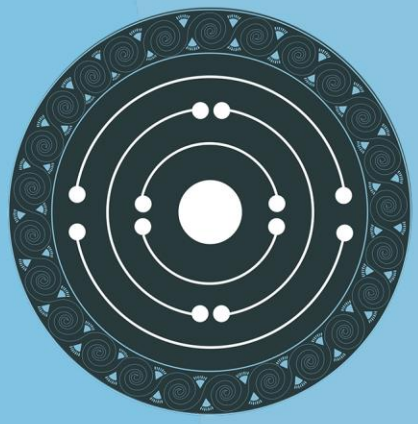


PARTHENOS

Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies

Report on Guidelines for Common Policies Implementation (1)

KNAW-DANS (overall coordination)
CLARIN
MIBACT-ICCU
KCL
PIN

PARTHENOS is a Horizon 2020 project funded by the European Commission. The views and opinions expressed in this publication are the sole responsibility of the author and do not necessarily reflect the views of the European Commission.

HORIZON 2020 - INFRADEV-4-2014/2015:

Grant Agreement No. 654119

PARTHENOS

Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies

Guidelines for Common Policies Implementation (1)

Deliverable Number D3.1

Dissemination Level Public

Delivery date 25 April 2017

Status Final

Author(s)

Hella Hollander (KNAW-DANS)
Francesca Morselli (KNAW-DANS)
Femmy Admiraal (KNAW-DANS)
Anders Conrad (CLARIN: RDL)
Thorsten Trippel (CLARIN: Univ. Tübingen)
Douwe Zeldenrust (CLARIN: KNAW-Meertens)
Paola Ronzino (PIN)
Sara Di Giorgio (MIBACT-ICCU)
Antonio Davide Madonna (MIBACT-ICCU)
Mark Hedges (KCL)

Klaus Illmayer (OEAW)
Vanessa Hannesschläger (OEAW)
Roberta Giacomi (SISMEL)
Maurizio Sanesi (SISMEL)
Emiliano Degl'Innocenti (CNR-OVI)
Lene Offersgaard (CLARIN: UCPH)
With contributions from all partners,
Eld Zierau (CLARIN: RDL)
Michael Svendsen (CLARIN: RDL)
especially:
Heiko Tjalsma (KNAW-DANS)
Valentijn Gilissen (KNAW-DANS)
Emlie Kraaikamp (KNAW-DANS)
Adeline Joffres (CNRS/Huma-Num)
Marie Puren (INRIA)
Claus Spiecker (FH Potsdam)

Project Acronym	PARTHENOS
Project Full title	Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies
Grant Agreement nr.	654119

Deliverable/Document Information

Deliverable nr./title	D3.1 Guidelines for Common Policies Implementation (draft)
Document title	PARTHENOS_D3.1_Guidelines for Common Policies Implementation (1)
Author(s)	Hella Hollander, Francesca Morselli, Femmy Admiraal, Anders Conrad, Thorsten Trippel, Douwe Zeldenrust, Paola Ronzino, Sara Di Giorgio, Antonio Madonna Davide, Mark Hedges
Dissemination level/distribution	Public

Document History

Version/date	Changes/approval	Author/Approved by
V 0.1 21.03.2017	Separate chapters from four Task leaders merged into one document	Femmy Admiraal Francesca Morselli
V 1.a 31.03.2017	Reviewed document by all tasks: content, bibliography, format, style and language	Hella Hollander Femmy Admiraal Francesca Morselli Sara Di Giorgio Anders Conrad Lene Offersgaard Mark Hedges
V 1.b 07.04.2017	Reviewed final first draft of the document by PIN: language, consistency	Sheena Bassett
V 2.0 12.04.2017	Document ready for internal submission to PIN	Hella Hollander
Final 25.04.2017	Final Version	Franco Niccolucci

Table of Contents

1. Introduction to the WP3 approach

1.1. Composition of WP3 and structure of this deliverable

The first goal of WP3 is to agree on and define the concepts of Policy, Guidelines, Best practice, their objectives and the target audience.

WP3 started at the beginning of the project, in February 2016, immediately following up the deliverable of WP2 on user Requirements (D2.1). The partners involved are KNAW-DANS (WP3 Leader and Task leader T3.2), CLARIN (Task leader T3.1), MIBACT-ICCU (Task leader T3.3), KCL (Task leader T3.4). All PARTHENOS partners (fifteen organisations, sometimes consisting of multiple institutes) contributed and were involved.

Their combined effort is visible in this deliverable, which gives an overview of existing policies concerning data management as well as policies concerning quality of data, metadata and repositories and IPR, open data and open access. The use and added value of common policies is presented by use cases throughout the deliverable, showing how different partners implemented these policies within their organisation.

There was intensive collaboration between WP2 and WP3: WP2 defines the requirements for shared policies as they represent the needs of the user community concerning data life cycle policies. The definition of common guidelines and best practices enabling cross-discipline data use and reuse, data policies to improve the data quality and long-term preservation, policies addressing sensitive data and privacy issues are expressed in comprehensible use cases in D2.1. WP2, together with WP3, will provide an inventory of existing policies from the different Humanities infrastructures. WP3 paid special attention to the request to set up a PARTHENOS Data Management Plan. This has resulted in a template (draft) which gives an overview of questions and answers addressing standards and guidelines in data management within the Humanities as a whole, keeping in mind domain-specific procedures and practices.

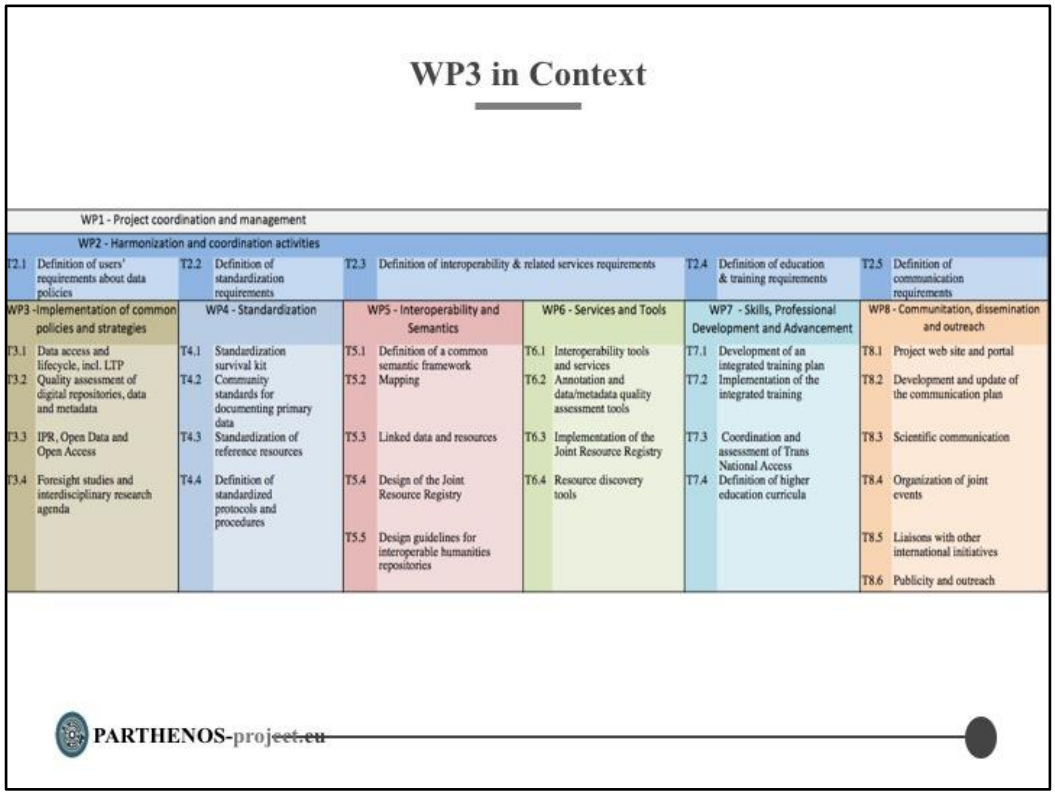


Figure1.1: WP3 in context with other WPs.

Each task in WP3 produced several recommendations which are integrated according to the FAIR principles for data quality, and presented in the conclusion as a set of high-level PARTHENOS recommendations. These guidelines for the user community will be tested by WP2, addressing questions such as: are the guidelines covering the needs of the community or are there gaps to be filled by the final deliverable D3.2? Which amendments are needed? WP2 will set up an expert panel which will evaluate the outcomes of this deliverable.

WP3 also deals with foresight studies, as it is anticipated that future virtual research environments will rely increasingly on data and hence will be heavily influenced by the availability, quality and access characteristics of this data. The result of this task will be a separate deliverable; however, to introduce this work an overview of the approach and methodology is presented in this deliverable.

The work of WP3 will not only result in a set of theoretical documents, as their content will also be disseminated by an easy-to-use tool called the PARTHENOS wizard. This tool is created to guide the user, who might be a researcher or a Research Infrastructure end user, or a policy maker, through the jungle of existing and relevant policies, guidelines and best practices in their

community or research discipline. The wizard will be connected to the data model and infrastructure of PARTHENOS.

The PARTHENOS high-level principles are offered as common guidelines to build bridges between different, although tightly interrelated fields and stakeholders within the Humanities by the harmonization of policy definition and their implementation. A coherent, well accepted set of policies, guidelines and tools will be presented to the user community.

1.2. WP3 Methodology

The aim of this document is to present to its stakeholders (see Section 1.4) a series of recommendations and guidelines about which policies to apply during and after their research or infrastructure work. “During their research work”, because policies on data and repositories guide the data creator to produce high quality data; “after their research work”, because policies on access and reuse help make the data more accessible and reusable.

WP3 analysed the requirements gathered by WP2 in the Deliverable D2.1 Report on User Requirements[1] shown according to a simplified Cockburn schema, which were gathered from the different research communities identified within the project.

In particular, Chapter 1 was used as a roadmap for the guidelines because it describes user requirements as regards data production, storage, management, curation and long-term preservation (Sub Task 2.1.1), as well as requirements concerning the quality assessment of digital repositories, individual data items and individual metadata items, as expressed by the research communities involved in the project (Sub Task 2.1.2) and requirements about IPR, Open Data and Open Access, both those expressed by the research communities involved in the project and others emerging from related national and European regulations (Sub-Task 2.1.3).

As mentioned in the Introduction, WP3 comprises four tasks. However, Task 3.4 “Foresight Studies” is not an integral part of the present document, as it focuses on

[1] PARTHENOS: Report on User Requirements (D2.1). 20 October 2016 (final version).

the future developments in terms of policies on data quality and data management. The outcomes of Task 3.4 will be fully elaborated in a separate deliverable (D.3.3).

A three-step methodology

When the work of WP3 and its tasks started in February 2016, the first decisions to be made were:

1) How to conduct the “fieldwork” research?
2) How to organize the information we would have gathered?
3) How to propose the outcomes and guidelines derived from the investigations?

The research area of WP3 covers a broad field of policies related to data quality and their management, as well as to the repositories in which they are preserved and accessed. This makes the identification of a shared and coherent methodology even more complex.

1) After a period of initial investigation, the first three tasks of WP3 agreed on the necessity to explore the current status of each policy field they are involved in. This translated to the creation of surveys among stakeholders, desk and literature research and interviews - when it was necessary to have direct contact with specific stakeholders. Through these different methodologies and tools, the aim was to identify the current policies in use by each field addressed by PARTHENOS WP3 ((meta)data quality; data management plans and IPR issues).
2) The three tasks collected and organized the information gathered in similar ways, mainly in Google spreadsheets. The complexity and importance of these tables were not considered to be simple ways of organizing data, but real heuristic instruments for the WP members to think and answer complex questions about data quality, data management and data accessibility.
3) As indicated in the PARTHENOS Description of Work document, this WP proposes concrete guidelines and best practices to its different stakeholders, in relation to data and repository quality (Chapter 2), data management plans (Chapter 3), as well as IPR and Open Access (Chapter 4). In order to make

the guidelines developed by each task more easily accessible and easily applicable, they are structured according to the FAIR principles.

The final chapter of the present deliverable merges the recommendations and guidelines from each task into a set of PARTHENOS high-level recommendations, in order to provide a useful and compact instrument for all the stakeholders involved in PARTHENOS addressing topics such as data and repository quality, data management plans, IPR and open access.

Case studies

In this deliverable, both established and recent theories concerning (meta)data and repository quality, accessibility of research data and research data management have been taken into consideration. However, it was agreed that it was also relevant to show how these challenges, as well as the proposed guidelines, are being addressed by the institutions involved in WP3. For this reason, each chapter includes relevant experiences of those institutions involved in this WP, represented as case studies. For example, Case study 2.2.1.4 describes how the University of Copenhagen received the DSA certification, and how this contributed to increasing the quality of their repository and data. Case study 4.3.4.2 about DANS shows how the implementation of a CC0 access policy on research data has enabled the researchers to access and reuse archaeological data more easily.

1.3. Overall framework of the deliverable: the FAIR principles

1.3.1. Introduction to the FAIR principles

In 2014, the FAIR guiding principles for individual datasets were formulated: Findable, Accessible, Interoperable and Reusable. These are data principles which were first published in March 2016[2] and quickly have become very popular.

The intent was, according to the creators, that these principles may act as a guideline for those wishing to enhance the reusability of their data holdings, rather than being a standard or specification. In other words, the FAIR principles provide a

[2] Wilkinson et al. The FAIR Guiding Principles for scientific data management and stewardship : *Scientific Data* **3**, Article number: 160018 (2016), DOI:10.1038/sdata.2016.18.

set of mileposts for data producers and publishers to help ensure that all data will be Findable, Accessible, Interoperable, and Reusable. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals.

In the FAIR Data approach, data should be:

Findable – Easy to find by both humans and computer systems and based on mandatory description of the metadata that allow the discovery of interesting datasets;
Accessible – Stored for long-term[3] such that they can be easily accessed and/or downloaded with well-defined license and access conditions (Open Access *when possible*), whether at the level of metadata, or at the level of the actual data content;
Interoperable – Ready to be combined with other datasets by humans as well as computer systems;
Reusable – Ready to be used for future research and to be processed further using computational methods.

The principles were designed to serve the community as a minimal scope approach, which focuses on the specification of minimally required standard protocols, lightweight interfaces and formats. To make them more concretely applicable, in the original proposal (also known as the FORCE11 approach[4]), the four principles were further segmented as follows:

To be Findable:
F1. (meta)data are assigned a globally unique and eternally persistent identifier.
F2. data are described with rich metadata.
F3. (meta)data are registered or indexed in a searchable resource.
F4. metadata specify the data identifier.

[3] The exact definition of *long-term* may vary. While the DSA considers >5 years as long-term, other institutions may interpret *long-term* as covering longer periods of up to 50 or 100 years.
[4] https://www.force11.org/group/fairgroup/fairprinciples.

To be Accessible:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.

A1.1 the protocol is open, free, and universally implementable.

A1.2 the protocol allows for an authentication and authorization procedure, where necessary.

A2 metadata are accessible, even when the data are no longer available.

To be Interoperable:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (meta)data use vocabularies that follow FAIR principles.

I3. (meta)data include qualified references to other (meta)data.

To be Re-usable:

R1. meta(data) have a plurality of accurate and relevant attributes.

R1.1. (meta)data are released with a clear and accessible data usage licence.

R1.2. (meta)data are associated with their provenance.

R1.3. (meta)data meet domain-relevant community standards

The FAIR principles now are widely used by many stakeholders in research data management. However, this does not mean that the framework has reached a fully crystallized final state. In fact, the principles are not intended to be static and the rationale behind them is constantly under reconsideration.[5] The FAIR principles are thus constantly revisited, updated and refined.[6]

When comparing the FAIR approach to other models of digital data curation and archiving, one major observation stands out: FAIR targets depositors (of whatever stakeholder category), not technical infrastructures. The principles deliberately do not specify technical requirements, but are a set of guiding principles that provide for a continuum of increasing reusability, via many different implementations.[7] This means that the model speaks to individual researchers

[5] http://datafairport.org/fair-principles-living-document-menu.
[6] Mons et.al http://content.iospress.com/articles/information-services-and-use/isu824#x1-50011.
[7] Mons et.al http://content.iospress.com/articles/information-services-and-use/isu824#x1-50011.

without a technical background or experience in digital data preservation, as well as to experienced and trained depositors, such as people working in data archives or Research Infrastructures. Since PARTHENOS targets different types of stakeholders (see Section 1.4), we feel that the FAIR principles are an excellent match the PARTHENOS approach.

A number of alternative models and standards for digital data curation and archiving are described briefly below.

Research Data Lifecycle, e.g. UKDA

A Data Life Cycle focuses on the processes which data might go through, from the data creation onto the final phases of accessing and re-using the data. In between stages are processing, analysing and preserving the data. As such, it is typically originated in and focused on the life cycle of data created and processed in research. For data in Cultural Heritage Institutions, it is possibly a less suitable model.

SCAPE Policy Framework

This Framework consists of three *preservation policy levels* supporting an organisation in creating their preservation policies. These levels are Guidance policies, Preservation Procedure policies and Control policies. The first level describes the general long-term preservation goals of the organisation for its digital collection(s). The second one describes the approach the organisation should take in order to achieve the goals as stated on the higher level. The third level describes the general long-term preservation goals of the organisation for its digital collection(s). From the general intention of this framework, to make the creation of a preservation policy for organisations more straightforward and better prepared for machine readable policies, it clearly follows that this framework is very much an organisational policies approach.

Reference Model: OAIS Reference Model

OAIS is the most widely used Reference Model. OAIS is a very elaborate model describing all the functions of the data management needed to ingest, describe, store and make available the data by a data repository, from the moment of the intake (ingest) of data on to the dissemination of data to users. OAIS is not a

blueprint, but a *conceptual framework*. With OAIS, a repository can describe its core archival functions and processes in standard terms for reference purposes. OAIS is destined for a clear defined Designated Community (or more Designated Communities). It has a rather strong IT architecture background.

Data Seal of Approval

The DSA, the Data Seal of Approval, contains criteria for the quality of trustworthy digital repositories and is a lightweight form of certification of Trustworthy Digital Repositories. It is a basic certification standard, based on a self-assessment of the requirements by the repository itself that is peer-reviewed by external reviewers. Its level of focus is a repository, not in the first place the data themselves.

Capability Maturity Model

A Capability Maturity Model can be set up for heuristic reasons. It can be used as a mean of determining and comparing how far existing or emerging data repositories are in meeting the requirements for being considered as a mature, fully-developed, repository. It is not a data model, but like a certification model, more on the level of a repository. The model also has a rather strong IT architecture background.

1.3.2. The FAIR principles in this deliverable

As stated above, the formulation and structuring of the FAIR principles should not be regarded as definitive, but rather as an ongoing collaboration aiming at high-level, yet accessible, recommendations. In this sense, the framework also allows for flexibility in highlighting one aspect or the other. Throughout this deliverable, referencing to the individual principles (F1, F2, etcetera) follows the numbering as proposed originally by FORCE11.[8]

Despite the fact that the FAIR principles are taken as the overall guiding framework, each task applied them slightly differently, according to the focus of that particular task. In Chapter 2, which reports on Task 3.2, the policies for the quality of data, metadata, and repositories were mapped onto the FAIR principles, in order to

[8] https://www.force11.org/group/fairgroup/fairprinciples.

formulate the high-level recommendations for stakeholders. The methodology for how the matching was applied is explained in detail in Section 2.3.2.

In Chapter 3, focusing on Task 3.1, good practices in data management are highlighted and organized according to the FAIR principles. Detailed recommendations are given on how to support each of the principles on a practical level. In addition, it is shown how the FAIR principles are reflected in the answers to the questionnaire that was carried out in this task.

Finally, in Chapter 4, which shows the results of the work carried out by Task 3.3, the guidelines under the F of FAIR are considered to be of limited importance. Therefore, this chapter only focuses on the principles subsumed under A(ccesible), I(nteroperable), and R(eusable).

1.4. Who are the stakeholders of WP3?

When attempting to define stakeholders, there are two possible perspectives to consider. The first one focuses on "user communities", based on research disciplines. Stakeholder groups have already been defined by PARTHENOS D2.1[9] from this perspective. Therefore, this is the project's standard schema. There are four main research areas of interest:

1) **History** (in a broad sense: including Medieval Studies, Recent History, Art History, Epigraphy, etcetera).
2) **Language-related Studies** (including Literature, Linguistics, Philology, Language Technology, etcetera).
3) **Archaeology, Heritage & Applied Disciplines** (including Cultural Heritage, Archives, Libraries, Museums, Preservation / Conservation experts, Digital curation / edition / publishing, etcetera).
4) **Social Sciences** (in a broad sense: Sociology, Political Science, Geography, Anthropology, Cultural Studies etcetera).[10]

[9] PARTHENOS: Report on User Requirements (D2.1). 20 October 2016 (final version), pp. 11f.
[10] PARTHENOS: Report on User Requirements (D2.1). 20 October 2016 (final version), pp. 12.

From this list, different perspectives can be derived, focusing on the requirements of certain research areas or, to go into even more detail, of single disciplines. This has been discussed in several follow-up chapters in PARTHENOS D2.1.

However, for our discussion of data and repository quality, we prefer to assume the second possible perspective on stakeholders. In this alternative mode of classification, “research communities” are merely one of several stakeholder types. We made this decision due to the fact that, when it comes to repository quality, it is not the research discipline that is the main factor shaping individual needs and requirements, but rather the *type* of stakeholder that is crucial.

The guiding principles of FAIR state that there are "multiple stakeholders" involved in the process of "enabling optimal use of research data and methods"[11]. The guiding principles document identifies the following stakeholder groups: researchers, professional data publishers, funding agencies, and the data science community[12]. Each of these stakeholders has an individual perspective on the question of data quality and FAIRness. For the purpose of this deliverable, it is necessary to refine this list and adapt it according to the most important stakeholders of the PARTHENOS project. This distinction is necessary because requirements for quality assessment differ greatly depending on who raises the question of quality.

A good example to illustrate these differences is the question of what the difference between data and metadata is. A researcher would regard general data on the project (such as project name, duration) they are working in as “metadata”, whereas a Research Infrastructure (e.g. CLARIN in the Virtual Language Observatory VLO) would consider them “data” (see also Section 2.1). In this context, WP3 experienced similar problems as WP2 (as outlined in D2.1) - “user communities” is often too broad and at the same time too narrow a term to describe the challenges that individual researchers or research projects face, and often, the identified “communities” overlap in some places, whilst leaving gaps elsewhere.

Several stakeholder types particularly relevant to the PARTHENOS project are described in the following section. In addition, the specific challenges that the stakeholder types face in the area of data storage are identified.

[11] FORCE11: Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing version b1.0 https://www.force11.org/fairprinciples.
[12] Ibid.

To be clear: one classification does not compete with the other. However, for the specific purpose of this deliverable, definition by stakeholder type rather than by research discipline is more suitable as it allows better insight and analysis of quality aspects as we expect to find a broader and more holistic perspective.

1.4.1. Our stakeholders: research communities, data archives, Research Infrastructures, Cultural Heritage Institutions

1.4.1.1. Research communities / researchers

"Research communities" are viewed in terms of broader disciplines (e.g. historians, linguists, social scientists, etcetera). However, these "communities" cannot always be regarded as a single "stakeholder" or "actor", since the communities rarely act as one entity or group; more often than not, individual researchers or research projects will be the actors whose needs and questions in the field of repository and (meta)data quality will have to be addressed. This conclusion is supported by the use cases developed in T3.2 (see Chapter 2). Still, communities as a whole can sometimes become powerful entities which function in a way similar to Research Infrastructures (see below) and develop binding standards. One example of this is the Text Encoding Initiative (TEI).

"Repository quality" can mean quite different things to different research communities or individual researchers and its interpretation strongly depends on the conventions within a certain community, the nature of the data or the goals of the project.

There are also different perspectives between disciplines and even within disciplines. They are, on the one hand, caused by different approaches and traditions as well as different levels of familiarity with digital methods, and on the other by the different types of data produced and used. An example would be the audio / video data produced by social scientists and linguists versus the data on artefacts.

It also makes a difference whether researchers work in smaller projects applying digital methods or whether their work is done in the context of (larger) Research Infrastructures. For smaller projects, the quality of the immediate outcome

at the end of the project may be more important than a long-term perspective for the data produced. This, in turn, might lead to larger amounts of lower quality data rather than smaller amounts of detailed data with proper metadata, and complex visualizations rather than proper archiving. For data archives and RIs, this can make the data difficult to store and to preserve; for the other members of the research community, it can discourage the reuse of existing data.

1.4.1.2. Data archives

"Data archives" can be located at / hosted by research institutions or Cultural Heritage Institutions and can be part of a Research Infrastructure. They are (or host) digital repositories, where research and/or cultural heritage data can be stored in a sustainable manner.

When it comes to repository quality, data archives are often simultaneously policy developers/providers as well as policy implementers. Concerning (meta)data quality, they are also often policy developers and have to deal with (meta)data which might not meet the criteria developed in the respective policies.

1.4.1.3. Research Infrastructures (RIs)

"Research Infrastructures", are officially established European RIs/ERICs[13] like CLARIN and DARIAH (and not informal infrastructures or even established communities such as e.g. the Text Encoding Initiative. In our understanding, these would be described by the term "research communities"; see above). Their core task is to supply researchers from a certain domain with infrastructure (e.g. repositories, tools, standards) and to take a mediating role.

Research Infrastructures can (and, in the best case, should) overlap with the stakeholder groups "data archives" and "research communities", as they ideally represent and support a certain research community and provide this community with the means to archive their data.

[13] European Research Infrastructure Consortium. https://ec.europa.eu/research/infrastructures/index_en.cfm?pg=eric.

1.4.1.4. Cultural Heritage Institutions (CHIs)

"Cultural Heritage Institutions" are defined as galleries, libraries, archives and museums (also known as GLAM institutions). They often face the problem that policies for the handling of the physical objects they host are in place, but these may be incompatible with the requirements of the digital space. In addition, the needs and requirements that Cultural Heritage Institutions have can differ from the needs of the researchers who want to work with their data. Concerning repository quality, Cultural Heritage Institutions (at least large ones) face a different situation than the other stakeholder groups, as they usually have their own repositories which are often built in cooperation with private firms.

1.4.2. Roles that stakeholders can assume

As the following graph and case study will illustrate, it is often difficult to draw a clear line between stakeholders, as one entity can be in the position of more than one stakeholder type at once: a CHI can simultaneously be a data archive, a researcher can be part of a RI, and so on.

Therefore, the *roles* stakeholders can assume add a third dimension to the discipline- / stakeholder type-focused typology developed above: a researcher (or an entire research community) will, at some point in their work, be a data end user, but at another time might become a data provider, or a data processor. The same is true for all other stakeholder groups. For this reason, the different chapters of this deliverable will not always refer to the stakeholders identified here in a uniform way, as the chapters' perspectives on the stakeholders may vary slightly according to the topics addressed in them.

1.4.3. The PARTHENOS WP3 stakeholders: a graph

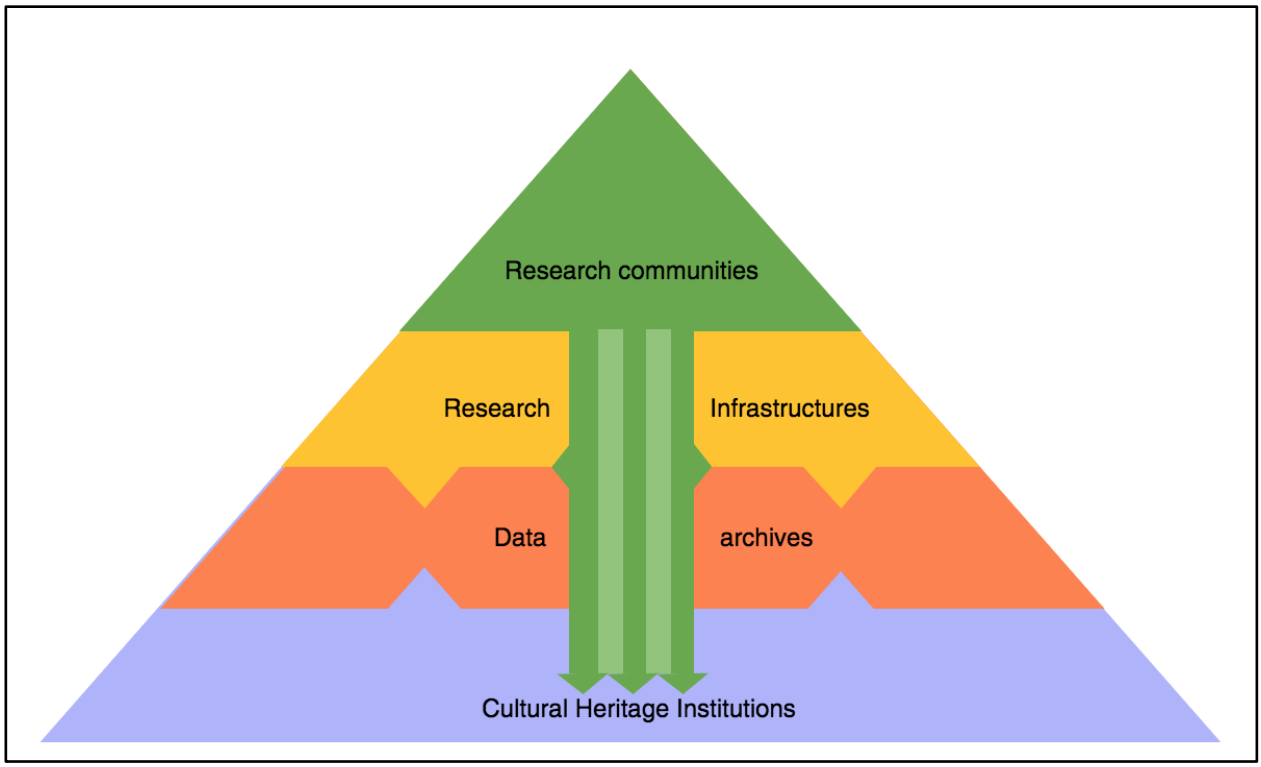


Figure 1.2: Stakeholders identified in PARTHENOS WP3.

This graph seeks to illustrate what stakeholder groups were identified as main target groups and actors of the PARTHENOS project and how they interact with each other. As the research communities are at the core interest of PARTHENOS, they are positioned at the top of the graph. They interact with all other stakeholder groups; however, they are sometimes isolated from each other (which is represented by the light green bars). RIs carry the research communities “on their shoulders”, and are often based on data archives. Data archives are positioned between RIs and CHIs, as they are often part of both RIs and CHIs or at least heavily interact with them (see case study below). CHIs finally are the basis of all research in Humanities studies and are also the base of the stakeholder pyramid.

1.4.4. Case study: an institute in various stakeholder roles - the case of ACDH-OEAW

The Austrian Centre for Digital Humanities of the Austrian Academy of Sciences (ACDH-OEAW) can serve as a case study that shows the difficulties that arise when

trying to formulate stakeholder groups and their relationships to each other. As is often the case, ACDH-OEAW cannot be identified with one single stakeholder role, but rather belongs to several groups: the Centre's aim is to serve the research community by providing it with and hosting its data in its role as a *data archive*. At the same time, ACDH-OEAW is the Austrian national coordinating office for the two European *Research Infrastructures* CLARIN and DARIAH. In this role, the Centre has to service and support several different research communities, maintaining their different standards. Various researchers at the Centre also carry out their own research, thereby representing *research communities* and sharing the needs and requirements of the communities they belong to. Thus, the Centre belongs to several of the stakeholder groups identified by PARTHENOS and consequently, when it comes to data and repository quality, has to fulfil the requirements of more than one stakeholder group. While this situation can cause lengthy development processes and make data creation more intricate, in the best case this situation leads to top-quality repositories and data suited to the standards requirements of manifold research communities. Under such circumstances, data sets are kept in use and thereby alive, to the benefit of individual research communities as well as the Humanities as a whole.

2. Quality assessment of data, metadata, and digital repositories

This chapter is the result of the work carried out by Task 3.2 within PARTHENOS WP3, entitled “Common Policies and Implementation Strategies”. It provides a first overview of the policy landscape at a national and international level, together with a first proposal of high-level recommendations on quality of data and metadata, as well as quality of repositories.

The chapter starts off with a series of definitions, followed by a brief introduction on certification of repositories, and assessment tools and models that may be of interest to institutions that wish to assess their current state-of-the-art with respect to the repositories’ quality.

From Section 2.3 onwards, the focus is on existing policies for the quality of data and repositories, which describes the main undertaking of Task 3.2. The methodology employed in this deliverable takes a strong heuristic perspective, by allowing researchers involved in such investigation to analyse in an innovative way existing policies as well as future developments.

Finally, this chapter will provide a series of high-level recommendations for the identified stakeholders in order to offer a practical instrument both for researcher and for institutions to achieve the best quality possible, even in small and non-institutional environment.

2.1. Defining data and metadata in the Humanities and Social Sciences

The vast availability of digital resources (i.e.: data and metadata, services and tools) in the Humanities research ecosystem fosters the need to raise awareness on the existence and relevance of e-infrastructures: these, in fact, support scholarly research promoting access, interoperability and reuse of Humanities and social science data.

Such practice should involve all the relevant stakeholders: research communities, Research Infrastructures, data repositories, and Cultural Heritage Institutions. As

was pointed out in Section 1.4.2, organizations may assume alternating stakeholders' roles. In this context, the notion of 'data' - and other related concepts like 'metadata' and 'research data' - are crucial, and have to be carefully defined, since they seemed to be too context dependent to be useful without further specifications. In this section, the point of view of the researcher acts as the guiding perspective, since research communities are the main stakeholders of PARTHENOS WP3, and individual researchers often are the actual depositors of the data.

2.1.1. Data

Data is the plural of the Latin word "datum". Luciano Floridi traces back the definition of datum from the term identified by the Greek Euclide "dedomena": dedomena is pure data, not derived from the environment, and represents the data before any cognitive interpretation and processing.[14] In other words, data is "something given", meaning information which is intelligible without being interpreted. This perspective is, without doubt, a bit more difficult to adopt for the Humanities, as what are usually sources for the Humanities (text, images, music, films) are semiotic systems in themselves, that need a first layer of interpretation before being used as the basis for other research and hypothesis.

The meaning of *data* can therefore encompass a range of different possibilities, such as:

- Facts and statistics collected together for reference or analysis (such as in *there is very little data available*).
- The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media (as in any modern computing device).

The term *data* is a crucial element for Information Technology and related disciplines, where it is often used to characterize the very notion of information. The

[14] Shannon, C. E., 1993, Collected Papers, edited by Sloane, N. J. A. and Wyner, A. D. New York: IEEE Press, p. 180; quoted in Floridi, L. 2010, Information: a Very Short Introduction, Oxford: Oxford University Press.

term *information* otherwise has many possible meanings and can be expressed and explained in several different ways, depending on the point of view and expectations of the observer:

> *The word 'information' has been given different meanings by various writers in the general field of information theory. It is likely that at least a number of these will prove sufficiently useful in certain applications to deserve further study and permanent recognition. It is hardly to be expected that a single concept of information would satisfactorily account for the numerous possible applications of this general field.*[15]

Both the terms *information* and *data* share the same kind of polysemy that forces us to carefully select our point of view and build working definitions based on our purposes. Generally, we consider information and data as interrelated concepts: *information* is the *meaning* of data as it is interpreted by human agents (in our case: researchers), *data* consists of facts, which become information when they are seen in context and convey meaning to people. Therefore, we can say that *information* is composed by *data* and *meaning*.

Thus, in order to be meaningful, *information* is made of *data* that should be matching a number of syntactical constraints (i.e.: well-formed) and should be meaningful for the recipient, being the latter a human or a machine.

2.1.2. Metadata

> *Consider how retailers store information about their products and their customers; employers about their employees and their operations; organizations about events they manage; research institutions about trends and notable people in their area; libraries, archives, and museums about the materials in their care; governments about their citizens, their allies, and their enemies –*

[15] Shannon, C. E., 1993, *Collected Papers*, edited by Sloane, N. J. A. and Wyner, A. D. New York: IEEE Press, p. 180; quoted in Floridi, L. 2010, *Information a Very Short Introduction*, Oxford: Oxford University Press.

> *this is all metadata [...] The information we create, store, and share to describe things, allows us to interact with these things to obtain the knowledge we need. The classic definition is literal, based on the etymology of the word itself – metadata is "data about data."*[16]

Cultural and memory institutions have a long tradition of setting up, publishing, and sharing vast amounts of metadata, such as library catalogues and archival finding, providing inventories of books and documents with detailed descriptions of individual items using many different formats and approaches (i.e.: bibliographic approach vs historical approach). There are various categories of metadata, used to support different use cases in the digital domain, and render resources that are Findable, Accessible, Interoperable, and Reusable:

- Descriptive metadata;
- Structural metadata;
- Administrative metadata;
    - Preservation metadata;
    - Technical metadata;
    - Rights metadata.

Among the most common purposes of metadata we consider discovery, identification and/or understanding of a given resource, interoperability, digital-object management, preservation, etcetera; metadata also supports digital object exploration (navigation within parts, pages and/or sections of a given item) and identifies different versions of a given object, by providing technical information (i.e.: resolution of a digital image).

Descriptive metadata include metadata that provides information about the content of a given Cultural Object (i.e.: title, author, publication, subject etcetera). Structural metadata describes the intellectual or physical elements of a digital object, such as information on page lay-out, and a table of contents, which enable search and retrieval as well as navigation of the digital object.

[16] Riley, J., 2017, Understanding Metadata. What is Metadata, and what is it for?, Baltimore, Maryland: NISO.

Administrative metadata, on the other hand, is carrying information needed for resource management. There are different types of administrative metadata. Preservation metadata are needed as a part of the strategies needed to support long-term sustainability of digital resources (i.e.: checksums, hashes, like MD5 etcetera). Technical metadata specifies the technical features of its digital representations, such as file types. Finally, the IPR context of a given digital resource, could be managed using rights metadata (licences, such as a Creative Commons, etcetera).

2.1.3. Datification

The term *digitization* usually refers to the process of encoding parts of the analogue reality - either physical (i.e.: objects, such as manuscripts, sculptures and other artefacts) and/or intellectual (i.e.: texts carried by manuscripts etcetera), to represent, process and store information. To digitize thus means to convert any analogue source of input (i.e.: a text, a sound etcetera) to a series of discrete units represented in a computer. The output of this process is the production, storage and retrieval of digital *data* considered as discrete (single, individual) items of information. The increasing use of IT in Humanities research - as proved by recent studies by Martin Hilbert[17] and other scholars - resulted in a large availability of digital resources that altogether span different periods, languages and documents types (i.e.: sources in manuscript and printed form, secondary literature and bibliographical records).

However, mere digitization - *turning analogue information into computer readable format* - by itself does not datify[18]. The main difference between *digitization* and *datification* lies in the fact that - through the latter - digital data (i.e.: words) stored somewhere are turned into knowledge, by the means of reuse (i.e.: computation, interpretation, etcetera), thus generating valuable information.

[17] Hilbert, M., 2012, How much information is there in the "information society? Significance, 9 (4), 8–12.
[18] Mayer-Schönberger V., Cukier, K., 2013, *Big Data*, London: John Murray, pp. 83-97.

2.1.4. Research data

Research data is a specific type of data that is collected, or created, for purposes of analysis to produce original research results. Research data can be generated for different purposes and through different processes, and can be divided into different categories. Each category may require a different type of data management plan.

- Observational: data captured in real-time, usually irreplaceable. For example, sensor data, survey data, sample data, or neurological images.
- Experimental: data from lab equipment, often reproducible, but can be expensive. For example, gene sequences, chromatograms, or toroid magnetic field data.
- Simulation: data generated from test models where model and metadata are more important than output data. For example, climate models, or economic models.
- Derived or compiled: data is reproducible but expensive. For example, text and data mining, compiled database, or 3D models.
- Reference or canonical: a (static or organic) conglomeration or collection of smaller (peer-reviewed) datasets, most probably published and curated. For example, gene sequence databanks, chemical structures, or spatial data portals.

Since the nature of research data is very diverse, and sometimes it is not immediately clear what research data may actually include, here is a short list of examples:

- Text documents, spreadsheets, slides;
- Laboratory notebooks, field notebooks, diaries;
- Questionnaires, test responses, transcripts, codebooks;
- Audiotapes, videotapes, photographs, films;
- Artefacts, specimens, samples;
- Database contents including video, audio, text, images, 3D models;
- Models, algorithms, scripts;
- Contents of an application such as input, output, log files for analysis software, simulation software, schemas;

- Methodologies, standard operating procedures, protocols.

For this reason, Johanna Drucker has suggested that data in the Humanities is considered as "capta", rather than data.[19] The difference between the two is that "capta" refers to information that is not given in the natural world, but is rather captured or gathered. Trevor Owens has expanded this definition indicating that Humanities data is "multifaceted objects that can be mobilised as evidence in support to an argument". He sees humanist data therefore as a threefold concept. It can be considered as:

- **Constructed artefacts**: data are always manufactured, created by someone. In fact, in the Humanities, the idea of "raw data" can be misleading. The creation of data requires precise choices of what to collect and encode.
- **Interpretable text**: we can think of data as an authored text. Humanists should interpret data as an authored work where the intentions of the author are worth consideration.
- **Processable information**: data can be processed by computers - differently from scientists, for humanists the results from the information processing, are open to the same kind of hermeneutic exploration and interpretations as the original data.

2.1.5. Case study: the CENDARI "data-soup"

The Collaborative European Digital Archive Infrastructure (CENDARI) project is one of the PARTHENOS participating e-infrastructures. CENDARI gathers curated data covering two research areas in the community of "Studies of the Past": WW1 and Middle Ages. It includes data from different sources (mostly across the GLAMs sector) both unique and deposited. The so-called CENDARI 'data soup', contains a wide range of formats and levels of description of data. Recognised and interoperable standards - in use in the different research domains involved - were used to encode data and describe cultural objects and collections (i.e.: EAD for Archival documents).

[19] Drucker J., Humanities approaches to Graphical Display, 2011, *DHQ: Digital Humanities Quarterly* 5, (1).

The CENDARI dataspace contains 829,087 descriptions, represented in several types of data formats. This information is stored in a repository called CKAN, an open source data portal platform developed and maintained by the Open Knowledge Foundation. The kind of file formats and standards, as well as the level of organization and accessibility of data provided by the Cultural Heritage Institutions in contact with CENDARI, vary from case to case: small archives are usually lacking resources for metadata standardization and data storage, therefore their archival descriptions are often accessible via spreadsheets and are not available online (*hidden archives*). National and international archives, instead, usually have a cataloguing and encoding department: nevertheless, they often lack both technical and political means to share their data with other institutions and projects.

Along with the aggregation work on data, CENDARI researchers have also encoded information related to archival descriptions and archival institutions, using the open source software ATOM ('Access to Memory'), promoted by the International Council for Archives and fully supporting all the archival descriptions standards. CENDARI established collaborations with international networks in Digital Humanities, in order to engage communities of scholars and digital humanists: thus, the risk that data collected in the context of research projects become obsolete and unusable is reduced.

2.1.6. Assessment of research data

2.1.6.1. Formal correctness of data

Research data should be understandable to other researchers for purpose of reuse and validation. This means that data formats and metadata would best be assessed from the perspective of a user who has not worked with the data before. Can they find what they are looking for, can they gather the data they need, can they open the data, and can they understand the content?

Metadata needs to be as complete as possible and fully transparent. If codes or variables are used, the explanation of those codes and variables needs to be directly available. In addition, users need to be able to determine which files contain

what kind of data. They need to be able to open the file format, as much as possible independently of the use of specific software or hardware.

The entire dataset could thus be assessed in the light of the FAIR principles:

- Does the dataset have a persistent identifier?
- Is there metadata or documentation available? Is the metadata sufficient for fully understanding the data content?
- Are the metadata accessible?
- Does the dataset have a user licence, are there clear conditions of reuse? Do user restrictions apply?
- Are the data files in a proprietary format, a well-supported 'acceptable' proprietary format, or are they in a preferred/open format?
- Does the data use a standardized coding scheme?
- Is the data linked to other data (how)?

2.1.6.2. Case Study: DANS-KNAW guidelines to preferred data formats

A working group within DANS-KNAW is responsible for monitoring deposited and archived file formats. File formats are evaluated with respect to long-term guarantees in terms of their usability, accessibility and sustainability. DANS published a Preferred Formats guide in September 2015,[20] which details the best options for long-term preservation per file type. These guidelines aim to make data available in file formats which are, as far as possible: open formats; frequently used; independent of specific software, developers or vendors. The working group continuously keeps an eye on new developments, and participates actively in international discussions regarding the subject.

There are still many files in the DANS online archive EASY which pre-date the policies based upon the Preferred Formats guidelines. DANS is spending effort on applying its guidelines and policies in retrospect, ensuring long-term preservation for datasets with files at risk of becoming obsolete. However, in past years, it has often proven to be difficult to properly identify and retrieve files: various file format

[20] DANS Preferred formats. September 2015, Version 3.0, DANS, URL: https://dans.knaw.nl/en/deposit/information-about-depositing-data/DANSpreferredformatsUK.pdf.

identification tools, including DROID and JHOVE, have been used to scan the content of EASY, but the yielded results were found to be lacking.

On the other hand, results for the identification of GIS-files are significant: DANS migrates GIS to the well-supported, text-based MapInfo Interchange Format: MIF and MID. The MID-file was always recognized as a MIDI music file, showing that the tools did not scan the files beyond their extension. Old WordPerfect/Word files are also problematic for scanning tools because of the need to go beyond the file extension: extensions such as '.1' or '.h1'/'.c1' are often seen denoting not the file format but 'Chapter 1'; similar with the use of the '.ind' extension for an 'Index' – variations abound and the logic behind the naming is not always obvious.

Over the last year, DANS made use of the newer tools FITS and Apache TIKA in a self-developed 'File Analysis and Research Module' (FARM), ultimately to great success. While format identifications may remain inconclusive for a number of files, the tools did allow for a thorough identification and retrieval of published but unconverted WordPerfect and Microsoft Word files out of EASY. A Python script (to be made available to the public, expected end of March 2017, when DANS will release its Preferred Formats guidelines as a wiki) was developed in order to mass migrate these files to the long-term preservation format PDF/A, according to the Preferred Formats guidelines.

At the moment, DANS is looking into the best ways in which to store and publish the newly migrated Preferred Formats with the datasets in EASY, including recording provenance administration. The same path can then be followed for other non-preferred file types published in EASY.

2.1.6.3. Content quality of data

The quality of the content of the data should ideally be the responsibility of the data creator. A repository has the task to make research information available for the long-term and has no liability regarding the content quality. Repositories can, however, assist in assessing levels of content quality and, in effect, actively promote high quality content or improving content quality. This could be done by enabling a platform for researchers to give insight into the quality of datasets, including the option to provide feedback on it. Quality assurance through peer review is a common practice as well. As an example, the Research Data Journal for the Humanities and

Social Sciences (RDJ) is a peer-reviewed journal, which is designed to comprehensively document and publish deposited datasets, facilitates their exploration, and contributes to the transparency of research.[21]

2.1.6.4. Case study: DANS-KNAW data reviews

Regularly, DANS asks the users of data to give their opinion on the quality of the archived data. A subjective opinion is given by users which may or may not recommend the data for reuse by others. In May/June 2016 the most recent review survey was organized. By using SurveyMonkey, 107 people were asked to review a total amount of 1,220 pre-selected datasets. This resulted in 82 completed reviews. Of the people who responded, 41% were researchers, 8% students, 23% archaeologists and 28% had another profession. Within the group of researchers, 43% was also working in the field of Archaeology, 21% in Social Sciences, 16% in Humanities, and 4% Life Sciences, Medicine and Health Care. In total, 12% belonged to the group “Other” (e.g. History, Economy).

The users were asked to score the data on a scale from 1 to 5 using 6 criteria: quality of the data, documentation, completeness, coherence, structure and usability of file formats. All the datasets were scored between 3.9 and 4.1 which translates as above average to good. To receive a better insight of the opinion of the participants, a few additional questions were posed concerning the recommendation for reuse: is the data meeting up to the expectations and was the data useful to answer the research question, was it used for a new publication?

The results showed that 79.3% of the respondents did recommend the data, while 8.4% did not. Regarding the usefulness of the data, 63.4% of the respondents could answer their research question, and 6.1% could not. Reasons mentioned for not being able to answer the research question were the fact that the data was not relevant or didn’t meet up the expectations. Finally, 12.2% used the data for a new publication, and 37.8% of the respondents is planning to publish their results based on the used data.

[21] http://www.brill.com/products/online-resources/research-data-journal-humanities-and-social-sciences.

In addition, reviewers were asked to indicate any strong and weak points of the data. Most positive remarks were about the innovative aspects of the data and the fact that the data was clear enough to understand for reuse. Completeness and structure of the data was sometimes criticized, together with the fact of missing codebooks or missing files. Compared to previous years, the quality of the datasets is scored at roughly the same level as in 2016. However, less people recommend the data for reuse (79% instead of 90%). Reasons mentioned were the fact that the data was not relevant or didn't meet the expectations. The fact that the number of people that responded was lower may have had an influence on this.

There is a growing demand for quality criteria for research datasets. The DSA (*Data Seal of Approval for data repositories*) and FAIR principles (Findable, Accessible, Interoperable and Reusable) get as close as possible to giving quality criteria for research data. They do not do this by trying to make value judgements about the content of datasets, but rather by qualifying the fitness for data reuse in an impartial and measurable way. By bringing the ideas of the DSA and FAIR together, the long-term preservation of data of high quality will be guaranteed (see also Section 2.2.1.3).

2.1.6.5. Case study: the ADS criteria for evaluating datasets

Archaeology Data Service (ADS)[22] is the mandated archive for data of many projects funded by the Arts and Humanities Research Council and the Natural Environment Research Council. ADS promotes standards and guidelines for best practices in the creation, description, preservation and use of archaeological information. As part of the ARIADNE infrastructure, ADS's has been selected to represent the best practices in the archaeological domain concerning the assessment of data quality. This case study concerns the guidelines developed by ADS to guide depositors to provide good quality data and shows what are the criteria adopted by ADS to evaluate data deposited in their archives.

The ADS accepts a wide spectrum of archaeological data types, including CAD files, databases, digital photography, geophysical and other survey data, GIS files, images and drawings, satellite imagery, spreadsheets, texts, and virtual reality files.

[22] http://archaeologydataservice.ac.uk/.

A rigorous process of peer review of materials proposed for accessioning is available. Especially, when the suitability of a dataset for archive or its potential reuse is unclear, the ADS refers to a Collections Evaluation Working Group drawn from its Management Committee to assist in evaluating datasets and maintaining the rigorous standards necessary for the effective development of a quality resource base. Data resources are evaluated according to the following criteria:
- their intellectual content and the level of potential interest in their reuse,
- how and whether they may viably be managed, preserved, and distributed to potential secondary users,
- the presence or absence of another suitable archival home.

The following guidelines are extracted from ADS' Collections Policy (6th Edition, 1st April 2014) and it is available at:
http://archaeologydataservice.ac.uk/advice/collectionsPolicy.

