D2.5 Report on Interoperability of Digitisation Platforms

Succeed

16/12/2014

Abstract

This task is about fostering the connection and technical interoperability of the interoperable service platform created during the IMPACT project (www.impact-project.eu) with other related services for digital libraries, such as for example OPF.
Moreover, a technical workshop has been held with participants from these and similar initiatives to review best practices and support the alignment of concepts and implementations.
This deliverable describes the outcomes of efforts to connect with related services and includes a set of recommendations based on the results of the workshop.

Document information

Deliverable number	D2.5	Start: M17	Due: M23	Actual: M24
Deliverable name	Report on interoperability of digitisation platforms			
Internal/External	External			
Activity type	Support			
Participant	UA, **KB**, INL, IAIS, PSNC, USAL			
Estimated person months for this deliverable	14.00			
Dissemination level[1]	PU			

Document history

Revisions

Version	Status	Author	Date	Changes
0.1	Draft	Sieta Neuerburg	15-11-2014	Created
0.2	Draft	Sieta Neuerburg Bob Boelhouwer	21-11-2014	Information added on Ageement of alliance between IMPACT and OPF and on the OAI/CMDI repository for digitisation tools
0.3	Draft	Joachim Köhler	08-12-2014	Comments and Review
0.4	Draft	Sieta Neuerburg Bob Boelhouwer Tomasz Parkola Stefan Eickeler	13-12-2014	Information added on CLARIN, dArceo, Contentus. Comments integrated.
1.0	Final	Sieta Neuerburg	16-12-2014	Final version consolidated

Approvals

Version	Date of approval	Name	Role in project	Signature
0.3	08-12-2014	Joachim Köhler	Internal Supervisor	
0.4	15-12-2014	Isabel Martínez	Technical Project Manager	

Distribution

Version	Date of sending	Name	Role in project
0.4	13-12-2014	Isabel Martínez Rafael Carrasco	Technical Project Manager Project Coordinator
1.0	17-12-2014	Cristina Maier	Project Officer

---

[1] PU Public; RP Restricted to other programme participants (including Commission Services); RE Restricted to a group specified by the consortium (including Commission Services); CO Confidential, only for members of the consortium (including the Commission Services)

Table of Contents

1. INTRODUCTION

Deliverable 2.5 is part of Succeed WP2 (Interoperability and infrastructure). The primary goal of this work package is to improve the visibility of research results and the transnational accessibility of research infrastructures by providing and maintaining a self-managed, interoperable, web-based platform with community facilities for the demonstration and testing of various digitisation software tools and technologies (see chapter 2).

Deliverable 2.5 meets the following objective in the project's Description of Work:
- O2.3 Contribute to the technical interoperability with other platforms and infrastructures.

In particular, D2.5 is an output of the following tasks in WP2:
- T2.3 Interoperability.
  The objective of this task is to foster the connection and technical interoperability of the interoperable service platform with other related services for digital libraries, such as for example those of Contentus, Clarin, OPF and dArceo (see chapter 3).

Furthermore, technical interoperability with related digitisation software platforms will be studied in order to support the adoption of common interfaces and open standards.

To achieve these goals, a technical workshop has been held with participants from these and similar initiatives to review best practices and support the alignment of concepts and implementations (see chapter 4).

2. SUCCEED INTEROPERABILITY PLATFORM

The Succeed Interoperability platform is a software suite which was first created by the EU funded IMPACT project (2008 – 2012). The platform provides online access to a set of tools for digitisation which would otherwise have to be installed and used independently. As such, the operational platform supports the interoperability of tools created by different institutions and simplifies the testing of these tools, working alone or in combination with others. The tools have been made available as web services by a number of different providers, which allows the user to try them out online without having to install any of these tools locally. The platform helps raise awareness of the methods and tools which are available for every digitisation step.

The platform also offers access to open datasets created by the IMPACT project, including 250,000 high-resolution images and 30,000 ground truth files. The service uses the protocol (API) provided by the University of Salford, which provides users with direct to access the IMPACT repository. An easy to use interface has been integrated in the repository management system, enabling system administrators to easily tag any image with a set of over 80 keywords, and thus allowing the enrichment of the current dataset at any time.

