Преглед НЦД 22 (2013), 75–80

Cezary Mazurek, Tomasz Parkoła, Marcin Werla
Poznan Supercomputing and Networking Center, Poznań, Poland

TOOLS FOR MASS DIGITIZATION AND LONG-TERM PRESERVATION IN CULTURAL HERITAGE INSTITUTIONS

Abstract: This paper describes the present state of digital libraries infrastructure in Poland, with focus on the tools crucial for daily digital libraries duties, such as online availability, long-term preservation and digitization workflow management. Currently, there are over 80 digital libraries in Poland, providing access to over 1 000 000 digital objects. Such a great amount of resources and digitization efforts requires specialized tools for the management and safety of the resulting assets. Poznan Supercomputing and Networking Center has been actively supporting digitization with innovative ICT technologies and R&D outputs to foster and improve digitization in Poland. Therefore, PSNC has developed several software tools for this purpose, including dLibra, dMuseion, dArceo and dLab.

Keywords: mass digitization, digital library, digitization workflow, tools, cultural heritage

1. Introduction

Digitization activities, including mass digitization, text recognition and long-term preservation play very significant role in the cultural heritage domain [1], especially in context of discovery and access to collections. Digital assets representing cultural heritage objects are important as they allow protecting original objects, and at the same time, give easy access for reuse in education, research and commerce. Successful projects related to digitized historical content like Europeana, American Memory, Trove and SCAPE prove this importance.

In Poland hundreds of cultural heritage institutions are involved in digitization activities, building together a network of over 80 Polish digital libraries, and making over 1 000 000 digital objects accessible via the web browser [2]. Over 75% of them use for this purpose dLibra Digital Library Framework (http://dlibra.psnc.pl/) which is focused on management and on-line presentation of large number of objects; it has been developed since 1999 by the Poznan Supercomputing and Networking Center (PSNC). Experience from multiple digital libraries shows that coordination of resources, tools and other aspects of digitization activities is complex: often the efficiency of such process is the most problematic barrier to mass digitization and online accessibility. Recent PSNC works in the digital libraries domain were focused on the above issues. As a result, PSNC provided a new kind of software infrastructure for cultural heritage institutions. Text recognition (OCR), cooperation with external digitization centres, automated conversion of the source data to various formats, long-term preservation and on-line access are all covered by the infrastructure, which is composed of dLibra, dMuseion, dLab and dArceo tools. This infrastructure opens a new perspective for the Polish cultural heritage institutions in terms of mass digitization.

The second chapter of this paper presents dLibra and dMuseion that are dedicated to the presentation and online availability of the digital resources, with dedicated features for integration with portals such as Europeana, DART-Europe or ViFaOst. The third chapter presents the innovative approach to long-term preservation with the focus on OAIS migration approach and flexible cooperation model. The fourth chapter presents the possible and

exemplary model for the usage of dLibra/dMuseion and dArceo tools that are managed by the digitization workflow management system dLab. The last chapter summarizes and concludes this article.

2. Online accessibility of the digital resources

dLibra digital library framework has been developed by PSNC since 1999 and is a base for more than 60 digital libraries in Poland. The functionality has been defined in course of multiple meetings, workshops and conferences, giving the possibility for institutional users to express their most needed features. The main users of the dLibra software in this context are the Polish scientific and cultural heritage institutions, including public and university libraries, museums, archives, foundations and even individuals. The broad spectrum of digital content provided by these institutions via the means of dLibra receive much of appreciation from Polish and foreign end-users, including scientists, students and hobbyists.

