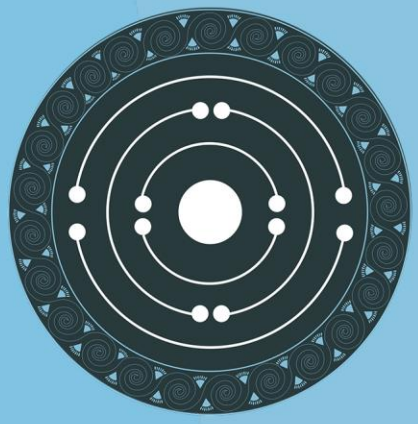

PARTHENOS

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

D2.3 REPORT ON THE ASSESSMENT OF DATA POLICIES AND STANDARDIZATION

CLARIN (WP leader)

MIBACT-ICCU (Task leader)

NIOD-KNAW (Task leader)

FHP

AA

CNR-OVI

Huma-Num

OEAW

SISMEL

UCPH

With contributions from all PARTHENOS partners

31-07-2017

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

REPORT ON THE ASSESSMENT OF DATA POLICIES AND STANDARDIZATION

Deliverable Number	D2.3
Dissemination Level	Public
Delivery date	31 July 2017
Status	Final
Author(s)	Steven Krauwer (CLARIN, WP Leader) Karolien Verbrugge (NIOD-KNAW, Task Leader) Sara Di Giorgio (MIBACT-ICCU, Task Leader) Claus Spiecker (FHP) Christos Chatzimichail (AA) Emiliano Degl'Innocenti (CNR-OVI) Hélène Gautier (Huma-Num) Roberta Giacomi (SISMEL) Lina Henriksen (UCPH) Klaus Illmayer (OEAW) Antonio Davide Madonna (MIBACT-ICCU) Maurizio Sanesi (SISMEL) With contributions from all PARTHENOS partners

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	D2.3 Report on the Assessment of Data Policies and Standardization
Document title	PARTHENOS_D2.3_Report on the Assessment of Data Policies and Standardization
Author(s)	Steven Krauwer (CLARIN, WP Leader) Sara Di Giorgio (MIBACT-ICCU, Task Leader) Karolien Verbrugge (NIOD-KNAW, Task Leader) Claus Spiecker (FHP) Christos Chatzimichail (AA) Emiliano Degl'Innocenti (CNR-OVI) Hélène Gautier (Huma-Num) Roberta Giacomi (SISMEL) Lina Henriksen (UCPH) Klaus Illmayer (OEAW) Antonio Davide Madonna (MIBACT-ICCU) Maurizio Sanesi (SISMEL) With contributions from all PARTHENOS partners
Dissemination level/distribution	Public

Document History

Version/date	Changes/approval	Author/Approved by
V 0.1 18.07.2017	First integrated draft	Steven Krauwer (integrator)
V 0.2 19.07.2017	Numbering system updated and sound (I think)	Steven Krauwer (integrator)

V 0.3 19.07.2017	Management summary and Introduction added	Steven Krauwer (integrator)
V 0.4 19.07.2017	Karolien's bits and comments added; man. Sum and intro enhanced; placeholders for Karolien's 10 sheets	Karolien Verbrugge
V 0.5 20.07.2017	Claus Spiecker's bits and last comments	Claus Spiecker
Final 31.07.2017	Final corrections and editing	Sheena Bassett

Table of Contents

Executive Summary

This document is the deliverable D2.3 *"Report on the Assessment of Data Policies and Standardization"*. It is part of the process to collect needs and experiences from the project constituencies (research communities, ESFRI and non-ESFRI Research Infrastructures, other projects) and to organize and feed them as requirements for implementation carried out by WP3, 4, 5, 6, 7 and 8 of the PARTHENOS project. After a first collection of requirements documented in deliverable D2.1 *"Report on User Requirements"*, published on 31 January 2016 and further updated on 20 October 2016, the content was taken up by WP3 (Common policies and implementation strategies) and WP4 (Standardization). The role of WP3 was to distil draft guidelines for Data Policy Implementation, for Quality Assessment, and for dealing with IPR, Open Data and Open Access, resulting in deliverable D3.1 *"Guidelines for Common Policies Implementation (draft)"*, published on 25 April 2017. The role of WP4 was the preparation of a minimum standardization package and the supporting tools for its use, resulting in deliverable D4.1 *"Standardization Survival Kit"*, published on 31 October 2016, and D4.2 *"Report on Standardization"*, published on 26 May 2017.

The task of WP2 on the basis of deliverables D3.1 and D4.1 was twofold, as described in the DoW:

- **Assessment of data policies**: The section concerns the assessment of the project outcomes concerning policies about data lifecycle, data curation and long-term preservation; access and authentication policies; quality assessment of digital repositories; and IPR management and privacy issues. It reviews the usability of the guidelines produced, and describes any amendment perceived as necessary.
- **Assessment of standardization**: The section concerns the assessment of the technical standardization solutions produced in WP4 comparing them with the communities' needs, and proposing amendments if necessary.

The structure of the document reflects this: after a general introduction in chapter 1, chapter 2 *"Part 1: Assessment of data policies / recommendations produced by WP3"* assesses the proposals made in deliverable D3.1, followed by chapter 3 "*Part 2: Assessment of standardization"*, which assesses the proposals made in deliverable D4.1

and D4.2. In chapter 4 one finds an alphabetic list of acronyms and abbreviations used throughout the report.

The review of the PARTHENOS Guidelines by the experts was very positive and showed that a lot of the relevant information was covered and the general structure based on FAIR principles was positively approved, too. It became clear that the Guidelines fulfil the intention to support several different types of stakeholders particularly relevant to the PARTHENOS project, during the process of data management and define policies concerning quality of data, metadata and repositories and IPR, open data and open access.

The specific recommendations concerning aspects such as formal presentation, avoiding buzzwords, presuppositions, weak and confusing definitions and to consider the differences between different stakeholders and their methodological approaches etc. are specifically detailed in chapter 2. The assessment of standardization in chapter 3 has shown that WP4 has made a comprehensive overview of the most important standards and resources used in different academic disciplines, although standards in social sciences were somewhat underrepresented and further research seems necessary to connect the needs of the researchers in the social sciences to the proposed technical standardization solution produced by WP4.

Furthermore, in this deliverable suggestions are made for resources and standards that are still missing. The annex contains the systematic result of this gap analysis. Besides this, recommendations are given about other issues that would be an asset to the SSK, for example, including research scenarios on legal and privacy issues that form a barrier to data use without restrictions, and including not only open standards, but also commercial standards if they are well-used in the communities.

The findings reported in D2.3 will be fed back into WP3 and WP4, where they will be taken into account when preparing the final versions of the recommendations for policies (D3.2) and for standardization (D4.3 and D4.4).

1. Introduction and structure of the document

1.1 Background

This document is the deliverable D2.3 *"Report on the Assessment of Data Policies and Standardization"*. It is part of the process to collect needs and experiences from the project constituencies (research communities, ESFRI and non-ESFRI Research Infrastructures, other projects) and to organize and feed them as requirements for implementation carried out by WP3, 4, 5, 6, 7 and 8 of the PARTHENOS project. After a first collection of requirements during months 1-9, laid down in deliverable D2.1 "*Report on User Requirements"*, published on 31 January 2016, and updated on 20 October 2016, the content was taken up by WP3 (Common policies and implementation strategies) and WP4 (Standardization). The role of WP3 was to distil draft guidelines for Data Policy Implementation, for Quality Assessment, and for dealing with IPR, Open Data and Open Access, resulting in deliverable D3.1 *"Guidelines for Common Policies Implementation (draft)"*, published on 25 April 2017. The role of WP4 was the preparation of a minimum standardization package and the supporting tools for its use, resulting in deliverable D4.1 *"Standardization Survival Kit"*, published on 31 October 2016, and D4.2 *"Report on Standardization"*, published on 26 May 2017.

1.2 Tasks to be carried out and structure of the document

The task of WP2 on the basis of deliverables D3.1 and D4.1 was twofold, as described in the DoW:

- **Assessment of data policies**: The section concerns the assessment of the project outcomes concerning policies about data lifecycle, data curation and long-term preservation; access and authentication policies; quality assessment of digital repositories; and IPR management and privacy issues. It reviews the usability of the guidelines produced, and describes any amendment perceived as necessary.
- **Assessment of standardization**: The section concerns the assessment of the technical standardization solutions produced in WP4 comparing them with the communities' needs, and proposing amendments if necessary.

The structure of the document reflects this: after a general introduction in chapter 1, chapter 2 *"Part 1: Assessment of data policies / recommendations produced by WP3"* assesses the proposals made in deliverable D3.1, followed by chapter 3 "*Part 2: Assessment of standardization"*, which assesses the proposals made in deliverable D4.1 and D4.2. In chapter 4 one finds an alphabetic list of acronyms and abbreviations used throughout the report.

1.3 Acknowledgements

On the whole, both deliverables D3.1 and D4.1 were received very positively by both the participants in WP2 preparing this deliverable D2.3 and by the external experts consulted, whose specific expertise has been extremely valuable for the identification of gaps and inconsistencies, for raising questions, for indicating priorities and for suggesting better ways to describe or explain the proposals made.

The authors of this report would like to thank many PARTHENOS colleagues and especially the following experts (plus a number of them who preferred to remain anonymous) from other institutions for their time and effort and their valuable contributions: Anna Maria Tammaro, University of Parma, Italy; Antonella De Robbio, University of Padua, Italy; Benjamin Stular, Research Centre of the Slovenian Academy of Sciences and Arts (ZRC SAZU), Slovenia; Birte Christensen-Dalsgaard, University of Aarhus, Denmark; David Nathan, Linguist at Groote Eylandt Language Centre, Northern Territory, Australia.; Dennis Zielke, Fraunhofer Information Centre for Planning and Building (Fraunhofer IRB), Germany; Federico Nurra, Institut national de recherches archéologiques préventives (INRAP), France; Gerasimos Chrysovitsanos, Academy of Athens, Greece; Jochen Klar, Leibniz Institute for Astrophysics (AIP), Potsdam, Germany; Kleopatra Kalafata, Humanities at Scale (HaS); Luca dell'Agnello, National Center of the National Institute for Nuclear Physic for Research and Development (INFN-CNAF), Italy; Marjo Bakker, Netherlands Institute for War Documentation (NIOD), Institute for War, Holocaust and Genocide Studies, Netherlands; Marzia Piccininno, Istituto Centrale per il Catalogo Unico delle Biblioteche Italiane (ICCU), Italy; Natascha Schumann, GESIS Leibniz Institute for Social Science Data Archive, Germany; Neil Jefferies, Bodleian Library, University of Oxford, United Kingdom; René van Horik, Data Archiving and

Networked Services (DANS), Netherlands; Silvia Trani, Sapienza University of Rome, Italy; Véronique Ginouvès, Centre national de la recherche scientifique (CNRS), France.

2. Part 1: Assessment of data policies / recommendations produced by WP3

This chapter starts with an explanation of the approach adopted, which includes a description of the methodology and a presentation of the general results. The three following sections address the quality of data, metadata and repositories (2.2), the implementation of data policies (2.3), and IPR, open data and open access (2.4). For the ease of reading by the authors of D3.1 the numbering in these sections follows the numbering in D3.1, prefixed by "2", the number of this chapter.

2.1 Introduction to the approach adopted for Part 1

2.1.1 Methodology adopted for Assessing the Guidelines about Data Policies

One of the many tasks in PARTHENOS project is the development of Guidelines for Common Policies Implementation, paving the way for common data policy implementation, shared methods for quality assessment repositories and clear policies on IPR, Open Data and Open Access for the different research communities represented by the PARTHENOS Consortium. A first draft of the Guidelines for Common Policies Implementation has been elaborated by WP3 and presented in the Deliverable 3.1, published in April 2017. The PARTHENOS Guidelines will be a comprehensive report formed by three distinct sections:

1) on Quality Assessment of Repositories, including data and metadata about the policies and methods for the quality assessment of repositories, data and metadata,
2) on Data Policy Implementation, covering policies for data lifecycle, curation and long-term preservation, and access and authentication,
3) on Intellectual Property Rights (IPR), Open Data and Open Access, a section that reports the policies about IPR and Open Data, and the frameworks enabling Open Access according to academic best practices and EU recommendations.

The Guidelines should be conceived as a ‘guide to good practice’, and as such it includes examples, practical cases and answers to users’ questions.

The Task 2.1, involved in the definition of user requirements on data policies[1], has also the responsibility for testing and validating the policies produced by WP3, to guarantee that the results and the methodology presented in the Guidelines can satisfy the needs of a large community of researchers represented by the partners.

For the assessment of the D3.1 Guidelines for Common Policies Implementation, Task 2.1 conducted an online consultation among experts selected by the partners. The consultation had the objective to evaluate the Guidelines and to gather possible amendments, suggestions, new requirements from experts with different profiles as potential stakeholders and users of the Guidelines, in order to improve the document and to share in our communities.
Each partner selected at least two experts to cover all the following different profiles:

- Policy Maker
- Data Consumer
- Content Provider / Research Information Manager
- expert on Data Management
- expert on Quality Assessment
- expert on IPR, Open Data, Open Access

Each expert was asked to review the Guidelines either as a whole or only a specific section relevant to them. For example an expert on IPR, Open Data and Open Access reviewed the principles and the Guidelines related to this specific topic.
It was very important to receive a descriptive and quality input from the experts as the answers of the questionnaire are part of this deliverable and will help WP3 to improve the Guidelines.

