Stanford Tracer Bullets

About the Stanford Tracer Bullets

Stanford's linked data production project focuses on technical services workflows. For each of four key production pathways we will examine each step in the workflow, from acquisition to discovery, to determine how best to transition to a linked data production environment. Our emphasis is on following each workflow from start to finish to show an end-to-end linked data production process, and to highlight areas for future work. The four pathways are: copy cataloging through the Acquisitions Department, original cataloging, deposit of a single item into the Stanford Digital Repository, and deposit of a collection of resources into the Stanford Digital Repository.

Stanford Project Proposal

Deliverables

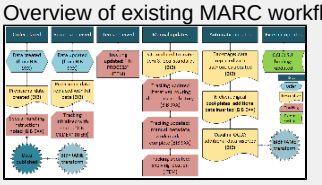
Overview presentations

- Technical Services Workflows: Overview of Tracer Bullet project, with details on Tracer Bullet 1: Conversion of vendor-supplied MARC records (PowerPoint)
- Screenshots of tools and data used in Tracer Bullets 2-4: local instance of Library of Congress BIBFRAME Editor, Blazegraph triple store (PowerPoint)

Tracer Bullet 1: Conversion of vendor-supplied MARC records to BIBFRAME

Existing MARC workflows

Overview of existing MARC workflows



Annotations for this chart

Annotations for metadata flowchart

A. Order placed

1. *Data created (from BIB 9XX)* -- For orders placed in vendor databases, the vendor sends order data in the bib record 9XXs, which are used to automatically generate the order record. For other kinds of orders, ACQ staff creates the order record manually in Symphony.
2. *Provisional data created (BIB)* -- For orders placed in vendor databases, the vendor sends the provisional bib records to be automatically loaded into Symphony. For other kinds of orders, ACQ staff creates provisional bib records in Symphony or imports OCLC copy.
3. *Special handling instructions noted (BIB 9XX)* -- For example, rush handling on receipt.
4. *Data published* -- Searchworks indexes and displays the provisional bibliographic data.
5. *BIBFRAME transform* -- Possible first BIBFRAME transform to allow discovery of on-order items.

B. Records received

1. *Data updated (from BIB 9XX)* -- N/A
2. *Provisional data replaced with full data (BIB)* -- Matches on title control number in Symphony; preprocessing required to match on order ID first for Casalini.
3. *Tracking initialized with status "ON ORDER" (ITEM)* -- Item record is automatically created during the full record load.

C. Items received

1. *Tracking updated: "IN PROCESS" (ITEM)* -- ACQ receiver wands item to pull up associated bib record, and changes the location to register that the item has arrived at SUL.

D. Manual updates

1. *Data updated to match item & local standards (BIB)* -- Depending on the state of the metadata, this work may be done by ACQ or CLASS. See the receiving workflow for more detail.
2. *Tracking updated (iterative): routing directions and history (BIB 9XX)* -- Each time the item changes hands, the 910 field is updated. This is used to track the location of the item, to troubleshoot any issues that arise during processing or later, and to direct where the item should go next.
3. *Tracking updated: manual metadata work marked complete (BIB 9XX)* -- The cataloging date in the 916 field is updated from "NEVER" to the current date. This registers that manual work on the metadata is done, and that the item will proceed to Binding and Finishing.
4. *Tracking updated: shelving location (ITEM)* -- This shows that processing is complete and that the item is ready for use.

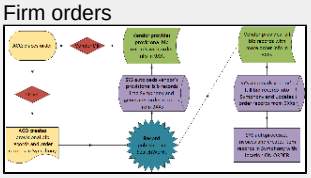
E. Automatic updates

1. *Backstage: data replaced with authorities updated (BIB)* -- MARC records in Symphony are overlaid with the updated records returned from Backstage.
2. *Nielsen, digital bookplates: additional data inserted (BIB 9XX)* -- Nielsen data is inserted as 905, 920, and 986 fields; these are indexed in SearchWorks as the corresponding 5XX fields. For digital bookplates, a 979 field is inserted with a link to the bookplate object in the SDR.
3. *Casalini, OCLC: additional data inserted (BIB)* -- Casalini TOCs are downloaded as PDFs, and a link to the local file is added as an 856. OCLC master record numbers are inserted as an 079.
4. *BIBFRAME transform* -- Possible later BIBFRAME transform to minimize the work required to keep both datasets in sync.

