European Holocaust Research Infrastructure Preparatory Phase

H2020-INFRADEV-2019-2

GA no. 871060

Deliverable 7.3

Data Management Plan EHRI RI

Author(s)

René van Horik

Elly Dijk

Louise Bezuidenhout

DANS-KNAW

Mike Bryant

Tobias Blanke

KCL

Start: December 2019 [M1]

Due: November 2021 [M24]

Actual: December 2021 [M25]

Document Information

Project URL	https://www.ehri-project.eu
Document URL	https://www.ehri-project.eu/deliverables-ehri-pp-2019-2022
Deliverable	D7.3 Data Management Plan EHRI RI
Work Package	WP7
Lead Beneficiary	1 - KNAW
Relevant Milestones	MS2
Dissemination level	Public
Contact Person	René van Horik (rene.van.horik@dans.knaw.nl)
Abstract (for dissemination)	This deliverable contains an initial Data Management Plan (DMP) for the future EHRI-Research Infrastructure (EHRI-RI), currently in its preparatory phase and due to be established by 2025. The DMP builds on the FAIR data principals, which provider guidelines for making data Findable, Accessible, Interoperable and Reusable, along with the policies and services of the European Open Science Cloud (EOSC) and the European research data policy and funding landscape in general.
Management Summary	This deliverable provides an initial Data Management Plan (DMP) for the permanent EHRI Research Infrastructure (henceforth EHRI-RI). Since the EHRI-RI is still in its preparatory phase and concurrent activities in both EHRI-PP and the EHRI-3 project will play a major role in defining its eventual organisational structure and the services that will be provided, this plan is both descriptive of EHRI's current data practices and prescriptive as to actions and processes that will be established in the future.

The scope of this document encompasses EHRI-RI first-order outputs, and does not at this time incorporate data generated or collected as a result of EHRI-PP project management activities. Outreach to project stakeholders and subsequent updates to this plan will be a part of the document lifecycle and scope will be |

subject to future reevaluation.

The building blocks of the plan are threefold: the FAIR data guidelines that seek to ensure research data is Findable, Accessible, Interoperable, and Reusable; the European Open Science Cloud (EOSC), which supports many projects upon which EHRI-RI can build; and Research Data Management (RDM) guidance provided by the European research data policy and funding landscape.

The DMP itself consists of an enumeration of the main types of data that will be in use and a brief description of their characteristics, followed by a description of how the FAIR guidelines are met and steps that will be taken to ensure compliance. Data types include but are not limited to: descriptions of institutions and archival material, reports on the archival situation in relevant countries, user-contributed and user-generated data, various geospatial information, and mixed-media material on the EHRI Document Blog and Digital Editions.

To ensure findability, all data types will reside in searchable repositories and assigned identifiers suited to appropriate level of surety, including Archival Resource Keys (ARKs) for descriptions of archival material and Digital Object Identifiers (DOIs) for material created by EHRI itself.

The EHRI-RI will endeavour to make material available in machine-readable structured forms and data made accessible this way will, except in cases that conflict with security, privacy or contractual requirements, have predictable and signposted deprecation timelines.

Interoperability will be ensured by the use of standardised metadata formats such as EAD, EAC, EAG, and SKOS, delivered where feasible using standard communication protocols such as the Open Archives Intiative's Protocol for Metadata Harvesting (OAI-PMH).

Open access is the default mode for material made available by EHRI activities and the majority of material can be reused either with no restrictions, or with attribution. In future, material made available by the EHRI-RI will be accompanied by standardised licenses, such as those made available by Creative Commons, clarifying reuse requirements where relevant.

Responsibility for data management and FAIR compliance will reside with the EHRI-RI technical leadership. A data management committee, comprised of relevant stakeholders, will be formed in 2022 and responsible for periodically reviewing EHRI-RI data management practices.

Since future EHRI-RI activities may involve data collection or digitisation a DMP template has been provided in Appendix A that is tailored for these use-cases, and those of affiliated efforts at the individual researcher level, to encourage effective RDM.

History

Version	Date	Reason	Revised by
0.1	21/09/2021	Initial version	René van Horik
0.2	18/10/2021	Integrated more EHRI-specific data management information about FAIR compliance	Mike Bryant

Table of contents

Glossary

API	Application Programming Interface
ARK	Archival Resource Key
CPA	Content Provider Agreement
CHI	Collection Holding Institution
DMP	Data Management Plan
DOI	Digital Object Identifier
EAD	Encoded Archival Description
EOSC	European Open Science Cloud
FAIR	Findability - Accessibility - Interoperability - Reusability
GDPR	General Data Protection Regulation
ISAD(G)	International Standard for Archival Description (General)
ISAAR	International Standard Archival Authority Record
ISDIAH	International Standard for Description of Institutions with Archival Holdings
LOD	Linked Open Data
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
ORCID	Open Researcher and Contributor IDentifier
ORDP	Open Research Data Pilot
PID	Persistent IDentifier
RDF	Resource Description Framework
RDM	Research Data Management
SKOS	Simple Knowledge Organisation System
TEI	Text Encoding Initiative
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
XML	eXtensible Markup Language

1. INTRODUCTION

The EHRI-PP project is responsible for designing the organisational, governance, and financial structures for a permanent EHRI Research Infrastructure (henceforth EHRI-RI), as well as its access and data management policies. A data management plan (DMP) is a formal document that outlines what data is used by an organisation and how it is handled. The goal of a DMP is to consider the many aspects of research data management, metadata generation, data preservation, and data analysis before the research activities begin to ensure that data is well-managed in the present, and prepared for preservation in the future.[1]

The main goal of the EHRI-RI DMP concerns the "long-term operation of EHRI" and will include an enumeration of the types of data the RI will handle based on current projections, as well as guidelines and policies that will facilitate the long-term usability of that data. The EHRI-RI DMP must be based on data policies and approved by the management of the EHRI-RI. As the EHRI-RI is currently in its preparatory phase the DMP in this deliverable must be aligned with the future data policies and governance models, which will result in an adaptation and extension of the practices detailed in this deliverable.

2. TIMELINE

EHRI is seeking to become a permanent research infrastructure. The most likely legal form will be the ERIC, the EU-designed European Research Infrastructure Consortium. In general, an ERIC will be sustained by fees paid by member countries. Each country that participates in the future ERIC will also need to develop a national node, which is essentially a consortium of Holocaust-related research and archival institutions in a given country. Currently, there are still many details to be finalised in the realisation of the EHRI-RI.

