Digital repository(ies) at Charles University

Jakub Řihák

Repositories, Digitization and Depository Department

CHARLES UNIVERSITY
Central Library

jakub.rihak@ruk.cuni.cz

Twitter
LinkedIn

This presentation is licensed under the Creative Commons: CC-BY-SA-4.0,
via http://www.nusl.cz/ntk/nusl-367302

Outline

- Starting point
- Basic principles
- Workflow...
- ... and its automation
- Current state of the Repository
- Short-term plans
- Long-term plans

Starting point

- various repositories for various kinds of digitized and digital-born content
  - electronic theses – Theses repository on top of SIS
    - lack of interoperability
  - other documents (digitized or digital-born) – DigiTool
    - expensive, licensing fees based on number of stored digital objects
    - outdated

Starting point

- demand for change
  - Pay own people, not for SW (or not as much)
  - joining the community
  - learning from existing solutions that are actually used
- 2014 / 2015
  - looking for new repository system
  - „On what content should we focus from the beginning?“

Starting point

- Which system to use? DSpace!
- What to base the repository on? Electronic theses!
- 1st half of 2016
  - installation / training / experiments
- 2nd half of 2016
  - let‘s begin with actual work (it‘s about time)

Basic principles

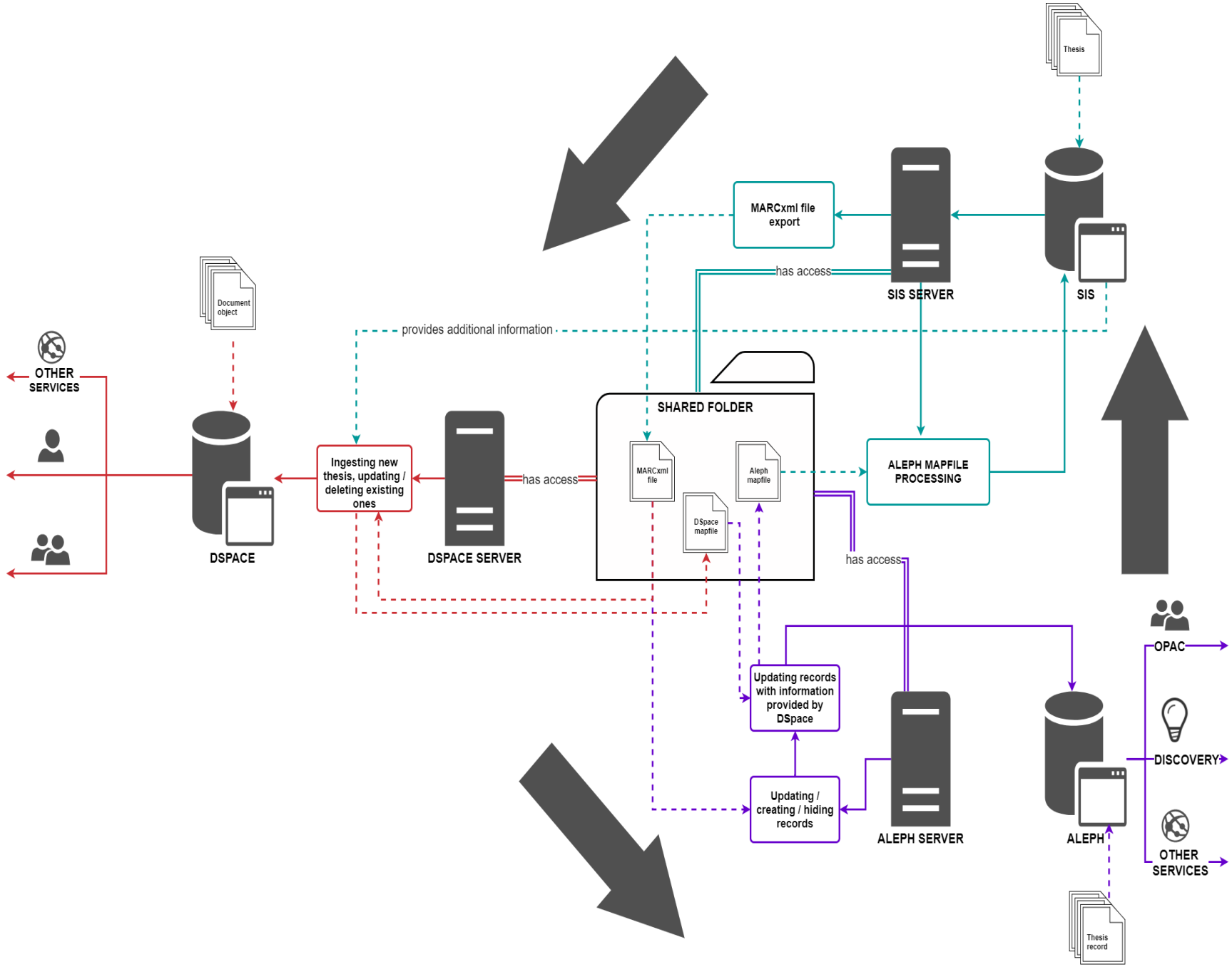
- theses should be ingested to DSpace directly from Study Information System (SIS)
- no unnecessary user interaction
- ingested theses have to have a permanent identifier and URL
- Access to ingested theses from:
  - OPAC
  - Discovery system

