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

Deliverable 13.2

Long-term access Infrastructure for preserving Holocaust research objects

René van Horik
DANS-KNAW

Ellen Leenarts
DANS-KNAW

Mike Priddy
DANS-KNAW

Michael Levy
USHMM

Jessica Knight
USHMM

Start: M1
Due: M44
Actual: M48

EHRI is funded by the European Union

Document Information

Project URL	www.ehri-project.eu
Document URL	n/a
Deliverable	D13.2 Long-term Access infrastructure for preserving Holocaust research objects
Work Package	WP13
Lead Beneficiary	DANS-KNAW
Relevant Milestones	MS4
Dissemination level	Public
Contact Person	Ellen Leenarts ellen.leenarts@dans.knaw.nl
Abstract (for dissemination)	This deliverable contains a description of components of a long-term access infrastructure for preserving Holocaust research objects. Key features of a long-term access infrastructure are given as well as components of a roadmap towards a long-term access infrastructure.
Management Summary	This report aims to provide essential information in relation to the realisation of a long-term access infrastructure for preserving Holocaust research objects. The introduction chapter consists of five parts. First, some general trends and principles are introduced that can act as a basis for a sustainable infrastructure for the future. The second section describes the context of this deliverable in relation to the EHRI project and its related deliverables. Next, main outcomes of a survey on the way archives cope with the management of digital assets are described. This provides important background information on the way archives deal with digital data objects. The fourth part of the introduction pays attention to the OAIS Reference Model, that defines archival terms in an unambiguous way and describes functions of archival systems. The FAIR data principles ("Findable, Accessible, Interoperable, Reusable") are introduced in the fifth part of this introduction. The FAIR data principles can be used to assess quality features of digital data objects and thus are relevant for a long-term access infrastructure for preserving research data objects. The second chapter consists of a description of prominent features of a long-term access infrastructure. They are the durability of the file format, the use of persistent identifiers, the storage of the data in a certified data repository, the quality of the documentation (or

	metadata), the usage licenses of the Information Objects and data protection and secured access to information. The third chapter covers aspects of the roadmap to create, operate and maintain a long-term access infrastructure for digital objects. A capability maturity assessment helps to find out how the realisation of a long-term access infrastructure can be organised in terms of required skills and competencies. Data management planning is essential to keep data understandable and organised in the long run. Data protection and legal issues is the third aspect of the roadmap, followed by attention for archival data storage and access to data objects. The last part of this chapter is a description of the digital preservation policies and practices of the United States Holocaust Memorial Museum (USHMM).

Table of Contents

Preface

This report is aimed at organisations that would like to create and maintain sustainable digital objects, such as digitized historical records, with an emphasis on organisations that curate digital objects on the Holocaust. The main goal of this report is to provide essential information in relation to the realisation of a long-term access infrastructure for preserving Holocaust research objects. The first chapter provides background and context information on digital preservation and long-term access to digital objects. The second chapter consists of a description of prominent features of a long-term access infrastructure and the third chapter covers aspects of the roadmap to create, operate and maintain a long-term access infrastructure for digital objects.

1 Introduction and background

This report is aimed at organisations that would like to create and maintain sustainable digital objects, such as digitized historical records, with an emphasis on organisations that curate digital objects on the Holocaust. Ultimately it is the aim of a long-term access infrastructure to provide the users of the archives (also called the "Designated Communities" a term explained further on) with optimal long-term access to its holdings.

The introduction consists of five parts. First, based on insights from the 1990s some general trends and principles are introduced that can act as a basis for a sustainable infrastructure for the future. The second section describes the context of this deliverable in relation to the EHRI project and its related deliverables. Next we describe the main outcomes of a survey on the way archives cope with the management of digital assets. This provides important background information on the way archives deal with digital data objects. The fourth part of the introduction pays attention to the OAIS Reference Model, that defines archival terms in an unambiguous way and describes functions of archival systems. The OAIS Reference Model provides guidance for the implementation of a long-term access infrastructure for preserving research data objects. The FAIR data principles ("Findable, Accessible, Interoperable, Reusable") are introduced in the fifth part of this introduction. The FAIR data principles can be used to assess quality features of digital data objects and thus are relevant for a long-term access infrastructure for preserving research data objects.

1.1 Back to the future

What must be done today to ensure that in about 20 year digital data created today are findable and usable in an authentic way? For this we need a long-term access infrastructure for preserving digital objects. Which standards, guidelines and principles are essential in this respect? The answer to this question is the main subject of this report. We cannot predict the future, but we can look back and see which past directions with respect to long-term access infrastructures still have value today can be carried forward into the distant future, say the year 2040.

More than twenty years ago, in 1996, "preserving digital information" was published by the Council on Library and Information Resources - CLIR (Waters, 1996). CLIR describes itself as "*an independent, non-profit organization that forges strategies to enhance research, teaching, and learning environments in collaboration with libraries, cultural institutions, and communities of higher learning*".[1]The report contains recommendations to help to develop reliable systems for preserving access to digital information. It is interesting to see which

[1] See: www.clir.org

insights and principles formulated in 1996 turned out to be future-proof, which means they “survived” until 2019, so they are obviously good candidates to be of value for the following decades.

The report states:

> “*Long-term preservation of digital information on a scale adequate for the demands of future research and scholarship will require a deep infrastructure capable of supporting a distributed system of digital archives. [...] A critical component of the digital archiving infrastructure is the existence of a sufficient number of trusted organizations capable of storing, migrating and providing access to digital collections. A process of certification for digital archives is needed to create an overall climate of trust about the prospects of preserving digital information*” (Waters, 1996, p. 40).

It can be observed that in the years after the report was published, this distributed system of digital archives became an important component of preservation infrastructures as we see them today. It can be expected that this principle will be a key component of a long-term access infrastructure for preserving digital objects. The internet, just emerging at the time, also brought expectations, formulated in the report as follows:

> “*Providing access to digital information in a distributed network environment means above all that digital archives are connected to networks using appropriate protocols and with bandwidth suitable for delivering the information under their control”.*

It can be stated that the internet infrastructure as it is available today provides both the appropriate basic protocols (to produce and consume data on the web) and suitable bandwidth.

Another key element concerns standards for describing and managing digital information. Back in 1996 the report stated:

> “*Descriptive information about the content of digital objects, their origins and provenance and their management over time is critical for both long-term preservation and future use of digital information. Standards and best practices for describing and managing digital information are needed to track changes in ownership or control over digital objects throughout their life cycle, to administer intellectual property rights, and to document any changes in the format and structure of digital objects that may ensue from migration. A responsible digital archive must provide to its users what it knows about the provenance and context of its objects so that users can make informed decisions about the reliability and quality of the evidence before them. Standards bodies, professional associations in the archival, library and information technology fields ... need to collaborate in an evaluation and expansion of descriptive standards and practices so that they satisfy the special requirements of digital preservation and access”* (Waters, 1996, p. 44).

The lifecycle approach was introduced to highlight that digital objects go through a sequence of stages from their initial creation to their eventual archival storage or deletion. Also metadata standards appear as a continuous factor in the realisation of a long-term access infrastructure. Collaboration between stakeholders is an obvious incentive for a successful implementation of a long-term access infrastructure.

Another element of a digital preservation infrastructure as it was foreseen in 1996 is related to the access of the digital objects and the corresponding need for durable data formats. This is illustrated by the following fragment.

> “*Digital archives have an obligation to maintain the information in a form so that users over the network can find it with appropriate retrieval engines and view, print, listen to or otherwise use it with appropriate output devices. In the descriptions of the resources they hold, responsible digital archives must provide to their users what they know about the provenance and context of their digital objects so that users can make informed decisions about the reliability and quality of the evidence before them. With respect to access, digital archives also have the responsibility to manage intellectual property rights by facilitating transactions between rights-holders in the information and users and by taking every reasonable precaution to prevent unauthorized use of the material”* (Waters, 1996, p. 26).

The 1996 report introduced important basic principles for a long-term access infrastructure for digital data such as a distributed, networked architecture, a key role for metadata, the importance of durable data formats and networked certified digital repositories. These principles were further developed and implemented in the decades that followed and became important foundations for contemporary systems and future implementations. In this report we build on these foundations to formulate the components of a long-term access infrastructure for preserving Holocaust digital objects.

1.2 Long-term access infrastructure in the context of the EHRI project

The aim of this report is to provide comprehensive information for organisations that manage Holocaust archives by means of an electronic catalogue and that would like to make this catalogue accessible in a sustainable way for the long-term. This deliverable also aims to cover the durability of digital Holocaust objects kept locally, such as digital images or multi-media files. The long-term access infrastructure as described in this deliverable consists of systems, standards, procedures and policies to preserve digital Holocaust research objects in the long-run.

The first section of this report shows that we can rely on digital preservation principles first formulated in the 1990s and further developed in the following decades. There is a high degree of probability that these principles will be effective in the long-run, at least until the year 2040. The EHRI project has implemented several of the principles that will be covered in more detail in this report. Examples are the distributed architecture of the EHRI portal, the prominent role of the EAD (Encoded Archival Description)[2] metadata standard and the communication protocols to populate, update and synchronize the common database of the EHRI-portal[3] with the information on inventories and institutions provided by the distributed archives in the project. Besides the durability of the distributed system as created in the EHRI project, this report also looks into long-term issues in relation to digital objects managed locally by the archives.

The work done in relation to this report is part of Task 13.2 of the EHRI project, “Secure Long-term Access Infrastructure for the Preservation of Holocaust Research Objects”. This task produced three deliverables of which this report is the last one. The deliverable “D13.3 Data management planning for long-term preservation”,[4] published in 2017, contains relevant input for a long-term access infrastructure. The main subject of the deliverable concerns data management. 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. A second deliverable “D13.4 Trusted Digital Repository” consisted of a workshop (organised in June 2018) and a report aimed

[2] The official EAD site can be found at: https://www.loc.gov/ead/

[3] https://portal.ehri-project.eu/

[4] D13.3 is available at: https://ehri-project.eu/ehri-deliverables

partners in the EHRI project that have the remit to preserve digital objects or will have this task in the future. The workshop introduced and discussed key issues in relation to the management of digital data by Holocaust archives. The workshop discussed attitudes towards the relevance of digital preservation, certification of digital repositories, the role of persistent identifiers in a digital preservation infrastructure, and methods to assess the capability of an organisation to publish metadata in a sustainable way.

This report builds upon the outcomes of both deliverables as they cover key features of a long-term access infrastructure that are presented in the next chapter. Figure 1 illustrates the relation between the three deliverables of Task 13.2.

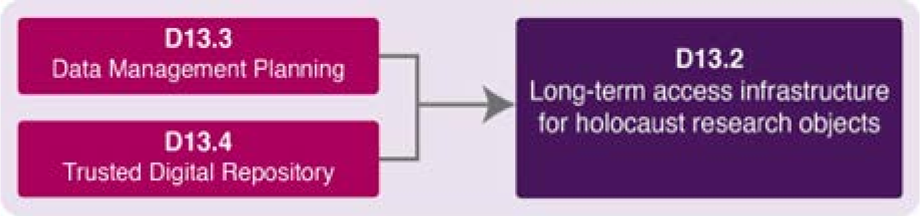

D13.3
Data Management Planning
D13.4
Trusted Digital Repository
D13.2
Long-term access infrastructure for holocaust research objects

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

1.3 The knowledge landscape of archives with respect to digital data

Archivists are uniquely placed within the discourse of data and its use (and non-use), with everyday practices and systems for managing collections, and the confluence of traditions of working with cultural heritage holdings and adaptation to emerging technologies, all in their purview. As such, cultural heritage practitioners are more than a vital link in the chain through which historical data are maintained and transmitted, viewing the knowledge landscape from their position of archival thinking offers insight into how this may render new forms of research engagement with the historical record. The report of the KPLEX project by Horsley and Priddy (2018) provides insights in the knowledge landscape of archives with respect to digital data and is the main source of information for this section.

