WOLFcon 2023

Alternative Import Workflow for Inventory Data beyond CBS

Antje Niemann, GBV Göttingen

Felix Hemme, ZBW Kiel / Hamburg

08/24/2023

CBS / CBS2FOLIO

Central Union Catalog (CBS) as common cataloguing tool

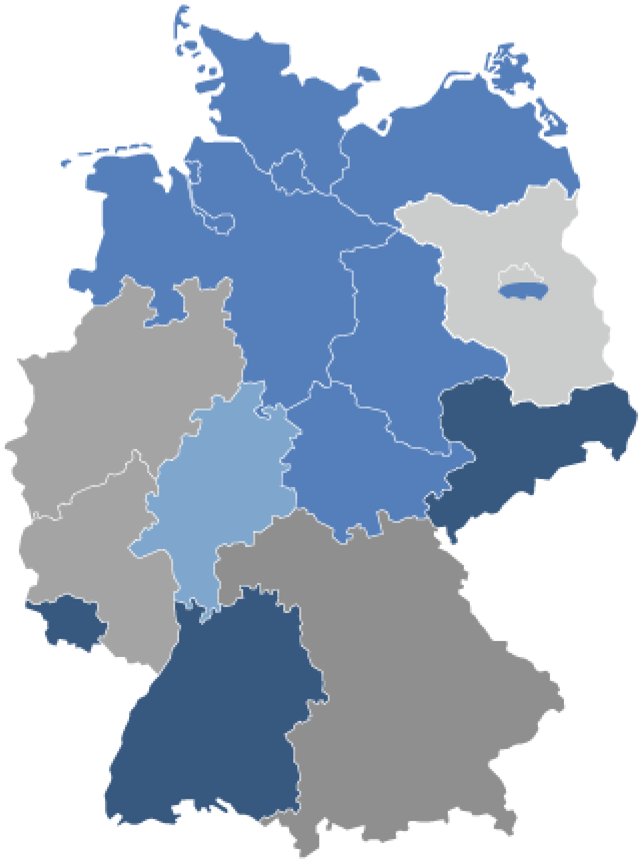
- ~ 500 German libraries
- ~ 80 million title records (~ 230 million ownerships)
- ~ 15 million authority records
- 2021: 1.3 million new records
- ~ 80 % takeover of third party bibliographic data
- In most cases just local information needs to be added (call number, barcode …)

CBS as data source for the local library systems

- Real-time update

CBS2FOLIO

- Set of components to populate FOLIO inventory storage (instances, holdings, items) with CBS metadata



GBV – Gemeinsamer Bibliotheksverbund
Common Library Network of the German States Bremen, Hamburg, Mecklenburg-Vorpommern, Niedersachsen, Sachsen-Anhalt, Schleswig-Holstein, Thüringen and the Foundation of Prussian Cultural Heritage (middle blue)

SWB – Common Library Network of the German States Baden-Württemberg, Saarland, Sachsen (dark blue)

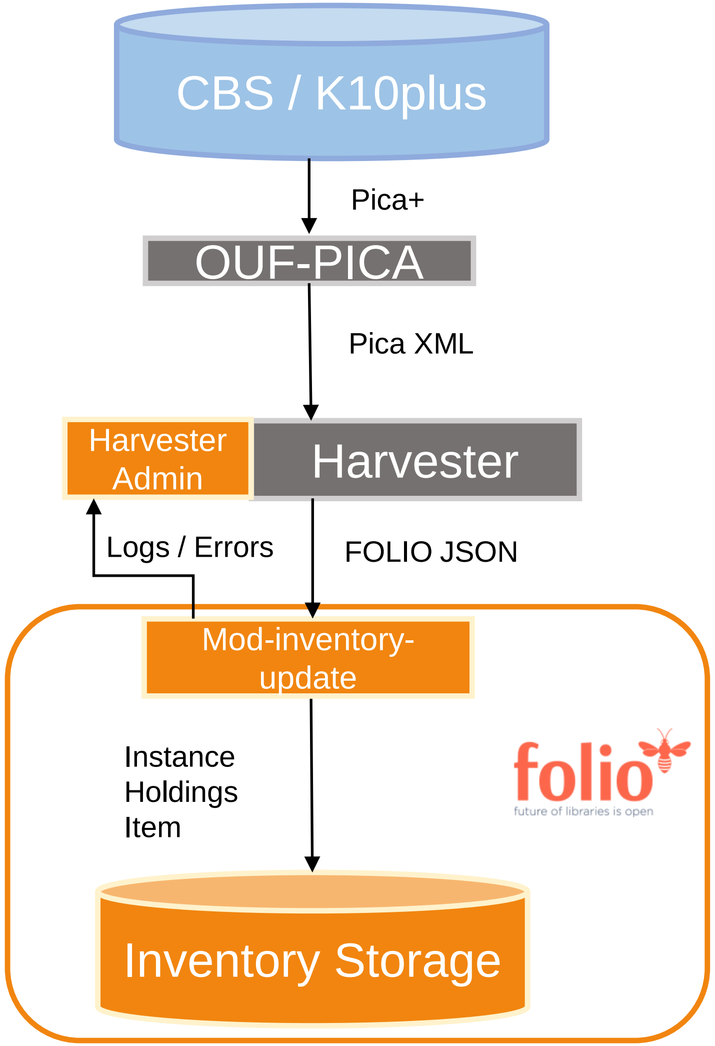
Library Systems in GBV Consortium

- CBS
  - Shared cataloguing
  - Interlibrary loan
  - Metadata source for Discovery

- Local Library System (currently OCLC LBS4, future FOLIO)
  - ERM
  - Acquisition
  - Circulation

- Not all CBS data is relevant for FOLIO
  - Selection of fields and records -> no authority records, no subject headings
  - Just a minimum of relations between different titles