Basic principles

- DSpace structure (communities, collections, etc.) should copy university structure
  - → Faculties (community level)
    - → document types (collection level)
      - → items
- Emphasis on automation (large quantities of theses defended each year)

Workflow

- usage of existing one
- No direct transfer from SIS to DSpace
- easier and faster to implement than using database (reinventing the wheel?)
- Could be used to connect new repository to Aleph using existing methods
  - Theses accessible from OPAC and discovery system


Thesis
MARCxml file export
SIS SERVER
SIS
has access
Document object
provides additional information
OTHER SERVICES
SHARED FOLDER
MARCxml file
Aleph mapfile
DSpace mapfile
ALEPH MAPFILE PROCESSING
Ingesting new thesis, updating / deleting existing ones
has access
DSPACE
DSPACE SERVER
has access
Updating records with information provided by DSpace
Updating / creating / hiding records
ALEPH SERVER
ALEPH
OPAC
DISCOVERY
OTHER SERVICES
Thesis record

Workflow automation

- Theses should be processed at least once a day
- Ingestion via command line tools or DSpace API
- Automated ingestion process should use already existing resources if possible
- But how exactly? What do we need?
  - metadata, files, ingestion method

Metadata

- DSpace uses Dublin Core format
- Two default metadata schemas
- Existing schemas can be extended, new schemas can be added
  - Our case → storing metadata used for additional sidebar facets and search filters

Metadata registry

The metadata registry maintains a list of all metadata fields available in the repository. These fields may be divided amongst multiple schemas. However, DSpace requires the qualified Dublin Core schema. You may extend the Dublin Core schema with additional fields or add new schemas to the registry.

	ID	Namespace	Name
	1	http://dublincore.org/documents/dcmi-terms/	dc
☐	2	http://purl.org/dc/terms/	dcterms
☐	3	http://dspace.org/eperson	eperson
☐	4	http://cuni.cz/schema/uk-results-schema/	uk
☐	5	http://www.ndltd.org/standards/metadata/etdms/1.0/	thesis

Delete schema

Add a new schema

Namespace: *

Namespace should be an established URI location for the new schema.

Hledejte v Digitálním repozitáři UK

Hledat

Sbírky fakult Univerzity Karlovy

Vyberte fakultu k procházení jejích sbírek.

1. lékařská fakulta [2641]
First Faculty of Medicine

2. lékařská fakulta [1038]
Second Faculty of Medicine

3. lékařská fakulta [2114]
Third Faculty of Medicine

Evangelická teologická fakulta [1133]
Protestant Theological Faculty

ODKAZY
Užitečné odkazy

PROCHÁZET
Celý repozitář

MŮJ ÚČET
Přihlásit se

PROHLÍŽENÍ
Fakulta
1. lékařská fakulta (2634)
2. lékařská fakulta (1037)
3. lékařská fakulta (2114)
Evangelická teologická fakulta (1133)
Fakulta humanitních studií (4332)

<subfield code="a">Univerzita Karlova.</subfield>
<subfield code="b">Katedra fyzikální a makromol. chemie</subfield>
    </datafield>

