

Digital Preservation Workflow Webinars 2023

eHumanities Workflow

Jaime Penagos



Outline

1. What is eHumanities / Discover?
2. Implementation of Workflows
3. Challenges & Further steps



Outline (detailed)

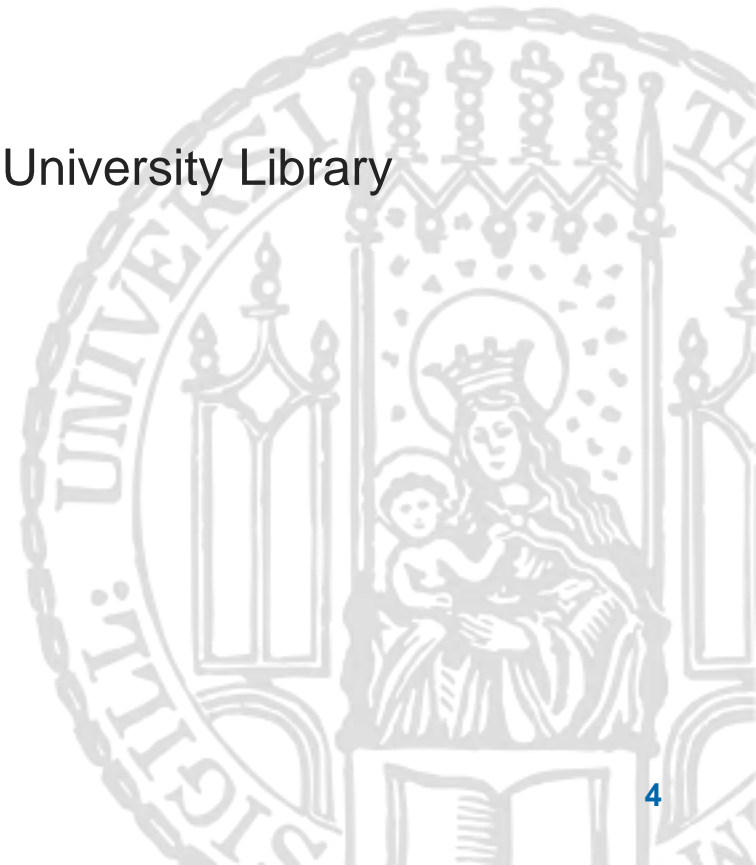
- Goal
- Data overview
- Design choices and tools
- Implementation



Research data management at University Library LMU Munich

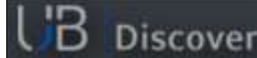
Context: Project "eHumanities - interdisciplinary"


- Project from the IT-Group Humanities LMU, University Library LMU and University Library of Erlangen-Nuremberg (FAU)
- Phase 1 (April 2018 – March 2021)
- Phase 2 (April 2021 – March 2023)



Discover

<https://discover.ub.uni-muenchen.de>





Quick Access

☐ SHOW ONLY LATEST VERSION


[Collection](#)

[Creator](#)

[Faculty](#)


[Hierarchy](#)

Discover provides access to research data from projects related to LMU Munich.

 If you need more information about how to browse on this site, take a look at our [help page](#).

[Learn more](#) about research data in general and further research data management services of the University Library of LMU Munich.


Discover...



Verba Alpina

Verba Alpina investigates the Alpine region, which is linguistically highly fragmented, in its historico-cultural and historical linguistic unity in a selective and analytical way.

[Search all](#)[Learn more about Verba Alpina](#)


Universitätsbibliothek


Open Data LMU

With Open Data LMU, the University Library LMU provides a platform for the publication of research data.

