PARTHENOS

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

GUIDELINES FOR COMMON POLICIES IMPLEMENTATION (FINAL)

KNAW-DANS (overall coordination)

CLARIN

MIBACT-ICCU

KCL

PIN

30 January 2019

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 (2)

Deliverable Number	D3.2
Dissemination Level	Public
Delivery date	30 January 2019
Status	Final
Author(s)	Hella Hollander (KNAW-DANS) Francesca Morselli (KNAW-DANS) Frank Uiterwaal (KNAW-NIOD) Femmy Admiraal (KNAW-DANS) Marnix van Berchum (CLARIN: Huygens ING) Thorsten Trippel (CLARIN: Univ. Tübingen) Lene Offersgaard (CLARIN: UCPH) Sara di Giorgio (MIBACT-ICCU) Paola Ronzino (PIN)

With collaboration from all partners, especially:

Anders Conrad (CLARIN: RDL)
Emiliano Degl'Innocenti (CNR-OVI)
Roberta Giacomi (SISMEL)
Valentijn Gilissen (KNAW-DANS)
Vanessa Hannesschläger (OEAW)
Mark Hedges (KCL)
Klaus Illmayer (OEAW)
Adeline Joffres (CNRS/Huma-Num)
Emilie Kraaikamp (KNAW-DANS)
Antonio Davide Madonna (MIBACT-ICCU)
Marie Puren (INRIA)
Maurizio Sanesi (SISMEL)
Claus Spiecker (FH Potsdam)
Michael Svendsen (CLARIN: RDL)
Heiko Tjalsma (KNAW-DANS)
Eld Zierau (CLARIN: RDL)

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	D3.2 Guidelines for Common Policies Implementation (Final)
Document title	Guidelines for Common Policies Implementation (Final)
Author(s)	Hella Hollander, Francesca Morselli, Frank Uiterwaal, Femmy Admiraal, Marnix van Berchum, Thorsten Trippel, Lene Offersgaard, Sara di Giorgio, Paola Ronzino.
Dissemination level/distribution	Public

Document History

Version/date	Changes/approval	Author/Approved by
V 0.1 25/01/2019	Review and corrections	Sheena Bassett (PIN)
V2 29/02/2019	Document ready for internal submission to PIN	Hella Hollander (KNAW-DANS)

Table of Contents

Executive Summary

This deliverable presents the PARTHENOS Guidelines for Common Policies Implementation, based on the work that was carried out by Work Package 3. This document is a follow-up of deliverable D3.1, submitted in April 2017. Whereas D3.1 aimed to give an overview of existing policies concerning data management, quality of data, metadata and repositories, and IPR, open data and open access, in the present document the results of this study are further consolidated, leading up to a coherent set of guidelines. In addition, this deliverable describes the concrete outcomes of the research that WP3 conducted: the PARHTENOS Policy Wizard and the DMP tool.

After an introduction to the topic (Chapter 1) and a description on the Methodology (Chapter 2), Chapter 3 starts with providing a background on the FAIR principles and argues why these principles are used throughout this deliverable as a reference point, by comparing the FAIR principles with a number of other models for data management. The chapter continues with an elaboration of how the FAIR principles can be operationalized, discussing in detail each of the particular aspects; Findability, Accessibility, Interoperability and Usability. At the end of each of these dedicated sections, the relevant Guidelines are presented in a textbox.

Chapter 4 first presents an overview of the legal frameworks that are relevant for access to research data and its re-use. It discusses intellectual property rights, and introduces two important legal regulations, namely the General Data Protection Regulation and the Public Sector Information directive. After that, the Chapter provides a discussion of Open Science, focusing particularly on Open Access and Open Data. Finally, four examples of commonly used licensing frameworks are presented; Creative Commons, Open Data Commons, RightsStatements, and the CLARIN licensing framework.

Chapter 5 focuses on the tools that were developed within WP3. It presents the PARTHENOS Policy Wizard, an online service that aims at helping researchers to discover which data policy applies best to their particular data. In addition, it discusses how the PARTHENOS DMP template for Archaeology was designed and tested, which resulted in an online tool for creating a Data Management Plan. This DMP template is then placed in a broader context of the work on a Domain Data Protocol carried out by a Science Europe working group.

1. Introduction

In the Humanities, it is often difficult to give a common definition of (research) data, because of the heterogeneous nature of the Arts and Humanities, and its versatile methodologies. For example, a linguistic corpus can be the result of collecting face-to-face interviews that have been further processed with specialised tools for creating annotations. A historian, on the other hand, may spend months in an archive in order to find those documents that shed light on her/his research hypothesis. For archaeologists, research data may also be 3D scans of digitized artefacts made accessible through archaeological databases.

In general, data can be described as the representation of observations, objects or other entities used as evidence of phenomena for the purpose of research or scholarship (Borgman 2015). Trevor Owens (2011) expanded on this definition by indicating that humanities data are those “multifaceted objects that can be mobilised as evidence in support to an argument”. He sees humanist data as a threefold concept:

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: data can be thought of 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.

When starting the work in WP3, we first wanted to understand what the concept of data means within each of the disciplines within the Arts and Humanities, and what the stakeholders'[1] personal understanding of data quality is. Therefore, we interviewed archaeologists, linguists, social scientists, historians, but also data archivists and data

[1] The stakeholders of WP3 were introduced and described in detail in Section 1.4 of D.3.1: Report on Guidelines for Common Policies Implementation.

specialists in Cultural Heritage Institutions (sometimes also referred to as GLAMs[2]) - as their activities interact and form a continuum with that of the researchers. As can be concluded from the summary below, we discovered that there is not a unique understanding of the concept of data among the Arts and Humanities, but rather a rich and diverse data landscape with different nuances and flavours for each discipline.

Archaeology

Archaeologists primarily deal with data from excavations and field surveys (such as descriptions, images, maps, GIS data). These data are most frequently created by the researchers themselves, in the sense that these data didn't exist before the discovery of the buried artefact. Specific research data can be represented by (among others):

- 3D models,
- Geomorphological models of earth surface processes and history,
- Dendrochronology and vegetation data: Tree-rings are also very useful not only for dating human artefacts, but also provide useful information such as temperature and precipitations,
- 14C (Radiocarbon) dating and carbon and nitrogen isotope analysis, to determine the age of an object.

Good quality data for archaeology follow well defined standards for measurements, excavation methodology and stratigraphical approach.

Language Studies

The field of language studies is difficult to define precisely as it includes a variety of practices such as text analysis or oral history (usually associated with History) that overlap with other disciplines. This is reflected in the type of data, used by researchers in this field:

- single word collections (dictionaries),
- complete/continuous texts,
- scans of manuscripts / typescripts,

[2] GLAMs stands for 'galleries, libraries, archives and museums'. Throughout this document, we will use the more general term 'Cultural Heritage Institutions'.

- OCR text + image files, mark-up (data & metadata),
- data for enrichment (geodata, person's, time, et cetera) & reference data,
- audio/video material,
- existing data (e.g. in excel / word files), legacy data; input data can be raw/unstructured or structured.

The creation of metadata is also seen as a crucial process during the data creation phase, as well as the use of standards and controlled vocabularies for mapping the fields of the metadata.

History (including Medieval Studies, Recent History, Art History, Epigraphy, et cetera)
Similar to the other disciplines in the Humanities, the types of data produced and analysed by historians vary greatly. They include multimedia sources like audio and video files, such as radio recordings or interviews (including testimonies). Historians are also frequent visitors of archives and libraries, as these institutions hold the heritage collections that they use as primary sources for their research. In this case, documents, folders or journal articles, represent the data that historians use to answer their research questions.

It is difficult to state where exactly the process of data creation starts for historians. When visiting an archive, they mainly collect, re-interpret or contextualize existing data. However, these data may be used to create something new, when for example merged with another datasets. Also, new layers of meaning are sometimes added to existing historical source material, such as transcriptions or new metadata.

In terms of quality, the notion of completeness is an important aspect for data used by historians. If not complete, the gaps in the source data should otherwise be clearly stated. A traceable provenance greatly enhances the reliability of data when used in historical research.

Social Sciences

Similar to the practices in other disciplines in the Humanities, social science researchers collect and create data in a variety of ways. The chosen approach largely depends on whether their research question relies on qualitative or quantitative data. In the first case, data are collected via qualitative methods such as interviews or focus groups: the resulting collected data are, therefore, an in-depth description of the interviewee’s history or experiences. In the second case, the data are collected via surveys or from already existing research, and come in the form of spreadsheets or databases.

In terms of quality, in the social science the validity (data is complete, accurate, secure and consistent) of the data also plays an important role. In addition, one of the main quality indicators in social science data is the anonymization of the data, with particular attention to sensitive data.

Data Archives

Data Archives (in the Humanities and Social Sciences) receive data from institutions such as universities as well as from single researchers. Data Archives consider all the information that researchers provide as data, as long as it is in one of the accepted standard formats so that they can be re-used by other researchers.

Data should also be as complete as possible: the metadata and the documentation should be coherent; finally, as is the case with any other stakeholders, data should be anonymized if they contain sensitive information.

Cultural Heritage Institutions

Cultural Heritage Institutions (CHI’s) include museums, archives, libraries and galleries. In different ways, researchers in the Humanities rely on them to collect their data, both first level (e.g. manuscripts, documents) and secondary sources (e.g. catalogues). Extensive literature and documentation (including recommended standards and best practices) about data and metadata exists for each of the CHI’s described, especially for libraries, archives and museums.

2. Methodology and relation to D3.1

The present document is the follow-up of D3.1, submitted in April 2017. Although these two deliverables reflect a continuum in the way they focus on common policies on data (metadata and repositories) quality in the humanities, they do have a slightly different approach, taking into account the ongoing developments in data management, as well as the feedback on D3.1. which was provided by a group of experts engaged by WP2. Based on this feedback, a list of "gaps" was produced, which will be addressed, and filled in whenever possible, throughout this document.

In addition, D3.2 focuses more on the concrete outcomes of the research that WP3 conducted over the last three years. While the first deliverable was conceived as a preliminary study on the topic of data quality, this second version aims at providing concrete tools for our stakeholders, and offering them best practices for the management, sharing and reuse of research data. This approach reflects the finding that the data management landscape is very fragmented and organised in disciplinary silos, which results in a myriad of best practices and guidelines, by making the management aspect of data very difficult to penetrate and to apply to the researchers' own data.

This document gives an overview of the research conducted in WP3, including an in-depth introduction to the FAIR principles (Chapter 3), IPR and open data recommendations (Chapter 4) and an overview of the tools for implementing common policies created by WP3 in the past three years (Chapter 5). The most relevant outputs are:

- the Policy Wizard, an interactive guide to the most relevant policies related to data quality, organized by disciplines but made available for all the humanities and social sciences researchers,
- A Data Management Template, developed in particular for the archaeology field, and a Domain Data Protocol, an ongoing effort of the organization Science Europe, whose aim is to propose a common framework for the data management for each domain (i.e. the Humanities in this case). The idea behind this initiative is that of reducing the fragmentation of data management initiatives by creating a shared layer on which each discipline can add its specific requirements.

In addition, an important part of this deliverable consists of the PARTHENOS Guidelines, that represent a series of recommendations and guidelines to our target audiences to FAIRify their data management during and after their research/infrastructure work, and to make data reusable. The PARTHENOS Guidelines were extracted from the work carried out by WP3 and described in D3.1, and are used as the most important take-away messages throughout Chapter 3 of the present deliverable. As they are envisioned as an accessible and actionable resource that is detachable from the main deliverable, they are also disseminated as a stand-alone brochure (see Appendix I).

3. The FAIR principles and their operationalization

As a cluster project, PARTHENOS is - *by design* - aware of the disciplinary uniqueness of the different communities involved. As outlined in Chapter 1, the various communities involved create and/or manage different types of research data. In Chapter 2, the aim of the project was presented which is to develop shared solutions which are of benefit to all the different strands of Humanities research, while also remaining mindful of differing disciplinary needs and requirements. This chapter will give background on these principles (Section 3.1), and discuss how these principles can be put into practice (Section 3.2).

3.1 FAIR - a starting point for common policies and implementation strategies

When drafting a set of common policies and implementation strategies around data management, we felt that a structure was needed. After considering various models, we decided that the FAIR principles provided the most useful framework to structure a set of guidelines and recommendations. The sections below provide insight into the considerations leading up to the decision to use the FAIR principles as a reference point (Section 3.1.1) as well as to present the principles in more detail (Section 3.1.2) and discuss the relation between the Guidelines and the FAIR principles (Section 3.1.3).

3.1.1 Models for data management

The landscape of models and best practices for data management is diverse. Generally, this is a desirable situation, as there will never be a one-size-fits all solution to this broad challenge, nor should there be. A number of alternative models and standards for digital data curation and archiving are described briefly below. Also, a brief explanation is provided on why they did not suit the PARTHENOS' project needs or - in case of the FAIR principles - why they do fit with the project rationale.

Most importantly, we felt that the chosen format should be relevant to all stakeholders involved. One of the possible dividing factors between the different stakeholders is that, while research methods may go hand in hand with data creation in some disciplines, others rely more heavily on existing data sets. For example, the latter holds for Cultural Heritage Institutions, which primarily manage existing data. Hence, the **UKDA Research Data**

Lifecycle model was not considered fit for all stakeholders in PARTHENOS. Vice versa, the **SCAPE Policy Framework** entails a best practice on preservation policy. It is primarily designed as an organisational model, and as such considered less suitable for research data management. **The OAIS Reference Model** is widely used to describe all the functions of the data management procedures to ingest, describe, store and make data available in a data repository. Using OAIS, a repository can describe its core archival functions and processes in standard terms for reference purposes. As OAIS has a rather strong IT architectural background though, it did not seem a suitable model for all stakeholders; preference was given to a more functionally oriented structure, rather than a technical one. The **CoreTrustSeal** and the **Capability Maturity Model** are frameworks that mainly focus on assessing data repositories - respectively focussing on quality and on maturity - but seem less suitable at the level of actual research data.

In 2014, the **FAIR guiding principles** for individual datasets were formulated: a set of principles that help stakeholders to make data Findable, Accessible, Interoperable and Reusable. These principles were first published in March 2016 (Wilkinson et al. 2016) and quickly have become very popular. The intent was, according to their 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 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.

When comparing the FAIR approach to the models above, 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. This means that the model speaks to the broad range of PARTHENOS stakeholders: from individual researchers without a technical background or experience in digital data preservation, to experienced and trained depositors, such as people working in data archives or Research Infrastructures.

3.1.2 What are the FAIR Principles?

According to 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 such that they can be easily accessed and/or downloaded with well-defined licence 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[3] 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.

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.

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

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 that they are constantly revisited, updated and refined.[4]

3.1.3 The PARTHENOS Guidelines and the FAIR principles

The Guidelines that feature throughout this deliverable are the result from the work of over fifty project members of the PARTHENOS WP3. The foundations of the Guidelines were laid out in the first version of this report, D3.1: Report on Guidelines for Common Policies Implementation (draft).[5] The list of Guidelines presented in the final Chapter of D3.1 is the result of an investigation which used the results from desk research, questionnaires, and interviews with selected experts. These high-level recommendations were mapped onto the FAIR principles to contextualize them in line with broader developments in the field of research data management.

After finalizing D3.1, the Guidelines were revisited and reorganised, on the level of the Guidelines as a coherent set, as well as on the level of individual Guidelines. Whereas some

[4] http://datafairport.org/fair-principles-living-document-menu.
[5] D.3.1: Report on Guidelines for Common Policies Implementation.

of the Guidelines were merged, others were split or re-ordered. One of the main developments is that the Guidelines were tailored to specific audiences. In the final version, distinction is made between recommendations that concern data producers and data users, and recommendations that concern research infrastructures and data archives. The result is a set of twenty Guidelines, consisting of one general Guideline (see text box below), and nineteen Guidelines that are classified as contributing to make data Findable, Accessible, Interoperable or Reusable, and organized under that specific FAIR. The following sections present the PARTHENOS Guidelines to Fairify data management and make data reusable), first by introducing them in a contextualized and extensive form, then in a graphical layout which shows the Guidelines as they have been printed for dissemination purposes.