The core functionality of dLibra is focused on the presentation and structuring of the digital resources [3]. dLibra can store digital objects in any format and can organize them in so called collections, supported by various indexes (e.g. authors), news feeds, community building features as well as advanced searching mechanism. The special function for grouping digital objects is widely used for periodicals to structure them according to the subsequent years, months and issues, but also for other multivolume objects. The software itself is composed of several services that can be set up on different servers, giving the scalability and flexibility of the overall solution. These services are utilised by two client application. The first one, Editor's and Administrator's Application is dedicated to the managers of the digital library. It is used to create the structure of the collections, submit digital objects either one by one or in a massive manner, create and edit or import (e.g. from MARC21) necessary metadata and manage access rights to the digital object or a set of objects. The second one, Reader's Application is a web-based front-end dedicated to the users of the digital resources, allowing for searching with narrowing feature, browsing digital object with the use of collections and groups of digital objects, as well as tagging and favourite functions.

dMuseion is very similar to dLibra in terms of the technical nature, as it is similarly composed of several services with same extendibility and flexibility capabilities. But dMuseion is very different on the user level, as the nature of the museum objects, usually visual arts holdings, has more demands related to the appearance rather that the textual content [4]. Therefore dMuseion features not only collections of objects, but also themed collections representing exhibitions with innovative function of printing a QR-Core for a certain holding in the digital museum. This QR-code encodes a website link that provides detailed information about this specific holding. Such a QR-code can then be attached to a real object in the museum hall, allowing museum visitors to get more detailed information about the holdings while visiting the museum. The other features, such as complex objects, support for galleries of pictures and 3D digital objects constitute a fully functional digital museum with advanced and innovative offer for any internet user who wants to investigate the digital resources with special focus on visual art holdings. dMuseion has been developed by PSNC since 2009 in cooperation with National Museum in Warsaw and is used by the Digital National Museum in Warsaw portal (http://cyfrowe.mnw.art.pl/).

Both dLibra and dMuseion can provide metadata of stored digital objects via OAI-PMH and OAI-ORE protocols, providing therefore means for Polish national metadata aggregator Digital Libraries Federation (http://fbc.pionier.net.pl,) to harvest metadata, make them consistent

across multiple digital libraries, and pass them to international aggregators such as Europeana, DART-Europe and ViFaOst.

3. Long-term preservation of the digital resources

One of the emerging challenges in the digital libraries in Poland is the long-term preservation, which aims to make the digital content available in a year, ten years, twenty years, etc. despite of the fact that software, hardware and formats change. This issue is important because the digital libraries are in fact only access points to the light versions (presentation versions) of the original digitized content (master files). This is even more important in light of the survey conducted in scope of the SYNAT project which proved that the need for the reliable, extendible and efficient software solution is a must in terms of long-term accessibility of the digital content. This is because almost none of the Polish institutions involved in the mass digitisation projects uses the long-term preservation tools in the context of their content. Therefore in the framework of the SYNAT project, funded by the Polish National Centre for Research and Development, PSNC has approached the issue of long-term preservation services in the context of cultural heritage resources. The idea is to advance the long-term preservation activities by giving the community means to build services attached to cultural heritage digital archives and digital libraries.

The newly developed prototype tool, called WRDZ (in Polish: *Wielofunkcyjne Repozytorium Danych Źródłowych*), has been enhanced by PSNC to a production level solution. This resulted in the dArceo software package that can be easily used in the context of the digitisation workflows and long-term preservation. dArceo consists of a number of network services primarily focused on the preservation of the texts, images and audio-visual content. It is an open-source solution, which is a complex offer for scientific and cultural heritage institutions, providing means to perform migration according to the OAIS model, conversion for the needs of building digital libraries, and advanced delivery for research purposes (e.g. transcription tools or master files viewer). The overall model of the dArceo deployments enables capability of sharing all the data manipulation services between institutions in a P2P-like manner.

Moreover, dArceo can be adjusted to utilize the results of the Polish national PLATON project (http://www.platon.pionier.net.pl/) and its archiving services PLATON-U4. The main goal of PLATON-U4 is to enable data archiving (including replicas located in geographically distant location) with the possibility to have on-line access to the backup and restoration in the case of basic data loss, 24 hours, 7 days a week. The archiving data nodes are located in 10 Polish cities and are connected via the high speed optical network PIONIER [5]. dArceo builds on top of the PLATON-U4 services and benefits from their functionality, therefore does not focus on the storage and replication issues, but rather on long-term preservation challenges, such as data format migration and monitoring.