[Search all](#)[Learn more about Open Data LMU](#)

5





Start Over
Collection Str > VerbaAlpina
✕

Limit your search

☐ SHOW ONLY LATEST VERSION

Collection

>

Contributor

>

Creator

>

DDC

>

Faculty

>

Format

>

Hierarchy

>

Keyword

>

Language

>

Location

>

Rights

>

Year of Publication

>

« Previous | 1 - 10 of 222,630 | Next »

Sort by relevance ▾

10 per page ▾

1. VerbaAlpina Version 19/1

Creator / Author:

Krefeld, Thomas
Lücke, Stephan

Year of Publication:

2021

Version:

19/1

DDC:

410 Linguistics
430 Germanic languages German
450 Italian, Romanian, Rhaeto-Romantic
004 Data processing computer science

Keywords:

High German
Shifted Western Romance
Western South Slavic
Sprachatlas
linguistic map (...)

DOWNLOAD

2. VerbaAlpina Version 19/2

Creator / Author:

Krefeld, Thomas
Lücke, Stephan

Year of Publication:

2020

Version:

19/2

DDC:

410 Linguistics
430 Germanic languages German
450 Italian, Romanian, Rhaeto-Romantic
004 Data processing computer science

Keywords:

High German
Shifted Western Romance
Western South Slavic
Sprachatlas
linguistic map (...)

DOWNLOAD

UB Discover

Search...

Q

Start Over







Back to Search

< Previous | 3 of 222,630 | Next >





VerbaAlpina instances for the morphological type "Teie(n) (gem.)"

Previous Version



MAIN INFORMATION

Creator/Author:	Krefeld, Thomas   [Institut für Romanische Philologie, Ludwig-Maximilians-Universität München]
	Lücke, Stephan   [IT-Gruppe Geisteswissenschaften (ITG), Ludwig-Maximilians-Universität München]
Contributors:	show 
Year of Publication:	2021
Version:	19/2
Funding:	Deutsche Forschungsgemeinschaft (DFG): 253900505 
Faculty:	Faculty for Languages and Literatures


RELATIONS

Is Part Of:	 START SEARCH
Has Part:	 START SEARCH
Is Identical To:	Find Record @ Verba Alpina 
Is Variant Form Of:	Find Record @ Verba Alpina 

CONTENT-RELATED INFORMATION

Abstract:	Contains 118 VerbaAlpina datasets, which are related to the morphological type "Teie(n) (gem.)".
Keywords:	ALMHÜTTE (GEBÄUDE, EINFACH, BEWIRTSCHAFTET, AUF DER ALM)  STADEL (HÜTTE, FÜR HEU, AUF DER ALM ODER AUF DER WIESE) SENNHÜTTE (GEBÄUDE, EINFACH, AUF DER ALM, ZUR VERARBEITUNG VON MILCH)  KÄSERAUM (RAUM ZUM LAGERN VON KÄSE, ANGEBAUT)


DOWNLOAD DATASET


RIGHTS: CC-BY-SA-4.0 

SIZE: 2485258

FORMAT: TEXT/CSV

CITATION

PID: [68fd5294-9077-3983-a20e-7f25c0](#) 

REQUEST 

CITE

DOWNLOAD BIBTEX

TOOLS

DOWNLOAD METADATA

E-MAIL RECORD

Discover (data overview)

- Verba Alpina (<https://www.verba-alpina.gwi.uni-muenchen.de/>)
 - CSV (Research Data), XML (raw metadata) (~ 450k files)
- OpenData LMU (<https://data.ub.uni-muenchen.de/>)
 - Platform for Research Data publication established in 2010 (based on EPrints) (~150 files)

Discover (data overview)



- Data has complex relationships and versioning across files
- Not a big set of files (~ 50 GB from text-based files), but each new version has around 250k new files (scalability)
- Connections between the files makes the ingest and pre-processing non trivial

Discover (design choices)



- Framework based on open source systems
- Search portal hosting different projects
- Modularity
- Scalability
- Own metadata schema


Discover (design choices)

Internal data format of Discover

- Based on DataCite, extending some properties that describe the specific needs of this project / our use cases.
- Fields like: *hierarchy*, *currentVersion*, *contentUrls*, *metadataUrl*, *checksum*, ... among others

- Structure:

```
<lmUB:rData>  
  <datacite:resource>  
    <datacite:identifier>  
    ...  
  </datacite:resource>  
  ...  
</lmUB:rData>
```



DataCite-Record

- Documentation: <https://github.com/UB-LMU/rdUB>

Discover (design choices)

Navigation through different versions of the same Dataset



VerbaAlpina instances for the morphological type "Teie(r
(gem.)"