Guideline 1: Invest in people and infrastructure

An important prerequisite to be able to implement the rest of the nineteen guidelines in this guide, is to invest in data infrastructures and in hiring and educating data experts.

Data producers and data users

Get acquainted with best practices in research data management. Check out the PARTHENOS training modules on data management or have a look at the CESSDA Data Management Expert Guide.

Research infrastructures and data archives

Invest in hiring and educating data experts and define a budget for making investments in technical infrastructure and staff.

3.2 Putting FAIR into practice

The current policy of most funders and research organisations is to rely on FAIR data. This also holds for the PARTHENOS partners. For example, the search tools for data provided by CLARIN, the Virtual Language Observatory (VLO)[6] and the Federated Content Search (FCS)[7], are good examples of making data FAIR. By assigning resolvable Persistent IDentifiers (PID's) the data is made accessible. If access is restricted, data centres provide access mechanisms, for example with the Identity Provider and Shibboleth architecture. In

[6] https://www.clarin.eu/content/virtual-language-observatory-vlo.

[7] https://www.clarin.eu/content/federated-content-search-clarin-fcs.

these tools, interoperability is achieved by the utilization of standards such as the ISO TC 37 SC 4 endorsed standards or those recommended by the Text Encoding Initiative (TEI)[8].

However, supporting FAIRification of data is a complex matter. For example, interoperability is dependent on the selection of standards as well as on the support of software available. In addition, interoperability and reusability are not only technological issues, but they are also addressed by policies, legal restrictions, etc. Reusability has the additional requirement that researchers are also allowed to reuse data and have access to the tools to do this. The many issues involved require a documentation of the policies that apply and decisions that were taken. Since it is virtually impossible to reconstruct this documentation when a data set is finished, ideally, these aspects are documented before data is collected or created. All of this is part of the data management plan, and as such, the data management plan (DMP) is essential to the successful implementation of a FAIR data policy.

3.2.1 Findable

Findability is the key for effective implementation of FAIR, since the proper way of locating data is a necessary condition for any other step. In order to comply with the ‘Findability’ principles, data providers will have to work on proper identification of their resources, and on providing a structured way of making the properties of data resources accessible.

3.2.1.1 Identification of data resources

To make a data resource findable and accessible, it is essential to provide it with a unique identifier. Though it may seem obvious that this is achieved by assigning file names, URI locations (e.g. http and ftp), stock numbers, cryptographic (md5) checksums, etc., most of these methods have downsides 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 be easy to interpret for third parties;
- checksums are unreadable for a human user and can be considered ‘not-writable’ by humans.

[8] http://www.tei-c.org.

Therefore, assigning a PID to data resources is a basic prerequisite for making data FAIR. Based on these identifiers it must be possible to cite data resources persistently and locate an authoritative copy. However, identifiers serve only to identify a resource. They do not need to authorize access to it, contain information on the content of an object, or provide any other form of semantics.

In terms of findability, problems can also arise with copies. Sometimes, copies have the same identifier as the original file, comparable to an ISBN being the same for each physical copy of the same edition of a certain book. However, in contrast to books, each copy of an electronic file needs to have its own unique identifier, comparable to a book signature that points to its location within a book collection. This is essential in the digital world, because copies of files can be identical in terms of size, content, etc., but can still be distinguished by their location. To ensure the integrity of a file and to check whether a file has been modified, copies should be compared 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, there is the problem of granularity: which set of elements is regarded as the object that needs identification? For example, research data created by measuring sensors often comes in multiple files; textual resources with various annotation layers can come in separate blocks; audio-visual data often consist of signal files together with transcriptions, notes and background information, sometimes in multiple files due to discontinued recording sessions or scene cuts. Regarding granularity, no clear guidelines exist about 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). The ISO 24619 standard 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.

Summary

I. 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 from the area of language resources is ISO 24619:2011

("Language resource management - Persistent identification and sustainable access (PISA)"). The choice of a persistent identifier schema must rely on careful assessment of advantages and disadvantages. Suitable example implementations for these are: handle systems including Digital Object Identifiers (DOI), and URNs.

II. The institution responsible for future access of the resources maintains digital preservation of the received authoritative copy of the data, including information on the identifier assignment.

III. For granularity, there is no clear guideline, but the recommendations from ISO 24619 are good to follow:

   a. 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.
   b. An identifier should be assigned if the resource is associated with the complete content of a digital file.
   c. 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.
   d. 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.

3.2.1.2 Findability by properties of a data resource

Identifiers allow different objects to be distinguished from each other and are a condition for findability in a digital world. However, the object's suitability for access and reuse also depends on other properties. 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. Based solely on the content of the object itself, a resource may not be findable. For example, three dimensional scans of artefacts consist of numeric representations of spatial vectors, often stored in proprietary and binary formats. These can hardly be searched for by humans. Similar issues arise with textual resources. For example, historical documents useful to the study of the history of religion don't necessarily contain the word 'religion'. If the only way to explore these texts is a full text search, it's possible the resource would not be found in a general search on the subject.

Therefore, in order to find a particular resource, it is necessary to have structured and meaningful descriptions of it, including descriptive and administrative metadata. This data can be indexed by general search engines, specialized search engines, or cataloguing applications. Cataloguing applications often require a distinct set of metadata for the archiving process. These catalogues are often very specific to the type of research data an institution archives and maintains, often tailored either to printed resources, like in libraries, or to artefacts, like 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 allow for better interoperability of resources. In general, the more complete the metadata are provided, the higher the quality is, even after conversion.

As was argued in the Introduction, in the domain of research data, there are very different types of resources, depending on the field of research and the domain of the scholars. Each of these requires particular classes of metadata to provide a meaningful description of the research data. A unification of all possible structured metadata sets would be extremely rich, but most metadata fields would remain empty. At the same time, some descriptive categories used for one type of resource may be inappropriate, irrelevant or misleading for another. Hence, it is required that the metadata schema in use suits 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 utilising 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, are often not able to interpret the structures of a metadata schema. Instead, these search engines 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 beyond HTML. For linked data, RDF is the most commonly used format. Though RDF is highly adjustable and metadata schemas can be described with it, using RDF as a principal 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.

Summary

I. Select an appropriate metadata schema for the type of resource being described. Metadata can have various functions, such as citation metadata, disciplinary metadata, preservation information, provenance, etc. The metadata intended for findability are the type of metadata used for citation and descriptive data in a catalogue. This should be the principal format for maintaining the descriptive metadata. Utilise 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). Using only less detailed schemas for describing research data, such as Dublin Core or Datacite MDS, is not recommended.

II. Provide different formats, this can include, for example, HTML to allow findability with standard internet search engines, Datacite MDS and Dublin Core for interoperability purposes with archives metadata, etc.

III. The metadata provided should be high-quality, i.e. as correct and complete as possible.

IV. Specify requirements about use of persistent identifiers for referencing and content retrieval of the metadata.

V. Select an appropriate persistent identification schema and assign a PID to every resource.

VI. Ensure semantic interoperability by referencing authority files in the metadata, for example, persistent author identifiers such as VIAF, ISNI, or ORCID.

Guidelines to make research data Findable

Research data should be easy to find by both humans and computer systems and based on mandatory descriptions of the metadata that allows the discovery of interesting datasets.

Guideline 2: Use persistent identifiers

Locating data is a necessary condition for any other step from access to reuse. To be findable, any data object and dataset should be uniquely and persistently identifiable over time with a persistent identifier (PID). A PID continues to work even if the web address of a resource changes. PIDs can take different forms, such as a Handle, DOI, PURL, or URN.

Data producers and data users

Reference the PID which was assigned to your dataset in your research output.

Research infrastructures and data archives

Select the appropriate form of persistent identification schema and assign a PID to every resource. Use the PID Guide from NCDD to decide on the right PID for your research infrastructure.

Guideline 3: Cite research data

If research data have a persistent identifier and are cited in accordance with community standards, the corresponding data objects or datasets are more easily found.

Data producers and data users

Get acquainted with data citation guidelines that are specific to your field or discipline and cite research data accordingly.

Research infrastructures and data archives

Provide information about best practices in data citation to research communities and make it easy for data users to cite data, e.g. by using a standardised button which says 'How to cite this dataset'.

Guideline 4: Use persistent author identifiers

A persistent author identifier (e.g. VIAF, ISNI or ORCID) helps to create linkages between datasets, research activities, publications and researchers and allows recognition and discoverability.

Data producers and data users

Distinguish yourself from any other researcher or research group. Apply for an author identifier if you do not already have one and reference it in your dataset.

Research infrastructures and data archives

Reference author identifiers in the metadata.

Guideline 5: Choose an appropriate metadata schema

Metadata is essential in making data findable, especially the metadata which is used for citing and describing data. A metadata schema is a list of standardised elements to capture information about a resource, e.g. a title, an identifier, a creator name, or a date. Using existing metadata schemas will ensure that international standards for data exchange are met.

Data producers and data users

To enable the discovery of content, describe research data as consistently and completely as possible. Include enough information for the data to be accessed and understood later on. If possible, use an existing metadata schema which fits the type of data object or dataset you are describing.

Research infrastructures and data archives

Clearly state which metadata schema you apply and recommend to the research community. To enrich datasets at data deposit, consider having a data submission form which collects additional metadata, e.g. about the provenance of the data.

3.2.2 Accessible

In contrast to findability, accessibility means that there is - at least technically - a way to access a resource based on the information provided when finding the resource. Basically, there are two criteria defining access, (1) a resource can be retrieved based on its identifier, and (2) the (descriptive) metadata is available, even if a resource itself is no longer accessible.

Creating FAIR data that is easily accessible needs to be a joint effort between data creators and policy makers at institutional and funder level. While researcher can make sure to structure and enrich their research output in such a way that data hosts can ingest this data as FAIR compatible as possible, data hosts themselves should invest in enrichment tools or user interfaces that help to make references in data objects syntactically parseable and semantically machine-accessible. Making descriptive metadata publicly accessible can be achieved by 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 it should be recorded as part of the documentation of the resource in restricted contexts. Data hosts should also publicize the protocol endpoint to suitable search providers. A good example is the CLARIN registry for endpoints[9] providing language related research data.

3.2.2.1 Accessibility of data resources based on their identifier

As was argued above, one of the fundamental criteria for findability is that each resource has a persistent and unique identifier. However, the existence of the identifier - and finding the identifier - does not mean that a resource is, in fact, accessible. To ensure accessibility, the identifier needs to be identified as such and there needs to be a (technical) procedure in place to retrieve the resource if only the identifier is known. For example, PID systems such as the Handle system or the DOI system allow for marking up an identifier and thus rewriting the identifier as a web address. For example, if the Handle hdl:11022/0000-0007-C5A7-E is marked as a handle, the openly available documentation for handles allows automatic rewriting of the handle to a web address http://hdl.handle.net/11022/0000-0007-C5A7-E which then resolves to the metadata, that also lists all parts or files that are constituents of the resource. Actionable PIDs that use a protocol, i.e. a technical specification on how to technically interpret a set of data, allow for retrieving a set of data.

If the data itself are not publicly available or when usage restrictions apply to the data, an actionable identifier cannot resolve in the same way. In such cases, the protocol needs to provide a way for authenticating and authorising individuals or agents such as computer

[9] https://centres.clarin.eu/oai_pmh.

programmes, and either grant or deny them access to the resource. According to the FAIR principles, this protocol also needs to be open and freely implementable. Very often, web-based systems with single sign-on mechanisms can be used to authenticate a user, and to allow authorisation either by properties of the authenticating institution or based on specific roles and access permissions for individual users. For example, using the Shibboleth procedure often used by academic institutions, a data centre could provide access to a resource that is open to academic users. It is good practice to describe the applied access and usage restrictions, and usage licences in human readable form in the metadata description of a resource alongside implementing an Authentication, Authorisation, and Access Infrastructure (often called AAAI or AAI).

Summary

I. Use persistent identifiers with established protocols, such as the Handle system or DOIs.

II. Make sure that the identifiers resolve to the metadata and/or resources to provide access to the resource.

III. Describe the access restrictions of a resource in the metadata.

IV. Implement an Authentication, Authorisation, and Access Infrastructure (AAAI or AAI).

3.2.2.2 Accessibility in the case of resources no longer existing

A special case of accessibility arises when resources no longer exist, for example if born-digital resources get deleted, or because of technical failures of storage media. The FAIR principles require that, in these cases, at least the metadata remains in an accessible form, publicly available using open and implementable protocols. It is good practice to modify the metadata to indicate that the resources no longer exist. The metadata should then remain accessible, for example, by OAI-PMH or SPARQL.

Summary

I. If a resource no longer exists, modify the metadata to indicate the changed status.

II. The metadata needs to be maintained in a publicly available, accessible location, for example via OAI-PMH or SPARQL endpoints that are made known in the community.

III. Make sure that the PIDs either resolve to the resource or the metadata directly, or indicate in the protocol that the resource no longer exists and point to the metadata.

Guidelines to make research data Accessible

Research data should be easily accessible and retrievable with well-defined access conditions using standardised communication protocols.

Guideline 6: Choose a trustworthy repository

A certified repository offers a trustworthy home for datasets. Certification is a guarantee that data are stored safely, and will be available, findable and accessible over the long-term. Examples of certification standards are CoreTrustSeal, nestor Seal, and ISO 16363 certification.

<u>Data producers and data users</u>

Make your data accessible through a trustworthy repository. In addition, if you follow the repositories' standards (on preferred file formats, metadata schemas etc.) you can make sure that all requirements for making data FAIR are met.

<u>Research infrastructures and data archives</u>

Clearly state the level of certification on your website. If you are not(yet) certified, state how you plan to ensure availability, findability, accessibility and reusability in the long-term.

Guideline 7: Clearly state accessibility

Access information specifies how a data user may access a dataset. When depositing data in a data repository, it should be clear which access options a data depositor can choose.

<u>Data producers and data users</u>

When choosing an access option, consider legal requirements, discipline-specific policies and ethics protocols when applicable. Choose Open Access when possible. When you collect personal data, ask yourself whether it contains any information which might lead to participants' identities being disclosed, what participants consented to and which measures you have taken to protect your data. If your data cannot be published in Open Access, the metadata should be, allowing data discovery.

<u>Research infrastructures and data archives</u>

Encourage (meta)data to be published in Open Access. Clearly state restricted access options for sensitive (meta)data that should not be part of the publicly accessible (meta)data. In this case, strive to make the (meta)data available through a controlled and documented access procedure.

Guideline 8: Use a data embargo when needed

During a data embargo period, only the description of the dataset is published. The data themselves are not accessible. The full (meta)data will become available after a certain period of time.

Data producers and data users

Clearly state why and for what period a data embargo is needed. Make the (meta)data openly available as soon as possible.

Research infrastructures and data archives

Specify whether a data embargo is allowed and what conditions apply.

Guideline 9: Use standardised exchange protocols

By using standardised exchange protocols, research infrastructures can make (meta)data publicly accessible and harvestable by e.g. search engines, vastly improving accessibility.

Research infrastructures and data archives

Use standardised protocols such as SWORD, OAI-PMH, ResourceSync and SPARQL. Convert metadata schemas into XML or RDF. Maintain a registry for protocol endpoints, the path at which research data can be accessed, and publish them.

3.2.3 Interoperable

Enabling interoperability is a great benefit for researchers and for the further processing of data in research projects. Therefore, data hosts should explain in detail how researchers can obtain data in their holdings, and how they combine such data with other repositories. It is also important to point out how to easily integrate the resulting and processed datasets back into the research data life cycle. 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.

