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

D13.3

Data management planning for long-term preservation

René van Horik
DANS-KNAW

Tonke de Jong
DANS-KNAW

Laura Brazzo
CDEC

Jessica Green
WL

Michael Levy
USHMM

Effi Neumann
YV

Annelies van Nispen
NIOD-KNAW

Frank Uiterwaal
NIOD-KNAW

Start: May 2015 [M1]
Due: October 2016 [M18]
Actual: August 2017 [M28]

EHRI is funded by the European Union

Document Information

Project URL	www.ehri-project.eu
Document URL	
Deliverable	D13.3 Data management planning for long-term preservation
Work Package	WP13 JRA6 Research data infrastructure for Holocaust material
Lead Beneficiary	01 DANS-KNAW
Relevant Milestones	
Dissemination level	Public
Contact Person	René van Horik / rene.van.horik@dans.knaw.nl / +31623297389
Abstract (for dissemination)	Data management planning (DMP) concerns the development, execution and supervision of plans, policies, programs and practices that control, protect, deliver and enhance the value of data and information assets. The certification of digital repositories (the subject of D13.4) is an important instrument to improve the quality of the data management infrastructure. Both DMP and certification of repositories are closely related and covered in this report. This deliverable contains the strategy and planning to create and disseminate expertise on DMP for the EHRI community. Input from the research data community on DMP is discussed with and assessed by policy makers from IT-savvy EHRI partners. This input consists of the FAIR data principles, the certification of repositories by means of a certification framework, the data management services provided by the EUDAT Collaborative Data Infrastructure, and the current relevant practices of the invited partners as well as the Dutch National Archives. All these will be used as building blocks for the final roadmap for a long-term access infrastructure of Holocaust digital objects (D13.2).
Management Summary	This deliverable contains a report on data management planning carried out in relation to Task 13.2 “Secure Long-term Access Infrastructure for the Preservation of Holocaust Research Objects”. A main activity concerns a workshop in which data management, repository certification principles and data management services are discussed with representatives from the EHRI consortium.

Table of Contents

1 Introduction 4
2 Context and coherence 5
2.1 Target audience 5
2.2 Strategy to achieve the goals of the task 6
2.3 Input from the research data community 6
2.3.1 FAIR data principles 7
2.3.2 Guidelines for Certification for Trustworthy Digital Repositories 8
2.3.3 Services provided by the European Research Infrastructure EUDAT 8
2.4 Towards a long-term access infrastructure 10
3 Workshop Outcomes 12
3.1 Program & Content Workshop 12
3.2 Workshop participants 12
3.3 Workshop Day 1: Data Management Planning 13
Introduction 13
DMP aspect 1: FAIR data principles 13
DMP aspect 2: Certification of TDR's 14
Information Management at the National Archives 14
DMP aspect 3: Data infrastructure services 15
3.4 Workshop day 2: Archiving, Access and Policies at the EHRI institutes 16
The Wiener Library 16
USHMM 18
NIOD 20
CDEC 20
DANS 21
3.5 Brainstorm and Discussion 22
4 Conclusion and next steps 25

1 Introduction

Deliverable 13.3 “Data management planning (DMP) for long-term preservation” is described in the DOW as follows: “*The report and corresponding workshop is for archives that wish to develop knowledge and capabilities in regard to the preservation of digital resources based on the EHRI data management policies. It will focus on data management planning, preservation policy, and access policy. It will be delivered in coordination with WP4”*.

This deliverable reports on the activities carried out in relation to Task 13.2 (Secure Long-term Access Infrastructure for the Preservation of Holocaust Research Objects). It includes details on the workshop that was held on 31 July and 1 August 2017 in The Hague. The outcomes of the workshop are included in *Chapter 3*.

The data and information assets of archives consist of a wide range of different types of digital objects. Examples are databases, text-files, websites, social media collections, digital images and multimedia files. These objects can both be digital surrogates of analogue originals (e.g. digitized photographs) or “digital born” (e.g. documentation of archival collections or digital recorded oral histories).

In a practical sense long-term preservation of data and information assets is related to questions like:

- What are the features of durable digital objects?
    - How to avoid file format obsolescence?
    - How to be sure that future users can use digital objects in a correct way?
    - How to be sure that digital objects remain authentic
- What kind of digital repository is required?
    - How to assess the quality of a digital repository?
    - How to implement a “Trusted Digital Repository”?
- What kind of services are required to provide long-term access to digital objects?
    - How to protect data objects?
    - How to share data objects?
    - How to archive data objects?

The aim of the activities carried out is to help to answer these kind of questions, taking into consideration the specific characteristics and requirements of the archives within the EHRI consortium. Within the consortium the number and variety of digital Holocaust objects fluctuates significantly, as well as the expertise and resources available to manage the digital assets.

2 Context and coherence

Task 13.2 “Secure Long-term Access Infrastructure for the Preservation of Holocaust Research Objects” is related to three deliverables:

- D13.3 Data management planning for long-term preservation
- D13.4 Trusted Digital Repository workshop
- D13.2 Long-term access infrastructure for preserving Holocaust research objects

Data management refers to the development, execution and supervision of (research) plans, policies, programs and practices that control, protect, deliver and enhance the value of data and information assets.

Data should be archived in a repository that complies to international standards and guidelines of trustworthiness: a certified ‘Trusted Digital Repository’.

Thus, the mission of the task (13.2) and deliverables (13.2, 13.3 and 13.4) can be characterized as: “Preserving digital Holocaust evidence for the future”.

The outcomes of both activities on “data management planning” (D13.3) and 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.

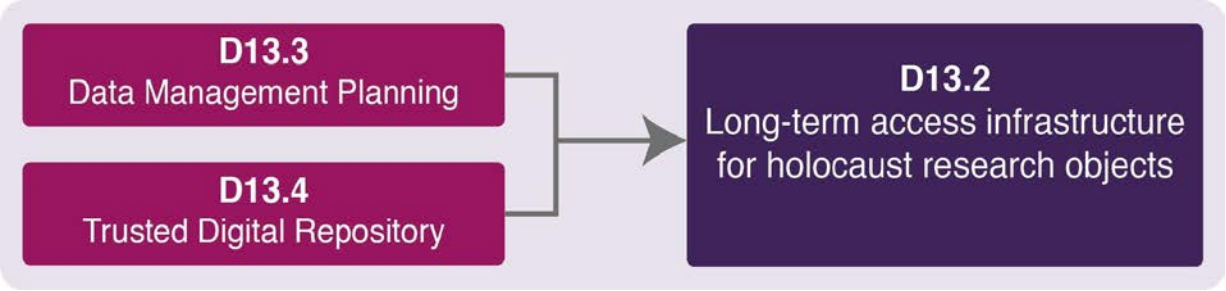


Figure 1: EHRI project Deliverables in relation to Task 13.2, “Secure Long-term Infrastructure for the Preservation of Holocaust Research Objects”.