[1] See D2.1 User Requirements Report: https://goo.gl/3IwI5J

For this reason task 2.1 created an online survey[2] with questions. Google Forms was adopted as platform for creating the survey and collecting the answers. The survey was designed to understand how far the Guidelines presented by PARTHENOS meet the needs of the organization or the user. The experts were asked to evaluate the guidelines and to make suggestions and point out best practices for improving the document. Each participant received an invitation letter with the entire Deliverable D3.1 as an attachment. More information on the structure of the document was enclosed to enable the experts to determine which section was most relevant to them for analysis. The survey itself started with a general introduction where information on the respondents was collected, followed by four main sections of questions that correspond to the structure of D3.1: the introduction and the three chapters on 1. Quality Assessment of Data, Metadata and Digital Repositories, 2. Data Policy Implementation with regards to data management, and 3. IPR, Open Data and Open Access. It was estimated that to answer the questionnaire it would take 15-20 minutes per section.

The structure of the questionnaire in detail:

- A GENERAL INTRODUCTION of the survey with the objective of the consultation and some instructions. Here it was possible to download the questionnaire so the respondent were able to read the guidelines and prepare the answers offline and then to complete the online survey.
- B SECTION SPECIFY YOUR PREFERENCE, where the experts were asked if they agree to the publishing of their contributions with their names and affiliation or if they preferred their contributions to be published anonymously.
- C INFORMATION ON THE RESPONDENT for collecting information about each expert such as name, research domain, skills and other.

The survey continued with four main sections of questions that corresponded to the structure of D3.1 as described previously.

2 https://docs.google.com/forms/d/e/1FAIpQLSfhnjLHUyPANqyEm1YpQAArk0LV5VU5ByAkk_Ff4r6eHb0FXA/viewform

All these sections had a similar structure: the experts were asked to rank the information, to comment and to point out best practices. The paragraphs about the WP3 methodology for identifying the commonalities and the guidelines weren't included in the assessment.
After each question the experts were asked if the information was helpful and the corresponding paragraphs and pages of the guidelines were indicated. The survey offered a closed answer with four different degrees of satisfaction ranging from not at all to completely. The questions were very simple and the answers were expected to give a clear statement from the expert as to if and how much the Guidelines met the needs of his organization.

After this, the experts were invited to contribute suggestions, additions, examples and best practices. At the end of each chapter they were asked about their general impression and if there were other issues that the chapter should examine additionally. The consultation started on the 11th of May and ended 31st of May 2017. A total of twenty two responses were received, from a very varied perspective, considering the nationality of the participants, the research domains and the institutions they belong to.

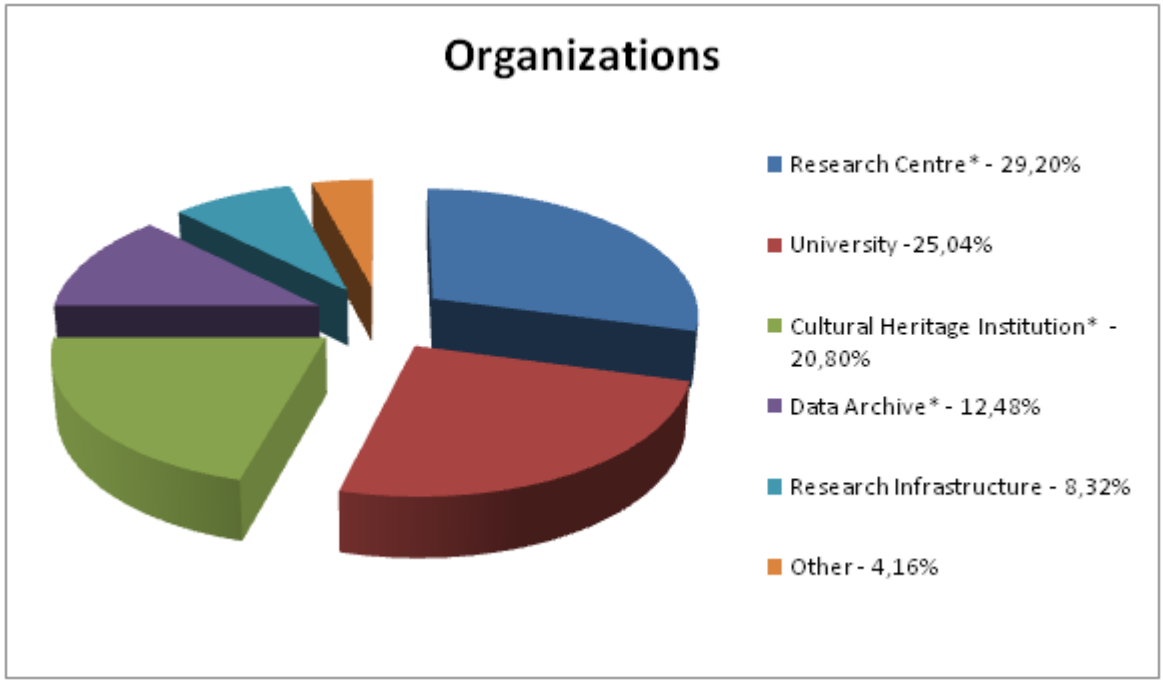

Fig. 1 – Most of the respondents belong to Research Centres, Universities and Cultural Heritage Institutions. 75% of the respondents in the Cultural Heritage Institutions belong to Libraries and the 25% to Archives.

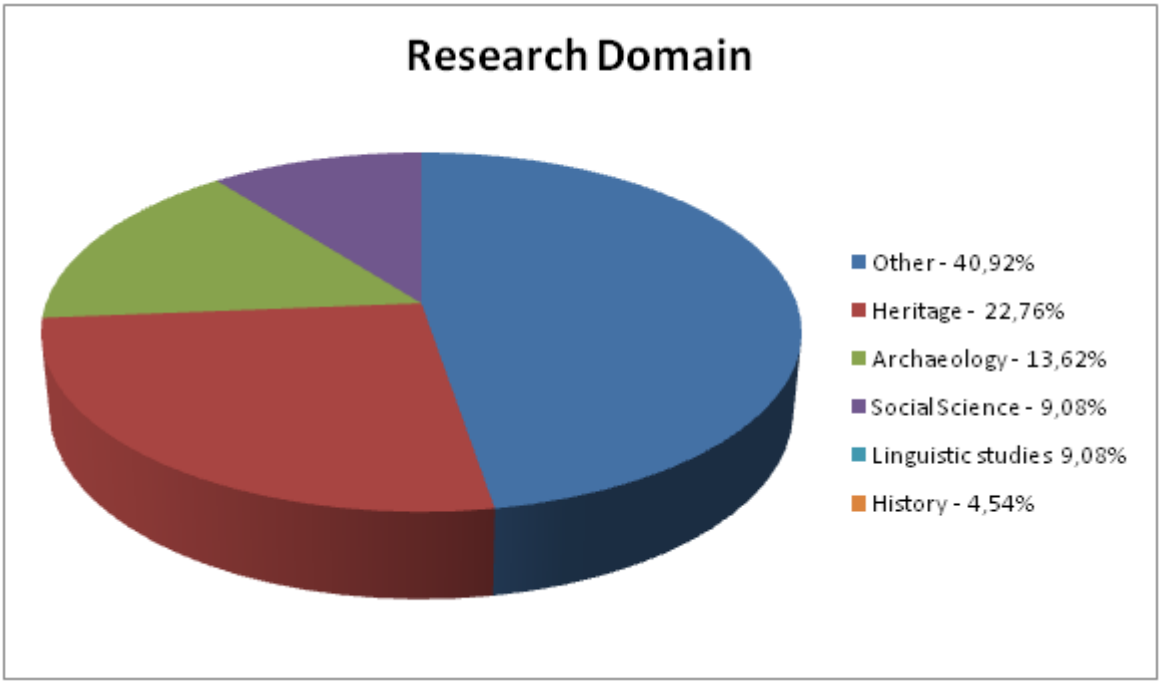

Fig. 2 – Most (40,92%) of the respondents declared 'other' when asked for their Research Domain and pointed out to be either involved in Archival Science (in particular born-digital archival), Computing applied to High Energy Physics, Digital Humanities, Digital Libraries and Scholarship. With 22,76% the Cultural Heritage Sector was the domain with the highest single coverage of all domains.

2.1.2 General results of the consultations

The Guidelines are intended to support several different types of stakeholders particularly relevant to the PARTHENOS project, during the process of data management and define policies concerning quality of data, metadata and repositories and IPR, open data and open access. The Guidelines should provide common recommendations, operative suggestions and best practice in order to enable cross-discipline data use and reuse, data policies to improve the data quality and long-term preservation, and policies addressing sensitive data and privacy issues. The aim of the consultation was to determine the relevance of the Guidelines by external experts, investigating if the most relevant issues have or have not been implemented as planned and asking for amendments and suggestions, to improve the final version of the Guidelines and to secure its optimal quality and impact.

In general, the reviews and the rating of the Guidelines were very positive, it became clear that a lot of relevant information was covered in the document and the general structure based on FAIR principles received positive feedback.

From the point of view of the formal presentation of the Guidelines, the most important recommendations and lessons learnt for improving the document were:

- to present the information contained in a better and clearly arranged form: the document could be vastly improved if each section was more clear, concise and to the point, moving the historical contexts and self-justifications to an appendix or another document;
- to check carefully, with native English speakers, the consistency of approach, style and expression of the language;
- to avoid jargon, buzzwords, presuppositions, weak and confusing definitions and acronyms.

Looking at the content of the introduction, most of the experts considered this part to be well balanced and useful for their institutions and the most relevant recommendations for this part were the following:

- to produce an introduction that is more focused on the Guidelines' purpose,
- to consider the differences between different stakeholders and their methodological approaches,
- to provide a better integration of principles, framework and stakeholders, to provide a better explanation for some topics about data (i.e. long term preservation, ethical issue of working data, differences between management and stewardship actions on data).

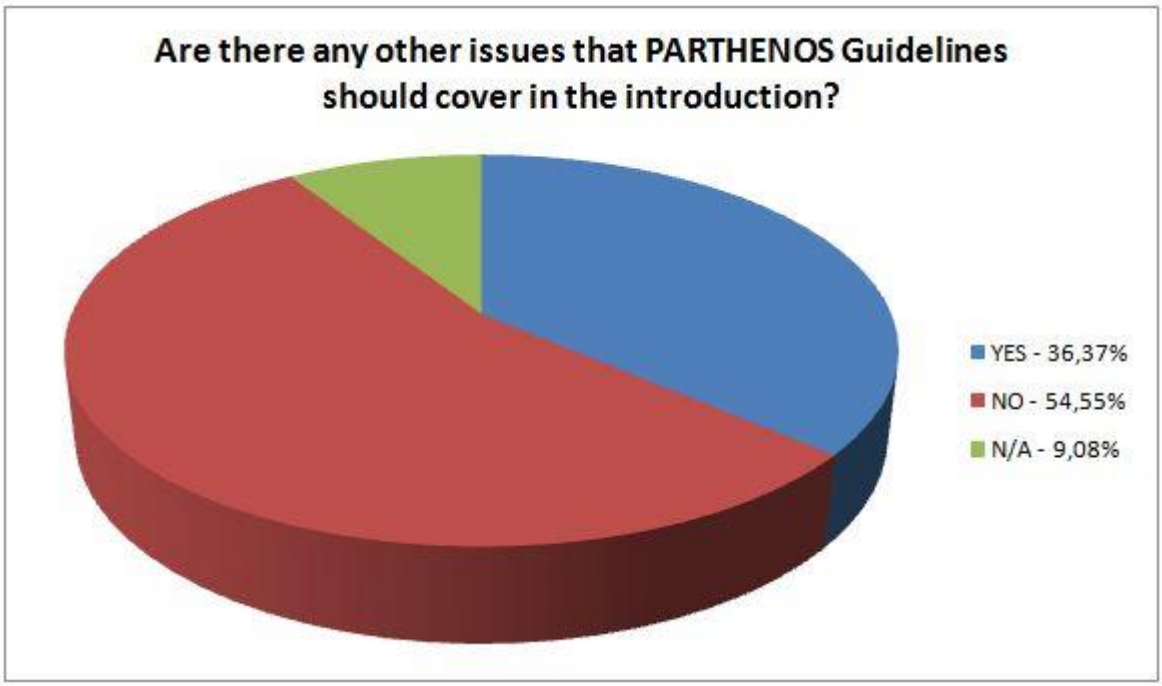

Fig. 3 – a large part of the experts agreed with the topics outlined in the introduction and some of them provided suggestions how to improve the section

2.2 Quality assessment of data, metadata, and digital repositories

The chapter on Quality assessment of data, metadata, and digital repositories in D3.1 was reviewed by eleven experts. Four of them are working for Research Centres or Data Archives, four for Universities, and two at Cultural Heritage Institutions and one for a Research Infrastructure (see Fig. 4). Most of them are researchers or scientists and as professors, lecturers, data managers, data analysts and data specialists they are well aware and prepared for the assessment of data, metadata and digital repositories. Together they cover almost all the PARTHENOS domains (see Fig. 5).