Deliverable D3.1 (pages 79-83) gives an overview of how the PARTHENOS partners have enabled interoperability. Two opposing approaches were identified in the data creation/ingestion phase, that strongly influence 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 tends to complicate interoperability, whereas the second one provides a clear framework, but contains the risk of technical obsolescence or a lack of acceptance. This is probably the reason why many opt for in-between approaches, recommending, or even pushing, data providers to use appropriate formats, while at the same time allowing ingestion of all other kinds of formats. As a consequence, a majority of data hosts hold a mixture of data with different interoperability capabilities. The best way to act in this respect is to point out which data fit the interoperability principle, and to motivate data providers in choosing formats and data structures with a high interoperability level.

The use of preferred formats is considered a best practice, although a list of accepted data formats does not reveal whether a specific (meta)data format boosts interoperability. More and more new formats are being developed, some of them claiming that they are the best solution for a specific domain or situation. Therefore, 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 strongly and actively a community supports it. In this sense, interoperability needs to mediate between 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. At the same time, data holders should also recommend changes in the practices of a community.

Summary

- Give an easy to find and detailed overview of accepted (meta)data formats, ideally in a single page that can be referenced directly.
- Present the possibilities for interoperability in a finely granulated and well-structured way, making use of up-to-date design and user interface methodology.
- Document and give easy access to the data model or models in use in a repository. Make clear which parts of the data model enable interoperability, and which parts are relevant when connecting datasets between projects.
- Develop, as a joint effort between repositories, scripts and tools for the (automatic) transformation of data in the ingest phase, enabling interoperability at an early stage.

3.2.3.1 Machine-actionable (meta)data

To support data reuse, humans as well as machines should be able to automatically find and use (meta)data. Machine-actionable means that machines act automatically when confronted with the wide range of types, formats, access mechanisms, and protocols, by registering provenance so that data collected can be adequately cited and reused. To make this happen, all actors in the data management process must provide information that allow machines 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 (Wilkinson et al., 2016).

Machine readability especially relies on a high level of (meta)data content, i.e. notably well-formed and predictive (meta)data. 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. Dedicated staff, responsible for data quality assurance and mediation between data creators and data hosts, help boost interoperability aspects. However, there is often no funding for such personnel, and data volumes continue to grow, complicating the work of data stewardship.

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 are a well-planned (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. In addition, it becomes increasingly important to establish automatic processes that clean (meta)data, derive metadata, and enrich data. Combined efforts in developing workflows and software solutions for such automatic processes are also necessary.

\Machine actionability also relies on clearly documented and stable endpoints, from where machines gather the (meta)data. APIs (Application Programming Interfaces) need to be readable with as few limits as possible 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 in other use cases.

Summary

I. Establish quality assurance processes, with a special focus on the data creation phase.
II. Combine and apply the push of data providers and automatic processes to boost (meta)data quality.
III. Invest in tools for cleaning up (meta)data and converting raw data into other, standardised and interoperable, data formats.
IV. Establish well documented machine-actionable APIs for the (meta)data.
V. Give more information on best practices for machine driven automatic data search and reuse.

3.2.3.2 Shared vocabularies and/or ontologies for (meta)data formats

Although good practices for using shared vocabularies and ontologies can be found in CLARIN's Concept Registry[10] (CCR) and Deliverable D3.4 of the ARIADNE project[11], researchers are often not aware of their existence or value, as shown in Section 3.2.3.2 of Deliverable 3.1 on the PARTHENOS inventory.[12] Thus, there is a need for an overview of shared vocabularies and/or ontologies in use for the different research domains and adoption needs to be encouraged more.

Summary

I. The description of metadata elements should follow community guidelines that use an open, well defined vocabulary.
II. 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.
III. Give pointers to vocabularies and ontologies that can be used, based on research domain specifics and tangible use cases.

[10] https://concepts.clarin.eu/ccr/browser/.
[11] ARIADNE D3.4: Final Report on Standards and Project Registry.
[12] PARTHENOS D.3.1: Report on Guidelines for Common Policies Implementation.

3.2.3.3 Syntactically parseable and semantically machine-accessible (meta)data

Syntactically parseable and semantically machine-accessible data are strongly dependent on established (meta)data formats in a community. It is important for semantic interoperability to have well-documented and communicated schemas. Reliability and permanent access are crucial when operating with shared semantics. Furthermore, harmonising such approaches on an international level is highly recommended. As shown by the CLARIN Concept Registry, such efforts seem to be more stable if agreements on semantics and the organisation of the descriptions of semantics are handled by higher level, international institutions, i.e. Research Infrastructures. The interoperability task of combining data is mostly done by researchers and research projects, but more documentation is needed on how to combine different datasets between projects and what are best practices.

3.2.3.4 Preferred formats for data stewardship and preservation

For disciplinary repositories or Research Infrastructures it may be possible and desired to prescribe a prioritised list of data formats, combined with support.[13] General guidelines for data formats, leading up to the best long-term sustainability and accessibility are:

- Formats are frequently used.
- Formats have open specifications.
- Formats are independent of specific software, developers, or vendors.

Based on these criteria, extensive lists of preferred and acceptable file formats are 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. 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.

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

[13] For an example, see the lists of preferred and acceptable file formats of DANS, September 2015, version 3.0: https://dans.knaw.nl/en/deposit/information-about-depositing-data/DANSpreferredformatsUK.pdf.

guidance, and maybe support, is still needed on the specific schema to be used (e.g. TEI), in order to best provide reusability. And even such discipline-specific standards will often have their own versioning and thereby raise their own challenges regarding preservation, as discipline formats will also need upgrading as part of preservation plans, separate from possible file format migrations.

Guidelines to make research data Interoperable

To speed up discovery and uncover new insights, research data should be easily combined with other datasets by humans as well as computer systems.

Guideline 10: Establish well documented machine-actionable APIs

Well documented and machine-actionable APIs - a set of subroutine definitions, protocols, and tools for building application software - allow for automatic indexing, retrieval and combining of (meta)data from different data repositories.

Research infrastructures and data archives

Document APIs well and make it possible to deliver the schema of the (meta)data model. Consider showing examples of how to successfully mine data from different endpoints and combine them into new data sets usable for new research.

Guideline 11: Use open well-defined vocabularies

The description of metadata elements should follow community guidelines that use open, well defined and well-known vocabularies. Such vocabularies describe the exact meaning of the concepts and qualities that the data represent.

Data producers and data users

Use vocabularies relevant to your field, and enrich and structure your research output accordingly from the start of your research project.

Research infrastructures and data archives

Give examples of vocabularies the research community may use, based on research domain specifics.

Guideline 12: Document metadata models

Clearly documenting metadata models helps developers to compare and make mappings between metadata.

Research infrastructures and data archives

Publish the metadata models in use in your research infrastructure. Document technical specifications and define classes (groups of things that have common properties) and properties (elements that express the attributes of a metadata section as well as the relationships between different parts of the metadata). For metadata mapping purposes, list the mandatory and recommended properties.

Guideline 13: Prescribe and use interoperable data standards

Using a data standard backed up by a strong community, increases the possibility to share, reuse and combine data collections.

Data producers and data users

Check with the repository where you want to deposit your data what data standards they use. Structure your data collection in this format from the start of your research project.

Research infrastructures and data archives

Clearly specify which data standard your institution uses, pool a community around them and maintain them especially with a perspective on interoperability. Good examples are CMDI (language studies) and the SIKB0102 Standard (archaeology).

Guideline 14: Establish processes to enhance data quality

To boost (meta)data quality and, therefore, interoperability, establish (automatic) processes that clean up, derive and enrich(meta)data.

Data producers and data users

Establish procedures to minimise the risk of mistakes in collecting data. E.g. choose a date from a calendar instead of filling it in by hand.

Research infrastructures and data archives

Invest in tools to help clean up (meta)data and to convert data into standardised and interoperable data formats. Combine efforts to develop workflows and software solutions for such automatic processes, e.g. by using machine learning tools.

Guideline 15: Prescribe and use future-proof file formats

All data files held in a data repository should be in an open, international, standardised file format to ensure long-term interoperability in terms of usability, accessibility and sustainability.

Data producers and data users

From the start of your research project think about future-proof file formats. Use preferred formats which are recommended by the data repository and are independent of specific software, developers or vendors.

Research infrastructures and data archives

Encourage the use of formats that are considered suitable for long-term preservation such as PDF-A, CSV and MID/MIF files. Provide an easy-to-find and detailed overview of accepted file formats.

3.2.4 Reusable

The recommendations in this section address aspects of future research practices and how current researchers and data archives can best accommodate and enable these, based on the experiences within the PARTHENOS consortium.

3.2.4.1 Clear and accessible data usage licence for (meta)data

For allowing data reuse, it is necessary to help the user understand the rights and responsibilities through an unambiguous statement of legal rights and policies retained by the rights holder(s). Standardised electronic statements regarding the legal rights retained, can support legal interoperability and help to make them understandable for a wide audience, and overcome national barriers.

3.2.4.2 Detailed provenance of (meta)data

It is common in the Humanities to set up workflows that transform raw or primary data into higher levels of processed data products (e.g. preparing a document for linguistic analysis by processing it with a chain of tools, such as part-of-speech tagger, lemmatizer, etc.). Each level builds upon the previous processing, which makes it essential to document the provenance for every data object. The provenance record lists the data that the resulting data object is based on, as well as the type of processing, tools, etc., the data was subjected to. This information is referred to as provenance metadata[14], which is crucial for reuse of processed data for scholarly purposes.

Apart from its importance for reusability, documenting provenance is seen as an integrated part of maintaining digital objects in a digital preservation repository. However, we have not been able to find any general recommendations on the format of provenance metadata. A good practice is the PREMIS object mode[15], otherwise we suggest that provenance metadata must be added or included in the metadata schemas used. To our best knowledge, provenance metadata is not discipline-specific, and ought to be applied in a general and interoperable way. This 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 to by their persistent identifiers, rather than by name, abbreviations, etc. This specifically requires that software tools must also be registered and persistently identified.

In the case of larger, and possibly heterogeneous, datasets, the question remains at which level of granularity provenance should be expressed. Ideally, 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 to enable reuse. This may, however, be difficult to achieve, depending on the supporting software of each particular repository, as well as on the file formats and object models in use. In practice, a rule of thumb is that provenance metadata should be provided at the level of object identification. In other words, if there is one persistent identifier for a complete dataset, there

[14] For definitions of provenance metadata, see http://smw-rda.esc.rzg.mpg.de/index.php/Provenance_metadata.
[15] http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf.

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.

3.2.4.2.1 Versioning

It is not uncommon that certain datasets or corpora are dynamic, rather than closed project data, and that they are continuously corrected, improved, developed and enhanced. The issue of versioning is closely related to that of provenance. Versioning metadata is in essence part of the provenance metadata, and includes identification of the version (a unique number or tag), a change log record, date, information about who performed the change, etc. New versions of existing digital objects are generally treated in two different ways:

- The new version is treated as a new object and as such it gets 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 the nature of the relationship to the previous version.
- The new version remains part of the existing digital object and therefore it retains the persistent identifier of this object. However, it is essential to the scientific integrity that it must be possible to refer to a specific version of an object. Therefore, 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 something similar.

Format migrations that are performed as part of digital preservation plans are considered equal to creating new versions of data objects, and thus, they must follow the general guidelines about versioning. In this case, the new version must contain a reference to the old file format that it builds upon, as well as information on the migration process, and tools that were used (c.f. Section 3.2.4.2.3 on workflows and tools). Depending on whether the original format is preserved alongside the new one, the reference to the previous version may be still retrievable or not.

3.2.4.2.2 Annotations

Annotation of resources is a common practice in the Humanities and it is often well supported by Research Infrastructures. However, even if the researcher is willing to share the annotations, it may not be possible to share the source that was annotated as well. These different situations lead to different requirements for provenance metadata for the annotation:

- If the source itself is open, and the researcher is authorised to do so, it may be possible to annotate directly in the source file. This case could be classified as generating a new version of the file, as described above. The researcher providing the annotations would need to supply provenance metadata describing his/her annotation/changes to the document.
- The annotation can be openly shared, whereas the source remains closed (often because of rights issues). 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, et cetera) as well as contain a reference (by persistent identifier) to the annotated object.
- 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 and a clear reference to the object being annotated.
- 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.

3.2.4.2.3 Workflows and tools

In the Humanities, as well as in the natural 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:

- Which tool was used (persistent identifier), and in which version?
- Possible references to algorithms (journal articles or other documentation).

- Who was initiating the process, when, which computing environment, etc.?
- Reference to the original data/object that was processed.

Depending on the file formats, this information may be included in the resulting files themselves, or it may 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, following the exact method of the original processing as much as possible. 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 research question.

It is important to note that software is 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 Domain-relevant community standards for (meta)data

With respect to metadata, the requirement to use discipline-specific metadata should be understood as supplementary to the metadata requirements already discussed in Section 3.2.1 under Findable. Here, we focus on metadata that specifically describe the type of resource in question, the manuscript, the excavation data, the corpus, et cetera, under study. This is the metadata of a more scientific nature 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 to facilitate specialised discovery options or search criteria that are required by the specific community. As a general guideline, data archives are advised to work with research communities on establishing standards that are relevant for their target community, and to offer support of such standards in their infrastructure as much as possible. Apart from metadata and data formats, this may also include support for specialised tools operating on agreed data formats or support for integration of data into a Virtual Research Environments (VRE). Discipline-specific repositories are often well-suited to offer such support, but even more general repositories may offer specialised support in certain fields.

In certain, particularly innovative, cases, using already existing community standards may not 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 that are created and used. The discussion of standards is an ongoing negotiation between researchers’ needs to define their own formats and the need for infrastructure support and data interoperability and reuse.

3.2.4.3.1 Competing standards

Sometimes, competing standards exist 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 that are followed, are also endorsed by the research community, and that general infrastructures must be flexible enough to accommodate the actual research that is performed in the various fields. This would also mean allowing some very generic types of data to accommodate research data in areas where no standards have been defined (yet).

3.2.4.3.2 Object and content models

Community standards will not necessarily follow the data-metadata separation as expressed in the FAIR criteria and may imply different object and content models, and representations. For example, the text community frequently uses TEI, which supports self-contained objects that encompass both data (body) and metadata (header), and suggests various content models, depending to the type of text being modelled. This is not necessarily easy to map into a data-metadata object model, and indeed TEI has been challenging for people implementing data repositories.

In the case of formats that originated from research practices rather than from repository and infrastructure builders, in many cases it will 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 broaden their paradigms into a more general infrastructural view.

Guidelines to make research data Reusable

Research data should be ready for future research and future processing, making it self-evident that findings can be replicated and new research effectively builds on already acquired, previous results.

Guideline 16: Document data systematically

To make clear what can and what cannot be expected in a dataset or repository, data should be systematically documented. Being transparent about what's in the data and what isn't facilitates trust and, consequently, data reuse.

Data producers and data users

Provide codebooks, including a description of methodology, a list of abbreviations, a description of gaps in the data, the setup of the database, etc.

Guideline 17: Follow naming conventions

Following a precise and consistent naming convention - a generally agreed scheme to name data files –makes it significantly easier for future generations of researchers to retrieve, access and understand data objects and datasets.

Data producers and data users

Consult the policies and best practices for your research discipline or domain to find the most suitable naming convention.

Research infrastructures and data archives

Clearly state best practices to create and apply specific file naming conventions.

Guideline 18: Use common file formats

By using standardised file formats that are widely used in your community, reusability is increased.

Data producers and data users

Use current popular file formats next to archival formats to share your data, e.g. Excel (.xlsx) and CSV or ESRI Shapefiles next to MID/MIF files.

Research infrastructures and data archives

Publish the data in popular formats next to the archival format if they are not the same.

Guideline 19: Maintain data integrity

Research data which were collected should be identical to the research data which are accessed later on. To ensure data authenticity, checks for data integrity should be performed.