Assessing intellectual content

A review process ensures that the content of datasets is of the highest intellectual quality: collected and recorded according to accepted archaeological standards. Assessing 'quality' is a subjective exercise and in this the ADS will be guided by the following principles:

Evaluating preservation potential and reuse value

The reusability of datasets is largely determined by community needs. Inevitably requirements continue to change and consequently the assessment of user needs forms an ongoing part of the activities of the ADS. Reuse value is also determined by the formats in which data are stored. If proprietary software packages form the basis of data entry/retrieval, and a majority of archaeologists do not have access to these proprietary systems, the dataset may be ranked low on the reuse value criterion.

Adequate documentation

The quality of datasets will be affected by whether or not they are accompanied by an appropriate level of documentation. This documentation should relate to both the content and the technical format of the resource. Documentation provides important

detail about the context in which data was created and maintained before archiving, and about the relationships between the dataset and other information sources.

Suitability for digital preservation

If the format in which a dataset is stored means that the digital resource is irrecoverably obsolete upon presentation to the ADS this will be sufficient reason for recommending that the dataset not be accessioned.

Determining need of primary archival home

There is no need to duplicate digital archiving services. If a resource, however, is deemed to be of particular value to its user community, the ADS will seek to enter into a partnership with a collaborating agency in order to provide access to it.

2.2. Assessment of repositories

The data and metadata as defined and described in the previous section, are ideally stored in a digital repository that complies with established quality requirements, to ensure the appropriate preservation, dissemination, and accessibility of research data. In this section, we focus on how to assess the quality of these digital repositories.

For the assessment of repositories, the established quality requirements largely deal with data management, not with the data itself. However, the assessment framework for repositories is not entirely unrelated to the actual data and metadata. In fact, when comparing the criteria for assessing repositories' quality to the FAIR principles dealing with data quality (see Section 1.3), it becomes clear that both standards rely on very similar underlying principles.

In order to assess the quality of a repository several tools and models have been developed over the past decades. These tools and models expound on the criteria against which the quality of a repository is assessed, providing detailed guidelines that help organizations improve the services related to their repositories.

Section 2.2.1 describes the framework that is commonly used to certify the trustworthiness of a repository and addresses the commonalities between the

assessment of repositories and the assessment of data. Section 2.2.2 discusses a number of tools and models that are used to assess the quality of repositories.

2.2.1. Certifications

Certification of a repository serves to show to its stakeholders that it lives up to a certain standard. By going through the certification process a repository critically assesses, or is critically assessed on, its workflow and procedures regarding preservation and dissemination. The extent to which a repository meets the standards determines whether or not a certain type of certification can be granted to it. The result is that stakeholders can easily assess the quality of a repository, or at least specific aspects related to preservation and dissemination. As well, the assessment process helps repositories gain insight into their current state and provides pointers for making improvements.

The key point here is the *trustworthiness* of a repository. This can be summarised as a set of requirements guaranteeing long-term access and preservation of the data in a repository, under clear rights and licences.[23] These rights and licences are dependent on national or discipline-oriented laws and codes of conduct. The legal side is, however, not the only dimension of trustworthiness, there is also the technical trustworthiness, in particular the authenticity and integrity of the data which have to be kept as well as their security.

2.2.1.1. Certification of "Trustworthy Digital Repositories"

For a number of years, a framework has existed for the certification of "Trustworthy Digital Repositories". Within this framework, a data repository is trustworthy if it has the mission to provide reliable, long-term-access to digital resources, now and in the future. In addition, the repository needs to indicate its understanding of the threats and risks to the data within its systems. Finally, a data repository can be considered trustworthy only if it is subject to a regular cycle of audit and/or certification.

This framework has *three* levels, in increasing trustworthiness:

[23] See also the webinar offered by DANS-KNAW "Core Trustworthy Data Repository Requirements": https://www.youtube.com/watch?v=VFLTJ7D2y5s&feature=youtu.be.

1) **Basic Certification** is granted to repositories which obtain DSA certification: Data Seal of Approval (DSA).
2) **Extended Certification** is granted to Basic Certification repositories which *in addition* perform a structured, externally reviewed and publicly available self-audit based on DIN 31644: nestorSeal.
3) **Formal Certification** is granted to repositories which *in addition t*o Basic Certification obtain full external audit and certification based on ISO 16363: the ISO-certification.

The first two levels are based on self-assessment, combined with external review. The third and highest level, the formal certification, however, is based on a full external audit.

2.2.1.2. The Data Seal of Approval and the World Data System

The DSA, the Data Seal of Approval is a lightweight form of certification, the basic certification. Recently, the DSA has been combined with the World Data System (WDS) certification of ICSU. These two, until now, independent certification systems have been merged. On November 25th 2016 the WDS and the DSA Boards have announced the availability of their unified and now "common" requirements. This involves some changes. These are, however, not of a fundamental nature: the DSA remains a basic certification standard for "Trustworthy Digital Repositories", based on a self-assessment of these guidelines by the repository that is peer-reviewed by external reviewers. There are still 16 guidelines; they are now, however, labelled as "requirements" to assess the trustworthiness of a repository. On the whole, these new requirements cover almost the same content as the old guidelines; they are, however, structured in a rather different way. The requirements are now broken down in three parts: organizational infrastructure (six requirements), digital object management (eight requirements) and technology (two requirements). This shows the difference with the old guidelines. These were divided into data producers (three guidelines), data users / consumers (three guidelines) and the repository (ten guidelines). Another change is that the self-assessments now will be reviewed by two reviewers instead of one as in the old DSA system.

2.2.1.3. Commonalities between the DSA Certification and FAIR Principles

It is noteworthy that the FAIR principles, directed at individual datasets, are remarkably similar to the underlying principles of DSA.

DSA principles (for data repositories)	FAIR principles (for data sets)
data can be **found** on the internet	**F**indable
data are **accessible**	**A**ccessible
data are in a **usable format**	**I**nteroperable
data are **reliable**	**R**eusable
data can be **referred** to	(citable)

Table 2.1: Commonalities between the DSA principles and FAIR principles (based on a presentation by Peter Doorn, DANS, International Open Access Week 2016).

The resemblance is not perfect as there are some features in both systems that do not seem to match. However, upon closer inspection, DSA and FAIR could be easily combined:

Usable format (DSA) is an aspect of interoperability (FAIR)

Interoperable in FAIR means “ready to be combined with other datasets by humans as well as computer systems”. In DSA usable (meaning preferred) format is an explicit and important item, covered in requirement R8 “Appraisal”. These are the data formats of which the repository can more or less guarantee the long-term preservation of the datasets.

Reliability (DSA) is a condition for reuse (FAIR)

Re-usable in FAIR means “ready to be used for future research and to be processed further”. In DSA this is part of requirement R7: “Data integrity and authenticity”. Authenticity should cover the degree of reliability of the original deposited data.

FAIR explicitly addresses machine readability

This is a basic characteristic of FAIR: it has been an explicit aim in formulating the FAIR principles that data should be designed, formatted and provided with metadata in such a way that they can be found and used automatically by *machine*s. In the DSA this is certainly not an explicit goal. However, in the requirement R13 “Data discovery and identification”, in which the repository is required “to enable users to discover the data”, one of the criteria is whether the repository facilitates machine harvesting of the metadata.

Citability is in FAIR an aspect of findability

Findable in FAIR means “easy to find by both humans and computer systems and based on mandatory description of the metadata that allow the discovery of interesting datasets”. In DSA requirement R13 “Data discovery and identification” pays attention to “proper citation” and the use of recommended data citations.

To conclude, for linking FAIR and DSA the following points should be considered:

- There is a growing demand for quality criteria for research datasets.
- The ideas of DSA and FAIR could relative easily be combined. The points on which they differ can be surmounted.
- Both principles can be used as quality criteria:
    - DSA for digital repositories.
    - FAIR for research data.
- Combining the two principles will make them easy to implement for any trustworthy digital repository.

The DSA assessment of repositories does not address the quality of the data sets, but it would be a desired goal for a repository to be able to declare to what degree each data set held is FAIR. Determining the FAIR-ness of a data set would currently require elaborating the FAIR principles towards automatic scoring of the FAIR-ness of a dataset. The repositories could work towards incorporating the FAIR principles in this way whilst making them implementable in any trustworthy digital repository.

2.2.1.4. Case study: DSA assessment of the CLARIN Repository at University of Copenhagen

The CLARIN Repository at University of Copenhagen is a centre in the CLARIN network of data repositories.[24] The CLARIN repositories can go through an assessment to get the CLARIN B-centre certification, which certifies that the centre provides sustainable access to resources and provides metadata and persistent identifiers (PIDs) for the resources that are accessible through the repository. The requirements for obtaining a B-centre certification are both a Data Seal of Approval (DSA) certification, and a successful assessment of the CLARIN B-centres checklist. Within CLARIN, B-centres are reassessed every three years, which corresponds to the validity duration of the DSA seal of approval. The CLARIN Centre at University of Copenhagen (UCPH) is currently going through reassessment including the new DSA assessment, as well as the CLARIN B-centre checklist assessment.

In 2012, the CLARIN community choose to use the DSA as an assessment criterion, since the DSA was widely recognised outside CLARIN as an assessment for data repositories. In 2016, the ICSU World Data System (WDS) and the Data Seal of Approval (DSA) Board announced a set of unified Requirements for Core Trustworthy Data Repository certification and released a new version of assessment criteria. With this collaboration, DSA certification is expected to become more widely recognised as validation of the trustworthiness of a repository.

The DSA has 16 requirements that cover a broad spectre of issues, from an organisational focus (mission, funding, continuity of access, skilled staff and guidance available), to data preservation focus (relevance, integrity and authenticity of the data, metadata, licences and access rights), as well as technical issues of the repository. A total of 90 sub-questions need to be answered. For all questions in the application form, externally accessible URL's and links should be provided to enable the reviewers to check that the statements are available for the users and covers the subjects asked for.

As an example, requirement 8 states that data and metadata should be based on defined criteria to ensure relevance and understandability for data users. Here the UCPH CLARIN group ensures the quality control checks of the repository. The data has to comply, technically, with the validation requirements of the UCPH

[24] CLARIN has different types of centres, and a list of centres can be seen at https://centres.clarin.eu/.

repository, which are specified on a web page. At ingest, all metadata will be validated by schemas. In addition, a guide for depositors has been written and it is also stated that the data providers are expected to deliver data meeting the academic standards specified in the Danish Code of Conduct for researchers.

During the preparation of the assessment application, the UCPH CLARIN group has made more information available on its website and has also identified areas where procedures needed improvements, such as information about backup and recovery. Currently, the group is awaiting the response from the DSA reviewers. It is expected that in some cases the documentation has to be clarified or extended. The feedback about where the procedures have to be more elaborated or better documented to show trustworthiness to the users will be most welcomed by the repository.

2.2.1.5. Nestor Seal

The Nestor Seal is an "Extended certification" which uses a plausibility-checked self-assessment. The Nestor procedure[25] is based on the specifications detailed in DIN 31644 and contains 34 criteria. Whereas the first 12 have to be implemented, for the others the repository can either describe why it is not applicable or explain to what extent the criterion is planned or implemented. When applying for assessment, the repository is expected to produce documentation for all criteria and submit the documents for a plausibility check by a Nestor reviewer. If the Seal is obtained, it is valid indefinitely. However, its relevance is likely to diminish after a number of years unless a further review is conducted. Similar to the DSA, the focus of the assessment is on the solutions used by the digital archive and not on the quality of the archived content.

2.2.2. Assessment tools and models

Part of the certification process is the assessment of the repositories' quality, either as a self-assessment or an external one. There are several tools and models in use that assist organizations in assessing the current state of quality of their repository.

[25] Nestor Certification Working Group: Explanatory notes on the nestor Seal for Trustworthy Digital Archives (Version 1, in English).

In that sense, they complement the certification requirements: while the requirements for obtaining a certain level of certification state the envisioned end state (see for example the DSA requirements), the tools provide a means of assessing the current state of development, indicating the areas that need to be addressed in order to meet the certification requirements. In addition, on the level of any organization dealing with data management, the tool can be used to raise awareness for data quality in general. By pointing to international tools and models, organizations may inform the data depositors about the many aspects that are relevant for assuring high quality digital preservation of data.

An extensive, up-to-date lists of digital preservation tools (including tools for ingest) is available from COPTR, the Community Owned digital Preservation Tool Registry: http://coptr.digipres.org/. A few models that can help preservation initiatives and maturing repositories are given below.

The CESSDA SaW Capability Development Model (CESSDA-CDM)[26][27]

This model provides both a starting point for emerging preservation initiatives and a reference tool for established archives that want to improve their services. It has been developed within the CESSDA Strengthening and Widening project (CESSDA SaW) and builds on the Reference Model for an Open Archival Information System (OAIS) as well as the European Framework for Audit and Certification (also known as Trusted Digital Repository EU (TDR-EU). “Although the main emphasis of the model is on social science research data, it is applicable for all organisations that have taken on the responsibility to preserve and keep data understandable for the long term, and make it available and accessible for a user community.” This means it is relevant for repositories in general.

The model has three distinct hierarchical levels, each looking at the characteristics of effective preservation processes and activities of an organisation. The maturity of activities is evaluated on a scale. Through careful consideration of the organization's goals, objectives and activities, an insight is reached on its state of

[26] CESSDA SaW Deliverable 3.1 *Heuristic Maturity Development Model*, [submission date 30-05-2016].
[27] CESSDA website: http://cessda.net/eng/CESSDA-Services/Projects/CESSDA-SaW/Work-Packages/WP3/CESSDA-CDM/Introduction.

development. Once an organization has gained sufficient insight in its current state of affairs, this provides the groundwork for improving the quality of a repository.

The CESSDA-CDM model applies the following three consecutive levels:

- Level 1: Capability Requirement Areas (CRA)
- Level 2: Capability Process Areas (CPA)
- Level 3: Objectives and associated Required or Expected Activities

On the first level, the main Capability Requirement Areas (CPA) are identified, that each address a core area of assessment: 1) Organisational Infrastructure, 2) Digital Object Management, and 3) Technical Infrastructure. Moving on to the next level, these three CRA's are further itemized into more detailed processes. This constitutes the second level of the model, the Capability Process Areas. Finally, on the third level, these processes are connected to concrete objectives. The objectives are then assessed based by analysing the corresponding activities that the organization carries out, that contribute to achieving the particular objective. Figure 2.1 gives a graphical representation of the connection between the different levels.

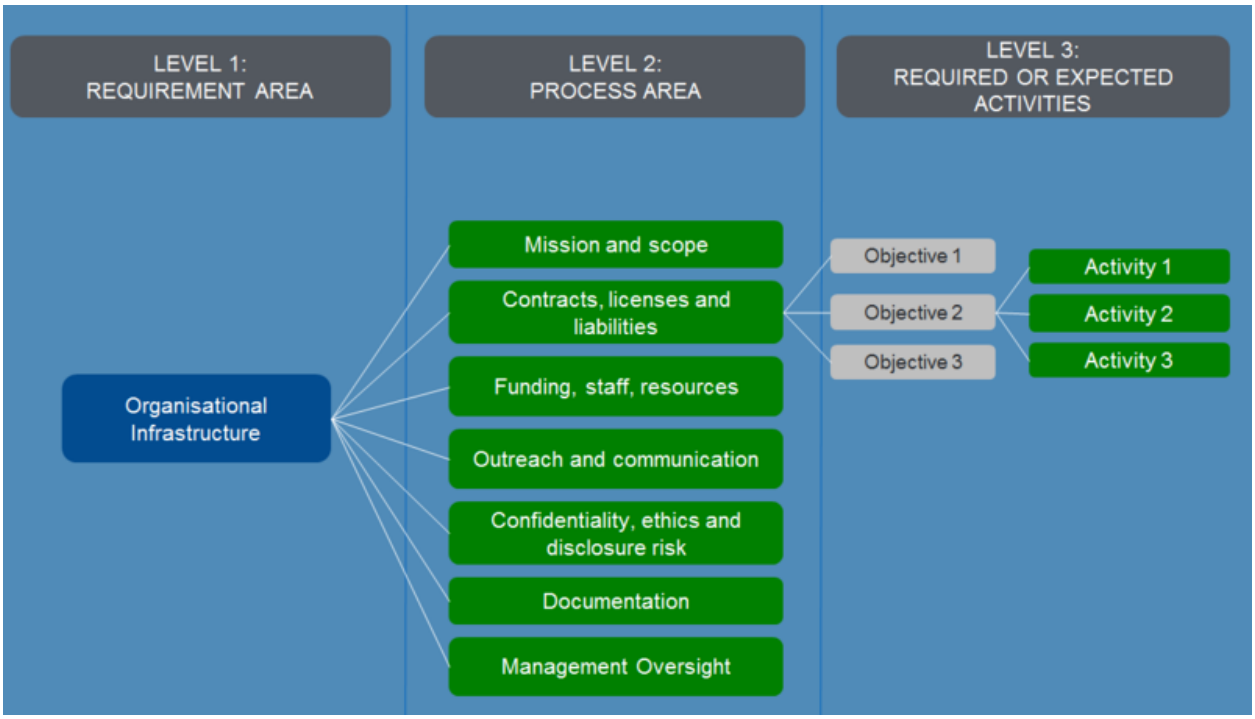


Figure 2.1: Three consecutive levels of assessment in the CESSDA-CDM model.[28]

Collaborative Assessment of Research Data Infrastructure and Objectives (CARDIO)[29]

This tool helps to achieve a good data management strategy and can be applied to repositories. It was developed to integrate popular features from other current digital curation tools. It focuses on three main aspects: Organisation, Technology and Resources and enables organisations to look critically at these aspects. Typical of this tool is the requirement to establish a diverse collaboration team that includes researchers as well as representatives from information services and other support services.

The Data Asset Framework (DAF)[30]

The Data Asset Framework (DAF), and the corresponding online tool, was created to assist organizations that seek to construct a registry of data assets. DAF was developed in a project led by Humanities Advanced Technology and Information Institute of the University of Glasgow, in close collaboration with the Digital Curation

[26] https://cessda.net/eng/CESSDA-Services/Projects/Current-projects/CESSDA-SaW/Work-Packages/WP3/CESSDA-CDM/Introduction/Model-Components/Objectives-Activities-and-Capability-Completeness.
[29] http://www.dcc.ac.uk/resources/tools/cardio.
[30] http://www.data-audit.eu/.

Centre of the University of Edinburgh. The DAF methodology identifies 4 consecutive stages in audits of research data assets:

1) Planning the audit
2) Identifying and classifying assets
3) Assessing management of data assets
4) Reporting and recommendations

The first stage intends to define the purpose and scope of the audit. Secondly, identifying and classifying the existing data assets helps to further detail the scope of the further audit activities. Since only selected aspects will be addressed in more detail, the classification of the research data serves to highlight those aspects that are most relevant to be assessed concretely in the third phase. Finally, based on the outcomes of the third phase, a final report is written and recommendations are formulated, which in turn can be used as a starting point for developing an organization's Data Management Plan (DMP).

2.3. Policies for the quality of data and repositories

This chapter takes into consideration policies related to the quality of data, metadata and repositories, in order to offer researchers, data archives, and policy makers a set of guidelines and recommendations on how to make research outputs widely available to research communities as well as to society at large.

2.3.1. Census of quality policies in the disciplines identified by PARTHENOS

While building an overview of the policy landscape among the Humanities disciplines in PARTHENOS, Task 3.2 decided to involve its own researchers in the *creation of a census of the policies related to data (including metadata) and repository quality in their own field of research* (see Section 1.4.1.1): Language Studies, History, Social Sciences, Archaeology, data archives and Cultural Heritage Institutions. The aim of this investigation is to collect a wide variety of policies for each PARTHENOS

discipline, therefore allowing our task to identify *commonalities* as well as *gaps* for the single disciplines.

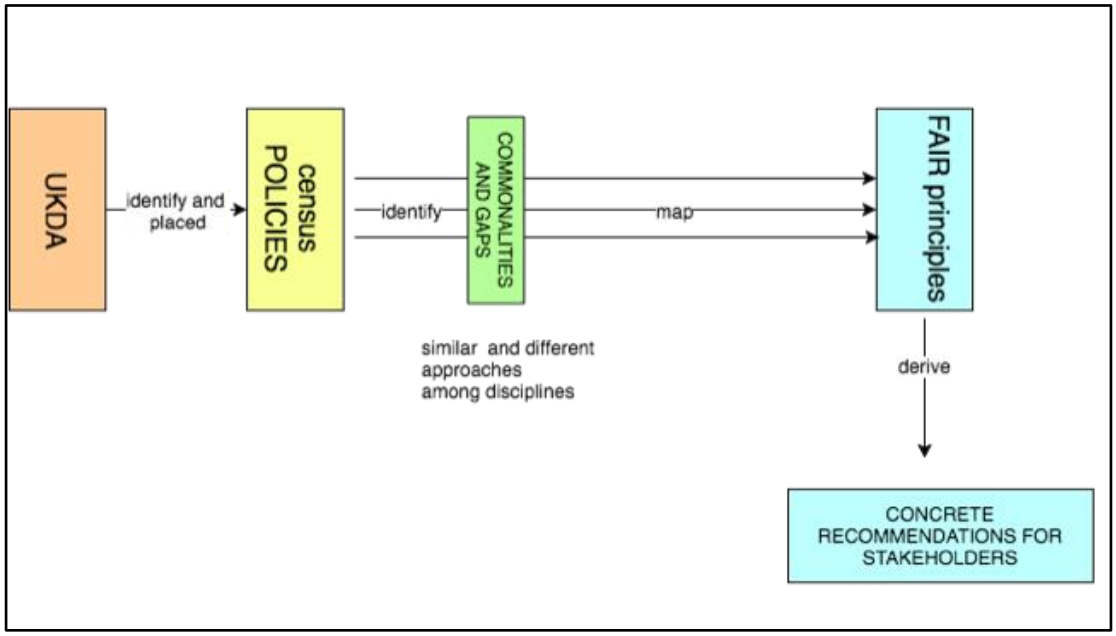


Figure 2.2: Approach and methodology applied in Task 3.2.

Figure 2.2 shows the process that Task 3.2 has followed from the design and methodology of the information retrieval, through the mapping of the collected policies to the FAIR principles. Finally, a set of concrete recommendations to the PARTHENOS Stakeholders were derived.

Our investigation began with the analysis of the *UKDA lifecycle*[31] *phases*, which is the data management data lifecycle that was taken into consideration for the research and the deliverable in WP2 – D2.1 Community involvement and requirements. According to the different phases of this data lifecycle, we identified *formal* as well as *tacit* policies that play a role in specific steps or moments of the data lifecycle. (see also Figure 2.4 below).

Two phases were crucial in the work of Task 3.2, namely the identification of *commonalities* and *gaps* (in green), and the mapping of the common policies to the *corresponding FAIR principles* (blue). These two phases are recognised as highly relevant in the activities of Task 3.2 as they represent the transition in our investigation from unidentified of policies in use by a certain discipline (if any at all) to

[31] http://www.data-archive.ac.uk/create-manage/life-cycle.

an organized set of high-level recommendations which ideally will be valid for all the disciplines under consideration by PARTHENOS.

Formal and tacit policies

In our investigation on existing policies for the Humanities, we realized that not all disciplines refer to policies for data quality in the same way. Some disciplines, for example, refer to a well-established corpus of rules for the creation, management and sharing of data. In this case, these "rules" are available in written form, are widely shared, well known and agreed upon among the same research community, and we refer to them as *formal policies*. This is the case of disciplines such as Archaeology and Linguistics. Archaeology in the last decades has in fact developed a wealth of resources and recommendations on how to best create data and metadata, both at a European and national level. Research institutes and data archives have also developed reliable exchange protocols and recommendations for data repositories.

Differently, disciplines as History have relied more (and still do) on *tacit policies* for data quality. This means that no written documentation and recommendation has been provided by and for the community, and that the main directions for data and metadata quality are disseminated via informal channels and informal media. This makes tacit policies quite unreliable, unstable and subject to interpretation.

2.3.2. Methodology of the quality assessment

The first step was that of ideating and designing a structured table where the researchers involved in Task 3.2 would be able to place relevant policies related to quality and in use by their own disciplines. The other requirement for this table was that it needed to allow collaborative work among the researchers. For these reasons, we opted for a spreadsheet in Google Drive, which has the advantage of being accessible by multiple people at the same time.

How did we recruit the researchers and other stakeholders for the data collection?

Task 3.2 is composed of a wide variety of researchers, in particular in the fields of contemporary History, Language Studies, Archaeology – and outside the research field – data archivists, as well as people involved in Cultural Heritage Institutions. In this way, it was deemed appropriate to assign one stakeholder to each tab. We also enriched the document with information from literature and desk research, as these also represent an important source of information on the stakeholder's policies landscape. When we felt that we were missing relevant information, we interviewed external experts, such as for Social sciences, Oral History and Language studies.

The table, which we called "Matrix Roles, Tasks, Quality" (see Figure 2.3 below) is the result of our initial investigation, as well as the starting point of our analysis; here we collected (and are still collecting) the necessary information to depict a rich scenario of policies on data and repository quality from stakeholders.

TASKS (aligned with UKDA) i.e. Archaeologist	DATA QUALITY	METADATA QUALITY	REPOSITORY QUALITY
DATA CREATION During the CREATION OF DATA, what are the POLICIES that result in good practice in respect to ...	KNA (NL - Dutch quality norm archaeology) Guide to good practice: ADS (UK - archaeology data service)	The Standard and Guide to Best Practice in Archaeological Archiving in Europe (ARCHES project, Europae Archaeologiae Consilium (EAC)	Archaeological exchange protocol, SIKB0102 (NL)
DATA MANAGEMENT During the DATA MANAGEMENT, what are the POLICIES that result in good practice in respect to ...	Guidelines to create 3D models of cultural heritage objects with no established digitization (3D Icons project)	N/A	DANS privacy regulation (anonymizing the data)
DATA PRESERVATION During the DATA MANAGEMENT, what are the POLICIES that result in good practice in respect to ...			Repository certification (general); DANS Preservation Policy; Persistent Identifiers (general)
DATA REUSE During the DATA REUSE, what are the POLICIES that result in good practice in respect to ...	Data contracts (in use by NWO); Persistent Identifiers (general); Data Citation Guidelines; Condition of use and Privacy	N/A	Repository certification (general); Persistent Identifiers (general)

PARTHENOS-project.eu 22

Figure 2.3: Matrix Roles, Tasks, Quality.

The table represented in Figure 2.3 is organized as follows: in the left column, the phases of the UKDA Data Lifecycle are listed (Data Creation, Data Management, Data Preservation, Data Reuse). In the top row, the main topics under investigation for Task 3.2 are listed: Data Quality, Metadata Quality and Repository Quality.

The questions that each stakeholder was asked to answer while working on this table were:

1) What are the policies during the phases of data creation, management, preservation and reuse that help me achieve better data and metadata quality?
2) What are the policies during the phases of data creation, management, preservation and reuse that help me in achieving better quality repositories?

Each cell in this table is meant to contain one or more policies, in use by the stakeholder and his/ her community.

For each PARTHENOS stakeholder (see Section 1.4) that participated in our policy census, a tab was created in the Google spreadsheet: Archaeology, Language Studies, Social Sciences, History, but also Cultural Heritage Institutions as well as data archives. Whilst not all of these are research disciplines, we have recognised them as crucial viewpoints and stakeholders, whose knowledge and practice in terms of data quality had to be collected and documented in our deliverable.

2.3.2.1. Considerations on the methodology

The approach outlined previously proved not to be an easy task for the researchers involved in the creation of the Matrix Roles, Tasks, and Quality. Probably the most difficult part was reflecting on the common policies in use by a certain discipline or type of stakeholder. In some cases, these policies are created and documented within a certain community, they are widely shared, updated and discussed (in the case of the archaeologists, for example). Under these circumstances, it was relatively easy to identify the reference policies. In other cases, (especially in the discipline History) there are no or very few formal policies related to the quality of data or repositories. This makes the formalization of quality policies even more difficult. In this last case, we proceeded with the identification of *tacit policies*, which we interpreted as a set of shared rules by a certain research community, but not formalised or written nor disseminated.

The mapping of quality policies for every stakeholder in the PARTHENOS network was not the main goal of this exercise, however. This was mainly a prerequisite to ultimately identify the *common policies* among the disciplines

analysed and to build a common policy framework for data, metadata and repositories from there, which could be applied to the Humanities. Internally we have called these common aspects and policies *commonalities* (see Section 2.2.1.3).

2.3.3. Overview of policies

This section gives an overview of the policies that we have collected so far, as of April 2017. The lists of policies, even though very rich, are not exhaustive yet: they will continue to be enriched until the submission of the final deliverable of WP3.

Even if not complete, this first census of quality policies (at a local, national and international level) gives us a blueprint and a methodology on which to base the work for the next two years. Table 2.2 shows a selection of policies identified in the field of Archaeology. For matters of space, we include in this document only a selection of the identified policies. The entire table collected with the detailed policies per discipline is available at the end of this document in Appendix II: Matrix 'Roles, Tasks, Quality'.

2.3.3.1. Research community - Archaeology

Policy	Link	Country
Archaeological Documentation	http://www.mnm-nok.gov.hu/wp-content/uploads/2013/01/b-ERD-szakmai-%C3%BAtmutat%C3%B31.pdf	Hungary
Guidelines for archaeological Measurements	http://www.bundesdenkmalamt.at/documents/621701608.pdf	Austria
DANS Data Management Plan for managing, documenting and sharing data	https://dans.knaw.nl/en/deposit/information-about-depositing-data/DANSdatamanagementplanUK.pdf	Netherlands
Specialist Recommendation for data and Metadata	https://wiki.de.dariah.eu/pages/viewpage.action?pageId=20058160	Germany, Europe
3D Icons Guidelines	http://3dicons-project.eu/eng/Guidelines-Case-Studies	Europe
Archaeological Exchange Protocol	https://dans.knaw.nl/en/deposit/information-about-depositing-data/archaeological-exchange-protocol	Netherlands
The Standard and Guide to Best Practice in Archaeological Archiving in Europe	http://archaeologydataservice.ac.uk/arches/attach/The%20Standard%20and%20Guide%20to%20Best%20Practice%20in%20Archaeological%20Archiving%20in%20Europe/ARCHES_V1_GB.pdf	Europe
A Framework for Transforming Archaeological Databases to Linked Ontological Datasets. In Computer Applications and Quantitative Methods in Archaeology	http://www.tracingnetworks.ac.uk/publications/CAA2010/paper.pdf	United Kingdom
Art & Architecture Thesaurus - AAT	http://www.getty.edu/research/tools/vocabularies/aat/about.html	United States
Quality management of 3D Cultural Heritage replicas with CIDOC CRM	http://ceur-ws.org/Vol-1117/paper6.pdf	Europe

Table 2.2: Examples of collected policies in the field of Archaeology.

2.3.3.2. Considerations on the current quality policy landscape

Some of the stakeholders and disciplines are more advanced than others in terms of policies. While this is well known in the field of research data management, it is interesting to note what are the disciplines and the stakeholders that are in more or less pressing need of improving their approach to data, metadata and repositories?

From an initial investigation, including interviews and desk research, the communities of archaeologists and data archives are those that have addressed and created policies for the achievement of high-level data, metadata and repositories most actively.

The probable explanation for this is the fact that Archaeologists are aware of the importance of archiving their data in a sustainable manner, as the data coming from an excavation is unique and can only be documented once. Archaeologists have been among the first to face the problem of finding a shared standard to describe archaeological artefacts. As for the data archives, their main aim of a data archive is to ensure that the quality of the data preserved is as good as possible. For this reason, the quality of data archives is often certified, in order to guarantee to the end user, the highest quality concerning the preservation of their data.

Archaeology

At a European level, communities of archaeologists and research centres have been involved in archaeological multiparty research projects and Research Infrastructures for the last years, such as ARIADNE.[32] The main goal of ARIADNE was to interconnect country specific experiences, by producing a European layer of models and standards that anyone in the community could adopt.

From our desk research and series of interviews, it has emerged that the community of archaeologists focused in particular on standards to be used for Archaeology-specific data creation processes (e.g. Lidar Documentation, 3D documentation, Dendrochronology standards), on vocabularies and on research data management plans. New trends are also emerging, such as the application of LOD and machine learning to large archaeological datasets.

[32] http://www.ariadne-infrastructure.eu/.

Language Studies

In Linguistics, the need to share and create international data repositories has increased over the past decades, and therefore the demand to create shared standards, such as the ISO standards for language code. Furthermore, the creation of a European Research Infrastructure for the linguistic community such as CLARIN[33] has certainly promoted the use of shared standards as well as the agreement on the use of shared and community best practices (e.g. the CLARIN on PID Policy Summary).

Social Sciences

Social Sciences have a long tradition in the definition of guidelines and policies, in particular in relation to research ethics and (high-level) guidelines for interviewing, documenting and archiving personal testimonies.

In relation to this last point, there is a strong need among this community to protect the personal information that may emerge during the interviews. This has resulted in national and international policies for data anonymization and development of access and reuse policies for social science data.

More recently, the community of Social Sciences has invested both in the data standardization topic (e.g. DDI alliance[34]), and in standardized exchange protocols for quantitative data (e.g. sdmx[35]). Also, the recent creation of the CESSDA[36] Research Infrastructure represents the willingness in the Social Sciences to develop and confirm shared research practices as well as to establish an international community of practice and interest.

History

The research community of historians is the community that seems to have the most conflictual relationship with recognised and formal policies. This doesn't mean that the community has refused "a-priori" to develop any data policy in their day to day work. Instead, this probably never emerged as a concrete need. There are a number of considerations to reflect on in the case of History and data policies.

---

33 https://www.clarin.eu/.
34 http://www.ddialliance.org/.
35 https://sdmx.org/?page_id=5008.
36 https://cessda.net/.

The phase of data creation is very different from that of a linguist and a historian. While the first collects data in the field, through interviews, user panels etcetera, the latter's investigation mainly takes place in archives, libraries and museums. This means that historians, rather than creating new data, tend to collect, organize and make sense of already existing data.

Historians don't create or organize data according to a shared data standard or according any formal policy. The way historians archive their data depends on the single researcher or on tacit policies shared by a certain community of practice. There are some exceptions, however. Oral History for example places itself at the convergence of a number of disciplines, like History, Language Studies, and Social Sciences. Oral History does create new data and does follow communities' policies on how an interview should be conducted, and stored, and on how sensitive data should be protected.

2.3.4. Strengths and weaknesses for each stakeholder and discipline

After completing the first phase, we were able to answer the following questions: *are there similar approaches to data and repositories quality among the investigated disciplines? Can these disciplines learn from each other by avoiding the fragmentation of efforts and by sharing more information between each other?*

The avoidance of the risk of fragmentation is, in fact, one of the main driving forces for the project PARTHENOS, and WP3 has made the need for policy integration between Humanities and social science disciplines one of its primary goals. How can a researcher in History reuse data policies in use by linguists, for example?

In order to answer this question, the second task was to investigate existing models that would enable us to organize the policies in a clear, immediate way and to make them shareable with the heterogeneous PARTHENOS Community.

Secondly, the mapping of the identified policies for each discipline to the FAIR principles (see Section 1.3) represents a crucial phase in our research, as it is the first step toward the accessibility and dissemination of the PARTHENOS high-level recommendation to the interested stakeholders.

Table 2.3 below shows a categorization according to how “strong” or “lacking” a certain discipline is in terms of policies related to data and repositories policies. This categorisation helped us to understand whether the disciplines and stakeholders analysed could exchange their expertise in terms of quality to other disciplines.

For example, how could the expertise of disciplines such as Archaeology or Language Studies in terms of PID (persistent identifiers) be transferred to historians when creating new data that hasn’t been described yet in the archives?

Discipline	**Strong in...**	**Weak in...**
Archaeology	Data standard, exchange protocols, enrichment, copyright and licence, sensitive data, data management plan	
Language Studies	Data management plan, data standards, exchange protocols, authority files, Citation Guidelines	Licences, copyright,
Social Sciences	Sensitive data, Data management plan, Data Anonymization, Exchange Protocols	Licence, copyright
History	Provenance, Rich Contextual Metadata,	Data Standards, Exchange Protocols, Annotation standards, Data Management Plan
Data archives	PID, preferred formats, Certification of repositories	
Cultural Heritage Institutions	Discipline specific data standards, Exchange protocols, IPR and copyright	Data Management Plan

Table 2.3: Discipline policies categorized related to data and repositories policies.

This comparative analysis as represented in Table 2.3 also allowed us to see what the policies are that are relevant and common to all the disciplines. We have thus

called them “commonalities”. The found commonalities have been then revisited and structured to guidelines, as we will explain in the next section.

2.4. From commonalities to recommendations

With the necessity to structure the commonalities and guidelines in a way that was easily understandable and accessible to our stakeholders, they were structured according to the FAIR principles in order to derive concrete recommendations (see also Figure 2.2). This step, in fact, represents the phase where the most relevant policies identified among the different stakeholders are mapped according to the structure of the FAIR principles, in order to make them accessible universally and not only to a restricted group of users.

2.4.1. First step: four high-level categories

During the information collection phase, we collected the policies for the different stakeholders for four fields of analysis, namely policies about:

1) Standards
2) Data completeness
3) Enrichment
4) Access

These four categories include high-level categories of policies, which are shared by every stakeholder. Task 3.2 decided to use these categories *as a general template in which to structure the recommendations for all our stakeholders, in order to improve (meta)data and repositories quality*.

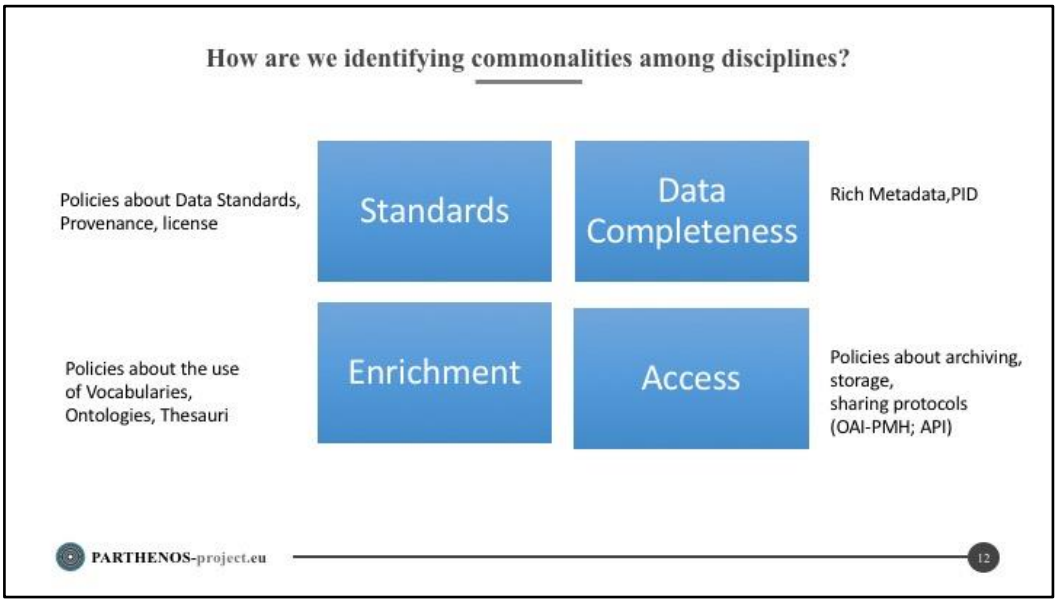


Figure 2.4: Identifying commonalities among disciplines.

In the first category “Standards” belong the policies for the use of data standards (which differs from the analysis of standards, which was undertaken by WP4). In the category “Data Completeness” belong those policies about the implementation of persistent identifiers and the creation of rich metadata. In the category “Enrichment” belong those policies about the use of vocabularies, ontologies and thesauri. Finally, in the category “Access” belong those policies about archiving, storage, sharing protocols.

2.4.2. Step two: mapping the high-level categories to the FAIR principles

While working on the four high-level categories mentioned above, it became clear that these categories corresponded well to the FAIR’s four letters/ principles (Findable, Accessible, Interoperable, and Reusable). This section shows how this mapping of the identified categories to the FAIR principles was performed and what is the rationale that we used in this process.

Data completeness = Findable

When collecting the policies in use by different communities, we categorised under “Data Completeness” the necessity for researchers to create metadata that is as rich

as possible, as well as supported by persistent identifiers. These two elements contribute greatly to make research data easily retrievable.

Similarly, the FAIR principles under “Findable” focus on the assignment of “eternally persistent identifiers” (F1) and on the creation of “rich metadata” (F2).[37]

Standards = Reusable

In our “Matrix” we have categorised under “Standards” those policies that refer to the use of data standards, information about data provenance as well as information about usage licence of research data.

Similarly, the FAIR principles under “Reusable” focus on similar topics, for example on the necessity to have “a clear and accessible data usage licence” (R1); that metadata are “associated with their provenance” information (R1.2); and that (meta) data meet domain-relevant community standards. (R1.3)

Enrichment = Interoperable

In Task 3.2 by enrichment we mean those instruments - such as vocabularies, thesauri and Authority Files that make research data not only retrievable, but interoperable with other data that might use different data formats or standards.

Similarly, the FAIR principles list under “Interoperable”, the use of “broadly applicable language for knowledge representation” (I1) and the use of “vocabularies that follow the FAIR principles” (I2).

Access = accessible

“Access” may include an almost infinite number of elements, but when collecting information on policies in use by the PARTHENOS stakeholders, we referred to a few number of elements, in particular to the presence of exchange protocols such as OAI-PMH or APIs, that would make research data more easily accessible.

Similarly, the FAIR principle use the term “accessibility” and list among others the following principles: (meta) data are retrievable using a standardized communication protocol (A1); such protocol is free and open (A1.1); and the protocol allows for authentication and authorization procedure, when necessary (A1.2).

[37] Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing version b1.0, available at https://www.force11.org/group/fairgroup/fairprinciples.

2.5. The PARTHENOS Guidelines and Best Practices to increase the quality of data, metadata and repositories.

The following guidelines provide a basic template for all the PARTHENOS stakeholders and can be applied by all the disciplines in the Humanities, as well as the closely related stakeholders: Cultural Heritage Institutions, data archives and Research Infrastructures.

The tables on the following pages list the elements reflecting the policies that are necessary to achieve high-quality (meta)data and digital repositories: as previously stated, they are organized according to the FAIR principles, as outlined by the FORCE11 guidelines.[38]

Findable

- **Persistent identifier**: each data and dataset should be identifiable by an eternal persistent identifier. This makes sure that a certain data object as well as an entire dataset is retrievable during time, when they are made available both via online and offline environments. Persistent identifiers can take different forms: handles, DOIs, PURL, URN
- **Rich metadata:** how rich and complete should metadata be? This is difficult to say, especially in the Humanities, where there "sufficient" and "not sufficient" can't be measured in details. In general, the Humanities agree on the principle that the more information and context connected to data - the better. - **Gaps in the data should be clearly stated**: historians, however, recommend that not only the context and richness of data should play a prominent role, but the gaps in data coverage as well. This makes clear what can be and what cannot be expected in a dataset or repository.
- **Discipline Specific Citation Guidelines:** each discipline in the Humanities has its own "best practice" for citing literature and other external data. Despite these different standards, each researcher in the Humanities should follow discipline specific citation standards.

Table 2.4: Elements reflecting the "Findable" principle.

Accessible

- **Exchange protocols:** in order to be fully accessible, research data should be fully accessible via (free) exchange protocols. In the last decades, and with the

[38] Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing version b1.0, available at https://www.force11.org/group/fairgroup/fairprinciples.

advent of digital archiving, repositories have implemented systems such as the OAI-PMH protocol, which refer to OAIS preservation scheme. In the last few years other protocols have become popular, such as APIs, which allow retrieval of data from other repositories without the need to set up any data repository.
- **Certification of repositories** (DSA, NESTOR, ISO): depositing research data in a certified repository means that the researcher can trust the preservation and dissemination policies adopted by such data archive, as they have been reviewed according to internationally agreed standards. - Similarly, for a digital repository to receive a certification (both formal - DSA- or formally attributed - ISO) means being attributed a recognition of trustworthiness and support of research.
- **Long-term preservation and archiving**: long preservation and archiving strategies are the ones that make sure that data are available for long time spans. However, the definition of how "long" the long term should be is quite difficult to quantify, as each discipline refers to different standards and definitions. Therefore, for an in-depth definition of this principle, we suggest to consult the policies and best practices for each specific discipline. - **Naming file convention**: following a precise and detailed naming convention allows researchers to retrieve and access their digital objects and data more easily; digital archives/ repositories have usually best practices in place to create and apply specific naming file conventions. We suggest to refer to the policies/ best practices for each discipline to find the most suitable naming convention for your research/ archive. - **Maintain the integrity and quality of data**: this is a general principle that emerged, in particular, from the interviews with historians. It refers to the necessity to maintain the richness and the context of the data created and collected during time.

Table 2.5: Elements reflecting the "Accessible" principle.

Interoperable

Controlled Vocabularies, Thesauri, Ontologies: these three reference objects all have different meanings and are used in different contexts. - A **controlled vocabulary** is a list of terms that have been enumerated explicitly. A taxonomy is a collection of controlled vocabulary terms organized in a hierarchical structure. - A **thesaurus** is a networked collection of controlled vocabularies: it uses associative relations in addition to parent-child relationship. - An **ontology** is a formal naming and definition of the types, properties, and interrelationships of the entities that really or fundamentally exist for a particular domain of discourse Despite the difference in use and the different meaning in the field of information science, these three "tools" all share the high-level principle of being able to enrich research data by linking them to information classification systems. This makes them "interoperable", so that they can be connected or referenced to other data from different knowledge systems, and unambiguous, so that their meaning

is explicitly referenced and disambiguated. Each discipline refers to different knowledge systems, therefore there are discipline-specific ontologies and controlled vocabularies, which we suggest are consulted separately.

Table 2.6: Elements reflecting the “Interoperable” principle.

Reusable

Data usage licence, legal information: information about the legal status of research data as well as about the possibility (or not) for reuse, is now considered as an essential part of data itself. For detailed information about IPR and legal information, see Chapter 4.
Use of standards shared by the community of practice: data and metadata standards can be considered the pillars of the process of data creation, and data preservation. During the last decades, many Humanities disciplines have created different data and metadata standards, suitable for their own communities. Not all disciplines we have covered have, however, a data standard to refer to, but mainly best practices that they share at the community level as guidance during the data creation process.
Provenance: data provenance documents the inputs, entities, systems, and processes that influence data of interest, by providing a historical record of the data and its origins. Provenance provides both to research and to cultural heritage data that information that is necessary to build a strong context and background of the data produced and disseminated.
Sensitive data: how to deal with sensitive data? This question was raised many times during the interviews with researchers in the Social Sciences area, but it can be shared also by other disciplines, such as Language Studies, for example. The policies we collected for the Social Sciences and Language Studies provide a useful guidance for researchers. On the other hand, the management of sensitive data is also a very relevant topic for data archives: information on the treatment of sensitive data for data archives are also included in the table in Appendix II: Matrix ‘Roles, Tasks, Quality’.
Data Management Plans: Data Management is a series of actions that aims to preserve and archive research data in the most effective way. Data Management Plans (DMP) have also become increasingly required from single researchers by funding bodies that want to ensure that the research outputs created by single researchers and institutes are not lost after a short time (see Section 3.3.1).

Creating documentation of the data creation, management and reuse: documentation of data creation, management and archiving is as essential as the creation of data itself. Every PARTHENOS stakeholder should cover this step in their activities in order to make their data available to other colleagues and stakeholders. Policies on documentation don't exist per se, but they are included in other policies developed for each discipline. We suggest to the readers check the policies available in the table "matrix" when mentioning "project documentation available".

Table 2.7: Elements reflecting the "Reusable" principle.

2.6. The PARTHENOS Wizard

Within PARTHENOS, intensive collaboration across work packages has taken place in order to create a PARTHENOS Interactive Guide in the form of a wizard. The main idea is to enrich the WP3 final deliverable with a tool that the PARTHENOS stakeholders can actively use for guidance when choosing and applying policies and best practices to their own research and activities.

From a content perspective, PARTHENOS WP3 identifies the relevant policies for each discipline addressed (see Section 2.3.3), and Deliverable 3.1 feeds the interactive guide with the policies identified by the three tasks of WP3. This information is structured in a matrix, as described above.

From a technical point of view, the information about policies and recommendations can be retrieved from the matrix and will be displayed by the PARTHENOS infrastructure through API streams. This ensures a sustainable solution as it makes the architecture very flexible and reusable for the dissemination of the information. The wizard is linked to the Data Model of PARTHENOS by a mapping tool called X3ML. All entities mentioned in the matrix, for example the entities *Data Management Plan, Protocol and Licence Agreement*, are compliant with the PARTHENOS entities as well as the CIDOC CRM Model. By mapping to the PARTHENOS entities which are compliant with the CIDOC CRM Data Model, the data produced will be findable in the Joint Resource Registry of PARTHENOS where the wizard is registered as a service. The focus of the PARTHENOS infrastructure is on retrieval and the wizard helps to retrieve protocols and best practices in use by each analysed discipline. The focus of the wizard is on giving advice on which policies to apply to research data. The matrix is stored in D4Science together with all

the mappings. This way the wizard will be created as an application for the PARTHENOS infrastructure and can be presented as a dissemination tool for the PARTHENOS community.

The wizard will be made available from the PARTHENOS website. The wizard web prototype was designed using a RESTful API approach and developed as a HTML5 widget application, which can be easily integrated and made accessible through different websites (e.g. CLARIN) in the future. Also, the information of the matrix could be published as Linked Open Data, for instance, by linking the spatial information to geo-names and transforming this into RDF. Not only information about common policies and high-level PARTHENOS guidelines will be disseminated, but they could also be integrated with the recommendations about standards or training modules used by PARTHENOS stakeholders.

In this phase of the project, a mock-up of the wizard was created to show the potential use of the tool: to guide people from the research community, working at digital repositories, Research Infrastructures, or Cultural Heritage Institutions, through the web of information towards relevant policies. Since the matrix with all the policies is stored and preserved in the PARTHENOS infrastructure, it can be modified if some policies change over time. During the ingestion process, the wizard application will download the most recent version of the matrix from the storage and extract all the updated information.