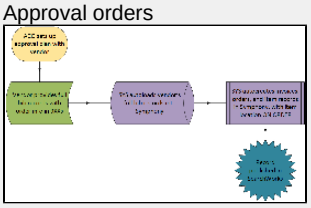
F. External updates

1. *OCLC*: SUL holdings updated -- N/A

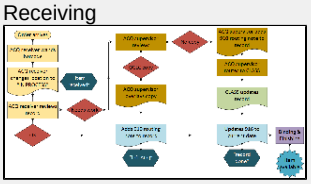
Firm orders



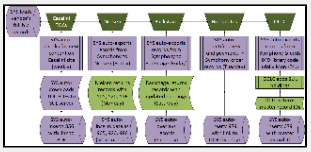
Approval orders



Receiving

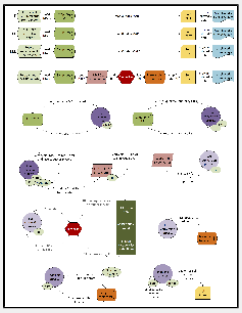


Automatic record enhancements



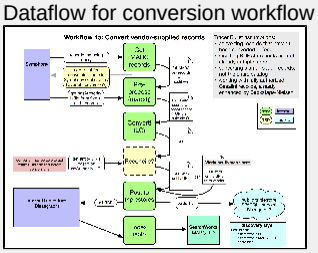
MARC-to-BIBFRAME conversion workflow implemented by LD4P

Conversion workflow overview



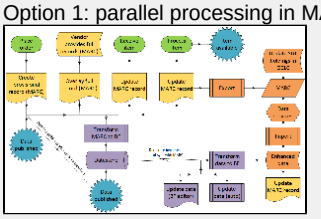
Described in Technical Services Workflows (PowerPoint)

Dataflow for conversion workflow

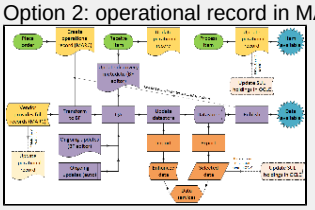


Exploratory MARC-to-BIBFRAME conversion workflows

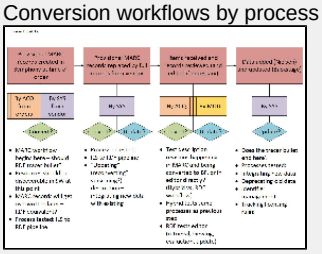
Option 1: parallel processing in MARC and BIBFRAME, MARC record primary



Option 2: operational record in MARC, discovery data in BIBFRAME, BIBFRAME data primary

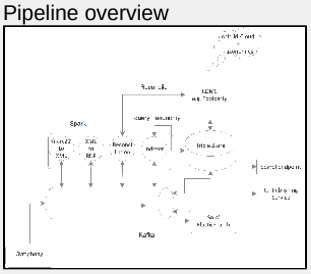


Conversion workflows by process

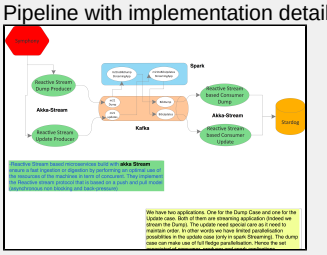


Reactive pipeline for larger-scale conversion of MARC records

Pipeline overview



Pipeline with implementation details



- Reactive pipeline demo (YouTube video)
- Scala/Kafka/Spark Linked Data Pipeline (github repository)