Import Workflow from CBS to FOLIO Inventory Storage



CBS / K10plus

Pica+

OUF-PICA

Pica XML

Harvester Admin

Harvester

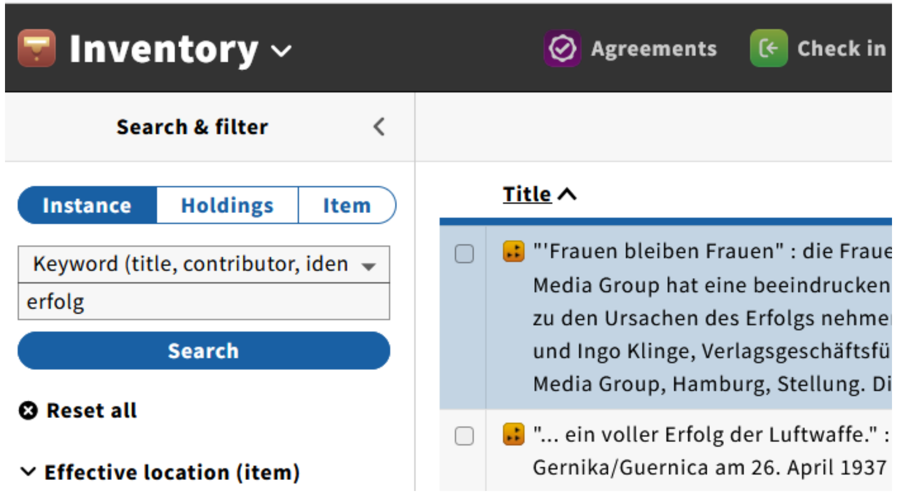
Logs / Errors

FOLIO JSON

Mod-inventory-update

Instance
Holdings
Item

Inventory Storage

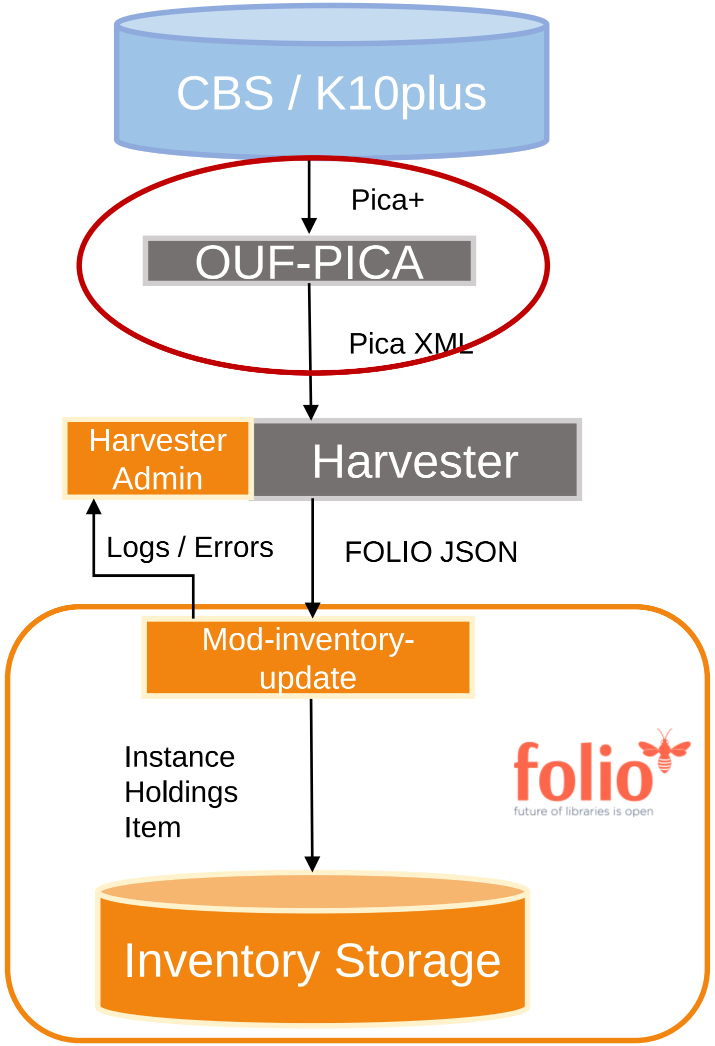


erfolg

□ "'Frauen bleiben Frauen" : die Fraue
Media Group hat eine beeindrucken
zu den Ursachen des Erfolgs nehme
und Ingo Klinge, Verlagsgeschäftsfü
Media Group, Hamburg, Stellung. Di

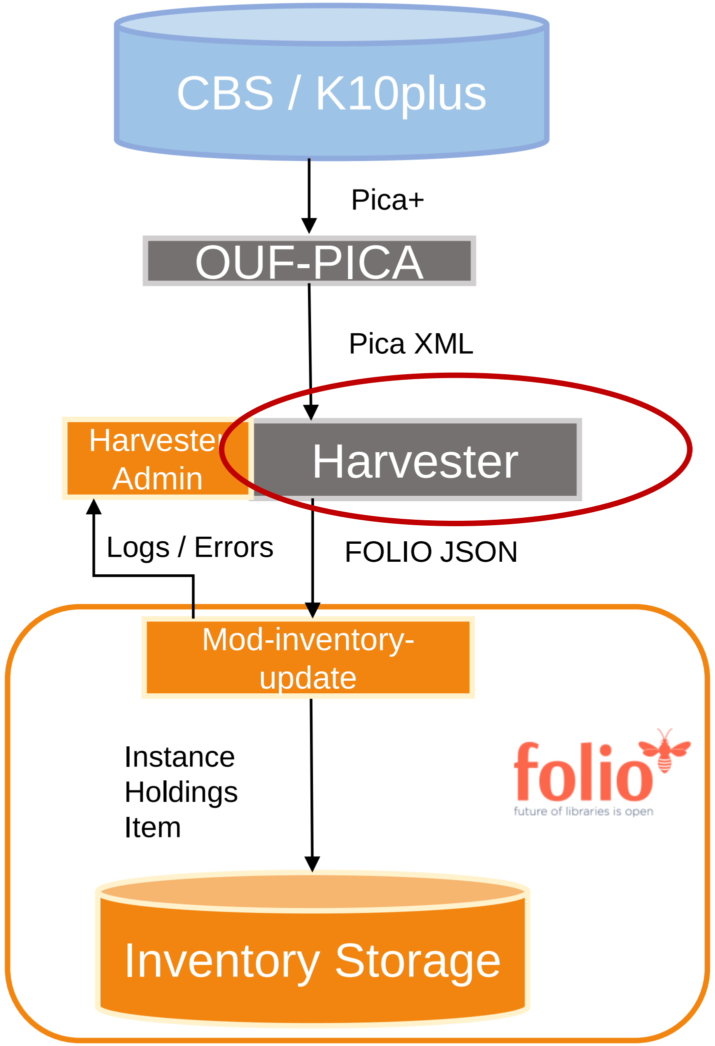
□ "... ein voller Erfolg der Luftwaffe." :
Gernika/Guernica am 26. April 1937

Import Workflow from CBS to FOLIO Inventory Storage



- OUF-PICA
  - fetches records from CBS using the CBS OUF tools
  - calculates the record status (upsert or delete)
  - converts Pica+ to XML
  - controls the Index Data Harvester