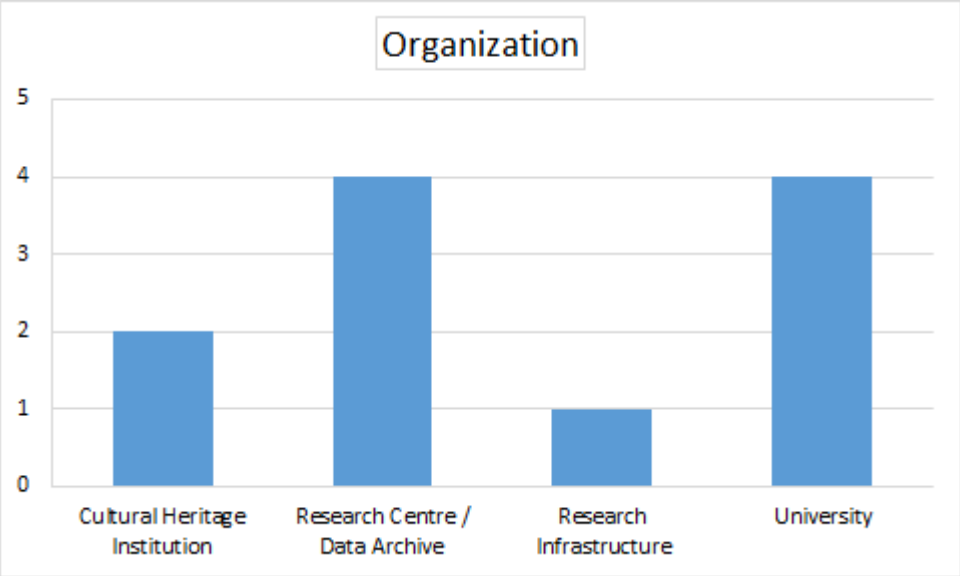

Fig. 4 – Experts work for Cultural Heritage Institutions, Research Centres, Data Archives, Research Infrastructure organizations and Universities

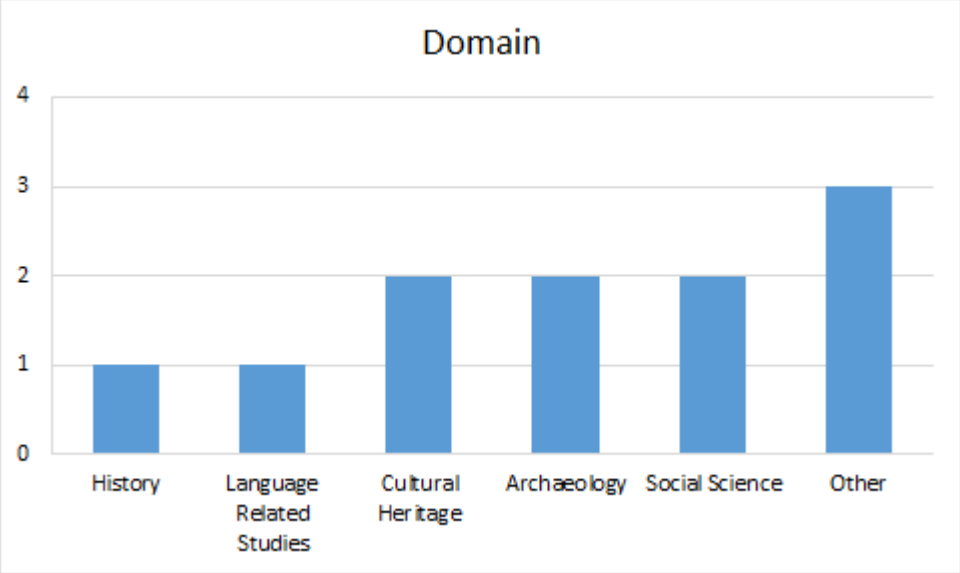

Fig. 5 – domains of experts cover almost all PARTHENOS communities, others include Digital Humanities, Digital Library and Archival Science

The themes in this chapter were discussed in a more controversial way than in all the other chapters and a lot of remarks and recommendation were made by the experts.

The chapter covers the topics of access and authentication policies and the quality assessment of digital repositories. It starts by first defining data and metadata in the Humanities and Social Sciences including the assessment of research data, followed by

the assessment of repositories looking at certification and assessment tools and models. The chapter then continues with policies for quality of data and repositories and the PARTHENOS Guidelines. Finally, the PARTHENOS Wizard and its relevance as well as other issues on Quality are addressed.

2.2.1 Defining data and metadata in the Humanities and Social Sciences

This section was assessed in two parts. Firstly, the definition of data, metadata, datification and research data and, secondly, the assessment of research data.

2.2.1.1 Definition of: Data / Metadata / Datification / Research data

Most of the experts considered the definitions as helpful, some had objections which led to many remarks and suggestions when asked for corrections or additions.
The experts rated the definitions very differently, one saw them as “adequate and balanced”, another expert thought that “the description is a mix of relevant and trivial information” and that “the definitions mixes objective and subjective arguments”. Another expert suggested to have a less theoretical approach and to add more examples of research data, comparing it with other kinds of data.

For the definition of data, one expert suggested to add the third meaning: “Fact, ideas or discrete pieces of information, especially when in the form originally collected and unanalysed. Data that is often used.” Another expert pointed out that data has different dimensions.

Several experts didn’t like the paragraph on “Datification” and remarked, for example, that it “made the distinction between data and metadata … less meaningful”.
When asked to add examples on data, metadata and research data that could be useful there were many different answers. The following are examples and list a few of them:
Indications about metadata standards:

a) the standards developed by the Moving Picture Experts Group (MPEG), a working group of ISO/IEC;
b) the Adobe Extensible Metadata Platform (XMP);
c) ISO 19115-1 and 19115-2 about geographic metadata;

d) Metadata Authority Description Schema (MADS), related to MARC authority elements;

e) Functional Requirements for Bibliographic Records by IFLA.

f) EAC that provides an XML-based metadata schema for encoding description of the creators of archival records;

g) EAD that provides an XML-based metadata schema for encoding archival finding aids.

Suggestions for additional metadata:

- Data: dimensions of artefacts, date of artefacts
- Metadata descriptive: description of a collection
- Structural: number of pages,
- Technical: format of picture,
- Rights: creative commons

Suggestion with links:

- Linked Data for Production (https://wiki.duraspace.org/pages/viewpage.action?pageId=74515029)
- reassembling the Republic of letters COST IS1310 (http://www.cost.eu/COST_Actions/isch/IS1310)
- Jisc Research Data Spring (https://www.jisc.ac.uk/rd/projects/research-data-spring)

2.2.1.2 Assessment of research data

After being informed that for the assessment of research data, PARTHENOS uses the FAIR principles as a general framework for verifying the quality and the correctness of data, the experts were asked for comments. One expert had doubts regarding this approach by pointing out that: “The correctness of data as such is not verified by the FAIR principles, I think.” Other remarks were very specific, e.g. “With regard to formats it may be helpful to explain why e.g. proprietary formats are not suitable for long-term preservation even when these formats are widely used by the designated community.” Another expert missed as most important aspect of re-use of data the absence of the knowledge about the context of the data creation.

Asked for examples about data quality that could be useful for the guidelines, the experts suggested Peter Kiraly's work (e.g. https://pkiraly.github.io/metadata-quality-project-

plan.pdf) and the International Image Interoperability Framework as it contributes to the quality of digital images (http://iiif.io).

2.2.2 Assessment of Repositories

Most of the experts rated the explanations on certification of repositories and assessment tools and models for assessing the repositories quality very positively. One suggested looking at similar comparisons like the Preservation And Archiving SIG (http://www.preservationandarchivingsig.org/). Another expert pointed out that "a schematic overview of the different methodologies and their advantage and disadvantage - and some hints of recommendations" would have been even more useful and added that he, as a potential user of a data repository, doesn't know what to look for. One expert gave very detailed comments stressing the importance of certification of repositories and recommended strongly to review the parts concerning certification before publishing it to be up-to-date as many things were happening right now.

When asked for further comments and additions, one expert suggested adding a definition of the term "digital repository" related to its use and meaning in the PARTHENOS Guidelines as the term might have different meanings in different domains and advised that the Extended and Formal Certifications should have as a Reference the DIN Standard 31644:2012 (Criteria for Trustworthy Digital Archives). Another suggested specifying if the assessment services are free or not.

2.2.2.1 Best practices

The experts suggested adding the following examples and best practices on repositories and assessment tools and models for assessment of the repositories quality that could be useful for the guidelines:

- http://www.dcc.ac.uk/resources/repository-audit-and-assessment/trustworthy-repositories
- assessment tools: DRAMBORA toolkit
- Unsworth, J. 2000. Scholarly primitives: what methods do humanities researchers have in common, and how might our tools reflect this?
    - Discovering
    - Annotating
    - Comparing

  - Referring
  - Sampling
  - Illustrating
  - Representing
- ISO Standard for audit and certification of trustworthy digital archives -> ISO 16363

2.2.2.2 Examples of other disciplines

PARTHENOS conducted an analysis of existing policies among the research communities involved in the project. Therefore, the experts were asked to look at the list and to add any missing policies for their discipline.

One expert pointed out that “all the policies listed are collection oriented and share an archival approach”. Another remarked that the list did not just include “policies but an assortment of standards” and if OAI-PMH is listed then OAI-ORE and OAI-RS should be, too. For preservation and provenance he suggested W3C PROV-O and PREMIS.

Further suggestions from the experts were:

ARIADNE Catalogue Data Model (http://ariadne-support.dcu.gr/)

- ARIADNE Reference Model (http://www.ariadne-infrastructure.eu/Resources/Ariadne-Reference-Model)
- Social Sciences: GESIS Digital Preservation Policy: http://www.gesis.org/fileadmin/upload/institut/wiss_arbeitsbereiche/datenarchiv_analyse/Digital_Preservation_Policy_1.4.8.pdf
- For History (and maybe others as well) the "preservation imaging" guidelines were seen as possibly being relevant in relation to the digitization of photographs, documents, etc. (https://www.metamorfoze.nl/sites/metamorfoze.nl/files/publicatie_documenten/Metamorfoze_Preservation_Imaging_Guidelines_1.0.pdf)

2.2.3 Policies for the quality of data and repositories and PARTHENOS Guidelines

Most of the statements agree with the overview on the state of the art of quality policies in the different disciplines that are covered by PARTHENOS (chapter 2.3.3.2). There is one concern that the overview may simplify practices in some research communities.

The analysis of strengths and weakness for each stakeholder and discipline (chapter 2.3.4) is also well received. There are only slightly different ratings regarding the table 2.3. One statement is that IPR and licences are weak points for every stakeholder and discipline.

There is one recommendation in this part that could be considered for the next deliverable: A general best practice to use W3C, IETF or ISO standards in favour of specialized discipline approaches wherever feasible.

2.2.4 From Commonalities to Recommendations

No specific comments

2.2.5 Guidelines and Best Practices to increase the Quality of Data, Metadata and Repositories

These recommendations were divided on the basis of the FAIR principles. The survey asked for a response to question in this section.

For the part “Findable” there was, in general, very positive feedback. There was one remark on missing policies on multilingualism especially for vocabularies. Another comment asked for explicitly referencing published material that is based on the research data found. There were also remarks on describing the used thesauri and ontologies. One expert mentioned that findability is related to “the functionality of search and retrieval systems”, which should be included in the recommendations. Additionally, there were some hints on standards that should be integrated. This includes the vocabulary of http://schema.org/ and the Memento framework that targets time specific versions of websites (https://tools.ietf.org/html/rfc7089).

The part “Accessible” had also positive feedback. One additional recommendation was, that, sometimes, a simple download of data would be useful. It should be also stated that, especially, accessibility is an “ongoing curation process”. Also, there was a hint about the connection between accessibility and reusability when it comes to legal issues and licences.

The interoperability part was seen as helpful by every expert. Further suggestions included explicitly mentioning linked open data (LOD), especially when it comes to expressing taxonomies and thesauri. It was seen as also necessary to capture and express context and provenance, where provenance “should cover both the physical and digital histories of data”. What was missing was a reference to “the trust chain for data”, where non-repudiation plays an important rule, because “not only can we assert that data came from a particular source but the source itself cannot deny its originating role” as this is important for machine-to-machine transactions. Additionally, it was suggested to include some information on how the ontologies “exist in a network of topics”. For the vocabularies, the multilingual aspect was also considered of great importance.

The last part on reusability received very positive responses from the experts. One expert advised to add some examples, e.g. some from chapter 3. There was also a valuable remark pointing out that opening data for reuse doesn’t mean losing control of the data. It could also be mentioned that there are economic benefits when allowing data to be reused.

2.2.6 The Parthenos Wizard

There was, in general, a very good response to the idea of an interactive guide in the form of a wizard. For a majority of the respondents such a PARTHENOS wizard was seen as helpful for their organization. Therefore, it was highly recommended to implement such a tool.

Some answers emphasized that the wizard needs to be user friendly. This includes not only to focussing on usability but also to allow user feedback. Both can help to improve the quality of the wizard.

One answer pointed to https://biosharing.org, which is a comparable approach for the life, environmental and biomedical sciences that is already in use. Not only can this be a stimulating example but it would also be a good idea to contact the developers with a view to sharing experiences and exploring opportunities for cooperation. Other research communities that are not in the scope of PARTHENOS may also have such tools and this should be investigated. Connecting these tools could be very helpful for cross-discipline approaches and visibility among diverse research communities.

2.2.7 Relevance and other Issues on Quality

Almost all experts considered the themes outlined in the chapter on quality assessment of data, metadata, and digital repositories relevant for assessing the impacts of the data workflow managed by their institution. One expert noted as missing the aspect of collaboration as “a key concept of RI development between targets developers, cultural heritage practitioners, researchers and managers”. Another agreed on the relevance but pointed out that his interests were much wider.

When asked for other issues on quality assessment of data, metadata, and digital repositories that the PARTHENOS Guidelines should examine, the experts had several valuable suggestions:

- analysis of the specific data producers, data consumers and data repository roles and responsibilities
- some general indications about necessary competencies related the PARTHENOS Guidelines context (both curricula academic higher education and continuing professional training)
- to recall the EU recommendation and directives on preservation, reuse, etc.
- as standards are evolving, current practices are a long way from perfect so this would need to be an ongoing, tracking exercise;
- to be able to travel across the borders of particular collections, institutions, languages, nations, in order to exchange ideas;
- services provided by research infrastructures (in relation to the PARTHENOS target group) such as EUDAT, EGI, CESSDA, DARIAH, CLARIN can be evaluated and assessed.