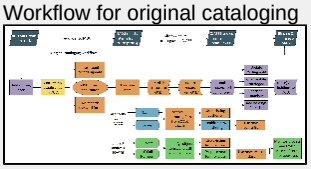
Supporting material

- User stories for converting MARC records to BIBFRAME (pdf)
- Plans for a Minimum MARC Bibliographic Record & Attachments (pdf)
- BIBFRAME-to-Solr mapping (Google doc)
- SPARQL queries for Solr mapping (pdf)

Tracer Bullet 2: Original Cataloging in BIBFRAME

See also Screenshots of tools and data used in Tracer Bullets 2-4: local instance of Library of Congress BIBFRAME Editor, Blazegraph triple store

Workflow for original cataloging

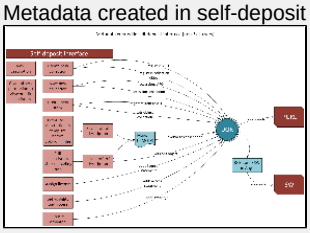


Tracer Bullet 3: Deposit item to digital repository with RDF metadata; deposit by item creator

See also Screenshots of tools and data used in Tracer Bullets 2-4: local instance of Library of Congress BIBFRAME Editor, Blazegraph triple store

Existing item-deposit workflows

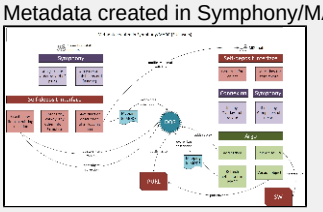
Metadata created in self-deposit interface (non-SUL users)



Description

This diagram represents metadata only (not file management or other admin tasks) for objects described entirely through the self-deposit tool Hydrus. The boxes under "Self-deposit interface" represent the metadata-related tasks a user can perform through that interface. The leftmost column of boxes are metadata tasks contained within the self-deposit tool. The right column of boxes involves writing data to DOR. Except where otherwise specified, these tasks apply to description of both collections and items. Currently this model is more commonly used for deposits originating from non-SUL users.

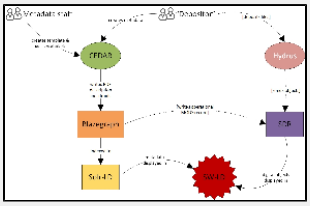
Metadata created in Symphony/MARC (SUL users)



Description

This diagram represents metadata only (not file management or other admin tasks) for objects that are accessioned through the self-deposit tool Hydrus by SUL staff and cataloged by the MDU in MARC. The boxes under "Self-deposit interface" represent the metadata-related tasks an internal user performs through that interface in this workflow. Currently this model is commonly used for digital files received by Acquisitions or acquired by curators. The self-deposit tool serves as a convenient way to get a file into the SDR, notify MDU staff that cataloging is needed, and pass information such as the catkey and purl for the object to MDU. New items are deposited to existing collections that have been set up with the appropriate rights, permissions, etc.

RDF-based item-deposit workflow



Description

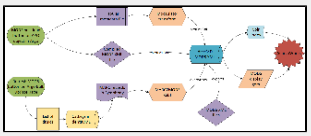
The tracer bullet focuses on the metadata flow rather than the file management portion of this scenario, as the upcoming adoption of Hyku will have a significant impact. For the purposes of the tracer bullet, we are working with digital objects already deposited into the SDR and described via Hydrus. The "depositor" (actually metadata staff) will describe the objects in CEDAR based on the records in SearchWorks, but will reformulate the metadata in CEDAR independently rather than doing a simple mapping from MODS. One possible step is to generate an operational MODS record for SDR use from the RDF description, but for the purposes of the tracer bullet this operational record will not actually be written to the repository.