Import Workflow from CBS to FOLIO Inventory Storage

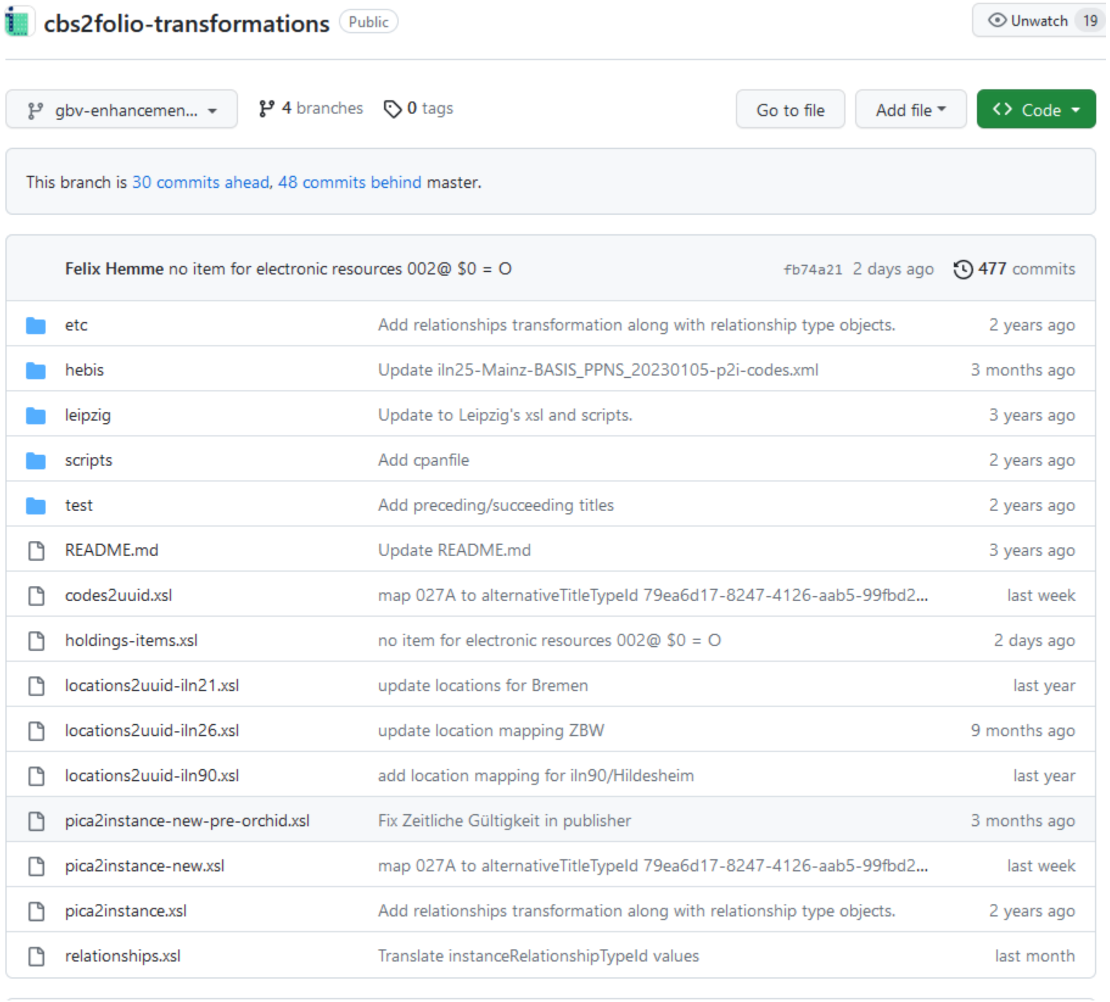


- Harvester / localindices
  - Its primary use is harvesting of bibliographic records and its holdings
  - Can read data from a variety of data sources
  - Transforms the data through highly configurable XSLT based transformation steps and pipelines
  - Stores the transformed data to storage systems like Solr databases or **FOLIO Inventory**
  - Harvest job definitions, scheduling, and transformation pipelines are configured in a MySQL database
  - https://github.com/indexdata/localindices

Harvester: Transformation via XSLT

Example for XSLT transformation steps

https://github.com/indexdata/cbs2folio-transformations



cbs2folio-transformations Public

Unwatch 19

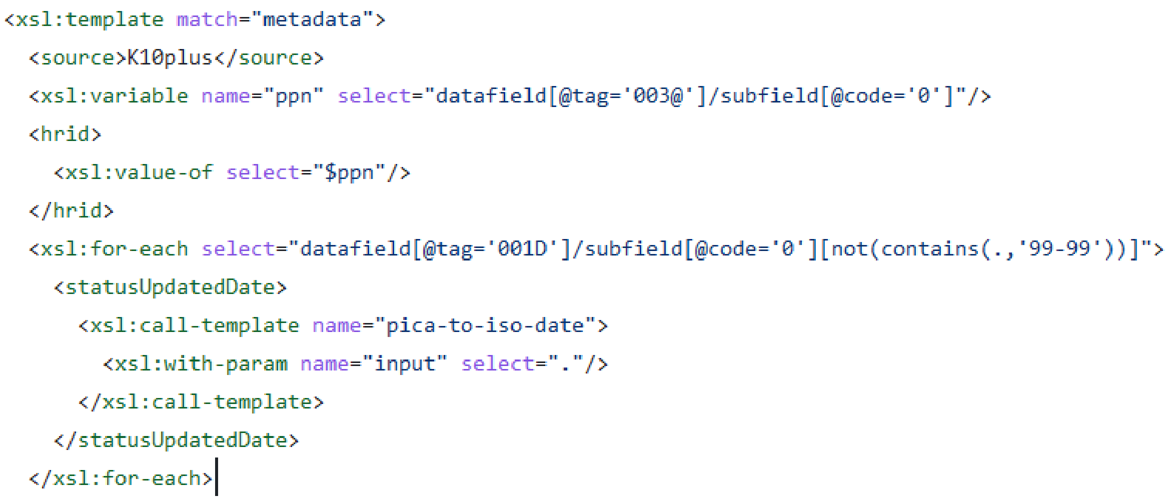
gbv-enhancemen... ▾ 4 branches 0 tags Go to file Add file ▾ Code ▾

This branch is 30 commits ahead, 48 commits behind master.

Felix Hemme no item for electronic resources 002@ $0 = O fb74a21 2 days ago 477 commits

