Преглед НЦД 22 (2013), 7–14

Tomas Foltyn
National Library of the Czech Republic

HAS IT BEEN ALREADY DIGITIZED? HOW TO FIND INFORMATION ABOUT DIGITIZED DOCUMENTS

Abstract: The Digitization Registry of the Czech Republic is the research project, which aim is to create a national registry of digitized documents that enables to avoid unwanted duplicities in the digitization as well to share the digitization results across the Czech Republic. This could make the digitization more effective and also save the financial resources.

Keywords: digitization registry, digitization, metadata, records, duplicities

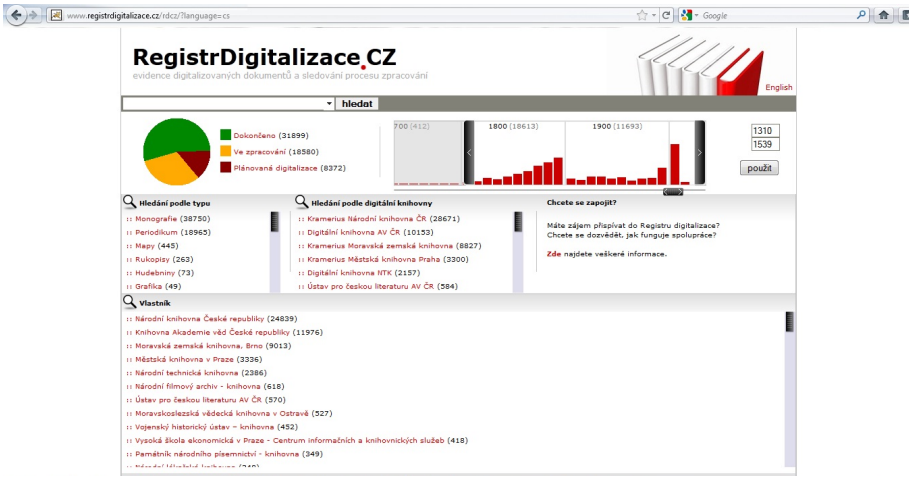
Introduction

The Digitization Registry of the Czech Republic (project acronym is RD.cz) [2] is a national project which is developed in the close cooperation of the National Library of the Czech Republic, the Academy of Sciences Library and the Incad private company. The project´s aim is to create national registry of digitized documents, which could help librarians to avoid unwanted duplicities and to enable sharing the digitization results across the Czech Republic. It is designed especially for the libraries or other memory institutions that digitize their collections. Because the system enables to avoid unwanted duplicities, it can save financial resources used for scanning and subsequent processing, thus making the digitization more effective.

1. Project background

The initiative connected with the establishment of the Digitization Registry project arose from the needs of two libraries – the National Library of the Czech Republic and the Academy of Sciences Library. The National Library increased its production of digitized documents due to the support of the Norway funds that allowed to digitize and process more than 2 400 000 pages of monographs, which had been endangered by the degradation of the paper. It was a great increase of digital content, because at the same time, the National Library is digitizing its collection of modern periodicals. Until July 2011, more than 8 100 000 pages have been is produced and available via the digital library [1, 3]. During the production of monographs, the National Library stressed the need to coordinate preparation of original documents with other heritage institutions across the Czech Republic. Special need for adequate digitization workflow was the second parallel issue. Almost the same ideas were pointed out by the Academy of Sciences Library (due to its own digitization workplace). For this purpose, a joint research project was submitted and afterwards approved.[1] The funding was coming from the Ministry of Culture. Research activities were done from 2008 to 2011. The web service was established already during the project development and it has been used since the beginning by many libraries across the Czech Republic.

[1] The project was called "Evidence of digitized documents, processing management and presentation system development (DC08P02OUK008)"

