Enterprise Information Management in Cultural Heritage Domain

Cezary Mazurek, Marcin Werla

Enterprise Information Management in Cultural Heritage Domain

Cezary Mazurek[1][0000-0002-8715-9326] and Marcin Werla[1][0000-0002-5874-9757]

[1] Poznań Supercomputing and Networking Center, Jana Pawła II 10, 61-139 Poznań, Poland
{mazurek,mwerla}@man.poznan.pl

Abstract. The aim of this paper is to present the complexity of information management in cultural heritage domain on the basis of real-life examples of distributed research infrastructures for the arts and humanities. The digitisation of cultural heritage artefacts is a process that is ongoing for many years in institutions all over the world and generates increasing amount of digital information. This generates challenges on the level of particular institutions which are holding heritage collections, but also on the national and international level, where such information is combined to provide end users unified access to distributed heritage datasets. The paper presents the flow of cultural heritage information from the level of a single institution up to pan-European data platform Europeana.eu via the level of regional and national cultural heritage data services. It is based on experiences of Poznań Supercomputing and Networking Center collected during the last 15+ years of involvement in the development of cultural heritage network services on all these levels.

Keywords: Digital Cultural Heritage, Metadata Harvesting, Metadata Aggregation, Digitisation Management, Digital Humanities, Interoperability.

1 Overview of Cultural Heritage Data Landscape

The most important information systems on the level of a single memory institution are those which are used to manage information about physical collections of the institution. They can be called library catalogues, integrated library systems, museum inventory management systems or archival information systems. Their key responsibility is to collect information about the cultural heritage assets and support management of these assets. To enable that, some of the systems are not only holding data records but are also supporting domain-specific procedures like inter-library book loans or museum objects exchange. In some domains, these systems are based on sets of well-defined rules or procedures, like for example, SPECTRUM standard [1]. Part of the information stored in such systems is describing the basic features of given physical object and is called descriptive data (for example title, author, date of creation etc.).

When such an object is digitised, it becomes data and the descriptive data becomes metadata (data describing data). Digital objects with their metadata are usually managed by another type of systems, specialised in digital assets management and called digital libraries, museums, archives or in general repositories. Such repositories

have to handle various types of data (as various as digitisation outcomes can be) and various types of metadata – descriptive, administrative, technical, structural etc.

Part of that information (both data and metadata) can be made available online for a wider audience through institution's web portal or online collections system. Such system is usually indexed by Google and other search engines, but in many cases, it also can actively distribute information (usually metadata) to other platforms, called metadata aggregators. Such platforms aggregate information from several sources, depending on their scope of interest. Moreover, aggregators can provide data to other aggregators (for example with wider scope), which makes them an interconnected network of cultural heritage data services.

Further in the chain of data flow are domain specific services, often very narrow, focused on re-use of cultural heritage data and metadata in a specific context, usually education, humanities research or tourism. Another re-use community is set up by so-called creative industries, like graphic designers, game developers or artists.

The following sections of this paper describe examples of consecutive stages of cultural heritage data flow, starting on a level of single cultural heritage institution, through regional data platforms, to national and international level metadata aggregators, to re-use environments. For each of these stages, real-life examples are provided and major challenges are emphasized. Section two focuses on an institutional level information management in the processes of digitisation, long termin digital preservation and provision of on-line access to digital cultural heritage collections. Section three provides an ovierview of regional cultural heritage data platforms and section three shows, how data from such platforms is aggregated and distributed on national and European level. The papers ends with a summary and conclusions.

2 From Offline Physical to Online Digital – Digitisation Workflow Management

The basic digitization workflow consists of following high-level steps: (1) selection of objects for digitisation, (2) preparation of objects and digitisation, (3) postprocessing, (4) digital archiving, (5) online publication. During all these steps digital information is created, processed and transferred. The type and amount of information, as well as the level of automation of the entire process, depends strongly on the context of the specific institution. Solutions used to support the process can vary from simple spreadsheet-based information management to sophisticated systems like dLab developed by PSNC [2].

