EUROPEAN HOLOCAUST RESEARCH INFRASTRUCTURE

European Holocaust Research Infrastructure
H2020-INFRAIA-2014-2015
GA no. 654164

D13.4

Trusted Digital Repository workshop

Ellen Leenarts
DANS-KNAW

Michael Levy
USHMM

Mike Priddy
DANS-KNAW

Linda Reijnhoudt
DANS-KNAW

Start: January 2017 [M21]
Due: February 2017 [M22]
Actual: September 2018 [M39]

EHRI is funded by the European Union

Document Information

Project URL	www.ehri-project.eu
Document URL	
Deliverable	D13.4 Trusted digital repository workshop
Work Package	WP13
Lead Beneficiary	DANS-KNAW
Relevant Milestones	MS2
Dissemination level	Public
Contact Person	Ellen Leenarts, ellen.leenarts@dans.knaw.nl
Abstract (for dissemination)	The EHRI workshop ‘Trusted Digital Repository’ provided an overview of available techniques to evaluate the maturity in digital preservation of a CHI and the process of obtaining a globally recognised Trusted Digital Repository (TDR) status. Prior to the workshop attendees could fill in a survey on digital preservation that was used as valuable input. During the workshop the attendees were first introduced to the topics and achieved a better understanding of these due to the break-out structure of the workshop. The workshop took place on 26 June 2018 in Vilnius, Lithuania.

Table of Contents

Table of Figures

1 EHRI Trusted Digital Repository workshop (June 2018)

The 4-hour workshop “Trusted Digital Repository” took place on June 26th 2018 in Vilnius, at the Vilna Gaon State Jewish Museum (VGSJM). The workshop targeted institutions that wish to gain insights into sustainable repositories to manage digital objects.

In total, twenty representatives of the following institutions attended the workshop: CDEC, Cegesoma, INSHR-EW, IFZ, ITS, Kazerne Dossin, KCL, ONTOtext, USHMM, Wiener Library and Yad Vashem.

1.1 Context of the EHRI Trusted Digital Repository workshop

The workshop “Trusted Digital Repository” was organised as part of Task 13.2 Secure Long-term Access Infrastructure for the preservation of Holocaust Research Objects. An important aspect in this task is to provide EHRI institutions holding digital material practical guidance on digital preservation. At the same time, it is important for WP13 to investigate what the current status is at these institutions with regard to digital preservation and access infrastructure and what their needs for support are.

The outcomes of two workshops, one on “Data Management Planning” (D13.3) and one on “Trusted Digital Repositories” (D13.4) are input sources for the “Long-term Access Infrastructure for Preserving Holocaust Research Objects” (D13.2). The three deliverables are the outcome of the activities carried out in Task 13.2 “Secure Long-Term Access Infrastructure for the Preservation of Holocaust Research Objects”. This secure long-term access infrastructure will consist of a set of guidelines, principles and services that enable organisations to provide durable access to digital Holocaust resources.

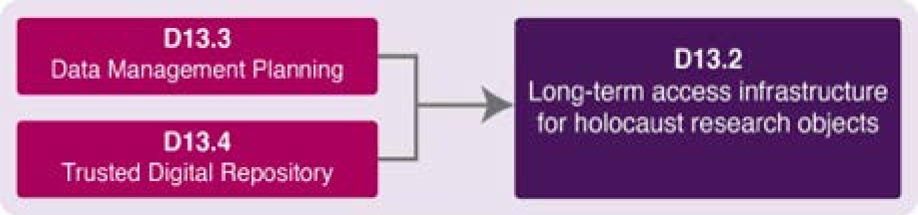

D13.3
Data Management Planning
D13.4
Trusted Digital Repository
D13.2
Long-term access infrastructure for holocaust research objects

Figure 1 EHRI project Deliverables in relation to Task 13.2, “Secure Long-term Infrastructure for the Preservation of Holocaust Research Objects”

The data management planning workshop described in D13.3 took place on 31 July and 1 August 2017. It was decided and approved by the PMB to organise the workshop “Trusted Digital Repositories” prior to the GPM in 2018 to give partners the opportunity to participate without much additional investment in travel costs and time. Therefor the deadline of the deliverable was moved to M39.

1.2 Objectives of the workshop

Digital objects must be managed, curated and archived in such a way that they remain meaningful into the future. This means that the data objects are properly documented, formatted, protected, stored and made available in digital repositories[1]. Sustainability of repositories raises a number of challenging issues in different areas: organizational, technical, financial, legal, etc. Certification can be an important contribution to ensuring the reliability and durability of data repositories. In the context of the workshop, WP13 aimed to describe, inform and support institutions on aspects of digital preservation, so that attendees can:

[1] “A digital repository is a mechanism for managing and storing digital content. Repositories can be subject or institutional in their focus.” http://www.rsp.ac.uk/start/before-you-start/what-is-a-repository/

- gain insights on aspects of sustainable repositories to manage digital objects,
- gain knowledge on standards, guidelines and tools to maintain digital objects for the long-term,
- contribute to the long-term access infrastructure of Holocaust digital objects.

The workshop introduced the assessment of digital repositories in a number of ways such as capability maturity modelling, levels of preservation according to the NDSA[2], and basic certification according to the Core Trust Seal[3]. Furthermore, the workshop provided the attendees with information why persistent unique identifiers are used by so many data repositories to make their data objects accessible and reliable.

Trusted digital repositories and archives have the skills and competences, IT infrastructure, policies, and work processes in place to enable digital preservation. Of course, the repository should be able to adjust its policies etc., depending on external factors and influences. Questions such as the following are very relevant:

- *What are your preservation plans for these kinds of resources to make sure they are maintained for the future too?*
- *In what way is your preservation plan different from traditional preservation planning?*
- *Where are the digital objects stored?*
- *How are they described?*
- *How can you update the content or the description?*
- *Are these processes documented?*
- *Is there a technology watch, or a disaster recovery strategy?*

1.3 Survey on digital preservation

The representatives of Collection Holding Institutions (CHIs) who registered for the workshop were invited to participate in a short survey on digital preservation. The survey was based on a much longer survey that was used in the Knowledge Complexity (K-PLEX) research project[4]. Only one response per institute was needed. In total we received 7 responses to the survey. As there were a limited number of CHIs planning to participate, the more technical partners were not supposed to fill in the survey. 7 responses to the survey can therefore be viewed as an acceptable number, and representatives of 9 Collection Holding Institutions (CHI) attended the workshop.

In the first presentation “Hidden by Not Sharing/Hidden by Sharing” by Mike Priddy (see appendix B) the responses to the survey by the EHRI partners were considered. The responses to a number of questions were used to illustrate that overall the results of the small group of 7 responses did not deviate significantly from the results of the K-PLEX project survey.