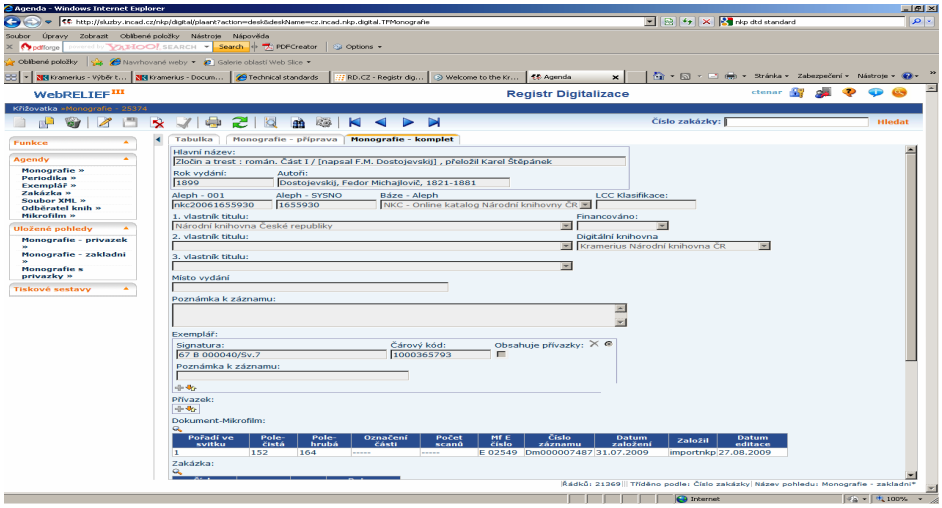


Picture 1: The front page of the webpage www.registrdigitalizace.cz

2. Interfaces and technical description

There are two user interfaces in the system. Public available interface is available on the website www.registrdigitalizace.cz, both in Czech and English language mutation. It could provide information about the location of the original document, the location in a digital library and library catalogues (including direct links to both systems), about the status of the process etc. The last information is very useful, because everyone knows immediately, whether the book has been already digitized, or it is in processing right now, or if an institution has chosen the document for digitization and will soon be processed. Afterwards, it is easily to contact the owner of the digital document and, where it is possible, to arrange conditions for replication and thereby reduce the cost of digital form. This solution is very welcomed from the final user´s perspective, especially from that one´s who want to have information about all digitized documents. To simplify the browsing, the multicolored semaphore was added to the front webpage – green color means "Document was digitized", yellow says "Document is processing" and finally red means "Document is going to be digitized soon".

From the workflow management perspective, the system is developed in the way to be able to record different working processes and different types of documents. For the evidence of the workflow the second interface is used. It is accessible only after a user's registration - with special user rights. Registered users could work with the system and insert information about microfilming, scanning, post-processing, OCR processing, metadata creation or publishing procedures.

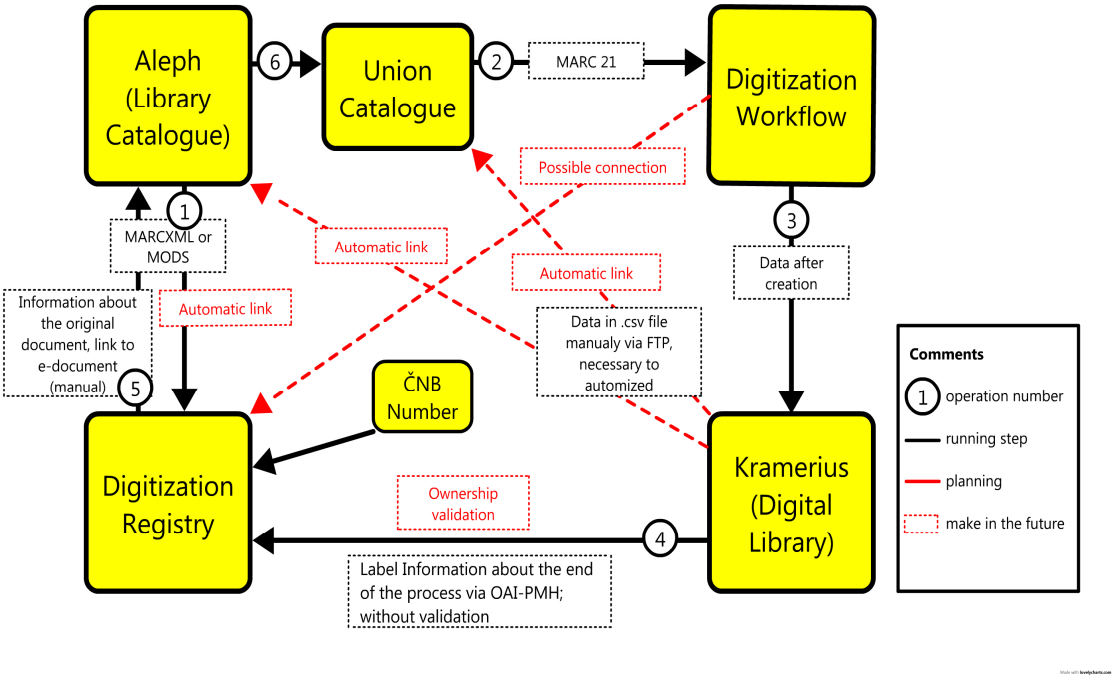


Picture 2: The digitization workflow management system (access only after personal login) [4]

The system is based on RIII application framework (J2EE) and all the data are stored in a relational database Oracle. For user access, search and retrieval FAST is used. All user accesses are implemented via web application. The Digitization Registry is independent of the operation system, and it is able to be adapted to various languages. The implementation team is working now on new version R4 which should be freely available as open source application based on FEDORA commons, Grails and Jasper.