In dLab system the core element of the information model is a digitisation task. It usually corresponds to all work related to and outcomes of digitisation of a single physical object. Such a single object can be a painting, a postcard, a book or a full year of issues of a journal which are bound together in one binding. Therefore an outcome of a single digitisation task may be one but also more digital objects which are at the end archived and made available online. The progress of execution of a single task is monitored by the dLab system and reported to managers of the digitisation process/project.

Each digitisation task is divided into activities. Activities correspond to single actions that have to be taken in order to perform entire task. Examples of such activities are: digitisation of an object, cropping and deskewing of output scans, running OCR processing, converting digitisation output from one file format to another. Usually, there are finish-to-start dependencies between tasks which determine the ordering of execution (e.g. digitisation -> postprocessing -> OCR), but there are also parts of the workflow, which can be executed in parallel (e.g. when all files are ready and quality checked, archiving and online publication can happen in parallel). Moreover, some activities can be executed automatically (like file format conversion, technical metadata extraction) while others require human involvement (quality verification). For some specific purposes, tasks can be grouped so that there is one activity that is shared for all grouped tasks. A typical scenario for such grouping is an activity of physical transportation of a number of books from library storage room to digitisation lab, and back. The transportation is done at once for a group of books (tasks), digitisation is done one by one, and transportation back is again done in a batch.

dLab system holds the full history of execution of a specific task, including all its activities. Each activity may be in a number of states, depending on the progress of the task (see Fig. 1). dLab system includes the following basic states:

- Waiting – cannot be executed because of dependencies to other activities.
- To do – can be executed.
- Doing – execution of the activity is in progress
- Done – was executed properly, but no quality assurance was made on the results.
- Accepted – the results of execution were accepted by the quality assurance person.

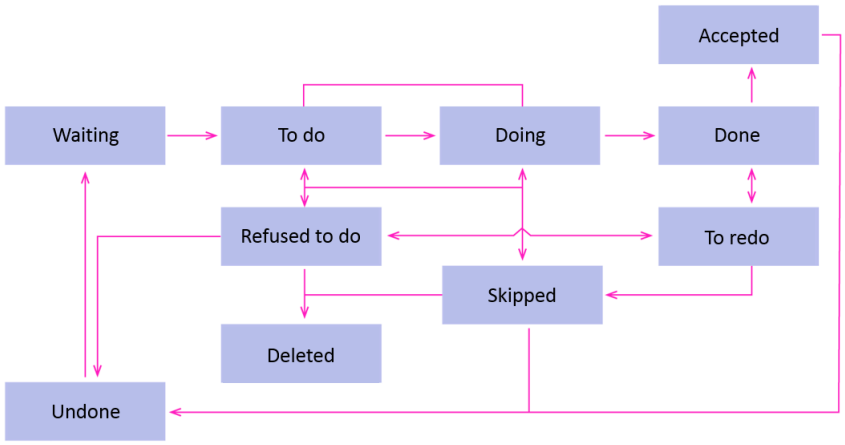

Fig. 1. State transitions of activities in a digitisation task in the dLab system.

Additionally, for exceptional situations, a number of additional states are available, including:

- Refused to do – when the task was supposed to be executed, but before the execution, it came out that it's not possible. For example, a book was supposed to be digitised, but it came out that it has damaged pages which should be fixed prior to digitisation.
- To redo – after execution, the results were not accepted by quality assurance and the activity has to be repeated.
- Undone – after proper execution of the activity it came out that one of the previous activities was executed improperly and an entire subset of activities has to be repeated. For example, OCR activity was executed and then it came out that few activities earlier, the digitisation was done wrong, therefore digitisation and all further activities have to be withdrawn and repeated.
- Skipped – used when an activity was optional and the system operator decided to skip it in particular case.

Users in the dLab system have various roles which determine their permissions, assignments to specific activities and operations which they can do on such activities (e.g. execution or quality assurance).