The platform is accessible through the following link:
http://succeed-project.eu/demonstrator-platform. For additional details on the architecture, deployment and licensing of the platform, please refer to D2.4 (published at http://www.succeed-project.eu/publications).

3. OVERVIEW OF CONNECTIONS ESTABLISHED WITH RELATED SERVICES FOR DIGITAL LIBRARIES

3.1 Agreement of alliance between IMPACT and OPF

One of the main objectives of Succeed was to foster the transfer of knowledge, by sharing best practices and setting up Centres of Competence for European innovation and research. The IMPACT Centre of Competence in Digitisation (http://www.digitisation.eu), an output of the IMPACT FP7-project, will sustain the actions beyond Succeed. In November 2014, IMPACT has signed an Agreement of Alliance with the Open Preservation Foundation (OPF) to strengthen their collaboration and enhance their respective membership benefits. The Open Preservation Foundation sustains technology and knowledge for the long-term management of digital cultural heritage.

The Agreement commits the two organisations to sharing resources and delivering at least one event each year to bring together expertise in digital preservation and text digitisation. The Benefits of the Agreement are:

- Access to each other's members-only web pages and discussion lists
- Attendance at each other's events for free or at member rates
- Participation in each other's interest or working groups
- Observer status at the Annual General Meeting and Planning Days

The alliance between IMPACT and OPF will further enhance collaboration and interoperability between initiatives and platforms in the field of digitisation.

3.2 CLARIN

CLARIN is one of the research infrastructures that were selected for the European Research Infrastructures Roadmap by ESFRI, the European Strategy Forum on Research Infrastructures. It is a distributed data infrastructure, with sites all over Europe, such as universities, research institutions, libraries and public archives. They all provide access to digital language data collections, to digital tools to work with them, and to expertise for researchers to work with them. The CLARIN Governance and Coordination body at the European level is CLARIN ERIC. CLARIN currently has nine full members, but more countries are expected to join in the near future.[2]

CLARIN aims to provide easy and sustainable access for scholars in the humanities and social sciences to digital language data (in written, spoken, video or multimodal form), and advanced tools to discover, explore, exploit, annotate, analyse or combine them, wherever they are located. CLARIN is building a networked federation of European data

---

[2] Information obtained from: http://clarin.eu/content/general-information

repositories, service centres and centres of expertise, with single sign-on access for all members of the academic community in all participating countries. Tools and data from different centres will be interoperable, so that data collections can be combined and tools from different sources can be chained to perform complex operations to support researchers in their work.[3]

The European data repositories provide metadata describing the items in the respective data collection. The data from each participating repository is regularly harvested by the central search portal. This centralized database with metadata can be queried by parties wishing to discover a particular dataset or a specific tool.

Succeed has made the list of tools (please refer to Succeed Deliverable D3.2[4]) available as metadata (OAI). This list will also become available in CLARIN's CMDI format. This will allow parties using the CLARIN portal to find tools from the Succeed tool list in the CLARIN portal. In the portal, users can find out about the tool's properties and where it can be obtained.

3.3 Contentus

The research and development project CONTENTUS (2007-2012 – THESEUS research program) was set up to explore ways of making Germany's digital cultural heritage accessible to as many people as possible and of preserving it for use by future generations.

During the project, technologies have been developed for the entire process chain - from digitisation to presentation of the cultural objects in a multimedia search environment. The work concentrated on automated quality control, cataloguing of digitised media, interlinking of the media objects (both with each other and with external data sources) and the innovative presentation of this vast range of multimedia data in a semantic search interface.[5]

In the first year of the Succeed project, a first connection to the Contentus platform was made by partner IAIS. However, since the Contentus project was finished in 2012, there were some problems with the maintenance of the Contentus platform. To improve the level of service for the user and reduce the processing delay, the Succeed project team transferred the functions of the Contentus platform to the Succeed platform.