Year	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
EHRI1	▨	▨	▨	▨	▨	▨										
EHRI2						▨	▨	▨	▨	▨						
EHRI3												▨	▨	▨	▨	
EHRI PP										▨	▨	▨	▨			
EHRI RI																▨

Table 1*: Timeline of EHRI projects*

The timeline in Table 1 provides a general overview of the duration of the EHRI projects that will form the basis for the establishment of a permanent organization in 2025. Concerning the formulation of a DMP the following dates and Deliverables are relevant.

● April 2021 -> D12.1 EHRI-3 Initial Data Management Plan

---

1 Definition from Science Europe Data Glossary:
http://sedataglossary.shoutwiki.com/wiki/Data_management_plan

- November 2021 -> D7.3 EHRI-PP Data Management Plan

- October 2023 -> D12.2 EHRI-3 Updated Data Management Plan

- 2025 -> EHRI-RI established

As the period between the provision of the EHRI-PP DMP and the establishment of the EHRI-RI is at least 3 years, the final EHRI-RI DMP will be an update of the DMP presented in this deliverable.

3. DOCUMENT SCOPE AND CHARACTERISTICS

This document is focused on first order EHRI-RI activities involving research (meta)data and digital objects connected to Holocaust-related archival material, as well as the operation of services and data collected about users of the future EHRI-RI services. Out of scope for this document are activities relating to the management, planning, design and implementation of the EHRI-RI itself, as well as the data management practices of EHRI partner institutions. The document scope may be reevaluated in future iterations of this DMP.

3.1. Relationship to EHRI-3 DMP

The EHRI-3 project will run from 2020 to 2024. It deepens the integration of Holocaust archives and research that has been undertaken by EHRI since 2010 when EHRI-1 was first established. The initial version of the EHRI-3 DMP provides an overview of the data objects used and created in the EHRI consortium and describes how they are managed. Moreover the DMP gives advice on how the management of the data objects can be organised and managed in a more durable way. The EHRI-3 DMP can be characterized as primarily descriptive.

The EHRI-PP project will run from 2019 to 2022 and aims to transform the EHRI project into a permanent organisation that will help secure the future of trans-national Holocaust research, commemoration and education. The EHRI-PP DMP can be characterized as primarily prescriptive as it aims to provide the building blocks on which future permanent RI data management policies can be based. These building blocks will be presented and discussed further on in the deliverable. The communication of the EHRI-PP DMP will be supported by workshops and training organised by WP5 in EHRI-PP.

4. UPDATING AND OUTREACH

A DMP is not a fixed, static document. New developments and new requirements appear that will influence the status of the current DMP. Periodically the DMP has to be updated to meet these new requirements. Moreover, outreach, training and support to ensure continued compliance with data management policies are important to make the DMP effective. Feedback on the training and support activities for stakeholders of the DMP will also result in a need to adjust and modify a DMP. This process is illustrated in FigureFigure 1. Reviewing and updating the DMP on a yearly basis is advised, and the update process should take into consideration both suggestions provided by the users of the DMP as well as emerging relevant tools, services and standards.

5. BUILDING BLOCKS OF THE EHRI-RI DMP

The EHRI-RI DMP will be based on 3 building blocks. These are the FAIR data principles, output from initiatives related to the European Open Science Cloud (EOSC) and research data management guidance (see Figure 1.) This section provides background information on these building blocks.
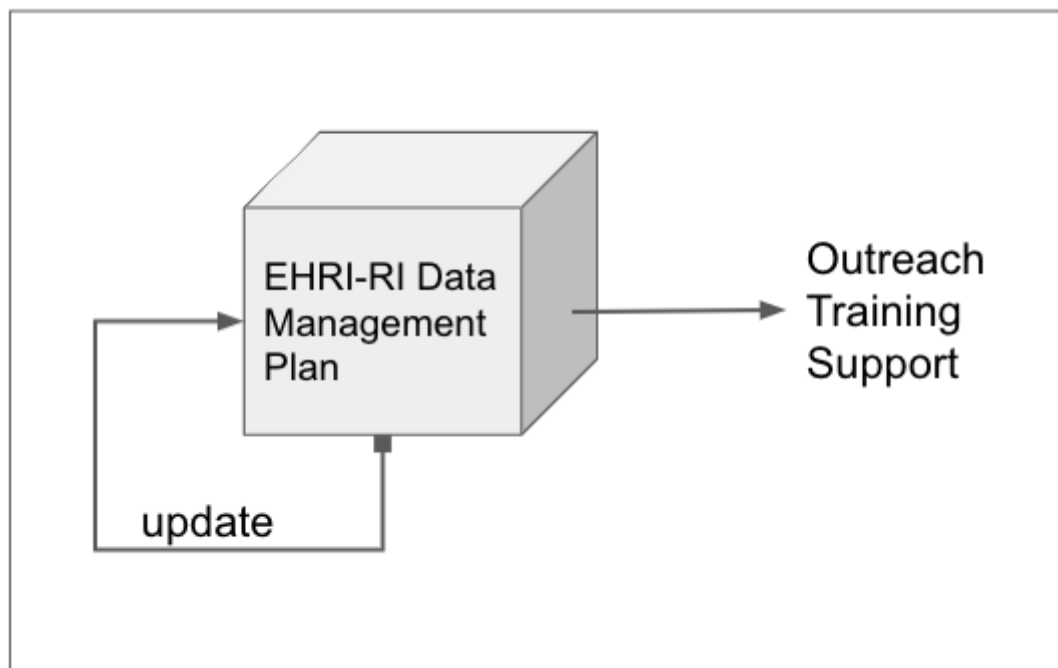
Figure 1*: EHRI-RI DMP update and outreach feedback process.*

5.1. FAIR Data Principles

The FAIR data principles intend to improve the findability, accessibility, interoperability and reusability of digital assets (see Figure 2.) For each of the four principles practical implementation guidelines are stated. The figure below can act as a reference of the meaning of the FAIR principles. The compliance list contains action points that help to assess the "FAIRness" of data collections.

Figure 2*: FAIR data principles. Source: https://ccafs.cgiar.org/open-access-and-fair-principles*

5.1.1. Findable

The first step in facilitating the (re)use of data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata is essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

- F1. (Meta)data are assigned a globally unique and persistent identifier
- F2. Data are described with rich metadata (defined by R1 below)
- F3. Metadata clearly and explicitly include the identifier of the data they describe
- F4. (Meta)data are registered or indexed in a searchable resource

5.1.2. Accessible

Once the user finds the required data, they need to know how they can be accessed, possibly including an authentication and authorisation procedure.

- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
- A1.1 The protocol is open, free, and universally implementable
- A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
- A2. Metadata are accessible, even when the data are no longer available

5.1.3. Interoperable

The data usually needs to be integrated with other data. In addition, the data needs to interoperate with applications or workflows for analysis, storage, and processing.

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

5.1.4. Reusable

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

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

The principles refer to three types of entities: data (or any digital object), metadata (information about that digital object), and infrastructure. For instance, principle F4 defines that both metadata and data are registered or indexed in a searchable resource, the infrastructure component.

5.2. European Open Science Cloud - EOSC

The European Open Science Cloud (EOSC) can support the realisation of the future EHRI-RI. The European Commission provides the following background information on the EOSC:

> *The ambition of the European Open Science Cloud (EOSC) is to provide European researchers, innovators, companies and citizens with a federated and open multi-disciplinary environment where they can publish, find and re-use data, tools and services for research, innovation and educational purposes. This environment will operate under well-defined conditions to ensure trust and safeguard the public interest. The European Open Science Cloud (EOSC) ultimately aims to develop a 'Web of FAIR Data and services' for science in Europe upon which a wide range of value-added services can be built. The implementation of the EOSC is based on a long-term process of alignment and coordination pursued by the Commission since 2015 with the many and diverse stakeholders of the European research landscape. The current phase of implementation (2021-2030), is taking place according to a Strategic Research and Innovation Agenda (SRIA) which is co-developed with the entire EOSC community. EOSC is transitioning to a more stakeholder-driven approach with a shared vision, common objectives and complementary contributions at European, national and institutional levels.*[2]

The implementation phase of the EOSC started in 2021, so at this stage it can be expected that the coming years new developments will influence the direction and appearance of the "EOSC-environment". The preparatory phase of EOSC (that started in 2015) has resulted in outputs relevant for the EHRI-RI DMP. The EOSC-portal[3] contains an actual overview of

---

2 https://ec.europa.eu/info/research-and-innovation/strategy/strategy-2020-2024/our-digital-future/open-science/european-open-science-cloud-eosc_en
3 https://eosc-portal.eu/

services, use cases, projects and other resources that feed into future iterations of the EHRI-RI DMP.

The development of the EOSC is supported by more than 50 projects funded by the EU.[4] Several of these projects have created outputs that are relevant for the EHRI-RI DMP. Table 2 contains a list of projects funded in the Horizon 2020 program that created results that can be used in the data management planning of the EHRI-RI. Most projects have already ended, but some of them will run until 2023. For each project in the table a short description is given of its benefit for EHRI-RI. The implementation phase of the EOSC will also be supported by EU funded projects in the coming years and they are potentially relevant for the EHRI-RI.

Project Acronym	Project Name	Start Date	End Date	Website
DICE	Data Infrastructure Capacity for EOSC	01/2021	06/2023	www.dice-eosc.eu
	DICE provides European wide data services integrated with community platforms. A future EHRI-RI might evolve into a community platform for which these data services are relevant.			
EOSC-Future	EOSC Future will establish a trusted platform with open and FAIR data, resources and services for all scientific disciplines.	04/2021	09/2023	https://eoscfuture.eu
	The main relevance of this project for EHRI is that the project will further develop the EOSC-portal (with potential relevant services for EHRI) and training.			
EOSC-hub	Integrating and managing services for the European Open Science Cloud	01/2018	03/2021	eosc-hub.eu
	The EOSC-hub project has contributed to the establishment of the EOSC-portal and EOSC-market place. It has developed business and sustainability models for data services and training material for data management and other issues related to the research life cycle.			
EOSC-	EOSC-Nordic	09/2019	08/2022	www.eosc-nordic.eu

---

4 https://www.eoscsecretariat.eu/eosc-projects-list

Nordic	EOSC-Nordic aims to facilitate the coordination of European Open Science Cloud (EOSC) relevant initiatives within the Nordic and Baltic countries.			
EOSC-Pillar	Coordination and Harmonisation of National Initiatives, Infrastructures and Data services in Central and Western Europe	06/2019	06/2022	www.eosc-pillar.eu
	EOSC-Pillar aims to coordinate national open science efforts in several European countries and to harmonize them by contributing to the EOSC. As EHRI also consists of organisations in different countries, the coordination efforts by EOSC-Pillar can be of value for the EHRI-RI. (Countries active in EOSC-Pillar: Austria, Belgium, France, Germany, Italy).			
EOSC-Synergy	European Open Science Cloud - Expanding Capacities by building Capabilities	09/2019	02/2022	www.eosc-synergy.eu
	ESOC-Synergy is, just as EOSC-Pillar, a cluster project that aims to harmonize efforts from different countries and to implement the EOSC at a national level (Countries active in EOSC-Synergy: Spain, Portugal, UK, Czech Republic, Germany, Slovakia, Poland and the Netherlands).			
EOSC pilot .	The European Open Science Cloud for Research Pilot Project	01/2017	05/2019	eoscpilot.eu
	The project has been funded to support the first phase in the development of the EOSC			
FAIRsFAIR	Fostering Fair Data Practices in Europe	03/2019	02/2022	fairsfair.eu
	FAIRsFAIR aims to supply practical solutions for the use of the FAIR data principles throughout the research data life cycle. Several guidelines, recommendations, training and services can be used in research data management aspects of EHRI-RI. An example is a service with which the "FAIRness" of a data collection can be assessed.			
FREYA	Connected Open Identifiers for Discovery, Access and Use of	12/2017	11/2020	www.project-freya.eu