This shows, how complex a digitisation workflow can be, especially if the memory institution is involved in a mass digitisation project during which several hundred thousands of objects are supposed to be digitised within the precise time period.

As mentioned above, usually two parallel final steps of digitisation are digital archiving and online publication. Digital archiving is often based to some extent on the OAIS reference model [3]. It assumes that information is contained within packages, that have structure and content dependent on the stage of archiving process. Three main types of information packages are submission, archive and dissemination. Submission information package is the input to the archiving process, and usually, at the same time, it is a full output of the digitisation activities. During ingestion of information package to the archive, it is transformed into archive information package, which should contain as much information as possible and be independent of the archival system, so that if the system is gone for some reason, the archive information package is so self-explanatory, that the information within can be accessed. This assumption, for example, makes encryption of archive information packages really deprecated, as the encryption key has to be stored somewhere outside of the package and makes access to the information inside the package dependent on some extra information stored somewhere outside of the package. Dissemination information package is obtained by processing of the archive information package to optimise the information for delivery to specific, authorized user. Usually, it may mean that only a subset of information from the archival information package is copied to the dissemination package (for example only part of the metadata or lower resolution images).

The second final step, online publishing of the digital object, is usually done by submitting a dedicated, optimised for online viewing, version of the digital object to the online collections system of the institution. The format of the object, as well as the set of metadata attached to it, depends on the possibilities and design of that online portal. In Poland, many regional level institutions decided to cooperate on that stage, by creating regional data platforms that are shared infrastructures for these institutions and allow them to optimise several aspects of the online delivery of cultural heritage

content. In order to support that PSNC developed the dLibra system [4], which together with dLab (for digitisation workflow management) and dArceo (for long-term digital preservation) create together DInGO toolset (https://dingo.psnc.pl/) used by several hundred institutions in Poland. The next section provides an overview of selected information management aspects related to online access to cultural heritage objects.

3 Regional Data Platforms

Systems used to manage physical collections are usually treated as internal to the institution. Sometimes they have a window to WWW which provides part of the functionality, and part of the information, which can be accessible to users external to the institution – to the general public or for example to registered readers of a given library. In case of smaller organisations, such systems may be in fact desktop installations, existing on (the only) one PC in the institution.

Portals which are used to make digital collections available online are usually more demanding in terms of hardware resources than systems used to manage institution's inventory. Reasons for that are twofold. First of all, the amount of data stored in digital repositories is usually several orders of magnitude bigger than the amount stored in library or museum catalogues. High-resolution scans of a thick book can take several gigabytes, while the catalogue record describing that book takes not more than few kilobytes. This difference requires not only more powerful hardware to store and process the data, but also significantly more network bandwidth, to make sure that the digital content may be transferred to online users. Secondly, the group of potential users of a digital library or museum is usually much bigger when compared to a basic online catalogue. Information available in the online catalogue is usually most interesting for users geographically close to the institution because if they will find an interesting position in the catalogue, they can go and access the physical object on site. In case of digitized objects, the access is ensured via online data transfer and can be utilized by users all around the world, as long as they have the network connection. It means that not only the servicing of a single user session requires more resources, but also there will be most probably more user sessions than in case of other institution's services of a more local nature.

It makes the creation of a sustainable digital library or museum a challenge, which in Poland is widely addressed be setting up of so-called regional digital libraries [5]. Such regional initiatives are playing the role of regional platforms for cultural heritage data and metadata, which are maintained by bigger institutions like university libraries or local research computing centres and used by many other institutions from the region, creating consortia of several tens of partners. From the information management point of view, it is interesting how this approach allows for a consensus between the benefits of shared approach and need for emphasis on the unique institutional identity of each consortium partner. Shared infrastructure approach requires all partners to agree on shared metadata schema and common layout of digital collections. The institutional identity remains visible thanks to three levels on which it can be expressed. The most basic level used in almost all regional platforms in Poland is related to a metadata

schema, which usually in such cases includes provenance field used to indicate to which institution a particular object belongs. The second level is implemented with institutional collections, sets of objects from a single institution, which are promoted in the web portal of the regional digital library. With such approach (see Fig. 2) anyone visiting the main page of the regional portal have a chance to see the institutions contributing to it and browse their digitised objects.