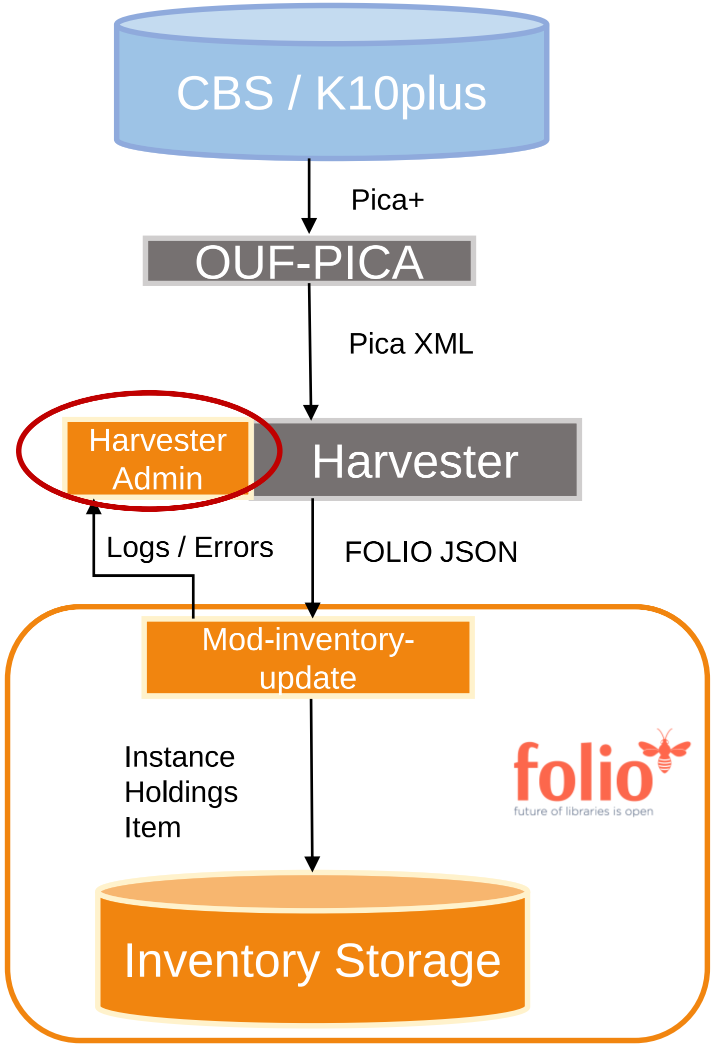
etc	Add relationships transformation along with relationship type objects.	2 years ago
hebis	Update iln25-Mainz-BASIS_PPNS_20230105-p2i-codes.xml	3 months ago
leipzig	Update to Leipzig's xsl and scripts.	3 years ago
scripts	Add cpanfile	2 years ago
test	Add preceding/succeeding titles	2 years ago
README.md	Update README.md	3 years ago
codes2uuid.xsl	map 027A to alternativeTitleTypeId 79ea6d17-8247-4126-aab5-99fbd2...	last week
holdings-items.xsl	no item for electronic resources 002@ $0 = O	2 days ago
locations2uuid-iln21.xsl	update locations for Bremen	last year
locations2uuid-iln26.xsl	update location mapping ZBW	9 months ago
locations2uuid-iln90.xsl	add location mapping for iln90/Hildesheim	last year
pica2instance-new-pre-orchid.xsl	Fix Zeitliche Gültigkeit in publisher	3 months ago
pica2instance-new.xsl	map 027A to alternativeTitleTypeId 79ea6d17-8247-4126-aab5-99fbd2...	last week
pica2instance.xsl	Add relationships transformation along with relationship type objects.	2 years ago
relationships.xsl	Translate instanceRelationshipTypeId values	last month

Excert from pica2instance.xsl (source and hrid)



```
<xsl:template match="metadata">
  <source>K10plus</source>
  <xsl:variable name="ppn" select="datafield[@tag='003@']/subfield[@code='0']"/>
  <hrid>
    <xsl:value-of select="$ppn"/>
  </hrid>
  <xsl:for-each select="datafield[@tag='001D']/subfield[@code='0'][not(contains(.,'99-99'))]">
    <statusUpdatedDate>
      <xsl:call-template name="pica-to-iso-date">
        <xsl:with-param name="input" select="."/>
      </xsl:call-template>
    </statusUpdatedDate>
  </xsl:for-each>
```

Import Workflow from CBS to FOLIO Inventory Storage

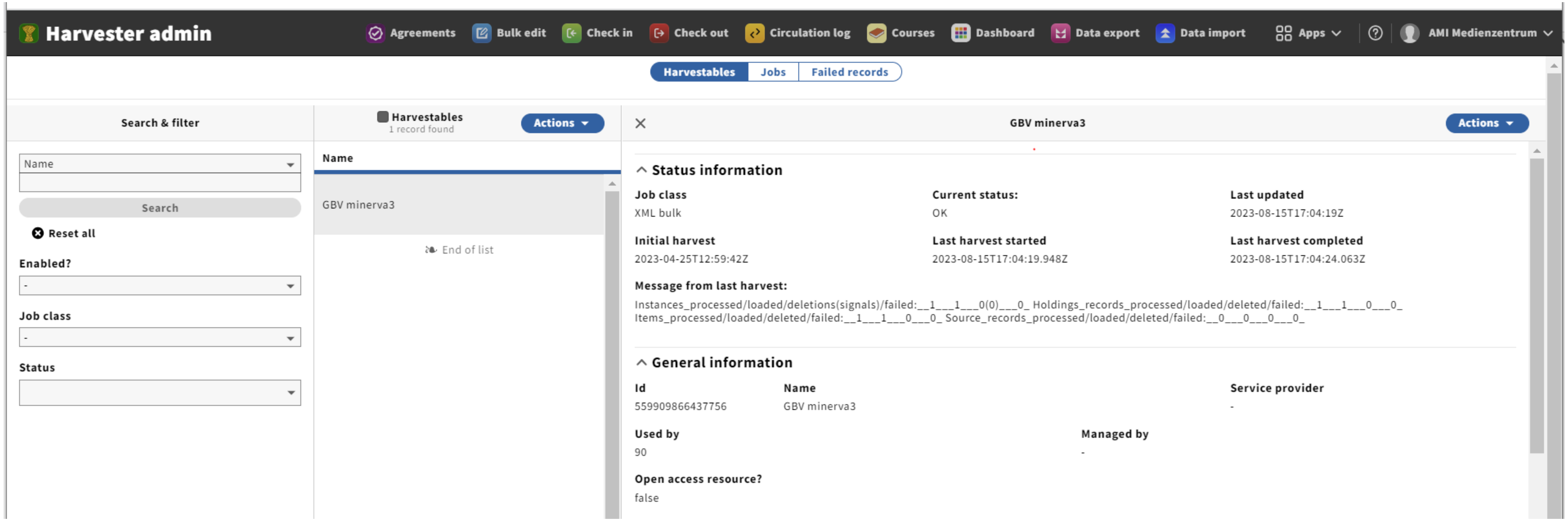


- mod-harvester-admin
  - Okapi service that can be put in front of Harvester
  - Provides FOLIO based access to control the Harvester
  - https://github.com/indexdata/mod-harvester-admin

- ui-harvester-admin
  - Provides an FOLIO/JSON based interface to the configuration database that FOLIO clients (like a Stripes UI) can use
  - https://github.com/indexdata/ui-harvester-admin

mod-harvester-admin / ui-harvester-admin

- Harvester-admin provides an FOLIO/JSON based interface to the **configuration** for managing harvest jobs (Harvestables)



Status information

Job class
XML bulk

Current status:
OK

Last updated
2023-08-15T17:04:19Z

Initial harvest
2023-04-25T12:59:42Z

Last harvest started
2023-08-15T17:04:19.948Z

Last harvest completed
2023-08-15T17:04:24.063Z

Message from last harvest:
Instances_processed/loaded/deletions(signals)/failed:__1___1___0(0)___0_ Holdings_records_processed/loaded/deleted/failed:__1___1___0___0_
Items_processed/loaded/deleted/failed:__1___1___0___0_ Source_records_processed/loaded/deleted/failed:__0___0___0___0_