2.1 Target audience

Data management planning, trustworthy digital archiving and issues relevant to realise a long-term access infrastructure is obviously relevant for archives in general, and not specific for archives that curate Holocaust research objects. What is specific though for Holocaust archival institutes is the subject or theme of the archival records, photographs, documentation and other objects they curate that require specific procedures (e.g. in the field privacy protection).

The activities in this task and its current deliverables are targeted at representatives of

archives within the EHRI consortium who wish to develop and extend knowledge and capabilities in regard to the management and long-term preservation of digital resources. Ideally, representatives of archives are involved in managing digital data objects or in formulating an appropriate strategy, such as policy makers.

Archives curating Holocaust objects (both inside and outside the EHRI consortium) have a couple of specific characteristics, such as:

1. Heterogeneous level of IT-savviness.
   The EHRI consortium contains 21 partners that curate documentation and/or archival material on the Holocaust, both in digital and/or analogue form. The level of sophistication concerning the application of information technology to curate digital assets ranges from basic to advanced. This determines to what extent an archive might be able to contribute to the workshop or learn from its results. Two groups of archives within the EHRI consortium are distinguished (1) "IT-Savvy" archives that have a policy on data management / long-term archiving (or intent to define a policy) and (2) archives that do not yet have a data management / long-term archiving policy. A representation of the first group will be actively involved in this task (to discuss data management planning issues).
2. A lot of the curated archival material contains personal data.
   This brings data management issues such as "privacy protection" and "user authorisation and authentication" to the forefront.
3. Analogue archival material is very vulnerable.
   The majority of the archival material originates from the period of the Second World War and the paper quality of this material is low. Digitization can be used to preserve the vulnerable originals. Issues like preservation imaging and long-term access to the images will influence data management policies applied by the archive.
4. The archive collection can contain copies of originals curated by other archives.
   Also specific for collection holding institutes curating Holocaust records is that their collections can contain copies of records. This can include copies of original archive sources, both in analogue (e.g. photocopy) and digital (e.g. digital images) format as well as copies of archival finding aids. This "copy-original" issue has specific implications for data management. The original and the copy, for instance, can be described in different ways and do often not have the same level of detail of description.

2.2 Strategy to achieve the goals of the task

The strategy chosen to ultimately arrive at a roadmap for a long-term access infrastructure for preserving Holocaust Research Objects (D13.2) consisted of three steps:

1. Principles, standards, procedures, etc. (in relation to long-term access infrastructure to preserve data) from the research data community were selected.
2. These were presented to policy makers (concerning the information architecture) in the EHRI consortium.
3. We assessed to what extent the standards etc. are relevant / of value for the curators for CHIs (in the EHRI consortium).

The workshop in summer 2017 (31 July - 1 August) played an important role in this assessment process. The summary of this workshop can be found in Chapter 3. The next section discusses the input from the research data community.

2.3 Input from the research data community

Management of digital assets is discussed in several communities, such as the cultural

heritage community, the records management community and the research data community. Although each community uses methods, standards, terminology and governance models specific to its needs, there are a number of lessons to be learned and applied to the development of a data management policy for EHRI. Since the assets managed by EHRI partners are primarily aimed at scholarly users, it is particularly useful for this work package to take a closer look at data management aspects provided by the research data community.

The following data management principles, standards and procedures from the research data community serve as input:

1. FAIR data principles, aimed at the quality of data objects.
2. Guidelines of Certification for Trustworthy Digital Repositories, aimed at the quality of repositories that curate digital objects.
3. Services provided by European Research Infrastructures (e.g. EUDAT), aimed at services that support data management.

Each of the above aspects (data, repositories, services) acts as a reference for the assessment of data management issues relevant to CHIs curating Holocaust data objects.

2.3.1 FAIR data principles[1]

The FAIR data principles are aimed at making data Findable, Accessible, Interoperable, and Reusable. Each principle is clarified below:

1. The Findable data principle.
   The findable principle concerns the assignment of persistent identifiers to digital objects, to provide rich metadata and to register the data in a searchable resource. To be findable:
   - F1. (meta)data are assigned a globally unique and persistent identifier
   - F2. data are described with rich metadata (defined by R1 below)
   - F3. metadata clearly and explicitly include the identifier of the data it describes
   - F4. (meta)data are registered or indexed in a searchable resource
2. The Accessible data principle.
   The accessible principle is related to the retrieval of objects by their identifier and the availability of metadata. To be accessible:
   - A1. (meta)data are retrievable by their identifier using a standardized communications protocol
   - A1.1 the protocol is open, free, and universally implementable
   - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
   - A2. metadata are accessible, even when the data are no longer available
3. The Interoperable data principle.
   Interoperability is realised by using formal, broadly applicable languages for knowledge representation and qualified references. To be interoperable:
   - I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
   - I2. (meta)data use vocabularies that follow FAIR principles
   - I3. (meta)data include qualified references to other (meta)data
4. The Reusable data principle.

[1] See: <https://www.force11.org/group/fairgroup/fairprinciples> [cited 8 May 2017].

The reusable principle involves the application of rich, accurate metadata, clear licenses, provenance and use of community standards. To be re-usable:

R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

2.3.2 Guidelines for Certification for Trustworthy Digital Repositories

A Trusted Digital Repository (TDR) has the mission to provide reliable, long-term access to managed digital resources to its so-called "designated community"[2]. A designated community is an identified group of potential consumers who should be able to understand a particular set of information. A number of criteria and guidelines have been established regarding the long-term sustainability of digital data.

A European Framework for Audit and Certification of Digital Repositories was set up to help organisations in obtaining appropriate certification as a trusted digital repository[3]. It has established three increasingly demanding levels of assessment: *Basic Certification*, consisting of self-assessment and external review of the criteria that are part of the Data Seal of Approval (DSA); *Extended Certification*, including the Basic Certification and additionally and externally reviewed self-assessment against a more fine-grained ISO standard (ISO 16363); and *Formal Certification*, the validation of the self-assessment through a third-party official audit based on the ISO standard.

The workshop presents the TDR assessment frameworks with an emphasis on the Data Seal of Approval (http://datasealofapproval.org)[4]. Fundamental to the DSA guidelines are five criteria, that together determine whether or not the digital research data may be qualified as sustainably archived:

- The research data can be found on the Internet.
- The research data are accessible, while taking into account relevant legislation with regard to personal information and intellectual property of the data.
- The research data are available in a usable format.
- The research data are reliable.
- The research data can be referred to.

2.3.3 Services provided by the European Research Infrastructure EUDAT

The EUDAT initiative aims at developing and supporting research data services for all scientific disciplines and that support the data lifecycle. The Humanities are an important target group for EUDAT. The EUDAT "B2Service Suite" consists of services to exchange,

[2] The concept "Designated Community" originates from: Reference Model for an Open Archival Information System (OAIS) (Consultative Committee for Space Data Systems), CCSDS 650.0-M-2, Magenta Book, June 2012 [cited 8 May 2017]. PDF format. Available from World Wide Web: <https://public.ccsds.org/pubs/650x0m2.pdf>.