2.3 Data policy implementation

The chapter on Data Policy Implementation of D3.1 was assessed by nine experts, working with different responsibilities (data managers, lecturers, professors, librarians, etc.) in Research Centres (4), Universities (2), Cultural Heritage Institutions (2) and Research Infrastructures (1 - see Fig. 6). The domains of the experts range from the PARTHENOS communities as Cultural Heritage and Archaeology to others such as Library and Information Science (see Fig. 7).

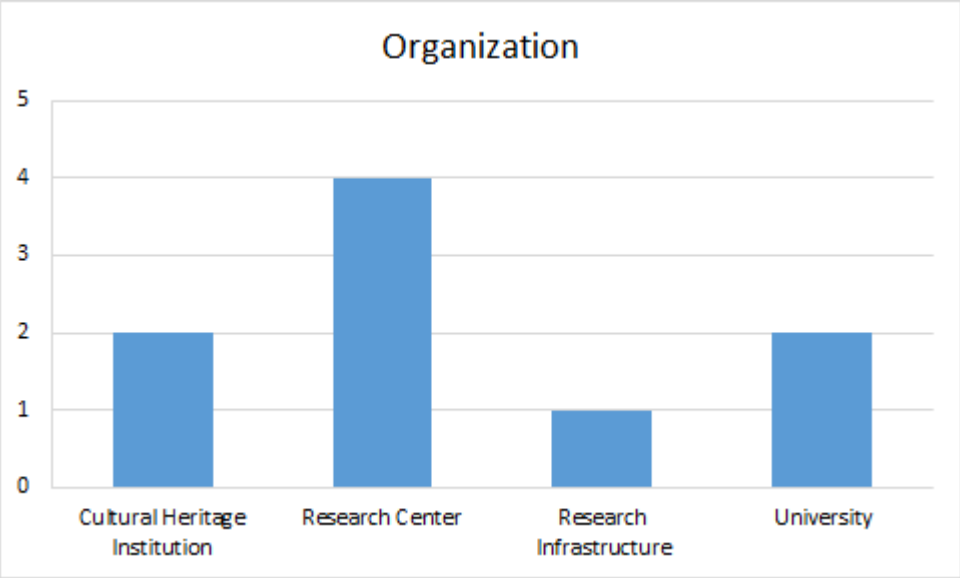

Fig. 6 – Experts work for Cultural Heritage Institutions, Research Centres, Research Infrastructures and Universities

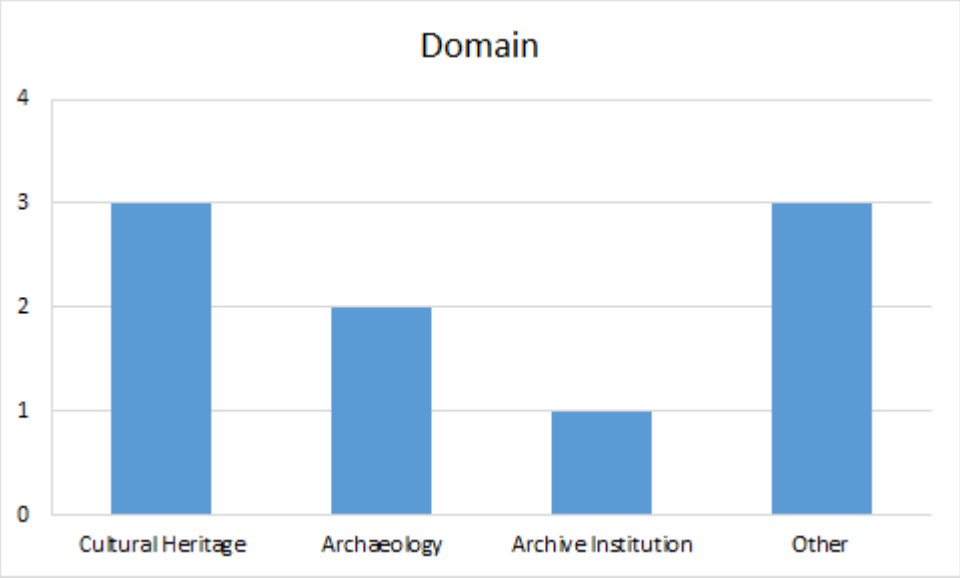

Fig. 7 – Domains of experts range from Cultural Heritage, Archaeology and Archive Institutions and include among "other" Library and Information Science, Publication Support and Computing applied to High Energy Physics

This chapter of D3.1 covers the policy implementation regarding the data lifecycle, data curation and long-term preservation. This is done by describing the method for assessing the current data management situation first. A detailed presentation of different aspects of good practice guidelines, based on the FAIR principles, will follow and - eventually -

supporting practices to FAIR data will be presented, with a specific focus on data management planning and long-term digital preservation.

2.3.1 Approach to assess the current situation

Most of the experts thought that the approach of using a questionnaire structured along the steps in the research data life cycle and following the FAIR principles was a good way to assess the current situation. However, some of them saw some weaknesses, e.g. the basis for the deduction of the Guidelines wasn't always clear and the use of FAIR principles to structure the presentation of the Guidelines led to redundancies and fragmented topic areas which could be presented more coherently (particularly the metadata topic).

2.3.2 Guidelines defining good practices

The Guidelines for good practices, provided by D3.1, are structured around the FAIR principles and aim to help research infrastructures and repositories to supply their services more efficiently. The experts were asked to review these Guidelines, focusing on each of the FAIR principles separately.

2.3.2.1 Findable

The acceptance of the Guidelines based on the *Findability* principle ranged from positive to very positive. The experts mainly commented about Persistent Identifiers, object identification and versioning (recommended addition: http://www.niso.org/publications/rp/RP-8-2008.pdf), suggesting that these concepts should be explained in more detail. They also proposed the addition of a Guidelines list "which would go some way to getting more interoperability as Geoname (location) or ISO 639 (languages) or Sachs Hornbostel (instruments)", along with a suggestion for considering the metadata schema of Google and its dataset mark-up.

2.3.2.2 Accessible

The assessment of the set of Guidelines related to the *Accessibility* principle was mainly very positive. Commentary in this section regarded requests for a better explanation of the restrictions in data-sharing practices and further references to OAI-ORE. The "Long-term accessibility of metadata" subchapter seemed to trouble the experts a bit more; one of

them pointed out that he found it difficult to navigate among the many existing standards, thus “it would be very helpful to have the standards listed by (one or more) preferences”.

2.3.2.3 Interoperable

Most of the experts seemed to strongly agree that the Guidelines related well to the *Interoperability* principle, sufficiently covering the most important aspects and considering the most relevant needs of their institutions. Again, some experts suggested adding a list of Guidelines which ensure a higher interoperability level as well as a list of “open source deep learning frameworks for data scientists”. Also, some elaboration on the subchapter “Metadata formats utilize shared vocabularies and/or ontologies” was recommended, putting emphasis on the importance of harmonization efforts in data curation and the risks deriving from not doing so.

2.3.2.4 Reusable

As for the good practices based on the *Reusability* principle, the majority of the experts ranked them, once again, positively. The main issue which was posted here is that “it is not clear how to cope with data generated by tools and written in proprietary (or custom) format for which you need some software to read it and documentation to use the software”.

2.3.3 Supporting practices to FAIR data

In D3.1, following the description of the different aspects of FAIR principles (p. 64), there is a section about data management planning and long-term digital preservation as important supporting practices in providing FAIR Data and which are, therefore, subject to specific recommendations.

2.3.3.1 PARTHENOS Data Management Plan template

The experts were, first of all, asked their opinion about the draft Data Management Plan (DMP) template. In general, the experts were positive but pointed out three issues that could improve the plan:

- It is stated that the DMP should provide guidance for researchers. However, part of the questions are aimed at institutions and researchers will not be able to answer these questions. In general, many researchers will find it hard to understand the questions, because they have not dealt with data management or FAIR data before.

Additional guidance, explanations and examples would be helpful to understand what is needed from them regarding data management.

- The recommended options for some of the questions are fairly limited, mixed up or not applicable at all for researchers (e.g.: Specify what methods or software tools are needed to access the data?). Some thought should be given to improve the lists with options.
- The topics of sensitive data and privacy issues are underrepresented in the DMP template.

As the PARTHENOS DMP template was initially drafted to meet general needs, the experts were asked if they saw additional needs for their specific scientific disciplines. Once again, the aspect of ethical/privacy issues was mentioned by several experts as an element in the DMP that didn't receive enough attention.

2.3.3.2 Long-term digital preservation

The second important supporting practice in providing FAIR data in the Guidelines is long-term digital preservation. The experts were asked if the chapter covered the most important aspects of long-term digital preservation. One expert found the chapter a bit superficial and therefore hard to understand for persons unfamiliar with the subject. Some general information (why is long-term preservation important? what kind of dangers/risks are linked to this issue?) were missing and the basic terms/concepts were seen as not sufficiently described. Other experts suggested incorporating best practices in the document to give the user more guidance.

2.3.4 General

All experts considered the themes outlined in the chapter on Data Management and Policy Implementation as relevant for assessing the impacts of the data workflow managed by their institution. One expert recommended additionally to look at practices on altmetrics, pointing to the outputs of the NISO Alternative Assessment Project (NISO RP-25-2016).

2.4 IPR, open data and open access

The chapter on IPR, Open Data and Open Access was reviewed by seven experts. It's interesting to notice that these experts cover the different PARTHENOS research domains and they have specific skills on legal issues on digital contents.

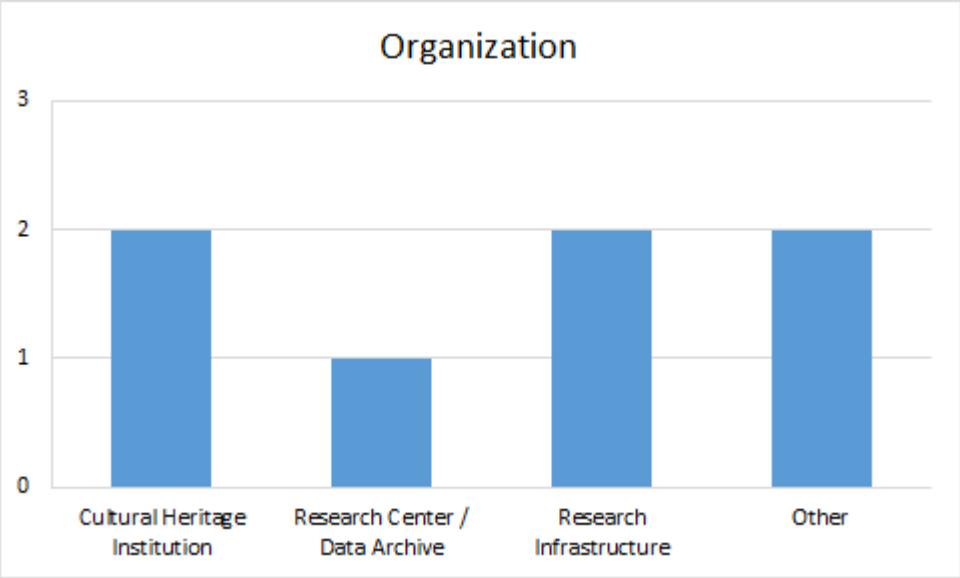

Fig. 8 – Experts work for Cultural Heritage Institutions, Research Centres, Data Archives and Research Infrastructures, for 'other' an expert is belonging to an Information Service, the other one to a Research Centre and Cultural Heritage Institute

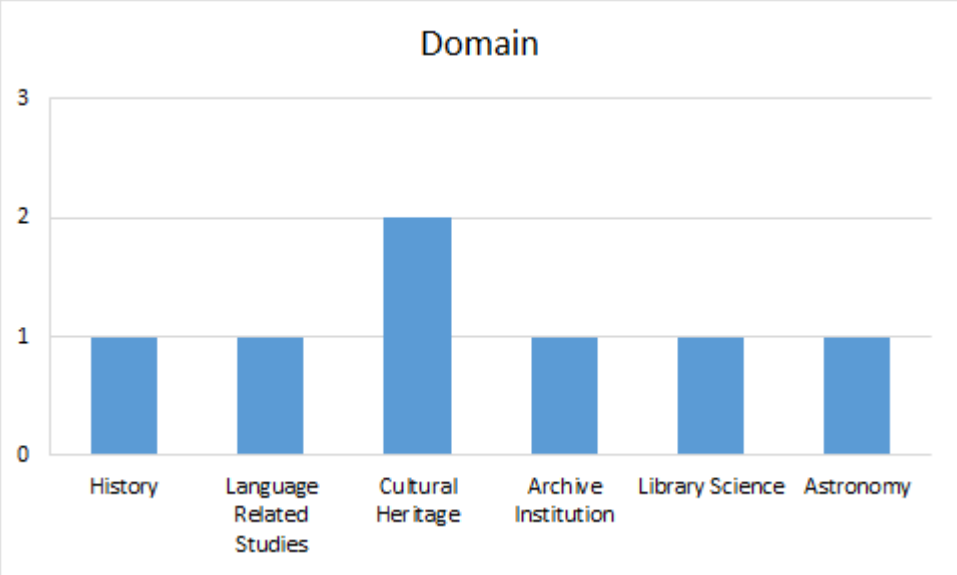

Fig. 9 – The seven experts come from different domains, and they cover the most important research area identified within PARTHENOS communities

They all agreed on the seven recommendations presented in the Guidelines and in general the themes and information outlined in this chapter were judged relevant for assessing the impacts of data workflow managed by their institution/research.

In regard to other issues that PARTHENOS should examine in this chapter, one expert indicated the economic benefits of Open Access. Another expert suggested making a clear distinction between the legal topic of copyright and licences and the scientific topic of attribution. Finally, no experts pointed out best practices.