1.4 Programme

09.00 – 09.15 AM	Welcome (Ellen Leenarts, DANS-KNAW)
09.15 – 10.00 AM	Hidden by Not Sharing/Hidden by Sharing (Mike Priddy, DANS-KNAW)
10.00 – 10.15 AM	Certification – CoreTrustSeal (Ellen Leenarts, DANS-KNAW)
10.15 – 10.30 AM	Persistent and unique identifiers (Linda Reijnhoudt, DANS-KNAW)
10.30 – 11.00 AM	Coffee / Tea break
11.00 – 11.15 AM	Capability development modelling (Mike Priddy, DANS-KNAW and Michael Levy, USHMM)

[2] National Digital Stewardship Alliance. https://ndsa.org/about/

[3] https://www.coretrustseal.org/

[4] https://kplex-project.eu/

11.15 – 12.15 PM	3 one-hour break-out groups: • Certification and CoreTrustSeal • Persistent and unique identifiers and • Capability development modelling
12.30 – 12.45 PM	Report back
12.45 – 13.00 PM	Round up

There were 3 break-out sessions that were all introduced centrally before they took place.

1.5 Summaries of the presentations

1.5.1 Hidden by Not Sharing/Hidden by Sharing by Mike Priddy (DANS-KNAW)

As part of the Knowledge Complexity[5] project DANS-KNAW undertook research into Hidden Data and the Historical Record[6] to better understand why humanities data[7] resists big data analysis techniques and its inclusion in big data projects. As part of this research DANS-KNAW conducted interviews with practitioners in cultural heritage institutions (predominantly archival institutions) about their practice, as well as an open online survey to support the knowledge gained from the interviews.

All the interviewees who worked in CHIs, 83% of the K-PLEX respondents and all of the respondents to the workshop survey, hold both digital and analogue resources. Therefore, understanding digital preservation is essential for archival practitioners in cultural heritage institutions.

It may be obvious that resources useful for researchers may be hidden through not making them shared and available for use, however, it is less apparent that sharing data resources can also ‘hide’ resources that may help to answer a research question. There are many factors involved in sharing and not sharing, both active and passive actions; a few were presented along with key recommendations from K-PLEX that affect cultural heritage institutions (Figure 2).

Archivists are well versed in handling knowledge complexity and in supporting and aiding researchers with their research questions. Embodied knowledge is a big challenge for the institution as well as the researcher using the resources. Expertise is built up over time, and although shared, it is often in person. A practitioner may help a researcher to recommend resources for a given research question. It is still a very much a process of human to human interaction, and this can be labour intensive in an environment of impact driven metrics and key performance indicators. Even with just basic online metadata researchers are now prepared with resource identifiers for specific documents before “they are even coming to the reading room”[8], and moreover, researchers are asking for access to data online so they can conduct their research at their desk. The demand for digital and digitised data is there, and the expectation that it should be available will continue to grow as new research methods are employed. Cultural heritage institutions are beginning to engage with the outside world rather than waiting for visitors, as they move from inward-looking decision making to an outward-looking approach.

[5] https://kplex-project.eu/

[6] https://kplexproject.files.wordpress.com/2018/06/kplex_deliverable-d3-1.pdf (30 March 2018)

[7] For the purpose of the KPLEX study the term ‘data’ was considered to encompass all sources of knowledge held by cultural heritage institutions that may be used by researchers.

[8] Interviewee of K-PLEX. https://kplexproject.files.wordpress.com/2018/06/kplex_deliverable-d3-1.pdf

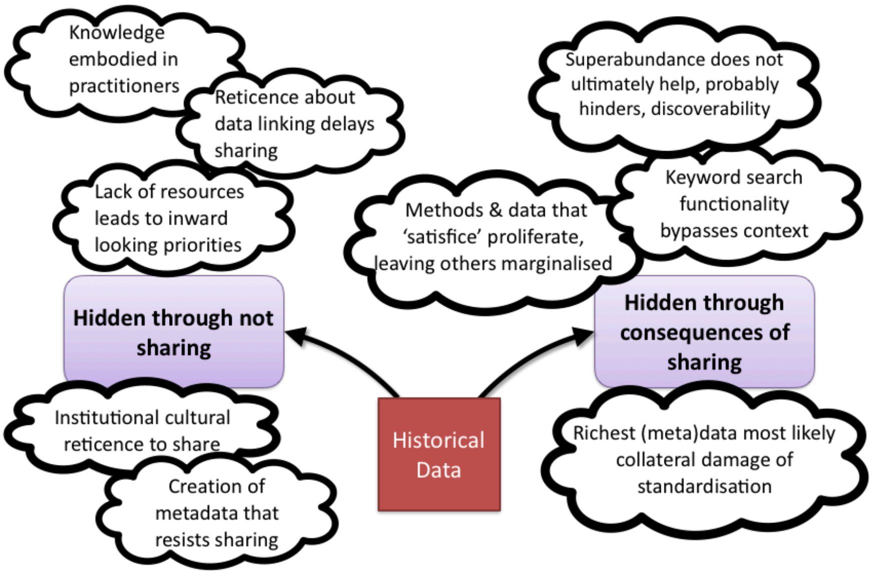

Knowledge embodied in practitioners
Reticence about data linking delays sharing
Lack of resources leads to inward looking priorities
Hidden through not sharing
Institutional cultural reticence to share
Creation of metadata that resists sharing
Historical Data
Methods & data that 'satisfice' proliferate, leaving others marginalised
Superabundance does not ultimately help, probably hinders, discoverability
Keyword search functionality bypasses context
Hidden through consequences of sharing
Richest (meta)data most likely collateral damage of standardisation

Figure 2 Hidden through not sharing and hidden through the consequences of sharing

The creation of metadata (archival catalogue records) can resist the sharing and publishing online, sometimes due to complexity of what is being described but also due to institutional and individual practices changing over time. Moreover, the fact that not all catalogue records are in a digital format or are not complete, is only part of the issue. The need for standardisation whilst publishing and aggregating may cause the richness in, and the context of, the data to be lost as it is often a minimal common set of information that is aggregated.

More and more cultural heritage institutions are placing their holdings catalogues online. As one K-PLEX interviewee put it: “you can’t stay in your own cocoon to do your own things.” However, only 9% of respondents to the K-PLEX survey have 100% of their holdings catalogue publicly online[9]. Cultural heritage institutions appear to have a cultural reticence to sharing online. In K-PLEX we were not able to investigate further what where the cultural challenges that influenced the reticence to share. Although, anecdotal evidence from projects such as CENDARI[10] & EHRI suggests there appears to be no broad drive to share knowledge in a digital form. However, there are a number of non-cultural factors involved, including the enormity of the challenge, but primarily it comes down to the lack of capacity, capability, and hence financial resources (Figure 3).