As part of the investigations of the “Knowledge Complexity Project” (KPLEX)[5] a survey of cultural heritage practitioners identified that over 90% held digital collections (figure 2). Yet despite the fact that over 50% felt it was very much their public duty to share data (figure 3), only 35% had established involvement in aggregation (or data infrastructure) projects (figure 4).

Dissemination of knowledge was also a perpetuation of institutional purpose, with visibility growing “enormously” through participation in aggregation networks, leading to increasing numbers of users both from afar and in reading rooms amongst KPLEX interviewees. Thus, the uses of archival data were found to be changing as a result of the increased visibility of descriptive metadata, research artefacts themselves and/or their underlying data. Practitioners had noticed that researchers were approaching them with more refined questions rather than seeking general guidance from archivists as to what the possibilities for narrowing their research questions might be.

A societal-level change could reshuffle institutions’ priorities, necessitating work they had not previously found time for, when the alternative was to slip into irrelevance or obscurity. Thus, shaking up institutional practice from the outside could therefore achieve significant change in a relatively short time. External influences were cited by many participants as the catalyst for adopting greater standardisation.

[5] See: www.kplex-project.eu

Archival practitioners, interviewed in KPLEX, therefore, saw themselves as the source of creative thinking about how to move practice forward, although the degree to which they felt empowered to act on their ideas varied. One development that had been embraced, at least at some level, across the sector, was the move from analogue to digital. This meant that even where the organisation of material had not really changed in recent memory, the digital revolution had mandated that institutions revisit their fundamental practices.

Researchers who had already refined their inquiries had been awaiting developments, as they had previously sought data that they knew existed but had been denied access as it was not available in an appropriate format. Ending the frustrations of researchers who had been able to discover but not access data was therefore an important step as well as promoting discoverability for other users; changing research was seen as an inevitable consequence of changing archival practice.

Archival practitioners are therefore tasked with facilitating a flow of knowledge without any bottlenecks caused by discrepancies between the expectations and practices of users and those of others acting on the data. This facilitation was dependent on a culture of decision making and practice conducive to breaking down barriers to knowledge flow.

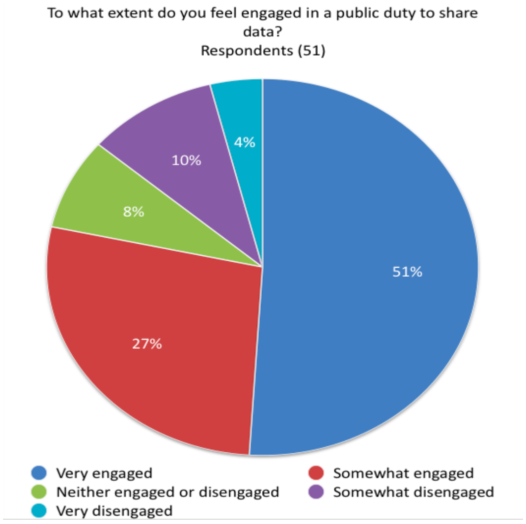


Figure 2. Holdings of CHIs (Cultural Heritage Institutes)

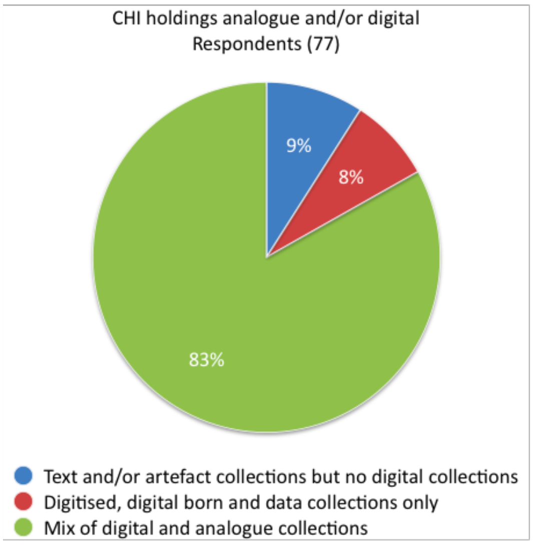


Figure 3. Level of engagement in public duty to share data

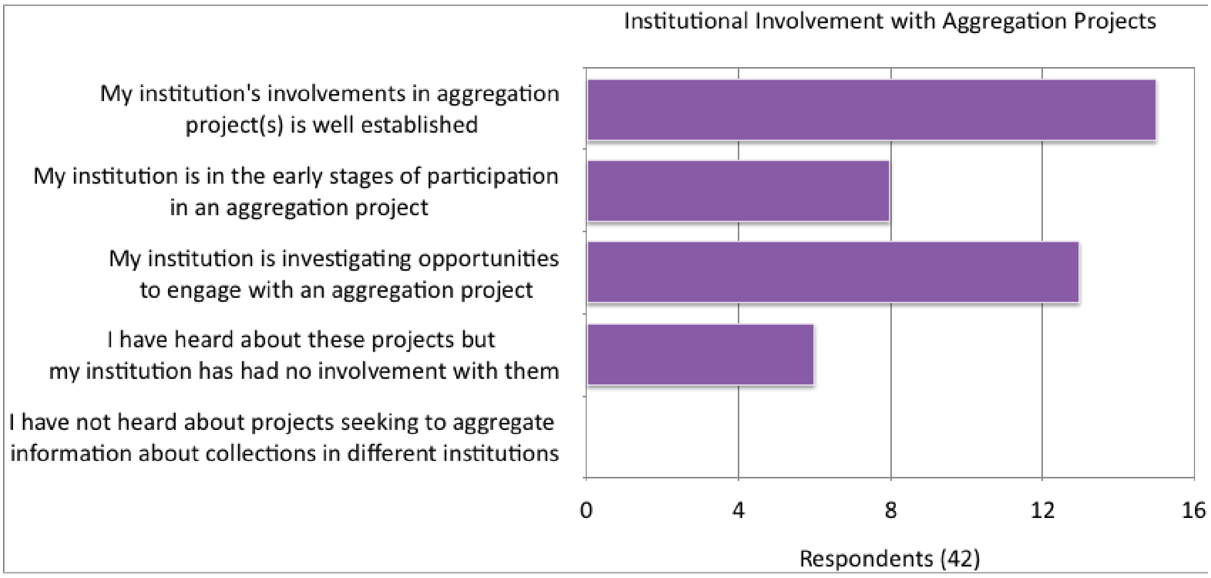


Figure 4. Institutional involvement with Aggregation Projects

There appears to be an assumption that the original physical object as some form of 'back-up' of the digital simulacra and thus there appears to be little investment in preserving the digital version. However, there is an investment of time, effort, and money in the process of digitisation that would be prohibitive should it need to be repeated at scale. Digitisation had also opened up opportunities for acquisition as donors were more likely to offer a digital copy of items than originals. In both cases digital long-term preservation is essential.

Far from developing practice being stymied by practitioners trapped in established habits of consulting physical materials, the ease of digital working has taken hold. Archivists as well as researchers are thus becoming more familiar with digital holdings as they eschew the troublesome non-digital. The current prioritisation of digital discovery for research by aggregators, begs the question from the researcher: if I can see it is there from my desktop, why can't I also get the resource on my desktop? However, there were fears amongst archival practitioners that adoption of new technologies and practices by institutional management was not governed by a long-term strategy and could be in thrall to passing trends. Furthermore, current working practices were often described as unsustainable as institutions struggled to keep up with changes in practice, and thus progress could be squandered.

For smaller institutions that had not previously enjoyed exposure to a wide audience, digitisation had expanded the proportion of material used, however smaller institutions were at risk of becoming marginalised as they drifted away from the orbit of standards used by better-resourced institutions. Archival practitioners expressed openness to changing their practice when they were confident of the benefits of sharing. Infrastructure projects were seen as both "a good way to see the importance of standards and norms" and "to have a larger view about our field and other scientific fields". However, stumbling blocks like differences in metadata schema continued to get in the way of closer cooperation but "evolving" with other institutions within an infrastructure stimulated a general harmonisation of goals.

The "two-way" advantages of sharing were widely acknowledged (ibid.). While conformity changed institutional practice, it also provided space for reaffirming institutional identities, which was a key motivation for joining infrastructure projects. Archival practitioners were largely optimistic about the profound changes to research they believed were afoot. They were also broadly supportive of the cosmopolitan, democratic spirit of sharing embodied by infrastructures. Their enthusiasm about archival holdings fueled their commitment to sharing their “hidden treasures”, which they saw as precious but also “common knowledge”, in the sense that such knowledge should be a commons. Participants in KPLEX were in unequivocal agreement that they were providing a service of public knowledge.

1.4 Open Archival Information System (OAIS) Reference Model

The knowledge landscape as described by the KPEX project makes clear that cultural heritage institute will benefit from a clear overview of the functions of a digital preservation system. The Open Archival Information System (OAIS) Reference Model has become the *lingua franca* of digital preservation helping to reach common agreement on the features and functions of digital preservation systems. The importance of the OAIS Reference Model lays in the fact that it provides fundamental concepts for digital preservation activities (that are testable) and that it provides fundamental definitions so people can speak without confusion. The reference model has formed the foundation of numerous architectures, standards, and protocols, influencing system design, metadata requirements, certification, and other issues central to digital preservation (Lavoie, p. 1). The OAIS model plays a role in the architecture and design of digital preservation information systems. The OAIS standard states: *“It is assumed that implementers will use this reference model as a guide while developing a specific implementation to provide identified services and content*” (OAIS, 2012, p. 1-3).

An OAIS is defined as *“an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community”* (OAIS, 2012, p. 1-11). A Designated Community is defined as “*an identified group of potential consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities*” (OAIS, 2012, p. 1-10). The OAIS reference model is being used by several memory institutes as a basis for activities in the field of digital preservation. The model contains requirements for an archive to provide long-term preservation of digital information.

The OAIS Reference Model distinguishes six mandatory responsibilities that an organisation must discharge in order to operate an OAIS archive (OAIS, 2012, p. 3-1). The archive must:

- Negotiate for and accept appropriate information from information producers.
- Obtain sufficient control of the information provided to the level needed to ensure long-term preservation.
- Determine which communities should become the Designated Community and, therefore, should be able to understand the information provided.
- Ensure that the information to be preserved is independently understandable to the Designated Community.
- Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, and which enable the information to be disseminated as authenticated copies of the original, or as traceable to the original.
- Make the preserved information available to the Designated Community.

1.4.1 OAIS Environment

Figure 5 illustrates the environment surrounding an OAIS. Producers are persons or client systems that provide the information to be preserved. Management’s responsibilities include formulating, revising, and in some circumstances, enforcing, the high-level policy framework governing the activities of the OAIS. Examples of functions carried out by Management include strategic planning, defining the scope of the archived collection, and articulating the preservation guarantee associated with items entrusted to the archive (Lavoie, 2014, p. 9). Consumers are individuals, systems or organisations that use the information preserved by the OAIS. The Designated Community (as defined above) are the primary users expected to independently understand the archived information in the form in which it is preserved and made available by the OAIS.

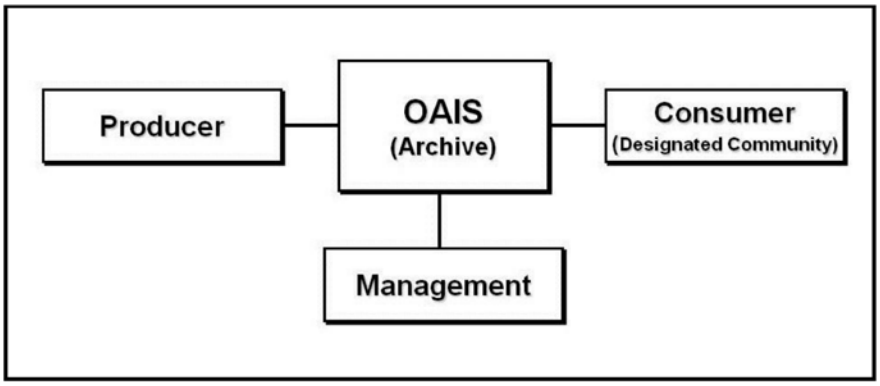


Figure 5. OAIS Environment[6]

A Holocaust archive should accept the responsibility to preserve information and make it available for its consumers or Designated Community in the long term. The Designated Community as distinguished by Holocaust archives, the target group of this report, is often described on the website. Four examples of a description of a Designated Community of a Holocaust archive are given below.