[3] See: <http://www.trusteddigitalrepository.eu/Welcome.html> [cited 8 May 2017].

[4] The repository certification requirements of the DSA and the WDS (World Data System) will merge into a "Core Certification of Trustworthy Data Repositories". See: <https://www.datasealofapproval.org/en/news-and-events/news/2016/11/25/wds-and-dsa-announce-uni-ed-requirements-core-cert/> [cited 8 August 2017]. Since September 2017 the CoreTrusSeal website is active: <https://www.coretrustseal.org> [cited 22 September 2017].

synchronize, store, share, replicate, protect and find data (See: http://eudat.eu).The EUDAT services suite or Collaborative Data Infrastructure (CDI) consists of seven services. They are briefly described below.

The B2DROP service can be characterized as a personal cloud storage service. It is a secure and trusted data exchange service[5].

The next service of the EUDAT Services Suite is the B2SHARE service to store and share small-scale research data from diverse contexts[6]. The service automatically assigns persistent identifiers to records. The B2SHARE service assigns handle PIDs[7]. Depositors can document their data objects and give the data a usage license, preferably an open access license.

The third service of the EUDAT CDI is the B2SAFE service. This service allows community and department repositories to implement data management policies on research data across multiple administrative domains.

The B2STAGE service enables the movement of large amounts of data between data stores and high-performance computing resources.

The B2FIND service can be characterized as a simple, user-friendly metadata catalogue of research data collections stored in EUDAT data centres and other repositories. The service provides access to resources that are also available in the EHRI portal. This is because a repository is harvested both by the B2FIND service and the EHRI-portal[8].

B2HANDLE provides an abstraction layer between a globally unique persistent identifier and a physical location of a data object allowing researchers to reliably cite and refer in the long term.

B2ACCESS provides an easy-to-use and secure authentication and authorization platform integrated in all other services. It provides various methods of authentication through the home organisation identity provider, but also allows social IDs like Google and Facebook as well as the EUDAT ID. Managers can specify authorisation decisions in the dedicated interface.

Figure 2 gives an overview of the services of the EUDAT Collaborative Data Infrastructure (CDI). The EUDAT project ends early 2018 and this of course obstructs the realisation of a sustainable trustworthy infrastructure. EUDAT, however, will be part of the “European Open Science Cloud” (EOSC) that is planned to emerge on the basis on several European data infrastructure initiative. Concerning data management services EHRI should keep an eye on this development as it can play a role in the long-term access to EHRI databases.

[5] See: <b2drop.eudat.eu>. [cited 22 September 2017]. The EHRI Project uses B2DROP. Archives can store example data to be incorporated in the EHRI portal. Based on the example data a more robust and sustainable data integration procedure can be created.

[6] See: <b2share.eudat.eu> [cited 22 September 2017].

[7] Background information on the Handle system can be found at: <https://en.wikipedia.org/wiki/Handle_System> [cited 22 September 2017].

[8] Interviews of survivors of Sobibor concentration camp, for instance, are available by the B2FIND service as well as by the EHRI portal. Both metadata records contain a persistent identifier to the data objects and this enables a trustworthy retrieval of the data.

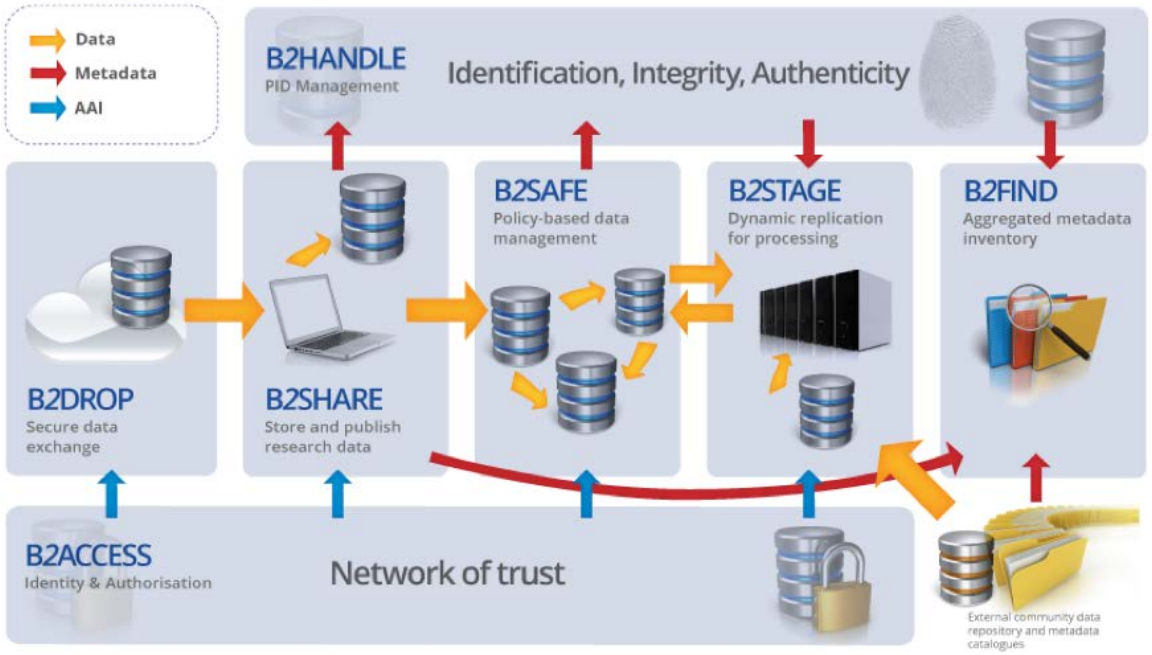


Figure 2: The services of the EUDAT Collaborative Data Infrastructure

2.4 Towards a long-term access infrastructure

The challenge in task 13.2, Secure Long-term Access Infrastructure for the Preservation of Holocaust Research Objects, is to collect, formulate and disseminate knowledge and expertise on the management and long-term preservation of digital objects of value to the 20 archives in the consortium and beyond. The sources of information come both from the EHRI partners and from the community that develops and maintains services in the field of research data management. This process is described below:

1. The collection phase

The main idea is to assess services, systems and procedures of the EHRI partners to curate digital objects in relation to reference models and data infrastructure services that have their origin in the research data management field. As not all EHRI partners have the same capability with regard to data management and long-term preservation, a consultation is carried out. An important part of the collection phase is the workshop on data management planning and becoming a trusted digital repository.

2. The formulation phase

State of the art information on research data management and long-term archiving is provided in the present report, presented in Chapter 3 as the outcomes of the workshop. This report is available to the EHRI consortium.

3. The dissemination phase

The main focus of the dissemination phase will be the formulation of a long-term access

infrastructure for preserving digital Holocaust objects that is scheduled to be delivered near the end of the EHRI project. The long-term access infrastructure for preserving Holocaust research objects is the subject of Deliverable 13.2. The formulation of this infrastructure will be based on input from the research data community. See Figure 1.

3 Workshop Outcomes

This chapter contains a report of an EHRI workshop on data management organised on 31 July and 1 August 2017.