For some cultural heritage institutions there is an attempt to out-do Google: “to jump Google and go directly to your portal”[11], in hoping to develop a community that regularly visits the institutional website. However, the knowledge is easier to discover through the Google aggregation methodologies. Though for the researcher the approach of a ‘Google search’ loses context amongst the considerable noise. For practitioners, context is very important and a lot of effort is put into curation of the archival holdings, which is why aggregators such as the EHRI portal, APEF[12], and CENDARI maintain the archival structure wherever possible. The challenge is to understand and manage both online routes to the institution’s resources. There is, perhaps, a reticence to share data online because of the ubiquitous nature of Google reducing the online relevance of the institution. By not having the entire catalogue of an institutions holdings available online it “skews research towards what's easily available, properly catalogued, easy to find and ideally available freely online because that's

[9] Compared to 16% of respondents whose institution have less then 20% of their holdings descriptions online.
[10] Collaborative European Digital/Archival Infrastructure http://www.cendari.eu/
[11] Interviewee of K-PLEX. https://kplexproject.files.wordpress.com/2018/06/kplex_deliverable-d3-1.pdf
[12] Archives Portal Europe Foundation http://www.archivesportaleuropefoundation.eu/

what researchers will go to because it's just the most convenient[13].” Thus delays and choices in what is shared online exacerbates a Matthew Effect.

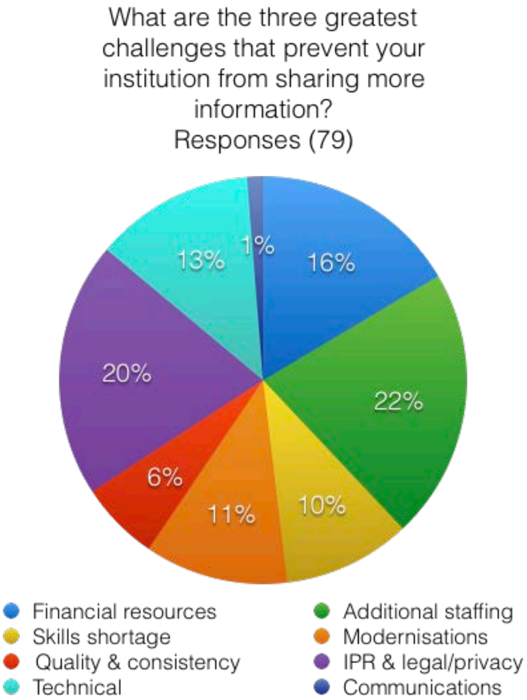


Figure 3 Challenges to sharing information (about holdings). Source: K-PLEX: Deliverable D3.1 Report on Historical Data as Sources[14].

Archives have yet to have their ‘industry’ fundamentally turned upside down like many other creative and knowledge-based domains by the transformative nature of the digital (third industrial) revolution. Thus, archival cultural heritage institutions are being left behind, in part because the digital preservation does not provide a performative gain or replace physical artefact preservation, but generally adds a novel set of processes and effort that need to be undertaken. As William Gibson put it: “The future is already here — it's just not very evenly distributed[15].”

K-PLEX has only scratched the surface in this area and although we are not entirely certain, no one has recently researched the practice and concerns of archivists and how they support research. This research reinforces what has been discovered through infrastructural projects and previous workshops and events earlier on the vector of this work.

1.5.2 Capability development modelling by Mike Priddy (DANS-KNAW)

This presentation was based upon outcomes developed from the workshops on Sustainable Digital Publishing of Archival Catalogues of Twentieth-Century History Archives (held in 2015)[16]. The main goal of these DARIAH “Open History” workshops was to enhance the dialogue between (meta-)data providers and research infrastructures, which arose from the experiences of both EHRI and CENDARI projects[17].

[13] Interviewee of K-PLEX. https://kplexproject.files.wordpress.com/2018/06/kplex_deliverable-d3-1.pdf

[14] https://kplexproject.files.wordpress.com/2018/06/kplex_deliverable-d3-1.pdf

[15] https://quoteinvestigator.com/2012/01/24/future-has-arrived/

[16] Veerle Vanden Daelen, Jennifer Edmond, Petra Links, Mike Priddy, Linda Reijnhoudt, et al. Sustainable Digital Publishing of Archival Catalogues of Twentieth-Century History Archives. *“Open History: Sustainable digital publishing of archival catalogues of twentieth-century history archives”*, Dec 2015, Brussels, Belgium. 2016. 〈hal-01281442v2〉

[17] Ibid.

The challenges to aggregating metadata describing institutional archival holdings are manifestly multifarious:

- The richness of metadata may be lost in standardisation to provide archival records.
- Each country appears to have its own archival cataloguing application(s).
- Each CHI has its own unique institutional (or even individual) approach to using cataloguing software to meet existing cataloguing practices.
- Exporting metadata from the catalogue is usually, at best, a secondary consideration for the developers of the software. If there is an export option, the software invariably exports metadata in a proprietary format unique to the software.
- Technical challenges are *assumed* as biggest hurdle to sharing archival description metadata, but often this is not the case.
- Many CHIs lack unique consistent identifiers usable online.

In response to these challenges the EHRI Capability Development Model (CDM) was developed to evaluate if a CHI can meet the requirements, in terms of capability, maturity and capacity for aggregation of their archival descriptions into the EHRI Portal[18].

Capability maturity modelling (CMM)[19], used to develop the EHRI CDM, is a methodology to evaluate the quality of services[20] provided by an institution and the processes and activities need to support the service delivery. It has two dimensions of evaluation: firstly, how complete are the processes and secondly how optimised (or mature) are the activities needed to provide the service. It is not necessary to aim for all capabilities to be at the highest possible maturity (Figure 4). In CMM there are normally five maturity levels: 1 initial, 2 repeatable, 3 defined, 4 managed and 5 optimised. In the CDM we use levels 1 to 3 only, plus level 0 for not applicable (Figure 5). Thus, the EHRI CDM is simplified and is used descriptively in a self-assessment process of evaluation targeted to the specific purpose of aggregation of archival descriptions[21] (Figure 5). The level that a CHI needs to attain, to become a sustainable publisher of metadata that can be aggregated by EHRI, is set by EHRI.

