Still to come I

- Open BIBFRAME data to public in some form
  - Bulk download? Searchable interface?

- Analyze data structures for Editor, vocabulary, conversion specs. improvements

- Loading BIBFRAME from ILS or elsewhere into Editor
  - eg., “copy cataloging”

- Ingest CIP and ONIX records

- Implement offset and limit in SPARQL queries
Still to come II

- More SPARQL queries for related works, translations
- Link MARC 7xx related works to existing descriptions.
- More flexible Editor
- New RDF display interface: pure SPARQL display?
- Nametitle authority Works: link translations on ingest
- Services at ID to support external users: picklists etc.
Useful Links

Compare side-by-side MARC/BIBFRAME
bib: http://id.loc.gov/tools/bibframe/compare-id/full-ttl?find=5226
authority: Work conversion

SRU BIBFRAME in Metaproxy
• BY Voyager bib id: (rec.id) Metaproxy for Snoopy on Wheels
• Add some Entity resolution : "bibframe2a" recordSchema
• by LCCN: (bath.lccn) Lookup using LCCN

ID label lookup for any authority/vocabulary
• http://id.loc.gov/authorities/names/label/Twain,%20Mark,%201835-1910.%20Adventures%20of%20Huckleberry%20Finn

Find docs by rdf:type in ID: http://id.loc.gov/search/?q=rdftype:NameTitle&q=
Documentation:
  o http://www.loc.gov/bibframe
  o https://github.com/lcnetdev
Questions?

- Nate Trail
- LS/ABA/NDMSO
- Library of Congress
- ntra@loc.gov