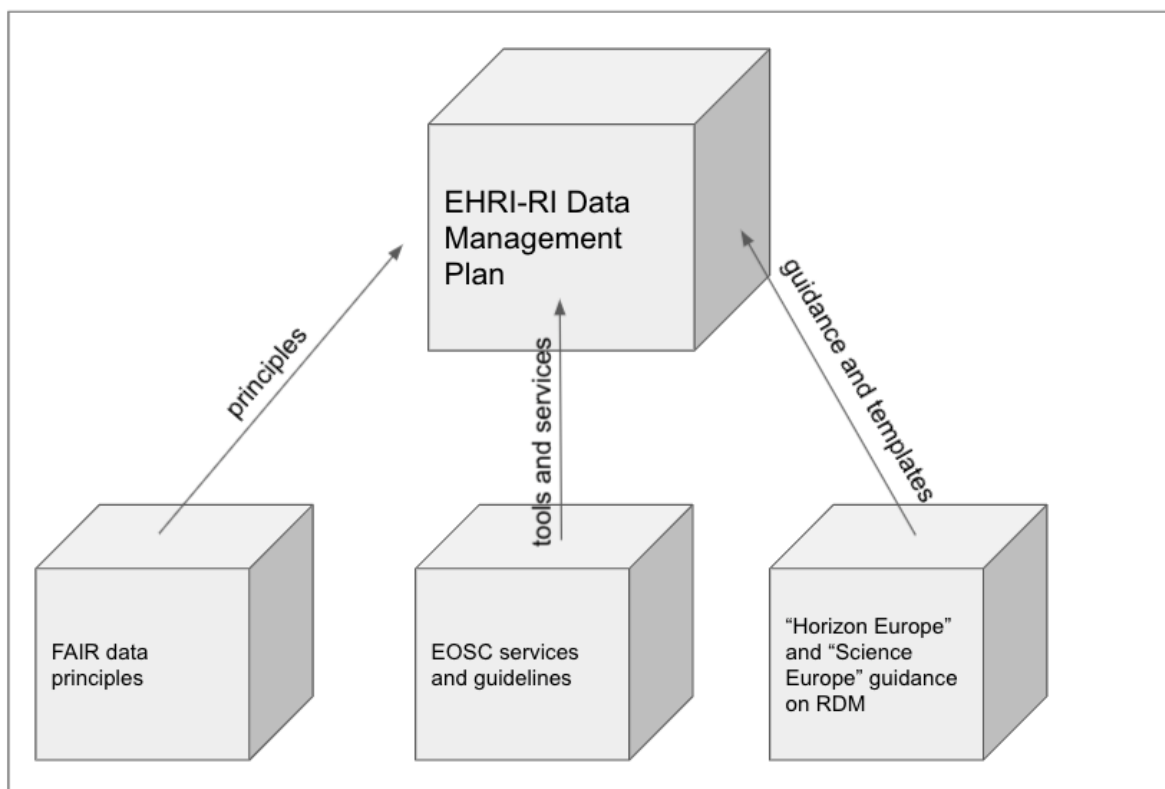|				
	Research Resources			
	The FREYA project aimed to extend the infrastructure for persistent identifiers (PIDs) as a core component of open research, in the EU and globally. PIDs certainly will be an important component of the EHRI-RI DMP.			
OpenAIRE Advance	Support Open Access/Open Data mandates of the European Commission, by sustaining current infrastructure, aiming to be a trusted e-Infrastructure within the realms of the European Open Science Cloud.	01/2018	02/2021	https://www.openaire.eu/openaire-advance-project
	OpenAIRE-Advance strived to empower National Open Access Desks (NOADs) so they become a pivotal part within their own national data infrastructures, positioning Open Access and Open Science onto national agendas. OpenAIRE provided RDM training and developed training materials in the field of RDM. The EHRI-RI can benefit both from the expertise of the NOADs and from the training materials developed in the project.			
RDA Europe 4.0	Research Data Alliance	03/2018	05/2020	www.rd-alliance.org
	The RDA is a global community driven initiative to enable the open-sharing and reuse of data. The outputs of several RDA Working Groups and Interest Groups can help to compile the EHRI-RI DMP.			
SSHOC	Social Sciences & Humanities Open Cloud	01/2019	04/2022	sshopencloud.eu
	SSHOC is a project that unites several international partner organisations in developing the social sciences and humanities area of the EOSC. As EHRI is active in the humanities the work done in SSHOC is relevant for the EHRI-RI DMP. Examples are activities in the certification of data repositories, the harmonization of metadata and ontologies, and tools to analyse humanities research data.			

Table 2*: EOSC-projects relevant for the EHRI-RI DMP*

5.3. Guidance on Research Data Management

The third building block of the EHRI-PP DMP concerns guidance activities on research data management provided by the European research data policy and funding landscape. There, consensus is formulated on the details of a data management plan.

The European research and innovation funding program until 2027, Horizon Europe[5], has created a DMP Template that is part of a project grant agreement.[6] This DMP template is relevant for EHRI-PP in two ways. First, in case EHRI-RI applies for funding in the Horizon Europe program this DMP template has to be followed. Secondly, the DMP template can be used to manage data that the EHRI-RI receives from data providers, such as collection holding institutes or researchers. For instance, data that will be provided by research that is funded or in other ways supported by the EHRI-RI. If, for instance, a research project, affiliated with EHRI-RI creates interview-data, a DMP has to be provided in which details on the storage and accessibility of the datasets is provided. In the Horizon Europe DMP template the FAIR data principles are a significant component.

Another source of guidance on research data management are the activities in this field of Science Europe. Science Europe is an association of major Research Funding Organisations (RFOs) and Research Performing Organisations (RPOs).[7] Research data management is a key policy priority addressed by Science Europe, for which guidance and support is provided.[8]

Figure 3. *Building Blocks of the EHRI-RI DMP*

6. EHRI-RI DATA MANAGEMENT PLAN

The preceding sections present the context and building blocks of the EHRI-RI DMP. This section provides an overview of the main data types used in the EHRI-RI and describes how the FAIR data principals will be complied with in practice.

Future digitisation and research data gathering activities taking place as EHRI-RI activities will ensure that the core requirements for data management planning enumerated in the EHRI-RI DMP Template (see Appendix A) are fulfilled in a satisfactory manner.

6.1. Types of data in the EHRI-RI

The EHRI-RI will consist of a Central Hub which provides a variety of services to EHRI-RI National Nodes and other upstream and downstream data consumers. While the precise make-up of the Central Hub is yet to be determined, it will to a considerable extent be derived from services established in the course of EHRI project activities, including those active in the concurrent EHRI-3 project. This section enumerates the types of data that make up current services and those planned to be established in either EHRI-3 or the EHRI-RI implementation phase.

6.1.1. EHRI Portal - archival metadata

The EHRI portal contains metadata for several types of archival entities and their corresponding contexts:

Country reports: EHRI publishes a report for countries that had significant involvement in the Holocaust or where archives now reside that have significant Holocaust-related holdings.[9] These reports take the form of textual summaries of the relevant history of the country as it relates to the Holocaust, plus a description of the archival situation, such as how the state archives are organised and other notable institutions. In some cases this will incorporate longer-form research conducted by EHRI.