NIOD, the Dutch institute for War, Holocaust and Genocide Studies states: *Issues related to war violence generate a lot of interest from society and demand independent academic research. NIOD conducts and stimulates such research and its collections are open to all those who are interested.*[7] The UK based Wiener Library for the Study of the Holocaust and Genocide has the mission *To serve scholars, professional researchers, the media and the public as a library of record*.[8] The Designated Community of Yad Vashem can be described in relation to the statement *Yad Vashem, ..., is the ultimate source for Holocaust education, documentation and research*.[9] A last example concerns the Institute for Contemporary History (IfZ) in Munich, which counts scientific researchers as its Designated Community. *IfZ is a non-university research institution that researches the entire German history of the 20th century up to the present day in its European context*.[10]

The Designated Community for Holocaust archives can be defined as people, active in research and education, and ranging from specialists to the general public. It is the scope of the Designated Community that determines both the contents of the OAIS and the forms in which the contents are preserved, such that they remain available to, and independently understandable by, the Designated Community (Lavoie, 2014, p. 10).

[6] Lavoie, 2014, p. 9.

[7] See: <https://www.niod.nl/en/about-niod> [cited 12 October 2018]

[8] See: <https://www.wienerlibrary.co.uk/Our-History> [cited 12 October 2018]

[9] See: <http://www.yadvashem.org/about> [cited 12 October 2018]

[10] See: <http://www.ifz-muenchen.de/das-institut/> translated from German [cited 12 October 2018]

1.4.2 OAIS Functional Model

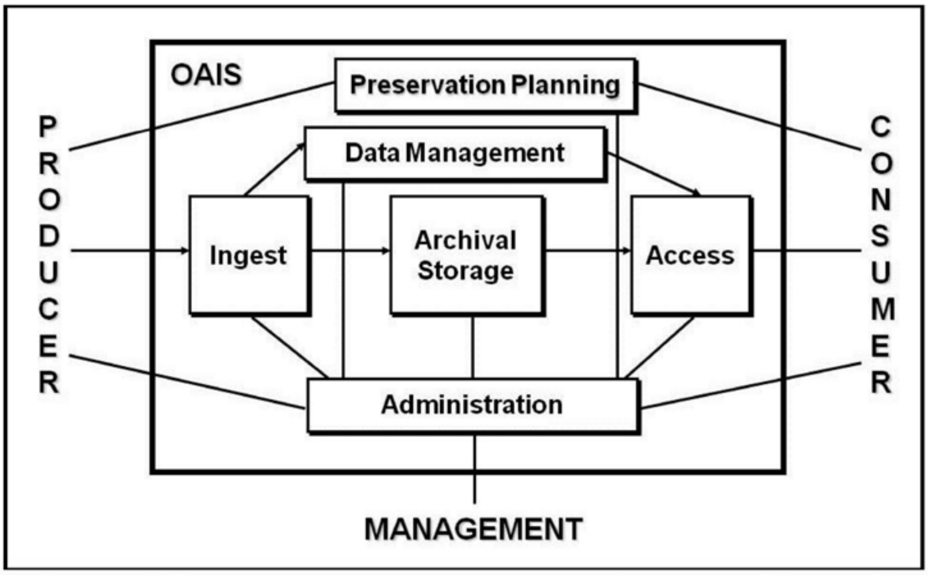


Figure 6. OAIS Functional model[11]

The OAIS Functional model, as depicted in figure 6, contains six high-level services or functional entities that facilitate the preservation and access of information stored in an OAIS.

- *Ingest*. Contains the services and functions that accept information objects (both content and associated description information) from producers, prepares the information for storage, and ensures that the information and their supporting Descriptive Information become established within the OAIS.[12]
- *Archival storage*. Contains the services and functions used for the storage and retrieval of information objects.
- *Data management.* Contains the services and functions for populating, maintaining and accessing a wide variety of information objects.
- *Administration.* Contains the services and functions needed to control the operation of the OAIS functional entities on a day-to-day basis.
- *Preservation planning.* Contains services and functions for monitoring the environment of the OAIS and providing recommendations to ensure that the information objects stored in the OAIS remains accessible to the Designated Community over the long term, even if the original computing environment becomes obsolete.
- *Access.* Contains the services and functions that make the archival information holdings and related services visible to Consumers.

An OAIS-type archive will implement each of these functional entities, in one form or another, in the course of building a complete archival system. Key components of this implementation process are presented in the next chapter of this report.

[11] See: Page 12 Lavoie (2014)

[12] The OAIS reference model provides a high-level description of the information objects managed by the archive. An information package consists of the object that is the focus of preservation, along with metadata necessary to support its long-term preservation, access, and understandability, bound into a single logical package. See OAIS (2012) for a detailed description of the information objects.

1.4.3 OAIS Information Model

The OAIS Reference Model consists of an Information Model that describes the types of information that are exchanged and managed within an OAIS. A basic concept is the Information Object that is composed of a Data Object (that is either physical or digital) and the Representation Information that allows for the full interpretation of the data into meaningful information. This model is valid for all types of information in an OAIS (OAIS, 2012, p. 4.21). Figure 7 illustrates the Information Object concept.

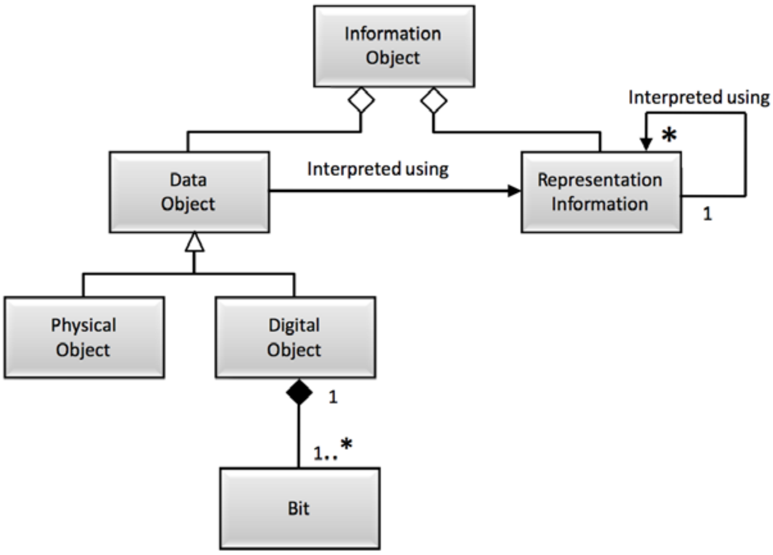


Figure 7. OAIS Information Object[13]

Representation Information is the information necessary to render and understand the bit sequences constituting the Data Object. Representation Information is required in order to make the Data Object available in a form that is independently understandable by the Designated Community.

Representation Information might include a description of the hardware and software environment needed to display the Data Object and/or access its contents. It might also summarize the appropriate interpretation of the Data Object. For example, if the Content Data Object is an ASCII file of numbers, Representation Information might indicate that the numbers correspond to average daily air temperature readings for London, measured in degrees Celsius, for the period 1972 – 2000. Representation information can be divided into two types: Structure Information and Semantic Information. Structure Information is most easily understood in the context of digital objects, and refers to mappings between digital bits and various concepts and data structures that render the bits into intelligible information – i.e., an image, text, an interactive program. Generally speaking, Structure Information describes the format of the digital object. Semantic Information, on the other hand, is information that clarifies the meaning or appropriate interpretation of the Content Data Object. A glossary, a data dictionary, and a software application's user documentation are all examples of Semantic Information that may be bundled with the Data Object as part of its Representation Information (Lavoie, 2014, p. 16).

[13] See page 4-21 of OAIS (2012)

1.5 FAIR Data Principles

The features of durable, sustainable data objects are closely related to the ‘FAIR data principles’ that are introduced in a 2016 paper (Wilkinson et. al. 2016) and since then gained considerable attention in the research data community. The acronym FAIR stands for “Findability”, “Accessibility”, “Interoperability” and “Reusability” of data objects. Importantly, data should not only be ‘FAIR’ for humans but also for machines, allowing, for instance, automated search and access to data. The principles have gained solid ground in the scientific community leading to many initiatives around improving the ‘fairness’ of research data objects. Also funders like the European Commission have drafted Guidelines on FAIR Data Management for the H2020 programme.[14] Good data management is one way to support the FAIR principles. One can make digital data more ‘FAIR’ by using persistent identifiers, adding sufficient documentation and metadata and adding clear licenses. When storing the data in a trusted digital repository, usually these services that improve the ‘fairness’ of data are provided to the depositor. The FAIR data principles are aimed at the assessment of quality features of digital data objects.

Each of the four FAIR principles are described.[15] 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. The Findable principle is explained in more detail by providing four statements that are given below.

> F1. data objects 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. data objects are registered or indexed in a searchable resource.

Digital preservation services described in the foregoing section of this article that can be used to implement the Findable principle are for instance the DOI as persistent identifier and the PREMIS metadata element set.

The Accessible principle is related to the retrieval of data objects by their identifier and the availability of metadata. Two statements support this principle:

> A1. data objects 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

The Accessibility principle does not necessarily mean 'open' or 'free', but rather gives the conditions under which the data objects are accessible. The A2 principle is related to the problem of 'broken links'. It can be useful to keep metadata on data objects available in order to inform users on the provenance of the data objects.

The Interoperability principle[16] is realized by using formal, broadly applicable languages for knowledge representation and qualified references. To be interoperable:

> I1. data objects use a formal, accessible, shared, and broadly applicable language for knowledge representation.
> I2. data objects use vocabularies that follow FAIR principles.
> I3. data objects include qualified references to other data objects.

[14] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[15] The FAIR principles are initiated by FORCE11, a community of scholars, librarians, archivists, publishers and research funders. See: https://www.force11.org/group/fairgroup/fairprinciples

[16] Interoperability can be defined as 'the ability of data or tools from non-cooperating resources to integrate or work together with minimal effect.” (Wilkinson, 2016, p. 2)

Within the EHRI project activities are carried out to compile shared controlled vocabularies and authority list.[17]

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

R1. data objects are richly described with a plurality of accurate and relevant attributes.

R1.1. data objects are released with a clear and accessible data usage license.

R1.2. data objects are associated with detailed provenance.

R1.3. data objects meet domain-relevant community standards.

A main idea of the principle is that the metadata must be detailed enough to determine whether the data objects are useful for a particular goal. The availability of additional information on legal aspects, provenance (the origin and history of the data objects) and community standards are specific issues to consider.

1.6 Towards a long-term access infrastructure

Digital preservation principles defined more than 20 years ago, presented in the first section of this chapter, turned out to be of value today and in the future. The survey carried out in the KPLEX project made clear that the collections of archival institutions increasingly contain digital objects and that they are confronted with the need to formulate a strategy to manage these digital objects. Management here means, sustainable archiving and providing access to the objects in an optimal way. The OAIS Reference Model provides an overview of services or functions and types of data objects that should be taken into consideration when creating a long-term access infrastructure for preserving Holocaust digital objects. OAIS is a model and not an implementation. The standard does not provide technical details of an archival system, such as system architectures, data storing and processing technologies, or database design. The OAIS Reference Model, however, can assess the quality of components of a digital preservation system, such as the metadata schema used or the way the data objects are formatted. Another example of the importance of the OAIS Reference Model concerns the assessment of the trustworthiness of a repository based on tasks and functions of the OAIS Reference Model. The FAIR data principles provide a framework to assess the sustainability of data objects.