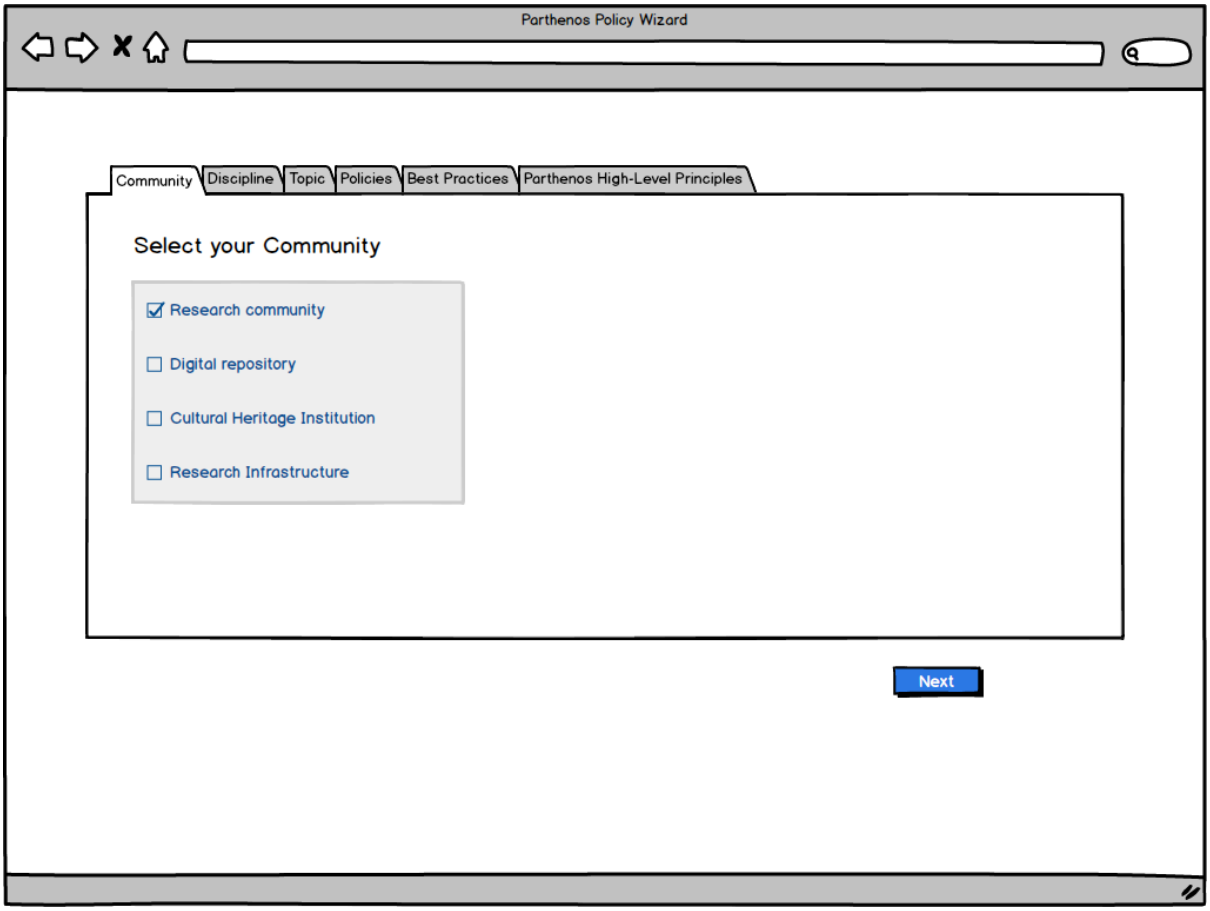


Figure 2.5: Mock-up of the PARTHENOS Policy Wizard: select community.

The innovative approach of this work is that it presents the commonalities between the different disciplines present in the PARTHENOS community and it offers common solutions to the users. In addition, gaps were identified when there was no policy available for a certain discipline and a comparable policy from another discipline is suggested as an example. A future goal is that by using linked data it would be possible to identify relevant parts *within* a certain guideline or protocol and give a direct link to this particular information.

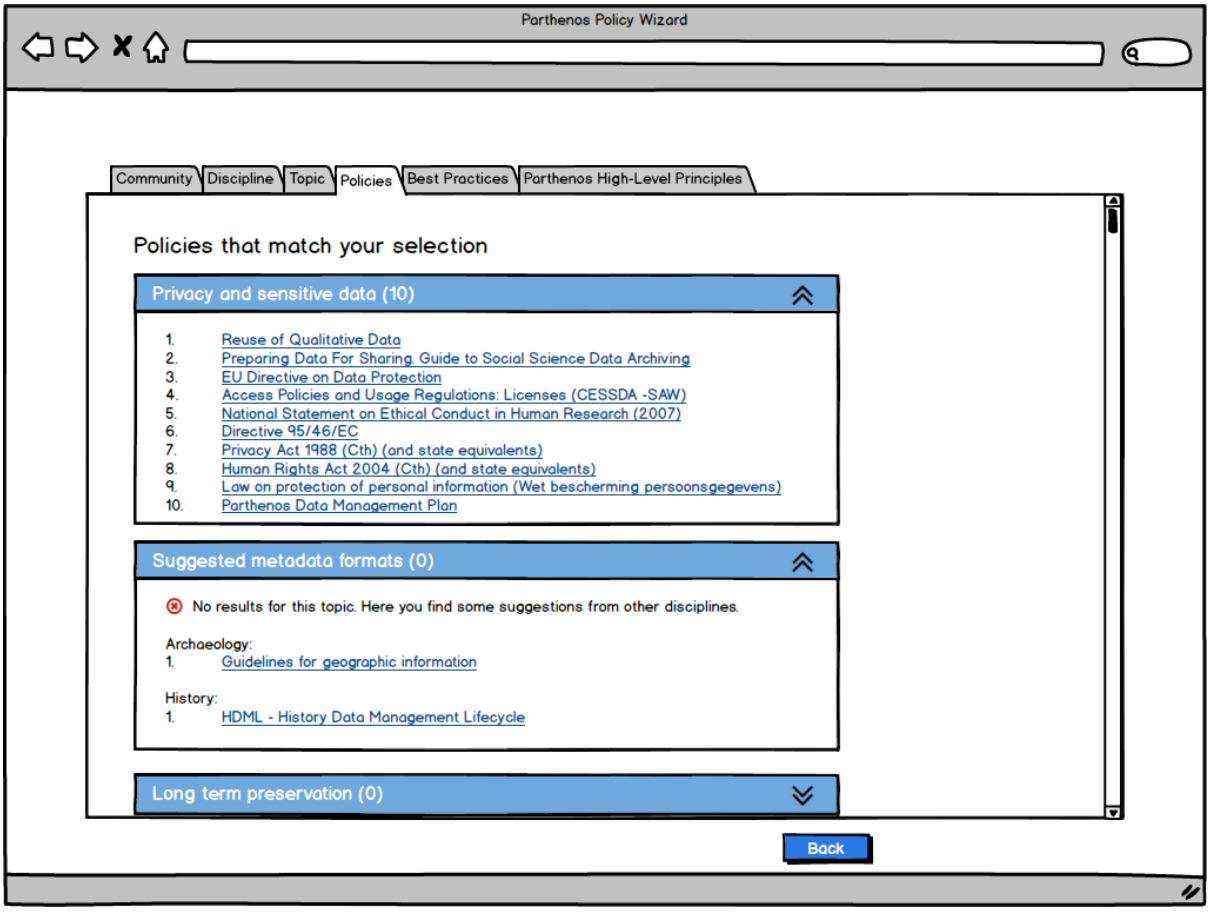


Figure 2.6: Mock-up of the PARTHENOS Policy Wizard: policies matching selection.

3. Data policy implementation

3.1. Current situation with regards to data management

PARTHENOS assembles partners from various projects, initiatives and infrastructures in the area of the Humanities and Social Sciences working with or relying on data management. This entails data centres providing services for data management, including assisting in data management activities, and individuals utilizing these services in their everyday practices. Some infrastructures are linked to archives and libraries. Others are rooted in communities of researchers joining forces to develop the services they lack in their daily work, especially with regards to sustainable data processing, archiving, and dissemination. Based on this heterogeneous field of participants, it seemed to be an essential task for the definition of guidelines and principles for data management to explore the current status among the various partners.

The method of choice for assessing the current situation within PARTHENOS partner organisations was to create a questionnaire asking about all kinds of aspects of data management. The questionnaire consisted of two sections: a general section and a more specific part. The general section was intended for data warehousing and organizing the answers later and to see possible differences based on the point of view. Therefore, information about the person filling in the questionnaire, e.g. their role, discipline and partner affiliation was collected. The data management specific part of the questionnaire was constructed as a matrix of two dimensions: one dimension followed the steps in the research data life cycle, the other the FAIR principles.

After an analysis of the various data lifecycle models, together with WP3 members, WP2 had decided to adopt the UKDA Research Data Lifecycle model[39] as the backbone for aligning the user requirements collected within the PARTHENOS community (see PARTHENOS D2.1 “Report on User Requirements”[40]). This decision was taken because of the completeness and clarity of the various steps offered by the model, which helped to identify a shared framework for the quality assessment of

[39] UK Data Archive. 2016. ‘Research Data Lifecycle’, available at http://www.data-archive.ac.uk/create-manage/lifecycle.
[40] PARTHENOS: Report on User Requirements (D2.1). 20 October 2016 (final version).

data and metadata, to identify common requirements and, finally, to produce guidelines defining common good practice for the research areas engaged in the project. As planned, these results were used for the assessment and the harmonization of the existing policies in use by the different disciplines. The questionnaire, therefore, was organized and labelled accordingly to the steps of the chosen Data Lifecycle model: Data Creation, Processing Data, Data Analysis, Data Preservation, Giving Access, Reusing Data.

A second dimension used in the questionnaire represented the FAIR-principle: Findability, Accessibility, Interoperability, Reusability. An extra column "other" was provided in case the answer was not deemed suitable for being classified according to the FAIR principles. The informants were asked to classify their answers according to these principles where possible. If the distinction was unclear, they were allowed to duplicate their answers. Empty fields and empty answers were also possible. An option to comment was provided for each question. Though the answers were not provided anonymously, it was indicated that for this first, general assessment in a cross section study the answers should not be related to the individual institutions, but summarized and generalized.

For each step of the data life cycle existing data management material was being used to develop the questions. This material is based on long-time experience to improve data management at and around an infrastructure to handle data. Among the underlying material were the Data Seal of Approval criteria.[41] Originally developed by DANS (2008) and handed over to an international board (2009), "the objectives of the Data Seal of Approval are to safeguard data, to ensure high quality and to guide reliable management of data for the future without requiring the implementation of new standards, regulations or high costs" now a seal of quality for data repositories. Other material like the data management plan template provided by CLARIN Germany[42] was developed to organize "the management of research data that is produced and analysed in the course of research projects" in the area of language related research data, which itself uses ideas from the DMP[43] online, which is a tool supporting scholars structuring and organizing their data.

Other material focuses more on supporting people on the edge of creating

[41] http://www.datasealofapproval.org.
[42] http://clarin-d.net/en/preparation/data-management-plan.
[43] https://dmponline.dcc.ac.uk/.

data which can be seen in the guidelines from the Radboud University for Data Management for Students[44] addressing especially students. Institutions like the Australian Data Service tried to adapt the CMM (Capability Maturity Model) to manage research data in a better way.[45]

Many papers focus on data management practice like the paper by Kevin Crowston, Jian Qin (Syracuse University) on a CMM for scientific data management (SDM) practices, with the goal of supporting assessment and improvement of these practices [46] or the ASIST paper (2016) "Workshop Building Capabilities for Sustainable Research Data Management Practices".[47] Other papers focus on more theory based approaches like the DCC benchmarking tool for data management strategy development in research environments[48] and UKOLN University of Bath, 2011, Community Capability Model for Data Intensive Research.[49]

Based on this previous work, a group of partners in PARTHENOS created a total of 47 questions to be answered, clustered in 6 sections, one per state in the data life cycle. Each question received a brief summary to explain its intended content. In total, we received 16 different completed questionnaires, reflecting 10 partners of PARTHENOS. Some partners here provided multiple questionnaires as they represent different roles and different parts of their respective organizations. The answers were provided by CLARIN (various institutions and partners), CNR (various institutions and partners), CNRS, INRIA, KCL, KNAW (DANS and NIOD), MIBACT-ICCU, OEAW (various institutions and partners), PIN (various institutions and partners), and SISMEL. The roles of the informants ranged from researchers providing data on data centres and archives, covering all major participants in the data management process. It turned out that the aspect of long-term archiving was underrepresented in the questionnaire, obviously as a result from the underlying material used to develop the questionnaire. The complete questionnaire, including the instructions for filling it in, is provided in Appendix III: Questionnaire.

---

44 http://libguides.ru.nl/datamanagement.
45 http://www.ands.org.au/guides/capability-maturity.
46 https://www.researchgate.net/publication/228401999_A_Capability_Maturity_Model_for_Scientific_Data_Management_Evidence_from_the_Literature.
47 https://www.asist.org/files/meetings/am16/Building_Capabilities.pdf.
48 http://cardio.dcc.ac.uk/.
49 https://communitymodel.sharepoint.com/Documents/CCMDIRWhitePaper-24042012.pdf.

3.2. Guidelines defining good practices

This section provides guidelines for good practices with regards to data management as developed by PARTHENOS partners with a focus on the stakeholders' Research Infrastructures and repositories. Whereas the analysis of the current state of the art was organised according to the research life cycle, the following sections of recommendations and guidelines will be structured around the FAIR principles (see Section 1.3). This is due to the momentum that FAIR principles have created during the first period of the project, especially as an essential part of the data management requirements for Horizon2020 projects. Additionally, the FAIR principles are more likely to produce general guidelines for data management.

Whereas FAIR relates to data as such, in the following, we provide guidelines for good practices that will help Research Infrastructures and repositories to supply their services in accordance with the FAIR principles. The guidance is based on existing practices, as established above, but will also address gaps, as well as needs anticipated by the much more data-intensive research practices that FAIR principles are trying to accommodate.

3.2.1. Findable

For the FAIR principles for data reuse, findability is the key for effective implementation. Though the least obvious when reusing the data, the proper way of locating data is a necessary condition for any other step. There are various aspects relating to findability which should be explored here together with recommendations for best practice.

3.2.1.1. Identification of data resources

To find and access a data resource, it is essential to identify the object of interest. Though it seems obvious to identify electronic resources in terms of file names, URI locations such as http and ftp, stock numbers, cryptographic (md5) checksums, etcetera, most of these systems have inherent problems when it comes to

identification: file names are not unique, nor unchangeable and persistent, resolvable URIs can change when servers move, stock numbers refer to specific locations and installations and may not easy to interpret for third parties, checksums are unreadable and can be considered ‘not-writable’ by humans.

Another problem is in the area of copies, i.e. if a copy receives the same identifier, something like an ISBN being constant for each book of the same edition, or if each location receives a different identifier, such as a number in a book collection implying a location and hence allowing direct access to an object based on the identifier. This is essential in the digital world, as copies of files can be identical in terms of size, content, etcetera, but can still be distinguished by their location. To ensure the integrity of a file and identify if a file has been modified, it is essential to compare the copy with the original, either by means of a fingerprint such as a checksum, or by a direct comparison of two files.

Last, but not least, is the problem of granularity, i.e. which set of particles is seen as one object requiring identification. Research data evolved from measuring sensors often comes in multiple files, textual resources with various annotation layers can also come in separate blocks, audio-visual data often consists of signal files together with transcriptions, notes and background information, sometimes in multiple files due to discontinued recording sessions and scene cuts.

Based on identifiers it must be possible to cite data resources persistently (i.e. even if the location changes the identifier stays persistent) and locate an authoritative copy (i.e. the authoritative copy is not altered, and the identifier can always be used for finding the current location of the copy). Identifiers do not need to authorize access to a resource, contain information on the content of an object or provide any other form of semantics besides identification.

3.2.1.2. Recommendations on identification

1) Each resource must be assigned a permanent and unique identifier which can be used for determining the location of the representation of the original authoritative copy. A suitable standard is ISO 24619:2011 (“Language resource management -- Persistent identification and sustainable access (PISA)”) from the area of language resources. The choice of a persistent identifier schema must rely on careful assessment of advantages and

disadvantages. Suitable example implementations for these are: the handle system[50], Digital Object Identifiers (DOI, also being a handle system), URNs.

2) The institution responsible for future access of the resources maintains digital preservation of the received authoritative copy of the data, including information of the identifier assignment.
3) For granularity, there is no sound recommendation, but we follow the recommendations from ISO 24619
4) The level of granularity of existing identifier schemes for a type of resources should be retained, for example for books there are ISBNs, so this level would be retained.
5) An identifier should be assigned if the resource is associated with the complete content of a digital file.
6) An identifier should be assigned if a resource is autonomous and exists outside a larger context, such as a collection of poems by one author being used independently of the collection of all works by the same author, hence the collection of poems is assigned a separate identifier despite the fact that it is also part of the larger unit.
7) An identifier should be assigned if a resource is intended to be citable apart from any larger unit. The intention is left vague and can be seen as part of the required negotiations between the depositor and the archive.

Regarding granularity, there should be guidelines for how to refer to smaller parts of a resource, e.g. individual files if the resource is composed of multiple files or the content of a file such as individual paragraphs or other structures marked up in an XML file. ISO 24619 suggests part identifiers for smaller units that are part of larger units. Such assets must be assigned persistent/permanent identifiers following a Persistent Identifier Scheme which enables future access of the asset.

3.2.1.3. Findability by properties of a data resource

Identifiers only allow different objects to be distinguished from each other and are a condition for findability in a digital world. However, they do not ensure findability as in

[50] http://www.handle.net/.

finding a suitable resource for access and reuse. For this purpose, the object's properties need to be taken into account. An object's properties can be resource internal, i.e. properties that are work inherent, or external, i.e. descriptions created outside of the object.

Digital content of objects is not necessarily sufficient for finding a resource. For example, three dimensional scans of artefacts are consisting of numeric representations of spatial vectors, often stored in proprietary and binary formats. These can hardly be searched for by persons. But also, textual resources are problematic, as a search for the textual content can only yield a full text search, like a concordance, rather than allow for a search for properties.

For finding data resources, it is necessary to have structured and meaningful descriptions of resources, including descriptive and administrative metadata (see also Section 2.1.2). This data can be indexed by general search engines, specialized search engines or cataloguing applications. Cataloguing applications often have a distinct set of metadata required in the archiving process. These catalogues are often very specific to an institution and the research data they archive and maintain, often targeted either in the direction of print, as in libraries, or artefacts, as in museums. Some metadata schemas can be translated into others, but in general this conversion is neither lossless nor yielding perfect results in the target formats. Nevertheless, the conversion can provide insights and allow for interoperability of resources. In general, it seems to be the case that the more complete the provided metadata, the higher the quality even after conversion.

In the domain of research data, there are very different types of resources, depending on the field of research and the domain of the scholar, ranging from texts with structural, grammatical, typesetting information, to artefacts and manuscripts with detailed descriptions of qualities, textures, and material, and data from questionnaires, sensors, including signal recordings. Each of these requires individual classes of metadata to provide a meaningful description of the research data. A unification of all possible structured metadata sets would be extremely rich and most data fields would remain empty. At the same time, some descriptive categories used for one type of resource may be inappropriate, useless or misleading for another. Hence, it is required that the metadata schema is suitable for the description of the type of data. Libraries and archives distribute their metadata with the help of the Open Archive's Initiative Protocol for Metadata Harvesting (OAI-

PMH). Metadata provided in such a way can be used by domain specific or research specific search engines, for example for faceted search applications utilizing the structure of the metadata schema. These search engines can also work with a variety of metadata schemas, depending on their implementation.

General search applications, such as Google, do not necessary interpret the structures of a metadata schema. These search engines basically require an HTML version of the metadata for indexing and searching, distributed by standard web server technology. Microformats in HTML can be utilized for conveying structural and semantic information going beyond HTML. For linked data, RDF is the most commonly used format. Though RDF is highly adjustable and metadata schemas can be described in RDFS, using RDF as a primary descriptive format is problematic. All recent metadata schemas can be converted into RDF, hence the metadata can be provided as data formats suitable for linked data using SPARQL endpoints. For metadata to be linked, common elements are required, such as identifiers for persons, institutions, and locations. Such linkable elements can be taken from authority files, often provided by national libraries.

Recommendations to support findability of resources

1) Select an appropriate metadata schema for the type of resource being described, fitting to the type of resource. Metadata can cover various aspects, such as citation metadata, disciplinary metadata, preservation information, provenance, etcetera. The metadata intended for findability is the type of metadata used for citation and describing data in a catalogue. This should be the primary format for maintaining the descriptive metadata. Utilize existing metadata schemas, such as schemas according to ISO 24622-1 (Component Metadata Infrastructure, adjustable to each type of resource), or MARC21 (if appropriate for the type of data). Dublin Core alone is not suitable for a detailed description of research data, nor is Datacite MDS.
2) Make requirements, about use of persistent identifiers for referencing and association with the referred contents, part of the metadata.
3) The metadata provided should be high quality, i.e. as correct and complete as possible, including enough information for later access and comprehensibility.

4) Select an appropriate persistent identification schema and assign a PID to every resource.
5) Ensure semantic interoperability by referencing authority files, for example ISNI, VIAF, ORCID.
6) Make descriptive metadata publicly accessible using standardized protocols, such as OAI-PMH, or SPARQL. Information that needs to be protected, for example for privacy reasons, should not be part of the publicly accessible metadata but should be recorded as part of the documentation of the resource in restricted contexts.
7) Publicize the protocol endpoint to suitable search providers, for example CLARIN maintains a registry for endpoints providing language related research data.
8) Provide different formats, this can for example include HTML to allow findability with standard internet search engines, Datacite MDS and Dublin Core for interoperability purposes with archives metadata, etcetera.

3.2.2. Accessible

Accessibility is addressing the topic of providing digital resources to a wider audience then the data providers, using metadata for dissemination and systems for granting access. These aspects have different implications and conditions, which will be discussed in the following sections.

3.2.2.1. Implications of accessibility for data providers

The issue for institutions: respecting laws with an incentive for making the data accessible

The research institutions that hold and give access to data are public institutions or public funded projects that are responsible for respecting national and international laws. The Directive on the Reuse of Public Sector Information is an incentive for open data policy and it generally encourages public sector institutions to make as much information available for reuse as possible and to foster the production and publication of interoperable open data sets, open standards, data formats, ontologies

and vocabularies.

Accessible data: definition/presentation

The FAIR guiding principles provide quite precise requirements for data to be considered as “Accessible”. It implies that (meta)data are retrievable by their identifier using a standardized communications protocol, which is open, free, and universally implementable, allowing for an authentication and authorization procedure, if required. The metadata should be persistent even if the data are no longer accessible[51].

Data access as part of data preservation

Data preservation is a phase of the data life cycle, which includes all activities needed to ensure continued access to digital resources and the information they enclose, by humans and machines.

The definition of a preservation plan is part of the strategic planning an institution/RI should develop to ensure the long-term preservation of the managed digital resources. Section 3.3.2 describes in more detail what digital preservation is about and what must be covered in preservation planning affecting policies, organisation and technology to carry out the preservation plans.

Necessity to go further than putting data online to make them accessible

Putting data 'on the web' is not enough. To be actually interoperable and reusable, Data Objects should not only be properly licensed, but the methods to access and/or download them should also be well described and preferably fully automated and using well established protocols.[52]

3.2.2.2. Retrieving metadata with an open, free and universally implementable communications protocol

Context: resources are findable

Except for three data repositories which systematically made the archived data freely

[51] See the specification of the FAIR principles at https://www.force11.org/fairprinciples: Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing version b1.0.
[52] https://www.force11.org/node/6062/#Annex6-9.

available, data repositories have many different policies with regards to access restrictions. Even if they generally recommend free or public access, repositories enable data providers to limit access to the archived data if necessary.

Most of the time, metadata is publicly available. Repositories can even choose to make findable metadata for non-public resources, through a Triple Store or an OAI-PMH server. Data repositories are OAI-PMH compliant for distributing metadata.

Resources can be retrieved, if allowed by the end-user licence with respect to the user in question (login required for non-public resources). To each resource/data object is assigned an ID that could be a PID, OAI Identifier and URI. Each resource/data object is associated with metadata such as title and the abstract description (e.g. from the Dublin Core metadata), to give users a clear overview of what the resource is about. Resources can be downloaded.

Assigning a Persistent Identifier and the preservation of the relation between this identifier and the contents it identifies to ensure findability over the long-term is also a strategy for sustainability.

Discovering online resources

Discovery services:

- Data are discoverable via the website of the data repositories, and sometimes via digital platforms like Europeana. ISIDORE provides both a REST API[53], a RDF 3Store[54] and a Web interface for metadata discovery.
- These portals enable users to search within the metadata fields associated with the data. They can offer (often a combination of) full text search[55], advanced search[56], faceted search[57], or a search by collections.[58]

Resources can be retrieved online through the institutions' portal. If users meet with the access conditions, they can have access to the resources. EASY (DANS) allows authorized users to directly open files supported by the browser (images, PDFs)

53 http://api.rechercheisidore.org/.
54 http://rechercheisidore.org/sparql.
55 http://www.mirabileweb.it/ricerca_globale.aspx.
56 https://easy.dans.knaw.nl/ui/advancedsearch.
57 http://portal.ariadne-infrastructure.eu/.
58 http://archaeologydataservice.ac.uk/archives/.

and/or to download selected datasets from the landing page.

Searching/browsing does not require users to log-in and all qualified Dublin Core metadata is publicly available. Some services such as ISIDORE, CulturaItalia provides an API and/or a Triple Store for metadata discovery.

To support data findability, it could be convenient to identify high-level facets for browsing the gathered information, such as the ARIADNE portal where it is possible to discover meta(data) selecting one or more of the following facets:

- Resource type: every resource in the portal is categorized with a resource type. The type can be any of the following options: Fieldwork archives, Event/intervention resources, Sites and monument databases or inventories, Scientific datasets, Artefact databases or image collections, or Burial databases.
- Native subject: subjects from a vocabulary used by the original owner of the resource.
- Derived subject: subjects derived from mapping native subjects to Getty AAT vocabulary terms.
- Keyword: keywords or tags describing the resource.
- Contributor: the agent responsible for describing the resource in the Catalogue.
- Publisher: the agent responsible for making the resource accessible.
- Place: place names the resource is connected with.
- Period: time periods the resource is associated with.
- Rights: access rights connected to the resource.
- Language: language of the resource.

Searching resources

The Federated Content Search is used by one infrastructure to retrieve publicly accessible data, by using the FCS API which is based on SRU/CQL. Search engine based on LUCENE and SOLR has been developed by one institution. In the other cases, very little information has been provided. However, we can mention the use of the FEDORA search interface.

OAI-PMH harvesting

OAI-PMH harvesting is also performed to make resources from repositories available via other portals/interfaces, such as the CARARE portal in Europeana. For instance, NAKALA[59] makes metadata accessible through OAI-PMH and by a Triple Store. One institution provides a SOAP web service and a number of warehouse management systems (WMS) which allows metadata to be incorporated within the Heritage Gateway.

The OAI-PMH standard is generally adopted as repository and discovery service, and metadata are openly available, that means that they are freely available to use, reuse and redistribute and the only restriction could be attribution and share alike. For discovering and finding meta(data) advanced search and faceted browse services which target the qualified Dublin Core metadata of the datasets are adopted. Metadata for non-public resources are made also available, for giving high-level information on protected data and content. OAI-PMH Harvesting is also adopted in order to make resources from the repository available via other portals/interfaces, such as the CARARE portal in Europeana.

3.2.2.3. Authentication and authorization procedure

The FORCE11 community strongly recommends publishing data in complete Open Access, whenever possible. Some exceptions to Open Access can be made, but they have to be carefully justified.[60] Institutions within PARTHENOS clearly support Open Access, but they need to be able to set up limitations for accessing data when necessary:

- National laws and regulations enforce protection of personal data and databases, sensitive information, intellectual property rights, and copyrights rights.
- When legitimate interests of the rights holders are at stake, data providers should be able to restrict access to data by defining an initial period of preferential use, due to confidential or contractual protection reasons (e.g. for data with commercial potential).

59 https://www.nakala.fr/.
60 https://www.force11.org/node/6062/#Annex1.

In order to give complete Open Access to data or to set up exceptions, PARTHENOS members should respect the following recommendations:

Licensing data

Most academics appear to believe that non-licensed data is fully open. Actually, non-licensed data is difficult to use, because its future users can't assess its usability. Therefore, licensing data is key to FAIR data publishing.[61]

PARTHENOS members should systematically license the data they wish to be made accessible, and precisely describe their conditions of use (academic and/or private/commercial). Such licences should also be cited with PIDs.[62]

The use of licences that are as open as possible is recommended. National and international licensing frameworks offer interoperable and open licences, adapted to all sorts of data. Preferably these licences should be machine-processable, as especially large repositories and archives can otherwise not maintain licence restrictions:

- Creative Commons;[63]
- Open Data Commons;[64]
- Europeana Licensing Framework;[65]
- Licence ouverte/ Open licence.[66]

Depositing data in a repository

To control access to data, it is necessary to store it in a data repository. Some institutions don't have locally developed systems, but they can find data repositories suitable for their data:

- re3data or "Registry of Research Data Repositories"[67] is a catalogue of data repositories.

---

61 Ibid.
62 Ibid.
63 https://creativecommons.org/share-your-work/licensing-types-examples/.
64 http://opendatacommons.org/licenses/.
65 http://pro.europeana.eu/files/Europeana_Professional/Publications/Europeana%20Licensing%20Framework.pdf.
66 https://www.etalab.gouv.fr/licence-ouverte-open-licence.
67 http://www.re3data.org/.

- The list “Data Repositories”[68] is a part of the Open Access Directory project. It provides a list of open data repositories.

Many solutions are available. Some important criteria have to be taken into account before choosing a data repository:

Features

- How do you evaluate the sustainability of the repository?
- Can you deposit data easily?
- Can you find data easily?
- How do you evaluate the accuracy of data description? (Precision and number of metadata fields)

Functionalities

- Is data preserved under a trustworthy preservation program?
- Is there a digital preservation strategy that fulfils the requirements to:
  - Bit level preservation
  - Logical preservation
  - Treatment of confidentiality issues
  - Preservation costs
- Is a PID systematically assigned to data and ensured to be linked to the data?
- Can you determine which version of your data is accessible?
- Is data provenance clear and precise?
- Does the repository provide usage statistics?
- Can access to data be controlled?
- Is the repository interoperable?

Requirements

- Is it a disciplinary or a multidisciplinary repository?
- Costs?
- Type of data accepted?

[68] http://oad.simmons.edu/oadwiki/Data_repositories.

- Accepted formats?
- Which licences are proposed?
- What is the current limit in terms of data volume?

The following are sample data repositories that some projects within the PARTHENOS context might use:

Name	URL	Type	Description
Zenodo	https://zenodo.org/	Public institution	Zenodo has been created by CERN (European Organization for Nuclear Research) and OpenAire. Zenodo collects all sorts of datasets and provides a DOI.
Dryad	http://datadryad.org/	Non-profit organization	Dryad is a multidisciplinary repository, but it especially collects medical datasets.
Datahub	https://datahub.io/fr/	Non-profit organization	Datahub has been created by the Open Knowledge Foundation. It collects Humanities and social science data.
Figshare	https://figshare.com/	For-profit organization	Researchers can deposit their data for free. But Figshare offers a commercial solution to institutions for managing their data.

Table 3.1: Sample of data repositories.

Defining categories of users

For resources with restricted access, institutions need different categories of users, for example:

- internal administration;
- public use;
- academic use;
- individual/private use;
- commercial use.

Metadata about access rights are generally generated from the chosen depositor licence.

Using authentication, authorization, and identification (AAI) procedures

For accessing access-restricted data, an authentication, authorization, and identification (AAI) infrastructure needs to be in place starting with local password protection, but ranging to single sign-on solutions. To manage access rights, it is recommended to (1) generate log files, and (2) use access control lists associated with Shibboleth authorization.

3.2.2.4. Long-term accessibility of metadata

Metadata are essential for the reuse of data and the reconstruction of results: they enable future users to understand the deposited research data. Therefore, it is essential that metadata accompany each digital dataset. Besides metadata should be sustainable, which means it is preserved and accessible, even when data itself is no longer available.

Defining responsibilities for metadata maintenance

Defining a workflow (for instance in a data management plan) is essential to maintain metadata. Institutions need to develop a strong supporting organisational structure, including metadata managers.

A minimum set of metadata should be required by the data repositories. Metadata is usually described by the data providers themselves. It also seems necessary to actively associate the data providers with the data stewardship: this cooperation enables data repositories and infrastructures to ensure continuous access to data over a longer time.

Licensing metadata

Open and completely public metadata is recommended. Therefore, institutions should apply a well-defined licence to metadata.[69]

[69] https://www.force11.org/node/6062/#Annex10-11.

Adopting standards

Metadata should be provided in a machine-readable format, which means "that there is an open standard for the format against which reliable parsing code can be written.[70] Therefore, metadata should refer to a standard.

The Component Metadata Infrastructure (CMDI, ISO 24622-1), which was developed with CLARIN involved, offers "a framework to describe and reuse metadata blueprints[71]". It enables to create an environment supporting different metadata schema. The most commonly adopted standards are the following:

- MAG and METS-MDI schemas: Dublin Core[72], VRA[73], NISO[74], MD5, METS;[75]
- ACDM;[76]
- CIDOC CRM;[77]
- PREMIS[78] for preservation metadata, technical metadata standards like MIX[79] for still images etcetera.

Referring to shared controlled vocabularies or ontologies

Metadata should systematically refer to shared controlled vocabularies or ontologies. It enables the mapping of metadata fields between heterogeneous resources.[80]

Assigning persistent identifiers

In order to ensure the long-term accessibility of metadata, PIDs should be systematically assigned to the deposited data (for instance by using the Handle System). PIDs could adopt multiple forms:

- OAI Identifier;
- URI;

---

70 Ibid.
71 https://www.clarin.eu/content/component-metadata.
72 http://dublincore.org/
73 https://www.loc.gov/standards/vracore/.
74 http://www.niso.org/standards/.
75 http://www.loc.gov/standards/mets/.
76 http://support.ariadne-infrastructure.eu/.
77 http://www.cidoc-crm.org/.
78 http://www.loc.gov/standards/premis/.
79 http://www.loc.gov/standards/mix/.
80 https://www.force11.org/node/6062/#Annex10-11.

- DOI;
- Permalinks;
- In some cases: library management system numbers (for instance, an Aleph System[81] number).

Access to metadata (for both humans and machines) should be also facilitated by standardized data citation.[82] It is recommended to use DataCite guidelines[83] as a standard.

3.2.3. Interoperable

The third FAIR principle states that data should be interoperable. There is a minimal set of definitions for the interoperability aspect:

To be Interoperable:

- I1. (meta)data uses a formal, accessible, shared, and broadly applicable language for knowledge representation
- I2. (meta)data uses vocabularies that follow FAIR principles.
- I3. (meta)data includes qualified references to other (meta)data.[84]

More precisely, Data Objects can be Interoperable only if:

- I3.1 (Meta) data is machine-actionable [...]
- I3.2 (Meta) data formats utilize shared vocabularies and/or ontologies [...]
- I3.3 (Meta) data within the Data Object should thus be both syntactically parseable and semantically machine-accessible [...][85]

The FAIR principles are meant as a “guide to FAIRness of data” and not as a specification[86] (see also Section1.3). Our approach is to extract best practices and recommendations for interoperability from the insights we get from the partners in

[81] http://library.harvard.edu/lts/systems/aleph.
[82] Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014, available at https://www.force11.org/group/joint-declaration-data-citation-principles-final.
[83] https://www.datacite.org/.
[84] https://www.force11.org/group/fairgroup/fairprinciples.
[85] https://www.force11.org/fairprinciples.
[86] Ibid.

PARTHENOS. We, therefore, can profit from the experiences of existing projects, especially from the perspective of already established workflows on interchanging data. This allows us, on the one hand, to analyse if and how these workflows fit to the FAIR principles. On the other hand, we abstract recommendations on how partners can press ahead with enabling more interoperability. As there are many partners in PARTHENOS we have a good sample for focusing on the potential and limits of data interchange. This is in line with one of the main aspects of the FAIR principles, to "enable a broad range of integrative and exploratory behaviours".[87]

Our first step was to identify technologies used by the partners which can be seen as enabling interoperability. This list is in no way exhaustive, as not every partner gave a clear answer about it. In addition, there are two opposed approaches in the data creation/ingestion phase that strongly influences the ability for interoperability: (1) an open approach, allowing any kind of format and data, and (2) a restricted approach, allowing only appropriate formats. The first one has a tendency to complicate interoperability, whereas the second one provides a clear framework, but has the risk of technical obsolescence or a lack of acceptance. It seems that this is the reason for many in-between approaches, which recommend or even push data publishers to use appropriate formats and at the same time allow ingestion of all other kinds of formats. The consequence is that a majority of data hosts comprise a mixture of data with different interoperability capabilities. The task is to identify the data that fits the interoperability principle and to motivate data providers in choosing formats and data structures with a high interoperability level.

As an example of best practices, DANS encourages depositors to provide their data in preferred or accepted formats according to the DANS Preferred Formats Guidelines.[88] For the formats that are not on the list, DANS archivists check if migration to preferred formats can be achieved (either by the depositor or by DANS). Where this is not possible, data is still stored, although less guarantees for the long-term can be given. In the case where the files are migrated by the archive, the original data will always be archived as well as the processed data, including a reference to the original data in the file metadata. As a general guideline, DANS considers preferred formats to be the ones that have open standards, are well

[87] Ibid.
[88] DANS Preferred formats. September 2015, Version 3.0, DANS, URL: https://dans.knaw.nl/en/deposit/information-about-depositing-data/DANSpreferredformatsUK.pdf.

supported and do not rely on the use of specific software or platforms.

One thing to mention here is that a list of accepted data formats does not tell us at first if one specific (meta)data format boosts interoperability. It is a hint that a repository supports such format (allow uploading it, doing analysis processes on it, and/or guaranteeing that it is properly stored). But more and more new formats are being developed, some of them claiming that they are the best solution for a specific domain or situation. It can be therefore be argued, that interoperability of (meta)data formats is not so much a technical issue, as it is more a community issue on how widely a format is accepted and how strong and active a community supports a format. Interoperability, in this sense, needs to find measurements that deal with technical claims and the concerns of communities, to find the best working solution in terms of (meta)data formats and the use of shared vocabularies and ontologies, but also to recommend changes in the practices of a community. By now, we have collected many use cases to gain an understanding of the different domains. The next iterative step will be to find measurements for the interoperability aspect (as well as for the other three FAIR principles) and to derive recommendations from there. This will be an important fundament for constituting "Data FAIRports"[89].

Our partners support many (meta)data formats. This list is quite big but still insufficient (and under permanent change as new formats derive). Instead, this is a good place to refer to PARTHENOS WP4 on Standardization who are working on a "Standardization Survival Kit", giving a broad overview on the different (meta)data formats in use. This includes descriptions of the different formats and domain-specific recommendations. An explanation of the expected outcome and a first overview on formats can be found in PARTHENOS Deliverable 4.1[90].

To get an overview of the different formats that are in use by the partners, some of them have referred to online sources where they documented the supported (meta)data formats. This is a selected list of these references:

[89] As explained in the Guiding Principles for FAIR data publishing, Version B1.0 (https://www.force11.org/node/6062/#Annex6-9), a "Data FAIRport" is a "repository of FAIR data" that implemented "a FAIR view on data" (ibid, § 2).

[90] The project deliverables of PARTHENOS can be found on the project's website: http://www.parthenos-project.eu/projects-deliverables/, direct link to Deliverable 4.1 entitled "Standardization Survial Kit": http://www.parthenos-project.eu/Download/Deliverables/D4.1_Standardization_Survival_Kit.pdf.

Reference description	Link
CLARIN – Standard recommendations	https://www.clarin.eu/content/standard-recommendations direct link to the document "Standards for LRT": https://www.clarin.eu/sites/default/files/Standards%20for%20LRT-v6.pdf
CLARIN Standard Guidance	http://clarin.ids-mannheim.de/standards/
DANS -File formats, preferred formats and accepted formats	https://dans.knaw.nl/en/deposit/information-about-depositing-data?set_language=en (overview) direct link to version 3.0of the document: https://dans.knaw.nl/en/deposit/information-about-depositing-data/DANSpreferredformatsUK.pdf
FACILE - Service de validation de formats	https://facile.cines.fr/

Table 3.2: Selected reference descriptions.

What the interoperable principle implies is a strong focus on (meta)data formats that are commonly used and backed by a strong community. Coming back to the definition of the stakeholders for this Deliverable, Research Infrastructures have the best capabilities to define such (meta)data formats, pool a community around them, and maintain them especially with a perspective on interoperability. There are already good examples for this, for instance ISO 24622-1 (CMDI, co-developed by CLARIN), or ARIADNE's ACDM.

Small projects sometimes don't have the insights into well supported (meta)data formats. Also, the perspective on interoperability is not a first level concern. It is, therefore, a good practice by funding institutions and data centres to insist on elaborating on these issues in a data management plan. Repository or data managers can assist in developing interoperability aspects for such projects. It would be also good to have a contact point where experiences on enabling interoperability between projects are shared and documented. This can help to avoid a mere project specific perspective in the data management plan and instead take interoperability issues more seriously. A first step would be to revise templates for data

management plans and implement questions that target interoperability.

General recommendations on interoperability

- Give an easy to find and detailed overview on accepted (meta)data formats. Ideally, in a single page that can be directly referenced and where the information on (meta)data formats is not hidden in an overwhelming document that covers all of the aspects of the repository. In general, a fine granulated and good structured documentation that uses modern aspects of design and user interface methodology can help to see on a glance possibilities for interoperability. It may be a good idea to structure such documentation along the FAIR principles.
- Document and also give easy access to the data model or models in use in a repository. Also, make clear which parts of the data model enable interoperability, and which parts are relevant when connecting datasets between projects.

On a technical level, the (automatic) transformation of data in the ingest phase of repositories can enable interoperability on the fly. That is an area where common developed scripts and tools should be developed through a joint effort and shared between repositories.

3.2.3.1. (Meta)data is machine-actionable

The FAIR principles focus on the ability of human and particularly of machines to automatically find and use the data through the provision of FAIR (meta)data, with the ultimate goal to support data reuse by the individuals. There is a strong focus on using standards for metadata, therefore, having in many cases metadata machine-actionable. This means that machines have to act automatically when confronted with the wide range of types, formats, access mechanisms, and protocols, by keeping record of provenance so that data collected can be reused and adequately cited. To make this happen all actors in the data management process, e.g. researchers, data producers, and data repository holders must comply with the FAIRness of data and provide information that will allow the system to identify the

type of object, determine its usefulness within the context of the metadata and/or data elements retrieval, and determine its usability, with respect to licence, rights, or other use constraints[91].

Many repositories that are part of the PARTHENOS consortium are already implementing various aspects of the FAIR principles using a variety of technology and methodology choices.

Interoperability of (meta)data can be increased by a high level of (meta)data quality. Machine readability especially relies on notably well-formed and predictive (meta)data content. Paying attention to quality from the beginning is the key to success. This implies having a strong focus on this issue in the data creation phase of the data life cycle. In small projects and departments, having dedicated staff responsible for data quality assurance and mediation between data creators and data hosts helps boost interoperability aspects. But there is often no funding for such personal and growing data volumes, complicating the work of data stewardship[92]. There are at least two interlocked approaches to make such a task more feasible. On the one hand, pushing data providers to deliver high quality metadata. Effective options, therefore, are a well-thought-out (meta)data input interface, validation of the input in a traceable way, comprehensive documentation of the data ingest process, well explained best practices, and offering training. On the other hand, establishing automatic processes that clean (meta)data, derive metadata, and enrich data. This is an approach that will increasingly become more important. Combined efforts in developing workflows and software solutions for such automatic processes are also necessary, e.g. machine learning tools.

Machine actionability in terms of interoperability relies on clear documented and stable endpoints, from where machines gather the (meta)data. APIs need to be readable with as few limits as possible. Such APIs need to be well documented and they should also deliver the schema of the (meta)data model on request. Best practices on how to successfully mine data from different endpoints and combine them into new data sets used for research questions may help in boosting interoperability use cases. As enabling interoperability is a great benefit for researchers and for processing data further in research projects, data hosts should

[91] Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3:160018, DOI: 10.1038/sdata.2016.18 (2016).
[92] Data stewardship is explicitly mentioned in the guiding principles for FAIR data publishing, available at: https://www.force11.org/fairprinciples.

explain more in detail how to get data from them and how to combine such data with other repositories or how to use the data in projects. It is also important to point out how to integrate the resulting and processed datasets back again into the research data life cycle easily. The establishment of a knowledge base on an international level where people can share experiences could help to lower the barrier for such interoperability approaches.

Recommendations

- Establish quality assurance processes, with a special focus on the data creation phase.
- Pushing data providers and establishing automatic processes to boost (meta)data quality and, therefore, interoperability should be combined and applied.
- Invest in tools that help in cleaning up (meta)data and converting raw data into other (standardised and interoperable) data formats.
- Establish well documented machine-actionable APIs for the (meta)data.
- Give more information on best practices for machine driven automatical data search and reuse (as it is emphasized in Chapter 1.4 of the FAIR principles).
- On a higher level support standard interfaces for exchanging metadata.

3.2.3.2. (Meta)data formats utilize shared vocabularies and/or ontologies

Shared vocabularies and/or ontologies are seldom mentioned in our analysis. One reason is that we didn't specifically ask for this in the questionnaire. It seems that this topic also needs more documentation and best practice examples on how to do this. Also, a compact overview on shared vocabularies and/or ontologies in use for the different research domains would be helpful.

CLARIN's approach with the Concept Registry[93] (based on SKOS) and the workflow around it could give helpful insights on the FAIR 3.2 principle.

The ARIADNE's approach can also provide good guidance on the adoption

[93] Also known as CCR, https://www.clarin.eu/ccr.

of shared vocabularies.[94] ARIADNE has developed an e-infrastructure which enables the integration of archaeological datasets from various different institutions, integrating resource discovery metadata using controlled vocabularies, thesauri, gazetteers and ontology (CIDOC CRM). In ARIADNE, the subjects to which the various datasets relate are described using terms drawn from the Art and Architecture Thesaurus (AAT) of the Getty Research Institute, which formed the spine for the whole framework of terms in ARIADNE. The use of a shared thesaurus required a mapping of each terminological resource, already in use by content providers, to the AAT concepts. This activity demonstrates the semantic and conceptual similarity between the different archives.

In general, it would be good to work on harmonizing the sharing and curation of data from vocabularies and ontologies.

Recommendations

- The description of metadata elements should follow community guidelines that use an open, well defined vocabulary
- Convince researchers to use FAIR compatible vocabularies and ontologies from the very start. Give recommendations on how to do this and how to integrate references in their research data and metadata.
- Give pointers on which vocabularies and ontologies can be used, based on research domain specifics and on the tangible use cases.

3.2.3.3. (Meta)data within the data object should be both syntactically parseable and semantically machine-accessible

Syntactically parseable and semantically machine-accessible data is strongly dependent on established (meta)data formats in a community. As an example, the use of TEI[95] in the LRT research community is enabling interoperability in this perspective.

It is important for semantic interoperability to have well-documented and communicated schema. A well-established approach is the CLARIN Concept

[94] ARIADNE D3.4 Final report on standards and project registry, available at http://ariadne-infrastructure.eu/Resources/D3.4-Final-Report-on-Standards-and-Project-Registry.
[95] Text Encoding Initiative, http://www.tei-c.org/index.xml.

Registry, where shared concepts are identified, described, managed, and given a persistent identifier. In connection with the CLARIN CMDI framework[96] and the CMDI Component Registry[97] this enables a strong interoperability potential. Indeed, such efforts can also be handled by single projects - although probably on a smaller technological level - but it seems to be more stable if an agreement on semantics and the organisation of the descriptions of semantics is handled by higher level institutions, e.g. Research Infrastructures like CLARIN. Reliability and permanent access is crucial when operating with shared semantics. Furthermore, harmonizing such approaches on an international level is highly recommended.

There are some technologies that are mentioned by our partners that work on the level of syntactic and semantic interoperability:

Description	**Link**
ARIADNE Dataset Catalogue Model (ACDM)	http://support.ariadne-infrastructure.eu/
CIDOC Conceptual Reference Model (CIDOC CRM)	http://www.cidoc-crm.org/
CLARIN Concept Registry (CCR) CLARIN Component Metadata Infrastructure (CMDI)	https://www.clarin.eu/ccr https://www.clarin.eu/content/component-metadata

Table 3.3: Technologies on the level of syntactically and semantically interoperability.

The ACDM have been developed to encode the descriptions of content from sparse datasets of archaeological data with the aim to produce a detailed representation of the archaeological information of the legacy archives made available by the consortium through its portal.[98] The Catalogue, and the detailed information it contains, represents the core of the entire integration process.

More documentation is needed on how to combine different datasets between projects and how this works best. We do have a lot of documented protocols and standards, but we lack examples from the other research communities on how to combine datasets from different sources. One reason may be that we

[96] For an intro to CMDI consult https://www.clarin.eu/content/component-metadata.
[97] https://catalog.clarin.eu/ds/ComponentRegistry/.
[98] http://portal.ariadne-infrastructure.eu/.

asked data holders and not data users, because combining data is not storing/providing different data. The interoperability task of combining data is probably mostly done by researchers and research projects.

Recommendations:

- Convince researchers to structure and enrich their research output in such a way that data hosts can ingest this data already as FAIR compatible as far as possible. This needs a joint effort between policy makers and data creators (Section 1.4 gives an overview of the stakeholders in this process).
- Invest into enrichment tools or user interfaces that help to make references in data objects syntactically parseable and semantically machine-accessible.

3.2.4. Reusable

Even though data may be findable, accessible, and interoperable, they are not automatically *reusable,* in the sense of reusable for new research. In order for potential future researchers - or computers - to assess and reuse data, it must be supplied with rich descriptions that precisely establish the scientific status of the data as well as the conditions for its reuse. The recommendations in this section address aspects of future research practices and how current researchers and data archives can best accommodate these.

The FAIR principles define reusable as:

R1. meta(data) are richly described with a plurality of accurate and relevant attributes

- R1.1. (meta)data is released with a clear and accessible data usage licence.
- R1.2. (meta)data is associated with detailed provenance.
- R1.3. (meta)data meets domain-relevant community standards.[99]

The essential requirements for findable, accessible, and interoperable are assumed to be fulfilled, meaning data is already supposed to be identified and equipped with metadata. Under “reusable” we focus on the “richly described” part that particularly enables data-based research.

[99] https://www.force11.org/group/fairgroup/fairprinciples.

3.2.4.1. (Meta)data released with clear and accessible data usage licence

Metadata contained within the Data Object should inform the consumer about the licence of the data elements; this metadata should be machine-readable to facilitate automated data harvesting while maintaining proper attribution. The Metadata contained within the Data Object should inform about any access-control policy, such that consumers can determine which components of the data they are allowed to access.[100]

For allowing data reuse it is necessary to inform the user in understanding the rights and responsibilities through unambiguous statement of legal rights and policies to provide sufficient notification of the legal rights (if any) retained by the rights holder(s). Standardized electronic statements regarding the legal rights retained can support legal interoperability and help in their comprehensibility by a wide audience, and overcome national barriers (see also Chapter 4).

3.2.4.2. (Meta)data associated with detailed provenance

Furthermore, in eScience, where pattern recognition in 'big' functionally linked or integrated data sets is becoming the norm, ***provenance*** *is* ***key****. In case a pattern emerges from the data analysis algorithms, rationalization and confirmational studies in the underlying data sources is a crucial next step. If the provenance of the Data Elements to their original Data Object and subsequently to the underlying resources (human readable text, data bases, raw data files etc.) is lost, researchers will not be able to track the evidence for what the pattern seems to suggest for a testable hypothesis.*[101]

[100] Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing version b1.0: https://www.force11.org/fairprinciples.
[101] Ibid.