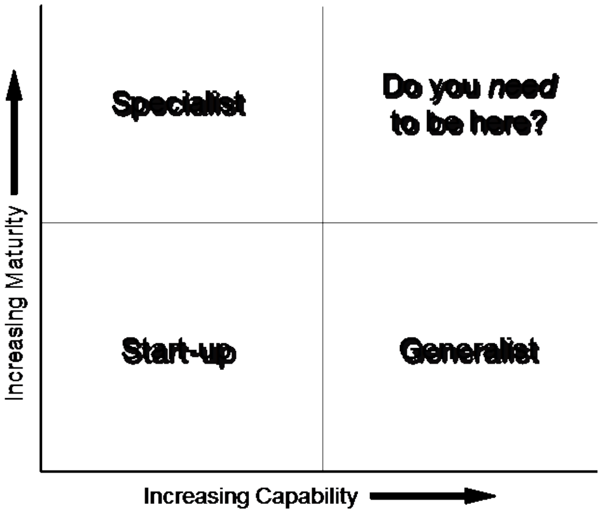


Figure 4 Capability vs. Maturity

[18] https://portal.ehri-project.eu/
[19] A short and simple description is available here: https://www.itgovernance.co.uk/capability-maturity-model
[20] Initially for, and more often used on, software-based services.
[21] There is an element of prescriptive usage in that the EHRI CDM self-assessment tool describes actions needed to improve the maturity of an activity.

The EHRI CDM has 5 main capabilities (or goals) to be assessed, each with a number of activities (sub-goals). These capabilities are self-assessed by the CHI in a spreadsheet[22], which includes minimum maturity level required per activity, action to be undertaken if minimum level is not met, assessment advice, and additional notes for EHRI integration.

	Goal		Sub-goal	1 - initial	2 - repeatable	3 - defined
2	A CHI describes its holdings	a	According to international standards accepted in your field	The institute is aware of standards for descriptions for holdings, such as ISAD(G), but does not apply them.	Standards have been considered and internal data formats are checked against the standards.	When considering changes to the cataloguing system, standards are part of the procedure.
2	A CHI describes its holdings	b	In a consistent way (training, policies, best practices) considering the choice of fields, the detail level of the descriptions and parallel descriptions in multiple systems/languages	Employees receive no guidance on what fields to use or how the data should be recorded. Different cataloguers use different rules to describe the holdings.	Cataloguers receive some instructions about how to catalogue items and collections, but this knowledge is transferred only verbally and is not enforced.	There are guidelines on which fields to use, information on how to use them and to what extent.
…	…		…	…	…	…

Figure 5 Example of goal and sub-goal assessment criteria from the CDM spreadsheet

There are many benefits to using CMM and a CDM to assess the quality of a service that an institute provides; be that from providing a common vocabulary to discuss quality, to identifying underused capabilities, gaps and weaknesses in service provision, or even identifying evidence to support certification[23]. However, within the EHRI community of CHIs adoption of the CDM is a challenge as not only is the value of the process not well understood, but also no single person has the insight to be able to evaluate all the self-assessment goals and sub-goals. Moreover, there is a wide variety on maturity in the community of CHIs, and thus different, perhaps bespoke, approaches may need to be employed to understand the circumstances of a CHI which are possibly not sustainable in the long term. The CDM can be used to help guide the CHI on how to improve its maturity when a CHI is not sure what to do next.

1.5.3 Introduction to NDSA levels by Michael Levy (USHMM)

In 2013, the National Digital Stewardship Alliance developed and published "Levels of Preservation," as an aid to assist those who are entrusted with digital materials that are worthy of long-term preservation. At that time there were other instruments that practitioners could use to assess their institutions and their practices with respect to their preservation activities and to help plan. These include such guidance documents, recommendations, and standards instruments as the Open Archival Information System reference model[24], Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC:CC)[25];

[22] The EHRI CDM spreadsheet is available here: https://tinyurl.com/y8kdtcv2

[23] For example, for a trusted digital repository certification such as CoreTrustSeal (see below)

[24] https://www.iso.org/standard/57284.html

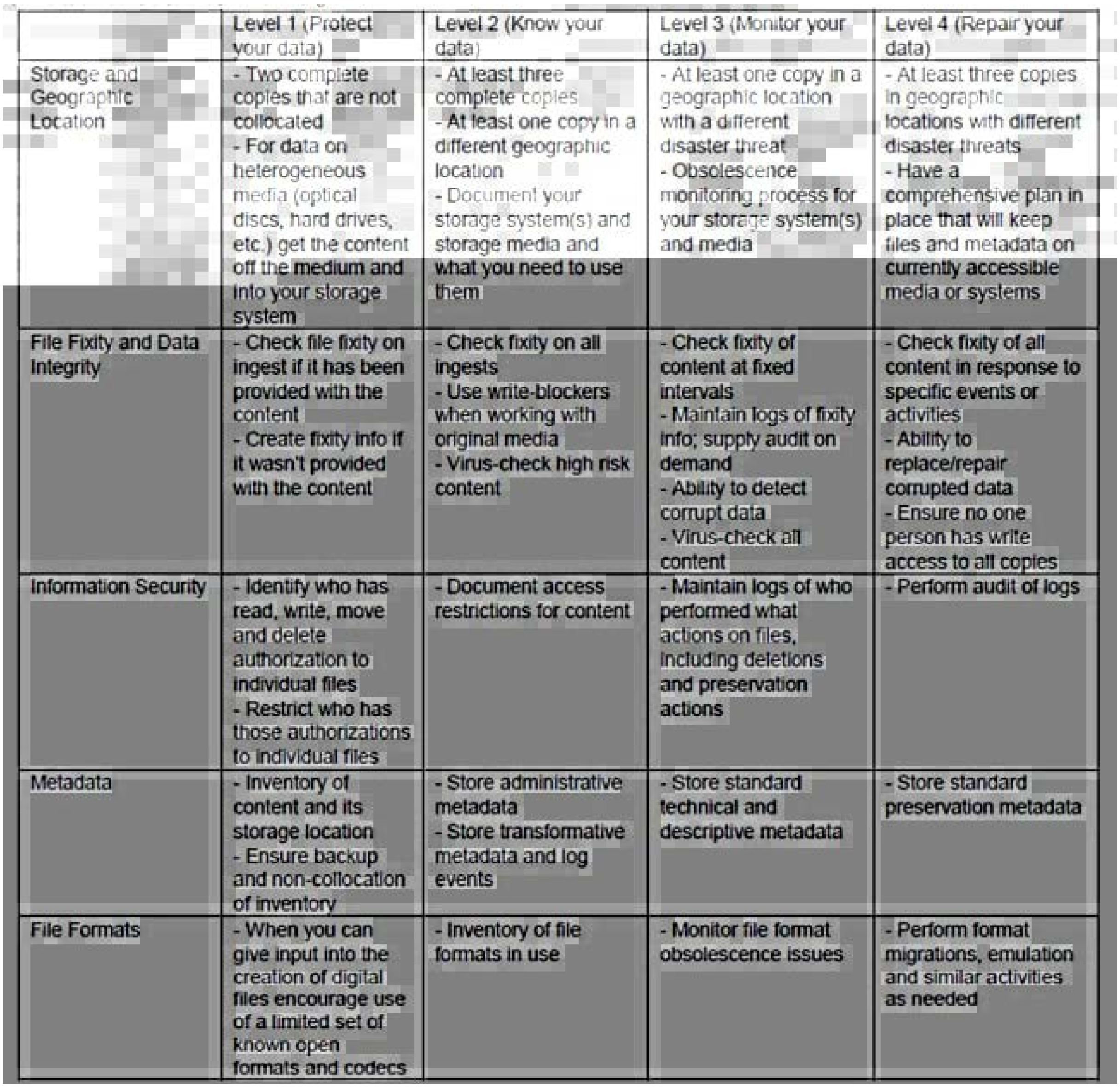
[25] https://en.wikipedia.org/wiki/Trustworthy_Repositories_Audit_%26_Certification