The next chapter provides key features of a long-term access infrastructure based on foundations laid in this introductory chapter.

[17] Riondet et al, *Report on standards* (2017). The Encoded Archival Description (EAD) metadata standard plays a central role in the application of metadata in the EHRI project.

2 Key features of a long-term data access infrastructure

The digital holdings of archives that curate Holocaust records 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. electronic documentation of archival collections). Data Objects together with its related Representation Information (Ref. OAIS Reference Model, see figure 7) must be preserved for access and use by a Designated Community. An assessment of the “FAIRness” of the Information Objects helps to determine which measures should be undertaken to guarantee the sustainability of the digital holdings. Prominent features of a long-term access infrastructure are described in this chapter. They are the durability of the file format, the use of persistent identifiers, the storage of the data in a certified data repository, the quality of the documentation (or metadata), the usage licenses of the Information Objects and data protection and secured access to information.

2.1 Standard file format

A file format specifies how information is encoded in digital form. Durable data file formats that are non-proprietary, that is they are independent of any specific software, developer or vendor. Its specification is openly available and it has a wide user community. Obsolescence of data file formats occurs when new generations of software phase out support for older formats or new versions of data file formats occur that are not compatible with older versions. Different content types have, over time, developed their own file format. Migration of files to a new commonly used format is a widely applied preservation strategy. This requires monitoring of commonly used file formats. File format obsolescence is less of a problem than was perceived some ten years ago. Many file format specifications are still supported and still usable today.

It should be noted that for some digital file formats different versions exist, e.g. PDF, DOC or TIFF. Tools for file format identification and file format verification can be used to manage and track file formats and versions of file formats. The PRONOM online information system, for instance, provides documentation on a wide range of data file formats.[18] The related DROID file format identification tool provides categories of format identification for unknown files in a collection.[19] Often data archives provide an overview of data file formats they consider durable. By using this durable or preferable data file format the depositor can be quite confident that the data will remain readable in the long run.[20]

The optimal choice for a data file format is also influenced by its function. It has been generally agreed, for instance, that the TIFF format is a suitable format for archiving master digital image files and that the JPEG format can be used for access copies of digital images. The WAV format is commonly used for audio archiving. In other areas a lack of agreement can be observed, e.g. in digital video, where different wrapper formats and delivery and encoding methods are apparent. Sustainability of file formats for cultural heritage institutes as well as open source tools to support its long-term usability is covered in detail in the PREFORMA handbook, entitled “validating formats, a prerequisite for preserving digital objects” (Preforma, 2017).

[18] See: http://www.nationalarchives.gov.uk/PRONOM/

[19] http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/

[20] See for instance overview of preferred file formats of the Dutch research data archive DANS: https://dans.knaw.nl/en/deposit/information-about-depositing-data/file-formats

The most prominent digital file types used to format digital objects, are text files, digital images and multimedia files, such as video files. By using standard file formats, the longevity of these files is supported. The table below provides an overview of file formats that can be considered as durable.

File Type	*Durable Data Format*
Digital images	JPEG (.jpg, .jpeg) / TIFF (.tif, .tiff) / PNG (.png) / JPEG 2000 (.jp2)
Plain text	Unicode
Text	PDF/a
Audio	BWF (.bwf) / MXF (.mxf) / Matroska (.mka) / FLAC (.flac) / OPUS
Video	MXF (.mxf) / Matroska (.mkv)

Table 1. Durable data formats[21]

Data formats do exist that are currently widely used, but cannot - given the criteria stated above - be considered as durable, e.g. because the format relies on proprietary software to be rendered. Examples are the file formats of Microsoft Office, such as Word. In this case the principle “Trust but verify” is applicable.

A strategy has to be followed to be certain that the data formats remain durable. The most suitable strategy seems to be the “migration strategy”. This means that file formats are repeatedly converted to keep up with the present technical generation (Preforma, 2017, p. 30).

2.2 Persistent identifiers

The next aspect that improves the sustainability of digital objects concerns the application of persistent identifiers (PID). PIDs are long lasting unique references to resources that were invented to address challenges arising from the distributed and disorganized nature of the internet.[22] They are an important digital preservation component as they enable the long-term location of data objects even when its location changes. PIDs aim to prevent 'link rot' (the web link to a resource is unavailable) and 'content drift' (the web link does not refer to the correct resource). A persistent identifier typically has two components: a unique identifier and a service that locates the resource (and that takes the changing of this location into consideration).

Persistent digital identifier systems do require resolution systems that create and maintain the link between the identifier and the object. These resolver systems must be supported by people and services and thus relies on an organizational sustainability policy. Several persistent identifier schemas have existed for quite some years and are used by a considerable community. For this reason they are good candidates for the persistent identification of resources.

So-called ‘Handles’ are widely used persistent identifiers, supported by the Handle System, a distributed system for assigning these persistent identifiers. Handles consist of a prefix which

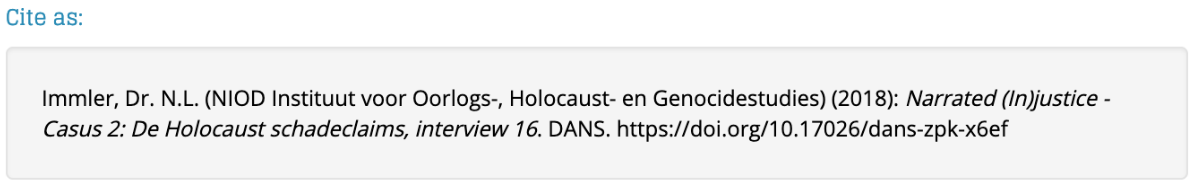
[21] Table based on list of Preferred data formats of DANS, https://dans.knaw.nl/en/deposit/information-about-depositing-data/before-depositing/file-formats

[22] For a discussion of the function and usage of PIDs see: McMurry et al. (2017)

identifies a 'naming authority' and a suffix which gives the 'local name' of a resource.[23] The International DOI Foundation (IDF) implements Handles and mints them as DOIs (Digital Object Identifier).[24] A DOI can be assigned to any physical, digital or abstract entity that one wishes to identify. In case an organization wants to link PIDs to objects it has to join a service provided by a DOI Registration Agency. Two examples of a DOI Registration Agency are CrossRef and DataCite. CrossRef is a Registration Agency initiated by publishers, mainly to facilitate the persistent identification of publications.[25] DataCite is the appropriate Registration Agency for the persistent identification of data objects, thus improving the citation of e.g. datasets.[26]

An example of a persistent identifier for person names is the ISNI (International Standards Name Identifier) that assigns identifiers (a 16 digit number) to names based on the ISO-27729 standard.[27] The ORCID initiative (Open Researcher and Contributor ID) uses the same ISO standard (so an identifier is unique) but focuses on the exclusive provision of PIDs for researchers.[28] The persistent identifier as such to an object (e.g. publication, data file, person) does not contain any information, or metadata, about the resource to which it refers. Metadata does play an important role in the digital preservation of resources in order to understand and assess the value of a digital object. Often the persistent identifier is a field in the metadata of an object.

Display guidelines are formulated by PID service providers that can be used to make PIDs easy to recognise and use, both by humans and machines.[29] Figure 8 shows an example of citation documentation of a dataset that contains a globally unique persistent identifier. The PID is a permanent link to the database.



Cite as:

Immler, Dr. N.L. (NIOD Instituut voor Oorlogs-, Holocaust- en Genocidestudies) (2018): *Narrated (In)justice - Casus 2: De Holocaust schadeclaims, interview 16*. DANS. https://doi.org/10.17026/dans-zpk-x6ef

Figure 8. Example of citation information of dataset that contains a machine actionable link

The landscape of available persistent identifiers is in development. The FREYA project[30] aims to extend the infrastructure for persistent identifiers (PIDs) as a core component of open research, in the EU and globally. The project works on improving discovery, navigation, retrieval, and access to research resources. New provenance services will enable researchers to better evaluate data and make the scientific record more complete, reliable, and traceable. By engaging with the global community through the Research Data Alliance (RDA)[31] and other research infrastructures, FREYA works together to realise the vision of fully accessible data.

[23] Information on the prefixes is stored in a Global Handle Registry. This Registry is operated by the DONA foundation. See: <https://www.dona.net/>

[24] See: <https://www.doi.org/>

[25] See: <https://www.crossref.org/>

[26] See: <https://www.datacite.org/>

[27] See: <http://www.isni.org>

[28] See: <https://orcid.org/>

[29] See for instance the display guidelines of DataCite at: <https://support.datacite.org/docs/datacite-doi-display-guidelines>

[30] www.freya-project.eu

[31] See: <https://www.rd-alliance.org/>

2.3 Certified Data Repositories

A Trusted Digital Repository (TDR) is a digital repository that is certified according to a set of requirements. It ensures that the digital objects will be archived and made available in a controlled environment. In general terms a TDR ensures that the data objects are accessible and usable according to the assessed features such as licenses, format and documentation. The Core Trust Seal (CTS) is an important reference to assess the trustworthiness of digital repositories.[32]

The CTS consists of 16 guidelines that a repository should adhere to in order to be certified. They concern for instance the way the repository maintains licenses, how the repository ensures ongoing access to and preservation of its holdings, how the funding and expertise of its staff is organised, how the repository the authenticity and integrity of the data guarantees, how the repository ensure sufficient documentation is available to ensure end-users can evaluate its quality, how the repository facilitates proper citation of the data in a persistent way, and how the repository protects its holding.

Figure 9. Seal of CTS to be used by CTS certified repositories

The CTS guidelines are:

1. The repository has an explicit mission to provide access to and preserve data in its domain.
2. The repository maintains all applicable licenses covering data access and use and monitors compliance.
3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.
4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.
5. The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.
6. The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either inhouse, or external, including scientific guidance, if relevant).
7. The repository guarantees the integrity and authenticity of the data.
8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.
9. The repository applies documented processes and procedures in managing archival storage of the data.
10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.
11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.
12. Archiving takes place according to defined workflows from ingest to dissemination.
13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.

[32] See: <www.coretrustseal.org>

14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.
15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.
16. The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

The CTS supports a low-threshold self-certification process. To what extend the guidelines of the CTS are implemented by an organisation is evaluated by an internal assessment in which the level of conformance is documented and presented to the CTS board. In case all guidelines are implemented the CTS board will give permission to put the CTS logo on the website of repository.

2.4 Standardised metadata

Metadata refers to data that supports the discovery, understanding and management of other data and information. A large number of metadata standards and schemas have been developed to describe, structure or manage objects. They are an important component of a long-term access infrastructure for preserving Holocaust research objects. By documenting data objects its value can be assessed, also in the future.

Three metadata standards are discussed:

1. Dublin Core Metadata Element Set (DCES), standard used for resource discovery on the internet.
2. Encoded Archival Description (EAD), standard to document archival finding aids. This standard is the basis of the EHRI portal.
3. Preservation metadata (PREMIS)

Dublin Core Metadata Element Set (DCES)

For one specific function of metadata, namely resource discovery, the Dublin Core Metadata Element Set is used on a large scale.[33] The fifteen core elements of DCES (see table 2) are applied in a large number of projects and initiatives in order to enable the discovery of objects on the Internet. In 2001 DCES became an official ANSI/NISO standard (Z39.85) and in 2003 DCES was issued as international ISO standard 15836.

Metadata Element	*Definition*
Contributor	An entity responsible for making contributions to the resource.
Coverage	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
Creator	An entity primarily responsible for making the resource.
Date	A point or period of time associated with an event in the lifecycle of the resource.
Description	An account of the resource.
Format	The file format, physical medium, or dimensions of the resource.

