Roadmap for Sustainable Centres of Competence to Support Digital Libraries

Support Action Centre of Competence in Digitisation (Succeed)

Document information

Deliverable number	D7.7	Start: M13	Due: M24	Actual: M24
Deliverable name (DoW)	Roadmap for Sustainable Centres of Competence to Support Digital Libraries			
Internal/External	External			
Activity type	SUPP			
Participant	UA, **KB**, INL, PSNC, BVC, BL			
Estimated person months for this deliverable	4.00			
Dissemination level	PU (public)			

Document history

Revisions				
Version	Status	Author	Date	Changes
0.1	Sources compilation	Irene Haslinger	20/09/2014	
0.2	Initial draft	Lotte Wilms	05/12/2014	
0.3	Draft document	Alicia Blaya, Rafael C. Carrasco, Neil Sandford, Lotte Wilms	30/12/2014	

Approvals				
Version	Date of approval	Name	Role	Signature
1.0	31/12/2014	Jean Philippe Moreux (BnF)	Internal supervisor	

Succeed is supported by the European Union under FP7-ICT and coordinated by Universidad de Alicante.

Distribution (this document was sent to)			
Version	Date of sending	Name	Role in project
0.3	29.12.2014	Jean-Philippe Moreaux (BnF)	Internal supervisor
1.0	02.01.2015	European Commission	
1.0	02.01.2015	All Succeed partners (Intranet)	

About this document

This document is a public deliverable (D7.7) of the Succeed project (FP7-ICT-600555). The title has been slightly modified for enhanced consistency with the deliverable description, as contained in the DoW.

Copyright statement

[1] http://creativecommons.org/licenses/by-sa/3.0/

Contents

1 Purpose . . . 3
2 Scope . . . 3
2.1 Capacity building . . . 4
2.2 Technology . . . 4
2.3 Sustainability . . . 4
3 Methodology . . . 6
3.1 Consultation . . . 7
3.2 Desk research and iteration . . . 7
4 Initial findings . . . 7
4.1 Finding shared interests . . . 7
4.2 Common ground . . . 8
4.3 Process improvement . . . 8
4.4 Challenging organisational and technological barriers . . 8
5 Detailed review of findings . . . 9
5.1 Acquisition issues . . . 9
5.2 Ingest issues . . . 11
5.3 Access issues . . . 13
5.4 Bit-stream preservation issues . . . 14
5.5 Content preservation issues . . . 15
5.6 Generic issues . . . 15
6 The digital libraries of the future . . . 17
6.1 Technical issues . . . 18
6.2 Organisational/political issues . . . 18
7 Conclusions and recommendations . . . 18
7.1 The Pareto principle . . . 19
7.2 Investment in change . . . 20
7.3 Role of a Centre of Competence . . . 20
7.4 Role of the Private Sector . . . 22
8 Role of the European Commission . . . 23
Appendix: List of consultation events . . . 25
Workshop: From European and national projects to high-quality services and products . . . 25
Panel: Virtual Research Community for the Preservation of DCH 25
Round table: On the future of research and funding in digitisation and the possible roles of Centres of Competence . 26
Technical workshop: on the interoperability of digitisation platforms . . . 27
Panel discussion: Digitisation, conservation and preservation in digitisation . . . 27
Hackathon: Succeed First Developers' Workshop . . . 28
Hackathon: Succeed Second Developers' Workshop . . . 28

Executive Summary

This roadmap provides insight into the capacity to support the evolution of European digital libraries through centres of competence. It summarises the main technical, capacity-related and economic challenges and the actions needed to advance the state of the art towards mass digitisation of cultural heritage in Europe.

The main conclusions include the need to support smaller cultural heritage institutions —remarkably for investment in change—, the role of Centres of Competence —in three aspects: capacity building, advance of technology and sustainability— as well as the role of the private sector.

It is recommended that the Centres of Competence should be made accountable for their own budget and responsible for their own future. The Centres may receive a conditional structural funding but they should acquire the rest of their income through participating in projects, consultancy, public-private collaboration, and membership fees.

Succeed is supported by the European Union under FP7-ICT.

1 Purpose

The purpose of the Succeed Roadmap is to provide insight into the capacity to support the evolution of European digital libraries through centres of competence. This roadmap summarises progress towards mass digitisation of cultural heritage in Europe in terms of the main technical, capacity-related and economic challenges and the actions needed to advance the state of the art. This is in response to the target of getting Europe's entire cultural heritage digitised by 2025 in line with the Commission *Recommendation* of 27 October 2011 *on the digitisation and online accessibility of cultural material and digital preservation* (2011/711/EU).[2] In regards to the organisation and funding of digitisation, it calls for consideration of "ways to optimise the use of digitisation capacity and achieve economies of scale, which may imply the pooling of digitisation efforts by cultural institutions and cross-border collaboration, **building on competence centres for digitisation in Europe**".[3]

The roadmap is also aligned with the Digital Agenda for Europe, one of the flagship initiatives under Europe 2020, whose action 51[4] is intended to "reinforce the coordination and pooling of resources", especially ICT-based resources capable of delivering online access to Europe's cultural heritage.

2 Scope

The Succeed Expert Advisory Board[5] (EAB) defined the scope of a centre of competence as follows:

1. Centres of competence can coordinate and guide local, regional and national initiatives and function as an information provider on all subjects concerning digitisation (including legal problems).

2. Centres of competence can also assist in the integration of complementary technologies to improve and enhance access to the digitised content. For example, effective access to content may require image processing and OCR to be linked to machine translation technology. Bridging the gap between technologies is a key role of the centres.

3. The centres should also:

    - study user requirements (for example, an observatory could facilitate the understanding of users' needs);

[2] http://ec.europa.eu/information_society/newsroom/cf/document.cfm?action=display&doc_id=160

[3] Our emphasis.

[4] http://ec.europa.eu/digital-agenda/en/pillar-v-research-and-innovation/action-51-reinforce-coordination-and-pooling-resources

[5] http://www.succeed-project.eu/sites/default/files/deliverables/Succeed_600555_WP1_D1.1_ReportActivitiesOutputsEAB_1.0_D.pdf

- lead innovation;
- establish guidelines for practice and standards;
- create networks of experts who share their experience, thus helping others to advance faster;
- promote the transfer of knowledge.

We initially considered this remit from three inter-connected perspectives: the capacity-building that a centre of competence can provide, the role of technology and the steps towards economic viability, or sustainability, that are necessary for survival of the network of centres. The relationships between the three factors is shown in the conceptual model of Figure 1.

