The Czech Digital Library and Tools for the Management of Complex Digitization Processes

Martin LHOTÁK

Library of the Academy of Sciences of the Czech Republic

lhotak@knav.cz

INFORUM 2012: 18th Conference on Professional Information Resources

Prague, May 22-24, 2012

Abstract

The main idea is to create solution which will aggregate content of digital libraries in the Czech Republic and provide access to this digital collections. Other tools will be also developed to support complex digitization processes in the frame of the CDL project. Included are processing and archiving of digital documents. The goal is to use these tools to increase the number of materials available in the Czech Digital Library. Using the same production and archiving tolls will enhance interoperability and data sharing between individual digitization projects.

Article

The goal of the project is to create the Czech Digital Library ("CDL") which will aggregate content of digital libraries in the Czech Republic. It will serve both as a uniform interface for end-users and as a primary data provider for international projects, especially for Europeana, the European digital library. It will also be an important source of digital data generally and is one of the main pillars needed to provide centralized digital services in the Czech Republic, as defined in the "Library Development Strategy of the Czech Republic for 2011 to 2015," approved by the Czech Ministry of Culture.

The open source Kramerius 4 system will be the initial software solution for the Czech Digital Library. Kramerius 4 is based on the Fedora repository and is widely used as a digital library system in the Czech Republic. It was developed jointly through the cooperation of the Library of the Academy of Sciences and the National Library of the Czech Republic with IT companies Qbizm and INCAD. Ensuring the interoperability with various types of digital libraries and institutional repositories is necessary. Besides data harvesting from different instances of the Kramerius 4 system, it is also necessary to arrange a connection with other systems (e.g., Dspace, Eprints, Digitool). Some special proprietary solutions must be worked out, such as for the digital library of the Institute for Czech Literature of the Academy of Sciences, which will be technically more difficult to hook up for cooperation.

The Czech Digital Library will also serve as OAI-PMH provider with the ESE profile support to share data with Europeana.

Some other OAI-PMH profiles might also be implemented to facilitate cooperation with other centralized international projects (e.g., World Digital Library).

The Digitization Registry, which was built formerly as a project of the Library of the Academy of Sciences and the National Library of the Czech Republic, will be used as an interconnecting system and relevant source of information. The Registry is publicly accessible online at http://www.registrdigitalizace.cz/. It holds a large amount of information about digitized documents in the Czech Republic. Included is the identification of original printed documents, owner and location of the digital library where the digital document is available, persistent identification (e.g., Czech National Bibliography Number) and other relevant entries. The main aim of the national registry of the digitized documents is to avoid unwanted duplication to enable the sharing of digitization results throughout the Czech Republic and not waste time or money by scanning the same documents. The Digitization Registry could also provide tools for digitization workflow management to simplify the process of monitoring the digitization. This solution could also serve end-users as the central access point to digitized documents. Very important also is the fact that it cooperates with library catalogue systems as well as with digital document repositories.

In the context of interoperability and cooperation with library information systems, the Registry is designed to communicate and cooperate automatically with other library information

systems as much as possible. It uploads bibliographic records of items chosen for digitization in batches exported from the Aleph catalogue in MARCXML. The Registry is able to harvest data from digital libraries via OAI-PMH to import data describing digitized documents. Finally, it sends information about completed digitization to library OPACs together with a link to digital documents. Information is subsequently sent from library OPACs to the Union catalogue of the Czech Republic.

The current solution for the Digitization Registry is heavily dependent on the National Library, which runs the central installation. The solution is based on commercial tools and it is not easily transferable to individual digitization centers. It is necessary to provide a persistent connection to the National Library server so as to monitor every production step in a particular digitization center. Therefore it is useful to develop a freely available open source solution which might be part of every digitization center infrastructure. The central Digitization Registry in the National Library will afterward harvest information from particular systems monitoring digitization workflow. The open source workflow monitoring solution, described above, should be developed as part of the Czech Digital Library project.

Other tools will be developed to support complex digitization processes in the frame of the CDL project. Included are processing and archiving digital documents. The goal is to use these tools to increase the number of materials available in the Czech Digital Library. Using the same production and archiving

tolls will enhance interoperability and data sharing between individual digitization projects. The solution developed in 2010 by the National Library and the Academy of Sciences Library, will be used as the basis of the new toll. The core of the solution is the Fedora repository. It enables the creation, storage, editing and exporting of digital documents compatible with the Kramerius system, which is used for dissemination. The rapid semi-automatic creation of the standard metadata will be enabled by the production system before its complete placement in the work process. It is comprised of structural, descriptive and archival metadata, OCR and conversion to specific graphical formats.

With regard to the archival part of the solution, standards for long term archiving, such as the OAIS model, will be implemented. Outputs from other projects, e.g., the Goobi project of the University Library in Goettingen, Germany, will be used to complete the solution.

Mutual interoperability between all developed systems and tools will be accented in the frame of the project as well as interoperability with solutions already existing on the market. The aim is to share, use and reuse digital content as easily and effectively as possible.

The project is funded by the Ministry of Culture of the Czech Republic under identification code DF12P01OVV002.