Previous Version

MAIN INFORMATION

Creator/Author: Krefeld, Thomas
[Institut für Romanische Philologie, Ludwig-Maximilians-Universität Münche

Lücke, Stephan
[IT-Gruppe Geisteswissenschaften (ITG), Ludwig-Maximilians-Universität M

Contributors: show ▾

Year of Publication: 2021

Version: 19/2

Funding: Deutsche Forschungsgemeinschaft (DFG): 253900505

Faculty: Faculty for Languages and Literatures

RELATIONS

Is Part Of: START SEARCH

Has Part: START SEARCH

Is Identical To: Find Record @ Verba Alpina

Is Variant Form Of: Find Record @ Verba Alpina

Discover (design choices)

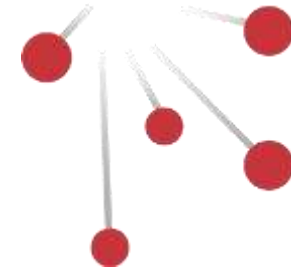


- Versioning
- Unique identifiers across all systems
- Automation of ingest tasks and index updates

Discover (tools)

FEDORA (Flexible Extensible Digital Object Repository Architecture)

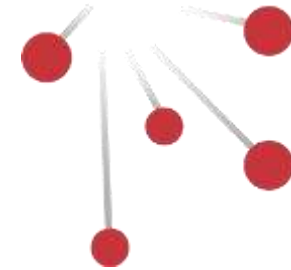
- Open source repository
 - REST interface
 - Linked Data Platform (LDP)
 - Web Access Control (Solid / WebAC)
 - Memento
 - Activity Streams 2.0
-
- Persistent content stored on disk using OCFL (Oxford Common File Layout)



Discover (tools)

FEDORA (Flexible Extensible Digital Object Repository Architecture)

- Open source repository
- REST interface
- Linked Data Platform (LDP)
- Web Access Control (Solid / WebAC)
- Memento
- Activity Streams 2.0
- Persistent content stored on disk using OCFL (Oxford Common File Layout)
 - Application independent approach to the storage of digital information in a structured, transparent, and predictable manner



Discover (tools)



Apache Camel

- Open source integration framework
- Enterprise Integration Patterns (EIP)
- Java