All three aspects of the work of a competence centre provide support to cultural institutions in order to keep pace with spiralling demand —increases in the volume of content that is to be digitised, increases in the quality of that content and increases in the sophistication of the access tools.

2.1 Capacity building

Centres of Competence can provide access to specialist skills, resources and know-how so that cultural memory institutions do not need to maintain permanent in-house facilities and specialist staff. A centre can develop and deliver training materials and events for people already working in memory institutions or planning to do so. They can advise both private and public sector memory institutions on a range of topics in their respective fields of expertise. Importantly, this creates additional revenue streams from commercial activities and provides flexibility in the kind of business models that can underpin the network of centres.

2.2 Technology

Centres of Competence can assess the capabilities of existing and emerging technological tools and standards, provide access to the most suitable tools for specific workflows and specific domains and provide training and consultancy. These commercial and quasi-commercial activities that provide cultural institutions with additional capacity give the centres revenue opportunities. Centres of competence can also act as catalysts that identify gaps in the market for software developers or act as developers —perhaps in conjunction with commercial partners such as independent software vendors— responding to unmet needs across a range of workflows such as scanning or OCR.

2.3 Sustainability

Centres of competence have a major role over the next decade in making the digitisation, access and preservation of cultural heritage more cost-effective,

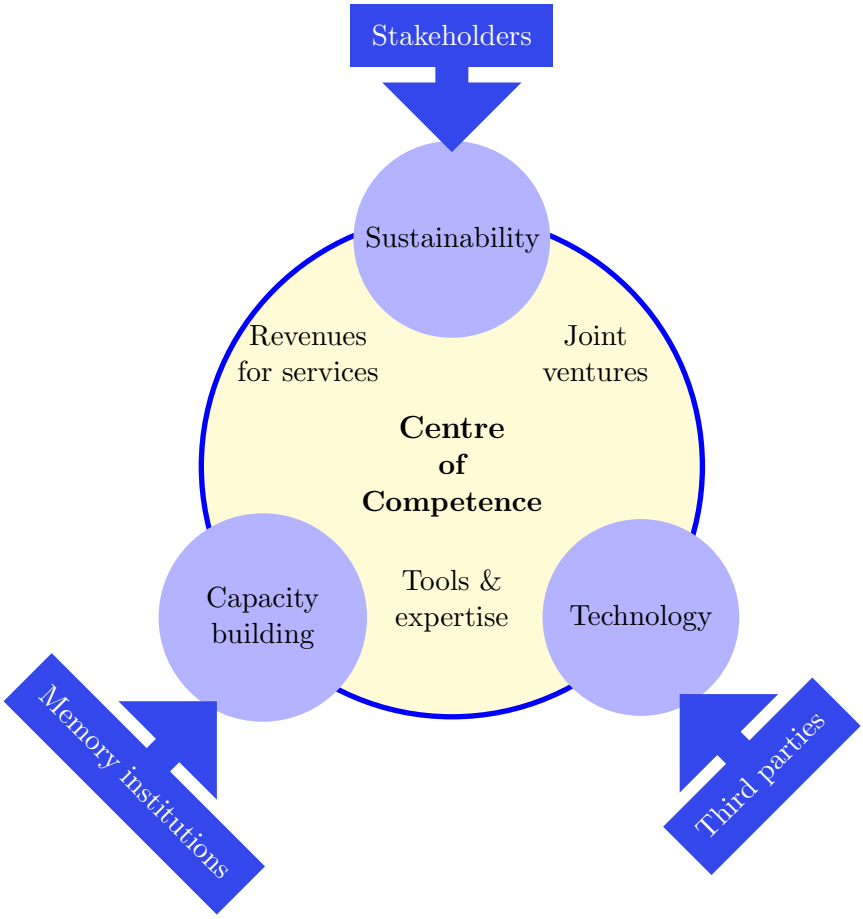

Figure 1: Conceptual model of a Centre of Competence.

as well as in facilitating the access to the digitised content. Economies of scale and improved workflow efficiency —including reductions in errors and wasted work— directly drive down costs. Re-use of objects once they have been made digital provides additional benefit by opening up new revenue streams. However, the Expert Advisory Board cautioned that the transition to self-sustaining operation had not yet completed and that for the foreseeable future the centres would have to operate a *mixed economy* with some direct or indirect state funding. This could take the form of partial reimbursement of digitisation costs (following the cost model produced by 4C) in the same way as Open Access publishing costs are now treated as eligible costs for Horizon 2020 projects.[6]

[6] `http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf`

Table 1: Organisations consulted

Organisation	Workflows	Formats
Open Preservation Foundation[8]	Preservation	Mainly text
IMPACT[9]	Digitisation	Text
VCC-3D[10]	All	3D
V-MusT.net[11]	Digitisation	Access (mainly 3D)
Presto Centre[12]	All	Audio-visual
APARSEN[13]	Preservation	All
4C[14]	Preservation	All
DCH-RP[15]	Preservation	All

3 Methodology

There are many published models[7] of the digital curation lifecycle of which the Digital Curation Centre (DCC) model is perhaps the most widely cited. The DCC model differentiates between curation (responsibility for the content of a digital object) and preservation (responsibility for the form of the content, its representation). This separation can also be found in the remits of the centres of competence that are the focus of this report.

Our approach to the construction of a generic roadmap for centres of competence has been based on a combination of desk research and consultation with key individuals from existing service providers including *networks* and *centres of excellence* as well as competence centres.

European centres of competence in digitisation and digital preservation can be expected to address different workflows in the lifecycle of digitised cultural heritage. They may also specialise in specific formats or media types. We consulted eight EU-funded projects supporting the workflows of digitisation, access and preservation and these are characterised in table 1. DCH-RP and C4 were included in the final phase of preparation of this roadmap. We did not enter into dialogue with the users of those services.

The goal of the consultation was to identify opportunities for convergence between those various stakeholders in order to make it easier for them to reach sustainability. This, in part, meant aligning the various areas addressed by

[7] http://wgiss.ceos.org/dsig/whitepapers/Data%20Lifecycle%20Models%20and%20Concepts%20v8.docx
[8] http://openpreservation.org
[9] http://www.impact-project.eu
[10] http://www.vcc-3d.com
[11] http://v-must.net
[12] https://www.prestocentre.org
[13] http://www.alliancepermanentaccess.org/index.php/aparsen
[14] http://4cproject.eu
[15] http://www.dch-rp.eu

the roadmap to the practicalities of operating a support service in order to establish a common vision across the spectrum. In practice, convergence was hard to find and this was reported in Succeed deliverable D7.6.[16] The implications are discussed below. The situation is similar to that expressed in a report[17] from the DCH-RP project (not a participant in our earlier consultations): "There is not an existing roadmap that the DCH-RP project could build on or progress further. The project has to develop its own roadmap for the specific domain and task that it is addressing."

