10th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2017. ISSN 2336-5021. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings/**.

DIGITAL REPOSITORY(-IES) AT CHARLES UNIVERSITY "WHERE ARE WE NOW AND WHERE ARE WE HEADING?"

Jakub Řihák

jakub.rihak@ruk.cuni.cz

Central Library, Charles University

Abstract

This paper describes recent activities of the Central Library of Charles University (based in Prague, Czech Republic) in regards to providing access to digitized and digital-born content, in particular theses and habilitation theses as well as additional varieties of electronic content. The paper also describes the process behind the creation of the digital repository of Charles University, current tasks and plans for the future development of this service. We attempt to answer two "simple" questions: "Where are we now?" and "Where do we want to be in the future?"

Keywords

DSpace; Digital Repositories; Automation; Library

Introduction

Since 2010, Charles University has had an internal regulation[1] that specifically targets the submission of theses in electronic form and makes it mandatory to submit theses to Study Information System (SIS) in the form of an electronic document. It also specifies that this electronic thesis has to be published online in the university repository. This task was previously fulfilled by ingesting theses into the Qualification works Repository system[2] created as a part of SIS.

In previous years, Charles University had provided access to most of its digitized and digital-born documents (small portion of digitized theses among them) in the DigiTool system developed by ExLibris. Even though this system is still running and is used to store and provide access to various types of digitized and digital-born materials, there was a demand for a change. The main reasons for this change were the following:

- high annual support fees
- licensing fees based on the number of digital objects stored in the repository
- demand for an open-source solution with big community support, both in the Czech Republic and abroad

The first analyses on the possibility to use a different repository system were carried out between the years 2014 and 2015. A special committee consisting of the university management, the faculty library management and a specialist from the field of librarianship and information science was established and entrusted with the task of comparing various digital depositories and digital library systems with the prospect of choosing the best possible solution to replace the expensive proprietary system with a more modern one with open source licensing.

In the meantime, it was decided that a new electronic thesis repository is needed, because the Qualification works Repository system didn't satisfy all the requirements for interoperability between other library systems (with the exception of the library catalogue) and services, e. g. the discovery system, the National repository of Grey Literature and other international indexes, databases, information services and service providers.

It was decided that a new digital repository will be created using DSpace repository system, which is used by many Czech universities[3], has an established international community[4] and is developed as open-source software[5]. As for the annual support fees and licensing, there is no additional cost for using this system, as its support and development is community driven, with the possibility of voluntary memberships[6].

After more than six months of work, the Charles University Digital Repository[7] (CU Digital Repository) was created. It was decided that it would be used primarily as a repository for

[1] Available from: **http://www.cuni.cz/UK-3470.html**
[2] Available from: **http://is.cuni.cz/webapps/zzp/**
[3] **http://www.dspace.cz/dspace-v-cr**
[4] **http://registry.duraspace.org/registry/dspace**
[5] **https://github.com/DSpace/DSpace**
[6] **http://duraspace.org/all_members/dspace**
[7] Available from: **https://dspace.cuni.cz**

newly defended theses due to the demand from university management and because theses offer a steady flow of new content to the repository. After nearly a year of successful operation, the Central Library now works on transferring other collections of digitized and digital-born documents from the DigiTool system and prospectively ending the use of the DigiTool system for storing and publishing digital materials.

In this article, I will try to describe the whole process by which the CU Digital Repository was created and the way it went from being an idea to a system that now stores and provides access to all publicly available theses of Charles University.