    <datafield tag="850" ind1=" " ind2=" ">

<subfield code="a">PRF</subfield>
    </datafield>

    <datafield tag="IDS" ind1=" " ind2=" ">

<subfield code="a">149396</subfield>
    </datafield>

<controlfield tag="repId">193071</controlfield>
<controlfield tag="didId">193071</controlfield>
<controlfield tag="func">insert</controlfield>
<controlfield tag="ds_dateAccepted">31-08-2017</controlfield>
<controlfield tag="ds_workType">Rigorózní práce</controlfield>
<controlfield tag="ds_academicTitle">RNDr.</controlfield>
<controlfield tag="ds_facultyName_cs">Přírodovědecká fakulta</controlfield>
<controlfield tag="ds_facultyName_en">Faculty of Science</controlfield>
<controlfield tag="ds_facultyAbbr">PřF</controlfield>
<controlfield tag="ds_publication_place">Praha</controlfield>
<controlfield tag="ds_finalGrade_cs">Prospěl</controlfield>
<controlfield tag="ds_finalGrade_en">Pass</controlfield>
<controlfield tag="ds_studyLevel">rigorózní řízení</controlfield>
<controlfield tag="ds_studyField_cs">Modelování chemických vlastností nano- a biostruktur</controlfield>
<controlfield tag="ds_studyField_en">Modeling of Chemical Properties of Nano- and Biostructures</controlfield>
<controlfield tag="ds_studyProgram_cs">Chemie</controlfield>
<controlfield tag="ds_studyProgram_en">Chemistry</controlfield>
<controlfield tag="ds_departmentName_cs">Katedra fyzikální a makromol. chemie</controlfield>
<controlfield tag="ds_departmentName_en">Department of Physical and Macromolecular Chemistry</controlfield>
<controlfield tag="ds_keywords_cs">molekulární dynamika, simulace spekter, kvantová chemie, chiralita, optická aktivita</controlfield>
<controlfield tag="ds_keywords_en">molecular dynamics, spectra simulations, quantum chemistry, chirality, optical activity</controlfield>
<controlfield tag="ds_work_availability">V</controlfield>
</record>

Ingestion method

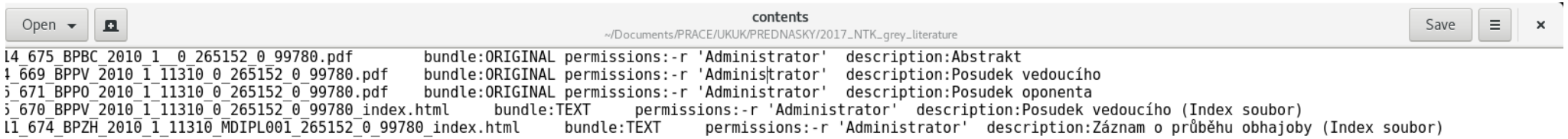
- Range of methods available
- Simple Archive Format package + command line import tool
- Simple way to check for errors in package structure and content - helpful during automation tool development

```
archive_directory/
    item_000/
        dublin_core.xml         -- qualified Dublin Core metadata for metadata fields belonging to the dc schema
        metadata_[prefix].xml   -- metadata in another schema, the prefix is the name of the schema as registered with the metadata registry
        contents                -- text file containing one line per filename
        collections             -- text file that contains the handles of the collections the item will belong two. Optional. Each handle in
                                -- Collection in first line will be the owning collection
        file_1.doc              -- files to be added as bitstreams to the item
        file_2.pdf
    item_001/
        dublin_core.xml
        contents
        file_1.png
        ...
```