[33] See: http://www.dublincore.org/documents/dces/

Identifier	An unambiguous reference to the resource within a given context.
Language	A language of the resource.
Publisher	An entity responsible for making the resource available.
Relation	A related resource.
Rights	Information about rights held in and over the resource.
Source	A related resource from which the described resource is derived.
Subject	The topic of the resource.
Title	A name given to the resource.
Type	The nature or genre of the resource.

Table 2. 15 Dublin Core Metadata Elements

Encoded Archival Description (EAD)

The Encoded Archival Description (EAD) is an international metadata standard to encode archival finding aids. The EAD is a key-component of the data-infrastructure of the EHRI project. It is used to aggregate distributed information to be incorporated in the EHRI portal. The role of the EAD and other standards in the EHRI project is addressed in Riondet (2017). EAD defines the structural elements of finding aids and their relationships. It accommodates the hierarchical structure of large collections of unpublished material and has more than 100 description elements.

PREMIS

With respect to long term access to digital objects the PREMIS Data Dictionary is of importance. PREMIS stands for “PREservation Metadata: Implementation Strategies”.[34] It is an international working group concerned with developing metadata for use in digital preservation. PREMIS is incorporated in a number of commercial and open-source digital preservation tools and services and aims to contain all information needed to support the preservation process.[35]

The PREMIS Data Dictionary is organized around a data model consisting of five entities or semantic units associated with the digital preservation process. The first one is the 'Digital Object', a discrete unit of information subject to digital preservation, for instance a digital image, a website or a database. The second entity is named 'Environment'. It contains details on the technology (software or hardware) supporting a digital object in some way (e.g. rendering or execution). The third entity, 'Event', is described as an action that involves or affects at least one 'Object' or 'Agent' associated with or known by the preservation repository. 'Agent', the fourth entity, is a person, organization, or software program/system associated with 'Events' in the life of an 'Object', or with 'Rights' attached to an 'Object'. It can also be related to an environment 'Object' that acts as an 'Agent'. The fifth and last entity in the PREMIS Data Dictionary concerns assertions of one or more 'Rights' or permissions pertaining to an 'Object' and/or 'Agent'.

The PREMIS Data Dictionary consists of almost 200 entries, divided over the five semantic units given above. It defines what a preservation repository needs to know. It is important to

[34] The official PREMIS website can be found at: <https://www.loc.gov/standards/premis/>

[35] See: <http://www.loc.gov/standards/premis/tools.html>

note that the focus is on the repository system and its management, not on the authors of digital content, people who scan or otherwise convert analogue content to digital, or staff who evaluate and license commercial electronic resources.

The primary uses of PREMIS are for repository design, repository evaluation, and exchange of archived information packages. Caplan states “Those designing or developing preservation repository software applications should use PREMIS as a guideline for what information should be obtained and recorded by the application or otherwise known to repository management. Those who are planning to implement a preservation repository should use the PREMIS Data Dictionary as a checklist for evaluating candidate software. Systems which can support the PREMIS Data Dictionary will be better able to preserve information resources in the long term” (Caplan, 2017, p. 3).

2.5 Usage licenses

The long-term access to data objects requires proper usage license management. A license agreement is a legal arrangement between the creator/depositor of a data object and the data repository, signifying what a user is allowed to do with the data. Stating clear reuse rights is an important aspect in making sure your data meet the R (Reusable) in FAIR data management.

A license agreement makes clear who, at a certain point in time, has access to data objects and under which conditions. It can also state that the data objects are protected and not accessible. The details of the license agreement can be stored either in the header of the data object (e.g. in the header of a digital image) or in the documentation record that describes the data object (e.g. in the Dublin Core data element “Rights”). A first step to towards proper access management concerns the establishment of ownership. Data objects can only be archived, published and made available for reuse by the owner or by permission of the owner.

Once the ownership and permissions are settled, the next step is to determine what actions are allowed with the data objects. The Creative Commons (CC) copyright licenses and tools[36] provide a comprehensive framework for this. Table 3 provides an overview of the CC licenses. A precondition of a CC license is that a data object can be copied and redistributed.

License	Can I copy & redistribute the work?	Is it required to attribute the author?	Can I use the work commercially?	Am I allowed to adapt the work?	Can I change the license when redistributing?
CC0	Y	N	Y	Y	Y
CC BY	Y	Y	Y	Y	Y
CC BY-SA	Y	Y	Y	Y	N
CC BY-ND	Y	Y	Y	N	Y
CC BY-NC	Y	Y	N	Y	Y

[36] See: <https://creativecommons.org/>

CC BY-NC-SA	Y	Y	N	Y	N
CC BY-NC-ND	Y	Y	N	N	Y

Table 3. Creative Commons licenses[37]

Figure 10, part of a website page, illustrates how a usage license can be attached to a digital object, in this case a digital image. The usage license provides details on the way the digital image may be used by others.

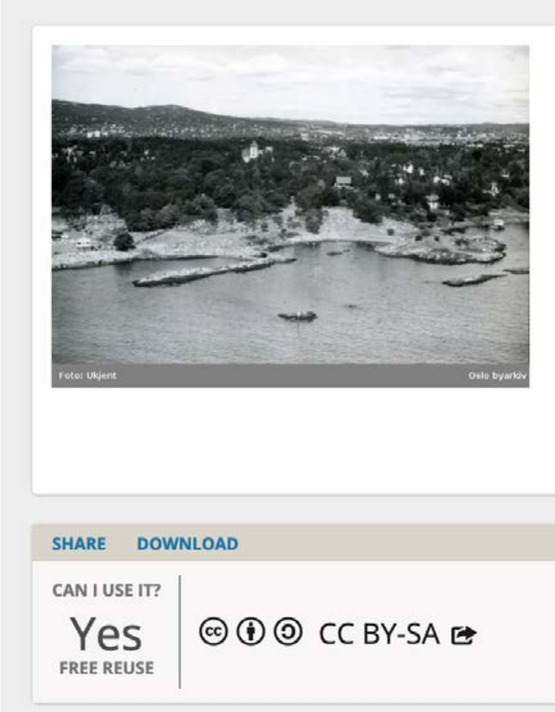


Figure 10. Example of Creative Commons usage license for digital image[38]

2.6 Data protection and secure access[39]

Publishing data in a data repository does not automatically make them openly accessible. (Sensitive) personal data can still be protected by limiting access to the data. Access controls can permit control down to an individual file level, meaning that mixed levels of access control can be applied to a data collection.

Many data repositories operate a three-tiered approach to data access:

- Open access
  Data that can be accessed by any user whether they are registered or not. Data in this category shouldn't contain personal information, unless consent is given.
- Access for registered users (safeguarded)
  Data that is accessible only to users who have registered with the archive. This data

[37] Table from CESSDA Data Management Expert Guide <https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/6.-Archive-Publish/Publishing-with-CESSDA-archives/Licensing-your-data>

[38] See: <https://www.europeana.eu/portal/record/2022608/BAR_A_20015_Ua_0003_055.html>

[39] This section is taken from the CESSDA Data Management Expert Guide. See: <https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/6.-Archive-Publish/Publishing-with-CESSDA-archives/Access-categories>

contains no direct identifiers but there may be a risk of disclosure through the linking of indirect identifiers.

- Restricted access
  Access is limited and can only be granted upon request. This access category is for the most sensitive data that may contain disclosive information.
  Restricted access requires long-term commitment of the researcher or person responsible for the data to handle the upcoming the permission requests.
- Embargo
  Besides offering the opportunity for restricted access 'for eternity' most data repositories allow you to place a temporary embargo on your data. During the embargo period, only the description of the dataset is published. The data themselves will become available in open access after a certain period of time.

Access conditions may differ slightly between data repositories.

3 Towards a roadmap for a long-term access infrastructure

In a long-term access infrastructure digital objects are formatted in a durable standard, have a persistent identifier, are stored in a certified repository, are documented according to an appropriate standard metadata format, have a suitable usage license and have appropriate access rights.

This chapter covers aspects of the roadmap to create, operate and maintain a long-term access infrastructure for digital objects. A capability maturity assessment helps to find out how the realisation of a long-term access infrastructure can be organised in terms of required skills and competencies. Data management planning is essential to keep data understandable and organised in the long run. Data protection and legal issues is the third aspect of the roadmap, followed by attention for archival data storage and access to data objects. The last part of this chapter is a description of the digital preservation policies and practices of the United States Holocaust Memorial Museum (USHMM).

3.1 Capability maturity assessment

The NDSA (National Digital Stewardship Alliance) has formulated a set of recommendations *“for how organisations should begin or enhance their digital preservation activities. This overview of levels of digital preservation allows organisations to assess the level of preservation achieved for specific materials in their custody. The guidelines are organised into five functional areas: storage and geographic location, file fixity and data integrity, information security, metadata, and file formats.”* The NDSA levels of digital preservation do not cover such things as digital preservation policies, staffing, or organisational support.[40]

Capability maturity modeling can help to establish the maturity level with respect to the application of digital preservation services, data management policies, and organisational support for digital preservation. It provides guidance for improving the situation. A capability maturity model (CMM) is a set of structured levels that describe how well the practices, processes and behavior of an organization can reliably and sustainably produce desired outcomes. Capability maturity modelling is discussed in (Daelen et al, 2016).

The assessment of the level of maturity in relation to data management helps organizations to formulate a strategy to provide long-term access to its digital assets. An assessment of the capability to apply specific digital preservation services and data management policies is important for Holocaust archives. It will enable them to become aware of the current state-of-art concerning the application of data management policies as well as required actions to improve the quality of the data management infrastructure.

The Digital Preservation Capability and Maturity Model (DPCMM) can be used to 'conduct a gap analysis of current digital preservation capabilities and to help practitioners and organizations delineate a multi-year roadmap of incremental improvements' (Dollar, 2015). The model helps organizations to proactively address digital preservation issues. The DPCMM has five stages that are briefly described below.

- Stage 1, 'Nominal Digital Preservation Capability'. Generally, there may be some understanding of digital preservation issues and concerns but this understanding is likely to consist of ad hoc electronic records management practices and digital continuity infrastructure and initiatives. Although there may be some isolated

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

instances of individuals attempting to preserve electronic records on a network or removable storage media, practically all electronic records that merit long-term preservation is at risk.

- Stage 2, 'Minimal Digital Preservation Capability'. A surrogate preservation repository for electronic records is available to some records producers that satisfies some but not all of the OAIS specifications[41]. There is some understanding of digital preservation issues and strategies but it is limited to a relatively few individuals. Most electronic records that merit long term preservation are at risk.
- Stage 3: 'Intermediate Digital Preservation Capability'. Describes an environment that embraces the OAIS specifications and other best practice standards and schemas and thereby establishes the foundation for sustaining enhanced digital preservation capabilities over time. This foundation includes successfully completing repeatable projects and outcomes that support digital preservation capabilities and fosters collaboration, including shared resources, between record producing units and entities responsible for managing and maintaining trusted digital repositories. In this environment many electronic records that merit long term preservation are likely to remain at risk.
- Stage 4: 'Advanced Digital Preservation Capability'. Characterized by an organization with a robust infrastructure and digital preservation services that are based on the OAIS specifications that are audited. At this stage the preservation of electronic records is framed entirely within a collaborative environment in which there are multiple participating stakeholders. Some electronic records that merit long-term preservation may still be at risk.
- Stage 5: 'Optimal Digital Preservation Capability'. This is the highest level of digital preservation readiness capability that an organization can achieve. It includes a strategic focus on digital preservation outcomes by continuously improving the manner in which electronic records lifecycle management is executed. Few if any electronic records that merit long-term preservation are at risk.

The DPCMM consists of fifteen components that are necessary and required for the long-term continuity, access, and preservation of authentic, accessible and reliable electronic records. A short description of the components is given in the table below. The level of digital preservation capability for an organization can be identified by associating the components with one of the five stages of the model.