Especially in the natural sciences, it is common to build workflows that transform raw or primary data into higher levels of processed data products. Each level builds upon the previous processing, making it very important for every data object to contain an exact provenance description, referring to the data it is based on and documenting what processing, tools, etcetera, that the data was subjected to. This information is referred to as provenance metadata, which is crucial for reuse of processed data for scientific purposes.

Such workflows exist in the Humanities as well, e.g. preparing a document for linguistic analysis by processing it with a chain of tools, such as part-of-speech tagger, lemmatizer, etcetera. But, furthermore, in the Humanities practices such as annotation and versioning call for provenance metadata. Finally, provenance metadata plays a role in digital preservation practices.

Provenance Metadata

The *Term Definition Tool* of Research Data Alliance mentions two definitions of provenance metadata:

1) Provenance information metadata concerning the creation, attribution, or version history of managed data.
2) Provenance metadata that indicates the relationship between 2 versions of data objects and is generated whenever a new version of a dataset is created.[102]

Apart from the importance with respect to reusability, documenting provenance is seen as an integrated part of maintaining digital objects in a digital preservation repository. PREMIS suggests doing this by linking Object entities and Event entities.[103] We have not, however, been able to find any general recommendations on the format of provenance metadata. If not following the PREMIS object model, we suggest that provenance metadata must be added or included in the metadata schemas used instead. To our best knowledge, provenance metadata is not discipline-specific, and ought to be applied in a general and interoperable way. This

[102] Provenance Metadata - DFT: http://smw-rda.esc.rzg.mpg.de/index.php/Provenance_metadata.
[103] Data Dictionary for Preservation Metadata: PREMIS version 3.0: http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf.

would require that:

- Creation and attribution metadata must be part of any bibliographic or citation metadata schema and must be included in all cases. It must be created at the time of deposit into a repository and must be mandatory and machine readable, including e.g. an ORCID for the creator if at all applicable. It is advised that repositories include checks, either manual or automatic, for sensible and correct attribution metadata at deposit time.
- All resources, whether human beings, research or data objects, or specific research tools or software must be referred by their persistent identifiers, rather than by name, abbreviations, etcetera. This specifically requires that software tools must also be registered and persistently identified.

In case of larger, and possibly heterogeneous, datasets, the question arises at which level of granularity provenance should be expressed. Ideally, and in accordance with the FAIR principles' permission to separate data and metadata, provenance could be expressed not only at metadata/dataset level, but for each individual file in the dataset. Especially in the case of heterogeneous datasets, this might indeed be necessary for reuse. This may, however, be difficult to attain, depending on supporting repository software, as well as the file formats and object models being used. It may be possible to develop methods that record changes of individual files precisely enough for reuse, even at the level of the overall dataset.

In practice, a rule of thumb could be that provenance metadata should be provided at the level of object identification. So that if there is one persistent identifier for a complete dataset, there must be, as a minimum, provenance data at dataset level. If each file of a dataset gets its own identifier, provenance metadata must accordingly be provided at file level.

Versioning

It is not uncommon that certain datasets or corpora are live texts, rather than closed project data, and therefore are being continuously improved and developed over time. Also, datasets may evolve, new data being added and errors being corrected, in some cases after the initial deposit into a repository. The issue of versioning is

closely related to provenance, and metadata about versioning, including identification of the version, such as a unique number or tag, a change log record, date, information about who performed the change, etcetera, must be considered as part of the provenance metadata.

New versions of existing digital objects are generally treated in two different ways:

1) The new version is treated as a new object and gets its own persistent identifier, separate from the earlier version. In this case, provenance metadata for the newly created object will need to contain a link and an indication of its relationship to the previous version.
2) The new version is contained within the existing digital object and therefore retains the persistent identifier of this object. It is essential to the scientific integrity, that it must be possible to refer to a specific version of an object. So in this case, the repository service must provide a mechanism to address different versions, for example by adding the version to the identifier as a search parameter or similar.

Format migrations that are being performed as part of digital preservation plans can be understood as creating new versions of data objects, and must follow the general guidelines about versioning. In the case of format migrations, the new version must contain a reference to the old file format that it builds upon, as well as information on the migration process, used tools, etcetera (see under Workflows/tools below). Depending on whether the original format is being preserved alongside with the new one, the reference to the previous version may be retrievable or not.

Annotations

Annotation of resources is a common practice in the Humanities and is often well supported by Research Infrastructures. Even if the researcher performing the annotation is willing to share this as deposited research data objects, it may not be possible to openly share the source being annotated. Different strategies for annotation lead to different requirements for provenance metadata for the annotation:

1) If the source is itself open, and the researcher is authorized to do so, it may be possible to annotate directly into the source file. This case could be classified as generating a new version of the file, as described above. The annotating researcher would need to supply provenance metadata describing his/her annotation/changes to the document.
2) The annotation can be openly shared, whereas the source remains closed (often for rights reasons). This case forces the researcher to create the annotation in a separate object - and possibly a separate repository - from the source. The provenance metadata must describe the annotation (creation, attribution, etcetera) as well as contain a reference (by persistent identifier) to the annotated object.
3) Even if the source file has a licence that allows further sharing, an annotated version may be deposited as a separate data object with its own identifier. This case requires that provenance metadata cover both the annotation (as in 1.) and a clear reference to the object being annotated.
4) In some cases, annotations are machine generated by processing data through a single tool or through a chain (pipeline) of tools. This scenario is described below.

Workflows/tools

In the Humanities as well as in the sciences, data, such as corpora or machine-generated annotations, are created through workflows utilising software or computer-based tools. The result of a workflow can be derived data, modified data, or annotated data. In all cases, the provenance metadata of the generated data object must contain an account of the process:

- What tool was used (persistent identifier), and in what version?
- Possible references to algorithms (journal articles or other documentation).
- Who was initiating the process, when, computing environment, etcetera?
- Reference to the original data/object being processed.

Depending on the file formats, this information may be included in the resulting files themselves, or be added to the metadata for the dataset or data file in question.

The purpose of this guidance is partially to allow researchers to verify and

recreate data objects from their sources, as much as possible following the exact method of the original processing. It will also allow for implementing error fixes and improvements of algorithms and to make it possible to identify parts of a workflow that could benefit from being rerun. Finally, it will allow future researchers to assess which stage of a workflow to use in the case of repurposing the data for a different kind of research question.

Please note that software is being considered as an object in itself that may require its own provenance record. As a minimum, it must be described and identified persistently with a reference to an authoritative source.

3.2.4.3. (Meta)data meet domain-relevant community standards

The FAIR principles consequently use “(meta)data” to indicate that the guidelines relate to both metadata and “data” in the sense of the research content of the digital objects. With respect to community standards, this guides us to look into domain-relevant metadata standards as well as standardised +data formats. The move towards creating more generic Research Infrastructures already tends to be a move towards defining and using community standards. This has been taking a step away from previous practices, where it was common that each research project invented its own formats. If the data and metadata being created fit reasonably into community standards, these must be applied, which leads to utilising existing infrastructure services.

On the other hand, it is still part of research to invent new things. In other words, there will be cases where using community standards is not going to be in the best scientific interest. As a result, it may be harder to follow FAIR principles, as there may not be a supporting infrastructure that is already supporting the formats being created and used. The discussion of standards constitutes an ongoing negotiation between researchers’ needs to define their own formats and the need for infrastructure support and data interoperability and reuse.

With respect to metadata, the requirement for using discipline specific metadata should be understood as supplementary to the metadata requirements already discussed under *Findable*. Here we focus on metadata that specifically describe the type of resource in question, the manuscript, the excavation data, the

corpus, etcetera, under study. This is the more scientific kind of metadata that will help future researchers to assess the usability of the data for specific research purposes. In repositories or infrastructures supporting particular research communities with particular types of data, such discipline specific metadata can also help facilitate specialised discovery options or search criteria being required within the specific community.

As a general guideline, data archives are advised to work with research communities to establish the relevant community standards for their target users, and to build as much support of such standards into their infrastructure as possible. Apart from metadata and data formats, this can include support for specialised tools operating on agreed data formats or integration into Virtual Research Environments. Discipline-specific repositories are often well suited to offer such support, but even more general repositories may offer specialised support in certain fields.

Competing standards

There can sometimes be competing standards within a community, which may cause repositories and infrastructures to support more than one standard for a given research community. It is recommended that the standards being followed are endorsed by the research community, and that general infrastructures which are not entirely dedicated to e.g. one particular format, must be flexible enough to accommodate the actual research being performed in the field. This would also call for allowing some very generic types of data to accommodate research data in areas where no standards have (yet) been defined.

Object and content models

Community standards will not necessarily follow the data-metadata separation in FAIR and may imply different object and content models, and representations. As an example, the text community is often using TEI[104], supporting self-contained objects that encompass both data (body) and metadata (header) and suggests various content models, according to the type of text being modelled. This is not necessarily easy to map into a data-metadata object model, and indeed TEI has caused headaches for people implementing data repositories.

[104] TEI: Text Encoding Initiative: http://www.tei-c.org/index.xml.

Even in a case like TEI, as well as other formats grown out of research practices rather than from repository and infrastructure builders, it will, in most cases, be possible to create and describe datasets appropriately, possibly by employing some automatic extractions of metadata from data files into repository metadata fields. Here the complications can sometimes seem to be human rather than technical, as researchers may be unwilling to relax their paradigms into a more general infrastructural view.

Preferred formats for data stewardship and preservation

Especially in the case of disciplinary repositories or Research Infrastructures, it may be possible and desired to prescribe a prioritized list of data formats, combined with graduated support. An example of this is DANS that offers lists of *preferred* and *acceptable* file formats, based on general guidelines for obtaining the best long-term sustainability and accessibility:

- Are frequently used.
- Have open specifications.
- Are independent of specific software, developers or vendors.[105]

Based on these criteria, extensive lists of preferred and acceptable file formats are being offered, with respect to long-term usability, accessibility, and sustainability. An approach like this can be very helpful to guide researchers towards sustainable data formats. Additional considerations for preservation formats can be found in Section 3.3.2.

It has to be noted though, that focusing on file formats does not necessarily cover the content of such files and may need to be supplemented by community standards for the content model. As an example, XML is considered a preferred preservation format, but you will still need to supply guidance and maybe support for the specific schema to be used (e.g. TEI), in order to best provide reusability. And even a schema such as TEI or other discipline-specific standards will often have its

[105] DANS Preferred Formats. September 2015, version 3.0: https://dans.knaw.nl/en/deposit/information-about-depositing-data/DANSpreferredformatsUK.pdf.

own versioning[106] and thereby raise its own challenges regarding preservation, as discipline formats will also need upgrading as part of preservation plans, separate from possible file format migrations.

Recommendations

- Use international standard formats, i.e. XML and RDF textual formats.
- All files held in the repository should be in an open, simple, standardised format that is considered likely to offer a degree of long-term stability (see also Section 3.3.2 about preservation formats). When a format is in danger of becoming obsolete, proper digital preservation actions must be performed.[107] Adopt the preservation by migration, if necessary.
- Use open source tools for generating metadata and for automatic validation.[108]

3.3. Supporting practices to FAIR data

While FAIR principles address data itself at a quite high level, actually providing FAIR data will depend on observing good practices in Data Management and Data Stewardship. Such good practices have to be observed both by researchers, creating and providing the data, and by repositories and Research Infrastructures providing services for the long-term stewardship and access of data. A few of these supporting practices provide well-established methods in their own right and will be subject to specific recommendations: data management planning and long-term digital preservation.

3.3.1. Data Management Planning

Data Management Planning is already becoming a well-established practice, due primarily to funders requiring researchers to provide a Data Management Plan

[106] TEI is currently version 5 (P5) and has required migrations from previous versions within the same XML specification.
[107] Archaeology Data Service (ADS), available at http://archaeologydataservice.ac.uk, section 3.2.
[108] MiBACT-ICCU.

(DMP) as part of funding applications. Various universities and institutions are providing support, guidance and tools for developing such plans, among them the Digital Curation Centre in the UK with its widely adopted DMP online tool. Furthermore, data management guides or actionable data management plans that could potentially enable some sort of automatic allocation of repository or storage resources, depending on requirements defined in the DMP, are being developed.

Lately, the European Commission has considerably accelerated this development by adding requirements for data management plans as project deliverables in Horizon 2020 funded projects. The Commission has provided guidance and a DMP template, strongly based on the FAIR principles. Generally, there is a strong sense that a good data management plan provided by the research team is a first step towards FAIR data.

Supplementary to this perspective, in which researchers are asked to provide DMPs in order to receive funding, repositories and infrastructure providers have also developed an interest in DMPs. Aside from those repositories that have received accreditation (e.g. from DSA, WDS, DIN and others), there is still a huge number of institutes managing repositories whose data policies are unknown. In fact, repository accreditation involves only a small number of repositories, perhaps due to the general attitude, which goes towards voluntary accreditation (bottom-up approach) in some countries, and prescribed by law (top-down approach) in others.

Although it is not a mandatory requirement for repository providers, the DMP could offer the right level of trustworthiness to repositories that have not undergone an accreditation procedure.

The survey presented in Appendix III: Questionnaire, shows that some service providers require a DMP to allow researchers to deposit their data. The repository may in itself have some requirements for data management that need to be included in the plan, in order for the resulting data to be accepted. This would seem to imply that the researcher makes some sort of agreement with a repository already at the time of application, based on the DMP. But it also adds possibilities for the repository to better adjust its services towards actual researcher needs.

3.3.1.1. PARTHENOS Data Management Plan - draft template

The PARTHENOS DMP template we propose aims at addressing the domain-specific procedures and practices within the Humanities, paying special attention to standards and guidelines used in data management that are relevant for this specific research community, which includes archaeologists, historians, linguists and social scientists. The PARTHENOS DMP describes the data life cycle of the data that is created, collected, archived and preserved by projects and Research Infrastructures in the Humanities, including information that makes data FAIR: findable, accessible, interoperable and re-usable.

The PARTHENOS DMP, which builds on the Horizon2020 DMP template, has been enriched and tailored with the specifications from the Humanities, which were derived from a survey carried out among the consortium's experts (see Appendix III: Questionnaire). To gather these specifications, we asked representatives of the PARTHENOS communities to describe their daily data management procedure in detail, structuring the questionnaire into the various phases of the data life cycle and then mapping them to the FAIR principles. Each respondent had the opportunity to choose his/her role (e.g. researcher creating data/ repository provider) and replied providing hints from their good practices.

There are many reasons that lead us to adopt the H2020 DMP template as a starting point to develop the PARTHENOS DMP. On the one hand, the template is already well consolidated and well-known by most researchers, which makes the researcher's approach to the DMP easier. On the other hand, as it is structured around the FAIR principles, it was easier for us to map the results of our analysis to the issues addressed by the template, and consequently to provide support to researchers, offering them solutions rather than uncertainties.

That said, the PARTHENOS DMP is the result of a first attempt to collect the high-level requirements that satisfy each community of researchers involved in the project, with the aim to provide guidance and support to Humanities researchers as they write their DMP, proposing a list of recommended answers that will facilitate them in compiling the DMP. A second stage of this work will concern the production of PARTHENOS community-specific DMP templates, which will be included in the final version of the deliverable on "Guidelines for common policies implementation"

(D3.2) in alignment with the discussion of the Science Europe Working Group on Research Data, which include representatives from the various communities.

Further work will concern the creation of a DMP template addressing institutions that manage the repositories. Since enabling interoperability is a great benefit for researchers, repository providers should be able to explain in depth how to provide data to them in the best way. Through the envisaged template, PARTHENOS will provide the right tools to repository providers to be able to offer standardised answers and guidance, and to liaise with researchers that are looking to ingest their data.

DMP component	Issues to be addressed	Guidance
1. Data summary	State the purpose of the data collection/generation	Please, include a brief description of the reason for collecting/generating data
	Explain the relation to the objectives of the project	☐ Data availability ☐ Data reuse ☐ Data interoperability ☐ Other, please specify
	Specify the types and formats of data generated/collected	☐ All data formats are collected ☐ Only open formats are collected ☐ List data formats: i.e. XML, RDF, TEI… ☐ Include link to preferred data format document ☐ Select format from the suggested lists: ○ CLARIN ○ Meertens Institute ○ CINES ○ KNAW-DANS ○ GAMS ○ ADS ☐ Other, please specify
	State the expected size of the data	☐ Number of files, please specify ☐ Number of digital objects, please specify ☐ Not available
	Specify the granularity of the collected data to be archived	☐ Single items (i.e. one page of a manuscript, one excavation report…) ☐ Datasets ☐ Collections ☐ Corpora ☐ Other, please specify

DMP component	Issues to be addressed	Guidance
	Specify if existing data is being reused (if any)	☐ Yes ☐ No ☐ I don't know
	Specify the origin of the data	☐ PhD data ☐ Project results ☐ Other, please specify
	Outline the data utility: to whom will it be useful.	Please, list possible stakeholders reusing your data
	Describe your strategy of data exploitation	Please specify if you: ☐ Plan agreements with other institutions ☐ Use policy on data reuse ☐ Use licence of use, specify which ☐ Other, please specify
2. FAIR Data 2.1. Making data findable, including provisions for metadata	Specify standards for metadata creation (if any). If there are no standards in your discipline describe what type of metadata will be created and how.	Please, select from the list: ☐ ACDM ☐ CARARE ☐ CCR ☐ CIDOC CRM ☐ CMDI ☐ DC ☐ DDI ☐ Europeana ☐ ICCD ☐ MIDAS ☐ OAI-ORE ☐ Other, please specify
	Specify if metadata are updated during the project and/or once the data are archived	☐ Yes, automatically ☐ Yes, manually ☐ No ☐ I don't know

DMP component	Issues to be addressed	Guidance
	Describe the mechanisms used to identify digital resources	☐ Digital Object Identifiers ☐ Aleph System number ☐ Data Cite DOI ☐ Digital resources ID ☐ Landing page ☐ URL/URI ☐ PID/handle ☐ OAI identifier ☐ Other, please specify ☐ No identification mechanism
	Outline the approach towards search keywords	☐ Full text search ☐ Advanced search ☐ Faceted search ☐ Search by collections ☐ Other, please specify
	Describe how resources are being retrieved from your repository, which interfaces and standards are supported (including API's for indexing and object retrieval)	☐ Resources can be downloaded from the landing page ☐ FLAT ☐ OAI-PMH ☐ FCS API ☐ Actionable APIs ☐ LUCENE ☐ SOLR ☐ IIF ☐ FEDORA ☐ Other, please specify
	Specify if metadata of non-public resources are publicly available	☐ Yes ☐ No ☐ I don’t know
	Outline naming conventions used in your project	Please, list below the most relevant ones

DMP component	Issues to be addressed	Guidance
	Outline which data publication workflow is followed	☐ No formally defined publication workflow ☐ DSA criteria ☐ According to a DMP, please specify ☐ OJS publishing platform ☐ OAIS reference model ☐ ElasticSearch ☐ Virtuoso Triple Store ☐ Registry API ☐ Other, please specify
2.2 Making data openly accessible	Specify which data will be made openly available? If some data is kept closed provide rationale for doing so	Specify if there are any restrictions on public accessibility and describe the exceptions to public and free access
	Specify how the data will be made available	☐ Deposition in a repository, please specify which ☐ Other, please specify ☐ I don't know yet
	Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?	☐ Component Metadata Infrastructure (ISO 24622-1) to create an environment that supports different metadata schema ☐ MAG and METS-MDI schemas ☐ Dublin Core ☐ VRA ☐ NISO ☐ MD5 ☐ METS ☐ ACDM ☐ CIDOC CRM ☐ (Qualified) Dublin Core metadata

DMP component	Issues to be addressed	Guidance
		fields. ☐ NAKALA or ISIDORE ☐ Dublin Core elements for collection/thematic metadata. ☐ GEMINI for spatial terms ☐ LOD terms such as LCSH, TGN and Heritage Data are used within metadata. ☐ Other, please specify
	Outline the method used to ensure that there is appropriate metadata available to ensure the understanding and reuse of data over time.	☐ Minimum set of metadata required ☐ Metadata is associated to each digital object ☐ Use of metadata standards, please specify which ☐ QA committee for metadata ☐ Other, please specify
	Specify where the data and associated metadata, documentation and code are deposited	Please, specify the repository you will deposit your data to
	Specify if the repository you will submit your data is accredited	☐ Yes, please specify: ○ DSA ○ WDS ○ DIN 31644 ○ ISO 16363 ○ Other, specify which ☐ No ☐ I don’t know
	Specify if the repository you will submit your data informed you about the recommended formats	☐ Yes ☐ No ☐ I don’t know
	Outline if reliability and service levels of the repository are	☐ DSA

DMP component	Issues to be addressed	Guidance
	specified, and which certificates and methods of assessment are acceptable	☐ NESTOR ☐ Self-assessment ☐ Internal standards and procedures relying on the Trustworthy Repositories Audit and Certification ☐ Other, please specify
	Specify if a (written) access policy to the archived data is available, which e.g. states when and under which conditions a resource become available to different actors: submitter; reviewer, collaborators; scientific community; general public, etcetera	☐ Yes, please specify ☐ No ☐ I don't know
	Specify if your archive is subject to national/European laws and regulations	☐ Yes, please indicate: ○ IPR ○ privacy regulations ○ database rights ○ Other, please specify ☐ No ☐ I don't know
	Specify how access will be provided in case there are any restrictions	☐ Authenticated access ☐ Scientific board ☐ No access provided ☐ Other, please specify
	Outline the process used to ensure the integrity and authenticity of the data stored by your organization/RI.	☐ No general policy ☐ Checksums ☐ FEDORA mechanism ☐ Scientific board ☐ Other, please specify
	Describe how metadata with restrict access are maintained	☐ Through policy (please, specify) ☐ Access Control Lists ☐ NAKALA system

DMP component	Issues to be addressed	Guidance
		☐ Granted access upon request ☐ Other, please specify ☐ Not applicable
2.3. Making data interoperable	Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.	☐ AAT ☐ ACDM ☐ CCR ☐ CIDOC CRM ☐ CMDI ☐ OAI-ORE ☐ Other, please specify
	Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?	☐ Yes ☐ No ☐ I will provide mapping to common ontology, please specify
	Specify the adopted standards or best practices for digital content creation (digitization). Specify the standards used (link to the URL, if online or attach a copy, if the standards are locally customized)	☐ Recommendations of research funding national organization, please specify ☐ Guide to good practices, please specify ☐ HTML ☐ XML-TEI ☐ JSON ☐ Link of the document (please, specify) ☐ Other, please specify ☐ Not applicable
2.4. Increase data reuse (through clarifying licences)	Specify how the data will be licenced to permit the widest reuse possible	☐ Open data policy ☐ Public Domain Mark ☐ CC0 ☐ CC-BY ☐ CC-BY-SA

DMP component	Issues to be addressed	Guidance
		☐ Other, please specify
	Specify when the data will be made available for reuse. If applicable, specify why and for what period a data embargo is needed	☐ No specific date ☐ Free access is subordinated to legitimate interests of rights holders and protection of confidentiality and personal information and protection of cultural resources ☐ Embargo date can only be handled when the technical framework allows it ☐ Date individually set with the data providers ☐ Other, please specify
	Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the reuse of some data is restricted, explain why.	Describe your strategy and licence policy if thirds parties reuse data: ☐ Creating revenue ☐ Combining data with other data ☐ CC NC ☐ Free reuse if appropriately cited ☐ Intellectual property rights ☐ Protection of confidentiality and personal information ☐ Protection of cultural resources ☐ Other, please specify
	Specify if a policy on data created on third parties' data is available	☐ Yes, please specify ☐ No
	Describe data quality assurance processes	☐ Scientific and technical committee ☐ Tools for automatic checks ☐ Conform to format specification ☐ Verifying consistency with data models and standards

DMP component	Issues to be addressed	Guidance
		☐ Other, please specify ☐ No formal QA process defined
	Specify if defined criteria ensuring relevance and understandability of the data for users are available	☐ No general policy ☐ Requiring a minimal set of metadata to be procured ☐ Panel of specialists for QA ☐ Formats, standards and certification models recognized by the scholarly community ☐ Use of metadata schemas that can be mapped onto the Virtual Language Observatory facets ☐ Collection level metadata ☐ Other, please specify
	Specify the length of time for which the data will remain re-usable	☐ 5 years ☐ 10 years ☐ Other, please specify
	Specify if the rights related to the data are documented	☐ Yes ☐ No
	Describe which information you gather on the rights holder, and how you make sure that nobody is left out	☐ The rights owner is recorded in the metadata form ☐ Adequate documentation / permissions are gathered from their holders ☐ Agreement with each content provider ☐ The data creator is responsible for recording any rights ☐ If rights are held by third parties, the creator is responsible for ensuring permissions are given, or content

DMP component	Issues to be addressed	Guidance
		removed ☐ Support standards for data citation ☐ Provide proper attribution and credit information in an external metadata record where a dataset is implemented by different individual contributors ☐ None of the above, because…
	Provide an example of how you ensure the availability of sufficient information (technical data and metadata) for end users to enable them to make reliable quality-related evaluations (if the data allows it)	☐ Staff with specialized education or training ☐ Detailed metadata ☐ Special training course to use specialized infrastructure ☐ QA working groups ☐ Domain experts collaborate with technical partners to ensure precise mappings from content providers schemas to project ontology ☐ Other, please specify
	Specify the licences covering data access and reuse, and describe how the compliance is checked	☐ Creative Commons ☐ Rights Management licence framework ☐ Shibboleth authorization ☐ Compliance is checked ☐ Compliance is not checked ☐ Other, please specify
	Specify if you consider copyright and intellectual property important concerns in managing digital materials when data are being reused	☐ Yes (select from the list below) ○ Permissions is granted for copyrighted material upon written request ○ Permission is requested to: a) to authors for online publishing;

DMP component	Issues to be addressed	Guidance
		b) to publishers for online republishing of printed works; c) to persons appearing on audio-visual materials; d) to reproduce places, monuments, artefacts in audio-visual and other media; e) to library owning copy of rare texts in public domain ○ Other, please specify ☐ No ☐ I don't know
3. Allocation of resources	Estimate the costs for making your data FAIR. Describe how you intend to cover these costs	☐ Funding provided by the project ☐ Collaboration with other projects ☐ Other, please specify
	Clearly identify responsibilities for data management in your project	Please, list the responsible actors/partners for every data life cycle activity
	Describe costs and potential value of long-term preservation	Describe cost for long-term preservation: (get help in calculating RDM cost with the Guide Research Data Management and Costs). Potential value of long-term preservation, please select from the list below: ☐ Data is potentially important for reuse by a larger community ☐ Data contributes to improve an open access publication ☐ Data was produced with a process that is difficult to repeat

DMP component	Issues to be addressed	Guidance
		☐ Data need to be archived because the financier requires it ☐ Other, please specify
	Clarify how you estimate the costs of archiving	Please, specify which is the "unit" of archiving: ☐ Price per megabyte ☐ Price per digital object ☐ Price per number of backups ☐ Price per authorized user ☐ Price per file ☐ The cost is covered by the archive ☐ The cost is covered by the project ☐ Other, please specify ☐ Not available
4. Data security	Address data recovery as well as secure storage and transfer of sensitive data. Specify if your organization/RI developed tools to control the risks associated with receiving, managing, processing and ingesting digital collection content	☐ Checking/syntactic parsing of data structures ☐ Mechanisms to secure the reception and storage of exact copies of the original files (ingestion phase) ☐ Tools for generating metadata and for automatic validation of the XML ☐ Virus scanner for scanning file uploads ☐ Technology vulnerability scan ☐ SLA with the data storage provider ☐ Procedure for file fixity checking ☐ DRAMBORA Risk Assessment ☐ Declaration of Confidentiality for employees ☐ Bespoke Content Management System (CMS) with Object Management

DMP component	Issues to be addressed	Guidance
		System (OMS) extension. ☐ FLAT: a repository solution based on Fedora Commons ☐ Other, please specify
	Specify if you have policies regarding the storage of intermediate results and temporary files	☐ No general policy ☐ Policies on IPR ☐ Licences policy ☐ Other, please specify
	Specify if your system uses automated backup processes, and if an automated monitoring processes of storage is available	☐ Yes, please specify ○ FLAT ○ Scheduled backup processes ○ Microsoft Cloud ○ SURFsara ○ FEDORA version control ○ Other ☐ No
	Describe the digital asset management system used. The system may be used to manage the full life cycle of your digital objects	The system includes: ☐ Management of data creation ☐ Metadata repository ☐ Image repository ☐ Registry of preservation metadata ☐ Tools providing access to users: ○ FEDORA ○ DSpace ○ Locally developed system ○ FLAT ○ Escidoc/Fedora Commons as DAMS ☐ Other, please specify
	Describe how the system supports preservation	☐ Snapshots on the NAS (Network

DMP component	Issues to be addressed	Guidance
		Attached Storage) for "hot data" ☐ Distributed copy on our distributed file system (Active Circle) for "lukewarm data" ☐ Backup on LTO tape drive for "cold data" ☐ Long-term preservation (+/- 20 years) ☐ DOI and URN Persistent Identifiers are assigned to a dataset ☐ All data streams are preserved in the original format as distinct files ☐ Other, please specify ☐ No preservation supported
5. Ethical aspects	Outline how your organization/RI ensures compliance with disciplinary and ethical norms	☐ Anonymising data where necessary ☐ Privacy constraints and applicable ethical norms ☐ Data accompanied by informed consent statements ☐ VSNU guidelines ☐ Privacy policies ☐ National laws ☐ ALLEA's European Code of Conduct for Research Integrity ☐ Other, please specify
	Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)	☐ My institution has a RDM, please specify ☐ No other procedures are used ☐ Horizon 2020 ☐ DCC DMP ☐ ZonMw

DMP component	Issues to be addressed	Guidance
		☐ Arts and Humanities Research Council ☐ DDI ☐ UK Data Archive ☐ Other, please specify
6. Other	Describe what your organization/RI does to enable long-term preservation of digital resources	☐ Incremental and periodic backup ☐ Updating of software ☐ Cooperation with national stakeholders to enable long-term preservation ☐ Preservation by migration ☐ Other, please specify ☐ No formal processes and systems for long-term digital preservation
7. Long-term preservation	Specify the workflow used to ensure long-term preservation	☐ Creation of metadata and documentation ☐ Data validation ☐ Registration of audit trails ☐ Bit integrity ☐ PID redirection ☐ Metadata conversion ☐ Moving digital objects ☐ Check the integrity of the copy by md5 checksum ☐ Preserving access restriction and access control lists ☐ Other, please specify
	Specify if different workflows for different data are available	☐ Yes ☐ No ☐ I don’t know
	Describe how you identify appropriate approaches and tools to prevent technological obsolescence.	☐ No formalised technology watch ☐ FLAT

DMP component	Issues to be addressed	Guidance
		☐ List of preferred formats ☐ Regular review cycles of the hardware and backend ☐ Preservation by migration ☐ MoRe ☐ Collaboration with other institutions for addressing new problems and solutions ☐ Other, please specify
	Outline how the heterogeneity of your digital content may influence your processes, e.g. in respect of operating systems or documentation	☐ Some data can only be stored as is for download ☐ Development of transcoding routines ☐ Digital works containing heterogeneous formats cannot be fully supported in our infrastructure ☐ Acceptable if in preferred formats ☐ Other, please specify ☐ No general policy
	Describe the process you follow (if any) to ensure continued authenticity and integrity of your digital resources throughout time	☐ Manual procedure, like peer-review ☐ Automatic procedures for fixity checking ☐ Format migration ☐ Deposition of new versions of datasets ☐ FLAT ☐ FEDORA mechanism ☐ geo-replication maintaining a complete copy of the data archive at a remote site ☐ Other, please specify
	Specify to what level your organization/RI does bit	☐ Number of copies

DMP component	Issues to be addressed	Guidance
	preservation	☐ How often are copies checked individually ☐ How often copies are checked for changes between copies ☐ Other, please specify
	Describe what kind of packaging your system uses for data under bit preservation. Specify if you support different levels of bit preservation	☐ Several replicas of data preserved ☐ Independence between the replicas ☐ Geographical independence ☐ Organizational independence ☐ Regular audit of the replicas being intact ☐ Only a restricted group of users is allowed to access the bit-streams ☐ Accessible and usable upon allowed demand ☐ Linkage of persistent identifier ☐ Other, please specify
	Specify how the preservation of legacy data is handled. Do you create updated metadata? Do you review IPR?	Preservation: ☐ Bitstream ☐ Preservation by migration ☐ Other, please specify Metadata ☐ Metadata updates are applied ☐ Metadata are not update IPR ☐ Review IPR in relation to Europeana Publication strategy ☐ Other, please specify ☐ No legacy data
	Specify if your institute/RI collaborate with other national and	☐ AAI

DMP component	Issues to be addressed	Guidance
	international institutions in digital preservation initiatives	☐ ARIADNE ☐ CESSDA ☐ CIDOC CRM SIG ☐ CINES ☐ Data Seal of Approval ☐ DCCD ☐ Digital Preservation Coalition ☐ EGI ☐ EHRI ☐ EOSC ☐ EUDAT ☐ INDIGO-DataCloud ☐ KNOWeSCAPE ☐ National CLARIN consortia ☐ OPENAIRE ☐ PARTHENOS ☐ Re-SEARCH ☐ ZIM-ACDH
	Specify if your organization/RI avails of existing policies on data preservation	☐ Yes, please specify ☐ No policies

3.3.2. Long-term digital preservation

In order to ensure a sustainable provisioning of access, there must be a preservation programme dealing with digital preservation issues covering the methods, organisation and systems which are needed to ensure access to digital materials over the long term. This programme covers the series of managed activities necessary to ensure continued access to digital materials for as long as necessary, also called digital preservation.

It should be noted that some aspects of digital preservation are hard to relate to specific FAIR principles, as it is the very foundation for finding, accessing and reusing of data, where interoperability is closely related to the way data must be preserved in order to be understood and processed over a long period. The repositories could work towards incorporating the FAIR principles into their everyday operations and making them implementable in any trustworthy digital repository.

The recommendations are to take into account the various aspects of digital preservation as they are described in this section. The various aspects of digital preservation must be taken into account in the full lifecycle of data, e.g. from validation of data at creation time to emulation of data at access time in the future.

Digital preservation has a wide span of topics, which cannot all be addressed here. Therefore, this should be seen as a summary of some the most important topics, where additional literature can be found e.g. on Digital Preservation Handbook,[109] the OPF website,[110] and various freely available papers from the iPres conferences.[111]

A known and much used standard within digital preservation communities is the Open Archival Information System (OAIS) reference model presented in ISO 14721:2012.[112] This reference model describes at an abstract level, which functional entities are required in digital preservation (both on organizational and technical levels). A very brief introduction of OAIS reference model is provided here, as it assists in describing where digital preservation must be taken into account. The

[109] DigitalPreservationCoalition (DPC), "Digital Preservation Handbook", Available at http://www.dpconline.org/handbook/.
[110] Open Preservation Foundation website, Available at http://openpreservation.org/.
[111] International Conference on Digital Preservation website, Available at https://ipres-conference.org/.
[112] ISO 14721:2012. Space data and information transfer systems - Open archival information system (OAIS) – Reference Model, see: https://www.iso.org/standard/57284.html.

functional entities and information packages are illustrated in Figure 3.1.[113]

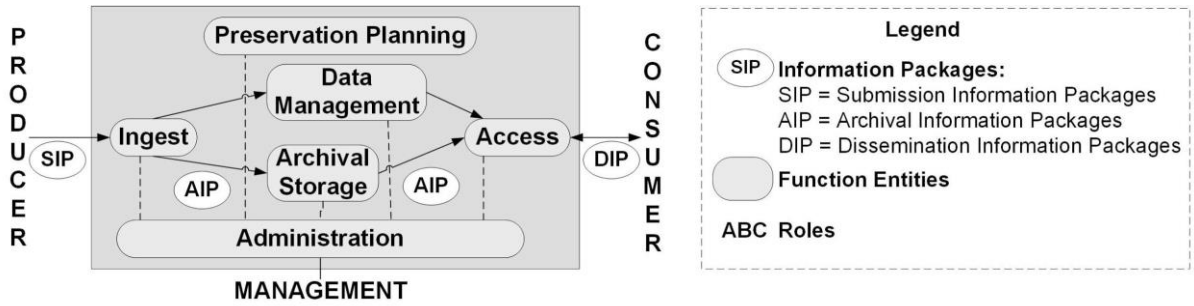


Figure 3.1: OAIS Reference Model - Functional Entities.

The SIP is the submitted data, which is processed to an AIP for archiving. An AIP may, for example, be a migrated version of the submitted data (into a preservation format) and enriched with metadata that enables future preservation actions and access. The Archived AIP can be accessed where a DIP is delivered in a dissemination form (which may be a migrated version from the archived version into format that is better for dissemination).

An important functional entity is the Preservation Planning, which must ensure that preservation plans are constantly updated and executed in order to fulfil preservation purposes based on existing standards and technology. The Preservation Planning is defined by Management and administered through the Administration functional entity and influences every other part of the functional entities of the entire technology and organisation constituting a repository with digital preservation.

It should be noted that OAIS is a reference model and not an implementation model, thus the functional entities represent functionality that should be addressed by the repository as a whole and information packages are not necessarily physical packages.

The ISO 16363:2012 [114] *Audit and certification of trustworthy digital repositories* standard is based on the OAIS reference model and basis for certifying trustworthy digital repositories with digital preservation. There are usually two levels

[113] Corresponds to Figure 4-1: OAIS Functional Entities in the OAIS Reference model.
[114] ISO 16363:2012. Space data and information transfer systems -- Audit and certification of trustworthy digital repositories, see: https://www.iso.org/standard/56510.html.

of preservation, although they are interrelated:

- *Bit preservation* to ensure that the bit-streams remain intact and readable
- *Logical preservation* to ensure that the bits remain understandable and usable according to preservation purposes.

In order to obtain a repository with a trustworthy *Sustainable digital preservation,* both levels of preservation must be supported by a well-funded organization with well-described and implementable digital preservation policies and strategies.

3.3.2.1. Bit preservation

Bit preservation is basically about preservation of bits as preservation of the *integrity* of bits. However, in order to obtain the optimal bit preservation, there are other aspects to take into account. For example, other information security aspects as defined in the ISO 27000 series are: *availability* and *confidentiality*. Additional requirements are: *linkage of persistent identifier* the bits it supposed to point to and sustainability (addressed separately)[115].

Bit preservation - *integrity* should include:

- Several replicas of data preserved.
- Independence between the replicas.
- Regular audit of the replicas being intact.

Independence between the replicas can be obtained on several levels, as geographical independence by placing replicas on different geographical places in order to mitigate risks of losing all copies in a fire or as consequence of a natural disaster. This form of independence was also seen in one case of the PARTHENOS partners. Other types of independence can be organizational in order to mitigate risk of the same person/procedure to destroy all copies of data by mistake, different hardware, operating systems, media to mitigate risks of the same errors occurring etcetera.

[115] More detailed information on “A Holistic Approach to Bit Preservation” can be found in: Library Hi Tech: Vol 30, No 3, pp.472 - 489, DOI:10.1108/07378831211266618.

Bit preservation - *confidentiality* is the property that information is not made available or disclosed to unauthorised individuals or processes (e.g. only a restricted group of users are allowed to access the bit-streams). Such aspects must be carefully evaluated against the priority between requirements to integrity and confidentiality.

Bit preservation - *availability* is the property of being accessible and usable upon allowed demand (e.g. ability to access and possibly to process the bit-streams in connection with reuse). As for confidentiality, such aspects must be carefully evaluated against the priority between requirements to integrity and confidentiality.

Bit preservation - *linkage of persistent identifier* is needed as a requirement as there are cases where the linkage is only part of metadata that are not bit preserved, and therefore can get lost. In OAIS terms, this requirement could be formulated as the persistent identifier needing to be part of the archived AIP.
In relation to bit preservation, the Preservation Planning must cover planning to ensure that the bit preservation policies and strategies are fulfilled.

3.3.2.2. Logical preservation

Logical preservation can be quite complex due to potential complexity in data structures. The EU Planets project[116] set out to provide technical support for logical preservation, based on a view of three main interrelated activities:

- Characterisation;
- Preservation planning;
- Preservation actions.

Common to all the above categories of activities is that they are either based on or contribute to the metadata of the digital material.

Logical preservation - *Characterisation* consists of finding characteristics of digital material and file formats. Characterisation is important in order to:

[116] More information can be found on http://planets-project.eu/.

- Make quality assurance of the data which for example may discover incompatibility with used file formats, and thus either reject or correct the data before further processing. Another example would be to match format against accepted formats and again reject data or migrate them into a preservation format.
- Produce primarily technical metadata (e.g. using MIX[117] for still images) and provenance metadata (e.g. using PREMIS[118]), which can be crucial for deciding on a preservation strategy for the data or for later discovery of obsolescence of formats that will result in enabling of preservation actions.
- Compare differences between characteristics of an original file and a migrated file as input to an evaluation of whether the losses from the migration are acceptable.

It is important to get as precise and standardised information form characterisation as possible. For instance, by using recognised format registries like PRONOM[119] for specification of the exact file format.

Logical preservation - *Preservation planning* involves specification of preservation plans as well as determining the best ***preservation strategy***. Preservation actions are initiated based on the preservation plans.

There are pre-conditions for planning and timely execution of appropriate functional preservation actions, for example that there is sufficient information, or access to retrieval of information, on which the planning is based. That means that the bit preserved digital material must be prepared for planning and execution of functional preservation actions. Furthermore, preservation planning must be executed in a way that ensures that preservation policies and strategies are followed.

Logical preservation - *Preservation strategies* are usually based on two main

[117] *Metadata for Images in XML Schema (MIX)*, Version 0.2 (draft), July 30, 2004; http://www.loc.gov/standards/mix/.
[118] http://www.loc.gov/standards/premis/.
[119] The National Archives Technical Registries PRONOM, Available at http://www.nationalarchives.gov.uk/PRONOM.

preservation strategies suitable for digital preservation:

- *Emulation*
  The emulation strategy consists of simulating the original environment that was used to render the digital material. The original bit-streams are then rendered in a new environment via the emulated environment
- *Migration*
  The migration strategy consists of migration of the data from one representation to another, i.e. from one structure and contents represented in a set of files to a possibly new structure and a new set of files with new file formats. Use of preservation formats may be important in order to gain best value of a migration strategy.

For both of these preservation strategies the success of preservation depends on the success of preserving the authenticity in an emulated environment or in the target file format of a migration. There will almost always be some sort of loss no matter which strategy is used, and it is, therefore, important that the preservation purposes and significant properties (properties we want to preserve) are defined in a way that can enable choice of the most suitable preservation actions.

Logical preservation - *Preservation actions* cover different actions initiated based on the preservation plans and are executed by tools and organisational procedures. Examples of preservation actions on the logical level are characterisation or file format migration of data. Although only defined on the logical level, there are also preservation actions on the bit level, such as integrity check of replicas and media migration.

3.3.2.3. Digital preservation policies and strategies

The basis for determining the right preservation planning for all levels of digital preservation planning, characterisations and actions is policies and strategies which are constantly maintained on basis of new technology and changing organisation (according to Monitor Technology described in the OAIS Reference Model). A selection of important issues that policies and strategies should address are:

- Persistent identifiers;
- Standards;
- Preservation formats;
- Sustainability;
- Audits.

Policies and strategies - *Persistent identifiers* are important in order to ensure persistent reference to data and thus long-term access of data. According to the Digital Preservation Handbook [120] choice of a persistent identifier scheme is described as follows:

> **Choosing a Persistent Identifier Scheme**
>
> There needs to be a social contract to maintain the persistence of the resolution service - either by the organisation hosting the digital resource, a trusted third party or a combination of the two. Each scheme has its own advantages and constraints but it is worth considering the following when deciding on a persistent identifier strategy or approach:
>
> *Advantages*
>
> - Critically important in helping to establish the authenticity of a resource.
> - Provides access to a resource even if its location changes.
> - Overcomes the problems caused by the impermanent nature of URLs.
> - Allows interoperability between collections.
>
> *Disadvantages*
>
> - There is no single system accepted by all, though DOIs are very well established and widely deployed.
> - There may be costs to establishing or using a resolver service.
> - Dependence on ongoing maintenance of the permanent identifier system."

[120] DigitalPreservationCoalition (DPC), "Digital Preservation Handbook - Persistent Identifiers", Webarchive copy at: Archive.org, archive timestamp: 2017-02-09 01:41:47, archived URI: http://www.dpconline.org/handbook/technical-solutions-and-tools/persistent-identifiers, Available at http://web.archive.org/web/20170209014147/http://www.dpconline.org/handbook/technical-solutions-and-tools/persistent-identifiers.

For composed digital assets, the persistent identifier must be basis for getting the necessary information to render the asset. In OAIS Reference Model terms, this mean that based on the persistent identifier, it must be possible to produce a DIP (Dissemination Information Package) based on the preserved AIP (Archival Information Package).

The aforementioned 'social contract' covers placement of the responsibility of a digital preservation programme in order to maintain the contents, as well as the relation between content and the persistent identifier.

Policies and strategies – *Standards* are important to address in order to enable continued understandability of the data and thus long-term access of the data, regardless of technological or organisational changes over time.

Use of standards can be related to both contents and metadata. For example, using standardised preservation formats for the contents (as described below), structuring contents and relations to metadata using standards like METS or RDF, using standards like PREMIS and MIX for description of specific metadata, and using standards standardised denotation of specific file formats by using registries like PRONOM.

Policies and strategies - *Preservation formats* are the file formats that are accepted for long-term preservation. Not all file formats are suited for long-time preservation, and the number of existing file formats is increasingly growing. Therefore, it will only be possible to guarantee logical preservation of formats that are monitored and with properties that makes it possible to perform preservation actions like migration or emulation at a later stage.

There are different choices of preservation formats for different institutions e.g. due to the significant properties that are the most valued. However, there are agreement for most of the requirements that should be fulfilled for preservation formats. These are that the formats must be:

- Standardised;
- Well documented;
- Open;
- Easy to understand;

- Widely used;
- Supported by existing tools.

It should be noted that preservation formats used for AIPs are not necessarily the same as the access formats used for DIPs in the dissemination. For example, JPEG is not regarded as a good preservation format (TIFF is) but is often used as an access format.

Policies and strategies - *Sustainability* covers a lot, but the cornerstones in sustainability is dedicated management to digital preservation, manageable costs and continued funding.

Dedicated management to digital preservation may be hard to measure, but a first step is expression of management commitment as part of the policies and strategies. An extra complexity can also arise for bit preservation based on collaboration between independent organisations.

Costs is a crucial factor, as sustainability will depend on the ability to respect budgets for preservation. The costs are also complex both for bit preservation and logical preservation, but some guidelines can be found from the results of the 4C EU project (Collaboration to Clarify the Costs of Curation)[121].

Policies and strategies – *Audits* are recommended in order to ensure quality and sustainability. Without audits, there will be a risk of weak links that are not discovered, such as lacking funding for monitoring technology, which could result in a critical delay in execution of relevant preservation actions.

There do exist different types of auditing standards (like ISO 16363). It is noteworthy, though, that an audit in itself does not necessarily ensure trustworthiness, and there is no real evidence that a formal audit certificate is better than a well-done self-audit. For ISO 16363, both formal audits certificate[122] and self-audits[123] exist.

[121] Description available at http://www.4cproject.eu/.
[122] See e.g. https://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/iso16363.
[123] See e.g. “Trustworthiness: Self-assessment of an Institutional Repository against ISO 16363-2012”, Available at http://www.dlib.org/dlib/march15/houghton/03houghton.html.

3.4. Further work

This draft version (D3.1) of the guidelines on Data Policies Implementation is largely based on the information that has been collected about existing practices, as documented in Appendix II below. The recommendations and guidelines try to encompass the Humanities and Social Sciences as broadly as possible and are to some extent general in nature. It is anticipated that this version of the document will be distributed to data archives and Research Infrastructures across Europe, as well as to the research communities covered by the PARTHENOS project. The work carried out by WP3 will be reviewed by an expert panel from members active in WP2. In addition, we hope that these communities will test and review the guidelines and recommendations with respect to actual practice and provide feedback.

Further work is going to incorporate such feedback into the final deliverable by early 2019 (D3.2), so that it will reflect practitioners' views and needs to a higher extent than this first deliverable (D3.1). We will establish dialogue with relevant partners, research communities and Research Infrastructures in order to evaluate the usefulness and relevance of these guidelines for actual implementation. This work is expected to follow three paths:

- The general recommendations and guidelines will be revised and refined in order to further reflect experience and requirements from actual practice.
- Potential issues regarding implementation of the guidelines will be addressed, this could include filling out gaps being identified by reviewers.
- Possible issues when implementing the guidelines into specific disciplinary practices will be addressed, in time that such issues are being discovered.

4. IPR, open data and open access

4.1. Introduction

This chapter represents the work undergone by Task 3.3, entitled “IPR, Open Data and Open Access”, that aims to investigate and examine the landscape of policies and practice for the management of intellectual property of data and the provision of open access to data and literature, in use by the Humanities and Social Sciences.

Task 3.3 worked in two stages: the first collecting and investigating operational information on tacit existing policies adopted in everyday practice by research communities involved in PARTHENOS (see Appendix II: Matrix ‘Roles, Tasks, Quality’) and analysing the requirements gathered in D2.1, that reviewed reports and similar documents containing requirements. In the second step, Task 3.3 identified commonalities and gaps to deliver common principles and practical guidelines for managing IPR, Open Data and Open Access. Then each principle was mapped into to the corresponding FAIR principle, adopted as a framework in D3.1.

The PARTHENOS Project identified clearly the need of researchers to work with large amounts of data that have copyright conditions presented in a clear way. Open data and open access are a challenge for a better research environment to promote innovation, development and to connect researchers from across disciplinary and countries. Nevertheless, there is a need expressed by the research communities, in the IPR field to manage restricted access to protected resources by users. Limitations for re-using data is generally due to personal data protection, copyright issues, database rights expressed by national laws and regulations.

The PARTHENOS analysis demonstrates that there are overarching issues that inhibit the diffusion of open data and open access in the research practices: lack of knowledge and guidance about legal issues concerning research data generally, lack of policies and recommendations for open data and open access for some research communities, lack of a shared and clear framework of licences and difficulties in applying the PSI directive correctly when data has commercial value or can be aggregated into works of value.

PARTHENOS’ common goal is to support the ability of the research communities to share, access, and reuse data, as well as to integrate data from

diverse sources for research, education, and other purposes. This requires effective technical, syntactic, semantic, and legal interoperability rules and practices.

These guidelines are presenting high-level recommendations which will help research funders, infrastructure managers, research and cultural institutions and researchers for all the disciplines in consideration by PARTHENOS in furthering the goal of open data and open access in their organization and network and establish a harmonized policy for sharing and reuse data.

4.2. How we collected the information

The first step was designing a table where the researchers involved in Task 3.3 would be able to record relevant policies related to IPR, open data and open access in use by their own disciplines. We adopted a spreadsheet in Google Drive for allowing collaborative work among the researchers.