Ingestion method

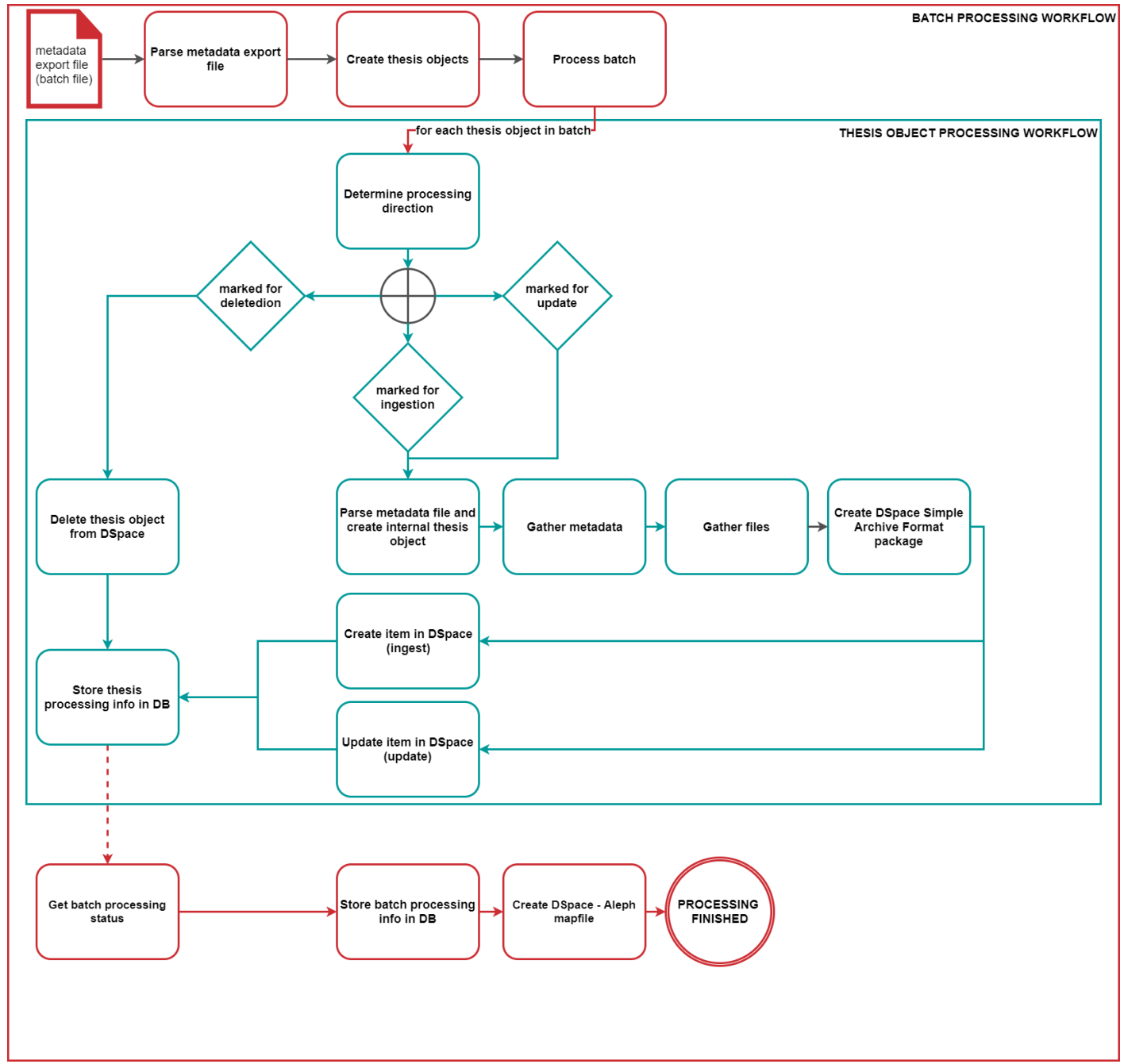
```
<dublin_core>
    <dcvalue element="title" qualifier="none">A Tale of Two Cities</dcvalue>
    <dcvalue element="date" qualifier="issued">1990</dcvalue>
    <dcvalue element="title" qualifier="alternative" language="fr">J'aime les Printemps</dcvalue>
</dublin_core>
```

(Note the optional language tag attribute which notifies the system that the optional title is in French.)