DPCMM Component	**Short description of the component**
Digital Preservation Policy	*The purpose, scope, accountability, and approach to the transfer of records and the operational management and sustainability of trustworthy preservation repositories.*
Digital Preservation Strategy	*How are risks associated with technology obsolescence addressed? E.g. conversion of files to preservation formats and monitoring of changes in technology.*

[41] For a discussion of the OAIS reference model, see section 1.4.

Governance	*Formal decision-making process that assigns accountability and authority for the preservation of electronic records, and that articulates approaches and practices to meet stakeholder needs.*
Collaborative Engagement	*The level of collaboration concerning different aspects among the many stakeholders an organization has.*
Technical Expertise	*Level of expertise in electronic records management and digital preservation. This may exist within internal or contracted staff or by external service providers.*
Open Standard Technology Neutral Formats	*Actions undertaken to mitigate file format obsolescence.*
Designated Community	*Formal documentation that defines the rights, obligations and responsibilities of the Designated Community (the group of potential users of the archive that should be able to understand a particular set of information).*
Electronic Records Survey	*The objective of the survey is to identify different types of categories of electronic records in order to support planning and preservation activities (e.g. records that require transformation to a durable file format).*
Ingest	*The specifications of the ingest process (consisting of agreements, processing, validation, etc.).*
Archival Storage	*Details on the archival storage, such as number of repository instances, details on metadata and generation of operational statistics.*
Device / Media Renewal	*Details on the monitoring and renewal policies of storage media to ensure that the bit streams remain readable over time.*
Integrity	*Ensuring the integrity ('fixity') of records to cope with accidental or intentional alterations (By using 'digital fingerprints').*
Security	*Techniques to block unauthorized access, protect the confidentiality and privacy of records and intellectual property rights.*

Preservation Metadata	*Preservation metadata describes preservation actions associated with the custody of permanent electronic records. Preservation metadata includes an audit trail that documents preservation actions carried out.*
Access	*Details concerning the access to records, e.g. the in relation to support to open standard technology neutral formats.*

Table 3. Short description of the 15 Components of the Digital Preservation Capability Maturity Model (DPCMM)

An online self-assessment tool can be used to benchmark capabilities to manage and preserve long-term electronic records, support the development of improvement plans, and promote collaboration and the exchange of information on good practices.[42]

The DPCMM is an excellent tool to assess the maturity level of organizations with respect to its capability to use digital preservation services and to formulate data management policies. It helps organizations that manage Holocaust archives formulate a suitable digital preservation strategy.

3.2 Data management planning

Data management refers to the development, execution and supervision of plans, policies, programs and practices that control, protect, deliver and enhance the value of data and information assets. Data management policies provide a broader context for the sustainability of digital collections. It is critical to make sure data are well-organized, understandable and reusable, also in the long term.

Some of the benefits of good data management are that it saves time in the long run. By providing rich documentation of data objects, for instance, users can find their way easily in the data collection. It makes sharing and preserving of the data easier. Moreover, it helps to avoid data loss in the event of a disaster or user error. In this section we discuss life cycle models for data management, the relevance of data management plans and the FAIR guiding principles for data management.

The continuity of digital materials can be ensured in case data management strategies are based on a lifecycle approach. Two data life cycle models, that both have their origin in the research data community, are presented. The 'Research Data Lifecycle' consists of six elements (Corti et al, 2014), see figure 10. Each element is described briefly with an emphasis on the potential value for the management of archival records. The first step, 'Planning', focuses on an exploration of the data sources that will be managed, with an emphasis on how these data sources can be made available for their intended users. The second step, 'Collecting Data', concerns the acquisition of the data as well as the capturing of metadata. The next step is 'Processing and Analyzing data,' which includes the creation of transcriptions and translations, the anonymization of data, as well as the management and storage of the data. The fourth element of the research data lifecycle, 'Publishing and Sharing data', consists of activities such as the establishment of copyright, the creation of user documentation and the realization of the access to the data. The fifth step is 'Preserving Data'. The monitoring of the durability of the data and, if applicable, the execution of preservation actions (such as the migration of files to a durable format) are part of this step.

[42] The DPCMM online assessment tool can be found at http://digitalok.org

The sixth element of the model, 'Reusing Data' involves activities such as usage analysis and related actions to stimulate the reuse of the data.

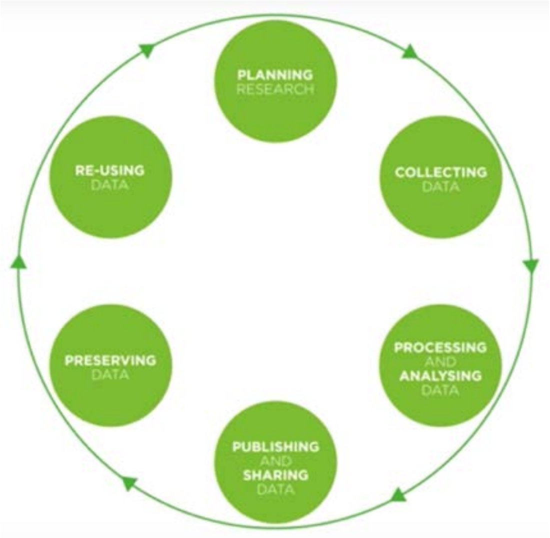


Figure 10. Research Data Life Cycle[43]

The 'Curation Lifecycle Model' as promoted by the Digital Curation Centre provides a high-level overview of the stages required for successful management of data (Higgins, 2008). Figure 11 contains the Curation Lifecycle Model. An example of a full lifecycle action of the model concerns the assignment and monitoring of administrative, descriptive, technical, structural and preservation metadata. Also, the preservation planning in relation to the lifecycle actions, the participation in and monitoring of the development of shared standards and tools are part of the model.

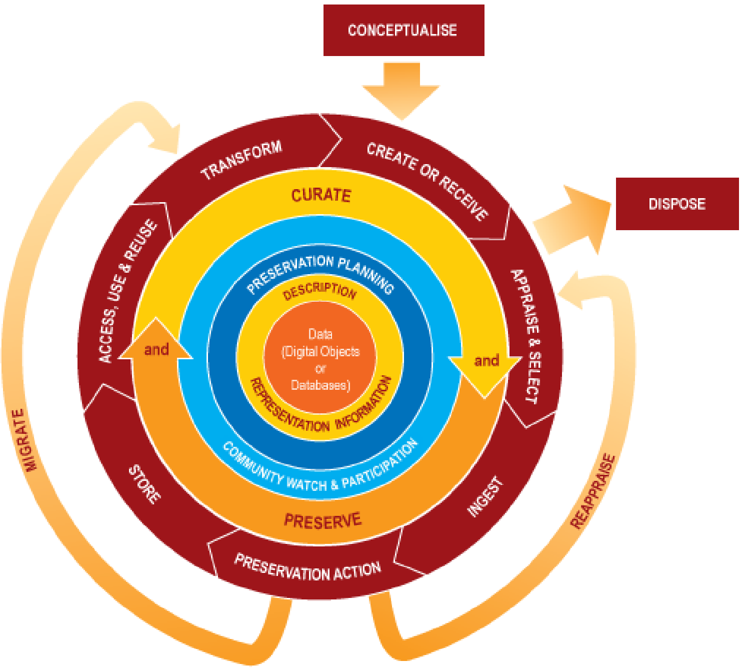


Figure 11. Curation Lifecycle Model

[43] Source: https://www.ukdataservice.ac.uk/manage-data/lifecycle.aspx

In the research data community, it is increasingly common that researchers create a Data Management Plan or DMP, often required by the funder of a research project. A DMP is a formal document that outlines how data are to be handled both during a research project, and after the project is completed. The goal of a DMP is to consider the many aspects of data management, metadata generation, data preservation, and analysis before the project begins; this ensures that data are well-managed in the present and prepared for preservation in the future.[44] Several templates and online DMP creation tools exist that can be used to formulate a DMP.[45] The main message of the DMP related activities is that it is important to act upfront and to keep the further lifecycle of the data into consideration.

3.3 Data protection, federated access and legal issues

In case archival material contains personal data, legal aspects in relation to data management policies must be taken into consideration. Oral History interviews, for instance, by definition contain personal data. Next to privacy protection also other legal issues are relevant, such as the formalization of licenses for using data.

With respect to data protection the EHRI report 'Digital handbook on privacy and access' contains valuable information (Luyten and Boers, 2013). The report relates to the EHRI portal that provides access to distributed Holocaust archives and connects information about Holocaust resources. In principle there are no restrictions to access the collection descriptions. However, special regulations should be followed in case the information contains personal data, that is information relating to a living individual who is or can be identified either from the data or from the data in conjunction with other information. In May 2018 the GDPR (General Data Protection Regulation) was established, a law applicable throughout the EU.[46] In general terms this law will protect personal data more strictly. For the use of personal data for scientific, historical and statistical research, the GDPR are supplemented by detailed rules that differ from country to country.

The approach described above takes the applicable laws and regulations as starting point and applies them to the data sources. It is also possible that data sources get a user license that enables its usage with as less as possible obstructions. An example is the formulation of an informed consent statement for data, preferably to make the data available as open access data, so without restrictions. Of course, the informed consent details must comply with all existing regulations and laws, be formulated by the appropriate stakeholders and signed by the interviewee. For instance, in case an interviewee who contributes to an oral history project is asked to give permission that the interview data may be used by a defined user community and this permission is expressed in an informed consent form and attached to the data objects. This will stimulate the access and usage of the data and minimize administrative actions required to open up the collection for usage. The Creative Commons initiative (see section 5.2) provides a list of types of licenses that can be used to formulate suitable rights obligations, such as details on the rights to distribute data objects and which obligations must be followed, for instance in relation to the acknowledgement of the rights holders.

In case legal aspects in relation to the protection of privacy and the assignment of a suitable user license are settled the actual access to the data should be supported by an Authentication and Authorization Infrastructure (AAI) framework. The authorized user of the

[44] Definition taken from the Science Europe Glossary, see: http://sedataglossary.shoutwiki.com/wiki/Data_management_plan

[45] See for instance DMPonline at: https://dmponline.dcc.ac.uk/ and https://easydmp.eudat.eu/plan/

[46] See: https://www.eugdpr.org/

data must be identified and right users should have access to the right data objects. eduGAIN[47] is an example of an international service interconnecting research and education identify federation services. It enables the secure exchange of information related to identity, authentication and authorisation between participating federations. The AARC project (Authentication and Authorisation for Research and Collaboration) is working on an integrated cross-discipline authentication and authorisation infrastructure.[48]

3.4 Archival data storage

The Digital Preservation Handbook states that *"the use of storage technology for digital preservation has changed dramatically over the last twenty years. During this time, there has been a change in practice. Previously, the norm was for storing digital materials using discrete media items, e.g. individual CDs, tapes, etc., which are then migrated periodically to address degradation and obsolescence. Today, it has become more common practice to use resilient IT storage systems for the increasingly large volumes of digital material that need to be preserved, and perhaps more importantly, that need to be easily and quickly retrievable in a culture of online access. In this way, digital material has become decoupled from the underlying mechanism of its storage"* (Digital Preservation Handbook, 2015). The average lifespan of different kinds of digital media varies, but it is clear it is much shorter than the lifespan of analog media such as paper and microfilm, provided that they are stored in an environment where the risk of damage and decay is reduced.

Storage for long term preservation concerns specific requirements above standard data storage solutions such as the creation of backups. Preservation storage systems require "*a higher level of geographic redundancy, stronger disaster recovery, longer-term planning, and most importantly active monitoring of data integrity in order to detect unwanted changes such as file corruption or loss"*. The Digital Preservation Handbook advises "*to apply a storage strategy that has the following characteristics. Multiple independent copies exist of the digital materials, geographically separated into different locations. The copies use different storage technologies and the copies use a combination of online and offline storage techniques. The storage is actively monitored to ensure any problems are detected and corrected quickly."*

The NDSA recommendation of digital preservation (see section 3.1) contains four levels of data storage in relation to its geographic location.