One of the lessons learned from this experience is that the coupling of small platforms is

---

[3] Information obtained from: http://www.clarin.eu/guest-portal

[4] Available at: http://www.succeed-project.eu/sites/default/files/deliverables/Succeed_600555_WP3_D3.2_FinalReportOnAvailableTools_v1.0_D.pdf

[5] Information obtained from: http://www.dnb.de/EN/Wir/Projekte/Abgeschlossen/contentus.html

problematic because each platform will act as a single point of failure. It would be good to strive for bigger platforms in the future and make sure that they are maintained by a consortium. These platforms can then be loosely coupled to other big platforms with good maintenance.

3.4 dArceo

dArceo is a system for long-term preservation of source data (e.g. master files), primarily focused on textual, graphical and audiovisual content. It makes migration of source data possible with respect to the OAIS model. Additionally, dArceo provides conversion and source data delivery functions, which may help both to build digital libraries and access source data by advanced users, e.g. humanists. dArceo can be configured to store data in the PLATON-U4 archiving services which has been developed in scope of the PLATON project and deployed in the Polish nationwide network PIONIER.[6]

dArceo is able to execute various migration and conversion services which are described with WSDL/OWL-S. dArceo uses semantic databases to determine migration or conversion paths for specific objects. For ingest and access, dArceo uses asynchronous REST interfaces and the OAI-PMH protocol for metadata harvesting.

Through the Succeed Interoperability platform, users may execute migration or conversion services that are available in dArceo. Moreover, users could make use of dArceo's REST interface to download a digital object from the dArceo preservation system to the Interoperability platform and execute OCR on it. If a selected object is already composed of images and GT, the platform could be used to perform an evaluation of specific OCR engines.

---

[6] Information obtained from: http://dingo.psnc.pl/darceo/

4. TECHNICAL WORKSHOP ON THE INTEROPERABILITY OF DIGITISATION PLATFORMS

On 2 October, 2014 the Succeed project organised a technical workshop on the interoperability of digitisation platforms at the National library of the Netherlands (http://www.kb.nl) in The Hague. Nineteen researchers, librarians, and computer scientists from several European countries participated in the workshop (see Annex I List of participants). The workshop's programme included a number of presentations and several interactive sessions focusing on the different dimensions and the benefits of interoperability, the main barriers to achieving interoperability, and the potential role of Centres of Competence in leveling these barriers. The current chapter contains the main results of the workshop and a set of ensuing recommendations for the future.[7]

4.1 Organisational aspects of interoperability

Interoperability is often defined as a technical (IT related) problem. However, the contributions of the workshop's participants made clear that there are many other aspects of interoperability that need to be taken into account, including:

1. Technical aspects
2. Organisational aspects
3. Legal (IPR) aspects
4. Financial aspects

The output of the interactive sessions made clear that organizational aspects are of the greatest importance. Despite the apparent predominance of technical issues and solutions, it is important to keep in mind that technical problems can only be solved if the respective users, programmers, institutions, and companies are willing to solve them.

In other words, one of the key conclusions from the workshop is that organisational aspects are central to achieve (technical) interoperability between different digitization platforms, tools and systems (see sections 4.3 and 4.4).

4.2 Benefits

During the workshop, we asked our participants why they, their institutions or companies, were actually interested in achieving interoperability.

It was agreed by all that interoperability through cooperation with partner institutions brings many benefits to the institutions themselves. Most importantly, this type of cooperation **increases synergy**, which allows all participating institutions to **allocate**

---

[7] A blogpost about the workshop can be found here: http://www.digitisation.eu/blog/succeed-interoperabily-workshop-report/#more-2340. Pictures taken during the workshop can be found here: https://www.flickr.com/photos/impactcoc/sets/72157648312310807/

their resources more efficiently and effectively. Moreover, cooperation helps to **reduce
costs for software maintenance**.

Additionally, the results of interoperability also offer an important benefit for end users: it
helps to **improve data accessibility and usability**.