Below are presented amendments and suggestions referred to the specific paragraphs of the chapter, in order to help the evaluation and eventual integration of these into the Guidelines.

2.4.1 Introduction and methodology (4.1, pp. 129-130)

One expert suggested considering also cultural knowledge and unrecognised knowledge rights as limitations for re-using data (p. 129, "Limitations for re-using ...").

2.4.2 How we collected the information (4.2, pp. 130-134)

Regarding the way the information was collected one expert pointed out that it should be made clear what “legitimate interests” means, i.e. who is entitled to decide when they are legitimate (p. 134, "Allow restricted access to the data for protection of legitimate interests of the rights holders ..."). It was seen also as crucial to more clearly define the concepts of "open" not as a simple and binary opposition with "closed", or, for instance of "free" vs "commercial".

2.4.3 Legal framework (4.3., pp. 134-136)

The presentation of the legal framework was in general found helpful by the experts. Only one expert saw it differently. This part could be improved by explaining in more detail the term “Open Science”. One expert suggested reviewing the style and logic of the second paragraph (p. 134 "Data infrastructures ...") and to use more balanced expressions in some cases in order to avoid the impression of ideological statements (p. 135, "Therefore, the legal restrictions sometimes unnecessarily imposed …”).

2.4.3.1 Intellectual property rights (4.3.1., pp. 136-139)

Similarly, the presentation of the intellectual property rights (IPR) was judged quite helpful although one expert thought that “a more schematic approach with charts” would have been better. Another expert pointed out that there is a distinction – at least in Germany – between IPR for design layouts and trademarks known as *Markenrecht* and those for

patents known as *Patentrecht*. He also considered that “Database protection rights” as a subcategory of “Copyright” and added that a part about the eligibility of data for copyright protection, i.e. threshold of originality, is missing. One expert suggested adding in the section on IPR also the limitation to access in relation to Traditional Knowledge (TK), ICIP (Indigenous Cultural and Intellectual Property), intangible heritage, communal/collective rights, and more.

2.4.3.2 Sensitive data (4.3.2., pp. 140-141)

Despite being rather brief compared to the other sections, the presentation of sensitive data appeared quite helpful to experts. An expert who evaluated this part less favourably suggested adding a paragraph to explain informal consent, anonymization and the problem of de-anonymization. Another expert recommended considering (p. 141, top of page) global and ethical concerns as a core part of rights issues, because sensitivity issues as in health data, surveillance or social media-acquired data are as important as the legal framework of rights.

2.4.3.3 PSI Directive (4.3.3., pp. 143-145)

The presentation of the PSI Directive was evaluated positively. However, the actual implementation of the PSI Directive is not presented very clearly. One expert proposed adding some good practices and provided a link:
(http://www.bl.uk/aboutus/stratpolprog/pubsect-info-regulations/). It was also suggested to specify that the PSI Directive was created within the Intellectual Property context.
It was also suggested adding that sometimes even components of metadata must have controlled access (p. 145, "Protected data and personal data must be available through a controlled procedure ...").

2.4.3.4 Open Access and Open Data (4.3.4, pp. 145-152)

The presentation of the Open Access and Open Data was evaluated positively with only one exception. An expert suggested not to treat IPR as in conflict with Open Science, i.e. to expound the relation between IPR and Open Science. Getting the right mechanisms and relationships is essential to trusted sharing and sustainability. It was suggested also to add “platinum” open access to the other models of OA (p. 146, "Open access publishing."). An expert asked for a change on page 146, to line 9 from "In other cases" to: “A lot of open access journal don't have APC's [, the cost of open access etc.]”: that would make it more

clear that some OA journals do not charge the authors/institutions at all. It was also requested to explain the concept "Digital Single Market".

2.4.3.5 Licensing Framework, Rights Statements, Creative Commons and licensing framework in PARTHENOS Community (4.3.5, 4.3.6, 4.3.7., 4.3.8., pp.152-157)

The presentations of Licensing Framework, Rights Statements, Creative Commons and licensing framework in PARTHENOS community were assessed rather positively with only one exception. The experts strongly advised providing a list of the different statements and licences, to give an introduction into the CLARIN licences like the Rights Statements and Creative Commons and to add a part about the Public Domain Mark. It was also asked to provide two different levels of licences for data and databases. So, it would be more appropriate talking about ODC Open Data Commons in all their different variations as Public Domain Dedication and License (PDDL) — “Public Domain for data/databases”; Attribution License (ODC-By) — “Attribution for data/databases”; Open Database License (ODC-ODbL) — “Attribution Share-Alike for data/databases”. In addition, it would be important to define the differences between CC0 and ODC Public Domain Dedication and License (PDDL) — “Public Domain for data/databases”. Furthermore, the Attribution Licence (ODC-By) - "Attribution for data/databases" adds to data an attribution which is very useful in the world of researchers.

It was also recommended to discuss the use of restrictive licences as they could be considered “harmful” by parts of the community. Another expert suggested correcting the gender specific language (p. 156, "a separate application allows HIM to send ...") and to enrich the discussion about accessibility of the rights holder, operationalising of decision making and scalability. It was asked that the abbreviations LRT, NORED, DEP (p. 157) be explained.

2.4.4 Authentication and authorization infrastructure (4.4. pp. 157-160)

While most experts found the presentation of authentication and authorization infrastructure helpful, one expert considered it as “a very technical topic” which should be moved to an appendix. Another expert suggested considering the closeness of an affiliation of institutions as possible limitation to open access.

2.4.5 Outcome (4.5 pp 161-173)

2.4.5.1 (Meta)data should be open as possible and Closed as necessary and Related guidelines (pp. 161-163)

All experts agreed that (meta)data should be as open as possible and as closed as necessary. They found the related Guidelines helpful. One expert considered that "publicly funded' research may, in some cases, be loaded with access limitations (research on health, welfare). On the other hand, commercial research that claims to be scientific may be pressed to share their data. Science itself drives to openness. He also suggested considering possible incoherencies in this paragraph (p. 162, “1d Standards ...”).

2.4.5.2 Protected data and personal data must be available through a controlled and documented procedure and Related guidelines (pp. 163-164)

All experts agreed that protected and personal data must be available through a controlled and documented procedure. Most of them found the related Guidelines helpful. One expert had a different opinion, highlighting that “the guidelines are not building up on what was written in the sections before” and that it should be made clearer to whom the data must be available.

One expert asked to specify who is entitled to ensure "legitimate interests" (p. 163, "must be subordinate to the legitimate interests of rights holder …"). It was recommended, also, to update the concept of informed consent to a more modern concept "Free, prior and informed consent" (p. 163, "2B. Obtain informed consent …") and to include the possibility that those seeking access can be the information providers or those who are otherwise stakeholders in the context of the data (e.g. persons recorded etc.) (p. 164 "2D Ensuring …").

2.4.5.3 (Meta)data licences framework should support legal interoperability fostering harmonization of rights (pp. 164-169)

All experts agreed that the (meta)data licences framework should support legal interoperability fostering harmonization of rights. One expert indicated the potential need for a “legal European framework more open in such direction”. Five experts gave a quite positive assessment of the relative guidelines while one evaluated less positively and

another even stated that the section was not ready for publication due to the lack of formatting and of captions for some figures. He also alluded to redundancies regarding the CLARIN license categories and the paragraph 3B.

2.4.5.4 (Meta)data should be licensed to permit the widest reuse possible (pp. 169-170)

All experts agreed that (meta)data should be licensed to permit the widest reuse possible. One expert suggested adding a case study about the OCLC licence. The related guidelines were assessed quite positively except by one expert because of too many typos and a too colloquial style.

2.4.5.5 (Meta)data rights holder should be identified before data publishing (pp. 170-171)

All experts agreed that (meta)data rights holders should be identified before the data is published and they found the related Guidelines helpful. Notwithstanding, one expert noticed that this part “seems to be all about copyright, which […] is just a part of IPR”.

2.4.5.6 (Meta)data rights statements should communicate the copyright and reuse status transparently and clearly (pp. 171-172)

All experts agreed that (meta)data rights statements should communicate the copyright and reuse status transparently, clearly and in machine-readable form and found the related Guidelines helpful.

2.4.5.7 Specify why and for what period a data embargo is needed (data should be made available as soon as possible) (pp. 172-173)

All experts agreed that it is necessary to specify why and for what period a data embargo is needed and found the related Guidelines helpful. The themes outlined in this chapter were judged relevant by the different experts for assessing the impacts of data workflow managed by their institution/research.

3. Part 2: Assessment of standardization

3.1 Introduction to the approach adopted for Part 2

This part of the deliverable concerns the assessment of the technical standardization solutions produced in Work Package 4 (WP4) comparing them with the communities' needs, and proposing amendments if necessary. It is a combined effort of NIOD-KNAW INRIA, CLARIN, TCD, CNR, CNRS, CSTC, FORTH, OEAW, MIBACT-ICCU, FHP, SISMEL, and AA.

The technical standardization solution produced by WP4 comprises the Standardization Survival Kit (SSK) which is conceived as a comprehensive online environment aiming at providing basic information, documentation, and resources concerning standards applicable to a wide scope of digitally based humanities and cultural heritage research activities. The SSK is a digital platform/guide that refers, with the help of research scenarios, to resources and standards on dedicated websites.

Although WP4 has delivered important components until now for the SSK (architecture, standards, and resources), the technical standardization solutions produced by WP4 were at the time of this deliverable too basic for comparison with the community's needs and for making a meaningful assessment. Consequently, WP2 and WP4 decided to split the assessment into two parts. The first part of the assessment is about the identification of missing standards and resources (gap analysis) and the enrichment of the already identified standards and resources. The second and most important part will be the actual comparison of the SSK with the community's needs. This part of the assessment is postponed till the beginning of 2018 when the initial implementation (design and resource pages) of the SSK will be ready. WP4 will use this feedback to improve the SSK even further.

3.1.1 Structure of this chapter

The introduction to the chapter about the assessment of standardization summarizes the link between the work conducted in WP2 and WP4 and the development of the SSK so far. Furthermore, the methodological process is explained and finally, the most important

conclusions based on the gap analysis are summarized and recommendations are given for further improving the SSK. The second chapter of Part 2 is the actual assessment. In the last chapter a specific component of the SSK is discussed, the 'Why Standards' leaflet. The chapter about the assessment is divided into four subchapters each dealing with one of the research fields (studies of the past; language-related studies; heritage, applied disciplines, and archaeology; and social sciences). The subchapters give an overview of the standards used in the distinctive research fields, with suggestions for standards and resources, and recommendations for the prioritization of the work. The annex contains the systematic result of the gap analysis.

3.1.2 Context

For the deliverable D4.1, WP4 produced a first outline of the SSK design. This initial concept of the SSK consisted of a sketch of the website and its components. Furthermore, the content and the process of the creation of the architecture and taxonomy of the SSK and the next steps were described.

For the identification of standards the use cases described in D2.1 were used. These use cases outlined the requirements for standardization expressed by the research communities involved in PARTHENOS. The use cases were revised and elaborated by WP4 with information about the data creation process. Also, extra use cases were added. Next, the use cases were divided into different steps. Each step represents a different task in the research process. After that, standards used in each stage of the research were identified. The next step was to find commonalities between use cases that dealt with the same issues and to create more general research scenarios that can be useful for different research communities. In the coming months, the standards and resources will be connected to the steps of the research scenarios.

The second deliverable (D4.2) of WP4 contained a systematic overview of the identified standards. The standards were elaborated by adding information about the scope of the standard: a technical overview (examples, technical contents); resources (data repository, Github, bibliography (Zotero), blog entries, etc.); and ongoing efforts. After D4.2 the enrichment of the knowledge base will be a continuous process. This is the basis for the content of the SSK and will provide support to researchers in using and contributing to these standards.

Besides these activities, WP4 also developed a leaflet about standards. This leaflet is meant to inform researchers that are not familiar with using standards about the importance of structuring data in certain formats.

3.1.3 Methodology adopted for Assessing the Standards

WP4 is in the phase of collecting as many standards, resources, and information for the SSK as possible. They requested if T2.2 can add extra standards and resources that are still missing in D4.2 and elaborate the standards with extra information. Besides this, WP4 asked if the members of T2.2 could prioritize the standards and resources for implementation in the SSK.

The use cases described by T2.2 regarding the standardization requirements in D2.1 and the research scenarios created by WP4 are structured around the four research communities defined by PARTHENOS. This structure was also used for this assessment. Although the structure on the SSK will not be centred around the research communities but around general research scenarios useful for researchers belonging to different communities, the process of identifying missing standards and resources was easier when using the research communities as structure.

A spreadsheet was created with four tables corresponding to the four research domains. In the spreadsheet, there were 14 columns corresponding to the 11 descriptive elements that will be used to describe standards within the SSK and three columns that were added to the purpose of this assessment. The content of the standards described in D4.2 was added to the spreadsheet. The members of T2.2 were asked if they would:

- indicate missing standards
- identify experts on the subject
- assess the content of each table by adding comments and remarks
- indicate missing resources (bibliographical references, tools, services, and samples)
- prioritize the resources for incorporating in the SSK.

The 14 description elements of the standards were:

- Name
- Standards involved
- Research field(s) (the disciplinary scope of use of the standard)
- TaDiRAH Activity(ies)
- Overview
- Working material
- Bibliographical references
- Tools, services, and samples
- Creator / Developer of the standard
- Last update
- Licence.

The elements added for the purpose of the assessment were:

- Prioritizing of the resources
- Experts on the standard
- Remarks and comments.