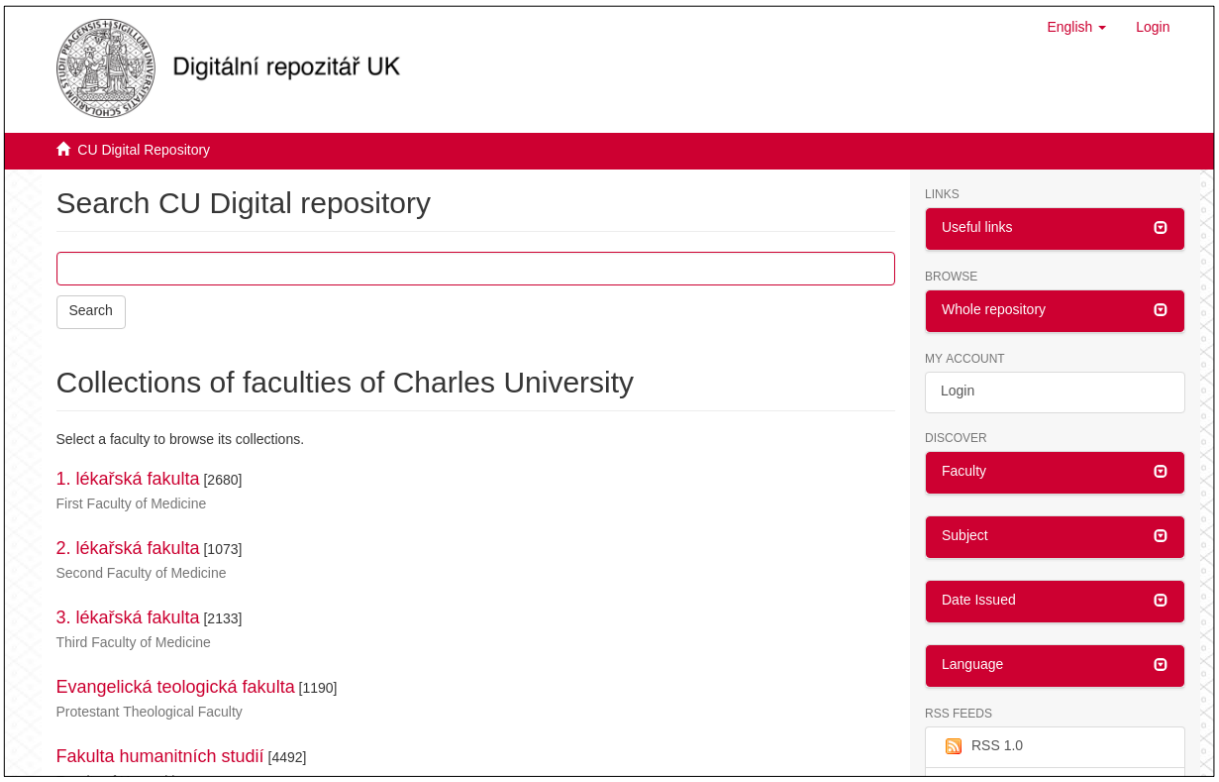

Figure 1: CU Digital Repository Homepage (https://dspace.cuni.cz)

Creation of the CU Digital Repository

Works on the new digital repository began in early 2016, and the whole repository should be ready to ingest, store and publish newly defended theses from 1 January 2017. The Central Library of the Charles University wanted to implement the following principles in order to minimize the time between submission of the finalized thesis to the Study Information System and its publication in the digital repository and reduce the possibility of any human error in the ingestion workflow:

- The thesis should be ingested into DSpace directly from the Study Information System (SIS)
- There should be no unnecessary user interaction
- The ingested thesis has to have a permanent identifier and URL that won't change when the new version is ingested
- The ingested theses have to be accessible from the electronic catalogue (OPAC)
- The ingested theses have to be accessible from the discovery system

SIS does not provide an Application Programming Interface (API) of any kind, so the idea was to connect directly to the underlying database and gather all the necessary data (bibliographic metadata, thesis files and embargo information) from there.

Together with discovery system, OPAC is one of the main resources for finding an electronic thesis in Charles University, so there has to be a process that would allow adding links to digital objects in the repository to the correct record in library information system. Links in OPAC have to be permanent so that they don't change in cases where a new version of a particular thesis is ingested or transferred to another location. This could be done with the support of handle identifiers that have built-in support in the DSpace system.