3.1 Consultation

The Succeed Roadmap is based on input from a variety of sources, including events either organised or participated in by the Succeed project. The key consultation events are listed in the appendix. The project also organised two hackathon events aimed at understanding the way tools can be validated.

3.2 Desk research and iteration

The first phase of desk research provided material for an initial consultation with stakeholders and lessons learned from that consultation were fed back into deliverable D7.6, concerning the sustainability of a centre of competence. There were fundamental differences between the various consortia and key barriers to convergence including the differences between the legal forms of each trading entity. Towards the end of the project, roadmaps appeared from the 4C and DCH-RP projects. These broadened out the debate in a number of areas, including the 4C recommendation that funding for digitisation should be subject to a business case justification.

4 Initial findings

At the end of the second consultation period, a number of recommendations were derived from the feedback received, focusing on collaboration to tackle shared barriers. These recommendations can be summarised as follows:

4.1 Finding shared interests

Since existing centres of competence address a variety of market segments, are at various levels of maturity, and have access to a diverse set of capabilities it is not surprising that they do not share a single business-model. However,

[16] http://www.succeed-project.eu/sites/default/files/deliverables/Succeed_600555_WP7_D7.6_BestPracticesAndManagementProceduresForCentresOfCompetence_v1.0.pdf

[17] http://www.dch-rp.eu/getFile.php?id=404

they do share an interest in the systematic removal of barriers to sustainability such as lack of appropriate technical standards. For that reason, we recommend establishing an ongoing consultative framework addressing those barriers across the entire range of centres.

The European Commission (EC) should invest in finding a transnational solution for aligning the legal status of the centres of competence to allow them to collaborate and share opportunities and risks.

4.2 Common ground

The main differences between the various existing competence centres are found in the primary tasks and domains being serviced, so the likely areas where common ground for pilot collaborations have been identified in generic issues such as protection of intellectual property (IP) and metadata standards.

In particular, it is in the best interests of the operators of competence centres and their clients that there is a formal recognition of the competence of staff and the quality of services. This would require investment for collaborative development of certification protocols and a coherent policy for training.

4.3 Process improvement

Given that process improvement within memory institutions is the most likely way of increasing their capacity and throughput of their digitisation programmes to meet increased demand (such as the 2025 target), there is concern in the current climate of austerity that institutions that have been successful in cutting unit costs will see their budgets reduced rather than having their throughput increased.

Clarification is required on the future commitment of resources from the EC budgets associated with the Digital Agenda.

4.4 Challenging organisational and technological barriers

There is a role for competence centres to provide compelling arguments for adherence to standards and best practices, including dissemination of success stories. The following barriers (identified initially during an interoperability workshop) need to be addressed:

- Staff with responsibilities for specifying, selecting, configuring and maintaining complex systems receive poor support from vendors such as inadequate documentation. There is a lack of commitment to capacity-building. These are symptoms of an immature market.
- Adoption of systems that are rapidly overtaken by the pace of technology evolution leading to dependence on archaic legacy systems and outmoded standards. These are symptoms of a technology-driven supply chain.

- Failure to develop 21st century skills and long-term thinking, lack of openness to sharing knowledge/innovation and decision-making that focuses on internal issues rather than partnership and perpetuates the "not invented here" syndrome. These are symptoms of an organisation driven by fear rather than trust.

- Reliance on output-based funding models rather than outcome-based investment models and a lack of models and tools for decision-making based on return on investment calculations. These are symptoms of an unsustainable business model.

5 Detailed review of findings

We present our detailed findings using a comprehensive model of curation and preservation workflows from the LIFE project[18] which provides a framework of workflows for five high level phases of the digital object lifecycle: acquisition, ingest, bit-preservation, content-preservation and access. These phases are recognisable in other roadmaps such as DCH-RP. We use this to identify and analyse the needs of the memory institutions. The workflows within this framework are generic and consistent regardless of media type although digitisation processes and tools will vary. We do not explicitly address each of these workflows, as some (such as deposit and refreshment), expose only generic concerns or issues that are well-known from prior work in earlier projects such as Planets[19], Impact[20] and Scape[21].

Digitisation is not explicitly referenced in the LIFE model, but references to digitisation can be found elsewhere as a separate workflow or as part of the acquisition (obtaining) phase. We adopt the approach used by Collections Trust for their 2010 report to the Comité des Sages[22] which differentiates between two workflows, both in the acquisition phase: scanning/photography of material for preservation, and the creation of surrogates (such as PDF and OCR files) for access. In this way we can consider the needs that can be addressed by a network of centres across the entire digital cultural object's lifecycle.

5.1 Acquisition issues

These include selection, submission agreement, IPR and licensing, ordering/invoicing, scanning/photography, creation of surrogates and check-in.

[18] http://www.life.ac.uk/3/docs/ipres2009v24.pdf
[19] http://www.planets-project.eu
[20] http://www.impact-project.eu
[21] http://www.scape-project.eu
[22] The Cost of Digitising Europe's Cultural Heritage, http://nickpoole.org.uk/wp-content/uploads/2011/12/digiti_report.pdf

5.1.1 IPR and licensing

Control of rights over intellectual property is essential but contentious and poorly-understood. Barriers include territorial differences, inconsistent and incomplete protection and lack of enforcement. Rights can sometimes be acquired quite simply, at other times a detailed and expensive investigation of rights ownership is necessary to simply establish whether a book is still protected by copyright, for example.

Current trends include public-private partnerships where a commercial licensee pays for the digitisation in exchange for exploitation rights, the content being locked behind a paywall for the duration of the licence agreement. Even though the license allows the original content-holder to access the digitised content, perhaps for research purposes, publication may be restricted.

The emergence of the Creative Commons licensing schemes provides a more flexible solution to exploitation reflecting the *mash-up* culture of the internet and its acceptance of derived works. One important barrier is that many collections are made available for digitisation and access under license agreements that pre-date the world-wide web.

It is a misconception to assume that such issues can be resolved by *market forces* alone.

5.1.2 Scanning and photography

At the Succeed final conference in November 2014, Daniel Pletinckx of the `vMusT.net` network described the challenge of introducing new technology into a memory institution, especially the type of visual/3D technologies available to museums. In a survey, 80% of museums declared themselves ready to take investment decisions but the reality is that they have limited capacity to undertake the development and support. He also pointed out that museums tend to use established, mature, technology. There is a clear role for specialist external consultants to help organisations establish a business case for investment in such technologies and to support development.