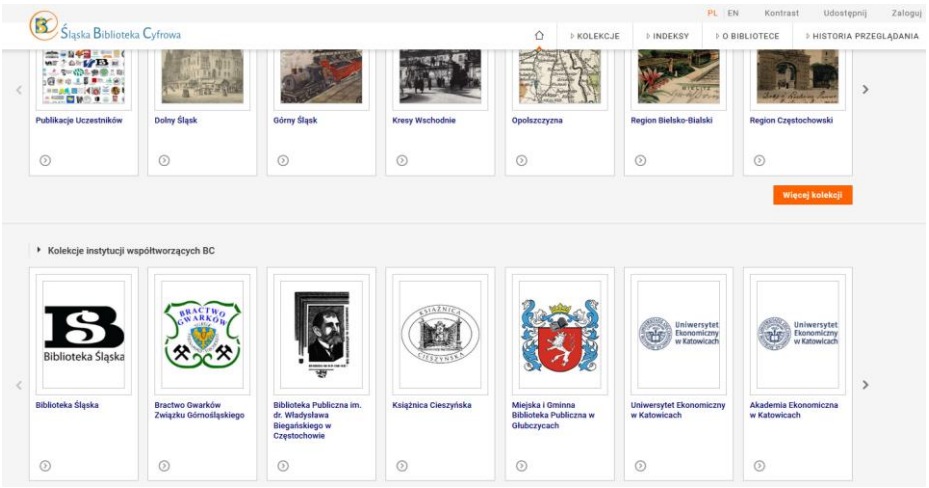

Fig. 2. The main page of Silesian Digital Library (http://sbc.org.pl/) with institutional collections promoted in the bottom row visible at the screenshot.

The ultimate way to ensure institutional visibility on such regional data platform is a setup of dedicated web portal based on that platform, which has a unique design with strong institutional identity and presents only objects which are provided to the platform from that specific institution. Such approach, called the virtual digital library (see Fig. 3), is not the most popular one, as it requires additional hardware resources, but it helps in some cases to liaise between regional cooperation and institutional ambitions.

The regional data platforms approach allowed hundreds of medium and small Polish institutions to start publishing their digitised collections online. Nowadays over 40 regional digital libraries exist in Poland and together with institutional digital libraries and repositories, they make available online around 5 million objects. Information about these objects is aggregated on the national level in a dedicated service called Polish Digital Libraries Federation. This portal and its European counterpart are described in the next section.

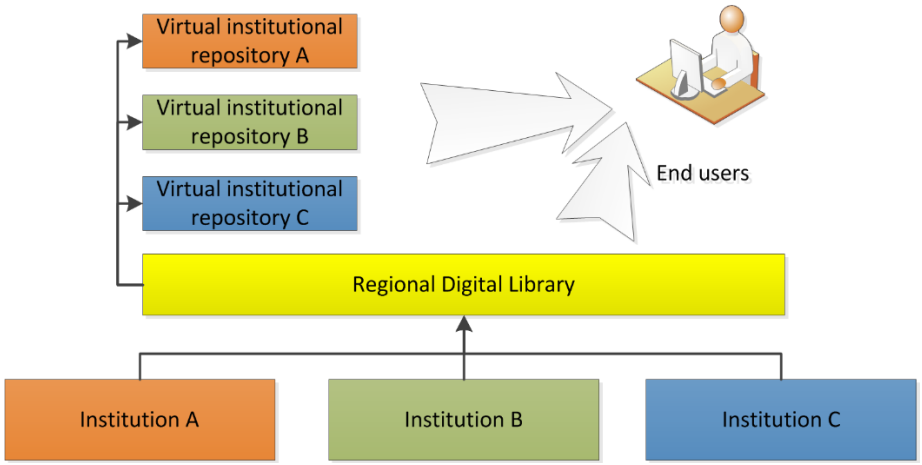

Fig. 3. Virtual repositories based on a regional digital library platform.