3. Application structure logic – the connection to other library systems used in Czech Republic

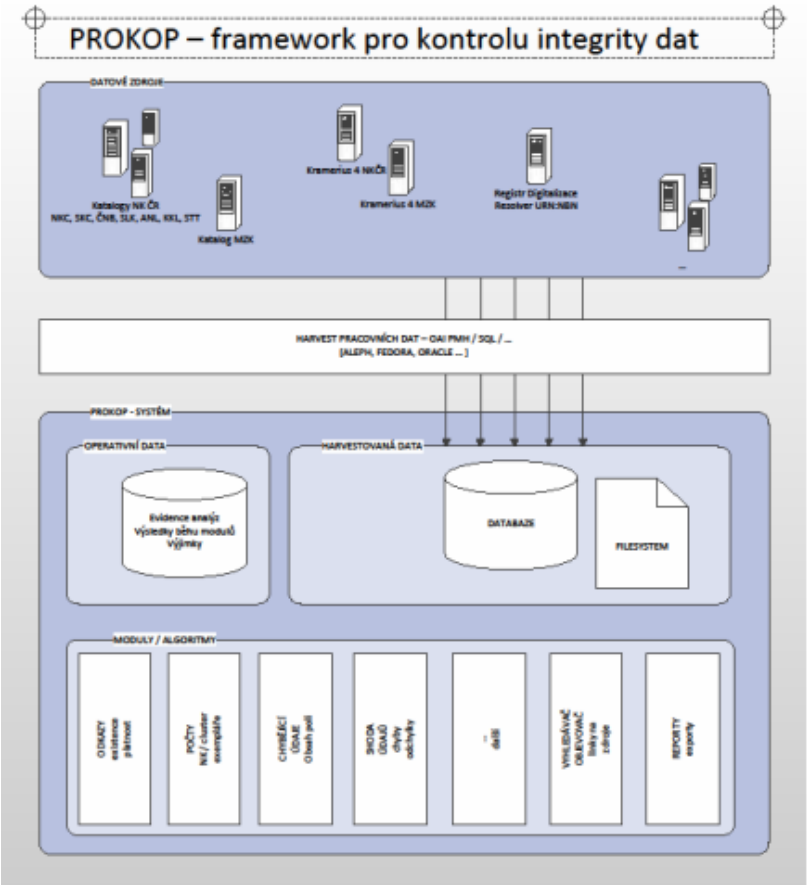
The first of the main subsystems is digitization workflow, where all metadata related to images are created. Bibliographic metadata, which are used in this workflow, are downloaded directly from the library catalogue Aleph in MARCXML format and transformed to MODS. That is why the library catalogue is the second important pillar of mentioned data flow. The third subsystem is digital library called Kramerius [3]. It needs a unique identification of digital document as well as the persistent URL. This keeps the connection between digital library and final user alive in the sphere of quotation. Digitization Registry is the last subsystem. It serves for evidence of all digitized documents in the Czech Republic. It ought to enable to avoid duplicities and to monitor life cycle of digitized documents from the planning stage to access.



Picture 3: Scheme of the library systems connection

4. Prokop – separate module for deduplication

From the digitization process perspective, it is quite easy to scan or post-process; however, to keep the metadata unique among all library systems is a big challenge. One of the biggest issues is a librarian´s effort in avoiding duplications (it is known that many wrong records were made in the past, especially in the Union Catalogue, which were re-used in other systems). The complete system used in the biggest Czech libraries is divided into four main subsystems (see Picture 3). The aim is to build a reasonable data flow.