DRAMBORA[26], and others. Such instruments were developed to be comprehensive and to lead an institution towards becoming a digital repository that is able to stand up to an auditing process. Those standards encompass issues such as organizational administration, funding and staffing models, legal issues, proper operating procedures, oversight, and many other issues. Such a comprehensive process may be daunting for people working in small institutions with limited resources, and/or institutions that are just beginning to grapple with digital preservation. Many institutions may not be ready to marshal sufficient resources to begin serious work towards certification. If the result of not attempting to engage in receiving a preservation certification is to do nothing, this is a very poor outcome. A lightweight approach such as the NDSA Levels may be what is needed to help practitioners take some practical steps that do not require heavy investment in time or financial resources.

The NDSA Levels were developed in an attempt to provide a set of guidelines that any practitioner can understand and that can be used as a self-assessment tool and that provides guidance for improvements that could be put into practice fairly readily. The Levels are organized into five functional areas: storage and geographic location; file fixity and data integrity; information security; metadata; and file formats. Each of the five functional areas are graded on 4 numbered tiers, each tier generally encompassing the one below and adding additional, more stringent measures. The Levels are often presented in a grid, 5 rows (functional areas) and 4 columns (levels) that can be presented on a single page[27].

[26] http://www.dcc.ac.uk/resources/repository-audit-and-assessment/drambora

[27] https://ndsa.org/activities/levels-of-digital-preservation/



	Level 1 (Protect your data)	Level 2 (Know your data)	Level 3 (Monitor your data)	Level 4 (Repair your data)
Storage and Geographic Location	- Two complete copies that are not collocated - For data on heterogeneous media (optical discs, hard drives, etc.) get the content off the medium and into your storage system	- At least three complete copies - At least one copy in a different geographic location - Document your storage system(s) and storage media and what you need to use them	- At least one copy in a geographic location with a different disaster threat - Obsolescence monitoring process for your storage system(s) and media	- At least three copies in geographic locations with different disaster threats - Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems
File Fixity and Data Integrity	- Check file fixity on ingest if it has been provided with the content - Create fixity info if it wasn't provided with the content	- Check fixity on all ingests - Use write-blockers when working with original media - Virus-check high risk content	- Check fixity of content at fixed intervals - Maintain logs of fixity info; supply audit on demand - Ability to detect corrupt data - Virus-check all content	- Check fixity of all content in response to specific events or activities - Ability to replace/repair corrupted data - Ensure no one person has write access to all copies
Information Security	- Identify who has read, write, move and delete authorization to individual files - Restrict who has those authorizations to individual files	- Document access restrictions for content	- Maintain logs of who performed what actions on files, including deletions and preservation actions	- Perform audit of logs
Metadata	- Inventory of content and its storage location - Ensure backup and non-collocation of inventory	- Store administrative metadata - Store transformative metadata and log events	- Store standard technical and descriptive metadata	- Store standard preservation metadata
File Formats	- When you can give input into the creation of digital files encourage use of a limited set of known open formats and codecs	- Inventory of file formats in use	- Monitor file format obsolescence issues	- Perform format migrations, emulation and similar activities as needed

Figure 6 Version 1 of the levels of digital preservation - NDSA

The NDSA Levels of Preservation Working Group (also referred to as the Levels Reboot Team) are currently working on updating the Levels and on developing case studies and materials for teaching and promulgating digital preservation concepts.

More information about the Levels and the Reboot program can be accessed at https://ndsa.org/working-groups/levels-of-preservation/

1.5.4 Introduction to certification - CoreTrustSeal by Ellen Leenarts (DANS-KNAW)

The framework of international trusted digital repository certification standards, as can be seen in figure 7, consists of several standards. By means of these assessment instruments digital repositories can improve the quality of their work processes and management systems to become a certified 'Trustworthy Digital Repository' (TDR). There are three certification instruments available, with increasing degrees of complexity and depth:

- CoreTrustSeal (CTS) (based on Data Seal of Approval (DSA), and World Data System (WDS))[28]
- Nestor Seal (verification according to DIN 31644)[29]

[28] https://www.coretrustseal.org/

- ISO 16363 certification[30]

The assessments vary in intensity from a peer review of completed documentation (self-assessment) to a prepared on-site visit by an external audit team. These instruments are used worldwide. Data sponsors, producers and re-users may trust any managing body that has been certified according to one of the above standards.

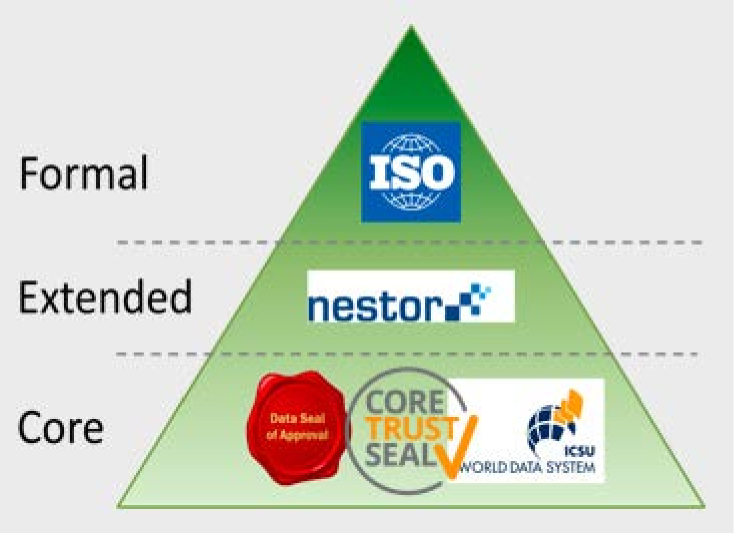


Figure 7 Global certification landscape Trusted Digital Repositories

The CoreTrustSeal is, as the name suggests, a core level assessment. The Data Seal of Approval and the World Data System merged their data certifications under the umbrella of the Research Data Alliance. It is a community-based standard that offers a certification tool and extended guidance[31]. When a repository applies for certification the assessment is reviewed by community peers. More than 130 repositories have been certified with the DSA, WDS and now CTS.