The table, which we called “Matrix on IPR, Open data and Open Access” (see image below) is organized as follows: in the left column, the main tasks related to implicit policy and implicit procedural activities on IPR management, open data, ethical aspect and privacy issues, usage restrictions, open access. In the top row, the name of the researchers that represent one of the four communities identified by PARTHENOS: History in a broader sense, Language Studies, Cultural Heritage, Applied Disciplines and Archaeology, and Social Sciences in a broader sense. Each tab in the Matrix corresponds to one of these PARTHENOS disciplines.

The information that each researcher had to provide in this table was:

Regarding IPR management: “Outline the relevance of IPR management in your policy or procedural activity in relation to:

- Identification of IPR status;
- Getting permissions;
- IPR policy statements;
- Licensing framework;
- Orphan works and out of commerce;
- Good practice.

Regarding Open data: “Outline the relevance of Open Data in your policy or procedural activity in relation to:

- Definition of a minimum set of data;
- Content reuse (images, text, video, audio, etcetera);
- Adopted standards;
- Good practice;
- Other (e.g.: data citation).

Regarding Ethical aspects and privacy issues: “Outline the relevance of Ethical Aspects in your policy or procedural activity in relation to:

- Procedures for identification of ethical aspects;
- Protection of sensitive personal data;
- Data processing and Big Data;
- Good practice.

Regarding usage restrictions: “Outline the relevance of Open Access in your policy or procedural activity in relation to:

- Methods and procedures;
- Relation with publishers;
- Business model;
- Good practice.

Regarding open access: “Outline the relevance of Open Access in your policy or procedural activity in relation to:

- Methods and procedures;
- Relation with publishers;
- Business model;
- Good practice.

Table 4.1 shows a screenshot of the working document for the discipline Archaeology, with related policies on IPR, Open Data and Open Access.

RESEARCHER - ARCHAEOLOGY		
Task	**Policy/complicit procedural activity**	**Description**
IPR management	**Outline the relevance of IPR management in your policy or procedural activity**	The E-Depot Dutch Archaeology being integrated in DANS makes use of mandatory deposit licences and conditions of use
	identification of IPR status	Deposit licences (various access categories) / Conditions of Use
	getting permissions	For Restricted Access data necessary
	IPR policy statements	https://dans.knaw.nl/en/about/organisation-and-policy/legal-information
	licensing framework	Legislation of the Netherlands on IPR/ Personal Data, Codes of Conduct for Academic Research and Open Access Initiative
	orphan works and out of commerce	No policy yet
	good practice	Open if possible, Restricted if obligatory
	other	
OPEN DATA	**Outline the relevance of Open Data in your policy or procedural activity**	
	definition of minimum set of data	Dublin Core metadata
	content reuse (images, text, video, audio, etcetera)	
	adopted standard	OAI-PMH for the repository
	good practice	
	other (e.g.: data citation)	Data citation always obligatory (Conditions of Use)
ETHICAL ASPECTS AND PRIVACY ISSUES	**Outline the relevance of Ethical Aspects in your policy or procedural activity**	
	procedures for identification of ethical aspects	Only for personal data, not on other ethical aspects
	protection of sensitive personal data	Privacy regulation: https://dans.knaw.nl/en/about/organisation-and-policy/legal-information/DANSprivacyreglementNL.pdf
	data processing and Big Data	Not specific
	good practice	
	other	

USAGE RESTRICTIONS	**Outline the relevance of Usage Restrictons in your policy or procedural activity**	
	Single Sign On (users and roles)	Yes (SURF Conext)
	AAI Infrastructure (EDUGain Federation)	Not yet
	good practice	
	other	
OPEN ACCESS	**Outline the relevance of Open Access in your policy or procedural activity**	
	methods and procedures	Essential element in licences
	relation with publishers	N/A
	business model	N/A
	good practice	See under IPR Management - good practice
	other	

Table 4.1: Inventory of policies on IPR, Open Data and Open Access in the field of Archaeology.

The consultation of PARTHENOS research communities on data reuse showed that research and cultural institutions can share research data in two main ways:

- Make the data available through open (meta)data and open access modality
- Allow restricted access to the data for protection of legitimate interests of the rights holders, for protection of confidentiality and for protection of cultural resources, as determined by law through the restriction or the control of the use of such data.

The general vision for publishing research data is 'Open if possible, Restricted if obligatory'.

4.3. Legal framework

During this last decade, the data from the research, from digitised literature and archives, from Archaeology and other disciplines applied to culture heritage and Humanities has created new possibilities to share knowledge, to carry out research and to develop and implement public policies. It is clear that much of the value of public research data lies in its wide dissemination and reuse, particularly in digital networks and e-Infrastructures. Policies for reuse and data sharing are supported by European and national agencies to improve research and education outcomes, enhance economic returns, promote social integration goals, or support innovative models for consuming and producing culture. Public research data has public good characteristics, and is often global public goods.[124]

Data infrastructures, which store and manage data, promote an easier data exploitation of this data across global markets and borders, and among institutions and research disciplines, thanks to interoperability and access services. A key part of this process is the change in the way scientific research is carried out, as we move rapidly towards Open Science.

Over the last decade there has been a body of literature statements, declarations, and principles in support of open access and reuse of data by various research organizations and disciplines, including the broader research community (Science

[124] Stiglitz, Joseph E., 1999, Knowledge as a Global Public Good. In GLOBAL PUBLIC GOODS. Inge Kaul, Isabelle Grunberg, and Marc Stern, available at http://web.undp.org/globalpublicgoods/TheBook/globalpublicgoods.pdf.

International 2015; RECODE 2015; LIBER 2014; CODATA PASTD 2014; Denton Declaration 2012)[125], international governmental research-related organizations (G8 2013; OECD 2007)[126] and many national governments and their agencies.

However, the ability to access and reuse data is compromised because there is a lack of clarity about the legal conditions under which the data can be reused and when restrictions are provided on the reuse of data. In most cases, the restrictions related to data reuse of collide with the obligation to make public research data widely available.

Restrictions may inhibit the reuse of data to a greater extent than originally assumed. In fact, for a set of data derived from the result of the combination of parts of two or more of other data sets, the most restrictive conditions of the underlying datasets will be transferred to the whole derivative dataset. Therefore, the legal restrictions sometimes unnecessarily imposed, can have widely and undesirable effects that limit the reuse of derived dataset in which most of the data may otherwise be in the public domain or other open licences.

Appendix IV: EU and national regulations to promote access and data reuse summarizes the policies collected by the partners. Project partners are in general agreement with the free circulation and reuse of data and usually provide an adequate licensing framework. However, from a regulation point of view, their main focus is related to open access. Analysing the results of the survey, in fact, there are many agreements that public institutions signed over the years to promote the free circulation of scientific publications.

This policy responds to some needs expressed by the institutions themselves: to ensure bigger visibility to the work carried out within the institution and improve the quality of the publications thanks to a peer-review process.

---

[125] Science International, 2015, Accord on Open Data in a Big Data World: http://www.icsu.org/science-international/accord; Uhlir, Paul F., 2015, "The Value of Open Data Sharing", CODATA report for the Group on Earth Observations: http://zenodo.org/record/33830 - .VwZfUYfmrIU; LIBER, Association of European Research Libraries, 2015, The Hague Declaration on Knowledge Discovery in the Digital Age: http://thehaguedeclaration.com/the-hague-declaration-on-knowledge-discovery-in-the-digital-age/; RECODE Project, 2015, Policy Guidelines for Open Access and Data Dissemination and Preservation. European Commission: http://recodeproject.eu/wp-content/uploads/2015/02/RECODE-D5.1-POLICY-RECOMMENDATIONS-_FINAL.pdf; Denton Declaration, 2012, Open Access Conference. Available at: https://openaccess.unt.edu/denton-declaration.

[126] G8, 2013, Open Data Charter (2013): https://www.gov.uk/government/publications/open-data-charter; Organization for Economic Co-operation and Development (OECD), 2007, OECD Principles and Guidelines for Access to Research Data from Public Funding: http://www.oecd.org/sti/sci-tech/38500813.pdf.

About open data, it is interesting that the institutions consider this as a part of open access and not as an independent field of research.

4.3.1. Intellectual property rights

Intellectual property rights (IPR) management is an important part of all data management plans and includes all the different aspects which allow researchers to access and reuse data / comparing to the national and international rules.

Intellectual Property Rights (IPR) can be described as rights acquired over any work created or invented with the intellectual effort of an individual: inventions; literary and artistic works; images and symbols, as well as discoveries, words, phrases, symbols, and designs. Intellectual property is divided into three categories:

- Industrial Property includes patents for inventions, trademarks, industrial designs and geographical indications, integrated circuits and design layouts and confidential information (trade secrets).
- Copyright involves a wide range of creative, intellectual, or artistic forms, or "works", literary works, films, music, artistic works and architectural design. This also deal with the creation of research data and plays a role when creating, sharing and re-using data. It is important to remark that copyright doesn’t impose any restrictions on the sharing of facts and ideas, procedures, methods of operation or mathematical concepts, which are part of the public domain, i.e. that it only applies to a physical manifestation.
- Database protection rights address the investment that is made in developing a database, even when this does not involve any creative aspect. In fact, selecting what data is included in a database or how to organize the data, are all creative decisions that may receive copyright protection.

Research data refers to information, in particular, facts or numbers collected to be examined and considered as a basis for interpretation, discussion, or calculation. Examples of research data are statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and

images. This includes research data underlying publications and/or other data such as curated but unpublished datasets or raw data.

In the framework of intellectual property laws, owners have certain exclusive rights, such as the possibility to publish in various markets, licence the manufacture and distribution of inventions, data exploitation and to sue in case of unlawful copying.

Generally, in most EU countries, the author or co-authors of a work are the first owners or co-owners of the copyright. Copyright law is different among different countries, but thanks to the rules deriving from international treaties and European legislation, most of the countries have similar rules about what is protected or not by copyright. Therefore, there is no need to ask permission to use a work or a resource if:

- … the work is no longer protected by copyright because the author died more than 70 years ago (in most countries the time period is 70 years after the death of the author. For some countries, it is 50 years after the death of the author). When this time period has expired, the work is said to be in the public domain.
- … the copyright belongs to you or your organisation. In some countries copyright laws provide that the first owner of copyright in a work is the author. Other countries state that this is the case except where the author is an employee acting in the course of employment.
- … some countries have provisions for orphan works. This is where the right holder cannot be found, then it is possible to use the work for certain purposes. However, it is necessary to conduct a diligent search for the rightsholder(s) to make sure that none could be identified or located.
- … there are measures in national legislation which provide that somebody can use a work protected by copyright for specific purposes, for example for study.

Some examples of the types of works protected by copyright for the life-time of the author plus 70 years in Member States of the EU are: literary Works including books, journals, emails, blogs, letters, newspaper clippings, song lyrics, musical works (recorded original musical work) including classical and popular music and

performing works, such as written scripts used for concerts and plays, films including recordings on any medium from which a moving image may be produced, artistic works including paintings, drawings, engravings, sculptures, photographs, greeting cards, postcards, diagrams, maps, works of architecture, hand-crafted works, medals. In the Member States of the EU, if a work falls into the following category it will be protected by copyright for 50 years from the end of the year in which it was made, or 50 years from the date it was first made publicly available: sound recordings including Oral History, sound effects, recorded lectures, recordings of literary, dramatic or musical works.

In some Member States of the EU, if the work falls into the following category it will be protected by copyright for 50 years from the end of the year of the making of the broadcast: broadcasts including the electronic transmission of visual images, sounds and other information such as streaming from website, TV.

However, the IP systems in the EU currently vary between Member States which maintain a system of institutional ownership, and those which maintain a system of professor's privilege (inventor ownership). In fact, in many countries, when a work is made by an employee during of his/her work, the employer will be the first owner of copyright of the work created under a contractual relationship.

Some countries of the EU, for example Italy or Sweden, have a specific form of "professor's privilege" regime according to which the researchers, PhD students, etcetera, are entitled by law to the ownership of the work they created in the course of their employment. Results of publicly-funded research, created or developed by researchers, does not belong to the academic institution but to the researchers.

The rights give to the owner exclusive economic rights for certain period of time to copy the work, issue copies to the public and to make an adaptation of it. The author also has moral rights concerning the right of integrity and of attribution being. A researcher can decide if he wants to share their data with others, since the benefits are so well known in order to promote research integrity and collaborative opportunities.

Scholars and researchers that want to reuse and share data and content need to know the terms of use for the database and the data content. What are the legal rights in data, who has these rights, under which conditions may use it and how does rights holder use the rights to share data in a way that allow productive and successive uses. Moreover, a researcher who wants to enrich data in their work with

data provided in part by others wants to be sure that any legal, ethical, and professional obligations that one may have to the provider of the data are met. IPR depends on national law, a allowing the users' rights to be modified by each country. This context defines a form of legal uncertainty that is serious impediment to the productive reuse of research data. It can be avoided if the repository managed by the research centre requires depositors to grant permission to downstream users or to give up any intellectual property rights they may have in the data. In order to solve this problem, international initiatives have been set up, such as licensing frameworks like the Creative Commons and Rights Statements.

Actually, there is an ongoing consultation for updating copyright rules at EU level in order to answer to the new challenges offered in the digital age. The European Commission has presented legislative proposals to guarantee a more cross-border access to content online[127], wider opportunities to use copyrighted materials in education, research and cultural heritage[128], a better functioning copyright marketplace. The objective is to support copyright industries to increase in a Digital Single Market[129] and European authors to reach new audiences, while making European works widely accessible to European citizens, also across borders. In the proposal, the EU wants to set a good balance between copyright and relevant public policy objectives such as education, research, innovation.

4.3.1.1. Case study: IPR management in the CENDARI project

One of the key tasks of the CENDARI project (www.cendari.eu) was to federate a large corpus of highly heterogeneous data and metadata from a range of over 1,200 institutions. For some of these institutions, data could be accessed via an aggregator, such as Europeana, which offers an open API for data sharing. In other cases, individual institutional data was either delivered via a file transfer or had to be created or curated by hand by the project researchers.

This landscape of partners, formats and datatypes resulted in an exceptionally complex IPR situation with many different licence types and restrictions already governing the data coming in which had to be preserved going out. In

127 https://ec.europa.eu/digital-single-market/en/modernisation-eu-copyright-rules#choiceandaccess.
128 https://ec.europa.eu/digital-single-market/en/modernisation-eu-copyright-rules#improvedrules.
129 https://ec.europa.eu/digital-single-market/en/digital-single-market.

addition, there were often competing voices and positions among the many communities and institutions the project was dealing with.

The approach taken by CENDARI was to work within the standard and recognised Creative Commons licensing system, which was applied as follows: a CC-BY licence was applied by default to all data in the system. This was in step with the Archives Portal Europe, a key partner in recruiting data, as well as with the DARIAH ERIC, the project's umbrella infrastructure. Data coming from Europeana, however, had to be flagged as reusable under the same licence it was acquired under, in most cases CC-0. Individual institutions contributing under CC-BY were also given the option to use CC-0, in particular for metadata that did not appear in the Europeana ecosystem, to facilitate its later presentation there. This exemption enabled sharing between CENDARI and Europeana in two directions, to the benefit of smaller partner institutions.

Finally, in some cases, specific licences were requested by institutions, such as the addition of an NC-SA clause for one particular US-based institution. This flexibility allowed the project to recruit data that might not have been available if a narrower approach to rights management had been applied. This did create additional system complexity, however, as metadata outlining the rights under which a specific dataset had been acquired and could be reused had to be applied at a far finer level of granularity.

4.3.2. Sensitive data

There are various definitions of sensitive data. Generally speaking, sensitive data is considered to be data that needs a high grade of protection. In most cases this is personal data. In particular, one should think of personal data identifying someone by his or her health, religion, political conviction, race, sexual orientation or personal identification number. This information is of a potentially very sensitive nature. Legally this kind of data is often defined as "special" personal data. Non-special or common personal data contains more elementary data such as name, address or telephone number. Special personal data is submitted to a stricter protection regime than the latter group, the common personal data. Besides personal data, there are other possible categories of sensitive data, such as secret data on state security and confidential business data.

There is a further category of highly sensitive data that needs even greater protection. In particular, the personal data of people who have witnessed or have been involved in circumstances such as (past) wars, armed conflicts, or have had medical or psychiatric treatment, or have handicaps (especially for children). Disclosure poses a risk for such people who would be very vulnerable without strong legal protection.

As a general rule, all this sensitive data need protection. Personal data is regulated in the current European data protection laws and from 2018, this will be regulated by a common European law, the GDPR – General Data Protection Regulation, which is focused on preventing the identification of living persons. This protection is required in order to hide the identity of the respondent / test subject.

4.3.2.1. Case study: open metadata for sensitive data

This use case describes how DANS handles sensitive data.

Handling sensitive data

How then does DANS, as a research data repository, handle sensitive data? All research data at DANS is stored in and made available by its online repository EASY: www.easy.dans.knaw.nl. A licence agreement is always agreed between DANS and the depositor of the dataset: the person or organisation depositing a dataset in EASY who is normally the rights holder. One of the most important parts of this licence agreement is the *access category* by which the access to the dataset can be specified.

DANS supports the *Open Access* movement. This means that DANS encourages research data and publications to be made freely available as much as possible, without any restrictions. However, substantiated reasons exist why research data is not, or not immediately, freely accessible. This can be due to the presence of personal data or a temporary embargo on data due to an impending Ph.D. thesis or other publication, contract obligations with third parties, etcetera, DANS therefore, provides along with open access, the possibility of restricted access to research data.

EASY offers two Open Access categories and one Restricted Access category.

The access categories are:

· *Open Access (CC0 Waiver)*

The dataset is, without any restriction, made available to all EASY users, both registered and unregistered, in accordance with the conditions of the *Creative Commons Zero Waiver.*

· *Open Access for Registered Users*

The dataset is only made available to all *registered* EASY users. Any existing copyrights and/or database rights are respected.

· *Restricted Access*

The dataset is only made available to those registered users that have obtained permission from the rights holder.

Datasets containing personal data are mostly placed in the category Restricted Access. Some datasets with personal data are made available in the Open Access categories. This is, however, only possible when explicit informed consent has been given by the persons involved. This is quite often the case with Oral History interviews. Besides from this open category, sensitive data can only be accessed by authorised users whose identities have been checked and who may be required to also sign special, additional, conditions of use.

Metadata

Metadata, being the information about the data, is always freely accessible in EASY. That means that no registration is needed for searching in the metadata, or harvesting it. In other words, all the information in the metadata is open for everyone. That means that the metadata of sensitive datasets can never contain confidential or identifying elements or characteristics, like names. When someone who is looking for data finds a description of a dataset containing sensitive data, he or she can submit a permission request to the rights holder for getting access to the data. If this is granted, possibly after further additional conditions are met, the dataset will be

available to download by this user. Even then the use is restricted, as the user is not allowed to make public the confidential data in this dataset. It is only permitted to refer to the data in an anonymised way as individual people should never be identified.

By operating in the way described here, DANS operates as a Trusted Digital Repository (TDR). It is effectively certified as such, both by DSA and nestor-seal. DANS has an infrastructure that supports both the security and the legal "storage and access" policies on sensitive data. This means that the data depositors, the data users and the repository staff all have to be aware of the rules and the risks. Securing confidentiality is as important to the DANS staff as to the data depositors or users, and for this reason, confidentiality declarations have been made mandatory for the staff at DANS.

4.3.3. PSI Directive

The European legislation on reuse of public sector information is a Directive which promotes the free circulation and reuse of data produced by public institutions without restrictions. The Directive is based on the principle that these resources should be free because the citizens have already paid (through taxes) for them and they should not pay twice for the same service or information. Moreover, after the original Directive 2003/98/EU[1][130], it was demonstrated that the creative industry obtained economic benefits from the free reuse of public data.[131]

However, this Directive, which was to be implemented by the Member States by 2015, has had some deviations in its application, mainly for two reasons:

1) Member States have their own policy on data reuse, usually produced before the Directive was implemented.
2) Several exceptions were established by the Directive itself.

[130] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:345:0090:0096:en:PDF.
[131] "Should inter alia allow European companies to exploit its potential and contribute to economic growth and job creation.." http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02003L0098-20130717&from=EN.

In fact, although the Directive was adopted by all EU countries, it is possible to identify three different ways to implement it:[132]

- Adoption of specific PSI reuse measures.
- Combination of new measures specifically addressing reuse and legislation pre-dating the Directive.
- Adaption of existing legislative framework to include the PSI.

This means that the Directive has been, in many cases, adapted (and/or amended) to an existing regulatory framework. At the same time, local public institutions each have their own interpretation of the Directive, based on their previous experiences. This has resulted in a heterogeneous application of the Directive producing, for examples, different levels of access for similar data produced by different Member States. At the same time, the presence of many exceptions (derogations) to the Directive has made its uniform interpretation and application ever more complex. For example, if data contains sensitive information, the IPR is still valid or the data relates to national security, it must not be shared according to the PSI Directive.

Moreover, the PSI Directive lays down an exception for data which, while falling within public domain, is sold by public institutions in order to support their activities. While stating that public institutions should charge only the marginal costs, the PSI Directive contains an exception for Archives, Libraries and Museums which allows the sale of data to generate an income. In a similar manner, also private–public partnerships are not covered by the PSI Directive.

Law makers recognised the difficulties in applying the PSI Directive in the European context and for this reason they considered it to be a sort of a minimum common regulatory framework. From this point of view, it is relevant that in the recommendations produced by PARTHENOS, great relevance has been assigned to the adoption of a standard licensing framework:

> *"In Member States where licences are used, Member States shall ensure that standard licences for the reuse of public sector documents, which can be adapted to meet particular licence applications, are available in digital format and can be processed*

[132] https://ec.europa.eu/digital-single-market/en/implementation-public-sector-information-directive-member-states.

electronically. Member States shall encourage all public sector bodies to use the standard licences."[4]

So, considering the approach of the institutions involved in PARTHENOS and the principles expressed by the PSI Directive, it is possible to define the following recommendations:

1) (Meta)data should be open as possible and only closed when necessary.
2) Protected data and personal data must be available through a controlled procedure.
3) (Meta)data rights should communicate the copyright and reuse transparently, clearly and be machine readable.

4.3.4. Open Access and Open Data

In the Budapest Declaration (2002)[133] and the Berlin Declaration (2003)[134] it is possible to find a clear definition of Open Access. The Declarations describe 'access' in the context of open access as including not only basic elements such as the right to read, download and print, but also the right to copy, distribute, search, link, crawl, and mine information. Open access (OA), in fact, can be defined as the practice of providing online access to scientific information that is free of charge to the user and that is re-usable. Generally, a distinction is made between OA to scientific peer reviewed and non-peer-reviewed academic publications, such as journal articles, conference papers, theses, book chapters, monographs, and research data. Two main models are emerging for open access to publications:

- Self-archiving (also referred to as 'green' open access) means that the published article or the final peer-reviewed manuscript is archived (deposited) by the author - or a representative - in an online repository before, alongside or after its publication. Repository software usually allows authors to delay access to the article ('embargo period'). Some publishers

[133] Budapest Declaration on World Heritage (2002) https://ec.europa.eu/digital-single-market/en/implementation-public-sector-information-directive-member-states.
[134] The "Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities" (2003) https://openaccess.mpg.de/Berlin-Declaration.

request embargo periods, arguing that these protect the value of the journal subscriptions they sell.

- Open access publishing (also referred to as 'gold' open access) means that an article is immediately provided in open access mode when published. In this model, the payment of publication costs is shifted away from readers (paying via subscriptions) to the author, also called 'Article (sometimes 'Author') Processing Charge' (APCs). These can usually be borne by the university or research institute to which the researcher is affiliated, or to the funding agency supporting the research. In other cases, the costs of open access publishing are covered by subsidies or other funding models.

The European Commission has developed the implementation of OA policies now spreading across Europe. The Horizon 2020 Programme provides requirements and guidelines[135] for guaranteeing Open Access to Scientific Publications and Research Data produced by funded projects. According to the Commission, there is no need to pay for information funded from public investment when it is accessed or used by researchers, innovative industries and the public, while it is important to preserve this information over the long term. This policy will increase the benefits to both European businesses and public knowledge.

4.3.4.1. Case study: Open Access

The transition from a print edition to a digital open access publication: the case of the journal *Lexicon Philosophicum.*

Lexicon Philosophicum. International Journal for the History of Texts and Ideas, http://www.lexicon.cnr.it, is an international Open Access electronic journal published by the Istituto per il Lessico Intellettuale Europeo e Storia delle Idee (ILIESI-CNR). The journal is the outgrowth of a previous traditional journal published on paper: since 1985 the ILIESI has published twelve volumes in the form of 'Cahier' under the

[135] H2020 Programme Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf.

same name *Lexicon Philosophicum*, appearing in the series “Lessico Intellettuale Europeo” published in Florence by Olschki.

The current *Lexicon Philosophicum* is an annual, open peer-reviewed, open access journal, with an interdisciplinary character. The journal provides open access to original, unpublished high quality contributions: critical essays, research articles, short texts editions, and critical bibliographic reviews on the history of philosophy, the history of science, and the history of ideas, with a special attention to textual and lexical data.

The new journal has been created within the activities of the European project Agora Scholarly Open Access Research in European Philosophy (2011-2014). The journal has been part of an evaluation experiment (see below) for which the goal was to determine and enhance standards in the field of open collaborative peer review in the Humanities and Social Sciences

The journal articles can be interlinked with a large collection of primary sources of Ancient and Early Modern Philosophy available in the portal Daphnet (http://www.daphnet.org/) and with the selected contributions contained in the Daphnet Digital Library platform (http://scholarlysource.daphnet.org/index.php/DDL). ILIESI always ask permissions for online publications in these platforms: a) in case of explicit authorization CC-BY-NC-SA is used; b) if the authorization for reuse is not present, “all rights reserved” is applied; c) if authorization is difficult to ask for, CC-BY-NC-SA (silence means consent) is used.

Adopting the Open Journal System (OJS), the journal adheres to the open access protocols to improve the quality and the dissemination of scholarly publishing in the field of philosophy. OJS is a journal management and publishing system that has been developed by the Public Knowledge Project. Its main features are: 1) It is installed and controlled locally. 2) Editors configure the requirements, sections, review process, etcetera. 3) There is online submission and management of all content. 4) A subscription module with delayed open access options. 5) Comprehensive indexing of content part of global system. 6) Reading Tools for content, based on field and editors’ choice. 7) Email notification and commenting ability for readers. 8. LOCKSS system to create a distributed archiving system among participating libraries and which permits those libraries to create permanent archives of the journal for purposes of preservation and restoration.

Lexicon Philosophicum uses DOI to guarantee the URL stability of its documents. *Lexicon Philosophicum* provides immediate open access to its content on the principle that making research freely available to the public supports a greater global exchange of knowledge. The contributions published in the journal are made available in Open Access under the Creative Commons General Public Licence Attribution, Non Commercial, Share-Alike version 3.0 (CCPL BY-NC-SA). Such a licence, while granting the paternity and integrity to the original author(s), permits public and unrestricted access to the works, their use, copy, reproduction, and redistribution, provided that such uses are not commercial. It also allows the creation of derivative works (such as translations and adaptations), provided that the derivative works are distributed under the same licence as the original works.

Open peer review experiment: *Lexicon Philosophicum* and *Nordic Wittgenstein Review* (NWR; www.nordicwittgensteinreview.com) took part in an Open Review experiment, in which double-blind peer review was supplemented with a session of Open Review or Preview online of the submitted articles accepted for publication for one month during which registered users were asked to comment on and discuss the accepted papers. Discussions were moderated by the editors and editor-in-chief.

Open access to research data, also known as open data, are the pillars of a modern research methodology, based on cooperative work and new ways of knowledge distribution using digital technologies. This new approach promotes the data sharing and a dynamic exchange of ideas and research results, improving the scientific research (through improved reproducibility), and accelerating also innovation. Compared to traditional research methods, where publishing and patenting are more relevant than collaboration and sharing, Open Science supports joint effort and sharing results in order to involve broader communities and face up to global challenges. So, effectively, no knowledge or any discovery is completely "owned", but rather is shared to benefit all society.

The May 2016 Council Conclusions on 'the transition towards an Open Science system[136]' call will be adopted by 2020 for a transition to open access to scientific peer reviewed publications and for the re-using of data. The principle is "as

[136] http://data.consilium.europa.eu/doc/document/ST-9526-2016-INIT/en/pdf.

open as possible as closed as necessary" and it welcomes the intention of the Commission to make research data produced by Horizon 2020 open by default, whilst recognizing the right of opting out.

The Open Science Agenda[137] defines by 2020 two high-level aims for OA: all peer reviewed scientific publications are freely accessible and FAIR data sharing is the default for scientific research.

While open access to scientific publications is growing, and increasing in terms of use, open access to research data is only recently beginning to be known. Although some disciplines have a culture of sharing research data, most researchers are hesitant to make their research data publicly available. This is mainly because there is the fear that it will facilitate competitors before publishing their data, improper or lack of attribution, or the risk of possible errors in the data or the analysis.

The Open Science practices can enhance a researcher's career through more citations and opportunities for cooperation. Within the research activities, journal citations assume a relevant importance while other outputs are generally not considered in the assessments of research impact. However, some forms of research outcomes can be precious. For example, negative results are commonly not published in journal articles, but making this data publicly available and citable would reduce the costs of duplicating failed experiments, allowing researchers to receive credit for their work. Sharing research data or preliminary analysis before publication would offer the researchers the opportunity to receive feedback and improve their work. Due to low quality, many studies are not reproducible but it is important to underline that reproducibility is a fundamental principle in scientific research. Only if researchers can replicate their results, it is possible to validate the research methods but it is difficult to reproduce research results based only on unclear methodology sections in journal articles. The reproducibility problem could be solved by sharing research data supporting good scientific conduct. This will provide a greater incentive for researchers because their work would stand up to scrutiny.

Both scientific research and economic growth will receive a considerable boost if research data is open and this will certainly also be significant for the Digital

[137] February 2016. DIRECTORATE-GENERAL FOR RESEARCH AND INNOVATION (RTD). Draft European Open Science Agenda https://ec.europa.eu/research/openscience/pdf/draft_european_open_science_agenda.pdf.

Single Market. The benefits that users receive from making scientific information freely available to the global life science community is widely demonstrated. Therefore, the same should apply to the Humanities.

However, open access is only one part of making data findable, accessible, interoperable and re-usable (FAIR) and, therefore, needs to be addressed in the wider context of 'Open Science'.

Open access is sometimes in conflict with Intellectual Property Rights (IPR). If researchers decide to commercially exploit the results of their research, they may decide to protect their IPR otherwise they should choose the open access route. Member States agree on the relevance of open research data and on policies and actions devoted to promoting the collection, curation, preservation and reuse of research data. Private research organizations wish to obtain concrete support to be able to commercialize innovative products and, on the other hand, researchers also request credit for their work. These aspects need a level of protection in relation to the original idea. In this light, IPR has a fundamental role to incentivize individual efforts and to encourage investment.

It is also important to observe that IPR, that is introduced in the scientific discovery process, may inhibit the collaborative model of Open Science. A correct level and type of protection should allow the right balance between the incentives needed for an initial creation, and the freedom to reuse and improve upon such creation, to be found.

4.3.4.2. Case study: Open Data

Implementing CCO licence on data

Both government and the scientific community increasingly emphasize the importance of open access to publicly funded data. The European Science Foundation and other leading European research funders have declared their support for the “Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities”.

As an early adopter of open access and open data, DANS is implementing this policy in practice. DANS has decided to no longer require registration for users as a

standard. The registration of users is considered as an obstacle to this "open access". Furthermore, computer applications (such as linked data apps, text and data mining) encounter barriers with registration and are not able to query archived data or edit them. By removing this registration requirement, the licence 'Open access for registered users' will change to an open licence, for which DANS uses CC0 Waiver of Creative Commons as the standard. The standard limits the legal and technical barriers for the reuse of data by waiving copyright and neighbouring rights, to the extent permitted by the law. DANS will continue to draw users' attention to the fact that, in accordance with the VSNU/KNAW Code of Conduct for Academic Practice, proper citation of research remains imperative.

The strategic decision was made to make the default setting Open access for everyone in EASY the online archiving system of DANS. The more practical phase of implementing this new standard required the following steps to:

- Update the DANS Licence agreement by including CC0.
- Update the guidelines on data depositing.
- Update the help texts in the archiving system EASY: the dataset files are accessible to all users of EASY and '**CC0 Waiver - No Rights Reserved**' applies. For more information please visit https://creativecommons.org/about/cc0.

In this category, all possible rights (such as copyrights and database rights) on the dataset files have been waived. Other actions undertaken by DANS are:
- Communication and disseminating activities by promoting them in DataLink, the newsletter of DANS, on the DANS website and in mailings.
- Contacting researchers who deposited their data in previous years to enquire if they objected to transforming their data to a CCO licence. If depositors didn't agree they could opt out by choosing a more restricted category. At the start, 405 (non-archaeological) depositors of one or more datasets were willing to change their data into CC0. This was followed by a number of archaeological organizations which agreed to change a collection of thousands of archaeological datasets into Open Access.

Implementing this change by software developers in the archiving system EASY by transferring thousands of datasets towards the status ‘Open for Everyone’ is still in progress. Not only the change of category but also the update of the old licence related to the archived data is needed.

The open access movement is an ongoing process and DANS likes to share its experiences on this. A related document is the IPR report comparing licences and access at Europeana and DANS (Heiko Tjalsma) which was presented at the PARTHENOS Workshop in Rome in November 2016.

4.3.5. Licensing frameworks

A licence is legal document that the rights holder applies to his (or her) work or resource that gives permission to do something with the work/resource. The licensing framework is an essential “tool” for anyone wishing to present their data to a wider audience, in order to make clear how to use (and reuse) it.

Of course, it is possible for each institution to define its own licensing framework (based on specific issues), but over the years some standard models have proved to be greatly successful, with a wide application. These standard models, in fact, provide several advantages. In particular, they clearly establish what the users can do and can't do with data; moreover, the licences are also machine readable, so a web browser or a computer system can read them automatically.

For this reason, the adoption of a standard licensing framework is a crucial aspect for each RI, keeping in mind the needs and the concerns of different members.

The Rights Statements and Creative Commons can be considered as two reference models for PARTHENOS community: they have all the characteristics required and have already been widely adopted in the field of Cultural Heritage Institutions.

Moreover, a survey carried out along the project partners has demonstrated that these two licensing frameworks are easy to map compared to others adopted by project partners.

4.3.6. Rights Statements (RightsStatements.org)

RightsStatements[138] is a project that was born from the collaboration of Europeana (the European Culture Portal) and the Digital Library of America and was a response to the increasingly important need for a licensing framework able to cover, in a clear and simple manner, the rights related to objects shared by Cultural Heritage Institutions. For this reason, the working group created twelve statements that "can be used by Cultural Heritage Institutions to communicate the copyright and reuse status of digital objects to the public".[139]

Because the RightsStatements licensing framework was developed very recently, it has the advantage of being able to seen as complementary and not a substitute for Creative Commons. In fact, it was developed especially for the resources that a) are still in rights, b) fall in the no copyrights area but still have some restrictions to their reuse or c) have an uncertain attribution. For this reason, it is divided in three main categories: in copyright, no copyright and other.

The five "in copyright" statements allow the reuse of resources for educational and non-commercial purposes and cover two particular cases: EU orphan works and rights-holder(s) that can't be identified or located.

The four "out of copyrights" statements, instead, focus on the resources that, although they are no longer in copyright, still have some restrictions that prevent their free reuse or whose rights have been ascertained only for a specific jurisdiction.

The last section, "other", is devoted to unclear rights statements and probably is the most critical to assign. These rights statements should be used only if is not possible to define a clearer rights statement or licence. A typical example is represented by "no known copyright", used only for the resources "for which the copyright status has not been determined conclusively, but for which the data provider has reasonable cause to believe that the work is not covered by copyright or related rights anymore".[140]

138 http://rightsstatements.org/en/.
139 http://rightsstatements.org/en/about.html.
140 http://rightsstatements.org/page/1.0/?language=en.

4.3.7. Creative Commons

The standard model with the widest application, until today, is the Creative Commons. This licensing framework, in fact, provides different levels of data sharing, being able, in this way, to cover a very large series of scenarios.

The Creative Commons was born in 2001 and it was partially inspired by the Free Software Foundation. The creators wanted to help those who wished to share their "works freely for certain uses, on certain conditions; or dedicate your works to the public domain"[141]. It offers a framework of standardized licences, some of which apply to data and databases. The adoption of this licence was very successful and it is now used by over one billion resources worldwide. Over the years, the Creative Commons group has also paid great attention to Science, Education and Global Infrastructures, trying to resolve some of the most common issues present in these fields.

The current development of the Creative Commons foresees three different levels to share data: resources available under the public domain, resources considered *free culture* and resources that are *not free culture*. The resources which fall under the public domain are the ones that, for different reasons, do not have any kind of limit to their reuse. It is possible, in fact, to apply the public domain licence in two cases: if the IPR is expired or the creator has voluntarily surrendered it.

The *free culture* licences, instead, while having the same possibilities of reuse of the public domain, are characterized to maintain some rights. In this case, the licence requires just the attribution to the owner of the resources, and that must be immediately recognizable. However, the *free culture* licences permit third parties to adapt the work, and also commercial reuse.

The *not free culture* licences, consequently, are all the licences that don't satisfy the conditions provided in the other two cases. This means that, despite that they are under the umbrella of Creative Commons licensing framework, the resources have several limitations to their reuse: for example, it is not possible to adapt or derive other works from the original ones and the commercial reuse is not allowed.

141 https://creativecommons.org/about/history/.

4.3.8. Licensing framework in PARTHENOS Community

A survey was carried among the project partners in order to determine if it was possible to map their licences to the Creative Commons and Rights Statements frameworks and to start a discussion on a common shared licensing framework.
The result has been very encouraging. Half of the partners, in fact, already use these licences for their own resources, while those who have a customized licensing framework have had no difficulty in mapping them.

However, the analysis of results was really useful for the development of PARTHENOS licensing framework. First of all, the partners are very willing to open up their data, when possible. In fact, the open licences, or those that allow free reuse of resources, although with slightly different modes, are the most commonly used. It is interesting to underline that, in most cases, the institutions wishing to keep track of the work they have done are requesting the use of licences with attribution. This willingness to open data is demonstrated also by the choice to use Creative Commons licences for resources that, while not allowing commercial reuse, allow free reuse for research and educational purpose.

Sometimes, it was not possible to apply a Creative Commons (or a Creative Commons like) licence, for which partners decided to use instead the rights statement “in copyright”. This choice is particularly important for the PARTHENOS community because it refers to resources that have data protection issues, in line also with the provisions of the PSI Directive (IPR, copyrights, sensible data, etcetera).

From this point of view, the work carried out by CLARIN, a Research Infrastructure for Language Resources and Technology, particularly active in the field of the Humanities and Social Sciences, is relevant. CLARIN has developed a licensing framework that is particularly rich, and that is able to respond to different requirements concerning copyright and/or personal data protection issues. These licences have been grouped into three broad areas: PUB (Public language resources), ACA (Academic language resources) and RES (Restricted language resources). The PUB resources are freely usable by everyone and they have no reuse limitations, not being presently in copyright and/or having data protection issues.

The resources that fall in ACA area, instead, are freely reusable only for research purposes. Even in this case, users do not ask for any kind of permission for the reuse, but they need access to resources via a Federated Identity Service.

Finally, the RES resources have the characteristic, unlike others, to be accessible just for research purposes and only available after having made a request. In this case, after the user has logged in via a federated login, a separate application allows him to send a request to the rights holder to get authorization for reuse of the data.

While PUB licences are in line with Creative Commons and Rights Statements, the ACA and RES licences deserve attention because they cover an area that these two standard licences are not able to cover, or only cover partially. The “in copyright” licence, in fact, does not exclude the public visibility of the resource, which does not apply to the RES resources that contain information that is not freely available.

Since several project partners have the same issue, within the PARTHENOS Community it will be fundamental to develop an authentication process (AAI) that allows profiling of users in the right way. At the same, it would be very useful to provide a separate application, such already happens for CLARIN, in order to facilitate contact between users and rights-holder to avoid inappropriate data reuse.

4.3.8.1. Case study: open licences supporting legal interoperability

CLARIN PUBLIC END-USER LICENSE (CLARIN-EULA-PUB-v.1.0)

Copyright holder:

Resource:

The Copyright holder grants the End-User a free, non-exclusive and perpetual (for the duration of the copyright) right to **use and make copies of the Resource, distribute copies and present the Resource in public** as such, as modified, or as part of a compilation or derived work. The permission applies to all known or future modes and means of communication and includes a right to make modifications enabling the use of the Resource on other devices and in other formats.

Additional license terms as defined in the Terms of Service Agreement:

1. Identification and Access conditions: [ID, -]
2. General Use conditions: [BY, NC, LRT, -]
3. Distribution conditions: [NORED, SA, DEP, ND, -]

This license has been made in compliance with copyright agreements by WIPO – the World Intellectual Property Organization. The rights granted in this license shall be so interpreted that in case applicable intellectual property laws grant rights not mentioned in this license, they are also regarded as part of the rights to be licensed; the purpose of this license is not to restrict any rights intended to be licensed within different legal systems. Additional rights to the Resource may be agreed separately in writing. The full agreement is available in Appendix V: CLARIN deposition license agreement.

The need to have an AAI is still more pressing since the number of resources made available by institutions is increasing rapidly, and consequently so is the risk of personal data issues also growing.

4.4. Authentication and authorization infrastructure

The need to access networked applications and remote/distributed data is evolving very fast with the development of shared virtual environments and authentication and authorization of users is considered a key aspect of digital data infrastructures. Federated access is a particularly desirable in a situation where services are offered across institutions to users that do not belong to the same institution that offers the service.

Authentication and authorisation Infrastructures [142] are based on technology underlying federated access to the research community, where:

- The user's credentials (typically organizational affiliations) are handled by the user's organisation, also called the Identity Provider (IdP).
- The user can log in using the same credentials to different resource

[142] http://www.geant.org/Services/Trust_identity_and_security/edugain.

providers that have agreed to accept those credentials.

This technology, while is widely used in Open Science domains, in the cultural heritage and Humanities sectors is not so well known, except in the CLARIN digital infrastructure, as reported below. To address the need for an AA Infrastructure in the PARTHENOS research communities it is important to clarify basic concepts on which AAI is based on:

- *Authentication*: the process of verifying the identity of an entity, either in person or electronically, where credentials are requested and checked to verify or disprove an entity's claimed identity;
- *AAI*: an infrastructure that supports Authentication and Authorisation Services. The minimum service components would be the management of identities and privileges specific to users or resources;
- *Authorisation*: the assignment of rights and capabilities granted to a specific principal (such as a person). Normally authorisation takes place when a user has been authenticated;
- *Federated AAI:* an AAI that supports multiple Identity and Privilege Providers, trusted by the members of the federation;
- *Service Provider*: or 'SP' is a resource or set of contents available to users via a login. This login may be to limit access to subscribers or specialist groups, or to provide personalisation features. In a federated environment, Service Providers do not hold identity information about users but instead rely on Identity Providers (i.e. the institution or organisation that a user belongs to) for sending relevant information to them.
- *Identity Provider*: or 'IdP' is a term used to describe any institution or organisation that manages information about users and wants to provide access to resources (SP) for these users.
- *A policy or agreement* – that IdPs and SPs sign up to, to agree how to interact with each other. These are typically implemented at a national level.
- *Registration* – a place to sign up and give to a federation information about your IdP or SP - also called your 'entity'.
- *Metadata* – the collected information about entities, brought together in one

place and typically digitally signed by a federation and published to its members.

- *Discovery service* – a tool used by Service Providers to allow users to select their own Identity Provider.

Authentication and authorisation are often separated from the application and the data: authentication of the users is done by the user's Identity Providers while the authorisation is done by the services based on the information received by the Identity Providers.

Federated access provides the technical and policy framework to allow for services to be shared in a trustworthy manner across borders. How authentication is carried out by the institutions and how rights management is carried out by the service provider is left up to the respective parties to decide and arrange. Federated access has advantages for both users and application developers:

- Users will be able to log in once (single sign-in) using their institutional credentials and access multiple services (sign on), Single Sign-On, whilst having the assurance that their personal data will not be disclosed to third parties.
- Researchers, digital cultural curators and cultural institutions participating will be free of the burden of user name and password administration, and will have access to more tools for managing data. For a large number of users this means reduced administration and service provisioning costs; and it avoids duplications of identity stores.
- Collaboration among different parties becomes easier.
- Institutions in a federated context can act both as IdPs and SPs, or they can act as either IdPs or SPs.

The first step to join a federation is to talk to the federation operator in a specific country. The list of existing federations is available online at: https://refeds.org/resources/resources_list.html.

CLARIN has developed a Federated Identity[143] to access protected resources (files, web applications) available to academic users from many EU-countries.
To be part of the Federation, each CLARIN centre signs an agreement, giving the power of attorney to the CLARIN ERIC. The CLARIN ERIC can then sign subsequent agreements with the national Identity Federations involved, to ensure that they give their users access to the CLARIN services. This construction avoids the situation where each CLARIN centre would need to sign an agreement with each Identity Federation. The currently operational Service Providers are listed on line by the Centre Registry.[144]

4.5. Outcome: principles and guidelines

Although these guidelines were produced according to FAIR principles, it is relevant to remember that for IPR, Open Data and Open Access, the Findable principle was not considered, because the point of view of this chapter is oriented to the legal and not to the technical aspects related to these topics.

The goal of this guideline is to provide clear indications for PARTHENOS' research communities, supporting them in the licences assignment and legal interoperability. Legal interoperability happens when:

- Two or more datasets provide the same legal rights, terms and conditions.
- It is possible to combine data by other users without compromising the legal rights of the original sources.

Before reaching the legal interoperability, however, it is necessary to consider carefully different issues that allow content providers to correctly make available their data from a legal point of view.

Moreover, because it is quite impossible to analyse in detail all the aspects of this topic, the following section aims to provide an overview of the legal issues related to data reuse, in order to give a point of reference for all the actors involved

143 https://www.clarin.eu/content/federated-identity.
144 https://centres.clarin.eu/spf.

in this process: policy makers, data consumer, content providers, Research Infrastructures and data managers.

PARTHENOS Principles for common policies on IPR, OD & OA implementation & FAIR

To be Accessible

1) (Meta)data should be open as possible and closed as necessary.
2) Protected data and personal data must be available through a controlled and documented procedure.

To be Interoperable

3) (Meta)data licences framework should support legal interoperability fostering harmonization of rights.

To be Reusable

4) (Meta)data should be licensed to permit the widest reuse possible.
5) (Meta)data rights holder should be identified before data publishing.
6) (Meta)data rights statements should communicate the copyright and reuse status transparently, clearly and machine readable.
7) Specify why and for what period a data embargo is needed (data should be made available as soon as possible).

Guidelines

To be Accessible:

1. (meta)data should be open as possible and closed as necessary

According to policies developed in recent years, especially with regards to public institutions, data produced should be available with the widest open licence in order to encourage reuse without limitations. This requirement is also more relevant if it is considered that the research communities believe that this is the best way to produce quality output.

1A. If a data creation process is carried out with public funding, it should be as open as possible.
Sometimes it is not possible assign a public domain licence to data, because in many cases there are a lot of limitations on their reuse; for this reason, usually, the principle adopted to assign a licence is called the 'minimum common denominator'.
This principle, however, can't be considered a good practice. Based on the idea that is appropriate to use the most restrictive licence for the entire dataset, this procedure assigns by default an inappropriate licence to a high number of resources that doesn't match the original one. So, when a data provider assigns a licence, they must be sure to give the right one to each resource, also combining resources in different datasets that have the same licence.

1B. (Public financed) Research projects must ensure open access to all peer reviewed scientific publications relating to its results.
In order to achieve this goal, it is just necessary that the resource(s) will be available online, downloadable and printable and that no sensitive data is present in the publication.

1C. Any data restriction must be justified by research and public institutions, according to existing legislation.

1D. Standards are useful to make data and metadata easily accessible and reusable. Therefore, it is recommended to have data as standardised as possible for a better fruition and to ensure data survival: for this reason, data should also be in a format that can easily be modified and updated, in order to have it always readable and usable.

1E. If data are not publically financed, it is up to the owner institution to assign a licence to its data: therefore, in order to have access to data, permission is often needed This means data is less accessible for users, and also the reuse of data can be affected by this condition: if data is closed, it is difficult for researchers to obtain it, and this can negatively affect, over the course of time, the survival of data.

The level of reuse of data depends on its accessibility. If data is open, it can be easily accessed and reused.

1F. It is important to pay attention to sensitive data, that should always be secure, independently from the type of licence used.

2. Protected data and personal data must be available through a controlled procedure

The free access to data and its reuse, which allow the legal interoperability, must be subordinate to the legitimate interests of rights holder and of the entire society in a broader sense (i.e. public security), according to existing legislation.

Of course, legitimate interests change based on the country in which they are produced and for particular situations, but usually they reflect laws that regulate Intellectual Property Rights, national and public security, person privacy and so on.

Rights that should be considered in the licensing data process are:

- Intellectual Property Rights;
- National security and public safety laws;
- Protection of confidentiality and personal information;
- Protection of cultural resources;
- Periods of exclusive use of research data.

2A. It is necessary to guarantee the protection of personal data taking into consideration the following procedures:

- Providing information about the nature of the research.
- Seeking the written consent of people directly interested.
- Anonymization of sensitive information.

2B. Obtain informed consent, also for data sharing and long-term preservation / curation:

- Protect identities e.g. anonymization, not collecting personal data.
- Regulate access where needed (all or part of data) e.g. by group, use, time

period.

- Securely store personal or sensitive data separately.

2C. Guarantee controlled access to sensitive data

It is recommended to adopt an AAI (Authentication and Authorization Infrastructure) for accessing registered users.

2D. Ensuring that only people with the right profile can have access to sensitive data

It is fundamental, before providing access, that people interested in sensitive data sign a written agreement; at the same time, it is necessary to avoid providing access to everyone who wants to exploit sensitive information for their work.

To be Interoperable

3. (Meta)data licences framework should support legal interoperability fostering harmonization of rights

A licensing framework for scientific data and associated metadata supporting legal interoperability among Research Infrastructures is currently not in place for participating institutions in the PARTHENOS project. The work done to provide a landscape of existing licences in use still needs thorough analysis in line with both national and institutional policies.

The goal of reaching such a framework would ease the sharing of data among research institutions providing data to users and society as such, by framing or adopting a common set of licence statements that are simple, understandable and can be easily implemented and reused in the metadata schemas of the data in question.

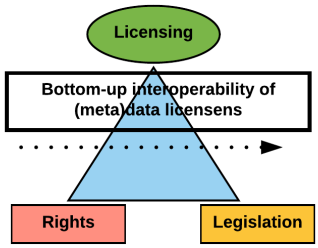
Technically that can be achieved by storing the URI of the rights statement in the metadata element or property associated with the digital object or data file to which the rights statement applies, thus ensuring machine readability for software agents or human readability for users wanting to get more information by following the link provided in often the “rights” label (see, Europeana Data Model[145] and

[145] http://pro.europeana.eu/share-your-data/data-guidelines/edm-documentation.

Datacite Metadata Schema[146]), but more correctly in the “licence” label in e.g. Dublin Core[147].