4.3 Barriers

After establishing the benefits of interoperability, we asked the participants to identify
the barriers which – in their view – stop us from actually implementing it.

The following four issues were selected as being the most important barriers to achieving
interoperability:

- Conflicts of interest between stakeholders; i.e.
  - Competing standards.
  - Commercial interests go against interoperability.
  - Issues related to Intellectual Property Rights.
- Maintenance
  Many digitization platforms and tools are not maintained after a research project
  ends.
- Funding of projects
  Until recently, the EU funding system for digitisation projects supported
  interoperability *within* rather than *between* projects.
- Insufficient expertise by users
  There seems to be a deficiency in "21st century skills" among researchers, which
  leads to a lack of understanding between technical experts on the one hand, and
  Social Sciences and Humanities (SSH) researchers on the other.

4.4 Recommendations for CoCs

Having identified the most important issues and barriers, we concluded the workshop by
a discussion, focusing on the questions: how to formulate an agenda for the future? And
who should address these issues?

The discussion made clear that the Centres of Competence (CoCs) have a leading role to
play in leveling the aforementioned barriers and in setting the agenda for the future.
The results of the discussion led to the following set of recommendation for the CoCs,
indicating how they could enhance interoperability.

To begin with, the CoCs should take up the role of **broker**; i.e.
- Facilitate experts in the field.
- Establish connections between technical experts and (SSH) researchers.
- Make results of interoperability efforts more visible for end users.

Secondly, the CoCs have a responsibility in organising **training** for technicians as well as researchers working with the data

And finally, performing the abovementioned roles will help the CoCs to gain authority and to increase their impact on **setting the agenda for future European funding**. It is clear that the EU has already started to focus more on concrete, maintainable and interoperable outcomes of projects, as is reflected by requirements of technology readiness. The CoCs should advance these developments by **coordinating proposal writing** and striving for **increased coherence between proposals and projects**.

5. CONCLUSIONS

The work performed in WP2 has resulted in the establishment of a number of successful connections with related services for digital libraries, including OPF, CLARIN, Contentus, and dArceo.

The interoperability workshop which was organised by WP2 was a success, in terms of active participation by the contributors, as well as in terms of the outcome of the respective sessions. The conclusions make clear why interoperability is important to institutions and what is stopping them from actually implementing it. Moreover, the outcomes of the workshop resulted in a number of recommendations with regard to the EU funded centres of competence.

This outcome is particularly relevant in relation to the Succeed project and its deliverable *D7.7 Roadmap for funding centres of competence in work programmes*. The results of the Succeed workshop on interoperability will accordingly be used as input for the roadmap.

ANNEX I LIST OF PARTICIPANTS

Name	Institution/ Project
Alastair Dunning	The European Library
Bert Lemmens	PREFORMA/ PACKED
Bob Boelhouwer	Instituut voor Nederlandse Lexicologie (INL)
Carl Wilson	Open Planets Foundation
Clara Martínez Cantón	National Distance Education University (UNED)
Enrique Molla	Universidad de Alicante (UA)
Gustavo Candela Romero	Biblioteca virtual Miguel de Cervantes
Irene Haslinger	National Library of the Netherlands (KB)
Jesse de Does	Instituut voor Nederlandse Lexicologie (INL)
Marc Kemps-Snijders	Meertens Institute – Nederlab
Marcin Pol	CLARIN PL – Language Technology Centre
Menzo Windhouwer	Data Archiving and Networked Services (DANS) – The Language Archive
Pavel Kats	Europeana
Rafael Carrasco	Universidad de Alicante (UA)
Rosette Vandenbroucke	DCH-RP/ Service Public Federal de programmation politique scientifique (BELSPO)
Sieta Neuerburg	National Library of the Netherlands (KB)
Stefan Eickeler	Fraunhofer-Gesellschaft zur Förderung der Angewandten Forschung e.V. (IAIS)
Tomasz Parkola	Poznań Supercomputing and Networking Centre (PSNC)
Xavier Espluga	Philobiblion