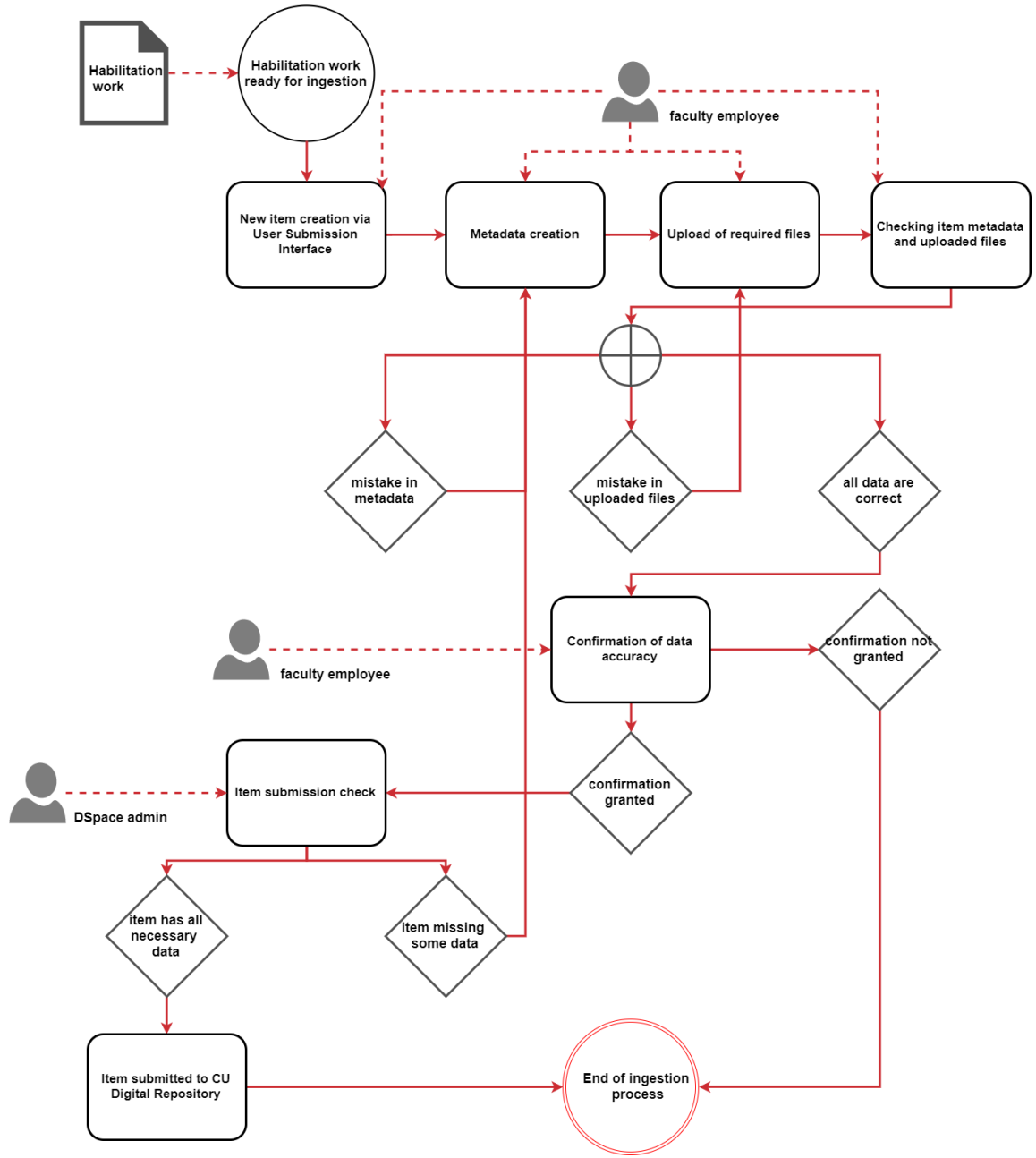
Automation Tool

- Python3 + PostgreSQL database
- Exported file from SIS → 'batch'
- Checks for new batch every 15 minutes
  - But there's only one export per day by default
- Processing workflow


BATCH PROCESSING WORKFLOW
metadata export file (batch file)
Parse metadata export file
Create thesis objects
Process batch
for each thesis object in batch
THESIS OBJECT PROCESSING WORKFLOW
Determine processing direction
marked for deletedion
marked for update
marked for ingestion
Delete thesis object from DSpace
Parse metadata file and create internal thesis object
Gather metadata
Gather files
Create DSpace Simple Archive Format package
Create item in DSpace (ingest)
Store thesis processing info in DB
Update item in DSpace (update)
Get batch processing status
Store batch processing info in DB
Create DSpace - Aleph mapfile
PROCESSING FINISHED

Current state of the Repository

- Available at https://dspace.cuni.cz
- Over 91k items
- Connected to National Repository of Grey Literature
  - OAI-PMH protocol
- Listed in OpenDOAR
- Habilitation works submission workflow
  - Cannot be done automatically (for now)
  - Using DSpace User Submission Interface
- New feature: automatically generated citations


Habilitation work
Habilitation work ready for ingestion
faculty employee
New item creation via User Submission Interface
Metadata creation
Upload of required files
Checking item metadata and uploaded files
mistake in metadata
mistake in uploaded files
all data are correct
Confirmation of data accuracy
confirmation not granted
faculty employee
confirmation granted
Item submission check
DSpace admin
item has all necessary data
item missing some data
Item submitted to CU Digital Repository
End of ingestion process

Short-term plans

- Shibboleth user authentication
- Horizon2020 – Open Access scientific publications
- Electronic books for students with special needs
- Content transfer from DigiTool repository
- Creation of a Central Digital Library for historical monographs, periodicals and maps

Long-term plans

- Central Access Point for digitized and digital-born content of the Charles University
  - Current state analysis (underway)
  - → strategic plan for development of CAP and possible long-term preservation of the digitized and digital-born content

Conclusion

- 6 months from start to finish
  - emphasis on automation
- Access to growing number of documents and document types
- Connection to national a international services
- Digital repository for digital-born content
  - Theses were just the start
- There will be a dedicated digital library for digitized historical documents
- Integration under singular Central Access Point

Thank you for your attention