3.1 Program & Content Workshop

The workshop consisted of two days: the first day's topic was Data Management Planning, whereas the second day's topic was Long Term Access to Holocaust Data. Three presentations were given the first day, according to the topics that are described in Section 2.3: FAIR Principles (by Peter Doorn), Certification of Trusted Digital Repository (by Heiko Tjalsma), and Data Infrastructure Services (by René van Horik). In addition, a presentation on digital information management at the Dutch National Archives was given by Margriet van Gorsel. For the second day, participants were asked to prepare slides about their view on Archiving, Access, and Policies. A summary of these discussions can be found in Section 3.4.

Figure 3: Program of the Workshop

3.2 Workshop participants

Within the EHRI consortium partners were approached that have experience with managing data assets, e.g. because they operate information management systems. Another activity

concerns data management policies. The workshop participants are thus able to evaluate the value data management policy issues have for the whole EHRI consortium and beyond. Six EHRI partners were represented: CDEC, WL, DANS, USHMM,fc Yad Vashem and NIOD. The following persons / EHRI partners contributed to the workshop:

Laura Brazzo	Fondazione Centro di Documentazione Ebraica Contemporanea (CDEC)
Jessica Green	The Wiener Library (WL)
René van Horik	Data Archiving and Networked Services (DANS)
Tonke de Jong	Data Archiving and Networked Services (DANS)
Michael Levy	United States Holocaust Memorial Museum (USHMM)
Effi Neumann	Yad Vashem (YV)
Annelies van Nispen	NIOD Institute for War, Holocaust and Genocide Studies (NIOD)
Frank Uiterwaal	NIOD Institute for War, Holocaust and Genocide Studies (NIOD)

3.3 Workshop Day 1: Data Management Planning

Introduction

The workshop started with an introduction in which the context and strategy of the activities related to this workshop were described. This introduction to a large extent contains the information as given in chapter 2. A couple of remarks were given by the workshop participants:

1. Data management planning (DMP) in first instance is directed towards the researchers who elaborate on how they deal with data they use and create. In EHRI the data provider perspective is more prominent than this data user perspective. In EHRI a great deal of historians are active who work in a traditional way. DMP in EHRI must be directed on the archives rather than on the users.
2. Also data-reuse and the management of licenses should be added to the topics (see page 6)
3. Participants have experienced that certification of repositories can be very expensive. It can, however, play an important role in educating / training the people in the organisation on policies with regard to long-term access to digital assets.

DMP aspect 1: FAIR data principles

Presenter: Peter Doorn (Data Archiving and Networked Services (DANS))

Title of presentation: “FAIR Data Assessment of Datasets in Trusted Digital Repositories”

Link to slides: https://b2drop.eudat.eu/s/atw1lonNKULP9yp

Remarks and comments by the workshop participants:

1. The assignment and management of persistent identifiers turns out to be a very important component of data that is “FAIR”. Several practical questions concerning this were exchanged. E.g. on how to get PIDs for objects. A solution of getting PIDs for publications is to become a member of DataCite[9] (or join a national member of DataCite). The importance of PIDs for the EHRI infrastructure was confirmed a couple of times.
2. Although the complete implementation of the FAIR principles is considered as too

[9] See: <https://www.datacite.org> [cited 5 October 2017].

much for the EHRI consortium as a whole, the principles can still be considered as good guidelines. The FAIR data assessment tool (currently as prototype) might be relevant for EHRI at a later stage.
3. Some EHRI partners manage / create Linked Open Data. This type of data is by definition of high quality in terms of FAIR criteria.
4. As privacy protection / license issues are very important in the EHRI consortium several FAIR criteria (e.g. Findable) will not be fully supported.
5. For EHRI the "data scope" is rather on archival collections than on research data sets. This perspective is probably new in the FAIR data community and EHRI might consider to put this perspective more to the forefront.

DMP aspect 2: Certification of TDR's

Presenter: Heiko Tjalsma (Data Archiving and Networked Services (DANS))

Title of presentation "Certification of TDRs"

Link to slides: https://b2drop.eudat.eu/s/Lsdi8LMuFSpPUMr

Remarks and comments by the workshop participants:

1. Outsourcing of services (e.g. archival storage) does not influence the certification situation for an archive. The service provider now has to "prove" that the certification requirements are met. So the certification guidelines might have to be assessed over several service providers.
2. It might be the case that not all aspects of the certification are public (e.g. for security reasons). The reviewer, obviously, must have access to all relevant information in order to assess the quality of the repository. This issue is certainly applicable in EHRI.
3. The certification requirements are of value in a "self-assessment" process. Repositories can evaluate several aspects of the organisation of the data without any consequence.
4. EHRI is in the process to become an ERIC. In other ERICs (e.g. CLARIN) the certification of repositories plays a role to improve the quality of the repositories involved. And still the majority of the repositories in CLARIN do not have a certification. The way other ERICs engage with repository certification should be taken into consideration by EHRI.
5. The EHRI Content Provider Agreement (CPA) was discussed in relation to some certification rules. The CPA is not signed by all EHRI partners and also the formulation of the agreement has undergone several versions. An important issue is the ownership of the data, especially after the transfer of the data from the archive to the EHRI portal. This issue is not unambiguously settled in the EHRI project.
6. The assessment of certification guidelines is not a trivial activity. Carrying out a full certification process is for most EHRI partners not possible. Within EHRI in first instance awareness raising on the importance to assess features of the data management infrastructure is of value as it will provide an overview of aspects that are relevant for the management of digital objects.

Information Management at the National Archives

Presenter: Margriet van Gorsel (Dutch National Archives)

Title of presentation: "Digital Information Management"

Link to slides: https://b2drop.eudat.eu/s/xCXBzYytcpqpgOt

Remarks and comments by the workshop participants:

1. The design and management of an "information chain" as presented is seen as

relevant and has some resemblance with the EHRI project (several data providers and one service provider). Process based on standards, but exceptions do occur.
2. The National Archives repositories have received the dataseal of approval (replaced by the “CoreTrustSeal” in the future).

DMP aspect 3: Data infrastructure services

Presenter: René van Horik (Data Archiving and Networked Services (DANS))

Title of presentation: “Data infrastructure services of the EUDAT CDI”

Link to slides: https://b2drop.eudat.eu/s/OclvzCfpts7nAgJ

Remarks and comments by the workshop participants:

1. It seems several tools and services are available for data management (e.g. the ones provided by the EUDAT CDI). The threshold to use some of them is quite high. One must be trained. Also the sustainability of the service is an issue: what will happen with the services once the EUDAT project is over? A follow-up initiative has started (European Open Science Cloud - EOSC) but it is not clear what its potential value is for EHRI.
2. It is important to make a distinction between a data infrastructure and a research infrastructure. Some services of EUDAT are of value. B2DROP is already used in the EHRI project as exchange service for sample data sets.
3. The discussion of the store, share, archive, etc. services of EUDAT gives also insight in the EHRI information architecture. E.g. on issues such as the updating of records, the long-term perspective of the EHRI portal. We should start with an inventory of what is needed, prioritise and then plan the next steps.
4. Services that deal with privacy protection and authorized access (identity provision by B2ACCESS) is relevant for EHRI.