4 National and International Metadata Aggregators

As the development of digital libraries in Poland was progressing, in mid-2000's it became visible that the increasing number of online cultural heritage collections need a unified access point. General purpose internet search engines were indexing the metadata and data from digital libraries, but it was not possible to run user queries only over a subset of websites which were providing high quality and trusted data. Also, the way of query formulation was not ideal, as it was not supporting advanced queries connecting several metadata fields. Therefore PSNC started developing a solution that could provide more homogeneous access to distributed cultural heritage resources.

The public opening of the service took place in 2007 and since then the Polish Digital Libraries Federation is available at http://fbc.pionier.net.pl/ [6]. At the beginning it was providing access to around 80 000 objects from 15 cultural heritage institutions, now it is indexing over 120 institutions and provides access to around 5 million objects. The aggregation of the information was in the beginning based on the OAI-PMH protocol [7] and was using Dublin Core metadata schema [8] as a common standard for storing and indexing of the data. The initial assumption was that digital libraries cooperating with the Federation will be able to map information from their internal schemas to the widely know Dublin Core Metadata Element Set. After a few years of development and maintenance of the service it came out, that in many cases the mapping is established once when the particular digital library starts cooperation with the Federation and then is no longer maintained – it is not updated when the digital library modifies its metadata schema. Beside the Dublin Core set of metadata field came out to be too narrow to hold the semantics of the aggregated data which was leading to a serious loss of data quality. For example, the place of publishing of books had to be mapped to a "Publisher" field in the Dublin Core schema – a field that should hold the name of the publishing entity.

Finally, the initial technical architecture [9] was not scalable enough to support the increasing amount of data sources and metadata volume.

This led to the redesign of the entire system and development of new aggregation platform based on cloud technologies [10]. The new system was from the beginning designed with a focus on scalability and high availability. Also, the approach to information integration was redefined. The main way of obtaining data is now still OAI-PMH protocol, but support for simple CSV file delivery was added. Besides, now the assumption is that the Federation should aggregate as rich data as possible and handle the mapping of information on its own side, not relying on the data provider doing the mapping. This requires more effort related to the daily operations of the Federation infrastructure but can be to some extent automated and leads to significantly better data quality. The architecture of the new system consists of the following layers (see Fig. 4):

- Data storage – a most central layer of the system used to exchange data between all other layers, based on NoSQL database cluster.
- Data aggregation – scalable set of agents which are responsible for communication with data sources and delivery of data to the data storage layer. Agents are managed by a central agents manager and are specialized in different types of data sources. New instances of agents can be created if new data sources are added and new types of agents can be implemented to provide support for new types of data sources.
- Data processing – scalable set of processing components executing predefined chains of (meta)data transformation, for example, mapping from one schema to another, cleaning, normalisation, enrichment etc. Data processing components load data from and store data to the data storage layer and are triggered automatically as new data arrives from agents to data storage.
- Data provisioning layer – used to provide read-only access to data storage with a REST API that can be utilized to implement applications on top of the aggregated data. Example of such application is the portal of the Polish Digital Libraries Federation.

Besides providing access to the aggregated (meta)data via the Federation portal, PSNC also cooperates with European-level cultural heritage data platform called Europeana.eu. This platform to some extent repeats the scenario of the Federation, but is working on an international level and is getting its data mostly from aggregation services like the Polish Federation. Europeana is operating on a scale larger by one order of magnitude than its national level partners, with dedicated ingestion team that is responsible for data processing and is ensuring the best possible data quality. Europeana is accepting the data in a dedicated schema called Europeana Data Model (EDM) [11].