General information

Id
559909866437756

Name
GBV minerva3

Service provider
-

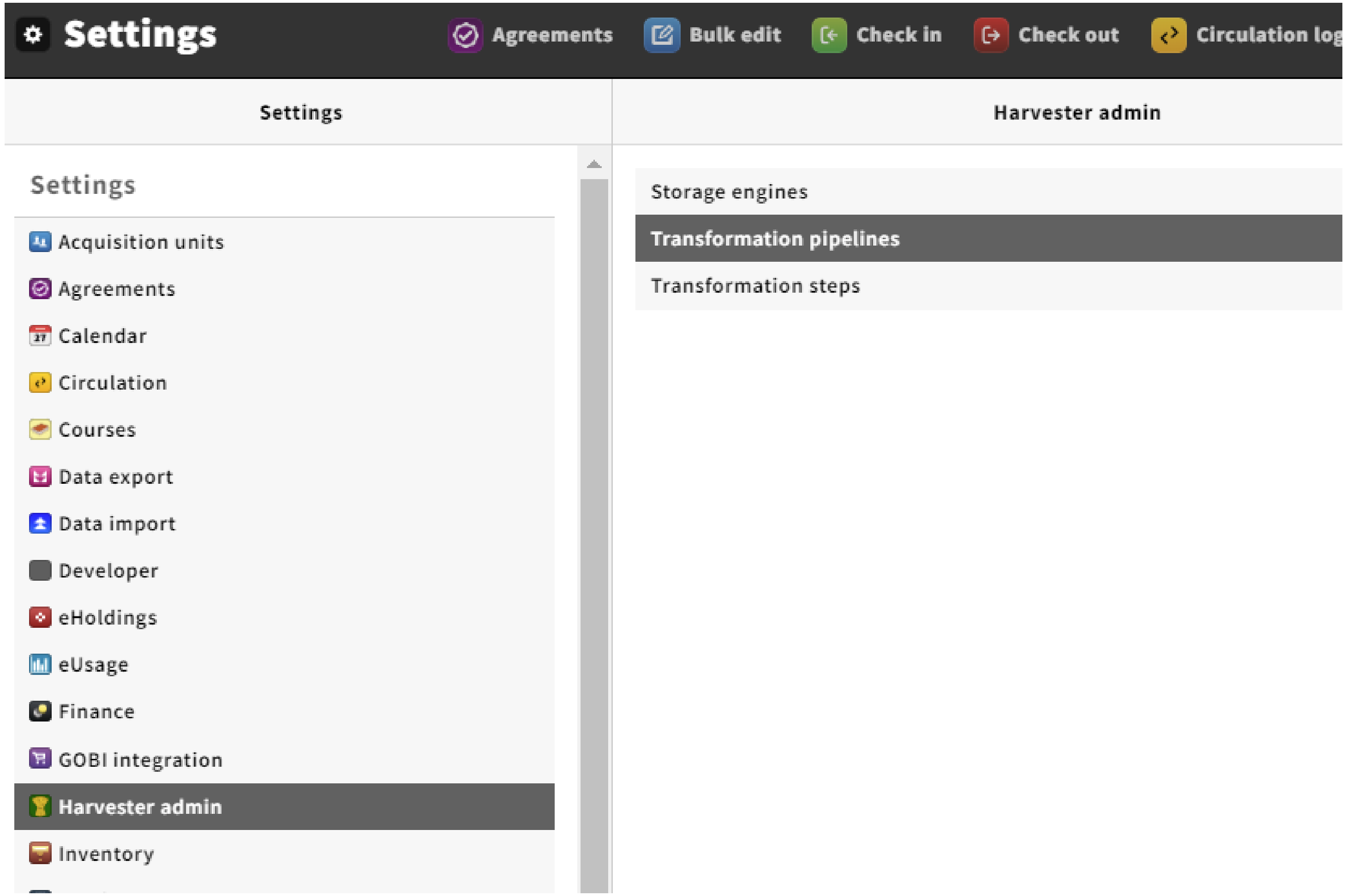
Used by
90

Managed by
-

Open access resource?
false

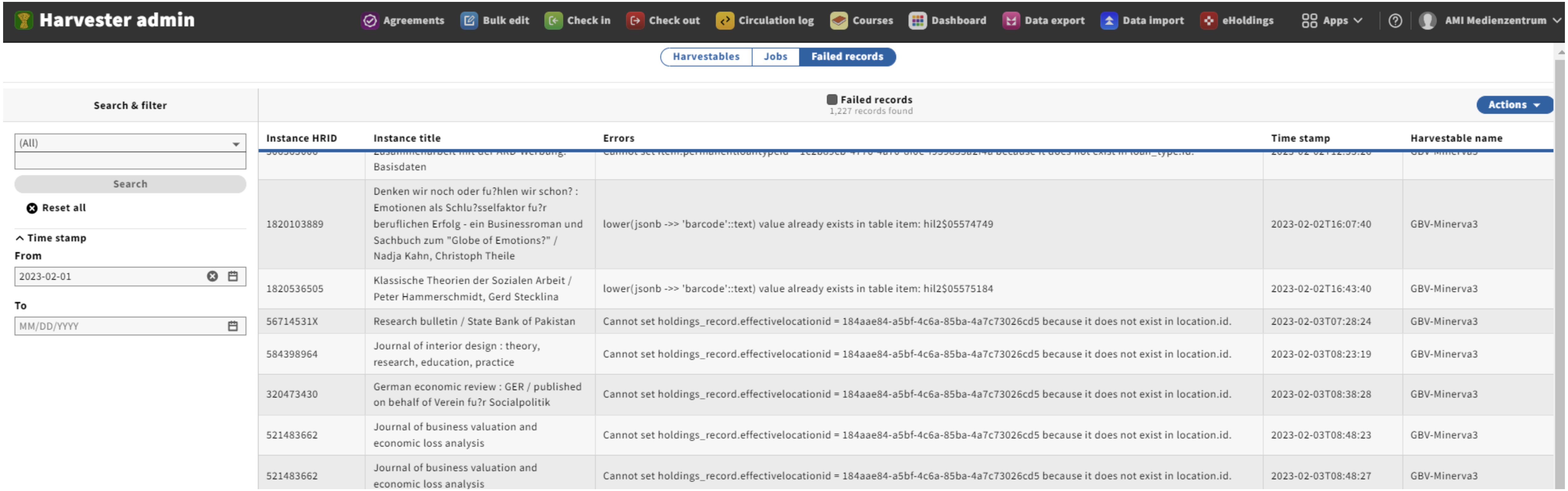
mod-harvester-admin / ui-harvester-admin

- … for managing the
  - Storage Engines
  - Transformation Pipelines and
  - Transformation Steps



mod-harvester-admin / ui-harvester-admin

- … for monitoring logs and error reporting



(All)