A huge emphasis has also been placed on automation. With an average of 8,274 graduates in the academic year 2015-2016 (HÁJEK & BOJAR, 2017), there is the prospect of large number of theses that need to be published in a digital repository each academic year. It was also decided that the CU Digital repository will have the following structure:

→ *faculties (community level)*

→ *document (work) types (collection level)*

→ *items*

This structure is common in several Czech DSpace repositories[8], and it allows the content to be structured in a logical way that copies the organizational structure of the university and allows the user to access all existing document types of each faculty which can be also used for promotional purposes by the university faculty, as a link to the faculty's own collection and can be provided to students on the faculty's website or in other promotional materials.

[8]For example: CTU DSpace repository (**https://dspace.cvut.cz/**), Pardubice University DSpace repository (**http://dspace.upce.cz/**) or VŠB – Technical University of Ostrava DSpace repository (**https://dspace.vsb.cz/**)

Defining workflow

After discussions with our library system administrators, it was finally decided that an existing SIS - Aleph workflow will be used to get a set of theses available for ingestion. This existing workflow is used to insert, update or delete (or rather hide) the record of the thesis bibliographic when a new thesis is available for publication. The DSpace thesis processing workflow could be inserted between those two steps with minimal changes in the existing SIS and Aleph processes. Dspace processes SIS exports, providing additional information about ingested theses to the Aleph library system. Aleph then processes the same metadata exports to insert, update or hide thesis records and the bibliographic record of each processed thesis[9] is enriched with the URL to the digital object in DSpace. The URLs and system numbers of processed theses are then passed back to SIS and stored in its database for future use. With the workflow set up in this manner, we can also ensure that all necessary data are identical in each of the connected systems as shown in Figure 2.[10]