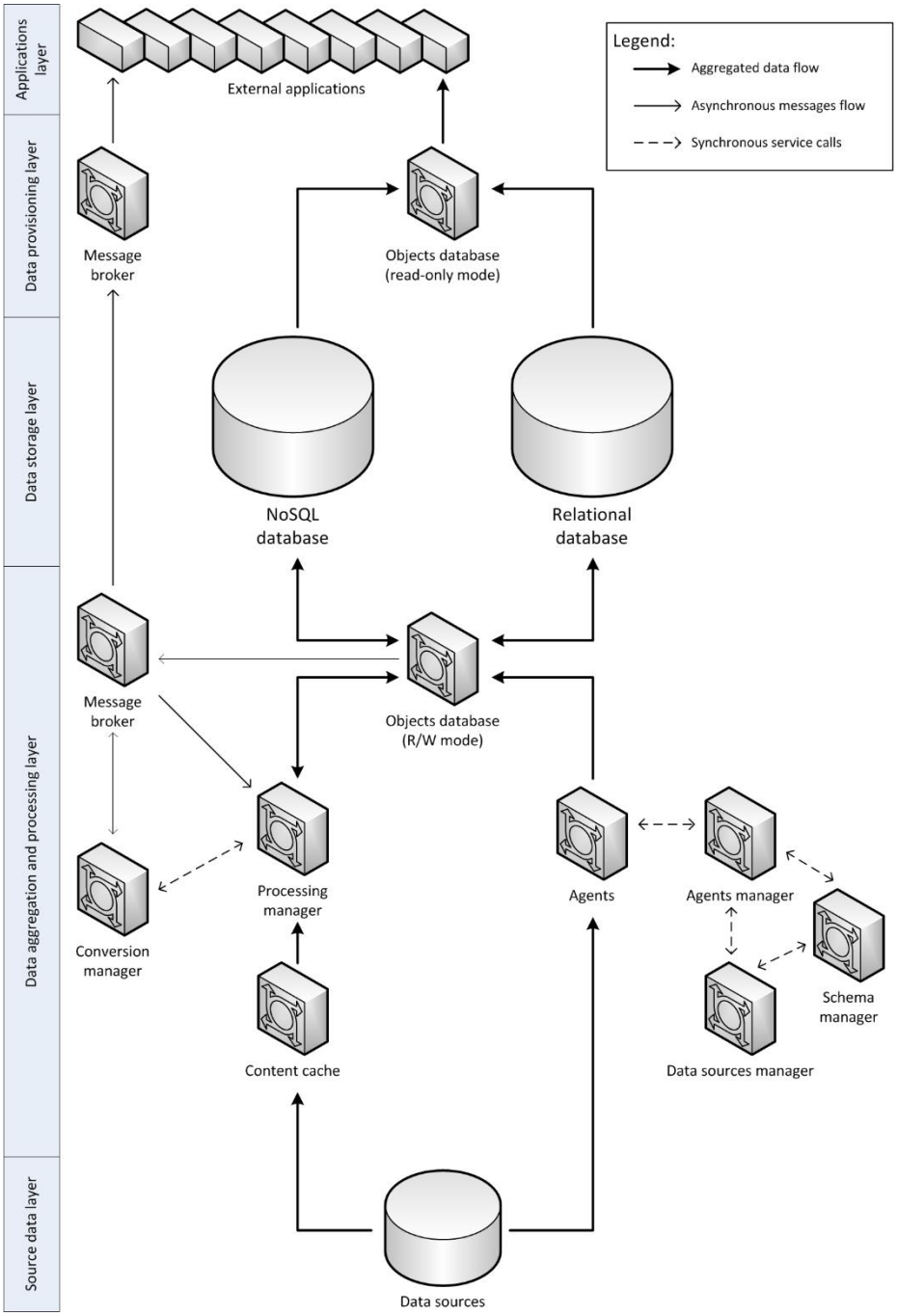

Fig. 4. Architecture of the Clepsydra system used in the Polish Federation of Digital Libraries.