5.1.3 Creation of surrogates

One of the concerns that is being addressed by the IMPACT Centre of Competence is that memory institutions wishing to enhance the value of their digitised collections are reliant on Optical Character Recognition (OCR) technologies to extract the meaning of text from the image. Existing OCR engines (such as Abbyy's FineReader, Omnipage, Readiris or Tesseract) provide accurate transcription of images containing modern text, not just as individual characters but as words. A 99.95% character recognition rate will result in one error about every 300 words or about one error every 20 sentences.

Lower accuracies are reported when those algorithms are applied to historical documents, such as those printed before the 20th century using an ornate Fraktur type-face.

A recogniser such as the one reported by Furrer and Volk,[23] with 99.5% success in recognising individual characters will statistically produce an error approximately every thirty words (assuming an average German word length of 6.26 characters), which is about once in every two sentences.

This may be acceptable for information retrieval purposes, where a non-perfect transcription can still be useful to identify the target of a search but is of questionable value where precise understanding of the text is required. Shift the performance by another order of magnitude, to 95% character recognition, and nearly four out of every five sentences will contain an error. Making a mistake in setting up or measuring a performance test or benchmark will potentially result in a significant distortion of the performance of a recogniser.

There is a clear role for a centre that can provide a well-informed and independent evaluation service that can establish the performance requirements of tools to address particular classes of problem. It can then compare the ability of specific tools to meet those requirements. If the evaluation process identifies gaps in the market, then tool-providers can accurately assess the scale of the market opportunity.

5.2 Ingest issues

These include quality assurance, metadata, deposit, holdings update and reference linking.

5.2.1 Quality assurance

OCR is not the only technology that is used in the acquisition and ingest phases of the digital object's lifecycle and the comments above regarding evaluation of the performance of OCR tools will apply equally to measurement of the fidelity of scanned images, automated extraction of metadata and so on. All of these aspects interact in complex ways and there is, as yet, no common approach shared across institution-types, domains, platforms or audience types although the 4C project is opening up the debate about how the cost of digital curation can be assessed more objectively and what that cost includes.

This makes evaluation of performance and delivery of high-level quality goals even more difficult. For example, the target audience for an iconographic collection such as the Wellcome Institute's History of Medicine is more concerned with the subject matter of the images than the aesthetics. This may lead to two different cost-benefit equations —one addressing the access workflow(s) and the other addressing long-term preservation and permanent access.

The quality parameters for an OCR digitisation project might include:

[23] http://www.aclweb.org/anthology/W11-4115

- Page level analysis
  - Image quality characteristics based on evaluation of cropping, skew, contrast, warping, noise, etc.
  - Text quality characteristics derived from analysis based on dictionaries, *named entities*, statistical analysis, etc.
- Document level analysis
  - Layout analysis characteristics, for example, checking if there are unexpected layouts within the document.
  - Inter-image analysis, for example, detection of duplicated images, ordering checks and keyword extraction for topic checks.
- Collection level analysis
  - Inter-document analysis, for example, duplicates detection.

5.2.2 Metadata

One of the crucial aspects of the ingest phase is the enrichment of the digitised object with descriptive and structural information (metadata) that enables various types of access to the content. This includes representation information that allows the bits and bytes generated by digitisation to be interpreted and the content reproduced. It also includes statements about the rights and restrictions associated with the object and its representations.

The third kind of metadata consists of descriptions of the object itself rather than its representation —transcription of the text contained in an image or a segment of an image (such as newspaper headlines, columns and pictures), descriptions (*tags*) of the subject matter, links between two objects. Increasingly sophisticated automatic techniques are being used to extract more and more meaning out of the object. This leads to five requirements where memory institutions would benefit from a more coherent approach:

1. In terms of capacity, automated enrichment may lead to more rather than less manual intervention —such as crowdsourcing based on validation and corrections made by user-communities—, resulting in validation of those corrections as part of the curatorial process.
2. While metadata standards exist (such as METS and ALTO), metadata produced outside of a framework of standards needs to be supported, just like any other legacy system. This requires both syntactical and semantic compatibility (speaking the same 'language' and assigning the same meaning to the words), adding access to domain-specific ontologies to the framework of standards.

3. As the quality of enrichment improves, so the number and variety of stakeholders with an interest in the content increases. Alignment must be achieved between collection-holders, their infrastructure providers, their domain experts, third-party researchers able to interpret and expand the scope of metadata, third-party tool-providers and open-source developer communities (such as the IMPACT hackathon participants, see the appendix) enabling more cost-effective metadata extraction.

4. The emergence of the *Linked Open Data* paradigm ("link your data to other data") and the Five Star Scheme[24] encourages publication of structured data on the web under open licenses without proprietary formats.

5. Enrichment ultimately leads to individualised access experiences for the various users and a memory institution acquires a variety of user communities such as researchers in the humanities and social sciences, linguists, genealogists, cultural commentators, journalists and bloggers.

5.3 Access issues

These include access provision, access control and user support.

5.3.1 Access provision

Access to digital cultural heritage objects requires a discovery mechanism and a presentation mechanism. The discovery mechanism may be based on a search facility or a browse facility. That is, an object may be retrieved through understanding of its content or context. The presentation mechanism focuses mainly on content. Language is of fundamental importance in search (being able to express what you are looking for) and presentation (being able to understand what you retrieve). Language-content may be automatically extracted and translated during the ingest process, generating searchable metadata from thesauri (controlled terms), or left in its original language which requires more challenging natural language processing within the search function. The emergence of open linked data provides the basis for semantic search for non-textual media based on the existence of searchable linked data —a hypermedia facility described in 1990 by Mark Dunlop.[25]

For a memory institution, this raises several problems. The ability to resolve user needs is dependent on natural-language processing and machine translation, both of which are far from perfect. Ingest and access workflows are often implemented using tightly-coupled front-end/back-end commercial systems which are not designed to be adapted using third-party tool modules.

[24] http://www.w3.org/DesignIssues/LinkedData.html

[25] https://personal.cis.strath.ac.uk/mark.dunlop/research/publications/thesis.html

Textual content most often comes from OCR used on scanned or photographic images. The shortcomings of OCR, especially on historic text, typewritten and handwritten material and objects with unstructured layout like newspapers have already been described. An additional challenge is provided by multimedia collections where the equivalent of OCR is speech-to-text, another immature technology.