The CTS assessment consists of 16 requirements[32] on context of the repository (e.g. designated community), organizational infrastructure, digital object management and technology. The assessment should be in English and references to public documents as evidence are strongly encouraged. When certified, the assessment becomes publicly available on the CoreTrustSeal site. Certification is for 3 years.

The effort that is involved in applying for the certification varies and depends on your maturity level of entry. In the case of DANS, the first self-assessment took 2 weeks, followed by several hundreds of hours to improve the work processes. A report[33] on a survey by the National Coalition of Digital Preservation in the Netherlands on the level of investment for the Data Seal of Approval (2016) shows that DANS' effort is not exceptional.

There are clearly benefits in investing the certification process. The external benefits are enhancing the reputation of a repository (or archive) and building stakeholder (funders, host organisations, publishers) confidence. Internally the assessment process raises awareness

[29] http://www.langzeitarchivierung.de/Subsites/nestor/EN/Siegel/siegel_node.html
[30] http://www.iso16363.org/
[31] https://www.coretrustseal.org/wp-content/uploads/2017/01/20180629-CTS-Extended-Guidance-v1.1.pdf
[32] https://www.coretrustseal.org/wp-content/uploads/2017/01/Core_Trustworthy_Data_Repositories_Requirements_01_00.pdf
[33] http://www.ncdd.nl/wp-content/uploads/2016/10/201611_DE_Houdbaar_Report_DSA-survey_2016.pdf

about digital preservation and improves communication within the repository and work processes.

1.5.5 Persistent Identifiers by Linda Reijnhoudt (DANS-KNAW)

Making your digital assets accessible and findable in the long-term over the Internet needs the use of persistent and unique identifiers that can be used in the URL.

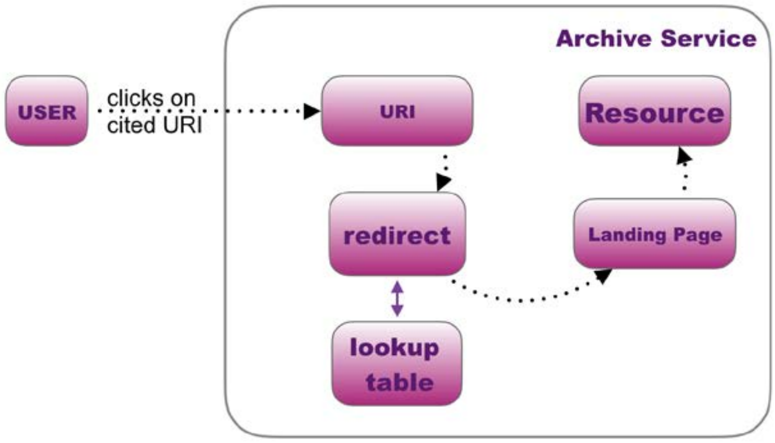


Figure 8 Resolving a HTTP redirects in action

As Figure 8 shows, when the Landing Page URL at the website of the Archive Service changes, the webserver needs to redirect the user to the new Landing Page URL. In order to keep all the old URLs viable, this lookup table will only expand over time. In case this lookup table is not kept synchronized, which can usually only be done by the webmaster, the old URL will result in a Page Not Found error.

One solution to this problem is the use of Persistent Identifiers (PID). A PID is a URL, that is kept persistent and resolvable through policies at the Archive Service. When a Landing Page URL changes, this change must be reported to the PID System so the PID Resolver will correctly redirect the PID to the new Landing Page URL (Figure 9).

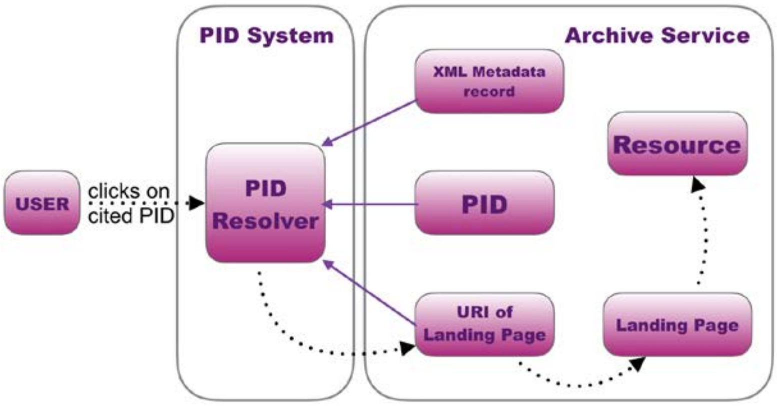


Figure 9 Resolving a Persistent Identifier

For an Archive Service to implement PIDs, it needs to have the ability to support persistent and unique identifiers. This means these identifiers need to be kept unchanged, regardless of software versions, data migrations, software suppliers and new insights.

The technical infrastructure of a PID System is in itself not enough to keep the PIDs of an Archive Service resolvable. Policies and practices must be instigated at the Archive System to make sure the changes in the URLs are shared with the PID System, so the PID remains resolvable.

Because of this promise to keep the PIDs indefinitely, CHIs must choose wisely the types of resources (historical person, collection, controlled vocabulary term) for which to implement PIDs.

2 Results of the EHRI Trusted Digital Repository workshop

In this chapter the results of the break-out groups during the workshop are described. These are followed by the recommendations of the K-PLEX project and overall recommendations to EHRI expressed at the round-up of the workshop.

At the workshop the participants could choose between 3 break-out groups:

1. Digital preservation levels and capability development modelling,
2. Looking more closely at the requirements of the CoreTrustSeal by evaluating available examples and compare the local situation at your CHI,
3. The use of persistent unique identifiers (PIDs).

As it turned out the participants either joined the 1st or 3rd group to acquire more knowledge on enhancing digital preservation and the use of PIDs. Therefore, there are no results below on the 2nd break-out group (CoreTrustSeal).

2.1 Results of break out group "Capability development modelling":

It could be considered that the digital revolution has done little to deliver performative or transformative benefits to the archival world. In fact, it has simply introduced more work to be done with managing and preservation both physical and digital objects. Moreover, digital preservation is quite different to physical preservation, not only in terms of workflows, but also conceptually, managerially and procedurally; requiring new policies and processes. Digitising physical materials does not mean that they are preserved; digital simulacra need preservation too and digitally-born objects have additional challenges to consider.