Project Blacklight

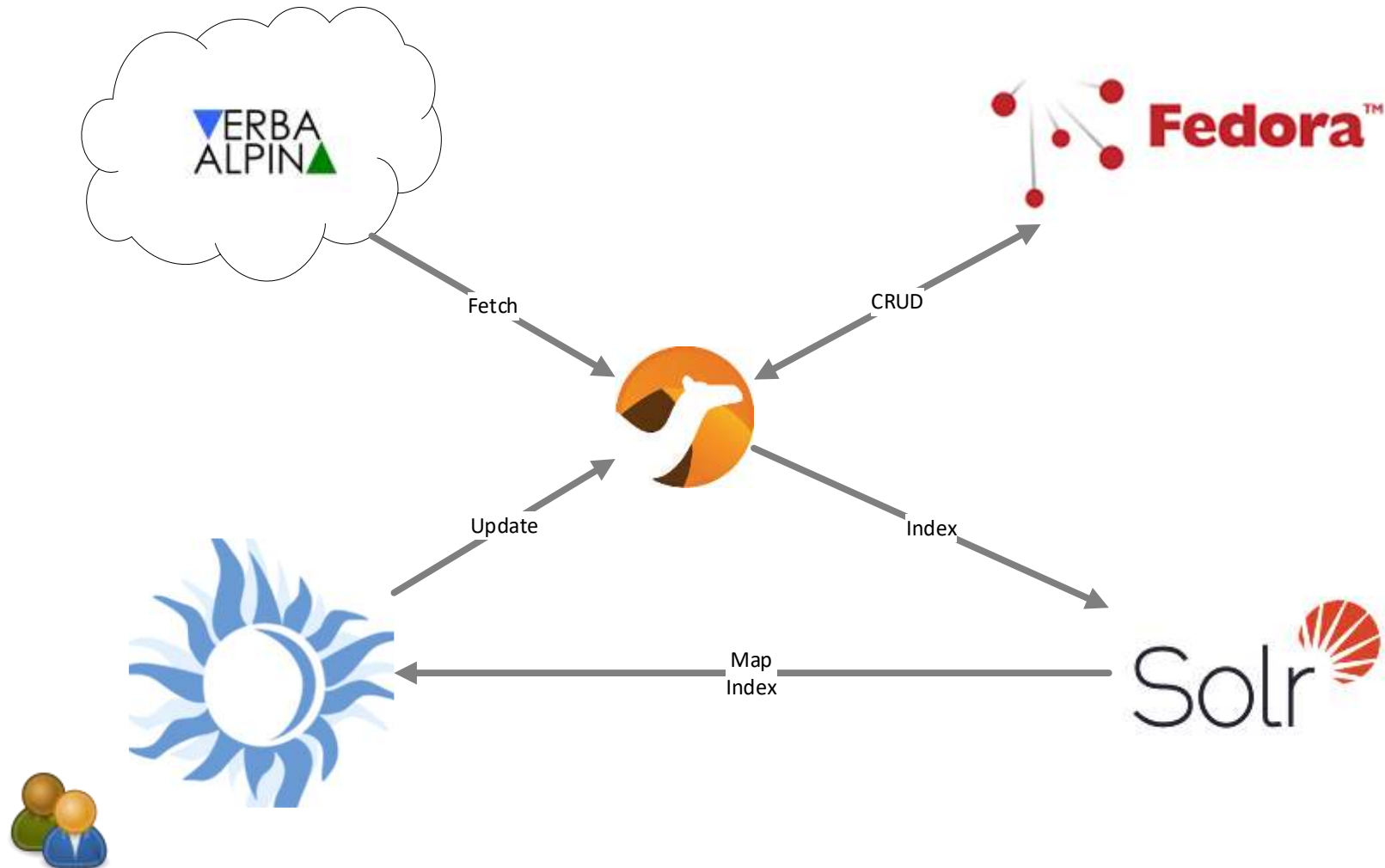
- Open source collaboration discovery platform framework
- Ruby on Rails