Flexible and extensible platforms and new services are needed to allow easy access, download and analysis of digitised/digital content and links to other digital collections and data. This can be achieved by developing and implementing technology infrastructure and environments which support open data. These digital environments need to support the application of new open tools and software to allow enhancement, augmenting and interrogation of the data. There is a clear demand for this by researchers who apply computational methods of analysis to data, such as location based search, language processing, text mining, image analysis and visualisations. Providing *Application Programming Interfaces* (APIs) to data can be very powerful ways to access data-sets, and can be used by software developers to build software applications on top of them. The *API economy* is a growing area and is stimulating the development of lots of applications for mobile phones and can thus support access to digitised content on handheld devices.

5.3.2 Access control

Access control can be applied at several levels of granularity, both in terms of who can access material and what level of detail can be presented. Current information-retrieval systems can impose constraints and levels of privilege on individual user accounts and differentiate between in-house and external users. Access to content can be controlled in terms of the level of detail available (e.g. low-high image resolution, abstracts or full-text), the number of items that can be retrieved, an expiry date or the status of the user.

In spite of the emergence of an authentication and authorisation infrastructure (AAI), current platforms/interfaces do not do a good job of allowing access to search and retrieval facilities across multiple collections, especially multi-format collections, and do not support user-defined workflows.

5.4 Bit-stream preservation issues

These include repository administartion, storage provision, refreshment, backup and inspection.

Bit-stream preservation is essentially about storage, where third-party data-centres provide an alternative to in-house storage. The role of a centre of competence is not to compete with these data-centres, although we envisage that a data-centre could function within the competence centre model where the focus is on an advisory and validation role. The APARSEN and DCH-RP

projects provide detailed discussion of bit-stream preservation practices based on principles enshrined in the OAIS[26] model.

5.5 Content preservation issues

These include preservation watch, preservation planning, preservation action, re-ingest and disposal.

Preservation of content is also addressed by the OAIS model and comes into the scope of OPF, APARSEN and DCH-RP. Here, we address the generic issues arising from use of in-house and third-party tools.

5.6 Generic issues

5.6.1 The innovation gap

There is a tension between providers of end-to-end digitisation and preservation platforms, developers of specialist tools, both in-house and out-house, and the memory institutions that provide a marketplace for those technology-suppliers. It is frequently claimed that the *GLAMs* marketplace (galleries, libraries, archives and museums) is *too small* to support development of step-changes in technology, yet Nick Poole estimated a total cost of digitising cultural material in the EU of over 105bn euro for the Comité des Sages report referred to earlier. This would seem to be an adequate commercial proposition for systems vendors, especially since the trend is for the institutions to initiate projects that provide the budget for the private sector to develop solutions that they then license back to the institutions, (as described by Ed Fay of the Open Preservation Foundation, at the Succeed final conference).

One of the undoubted problems with project-based tool development is that it is not the primary expected outcome of the project and the goal is often to do "just enough" development work to complete the project, using beta-tested versions of the tools. The result is a catalogue of potential tools that are hard to sell, hard to maintain and hard to support with little or no potential for re-use. This is particularly true when the intended market is constrained by the small number of potential users and barriers to uptake such as the weakness of a cost-benefit equation (for example, for historic texts).

Yet there are tools available: DCH-RP has catalogued around 140 preservation tools and the IMPACT Centre of Competence can advise memory institutions that want to set up a digitisation workflow for text material based on knowledge of some 250 tools. The value of such a catalogue can be enhanced if the tools have been independently evaluated by collection holders and the Succeed project has produced a set of guidelines and training to support tool-use and evaluation coupled with recommended practices for identifying appropriate standards (formats and licensing schemes, for example). Competence

[26] http://www.ccsds.org

centres that include user organisations are well-placed to continue that effort. However, grass-roots entities such as centres of competence cannot be expected to provide answers to all the problems. Take, for example, the case of CLARIN centres which have been asked to host tools from other partners at their own expense.

Succeed has identified two specific problems that need to be addressed if the results of research are to be taken into the mainstream. The first is the lack of co-operation between the user community and the tool creator in defining requirements. This typically manifests itself in the following:

- Tools that are not built to fit into the target library workflow or not tested in all relevant workflows.
- Tools that are for a niche market and require specialist skill to install and use them.
- Users that do not have access to the necessary technical know-how.
- Lack of support for those users —especially in reliance on inappropriate documentation and training. Too little information to get the tool operational or too much to be digested.

The second issue is that the development teams producing tools for collaborative projects do not have relevant experience of issues such as configuration management, making the deployment of the tools more difficult for the target users. When experienced developers are assigned to the project, this is often on a part-time basis and results in lack of day-to-day supervision.

5.6.2 Interoperability

The ability to exchange information and knowledge between systems or components of a system is an inherent requirement for the advance of digital libraries. Three of the projects consulted for this report identify the need for action to improve interoperability and promote the adoption of best practices in this area:

- The DCH-RP roadmap for digital preservation associates lack of interoperability with “digital silos” that are unable to exchange content (due to legal constraints or incompatible file formats) or metadata (due to semantic incompatibility) and recognises the need for adoption of best practices that address those technical, semantic and legal issues.
- APARSEN states that this ability to “exchange and use information and knowledge between independent systems” is inherent to long term preservation of the digitised content and that content and functionality may be lost if standards that support interoperability are not followed.

- PrestoPrime, the predecessor of the PrestoCentre, discusses (in its 2010 digital audio-visual status report) the importance of exchanging metadata as the key to ensuring access to audio-visual collections, establishing interoperability among audio-visual collections, and integration of audio-visual collections with other forms of cultural heritage.

At a technological level, the scope of best practices in this area ranges from safeguarding modularity of a workflow (so that improvements to specific components do not bring with them unwanted side-effects) to bringing together different objects, different collections and different institutions. The Succeed *Workshop on Interoperability* concluded that with the emergence of the *Internet of Things*, the interconnection of devices will require greater autonomy and interoperability emphasising the need for adoption of good practices such as simplicity and clarity of design, and test early/test often approaches.

That degree of interoperability also requires organisational commitment to standards in an area where rapid evolution of technology carries the risk of obsolescence and the burden of legacy systems. This risk may be perceived as acceptable in the context of an externally-funded project but not something that should be carried into operational systems. As previously stated, competence centres can play a useful role in promoting standards, thus reducing the risk.

5.6.3 Fragmentation

One of the challenges of trying to reach the 2025 target of having the entirety of cultural heritage available online is the dependence on a wide range of actors. This report has identified a range of stakeholders including different types of memory institution and their suppliers, users and project partners. They include large and small institutions with global, national, regional and sectorial remits and specialist research libraries, research teams and commercial organisations, all with various levels of previous experience.