The aggregation and processing of that data are happening in a cloud-based system called Europeana Cloud, that is a shared development of the Europeana Foundation and PSNC. This cloud infrastructure consists of the following components [12]:

- Metadata and Content Service (MCS) – used to provide authorized R/W access to all data stored in the Europeana Cloud platform, based on NoSQL database and object storage.
- Unique Identifier Service (UIS) – used to provide unique identifiers to all data stored in the MCS service.
- Data Processing Service (DPS) – used to process data stored in MCS in a scalable way. Based on streaming data processing framework, allowing to deploy and scale various data processing topologies.
- Image Service (IS) – used to provide HTTP-based access to high-resolution images stored in MCS services, based on IIIF image provisioning protocol [13].

Besides, there are several backend services used for internal asynchronous communication of the services and for services monitoring.

The information model used in the system is hierarchical. The basic component is a record which may have one or more representations. Each representation may have one or more versions and each version may have one or more files. A book may correspond to a record. Several representations inside that record may correspond to the digital version of that book in several formats (for example TIFF, JPEG2000 and PDF) and to the metadata of that book in several formats as well (for example in Dublin Core and EDM). Each such representation will start with one (the first) version including some files, but in case of updates in the future, new versions will be created accordingly.

Europeana Foundation, which is the operator of Europeana.eu platform, and many other institutions that are publishing cultural heritage collections online, are focusing not only on exposing as much data as possible but also on facilitating and promoting further reuse of that data in many areas. The next section describes briefly selected scenarios of such re-use.

5 Re-use of Cultural Heritage Data

For most of the people, it's obvious that cultural heritage is a key element for the identity of any society. For most of the people, cultural heritage in physical form is something that they can passively experience in dedicated spaces like museums, galleries or archives. The main reason for such situation is obviously the value of heritage objects associated with their age, history, beauty, uniqueness etc. Lack of possibility to physically interact with the heritage always was and still is one of the key means of its protection. High-quality digitisation changes that, because the uniqueness of the physical object is replaced with infinite possibilities of replication of information in digital form. Of course, the physical interaction is replaced with a virtual one, but with the current state of development of visualisation technologies, the difference between those two becomes smaller and smaller.

One of the domains for which wide access to high-quality digitised cultural heritage is a gamechanger is humanities research. Such wide digital access to information from memory institutions opens new research possibilities both in terms of the range of available historical sources and in terms of how they can be processed. New research approaches, such as "distant reading" [14] new digital research tools are created.

Example of such research tool, a virtual research environment, is Virtual Transcription Laboratory (VTL) developed by PSNC, and available at http://wlt.pcss.pl/. This service allows end-users to import a historical object from a digital library, run automated text recognition on it (with support for old prints recognition), and then perform team correction, transcription or annotation of the historical document [15]. The outcomes of such process can be exported in several formats and used further for other research activities. To achieve high interoperability with digital content platforms, this tool utilizes the services described in previous sections of the paper. When the user enters an identifier of the object to be imported, the VTL backed connects to the API of Digital Libraries Federation to locate from which service the imported object comes from. Then it connects to the API of the source digital library and imports metadata and data of the digital object. All that process is performed automatically and is a seamless integration of several services, not visible to the end-user.

6 Conclusions

In this paper an overview of a complex and large-scale cultural heritage information flow was provided, basing on real-life examples from the experience of Poznań Supercomputing and Networking Center. This flow starts on the level of a single institution, which is running a (mass) digitisation project and producing digital representations of its collections. After such digital collections are archived and made available online, they become part of a larger ecosystem which integrates regional data platforms with national and international data aggregation services. Such an ecosystem is a perfect environment for the development of research tools which are providing new possibilities in the art and humanities research.

In order to make creation and sustainability of such complex organism possible, several technical and non-technical factors are crucial. The most basic requirement is openness for cooperation between cultural and research institutions on various levels, also across national borders. Such openness can be then translated into a high degree of interoperability on the level of data models, APIs and licenses under which all that information is made available. Current stage of development of cultural heritage data platforms in Europe is a great example how widely distributed and independent organisations can establish cooperations and improve such important aspects of information societies as the access to its rich cultural heritage.