Archival institutions: EHRI publishes information about archival or other institutions that hold some form of Holocaust-related material. These descriptions correspond to the ISDIAH conceptual standard, incorporating names, street addresses and contact information, along with predominantly textual accounts of their historic context and holdings.[10]

Archival descriptions: EHRI publishes archival descriptions in a hierarchical form broadly corresponding to the ISAD(G) conceptual standard.[11] Top-level, or collection-level, descriptions are grouped by holding institution and can contain an arbitrary number of child items, which likewise may be nested to arbitrary depths. While these levels may conform to those outlined in ISAD(G), the degree of variance among institutions in how nesting is expressed organisationally means that they are not explicitly labelled in the EHRI portal in order to de-emphasise a frequent source of naming confusion.

Archival descriptions are created by both manual and automated processes. For the manual process, EHRI staff enter collection metadata via the portal's administration interface.

9 A general introduction to the country reports can be found at: https://www.ehri-project.eu/country-reports
10 International Standard for Describing Institutions with Archival Holdings, see: https://www.ica.org/en/isdiah-international-standard-describing-institutions-archival-holdings
11 General International Standard Archival Description, see: https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition

Automated and semi-automated processes employ a more complex workflow which varies depending on the type of data that requires an ingest process.

Authorities records: Authority records consist of historical information concerning corporate bodies, persons and families that have some function in the creation of or relate to archival material, described in accordance with the ISAAR standard.[12]

Controlled vocabularies: EHRI currently manages three public controlled vocabularies[13]:

- EHRI Terms: a hierarchically organised, multilingual set of subject headings for describing Holocaust-related material
- EHRI Camps: a list of camps based on the USHMM Encyclopedia of Camps and Ghettos
- EHRI Ghettos: a list of ghettos based on the Encyclopedia of Ghettos of Yad Vashem and the Encyclopedia of Camps and Ghettos of the USHMM

Controlled vocabularies are structured and described in a manner corresponding to the Simple Knowledge Organisation System (SKOS) specification.

6.1.2. User-contributed and user-generated data

The EHRI portal incorporates functionality only available to registered users. For EHRI staff and project subcontractors this functionality extends to the administration of the portal itself including the creation, modification and deletion of archival metadata; the underlying account system is the same.

6.1.2.1. The user registration process

Users can register on the EHRI portal either through an external authentication provider (such as Google, Facebook or Microsoft) or directly by providing a name, email and password.[14] When they opt to use an external authentication provider the portal asks for access to their profile information from which the following attributes are stored:

- name
- email address
- profile image URL

When creating an account on the EHRI portal, either directly or via an external provider, users are required to agree to abide by the terms and conditions and the EHRI data policy.

6.1.2.2. Email address validation

When registering directly with EHRI (not via an external authentication provider) users are asked to validate their email address by visiting a unique, one-time URL sent to them at registration time. Non-verified accounts which have not signed-in for over a 1-year period will be subject to automated removal.

6.1.2.3. User-contributed data

Registered users are able to contribute several types of data to the portal. This consists of:

- profile information, including some or all of the following fields that enable users who wish to do so to advertise their professional affiliation and/or research interests:

---

12 International Standard Archival Authority Record, see: https://www.ica.org/en/isaar-cpf-international-standard-archival-authority-record-corporate-bodies-persons-and-families-2nd
13 The controlled vocabularies are accessible via: https://portal.ehri-project.eu/vocabularies
14 For external authentication the OAuth2 protocol is used: https://oauth.net/2/

- ○ name
- ○ title
- ○ location
- ○ institution
- ○ role
- ○ personal website
- ○ work website
- ○ languages spoken
- ○ about
- ○ area(s) of interest
- ○ a profile image
- watched items: the set of metadata objects a user has expressed an interest in by clicking the "watch" button
- notes: textual annotations of metadata objects which may either be marked as public, in which case they are visible to all users of the portal after moderation by a member of EHRI staff, or private, which are visible only to the creator

Users are able to delete both their accounts and/or any information they have contributed to the project, such as annotations, from their personal account page. If users choose to delete their account any profile information will be automatically deleted; for the purposes of data integrity, however, annotations (which may be public) remain on the portal associated with a "stub" anonymous profile, as does information connected with the audit trail of archival objects.

Information contributed and later removed by registered users is still present in backups taken during the intervening period.

6.1.3. Tertiary archival metadata

In addition to the primary archival entities, EHRI's database incorporates tertiary entities that relate to the permission system and the audit trail. These consist of users of the archival system and groups into which they can be organised for the purposes of hierarchical or role-based access control. When archival entities are created, updated or deleted entries are added to the audit trail recording which user instigated the change. The audit trail therefore consists of an ordered list of administrative events within the database that can assist in understanding the provenance of archival descriptions.

6.1.4. EHRI Project Website

The project website provides an overview of and information about the project's mission and activities, and includes information about partners, staff and the advisory board. The project site also makes available public project deliverables.

6.1.5. Training Website

The training website hosts a number of self-directed learning resources covering a selection of Holocaust-related topics, along with related material focussed on the use of the EHRI portal and data analytics for the humanities.

6.1.6. EHRI Document Blog

The EHRI Document Blog[15] is a forum for the publication of exploratory research on topics related to the Holocaust and the research process, including digital research techniques. The site runs on the popular Wordpress platform, with custom plugins that facilitate integration

---

15 https://blog.ehri-project.eu

with the EHRI portal via its Search API (discussed below) and a visualisation tool based on the Omeka publishing platform that provides image annotation and mapping capabilities.[16]

6.1.7. Digital Editions

EHRI publishes a number of digital editions - standalone websites that explore Holocaust-related topics through the lens of a particular set of documents. For each edition a set of documents represented by scanned images are transcribed and annotated using the TEI XML format. Uploading the image and accompanying TEI to the EHRI Editions platform generates a web representation of the document and their metadata. Each edition typically also includes additional textual content to accompany and explain the content of the documents and their context.

6.1.8. Geospatial Repository

A component being tested in EHRI-3 and likely to become part of the EHRI-RI Central Hub is a repository of geospatial data such as historical borders and the locations of Holocaust-related places and events. The repository will be searchable and both datasets and their associated metadata will be accessible in structured form.

6.1.9. EHRI service registry

EHRI's service registry will provide users with an overview of EHRI-RI services and descriptions of their capabilities. It is currently planned that the services registry will be hosted by the EOSC Marketplace infrastructure.[17]

6.1.10. Marketing activities

Distinct from registered users of the portal are members of the public who choose to subscribe to EHRI publicity material via its mailing list. This list is managed by Mail Chimp.[18] Users can unsubscribe from the list at any time via a link in received communications.