Data producers and data users

Implement a method for version control. The guarantee that every change in a revised version of a dataset is correctly documented, is of integral importance for the authenticity of each dataset.

Research infrastructures and data archives

To identify if a file has been modified, it is essential to record provenance- the origin of the data plus any changes made over time - and to compare any copy with the original. A data integrity check can be performed by means of a fingerprint such as a checksum, or by a direct comparison of two files. Provide a mechanism to address different versions, for example by adding the version to the identifier as a search parameter.

Guideline 20: Licence for reuse

To permit the widest reuse possible of (meta)data, it should be clear who the (meta)data rights holder is and what licence applies.

Data producers and data users

Make sure you know who the (meta)data rights holder is before publishing your research data.

Research infrastructures and data archives

Communicate the (meta)data licence and reuse options transparently and in a machine-readable format. To improve interoperability, try to map your licences to frameworks which are already widely adopted such as Creative Commons.

3.3 Case Study: Comparison of the PARTHENOS recommendations and the CLARIN B-Centre Checklist

From its founding in 2012, CLARIN has focused on findability and accessibility of research data, and its mission is to enable and facilitate re-use of data resources. CLARIN also has a well-described procedure for assessing data centres, so-called CLARIN B-Centres that provide access to data resources and their metadata in a sustainable and secure way. In the CLARIN community, data centres can be certified as CLARIN B-Centres when they comply with the CLARIN B-Centre Checklist[16] and achieve the CoreTrustSeal. CLARIN has decided to reformulate the B-Centre checklist in 2019 to include clear references to the FAIR principles, and in this way also appreciate and support the promoting of FAIR as done by many parties, as adhering to the working for FAIR principles is in-line with the vision of CLARIN.

In this case study, we outline to which extent the CLARIN B-Centres certification includes the same recommendations as stated in the summaries of Sections 3.2.1-3.2.4 above, and how these issues are addressed.

[16] https://www.clarin.eu/content/checklist-clarin-b-centres.

Findability

Making resources findable is a core issue for CLARIN. The recommendations about findability are included as central checks in the CLARIN checklist. The recommendations about granularity are in principle supported by CLARIN, but are currently only partly addressed in the checklist. The recommendations mentioned in Section 3.2.1.2 Findability by properties of a data resource are seconded by CLARIN, but there is no strict checking of items against the fact that metadata provided should be high-quality. The recommendation about ensuring semantic interoperability by referencing authority files is partly covered in the CLARIN checklist by requiring references from metadata schemas to openSKOS.[17]

Accessibility

The CLARIN checklist requires accessibility of data resources based on their persistent identifiers. Concerning the recommendations about resources that disappear, it is important to note that in the CLARIN community resources should, in principle, not disappear. However, in some cases, it might happen that access to a resource has to be withdrawn because of legal issues, or cases in which it is better to guide users to an updated or corrected version of the resource. In all cases, the metadata and identifier of the earlier version needs to be kept available to give a message about a deprecated or withdrawn resource. This requirement is currently not included in the checklist, but the Standing Committee for CLARIN Technical Centres has ongoing discussions on how to implement this, and a compliance statement of this recommendation is expected to be included in a later version of the checklist.

Interoperability

In CLARIN the focus has been mostly on issues of findability and accessibility, but now that procedures are in place to make resources accessible, more emphasis can be put on issues around interoperability. For years, CLARIN has prioritised machine-actionable (meta)data and the use of a standardized harvesting protocol OAI-PMH, and the checklist addresses this. In addition, CLARIN has developed a central curation module[18] that verifies the format of metadata and the compliance to metadata schemas for harvestable resources. This module is used when centres are assessed. An ongoing effort is to align metadata from different data providers within the search interface of the Virtual Language Observatory

[17] http://openskos.org/.
[18] https://curate.acdh.oeaw.ac.at/#!Collections.

(VLO)[19] to enable users to find more relevant resources despite the fact that the various data providers use different metadata labels for comparable information. Since this is still an area in development, this requirement is not addressed in the checklist.

The recommendation *“Present the possibilities for interoperability in a finely granulated and well-structured way, making use of up-to-date design and user interface methodology.”* can perhaps inspire CLARIN to work more on common CLARIN guidance about formats and known options for interoperability. Currently, the CLARIN VLO – repository of metadata harvested from all CLARIN centres – links resources directly to the CLARIN language switchboard as an inspiring example of interoperability.

Currently, the CLARIN checklist does not include recommendations about data formats, leaving it up to the research communities to choose what is most convenient. Encouraging data centres to provide guidance on recommended formats could, perhaps, inspire users to select data formats that are already supported by existing tools, viewers or applications. As the checklist is tailored to data centres, it does not directly address the recommendations about the data creation phase.

Reusability

The CLARIN checklist follows the recommendation to *“request clear and accessible data usage licence for (meta)data”* for data. However, in CLARIN the metadata are always open and harvestable via OAI-PMH without a specific licence.

The CLARIN checklist requires persistent identifiers for the data in the metadata, but besides this requirement, it is not specified in detail which information should be included in the metadata. In general, the CLARIN community agrees with the recommendation that provenance of meta(data) should be available whenever possible, but the CLARIN approach allows for user-defined metadata schemas as long as the schema for the metadata is defined in CMDI format. This variation in metadata schemas makes it very difficult to check whether the specified metadata elements cover the needed provenance of (meta)data, and checking the metadata for detailed information of provenance will, therefore, be beyond the scope of the assessment. Therefore, the checklist does not address either the quality or the coverage of the metadata. The guidelines also suggest use of domain-relevant community

[19] https://www.clarin.eu/content/virtual-language-observatory-vlo.

standards for data. In CLARIN, this is not addressed in the checklist, but it is work in progress for a thematic committee on standards.

4. IPR and open data recommendations

The PARTHENOS Project identified the need of researchers to work with large amounts of data that have terms of use and re-use conditions presented in a clear way.[20] On the one hand, open data and open access are an opportunity to promote innovation and development, and to connect researchers from across disciplinary and countries. On the other hand, there is a need expressed by the research communities to manage restricted access to protect certain resources. Limitations to re-using data are generally due to personal data protection, copyright issues, or database rights expressed by national laws. PARTHENOS' common goal is to support 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.

Research infrastructures play a key role in promoting Open Science and Open Innovation, which are viewed as two fundamental challenges in innovation and economic growth by the European Commission (see Section 4.2). They support the diffusion of open data and open access in the research practices, offer a guidance about legal issues concerning research data generally, and share policies and recommendations for open data and open access, in order to allow researchers to access different databases and tools. Research institutions and Cultural Heritage 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 recommendation for publishing research data is 'as open as possible, as closed as necessary'.

[20] PARTHENOS D2.1: User Requirements Report.

4.1 Legal framework

Data infrastructures promote easier exploitation of their data across global markets and borders, and among institutions and research disciplines, thanks to interoperability and access services. However, this increasing exchange of data also creates the need for new public policies. European and National Agencies support policies for re-use and data sharing to improve research and education outcomes and enhance economic returns.

Public research data have public good characteristics, and are often global public goods (Stiglitz 1999). The obligation to make public research data widely available, however, may collide with legal restrictions of reuse. On the one hand literature statements, declarations, and principles by various research organizations and disciplines [21], international governmental research-related organizations[22] and national governments support open access and reuse of data. On the other hand, a lack of clarity about the legal conditions and restrictions on the reuse of data compromises access and reuse of data.

This section presents an overview of the legal frameworks that are relevant for access to research data and its re-use. In Section 4.1.1 Intellectual property rights are discussed which are illustrated with a case study in Section 4.1.2. Section 4.1.3 focuses on the General Data Protection Regulation and handling sensitive data in particular, and Section 4.1.4 briefly introduces the Public Sector Information directive. Appendix IV of D3.1[23], the deliverable that preceded the present document, summarizes the EU and some National regulation to promote access and data re-use, collected by the Partners.

4.1.1 Intellectual property rights

Intellectual property rights (IPR) can be described as rights acquired over any work created or invented with the intellectual effort of an individual. Intellectual property can be divided into three categories: industrial property, copyright and database protection rights. Copyright law differs among countries, but thanks to rules derived from international treaties and European legislation, most countries have similar rules about what is protected or not by

[21] Science International (2015): Open Data in a Big Data World; CODATA report for the Group on Earth Observations (2015): The Value of Open Data Sharing; LIBER (2014): The Hague Declaration on Knowledge Discovery in the Digital Age.

[22] G8 (2013) Open Data Charter; Organization for Economic Co-operation and Development (2007): OECD Principles and Guidelines for Access to Research Data from Public Funding.

[23] D.3.1: Report on Guidelines for Common Policies Implementation.

copyright. National differences include types of works that are protected or the time period for which a certain protection is valid.[24]

Copyright plays a role when creating, sharing and re-using research data; database protection rights address the investment that is made in developing a database, e.g. the selection or organisation of the data. Scholars and researchers that want to re-use and share data need to know the terms of use for the database and the data content. Productive and successive uses are defined by what the legal rights are, who has these rights, under which conditions the data may be used and how the rights holder uses the rights to share data. Moreover, a researcher who wants to enrich data 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.

Since IPR depends on national law, each country may modify the users' rights. This context leads to legal uncertainty that is a serious impediment to the transnational re-use of research data. The legal uncertainty may be overcome if repositories require depositors to grant explicit permission to downstream users or to give up any intellectual property rights they may have in the data. In order to homogenize the approaches, international initiatives have been set up, such as licensing frameworks like the Creative Commons and RightsStatements.org. Furthermore, an ongoing consultation for updating copyright rules at EU level addresses these new challenges of the digital age continuously.[25]

4.1.2 Case study: Transnational 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 twelve hundred 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 file transfer or had to be created or curated by hand by the project researchers.

[24] See D3.1 Report on Guidelines for Common Policies Implementation, pp. 137-139 for national differences and European commonalities.

[25] https://ec.europa.eu/digital-single-market/en/modernisation-eu-copyright-rules#improvedrules.

This landscape of partners, formats and data types resulted in an exceptionally complex IPR situation, since many different licence types and restrictions already governed the data coming in, and these had to be roughly preserved going out. In addition, there were often competing voices and positions among the many communities and institutions the project was dealing with.

The final 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.

4.1.3 The General Data Protection Regulation and personal data

As a general rule, all personal data need some kind of legal protection. Since May 2018, this is regulated by a common European law, the General Data Protection Regulation (GDPR), to which all the European countries are bound, including any private and public stakeholders that process and/or store personal data. This regulation replaces the Data Protection Directive of 1995[26] and focuses on preventing the identification of living persons. As it emerges, the GDPR is a legislation entirely dedicated to the treatment of personal and sensitive data, making it an even more prominent issue than it used to be in the past

[26] https://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX:31995L0046.

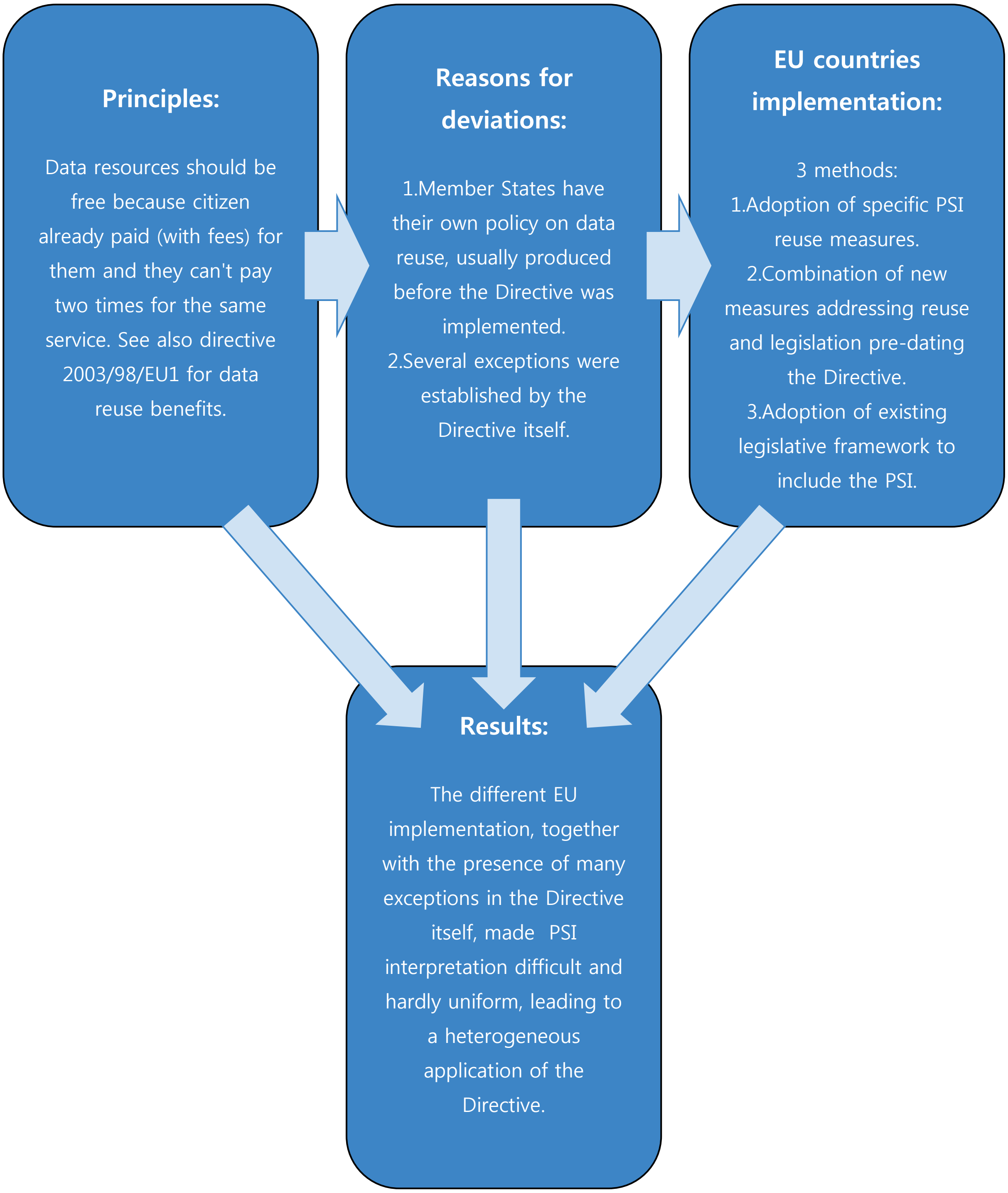
decades. This is increased by the exponential production of data in the last decades, given the informatization of public systems, the growth of social media platforms, as well as the mass collection of data of millions of users.

The new legislation determines that the processing of the personal data can be performed only when it is necessary for contracts or legal obligations in which the subjects is involved, or in case the processing is made in the interest of the subject. In all cases, the subject always needs to give her/ his consent to use the personal data by means of informed consent. As for the archiving of personal data, the GDPR allows the storage and archiving of personal data, only where the reference to the subject is limited as much as possible or anonymized, and only if the subject has given consent or when the archiving is made in the public interest.

In addition to addressing personal data in general, the GDPR also regulate the processing of sensitive data in particular. Generally speaking, sensitive data are data that need a high grade of protection, and in most cases, these are personal data. In short, not-special or common personal data are containing elementary data such as name, address or telephone number, whereas sensitive data are those that provide information of a potential sensitive nature, such as health, religion, political conviction, race, or sexual orientation. Legally, this kind of data is often defined as “special” personal data. Special personal data are subjected to a stricter protection regime than the latter. Some particular types of sensitive data even need more protection than on average, because disclosure may form a serious risk for particularly vulnerable people, as for example in the case of personal data of people who have witnessed or have been involved in circumstances as (past) wars, armed conflicts, medical or psychiatric treatment and handicaps (specially for children).