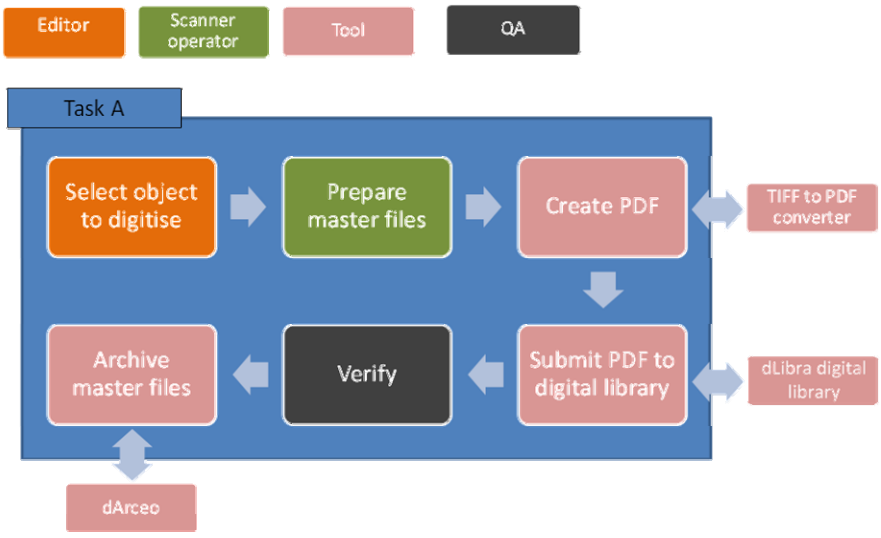
4. Complex software solution for mass digitization projects

With the growth of the digitization activities and digital resource, the need for specialized tools dedicated to digitization workflow management has appeared. In order to fill this gap PSNC in cooperation with Digital Repository of Scientific Institutes (DRSI) has developed dLab – a software tool for digitization workflow management, which is capable of handling digitisation tasks composed of certain activities related to particular stages in the digitisation workflow. DRSI is a consortium of 16 institutes of the Polish Academy of Sciences with an ambitious plan to

digitise millions of objects from their library collections. In the framework of DRSI there are several digitisation centres specialised in digitisation of particular types of documents, multiple libraries utilising these centres for massive digitisation and one portal (http://rcin.org.pl/) for presentation versions of the digital objects, being an effect of the digitisation activities.

The idea of dLab is that a user creates a digitization task which is composed of activities. For example, a very simple list of activities for a digitization task could include: select object to be digitised, prepare master files, create PDF, submit PDF to digital library, verify and archive master files in the long-term preservation system (see Picture below). Additionally, each activity can be performed by a user (human) or by a machine (automated tool). In the default configuration dLab uses OCR engine to perform text recognition, Document Express or FineReader packages to create presentation version in DjVu or PDF format, dLibra or dMuseion to make the presentation version available online and dArceo services to preserve master files. Moreover, dLab is also fully configurable in terms of the activities to be performed in the scope of a digitization task, it can be extended by plugins therefore integrated with multiple tools, e.g. ImageMagick, PDFBox or Tesseract.

The first dLab deployment, currently in the test phase, has been deployed in scope of the DRSI portal in 2011. An exemplary digitisation workflow handled by dLab tool is depicted on figure below. As the dLab can be configured in such a way that certain activity is assigned to and performed by a particular actor (a user or a tool), the colours on the figure represent different types of actors, including: Editor, Scanner operator, Tool and QA (quality assurer). In the example task (Task A) there are six activities and four types of actors. In order to perform digitisation task, all of its activities has to be executed. First of all the Editor has to select an object to digitise. As soon as this is done Scanner operator is responsible for digitisation and preparation of the master files. Next, an external tool is responsible for creation of the PDF based on the master files. The PDF is a presentation version of the digital object and is submitted by another tool, during the execution of the next activity, to dLibra-based digital library. After the digital object is submitted to a digital library, quality assurer verifies all activities performed in scope of this digitisation task. After a positive verification the master files are archived by yet another tool in the dArceo services. The archiving process is done by the dLab extension which communicates with dArceo services and creates in it a new digital object.