6.1.11. Survey data

The EHRI-PP project has conducted surveys to gather information about the current and potential users of the EHRI-RI. The EHRI User Needs and Innovation Strategy Survey[19] was jointly developed by EHRI-PP WP5 and WP6 and was distributed to employees and those affiliated with EHRI-PP and EHRI-3 partner institutions. The User Access and Training Strategies Survey was aimed at a wider group of current and potential users of EHRI services and distributed via the EHRI website and social media platforms. Both surveys were conducted in a manner that ensured responses were collected and subsequently processed in an anonymous manner.

6.2. Implementation and compliance with FAIR data principles

This section will review how EHRI-RI data will implement the FAIR data principles with respect to its collection of data objects described in section 6.1.

---

16 https://omeka.org/classic/
17 https://marketplace.eosc-portal.eu
18 https://mailchimp.com
19 https://www.ehri-project.eu/sites/default/files/downloads/D5.1%20-%20User%20Needs%20Analysis.pdf

6.2.1. Findability

Findability refers to the ability of EHRI-RI users to easily locate and reference in a precise manner relevant units of (meta)data. Locating items is facilitated by effective search functionality and referencing by the use of consistent and clear identifiers.

6.2.1.1. Identifiers

EHRI-RI will manage material with varying characteristics as relates to provenance: born-digital material, such as grey literature and other publications, are directly managed by EHRI, whereas others are digital proxies, or descriptions of material managed by third-party individuals or institutions. As a result, identifiers with different characteristics are required.

Indigenous identifiers: Descriptions of archival material in the EHRI portal incorporates identifiers assigned by the original holding institution. Some institutions assign fully unique identifiers to archival units at all levels of description (such as fonds, subfonds, and item); others are only unique within their given fonds or parent scope. In order to facilitate cross-referencing with the source, the existence of identifiers at all levels of description is required as a necessary condition for descriptions of archival material to be incorporated into the EHRI portal. To achieve a fully sustainable connection where automatic updating is possible it is required that the holding institution assigns unique identifiers to archival units within the scope of their top level container unit.

Local identifiers: Local identifiers will be assigned for all entities described above. Where entities are accessible over the web, the local identifier will form a component of the items' URL. The composition of EHRI-RI local identifiers varies depending on the software platform used to generate them. In the EHRI portal, identifiers are unique but also designed to be informational and readable (incorporating in many cases a transliterated form of the indiginous identifier), and thus not fully opaque.

Archival Resource Keys (ARKs): Descriptions of archival material held by third-parties will be assigned Archival Resource Keys (ARKs).[20] ARKs provide a solution to the technical requirement for opaque identifiers which can be made to resolve to different locations (i.e. that are not tied to a given domain name), but provide persistence guarantees more aligned with the requirements for data in the custody of partner institutions and outside of EHRI's direct control.

Digital Object Identifiers (DOIs): Finally, material for which EHRI is the primary custodian will be assigned Digital Object Identifiers (DOIs). This will include textual content available through the EHRI portal, such as country reports, grey literature such as posts on the EHRI Document Blog, and EHRI Digital Editions. EHRI-RI will endeavour to ensure all DOI requirements are fulfilled and responsibility for metadata maintenance is clearly assigned.

Uniform Resource Identifiers (URIs): EHRI-RI will also make use of Uniform Resource *Identifiers* (URIs) for a subset of data types, principally those that maintain compatility with the Simple Knowledge Organisation System (SKOS) and are available in Resource Description Format (RDF).

As of the conclusion of EHRI-2, indigenous, local and URI identifier types are implemented for EHRI metadata and controlled vocabularies. Both ARKs and DOIs are planned for implementation in either EHRI-3 or the EHRI-RI implementation phase. Table 3 provides an overview of the identifier types that will be used to facilitate the "findability" of data types prominent in the EHRI-RI.

---

20 https://arks.org/

Data type	Identifier type
Archival descriptions	Local, ARK
Archival authority file	Local, ARK
Archival institution description	Local, ARK
Country report	Local, DOI
Document blog post	Local, DOI
Online editions document	Local, DOI
Controlled vocabularies and terms	Local, URI

Table 3*: Data types and the identifiers that will be used, reflecting different surety characteristics.*

All EHRI-RI datasets will be stored in searchable repositories and be retrievable by identifier or text-based descriptive metadata appropriate to specific data types. Findability of descriptions in the EHRI portal will also benefit from ongoing efforts to coreference indigenous index terms (assigned by individual holding institutions) to the EHRI-specific controlled vocabularies described above.

6.2.2. Accessibility

EHRI-RI will support the Open Science paradigm, following the principle "as open as possible, as closed as necessary". This means in practice that users will not require an institutional affiliation, or have to pay to access any EHRI-RI related data. A subset of user-centred information on the EHRI portal, such as that detailed in section 6.2.3 will require registration to access.

For access via programmatic means, the EHRI-RI Service Integration Framework (SIF) includes a number of structured data interfaces, or APIs, tailored to different purposes and use-cases. More information on the SIF will be available in EHRI-PP D7.4 Technical Report[21] but an overview is provided below. Table 4 provides an overview of the data access facilities part of the EHRI infrastructure.

Search API: The Search API allows basic search and retrieval tasks, similar to those facilitated by the EHRI portal's web interface, to be performed with structured data, specifically JSON over HTTP. The request and response formats conform to those of the JSON API[22] specification. Example use-cases include:

- searching for items containing a specific phrase, among all supported data types
- retrieving JSON data describing country reports, institutions, documentary units and authority file using their EHRI identifiers
- searching documentary unit items held by a specific institution

21 D7.4 will be available from month 30
22 https://jsonapi.org

● searching institutions by proximity to a given geographical point

The Search API facilitates integration with third-party federated search tools.

GraphQL API: The GraphQL API is intended primarily for more advanced and comprehensive retrieval tasks, compared to the Search API. Described as a "query language for your API" GraphQL provides a machine-readable schema and mechanism for exploring deeply-nested or hierarchical data, over HTTP. Use cases include:

● retrieving a wider range of data than supported than the Search API, included user-generated data and information relating to archival provenance
● streaming larger amounts of data, without pagination
● fetching very specific data attributes for, e.g., user-interface development

OAI-PMH 2.0: The Open Archives Initiative Protocol for Metadata Harvesting 2.0 is an XML-over-HTTP interface that allows supporting tools to harvest metadata about the collections in the EHRI portal in either Dublin Core or Encoded Archival Description (EAD) format. The portal's implementation is fully-conformant with the version 2.0 specification and supports all protocol features, including deletion markers, country- and institution-level sets, and granular time-based harvesting.

