DLC Working Group on Collection and Discovery Services

Briefing on Batch Processing

Last Revision: January 13, 2020

Contents

Purposes and Intentions

- LTS needs to innovate and experiment in order to improve our current services and products and to develop new ones for the FDLP community, support our programs, and fulfill our obligations and responsibilities to our constituents.
- LTS must prepare to adapt to new demands and obligations and to adopt an agile and proactive approach to both expected and unexpected developments.
- In order to meet our Title 44 cataloging and indexing obligations, we need to seek alternative methods for obtaining bibliographic records for ingest into the CGP.
- We propose the approach that some metadata, even though incomplete, partially validated, or not reviewed, is preferable to no metadata at all.
- Batch processing does not preclude review, updating, and enhancing records at a later date
  - We will explore methods for the submission of requests for enhancements through AskGPO or in the CGP, such as a demand-driven cataloging process.

Major Policy Issues and Questions

Analysis and Enhancement of Record Sets

- To what extent do we review, analyze, update, correct, and enhance each record in the sets?
- We have drafted four analysis and editing plans.
    - High Level
        - Exhaustive review of these characteristics of the records and very refined and customized global edits depending on the findings from the record set analysis
        - Thorough examination of the set as a whole
        - Individually review a representative sample of the records (10-25%)
        - Global edits, additions, and deletions of metadata elements and data
        - Research, addition, and correction of SuDoc and item numbers
        - Clear public notice of GPO's actions and status of the records within the CGP
        - 500 note Contributed record: Metadata reviewed, not verified. Some fields updated by batch processes.
        - 922 BATCHPROCESSED
    - Medium Level
        - Expeditious analysis of the records using mainly automated means and a small sample group of the records
        - Thorough examination of the set as a whole
        - Individually review a small sample of the records (1-5%)
        - Utilize the results of the analysis to build a routine of global changes
    - Low Level
        - No examination of the records at all beyond the information solicited from and provided by the partners or contributors
        - Brief examination of the set as a whole
        - No individual record review
        - Make very limited global changes
        - Check and change the SuDoc numbers and item numbers as needed
    - No Analysis or Modifications
        - Accept the records as is and do not make any changes
        - When the records do not contain SuDoc numbers, add them as needed, such as by request of Federal depository libraries or preservation stewards, as part of projects, etc.
        - When the records do not contain item numbers, any actions taken depend on the outcomes of the CGP on GitHub Project Team and other endeavors to assess the item number system.
- We propose to apply the plans based on the circumstances and characteristics of each record set to be batch processed, such as:
    - Specific requests from Federal depository libraries and GPO partners
    - Research potential of the materials
    - High current topical interest of the publications
- We will mainly use automated processes.

- We will focus on access points:
    - 1XXs
    - 24Xs
    - 6XXs
    - 7XXs
    - 8XXs
- We will attempt to identify duplicate records and remove them from the record sets.

SuDoc Classification Numbers

- Should we ensure that all records contain accurate and complete full SuDoc numbers?
- Should we ensure that all records contain at least SuDoc stems?
- What would be the consequences if we do not review existing SuDoc numbers in records?
- What would be the consequences if some records did not have SuDoc numbers or stems?

Item Numbers

- Should we ensure that all records have item numbers?
- What would be the consequences if we do not review existing item numbers in records?
- What would be the consequences if some records did not have item numbers or stems?
- Would posting sets of records without item numbers be an effective alternative method of distribution?

Authority Work and Processing

- We propose to focus on automated authority processing to validate and update access points.
- We would create and update authority records individually as needed and identified through our data validation review and procedures.

Holdings and Item Records in the CGP

- What would be the consequences if we do not create holdings and item records in the CGP for each batch processed bibliographic record?

Brief Descriptions of the Projects

National Institute of Standards and Technology (NIST) Collection Project

The National Institute of Standards and Technology (NIST) Collection Project was selected to act as a model for the development of processes and procedures for the exchange of metadata records between **govinfo** and the CGP and for the review, analysis, enhancement, and ingest of bibliographic records from GPO partners, agencies, Federal depository libraries, and others into the CGP.

To view the NIST records in the CGP, click on the canned expert search wlts=NIST-1.

Here is a brief outline of the batch processing steps of this project:

1. Discussing the records with the agency and asking questions about their cataloging process

2. Analyzing a substantial sample of the records
3. Comparing the record set with the CGP
4. Evaluating metadata problems and issues
5. Searching for and obtaining records in OCLC to identify records with higher quality metadata
6. Developing a plan to revise and enhance the records
7. Running the global update routine
    a. Including a MARC 500 note to indicate that the records have been batch processed
    b. Including a local MARC 922 fields that tag the records as batch processed
8. Batch creating PURLs
9. Deduplicating the finalized record set with the CGP
10. Loading the record set into the test environment
11. Releasing the records in the CGP
12. Soliticing feedback and recommendations from the FDLP community and others

govinfo API Process

LTS staff wrote a script to use the **govinfo** API to create preliminary MARC records for newly ingested hearings. The script obtains the **govinfo** packages and converts the MODS metadata into MARC records. We import the preliminary MARC records into the OCLC online save file and follow our standard cataloging procedures to complete the records. We include the following fields and data in the online version records, and, in some cases, the print version records:

- 024 - GPO jacket number
    - The GPO identifier for each printing job
- 500 - Access ID (**govinfo**)
    - The unique identifier in **govinfo**
- 511 – Names of witnesses and their affiliations
- 518 – Hearing date

To view the completed records produced through the **govinfo** API process in the CGP, click on the canned expert search wlts=govinfohrg.

- We are considering expanding the scope of the process to types of publications that we do not currently catalog, such as bills, due to the high volume of materials.
- We could generate MARC records, manually edit individual records only for known issues, such as correcting capitalization problems, and load them into OCLC at the appropriate encoding level.

Document Discovery

Federal agencies are required by statutory mandate to provide Federal publications to the FDLP (44 USC 19). For online publications, the agencies notify us by submitting metadata about the publications and URLs. This is the Document Discovery process.

We have been testing a workflow to convert the spreadsheets of publications into preliminary MARC records. We import the records into the OCLC Connexion online save file. LTS staff members then complete the records in Connexion and export the records to the CGP.

To view the completed records produced through the **govinfo** API process in the CGP, click on the canned expert search wlts=docdiscovery.

- As with the **govinfo** API process, we could generate MARC records, manually edit individual records only for known issues, such as correcting capitalization problems, and load them into OCLC at the appropriate encoding level.

Data from Various Applications of Batch Processing Methods

- NIST Collection = 3,445
- Document Discovery = 589
- **govinfo** API Process = 102

Questions and Issues for the Working Group

- Have you obtained or looked at the NIST Collection records on the CGP on GitHub repository?
- How do you think the batch processed records might impact your procedures and practices, online catalog, and access to your collections for your users?
- What would be your response to a proposal to include only SuDoc stems, not complete SuDoc numbers, in batch processed records?
- What would be your response to a proposal to omit all SuDoc forms in batch processed records?
- What would be your response to a proposal to omit item numbers in batch processed records?
- Do you think you would download sets of batch processed records from GitHub?

Documents

- Cataloging Guidelines, Bibliographic Cataloging: Overview: Cataloging/Metadata Encoding Levels Policies

Next Steps

- Finalize the batch processing standard operating procedure
- Finalize the cataloging priorities guidelines
- Test the batch processing method with several record sets, such as the Congressional Serial Set publications and the digitized historical hearings