Search

Reset all

Time stamp

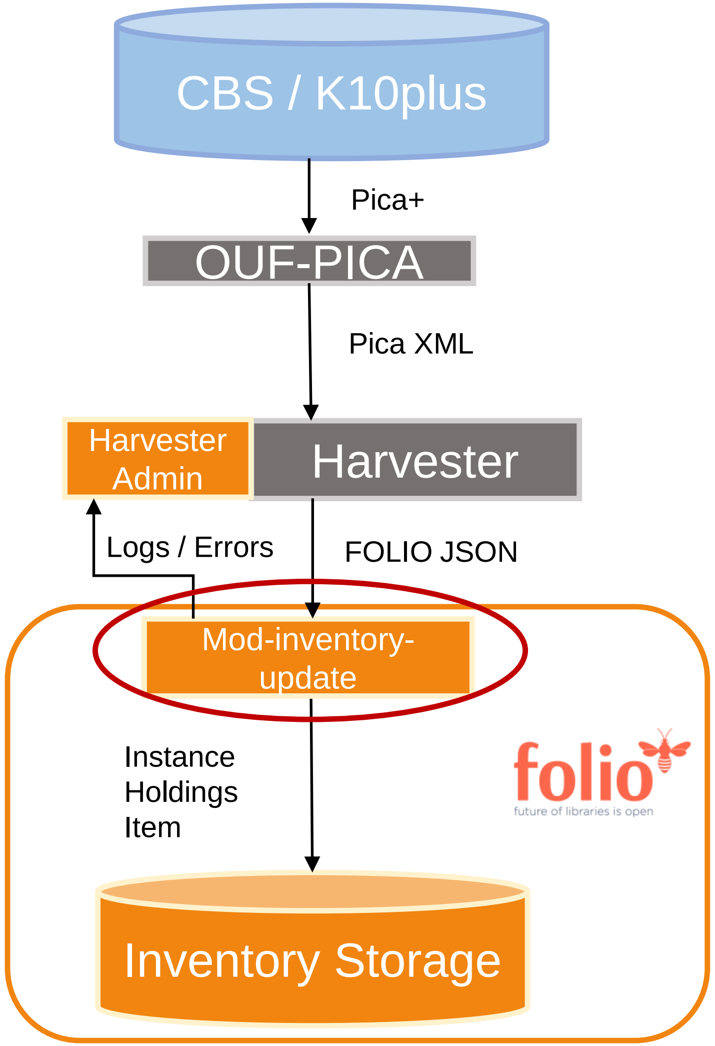
From: 2023-02-01

To: MM/DD/YYYY

Failed records
1,227 records found

Instance HRID	Instance title	Errors	Time stamp	Harvestable name
	Basisdaten			
1820103889	Denken wir noch oder fu?hlen wir schon? : Emotionen als Schlu?sselfaktor fu?r beruflichen Erfolg - ein Businessroman und Sachbuch zum "Globe of Emotions?" / Nadja Kahn, Christoph Theile	lower(jsonb ->> 'barcode'::text) value already exists in table item: hil2$05574749	2023-02-02T16:07:40	GBV-Minerva3
1820536505	Klassische Theorien der Sozialen Arbeit / Peter Hammerschmidt, Gerd Stecklina	lower(jsonb ->> 'barcode'::text) value already exists in table item: hil2$05575184	2023-02-02T16:43:40	GBV-Minerva3
56714531X	Research bulletin / State Bank of Pakistan	Cannot set holdings_record.effectivelocationid = 184aae84-a5bf-4c6a-85ba-4a7c73026cd5 because it does not exist in location.id.	2023-02-03T07:28:24	GBV-Minerva3
584398964	Journal of interior design : theory, research, education, practice	Cannot set holdings_record.effectivelocationid = 184aae84-a5bf-4c6a-85ba-4a7c73026cd5 because it does not exist in location.id.	2023-02-03T08:23:19	GBV-Minerva3
320473430	German economic review : GER / published on behalf of Verein fu?r Socialpolitik	Cannot set holdings_record.effectivelocationid = 184aae84-a5bf-4c6a-85ba-4a7c73026cd5 because it does not exist in location.id.	2023-02-03T08:38:28	GBV-Minerva3
521483662	Journal of business valuation and economic loss analysis	Cannot set holdings_record.effectivelocationid = 184aae84-a5bf-4c6a-85ba-4a7c73026cd5 because it does not exist in location.id.	2023-02-03T08:48:23	GBV-Minerva3
521483662	Journal of business valuation and economic loss analysis	Cannot set holdings_record.effectivelocationid = 184aae84-a5bf-4c6a-85ba-4a7c73026cd5 because it does not exist in location.id.	2023-02-03T08:48:27	GBV-Minerva3

Import Workflow from CBS to FOLIO Inventory Storage



- mod-inventory-update (MIU)
  - Okapi service in front of mod-inventory-storage (Inventory Storage) for populating the storage with instances, holdings and items
  - https://github.com/folio-org/mod-inventory-update

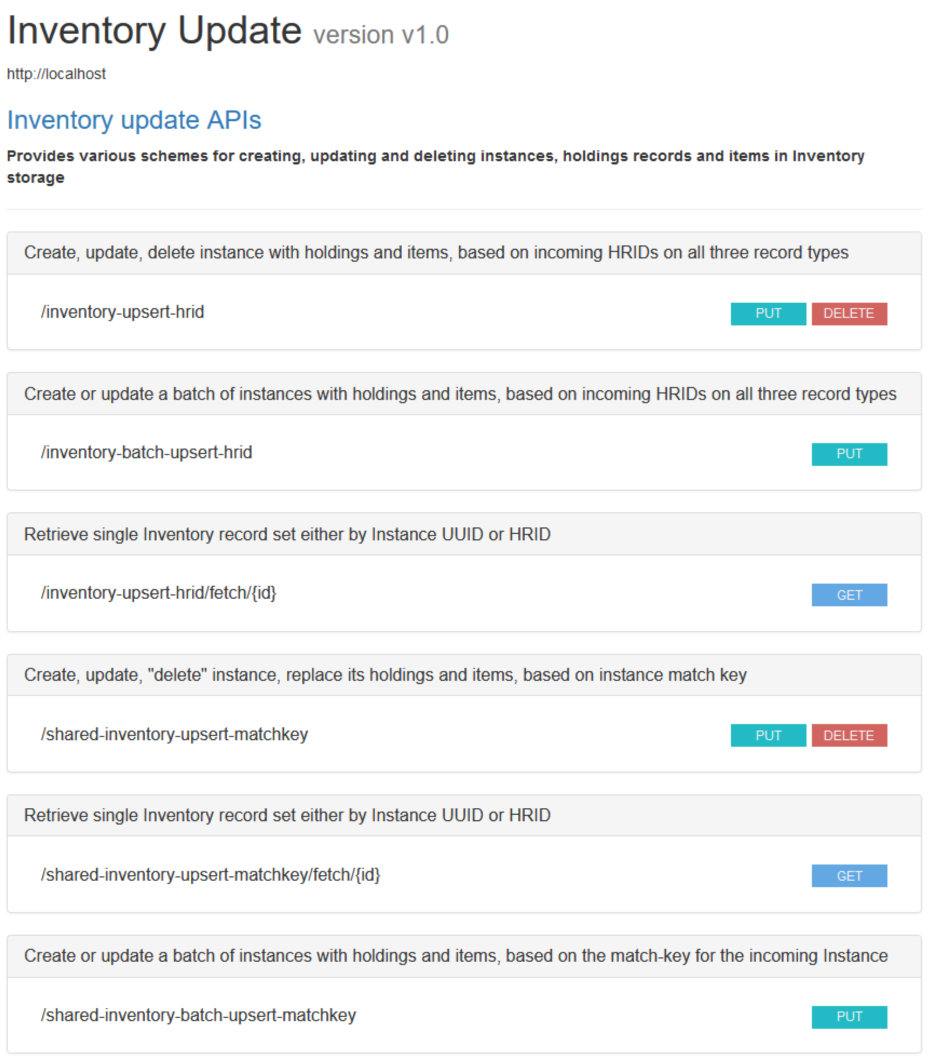
mod-inventory-update (MIU)