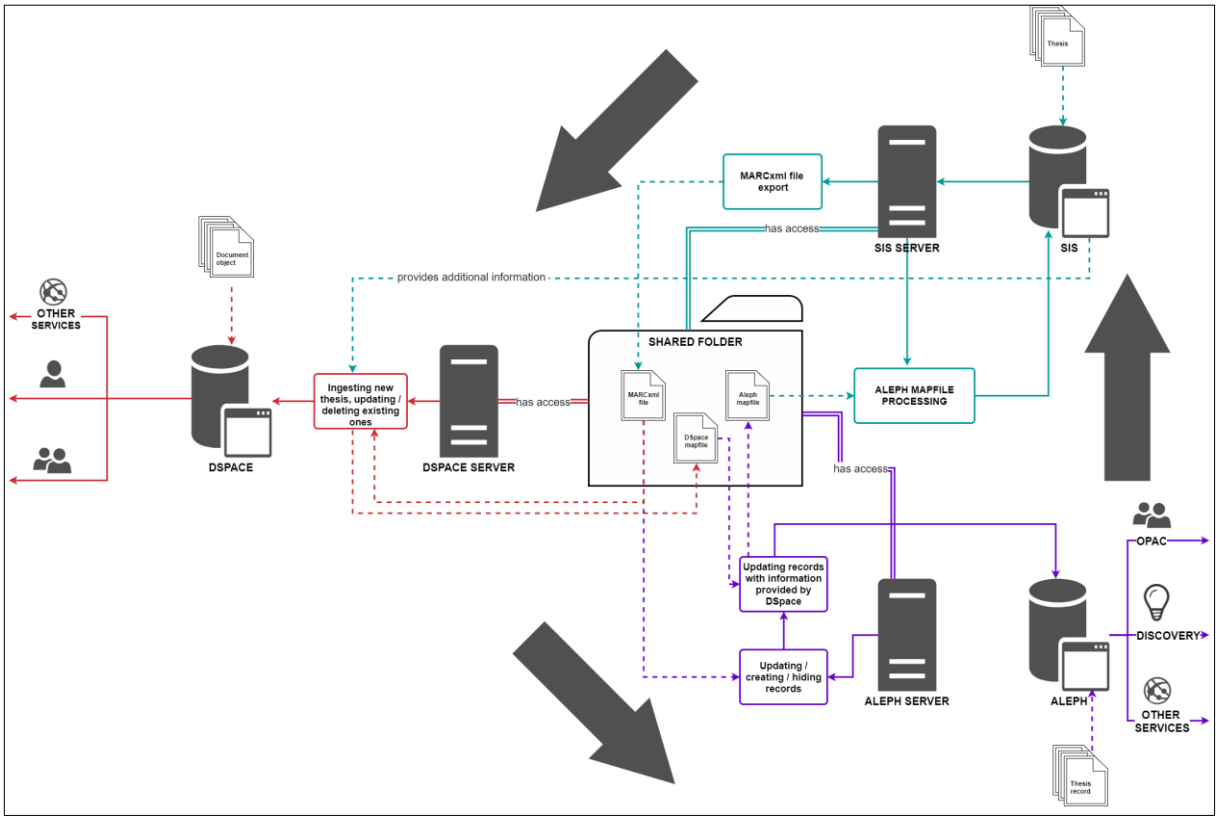

Figure 2: Thesis processing workflow diagram

[9] Of course, this does not apply to theses marked for deletion.

[10] Except for Aleph system number (unique bibliographic record identifier). This identifier can now only be added to the thesis record in DSpace after it is updated, since newly submitted theses are first processed by DSpace, not Aleph, which creates system numbers during the creation of the bibliographic record. This issue will be addressed in the future.

Workflow automation – basic considerations

As has already been mentioned, preferably the whole thesis ingestion workflow should be automated to prevent possible human errors and to save time. There were the three following premises regarding thesis processing:

- thesis processing should take place at least once a day, but the program should check for new exports regularly several times a day
- preferably, ingestion should be done via command line tools or DSpace API
- automated ingestion should use resources that already exist if possible

For the purpose of workflow automation, the Python3 programming language is used. However, before the programming work started, it was necessary to consider which metadata we would like use to describe an electronic thesis in DSpace, which DSpace ingestion method we should use and what changes in DSpace will be necessary to ensure sufficient accessibility of the final digital object in DSpace.

Metadata selection

The DSpace 5 system uses Dublin Core metadata format by default. There are two existing metadata schemas available [11] for item description in DSpace. Those schemas can be extended, or a new metadata schema can be created. This was the case with the CU Digital Repository, as additional metadata was required for creating custom search fields and sidebar facets that would help in making ingested theses more accessible and the whole DSpace user interface more user-friendly.