3.1.4 General results of the assessment and recommendations

This assessment has shown that WP4 has made a comprehensive overview of the most important standards and resources used in different academic fields. In this deliverable, suggestions are made to include other valuable standards and resources. Some standards already described in the deliverables of WP4 are mentioned in this deliverable also as to highlight the importance of these standards. One of those is CIDOC-CRM because it is the model on which a large part of the data cloud is shaped. In other research areas, commonly used standards don’t exist yet but are under development. In the PARTHENOS community, for example, there are activities for creating a comprehensive environment centred around the researchers' practices on and with 3D digital objects in arts and humanities.

Different academic communities sometimes identified the same standards. This is not problematic but emphasizes that more and more standards are not bound to one community. The Europeana Data Model is both mentioned in the field of studies of the

past as in cultural heritage. Also, in social sciences, some ‘general’ standards such as TEI, Dublin Core, and METS are in use and integrated into the research activities.

This gap analysis has been especially helpful for identifying standards and resources for the social sciences because there has been less attention on the social sciences than for the other academic fields until now. Further research is likely to be necessary for a more comprehensive overview of the standards used in the social sciences. Another suggestion is to include at least two research scenarios in the SSK that cover a quantitative and a qualitative research approach. This will probably help the SSK to better connect to the social sciences researchers.

Other considerations are the legal, such as Intellectual Property Rights, and privacy issues that form a barrier to data use without restrictions. An added value for users of the SSK could be a scenario on procedures and standards which helps them with these aspects during their research.

A final recommendation would be to not only include open standards, although these are to be preferred, but also to include commercial standards if they are well-used standards in the communities. A couple of well-known and often used commercial standards are mentioned in this deliverable. However, this overview is not exhaustive and other commercial standards could be added

3.2 Gap analysis of standards and resources

3.2.1 Studies of the past

3.2.1.1 Overview

The community of the studies of the past needs to be able to find the digital sources created by historians in different phases of their research practice, in order to use various distributed datasets and tools as an integral component of their research methodology. The current situation of the historical digital data is characterized by a high degree of fragmentation that prevents resource discovery and access. A large part of the available historical datasets form a vast and fragmented corpus: their potential is thus constrained by difficult access and lack of interoperability. Historical digital resources currently

available, as well as authority files for persons, objects, events and places, are often characterized by the use of a plethora of different standards (i.e.: Text Encoding Initiative (TEI), Encoded Archival Description (EAD), Metadata Object Description Schema (MODS), Metadata Encoding & Transmission Standard (METS), Encoded Archival Guide (EAG), Europeana Data Model (EDM), Data Model 2 Europeana (DM2E), Portable Document Format (PDF)), preventing their full discoverability. A key issue would be to bridge the gap between tangible and intangible aspects of the Cultural Heritage Objects (CHO), allowing researchers to follow truly innovative research paths. A suggestion is to add a scenario in the SSK where interoperability of standards is addressed and resources like mapping tools or converts are offered.

3.2.1.2 Standards and resources

EDM has been developed together with technical experts from the library, museum, archive and audio-visual collection domains, and has been designed to accommodate standards such as Dublin Core (DC), EAD, and Lightweight Information Describing Objects (LIDO) with the help of experts in these fields. Developed within the Europeana v1.0 project, EDM is a Resource Description Framework (RDF)-based data model for describing rich metadata records for Europeana, the European Digital Library. It can handle huge metadata record collections represented by heterogeneous metadata standards. EDM covers CHOs that are collected and delivered to Europeana by diverse cultural heritage institutions.

The DM2E model is a specialization of the Europeana Data Model (EDM) for the manuscript domain.

The CENDARI Collection Schema (CCS) was developed to encode detailed descriptions for collections housed by the associated cultural heritage institutions. Within the CENDARI metadata strategy, the concept of the collection is positioned between the institution and the item. In most cases, each collection will be associated with one institution that is responsible for it, and each collection record may also be associated with any number of item records providing detailed descriptions of items within the collection. CCS was designed to better meet the requirements of CENDARI users than existing standards by:

- extending the standard collection-level description metadata that would be found in encodings such as EAD;
- overcoming the semantic limitations of highly descriptive elements.

The CIDOC Conceptual Reference Model (CIDOC-CRM) is an essential standard for SSK. It is developed for domain experts in cultural heritage and related domains, providing a common and extensible semantic framework, with definitions and a formal structure to describe the implicit and explicit concepts and relationships used in cultural heritage documentation, to map and describe relevant information on cultural heritage objects, and to formulate requirements for information systems.

3.2.1.3 Prioritizing work

CIDOC-CRM should be included in the SSK as soon as possible. After that, it would be useful to add also DM2E, the standard schema for the manuscript datasets which comprise a significant portion of the studies of the past data.

3.2.2 Language-related Studies

3.2.2.1 Overview

The analysis is based on the use cases mentioned in D4.1 together with the standards, formats, and services mentioned in D4.2. Besides these, the CLARIN standard guidance website listing all standards relevant for the CLARIN community within the fields of language related studies (https://www.clarin.eu/content/standards-and-formats) has been studied. This CLARIN list is non-exhaustive, but still rather comprehensive. The CLARIN list is used as a source of inspiration and the chosen standards considered most important. Basically, language-related studies, as an academic field, are well covered by the standards already mentioned in D4.1 and D4.2 - and there are only a few suggestions for additions. Some of the most important additions are described in the following.

3.2.2.2 Standards and resources

The IASA's (International Association of Sound and Audiovisual Archives) Guidelines on the Production and Preservation of Digital Audio Objects gives valuable information about (de facto) standards, formats and schemas applied in relation to work in the sound archiving field. IASA supports international cooperation between audiovisual archives, e.g.

within the areas of acquisition and exchange, documentation and metadata, copyright, and digitization. Most of the standards mentioned in these guidelines already appear in the lists of standards from D4.1 and D4.2, but the guidelines provide a nice overview of the field and give recommendations for different solutions in diverse contexts.

Language-related studies sometimes include semantics in terms of knowledge structures and information management and this is the reason why the ISO/IEC standard for Topic Maps should be included in the SSK. Topic Maps ISO/IEC 13250 is a standard for the representation and interchange of knowledge, with an emphasis on the findability of information. The concept Topic Maps is also often described as a paradigm for description and interchange of complex relationships between abstract concepts and real-world resources by the use of a standard XML syntax.

The RDF-based Web Ontology Language (OWL) is a semantic web language designed to represent complex knowledge about things, groups of things, and relations between these things when this knowledge must be processed by applications (as opposed to humans). This representation of terms and their interrelationships is called an ontology and can be published on the World Wide Web and may refer to or be referred from other OWL ontologies. An expert-centre within this field is CLARIN-DK at the University of Copenhagen.
We also included the Open Language Archives Community (OLAC) and ISLE Meta Data Initiative (IMDI) metadata schemas even if OLAC (an extension of Dublin Core) is often considered too superficial and IMDI is often regarded as too much tailored towards specific research communities. The CLARIN community recommends the use of the Component Metadata Infrastructure (CMDI), which is not just a schema, but also a meta-model providing the framework necessary to define and use an individually tailored schema. Still, OLAC and IMDI are used in connection with many resources and should be included in the SSK. Experts of OLAC and IMDI are e.g. the CLARIN centres, CLARIN-PL and Max Planck Institut für Psycholinguistik, Nijmegen.

3.2.2.3 Prioritizing work

The standards suggested here for the SSK are already a shortlist of all standards within the field of language-related studies so most of the standards are really needed and

should be included in the SSK as soon as possible. However, the IMDI standard and stand-off annotation (which is not a standard as such) could wait till later.

3.2.3 Heritage, applied disciplines and Archaeology

3.2.3.1 Overview

Cultural Institutions, both public and private, support through national and international programmes, the converting of the 'physical cultural heritage' into 'digital'. This happened not only for documents and images but also for audio/video resources, for the performing arts and for the monuments, artworks, and archaeological finds. Also, intangible cultural heritage such as oral memories, food and drink, local traditions have been made digital so that no aspect fails to be present in the digital world. After digitizing, the content is also made available on the web, managed and collected by digital libraries, aggregators and portals for possible reuse and enjoyment by the whole community. When cultural information becomes digital there are often barriers to its open dissemination, and therefore reuse, for research and for portals and aggregators. One of them is the Intellectual Property Rights barriers, in particular of digitized cultural information with a heavy burden on proprietary attitudes and policies. Another is that cultural institutions have problems in implementing internationally accepted digitization standards that support interoperability and openness.

It is strategic to digitize collections and make them available along internationally and well-accepted standards for interoperability and openness, to enhance the digital cultural heritage and its reuse for research.

Europeana, the great European digital library, that contains over 53 million objects from 3,500 museums, archives, and libraries from across Europe encourages and supports the dissemination of open standards for digitizing and disseminating digital cultural contents (http://pro.europeana.eu/share-your-data/data-guidelines).

There are also other international associations in the cultural heritage sector such as IFLA (International Federation of Library Associations and Institutions), that has disseminated standards over the last fifty years, in all fields of library and information services. IFLA standards are internationally reviewed, published and regularly updated. Each IFLA

standard reflects current consensus on rules, principles, guidelines, best practice or models for a particular activity or service. A complete list of standards is available online (https://www.ifla.org/node/8750).

3.2.3.2 Standards and resources

The EDM should be considered and well presented in SSK. EDM was developed by Europeana in order to aggregate resources coming from different cultural heritage fields. For this reason, several experts of library, museum, archive and audio-visual collections contributed to building this data model that replaced the ESE (Europeana Semantic Elements), the first model adopted that wasn't able to preserve the richness of original data. The EDM accommodates some of most important international standards: The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), Open Archives Initiative Object Reuse and Exchange (OAI-ORE), Dublin Core, Simple Knowledge Organization System (SKOS) and CIDOC-CRM. The actual structure of EDM contains three main classes: ProvidedCHO (it provides information on the physical object), Web Resource (it provides information on web resource) and AggregationCHO (it provides information on the data provider). Thanks to these three classes, it is possible to avoid the overlapping of information between physical and digital object. EDM: rights support the 14 Rights statements to express the copyright status of a Digital Object, as well as information about how the users can access and reuse the objects. The list of available rights statements for Europeana are based on Creative Commons and Rightsstatements.org licensing framework that could be adopted also by the other research communities. Recently Europeana created an IIIF (International Image Interoperability Framework) profile for the Europeana Data Model that makes EDM interoperable with the IIIF standard, that develops and documents shared technologies, such as image servers and web clients that provides a world-class user experience in viewing, comparing, manipulating and annotating images.

Note that OAI-PMH, OAI-ORE, Dublin Core, SKOS and CIDOC-CRM, Creative Commons, Rightsstatements.org, and IIIF are standards widely used by the cultural heritage institutions, also not only strictly related to Europeana, so they should be described in the SSK separately.

For the libraries domain we suggest to include two fundamental standards developed and managed by IFLA:

- Functional Requirements for Bibliographic Records (FRBR): a conceptual schema, that provides a model entity-relationship to give a representation of bibliographic information. It distinguishes the essential entities for a different type of users, the attributes of these entities and the relationship between them.
- International Standard Bibliographic Description (ISBD): a standard to make a format accepted at international level, to make bibliographic data universally and easily available. It aims to give uniform criteria for sharing bibliographic information between national bibliographic agencies on one hand and the libraries on the other hand. The ISBD establishes a registration format composed of nine areas. This standard was adopted by national cataloguing codes and used by most of the national bibliographies. The descriptors are the following: 0: Content form and media type area 1: Title and statement of responsibility area 2: Edition area 3: Material or type of resource specific area 4: Publication, production, distribution, etc., area 5: Material description area 6: Series area 7: Notes area 8: Resource identifier and terms of availability area.

In the heritage science domain there is a need to provide standards regarding:

- Multispectral imaging for surface mapping of pigments, that describes a method to record multispectral images of colour painted materials and artworks, which is a commonly used technique currently available to the scientist, conservator, archaeologist and art historian for the non-invasive investigation of works of art.
- Digital 3D Objects in art and humanities. In the PARTHENOS working groups there are interesting studies that could contribute to formulating standards in those fields.

In particular, the PARTHENOS experts' working group on 3D is busy laying the foundations for a comprehensive environment centred around the researchers' practices on and with 3D digital objects and the White Paper 'Digital 3D Objects in Art and Humanities challenges of creation, interoperability, and preservation'.[3] The publication gathers contributions from more than 25 experts in 3D imaging, modelling, and processing, as well as professionals concerned by interoperability and sustainability of research data.

[3] https://hal.inria.fr/hal-01526713

The topics addressed in this document are meant to help to ensure the development of standardized good practices relating to the production, the handling, the long-term conservation and the reuse of 3D objects. Therefore, even if the focus is put on technical questions (formats, processing, annotation), the White Paper also addresses the need to clarify the legal status of 3D objects, in order to facilitate their reuse in non-research contexts, in particular in museums. This White Paper is the result of a workshop organized by CNR (Italy), CNRS (France) and Inria (France) within in the scope of WP4 on Standardization, with support from the technical partners and on behalf of the PARTHENOS research infrastructure.

3.2.3.3 Prioritizing work

Most of the standards for the cultural heritage sector are shared with other sectors, such as the community of studies of the past (EDM, CIDOC-CRM, Dublin Core, Creative Commons, OAI-PMH, OAI-ORE, SKOS and IIIF) so they should be easily included in SSK. Regarding the libraries sector, it is suggested to include the cited standards because libraries are not only belonging to the cultural heritage sector but also to the research and academic centres. Standards on heritage science are under development, but it is strategic for PARTHENOS to propose the first building blocks for standards for the 3D objects.