Level 1 (Protect your data)

- 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

Level 2 (Know your data)

- 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

Level 3 (Monitor your data)

- At least one copy in a geographic location with a different disaster threat
- Obsolescence monitoring process for your storage system(s) and media

Level 4 (Repair your data)

- 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

[47] See: https://edugain.org/

[48] See: https://aarc-project.eu/

3.5 Access to digital objects

Digital preservation is about access to digital objects in the future. But access today is also important as collecting and preserving digital objects without access is difficult to justify. The information system used by the institution to manage digital content determines how digital content can be accessed by a user (both humans and machines) e.g. via an on-site access service.

An often-applied strategy to provide long-term access to digital objects is to implement an institutional data provision service that facilitates that other services can get access to metadata and related digital objects.

This multimodal access principle is very well described by Owens. *"Given many of the inherent affordances of digital information, it makes sense that we will see a variety of assemblages, aggregations, interfaces, and methods for accessing and using digital collections. In this context, it makes sense to think about ways you will provide access to content instead of the way you will provide access. This increasingly involves thinking across a spectrum of access modes. One end of that spectrum is wholesale; methods for providing bulk access to collections as data sets and for thinking about consistent methods for uniformly providing access to all your content. On the other end of that spectrum is the boutique; providing highly curated sets of materials, special custom interfaces or modes of remediating or enhancing content for special use cases".* (Owens, 2017 p.105)

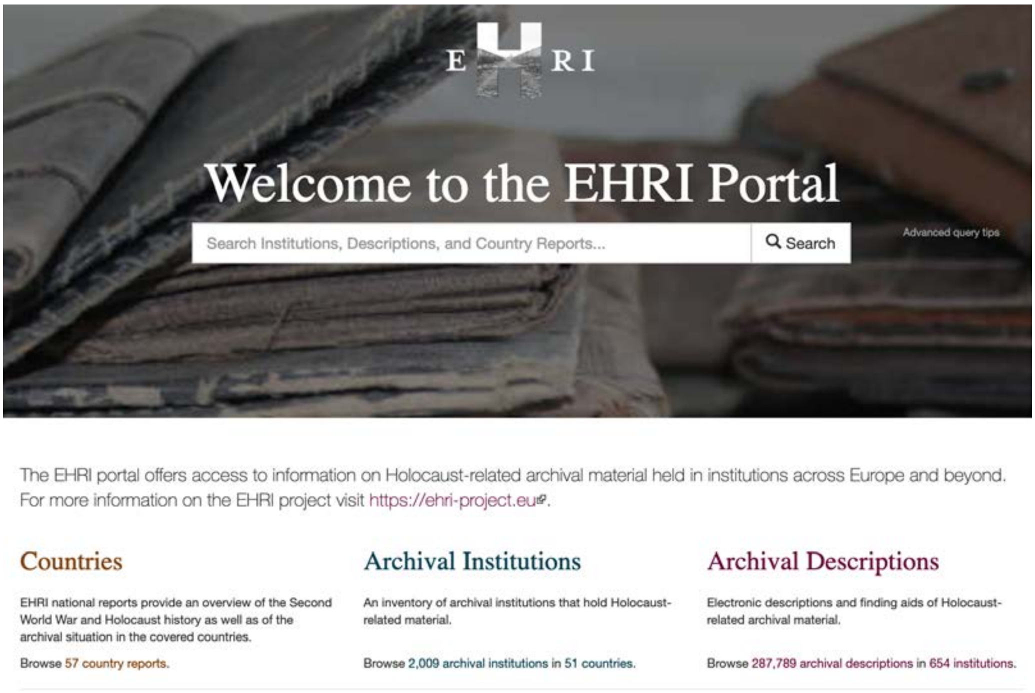
Chapter two of this report covers the key features of digital assets in order to be suitable for this multimodal way of access. This means the data files are stored in a standard data format, have a persistent identifier, are stored in a certified repository, are documented according to a standard metadata format, have a suitable usage license and - if applicable - are accessible by means of secured access service. Obviously, the most optimal access is realised with data objects that have an open access license. It is also possible that only the metadata is openly available and that the access to the related data objects is regulated by a specific usage license.

Data aggregation services have in common that they combine and enrich information that is provided by distributed institutions. In this way they enable the interoperability of digital collections of distributed archives. The advantage of this approach is that existing services and protocols can be used so it is not necessary to develop and maintain a dedicated one-off system that only disseminates "local" content. Of course there can be reasons to create a dedicated, proprietary access system to a specific collection e.g. to present the digital assets on a website with rich context information.

In general terms an access system based on the aggregator principle uses a harvesting protocol to collect information such as metadata from several sources and makes this available via a portal. In this information access model synchronisation between data provider and service provider is essential to keep the information up-to-date. Mutations in the digital collection of the data provider (such as new records added, records changed or records deleted) must be processed in the portal. A protocol that enables the harvesting of metadata from distributed archives is the OAI-PMH protocol.[49] The data provider (this is the local institute) must implement and facilitate the harvesting of the collection by supporting the OAI-PMH protocol. Service providers can automatically harvest the metadata of several institutes and create an online service based on the aggregated metadata.

[49] See: OAI-PMH stands for Open Archive Initiative Protocol for Metadata Harvesting https://www.openarchives.org/pmh/

The EHRI Portal[50], see figure 10, is an example of a portal that harvests information from distributed archives and that provide access information such as archival descriptions on the Holocaust. Other examples of aggregators are Europeana[51], aimed at providing access to European cultural heritage assets, Archives Portal Europe[52] for access to archival material from different European countries and the B2FIND[53] service, that harvests repositories with research data sets in a wide range of disciplines.



E RI
Welcome to the EHRI Portal
Search Institutions, Descriptions, and Country Reports...
Search
Advanced query tips

The EHRI portal offers access to information on Holocaust-related archival material held in institutions across Europe and beyond. For more information on the EHRI project visit https://ehri-project.eu.

Countries

EHRI national reports provide an overview of the Second World War and Holocaust history as well as of the archival situation in the covered countries.

Browse 57 country reports.

Archival Institutions

An inventory of archival institutions that hold Holocaust-related material.

Browse 2,009 archival institutions in 51 countries.

Archival Descriptions

Electronic descriptions and finding aids of Holocaust-related archival material.

Browse 287,789 archival descriptions in 654 institutions.

Figure 12. The EHRI Portal. Aggregator for metadata provided by distributed archives.

3.6 Use-case: Digital Preservation at the United States Holocaust Memorial Museum (USHMM)

This section contains a description of the digital preservation practices and policies of the United States Holocaust Memorial Museum.[54]

The case study provides a generally chronological narrative of some of the internal processes and decisions made by USHMM National Institute for Holocaust Documentation in managing, preserving, and making accessible a growing set of digital collections. At the current time the USHMM historical collection comprises over 85 million files and nearly a petabyte of data. The authors hope that others may benefit from our sharing some of the inner working processes that led to gradually improved processes and procedures and evolving maturity of our digital preservation activities over the last decade.

Early digital collecting

[50] See: https://portal.ehri-project.eu/

[51] https://www.europeana.eu/portal/en

[52] https://www.archivesportaleurope.net/

[53] http://b2find.eudat.eu/

[54] See: https://www.ushmm.org/

By 2006 it became obvious that magnetic storage media was dying. VHS, Betacam, and audio cassette tapes produced in the 1980s and 1990s were beginning to fail through physical degradation of the magnetic media. Prior to around that time, a common practice was to copy magnetic media to another similar physical medium, accepting the signal degradation inevitable with this type of operation with analogue media. However, it became evident that digitization would produce a single best copy that, at least in theory, would not degrade over time. The USHMM brought in an outside vendor to digitize over 7.000 hours of Oral Testimony. This (at the time) state-of-the-art system created around 10.000 Motion JPEG 2000 files (one file per physical tape medium), wrapped in the “Material Exchange Format” (MXF) and an equal number of MPEG-2 derivatives stored on four (4) 42-TB SAN units in the USHMM data center and replicated to nine (9) 18-TB storage units, which were dismantled and stored offsite. This early project established what would become the USHMM’s digital backup strategy for the next ten years: a single accessible location on spinning disk and one replication of the data on “dark” (infrequently accessed) storage media stored securely off site.

The success of the oral history digitization project prompted the Museum to launch another large-scale digitization initiative to duplicate the Museum’s microfilm collection. In the following five years the Museum digitized an estimated 20.000 reels of microfilm, producing approximately 20.000.000 digital files, one per microfilm frame. In addition to these large data producing projects driven internally by the IT Department and the Collections Management Division, smaller projects to digitize high quality versions of historical film, paper, and photographs were launched by work units throughout the Institution.

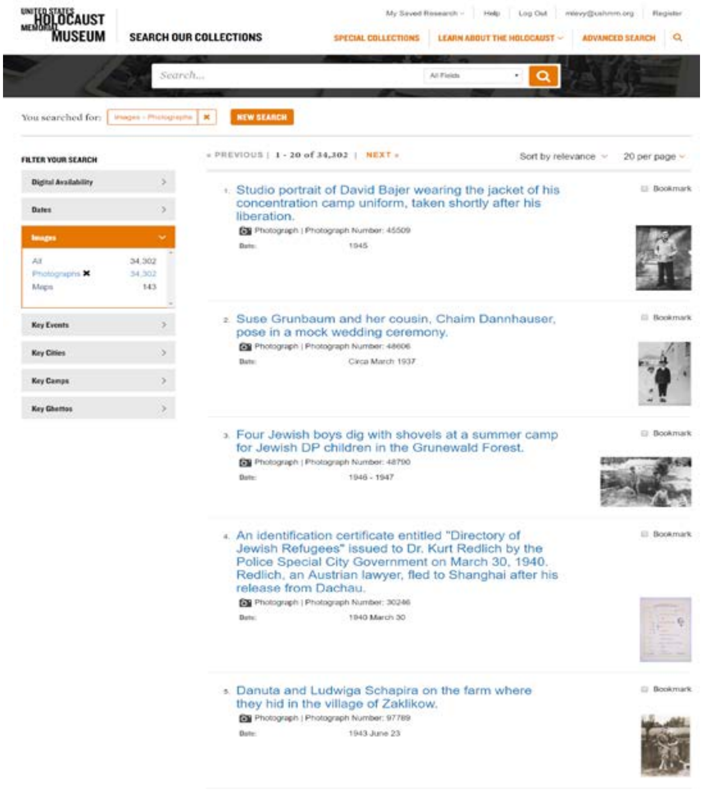
One of the Museum’s highest priorities has been to collect and make available reproductions of Holocaust evidence scattered throughout the globe to scholars, survivors, and the general public. Central to this effort is the Museum’s International Archival Program (IAP) division, which travels the world to locate and evaluate original documentation and arrange for its reproduction and acquisition. The IAP division accomplishes this through joint projects, purchases, and content exchanges. As the rest of the world became more digital, IAP acquisitions followed suit. Where previously USHMM staff would have acquired microfilm or photocopies of original holdings, increasingly they collaborated with source repositories in their digitization efforts and acquired copies that could be accessed both at the source repository and at USHMM. In the early years of 2005-2007, digital acquisitions represented a small percentage of total page acquisitions. Over the years, the percentage of digital acquisitions has grown, and lately digital acquisitions have largely replaced microfilm or photocopy duplication efforts. Today, the USHMM acquires around 10 million digital image files per year from other archival repositories. In addition to digitized archival paper collections, USHMM also acquires digitized copies of oral history, historical film, audio recordings, and photographs, and also continues to produces born-digital oral histories.