3A. (meta)data harmonization for reasons of interoperability

In order to being able to provide guidelines for good data curation practice of meta(data) licensing to data providers participating in PARTHENOS infrastructures, there needs to be a clear focus on the harmonization of the exact licensing of the digital data and objects according to interoperability standards. And preferably, there should be drawn a distinct line to the legislation (copyright) and rights (IPR) that is too complex and differentiated among countries to be solved legally within this project.



Though, Research Infrastructures should optimally share and provide a licence framework for all types of reuse for research data. At best, this would at least, but not be limited to, imply agreement templates for public, academic and restricted licences.

An example of a well-established framework using standard, interoperable and machine-readable licences to allow the interoperation between applications and services is the CLARIN License Categories[148] covering both deposition licence agreements (DELA) and end-user licence agreements (EULA).[149]

146 https://schema.datacite.org/meta/kernel-4.0/doc/DataCite-MetadataKernel_v4.0.pdf.
147 http://purl.org/dc/terms/license.
148 https://www.clarin.eu/content/license-categories.
149 https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/CLARINSA.

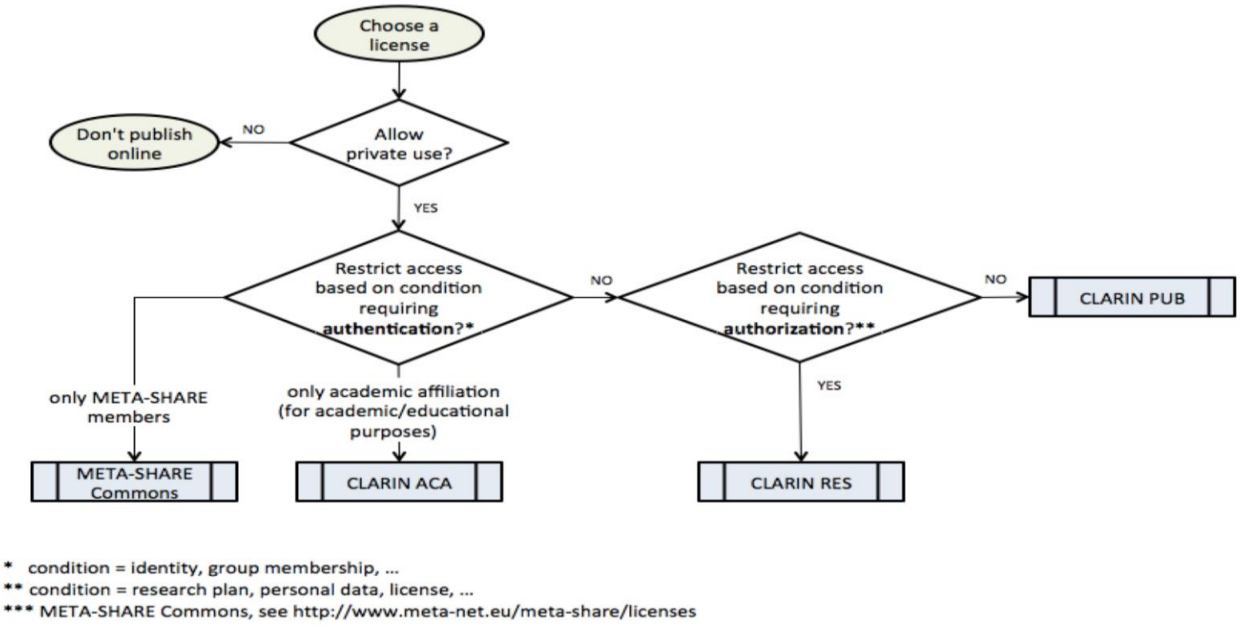


Figure 4.1: Workflow chart by Anje Müller Gjesdal & Gunn Inger Lyse, UiB.[150]

- CLARIN PUB (public domain use; licences for public resources):
  PUB resources can be distributed publicly. The distribution of these materials is not restricted by copyright or personal data protection issues. The same resource can have different licences depending on the end user's role or intended use.
- CLARIN ACA (academic use; licences for academic resources):
  ACA resources can be accessed only for research purposes. The end-user does not need to ask for usage permission but can access the resources via e.g. federated login[151].
- CLARIN RES (restricted use; licences for restricted resources):
  RES language resources have additional restrictions, which require permission from the rights holder. These resources may contain material whose usage is limited due to copyright and/or personal data protection issues. In practice, these resources require both using federated login to authenticate the end-user and sending a separate application to the rights holder for authorization possibly including a research plan with the resource.

150 https://clarin.b.uib.no/files/2014/02/CLARINO_Presentation_IPR_Solstrand_20130912.pdf.
151 https://www.clarin.eu/content/service-provider-federation.

The CLARIN license framework is set up with a clear distinction to rights and legislation and to legal responsibilities of data providers giving deposition licences (DELA) and service providers giving end-user licences (EULA) for licensing access to usage of resources. The figure below is a simplification of the "Authentication and Authorization overview" of the general framework.[152]

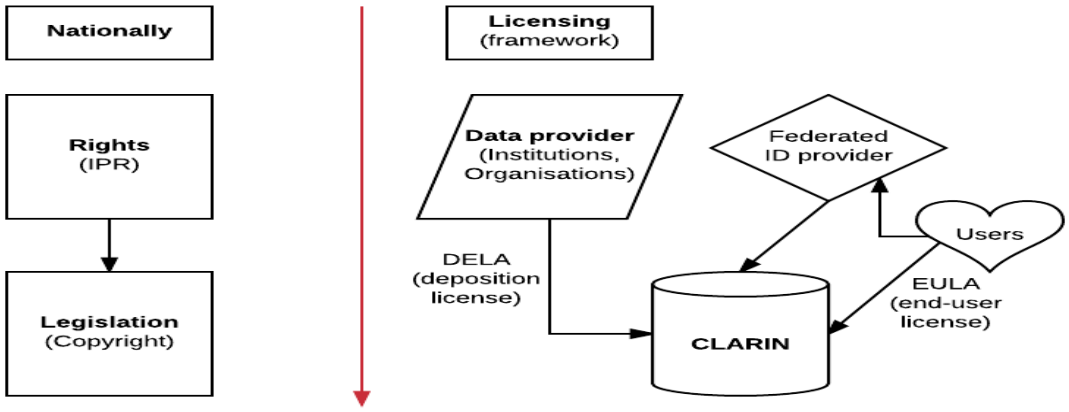


Figure 4.2: Simplified Authentication and Authorization overview.

CLARIN also provides a calculator tool for classifying licences into known CLARIN categories ("laundry tags")[153]. It would be worth considering a similar approach to map different licences onto known standards like Creative Commons or the more comprehensive initiative of the Europeana Licence Framework[154] governing the relationships of Europeana, it's data providers and users.

Relevancy for harmonizing interoperability of metadata licences for a research use case:

E.g. a Digital Humanities research group within a comparative literature department at a university wanting to pool textual digital data from various international literary archives in order to do big text and data mining of marked-up objects relevant for collecting data to answer basic research questions in their funded project.

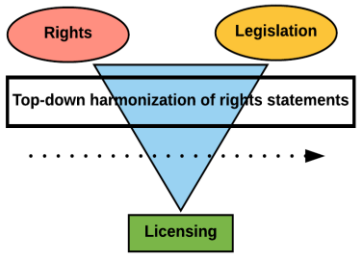
3B. Adopting top-down harmonization

[152] http://www-sk.let.uu.nl/u/m7s-2.1.pdf.
[153] https://www.clarin.eu/content/clarin-license-category-calculator.
[154] http://pro.europeana.eu/get-involved/europeana-ipr/the-licensing-framework.

Negotiating rights to copyrighted data is time consuming, but essential for data reuse allowing data and service providers to reach agreements for rights statements across both national and international legislation. The latter is particularly difficult to harmonize and should involve legal advice and liaison with legal entities or committees. Ideally, agreements and/or restrictive laws should be adopted and used as top-down harmonization tools, in order to ensure broader harmonization of data. One such successful example is The Rights Statements project led collaboratively by Europeana, Creative Commons (CC) and DPLA in the US as a joint initiative providing interoperable standardized rights statements for reuse of digital online objects aggregated by online cultural heritage platforms155. An international working group has composed a common framework consisting of 12 different rights statements for both use and reuse of national and international cultural digital objects.156 The framework builds on CC, which provides access to usage licences, where users are can associate a licence for materials on the web by referencing its URI with persistent links to representations of the statements. Rights statements are subordinated to three categories: In Copyright (InC), No Copyright (NoC) and Copyright Not Known (NKC) / Undetermined (UND), Not Evaluated (CNE). Where the Europeana Licensing Framework has been able to standardize rights across the European Union from the top down by developing a high-level metadata rights field in the Europeana Data Model (edm:rights), there are still areas of important differences in copyright law both within Europe and certainly in the rest of the world that needs harmonization. The EU, for example, has orphan works legislation, for where these need a clear status. Only EU partners are likely to be using the orphan works statement, since the Not Known Copyright (NKC) doesn't make sense to use within the European system. When either adopting or developing interoperable rights or licensing frameworks, collaborating infrastructures must provide clear guidance, educational campaigns and support for participating institutions and build up capacity for community engagement to avoid risks and

155 http://rightsstatements.org/en/about.html
156 http://rightsstatements.org/page/1.0/?language=en

confusion, especially between clearly distinct areas of licensing, rights and legislation and to help facilitate such processes for alignment across complex legal issues.



Relevancy for harmonizing rights statements for a research use case

E.g. a cross-disciplinary media and communication research centre studying political rhetoric and news in TV, who wants to be able to collect and analyse priority news broadcast with political interviews for studying current practices of gestures across countries and cultures.

To be Reusable

4. (meta)data should be licenced to permit the widest reuse possible

4A. Public Institutions and data paid with public money should be open

In recent years, both at National and European level, the free circulation of research data produced by public administrations has been encouraged. This approach stems from the idea that public research data was already paid for by taxpayer and should be available without any further payment. According to this vision, Public Institutions are no longer data owners, but they are a sort of “keeper” of digitized data. The only payment that Institutions may charge concerning digitization are the marginal costs, in order to ensure (partially) the financial sustainability of the Institution itself.

Generally, the best solution to foster the free circulation of data is represented by the application of a public domain licence. However, sometimes it is not possible to apply this kind of licence, due to several reasons (such as an author still living etcetera). In these cases, it is still possible to apply other Creative Commons licences that are considered 'free culture'. These licences (only attribution, attribution and share alike, over the public domain licences), in fact, aren't restrictive for end users and can allow reuse freely.

The main goal of creative commons licences is to spread creative works. For this reason, using the 'free culture' licences is an objective that public institutions

should pursue. For this reason, it is necessary to apply the available licences properly; several times, in fact, to avoid any issue, public institutions prefer to assign more restrictive licences, even if they are not correct. An easy example is represented by a large collection. Typically, the licence used is the most restrictive, without considering the differences between the singles items.

4B. Data users must reuse data according to local jurisdiction
It is the responsibility of data users to apply the provided licence in the right way, according to user's rights in the country in which data are used.

5. Metadata rights holder should be identified before data publishing

Usually, a Research Infrastructure aggregates big quantity of data from different data providers. This means that is really difficult assign for each data provider the right licence. Moreover, several times data providers are not able to define, without any doubt, the right licence to apply. So, especially for public institutions (who must share their data under the open by default) there is a high risk of providing a wrong licence. And last, but not least, there is the risk that end users, who are looking for data in a Research Infrastructure, don't have a clear idea of who is the real rights holder.

5A. The institution who manages data must have a clear idea of who the data rights holder is. Before data publishing, institutions need to know which people/organization can claim rights on data they want to share.

5B If the work or the resource is protected by copyright but, after a diligent search, it’s impossible to find any rights holder, then this is an orphan work. In this case, a disclaimer could be added where eventually copyright owners are invited to contact the repository manager or the researcher that used that resource. However, adding this disclaimer will not absolve the repository manager or the researcher from liability for copyright infringement; It will demonstrate that stringent efforts were made to find the copyright owner(s).

5C. Express clearly the rights holder of a data collection

Users that want to reuse data must have a clear vision of who is the rights holder, in order to address all their requests to the real data owner, if data is protected by copyright.

5D. Research organization and cultural institutions should have clear contracts and grant agreements that establish what are the rights in data managed or produced by that organization, stating clearly who is or are the rightsholder(s) of any resource and who is or are the rights holder(s) of the datasets containing these resources.

5E Provide proper attribution and credit

The right to attribution is considered a fundamental value for research data and it is a practice that allows the establishment of traceability and correct provenance. Significant progress has been made through the development of standards for data citation, however, especially for those datasets, produced by different contributors, it would be useful to provide information on proper attribution and credit in an external metadata record.

6. (meta)data rights statements should communicate the copyright and reuse status transparently and clearly

After the adoption of a standard licence for its data, the rights holder must be sure that all users will be able to understand the copyrights status applied.

It is necessary, in fact, to remember that not all users have the same level of knowledge of the law. For example, there is a huge difference between a researcher and a lawyer. For this reason, it is appropriate to provide different levels of licence explanations, in order to give everyone the best possible means for understanding them.

6A. Adoption of standardized licences helps their comprehensibility

Before adopting a standard licence, the rights holder must be sure it is expressed also as plain text understandable by users who don't have experience in the legislative field.

6B. Any information on data reuse must be declared clearly

If there are special terms and conditions to reuse data, the rights holder should inform clearly end users about them, publishing an explanation.

6C. Licences must be provided in different ways, in order to be understandable by everyone (humans and machine)
From a technical point of view, for the online licences, it is necessary to consider that exist three different readable levels with which share data:

- Human readable: this procedure provides a description of the licence that is clearly understandable by human who don't have skills in the legal field.
- Machine readable: this procedure provides a HTML code that is read by machine via right expression language.
- Legal code: this procedure provides a description of licences according to a traditional legal tool that lawyers understand.

7. Specify why and for what period a data embargo is needed (data should be made available as soon as possible)

Even though the goal is to make data available for all immediately there might be good reasons for an embargo period. Bearing in mind that the decision to grant public access has been made and it is just a matter of time to implement it. A good reason for a temporary delay of access to data may be to protect sensitive or personal data, e.g. the exact location of underwater archaeological sites for protection from unscrupulous thieves until they can be secured and that data can be made publicly available. Another reason may result from the point of view of financial exploitation. A temporary exclusive access to certain data could enable a public institution to get some return on the money invested to create that data and save tax payers money by using it in a responsible way. Therefore, consider the following when thinking of an embargo:

7A. Embargo periods must have a specific, clear stated end.

7B. Embargo periods should be as short as possible.

7C. The affected (meta)data should be as limited as possible and clearly described.

7D. The reason for the embargo period should be stated clearly in a comprehensible way.

7E. The (meta)data should be made immediately accessible after the embargo has ended.

7F. Embargo period specifications (duration, affected data, reason for embargo, etcetera) of data should be included in their metadata if possible.

5. Foresight study and interdisciplinary research agenda

5.1. Objectives and nature of the task

Task 3.4 is entitled “Foresight study and interdisciplinary research agenda”. The associated deliverable D3.3 will be “…a report that analyses current trends and outlines possible progress in the interdisciplinary sector addressed by the project. It provides insight into opportunities for and threats to innovation, as well as forecast developments concerning careers, research topics and funding opportunities. It consists of a detailed report, including a self-consistent summary detachable part available for separate use, e.g. as a political brief.” This chapter provides an introduction and an overview of setting up the work.

In short, the ‘foresight study’ will address how digital research methods in the domains addressed by PARTHENOS – that is to say, the (Digital) Humanities, heritage, and so on – may develop over the next 5-10 years, examining the current state of the art, identifying emerging trends and requirements, risks – and at the consequences of these emerging approaches/methods – and organisations such as universities and funding bodies could help the potentiality become actual. Although the task description doesn’t state this explicitly, we are assuming that we are in particular looking at these methods in relation to the data life cycle, policies, IPR and so on. Future developments in digital methods and virtual research environments will be constrained by existing data infrastructures, policies and frameworks, and will drive both the evolution of these infrastructures, policies and frameworks and the development of new ones.

5.2. Frameworks for foresight studies

5.2.1. Introduction

Foresight research is a key mechanism for the development and implementation of research and innovation policy in the medium to long-term, enabling policy-making bodies (such as government agencies) to set research priorities and influence the progress of research. It is important to understand that foresight research is not simply ‘future gazing’, nor is it about forecasting by experts (although experts may,

and should, participate). Rather, it is a way of facilitating structured thinking and debate about long-term issues and developments, and of broadening participation in this process of thinking and debate, for example through networks involving different stakeholders, to create a shared understanding about possible futures and to enable them to be shaped or influenced.

To this end, systematic frameworks, instruments and tools have been developed for carrying out foresight research – here we are following Georghiou et al. (2009)[157], in particular Chapters 1-3. This framework is a generic one, addressing foresight studies in different domains and contexts, so the terminology needs in some cases to be reinterpreted for our particular context.

5.2.2. What is 'foresight'?

There are multiple definitions in the literature – this is one that captures the key aspects:

> *[Foresight is] a process which involves intense iterative periods of open reflection, networking, consultation and discussion, leading to the joint refining of future visions and the common ownership of strategies ... It is the discovery of a common space for open thinking on the future and the incubation of strategic approaches.*[158]

[157] Georghiou, L., Cassingena Harper, J., Keenan, M., Miles, I., Popper, R.. The Handbook of Technology Foresight: Concepts and Practice (PRIME Series on Research and Innovation Policy in Europe), 2008.
[158] Cassingena Harper, J. (ed.) (2003). Vision Document, eFORESEE Malta ICT and Knowledge Futures Pilot, http://forlearn.jrc.ec.europa.eu/guide/7_cases/EforeseeMalta.htm.

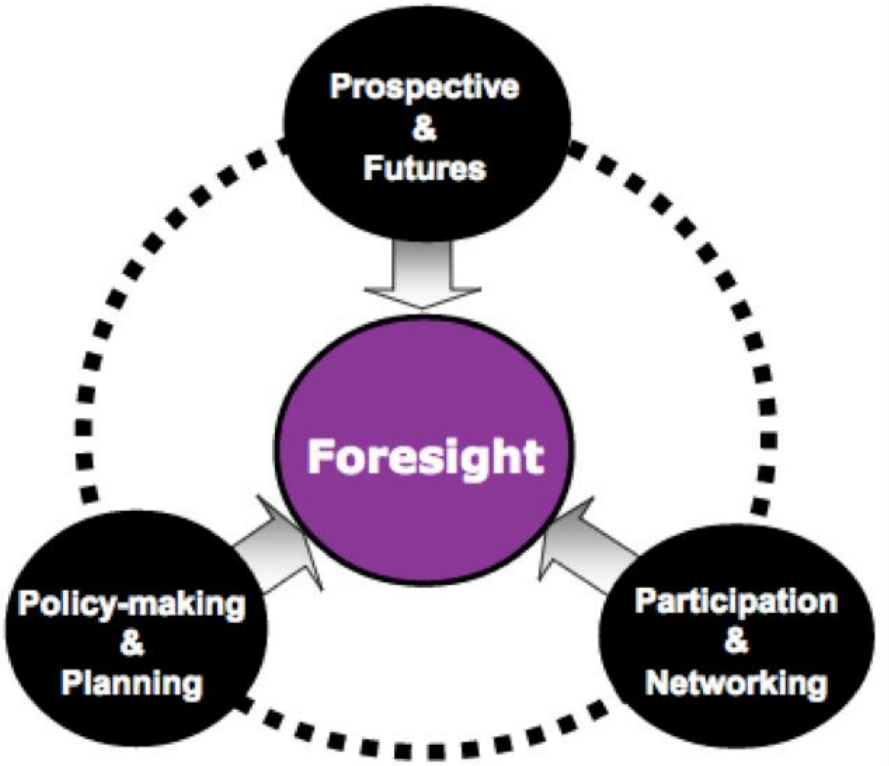


Figure 5.1: What is foresight?

The central aim is to develop an understanding of the future – or more precisely of prospects, that is to say potential futures – but this is essentially a shared vision. A key component is the participative aspect. The vision is not that of a small number of experts, but is based on engagement with and involvement of a broad range of key stakeholders, including decision- and policy-makers, but also ‘citizens’ (in the terminology of the framework) of the community in question, which in our case would include both potential users, infrastructures, and other stakeholders (such as infrastructure providers, data curators).

Engaging a representative range of relevant and informed stakeholders in the dialogue brings several benefits: it extends the breadth and depth of the knowledge base created by the foresight process, by drawing on distributed knowledge (different stakeholders have access to different information), and thus enriches and improves decision making; it increases the ‘democratic basis and legitimacy’ of the study report by avoiding a top-down, expert-driven analysis; it helps to spread the message about foresight activities and to embed it within participating organisations, thus improving sustainability.

It thus draws upon existing knowledge networks and stimulates new ones – in addition to any reports, these embedded networks are an important output of foresight activities, facilitating a longer-term thinking process that extends beyond the period of the study itself.

Finally, bringing longer term considerations into decision-making facilitates higher-level policy making and strategic planning. Here we draw together the various threads that we identify in our activities, and make recommendations to our audience – this corresponds to the 'research agenda' aspect of Task 3.4, which complements the 'foresight' aspect.

5.2.3. Foresight as a process

It must be emphasised that foresight is a process, which in the model that we are following[159] is analysed into five broad stages, as illustrated in Figure 5.2. The process is iterative and cyclical, both within stages and as a whole.

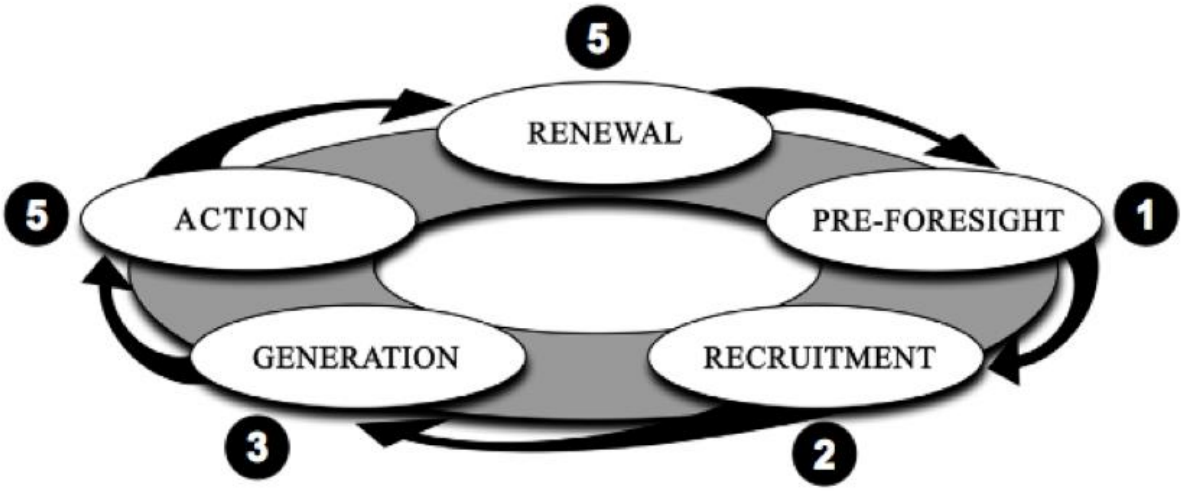


Figure 5.2: The foresight process.

The nature of these stages is shown in more detail in Figure 5.3. To examine these in the context of the PARTHENOS study:

[159] Georghiou, L., Cassingena Harper, J., Keenan, M., Miles, I., Popper, R.. The Handbook of Technology Foresight: Concepts and Practice (PRIME Series on Research and Innovation Policy in Europe), 2008.

Pre-foresight: This stage has already been carried out to a great extent, at least in draft form. The rationale and objectives of the activity have been identified (indeed in outline this was already in the DoW); the team has been assembled and the methodology defined; and the work to be carried out has been scoped, through the research areas identified by the team, and the initial analysis of literature.

Recruitment: This stage involves identifying and engaging with key stakeholders or 'citizens' of relevant communities (as discussed above, not just 'experts'), through workshops, panels and interviews.

Generation: This is the heart of the foresight process, in which the knowledge base is constructed ('generated') by 'exploration, analysis, anticipation of the possible futures'. Existing knowledge (including opinion) is collected together, analysed and synthesised; tacit knowledge is identified and made concrete; and new ideas about where we are going are developed. The task has already started to develop this knowledge base, although it is as yet still in the form of informal, human-readable documents.

Action: This is the stage in which the knowledge base developed is used as the basis for decisions and for planning change and innovation. In the case of our task, this will be a matter of making recommendations in our report that will hopefully be acted upon.

Renewal: The Renewal stage covers follow-on activities, including sustainability issues, embedding foresight in organisations (so that it continues), and evaluation of what we produce during the project, which will need to continue after the project is complete as the landscape will be ever changing.

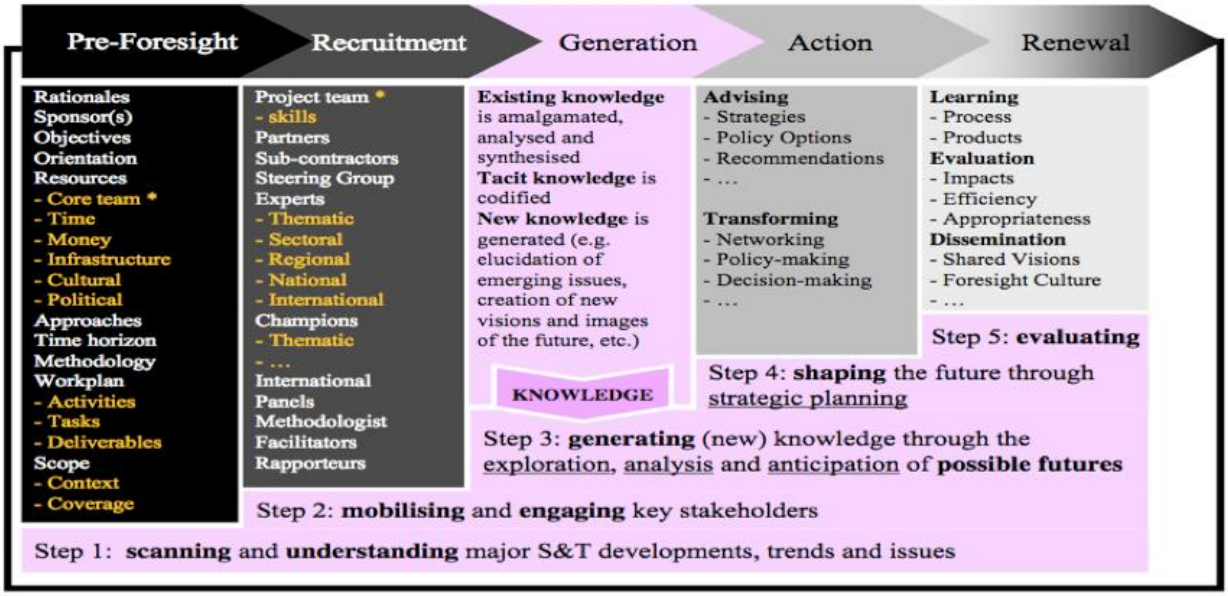


Figure 5.3: The stages of foresight.

5.2.4. Methods for foresight: the foresight diamond

Another aspect of this framework is the selection of specific methods for constructing the knowledge base, during the 'Generation' stage. Figure 5.4 shows the 'foresight diamond', a representation of the most relevant methods (the framework identifies substantially more) in terms of what the framework calls 'knowledge source': 'creativity'-based methods require more original, imaginative and open-ended thinking; 'expertise'-based methods make use of the skill and knowledge of people expert in specific areas; 'interaction'-based methods bring together knowledge from multiple people (not necessarily experts); 'evidence'-based methods are those based more on relatively 'hard' data, such as literature reviews, statistics, etcetera.

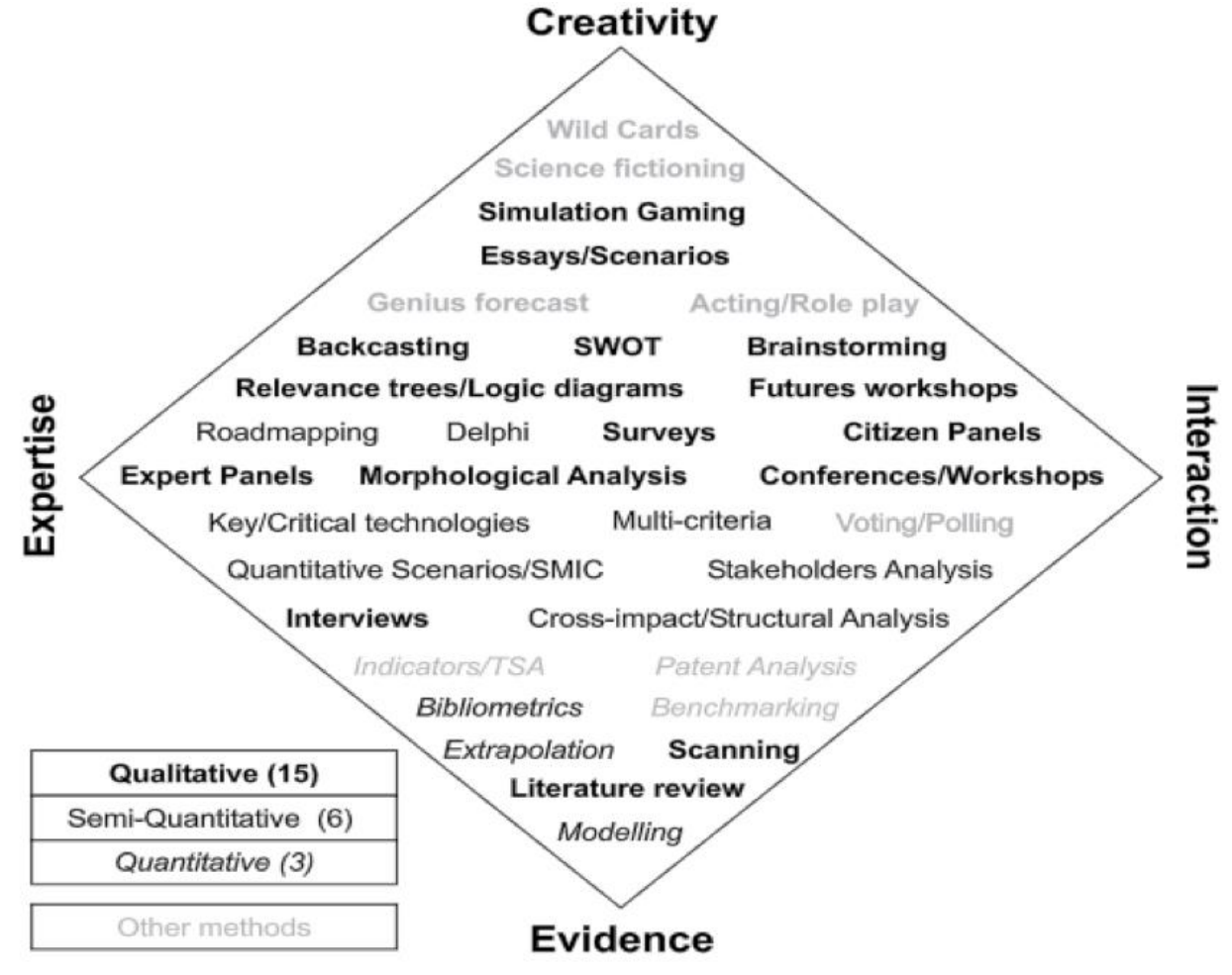


Figure 5.4: The foresight diamond.

If you examine the methods that we plan to follow in Task 3.4, as highlighted in Figure 5.5, they mostly fall into the bottom half of the diamond – this may be because the more 'creative' methods are used to look further forward into the future, whereas the timescale for T3.4 is much shorter, namely 5-10 years. Research[160] suggest that most foresight studies use approximately 5-6 different methods, so our planned approach fits this pattern.

[160] Georghiou, L., Cassingena Harper, J., Keenan, M., Miles, I., Popper, R.. The Handbook of Technology Foresight: Concepts and Practice (PRIME Series on Research and Innovation Policy in Europe), 2008.

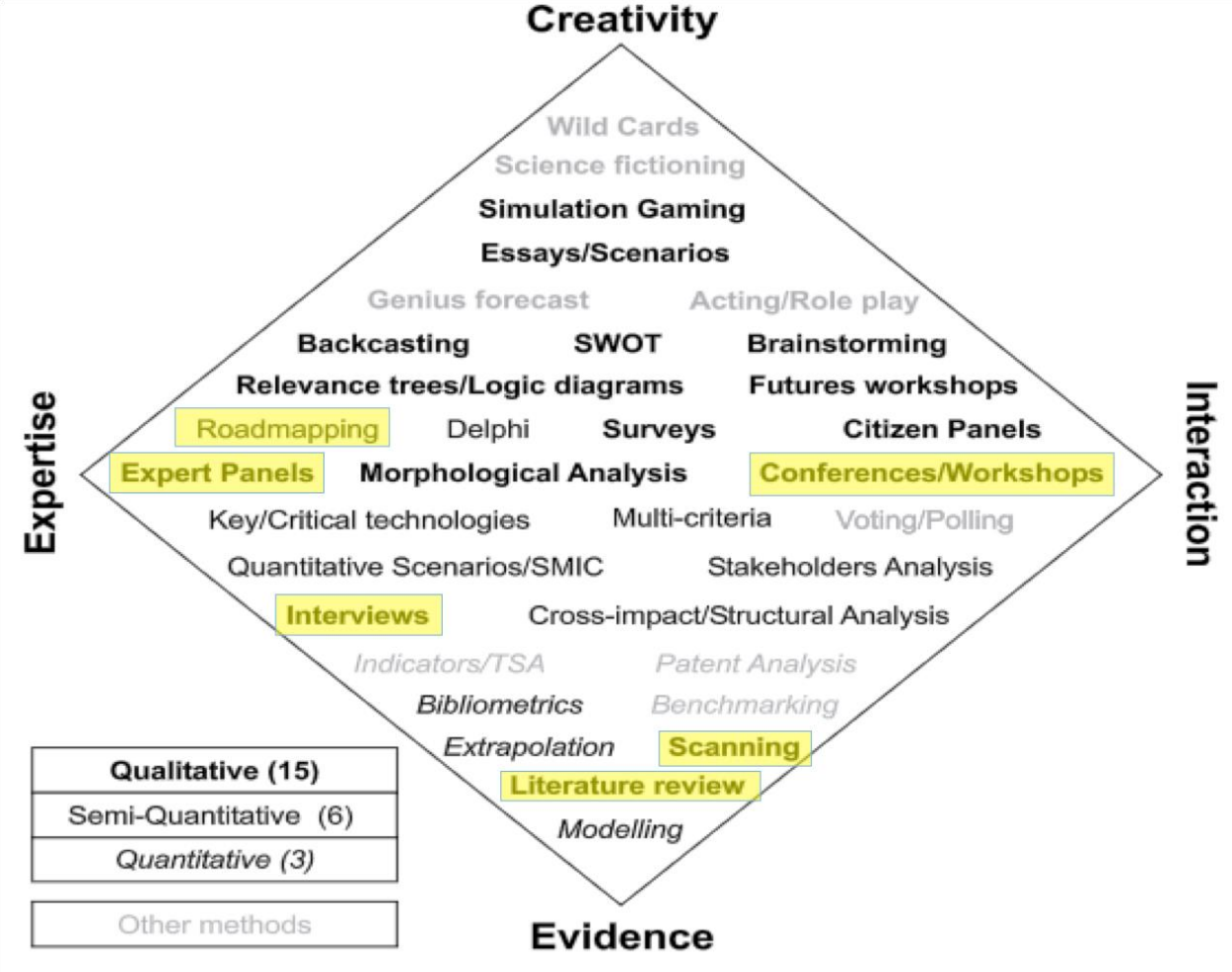


Figure 5.5: Foresight methods in PARTHENOS.

5.3. Overall approach to task

Within this overall framework, T3.4 is following a thematic approach, structuring the work around broad research areas (in terms of methods, approaches and issues, rather than subject disciplines) in which the partners involved in the task are interested and/or active. There were two rationales for this: firstly, it is these research methods and themes that drive the data- and policy-related issues that PARTHENOS is investigating; secondly, focusing on partner interests enables us to make the best use of people's time and expertise. The initial 'themes' are: Public Humanities (including crowdsourcing); big data; data curation and preservation; linked open data; Geospatial Humanities; socio-technical issues (include interdisciplinary issues).

As indicated in Figure 5.5, various activities (such as interviews and workshops) will be undertaken; however, the initial activity for the task will be a review of the existing literature and research projects in the research areas being addressed, with a view to identifying the current landscape and emerging trends, requirements, and issues. 'Literature review' is interpreted broadly, including not only academic publications but also 'grey literature' such as project reports, websites, and existing reports on Research Infrastructures, frameworks, policies, etcetera (e.g. http://www.jpi-culturalheritage.eu/wp-content/uploads/SRA-2014-06.pdf). In addition, it will take note of outputs from other PARTHENOS tasks, for example D2.1 (in particular, Section 2.3 on policies), and 'gaps' in provision identified by Tasks 3.1-3.3 (Chapter 3 and Chapter 4).

6. PARTHENOS high-level recommendations

The present deliverable is a product of the combined efforts of the different tasks in WP3 and gives an overview of existing policies concerning data management as well as policies concerning quality of data, metadata and repositories and IPR, open data and open access. In this final chapter, the recommendations from the previous are revisited and mapped onto the FAIR principles. The result is a set of PARTHENOS high-level recommendations.

WP3 PARTHENOS recommendations: Findable

- Select an appropriate metadata schema for the type of resource being described, fitting to the type of resource. Metadata can cover various aspects, such as citation metadata, disciplinary metadata, preservation information, provenance, etcetera. The metadata intended for findability is the type of metadata used for citation and describing data in a catalogue. This should be the primary format for maintaining the descriptive metadata. Utilize existing metadata schemas, such as schemas according to ISO 24622-1 (Component Metadata Infrastructure, adjustable to each type of resource), or MARC21 (if appropriate for the type of data). Dublin Core alone is not suitable for a detailed description of research data, nor is Datacite MDS (see Section 3.2.1.3).
- Make requirements for the use of persistent identifiers for referencing and association with the referred contents part of the metadata (see Section 3.2.1.3). Select an appropriate persistent identification schema and assign a PID to every resource (see Section 3.3.1). Each data-object and dataset should be identifiable by an eternal persistent identifier. This makes sure that a certain data object as well as an entire dataset is retrievable during time, when they are made available both via online and offline environments. Persistent identifiers can take different forms: handles, DOIs, PURL, URN (see Chapter 2, especially Section 2.5).
- The metadata should be provided in high quality, as correct and complete as possible, including enough information for later access and understandability (see Section 3.3.1 and Section 3.2.1.3,. Describe your metadata as richly as

possible (see Chapter 2, especially Section 2.5).

- Ensure semantic interoperability by referencing authority files, for example ISNI, VIAF, ORCID (see Section 3.2.1.3).
- Make descriptive metadata publicly accessible using standardized protocols, such as OAI-PMH, SPARQL (see Section 3.2.1.3 and Section 3.3.1).
- Publicize the protocol endpoint to suitable search providers, for example CLARIN maintains a registry for endpoints providing language related research data (see Section 3.2.1.3).
- Provide different formats, this can, for example, include HTML to allow findability with standard internet search engines, Datacite MDS and Dublin Core for interoperability purposes with archives metadata (see Section 3.2.1.3).
- Gaps in the data should be clearly stated: e.g. Historians recommend that not only the context and richness of data should play a prominent role, but the gaps in data coverage as well. This makes clear what can be and what cannot be expected in a dataset or repository (see Section 2.5).
- Apply Discipline Specific Citation Guidelines. Each discipline in the Humanities has its own “best practice” to cite literature and other external data. Despite these different standards, each researcher in the Humanities should follow discipline specific citation standards (see Section 2.5).

WP3 PARTHENOS recommendations: Accessible

- (Meta)data should be open as possible and closed as necessary (see Section 4.3.3).
- Protected data and personal data must be available through a controlled and documented procedure (see Chapter 4, especially Section 4.3.3). Information that needs to be protected, for example for privacy reasons, should not be part of the publicly accessible (meta)data but should be recorded as part of the documentation of the resource in restricted contexts (see Section 3.2.1.3 and Section 3.3.1).
- In order to be fully accessible, research data should be fully accessible via (free) exchange protocols (see Chapter 2, especially Section 2.5).
- Make your data accessible through a trustworthy repository. Depositing

research data in a certified repository (DSA-WDS, NESTOR, ISO): means that the researcher can trust the preservation and dissemination policies adopted by such data archive, as they have been reviewed according to internationally agreed standards. Data repositories should be trustworthy and therefore certificated. As a data archive apply for a quality assessment to receive a certification (both formal - DSA- or formally attributed - ISO) in order to be attributed recognition of trustworthiness and support of research (see Section 2.5).

- Long-term preservation and archiving strategies are the ones that make sure that data are available for long time spans. However, the definition of how “long” the long term should be quite difficult to quantify, as each discipline refers to different standards and definitions. Therefore, for an in-depth definition of this principle, we suggest consulting the policies and best practices for each specific discipline (see Section 2.5).
- Follow a precise and detailed naming convention which allows researchers to retrieve and access digital objects and data more easily; digital archives/ repositories usually have best practices in place to create and apply specific naming file conventions. We suggest referring to the policies/ best practices for every discipline to find the most suitable naming convention for your research/ archive (see Section 2.5).
- Maintain the integrity and quality of data. This is a general principle, that emerged in particular from the interviews with historians. It refers to the necessity to maintain the richness and the context of the data created and collected during time (see Section 2.5).

WP3 PARTHENOS recommendations: Interoperable

- Give an easy to find and detailed overview on accepted (meta)data formats. Ideally, in a single page that can be directly referenced and where the information on (meta)data formats is not hidden in an overwhelming document that covers all of the aspects of the repository. In general, a fine granulated and good structured documentation that uses modern aspects of design and user interface methodology can help to see at a glance possibilities for interoperability. It may be a good idea to structure such

documentation along the FAIR principles (see Section 3.2.3).

- Document and also give easy access to the data model(s) in use in a repository. Also make clear which parts of the data model enable interoperability and which parts are relevant when connecting datasets between projects (see Section 3.2.3).
- On a technical level the (automatic) transformation of data in the ingest phase of repositories can enable interoperability on the fly. This is something where common developed scripts and tools should be developed in a joint effort and shared between repositories (see Section 3.2.3).
- Establish a quality assurance processes, give a special focus on the data creation phase (see Section 3.2.3.1).
- Pushing data providers and establishing automatic processes to boost (meta)data quality and therefore interoperability should be combined and applied (see Section 3.2.3.1).
- Invest in tools that help cleaning up (meta)data and converting raw data into other (standardised and interoperable) data formats (see Section 3.2.3.1).
- Establish well documented machine-actionable APIs for the (meta)data (see Section 3.2.3.1).
- Give more information on best practices for machine driven automatically data search and reuse (see Section 3.2.3.1).
- On a higher level support standard interfaces for exchanging metadata (see Section 3.2.3.1).
- The description of metadata elements should follow community guidelines that use an open, well defined vocabulary. Convince researchers to use FAIR compatible vocabularies and ontologies from the very start. Give recommendations on how to do this and how to integrate references in their research data and metadata. Give pointers on which vocabularies and ontologies can be used, based on research domain specifics and on the tangible use case (see Section 3.2.3 and Section 3.3.1). Each discipline refers to different knowledge systems, therefore there are discipline-specific ontologies and controlled vocabularies, which we suggest are consulted separately (see Chapter 2, especially Section 2.5).
- Convince researchers to structure and enrich their research output in such

matters that data hosts can ingest this data already as FAIR compatible as far as possible. This needs a joint effort between policy makers and data creators (see Section 3.2.3.3).

- Invest in enrichment tools or user interfaces that help to make references in data objects syntactically parseable and semantically machine-accessible (see Section 3.2.3.3).
- (Meta)data licences framework should support legal interoperability fostering harmonization of rights (see Section 4.5).

WP3 PARTHENOS recommendations: Reusable

- All files held in the repository should be in an open, simple, standardised format that is considered likely to offer a degree of long-term stability. When a format is in danger of becoming obsolete, proper digital preservation actions must be performed. Adopt the preservation by migration, if necessary (see Section 3.2.4.3).
- Use international standard formats, i.e. XML and RDF textual formats (see Section 3.2.4.3).
- Use open source tools for generating metadata and for automatic validation (see Section 3.2.4.3).
- (Meta)data should be licensed to permit the widest reuse possible (see Section 4.5).
- (Meta)data rights holder should be identified before data publishing (see Section 4.5).
- (Meta)data rights statements should communicate the copyright and reuse status transparently, clearly and machine readable (see Section 4.5).
- Specify why and for what period a data embargo is needed. Data should be made available as soon as possible (see Section 4.5).