Picture 4: Functional schema of the PROKOP system

For the connections among all the above mentioned systems many various identifiers are used e.g. Aleph System Number, MARC field 001, Czech National Bibliography Number, ISSN and ISBN etc. The main objectives of the researchers are to choose right identifier for every step, connect them together and simplify the way of sharing metadata among all systems. Based on the analysis from 2010, it is now clear that new sets of tools for data gathering from all mentioned systems shall be developed. Round-up data will be unified to one profile which would enable the analysis via multiple algorithms and subsequent evaluation. The tools shall be composed of DB/FS working space and parallel special modules compounded of several JAVA scripts, SQL procedures, etc. The output will be the file which will compile all results, and all eventual inconsistencies will be highlighted. The process will be finalized by data optimization among all the systems.
The development of the system called PROKOP started in May 2011 under one year funding by the Ministry of Culture. First results shall be available during April 2012. If it turns out that the activities concerning this module are not completed and should continue, the research team of the National Library of Czech Republic is ready to submit another project.

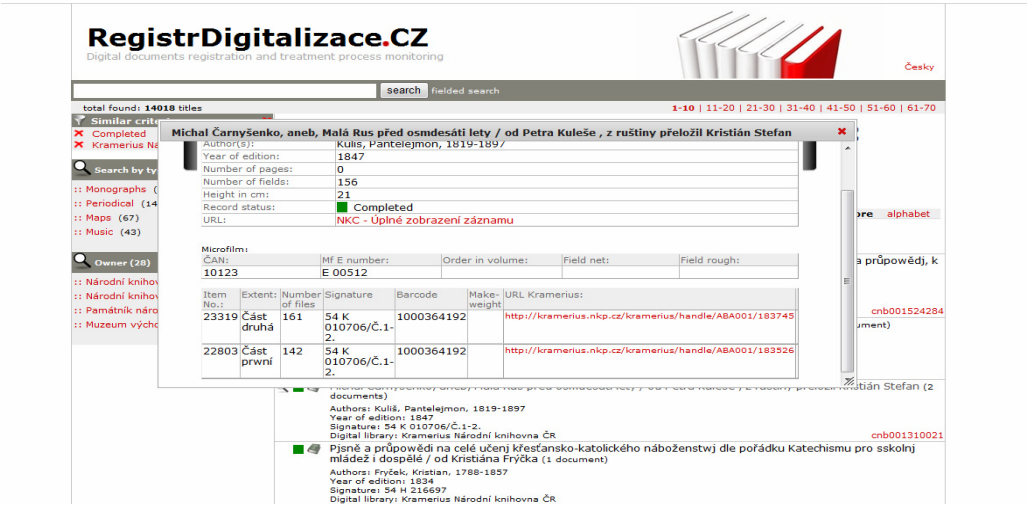
The analysis of the PROKOP system revealed that this tool can be used in broader context as well. It could be able to harvest the data from various numbers of library catalogues and to verify these records. In the Czech Republic there are three libraries with permanent store duties of the legal deposit (National Library of Czech Republic, Moravian Library Brno and Research Library Olomouc); thus PROKOP allows accurate verification of the number of Bohemian documents stored in their preservation collections. This enables to establish a virtual deposit library system for registering Bohemian documents in the reserve duplicate collections of the above mentioned libraries, and to store them in one place in the near future. Documents which will be lost from their regular collections could be replaced from this reserve collection. The next step

will be the automated offer and demand system which serves for publishing offering lists of documents sorted by libraries from their collections or required by them.. To accomplish this, a new national research project called “Building the cooperative system for the creation and management of modern preservation book´s collections in the Czech Republic and the development of needed tools” was submitted; it has been approved by the Ministry of Culture of the Czech Republic.

5. Project partners and how they can enter

Until March 2012, there were almost 60 000 records in the web application; the majority were from research collections and from the largest project partners – the Moravian State Library, the Municipal Library in Prague or the National Technical Library.[2] The owners of published documents are also other institutions. Every institution wishing to enter and insert its records, could have a brief look at the webpage section *“Do you want to join us* and to read the conditions and user guide and start cooperation.

To be able to cooperate properly, they should also learn how to browse the records. There are many ways to browse on the web service. Of course, it is possible to use a simple question by the “Google box”, or the advance searching based on metadata knowledge. Everybody could also search via the list of involved libraries (if he knows their collections and digitization activities), via list of digital libraries or via types of documents. The firs page offers also time line searching.



Picture 5: Record of requested document

The second step involves using context searching and faceted navigation. A user not satisfied with the result, can use faceted navigation; e.g, in which digital library is the document available, who is the owner of digital document or ask the most important question - whether the document has been completed, is it in the process of digitizing, or has been chosen for scanning and is waiting for production. Faceted navigation is created automatically according to displayed results.

[2] But there are many more libraries included. Among them are special libraries (National Medicine Library), regional libraries (Research Library Usti nad Labem) or foreign libraries held Bohemian collections (National Library of Slovakia).