Picture 1: Examplary digitsation task in the dlab tool

As seen on the example dLibra, dArceo and dLab are used as the complex software solution for this particular digital library infrastructure, where multiple tools need to be involved. This solution is used by the Digital Repository of Scientific Institutes for the needs of: long-term preservation of the master files, digitization workflow management as well as online presentation of the digital object.

5. Summary

dLibra and dMuseion are both dedicated to online presentation of digital assets. While dLibra is focused on library documents with strong support for texts, periodicals and document collections, dMuseion is related to museum objects with focus on themed collections and visual presentation aspects. dLibra and dMuseion are capable to import metadata from bibliographic records or inventory system records, making communication with external systems easier and faster and allowing to reuse data. Moreover, both are able to provide metadata to external systems via the OAI-PMH and OAI-ORE protocols.

dArceo is composed of multiple services responsible for realization of the long-term preservation idea, primarily for the textual, graphical and audio-visual content. The basic idea is an OAIS transformation approach to migration with support for conversion and advanced delivery services. Additionally all the data manipulation services can be shared, so that various institutions can benefit from already available migrations, conversions or advanced delivery techniques. dArceo has been developed in frame of the SYNAT project, financed by the Polish National Centre for Research and Development.

Finally, dLab is a dedicated software tool for the management of simple and complex digitization workflows, featuring digitization tasks management and pluggable architecture for the needs of integration with external tools.

The composition of the dLab tool and dArceo, together with dLibra digital library system or dMuseion digital museum system creates a complex software solution for all scientific and cultural institutions that are focused on mass digitisation activities.

Currently dLab and dArceo are used by the Digital Repository of Scientific Institutes – consortium of 16 institutes of the Polish Academy of Sciences (http://rcin.org.pl/). dLibra is a well-known digital library software, used by over 80 digital libraries in Poland, with several deployments abroad. dMuseion is successfully used by the Digitial National Museum in Warsaw (http://cyfrowe.mnw.art.pl/).

References

[1] Comité des Sages, *The New Rainessance*, http://ec.europa.eu/information_society/activities/digital_libraries/doc/refgroup/final_report_cds.pdf

[2] A. Lewandowska, C. Mazurek, M. Werla, *Enrichment of European Digital Resources by Federating Regional Digital Libraries in Poland,* In Research and Advanced Technology for Digital Libraries. 12th European Conference, ECDL 2008, Aarhus, Denmark, 14 – 19 września 2008, Proceedings. LNCS vol. 5173, str. 256–259.

[3] C. Mazurek, M. Werla, *Digital Object Lifecycle in dLibra Digital Library Framework.* Proceedings of DELOS 9th Thematic Workshop: Digital Repositories, 11 – 13 maj, 2005, Heraklion, Grecja.

[4] P.P. Czyż, M. Romeyko-Hurko, *dMuseion: od bazy danych do muzeum cyfrowego*. Konferencja „Polskie Biblioteki Cyfrowe", 9 grudnia 2009, Poznań. Materiały konferencyjne, str. 21-29. ISBN 978-83-7712-020-0.

[5] Brzeźniak Maciej, Meyer Norbert, Mikołajczak Rafał, Jankowski Gracjan, Jankowski Michał. *Popular Backup/Archival Service and its Application for the Archival of the Network Traffic in the PIONIER Academic Network*. In Computational Methods in Science and Technology, 2010, Special Issue, p. 109-118.

mazurek@man.poznan.pl
tparkola@man.poznan.pl
mwerla@man.poznan.pl