[11] Available at **https://goo.gl/BsX8hH**

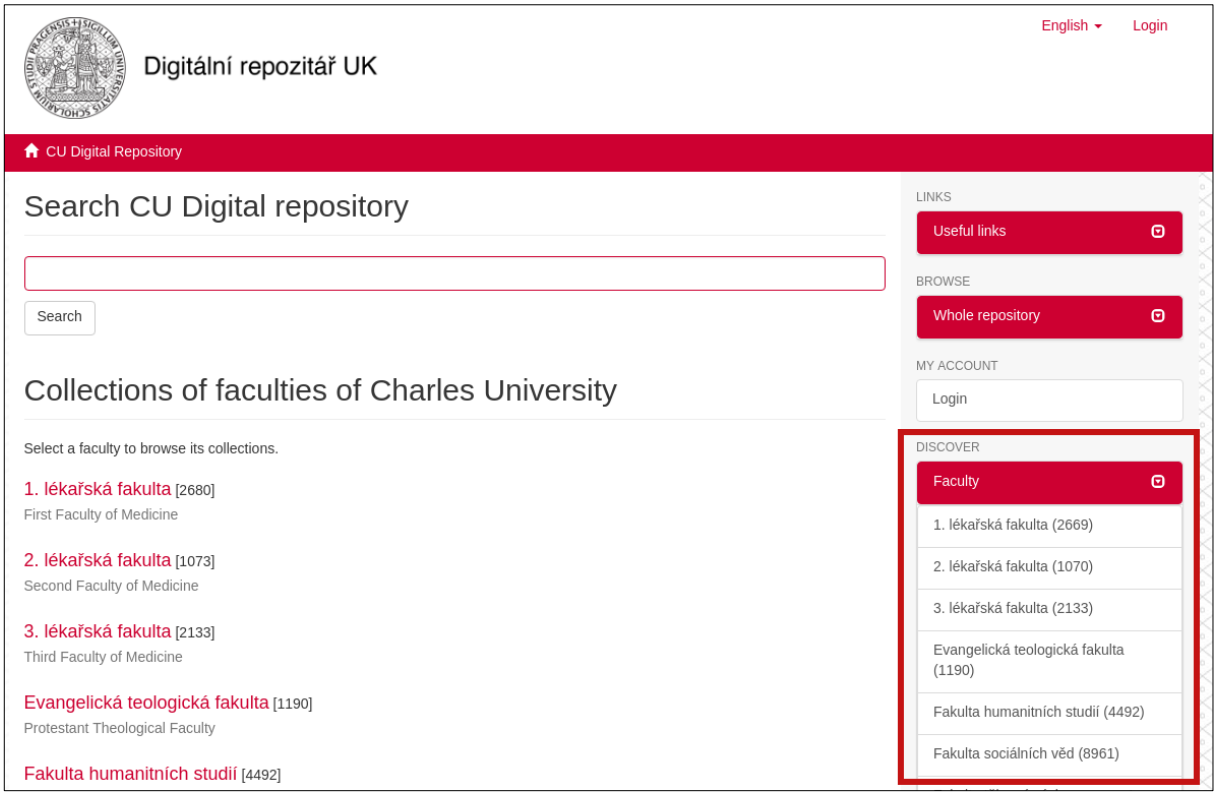

Figure 3: Example of custom metadata used in sidebar facet

For custom descriptive metadata that are not part of the standard bibliographic record, control fields are used. These are not used as a data source for the document's bibliographic description during Aleph processing and are generated just for the purpose of the DSpace ingestion workflow. An example of this part of the metadata export is shown in Figure 4.

```
<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>
```

Figure 4: Custom thesis metadata in MARCxml export

Ingestion method

DSpace offers multiple methods of content and metadata ingestion.[12] After discussions and meetings with colleagues from other universities that are using DSpace as their repository system (mainly Tomas Bata University in Zlín and Pardubice University), it was decided that Simple Archive Format packages will be used. A Simple Archive Format package is "an archive which is a directory containing one subdirectory per item. Each item directory contains a file for the item's descriptive metadata, and the files that make up an item." (DONOHUE, 2017) The basic structure of the DSpace Simple Archive Format is shown in Figure 5 (DONOHUE, 2017).

```
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
        ...
```