Apache Solr

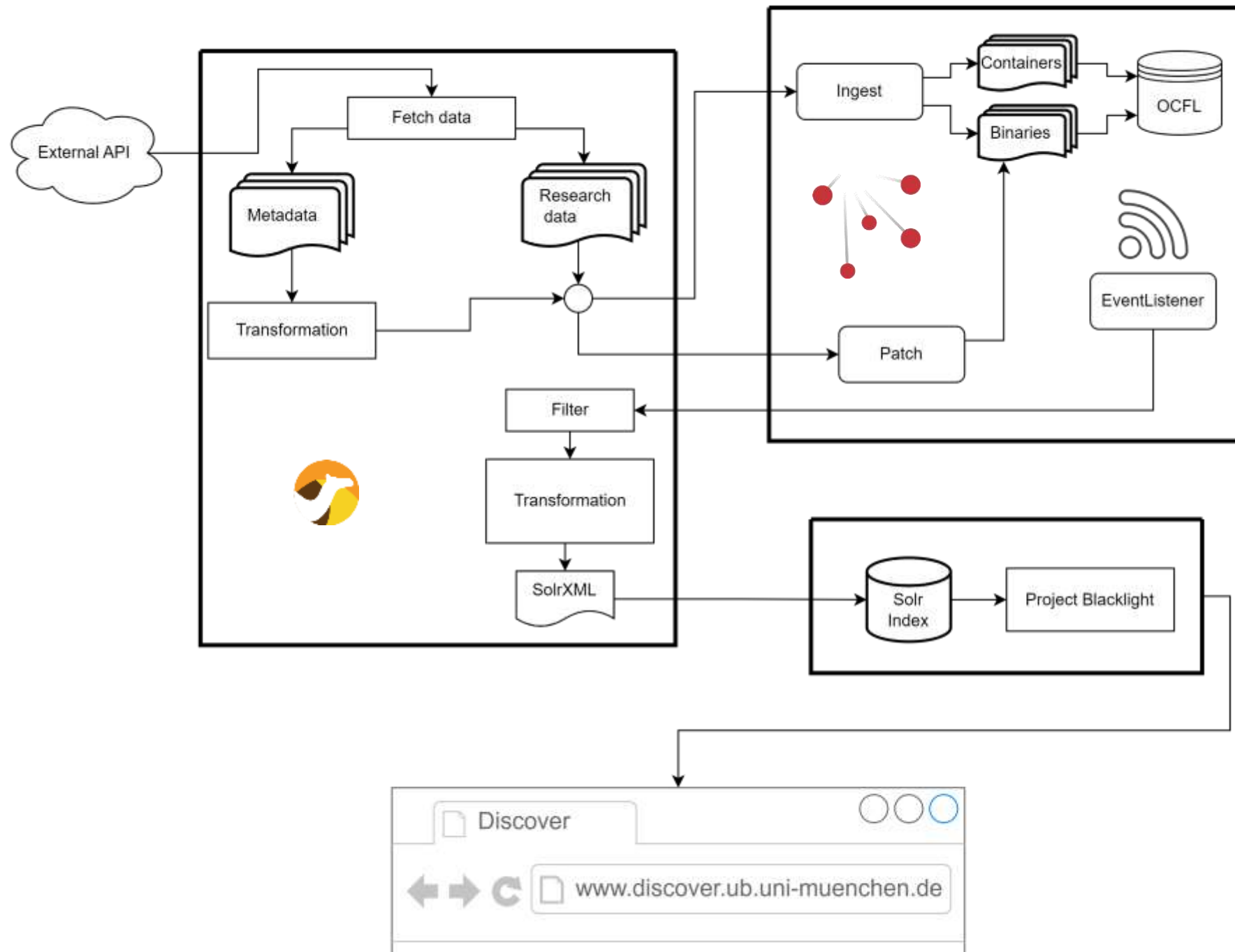
- Open source enterprise search platform (based on Apache Lucene)