3.4 Workshop day 2: Archiving, Access and Policies at the EHRI institutes

In relation to long-term access to digital objects three aspects were discussed: (1) archiving, (2) access and (3) policies. Each workshop participant provided input on each aspect.

The "archiving" aspect concerns issues related to the storage of digital objects by organisations that curate digital Holocaust objects. Questions related to this aspect are:

1. What kind of digital objects do you manage?
2. Where do you store these objects?
3. How do you monitor the quality of the digital objects?
4. Details on the information systems you use

The "access" aspect concerns issues related to the management of access to digital objects. Questions related to this aspect are:

1. How do you manage the access to the digital objects?
2. How do you protect the objects?
3. Details on licenses / legal issues

The "policies" aspect concerns issues related to the formulation of policies to provide long-term access to digital objects. Questions related to this aspect are:

1. Which stakeholders are involved in managing the digital collection?
2. Details on the business model to manage digital assets
3. With whom do you cooperate?
4. How do you check / monitor the quality of your assets?

Each workshop participant (representing an organisation that manages digital Holocaust objects) provided input on the long-term access aspects. See below.

The Wiener Library

(Jessica Green)

Concerning <u>archiving</u>:

The Wiener Library is currently undertaking an ambitious digital transformation, as we move towards more sustainable and efficient processes for creating, managing, preserving and accessing digital records.

Over the last few decades, the Library's digital holdings have grown to include approximately 18 TB of digital material, including audio and video files, databases, photographs, document scans, and more. Some of this material has been digitised in-house or by third-parties from their original analogue or paper formats, while a growing percentage of it is born-digital. In addition to being a digital copy holder of the International Tracing Service (ITS), the Library has accepted a number of large digital collections over the last year, including, most notably, a digital copy of the UN War Crimes Commission Archive. The Library expects a growth of digital collections donations over the next few years in a range of different formats, including large databases, audiovisual collections, and sets of PDFs/TIFF files with a corresponding Excel spreadsheet of metadata. In order to support this type of growth, the Library is taking steps towards improving and standardising methods for accessioning, cataloguing, making accessible, and preserving this digital collections.

One of the Library's short-term goals is to finalise a digital preservation policy to ensure long-term preservation of digitised materials, born-digital material, and large digital collections. There are currently a range of different file formats stored in a number of different locations on our shared servers, individual email accounts, a few hard drives and digital tape. The

Library is currently working towards gathering all scans of collection items, born-digital material, and digital collections into a separate shared drive. Files in this drive will then be organised by collection type and renamed according to their collection ID number. This will make linking between catalogue records and their digital objects more straightforward. In addition, the files will be moved to a dedicated server in a secure data centre this October.

In order to ensure files are accessible for the long term and to mitigate the effects of obsolescence, work is being done to forward-migrate existing files to file formats that are considered better for long-term preservation. This includes converting JPEGs to TIFFs and VLC media files to MP4s. The digital preservation policy will also cover future forward-migration of preservation files, multiple methods for backups, and running checksums. The Library recognises that digital preservation is an ongoing activity that is never complete; taking these steps and continuously reviewing/updating our policy will help the Library bring itself in line with best practice for long-term digital preservation. Since the Library can learn from others and help others learn, we are using guidelines from the Digital Preservation Coalition (DPC) and other professional bodies to inform our decisions and plan to share our policies and procedures with EHRI and other interested bodies. The end goal is for digital preservation practices to be as embedded into the Library's daily work as the physical preservation of its collections.

Concerning access:

The Wiener Library currently provides access to our digital holdings through a range of different information systems and is working to improve the searching, retrieval, and display of these digital objects for its staff and users. Access to ITS is provided to researchers on two dedicated terminals in our Reading Room using the OusArchiv database. Before using these terminals, researchers have to attend a training instruction and make appointments in advance. Our ITS Archive Researcher is on hand to help people use and search this database, as well as to conduct research for people unable to visit the Library themselves.

Three other dedicated terminals provide researchers with read-only access to some of our digital collections (including the UNWCC archive), as well as video and audio recordings via segmented drives on our server. To prevent people from downloading files, the Library has disabled internet access and USB ports on the Reading Room terminals that have access to this segmented drive. A small number of photographs are available to view on our online catalogue, Soutron. Currently the only way to attach images to our catalogue records is to upload thumbnail versions of the images into the catalogue database directly. This is an unsustainable model for providing online access to our digital materials, since the more files, and the larger the files, the slower the entire catalogue becomes.

In order to provide more materials online and at our dedicated terminals, the Library is exploring the implementation of a digital viewer that would link high-res digitised images of documents and photographs to their relevant catalogue record. This would allow for a richer user experience, including zooming in and out of a photograph, as well as help to comply with copyright and data protection regulations by restricting access to digital files based on IP address or logins.

The main system Library staff use to access our digital objects currently is Adobe Bridge. As materials were digitised, descriptive information was added directly into the embedded metadata of the images. The main method for finding a digital object is to search Bridge for keywords and search terms that might be included in this embedded metadata. The Photo Archivist is currently working on turning this embedded metadata into ISAD(g) compliant catalogue records in Soutron, so that our users can access, search and find these images as well.

All Library staff have recently undergone a mandatory copyright training session. Our Collections Team is evaluating our policies around granting licenses for use of digital images

as part of the follow-up to this. In developing our access policies, the Library aims to make accessible as many of our digital objects as possible, taking into account financial, technical, and copyright/data protection restrictions. This is based on the belief that digital users are just as valid and important as physical users to the Library; as such, they should have access to rich material and metadata online as well as onsite.

Concerning policies:

The Wiener Library is currently developing its policies around management and long-term access to digital objects. The requirements for digital objects are just as important as for physical objects, but more challenging to embed into our daily practices due to technological and budgetary constraints, as well as a deficit of digital skills. Although this shift towards managing and providing access to more digital objects brings with it certain challenges, it is something that benefits our staff and users greatly and is becoming more and more expected over time. Policies are being developed with both the needs of our users and staff in mind, as well as keeping an eye towards future trends and best practices. Sharing policies and business models for managing digital assets across like organisations would be helpful to this process.

USHMM

(Michael Levy)

Concerning archiving:

USHMM digital archiving practice

- Intensive digitization began 2006 with magnetic media, oral history. International copy archive projects more and more digital. Digitization of microfilm, historical film, paper archives, photos
- 50 million+ digital files, growing
- ~800 TB, growing
- State-of-the-art NAS w/erasure coding
- Tape backup. Offsite secure storage of backup
- Inventories, checksums. Currently using open source tools to crawl and store. md5 and sha1 are utilized. The process takes many months
- Compare recalculated inventory and checksum to store checksums every ~24 months
- Currently engaged in RFP process intended to supply commercial digital preservation platform, to automate digital preservation activities. Follow OAIS model ISO 14721:2012
- Web archiving of our own institution's digital output
- Informal relationship with US Government Publication Office. GPO uses the nonprofit Internet Archive's “Archive-It” service for archiving web and social media output

What are the requirements for Holocaust archives in general?

