

FOLIO Data Import Revisited

ARLEF Exploration | Jan-Feb 2024

Zoom: <https://stanford.zoom.us/j/91620626467?pwd=dVhRU1Fwb3lDWElTVHhZxOHVlWmRZz09>

See also [2024-02-28 ARLEF Data Import discussion](#)

[2023 ARLEF issues on Data Import](#)

[2023 ARLEF Call | Slide Deck](#)

[2024 January EBSCO Deck on Data Import](#)

[Data Import Topic Tracker](#)

ARLEF Institutional Reports

1. Is Data Import still an issue for you?
2. What are the specific pain points?
3. What are the highest priorities to address
 - a. Either symptoms
 - b. Or specific tickets

5 Colleges

Is Data Import still an issue for you?

Data Import continues to be an issue for the Five Colleges ([detailed report](#)).

What are the specific pain points?

1. *Performance, reliability, and stability have improved. However, we remain limited in the number of records we can reliably import (during the day 1000-2000 and after work hours 3000-6000). We are waiting for the split method to be applied which will allow us to import larger files. Even with this method, we continue to see the creation of corrupted and/or junk data, timeouts, and closed connection errors.*
2. *Matching to fields that repeat. Our pain point concerns both MARC-MARC matching and MARC to Instance matching with regard to repeating fields like 035 and Instance System Control number.*
3. *MARC modifications continue to be unreliable or lack the functionality that we need. Marc modifications can sometimes break records or not work as expected on updates. They can't be done on any fixed fields or the leader. We would like to be able to set up MARC modifications with our single record imports and haven't been able to do this on update because of the limitations.*

4. *Lack of development time and need for additional Product Owner. There is a need for at least one additional PO for Data Import. Can ARLEF sponsor a full-time PO? Also, we need a development team solely dedicated to Data Import.*

What are the highest priorities to address, either symptoms or specific tickets?

- *The lack of reporting and unreliable logs hinders our ability to troubleshoot.*
- *Add functionality to field mapping profiles where anything can be mapped from an incoming MARC file.*
- *Prioritizing importing EDIFACT invoices when using the slicing method. Currently only single record imports are prioritized.*
- *Fixing invalid SRS ids: <https://issues.folio.org/browse/MODSOURCE-300>*
- *MARC to Instance Matching [RRT]: <https://issues.folio.org/browse/MODINV-882>*
- *MARC to MARC Matching: <https://issues.folio.org/browse/MODINV-2742>*
- *Adding MARC modifications breaks records: <https://issues.folio.org/browse/MODDATAIMP-710>*
- *MARC Modifications for fixed fields, LDR, 001: <https://issues.folio.org/browse/UXPROD-4442>*
- *MARC modifications to single import overlay doesn't respect field protections: <https://issues.folio.org/browse/MODDATAIMP-897>*
- *Jobs run after canceling takes excess time: <https://issues.folio.org/browse/MODDATAIMP-970>*
- *Adjust link/unlink logic for Action profiles: <https://issues.folio.org/browse/MODDICONV-361>*
- *Improvements to user interface for job profiles*

Caltech

Is Data Import still an issue for you?

Yes. Summary below. More details [here](#).

What are the specific pain points?

- *Complexity*
- *Profiles for single record import*
- *Export OCLC Connexion records to FOLIO (we had trouble with this during implementation but will try again after Poppy)*
- ~~*Would like not to have to use UUIDs for location names and material types*~~
- *Clarity on difference between Update and Modify in interface*

What are the highest priorities to address, either symptoms or specific tickets?