- Accepts GET, PUT and DELETE requests

- 2 matchkeys (each also with batch process)
  - inventory-upsert-hrid /
    inventory-batch-upsert-hrid (GBV)
  - shared-inventory-upsert-matchkey /
    shared-inventory-batch-upsert-matchkey

- Provisional instance created when related instance does not exist yet

- Control record overlay on updates / Prevent MIU from overriding existing values



Inventory Update version v1.0

http://localhost

Inventory update APIs

Provides various schemes for creating, updating and deleting instances, holdings records and items in Inventory storage

Create, update, delete instance with holdings and items, based on incoming HRIDs on all three record types

/inventory-upsert-hrid PUT DELETE

Create or update a batch of instances with holdings and items, based on incoming HRIDs on all three record types

/inventory-batch-upsert-hrid PUT

Retrieve single Inventory record set either by Instance UUID or HRID

/inventory-upsert-hrid/fetch/{id} GET

Create, update, "delete" instance, replace its holdings and items, based on instance match key

/shared-inventory-upsert-matchkey PUT DELETE

Retrieve single Inventory record set either by Instance UUID or HRID

/shared-inventory-upsert-matchkey/fetch/{id} GET

Create or update a batch of instances with holdings and items, based on the match-key for the incoming Instance

/shared-inventory-batch-upsert-matchkey PUT

https://s3.amazonaws.com/foliodocs/api/mod-inventory-update/r/inventory-update.html

Performance and Scalability

- Stable and sufficiently fast processes for the initial loading of a tenant’s data and the real-time update

- Example: State and University Library Bremen

  - Initial loading: 6.75 hours for 18.8 million instances, holdings and items
    Average: 773 records / second

  - Real-time update: 2 hours for 1,5 months of changes in the CBS production system

CBS2FOLIO in a nutshell

- Functionality to import non-marc records into inventory
  - Record types such as json or xml are possible
  - 24/7 real-time update
  - Consistent results / controlled overriding of existing values
- Good Performance and Scalability
- Flexible Mapping by XSLT
- 2 matching processes (HRID or matchkey, each also with batch process)
- All CRUD processes are implemented (create, read, update, delete)
- Logging that allows the user to troubleshoot
  - Identifiers are given
  - Error messages are understandable
  - Clean up of the log file planned
  - Information about a hanging job and the last loaded record
- No connection to SRS yet, no authority records, minimum of relations between records

CBS2FOLIO -> thirdPartySystem2FOLIO?

- The software we developed provides the ability to connect a CBS based union catalog to FOLIO Inventory, but is not limited to CBS
- Let’s take a deeper look at some of the components and possible scenarios

Import scenario

- Functionality to import MARC and non-MARC records
- MARC should be provided in XML serialization
- No limit to file size. Harvester can be configured to spilt large files at defined number of records
- Can load in parallel using multiple *harvestables* at once
  - A *harvestable* is a job configuration that holds information about the transformation pipeline, storage, log level, URL to monitor
  - Can be used just once or multiple times, depending on the use case
- Create, update, and delete *Inventory record sets*
  - An *Inventory record set* is a set of records including an Inventory instance, and an array of holdings records with embedded arrays of items

Managing transformations

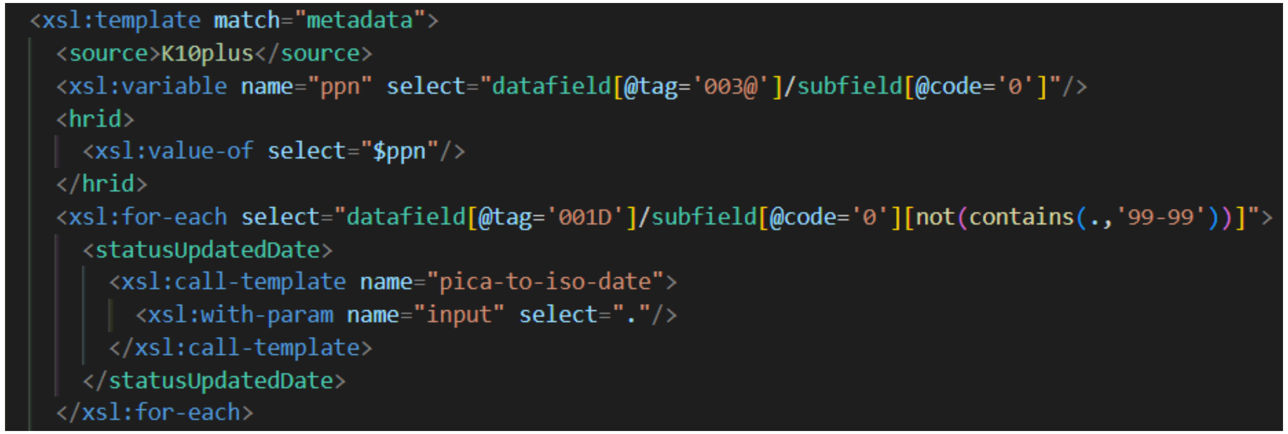
- Create XSLT mappings for MARC files
- XSLT can be shared and reused across libraries using external services like GitHub or GitLab, including version control
- This enables a collaborative workflow of managing mappings and the technical conversion