- Desirable for every Holocaust archive to begin to engage with digital preservation activities
- Recognize that digital assets with long term value require active preservation measures

Suggestions for implementation:

- Training, workshops, other educational processes

- Blog posts
- Other educational efforts
- Start with small steps: “Do something”
- Self-assessment
- “NDSA Levels of Digital Preservation” may be a good place to start for many organizations and practitioners
    - Storage and Geographic Location (Levels 1-4)
      File Fixity and Data Integrity (Levels 1 to 4)
    - Information Security (Levels 1 to 4)
    - Metadata (Levels 1 to 4)
    - File Formats (Levels 1 to 4)

Concerning <u>access</u>:

- Collections Search
    - Provide access to all catalog records and to “use copies” of media –- on web if possible, locally-only if not allowed; certain materials are highly restricted or embargoed
    - Ensure everything on web is searchable by web crawlers e.g. Google
    - User studies, improve UX
- Permanent URLs or HTTP 301 redirects (so old links to records still work)
    - Handles –- contracts with users for a permanent place on the web (we hope to implement at USHMM eventually)
    - Constant improvement
        - Technology changes. Obsolescence. Security standards. User expectations continually increase

What are the requirements for Holocaust archives in general?

- Broadest allowable access is desirable: increased access leads to broader use and interest, strengthens the field of Holocaust research overall, and thus leads to continued or increased support

Suggestions for implementation

- Continued support, strengthening and broadening the EHRI Portal and digital tools

Concerning <u>policies</u>:

USHMM Digital Preservation and Access Policies:

- Competing priorities include:
    - Quality (e.g. file formats, resolutions, bit rates)
    - Quantity / funding
    - Preservation, redundancy, resources
    - Every institution is unique
    - Quantity: File sizes affect preservation operations (e.g.time) & redundancy
    - USHMM distinguishes between “asset” and “instance” to help balance resource priorities
        - Asset = only trustworthy copy (e.g. magnetic media or fragile originals, born digital); irreplaceable, therefore warrants utmost attention
        - Instance = surrogate of a durable physical item, or a derivative; replaceable at a cost

What are the requirements for Holocaust archives in general?

- Institutions are each unique and must develop digitization policies according to their institutional responsibilities and requirements

NIOD

(Annelies van Nispen & Frank Uiterwaal)

Concerning archiving:

- NIOD has digitized archives, digitized photo's, research databases, audiovisual material and manages this with Dutch partners;
- The material is stored at KNAW's infrastructure or at dedicated partners as DANS or Netherlands Institute for Sound and Vision;
- The partners are specialized in Digital Preservation and we trust them. The KNAW infrastructure needs to be of high quality. (SLA: backup, disaster recovery);
- Specifications on the quality of the objects are project-based (Metamorfoze Digitisation Programme)

What are the requirements for Holocaust archives in general?: Hybrid archives with multiple sorts of digital objects. One-solution-fits-all approach will not work.

Concerning access:

- Different digital objects are managed by specialized partners, eg. Research data/Testimonials (DANS); Audiovisual material (NIBG);
- External Access is controlled by:
    - Archiefwet;
    - Privacy Protection Laws;
    - Copyright.
    - Internal access to the storage of digital objects is controlled (based on someone's role within the organisation);
- Persistent Identifiers need to be implemented within a few years.

CDEC

(Laura Brazzo)

CDEC has started a project to integrate its archival description databases, research databases, digitized (and digital born) material (papers, photographs, tape and video recordings) objects, as well as its library catalog. The goals of the project are to overcome, fragmented information, to avoid duplicated resources and to implement lacking measures for the long term preservation of data. The project can be summarized as "integration", "interoperability" and "preservation".

The digital asset management system used is the Linked Data platform "openDams/Bygle" (created by the company "regesta.exe"), based on W3C recommendations. It works as a data integration layer allowing the integration of heterogeneous data sources. Data from research databases and the library's catalog have been converted into and are imported in the system[10].

[10] The data is formatted in the RDF (Resource Description Framework) standard. The locations

Archival descriptions and digitized material are managed by xDams, an OS XML web-based platform, that is integrated inopenDams. It works as the main data provider of openDams. The metadata are stored in xDams in native XML databases. To encode the metadata in XML format, the EAD data model is used for the description of the archival resources. The EAC-CPF data model is used for the description of authority records.

Authority files (managed through openDams) are Uniform Resource Identifiers (URIs) linked to the archival descriptions by a lookup function. Linked resources are encoded in the related XML EAD files

Digitized material (papers, photographs, audios and videos) are attached to the appropriate archival description records. The xDams Platform enables the creation of an XML repository containing the metadata needed to the description of all the digital attachments referenced in the databases. To ensure consistency and validity to the referenced digital objects, the METS[11] standard is adopted. EAD and EAC-CPF are used in conjunction with METS in descriptive and administrative metadata contexts. The next, scheduled, step forward is the transforming of archival descriptions currently in XML, into the RDF format (using the OAD - Archival Description Ontology as data model) in order to support interoperability as well as a sustainability of data.

Compared to the beginning of the project (in 2013), significant progress has been made. Firstly we have set up a basic architecture for the data management. A lot of work, though, still needs to be done in order to not-to-lose resources in the coming years. One of this is definitely the upgrade of the Quest online journal[12] with a better set of descriptive metadata and the attribution of DOI as persistent identifiers of each published article. High resolution master images of the digitized materials are stored in the CDEC storage system. An agreement between CDEC and UCEI (Union of the Italian Jewish Communities) about the keeping of the backup copy in Rome is under review.

High resolution copies (TIFF, MOV, WAV) are stored in a “Networked Attached Storage” (NAS) storage (RAID5 Hot Spare, Redundant, Access controlled) localized at the CDEC Foundation. Data, metadata and low-resolution copies of digitized materials attached to the archival descriptions (JPEG, Mp3, mp4) are stored directly by Regesta on a remote Virtual Machine localized off-site.

DANS

(René van Horik)

Concerning <u>archiving</u>:

(expresses as so-called URIs) are imported in openDams/Bygle and published in a triplestore accessible through an intranet Sparql endpoint. Every new item created by openDams is natively a URI (see for example: <http://dati.cdec.it/lod/shoah/person/251> or <http://dati.cdec.it/lod/shoah/place/Milano> [cited 5 October 2017].

[11] Metadata Encoding and Transcription Standard. See: <http://www.loc.gov/standards/mets/> [cited 4 October 2017].

[12] Quest. Issues in Contemporary Jewish History. Published by CDEC. See: <www.quest-cdecjournal.it> [cited 5 October 2017].

The collection "World War II" has been created by DANS[13]. The collection consists of datasets that have been deposited by researchers. Keywords provided by the data depositors have been used to create the thematic collection.