2. Implementation and Workflows



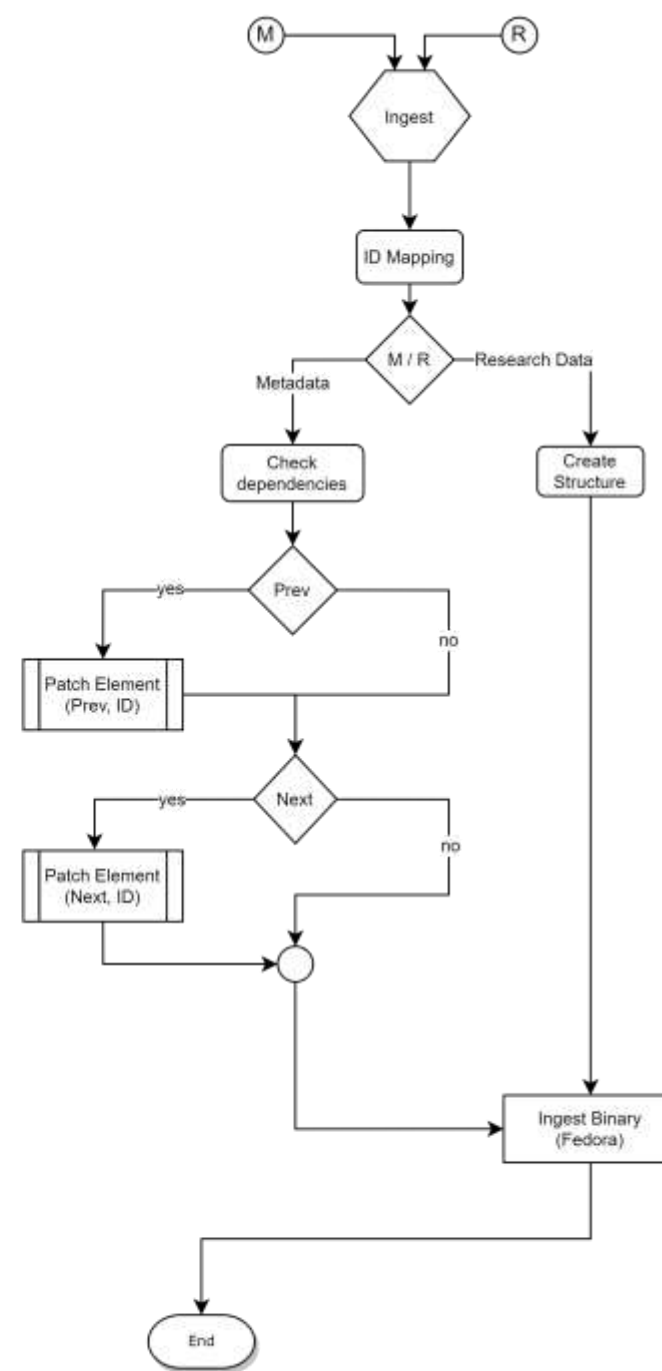
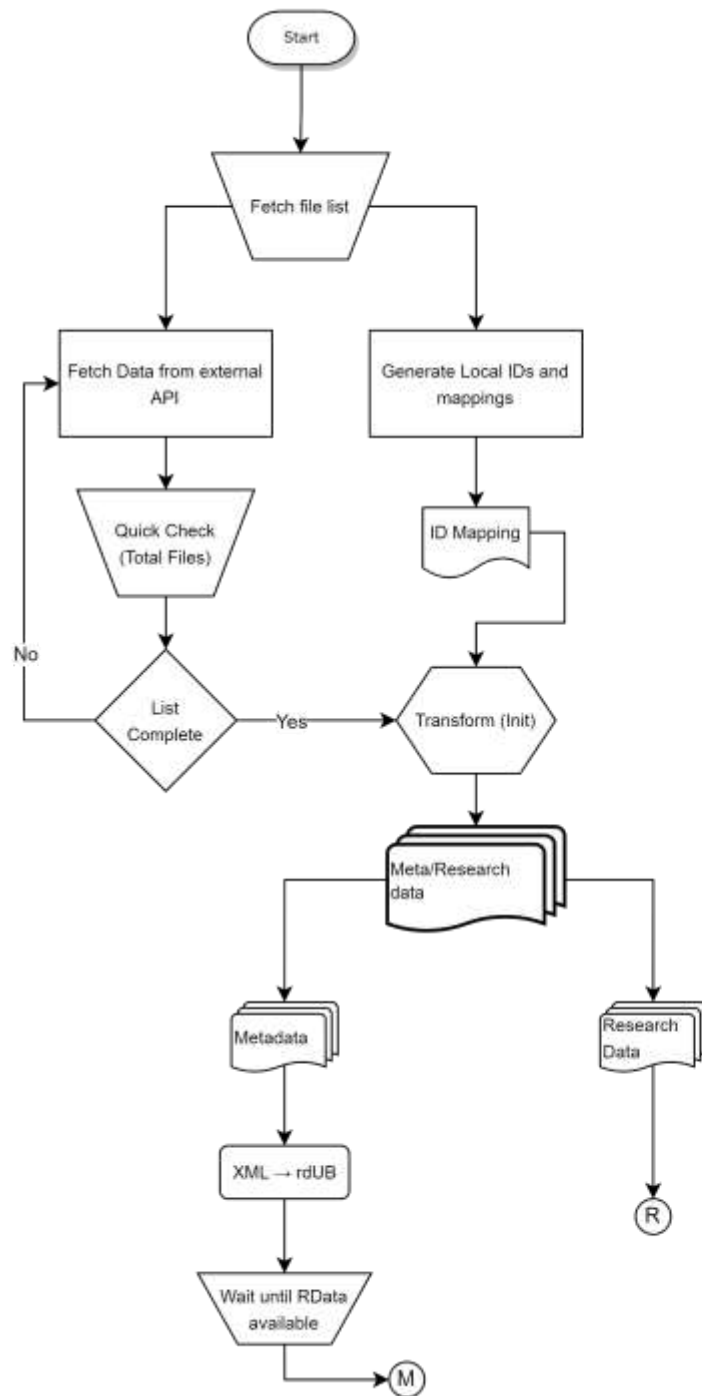