Figure 5: Simple archive format structure example

The Simple Archive Format package can be used for batch import of new items to DSpace, similarly to CSV import, but offers easy navigation in the content of each item and its descriptive metadata. Its simplistic nature is helpful in the development of an automation tool, because it allows possible errors in the package structure or content to be checked and corrected in very simple way, as can be seen in the following Figure 6, depicting a sample metadata file in the Dublin Core metadata schema.

```
<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.)

Figure 6: Simple Archive Format metadata example

Automation tool

The workflow automation tool was developed in 4 months. It uses the PostgreSQL database, where information on processing individual export files and theses is stored. The database is used for the purpose of determining whether or not the given export file or thesis entered the workflow in the past, to determine its processing 'direction' based on this information, and to store information on the processing state. Metadata exports are processed once a day, and each metadata export file represents a 'batch'. However, the automation tool checks for new

[12] **https://goo.gl/pFv9vF**

metadata export files every 15 minutes and is able to process failed 'batches' or just individual theses for which the processing has failed.

The automation tool is able to gather the necessary bibliographic and other descriptive metadata and thesis files and to create a Simple Archive Format package and import it to DSpace using a standard command line importer[13].

The test of actual live data revealed an issue with an improper character escaping during metadata export file creation, resulting in the metadata export file not being processed. There were also some minor issues with displaying the additional metadata values in the DSpace user interface. However they were solved by customizing the affected parts of the DSpace user interface using a combination of XSLT, HTML and CSS. With these issues solved, the ingestion of theses to the production repository began in December 2016.

Current state

The CU Digital Repository grows nearly every day. New these are ingested regularly and a small amount of habilitation works is already stored and published. There are currently over 90 000 items stored and available to the public. This also includes theses previously published in the Qualification works Repository that were moved to the CU Digital Repository during this year.

[13] See **https://goo.gl/j1vEph** for details.

In March 2017, the CU Digital Repository also began to receive habilitation works from individual faculties. At the beginning of February 2017, the Central Library was tasked with providing access to habilitation works according to Act no. 11/1998 Coll., on universities[14], and the CU Digital Repository had to be ready for their ingestion in one month.

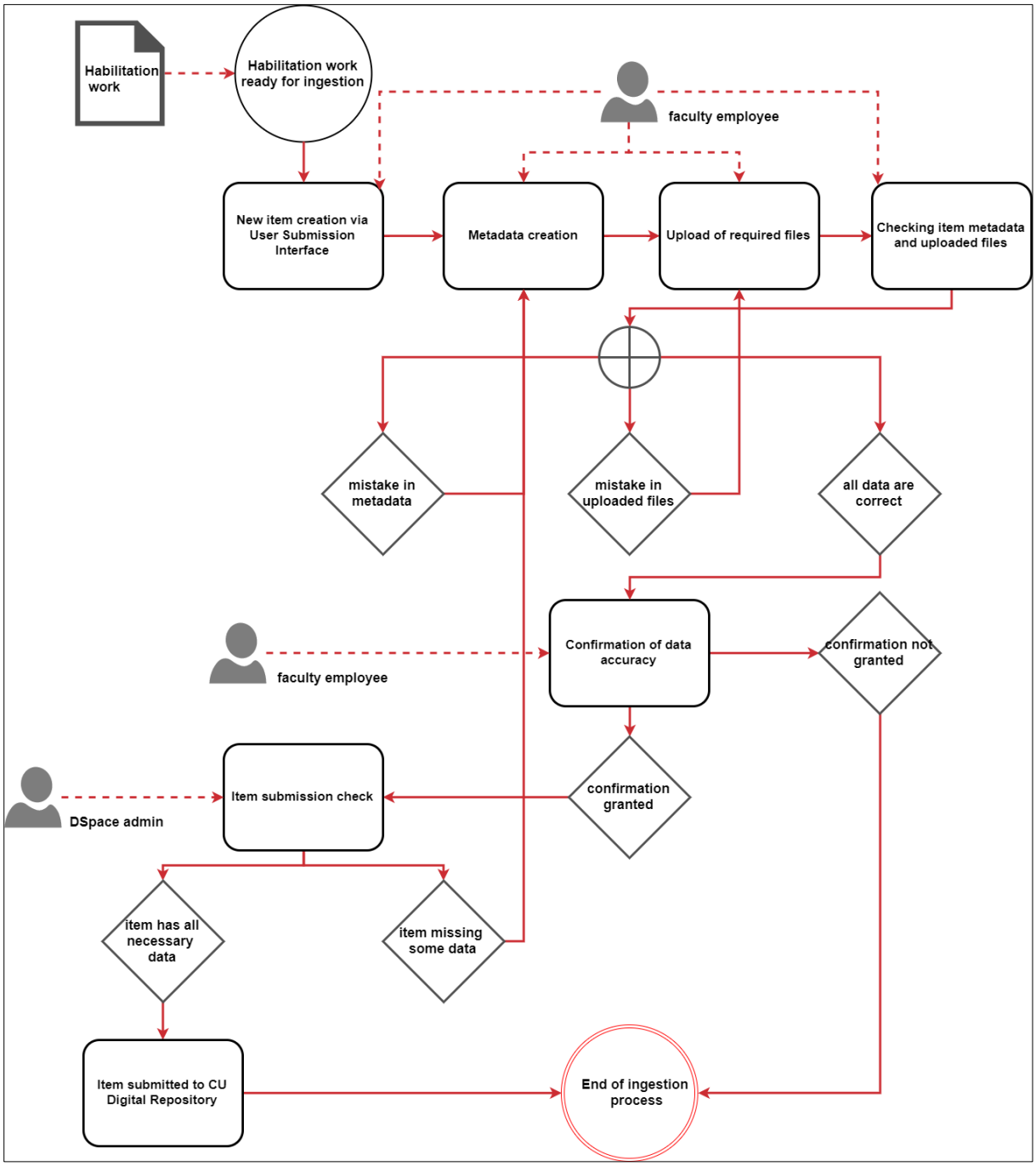

Figure 7: Habilitation works submission workflow

[14] Available at http://www.msmt.cz/vyzkum-a-vyvoj-2/zakon-c-111-1998-sb-o-vysokych-skolach.

As habilitation works are not stored in any electronic system, an ingestion workflow similar to the one used for theses could not be set up. Instead, it was decided that the internal DSpace tool - User Submission Interface[15] - will be used to gather all necessary metadata and files and publish habilitation works through the standard DSpace submission workflow.

New collections were created within the existing CU Digital Repository structure to hold habilitation works, and authorized faculty employees were given administrative rights to these collections, allowing them to submit new items and change items that have already been published. The CU Digital Repository administrators have the right to accept or reject submitted items. This provides repository administrators with a way to check submitted works and make it impossible to a submit habilitation work that does not follow the defined standards of bibliographic description or other content described in the Habilitation work submission methodology.[16] The habilitation work submission workflow is described in Figure 7. This workflow is not ideal for ingesting large amount of items, because it relies on manual work to a great extent, which could be very time-consuming when done for large quantities of documents. It is also prone to human error. However, it was designed with that in mind and offers a way to control the data quality of ingested items.