3.2.4 Social Sciences

3.2.4.1 Overview

Social sciences cover several disciplines focused on societies and individuals among which are anthropology, ethnology, political science, sociology or psychology. Due to the variety of disciplines, the differences in research practices and the scattering of the resources, it is difficult to find standards especially dedicated to social sciences and shared by all. Nonetheless, some standards are in use in social sciences. They firstly concern metadata. Thus, the Metadata Encoding & Transmission Standard (METS) is a standard for encoding metadata of objects in digital libraries whereas the Data Documentation Initiative (DDI) which complies with metadata standards of Dublin Core aims more specifically to describe observational data. In the same way, the DataCite Metadata Schema focuses on bibliographic data while the Statistical Data and Metadata eXchange (SDMX) is designed for statistical data and metadata.

Concerning the data, the Text Encoding Initiative (TEI) presents a set of guidelines for encoding different kinds of documents including surveys. Moreover, some institutions have created standard classifications which help researchers to assemble, compile and analyse comparable data across countries. For example, UNESCO and ILO (International Labour Organization) have respectively developed the International Standard Classification of Education (ISCED) and the International Standard Classification of Occupations (ISCO). In a similar way, efforts have been made to provide internationally recognized frameworks by establishing clear terms, definitions and service requirements thanks to vocabularies like the ISO 20252:2012 Market, opinion, and social research – Vocabulary and service requirements.

Social sciences researchers often collect and analyse heterogeneous data. There can be many steps involved in getting a clean result. For reviewing and reusing these data, it is crucial that they are accompanied by a detailed documentation including amongst others the research questions, methods in use, conventions used, and provenance of material. Such documentation can be encoded with TEI. It is unclear if the TEI is commonly used by social science researchers. Therefore, it could be a starting point to find out how suitable TEI is for social sciences research. Besides, it is essential that all of the data, documentation, and research results are bundled into a collection, particularly for longitudinal surveys, where data over long periods are collected and compared. Handling the research data cycle can be organized by different tools, e.g. the Open Science Framework that supports the entire research lifecycle (https://osf.io/).

Finally, some tools are widely used for collection, further data refinement, and analysis. Thus, in order to perform a statistical analysis which is a prominent approach in social sciences, the researchers can use for example SPSS (Statistical Package for the Social Sciences) or STATA which are leading but commercial statistical software or free alternatives like R, PSPP or JASP (Jeffreys's Amazing Statistics Program). The researchers can be bound by the data output they get from the tools or from the data providers. Especially social media research depends on the availability of data samples, e.g. via an application programming interface (API). When the tool produces proprietary file formats, then it is recommended that data are additionally exported or compiled in open data formats like CSV or XML where possible.

3.2.4.2 Standards and resources

For now, because of the weaker representation of these communities in the PARTHENOS project, and recognition that of all the SSH disciplines, social sciences is the most different, the social sciences have received rather less attention in the different deliverables with very few standards or resources especially dedicated to them. Despite that, some 'general' standards such as TEI, Dublin Core, METS, etc. are also in use in social sciences and are already integrated.

The Data Documentation Initiative (DDI) is a widely used standard in social sciences with a strong community. So, it is recommended to integrate it into the SSK. The DDI as a metadata format describes the data but for reusing and reviewing the research results it is necessary to have insights into the complete creation process of these results. Therefore, well-established workflows should be documented in the social science scenarios of the SSK.

As a resource, it would be useful to have a list of tools that produce reusable output in terms of recommended standard formats. Ideally, there should also be a scenario that deals with social media research and points out recommended data formats for API output, e.g. JavaScript Object Notation (JSON, http://json.org/).

Ethical aspects are crucial when working with data in social sciences. Data from surveys, interviews, ethnographic observations or social media may involve personalized or sensitive information. Handling of such data depends on legal issues and on ethical implications. It would be of great value if a scenario that covers policies for such ethical aspects, e.g. protection of research subjects, was provided by the SSK.

3.2.4.3 Prioritizing work

There is not a common standard in social sciences in terms of data formats but there are best practices in place for data stewardship. Accordingly, PARTHENOS could provide valuable input if not by promoting specific standards then by disseminating documentation and by providing examples of best practices and standard methods. As there are many approaches in the field of social sciences with different needs, there should be at least two scenarios in the SSK that cover a quantitative and a qualitative approach. An additional

benefit could be achieved if the design of these scenarios focuses on combining approaches from the social sciences with other research communities in PARTHENOS.

3.2.5 'Why Standards' leaflet

WP4 of PARTHENOS, in collaboration with the DARIAH Guidelines and Standards working group, created the 'Why standards?' leaflet as a partial component of the Standardization Survival Kit (SSK) and, consequently, of the PARTHENOS website (www.parthenos-project.eu). Its objective is to highlight the importance of structuring data in certain formats and raise the research community's interest in standardization, aiming especially at scholars with limited technical background.

Several representatives of the humanities, experts in standardization issues and graphic designers worked closely, from May to October 2016, to produce the leaflet. The first stage of their work was to investigate and catalogue existing standards, accompanied by real case scenarios. They also considered the use of cartoons in a metaphorical framework, to vividly elaborate the necessity of using standards in research activities. Then, the working groups prepared a draft in user-friendly format, containing the title (Why standards?) and other important information, such as why a scholar should read this leaflet (motivation and what the leaflet offers), the role of standards and why to use them (both generally and in humanities particularly), how research would be without standards (reusability, compatibility problems etc.), how the use of digital standards can affect the research community (that is, why we actually need standards), and links to PARTHENOS and DARIAH Working Group Guidelines and Standards.

The complete leaflet[4] consists of the comic, the PARTHENOS logo title page (maintaining the PARTHENOS graphic charter) and a shortened version of the initial (longer) brochure, which concludes that the use of standards not only ensures the data quality but also enables researchers to benefit from each other's work.

[4] http://www.parthenos-project.eu/Download/Flyer-Parthenos_standards_ls.pdf

3.3 Annex to the standards chapter – systematic gap analysis

Studies of the Past

Name	Standards involved (separated by ";")	Schema Format	Research field(s) (the disciplinary scope of use of the standard)	TaDiRAH Activity(ies)	Overview	Working material	Bibliographical references	Tools, services and samples	Prioritizing of the resources	Creator / Developer of the standard	Experts on the standard	Last Update	License	Remarks and comments
Project oriented EAD customization	EAD;TEI;Schematron; ISO 639;ISO 3166;ISO 15511;ISO 15924	XML	Studies of the past	Interpretation /modeling	TEI ODD can be used to document data models external to the TEI environment. Several projects working with archival standards (in particular EAD) use it as well. Parthenos created and maintain an instance of the EAD specification in ODD, that can be used to create project oriented customizations. With ODD, semantic and structural consistency is ensured as we encode and document best practices in both machine and human-readable format. ODD was created at first to give TEI users a straightforward way to customize the TEI schema according to their own practices and document this customization. But it is possible to describe a schema and the associated documentation of any XML format. ODD can be processed to generate an actual schema (a DTD, an RelaxNG XML, with embedded schematron, or compact schema and an XML schema), and documentation in various formats (XHTML, PDF, EPUB, docx, odt). We used ODD to encode completely the EAD standard, as well as the guidelines provided by the Library of Congress, and then derived a specific customization using Schematron rules, also described with ODD. The solution we propose is based on a flexible and customizable methodology : It combines the complete description of the specifications in a machine-readable way, and customization facilities, easy to understand for the end-user. More important, this solution doesn't change the core EAD schema, but add more specific rules in a comprehensive and human-readable format, by combining the EAD schema (expressed in RelaxNG) with ISO Schematron rules. Schematron is an ISO/IEC Standard (ISO/IEC 19757-3:2016) that parses XML documents and makes "as-ser-tions about the pres-ence or ab-sence of pat-terns". It can be used in conjunction with a lot of grammar languages such as DTD, relaxNG, ...	continuing/catalogs		https://github.com/ParthenosWP4/standardsLibrary/tree/master/archivalDescription/EAD/odd https://github.com/EHRI/data-validations/tree/master/ODD-RelaxNG/EAD		EHRI, Parthenos, Inria		March 2017	CC-by	
Omeka plugin: management of authority files	EAC-CPF; Dublin Core	XML	Studies of the past	Interpretation/	The solution described aims at simplifying the access, management and interoperability of prosopographical data: a file management tool, with publishing and interoperable capabilities that can be handled without a steep learning curve. It uses Omeka, the CMS for scholarly content, digital collection and exhibits. This system is able to ingest and produce authority files in different formats (XML, HTML, CSV, etc) supporting different standards (Dublin Core, FOAF, TEI, EAC-CPF, etc.) without requiring any special operation from the users. The authority records are ingested in XML markup following EAC-CPF (Encoded Archival Context - Corporate bodies, Persons and Families) convention, a quite complete format that allows to structure communities descriptions, individuals or families. It follows the indications of the second edition of ISAAR (CPF), the international standard for the description of archival producers.	?	https://www.zotero.org/groups/parthenos-wp4/items/collectionKey/I9X3MUTP; https://www.zotero.org/groups/parthenos-wp4/items/collectionKey/Z3ABBMDH	https://github.com/sgraziella/prosopography_LJP		PARTHENOS, Inria		September 2016	?	
ALTO-XML	METS	XML	Studies of the past	Interpretation /modeling	ALTO (Analyzed Layout and Text Object) is an open XML Schema developed by the EU-funded METAe project group for use with the Library of Congress' Metadata Encoding and Transmission Schema (METS). However, ALTO instances can also exist as a standalone document used independently of METS. The standard was initially developed for the description of text OCR and layout information of pages for digitized material. The goal was to describe the layout and text in a form to be able to reconstruct the original appearance based on the digitized information - similar to the approach of a lossless image saving operation. ALTO stores layout information and OCR recognized text of pages of any kind of printed documents like books, journals and newspapers. METS provides metadata and structural information while ALTO contains content and physical information. CCS Content Conversion Specialists GmbH maintained the ALTO standard until 2009. This company was involved with ALTO during the METAe project. From 2009, the Library of Congress (LC) Network Development and MARC Standards Office became the official maintenance agency for the ALTO XML Schema. At that time LC set up an Editorial Board to help shape and advocate for ALTO. The Board thus oversees maintenance of the ALTO XML Schema and helps foster usage in the digital library community. ALTO XML is also being used by Archival institutions to enhance access and fulltext findability of digitized Archives. When large digitized corpora with ALTO-XML become full-text searchable, Historical Research, Language studies profit. It eg. makes Named Entity Recognition and Text-mining possible.	Computer Files/Document	http://www.loc.gov/standards/alto/	https://github.com/altoxml		Library of Congress		2016-01-25	?	
EAC-CPF	EAC-CPF	XML	Studies of the past	Interpretation /modeling	Since long time, researchers have been discussing the need for a standard structure for the recording and exchange of information about the creators of archival (and - possibly - other kind of) materials. A group of archivists has defined the model "Encoded Archival Context - Corporate Bodies, Persons, and Families" (EAC-CPF), emphasizing its important role in archival description and its relationship with the Encoded Archival Description standard. This standard would provide a communication standard for the exchange of authority records based on International Standard for Archival Authority Records—Corporate Bodies, Persons, Families (ISAAR(CPF)) and would parallel the standard for encoding archival record finding aids that was found in Encoded Archival Description (EAD). A separate standard would pave the way to eliminating some practical problems found in the use of EAD, which had been developed as a comprehensive solution for encoding standalone finding aids which held all forms of descriptive data about archival records.	?	Data repository http://eac.staatsbibliothek-berlin.de/ Data Schema http://eac.staatsbibliothek-berlin.de/eac-cpf-schemas.html Publications http://eac.staatsbibliothek-berlin.de/tag-library/publications.html			EAC working group, Technical Subcommittee for EAC-CPF		2011	?	
MEI - Music Encoding Initiative	MEI Schema	XML	Studies of the past	Interpretation /modeling	The Music Encoding Initiative (MEI) is "an open-source effort to define a system for encoding musical documents in a machine-readable structure". This encoding format, commonly (and somewhat confusingly) also referred to as MEI, is one of the many music encoding standards existing today. The format, which dates back to 1999, is based on existing encoding standards - in particular, it is modeled on the Text Encoding Initiative (TEI) DTD, meaning that it is XML-based. It brings together researchers from various communities, including computer scientists, information scientists, musicologists, music theorists, librarians, and historians, and aims to define best practices for the representation of a broad range of musical documents and structures - thus facilitating the exchange, examination, validation and comparison of such documents. MEI is primarily catered towards an academic audience; as such, it distinguishes itself from the other XML-based music encoding format currently at the forefront, MusicXML, which has a strong commercial interest. The primary reference point for researchers or others interested wanting to engage with MEI is the official website, http://www.music-encoding.org. Here one can find, among many other things, a "Gentle introduction to MEI", various more in-depth tutorials, an extensive bibliography covering the history of the project from its conception to the latest developments, the proceedings of the annual conference (see below), guidelines providing extensive documentation of the different components of the MEI model as well as best practice suggestions, and an overview of tools and projects that utilise MEI (more on tools and projects below). The MEI community maintains an official mailing list, MEI-L, which is used as its general communication channel. Through this list, community members are informed about relevant events; moreover, it functions as a discussion platform. One such event is the annual Music Encoding Conference (MEC), which since 2013 has taken place alternately in Europe and in North America.	Music	Official website: http://music-encoding.org. Bibliography: an extensive bibliography can be found at http://music-encoding.org/community/bibliography.	MEI GitHub repository, containing the MEI schema (as well as various customisations), the MEI guidelines, sample encodings, stylesheets, and the source code and documentation for a number of tools: https://github.com/music-encoding. Several tools have their own GitHub repository; see the links at http://music-encoding.org/tools. Customization Service Sibelius to MEI Plugin Verovio MEI to Music21 Converter MEItoVexFlow LibMEI MEISE MerMEId		the Music Encoding Initiative Board		2016	Licensed under the Educational Community License, Version 2.0	