Most of the datasets contain oral history material. Many of these datasets are the result of the national program "Heritage of the War" (2005-2009) that improved access to the large variety of WOII collections and thereby contributed to the advance of knowledge of WOII. The metadata is open accessible and harvestable (by OAI-PMH). The datasets are harvested by several organisations and enriched (for instance by the B2FIND service of EUDAT (see section 2.3.3)). Authentication of the datasets is realised by its persistent identifier and metadata. If applicable an additional streaming service is available so the interview can be seen[14]. The average size of an interview is 2 GB. For the archival storage of the data a third party is used.

Concerning access:

The access to data objects is managed by an information system (http://easy.dans.knaw.nl) and is based on a user license. Three types of access can be distinguished. In the first place "Open Access" in two versions. One version requires registration of the user whereas the other version does not. The second version of access is "Restricted Access" and this implies that the user has to ask the owner of the data permission to get access. This request for access can be made via the information system. After the permission is granted the user can get access to the data. The third type of access is classified as "Other Access" and can have several appearances, e.g an embargo on the access special conditions for access.

Secure storage of the data sets is based on a data management policy in which third party services (e.g. data centre) are involved. All metadata of the datasets can be harvested.

Concerning <u>policies</u>:

The policy concerning data management at DANS can be characterized as ""Open if possible, closed if necessary". The repository has a DSA seal as well as a NESTOR seal[15]. Concerning data formats, DANS has formulated a "preferred format policy". Preferred formats are file formats of which DANS is confident that they will offer the best long-term guarantees in terms of usability, accessibility and sustainability. Depositing research data in preferred formats will always be accepted by DANS. Acceptable formats are file formats that are widely used in addition to the preferred formats, and which will be moderately to reasonably usable, accessible and robust in the long term. DANS favours the use of preferred formats, but acceptable formats will in most cases also be allowed[16].

3.5 Brainstorm and Discussion

The workshop participants decided to create a mindmap to formulate the main outcomes of the workshop[17]. The main topic of the discussion concerned the value of the the three data

[13] This collection can be found at: <https://easy.dans.knaw.nl/ui/browse>, select the collection "World War II" in the "Refine" screen at the right of the window. [cited 4 October 2017].

[14] See for instance interview with survivor of camp Buchenwald: <https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:60686/tab/6> [cited 3 October 2017].

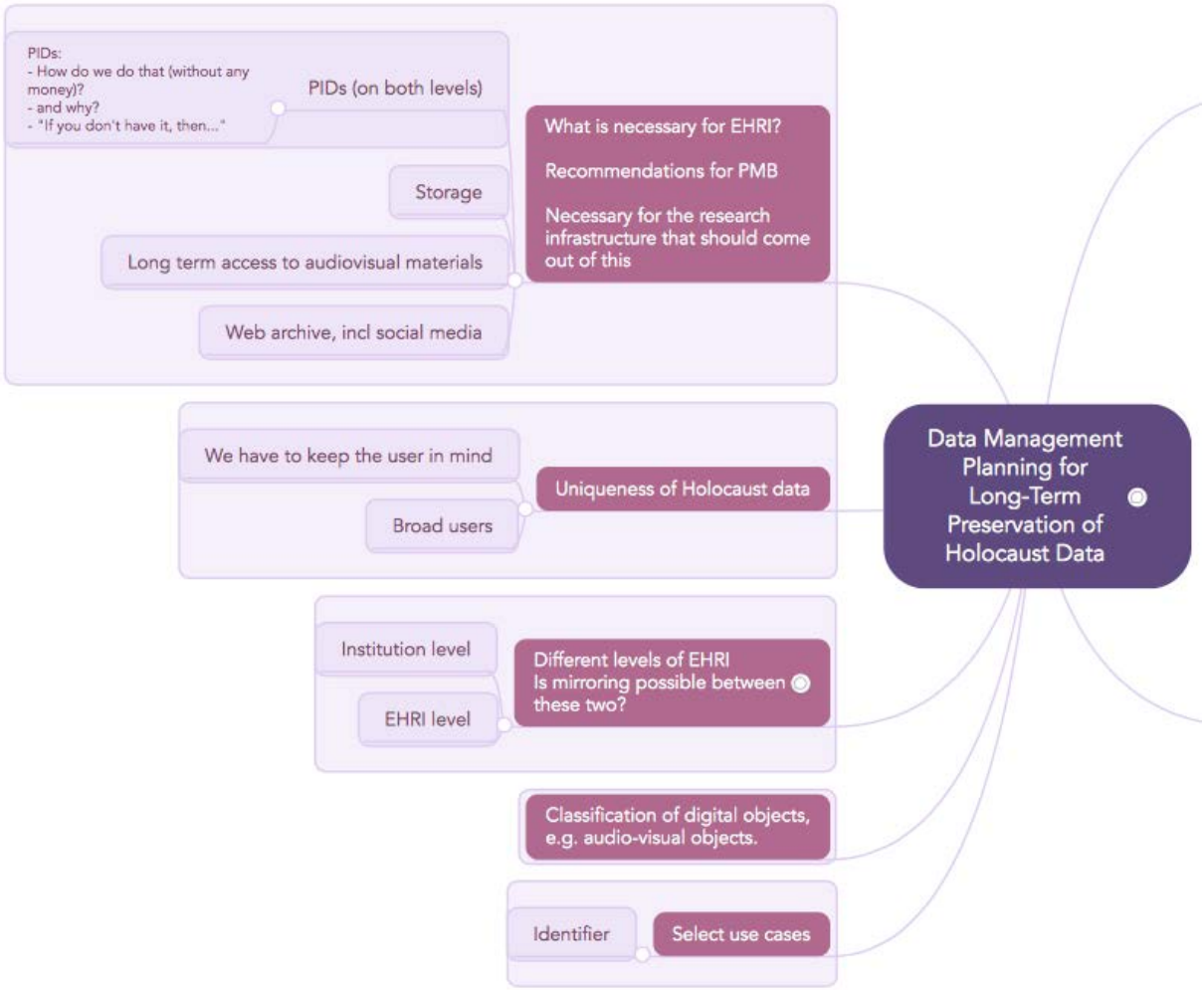
[15] Details on the certification policy of DANS can be found at: <https://dans.knaw.nl/en/about/organisation-and-policy/certification> [cited 3 October 2017].

[16] More details on the preferred format policy can be found at: <https://dans.knaw.nl/en/deposit/information-about-depositing-data/file-formats/file-formats> [cited 3 October 2017].

[17] For the creation of the mindmap Mindmeister was used, see: <http://www.minmeister.com> [cited

management planning aspects presented at the first day (FAIR data principles, repository certification and data management services of the EUDAT CDI) for the formulation of a data management policy for the long-term preservation of Holocaust data. The output of the discussion will be used to work on a long-term access infrastructure (Deliverable 13.2) to be delivered in early 2019.

The mindmap created can be found in Figure 4.