Discover



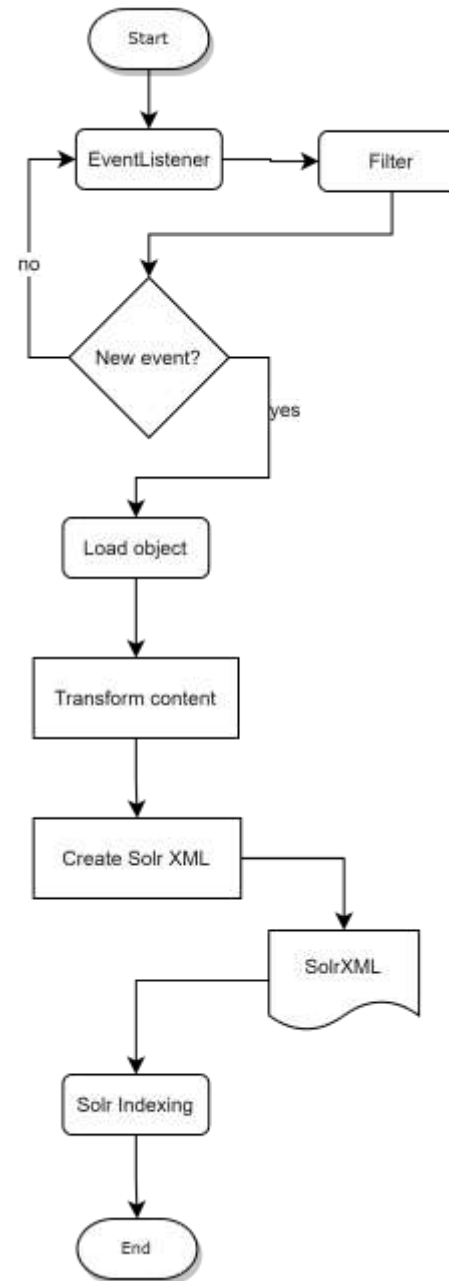
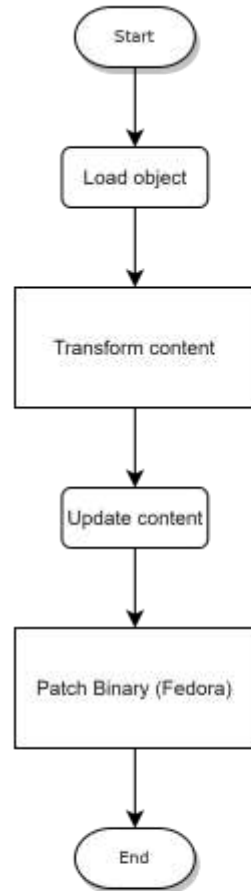
Discover