4.1.4 The Public Sector Information Directive

The European legislation on reuse of Public Sector Information (PSI) is a Directive which aims the free circulation and reuse of data produced by public institutions without restrictions. The following schema presents an overview of the PSI Directive situation in EU countries.



Principles:

Data resources should be free because citizen already paid (with fees) for them and they can't pay two times for the same service. See also directive 2003/98/EU1 for data reuse benefits.

Reasons for deviations:

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.

EU countries implementation:

3 methods:
1.Adoption of specific PSI reuse measures.
2.Combination of new measures addressing reuse and legislation pre-dating the Directive.
3.Adoption of existing legislative framework to include the PSI.

Results:

The different EU implementation, together with the presence of many exceptions in the Directive itself, made PSI interpretation difficult and hardly uniform, leading to a heterogeneous application of the Directive.

Figure 1: Schematic overview of the PSI Directive situation in EU countries.

The difficulties in applying the Directive in the European context led to a series of proposed revisions.[27] For the time being, the PSI is regarded as a minimum common regulatory framework.[28]

[27] Proposals for revisions: https://ec.europa.eu/digital-single-market/en/proposal-revision-public-sector-information-psi-directive.

[28] Example of good practice: http://www.bl.uk/aboutus/stratpolprog/pubsect-info-regulations/.

However, in the PARTHENOS Guidelines, great relevance is assigned to the adoption of a standard licensing framework as was mentioned above in Section 3.2.4.1. Taking into consideration the approaches of the different institutions involved in PARTHENOS, as well as the PSI Directive, it is possible to define the following recommendations:

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

A number of standardised electronic statements regarding the legal rights are discussed in more detail in Section 4.3 below.

4.2 Open Science

In 2014, the European Commission launched a policy on Open Innovation and Open Science with the goal of supporting the development of research through collaboration between people from different countries, sectors and disciplines. The overall objective of Open Science is to make scientific research data accessible under terms that enable reuse, redistribution and replicability of the research and its underlying data and methods. Open Science is thus supporting steering away from the standard practices of merely publishing research results in scientific publications, and towards sharing and using all available knowledge at the earliest stage of the research process. Open Science includes:

- Open data (available, intelligible, accessible, interoperable and re-usable data);
- Open access to academic, scientific, educational publications;
- Open source (for promoting the reproducibility of computer software code, using licences in which the copyright holder provides the rights to study, change, and distribute the software to anyone and for any purpose).
- Open reproducible research (Open Science workflows, transparency throughout the research lifecycle, shared protocols and standards)
- Open scientific evaluation (open peer review, alternative metrics for impact evaluation)

- Open policies and mandates (transnational, national, or institutional Open Access, Open Science or Open Data policies, mandates, recommendations or guidelines)[29]

To achieve these objectives and establish fundamentally new ways of how research is designed, performed and evaluated and how knowledge is shared, researchers have to re-organize the entire life cycle of research, from project design to publication of the results. Developing more sustainable, more connected and community-driven models of scholarly production is also meant to open science practice, and to generate projects of a greater socio-economic impact, primarily because of the increased accessibility to the scientific results.[30]

4.2.1 Open Access and Open Data

Open Access (OA) and Open Data are the pillars of Open Science. The Council of the European Union, on May 2016, in the final observations on “The transition towards an Open Science system” stresses that “open access to scientific publications and optimal reuse of research data are of utmost importance for the development of open science”. The Open Science Agenda defines two high-level aims for OA to be achieves by 2020: all peer reviewed scientific publications are freely accessible and FAIR data sharing is the default for scientific research.

4.2.1.1 Open access

Open Access (OA) is the practice of providing online access to scientific information that is free of charge to the user and is meant to be re-usable. According to the Budapest Declaration (2002) and the Berlin Declaration (2003), OA is the right to read, download and print scientific publications, as well as the right to copy, distribute, search, link, crawl, and mine information. Two main models for OA publications are emerging:

[29] For an exhaustive taxonomy of Open Science, see https://www.fosteropenscience.eu/resources.

[30] European Commission Directorate-General for Research and Innovation (2016): Open Innovation Open Science Open to the World.

Green open access: the published article or the final peer-reviewed manuscript is deposited by the author (self-archiving) in an online repository before, alongside or after (in presence of a 'data embargo') its publication.

Gold open access: the article is immediately provided in open access mode when published, by a commercial or institutional publisher. The payment of publication costs (APC) can usually be borne by:

- the university or research institute to which the researcher is affiliated
- the funding agency supporting the research
- subsidies or other funding models.

The Horizon 2020 Programme provides requirements and guidelines for guaranteeing OA to Scientific Publications and to Research Data produced by funded projects. According to the European Commission rules, OA is an obligation.

4.2.1.2 Case study: Lexicon philosophicum

Lexicon philosophicum. International Journal for the History of Texts and Ideas, http://lexicon.cnr.it, edited by ILIESI-CNR, represents a suitable case study of how a transition from a printed edition (Olschki 1985-2011) to a digital open access publication may work. Relaunched a new online journal, it has been created within the activities of the European Project Agora Scholarly Open Access Research in European Philosophy (2011-2014) and since then continues as institutional publication of ILIES-CNR.

Adopting the journal management and publishing system Open Journal Systems (OJS), the journal adheres to the open access protocols to improve the quality and the dissemination of scholarly publishing in the field of philosophy. The contributions published in the journal are made available in Open Access under the Creative Commons General Public License Attribution, Non-Commercial, Share-Alike version 3.0 (CCPL BY-NC-SA).

4.2.1.3 Open Data

Open data is open access to research data. Open access to research data refers to the right to access and reuse digital research data under the terms and conditions set out, for example, in a Grant Agreement. Research data refers to information, in particular facts or numbers, collected to be examined and considered as a basis for reasoning, discussion, or calculation.

As was mentioned in the Introduction, in a research context, examples of data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images. The focus is on research data that is available in digital form. Both scientific research and economic growth receive a considerable boost when research data are open and this will certainly also be significant for the Digital Single Market.[31] The benefits that users receive from making scientific information freely available to the global life science community is widely demonstrated.[32] Therefore, the same should apply to the Humanities. According to the Open Research Data Pilot[33], there two main types of Open Data:

- underlying data (the data needed to validate the results presented in scientific publications), including the associated metadata;
- any other data (for instance curated data not directly attributable to a publication, or raw data), including the associated metadata, as specified in the DMP – that is, according to the individual judgement by each project/grantee.

4.2.2 Case study: Implementing CCO licence on data: the case of EASY, the online archiving system of DANS

As an early adopter of open access and open data, DANS has decided to no longer require registration for users as standard. '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'

[31] https://ec.europa.eu/digital-single-market/en/policies/shaping-digital-single-market.

[32] See e.g. Open innovation, open science, open to the world : reflections of the Research, Innovation and Science Policy Experts (RISE) High Level Group and Munafò et al. (2017): A manifesto for reproducible science.

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

attention to the fact that, in accordance with the VSNU/KNAW Code of Conduct for Academic Practice[34], proper citation of research remains imperative.

The strategic decision was made to make the default setting Open access for everyone in EASY: the dataset files are accessible to all users of EASY and 'CC0 Waiver - No Rights Reserved' applies. DANS has contacted researchers who deposited their data in previous years to enquire if they objected to transforming their data to a CC0 licence. If depositors did not agree, they could opt out by choosing a more restricted category.

4.3 Licensing frameworks

A licence is legal document that the rights holder attaches to his/her work or resource for defining how to use and re-use it. It should be noted that licences designed for one type of subject matter (e.g. code, content or data) are not always best suited to licensing another type of subject matter.[35] For example, a database and its content may have separate rights and require different licences.[36] Over the years, some core standard models emerged like Creative Commons, Open Data Commons and RightsStatements that have already been adopted by many Humanities and Cultural Heritage institutions. Moreover, some communities have built upon these models to produce their own licensing framework such as the research infrastructure CLARIN for language resources.

In the following sections, four different licensing models are presented, each of which provides open as well as more restricted licences. A survey that was carried out among the PARTHENOS project partners shows that, in practice, a preference is given to open licences, or those that allow free re-use of resources, at least for research and educational purposes. However, in most cases the institutions request the use of licences with attribution. Sometimes, anyway, partners decided to use the rights statement “in copyright” mainly for resources that have data protection issues, in line with the provisions of the PSI Directive.

[34] http://www.vsnu.nl/files/documenten/Domeinen/Onderzoek/The_Netherlands_Code_of_Conduct_for_Academic_Practice_2004_%28version2014%29.pdf.

[35] https://opendatacommons.org/faq/licenses/#Why_Not_Use_a_Creative_Commons_or_FreeOpen_Source_Software_License_for_Databases.

[36] https://opendatacommons.org/faq/licenses/#Why_Do_You_Distinguish_Between_the_8220Database8221_and_its_8220Contents822.

4.3.1 Creative Commons[37]

The licensing model with the widest application these days is the Creative Commons licensing framework. Creative Commons was launched in 2001 and it was partially inspired by the Free Software Foundation. The creators wanted to help those who want to share their "works freely for certain uses, on certain conditions; or dedicate your works to the public domain".[38] It offers a framework of standardized licences, and some of them apply to data and databases. It provides different levels of data sharing, and is able, in this way, to cover a very wide variety of scenarios. The current development of the Creative Commons licensing framework takes into account 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 do not have any kind of limit to their re-use. It is possible, in fact, to apply the public domain licence in two cases only: if the IPR is expired or the creator has voluntarily surrendered it.

The free culture licences, instead, while having the same possibilities of re-use in the public domain, are characterized by maintaining some rights. In this case, the licence states only the attribution to the owner of the resources, that must be immediately recognisable. However, the free culture licences consent to third parties adapting the work, also for commercial re-use.

With the licences for resources that are not free culture, the resources have several limitations to their re-use: for example, it is not possible to adapt or derive other works from the original ones and commercial re-use is not allowed.

4.3.2 Open Data Commons[39]

Supported by the Open Knowledge Foundation since 2009, Open Data Commons provides a set of open data licences that enable to make one's data open and easily reusable. Unlike

[37] https://creativecommons.org/.
[38] https://creativecommons.org/about/history/.
[39] https://opendatacommons.org/licenses/.

the Creative Commons licences, they are specifically designed for data and databases. The Open Data Commons model has three different types of licences, that allow users to freely share, modify, and use the data(base):

- Public Domain Dedication and License
- Attribution for Data/Databases
- Open Database License

While the Public Domain Dedication and License (PDDL) imposes no restrictions, the other two do have attribution and sharing-alike requirements. The DDPL places the data(base) in the public domain waiving all rights. The Attribution License (ODC-BY) and the Open Database License (ODbL) permit sharing, creating and adapting the material. In addition, the ODbL credits the rights holder, it keeps licence and any original notices intact, and requires redistribution under same licence. Technological measures that restrict the work (such as DRM) are allowed as well as redistribution of a version without such measures.

4.3.3 RightsStatements[40]

Originating from the collaboration of Europeana and the Digital Library of America, RightsStatements have been specifically designed for Cultural Heritage institutions to provide a licensing framework able to cover the rights related to digital objects that they make available online in situations where the Creative Commons licences and other legal tools cannot be used. They are not intended to be used by individuals to license their own creations. RightsStatements can thus be seen as complementary of Creative Commons. They are divided in three main categories: in copyright, no copyright and other.

- The five “in copyright” statements allow the re-use of resources for educational and not commercial purposes and cover two particular cases: EU orphan works and rights-holder(s) not localizable or unidentifiable.
- 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 re-use or whose rights have been ascertained only for a specific jurisdiction.

[40] http://rightsstatements.org/en/.

- The last section, “other”, is devoted to unclear rights statements and probably is the most critical to assign. These rights statements, anyway, should be used only if is not possible to define a clearer rights statement or licence.

4.3.4 CLARIN Licensing Framework[41]

CLARIN, a research infrastructure for Language Resources and Technology, has formulated a licensing framework able to respond to different requirements concerning copyright and/or personal data protection issues by dividing licences into three categories:

- Public use (PUB);
- Academic use (ACA);
- and Restricted use (RES).

The PUB resources are freely usable and without re-use limitations. The resources that fall in ACA area, instead, are freely reusable only for research purposes. Users need access to resources via a Federated Identity Service. Finally, the RES resources are accessible just for research purposes and available only after having made a request to the rights holder via a separated application. Thus, the ACA and RES licences cover an area that the other standard licences are not able to cover, or cover partially.

4.4 Authentication and Authorisation Infrastructure

Due to the digital turn in Scientific Research and the availability of shared virtual environments, the need to access networked applications, remote and distributed data and services has become standard. As was mentioned above in Section 3.2.2.1, authentication and authorisation of users is a key feature for digital infrastructures. An authentication and authorisation infrastructure (AAI) is an infrastructure that provides support for authentication and authorisation[42] services. While authentication involves verifying the identity claimed by or for that particular entity, authorisation services manage the granting of approval to a system entity to access a system resource. Authentication and authorisation are often separated from the application and the data themselves. Authentication of the users is done

[41] https://www.clarin.eu/content/clarin-licensing-framework.

[42] RFC 4949 Internet Security Glossary (Version 2 August 2007): https://www.rfc-editor.org/rfc/pdfrfc/rfc4949.txt.pdf.

by the user’s Identity Providers, while the authorisation is done by the services based on the information received by the Identity Providers.

A Federated Identity Management System - enabling identity information to be shared among different contexts, entities and domains - is required to provide federated access to resources and services to users across different institutions. Whereas AAIs have been widely used in infrastructures within life and physical sciences, in the Social Sciences, Humanities and Cultural Heritage sectors AAI has become a key infrastructural component only recently. Both CLARIN[43] and DARIAH[44] - the two Landmarks in the S&CI ESFRI sector - developed and documented AAI components to manage the above issues.

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 login only once using their institutional credentials and access multiple services (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 not have to run username and password administration, and will have access to more tools for managing data. For a large scale 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 Identity Providers and Service Providers, or they can only act as one of the two.

[43] https://www.clarin.eu/node/3740.
[44] https://wiki.de.dariah.eu/display/publicde/DARIAH+AAI+Documentation.

5. Tools for implementing common policies

5.1 PARTHENOS Policy Wizard

In the course of the project, WP3 has worked towards creating the PARTHENOS Policy Wizard.[45] This online service aims at helping researchers to discover which data policy applies best to their particular data. The PARTHENOS Policy Wizard shows that many disciplines in the humanities are supported by a range of policies suggesting how data should be collected, processed, stored, and shared with other researchers. Some of these policies operate on a country level, because they depend on national regulations, while others are based on EU regulations and operate at a European level.

While most policies have been developed as discipline-specific guides, the PARTHENOS Policy Wizard also offers to search by topic, thus allowing researchers to explore not only those policies that apply to their particular discipline, but policies from neighbouring disciplines in the Humanities and Social Sciences as well. This innovative approach points to the commonalities between the disciplines, and offers the researchers a set of common solutions. To encourage this exchange even further, when there is no policy available for a certain discipline, a comparable policy from another discipline is suggested as an example. Should the researcher be aware of a policy that is not listed in the Policy Wizard, she/ he can add it through a simple interface that feeds the information about the new policy directly into the Policy Wizard backend.

Another interesting feature of the wizard is that it does not only share formal[46] policies, but aims to showcase those practices that have been adopted by a community of researchers, without being formalised or endorsed by any institution. These "best practices", are powerful normative instruments, in the form of a set of do’s and don’ts, that a community uses to regulate itself.

[45] Tykhonov et al. (2018): PARTHENOS Policy Wizard.

[46] Here, "formal policies” refers to those policies that are formally agreed upon a community and are often supported by an institute or research infrastructure that is leading in that particular field.