When the requested document is found, it is possible to open the complete record about it. The basic bibliographic information is available as well as the information about the size of the original document, both in URL, in the digital library and the library catalogue etc.

6. Utilization for mass digitization projects and future development

The National Library of the Czech Republic will start the program of mass digitization together with the Moravian Library Brno. The project called "National Digital Library" was submitted in June 2010. The actual production should have begun in early 2011. The project is financed with roughly 300 million CZK (85% contribution from the ERDF structural fund and 15% co-financed from the state budget) [5].[3] Within the project, both libraries should together create about 26 000 000 pages of digitized documents.

Because of such a huge number of pages, a close cooperation and coordination is essential. Digitization Registry needs to be one of the key parts of the project. To this purpose, the development never stops. At the end of 2010 and at the beginning of 2011, new functionalities connected with mass digitization needs were adopted. New data models were included, such as a deeper automatic connection between digital library and registry, and especially the unique web service solution for sharing the metadata between almost every digitization workflow system and the registry. This web service uses the technology of SOAP/REST questions and fully supports four most used file formats, including MODS standard.

In the late 2011, the new utilities for Digitization Registry web service were added. These functionalities were also requested by other Czech libraries. All of them were connected with:

a) better way for presenting a content from other libraries

b) new methodology of libraries accesses' to the registry – every library shall can update its data in its own way

c) deeper connection to library catalogue ALEPH based on persistent identification

d) new modules for the harvesting and the possibility of previews that enables better orientation in the registry content

e) federated searching of important public resources obtained also from abroad (e. g. National Library of Austria digitization activities)

f) new form creation which enables to browse the registry via the information about funding of every document

Together with mentioned mass digitization activities of the National Library of the Czech Republic and the Moravian Library Brno, new workplaces for digitization in Czech districts will be established. Their production will be aimed not only to the regional library collections, but to the archival documents, Civil Service materials or to the medical resource as well. The digitization registry will be the main and the most important management point where all records and information about digitization of the library collections will be stored.

7. International scope of the registry and future development

Because of the fact that the described project is unique in the European context, a project proposal for the creation of the Central Registry of the Central European countries was submitted in April 2010. Besides the Czech partners, the important institutions from Slovakia, Slovenia, Poland, Hungary, Germany and Austria were invited. But unfortunately the project was not approved by review committee, although the project was in the final round of the competition. Now, the Czech researchers are waiting for other EU program to submit a

[3] More information about the project are available on the webpage.

new project entitled Towards the Central European Countries Digitization Registry. In the case that the funding will be found, the research team ensures the creation of registries at the national level and harvesting them to a central registry, which will be developed in parallel.

During May 2011, the Digitization Registry project was approved by the members of the Czech Union Library Committee; it is responsible for library policy in the Czech Republic and could address /establish/ the recommendation directly to the Minister of Culture. The committee fully agreed with the importance of the system, asking the National Library of the Czech Republic to be the official head of the project, and to persuade the management of the National Library to employ a full-time worker responsible for the content of the registry and for the cooperation between the Czech cultural heritage institutions. That was a very important step for the registry because the research team could ask the ministry for the support for future development.
By mid - 2012, the research team will be working on the stabilization of the system and parallel highlights by testing to verify the robustness of the application. After that, the team will be ready for the first mass digitization inputs to determinate where the potential problems could arise and solve them immediately.

8. Conclusion

For libraries looking for a good, sophisticated system for their digital documents monitoring, -which should be also user friendly-, the Digitization Registry could be the right solution. The credibility of the system is already proved by its wide use in the Czech Republic by millions of digitized pages; it is available to all interested institutions and digitization projects. There are no limits to distribute and use it worldwide, and manage the digitization activities not only on the national level. The closest cooperation, especially in the case of mass digitization activities, could save a lot of money, time and personal resources.

References

Articles:
[1] Tomas Foltyn, *The Kramerius System – Open Source Solution for Digital Libraries*, Proceedings of the Third Workshop on Very Large Digital Libraries; Glasgow, Scotland (UK) September 10, 2010, Pisa 2010, ISBN 978-88628015-7

Webpages:
[2] www.registrdigitalizace.cz
[3] http://kramerius.nkp.cz/kramerius/Welcome.do?lang=en
[4] http://sluzby.incad.cz/nkp/digital/plaant?action=login
[5] www.ndk.cz

Tomas.Foltyn@nkp.cz