Processing Digital Collections

By 2008, the digital collection had grown to over 300 terabytes (TB) and, was distributed across multiple storage devices, and the collection was growing rapidly through new digital acquisitions and digitization efforts engaged in by various work units. The National Institute for Holocaust Documentation formed the Digital Asset Management and Preservation Division (DAMP) to serve two main purposes. First, DAMP Division provides centralized oversight and harmonization of digitization and digital acquisition activities and is responsible for management and preservation of all digital files. Second, DAMP develops and maintains Collections Search, which is the search and discovery tool for accessing all Collections catalogs and descriptions and of accessing digital files. The public web version of Collections Search is available on the web at https://collections.ushmm.org. Figure 13 contains a screen of the web interface of Collections Search. An internal version provides access to certain additional materials that cannot be published on the public web due to restrictions. These restrictions may be due to a contractual relationship in the case of an archival collection copied from another institution, or a copyright restriction, or a restriction related to privacy.

The Digital Collections Division (now the Digital Asset Management and Preservation Division, DAMP) was formed in 2008 within the Office of Collections (now the National Institute of Holocaust Documentation). Within the first 4 years, the new division started with basic operations to organize and take better control of the digital holdings. They secured the filestore ensuring only a limited number of users had write-access, took over incoming copy procedures, created a database to track digital objects, introduced format and validity checking, and re-organized researcher access points into a user-friendly logical arrangement. Through these initial efforts, it became clear that digital repository software would be essential to systematically validate file characterization, extract technical metadata, capture file format information, validate checksums, assign system neutral unique identifiers, manage storage migrations, and other critical operations better performed by a machine than human.

During this period, division staff continued to use command-line and desktop tools to produce inventories and checksums of disk files, and they developed policies and procedures to compare the output of these files on a regular basis to confirm that files were not lost or modified. They also began assessments of well-known commercial and open-source tools against the current and planned future needs of the institution, and began planning for a future enterprise-level preservation system.



You searched for: Images › Photographs

« PREVIOUS | 1 - 20 of 34,302 | NEXT »   Sort by relevance   20 per page

All	34,302
Photographs ✖	34,302
Maps	143

1. Studio portrait of David Bajer wearing the jacket of his concentration camp uniform, taken shortly after his liberation.
   Photograph | Photograph Number: 45509
   Date: 1945
2. Suse Grunbaum and her cousin, Chaim Dannhauser, pose in a mock wedding ceremony.
   Photograph | Photograph Number: 48606
   Date: Circa March 1937
3. Four Jewish boys dig with shovels at a summer camp for Jewish DP children in the Grunewald Forest.
   Photograph | Photograph Number: 48790
   Date: 1946 - 1947
4. An identification certificate entitled "Directory of Jewish Refugees" issued to Dr. Kurt Redlich by the Police Special City Government on March 30, 1940. Redlich, an Austrian lawyer, fled to Shanghai after his release from Dachau.
   Photograph | Photograph Number: 30246
   Date: 1940 March 30
5. Danuta and Ludwiga Schapira on the farm where they hid in the village of Zaklikow.
   Photograph | Photograph Number: 97789
   Date: 1943 June 23

Figure 13. USHMM Collections Search

Difficult Engineering Tradeoffs

By 2013, the Museum had half-a-petabyte of data (a petabyte including redundant copies). The digital collection consisted of some 150 TBs of digitized Oral Histories (and had begun a new project to digitize another 8.200 hours, which would add to that figure considerably); 300 TBs of microfilm digitized as uncompressed TIFF, and 50 TBs of digital acquisitions in various formats. Existing storage area network (SAN) storage units which housed all the digital content were approaching their end-of-life. Assessing the Institution's 5 year growth projections to anticipate future needs and evaluate storage replacement options. At that growth rate, the pace of acquisition and compression choices were not sustainable. The Museum would have produced over 3 petabytes of single-copy data in only 5 years, well beyond the ability to manage them. The Museum started seriously evaluating digitization format and compression choices and developed criteria to determine long-term preservation responsibility.

DAMP staff introduced the concept of “asset" and "instance" in order to help classify materials according to their preservation requirements. In short, an "asset" is a digital object that the Museum has preservation responsibility over and should therefore be captured and maintained at the highest foreseeable and practical quality and the lowest practical compression. This is contrasted with a digital “instance,” which is defined for this context as a digital object that that the Museum holds essentially for researcher convenience and access. An "asset" is irreplaceable and if lost is essentially lost for all time. An "instance," if lost, could realistically be replaced, even if doing so would require some effort. The reason the differentiation between "asset" and "instance" is that it was not practical economically to preserve every digital file held by the Museum at the very highest quality levels. There are common analogs in the in the physical world where some museums may distinguish between their accessioned collection objects, which are subject to the highest practices of conservation and protected against any threat to their preservation, versus their "study collection" which comprise objects considered less uniquely valuable and are allowed to be handled and may be used for various purposes even at the risk of some wear and tear. The "asset/instance" distinction led to the retirement of 300 TBs of digitized TIFF Microfilm with a 20% compression JPEG retained for users. This shift cut 5-year storage projections by 60%.

In addition, preservation responsibility criteria was introduced for digital copy collections (see figure above) to help guide acquisition staff in their format choices. With the reduction in overall storage infrastructure required, the museum secured about 800 TB of high-end network attached storage units featuring distributed redundant file storage and using Reed–Solomon error correction coding. By 2014, the Museum had migrated all digital content off the aging SAN units.

Bring in the Experts

In late 2014, the Museum secured funding and released a request for proposals to do an environmental scan of the USHMM and develop a roadmap for enterprise digital preservation and management.

In 2015, the Museum awarded a contract to an outside consultancy specializing in cultural heritage digital preservation and digital asset management to study the digital landscape at the Institution. In early 2016 the consultants produced a detailed findings report, outlined mission critical systems such as a digital preservation system, a digital asset management system, and a document management system, which the Museum would need to select and implement, as well as various scenarios for prioritizing future work.

The Museum determined that despite the need for all the systems outlined in the findings report (most notably digital asset management, document management, and digital preservation), securing a digital preservation system was the critical next step toward safeguarding digital content well into the future. Armed with reams of evidence, talking points, and a solid vision, the Museum started internal and external educational and fundraising efforts and began the detailed process of producing requirements to contract a digital preservation system. The museum released a public RFP (request for proposal) and began evaluating solutions bid by various vendors. By October 2017, the Museum secured preservation repository software and awarded a contract to a commercial provider of digital preservation systems that can be implemented to become OAIS and ISO 16363:2012 (Trustworthy Digital Repository) compliant.

The Museum and commercial digital preservation system vendor have spent the last year configuring the software, establishing an isolated network storage architecture, purchasing storage systems, creating preservation plans and policies by material type, and preparing content for ingest into the system.

Implementation of Digital Preservation System

Those implementing the digital preservation system were required to incorporate many factors and practices. The network and routing was designed around maximum safety and isolation of the preservation system and network traffic between and among the system components. System user identification and rights were designed with safety in mind, and system monitoring systems were installed and tested so that proper operation is ensured and any exceptions cause an alarm. The system is planned to have three copies of each object in the system: one each in two physical systems geographically dispersed plus one in a cloud environment.

Staff who were most familiar with the nature and type of each of the digital subcollections established patterns that were acceptable for each type of collection. A so-called "Technical Preservation Plan" (TPP) is a feature of the digital preservation system and one was developed for each subcollection. At the time of writing, there are 13 TPPs in the USHMM digital preservation system, and this may expand as future subcollections are required. The TPP can be used to approve or reject any digital file at the time of ingest and includes several features. At the time of ingest, for each object the system queries the Museum cataloging system to obtain metadata elements including such items as collection title; collection identification strings (such as Accession Number or Record Group Number or Photo ID Number); metadata relating to rights to access, use, and publish; and several other metadata elements depending on the active TPP. Any changes may be migrated and tracked over time.

Some digital subcollections are very well organized and controlled at the time of their creation and throughout their lifecycle, and were digitized using very orderly methods. Such practices include file naming conventions so that the name of each file could be easily interpreted to provide information such as, for example, the identifier of the collection and the ordinal number of the tape within an interview, or the collection identifier, microfilm reel, and frame number of digitized microfilm collection. In such cases, files could be rejected as problematic if the file name does not match a certain pattern. In those cases, the TPP can include file name restrictions. In addition, the number of file types are very limited and the TPP also engages file type identification and verification steps.

In other TPPs, for example where the Museum acquired collections digitized by other institutions, the digital objects will have file type and file name characteristics determined by the originating institutions, and the Museum has no control over these. In such cases the TPP for that type of collection must be much more relaxed, allowing a wider range of file name patterns and a wider range of file types. Such flexibility in the system allows the greatest control and quality assurance that is practical for each subcollection.

Today, the Museum is well underway toward a supported trustable digital preservation environment. At the time of writing, the USHMM has ingested about half of its 850 TBs. Staff project the ingest and validation of all existing digital content to complete in the summer of 2019.

Summary

Digital preservation can be a complex concept for many to grasp and, although it is associated in some ways with providing access to users, many aspects of the risks and the resources required to mitigate risks are often not immediately obvious. Just as in preservation of physical items, digital preservation is never completed and can never be spoken of in the sense of having been completed. Rather, digital preservation is an ongoing journey of eternal vigilance and constant evaluation of practices against threats of all kinds.

References

Caplan, Priscilla, Understanding PREMIS, https://www.loc.gov/standards/premis/understanding-premis-rev2017.pdf, Library of Congress, 2017

Corti, L., V. van den Eynden, L. Bishop, M. Woollard, Managing and sharing data: a guide to good practice. Sage, 2014

Daelen,Veerle Vanden, 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 (2016) 7-8. https://hal.inria.fr/hal-01281442v2, 2016

Digital Preservation Handbook, 2nd Edition, http://handbook.dpconline.org/, Digital Preservation Coalition, 2015

Dollar, Charles and Lori Ashley, Digital Preservation Capability Maturity Model (DPCMM), version 2.7. https://www.statearchivists.org/resource-center/resource-library/digital-preservation-capability-maturity-model-dpcmm/, 2015

Higgins, Sarah. “DCC Curation Lifecycle Model”, International Journal of Digital Curation. Vol. 3/1, 2008, 134-140. http://dx.doi.org/10.2218/ijdc.v3il.48

Horik, René van; Jong, Tonke de; Brazzo, Laura; Green, Jessica; Levy, Michael; Neumann, Effi, Nispen, Annelies van; Uiterwaal, Frank. D13.3 Data management planning for long-term preservation, Deliverable EHRI-project. https://ehri-project.eu/ehri-deliverables, 2017

Horsley, Nicola; Priddy, Mike. D3.1 KPLEX – Report on Historical Data as Sources https://kplex-project.eu/deliverables/, 2018

Lavoie, Brain. The Open Archival Information System (OAIS): Introductory guide (2nd edition) DPC Technology Watch Report 14-02 October 2014. DOI: http://dx.doi.org/10.7207/twr14-02

Luyten, Dirk and Hans Boers. D3.2 Digital handbook on privacy and access (2013). Deliverable EHRI-project. Online available at: https://ehri-project.eu/ehri-deliverables

McMurry J.A., Juty N., Blomberg N., Burdett T., Conlin T., Conte N., et al. Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. PLoS Biol 15/6v, 2017. See: https://doi.org/10.1371/journal.pbio.2001414

OAIS. Reference model for an Open Archival Information System (OAIS) Magenta Book, June, 2012 https://public.ccsds.org/Pubs/650x0m2.pdf

Owens, Trevor. The Theory and Craft of Digital Preservation. Manuscript submitted to Johns Hopkins University Press, 2017. https://doi.org/10.31229/osf.io/5cpjt

PREFORMA Handbook. Validating formats. A prerequisite for preserving digital objects. http://www.preforma-project.eu/uploads/dissemination/pfo_roadmap_print.pdf. PREFORMA Project, 2017

Riondet, Charles, et. al. D11.4 Report on standards, 2017. Deliverable EHRI-project. Online available at: https://ehri-project.eu/ehri-deliverables

Waters, Donald; Garrett, John. Preserving Digital Information, Report of the Task Force on Archiving of Digital Information (5/1996, 59 pp.) ISBN 1-88733450-5. Online available at: https://www.clir.org/pubs/reports/pub63/

Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18, 2016