- CEDAR template for self-deposit item metadata (create CEDAR account to view)
- App to fetch folder contents from CEDAR and post them to a triplestore (github repository)

Tracer Bullet 4: Deposit set of items to digital repository with RDF metadata

See also Screenshots of tools and data used in Tracer Bullets 2-4: local instance of Library of Congress BIBFRAME Editor, Blazegraph triple store

Existing bulk-deposit workflow



MODS-to-BIBFRAME bulk-deposit workflow

- Description of automated workflow to convert tabular MODS-based metadata for a digital collection to BIBFRAME (pdf)
- Sample tabular MODS-based metadata for digital collection (tsv)
- Mapping from tabular MODS-based metadata to BIBFRAME for sample collection (txt)
- Ruby script to convert tabular MODS-based metadata to BIBFRAME
- Resulting BIBFRAME metadata for digital collection (ttl)

Modeling authority data

- Publishing Library Authoritative Information as Linked Data (PowerPoint)

Completed Work

Analysis/Modeling

- Mapped Stanford's vendor-supplied copy cataloging and original cataloging workflows
- Mapped workflow for converting vendor-supplied records to linked data
- Generated requirements for work-based discovery environment, to take advantage of RDF
- Evaluated BIBFRAME profiles for original cataloging

Linked Data Creation

Team

Project Co-Managers

- Philip Schreur, Assistant University Librarian for Technical and Access Services
- Tom Cramer, Assistant University Librarian, Chief Technology Strategist, and Director of Digital Library Systems and Services

Acquisitions Department

- Alexis Manheim, Head of Aquisitions Department
- Linh Chang, Receiving and Access Librarian

Metadata Department

- Nancy Lorimer, Head of Metadata Department
- Joanna Dyla, Head of Medata Development Unit
- Vitus Tang, Head of Data Control and E-resources Unit
- Arcadia Falcone, Metadata Coordinator

Digital Library Systems and Services

- Darsi Rueda, Head of Library Systems Department
- Naomi Dushay, Digital Library Software Engineer
- Joshua Greben, Library Systems Programmer / Analyst
- Darren Weber, Digital Library Software Engineer

- Worked with vendor on improvements to supplied MARC data to enhance conversion to BIBFRAME
- Tracer Bullet 1: Converted set of 38,000 MARC records from Symphony to BIBFRAME using Library of Congress converter, loaded to Blazegraph triplestore, and indexed to Blacklight Solr environment via automated scripts
- Tracer Bullet 2: Created original descriptions of 50 items with local instance of BIBFRAME 2.0 Editor
- Tracer Bullet 3: Created original descriptions of about 30 digital assets using CEDAR RDF editor
- Tracer Bullet 4: Converted tabular MODS-based metadata for a digital collection to BIBFRAME using Ruby
- Piloted automated pipeline approach for conversion of MARC records to BIBFRAME, loading to triplestore, and indexing to Solr

Discovery Environment Creation

- Created Blacklight/Solr instance-based discovery environment with source data a mix of linked data and MARC data
- Developed a mapping from BIBFRAME 2.0 to Solr document for book materials
- Developed a mapping from RDF to Solr for digital assets

Tool Exploration / Requirements Definition

- Gathered requirements for conversion and editing tools
- Set up Registry of Tools
- Evaluated CEDAR template creation and metadata editing tool
- Developed a validation suite for MARC-to-RDF converters
- Created local instance of Library of Congress BIBFRAME 2.0 Editor

Presentations

- Technical Services Workflow Pipeline, Arcadia Falcone, LD4P Community Input Meeting 2017, Stanford, CA
- Linked Data for Production (LD4P): Technical services workflow evolution through tracer bullets (Stanford projects), Arcadia Falcone, Josh Greben, Nancy Lorimer, DCMI 2017, Washington, DC
- LD4P Tracer Bullet 1: an RDF-Based Copy-Cataloging Pipeline, Philip Schreur, ALA Annual 2017: LC BIBFRAME Update Chicago , IL
- The Shot Heard Round the World: Linked Data for Production's Tracer Bullet 1, Practical Copy-Cataloging in RDF, Philip Schreur ALA Annual 2017: Library Linked Data Interest Group Chicago, IL

April 2017 Update