- *In terms of immediately impacting everyday workflows, allowing single record import in Inventory to use non-default profiles would be most beneficial for us.*
- *In terms of developing profiles for use, clarifying Update/Modify differences in the interface would be extremely useful.*

Cornell

- Top issues:
 - <https://issues.folio.org/browse/UXPROD-2742> - Matching
 - <https://issues.folio.org/browse/MODDATAIMP-710> - Mods to MARC data as load records
 - Logging
 - Poppy migration issues including <https://issues.folio.org/browse/MODDICONV-361>

CU Boulder

- Top issues
 - Overlay MARC holdings - records can exported but not reimported
 - No bulk edit for Inventory
 - Matching - doesn't work on OCLC numbers, 035
 - Job profiles
 - Inability to use one load profile for multiple formats
 - Often freezes
 - Unable to edit/delete profiles

MSU

- Matching and Modification to MARC records through to the SRS data is the highest priority.
- Overlay MARC holdings - records can be exported but not reimported.
- Bulk Edit would be better served with a connection to Lists since that is how we envision collecting the records that need updating.

Stanford

- Full [Stanford report](#)
- Data issue is a 7/10 on our pain scale right now.
 - One of two top issues
- Top pain points:
 - matching (especially partial matches)
 - reporting / logging on data import activities

- MARC modification functionality for updating records
- Specific tickets (note that not all Stanford issues necessarily correlate to specific tickets)
 - <https://issues.folio.org/browse/UXPROD-2741>
 - <https://issues.folio.org/browse/MODSOURCE-300>
 - <https://issues.folio.org/browse/UXPROD-4348>
 - <https://issues.folio.org/browse/UXPROD-3556>
 - <https://issues.folio.org/browse/UXPROD-2741>

TAMU

- Data issue is a 3/10 on our pain scale right now.
- We spent a few years working through our process for Data Import while we were migrating from VOYAGER to FOLIO. At that time FOLIO Data Import was not a real or viable product so we wrote our own.
- For the majority of our Data Import now we use a little home grown utility that grew out of our Data migration project. But even those data import jobs are not a daily / weekly occurrence.
- We are aware of the deficiencies of Data Import and do want it to be a usable product. Especially since we are not planning on maintaining our “little home grown utility” and expect FOLIO to work.

UChicago

At Chicago, our biggest concerns are around data import infrastructure for development and support. The current model is not sustainable and the model needs to evolve in order to make data import a trusted and reliable solution for ingest of data at scale for FOLIO.

Our current data import strategy is to minimize the lift of data import using external tools to pre-process records whenever possible, thus avoiding the use of marc modifications and update features that are not a full replacement of the marc record. We have also developed a strategy to use the 035 \9 \$a to create a unique identifier constructed according to local rules that can be used for known item matching, reducing the need for complex matching at the instance / srs marc bib level. We still struggle with the ability to match and update a holdings or item record when multiples are present but are rethinking our workflows where we can to limit the need to work with multiple holdings and item records on a single instance in a single batch import job. We have still not implemented workflows that require regular data imports of large files (more than in the tens of thousands of records).

Highlights of our pain points include

- Lack of functional specifications and roadmap does not allow users to understand what functionality is available and how it is intended to work according to the development. (In

practice this has not always mapped predictably to the functional requirements outlined in the data import subgroup).

- Development continues on new functionality while not being able to support the current functionality. It is difficult to troubleshoot when data imports go awry and any fixes require a long wait since some bugs are not found until the software is installed on local systems and tested against a workflow.
- Local requirements are dismissed as out of scope for the project because they are not required by 80% of the users as determined by ????
- Specific missing or buggy functionality that continues to limit or prevent work flows that require the use of data import include the following.
 - MARC modifications not working with all imports, and does not work as expected according to use cases developed in data import subgroup.
 - job profile logic not implemented in a way that supports required matching / action combinations demanded by our work.
 - job profiles allow for non-functional profiles to be created leading to failed jobs without error messages or corrupt data without reporting it out. The UI should only allow expected and functional combinations of matching, action, and modification profiles.
 - Complex interactions with protected fields and deduplication of inventory data results in deletion of MARC data that is not intended and out of scope for what the system should do without explicit instructions from the user. (For example, <https://folio-org.atlassian.net/browse/MODDATAIMP-879>)