As a consequence, there is inadequate "glue" to bring those actors together in a coherent way and initiatives rely on relationships between individual players. Suggestions for reducing fragmentation include involving preservation experts at the beginning of the digitisation workflow, even though they may be in different departments of the organisation, promoting and supporting user groups across a number of user organisations and initiatives that bring together complementary communities (such as collection holders and research centres with domain-specific interests).

6 The digital libraries of the future

In order to overcome the issues related to the digitisation, access and preservation of cultural heritage, we first need to have a vision of what we are trying

to achieve. For this, the digitisation community has provided input for what they see as the *ideal* digital library of the future if there are no impediments. This ideal scenario can be broken down into two categories; technical and organisational/political.

6.1 Technical issues

A perfect digital library should be able to provide real time and reliable access to data providers, support collaboration and provide incentives for data publishing for researchers. It should also be concerned about (meta)data quality and standards, ideally using a standard data model, but still be flexible in terms of the types of models that users require. It should use only one exchange protocol, have an automated transformation mechanism and unlimited computing capacity. The tools that can be used with the digital libraries should be plug and play, as they should be as simple as possible. This comment extends to *crowdsourcing* technologies in particular. Finally, there should be the possibility to analyse the content by means of visualisation.

6.2 Organisational/political issues

The digital libraries of the future should be able to tap into the right skills of the right people at the right time. There should be brokers involved, both machine and human. Adding to the technical need for the digital libraries to be as simple as possible, there should be a federated identity if necessary, or at least a single sign on.

This ideal situation was painted by the current users and providers of digital libraries and much has been achieved by this community to reach this goal already. However, as we stated in this document, there are still barriers that need to be resolved. Some of these can be influenced by the various stakeholders in the community, but others cannot. The most important ones are described in the previous sections of this roadmap. To solve these issues and reach the ideal digital library, we see an important role for centres of competence in the digitisation and preservation community.

7 Conclusions and recommendations

This section, summarised in figure 2, includes recommendations made by the panel on *Digitisation, conservation and preservation in digitisation* at the Succeed final conference[27] (Hildelies Balk, Impact Centre of Competence; Ed Fay, Open Preservation Foundation; Daniel Pletinckx, vMusT.net and Stéphane Reecht, Bibliothèque nationale de France).

[27] `http://vimeo.com/114870109`

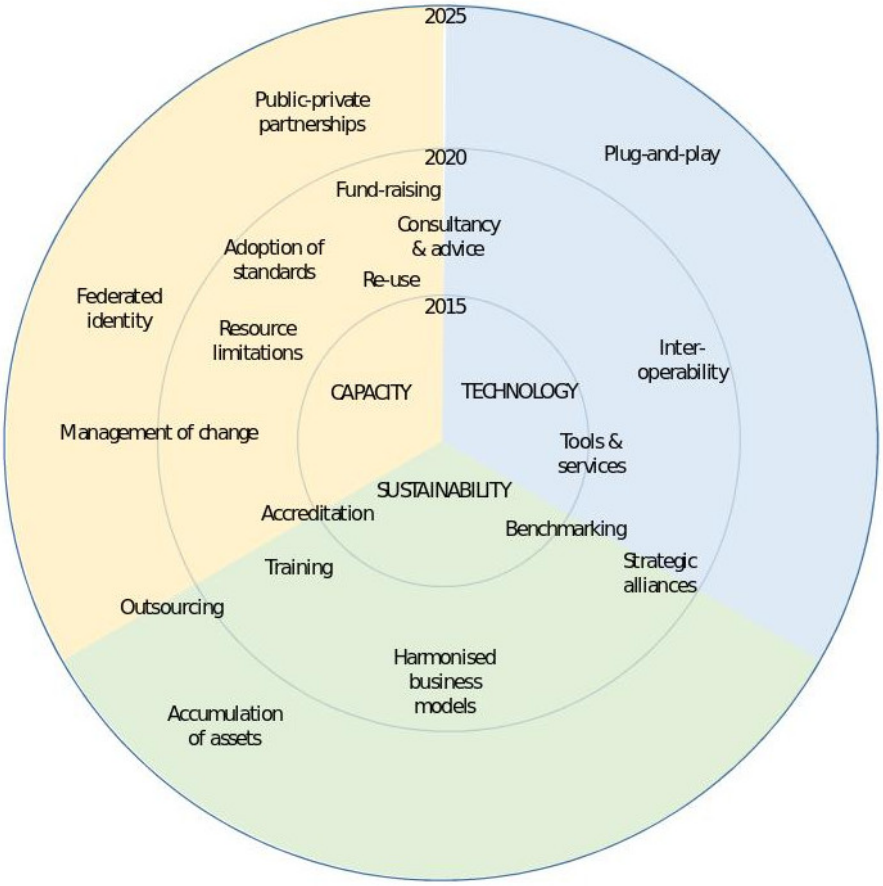

Figure 2: Succeed Roadmap (2015–2025)

7.1 The Pareto principle

We are coming to the end of a period of intensive investment in research related to digitisation, access and preservation of content, perhaps 100M–150M euro, with a substantial part of that going to libraries and other types of memory institutions. The transition from the EU 7th Framework Prgramme (FP7) to Horizon 2020 marks the beginning of the next phase, where the results of those projects are to be made available as part of the research e-infrastructures ecosystem. One of the factors that will determine the outcome is the way the stakeholders address the Pareto principle[28] which in this context suggests that 80% of memory institutions have access to just 20% of the capacity (resources). A support mechanism needs to be found to make a substantial proportion of that 80% sustainable.

[28] http://arxiv.org/PS_cache/cond-mat/pdf/0412/0412004v3.pdf

7.2 Investment in change

Large national libraries such as the British Library, Koninklijke Bibliotheek and Bibliothèque nationale de France have the research capability to implement their own systems whereas in the museum world (where 80% of institutions are ready to invest in digital technology but have little or no capacity for research and development) museum managers need support and advice in making a business case, in procuring suitable platforms and in developing suitable content. This leads them to rely on what the market offers as industry-standards, reducing the need to develop and maintain things in-house. One of the benefits of following industry standards is that digitised content can be re-used in different environments and the digitisation budget can be seen as an investment rather than a sunk cost. As well as investing in content, the smaller memory institutions will have to address the implementation of a change management programme. This may have to be achieved through structural rather than project funding.

7.3 Role of a Centre of Competence

For this section of the discussion we revert to the conceptual model introduced in Section 2 and discuss the ways in which a network of centres of competence contribute to the vision of the digital libraries of the future under the headings of capacity-building, technology and sustainability.