The wizard web prototype is developed as a HTML5 widget application, which can be easily integrated and made accessible through different websites (e.g. CLARIN, DARIAH) in the future. At the moment, the Policy Wizard is available via the PARTHENOS website.[47]

The Policy Wizard has a modular architecture and shares all available data via a REST API endpoint. All the information about the available policies is collected and curated in a Google spreadsheet, structured as a matrix. The ingest process of the PARTHENOS Policy Wizard (from the spreadsheet to the widget) is automated; changes in the matrix will be recognised directly in the wizard. 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, 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.

5.2 The PARTHENOS Data Management Plan for Archaeology

As was mentioned in Section 3.2, elaborating a carefully designed data management plan (DMP) is the key to the successful implementation of a FAIR data policy on the level of a research project. There is, however, no general consensus yet on what a DMP should specify exactly, and funder requirements for DMPs vary considerably. The requirements range from no formal requirements about the contents and structure of the DMP (e.g. German Research Foundation) to detailed templates provided by the funder itself (e.g. Horizon 2020).[48] Jones (2012) provided a summary of eight national funding organizations and their requirements for the United Kingdom. Some academic institutions have their own requirement of good scholarly practice, like the University of Edinburgh,[49] which sets additional requirements. To ease the situation, the Data Curation Centre (DCC) in the UK provides a website with an interactive template for various funders, called DMPonline.[50]

[47] https://parthenos.d4science.org/parthenos-wizard/.
[48] See Guidelines on FAIR Data Management in Horizon 2020 (2016).
[49] Research Data Service. Website of the University of Edinburgh's research data services for local data management services, The University of Edinburgh.
[50] Data Curation Centre, https://dmponline.dcc.ac.uk/.

The first draft of the PARTHENOS Data Management Plan template was presented in the deliverable D3.1,[51] and was based on the detailed description of the proliferation, complexity and issues with regards to data management in that same document. The task force in charge of the drafting of the template included researchers, data managers, and data curators. After a first round of internal reviews, a template to collect feedback from experts in the archaeological and heritage science domain, was drafted and circulated. The general comment was that the level of detail was too much is several sections of the template and that the structure of the DMP needed to be improved with guidance, examples, and links to guidelines. In the course of the past two years, this Data Management Plan was further tailored to meet specifically the needs of researchers in the field for Archaeology, which is described in this section.

5.2.1 Background to the work

The PARTHENOS Data Management Plan template for Archaeology is based on the DMP model developed within the Horizon 2020 framework[52] and it takes into account the PARTHENOS guidelines in the present document, as well as cross-references with the content of the other services developed by the project (i.e. the Wizard and the Standardization Survival Kit). The work joins up with European initiatives around Open Science and the FAIR principles[53], whose goal is to make scientific research data accessible and to guarantee the reuse, redistribution and replicability of the data. It anticipates the recommendations and actions promoted by the "European Commission Expert Group on FAIR data"[54], providing a tool that supports the compilation of a DMP. In addition, the need to offer online support for the compilation of a DMP was also strongly emphasized in a survey conducted by the Data Management team of the OpenAIRE project and the expert group on FAIR data.[55] The objective of this survey was to gather feedback from a group of experts who evaluated the H2020 DMP template with the aim of identifying any gaps and gather suggestions for improvement.

[51] See Section 3.3.1.1 of D.3.1: Report on Guidelines for Common Policies Implementation for the DMP description.

[52] See Data Management Template in Horizon 2020.

[53] See the website of the European Commission, Research and Innovation: https://ec.europa.eu/research/openscience/index.cfm.

[54] S. Hodson, et al. (2018): FAIR Data Action Plan: Interim recommendations and actions from the European Commission Expert Group on FAIR data (Version Interim draft).

[55] M. Grootveld et al. (2018) OpenAIRE and FAIR Data Expert Group survey about Horizon 2020 template for Data Management Plans (Version 1.0.0).

Taking the Horizon 2020 DMP model as a starting point, PARTHENOS developed a DMP template[56] that was extensively tested by the archaeological community, represented in PARTHENOS by ARIADNE.[57] A survey conducted among members of the ARIADNE consortium and subsequently extended to other experts in the domain, provided PARTHENOS with a comprehensive framework of standards and best practices for creating, storing and sharing data used by the archaeological community. These structure the DMP model that underlies the online tool.

5.2.2 Structure of the template

The PARTHENOS DMP template for Archaeology maintains the same structure as the Horizon 2020 DMP, to offer researchers a familiar model. In addition, the PARTHENOS DMP template provides various lists where the user can select standards and operational flows that are relevant particularly to projects in the field of Archaeology. Moreover, the DMP template will offer links to other relevant tools, such as the PARTHENOS Wizard and the Standardization Survival Kit (SSK), offering a wide range of background documentation. This makes the DMP tool also suitable for less experienced users, thanks to the combination of the rich content provided by the various tools.

The PARTHENOS DMP offers clear guidance to its stakeholders, mainly researchers and data repositories, to plan the life cycle of data. It will offer a long-term perspective by outlining how data will be generated, collected, documented, shared and preserved, taking into consideration commonalities and specific requirements of the archaeological and heritage science community. The PARTHENOS DMP template comprises sections about data collection and documentation, ethics, legal and security issues, data storage and preservation, and data sharing and reuse.

To facilitate the creation of the PARTHENOS DMP for Archaeology, PIN created an ad hoc application, which will be available on the PARTHENOS website.[58] The tool currently allows the compilation of the form and the ability to download a copy of the document in PDF and

[56] S. Bassett et al. (2017): A DMP template for Digital Humanities: the PARTHENOS model.

[57] ARIADNE (Advanced Research Infrastructure for Archaeological Dataset Networking in Europe) is a European network that was funded between 2013 and 2017 and which has been re-funded for the years 2018-22. See www.ariadne-infrastructure.eu.

[58] http://www.parthenos-project.eu/portal/dmp.

JSON. The ultimate goal is to obtain a machine-actionable DMP,[59] i.e. DMPs whose information can be processed and automatically understood by computers, and to produce documents that are interoperable and shareable within the stakeholder community.

5.2.3 The online tool

DMPs are generally created by researchers manually or by the use of online tools. When digital tools are used, the resulting files are often in textual formats (e.g. PDF or DOC) oriented towards human readability. Only in some cases the available tools provide a machine-readable format, i.e. the data is encoded to be processed automatically by the computers (e.g. JSON or XML).

The interface of the PARTHENOS DMP tool was designed to facilitate the creation of DMPs through the use of intuitive and user-friendly solutions. The questions are divided into subsequent pages, enriched by a common progress bar that presents itself as the main point of reference for the user. The overall view of the various parts that make up the model guides the user step by step, indicating the time approximately needed to complete each one. Each group of pages with similar thematic questions, divided between compulsory and optional, is enriched by informative pop-ups. If some of the points deemed mandatory for submitting the DMP have not been completed, this is displayed in red in the progress bar. At the end of the compilation procedure it is possible to download the completed form in PDF format, or in JSON.[60]

In the future, the application aims to allow interoperability and sharing of the DMPs, among those research communities that adopt common solutions to facilitate the cooperation between their systems. Therefore, it is necessary to consider both the syntactical and the semantic aspects of the data. Computers can interpret most of the information in a syntactic way, if these are encoded in standard formats like XML or JSON, but they are not able to understand it if they are not using controlled vocabularies and shared standards. The DMPs generated in the first version of the application already meet the requirements for syntactic

[59] T. Miksa et al. (2018): Ten simple rules for machine-actionable data management plans.

[60] The JSON file is fundamental within the application, as it offers users the possibility to save a copy of their work. In fact, the process of completing the questionnaire can be interrupted at any time by downloading the JSON file that contains the current data. This file can be reloaded into the interface, allowing to continue and finish the work.

interoperability, thanks to the encoding in JSON format. They will subsequently be adapted at the semantic level, using controlled vocabularies, standards and data models that are open and shared between the research communities, in particular the CIDOC CRM PARTHENOS entities [CRMpe] semantic model, developed within PARTHENOS.[61]

Further activities that are planned in the near future include the translation of the PARTHENOS DMP model and the related guidelines into different languages to provide national versions to those countries that have not yet developed their own model (i.e. versions in Italian, Spanish, Greek, German, etc.). At the moment, only the archaeological model has been developed and tested within the research community of PARTHENOS. Training seminars and consultancy services will be organized to promote and disseminate the PARTHENOS DMP model and its guidelines to the archaeological research communities, and to ensure consistent dissemination of PARTHENOS results with the aim of increasing awareness of open results in archaeology.

5.3 A Community Driven Research Data Management Protocol for Humanities

In the course of the past year the *research data* working group at Science Europe[62] has worked on the development of a data management protocol that is domain specific, the *Domain Data Protocol (or DDP)*. This protocol outlines the good practices for data management in a specific discipline or community sharing common data practices, standards and cultures. Whereas the various DMP templates, such as the DMP template for Archaeology described in the previous section, are targeting research data management at the level of individual research projects in a particular research field (e.g. Archaeology), the domain data protocols operate at the domain level (e.g. Humanities) and are envisioned to be used at different moments of the data lifecycle and by different stakeholders.

This section introduces the development of a *domain data protocol for the humanities* based on a framework for RDM by Science Europe.[63] The protocol is still in development, therefore we will focus here on its rationale as well as the methodology, both in relation to the

[61] PARTHENOS D5.1: Report on the common semantic framework.

[62] Science Europe is the association of European research funding and research performing organisations, see: https://www.scienceeurope.org.

[63] P. Doorn (ed., 2018): Science Europe Guidance Document Presenting a Framework for Discipline-specific Research Data Management.

community of researchers in the Humanities as well as with the Science Europe working group, where all the domains are represented.

5.3.1 Core Requirements

The very first step taken toward the development of Domain Data Protocols for all the domains, and the Humanities in particular, is to formulate a set of *Core Requirements* that are relevant for all stakeholders. These requirements are the result of a series of consultations between the representatives of each domain (Humanities, Archaeology, Language Data, Social Sciences Survey Research, Social Sciences - Psychology, Natural Sciences, Bioinformatics, Biology - Plant Science, Climate Research, Technical Science). The core requirements lie at the heart of the future development of the various Domain Data Protocols, and were presented in the Practical Guide to the International Alignment of Research Data Management (Science Europe 2018: 9-10).[64] The requirements were approved by the Science Europe General Assembly and were launched at the end of January 2019. Based on the following core requirements each discipline will develop its own Domain Data Protocol.

1. Data description and collection or re-use of existing data

a. How will new data be collected or produced and/or how will existing data be re-used?

b. What data (for example the kinds, formats, and volumes) will be collected or produced?

2. Documentation and data quality

a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany data?

b. What data quality control measures will be used?

3. Storage and backup during the research process

a. How will data and metadata be stored and backed up during the research process?

b. How will data security and protection of sensitive data be taken care of during the research?

[64] Science Europe (2018): Practical Guide to the International Alignment of Research Data Management.

4. Legal and ethical requirements, codes of conduct

a. If personal data are processed, how will compliance with legislation on personal data and on data security be ensured?

b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

c. How will possible ethical issues be taken into account, and codes of conduct followed?

5. Data sharing and long-term preservation

a. How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

b. How will data for preservation be selected, and where will data be preserved long-term (for example a data repository or archive)?

c. What methods or software tools will be needed to access and use the data? d. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

6. Data management responsibilities and resources

a. Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?

b. What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

5.3.2 The Core Requirements, the Domain Data Protocols, and DMP templates

As mentioned above, the DDPs operate on the level of the various domains, whereas DMP templates are tailored to specific research fields. As such, they act as complementary building blocks and come into play in different stages of the planning and implementation of a research project. The Domain Data Protocol will probably be most useful at the very beginning of a research project when knowledge of data management is not defined in detail yet. In addition, it might be particularly useful for researchers whose data is very interdisciplinary and doesn't fully apply to one discipline or the other (e.g. oral history). Since the main idea is that research funders will endorse the Domain Data Protocol for the Humanities once it is established, researchers adhering to this DDP will have an easier job when preparing a DMP for a particular project, as many aspects of it will already be covered in the protocol.

5.3.3 Outlook to future developments

Ultimately, the scope of the work with the core requirements will be that of building up different Domain Data Protocols (as different are the research domains) on top of the core requirements. However, this is seen as the very last step of a process that in the end lead will to domain-specific data protocols. In the short term, Science Europe will convene with stakeholders from every domain - in our case from the arts and humanities domain (e.g. Research Infrastructures like DARIAH, CLARIN, or projects like PARTHENOS) in order to make sure that the core requirements will be adopted by the related communities and recommended for adoption to their networks of researchers in the form of best practices.

6. Conclusion

The present deliverable is a product of the combined efforts of the different partners that worked together in WP3: 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). In collaboration with all PARTHENOS partners (sixteen organisations, sometimes consisting of multiple institutes) the four tasks aimed to give 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. After the first results of this effort were published in D3.1, the work was evaluated by a group of experts engaged by WP2. This feedback, combined with ongoing research and insights, resulted in the final version of the present document. Apart from this deliverable that concludes the work of WP3, three concrete and tangible outputs are delivered:

- the Guidelines to FAIRify data management and make data reusable
- the PARTHENOS Policy Wizard[65]
- the PARHTENOS Data Management Plan Tool[66]

Each of these outputs originates from a particular topic addressed in WP3 theoretically first, and then turned into practical, user-friendly, applications. As they are based on the research carried out within the community, the applications are tailored specifically to the PARTHENOS stakeholders' needs. To enhance their discoverability and impact, all of them are also integrated into the PARTHENOS Training Suite.

During the work in WP3, we found that the data management landscape is very fragmented and organised in disciplinary silos. This results in a myriad of best practices and guidelines, by making the management aspect of data very difficult to penetrate and to apply to the researchers' own data. However, after studying this heterogenous landscape in more detail, the higher-level commonalities became evident, which are, in turn, reflected throughout this deliverable as well as in the tangible outputs mentioned above.

[65]https://parthenos.d4science.org/parthenos-wizard/.
[66] http://www.parthenos-project.eu/portal/dmp.

7. References

Borgman, C. 2015. *Big Data, Little Data, No Data*. *Scholarship in the Networked World*. Cambridge, Massachusetts: The MIT Press.

Bassett, Sheena, Sara Di Giorgio, Franco Niccolucci and Paola Ronzino. 2017. A DMP template for Digital Humanities: the PARTHENOS model. Available at: https://www.garr.it/en/docs/4014-conferenza-2017-selected-papers-14-bassett.

Doorn, Peter (ed.). 2018. Science Europe Guidance Document Presenting a Framework for Discipline-specific Research Data Management. D/2018/13.324/1. Available at: https://www.scienceeurope.org/wp-content/uploads/2018/01/SE_Guidance_Document_RDMPs.pdf

Felicetti, Achile, Carlo Meghini, Christos Papatheodorou and Julian Richards. 2016. Final report on standards and project registry. ARIADNE Project Deliverable D3.4. Available at: http://www.ariadne-infrastructure.eu/Resources/D3.4-Final-Report-on-Standards-and-Project-Registry.

Grootveld, Marjan, Ellen Leenarts, Sarah Jones, Emilie Hermans and Eliane Fankhauser. 2018. OpenAIRE and FAIR Data Expert Group survey about Horizon 2020 template for Data Management Plans (Version 1.0.0). DOI: 10.5281/zenodo.1120244.
Available at: https://zenodo.org/record/1120245#.XE8QSaeZMmp.

Hodson, Simon, Sarah Jones, Sandra Collins, Françoise Genova, Natalie Harrower, Daniel Mietchen, Ruta Petrauskaité and Peter Wittenburg. 2018. FAIR Data Action Plan: Interim recommendations and actions from the European Commission Expert Group on FAIR data (version Interim draft). DOI: 10.5281/zenodo.1285290. Available at: https://zenodo.org/record/1285290#.XE8O76eZMmp.

Hollander, H. et al. 2017. Report on Guidelines for Common Policies Implementation. PARTHENOS Project Deliverable D3.1. Available at:

http://www.parthenos-project.eu/Download/Deliverables/D3.1_Guidelines_for_Common_Policies_Implementation.pdf.