Connecting to the National Repository of Grey Literature and OpenDOAR

The CU Digital Repository was connected to the National Repository of Grey Literature (NRGL) through the OAI-PMH protocol in April 2017. Thanks to this, Charles University is the biggest data provider for NRGL, with nearly 90,000 available records. This allows the CU Digital Repository to be more discoverable and allows Charles University to fulfil its vision of "taking active part in the development of the branches and subjects it teaches; [to be] a modern university open to the world" (Charles University, 2015) and also Strategic plan of Cantral Library of Charles University to a greater extent.

The CU Digital Repository is also registered in OpenDOAR – Directory of Open Access Repositories[17] and is indexed by Google Scholar on a regular basis. Registration in OpenDOAR is also one of the prerequisites for becoming a data provider for the OpenAIRE repository.

Automatically generated citations

The most recent change in the CU Digital repository is the addition of the item citation to the item record view. The item citation is generated using a built-in OAI-PMH provider and Citace.com API. When the user displays an item record, a query is sent to the OAI-PMH provider, which returns the necessary data in a Dublin Core format and sends it to Citace.com. This data is coverted to the correct citation format according to the ČSN ISO 690 standard and then embedded in item record page. To implement this feature, it was necessary to create a customized OAI-PMH metadata schema that would hold all the necessary information, and it was done in cooperation with Citace.com employees.

[15] See **https://wiki.duraspace.org/display/DSDOC5x/Submission+User+Interface for details**.
[16] **https://knihovna.cuni.cz/rozcestnik/repozitare/metodika-vkladani-habilitacnich-praci-do-repozitare/**
[17] Repository record available at: **http://opendoar.org/id/3873/**

Short-term and long-term plans

Short-term plans include:

- Enabling user authentication using Shibboleth connected to Central Authentication Service (CAS) identity provider,
    - Allowing the CU Digital Repository to dynamically assign roles to its users based on their user attributes provided by CAS and thus grant access rights to special collections in the CU Digital Repository.
- Ingestion of Open Access scientific publications from the Horizon 2020 programme,
    - fulfilling the requirements for research projects financed by the Horizon 2020 programme, according to which "each beneficiary must ensure open access to all peer-reviewed scientific publications relating to its results" (European Commission, 2017) by depositing publications in repositories. This is currently not possible on the institutional level, because the Register of Research Publications (OBD) currently being used does not provide access to actual files, and by connecting the CU Digital Repository to OBD, this access to research publications can be granted.
- Providing access to electronic books for disadvantaged students of Charles University,
    - which is in compliance with the Strategic plan of the Central Library of Charles University for the years 2015 – 2018. The Central Library is now working in close cooperation with Information and Advisory Services Centre (IASC) to provide access to these study materials and e-books via the CU Digital Repository.
- Transferring collections from the DigiTool repository,
    - collections of historical value, mainly digitized monographs, periodicals and maps, should be moved to the Kramerius digital library, which is currently being tested.
    - other collections, mainly of digital-born documents, could be moved to the CU Digital Repository. In the case of the collection of digitized theses, this transfer has already begun and is currently 80 % finished.
- Creating a digital library for historical monographs, periodicals and maps using the Kramerius digital library system.
    - The Kramerius digital library18 is, in our opinion, more suitable for providing access to digitized historical materials then DSpace and with the addition of ProArc19 software. It also has some of the long-term preservation capabilities.

[18] More details available at **https://github.com/ceskaexpedice/kramerius**