7.3.1 Capacity-building

Commercial and technical advice and support for change, including evaluation of the market, selection of platform and tools, training existing staff and recruiting new posts, can come from centres of competence which will become an increasingly important part of the ecosystem over the next ten to fifteen years. Among other roles, they will facilitate sustainability of the memory institutions and stewardship of content as well as supporting the discovery and adoption of appropriate tools and systems. Centres of competence from the public sector need to exploit the demand for their services from private sector archives as well as other libraries, archives and museums. The purpose of a competence centre in this sense is to ensure the sustainability of its user-community. If that is achieved, the competence centre will survive. Subsidising the operation of a competence centre does not guarantee the survival of its clients. This is a fundamental difference.

Advice on selection of digitisation, access and preservation technologies and management of change are at opposite ends of the consultancy spectrum and it is unlikely that the same individual will be able to fill both functions. It is important to see the lifecycle as a joined up whole, especially in terms of knowing the scale of operating costs *end to end*. Again, the project-funding mentality is unhelpful here as it leads to starting up of projects for which there

is no down-stream (preservation) budget. And if there is no budget then here is probably no project-plan either.

The network of existing centres has a vast knowledge base in its various fields of expertise and should be seen as the first point of access in the user community. They should contribute to the knowledge exchange within their network and pool resources with other centres when necessary to address complex problems. Two recent developments in this area have been the declaration of an Agreement of Alliance[29] between the OPF and the Impact Centre of Competence and a Memorandum of Understanding[30] between DCH-RP and the e-Infrastructures project EUDAT.

Staff in heritage institutions should provide the right academic and technical competences to further the development of the digital library. They must be open to working in teams across institutional boundaries Academics need to understand technology but they do not have to apply it. Technological experts need to understand the way scholars think and operate but they do not have to conduct the research themselves. Competence centres should take it upon themselves to spread their expertise and knowledge amongst the community they serve, in order to facilitate the education of the people working towards the digital libraries of the future.

7.3.2 Technology

Mechanisms for digitisation, access and preservation of content are usually fragmented. A support activity such as DCH-RP or Succeed is often domain and workflow-specific. For example DCH-RP is addressing preservation workflows (starting with ingest) in the specific context of digital cultural heritage. There are frequently language barriers and rhetorical clashes between domains. There is no reason to raise additional artificial barriers between digitisation and preservation functions in the way that is found in larger libraries where they are often in separate departments. These barriers collectively lead to the "not invented here" mind-set where unnecessary work results in "reinventing the wheel". The existence of a formal Centre of Competence will not prevent that mentality but it is important to stimulate open-mindedness, especially where similar problems can be shown to exist across domains and across workflows, such as quality control issues and resourcing.

The installed base of a vendor's digitisation and preservation platforms can often be counted on the fingers of one hand, yet often a significant proportion of institutions is known to be using a specific tool such as JHOVE or Tesseract-OCR, implying that the core platform is being used primarily for integration. Small, configurable tools that can be customised easily (e.g.

[29] `http://openpreservation.org/news/`the-open-preservation-foundation-and-impact-centre-of-competence-sign-agreement-of-alliance

[30] `http://www.digitalmeetsculture.net/article/dch-rp-and-eudat-a-joint-bet`

support for minority languages) and integrated into existing workflows will reduce the incidence of wheels being re-invented.

7.3.3 Sustainability

Centres of competence, properly funded, can play a crucial role in helping memory institutions achieve the critical mass necessary to deliver the digital libraries of the future as envisaged above. Taking the lifecycle as a whole, the various centres of competence should be made accountable for their own budget and responsible for their own future. A system of conditional structural funding seems promising whereby the centres receive a level of funding that enables them to run a small office and engage with potential users. They then have the responsibility to acquire the rest of their income through participating in projects, consultancy, public-private collaboration, and membership fees.

This means that they will have to identify and manage several types of funding from different sources and their legal structure must take account of this. Mechanisms such as voucher schemes can be administered through the centre as a way of incentivising memory institutions to use consultants and embark on their own digital access programmes.

We are confident that centres which deliver excellent service to a broad base of customers will survive.

7.4 Role of the Private Sector

So far, we have not explicitly defined the differences between public sector and private sector stakeholders, both of which can act as technology providers, service providers such as data centres, content-holders and content re-users. Indeed, a centre of competence may itself be privately owned or managed. There are, however, significant differences between private and public-sector cultures in terms of things like attitude to risk, timescales and investment in infrastructure and skills, just as there are differences between domains and between SMEs and large industrial organisations.

Not surprisingly, there are several ways that the private sector can contribute to the viability of centres of competence in the digitisation and preservation field, some of which are more suited to particular types of organisation more than others. Here, we characterise just four of those organisation types: suppliers of platforms/systems addressing specific workflows (such as scanning, OCR, ingest, storage), third-party providers of tools and services to those suppliers or their customers, corporate users of those platforms and tools (such as data warehouses and archives) and re-users of content in the creative sector. Depending on circumstances, they may be willing to pay or make an *in-kind* contribution to a competence centre in exchange for services such as the following:

- A platform supplier may provide a member of staff on secondment to a centre to begin the process of diffusing know-how into its customer-base
- A software developer may use a centre to specify, test or validate a new package on behalf of a user community, or to gain access to that community
- A private archive may seek consultancy support in areas where it lacks specific specialist skills in the same way as a public sector institution would.
- A re-user of content may use a centre for quality-control purposes or to validate its accounting system on behalf of the content-provider.
- A digital publishing company needs expertise on OCR.

From that short set of examples, we can see that a private sector organisation may be a customer of a competence centre, a collaborator within a centre or a business partner of a centre. It may build capacity within a centre by providing it with demonstration facilities or by asking it to participate in a development project.

8 Role of the European Commission

The European Commission has already made a substantial contribution towards the digitisation and preservation of Europe's cultural heritage. We see the EC continuing to take a leading role, certainly through the lifetime of the Digital Agenda initiative. This should include stewardship of the centres of competence to ensure harmonisation at a policy and strategic level and promotion uptake of the services of the network. By stewardship, we mean:

- Reducing the fragmentation of the digitisation and preservation ecosystem by identifying and initiating opportunities for collaboration such as:
  - Self-accreditation of the services of a centre
  - Collation and re-use of guidelines, recommended practices and inventories of software tools
- Establishing a mechanism for removing the barrier of incompatible governance models.
- Overseeing the next phase of the life of the network, including the transition to structural or infrastructure-based funding of the network and its core services.

Opportunities for acting as a catalyst for uptake of services from the network of competence centres include requiring:

- compliance with digitisation and preservation guidelines where they exist;
- evaluation and dissemination of EC-funded open-source software through the network's inventories of tools.