Jones, S. (2012): Summary of UK research funders' expectations for the content of data management and sharing plans. Digital Curation Centre (DCC). Available at http://www.dcc.ac.uk/sites/default/files/documents/resource/policy/FundersDataPlanReqs_v4%204.pdf.

Miksa, Tomasz, Stephanie Simms, Daniel Mietchen and Sarah Jones. 2018. Ten simple rules for machine-actionable data management plans. DOI: 10.5281/zenodo.1172672. Available at: https://zenodo.org/record/1434938#.XE8Sh6eZMmo.

Munafò, Marcus R, Brian A. Nosek, Dorothy V.M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J . Ware, and John P.A. Ioannidis. 2017. A Manifesto for Reproducible Science. Nature Human Behaviour 1(1):0021, DOI: 10.1038/s41562-016-0021.

Owens, T. 2011. *Defining Data for Humanists: Text, Artifact, Information or Evidence?* in *Journal of Digital Humanities*, Vol.1. Available at: http://journalofdigitalhumanities.org/1-1/defining-data-for-humanists-by-trevor-owens/.

Pennock, M. (2007): Digital Curation: A life-cycle approach to managing and preserving usable digital information". In: Library & Archives Journal, (1). Available at: http://www.ukoln.ac.uk/ukoln/staff/m.pennock/publications/docs/lib-arch_curation.pdf.

Research Data Service. Website of the University of Edinburgh's research data services for local data management services., The University of Edinburgh.
Available at: https://www.ed.ac.uk/information-services/research-support/research-data-service.

Stiglitz, Joseph E. 1999. Knowledge as a Global Public Good. In *Global Public Goods*, Inge Kaul, Isabelle Grunberg, and Marc Stern (eds.), New York: Oxford University Press.

Trippel, T.; Zinn, C. (2015): DMPTY - A Wizard For Generating Data Management Plans. In: Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland, (123), 71-78.
Available at: http://www.ep.liu.se/ecp/123/006/ecp15123006.pdf.

Tyknonov, Vyacheslav, Hella Hollander, Jerry de Vries, and Francesca Morselli. 2018. PARTHENOS Policy Wizard. Poster presented at the DH Benelux 2018, June 7-8, Amsterdam.

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

Wittenburg, P. et al. 2013-2018. CLARIN B Centre Checklist, Version 6, Last Update: 2018-02-07, Status Approved by the Centre Committee, https://office.clarin.eu/v/CE-2013-0095-B-centre-checklist-v6.pdf.

Appendix I: Guidelines to FAIRify data management and make data reusable

GUIDELINES to FAIRify data management and make data reusable

PARTHENOS

ABOUT THIS GUIDE

This guide offers a series of guidelines to align the efforts of data producers, data archivists and data users in humanities and social sciences to make research data as reusable as possible.

The guidelines result from the work of over fifty PARTHENOS project members. They were responsible for investigating commonalities in the implementation of policies and strategies for research data management and used results from desk research, questionnaires and interviews with selected experts to gather around 100 current data management policies (including guides for preferred formats, data review policies and best practices, both formal as well as tacit).

With a focus on (meta)data and repository quality, the PARTHENOS team extracted a set of twenty guidelines which different disciplines have in common.

For easy reference, the team assigned each of the guidelines to making data Findable, Accessible, Interoperable or Reusable. This subdivision is based on the FAIR Data Principles which were first published by FORCE11 (2016) and are intended to guide those wishing to enhance the reusability of research data. Each of the PARTHENOS guidelines is accompanied by specific recommendations for data producers and data users on the one hand and for data archivists on the other hand. The icons below the guidelines visualise which stake-holder is addressed.

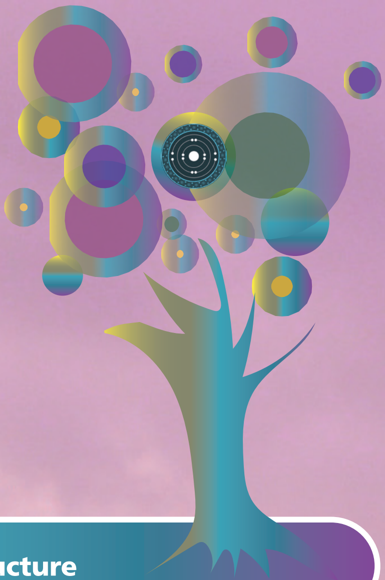
The lamp icon shows recommendations for data producers and data users such as researchers and research communities in history, archaeology, language studies and social science studies.

The wheel icon shows recommendations for research infrastructures and data archives in research institutes and cultural heritage institutions.

PARTHENOS is a consortium of sixteen European research institutions and infrastructures. PARTHENOS members aim to increase the reusability of research data by building bridges between the data life cycles of research communities, data repositories, research infrastructures and cultural heritage institutions in the interrelated fields of the humanities and social science.

20 GUIDELINES

to FAIRify data management and make data reusable



1 Invest in people and infrastructure

An important prerequisite to be able to implement the rest of the nineteen guidelines in this guide, is to invest in data infrastructures and in hiring and educating data experts.

Get acquainted with best practices in research data management. Check out the PARTHENOS training modules on data management or have a look at the CESSDA Data Management Expert Guide.

Invest in hiring and educating data experts and define a budget for making investments in technical infrastructure and staff.

FINDABLE

Research data should be easy to find by both humans and computer systems and based on mandatory descriptions of the metadata that allows the discovery of interesting datasets.

2 Use persistent identifiers

Locating data is a necessary condition for any other step from access to reuse. To be findable, any data object and dataset should be uniquely and persistently identifiable over time with a persistent identifier (PID). A PID continues to work even if the web address of a resource changes. PIDs can take different forms, such as a Handle, DOI, PURL, or URN.

Reference the PID which was assigned to your dataset in your research output.

Select the appropriate form of persistent identification schema and assign a PID to every resource. Use the PID Guide from NCDD to decide on the right PID for your research infrastructure.

3 Cite research data

If research data have a persistent identifier and are cited in accordance with community standards, the corresponding data objects or datasets are more easily found.

Get acquainted with data citation guidelines that are specific to your field or discipline and cite research data accordingly.

Provide information about best practices in data citation to research communities and make it easy for data users to cite data, e.g. by using a standardised button which says 'How to cite this dataset'.

4 Use persistent author identifiers

A persistent author identifier (e.g. VIAF, ISNI or ORCID) helps to create linkages between datasets, research activities, publications and researchers and allows recognition and discoverability.

Distinguish yourself from any other researcher or research group. Apply for an author identifier if you do not already have one and reference it in your dataset.

Reference author identifiers in the metadata.

5 Choose an appropriate metadata schema

Metadata is essential in making data findable, especially the metadata which is used for citing and describing data. A metadata schema is a list of standardised elements to capture information about a resource, e.g. a title, an identifier, a creator name, or a date. Using existing metadata schemas will ensure that international standards for data exchange are met.

To enable the discovery of content, describe research data as consistently and completely as possible. Include enough information for the data to be accessed and understood later on. If possible, use an existing metadata schema which fits the type of data object or dataset you are describing.

Clearly state which metadata schema you apply and recommend to the research community. To enrich datasets at data deposit, consider having a data submission form which collects additional metadata, e.g. about the provenance of the data.

ACCESSIBLE

Research data should be easily accessible and retrievable with well-defined access conditions using standardised communication protocols.

6 Choose a trustworthy repository

A certified repository offers a trustworthy home for datasets. Certification is a guarantee that data are stored safely, and will be available, findable and accessible over the long-term. Examples of certification standards are CoreTrustSeal, nestor seal and ISO 16363 certification.

Make your data accessible through a trustworthy repository. In addition, if you follow the repositories' standards (on preferred file formats, metadata schemas etc.) you can make sure that all requirements for making data FAIR are met.

Clearly state the level of certification on your website. If you are not (yet) certified, state how you plan to ensure availability, findability, accessibility and reusability in the long-term.

7 Clearly state accessibility

Access information specifies how a data user may access a dataset. When depositing data in a data repository, it should be clear which access options a data depositor can choose.

When choosing an access option, consider legal requirements, discipline-specific policies and ethics protocols when applicable. Choose Open Access when possible. When you collect personal data, ask yourself whether it contains any information which might lead to participants' identities being disclosed, what participants consented to and which measures you have taken to protect your data. If your data cannot be published in Open Access, the metadata should be, allowing data discovery.

Encourage (meta)data to be published in Open Access. Cleary state restricted access options for sensitive (meta)data that should not be part of the publicly accessible (meta)data. In this case, strive to make the (meta)data available through a controlled and documented access procedure.

8 Use a data embargo when needed

During a data embargo period, only the description of the dataset is published. The data themselves are not accessible. The full (meta)data will become available after a certain period of time.

Clearly state why and for what period a data embargo is needed. Make the (meta)data openly available as soon as possible.

Specify whether a data embargo is allowed and what conditions apply.

9 Use standardised exchange protocols

By using standardised exchange protocols, research infrastructures can make (meta)data publicly accessible and harvestable by e.g. search engines, vastly improving accessibility.

Use standardised protocols such as SWORD, OAI-PMH, ResourceSync and SPARQL. Convert metadata schemas into XML or RDF. Maintain a registry for protocol endpoints, the path at which research data can be accessed, and publish them.

To speed up discovery and uncover new insights, research data should be easily combined with other datasets by humans as well as computer systems.

INTEROPERABLE

10 Establish well documented machine-actionable APIs

Well documented and machine-actionable APIs - a set of subroutine definitions, protocols, and tools for building application software - allow for automatic indexing, retrieval and combining of (meta)data from different data repositories.

Document APIs well and make it possible to deliver the schema of the (meta)data model. Consider showing examples of how to successfully mine data from different endpoints and combine them into new data sets usable for new research.

11 Use open well-defined vocabularies

The description of metadata elements should follow community guidelines that use open, well defined and well known vocabularies. Such vocabularies describe the exact meaning of the concepts and qualities that the data represent.

Use vocabularies relevant to your field, and enrich and structure your research output accordingly from the start of your research project.

Give examples of vocabularies the research community may use, based on research domain specifics.

12 Document metadata models

Clearly documenting metadata models helps developers to compare and make mappings between metadata.

Publish the metadata models in use in your research infrastructure. Document technical specifications and define classes (groups of things that have common properties) and properties (elements that express the attributes of a metadata section as well as the relationships between different parts of the metadata). For metadata mapping purposes, list the mandatory and recommended properties.

13 Prescribe and use interoperable data standards

Using a data standard backed up by a strong community, increases the possibility to share, reuse and combine data collections.

Check with the repository where you want to deposit your data what data standards they use. Structure your data collection in this format from the start of your research project.

Clearly specify which data standard your institution uses, pool a community around them and maintain them especially with a perspective on interoperability. Good examples are CMDI (language studies) and the SIKB0102 Standard (archaeology).

14 Establish processes to enhance data quality

To boost (meta)data quality and, therefore, interoperability, establish (automatic) processes that clean up, derive and enrich (meta)data.

Establish procedures to minimise the risk of mistakes in collecting data. E.g. choose a date from a calendar instead of filling it in by hand.

Invest in tools to help clean up (meta)data and to convert data into standardised and interoperable data formats. Combine efforts to develop workflows and software solutions for such automatic processes, e.g. by using machine learning tools.

15 Prescribe and use future-proof file formats

All data files held in a data repository should be in an open, international, standardised file format to ensure long-term interoperability in terms of usability, accessibility and sustainability.

From the start of your research project think about future-proof file formats. Use preferred formats which are recommended by the data repository and are independent of specific software, developers or vendors.

Encourage the use of formats that are considered suitable for long-term preservation such as PDF-A, CSV and MID/MIF files. Provide an easy-to-find and detailed overview of accepted file formats.

Research data should be ready for future research and future processing, making it self-evident that findings can be replicated and new research effectively builds on already acquired, previous results.

REUSABLE

16 Document data systematically

To make clear what can and what cannot be expected in a dataset or repository, data should be systematically documented. Being transparent about what's in the data and what isn't facilitates trust and, consequently, data reuse.

Provide codebooks, including a description of methodology, a list of abbreviations, a description of gaps in the data, the setup of the database, etc.

17 Follow naming conventions

Following a precise and consistent naming convention - a generally agreed scheme to name data files - makes it significantly easier for future generations of researchers to retrieve, access and understand data objects and datasets.

Consult the policies and best practices for your research discipline or domain to find the most suitable naming convention.

Clearly state best practices to create and apply specific file naming conventions.

18 Use common file formats

By using standardised file formats that are widely used in your community, reusability is increased.

Use current popular file formats next to archival formats to share your data, e.g. Excel (xlsx) and CSV or ESRI Shapefiles next to MID/MIF files.

Publish the data in popular formats next to the archival format if they are not the same.

19 Maintain data integrity

Research data which were collected should be identical to the research data which are accessed later on. To ensure data authenticity, checks for data integrity should be performed.

Implement a method for version control. The guarantee that every change in a revised version of a dataset is correctly documented, is of integral importance for the authenticity of each dataset.

To identify if a file has been modified, it is essential to record provenance - the origin of the data plus any changes made over time - and to compare any copy with the original. A data integrity check can be performed by means of a fingerprint such as a checksum, or by a direct comparison of two files. Provide a mechanism to address different versions, for example by adding the version to the identifier as a search parameter.

20 License for reuse

To permit the widest reuse possible of (meta)data, it should be clear who the (meta)data rights holder is and what license applies.

Make sure you know who the (meta)data rights holder is before publishing your research data.

Communicate the (meta)data license and reuse options transparently and in a machine-readable format. To improve interoperability, try to map your licenses to frameworks which are already widely adopted such as Creative Commons.

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.

Appendix II: Glossary[67]

TERM	ABBREVIATION	DEFINITION	LINK
Archive		A place or collection containing records, documents, or other materials of historical interest. The Free Dictionary An archive may contain digital or analogue materials or both.	
Art and Architecture Thesaurus	**AAT**	The Art & Architecture Thesaurus (AAT) is a controlled vocabulary used for describing items of art, architecture, and material culture. The AAT is used by museums, art libraries, archives, cataloguers, and researchers in art and art history. The AAT is a structured vocabulary of around 44,000 concepts, including 131,000 terms, descriptions, bibliographic citations, and other information relating to fine art, architecture, decorative arts, archival materials, and material culture. Wikipedia	http://www.getty.edu/research/tools/vocabularies/aat/
Article Processing Charge	**APC**	Is a fee which is sometimes charged to authors to make a work available open access in either an open access journal or hybrid journal. This fee is usually paid by an author's institution or research funder rather than by the author themselves. Wikipedia	
Berlin Declaration		Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (2003).	https://openaccess.mpg.de/Berlin-Declaration.
Bit preservation		A baseline preservation approach that ensures the integrity of digital objects and associated metadata over time in their original form, even as the physical storage media which houses them evolves and changes. National Digital Stewardship Alliance Glossary Is defined as the required activities to ensure that the bit-streams remain intact and readable. Scalable Preservation Environments – Policy Elements	https://blogs.loc.gov/thesignal/2011/09/b-is-for-bit-preservation/ http://wiki.opf-labs.org/display/SP/Further+Reading#FurtherReading-FurtherReadingBitPreservation

[67] The present glossary serves as an addition to the Glossary in Appendix I of D.3.1: Report on Guidelines for Common Policies Implementation. Terms that are mentioned in D3.1, are not repeated here.