Harvester: Transformation via XSLT

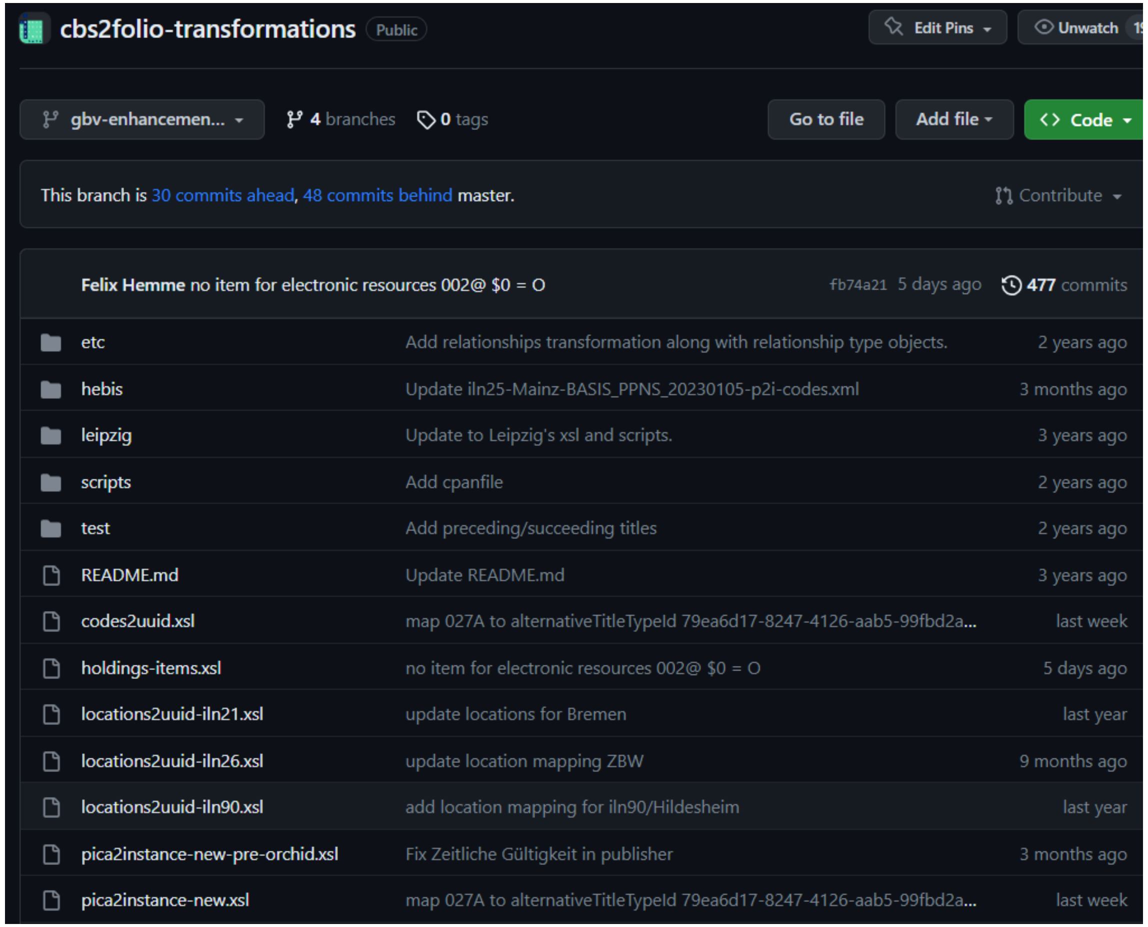
Example for XSLT transformation steps

https://github.com/indexdata/cbs2folio-transformations

Excerpt from pica2instance-new.xsl (source and hrid)



```
<xsl:template match="metadata">
  <source>K10plus</source>
  <xsl:variable name="ppn" select="datafield[@tag='003@']/subfield[@code='0']"/>
  <hrid>
    <xsl:value-of select="$ppn"/>
  </hrid>
  <xsl:for-each select="datafield[@tag='001D']/subfield[@code='0'][not(contains(.,'99-99'))]">
    <statusUpdatedDate>
      <xsl:call-template name="pica-to-iso-date">
        <xsl:with-param name="input" select="."/>
      </xsl:call-template>
    </statusUpdatedDate>
  </xsl:for-each>
```



This branch is 30 commits ahead, 48 commits behind master.

Felix Hemme no item for electronic resources 002@ $0 = O fb74a21 5 days ago 477 commits

etc	Add relationships transformation along with relationship type objects.	2 years ago
hebis	Update iln25-Mainz-BASIS_PPNS_20230105-p2i-codes.xml	3 months ago
leipzig	Update to Leipzig's xsl and scripts.	3 years ago
scripts	Add cpanfile	2 years ago
test	Add preceding/succeeding titles	2 years ago
README.md	Update README.md	3 years ago
codes2uuid.xsl	map 027A to alternativeTitleTypeId 79ea6d17-8247-4126-aab5-99fbd2a...	last week
holdings-items.xsl	no item for electronic resources 002@ $0 = O	5 days ago
locations2uuid-iln21.xsl	update locations for Bremen	last year
locations2uuid-iln26.xsl	update location mapping ZBW	9 months ago
locations2uuid-iln90.xsl	add location mapping for iln90/Hildesheim	last year
pica2instance-new-pre-orchid.xsl	Fix Zeitliche Gültigkeit in publisher	3 months ago
pica2instance-new.xsl	map 027A to alternativeTitleTypeId 79ea6d17-8247-4126-aab5-99fbd2a...	last week

A look at the Harvester admin FOLIO app

Improvements

Some areas of interest might be:

- SRS connection?
- Matchkeys?
- Testloads?
- Use cases?

SRS connection

- Reminder: No connection to SRS yet
- MIU populates into mod-inventory-storage directly
- The Harvester can store original MARC records in a given storage
- SRS records link to their Inventory equivalent by storing their UUIDs in 999's – would need to look up the UUIDs after an import
- Unknowns:
  - Performance when populating SRS
  - Actions taken by SRS on Inventory records

Matchkey methods

- Two matchkeys implemented to match on HRID or an matchkey
- Would potentially need to enhance matchkeys to support matching on ID‘s like OCLC ID in MARC 035$a/$z and on other fields in the Inventory records
- Investigate the need for multiple matchkeys with if/else conditions

- Example:

```
if 035$z matches instanceIdentifierTypeId abc
  then update the instance
else if 001 matches hrid
  then update the instance
otherwise do nothing
```

Testloads

- Investiage a dry run functionality
- Perform a testload to see if record matching would work as expected
- Preview some statistical data, e.g.
  - Records matched
  - Records created
  - Records updated
  - Records deleted
  - Errors

Summary

- MIU and MHA can be used to connect a CBS based union catalog to Inventory, but they are not limited to CBS as a source.
- The tools are format agnostic; they rely on XSLT transformations and can convert data that is provided in an XML (MARC XML, DC XML, PICA+ XML etc.) or JSON format
- MIU has proven to be reliable when it comes to loading millions of instances, holdings, and items during migration. It is also performant when loading batches of records on a daily basis.
- There is no connection to SRS yet. MIU is connected to mod-inventory-storage directly. If there is interest in pushing (MARC) data into SRS, the workflow has to be enhanced.

Thank you!

antje.niemann@gbv.de
f.hemme@zbw-online.eu

Decisive for this presentation is the spoken word.