In the change management that is essential to transition to a physical and digital archive[34] there were clear challenges identified. For change to occur in institutions there needs to be a common language, for example between IT services and archival services where domain terms do not have the same meanings: a corrupted file is not the same as an altered file. When it comes to assessing an institution's capacity and capabilities one needs management buy-in as it will take resources to do so, and therefore, there needs to be support and understanding of the process. Thus, the use of Capability Development Modelling to aid EHRI archives to become more digitally mature should be piloted with a few institutions; "It would be good if it would be in a EU project to get some initial resources for the first pilot group."

Participants in the breakout session came from institutions that vary in size and level of technical sophistication and resources. There seemed to be a general recognition among the participants of the need to pay continual attention to long-term preservation of digital assets, and that each institution faces similar challenges. Among these challenges are the need for additional resources. These resources are different than those required for preservation of physical materials, which are often unfamiliar to management. Resources, and especially technical resources, are often scarce in cultural heritage institutions. Securing the essential resources will require communicating the needs for digital preservation activities, along with the costs, risks, and benefits to institutional management. In addition, digital preservation concepts tend to be very unfamiliar with IT management, who are prone to think of it in terms of business continuity and not in terms of very long-term access to digital materials. The emergence of internationally-recognized digital preservation standards and guidelines, and the existence of digital long-term preservation certification, may be useful in helping staff to persuade and educate institutional top management, including IT management, that attention must be paid to ongoing preservation activities.

[34] All workshop survey respondents and over 80% of the K-PLEX survey respondents held both physical and digital objects.

There are however easy wins, digitised video, for example. Video tape formats and old video recorders are no longer being supported and maintained, and it takes considerable time (and money) to digitise libraries of video recordings. Therefore, the digitised content requires a preservation procedure and policy as a matter of urgency. Moreover, it is then easier to put the digitised material online and the process and impact can be demonstrated. There is an evident need to doing this carefully due to possible privacy issues.

This raises the question of how do archives (and hence archivists) share and link data, especially privacy sensitive data, between institutions? One solution is a secure remote access network (RAN), in which researchers can request access to data from one or more CHIs in an environment that meets the requirements for privacy of the data, set by the data creator/owner. This could be anything from a safe room dedicated to the purpose, to another CHI's reading room, or to a registered computer with additional security access features. For access to the data the researcher must be accredited and an agreement to mutually recognise other CHIs accreditation procedures, or a common accreditation, is required. There are many other agreements and policies that need to be in place for a RAN to operate, but the technology can be straightforward.

Knowledge about archival collections does not only reside with the archivists of the CHI; researchers studying archival material in depth will have more topic specific comprehension of the contents. Therefore, CHIs may wish to consider linking the work of researchers to their holdings, be it through annotations, transformed or derived data, or published papers. The International Image Interoperability Framework (IIIF)[35] is an example of a standard that may be used to facilitate the addition of information about an image that is not originated in the CHI.

2.2 Results of break out group "Persistent Identifiers":

The break-out group discussed the current use of identifiers in the CHIs.

In order to share data in a long-term reliable way, the use of persistent and unique identifiers is paramount. We all know that things will change: CHIs switch cataloguing systems, move storage, rearrange the collections etc. Even then, these identifiers must be kept as-is.

One way to make change part of the solution is to only expose identifiers that do not have any knowledge encoded in them, like the software used, or the hierarchical order of the collection.

Only when the identifiers are guaranteed to be persistent and unique can a CHI consider implementing PIDs. Not all types of resources might be good candidates, considering the amount of extra administration required when managing PIDs, so this decision must not be taken lightly.

In the break-out group, we note that best practices for publishing on the web are difficult to grasp for CHIs. This expertise is often lacking in the institute. Participants hope to improve their practices by participating in projects like EHRI.

2.3 Recommendations K-PLEX project (2018)

> *"The Knowledge Complexity (or KPLEX) project was created with a two-fold purpose: first, to expose potential areas of bias in big data research, and second, to do so using methods and challenges coming from a research community that has been relatively resistant to big data, namely the arts and humanities. The project's founding supposition was that there*

[35] http://iiif.io/

> *are practical and cultural reasons why humanities research resists datafication, a process generally understood as the substitution of original state research objects and processes for digital, quantified or otherwise more structured streams of information.*[36]*"*

Although K-PLEX research was specifically in the context of big data research, there are a number of findings and recommendations that are relevant to EHRI CHIs and the long-term preservation of data. Especially when one considers that archivist are the original custodians and managers of big data, where complex knowledge is measured in kilometres rather than kilobytes.

Some of the findings[37] (in brief) regarding big data that are relevant to archives that hold data are:

- Big data is ill-suited to representing the complexity of information and knowledge found in cultural heritage data and archives.
- Big data compromises the richness of information in cultural heritage collections.
- Standards are both useful to bring together data from disperse and disparate sources, and harmful in that they may not capture the richness and complexity of collections.
- The appearance of openness can be misleading since data can still become hidden by making it openly available.
- Research based on big data is overly opportunistic; finding patterns in data before formulating the research question. Archivists can help to temper this approach.
- How we talk about big data matters, and the archivist needs to both comprehend and participate in the discussions.
- Big data research is about narrative, and lacks inherent objectivity or truth value.
- Dark linking and de-anonymization are real threats and undoubtedly a concern for EHRI and CHIs.
- Organisational and professional practices are being forced to change.

Based upon the K-PLEX research into cultural heritage institutional practitioners' praxis a number of recommendations relevant to the EHRI CHI community are of interest when considering long-term preservation of data. These recommendations are quite simple, not ground breaking, but necessary to move forward.

Researchers should be supported to address any training needs for meaningfully discovering and engaging with data complexity at the point of access. Cultural heritage institutions have historically borne the weight of guiding researchers through their collections but the changing nature of researchers' contact with institutions, with self-guided use of technologies augmenting or replacing personal contact, presents new challenges for ensuring the optimal use of cultural heritage knowledge[38]. This also relates to the CTS self-assessment question VI 'Expert guidance.'

Cross-sectorial training should be considered to enable and encourage understanding and knowledge exchange between cultural heritage, ICT practitioners and researchers. This may better support research questions when the various stakeholders can contribute with mutual apprehension of the complexity of knowledge and the application of 'black-box'

[36] Edmond, J., Horsley, N., Huber, E., Kalnins, R., Lehman, J., Nugent-Folan, G., Priddy, M., (2018). Big Data & Complex Knowledge: Observations and Recommendations for Research from the Knowledge Complexity Project. Trinity College Dublin, Dublin. https://kplexproject.files.wordpress.com/2018/06/trinity-big-data-report-jklr_04-1.pdf

[37] Ibid.