7. Appendix I: Terminology used by WP3

7.1. Abbreviations

7.1.1. Institutions

CESSDA	Consortium of European Social Science Data Archives
CINES	National Computing Center for Higher Education
CLARIN	Common Language And technology Research INfrastructure
CNR-ILC	Centro Nazionale Ricerche - Istituto Linguistica Computazionale
CNR-ILIESI	Centro Nazionale Ricerche - Istituto per il Lessico Intellettuale Europeo e la Storia delle Idee
CNR-OVI	Centro Nazionale Ricerche - Opera del Vocabolario Italiano
CNRS	Centre National de la Recherche Scientifique
EGI	European Grid Infrastructure
EUDAT	European Association of Databases for Education and Training
FHP	Fachhochschule Potsdam
ICCU	Istituto Centrale per il Catalogo Unico delle biblioteche italiane e per le informazioni bibliografiche
INRIA	Inventeurs du monde numérique
KNAW-DANS	Koninklijke Nederlandse Akademie van Wetenschappen - Data Archiving and Networking Service
OPF	Open Preservation Foundation
SISMEL	Società Internazionale per lo Studio del Medioevo Latino
TCD	Trinity College Dublin
ZIM-ACDH	Zentrum für Informationsmodellierung - Austrian Centre for Digital Humanities

7.1.2. Domain and technology abbreviations

AAI	Authentication and Authorization Infrastructure
AAT	Art and Architecture Thesaurus
API	Application Programming Interface
CCR	CLARIN Concept Registry
CMDI	Component MetaData Infrastructure
CMM	Capability Maturity model
CMS	Content Management System
DMP	Data Management Plan
DOI	Digital Object Identifiers
DSA	Data Seal of Approval
FCS (API)	Flow Cytometry Standard
IPR	Intellectual Property Rights
ISBN	International Standard Book Number
ISNI	International Standard Name Identifier
NAS	Network Attached Storage
NGI	National Grid Initiative
NREN	National Research and Education Network
OA	Open Access
OD	Open Data
OAI-PMH	Open Archive Initiative - Protocol for Metadata Harvesting
OMS	Object Management System
ORCID	Open Researcher and Contributor ID
OAIS	Open Archival Information System
PID	Persistent IDentifier
PISA	Persistent Identification and Sustainable Access
PREMIS	PREservation Metadata Implementation Strategies

PSI	Public Section Informative directive
RDF	Resource Description Framework
REST (API)	Representational State Transfer
RI	Research Infrastructure
SKOS	Simple Knowledge Organization System
SPARQL	SPARQL Protocol and RDF Query Language
SSH	Secure SHell protocol
TDR	Trusted Digital Repository
TEI	Text Encoding Initiative
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
URN	Uniform Resource Name
VIAF	Virtual International Authority File
WMS	Warehouse Management Systems

7.2. Domain terms definition

Annotation
An annotation is a form of metadata (e.g. a comment, description, explanation) attached to a piece of text, an image, or other data type. Science Europe Data Glossary
Archive
A place or collection containing records, documents, or other materials of historical interest. - archive. (n.d.) American Heritage® Dictionary of the English Language, Fifth Edition. (2011). Retrieved August 11 2016 from http://www.thefreedictionary.com/archive An archive may contain digital or analogue materials or both.
Business model
Is an "abstract representation of an organization, be it conceptual, textual, and/or

graphical, of all core interrelated architectural, co-operational, and financial arrangements designed and developed by an organization presently and in the future, as well as all core products and/or services the organization offers, or will offer, based on these arrangements that are needed to achieve its strategic goals and objectives." Defining a business model in the new world of Digital Business
Copyright
is a legal right created by the law of a country that grants the creator of an original work exclusive rights for its use and distribution (see also Chapter 4). Wikipedia
Data archive
A data archive is a professional institution for the acquisition, preparation, preservation, and dissemination of research data (see also Section 1.4.1.2). Science Europe Data Glossary
Data centre
A data centre is a facility used to house computer systems and associated components, such as telecommunications and storage systems. Science Europe Data Glossary
Data embargo
A data embargo means that resources, even if they are submitted to a public repository, are not available for download to save the investment made by the resources producers.
Data Management plan
A formal document that outlines how data are to be handled both during a research project, and after the project is completed (see also Section 3.3.1). (Wikipedia)
Data Management Policy
A data management policy is a directive providing language that encourages or requires researchers to handle their research data in such a way that they fulfil institutional, grant or other types of funding expectations. Science Europe Data Glossary It has the purpose to ensure that research data is stored, retained, made accessible for use and reuse, and/or disposed of, according to legal, statutory, ethical and funding bodies' requirements. (Monash University Policy)
Data provider
An organisation which produces data or metadata. (Glossary of statistical terms)
Data Quality Policy
A Data Quality Policy can be described as a set of formal directive and recommendations to ensure researchers to produce data that are of the highest

quality possible, for the purpose of findability and reuse. Many universities and research institutes indicate to their researchers the guidelines for producing good quality data. Despite some variations in what can be considered “good data” for different institutions, the following are generally recognised as indicators of good quality data: accuracy, validity, reliability, timing, relevance, completeness (see also Chapter 2).
Data reuse
“Reuse can mean re-analysing data from a new research perspective, based on general advances made in science. It can also mean combining/re-combining or simply comparing older data with new data or model outputs in order to obtain a fuller picture or a longitudinal series of data.” - DANS Studies in Digital Archiving 6 - Selection of Research Data, Guidelines for appraising and selecting research data (pdf). 2011. Tjalsma, Rombouts.
Data stewardship
Data stewardship is the management and oversight of an organization's data assets to help provide business users with high-quality data that is easily accessible in a consistent manner. Data Search Management
Dataset (definition from PARTHENOS entities V.1.12)
A dataset is any set or collection of data, records or information kept as a persistent unit of information in the knowledge generation process from primary records up to any level of aggregation or integration. The identity of a dataset is given by its content on the bit-level of encoding and its provenance. Since large datasets have a very small chance to be “reinvented” with another meaning, it is often practical to base the identity of a dataset on the content only, and apply a respective disambiguation of provenance only in case of obviously accidental identity. Different versions of a dataset are regarded as different datasets. Their relation should be defined by metadata describing the derivation process, rather than by version numbers. In general, a dataset may be integrated from different sources of provenance, such as a corpus of inscriptions compiled from different publication or a snapshot of a complete digital library. The integrated dataset may preserve the units of information of the source from which it has taken components. The content of knowledge organization systems, such as gazetteers, author lists, thesauri and formal ontologies of terms at a particular point in time, fall under datasets.
Digital preservation
Designates the methods, organisation and systems which are needed to ensure access to digital materials over time. It covers the series of managed activities necessary to ensure continued access to digital materials for as long as necessary.
Intellectual Property Rights
Refers to creations of the intellect for which a monopoly is assigned to designated owners by law (see also Chapter 4). Wikipedia

Legal interoperability
Is the legal rights, terms, and conditions of databases from two or more sources are compatible and the data may be combined by any user without compromising the legal rights of any of the data sources used. The legal interoperability of data
Licensing framework
Is a standardized and harmonized set of licenses that provide an overview for use and reuse of data.
Long-term preservation
The act of maintaining information, Independently Understandable by a Designated Community, and with evidence supporting its Authenticity, over the long term. The medium-term (three- to five-year), long-term (> five years). OAIS
Open Access
Open access is the practice of providing (unrestricted) on-line access to scientific information that is free of charge to the end-user and free of most copyright and licensing restrictions. Science Europe Data Glossary
Open Data
Open data is the idea that (some) data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control. Science Europe Data Glossary
Orphan works
Orphan works are works like books, newspaper and magazine articles and films that are still protected by copyright but whose authors or other right holders are not known or cannot be located or contacted to obtain copyright permissions. Directive 2012/28/EU
Preservation strategy
Is the strategy used for preservation. A strategy for functional preservation can e.g. be a migration strategy or emulation strategy, which is an example of an overall strategy. For bit preservation, a strategy can be chosen in form of a specific solution for how the bits are preserved, which is an example of a detailed strategy.
Public Section Informative
The Directive on the reuse of public sector information provides a common legal framework for a European market for government-held data (public sector information). European Commission
Significant properties
Are those aspects of the digital material which must be preserved over time in order for the digital object to remain accessible and meaningful. (Wikipedia)

7.3. Technology terms definition

Application Programming Interface (API)
An API is a set of subroutine definitions, protocols, and tools for building application software. Wikipedia
Authentication and Authorization Infrastructure (AAI)
The Authentication and Authorization Infrastructure is a service and a process that allows the access to protected data of different organizations.
Big data
Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. Science Europe data Glossary
Bit preservation
Is defined as the required activities to ensure that the bit-streams remain intact and readable.
Capability Maturity Model (CMM)
Generally speaking, a Capability Maturity Model (CMM) is a technical and cross-discipline methodology used to facilitate and refine software development processes and system improvement. CMM is a benchmark used to compare organizational processes. It is routinely applied to the fields of IT, commerce and government to facilitate business area processes, such as software engineering, risk management, project management and system engineering. In the field of archives and repositories (also analogues archives, such as CHIs) CMM has been introduced as a way to assess the “maturity” of a repository.
Certification
In general, a certification (see also Section 2.2.1) is the assignment of a certificate to a body or system related to a standard. In the case of ISO certification, third parties offer these services. ISO does not offer certification through its committee on Conformity Assessment (CASCO) has produced a number of standards defining international consensus on voluntary criteria in certification good practice (http://4cproject.eu/community-resources/glossary). Applied to digital repositories, a certification testifies the quality of a repository in relation to its stability, reliability, preservation and dissemination capability. The DSA (Data Seal of Approval) is in fact a certification standard for digital repositories, based on on peer review of a set of 16 quality guidelines for the creation, storage and use of data. (http://sedataglossary.shoutwiki.com/wiki/Data_seal_of_approval)

Checksum
Mathematical value computed from a group of data being transmitted, and transferred with the data. The receiving device compares the checksum with its own computation and, if it differs from the received checksum, requests the transmitting device to resend the data. Business Dictionary
Data
Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship (see also Section 2.1). - *C.L. Borgman (2015). Big Data, Little Data, No Data: Scholarship in the Networked World. MIT Press.*
Data annotation
A type of metadata added to the original data, or part of it, pertaining and aiming to adding information or making information explicit.
Digital repository
“A digital repository is a place that holds digital resources, makes digital resources available to use, and organizes them in a logical manner” (see also Section 1.4.1). Science Europe data Glossary
Legacy data
Information stored in an old or obsolete format or computer system that is, therefore, difficult to access or process. Business Dictionary
Logical preservation
Is the part of digital preservation that ensures that the bits remain understandable and usable according to preservation purpose?
Machine readable
Is data (or metadata) which is in a format that can be understood by a computer. Wikipedia
Metadata
“Metadata is the data that describes an item such as a data set” (see also Section 2.1). Science Europe Data Glossary
Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)
The OAI-PMH it's a protocol developed by Open Archive Initiatives as a communication infrastructure. It is used to harvest metadata from an archive and provide them to an external source.
Persistent identifier

Is “an identifier is [that is] valid for long enough” (Hunter, 2005). This means as long as the referred resource is relevant, - and that the resource pointed to is under a digital preservation program in order to maintain the contents, - and that relation between content and the identifier is maintained in order to make the content accessible in the future via the identifier.
Preservation format
Is a file format which fulfils requirements for the chosen preservation strategies, which usually cover requirements like e.g. openness of the format.
Raw data
Raw data is data that has not been subjected to processing, analyses or any other manipulation. Science data glossary
Repository
A repository is generally defined as “place where things may be put for safekeeping” (see also Section 2.2) - repository. (n.d.) *American Heritage® Dictionary of the English Language, Fifth Edition*. (2011). Retrieved August 11 2016 from http://www.thefreedictionary.com/repository
Research data
Research data is any (digital or analogous) object or evidence that is needed to underpin research. Science data glossary
Secure SHell protocol (SSH)
SSH is a network protocol that provides administrators with a secure way to access a remote computer (http://searchsecurity.techtarget.com/definition/Secure-Shell)
Trusted Digital Repository (TDR)
“[A] trusted digital repository (TDR) is a repository whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future.” Science Europe data Glossary

7.4. Best Practice, policy, procedure, standard

Best practices
A method or technique that has consistently shown results superior to those achieved with other means, and that is used as a benchmark. Business Dictionary
Policy

The set of basic principles and associated guidelines, formulated and enforced by the governing body of an organization, to direct and limit its actions in pursuit of long-term goals. Business Dictionary
Data policy
Generally speaking, a policy is a system of principles to guide decision and achieve rational outcomes (Wikipedia). Data policies are norms regulating the data management and publications of research data. They range from recommendations to enforcement (ifdo.org - http://ifdo.org/wordpress/open-accessdata-policies/).
Procedure
A fixed, step-by-step sequence of activities or course of action (with definite start and end points) that must be followed in the same order to correctly perform a task. Repetitive procedures are called routines. Business Dictionary
Standard
Universally or widely accepted, agreed upon, or established means of determining what something should be. Business dictionary

8. Appendix II: Matrix ‘Roles, Tasks, Quality’

8.1. Collected policies in the field of Archaeology

Policy	Policy link	Country
Sustainability of Digital Formats. Planning for Library of Congress Collections	http://www.digitalpreservation.gov/formats/index.shtml	United States
Guidelines for Web-based data publication in Archaeology. A working document to inform archaeologists about sharing data, from the field to the Web.	http://104.197.134.73/wp-content/uploads/2011/06/Guidelines_Jan2011.pdf	United States
3D Icons Guidelines	http://3dicons-project.eu/eng/Guidelines-Case-Studies	Europe
Archaeology Data Service/ Digital Antiquity. Guide to Good Practice.	http://guides.archaeologydataservice.ac.uk/g2gp/	United Kingdom
Romanian Archaeologists Conduct Code	http://www.cultura.ro/uploads/files/CodDeontologicArheologi.pdf	Romania
A standard and guide to best practice for archaeological archiving in Europe	http://archaeologydataservice.ac.uk/arches/attach/The%20Standard%20and%20Guide%20to%20Best%20Practice%20in%20Archaeological%20Archiving%20in%20Europe/ARCHES_V1_GB.pdf	Europe
The Standard and Guide to Best Practice in Archaeological Archiving in Europe	http://archaeologydataservice.ac.uk/arches/attach/The%20Standard%20and%20Guide%20to%20Best%20Practice%20in%20Archaeological%20Archiving%20in%20Europe/ARCHES_V1_GB.pdf	Europe
Quality management of 3D Cultural Heritage replicas with CIDOC CRM	http://ceur-ws.org/Vol-1117/paper6.pdf	Europe
Guidelines for geographic information	http://www.athenaeurope.org/getFile.php?id=787	Europe
Richards J., Niven K. & Jeffrey S. (2013): Preserving our Digital Heritage: Information Systems for Data Management		

and Preservation, pp. 311-326, in: Ch'ng E. et al. (eds.): Visual Heritage in the Digital Age. Springer Series on Cultural Computing,		
Guidelines for archaeological Measurements	http://www.bundesdenkmalamt.at/documents/621701608.pdf	Austria
CCO Cataloguing Cultural Objects (CC0 Toolkit)	http://cco.vrafoundation.org/index.php/toolkit/	United States
KNA Kwaliteitsnorm Nederlandse Archeologie	http://www.sikb.nl/doc/KNA33/defitief/0.%20Voorblad%20KNA%20versie%203.3.pdf	Netherlands
idai.vocab		Germany
IT Empfehlungen - Dateiformate, Forschungsmethoden, Projektphasen	http://www.ianus-fdz.de/it-empfehlungen/	Germany
DANS General Conditions of Use	https://dans.knaw.nl/en/about/organisation-and-policy/legal-information/DANSGeneralconditionsofuseUKDEF.pdf	Netherlands
DANS Preservation Policy	https://dans.knaw.nl/en/deposit/information-about-depositing-data/DANSpreservationpolicyUK.pdf	Netherlands
DANS Data Management Plan for managing, documenting and sharing data	https://dans.knaw.nl/en/deposit/information-about-depositing-data/DANSdatamanagementplanUK.pdf	Netherlands
DANS Research Data Selection guidelines	https://dans.knaw.nl/nl/over/organisatie-beleid/publicaties/DANSselectionofresearchdata.pdf	Netherlands
DANS Preferred Formats	https://dans.knaw.nl/en/deposit/information-about-depositing-data/DANSpreferredformatsUK.pdf	Netherlands
Specialist Recommendation for data and Metadata	https://wiki.de.dariah.eu/pages/viewpage.action?pageId=20058160	Germany, Europe
Referentienetwerk erfgoed - Erfgoedthesaurus (Reference Network heritage - Heritage Thesaurus)	http://www.erfgoedthesaurus.nl/	Netherlands
Validating the Digital Documentation of Cultural Objects. In International Conference on Information Technologies for Performing Arts, Media Access and Entertainment (ECLAP). Porto, Portugal, pp. 104–117.	http://www.springer.com/us/book/9783642400490	Europe

RECODE (2015): Policy guidelines for open access and data dissemination and preservation	http://recodeproject.eu/wp-content/uploads/2015/01/recode_guideline_en_web_version_full_FINAL.pdf	Europe
DCH-RP - Digital Cultural Heritage Roadmap for Preservation - Open Science Infrastructure for Digital Cultural Heritage in 2020	http://www.dch-rp.eu	Europe
iDAI Thesauri	http://archwort.dainst.org/	Germany
Art & Architecture Thesaurus - AAT	http://www.getty.edu/research/tools/vocabularies/aat/about.html	United States
PACTOLS Thesauri	http://frantiq.mom.fr/fr/thesaurus	France
Archaeological Documentation	http://www.mnm-nok.gov.hu/wp-content/uploads/2013/01/b-ERD-szakmai-%C3%BAtmutat%C3%B31.pdf	Hungary
NWO Data Contracts	https://goo.gl/ewL2rL	Netherlands
A Framework for Transforming Archaeological Databases to Linked Ontological Datasets. In Computer Applications and Quantitative Methods in Archaeology	http://www.tracingnetworks.ac.uk/publications/CAA2010/paper.pdf	United Kingdom
Mapping Methods Metadata for Research Data	http://www.ijdc.net/index.php/ijdc/article/view/10.1.82/382	United States
Metadata for Research data: current practices and trends	http://dcpapers.dublincore.org/pubs/article/viewFile/3714/1937-	Canada

8.2. Collected policies in the field of Language Studies

Policy	Policy link	Country
CLARIN-D standards of language resources	http://de.clarin.eu/en/language-resources-and-services/standards	Germany
CLARIN Data Management Plan	http://de.clarin.eu/en/preparation/data-management-plan	Europe
ISO/TC37 Terminology and other language and content resources	http://www.iso.org/iso/home/store/catalogue_tc/catalogue_tc_browse.htm?commid=48104&published=on&includesc=true	World
ISO 639 International Standard for language codes		World

ISO 3166 International Standard for country codes		World
ISO 15924 Codes for the representation of names of scripts		World
TEI P5 Guideline- P5: Guidelines for Electronic Text Encoding and Interchange	http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html	World
DTA Basis-format (DTABf)	http://www.deutschestextarchiv.de/doku/basisformat_en	Germany
TEI Dictionaries	http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html	World
TEI-Markup	http://www.tei-c.org/Support/Learn/mueller-index.htm	World
TEI-HEADER	http://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html	World
CMDI	https://www.clarin.eu/content/component-metadata	
VIAF Virtual Authority File	https://viaf.org/	World
Data seal of approval	http://www.datasealofapproval.org/en/	World
OAI-PMH	https://www.openarchives.org/pmh/	World
CLARIN-D legal helpdesk	http://clarin-d.de/en/training-and-helpdesk/legal-helpdesk	Germany
CLARIN Legal Information Platform	https://www.clarin.eu/content/legal-information-platform	Europe
DARIAH-DE working papers on legal issues (working paper 6 and working paper 12: both in German)	https://de.dariah.eu/working-papers-beitraege	Germany
Creative commons	https://creativecommons.org/	World
Open Definition	http://opendefinition.org/od/2.1/en/	World
CLARIN: PID (Persistent Identifier) policy summary	https://www.clarin.eu/content/pid-policy-summary	Europe
FORCE11: Joint Declaration of Data Citation Principles	https://www.force11.org/group/joint-declaration-data-citation-principles-final	World

CLARIN: FAQ - Metadata in CLARIN: harvesting and VLO	https://www.clarin.eu/faq-page/275	Europe
European Commission: Policy on Research Ethics	http://ec.europa.eu/research/swafs/index.cfm?pg=policy&lib=ethics	Europe
CLARIN: Centre requirement (revised edition)	http://hdl.handle.net/1839/00-DOCS.CLARIN.EU-77	Europe
Checklist for CLARIN B centres	https://www.clarin.eu/content/checklist-clarin-b-centres	Europe
META-SHARE Licenses	http://www.meta-net.eu/meta-share/licenses	
META-SHARE Metadata Schema		
META-SHARE Policies	http://metashare.elda.org/info/	
Lexical Markup Framework (LMF): ISO-24613:2008	http://www.lexicalmarkupframework.org/	
GALA CRISP Standards and Guidelines for the Language Industry	http://lsrp.galacrisp.org/	
Simple Knowledge Organization System (SKOS)	https://www.w3.org/2004/02/skos/specs	
GÉANT Data Protection Code of Conduct	http://geant3plus.archive.geant.net/uri/dataprotection-code-of-conduct/v1/Pages/default.aspx	
DFG Practical Guidelines on Digitisation	http://www.dfg.de/formulare/12_151/	Germany

8.3. Collected policies in the field of Social Sciences

Policy	Policy link	Country
VSNU code wetenschapsbeoefening	http://vsnu.nl/files/documenten/Domeinen/Onderzoek/Code_wetenschapsbeoefening_2004_(2014).pdf	Netherlands

Ethical Guidelines from Association of Social Anthropologists of the UK and the Commonwealth (ASA)	http://www.theasa.org/downloads/ASA%20ethics%20guidelines%202011.pdf	United Kingdom, also used in Denmark
Statement on Ethics: Principles of Professional Responsibilities (American Anthropological Association)	http://www.aaanet.org/profdev/ethics/upload/Statement-on-Ethics-Principles-of-Professional- Responsibility.pdf	United States, also used in Denmark
Law on protection of personal information (Wet bescherming persoonsgegevens)	http://wetten.overheid.nl/BWBR0011468/2016-01-01	Netherlands
EU Directive on Data Protection	http://ec.europa.eu/justice/data-protection/	Europe
NWO Regeling Subsidies	http://www.nwo.nl/documents/nwo/juridisch/nwo-regeling-subsidies-1-december-2015	Netherlands
Reuse of Qualitative Data	http://www.socresonline.org.uk/12/3/1.html	United Kingdom
Preparing Data For Sharing. Guide to Social Science Data Archiving.	https://dans.knaw.nl/nl/over/organisatie-beleid/publicaties/DANSpreparingdataforsharing.pdf	Netherlands
Access Policies and Usage Regulations: Licenses (CESSDA -SAW)	http://cessdasaw.eu/calendar/webinar-access-policies-and-usage-regulations-licenses-30-06-2016-1100-am/	Europe
National Statement on Ethical Conduct in Human Research (2007)	https://www.nhmrc.gov.au/book/national-statement-ethical-conduct-human-research	Australia
Privacy Act 1988 (Cth) (and state equivalents)		Australia
Human Rights Act 2004 (Cth) (and state equivalents)		Australia
Freedom of Information Act 1982 (Cth) and amendments in the Freedom of Information Amendment (Reform) Act 2010 (Cth) (and state FOI and Right to Information (RTI) equivalents)		Australia
Australian Code for the Responsible Conduct of Research	https://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/r39.pdf?	Australia

NSF policy on the dissemination and sharing of research results	https://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4	United States
Data Documentation Initiative (metadata format)	http://www.ddialliance.org/	
Statistical Data and Metadata Exchange (SDMX 2002):	https://sdmx.org/?page_id=5008	
	http://www.dcc.ac.uk/resources/how-guides/license-research-data	
Directive 95/46/EC	http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:31995L0046	Europe

8.4. Collected policies in the field of History

Policy	Policy link	Country
HDML - History Data Management Lifecycle	**http://port.sas.ac.uk/mod/book/view.php?id=1220&chapterid=720**	United Kingdom
The Netherlands Code of Conduct for Scientific Practice (VSNU, 2014, NL)	**http://www.vsnu.nl/files/documenten/Domeinen/Onderzoek/The_Netherlands_Code%20of_Conduct_for_Academic_Practice_2004_(version2014).pdf**	Netherlands
Best practices for Oral History interviews	**http://www.oralhistory.org/about/principles-and-practices/**	United States
Policy for recording of Oral History interviews		United Kingdom (Royal airforce museum)
Interviewing Guidelines	**http://oralhistory.library.ucla.edu/interviewGuidelines.html**	United States (UCLA - Center for Oral History Research)
"Interviewer Stappenplan"	not online	Netherlands

"Veteranen Vertellen"		Netherlands

9. Appendix III: Questionnaire

The following questionnaire was utilized for the assessment of the current situation with regards to data management at the PARTHENOS partner institutions.

9.1. Structure of the questionnaire

General section

The general section is for data warehousing and organizing the answers later. This questionnaire is not supposed to be anonymized, but will be used for PARTHENOS Task 3.1 only. The consolidated version should not allow pointing answers to individual questionnaires or persons filling them in.

Specific part of the Questionnaire

The questionnaire has two dimensions: vertically you find the states in the life cycle, horizontally the FAIR principles. The Data Life cycle consists of Data Creation, Processing Data, Data Analysis, Data Preservation, Giving Access, Reusing Data. The columns represent the FAIR-principle: Findability, Accessibility, Interoperability, Reusability. We allowed for a column “other” if none of them applies but you still have an answer.

The answers should be inserted into the column which seems to be most appropriate to be the FAIR principle affected by the answer. If an answer affects more than one FAIR principle, the answer should be copied also into that field.

Duplication of information in the columns is allowed; if a FAIR principle does not relate to an answer to the question, the column is left blank. If an answer to a question is not available with regards to the organization or institution within PARTHENOS, the value should be n/a (“no answer”) in the “remarks” column. For additional remarks, the “remarks” column is being used.

9.2. Questionnaire

	Keyword for Question	Instruction (with explanation)
General/Basic Questions	Institution name	Name of your institution
	Your name	contact (who filled it in?)
	Your role	Are you filling in this questionnaire as: a Research Infrastructure provider such as CESSDA, CLARIN, DARIAH, ...; a repository and archive provider; a researcher creating data; in some other role.
	Disciplines	Which disciplines or subject fields do you cover with your institution? (Archaeology, History, Social Studies, Language based research, ...)
	Written data management policy	Does your infrastructure / institution have a written policy and/or written procedures that addresses data management? Please provide link to policy.
	Training of staff	Does your institution train and instruct employees on their responsibility for data management procedures?
Data Life Cycle State		
Data creation		
	Deadlines for data providing data	Do you ask for specific dates for when research data will be archived and when the research data becomes available? For the submitter, invited groups such as reviewers and associated researcher, the scientific community; the general public? (F; A)
	Responsibility for metadata maintenance	Do you define the responsibility of maintaining the metadata for cataloguing the data? Is it the data provider, the archivist, an infrastructure? Which metadata schemas do you support? (F;A) Are technical metadata or other metadata needed for preservation collected? (e.g. program + version generating data which may be needed for future emulation)
	Documentation of IPR to data	Do you make sure that you document the rights related to the data? Which information do you gather on the rights holder? How do you make sure that nobody is left out? Who has the right to modify access restrictions? (A)
	Accepted data formats	Do you have a list of accepted data formats? Which formats are these? How do you communicate this to your users? (I)
	Estimate size of data	Do you create an estimation of the data size to be created during the planning phase? Which units of measuring to you employ, such as file sizes, number of files, number of digital objects, etcetera?

		(I)
	Define granularity	Do you define the granularity of your archived data? I.e. sometimes data can be bundled in different forms, to use the book paradigm, it could be individual pages, paragraphs, chapters, parts, whole books, sets of books, complete works by author, works by year, etcetera (I)
	Quality Assurance specification	Do you specify Quality Assurance processes for data to be stored? Which QA processes are being used? (R)
	Specify required service level of repository	Do you specify reliability and service levels of a repository? Which certificates and methods of assessment are acceptable to you? (R)
	Budget	Do you estimate the costs of archiving? By which "units" of archiving, for example price per megabyte, price per digital object, price per number of backups, price per authorized user, price per file, ...? Who will be charged? (Other)
	Data management plan template	Do you have a template for a data management plan? Is your data management plan compliant to your disciplinary and ethical norms? Can you point to the template? (A, F)
	Adopted standards for digital content creation?	Have you adopted standards or best practices for digital content creation (digitization)? If so, please specify the standards used, link to the URL if online or attach a copy if the standards are locally customized or if best practices or guidelines have been developed.
	Data based on third party data	Do you have a policy on data that are created based on data of third parties?
Processing data		
	Disciplinary and ethical norms	Does your infrastructure ensure compliance with disciplinary and ethical norms? If yes, please state how (e.g. by anonymising data where necessary, etc).
	Relevance and understandability	Do you use defined criteria to ensure relevance and understandability of the data for data users? Please indicate how (e.g. by requiring a minimal set of metadata to be procured, ...).
	Integrity and authenticity	Do you ensure the integrity and authenticity of the data stored in your organization? Which processes do you use? / What kind of documentation or specification do you follow? Please give examples (e.g. using checksums, ...), (F, A)
	Legacy data	How do you handle legacy data provided to your institution? Do you create updated metadata? Do you review IPR? ...

Data analysis		
	Data publication workflow	Do you follow any data publication workflow? Can you provide the documentation of the workflow?
	Sufficient information for evaluation	How do you ensure the availability of sufficient information (technical data and metadata) for end users to enable them to make reliable quality-related evaluations, if the data allows it? If so please give examples (e.g. staff with specialized education or training, detailed metadata, special training course to use specialized infrastructure, ...)
	Managing IPR in the analysis process	Does your organisation consider copyright and intellectual property concerns in managing digital materials when data are being reused? How?
	Temporary and intermediate result storage	Do you have policies regarding the storage of intermediate results and temporary files? Possible aspects included for example privacy, licences, IPR
	Effects of analysis policies	Do you have policies in your analysis process having consequences for storing, preserving, citation, accessing, etcetera, of results?
Data preservation		
	Long-term preservation	Does your digital repository store and preserve your collections for the long term? Do you have policies for preservation? Describe what you do to enable long-term preservation of your digital resources
	Sustainability commitment and policies	Is digital sustainability considered as a clear organizational commitment and resources in your institution? Did you develop relevant policies on this topic? How are resources for preservation ensured over the long term?
	Collaboration with digital preservation initiatives	Does your institute/RI collaborate with other national and international institutions (including NREN, NGIs, data infrastructures and research projects) in digital preservation initiatives? Do you avail of existing policies on data preservation?
	Defined workflows for long-term preservation	Do you use defined workflows to ensure long-term preservation? Do they involve characterisation and how? Do they create metadata and documentation and which? Do they involve validation and how? Do they register audit trails and for what? Do they ensure bit integrity through workflow and how? Are their different workflows for different data and how? How are data related to a preservation plan?

	Tools to control risks while processing content	Did your organization/RI develop tools to control the risks associated with receiving, managing, processing and ingesting digital collection content?
	Preservation planning	How do you identify appropriate approaches and tools to prevent technological obsolescence? Do you have formalised technology watch to support this and how? How do you make plans for preservation actions for formats in risk? Do you have tools to support preservation planning and which? Do you rely on format registries and which? Do you rely on information collaborate efforts or other institutions (e.g. technology watch reports) and which?
	Metadata for usability	How do you ensure that there is appropriate metadata available to ensure the understanding and reuse of data over time? What types of metadata does it cover? What types of identifiers are used and how are they related to data? What type of standards are used? What type of Identifiers are used? How is structural metadata represented? How are descriptions of needed applications for accessing data represented?
	Influence of different formats	If your digital content is heterogeneous in format, type or in some other aspect, how does this influence your processes, e.g. in respect of operating systems, documentation? What are your considerations about which formats can be preserved? Are there different preservation strategies for different formats and which? Do you migrate to preservation formats before preserving and which?
	Data integrity (bit level)	Do you ensure continued authenticity and integrity of your digital resources throughout time? What processes do you follow? To what level do you do bit preservation? - how many copies, how independent are the copies (geographically, organisationally, hardware software etcetera), how often are copies checked individually, how often are checked for changes between copies? Do you support different levels of bit preservation? Do you include forensic aspects? What kind of packaging do you use for data under bit preservation?

	Data integrity (logical level)	What preservation strategies do you use? format migration? emulation? How is it decided what losses/risks of loss (in significant properties) that can be accepted (and thus which strategy to use)?
	Backup and monitoring	Does your system use automated backup processes? And if yes is an automated monitoring processes of storage available?
	Repository backend	What digital asset management system do you use? (This system may be used to manage the full life cycle of your digital objects including management of data creation, metadata repository, image repository, registry of preservation metadata, and a means of providing access to users, such as FEDORA or DSpace or a locally developed system) How does it support preservation?
	Legacy data preservation	How do you handle the preservation of legacy data? Do you create updated metadata? Do you review IPR? ...
Giving Access		
	Licenses and access	Do you use licenses which cover data access and use and how do you monitor the compliance? (What licence(s), which methods?
	Licenses and access	Do you and if so how do you ensure that users discover the data and refer to them in a persistent way, e.g. through proper citation and/or use of persistent identifiers?
	(general) stated access policy	Do you have a (written) access policy to the archived data, which e.g. states when and under which conditions a resource become available to: submitter; invited group (reviewer, collaborators); scientific community; general public?
	Access restrictions	Does your system have any restrictions on (public) accessibility? If yes, please describe the exceptions to public and free access.
	Laws and regulations	Is your archive/repository subject to national/European laws and regulations? If so, please indicate which and give examples of implications if there are any.
	Access control	Does your system allow to record and maintain metadata to restrict access and delivery of collections to authorized users? How?
	Discovery	Do you use a discovery service for the data? Describe what kind of discovery services your repository/infrastructure is offering, what resources are available, which metadata are searchable, etcetera. Also, are metadata of non-public resources publicly available?

	Resource identification	Do you use a mechanism or procedure for identifying resources? Describe how resources are identified, and what kind of landing page is being provided.
	Resource retrieval	Are resources retrievable from your repository? Describe how resources are being retrieved from your repository/ infrastructure, which interfaces and standards are being supported, including API's for indexing and object retrieval.
Reusing Data		
	Maintain tools for accessing data	Did you identify appropriate approaches and tools to prevent technological obsolescence?
	Managing IPR for reused data	Does your organisation consider copyright and intellectual property concerns in managing digital materials when data are being reused? How?
	Exploitation of data	What is your strategy on exploitation of data? Do you have a strategy and licence policy if third parties reuse data (e.g. creating revenue, combining it with other data, etcetera)?
	Citation	Do you require parties reusing data to credit the data creator, provider and archive of the data? Do you have recommended formats for that?

9.3. Consolidated answers

9.3.1. Data creation

For the data creation phase it can be observed, that there are not really commonly used general policies. As an example: one survey participant states for nearly all questions to have no general policy. In addition, there are many questions where survey participants give no answer. Probably this has something to do with the highly specific character of data creation and with the structure and rule of repositories in this data life cycle phase. That is why instead of policies a lot of commonly used practices are in place. They allow reacting on the needs of projects that deposit their data in a repository. Another point here is a general negotiation between a low-barrier integration of data, that often tries to simplify the integration process for data creators by freely allowing to put whatever data into the repository and only giving recommendations, and a more strictly regime that tries to control the data creation process by giving explicit policies to shape the data beforehand. That are two

opposed positions, that affects the work to be done in the following data life cycle phases. A third position is the one of data aggregators, that are not strongly involved into data creation, as this is done by the aggregated data holders. This implies to rely on this data holders and their data creation policies and workflows.

One important practice that often comes into play is an agreement on a data management plan (DMP; sometimes more specific: research data management plan) between the repository and a project. The DMP should be defined before data creation and allows with respect to the technical limits of the repository a project specific handling of many parameters of the data creation phase.

What may be missed in our survey is a question on policies handling references to data that is already deposited in another repository/archive/database. Is it possible to link to this data or is it necessary to re-deposit the data? We also didn't ask about differences between born-digital data and digitized data. Another interesting point would have been to ask how the logical structure of a data collection is documented and if it is possible to reconstruct this logic from zero.

It is in general striking that many answers emphasise the technical restrictions. In overall, policies and practices in the data creation phase do have a high tradeoff between the needs of projects – especially when there is a focus on specific disciplines – and technical motivated constraints.

Deadlines for data providing data

Technical restrictions are important for definitions of the **deadlines when do make research data available and archive this data**. When there is a DMP, then there are usually deadlines specified between the data provider and the data host. However, it is very different how far project specific adjustments are possible. That is because deadlines for data provision are strongly connected with the behaviour of the underlying technological stack. Data is often provided and archived immediately after pushing it into the repository. When there is the need of an embargo date, this can only be handled when the technical framework allows it. That is probably the reason why there is often no general policy on this topic, as it is implicit part of the technical workflow or individually done with the data providers.

Responsibility for metadata maintenance

Organisational structures are another important background for different approaches in the data creation phase. This comes often hand in hand with technical settings. Looking into **responsibilities for metadata maintenance** there are two different positions: the more common one claiming that the data provider is responsible and the other that the infrastructure takes over. The second point is done either by having trained staff in the organisation structure or on a technical level by only allowing distinct defined metadata schemas. Narrowing down to distinct defined metadata has the side effect of only supporting a small set of schemas, whereas other approaches enable the use of a wide set of metadata schemas. But it seems that this openness to schemas comes with the need of a strong supporting organisational structure, e.g. having repository/metadata managers. Additionally, it is important to have a defined workflow who is when taking over (sometimes defined by the DMP). Usually that means that data providers describe their datasets (sometimes by using specific tools) and the data holder does a quality check applying improvements or calling back to the data creator. One answer in the survey made the good point to credit all involved persons in this maintenance process. That could help visitors to rank the metadata quality.

Documentation of IPR to data

The metadata records are also the place where the **IPR to data is documented**, in most cases this means documentation of the rights holder and the rights conditions. It is clear, that the data creator has to submit this data and to take care of it. This commitment is defined in the DMP or via specific agreements and/or in the metadata capture form. Often there are documentations and guidelines for data creators on IPR issues or even consulting personal. Some data holders insist on permissions or intervene actively. In any case, the answers do not give specific information how – besides the commitments – it is made sure that nobody is left out. Modification of access restriction is usually done by the data creator, seldom by the right holder (probably due to technical reasons). Some data holders only allow open data, many give a recommendation on open data.

Accepted data formats

Accepted (or more precisely recommended) **data formats** are in the most cases documented in a list online. Nevertheless, there are still cases where such a list is

not available. This may be because a majority of the data holders accept any format, probably because of a low-barrier integration of the data. Instead there are strong attempts to encourage data creators using recommended formats, especially open formats. This is done with the help of the DMP, guidelines, training, in the preliminary project phase or by advising data creators that using not standard formats have the risk of limited reusability. Only a minority of data holders reject not accepted data formats, using a client, the data input form or a validator.

Estimate size of data

The **estimated size of data** that will be deposited by a data creator is in most cases not relevant for the data holders. Only 1/3 of the survey participants asks specifically for such data in the planning phase and/or in the DMP. This is almost information on the likely file sizes, usually number of digital objects, and sometimes number of files. More seldom there are restrictions on the file format, that implicit restricts the file sizes (resolution of images). It can be assumed, that the amount of data in file sizes is often not so relevant for projects in the covered disciplines and that exceptions are handled individually.

Define granularity

Granularity definitions of archived data comes often into play for structuring the data. Some of the participants do not have the possibility or need for granularity, but the majority has with very different approaches. Most of these approaches rely on the target disciplines of the data holder, e.g. for text documents that could be on a basic level a page that is then bundled in a book. Often collections are in use, connecting digital objects that are belonging together. A commonly used structure for granularity is a triple: single items that are fold together into a more generic digital resource, these generic digital resources attributed to collections. There is in many cases also the possibility of free decision on the granularity, based on the needs of the data provider. In one case the granularity is connected with the allocation of persistent identifiers (PID), something that was not in the focus of the question but would be interesting to ask further.

Quality Assurance specification

Quality assurance processes and specifications have a strong relationship to correct metadata and the use of standards like preferred data formats. That means that defined responsibilities for metadata maintenance and urging on accepted data formats have a strong effect on a quality assurance process. Only some of the survey participants do not have a quality assurance in use. For the others, it is connected with the organisational structure. If there is trained staff, e.g. data collection managers, then there is high commitment to quality assurance processes, as it seems that automatic processes – as some has in use – do not have a satisfying effect (especially if metadata is incomplete or wrong and data format standards are not used). An interesting approach – that relies probably on the availability of resources – uses prototypes or a review process before the data is stored, which guarantees a good handling of possible quality problems with data. This approach can be stretched when it combines automatic checks on a technical level with manual checks on a more content-related perspective.

Specify required service level of repository

For the specification of **reliability and service levels of a repository**, it can be observed that survey participants that are members of a Research Infrastructure like CLARIN, do apply their service levels from there (if such assessment is available by the Research Infrastructure). Others do not have such procedure or rely on established certificates and assessments like the data seal of approval (DSA). The DSA is widely accepted as a guarantee for reliability of a repository. Also, the NESTOR seal is in use. For members of CLARIN the CLARIN centre assessment is significant. Data collectors establish reliability by only allowing interoperability and exchange of data with repositories of dedicated and trusted partners. This is often combined with support of self-assessment. Only one survey participant has established internal standards and procedures that are used for self-certification and the raise of trustworthiness (relying on the Trustworthy Repositories Audit and Certification – TRAC document by the US Centre for Research Libraries). Some of the survey participants also indicate to audit ISO 16919 (Space data and information transfer systems – Requirements for bodies providing audit and certification of candidate trustworthy digital repositories, http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=579

50). For the reliability aspect, the need of strong defined and widely accepted certificates and assessments is obviously and already covered by some players.

Budget

It is quite astonishing that a majority of survey participants didn't give an answer on **costs of archiving and charging** this cost. This could be because there is no archiving, as it is the case for data harvesters. It could be also that as public funded institution the archiving service is without charge. More likely, this is something that is done very project specific. From the answers we got, there are in nearly every case limitations to free charging, be it limitations to a specific scope of projects, limitations to national research, or limitation that depend on the volume and size of the deposited data. The last aspect is affiliated with the question on an estimated size of data that will be deposited and requires discussion in the project planning phase and/or in the DMP. Missing information on charging of cost may be connected with time limitations of projects that only allow for a one-time charge of this project (at the expense of the project budget) or a charging of the institution, where a project is located, e.g. a University. Only one survey participant, who established a graduated contract system, discusses this issue. Also, only one survey participant gives information on a regular archiving cost evaluation. Improvements need to be done on communicating charge of costs, because here again only one survey participant has a published Charging Policy with a transparent cost list (based on number of files, size of files and complexity of the files).

Data management plan template

A small majority of the survey participants has a **data management plan template** or at least refers to one. That is a pleasant development. At the same time, there is potential to convince even more data hosters to offer DMP templates. Half of the positive survey answers point to an individual template that can be used online (most of them in English). A smaller set uses wizards like DCC DMPOnline. Other DMP references involves individual DMP generation without having a template online and referring to DMP templates of grant agencies. We missed to ask if a DMP is obligate, recommended, or not necessary for depositing data. This would have given us some hint on the importance of DMPs. At least one survey participant indicated to require a DMP as part of a project agreement.

Adopted standards for digital content creation?

We asked for **adopted standards or best practices in the process of digital content creation/digitization**. One third of the survey partners did not answer or have not adopted standards/best practices. That could be because digitization is not available or relevant for these survey partners. On the contrary, one third of the survey partners refer to a digitization workflow or to guidelines on digitization, all of them online whereupon not always in English (not necessary institutional guidelines but also guidelines on national level or by Research Infrastructures like CLARIN). The last third of the survey partners prefer adopted standards, most of them implicit or not documented online. For the digitization workflows/guidelines and adopted standards it can be seen that some distinguish on the foundation of data formats (e.g. text, audio) others between domains (e.g. museums, libraries).

Data based on third party data

The last question from the data creation section deals with **policies on data that is created on the basis of third party data**. A majority of the survey participants do not have such policies in place or do not see a reason for this, e.g. because such data is not accepted or it is assumed that the data creator is responsible for clearing the rights (this goes in accordance with IPR policies on data). If there is a documented policy then either a permission by this third party has to be submitted (including license agreements) or data access is restricted (or deleted depending on the rights situation). The workflow for such cases is often documented in the DMP. One survey participant has established a guidance workflow to find solutions for data that is created on base of third party data (which is part of a general guidance on IPR).

9.3.2. Processing data

With respect to the research data life cycle, processing data is generally understood as the steps a researcher or research team would do in preparation for their analysis.[161] In the context of this report, where the question has been asked to data

[161] UK Data Archive - Research Data Lifecycle: http://www.data-archive.ac.uk/create-manage/life-cycle.

archives, the focus has rather been on how data is being processed and handled within the archive. The questions aimed to uncover ways in which the data service providers target a number of essential qualities in their data processing, as part of delivering their services. The questions have included the following topics:

- Disciplinary and ethical norms
- Relevance and understandability
- Integrity and authenticity
- Legacy data

Disciplinary and ethical norms

The guiding question for this area was how and to which extent the data archive in question ensures compliance with disciplinary and ethical norms. The respondents follow a number of different approaches and refer to various sources of authority on the norms in question. With respect to disciplinary norms, Archaeological Data Service is offering extensive guidelines on handling of sensitive data, defining concepts and describing examples being of relevance for archeology. In other cases, respondents refer to national or European Code of Conduct documents, which could be said to establish a broader norm for a good scientific practice. With respect to ethical norms, these seem in most cases to be handled by referring to the legal framework (laws, licenses, agreements) that has been established for their enforcement.

Strategies in order to ensure compliance differ from leaving it all up to the user to comply, over guidance and policies, legal contracts, technical control mechanisms, and to completely avoiding to accept any data that could potentially imply legal or ethical issues. The legal concerns relate to two main areas: personal data which may need special treatment and legal agreements and intellectual rights and copyright. Unexpressed behind some of the replies, is the question of who will be held responsible in case of misuse. Some of the strategies seem to try to limit any possible responsibility on the service provider. Specific strategies that are being used, comprise:

- Helping researchers to uncover and deal with ethical and privacy issues in

their data management plan, by including these areas in a DMP template.

- Providing a full set of legal documents to use for deposition and end user licenses, consent forms, data processing agreements, etcetera.
- Support for anonymisation of personal data.
- Constraining access to potentially sensitive data.
- Restricting the possibility to deposit potentially rights protected data.
- Referring researchers to relevant Code of Conduct documents with guidance for good practice.
- Offering comprehensive guidance on deposition of sensitive digital data.

Relevance and understandability

The guiding question in this area concerns whether and how defined criteria are used to ensure data relevance and understandability for the research communities. One group of replies relate to how data can be described and made discoverable. This set of replies revolves around various types and formats of metadata and how they are being used. Another group of replies relates to the quality of data, dealing with various kinds of review, quality assurance and repository certification mechanisms.

A lot of the user's options to assess data relevance and understandability depend on metadata, how data is being described. The different data service providers offer descriptions by a number of different categories of metadata: some use minimal descriptive metadata such as Dublin Core, while some offer description of data according to disciplinary metadata standards or vocabularies. Using standards or ontologies can provide one of the paths towards data interoperability and are being used for that purpose. Another strategy towards interoperability is mapping different metadata formats to a common format within the infrastructure. Some partners emphasise the use of rights holder and license metadata as well.

Metadata are being used in various ways. In some cases, recommendations are provided, while leaving it up to the depositor to decide how to use metadata, "the data producer is in charge". Others are requiring a minimal set of descriptive metadata, often enforced to some extent by the deposition workflow. This can include automatic checks for schema or standards compliance, well-formedness and the like.

One partner has a policy for relevance, depending on distinct criteria on intellectual content, the preservation potential and reuse value for potential future users, as well as the likelihood that the data might not already be preserved and accessible from other sources.

Quality assurance of some kind plays an important role in securing relevance and understandability. This ranges from entirely manual control by a board of specialists or domain experts in order to guarantee the quality and scholarly standard, to automatic procedures. Manual quality checking can comprise securing completeness, clarity and documentation of data. Some partners rely on automatic tests being performed during the deposit workflows, in some cases followed by manual inspections.

Integrity and authenticity

The guiding questions for this topic were whether the integrity and authenticity of the data stored by the service providers, is being ensured, which processes are being used for that, and what kind of documentation or specification is being followed. Again, the replies range from applying entirely manual procedures like peer-review of deposited data to various automatic procedures for particularly fixity checking. Adding to this are some more long-term related strategies relating to cases of format migration and deposition of new versions of datasets which have already been archived.

Integrity and authenticity are to some extent being understood as related to repository internals, and some replies simply relate to DSA conformance or using unspecified Fedora mechanisms. A number of service providers apply regular fixity checking according to MD5, SHA-1 or SHA-256 hashes. This is known as a method for discovering possible “bit rot”, but is also being used at various stages of the deposit workflow, to secure correct transferral of data files to the repository, as well as to secure intermediate steps such as internal format translations. One partner is supplementing regular fixity checking with geo-replication, maintaining a complete copy of their data archive at a remote site.

In one case, filenames and metadata are being listed as part of the license agreement, allowing end users to check for completeness. Some repositories have mechanisms for versioning of data and particularly also strategies for always keeping

the originally deposited version, in case format translations or deposition of new versions of data should fail at some point.

Interesting enough, nobody mentions the specific use of provenance metadata, beyond versioning, or any automatic methods for verifying authenticity of the author and data. This area still seems to be entirely relying on manual evaluation and assessment - or to missing information, if automatic methods exist.

Legacy data

Legacy data can be a challenge to repositories, as they may not adhere to current standards of metadata and data formats, licensing, etcetera. We asked the participating service providers how they handle legacy data that gets transferred to them, and whether they create updated metadata and review IPR. Replies on this question reveal a high number of different approaches, from leaving it all up to the users/data providers to take care of, to involving archivists in regular preservation activities, according to specific criteria.

The issues mentioned mainly revolve around IPR, and to what extent it is possible to give access to the data, when to create updated metadata and when to perform preservation actions on data themselves. A number of archives have procedures in place for updating all these three mentioned components. In some cases, it is mentioned that the data provider must take part by agreement. For some archives, they offer several tiers of service, depending on data's compliance to pre-defined lists of file formats and their suitability for long-term preservation. The lower the tier of the service, the less guarantee is provided for the long-term.

A number of the archives that offer support for legacy data, perform translations of metadata into either disciplinary formats or infrastructure specific formats. Also, IPR information is being updated in some cases, others prefer to keep access to legacy data very restricted in case of any doubt.

As a general observation, it would seem that the more the data provider keeps participating actively in the data stewardship, the more it would seem likely that data repositories and infrastructures will be able to provide continuous access and preservation over a longer time.

9.3.3. Data analysis

The data analysis has been discussed in terms of five different aspects: data publication workflow, sufficient information for evaluation, managing IPR in the analysis process, temporary and intermediate result storage, and effects of analysis policies. It was understood that data analysis targets not only the original analysis, but a secondary use of data and the reconstruction of analysis results. This requires the publication of data in some form, storing underlying data plus intermediate results.

Data publication workflow

The discussion of the data publication workflow was guided by the question if a data publication workflow exists and is documented with the participating institutions in PARTHENOS.

The projects within PARTHENOS follow different approaches for a data publication workflow. In general, the following classes can be distinguished:

- No policy and data publication workflow
- Data publication workflow is a requirement but is developed or described on a project by project basis
- The workflow is procedurally defined by archiving workflow specifications and policies such as OAIS and Data Seal of Approval. The concrete implementation is still on a project-by-project basis.
- The publication of the data is technically integrated in the archiving workflow system, i.e. each data set being archived undergoes the publication procedure

Within PARTHENOS the technical integration and the fixed workflow defined as a policy are the minority.

Sufficient information for evaluation

For the reuse of data and reconstruction of results it is obviously essential to understand the research data. The information on the data is usually provided in the form of structured metadata or detailed description. The policies to ensure the quality

and coverage of the metadata was part of the questionnaire. The answers can be grouped as follows:

- The metadata intended to be provided for data management needs to be described by the data providers.
- Archivists receive extensive training on how to provide metadata, which should result in consistency, high quality and allow the archivists to assist the data providers (for example in the context of ISO 24622-1).
- The archive requires fixed metadata schemas (for example METS-MDI) with some mandatory fields, usually technically evaluated if these fields are filled in
- A technical scoring mechanism is applied to measure the quality of metadata, based on ISO 24622-1.

Most partners focus on training and procedural aspects, most technical solutions address the existence of values for mandatory fields only.

Managing IPR in the analysis process

In the context of reuse and reconstructing results, a number of legal and ethical issues may be relevant, starting with the intellectual property rights of the creator, contributor, owner of the primary data, creator of the analysis, interviewee, archive, etcetera. Traditionally, researchers were not afflicted, when they cited their sources and used the short quotations, and did not share their large collection of data of which there were other rights holders. With the digital turn, these become relevant also for archives, infrastructure providers and researchers in the Humanities and Social Sciences. The rights include not only the redistribution of data, publication of data on the web or elsewhere, but also possible restrictions to additional analysis as they might constitute derived work prohibited by some otherwise open licences. Hence dealing with IPR is an issue not only for publication but also for analysis and the institutions need to be aware of this.

The complexity of the problem is well represented in the variation of answers to this question, showing that most institutions are aware of the problem but a general direction is far from available and clear.

a. Attempt at generalization
    i. References to the national IPR legislation
    ii. Per institution definition of policy for open licences (not always obvious which)
b. Expert and competence pool
    i. working groups
    ii. training of experts
    iii. Documentation of best practices and recommendations
c. Individual specification on the level of digital objects
    i. Specification of restrictions in the metadata for each resource
    ii. Granting of access and usage rights on a case by case basis (“upon request”)
    iii. Standardized end user licence agreements to be accepted for each resource with details on reuse and distribution restrictions
    iv. Contractual relations between depositor and archive with indemnification and liability clauses

Temporary and intermediate result storage

Intermediate results in an analysis process are often not visible in publications, which usually refer to the underlying data and the results. However, often the analysis performed by a researcher can be rather complex with intermediate results being produced. The individual steps can also be the source of problems when reconstructing results as a change in the process, adjustment of parameters for automated processes. Storing intermediate results may enhance the possibility to reconstruct results.

In general, it turned out that there are basically two groups of institutions in PARTHENOS: those not being part or operating a virtual research environments, so they do not deal with temporary results, and those who operate analysis tools and virtual research environments. The latter institutions are aware of some issues, but there are no general rules. Intermediate results are not addressed in terms of archiving.

Effects of analysis policies

The application of some analysis tools may come with their own usage restrictions (“not for military use”, “results may not be commercially applied”), which may affect the applicable restrictions for the result as well. One solution proposed was to specifically ask users not to process privacy sensitive data and to evaluate possible legal constraints when applying analysis tools on a case by case basis. Mechanisms to automatically assembling the constraints, for example based on the metadata, are not in place anywhere.

9.3.4. Data preservation

Data preservation is a phase of the data life cycle, which includes all activities needed to ensure continued access to digital resources and the information they enclose. Policies, procedures and best practices adopted within the PARTHENOS consortium, have been investigated through a series of specific questions. The questions, addressed to Research Infrastructure providers, data creators, data managers and so forth, focussed on aspects regarding organizational matters and technological concerns.

The questionnaire included the following main points:

- Long-term preservation
- Sustainability commitment and policies
- Collaboration with digital preservation initiatives
- Defined workflows for long-term preservation
- Tools to control risks while processing content
- Preservation and planning
- Metadata for usability
- Influence of different formats
- Data integrity (bit level)
- Data integrity (logical level)
- Backup and monitoring
- Repository backend
- Legacy data preservation

Long-term preservation

The investigation about the long-term preservation process was conducted asking partners if their digital repository can store and preserve the collections for the long term and what is the process they follow to enable the long-term preservation of their digital resources.

From the analysis of the answers received it comes up that some of the respondents do not have a repository or in other cases formal processes and systems for long-term digital preservation are absent.

The majority of the respondents declared that their infrastructure/institution own a repository and that long-term preservation of digital resources is ensured by keeping the software update, performing incremental and periodic backup, and adopting specific technical requirements. Strategies to ensure the long term include:

- asking the maintainers of the repository to provide sufficient funding for running it for at least 10 years
- accepting only formats that are considered to be suitable for long-term preservation, as for example:
  - XML
  - RDF
  - OASIS
- offering rudimentary preservation metadata created and recorded using a semi-manual process
- cooperating with university libraries and other national stakeholders

One institution only has a DSA, while four of the institutes interviewed adopt a Preservation Policy which are formally documented here:

- https://dans.knaw.nl/en/deposit/information-about-depositing-data?set_language=en
- https://www.cines.fr/en/long-term-preservation
- http://archaeologydataservice.ac.uk/advice/preservation

Sustainability commitment and policies

Sustainability of digital resources is an important aspect as it has a direct impact on data preservation. The question asked in this part of the questionnaire aimed at

understanding whether the development of policies on digital sustainability is considered a clear organizational commitment by their institutions/Research Infrastructure.