12 September 2017].

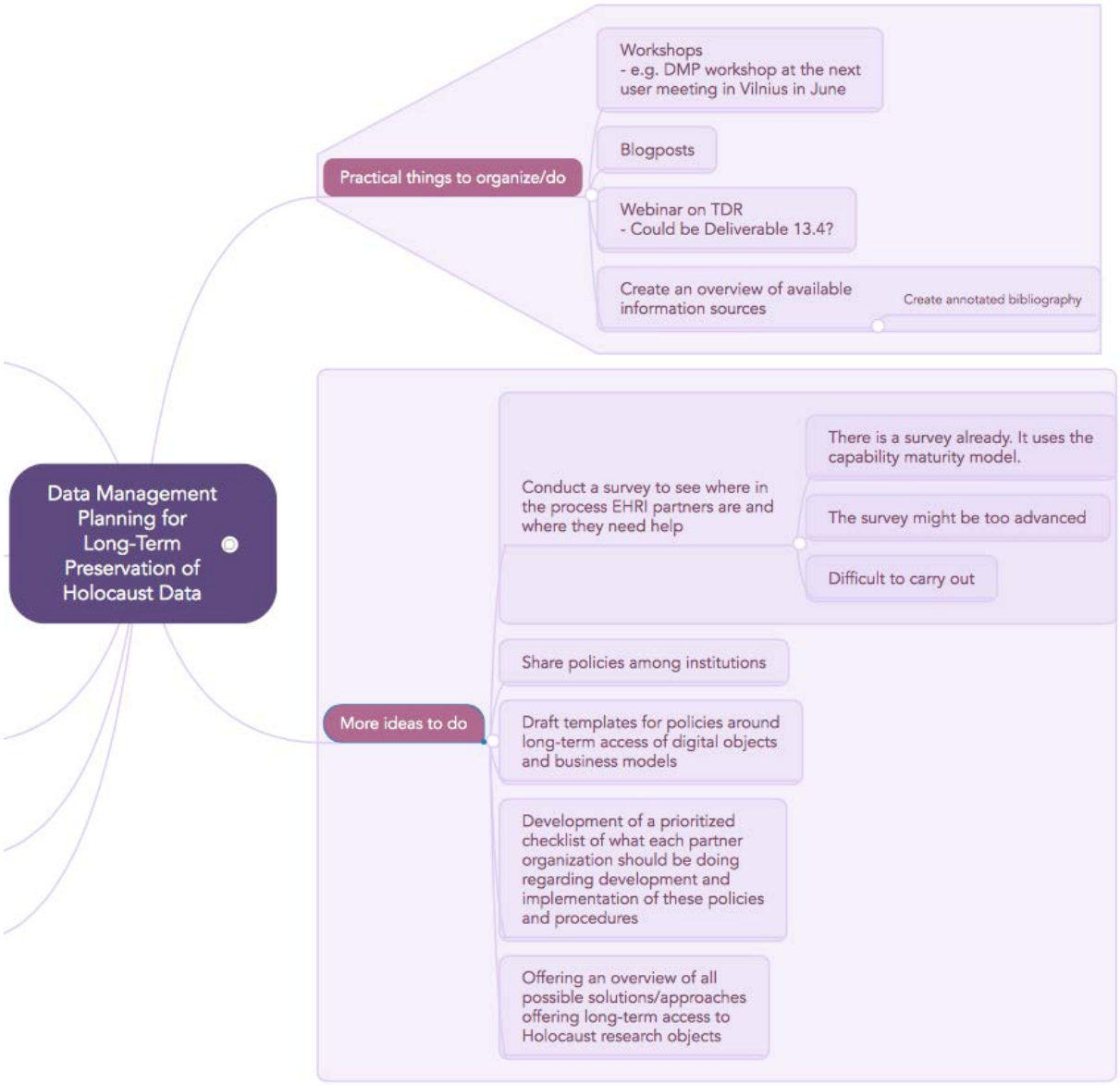


Figure 4: Mindmap of the discussion

4 Conclusion and next steps

All in all, a fruitful workshop was held on data management planning for long-term preservation of Holocaust objects, attended by representatives of organisations in the EHRI consortium that manage and curate digital objects. The first day's goal was to get all participants acquainted with data management building blocks from the research perspective. The second day's goal was to discuss current data management practices at the different institutes.

In addition, the roadmap for a long-term access infrastructure (LTA) was discussed. The participants of the workshop agreed on a structure for the LTA. The roadmap consists of three parts: archiving of data, access to data and policies to implement the LTA. Each part is elaborated on below. The activities carried out to establish EHRI as a permanent legal entity by becoming an ERIC obviously will play an important role in the formulation of the LTA. The activities carried out will be complementary to the work done in relation to the development of the ERIC. The LTA for preserving Holocaust Research Objects will be published as Deliverable 13.2 and is foreseen for month 44 in the EHRI project.

Concerning "archiving" the work on the roadmap for a LTA consists of five aspects:

1. Classification of the data that has to be curated by the LTA. It also concerns the selection of the data objects that have to be archived for the long-term.
2. Alternatives for the archival storage of data objects.
3. An assessment of available data formats for the data objects with respect to its durability.
4. The role of persistent identifiers in the LTA. Which identifier scheme can be used in which situation as well as practical implementation issues.
5. The monitoring of the LTA concerns the periodic evaluation of the quality of the components of the LTA and possible procedures to keep the archiving services of the LTA uptodate.

Concerning the "access" to data objects the LTA roadmap will pay attention to three issues:

1. Legal issues, such as licensing models and ways to protect sensitive Holocaust archives and personal information according to the latest European (and American for USHMM/Israeli for Yad Vashem) legislation.
2. Secured access is the next topic. This consists of AAI (Authentication and Authorisation Infrastructure) services such as the identity management of users of data objects.
3. Search engine optimization (SEO) that uses web-analytics (statistics) to improve the access to data objects.

Concerning "data management policies" for the LTA roadmap five aspects will be covered.

1. Security policies. Coverage of aspects such as data protection, data access conditions and management of user data.
2. Cooperation models. This activity will evaluate how the current EHRI Content Provider Agreement can be used to formalise cooperation with stakeholders.
3. Attention is paid to business models that facilitate the long-term operation of the LTA.
4. The role of certification frameworks is part of the quality control of the LTA.
5. Long-term viability of the LTA is the last aspect of data management policies for the LTA roadmap.

The workshop participants discussed the process to define the LTA roadmap and came to a

proposal for a timeline. The proposed activities are:

- The certification of data repositories is considered an important component of the LTA roadmap. This workshop has paid attention to the topic Trusted Digital Repository by discussing certification frameworks, such as the “Data Seal of Approval” (that is succeeded by the “Core Trust Seal”). The topic Trusted Digital Repository is the subject of deliverable D13.4. We consider to organise a workshop at the EHRI general partner meeting in June 2018 so all partners in EHRI can be informed on aspects of repository certification.
- Based on the outcomes of the DMP and TDR workshop the LTA roadmap will be defined and created in December 2018 as deliverable D13.2. All participants of the workshop have confirmed that they will, given available resources, contribute to the content of the roadmap.