It is hoped that support for these activities will be envisioned in the 2016 H2020 Work Programme.

Appendix: List of consultation events

Workshop: From European and national projects to high-quality services and products

Venue: The sustainability of competence centres, Digital Heritage, 29 October 2013, Marseille, France

Organised by: Daniel Pletinckx (V-MusT.net), Halina Gottlieb (NODEM, Nordic Digital Excellence in Museums).

Participants: Daniel Pletinckx (V-MusT.net), Halina Gottlieb (NODEM, Nordic Digital Excellence in Museums), Rafael C. Carrasco (Impact Centre of Competence) and Mohamed Farouk (Center for Documentation of Cultural and Natural Heritage, CultNat)

Contents: The participants expressed their concern that the current reductions in the budget of cultural heritage institutions could lead to insufficient funds for a sustained activity at the centres. Interestingly, museums prefer lower membership fees than libraries but, in contrast, they are more open to contracting specific services (which leads to slightly different funding approaches).

Several approaches were presented but, in all cases, the centre of competence acts as a broker for requests from cultural heritage institutions, identifies the most appropriate contractor or consortium, and commissions the work accordingly. The revenues for the centre come from a mixture of fees paid by associated institutions, a brokerage fee for the contracts and opportunities for participation of the centre in new research and development projects.

This business model requires, first, a sufficient number of members and associated partners, and, second, a stable legal status which allows the centre to tender. Cooperation between centres with non-overlapping lines of activity can progress in at least two lines: exchange of information (for example, about requests outside the scope of the centre), and new joint R&D proposals (such as semantic cross-linking between the descriptions of content).

Programme: `http://v-must.net/sites/default/files/Workshop_Competence_Centres.pdf`

Panel: Virtual Research Community for the Preservation of DCH

Venue: ICT 2013 Lithuanian Exhibition and Congress Centre, 7 November 2013, Vilnius (Lithuania)

Organised by: ICT2013 was organised by the European Commission in partnership with the Lithuanian Presidency of the Council of the EU, and the official sponsors of the Presidency.

Participants: representatives from several EU projects, among which:

- DCH-RP: Digital Cultural Heritage Roadmap for Preservation
- SCIDIPES: SCIence Data Infrastructure for Preservation — Earth Science
- APARSEN: Alliance Permanent Access to the Record of Science in Europe network
- EUDAT: Towards a European Collaborative Data Infrastructure
- CHAIN REDS: Coordination and Harmonization of Advanced e-Infrastructures for Research and Education Data Sharing
- DARIAH: Digital Research Infrastructure for Arts and Humanities
- DASISH: Data Service Infrastructure for the Social Science and Humanities
- CLARIN: Common Language Resources and Technology Infrastructure
- SCAPE: SCAlable Preservation Environments
- SUCCEED: Support Action Centre of Competence in Digitisation
- PREFORMA: a joint Pre-Commercial Procurement project

Report: No documentation of the meeting has been published yet. See also `http://ec.europa.eu/digital-agenda/events/cf/ict2013/item-display.cfm?id=10659`

Round table: On the future of research and funding in digitisation and the possible roles of Centres of Competence

Venue: Digitisation Days in Madrid, 20 May 2014, Biblioteca Nacional de España, Madrid, Spain

Organised by: Succeed and the IMPACT Centre of Competence in Digitisation with the cooperation of Biblioteca Nacional de España

Content: A round table meeting was organised at Digitisation Days, on 20 May, to obtain input from stakeholders regarding future research to further the development of the digital libraries. In addition, stakeholders were asked to give their view on the role of Centres of Competence in this. The discussion was structured around the topics USERS, ACCESS and DATA.

Participants: Some 30 librarians and researchers joined this workshop, and discussed the following topics:

1. The current state of affairs.
2. What stops us from making progress.
3. What would help to make progress.
4. What roles competence centres can play to move the issues forward.

The outcomes to date are, on the one hand, that *users* is not a generic term, but there are various groups of users with different needs. Centres of competence should try to liaise with the different groups and understand their needs. For instance, digital humanities scholars want better OCR, but researchers who develop OCR correction tools want good ground truth and standards for testing their tools. On the other hand, another outcome is that users and memory institutions should interact more: "From conservation to conversation".

In addition, participants have discovered that IPR issues are still huge boundaries as regards they also believe that it is important that data should originate from trusted repositories and be available for the long term.

To sum up, Centres of Competence can play a role by showcasing good practice, organising the evaluation of tools and standards, coordinating initiatives for new research proposals, networking and teaching.

Technical workshop: on the interoperability of digitisation platforms

Venue: 2 October 2014, National Library of the Netherlands (KB), The Hague, The Netherlands.

Organised by: the Succeed project

Participants: 19 researchers, librarians, and computer scientists from Germany, Poland, Belgium, the United Kingdom, the Netherlands and Spain

Reports: `http://www.digitisation.eu/blog/succeed-interoperabily-workshop-report` `http://www.digitisation.eu/blog/succeed-interoperability-workshop-report-2`

Panel discussion: Digitisation, conservation and preservation in digitisation

Venue: Succeed Final conference, Paris: 28 November 2014 Bibliothèque nationale de France, Paris, France.

Organised by: the Succeed project

Participants: Irene Haslinger (National Library of the Netherlands, host), Hildelies Balk (Impact Centre of Competence), Daniel Pletinckx (V-MusT.net), Ed Fay (Open Preservation Foundation), Stéphane Reecht (Bibliothèque nationale de France)

Report: http://www.digitisation.eu/blog/vimeo-video18_panel-discussion-digitisation-conservation-preservation-digitisation

Hackathon: Succeed First Developers' Workshop

Venue: 19-20 September 2013, National Library of the Netherlands, The Hague, The Netherlands

Organised by: Succeed

Participants: some 15 software developers from Austria, England, France, Poland, Spain and the Netherlands, experts from the digitisation as well as digital preservation communities, with some additional Taverna expertise. About half of the participants had participated in the Planets, IMPACT or SCAPE projects; the other half of them were new to the field. There were also some developers beyond the MLA community, i.e. a team from the Leiden University Medical Centre attended the hackathon to learn how they could use the T2-Client for their purposes.

Report: http://www.digitisation.eu/blog/1st-succeed-hackathon-kb

Hackathon: Succeed Second Developers' Workshop

Venue: 10-11 April 2014 Department of Software and Computing Systems of the University of Alicante, Spain

Organised by: Succeed

Participants: Some 15 programmers and researchers from Germany, Poland, the Netherlands and Spain.

Report: http://www.digitisation.eu/blog/succeed-2nd-hackathon