Some respondents stated that they have no repository. In cases where a repository exists, two of the respondents said they are still working on developing a digital sustainability plan.

Institutions replying that the digital sustainability is a clear organizational commitment, adopt international standard formats (XML, and RDF textual formats, XML format according to the METS-MDI schema, and the MAG schema), use preferred formats guidelines, perform checks on submitted metadata and execute periodic backup and software update.

Assigning a Persistent Identifier to ensure findability over the long-term is also a strategy for sustainability. Another institution provides different level of digital preservation:

- Snapshots on the NAS (Network Attached Storage) for "hot data".
- Distributed copy on their distributed file system (Active Circle) for "luke warm data".
- Backup on LTO tape drive for "cold data".
- Long-term preservation (+/- 20 years).

According to this questionnaire, only two archaeological institutions developed policies on digital sustainability.

Collaboration with digital preservation initiatives

Collaborating with national and international institutions that face the challenge of digital preservation may benefit an institution or Research Infrastructure from similar tools developed or solutions achieved. The replies to this question show that (most of) the institutions that replied to the questionnaire are collaborating with other initiatives or institutions that focus on the preservation aspect.

PARTHENOS partners are collaborating with national and international initiatives, including national research and education network (NREN), Grid providers (National Grid Initiative – NGI), federated infrastructures, international projects and infrastructures contributing to a sustainable access to research data, special interest groups.

The list of initiatives Research Infrastructure and institutions of the PARTHENOS consortium are collaborating with, includes:

- EUDAT (a Service-oriented, Community driven, Sustainable and Integrated initiative)
- CINES (National Computing Center for Higher Education)
- ZIM-ACDH (member of ICARUS that contributes to DARIAH-EU working group on preservation)
- Data Seal of Approval
- CIDOC CRM SIG to ensure the maximum standardization and preservation of data, by developing international core and domain ontologies and standards
- National CLARIN consortia
- EGI
- Federation in the Authorization and Authentication Infrastructure (AAI)
- Digital Preservation Coalition. This group is responsible for collating experience and advice from partners, and developing tools and guidance on all aspects of Digital Preservation
- International projects like: ARIADNE, CESSDA, DCCD, EHRI, EOSC, HAS, KNOWeSCAPE, OPENAIRE, Re-SEARCH, INDIGO-DataCloud

Defined workflows for long-term preservation

Analysing the replies to the question: “do you use defined workflows to ensure long-term preservation?” it is possible to identify three different approaches. There are institutions/Research Infrastructures that have no general policy for the long-term preservation, others that have policies at an embryonic form (with workflows compliant with OAIS and deflected by FEDORA mechanisms), and others which follow a defined long-term preservation workflow.

The strategies followed by the different RI/domain are reported below:

- PID redirection, metadata conversion (including transformation into MARC 21 and MODS), moving digital objects, checking the integrity of the copy by md5 checksum; preserving access restriction and access control lists

- CMDI for metadata. The integrity of the data is checked using FLAT.
- Use of a defined long-term preservation workflow, that includes: aggregation of metadata from data providers, mapping to ACDM, metadata enrichment via MoRe, metadata validation and publication. All these processes are driven by the solid framework defined by the Synergy model developed by the CIDOC CRM SIG.
- Perform actions according to internal protocols, related to the Preservation Policy. The actions are registered via checkboxes. Upon checking a checkbox, the archivist's user ID is registered with the action along with a timestamp.
- One partner has a long-term preservation strategy for libraries and museums digital collections. The metadata profiles integrate technical information on digital objects, useful for ensuring bit integration through MD5 file integrity check. Tools for creation and automatic validation of the metadata are available on-line. Recommendations, applications and information are available on-line.

Tools to control risks while processing content

Identifying and managing risks concerning stabilization and quality assurance of data stored into digital repositories is a fundamental activity in the preservation phase. We have asked organization/RI if they have developed tools to control risks while processing content. From the analysis of the results it is possible to identify situations where there are no policies or tools developed yet, and others with a solid workflow and tools developed to overcome the risks linked to receiving and storing digital content. The tools adopted by the RI/institutions include the following features:

- checking/syntactic parsing of data structures
- mechanisms to secure the reception and storage of exact copies of the original files (ingestion phase)
- tools for generating metadata and for automatic validation of the XML
- virus scanner for scanning file uploads
- technology vulnerability scan, the SLA with the data storage provider, a procedure for file fixity checking, an annual DRAMBORA Risk Assessment

as well as the Declaration of Confidentiality for employees and a periodical safety inventory

- bespoke Content Management System (CMS) with Object Management System (OMS) extension. This includes a number of applications (primarily in Java) that help the Digital Archivists through every stage of the ingest and preservation process. At a simple level, these are checklists which ensure specific tasks (e.g. validity checks, virus checks) are carried out before and archive can be formally accessioned. A more advanced application is responsible for comparing checksums stored within the OMS database, and reporting on any discrepancies encountered
- FLAT: a repository solution based on Fedora Commons

Preservation planning

The definition of a preservation plan is part of the strategic planning an institution/RI should develop to ensure the long-term preservation of the managed digital resources. We have asked representatives of the different domains involved in PARTHENOS whether planning a preservation strategy is an approach followed in their daily management practices. Particularly, researchers and experts of the various domains have been asked if they identify appropriate approaches and tools to prevent technological obsolescence and formats in risk, and if these are monitored by formalised tools or technologies.

In the minority of the cases there are no specific policies developed or followed to define a preservation plan. In other cases, when some planning is done, it does not include the use of specific formalised tools.

The analysis of the feedback from the respondents who declared that some preservation planning is done in their Institution/RI, shows that minimizing the risk of technological and formats obsolescence is a major concern, and consist in:

- using FEDORA
- relying on OAIS
- enforcing standardized data formats
- ensuring stability of the technological framework
- using long-term, large community supported backends and cycle review of

the hardware and backend (for the repositories)

- technology to avoid the risk of technological obsolescence
- uses of scanning tools to identify files using FITS and Apache TIKA
- use the DPC Technology Watch reports to monitor for preventing formats obsolescence

In other cases, specific tools and services are used to prevent technological obsolescence. In particular, in the archaeological domain, the ARIADNE project avails of the MoRe service for data mapping, conversion, enrichment and validation, providing advices and monitoring against possible risks. Technological obsolescence is overcome by keeping the modules that compose the infrastructure constantly updated.

The use of preferred formats is common in few repository providers. One of them has a Preferred Formats workgroup which keeps the Preferred Formats guidelines up-to-date, by keeping informed with developments, actively researching new formats upon contact and by regularly performing guideline reviews.

Another institution, follows the principle of preservation via migration, standardising formats during ingest and then maintaining a technology watch. When a format is in danger of becoming obsolete - or another format offers a more sustainable solution - then all instances are migrated to this new format. Particularly, when a format is identified as being 'at risk', a member of curatorial staff compiles a report on the number and location of all instances of that format. Plans for migration are then made, including tools and staff time necessary.

Collaboration with other institutions is another strategic point followed by some of the respondents.

For one of the respondent common procedures and workflows, shared internationally, would reduce the cost both in terms of time and money to be allocated for storage and long-term preservation and would contribute to the general interoperability and openness of scientific digital cultural heritage data. The so-called ‘hard sciences’ are already demonstrating that research can advance its capability by the use of e-Infrastructures offering high-speed connections, shared computing and storage resources, sophisticated authentication and authorisation mechanisms etcetera.

Metadata for usability

Metadata are important information about digital resources, fundamental for understanding and reusing data over time. Metadata may refer to descriptive, administrative and structural information of the digital object. We have asked partners how they make sure that appropriate metadata are available to ensure the understanding and reuse of data over time, what type of metadata are requested and if they use identifiers and of what type.

The replies can be divided in two groups. A group of respondents did not reply to this question, probably because there is no such approach in their institution. The rest of the replies shows that very much attention is paid to this matter and that metadata must accompany each digital resource. In particular, most of the institutions require a minimum set of metadata and require a persistent identification to ensure that data is understandable and reusable. A special committee for the quality assurance of metadata is available at three of the interviewed institutions, which is in charge of checking the quality, ensuring integrity and authenticity of metadata. A metadata capture form with detailed advice and guidance is used to capture metadata by another partner.

To ensure the understanding and reuse of data over the time some institutions adopt the following standards:

- Component Metadata Infrastructure (ISO 24622-1) to create an environment that supports different metadata schema.
- MAG and METS-MDI schemas: Dublin Core, VRA, NISO, MD5, METS.
- ACDM
- CIDOC CRM
- (Qualified) Dublin Core metadata fields.
- NAKALA or ISIDORE
- Dublin Core elements for collection/thematic metadata.
- GEMINI for spatial terms
- LOD terms such as LCSH, TGN and Heritage Data are used within metadata.

Most of the queried institutions/RI recommend the handle system to assign a persistent identifier to the digital resources; others adopt URI, Digital Object Identifiers etcetera.

Influence of different formats

By and large, partners were asked if the type of data's format they wanted to preserve (or that they were asked to preserve) had some influence on their process. They were also questioned about existing preferable format(s) for preservation according to the type of data which has to be processed.

In general, the partners who answered about their preservation policy according to the format of the data concerned deal with heterogeneous contents which implies to have several preservation strategies. Most of the partners mentioned some format and technical requirements, some do have a list of, or guidelines on preferred formats, but sometimes did not clarify them. Anyhow the main impact apart from possible technical problems is time to process quantity of data in heterogeneous format and the guarantee of the accessibility of the data after years. That is why some partner preferred formats which have open standards, are well supported and do not rely on the use of specific software or platforms.

If we get into details, some infrastructures which manage (but don't own) repositories have specific requirement on metadata provided which are assessed by a quality assurance committee for that. Some others have a Data Seal of Approval (DSA) which clearly indicates the preferred formats which they accept to deal with[162]. In order to encourage data format standards and stability, some Research Infrastructures also encourage the use of Handle system and PID policies. To preserve the quality of data, another disciplinary infrastructure prevent from using any data with compression and encourage open standards use (like ISO) and XML-based and lossless format.

Then, we had the case of a national infrastructure who set up a partnership with the National Centre in charge of providing high calculation level, preserving digital data in the long term and hosting national computing platforms. The strategy of this Centre is to have a "validating tool" which will indicate if the format and the format's version of the data you want to preserver and you entered in the system is

[162] E.g https://assessment.datasealofapproval.org/assessment_100/seal/html/.

correct or not according to the standards that the Centre chose to support. What is more, it is a migration tool. That is to say that if a user's data (files, videos, pictures, etcetera) are in an inappropriate format, the tool allow the user to migrate in a format that the Center could process. We also find this kind of existing tool in another disciplinary infrastructure who made their own in-house tools and procedures for each data type.

Data integrity (bit level)

In that section of the questionnaire, we mainly asked the partners how do their institution was ensuring the continued integrity of the data preserved throughout time, to what level did they do bit preservation and do they support different levels of bit preservation. In general, partners have different level of bit preservation and various strategies linked to it.

Some partners adopted a partnership's strategy to be more efficient. For example, they use a client to check well-formedness of XML-formats and validate against the referenced schema, also applies on metadata and to ensure the conformity to the DSA. The service of file integrity monitoring and risk management plan can also be managed by a national Grid Infrastructure as a partnership named that offers hosting service to the platforms concerned.

Some others built in-house processes which generally include systematic backups of digital resources which can requires integrity checks for the DSA and backup facilities. However, the bit preservation can be done using different types of supports (disks, tapes) and technologies like Irods[163] or geographically distributed systems[164].

Finally, there are also hybrid strategies that store checksums within their OMS database with regular validation and where the validation of checksums is undertaken through a bespoke tool (in Java) within interfaces with the OMS. Integrity checks (checksums) are timetabled to ensure that all files are validated prior to re-commencement of the cycle.

[163] http://cc.in2p3.fr/spip.php?page=article&id_article=2062&lang=en.

[164] Active Circle (http://www.active-circle.com) mainly between Paris & Lyon for CNRS. for DANS, a copy of the original dataset is always archived as submitted, even if files are migrated or metadata is added/changed. Data is stored on two separate remote locations with daily back-ups and checksums are used to verify authenticity of data.

Data integrity (logical level)

We then focused on data integrity at logical level asking questions about the processes of risk of loss evaluation and the ones of decision of accepted losses.

In general, software updates and migration processes are carried out by the partners who answered (or by their own partners) the questions. So far, they do not always enter in details on the ways the format is preserved and how long. Some of them preserved data in state of the art or accessible XML formats, if possible but leave format conversion into newer versions to the individual archive.

Contrary to some partners who don't have general policy on data integrity at logical level, some partner centres use open source solutions like FEDORA version control or develop other strategies.

One partner enlightens its migration processes and choices: data are managed per general type of data, it is assessed what would be the most open and well-supported format for the data which would still retain the file's significant properties, this assessment is guided by the Preferred Formats guidelines.

If an open variant would exist but would cause a loss of significant properties, it may be deemed relevant to still migrate and provide both the export and the original file; or to check whether the properties could be retained in a different manner. Anyhow, the original file is always archived as submitted, even if it is migrated to a different format.

Another partner also use migration leaded by a series of internal procedures for different data types. For example, their procedures for geospatial files list Coordinate reference system information, Geometry, Attribute fields and source elevation model, bit-type, colour map, pixel type (for rasters) as the key elements to be preserved. In special cases where content varies from a typical file, an individual may have to make a decision about procedures to take. Quite often this is reported back to their colleagues and if viable, used to inform future events.

Backup and monitoring

At this stage, partners were asked if they use automated backup and if so, if an automated monitoring processes of storage was available. In general, the partners who answered the questions on backup and monitoring do have in-house automated backup and schedule it on their own or developed partnerships to do so. Some others do it too and their SharePoint sites are backed up to Microsoft cloud.

Then, two examples can be developed. Firstly, one partner provides backups with "snapshots" and "policies" for distributed backups and tapes storage.
For another partner, that is a repository and data provider, integrity checks (checksums) are timetabled to ensure that all files are validated prior to re-commencement of the cycle.

Repository backend

In this section, we asked the partners what digital asset management system they were using. We specified that this system may be used to manage the full life cycle of their digital objects including management of data creation, metadata repository, image repository, registry of preservation metadata, and a means of providing access to users, such as FEDORA or DSpace or a locally developed system. On this particular point, we noticed that few of the partners who answered have no established policy concerning repository backend.

Several partners build their repository backend on solutions totally or partly based on FEDORA[165], in particular:

- Platforms are based on FEDORA
- objects are stored in FOXML format containing all binary data streams in base64 encoding and all data streams are preserved in the original format as distinct files;
- Fedora-Commons, DSPACE or self-developed back ends
- Escidoc/Fedora Commons
- EASY, a self-digital depositing system certified with both a Data Seal of Approval and a NESTOR Seal, which is built on FEDORA. DOI and URN Persistent Identifiers are assigned to a dataset upon submission.

Then, other developed in-house solution or partnerships include:

- locally developed system called NAKALA[166]. Preservation of digital data is made by the underlying infrastructure through various tool, at different levels[167];
- locally developed SharePoint based system, OAI-PMH compliant, integrating

[165] See:http://www.fedora-commons.org/.
[166] https://www.nakala.fr/.
[167] See above the “Sustainability commitment and policies” section of this document.

with Aleph and Primo library management system but does not support preservation;

- software GATTO and locally developed systems;
- bespoke Collections Management System (CMS) with Object Management System (OMS) extension. The OMS holds details of all their files, and has been developed to may bitstreams (i.e. files) to larger objects - this helps maintain the links between objects such as ESRI shapefiles that may comprise multiple elements.

Legacy data preservation

Finally, the questionnaire on the long-term preservation topic focused on how partners handled the preservation of legacy data. Partners were also asked if they created updated metadata and if they did IPR review. From a general point of view, those topics seems more complicated for the partners and may need some clarifications, in particular regarding the IPR review.

On one hand, several partners did not mention their policy on legacy data or answered that they did not have (or not yet). On the other hand, few others adopted a clearer strategy.

9.3.5. Giving access

By answering nine questions, institutions provide a complete overview of their strategy for giving access to data within the PARTHENOS project. “Giving access” has thus been discussed in terms of eight different aspects:

- Licenses and access
- General stated access policy
- Access restrictions
- Laws and regulations
- Access control
- Discovery
- Resource identification
- Resource retrieval.

To give online access to their data, institutions use preferably open source solutions like the repository softwares DSpace[168] or Fedora Commons[169] - associated with locally developed application on top. Only the King's College London is deploying a commercial solution, the web-based application Microsoft SharePoint. Fedora is sometimes implemented in the framework of the e-research environment e-SciDoc[170] (which includes the Fedora repository as a core functionality). It also serves as a baseline to develop new repository solutions like FLAT[171] (CLARIN) or EASY[172] (DANS). Some partners use locally developed systems based on open source technologies, like GATTO[173] (CNR-OVI) or NAKALA[174] (CNRS-HumaNum).

Licenses and access

In terms of licenses, institutions clearly support the Open access movement. Most often, they don't impose specific licences before archiving data, but recommend - in their terms of use[175], a depositing agreement or a Data Seal of Approval[176] - that data be as open as possible. Whenever possible, they nurture the use of Creative Commons licenses or equivalent and interoperable licenses[177]. However they underline that repositories assume responsibility from the data providers for access and availability of their data. Therefore, they don't monitor that users comply with their requirements in terms of use. Thus data providers have to respect national laws and regulations on their own initiative.

In terms of access, search engines like CLARIN VLO[178] or ISIDORE[179] allow data to be findable by end-users. Institutions generally ensure that their users can refer to data in a persistent way with PID such as Digital Object Identifier (DOI), or at least stable URLs or permalinks. Proper citation is recommended: using DataCite guidelines as a standard and at least citing the data creator(s) are seen as good scientific practices.

168 http://www.dspace.org/.
169 http://fedora-commons.org/.
170 https://www.escidoc.org/.
171 https://www.clarin.eu/sites/default/files/trilsbeek-windhouwer-CLARIN2016_paper_16.pdf.
172 https://easy.dans.knaw.nl/ui/home.
173 http://www.ovi.cnr.it/index.php/en/il-software-2/scarica-il-software-gatto.
174 https://www.nakala.fr/.
175 http://archaeologydataservice.ac.uk/advice/termsOfUseAndAccess.
176 https://assessment.datasealofapproval.org/assessment_100/seal/html/.
177 https://www.clarin.eu/content/license-categories.
178 https://vlo.clarin.eu/;jsessionid=7010394F408F0E8A4FF5CC9C25EDC444?0.
179 http://www.huma-num.fr/ressources/videos/isidore-une-plateforme-de-recherche-pour-les-shs.

General stated access policy

Less than a half of our partners have a written access policy to the archived data. These policies rely on three main aspects:

- Content providers decide if they make their data freely available or not.
- Setting up specific restrictions on data are possible. Data providers are able:
    i. to define an embargo period on data;
    ii. to limit access to specific data.
- Licences of use define different access rights according to:
    iii. the type of users: different rights can be granted to the submitter, the scientific committee, and the general public.
    iv. the openness of archived data: they are available through open access, or only for registered users, or in restricted access.

Access restrictions

Except three of them which systematically made the archived data freely available, data repositories have many different policies with regards to access restrictions. Even if they generally recommend free or public access, repositories enable data providers to limit access to the archived data if necessary. To access restricted data, users must be authenticated (for instance, with Shibboleth authorization system), must register (for instance, access with a password), or must wait until the end of the embargo period. Besides users are not able to access some stored data whose metadata records are not public. Access can be restricted because of:

- the nature of data: legal, ethical, or IPR protection reasons.
- data providers' needs: initial period of preferential use; confidential, contractual protection reasons.
- the type of users: academic access or personal use.

Laws and regulations

All data repositories are subject to national laws and regulations. MIBACT-ICCU mentions that it is also subject to the European legislation on reuse of public sector information, which fosters the production and publication of interoperable open data

sets, open standards, data formats, ontologies and vocabularies[180]. These national laws and regulations focus on:

- protection of personal data,
- IPR/ copyright,
- protection of databases,
- freedom of information/public access to government information.

Access control

The multiple solutions adopted to control access to data reflect the complexity of this question. However, it is possible to identify some key elements.

- Metadata about access rights can be generated from the chosen depositor license.
- To manage access rights, institutions generate log files[181]. Some of them use access control lists associated to Shibboleth authorization.
- A password is generally required to access restricted data.
- Users are administered internally.
- Different access modes are possible: for instance, public, academic or personal use.

Discovery

The discovery services offered by the data repositories within PARTHENOS rely on three main aspects:

1) Metadata: m
    - Most of the time, they are publicly available. Repositories can even choose to make findable metadata for non-public resources, through a Triple Store or an OAI-PMH server.
    - Metadata formats are generally not specified, but we can note the use of the Dublin Core schema by DANS.
    - Data repositories are OAI-PMH compliant for distributing metadata.
2) Discovery services:

180 https://ec.europa.eu/digital-single-market/en/european-legislation-reuse-public-sector-information.
181 http://www.meertens.knaw.nl/cms/nl/collecties/bescherming-gebruikersgegevens.

- Data are discoverable via the website of the data repositories, and sometimes via digital platforms like Europeana. ISIDORE provides both a REST API[182], a RDF 3Store[183] and a Web interface for metadata discovery.
- These portals enable users to search within the metadata fields associated to the data. They can offer a full text search[184], an advanced search[185], a faceted search[186] or a search by collections[187].

3) Search engines: very little information is available. For instance, MIBACT-ICCU's platform use search engine based on LUCENE and SOLR.

Resource identification

Persistent identifiers are recommended by institutions to identify resources. Institutions suggest to use the Handle System to assign these persistent identifiers.

PIDs can adopt multiple forms: OAI Identifier, URI or DOI. Some institutions identify resources with permalinks or library management system numbers (for instance, an Aleph System[188] number). Users can be instructed to cite data in a standardized manner referencing creator, organization, date, title, and PID.

Resources' landing pages can be delivered by a specific software, like a library resource management (for instance, Primo[189]). Landing pages are based on the detailed metadata, and may display:

- A title;
- An abstract description;
- The resource type;
- Subjects;
- Keywords;
- Contributor(s);
- Publisher(s);
- Place(s);

182 http://api.rechercheisidore.org/.
183 http://rechercheisidore.org/sparql.
184 http://www.mirabileweb.it/ricerca_globale.aspx.
185 https://easy.dans.knaw.nl/ui/advancedsearch.
186 http://portal.ariadne-infrastructure.eu/.
187 http://archaeologydataservice.ac.uk/archives/.
188 http://library.harvard.edu/lts/systems/aleph.
189 http://www.exlibrisgroup.com/category/PrimoOverview.

- A date or a time period;
- Rights associated to the datasets;
- Language.

Resource retrieval

The aim of this question consisted in understanding how resources are being retrieved from repositories. But institutions' answers mostly vary according to their understanding of this question. Nevertheless, some key elements have been identified.

1) Discovering online resources

Resources can be retrieved online through the institutions' portal. If users meet with the access conditions, they can have access to the resources. EASY (DANS) allows authorized users to directly open files supported by the browser (images, PDFs) and/or to download selected datasets from the landing page.

2) OAI-PMH harvesting

OAI-PMH harvesting is also performed to make resources from repositories available via other portals/interfaces, such as the CARARE portal in Europeana. For instance, NAKALA makes metadata accessible through OAI-PMH and by a Triple Store. One institution provides a SOAP web service and a number of warehouse management systems (WMS) which allows metadata to be incorporated within the Heritage Gateway.

3) Searching resources

The Federated Content Search is used by one infrastructure to retrieve publicly accessible data, by using the FCS API which is based on SRU/CQL.

Search engine based on LUCENE and SOLR has been developed by one institution. In the other cases, very little information has been provided. However we can mention the use of the FEDORA search interface.

Reusing Data

The research institutions that hold and give access to data are public institutions or public founded projects that are responsible for respecting national and international laws. The Directive on the reuse of public sector information is an incentive for open data policy and it encourages public sector institutions to make as much information available for reuse as possible and to foster the production and publication of interoperable open data sets, open standards, data formats, ontologies and vocabularies.

Limitations for re-using data is generally due to personal data protection, copyright issues, database rights expressed by National laws and regulations.
By analysing the best practices in the PARTHENOS consultation on research data management emerged a clear awareness that data sharing is necessary to promote research integrity and collaborative opportunities.

For allowing (meta)data reuse it's necessary the publication of rich metadata to describe these data and to enable discovery the content. Metadata should also support data citation and include information about provenance to facilitate verifying that the specific version and/or granular portion of data retrieved subsequently is the same as was originally cited. The data fields and metadata schema should be accessible, together with the details of any access restrictions, whether or not the underlying data can actually be accessed.

By publishing and sharing datasets or descriptions of datasets, researchers:

- Are aware of the expansion to the scale and impact of their research.
- Can cite the research data and other researchers can refer to it, which can exploit research impact.
- Support increased collaboration and reduced duplication: research datasets become more findable and discoverable.
- Foster research integrity: the validity of research results can be recognized.
- Allow for innovative applications: research data preservation allows for the application of developing analytical technologies within a field of research.

Every dataset deposited in organization repositories must be accompanied by a formal deposit licence for allowing the dissemination under specific Terms and Conditions of use which are explained in a Copyright and Liability Statement and Common Access Agreement. Different reuse condition for teaching, learning and

research and providing material is appropriately cited.

Cultural institutions and research centres need to know copyright and intellectual property of digital materials when data are being reused and (meta)data rights holder should be identified before data publishing. In fact both research data producers and consumers should know their rights and responsibilities accordingly to laws and policies, and follow the policies process related to data sharing principles and legal interoperability of their institution. This process is strictly necessary in the workflow for publishing “open data” where data providers make their data available and usable to others within the rules, and data users take advantage of the data that are made lawfully accessible and usable.

To establish who or what entity/person has the rights to any given collection of data before the data are used or disseminated to others, Cultural Heritage Institutions, Research Infrastructures, research centres or researchers may use different procedural activities to grant or request permission:

- To grant permissions for copyrighted material upon written request.
- To request permission:
  - to authors for online publishing;
  - to publishers for online republishing of printed works;
  - to persons appearing on audio-visual materials;
  - to reproduce places, monuments, artefacts in audio-visual and other media;
  - to library owning copy of rare texts in public domain.

However, it could be possible that cultural institutions and research centres repositories deals with digital objects or data where the copyright status has not been determined with certainty. Those resources could be published with appropriate statements for explicating that the data provider has not undertaken an effort to determine the copyright status of the work.

The consultation between PARTHENOS research communities on data reuse showed that research and cultural institutions can share research data in two main ways:

- make the data available through open (meta)data and open access modality
- allow restricted access to the data

For allowing data-reuse some organizations developed a plan for reviewing tools and programming languages. Where there may be danger of obsolescence (e.g. updates of Apache) plans are made for possible upgrades/impacts/mitigations etcetera.

Open meta(data) sharing

There are services, such as CLARIN discovery service, ARIADNE Catalogue or Easy managed by DANS, which allows researchers and research organisations to publicise the existence of research data and collections online. Those services help researchers to find, access, and reuse data and content from different research organisations, and cultural institutions.

The OAI-PMH standard is generally adopted as repository and discovery service, and metadata are open available, that it means that are freely available to use, reuse and redistribute and the only restriction could be attribution and share alike.

For discovering and finding meta(data) advanced search and faceted browse services which target the qualified Dublin Core metadata of the datasets are adopted. Metadata for non-public resources are made also available, for giving high-level information on protected data and content.

Searching/browsing does not require users to log-in and all qualified Dublin Core metadata are publicly available. Some services such as ISIDORE, CulturaItalia provides an API and/or a Triple Store for metadata discovery.

To support data findability, it could be convenient to identify high-level facets for browsing the gathered information, such as the ARIADNE portal where is possible to discover meta(data) selecting one or more of the following facets:

- Resource type. Every resource in the portal is categorized with a resource type. The type can be any of the following options: Fieldwork archives, Event/intervention resources, Sites and monument databases or inventories, Scientific datasets, Artefact databases or image collections, or Burial databases;
- Native Subject. Subjects from a vocabulary used by the original owner of the resource.
- Derived Subject. Subjects derived from mapping native subjects to Getty AAT vocabulary terms;
- Keyword. Keywords or tags describing the resource;

- Contributor. The agent responsible for describing the resource in the Catalogue;
- Publisher. The agent responsible for making the resource accessible;
- Place. Place names the resource is connected with;
- Period. Time periods the resource is associated with;
- Rights. Access rights connected to the resource;
- Language. Language of the resource.

Restricted access

Free access and reuse of research data must be balanced against legitimate interests of the rights holders or for protection of confidentiality and for protection of cultural resources, as determined by law through the restriction or the control of the use of such data.

Restricted access can be applied to research data which is stored in a data repository. Researchers can access data through a password-controlled access to the research data whilst allowing for discoverability and global awareness of their research. Restricted access is general applied to:

- Research data or cultural content with commercial potential;
- Research data containing personal data;
- Research data containing culturally sensitive information;
- Third party data on are active intellectual property rights or other limitation due to contractual agreements.

There could be different categories of users for the restricted access: public, internal administration, academic and individual; organized by identity federation system (Shibboleth on base of academic institutions network including AAI such as Geant)

Metadata about access rights generally are generated from the chosen depositor license.

Data exploitation

Resources can be retrieved, if allowed by the end-user license with respect to the user in question (login required for non-public resources). To each resource/data object is assigned an ID that could be PID, OAI Identifier and URI. Each

resource/data object has a landing page, that displays title, as well as the abstract description from the Dublin Core metadata, to give users a clear overview of what the resource is about. Resources could be downloaded from the landing page.
If data are protected, indexing is in general not possible from external resources due to legal restrictions and in this case exploitation requires detailed investigation of the individual licence agreements.

For data repositories such as EASY, users see the files of a dataset of which the 'visibility' is set to 'anonymous' - it is possible for an archivist to change the setting to 'none', should files be kept with the dataset for archiving purposes but not for publication (for example: non-anonymised privacy sensitive data; original files where a migration is provided for accessibility). If users meet with the access conditions and agree to the General Conditions of Use (in a pop-up window upon an Access request), they can either directly open files which are supported by the browser (images, PDFs) or select one or more files by using checkboxes with the files, then 'download' the selection. The download is provided as a ZIP and includes the Institution General Conditions of Use, Checksum information, and file-specific metadata in XML (if present).

OAI-PMH Harvesting is also adopted in order to make resources from the repository available via other portals/interfaces, such as the CARARE portal in Europeana.

Researchers are stimulated to provide their data in preferred or accepted formats according to the Institution guidelines. The Guidelines may include approaches and tools identified for accessing data, depending on the type of file used.

EASY is a self-depositing system. Upon submitting a dataset a depositor has to agree to the terms of our licence-agreement. In this agreement rights such as copyright are covered. The depositor carries the responsibility. End users need to agree on the DANS terms of use before using data, with one exception: data deposited under CC0 licence. For data not licensed under CC0 the DANS terms of use apply. For CC0 datasets, all forms or reuse is allowed. The use of CC0 as a license is stimulated in the depositing process: it is the standard setting in the module on access options. Reuse and depositing with a CC0 license is in general strongly stimulated by DANS.

For data not deposited under CC0 license the terms of use oblige users to cite the data creator and archive. A citation suggestion is given, DataCIte. See: https://dans.knaw.nl/en/search/about-reusing-data?set_language=en

The same citation is also requested for data licensed under CC0 as a good scientific practice.

10. Appendix IV: EU and national regulations to promote access and data reuse

Title	Summary Description
A. Open data	
European Commission: Communication on Open data An engine for innovation, growth and transparent governance (COM(2011) 882) http://www.europarl.europa.eu/RegData/docs_autres_institutions/commission_europeenne/com/2011/0882/COM_COM(2011)0882_EN.pdf	In December 2011, as part of the digital agenda for Europe, the Commission presented a communication on open data, presenting its vision on providing favourable framework conditions for the use and reuse of data.
EU Directive on the reuse of public sector information (the PSI Directive, 2003 and 2013) Directive 2003/98/EC http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:175:0001:0008:EN:PDF	The PSI Directive, sets out the general legislative framework at European level for government data. It provides for a minimum degree of harmonisation. It triggered a shift in the culture inside public administrations towards greater openness and is a key pillar of the open data policy. In June 2013 a revision of the PSI Directive was adopted. The revised PSI Directive (Directive 2013/37/EU) brings about important improvements. The reuse of public sector data, whether for commercial or non-commercial purposes, should fully respect EU and national privacy legislation as well as the intellectual property rights of third parties. Member States were obliged to transpose Directive 2013/37/EU by 18 July 2015.
The EU Inspire Directive (2007) http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32007L0002	The Inspire Directive entered into force in May 2007, applies to geographical information. It requires EU Member States to make such information available, provide descriptions of it in the form of metadata and enable its

	reuse by means of open standards (http:// inspire.ec.europa.eu/index.cfm/pageid/2/list/7). The directive has established a European spatial information infrastructure to support an integrated approach to European environmental policy.
Decreto Crescita 2.0 (Decreto Legge n. 179/2012) http://bit.ly/1xZrpAX	This law aims to strengthen the concept of transparent administration, fostering the "open by default" strategy. According to this law, in fact, all data published by public administration, in absence of a specific license, must be considered freely and reusable by everyone. Moreover, data published online must be issued with a licence, at most, that mention the attribution, without any other limitation.
B. Open access	
Denmark´s National Strategy for Open Access Open Access is a matter of getting maximum value for research Ministry of Higher Education & Science http://bit.ly/1RsTXLq	The strategy states that the implementation of Open Access is to take place through the green model – i.e. parallel filling of quality-assured research articles in institutional or subject-specific archives (repositories) with Open Access. However, the strategy does not exclude the use of the golden model as long as it does not increase the expenses for publication. Two central principles form the basis for the strategy. The implementation of the Open Access is to support the possibility for Danish researchers to continue to publish in the most recognised national and international journals, and also the possibility to publish. For the sake of research and society, it is stressed that it is crucial that the aggregate public expenditure to research publication is not increased significantly because of the implementation of Open Access
Open Access Policy for Public Sector Research Council and Foundations of 21 June 2012 (Denmark) http://bit.ly/2mMd10r	"This policy means that published scientific articles which are the result of full or part financing by research council and foundations must be made freely available to everybody via Open Access with the permission of the

	magazine. Requirements for the grant holder: The grant holder is - if the magazine allows it - requested to parallel-publish a digital version of the final, peer-reviewed scientific article which has been accepted by a scientific magazine. The article which is a result of full or part financing by research councils and foundations must be parallel-published in an institutional or subject-specific repository, i.e. a digital archive. The parallel publishing of the scientific article can - at the request of the magazine - take place after an embargo period, i.e. a period in which the article is only available in the scientific magazine, of up to six or twelve months after publication in the scientific magazine. The waiting periods for the specific research areas must be as follows: Health science - 6 months Natural science - 6 months Engineering science - 6 months Agricultural and veterinarian science - 6 months Social sciences - 12 months The Humanities - 12 months The final, peer-reviewed scientific article which is subject to parallel publication must include all graphic and other materials prepared for the article. Research data shall be excepted. The grant holder is responsible for making sure that relevant publication or copyright agreements with publishers are in accordance with the conditions for grants laid down by research councils and foundations in connection with parallel publication. Such conditions shall be observed according to current copyright rules. Which types of publication are included? The request for parallel publication only includes articles in magazines, i.e. serial publications or series with a scientific aim and which are published through an analogue or digital publication channel with routines for quality assurance through peer review. This means that the request for parallel publication does not include:

	monographs anthologies books popular science articles, i.e. articles processed by journalists without quality assurance through peer review."
Change of copyright: right of secondary exploitation (?) (Germany) https://www.gesetze-im-internet.de/urhg/__38.html	Der Urheber eines wissenschaftlichen Beitrags, der im Rahmen einer mindestens zur Hälfte mit öffentlichen Mitteln geförderten Forschungstätigkeit entstanden und in einer periodisch mindestens zweimal jährlich erscheinenden Sammlung erschienen ist, hat auch dann, wenn er dem Verleger oder Herausgeber ein ausschließliches Nutzungsrecht eingeräumt hat, das Recht, den Beitrag nach Ablauf von zwölf Monaten seit der Erstveröffentlichung in der akzeptierten Manuskriptversion öffentlich zugänglich zu machen, soweit dies keinem gewerblichen Zweck dient. Die Quelle der Erstveröffentlichung ist anzugeben. Eine zum Nachteil des Urhebers abweichende Vereinbarung ist unwirksam."
National Principles for Open Access Policy Statement (Ireland) http://bit.ly/2nVzCrF	"Common Principles within this policy: a) Policy confirms the freedom of researchers to publish where they feel most appropriate; b) this policy is intended to increase visibility and access to outputs of research funded by the Irish State; c) the policy is designed to support the free flow of information across national and international research communities; d) Policy is based on recognised best practice in keeping with original recommendations of the EURAB Policy on Open Access in relation to scientific publications (2006) General Principles 1. Peer reviewed journal articles and other research outputs resulting in whole or in part from publicly-funded research should be deposited in an Open Access repository and made publicly

	discoverable, accessible and re-usable as soon as possible and on an on-going basis. 2. Repositories shall release the metadata immediately upon deposit. Open access to the full text paper should be made immediately upon deposit or upon the publication date at the latest. 3. Researchers are encouraged to publish in Open Access Journals but publishing through Open Access Journals is not necessary to comply with this Open Access policy. Payment of additional Open Access charges through the 'Gold' Open Access model is not necessary to comply with this policy. 4. A repository is suitable for this purpose when it provides free public access to its contents, supports interoperability with other repositories and with other research information and reporting systems, is harvestable by national portal/s and international aggregators and takes steps toward long-term preservation. 5. Research data should be deposited whenever this is feasible, and linked to associated publications where this is appropriate. "
Position statement on Open Access to research outputs in Italy (Italy) http://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legge:2013;91	CRUI and EPR, being aware of the benefits of Open Access for National Research in terms of visibility, promotion and dissemination, commit themselves to act co-ordinately in order to achieve the success of Open Access in Italy by: a) encouraging the creation of OA repositories and technological infrastructures. These infrastructures should be implemented according to International interoperability standards and will capitalize on the OpenAIRE portal or other initiatives to be developed within the European Research Area (ERA) in order to maximize the visibility of European research; b) encouraging researchers to make their research results (data and publications) available in OA journals or institutional or subject

	repositories. Research results deposited in Open Access repositories should be made available in their post-print or publisher's version upon publication, and no longer than 12 months after their publication; c) contributing to an effective fulfilment of Open Access principles through the adoption of Institutional policies and rules asking researchers to deposit their publications in Institutional OA repositories. If such a repository does not exist, researchers should use other institutions' or subject repositories to deposit their publications and data.
Letter to Parliament on Open Access http://bit.ly/2ne1Al8	This is not a regulation, but a letter of the Minister to Parliament announcing the intention to make Open Access mandatory by 2016 (?) applying the Golden Road
CRNS - A guide to promote a complete and responsible research (France) http://www.cnrs.fr/comets/spip.php?article91	This guide is mainly informative. It develops a brief analysis of the difficult that the researchers (researchers, teachers-researchers, accompanying researchers) can be find during their work. It also makes recommendations on, among other things, good practices in terms of: publications, data processing, opening of results to the scientific community, communication.

11. Appendix V: CLARIN deposition license agreement

AGREEMENT FOR ADDING OPEN LANGUAGE RESOURCES TO THE CLARIN SERVICE – DEPOSITION LICENSE AGREEMENT (CLARIN-DELA-PUB-v.1.0)

1. Parties

<CLARIN CENTRE>, <CLARIN CENTRE CONTACT INFORMATION>, throughout this Agreement “Copyright curator”, and

<COPYRIGHT HOLDER NAME>, <COPYRIGHT HOLDER CONTACT INFORMATION>, throughout this Agreement “Copyright holder”

2. Scope and Intention of the Agreement

With this Agreement, the Parties regulate their rights and obligations concerning the use and distribution in the CLARIN Service of the Resource of the Copyright holder.

3. Definitions

“Resource” means material owned by the Copyright holder as defined in this Agreement, including software, applications and/or databases.

“Specifications” are any functional, technical or content-related requirements on the Resource, as defined in Appendix 1 of this Agreement.

“Update” means making the content of the Resource up to date by, e.g., correcting, amending or substituting data with new content to adapt the Resource to the technical infrastructure.

“CLARIN” means all parties representing national consortia according to paragraph 6.2 in the Statutes of CLARIN ERIC, EC decision 2012/136/EU, including <CLARIN CENTRE> representing <NATIONAL CLARIN CONSORTIUM>.

“Trusted Centre” means a CLARIN technical Service Provider which supports a reliable authentication and authorization interface such as an A or B level Centre specified in the CLARIN ERIC Technical and Scientific Description.

“CLARIN Service” means the distribution of Resources to users via Trusted Centres by CLARIN.

“End-User” means a user of the CLARIN Service.

4. Resource Subject to the Agreement and its Deposition

4.1 Identification of the Resource

This Agreement applies to the Resource described and specified in Appendix 1.

4.2 The obligations of the Copyright holder

The Copyright holder is responsible for depositing the Resource in compliance with the Specifications.

5. Delivery and Approval of the Resource

5.1 Delivery of the Resource

The Resource is delivered to the Copyright curator in the electronic form defined in the Specifications.

5.2 Verification and Approval of the Resource

After receiving the Resource, the Copyright curator validates the Resource within reasonable time and notifies the Copyright holder about the approval of the Resource for distribution. Should the Resource fail to comply with the Specifications, the Copyright curator either corrects the detected errors or requests a new version of the Resource from the Copyright holder.

5.3 Ownership

The ownership of the Resource remains with the original Copyright holder or holders. A copy of the Resource and the ownership of its physical carrier deposited by the Copyright holder are transferred to the Copyright curator at the time of delivery.

6. Maintenance and Updates

The Copyright holder has the primary right to update and maintain the Resource. Should the Copyright curator and the Copyright holder fail to agree on the maintenance of the Resource, the Copyright curator has the right to update the Resource or employ a third party to maintain and update the Resource for technical purposes.

After the termination of the Agreement, the Copyright curator has the right to update the Resource or employ a third party within the scope of this license to maintain and update the Resource for technical purposes.

7. Intellectual Property Rights and Access Rights

7.1 The intellectual property right and/or other rights governing the Resource subject to this Agreement belong to the Copyright holder or his licensors. Any third-party content of the Resource is identified in Appendix 2.

7.2 The Copyright holder makes the Resource available according to one or several of the licenses enclosed in Appendix 3 and identified below:

[ ] The latest version of the Creative Commons ZERO.

[ ] The latest version of the Apache license.
[ ] The BSD-2 license.
[ ] The BSD-3 license.
[ ] The GPL v.2 or later.
[ ] The LGPL v.2 or later.
[ ] The EUPL license.
[ ] The Eclipse Public license.
[ ] The MIT License.
[ ] The Microsoft Public License (MS-PL).
[ ] Princeton Wordnet

[ ] The latest version of the Creative Commons BY.
[ ] The latest version of the Creative Commons BY-SA.
[ ] The latest version of the Creative Commons BY-ND
[ ] The latest version of the Creative Commons BY-NC-SA.
[ ] The latest version of the Creative Commons BY-NC.
[ ] The latest version of the Creative Commons BY-NC-ND

[ ] META-SHARE Commercial No Redistribution
[ ] META-SHARE Commercial No Redistribution No Derivatives
[ ] META-SHARE Noncommercial No Redistribution
[ ] META-SHARE Noncommercial No Redistribution No Derivatives

Additional rights to the Resource may be agreed separately in writing.

7.3 Information about the license is to be published in conjunction with the Resource in accordance with the terms of the license. A sample End-User license agreement is enclosed in Appendix 4.

If the Resource is made available by the Copyright holder with the Creative Commons ND condition, the following still holds: “The Resource can be modified for the personal use of the End-User or research group of the End-User, but such a modified Resource may not be distributed.”

If the Resource is made available with the Creative Commons NC condition, the following interpretation is made: “Government-funded or non-profit research projects, e.g. projects funded by <NATIONAL RESEARCH FUNDING AGENCIES>, are not regarded as gaining economic benefit even if a portion of the financing is contributed by companies.”

8. Compensation

8.1 Compensation

For licensing the Resource,
[ ] no compensation is paid to the Copyright holder.
[ ] the Copyright holder is paid _____ euro as non-recurrent compensation excl. VAT.
[ ] the Copyright holder is paid _____ euro as other compensation excl. VAT.

8.2 Payment

The compensation shall be paid within thirty (30) days from the date of the invoice. The date of the invoice is the date of the acceptance by the Copyright curator of the Resource.

Payment overdue will be subject to an interest on overdue payments in accordance with the interest law.

9. End-User Rights

The Copyright curator commits to informing the End-Users about the terms under which the Resource is licensed to the End-User and about the rights and obligations that follow from the End-User License Agreement.

10. Legal Obligations

10.1 The Copyright holder shall be responsible for holding a copyright or a sufficient license and/or other rights based on intellectual property law to the Resource and that any use of the Resource for the purposes compliant with this Agreement does not in any form violate any third party copyright or any other rights based on intellectual property law or other incorporeal right.

10.2 The Copyright holder is held liable for all damages and costs he causes CLARIN or the Trusted Centres in the CLARIN Service by breaching any of the obligations in 10.1.

10.3 Should a third party present a justified claim that the Resource violates the obligations in 10.1., the Resource can be removed from the CLARIN Service.

11. Liability for Damages

Each Party is liable for the damages it causes. The Copyright curator is also responsible for the damages caused by a Trusted Centre. The liability is limited to the direct costs and damages caused to the other Party. The liability limitation does not apply to damages caused by an intentional infringement or gross negligence.

12. Effectiveness, Termination and Legal Consequences of Termination

12.1 This Agreement takes effect when signed by both Parties and remains in effect until the Parties have fulfilled all their obligations in the Agreement, unless the Agreement is terminated in advance in accordance with section 13 of this Agreement.

12.2 The following terms of the Agreement shall remain in effect after the termination of the Agreement:

Section 6. (Maintenance and updates)
Section 7. (Intellectual Property Rights and Access Rights)
Section 10. (Legal obligations)
Section 17. (Applicable law and settling disputes)

as well as all other terms of the Agreement that the Parties have indisputably intended to remain in effect in order to distribute the Resource subject to the Agreement.

13. Termination of the Agreement

13.1 Both Parties have a right to terminate the Agreement with immediate effect upon written notice of termination in case the other party is in material breach of the Agreement and has failed to take corrective action within thirty (30) days after receiving written notice.

13.2 Effect of the Termination

Should the Agreement be terminated because of material breach of the Agreement by the Copyright holder, the Copyright curator has a right to continue to use the Resource as specified in this Agreement even after the termination.

Should the Agreement be terminated because of material breach of the Agreement by the Copyright curator, the Copyright curator must end all use of the Resource and return or delete the copies of the Resource in his possession.

14. Appendices of the Agreement

14.1 The appendices of the Agreement are:

Appendix 1: Description and Specifications of the Resource as well as its proper reference
Appendix 2: Resources licensed by third parties
Appendix 3: Open-source licenses selected in the agreement
Appendix 4: Optional appendices, e.g., a sample End-User license agreement

14.2 Should the text in this Agreement and the text in the appendices be contradictory, the Agreement prevails. Should the Appendices in this Agreement be contradictory, the Appendices apply in the following order:

1. Appendix 1
2. Appendix 3
3. Appendix 2

15. Agreement and its Amendment and Severability

15.1 This Agreement supersedes and terminates all previous agreements and understandings between the Parties, whether oral or written, with respect to the subject matter of the Agreement.

15.2 The Parties may amend this Agreement by mutual written agreement only. Other amendments are void. The amendments take effect when signed by both Parties.

15.3. If a provision of this Agreement is or becomes illegal, invalid or unenforceable in any jurisdiction, that shall not affect the validity or enforceability in other jurisdictions of that or any other provision of this Agreement.

16. Contact Persons, Notifications and Reports

16.1 The contact person for the Copyright curator is: <CLARIN CENTRE CONTACT PERSON>, <CLARIN CENTRE EMAIL ADDRESS>

16.2 Notifications or reports by the Parties concerning this Agreement are considered valid when they have been made in writing or by email to the following addresses:

Copyright holder: <COPYRIGHT HOLDER>, <COPYRIGHT HOLDER CONTACT INFORMATION>, <COPYRIGHT HOLDER EMAIL>

Copyright curator: <CLARIN CENTRE CONTACT INFORMATION>, <CLARIN CENTRE EMAIL>

16.3 The Parties can change the Contact persons or Contact information defined in this Agreement by informing the other Party of the change.

17. Applicable Law and Settling of Disputes

This Agreement shall be governed by the law of <COUNTRY>.

Disputes concerning this Agreement will primarily be settled through mutual negotiations between the Parties. Should the Parties fail to find a solution through negotiation, the dispute shall be submitted to the district court in <CITY>.

18. Copies of the Agreement

This Agreement has been made in two identical copies, one for each Party.

19. Place, Date and Signatures

_____ _____ 20<XX>

Copyright holder Copyright curator

_____ _____

12. References

DANS. 2015. “Preferred Formats. September 2015, Version 3.0.” Data Archiving and Networked Services (DANS). https://dans.knaw.nl/en/deposit/information-about-depositing-data/DANSpreferredformatsUK.pdf.

FORCE11. 2014. “Guiding Principles for Findable, Accessible, Interoperable and Re-Usable Data Publishing Version b1.0.” September 10. https://www.force11.org/fairprinciples.

FORCE11. 2014. “Data Citation Synthesis Group: Joint Declaration of Data Citation Principles.” *Joint Declaration of Data Citation Principles - FINAL*. https://www.force11.org/group/joint-declaration-data-citation-principles-final.

Hilbert, Martin. 2012. “How Much Information Is There in the ‘information Society’?” *Big Data, Significance*.

Mayer-Schönberger, Viktor, and Kenneth Cukier. 2013. “Big Data: A Revolution That Will Transform How We Live, Work, and Think.” In *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, 83–97. London: Mariner Books.

Mons, Bared, Cameron Neylon, Jan Velterop, Michel Dumontier, Luiz Olavo Bonino da Silva Santos, and Mark D. Wilkinson. 2017. “Cloudy, Increasingly FAIR; Revisiting the FAIR Data Guiding Principles for the European Open Science Cloud.” *Information Services & Use* 37 (1): 49–56. http://content.iospress.com/articles/information-services-and-use/isu824#x1-50011.

OECD. 2007. “OECD Principles and Guidelines for Access to Research Data from Public Funding.” OECD. Organisation for Economic Co-Operation and Development. https://www.oecd.org/sti/sci-tech/38500813.pdf.

PARTHENOS. 2016. “Report on User Requirements.” D2.1. https://goo.gl/3lwI5J.

Riley, Jenn. 2017. *Understanding Metadata What Is Metadata, and What Is It For?* NISO Primer Series. Baltimore, Maryland: National Information Standards Organization (NISO). http://www.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf.

Shannon, Claude Elwood, and Neil J.A. Sloane. n.d. *Collected Papers*. Edited by A. D. Wyner. New York: IEEE Press.

Wilkinson, Mark D., Michel Dumontier, Ijsbrand Jan Aalbersberg, Gabrielle Appleton, Myles

Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data*3 (March): 160018. doi:10.1038/sdata.2016.18.