References

1. McKenna, G., & Patsatzi, E. (Eds.). (2007). *SPECTRUM: The UK museum documentation standard*. Museum Documentation Association.

2. Mazurek, C., Parkoła, T., & Werla, M.. (2012). Tools for mass digitisation and long-term preservation in cultural heritage institutions. In: *7th SEEDI, 2012: Digitisation of cultural and scientific heritage, Ljubljana (Slovenia), 17-18 May 2012* . Ljubljana, Slovenia: National and University Library.
3. Lee, C. A. (2010). Open archival information system (OAIS) reference model. *Encyclopedia of Library and Information Sciences*, 4020-4030.
4. Mazurek, Cezary ; Heliński, Marcin ; Werla, Marcin. *Distributed Digital Library Architecture for Business Solutions*. EUROSIS, 2005. ISBN 90-77381-171-1.
5. Mazurek, C., & Werla, M.. (2011). Network of Digital Libraries in Poland as a Model for National and International Cooperation. In: IATUL 2011 Conference: Libraries for An Open Environment: strategies, technologies and partnerships .
6. Lewandowska A., Werla M. – "PIONIER Network Digital Libraries Federation – Interoperability of Advanced Network Services Implemented on a Country Scale", Computational Methods in Science and Technology, Special Isuue (2010). str. 119 – 124 ISSN 1505-0602.
7. Sompel, H. V. D., Nelson, M. L., Lagoze, C., & Warner, S. (2004). Resource harvesting within the OAI-PMH framework. *D-Lib Magazine; 2004 [10] 12*.
8. Weibel, S., Kunze, J., Lagoze, C., & Wolf, M. (1998). *Dublin core metadata for resource discovery* (No. RFC 2413).
9. Mazurek, C., Stroiński, M., Werla, M., & Węglarz, J.. (2011). Distributed Services and Metadata Flow in the Polish Federation of Digital. In: 2011 International Conference on Information Society (i-Society) (pp. 39-46). ISBN 978-0-9564263-8-3.
10. Mazurek, C., Mielnicki, M., Nowak, A., Stroiński, M., Werla, M., & Węglarz, J.. (2013). Architecture for Aggregation, Processing and Provisioning of Data from Heterogeneous Scientific Information Services. In: Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions (pp. 529-546). ISBN 978-3-642-35646-9. Springer Berlin Heidelberg.
11. Knepper, M., & Charles, V. (2016). *Making library linked data using the Europeana Data Model*. Universitätsbibliothek Johann Christian Senckenberg.
12. Werla, M., Mamakis, G., Muhr, M., Knoth, P., Mielnicki, M., & Kats, P.. (2014). Design of Europeana Cloud technical infrastructure. In: IEEE/ACM Joint Conference on Digital Libraries, JCDL 2014, London, United Kingdom, September 8-12, 2014 (pp. 491-492). ISBN 978-1-4799-5569-5. IEEE.
13. Snydman, S., Sanderson, R., & Cramer, T. (2015, May). The International Image Interoperability Framework (IIIF): A community & technology approach for web-based images. In *Archiving Conference* (Vol. 2015, No. 1, pp. 16-21). Society for Imaging Science and Technology.
14. Jänicke, S., Franzini, G., Cheema, M. F., & Scheuermann, G. (2015). On close and distant reading in digital humanities: A survey and future challenges. In *Eurographics Conference on Visualization (EuroVis)-STARs. The Eurographics Association*.
15. Dudczak, A., Kmieciak, M., Mazurek, C., Stroiński, M., Werla, M., & Węglarz, J.. (2013). Improving the Workflow for Creation of Textual Versions of Polish Historical Documents. In: *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions* (pp. 187-198). ISBN 978-3-642-35646-9. Springer Berlin Heidelberg.