[19] More details available at **https://github.com/proarc/proarc/wiki**

Long-term plans include:

- Carrying out an analysis on the current state of the digital repositories and digital libraries used at Charles University and on the current state of publishing and preservation of digitized and digital-born documents,
    - that will serve as a foundation for the creation of a strategic plan for the development of services for providing access and the long-term preservation of digitized and digital-born content at Charles University, and should allow the Central Library to determine what the right direction of further development could be.
- Creating a strategic plan for the development of services for providing access to the digitized and digital-born content of Charles University.
    - The idea behind this strategic plan is to create a singular access point to the digitized and digital-born content of the university that can be promoted to the public more easily and guide users to the content instead of confusing them. Another advantages might be: more focused allocation of financial, technical and 'human' resources and future investments and development of any kind.
- Creating a central installation of the Kramerius digital library.
    - The Central Library would also like to create a centralized Kramerius digital library installation in which the digitization outputs of individual faculties could be published and which would serve (together with the already-implemented DSpace repository system) as a basis for this 'singular access point'.

Conclusion

The creation of the CU Digital Repository started in June 2016 after several years of discussions. Its primary objective was to provide access to electronic theses defended from January 2017 to date. This objective was fulfilled in time thanks to the emphasis that was placed on automated processing and the focus on extending the already-existing workflow and its resources. After nearly a year of successful operation, the content of the CU Digital Repository has grown both in size and in the variety of the content provided. CU Digital Repository now also provides access to habilitation works and is prepared for the ingestion of research publications from the Registry of Research Publications (OBD) and electronic books for the disadvantaged students of Charles University. Even though errors and mistakes were made during the creation of the CU Digital Repository , we would describe its development as successful.

The CU Digital Repository is connected to the NRGL repository and OpenDOAR, which makes it possible to share the information stored in this repository with a broader audience. The repository will be continuously developed to provide better services for its users. The Central Library also aims to create a dedicated repository for digitized historical monographs, periodicals and maps. These two repositories should, in time, replace the DigiTool repository system currently being used to store the majority of digitized and digital-born materials and provide a basis for the creation of a singular access point to the digitized and digital-born materials of Charles University. In doing so, they will provide users with better access to these materials,enable the better promotion, the better allocation of financial, technical and human resources and make the long-term preservation of digitized and digital-born materials possible.

References

DONOHUE, Tim. Importing and Exporting Items via Simple Archive Format. In: *DuraSpace Wiki: DSpace 5.x Documentation* [online]. San Francisco (CA): Atlassian, 2017 [Accessed 3 October 2017]. Available from: **https://wiki.duraspace.org/x/0QK3Ag**

Charles University Strategic Plan 2016–2020 [online]. Prague: Charles University in Prague, 2015 [Accessed 3 October 2017]. Available from: **http://www.cuni.cz/UKEN-110-version1-charles_university_strategic_p.pdf**

H2020 Programme: Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020 [online]. Brussels: European Commission, 2017 [Accessed 3 October 2017]. Available from: **http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf**

HÁJEK, Václav and Štěpán BOJAR, ed. *Výroční zpráva o činnosti Univerzity Karlovy v Praze za rok 2016* [online]. Praha: Univerzita Karlova, 2017 [Accessed 3 October 2017]. ISBN 978-80-246-3726-6. Available from: **http://www.cuni.cz/UK-8533-version1-vzc_2016_web.pdf**