Budapest Declaration		Budapest Declaration on World Heritage (2002)	https://ec.europa.eu/digital-single-market/en/implementation-public-sector-information-directive-member-states.
Centre informatique national de l'enseignement supérieur	**CINES**	Offers computer services for research and higher education in France.	https://www.cines.fr/en/
Centre National de la Recherche Scientifique	**CNRS**	Is a public organization under the responsibility of the French Ministry of Education and Research and carry out all research capable of advancing knowledge and bringing social, cultural, and economic benefits for society.	http://www.cnrs.fr/
Certification		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 though its committee on Conformity Assessment has produced a number of standards defining international consensus on voluntary criteria in certification good practice. 4C Project - Glossary terms. Applied to digital repositories, a certification testifies the quality of a repository in relation to its stability, reliability, preservation and dissemination capability.	
CLARIN Concept Registry	**CCR**	Offers a collection of concepts, identifiable by their persistent identifiers, relevant for the domain of language resources. CLARIN Concept Registry	https://www.clarin.eu/ccr
Common Language And technology Research INfrastructure	**CLARIN**	Is a European research network working in the field of archiving and processing of language-related resources in the humanities and social sciences.	https://www.clarin.eu/
Component MetaData Infrastructure	**CMDI**	Provides a framework to create and use self-defined metadata formats. CLARIN-D User Guide	https://portal.clarin.nl/node/4061
Consiglio Nazionale delle Ricerche - Istituto Linguistica Computazionale	**CNR-ILC**	Institute of the National Research Council of Italy (CNR). Carries out research activities in strategic scientific areas of the Computational Linguistics.	http://www.ilc.cnr.it/
Consiglio Nazionale delle Ricerche - Istituto per il Lessico Intellettuale Europeo Storia delle Idee	**CNR-ILIESI**	Institute of the National Research Council of Italy (CNR). Is dedicated to the history of cultural and scientific terminology.	http://www.iliesi.cnr.it/
Consiglio Nazionale delle Ricerche - Opera del Vocabolario Italiano	**CNR-OVI**	Institute of the National Research Council of Italy (CNR). Responsible for the development of the Historical Dictionary of the Italian language.	http://www.ovi.cnr.it/
Consortium of European Social Science Data Archives	**CESSDA**	Is a consortium for promoting the results of social science research and supporting international research cooperation.	https://www.cessda.eu/

Content Management System	**CMS**	Manages the creation and modification of digital content. It typically supports multiple users in a collaborative environment. Wikipedia	
Creative Commons	**CC**	Is an American non-profit organization devoted to expanding the range of creative works available for others to build upon legally and to share. The organization has released several copyright-licences known as Creative Commons licences free of charge to the public. Wikipedia	https://creativecommons.org/
Creative Commons 0	**CC0**	Besides licences, Creative Commons also offers through CC0 a way to release material worldwide into the public domain. CC0 is a legal tool for waiving as many rights as legally possible. Wikipedia	https://creativecommons.org/about/cc0
Creative Commons Public License	**CCPL**	Is one of several public copyright licences that enable the free distribution of an otherwise copyrighted work. A CC licence is used when an author wants to give people the right to share, use, and build upon a work that they have created. Wikipedia	https://creativecommons.org/share-your-work/licensing-types-examples/
Cultural Heritage Institutions	**CHIs**	Museums, galleries, libraries and archives. Europeana	
Data		Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship. - *C.L. Borgman (2015). Big Data, Little Data, No Data: Scholarship in the Networked World. MIT Press*.	
Data annotation		Is a type of metadata added to the original data, or part of it, pertaining and aiming to adding information or making information explicit.	
Data embargo		is a period during which access to academic journals is not allowed to users who have not paid for access. The purpose of this is to ensure publishers have revenue to support their activities. Wikipedia	
Data policy		Are norms regulating the data management and publications of research data. They range from recommendations to enforcement. IFDO - International Federation of Data Organizations	
Data Quality Policy		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.	
Digital Object Identifiers	**DOI**	Is a persistent identifier or handle used to uniquely identify objects, standardized by the International Organization for Standardization. Wikipedia	https://www.doi.org/

Digital preservation		A formal endeavour to ensure that digital information of continuing value remains accessible and usable. It involves planning, resource allocation, and application of preservation methods and technologies, and it combines policies, strategies and actions to ensure access to reformatted and "born-digital" content, regardless of the challenges of media failure and technological change. The goal of digital preservation is the accurate rendering of authenticated content over time. Wikipedia	
Digital Single Market	**DSM**	Is a policy belonging to the European Single Market that covers digital marketing, E-commerce and telecommunication. It was announced in May 2015 by the Juncker Commission. Wikipedia	https://ec.europa.eu/digital-single-market/
Directory of Open Access Journals	**DOAJ**	Is a community-curated online directory that indexes and provides access to high quality, open access, peer-reviewed journals.	https://doaj.org/
Electronic Archiving System	**EASY-DANS**	Is an online archiving system for depositing and downloading scientific research data. EASY was launched in the spring of 2007, with the aim of offering an archive system that was simpler than other archiving systems at the time, and with the distinctive feature that the user can upload his data himself. Wikipedia	https://easy.dans.knaw.nl/ui/home
European Association of Databases for Education and Training	**EUDAT**	EUDAT is an European project, co-funded within the 7th Framework Programme. It includes funding agencies that invest in research infrastructures and programmes of research, infrastructure operators and research communities who rely on the availability of data-management services, national data centres and providers of connectivity and the users who rely on the availability of data and services.	https://www.eudat.eu/
European Grid Infrastructure	**EGI**	Is a federated e-Infrastructure set up to provide advanced computing services for research and innovation.	https://www.egi.eu/
European Open Access Agenda		Presents five broad lines for actions: Foster Open Science, Remove barriers to Open Science, Develop research infrastructures for Open Science, Mainstream Open Access to research results, Embed Open Science in society	http://ec.europa.eu/research/openscience/pdf/draft_european_open_science_agenda.pdf
Fachhochschule Potsdam	**FHP**	University of Applied Sciences Potsdam	https://www.fh-potsdam.de/
FAIR	**FAIR**	Guidelines to improve the findability, accessibility, interoperability, and reuse of digital assets. GO FAIR Initiative	https://www.nature.com/articles/sdata201618

Federated Content Search	**FCS**	is an information retrieval technology that allows the simultaneous search of multiple searchable resources. A user makes a single query request which is distributed to the search engines, databases or other query engines participating in the federation. The federated search then aggregates the results that are received from the search engines for presentation to the user. This is often a technique to integrate disparate information resources on the web. It can also be a technique to integrate multiple data sources within a large organization or "enterprise." Wikipedia	
Federated Content Search Clarin	**FCS-CLARIN**	an *interface specification* that decouples the *search engine* functionality from its *exploitation*, i.e. user-interfaces, third-party applications, and to allow services to access heterogeneous search engines in a uniform way. CLARIN FCS API	https://www.clarin.eu/
Gold Open Access		immediate open access that is provided by a publisher. OpenAIRE	https://www.openaire.eu/oa-policies-mandates
Green Open Access		Immediate or delayed open access that is provided through self-archiving. OpenAIRE	https://www.openaire.eu/oa-policies-mandates
Horizon 2020 Programme		Horizon 2020 is the eighth framework programme funding research, technological development, and innovation. The programme's name has been modified to "Framework Programme for Research and Innovation". Wikipedia	https://ec.europa.eu/programmes/horizon2020/
Hybrid open-access journal		Is a subscription journal in which some of the articles are open access. Wikipedia	
Institut National de Recherche en Informatique et en Automatique	**INRIA**	The French National Institute for computer science and applied mathematics, promotes “scientific excellence for technology transfer and society”.	https://www.inria.fr
International Organization for Standardization	**ISO**	Is an independent, non-governmental international organization that brings together experts to share knowledge and develop voluntary, consensus-based, market relevant International Standards that support innovation and provide solutions to global challenges. ISO	https://www.iso.org/
International Standard Book Number	**ISBN**	Is a unique numeric commercial book identifier. Publishers purchase ISBNs from an affiliate of the International ISBN Agency. Wikipedia	https://www.isbn-international.org/
International Standard Name Identifier	**ISNI**	Is an identifier for uniquely identifying the public identities of contributors to media content such as books, television programmes, and newspaper articles. Such an identifier consists of 16 digits. It can optionally be displayed as divided into four blocks. Wikipedia	http://www.isni.org/
Istituto Centrale per il Catalogo Unico delle biblioteche italiane e per le informazioni bibliografiche	**ICCU**	Carries out coordination functions with due respect to library autonomy in the field of National Library Service as well as cataloguing projects accomplished through the use of new information technology.	http://www.iccu.sbn.it/

Koninklijke Nederlandse Akademie van Wetenschappen - Data Archiving and Networking Service	**KNAW- DANS**	Promotes quality in science and scholarship, innovation and knowledge valorisation. As a research organisation, the Academy is responsible for a group of outstanding national research institutes.	https://www.knaw.nl/
Licensing framework		Is a standardized and harmonized set of licences that provide an overview for use and reuse of data.	
Logical preservation		Is the part of digital preservation that ensures that the bits remain understandable and usable according to preservation purpose.	
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). Curation Costs Exchange	https://public.ccsds.org/Pubs/650x0m2.pdf
Machine readable		Is data (or metadata) which is in a format that can be understood by a computer. Wikipedia	
National Grid Initiative	**NGI**	Are organisations set up by individual countries to manage the computing resources they provide to the European e-Infrastructure. EGIWiki	
National Research and Education Network	**NREN**	Are specialised internet service providers dedicated to supporting the needs of the research and education communities within their own country. GÉANT	
Network Attached Storage	**NAS**	is a file-level computer data storage server connected to a computer network providing data access to a heterogeneous group of clients. Wikipedia	
Object Management System	**OMS**	In an Integrated Project Support Environment, the system which maintains information about the system under development. The Free On-line Dictionary of Computing	
Online Computer Library Center	**OCLC**	Is an American non-profit cooperative organization "dedicated to the public purposes of furthering access to the world's information and reducing information costs". Wikipedia	https://www.oclc.org/
Open Archival Information System	**OAIS**	Is a conceptual framework for an archival system dedicated to preserving and maintaining access to digital information over the long term. Brian Lavoie, *Meeting the challenges of digital preservation: The OAIS reference model*	
Open Archives Initiative Protocol for Metadata Harvesting	**OAI-PMH**	Is 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. Wikipedia	https://www.openarchives.org/OAI/openarchivesprotocol.html
Open Journal System	**OJS**	Is a journal management and publishing system that has been developed by the Public Knowledge Project.	https://pkp.sfu.ca/ojs/
Open Research Data Pilot		Aims to make the research data generated by Horizon 2020 projects accessible with as few restrictions as possible, while at the same time protecting sensitive data from inappropriate access. OpenAIRE	http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-

			access-data-management/open-access_en.htm
Open Researcher and Contributor ID	**ORCID**	Is a non-proprietary alphanumeric code to uniquely identify scientific and other academic authors and contributors. Wikipedia	https://orcid.org/
Open Science	**OS**	Is the movement to make scientific research, data and dissemination accessible to all levels of an inquiring society, amateur or professional. Wikipedia	http://data.consilium.europa.eu/doc/document/ST-9526-2016-INIT/en/pdf https://ec.europa.eu/research/openscience/pdf/draft_european_open_science_agenda.pdf.
Open Preservation Foundation	**OPF**	Sustains technology and knowledge for the long-term management of digital cultural heritage, in all its forms. Open Preservation Foundation	http://openpreservation.org/
Persistent Identification and Sustainable Access	**PISA**	Is the international standard ISO 24619:2011 that specifies requirements for the persistent identifier (PID) framework and for using PIDs as references and citations of language resources in documents as well as in language resources themselves. International Organization for Standardization	https://www.iso.org/obp/ui/fr/#iso:std:iso:24619:ed-1:v1:en
Persistent identifier	**PID**	Is a long-lasting reference to a document, file, web page, or other object. Wikipedia	http://www.athenaeurope.org/getFile.php?id=779
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.	
PREservation Metadata Implementation Strategies	**PREMIS**	Is an international working group concerned with developing metadata for use in digital preservation. Wikipedia	https://www.oclc.org/research/activities/pmwg.html
Preservation strategy		Is the implementation of proactive, scalable and sustainable strategies ensuring that digital data remain accessible and reusable over time. Digital Curation Centre	http://www.dcc.ac.uk/ http://www.nationalarchives.gov.uk/archives-sector/advice-and-guidance/managing-your-collection/preserving-digital-collections/developing-a-digital-preservation-strategy-and-policy/
Public Domain	**PD**	The public domain consists of all the creative works to which no exclusive intellectual property rights apply. Those rights may have expired, been forfeited, expressly waived, or may be inapplicable. Wikipedia	https://creativecommons.org/share-your-work/public-domain/
Public Domain Mark	**PDM**	Is a symbol used to indicate that a work is free of known copyright restrictions and therefore in the public domain. It is analogous to the copyright symbol, which is commonly used to indicate as copyright notice that a work is	https://creativecommons.org/share-your-work/public-domain/pdm/

		copyrighted. The Public Domain Mark was developed by Creative Commons. Wikipedia	
Public Section Informative	**PSI**	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	
Raw data see also **Underlying data**		Is data that has not been subjected to processing, analyses or any other manipulation. Science data glossary	
Repository		Is a place where things may be put for safekeeping. The Free Dictionary	
Representational State Transfer	**REST**	Is an architectural style that defines a set of constraints and properties based on HTTP. Web Services that conform to the REST architectural style, or RESTful web services, provide interoperability between computer systems on the Internet. Wikipedia	
Research Infrastructure	**RI**	Are facilities, resources and services used by the science community to conduct research and foster innovation. European Commission	
Resource Description Framework	**RDF**	Is a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed. W3C	https://www.w3.org/TR/rdf-schema/
Significant properties		Are the characteristics of digital objects that must be preserved over time in order to ensure the continued accessibility, usability, and meaning of the objects. A. Wilson, Significant Properties Report, 2007, p. 8	
Simple Knowledge Organization System	**SKOS**	Is a W3C recommendation designed for representation of thesauri, classification schemes, taxonomies, subject-heading systems, or any other type of structured controlled vocabulary. Wikipedia	https://www.w3.org/2004/02/skos/
Società Internazionale per lo Studio del Medioevo Latino	**SISMEL**	Is a cultural institute for research, training and scientific promotion purposes.	http://www.sismelfirenze.it/
SPARQL Protocol and RDF Query Language	**SPARQL**	Is a semantic query language for databases, able to retrieve and manipulate data stored in Resource Description Framework (RDF) format. Wikipedia	https://www.w3.org/TR/sparql11-protocol/
Text Encoding Initiative	**TEI**	Is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics. TEI	http://www.tei-c.org/
Trinity College Dublin	**TCD**		
Underlying data see also **raw data**		the data needed to validate the results presented in scientific publications. Horizon 2020	
Uniform Resource Identifier	**URI**	Is a string of characters designed for unambiguous identification of resources and extensibility via the URI scheme. Wikipedia	http://www.iana.org/
Uniform Resource Locator	**URL**	Is a reference to a web resource that specifies its location on a computer network and a mechanism for retrieving it. Wikipedia	http://www.iana.org/

Uniform Resource Name	**URN**	Is an Internet resource with a static name that remains valid even if its data is moved to another location. Techopedia	http://www.iana.org/
Virtual International Authority File	**VIAF**	Combines multiple name authority files into a single OCLC-hosted name authority service. VIAF	https://viaf.org/
Warehouse Management Systems	**WMS**	Is a software application, designed to support and optimize warehouse or distribution centre management. They facilitate management in their daily planning, organizing, staffing, directing, and controlling the utilization of available resources, to move and store materials into, within, and out of a warehouse, while supporting staff in the performance of material movement and storage in and around a warehouse. Wikipedia	
World Wide Web Consortium	**W3C**	Is an international community that develops open standards to ensure the long-term growth of the Web.	https://www.w3.org/
Zentrum für Informationsmodellierung - Austrian Centre for DigitalHumanities	**ZIM-ACDH**	The centre's focus is on applied research in the area of information and data processing in the humanities, with special emphasis on the theory of data-modelling, and the practical implementation of this research topic in teaching and projects.	https://informationsmodellierung.uni-graz.at/