- ETL
- Ingest



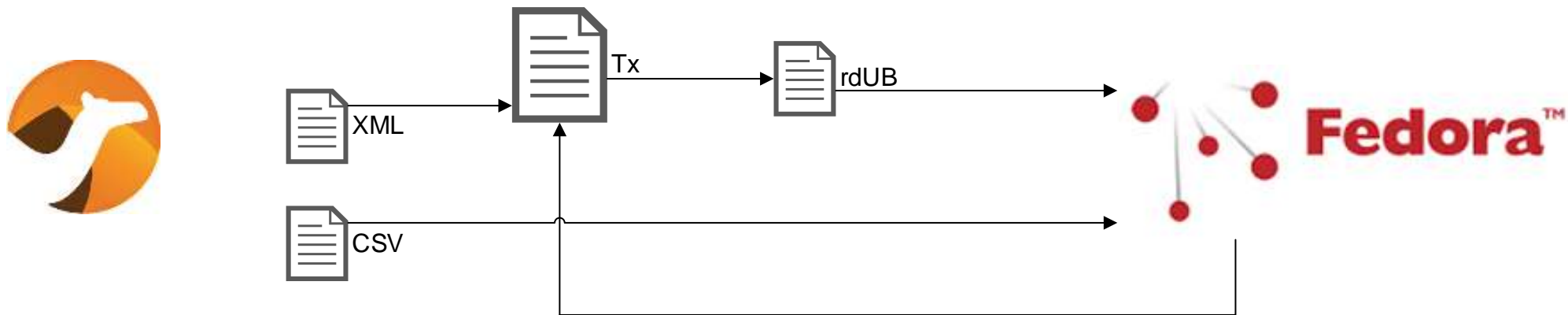
Discover

- Update
- Event Listener



Integration of the components with Apache Camel

- Components to transform XML to rdUB will transform the objects
- Generation of unique IDs for objects within the framework (ImUB)
- Determine relationships between objects (isPartOf / hasPart) based on the research data



Automatic creation of the objects in Fedora, with informations provided from the files in Verba Alpina.

- Minting of ID and container structure in Fedora will be created.
- Optimization of the ingest process into the repository.
- Objects will be checked before ingest and after the ingest the related objects will be patched.
 - IsPreviousVersionOf, IsNewVersionOf
 - HasPart, IsPartOf
- Any information missing? These properties could be ingested after the workflow is done.

Fedora → Solr

Integration using Apache Camel and the transformation, JMS (Java Message Service), and http components, along the REST API from Fedora to fetch and process the events.

- Fedora Events >>> Fedora Object >>> XSLT Transformation >>> Solr XML Document



Fedora → Solr



- Events in Fedora trigger Camel routes
 - JMS will be checked, object ID will be retrieved
 - Object will be loaded
 - Results will be transformed into Solr XML
 - Solr XML will be sent to Solr
 - After queue is done processing a Solr commit will be done

Project Blacklight



Blacklight loads the Solr index and is able to show the information to the end user through a GUI.

Includes the following features:

- OAI interface (formats: rdUB, DataCite, Dublin Core).
- Download the research data.
- ID information will be linked to the respective platforms (ORCID, GND, Wikidata, Glottolog)
- Other options (mail the record, request a DOI, ...)

Challenges



Metadata modelling: from *this looks easy and doable* to *oh my, this is impossible to model*

Performance and scalability on the ingest process: how can I ingest everything as fast as possible without collapsing our servers?

The data preservation and the independence of the persisted data: OCFL and Fedora.

Challenges and further steps



Extend the functionalities to support new projects into the framework (WIP)

How to handle larger datasets in a more automatic approach (WIP)

Extend our scope use of Fedora to be the backbone of other use cases and services



DIGITAL SERVICES



Thank you!

researchdata@ub.uni-muenchen.de
jaime.penagos@ub.uni-muenchen.de