XML export: The EHRI Portal also provides the means to download XML representations of data types not supported by the OAI-PMH protocol, directly from item web pages. These include:

● Encoded Archival Guide (EAG) 2012 encodings of metadata about institutions
● Encoded Archival Context (EAC) encodings of metadata about corporate bodies, families and persons (CPF)
● RDF/XML representations of controlled vocabularies in the Simple Knowledge Organisation System (SKOS) format

RSS: Abridged entries for the EHRI Document Blog are accessible via the Really Simple Syndication (RSS) XML format. Full post text is secondarily available via the non-standardised Wordpress JSON REST API.

EHRI Portal - datasets	Search API	GraphQL API	OAI-PMH	Other
Country reports	✔	✔		
Institution metadata	✔	✔		
Archival descriptions	✔	✔	✔	
Controlled vocabularies		✔		TTL RDF
Authority files	✔	✔		

Annotations		✔		
Tertiary Metadata		✔		

Table 4*: data access facilities to EHRI-RI collections*

6.2.2.1. Sustainable access

The EHRI-RI will commit to sustainable access to data resources and the services that provide them where this is compatible with security requirements, the EHRI Privacy Policy, and relevant Service Level Agreements (SLAs). Backwards-incompatible changes to production services that alter or otherwise restrict access to classes of data will be signposted on a predictable timeline of no less than 6 months duration.

6.2.3. Interoperability

While genuine interoperability of metadata resources remains a considerable challenge within the archival space — due to the diversity of descriptive approaches and the complexity of hierarchical description — the EHRI-RI will endeavour to fulfil best-practice and to implement technical solutions that facilitate meaningful data exchange. Use of standardised metadata formats is one component of this overall effort.

Table 5 outlines the metadata standards used by EHRI-RI data collections.

Metadata standard	Applied at
EAD (2002/3) - Encoded Archival Description	EHRI Portal - Archival Descriptions EHRI publishes archival descriptions in a hierarchical form broadly corresponding to the ISAD(G) conceptual standard. Top-level, or collection-level, descriptions are grouped by holding institution and can contain an arbitrary number of child items, which likewise may be nested to arbitrary depths. EHRI-RI will support import and export of both EAD 2002 and the newer EAD-3 revision.
EAG (2012) - Encoded Archival Guide	EHRI Portal - Archival Institutions EHRI publishes information about archival or other institutions that hold some form of Holocaust-related material. These descriptions correspond to the ISDIAH (International Standard for Describing Institutions with Archival Holdings) conceptual standard, incorporating names, street addresses and contact information, along with predominantly textual accounts of their historic context and holdings.
EAC (2010) - Encoded Archival Context of metadata about corporate bodies, families and persons (CPF)	EHRI Portal - Authority Files Authority records consist of historical information concerning corporate bodies, persons and families that have some function in the creation of or relate to archival material, described in accordance with the International Standard Archival Authority Record (ISAAR) and downloadable as Encoded Archival Context

	(EAC) XML.
RDF/XML representations of controlled vocabularies in the Simple Knowledge Organisation System (SKOS) format.	**EHRI Portal - Controlled Vocabularies** EHRI currently manages three public controlled vocabularies: *EHRI Terms*: a hierarchically organised, multilingual set of subject headings for describing Holocaust-related material *EHRI Camps*: a list of camps based on the USHMM Encyclopedia of Camps and Ghettos *EHRI Ghettos*: a list of ghettos based on the Encyclopedia of Ghettos of Yad Vashem and the Encyclopedia of Camps and Ghettos of the USHMM
TEI - Text Encoding Initiative (encoded as XML)	**EHRI Online Editions** Three online editions are published at the time of writing with a further two in development. Each edition is based around a specific topic explored via the use of primary sources for which scanned images have been encoded as TEI XML. The TEI documents typically embed references to corporate bodies, persons or places for which data can be retrieved for additional context via the EHRI APIs or Geonames.
BibTex / RFT / Tagged / XML format for bibliographic records	**EHRI Bibliography** A selection of publications relating to the EHRI project and Holocaust research.
IIIF Image API v2.0	**EHRI Visualisations platform & Online Editions** Image scans that are available on EHRI's online editions will be accessible via a IIIF Image API server, allowing them to be freely scaled, panned and zoomed by compatible viewers. A IIIF implementation is currently being evaluated in EHRI-3.
IIIF Presentation API	**EHRI Visualisations platform & Online Editions** Image scans and sets of images (corresponding to e.g. pages of a document) will be accessible using the IIIF Presentation API, allowing them to be navigated by compatible viewers. A IIIF implementation is currently being evaluated in EHRI-3.
WMS (Web Map Services)	**EHRI Geospatial repository** The geospatial repository will facilitate HTTP-based access to tile-based map base layers that can be incorporated into distributed map visualisation systems. A WMS server implementation is currently being evaluated in EHRI-3.
WFS (Web Feature	**EHRI Geospatial repository**

Services)	Datasets stored in the EHRI geospatial repository will be accessible via a WFS API, providing interoperability with common geospatial tools such as QGIS and ArcGIS. A WFS server implementation is currently being evaluated in EHRI-3.
RSS (Really Simple Syndication XML)	EHRI Document blog Headers and content summaries for articles on the EHRI Document blog are available via an RSS feed, providing interoperability with content syndication systems.

Table 5. *Metadata standards used in EHRI-RI data collections*

6.2.4. Reusability

Portal metadata resources are intended for dissemination and are thus open with no restrictions. The EHRI portal explicitly facilitates re-use of metadata by providing an OAI-PMH harvesting API, in addition to other APIs described above.

Non-metadata resources such as the EHRI country reports, blog posts, commentary accompanying EHRI Digital Editions and user-generated content may be republished providing that acknowledgment of the original author is given, as specified in the EHRI portal terms and conditions.[23] In the context of the future EHRI-RI organisation post-implementation, clarifying licenses such as Creative Commons BY 4.0[24] will be added to ensure visibility for reuse and attribution requirements.

When operational, EHRI-RI repositories that retain first-party or primary data objects — such as the geospatial repository — will aim to fulfill CoreTrustSeal requirements and eventually seek CoreTrustSeal certification.[25]

7. ALLOCATION OF RESOURCES

Responsibility for ensuring that data managed by the EHRI-RI is FAIR compliant will reside primarily with the technical leadership of the Central Hub. An EHRI-RI data management committee will be formed in 2022 and meet periodically to review RDM practices and assess compliance with FAIR and other relevant guidelines. EHRI-RI National Nodes will be responsible for stewardship and curation of data hosted by participating institutions but will abide by the overall EHRI access policy and FAIR data guidelines.