[38] Horsley, N., Priddy, M., (2018) D3.1 KPLEX – Report on Historical Data as Sources – 2018-03-30 https://kplexproject.files.wordpress.com/2018/06/kplex_deliverable-d3-1.pdf

computational methods[39]. Again, this also relates to the CTS assessment question VI 'Expert guidance.'

Institutions should be supported to introduce meaningful measurements of the use of their collections, to overturn the current norms of 1) an absence of data on what gets used; 2) an unmanageable blanket collection of this data, which institutions do not have the resources to analyse or 3) the collection of this data to monitor and support the popularity of collections, which risks becoming a tail that wags the dog. Instead, monitoring of use should be integrated into institutions' practices in a way that is meaningful and useful to them, for example helping to flag collections or data that are 'at risk' of becoming hidden[40]. This is also an aspect of CTS assessment questions VII 'Data integrity and authenticity' and XIII 'Data discovery and identification.'

Further research is required to deepen understanding of practitioners' fears about the possibilities of data linking – and to examine the validity of these concerns amid the uncertain future of the use of big data. It is entirely reasonable that practitioners are worried about the potential for identifying individuals and for sensitive data to become public through data linking when it is not yet certain that current regulations and best practice preclude this[41]. Investigating and perhaps applying procedures employed by social science data archives and national statistical institutions could help to limit the risks. Question IV of CTS 'Confidentiality/Ethics' looks for evidence of mitigation of disclosure risk.

ICT projects aimed at fostering increased sharing through data aggregation and infrastructures should provide long-term support to institutions to ensure developments do not stall, knowledge is not isolated in individual practitioners and technological obsolescence does not threaten progress, endanger data or discourage future participation in such projects[42].

2.4 Recommendations from the round-up of the workshop

The EHRI-2 project has shown that EHRI CHI partners differ widely in size (number of collections, number of employees), the type of digital assets, their IT capabilities, awareness about and level of digital preservation. A future ERIC or funded infrastructure development project involving CHIs that hold Holocaust materials would profit from being supported by digital preservation 'consultants'. There are many topics where CHIs could profit from more guidance.

To know what the capabilities regarding the publishing of digital (meta)data of a CHI is valuable, the EHRI Capability Development Model is a useful tool. Not only does it show the maturity level for the capabilities of a CHI, but it also provides pointers on what to implement or improve in order to advance a capability. The participants of the workshop would like to see a future pilot project where this type of assessment is used for one or more CHIs as a guidance to plan future improvements.

The participants of the workshop indicated they would like to see many more of these face-to-face workshops introducing new concepts and/or explaining how to put existing concepts in practice. The CHIs would also profit from sharing knowledge and practices. This would over time certainly result in increased maturity of the CHI with regard to digital preservation and publishing on the web.

[39] Ibid.
[40] Ibid.
[41] Ibid.
[42] Ibid.

3 Appendix A: Survey prior to the workshop ‘Trusted Digital Repository”

This appendix contains the list of survey questions that were presented to the registered participants of the workshop. The input provided by the CHIs is not presented here, as it would be impossible to do this anonymously. The input was used during the workshop.

1. Please select the collection types that your institution holds
   - Text-based sources
   - Other artefacts and art works
   - Digital data (also digitised resources, digital-born sources)
2. To what extent do you think provisions for data accessibility, such as those contained within Data Management Plans and the Trusted Digital Repository status, achieve sustainable access to, and re-use of, data?
3. To what extent are you confident that your institution's preservation planning and/or the metadata held ensures that digital objects are independently understandable in the long term?
4. How well equipped do you think your research data archive is for disaster recovery in the long term?
5. What barriers to FAIR access do you think users of your collections experience?
   - Findability
   - Accessibility
   - Interoperability
   - Reusability
6. What has been the most influential change to your working practices over the time you have worked in cultural heritage? What impacts did it have?
   - What was the most influential change?
   - What was the impact of this change on your collections?
   - What was the impact of this change on your role?
   - What was the impact of this change on the training required for your role?
7. How would you describe your user community?
   - Academic researchers
   - Student researchers
   - Businesses
   - Genealogists
   - School children
   - Other members of the public
8. To what extent do you feel engaged in a public duty to share data?
9. What are the main descriptive standards your institution follows?
10. How does your institution communicate information about its collections to researchers?
    Choose between: Occasionally, Regularly, Regularly, according to a documented policy
    - Public engagement activities
    - Researcher engagement activities
    - Institution's own website
    - Infrastructure's website
    - Finding aid published as a book
    - Card catalogue in reading room

Staff responding to face-to-face enquiries on site

11. Does your institution monitor what percentage of its collections is used? If so, how?

    Choose between: Yes, No, Don't know

    How is this monitored?

12. What percentage of your institution's collections is used?

    Text-based sources

    Non-text artefacts and art works

    Digital data (also digitised resources, 'digital born' sources)

13. How does your institution handle user access requests?
14. Through which media do users access your institution's collections? [tick all that apply]

    Online open access

    FTP (file transfer protocol) networked access

    Offline media (including post)

    Reading room access

    Public (physical) access

    Other (please specify)

15. If your institution requires accreditation of users, what does this involve and how long does it take?

    What does accreditation involve?

    How long does it take? (in days and hours)

16. Has your institute been involved in other projects that asked for making data accessible (like CENDARI)?
17. What percentage of the information describing your collections is available online to the general user?
18. Is your institution part of a formal/informal infrastructure (other than EHRI) that makes your collections findable? [tick all that apply]

    At a local level

    At a national level

    At an international level

    Only metadata is shared

    With an internally searchable catalogue

    With an externally searchable catalogue

    With collections that are remotely accessible

    No

19. Does your institution provide information [metadata] about your collections to an external portal/aggregator? If so, why?
20. How relevant do you think the aggregation of information from different cultural heritage institutions is to your institution's current operation and future goals?
21. What are the three greatest challenges that prevent your institution from sharing more information?
22. What might be done to provide greater opportunities for sharing digital resources?
23. For our workshop on 'Trusted Digital Holocaust Archives for the Future" we are planning to have 3 break-out groups, can you indicate which one would interest you most at this stage?

4 Appendix B. List of available presentations

- Hidden by Not Sharing/Hidden by Sharing by Mike Priddy, DANS-KNAW
- Capability development modelling by Mike Priddy, DANS-KNAW
- Introduction to NDSA Levels by Michael Levy, USHMM
- Introduction to certification – CoreTrustSeal by Ellen Leenarts, DANS-KNAW
- Persistent Identifiers by Linda Reijnhoudt, DANS-KNAW