While the financial model for the EHRI-RI is still under development we anticipate that funds for making data FAIR compliant will come from the budget allocated for general technical development and maintenance of the RI.

8. ETHICAL ASPECTS

EHRI-PP confidential deliverable D10.1 POPD - Requirement no. 1 [3] specifies procedures in regard to ethical data collection in project activities.[26] Work packages that gather survey data will abide by the personal data restrictions of the GDPR (and other relevant national statutes) and where possible anonymise individual responses, though due to the nature of the project it is unlikely that significant quantities of privacy-relevant data would be involved.

---

23 https://portal.ehri-project.eu/terms
24 https://creativecommons.org/licenses/by/4.0/
25 https://www.coretrustseal.org/
26 D10.1 POPD – Requirement No. 1 [3] (Confidential)

Raw survey data will not be disseminated beyond those conducting the research and, on conclusion of the research, will be either deleted or archived in secure storage for a period of no more than five years.

9. SECURITY

Management of Holocaust-related archival objects described by metadata available via EHRI-RI services will remain the responsibility of participating content providers. In cases where EHRI-RI will have primary custody of digital proxies of archival objects (such as document scans) it will follow best-practice to ensure the integrity of such objects are secure from accidental misuse or malicious tampering.

All active databases are hosted in data centers located within the EU. Since EHRI-PP project partners reside in the United States, United Kingdom and Israel, archival metadata including user-generated material may be transferred outside the EU for research and development purposes.

Personal data is protected according to EU regulations including the General Data Protection Regulation (GDPR).

EHRI-RI will endeavour to follow reasonable security best-practices in the maintenance of information technology infrastructure and data management. While this is always a moving target that requires constant vigilance we undertake to:

- Enforce 2-factor authentication for infrastructure administration tasks;
- Reduce vulnerability to brute-force attacks by enforcing the use of public/private key authentication — as opposed to password authentication — where possible;
- Ensure any stored passwords are hashed with modern cryptographic algorithms such as BCrypt;
- Frequently update IT resources to ensure that they run the latest security patches.

EHRI-RI will make use of IT automation, deployment and provisioning tools that will allow components to be rapidly and reliably redeployed in the event of a security breach or other compromise situation.

Appendix A: DMP Template for EHRI-RI generated data

Researchers use data provided by the EHRI-RI (and its related organisations). They also create data, such as publications or data collections. The EHRI-RI can ask researchers to comply with the EHRI-RI data management policy, aimed at this specific target group. For this a DMP has to be created. The EHRI-RI will provide a DMP template that researchers can use. The table below contains all core requirements.

EHRI-RI Data Management Template for Researchers. Version 1.

DMP Topic	Question
1. Data description and collection or re-use of existing data	
	How will new data be collected or produced and/or how will existing data be reused?
	What data (for example the kinds, formats, and volumes) will be collected or produced?
2. Documentation and data quality	
	What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany data?
	What data quality control measures will be used?
3. Storage and backup during the research process	
	How will data and metadata be stored and backed up during the research process?
	How will data security and protection of sensitive data be taken care of during the research?
4. Legal and ethical requirements, codes of conduct	
	If personal data are processed, how will compliance with legislation on personal data and on data security be ensured?
	How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?
	How will possible ethical issues be taken into account, and codes of conduct

followed?	

5. Data sharing and long-term preservation	
	How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?
	How will data for preservation be selected, and where will data be preserved long-term (for example a data repository or archive)?
	What methods or software tools will be needed to access and use the data?
	How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?
6. Data management responsibilities and resources	
	Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?
	What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

Below the EHRI-RI DMP Template is populated with content, based on an imaginary use-case in which researchers digitize a collection of historical documents. They create a digital image collection for which a DMP has to be created.

EHRI-RI DMP example

Project summary: The aim of the project is to digitize and analyse a collection consisting of photographs, letters and other vulnerable historical documents. The collection consists of about 1.000 objects. Digitisation is part of the conservation strategy of the collection.

DMP Topic	Question
1. Data description and collection or re-use of existing data	
	How will new data be collected or produced and/or how will existing data be reused?
	- The project will produce about 1.000 digital files, each representing a historical photograph, letter, print, etc.

	What data (for example the kinds, formats, and volumes) will be collected or produced?
	- The objects will be digitised according to "preservation imaging" guidelines (high resolution, calibrated scanners, uncompressed, etc.) - The "digital masters" will be stored in TIFF uncompressed format. - For each digital image master file derivatives will be created that support the IIIF standard.

2. Documentation and data quality	

	What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany data?
	- Documentation on image characteristics will be stored in the "tags" of the image files (TIFF-tags) (e.g. number of pixels, resolution, date of scanning, equipment used, etc.) - Documentation on the image (content, dating, location) based on Dublin Core standard,
	What data quality control measures will be used?
	- Quality of the digitisation process is managed by using calibrated (colour) targets and software to check conversion quality (preservation imaging guidelines) - Quality of metadata checked by managed editorial process (e.g. two persons review entered documentation)

3. Storage and backup during the research process	

	How will data and metadata be stored and backed up during the research process?
	- During the digitisation process the digital images will be stored on hard disks for which back-ups are created on a daily basis.
	How will data security and protection of sensitive data be taken care of during the research?
	- The material does not contain sensitive data.

4. Legal and ethical requirements, codes of conduct	

	If personal data are processed, how will compliance with legislation on personal data and on data security be ensured?
	- The project will follow the EHRI-RI regulations on data processing.
	How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

	- The digital images will be made available according to the CC-BY-NC-ND license (allows reusers to copy and distribute the material in any medium or format in unadapted form only, for noncommercial purposes only, and only so long as attribution is given to the creator).
	How will possible ethical issues be taken into account, and codes of conduct followed? - N/A

5. Data sharing and long-term preservation

	How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons? - The data collections will only be shared when the complete collection is processed (scanned, documented) - End users will have access to derivatives (in an IIIF environment) of the digital masters.
	How will data for preservation be selected, and where will data be preserved long-term (for example a data repository or archive)? - The digital masters will be archived in a certified digital repository.
	What methods or software tools will be needed to access and use the data? - Standard web browsers will be needed to access the data (using the IIIF APIs)
	How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured? - Identifiers will be used as part of the documentation of the digital images.

6. Data management responsibilities and resources

	Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)? - The project team will be responsible for data management.
	What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)? - The project budget will contain resources for data management.