Prosopography	EAC-CPF;FOAF;PND (now GND);VIAF	XML, RDF	Studies of the past	Analysis, Interpretation, Storage	Prosopography is the investigation of the common background characteristics of a group of actors in history, making a collective study of their lives. Prosopography is mostly used by historians to address two main research questions: roots of political action: e.g. the interests beneath the rhetoric of politics or the social and economic affiliations of political groupings; social structure and social mobility: e.g. the role in society, the degree of social mobility and the correlation of intellectual or religious movements with other factors. Among the typical products of researchers working on prosopography there are various kinds of repertoires, hand lists and other reference tools, such as: lists of names, holders of certain offices or titles or educational qualifications; family genealogies; full biographical dictionaries, which are usually built up in part from the first two categories and in part from an infinitely wider range of resources. With the digital turn in the humanities traditional (printed) reference tools have been digitized, and new ones have been produced ex-novo: at first on CD-ROMS and DVDs and - eventually - published online. A wide range of disciplines in the Humanities and Social Sciences are represented in PARTHENOS: along with authority lists of persons and places names, a wider set of thesauri, produced in different research areas - will be available in the project content cloud. For this reason a specific VRE named RubRIcA (see infra for a detailed description) - is under development, to address all the integration needs of a complex digital research infrastructure. RubRIcA is developed in collaboration with WP2 (requirements), WP5 (modeling and mapping) and WP6 (Implementation) and will be supporting a specific use case based on integration and standardization of reference resources about prosopography. In the supported workflow the researcher has to establish a universe to be studied, and answer to a set of uniform questions (e.g. birth, death, family, social origins, economic position, place of residence, education, amount and source of personal wealth, occupation, religion, experience of office and so on). The various types of information gathered about individuals in this universe should be then compared, combined, and examined for significant variables. Finally, these types of information are tested for internal correlations and for correlations with other forms of behavior or action. At the end of the process, the researcher should be able to use the information obtained to address specific research questions (for example): make sense of political action, in order to help explain ideological or cultural change, to identify social reality and to describe and analyze with precision the structure of society and its movements.	Biography	http://www.sismelfirenze.it/index.php/banche-dati/bibliotheca-scriptorum http://www.sismelfirenze.it/index.php/banche-dati/compendium-auctorum https://viaf.org/ http://www.getty.edu/research/tools/vocabularies/ulan/index.html http://rameau.bnf.fr/utilisation/liste.htm		.		.		
Project Oriented EAG Customization	EAD;TEI;MODS	XML	Studies of the past	Interpretation /modeling	Most CENDARI holdings schemas are mapped to EAD (Encoded Archival Description), the core standard for collection-level descriptions, but some components are used to generate the EAG (Encoded Archival Guide) records which lie above EAD in the overall hierarchy. EAG (CENDARI flavour) is a version of EAG designed to meet the needs of CENDARI regarding Archival Guides. The CENDARI Collection Schema (CCS) was developed to encode detailed descriptions for collections housed by the associated cultural heritage institutions. Within the CENDARI metadata strategy collection is conceptualized as being positioned between the institution and the item. In most cases each collection will be associated with one institution that is responsible for the collection, and each collection record may also be associated with any number of item records providing detailed descriptions of items within the collection. CCS was designed to better meet the requirements of CENDARI users than existing standards by: extending the standard collection-level description metadata that would be found in encodings such as EAD; overcoming the semantic limitations of highly descriptive elements; The schema is written in XML (eXtensible Markup Language), a widely-used standard for metadata encoding and interchange. It aims to provide a structure to allow the most important components of collection information to be collocated and linked up as necessary. The schema defines 16 top-level components and a mechanism for linking these together using XML identifiers: in addition, every component may be identified by an Universal Resource Identifier (URI) by which it may be linked to external resources (such as the controlled vocabularies and ontologies).	Archival domain: collections, records, documents, holdings...	EAG Index of elements: http://apex-project.eu/images/docs/APEx_EAG_2012_table_20130527.pdf EAG Schema (XSD file): http://www.archivesportaleurope.net/Portal/profiles/eag_2012.xsd EAC-CPF: Schema http://eac.staatsbibliothek-berlin.de/schema/cpf.xsd EAC-CPF Diagram: http://eac.staatsbibliothek-berlin.de/Diagram/cpf.html EAG(CENDARI): customising EAG for research purposes, official document: https://hal.inria.fr/hal-	CENDARI Item Descriptions: for item-level descriptions, CENDARI uses the MODS (Metadata Object Description Schema), supplemented by elements from the TEI P5 Manuscript Description Schema and a small number of additional elements created by CENDARI. An example of an item level description is available here: https://wiki.de.dariah.eu/download/attachments/15409655/cendari-item.xml?version=1&modificationDate=1372253948363&api=v2 A skeletal draft of the documentation is available here: https://wiki.de.dariah.eu/download/attachments/15409655/item-level-documentation0-1.doc?version=1&modificationDate=1372253948370&api=v2		CENDARI				

Heritage, applied disciplines and Archaeology

Name	Standards involved (separated by ";")	Research field(s) (the disciplinary scope of use of the standard)	TaDiRAH Activity(ies)	Overview	Working material	Bibliographical references	Tools, services and samples	Prioritizing of the resources	Creator / Developer of the standard	Last Update	License	Remarks and comments
Multispectral imaging for surface mapping of pigments		Heritage and applied disciplines	Capture/imaging	This standard describes a method to record multispectral images of colour painted materials and artworks, which is a commonly used technique currently available to the scientist, conservator, archaeologist and art historian for the non-invasive investigation of works of art. This document will concentrate on the wavelength range that can be observed using modified commercially available cameras, which typically employ silicon based sensors sensitive from approximately 350 nm to 1100 nm. Cameras based on InGaAs sensors, which can record infrared radiation from approximately 700 nm to 1700 nm, can be used regularly in cultural heritage applications but due to their specialized technology they are out of the scope of this standard. Concerning the choice of material/artefact, our suggestion, is to consider, for example, painted stone sculpture. This method may be applied to: - painted artefacts either untreated or subjected to any treatment or ageing - representative surfaces of objects, indoors or outdoors.	Visual materials/art original	http://cool.conservation-us.org/coolaic/sg/emg/dtf/DTF_Online_Weblinks.pdf; http://www.vips.ecs.soton.ac.uk/index.php?title=VIPS http://www.labsphere.com/support/datasheets-library/; https://www.britishmuseum.org/pdf/charisma-multispectral-imaging-manual-2013.pdf; http://libvips.blogspot.com.es/?view=magazine			FORTH	?	?	
Digital 3D objects		Heritage and applied disciplines	Capture/imaging	With the White paper on "Digital 3D Objects in Art and Humanities: challenges of creation, interoperability and preservation", which gathers contributions from more than 25 experts of 3D imaging, modeling and processing, as well as professionals concerned by interoperability and sustainability of research data, the PARTHENOS project aims at laying the foundations of a comprehensive environment centered around the researchers' practices on and with 3D digital objects. The topics addressed in the document are meant to help ensuring the development of standardized good practices relating to the production, the handling, the long-term conservation and the reuse of 3D objects. Therefore, even if the focus is put on technical questions (formats, processing, annotation), the White Paper also points the need to clarify the legal status of 3D objects, in order to facilitate their reuse(s) in non-research contexts, in particular in Museums. Today, the digital model has become essential for scientific documentation and analysis. However, with the rapid development and spread of 3D technology, there is an urgent need to integrate and customize the related visualization and analysis tools to support the specific needs of users within the Arts and Humanities research communities. Since the number of models produced increases exponentially, the need of efficient archival systems able to provide effective search and retrieval functionalities is also arising. This White Paper is the result of a workshop organized by CNR (Italy), CNRS (France) and Inria (France) within in the scope of Work Package 4 on Standardization, with support from the technical partners and on behalf of the PARTHENOS research infrastructure. It was held in Bordeaux (France), from November 30th to December 2nd, 2016, and entitled "Digital 3D objects in Art and Humanities: challenges of creation, interoperability and preservation". The workshop was also supported by the work of Huma-Num's 3D-SHS consortium. The workshop has been attended by selected PARTHENOS partners as well as some external experts, representative of both the technological and humanities domains (see program in Appendix). It aimed to enrich technical knowledge about 3D models, standards and tools in the Parthenos framework, addressing the common issues and epistemological questions related to the creation, use, reuse and preservation of 3D models. More precisely, the objectives were to: Identify best practices and standards to ensure interoperability and sustainability; Expand knowledge for scholars and researchers to support 3D projects in arts, social science and humanities; Bridge the gap between technical people and humanities scholars (contributing to a better understanding of technologies potential and user needs); Share general and targeted knowledge on 3D objects issues in Art and Humanities; Contribute to best practices in the digitization domain for archaeologists and human sciences scholars (including 3D preservation issues: representation schemas, viewers, etc). We selected four main topics to focus on during the workshop, corresponding to the life cycle and the various uses of 3D objects in the Humanities: (a) production and processing, (b) visualization and analysis, (c) description and preservation, and (d) bridges between Cultural Heritage and Museology. For each one of those, a number of sub-topics and issues have been discussed by domain specialists in brief presentations followed by a free discussion. Those topics are the basis of the core chapters of this white paper.	Visual materials/art original	See D4.2 (section on "Digital 3D objets in Arts and Humanities")			.	.	.	
Raman microspectrometry		Heritage and applied disciplines	Capture/imaging	Scope of the standard The specific standard describes a detailed methodology to record Raman spectra of colour painted materials and artworks for the non-invasive identification of organic and inorganic pigments. This document will be present standard protocols that can be applied in different types of Raman instruments (bench-top and portable) and in various types of laser sources. Concerning the choice of material/artefact, our suggestion, is to consider, for example, painted stone sculpture. This method may be applied to: painted artefacts either untreated or subjected to any treatment or ageing representative coloured surfaces of objects.	Visual materials/art original	Analytical Methods Committee, AMCTB No 67, Raman spectroscopy in cultural heritage: Background paper, Anal. Methods, 2015,7, 4844-4847. DOI: 10.1039/c5ay90036k I. M. Bell, J. H. Clark, P. J. Gibbs, "Raman spectroscopic library of natural and synthetic pigments (pre-~1850 AD)", Spectrochim. Acta A 53, 2159-2179 (1997). Also at: http://www.chem.ucl.ac.uk/resources/raman/index.html L. Burgio, R. J. H. Clark, "Library of FT-Raman spectra of pigments, minerals, pigment media and varnishes, and supplement to existing library of Raman spectra of pigments with visible excitation", Spectrochim. Acta A 53, 1491-1521 (2001). P. Vandenabeele, B. Wehling, L. Moens, H. Edwards, M. DeReu, G. Van Hoydonk, "Analysis with micro-Raman spectroscopy of natural organic binding media and varnishes used in art", Analytica Chimica Acta 407, 261-274 (2000). P. Vandenabeele, L. Moens, H. G. M. Edwards, R. Dams, "Raman spectroscopic database of azo pigments and application to modern art studies", J. Raman Spectrosc. 31, 509-517 (2000). P. Colomban, G. Sagon, X. Faurel "Differentiation of antique ceramics from the Raman spectra,of their colored glazes and paintings" J. Raman Spectrosc. 32, 351-360 (2001) M. Bouchard, D. C. Smith, "Catalogue of 45 Raman spectra of minerals concerning research in art history or archaeology, especially on corroded metals and coloured glass", Spectrochim. Acta. A 59, 2247-2266, (2003) California Institute of Technology, Division of Geological and Planetary Sciences (USA) (http://minerals.gps.caltech.edu/files/raman/) The RRUFF project, Univ. of Arizona (USA) (http://rruff.info/) The Infrared and Raman Users Group (IRUG) Spectral Database (http://www.irug.org) e-VISART Database, Univ. of the Basque Country, Dept. of Analytical Chemistry (Spain) (http://www.ehu.es/udps/database/database.html) Spectral Database for Organic Compounds, AIST (Japan) (http://riodb01.ibase.aist.go.jp/sdbs/)			?	?	?	
EDM	CIDOC-CRM, DUBLIN CORE; OAI-ORE	Heritage and applied disciplines	Interpretation / Modeling	The EDM (Europeana Data Model) was developed by Europeana in order to aggregate resources coming from different cultural heritage fields. For this reason, several experts of library, museum, archive and audio-visual collections gave their contribution to build this data model that replaced the ESE (Europeana Semantic Elements), the first model adopted that wasn't able to preserve the richness of original data. The actual model is able not only to cover different fields of the cultural heritage (ie. Museums and Libraries) but also four different types of resource: image, text, video, sound. The EDM basis are some of most important international standards: OAI-ORE, Dublin Core, SKOS and CIDOC-CRM. The actual structure of EDM contains three main classes: ProvidedCHO (it provides information on the physical object), Web Resource (it provides information on web resource) and AggregationCHO (it provides information on the data provider). Thanks to these three classes, it is possible to avoid the overlapping of information between physical and digital object. Moreover, great attention has been given on the semantic. EDM, in fact, "is a framework for collecting, connecting and enriching metadata. It does this by adhering to the modelling principles that underpin the approach of the Web of Data ("Semantic Web") connecting to generate new knowledge between nodes in the cultural heritage sector."		http://pro.europeana.eu/page/edm-documentation			Europeana Foundation			
FRBR		Heritage and applied disciplines	Interpretation / Modeling	It a conceptual schema developed by IFLA (International Federation of Library Associations and Institutions), produced through a model entity-relationship to give a semi-formal representation of bibliographic information. It was born at the end of XX century. The main goal of FRBR is to develop a conceptual model that allows to identify the essential requirements of bibliographic record, defining its structure and purpose. It is interesting to underline that the structure of FRBR consider not only the point of view of readers and/or staff libraries, but also of editor, publisher and so on. This analysis has made possible to distinguish the essential entities for different type of users, the attributes of these entities and the relationship between them. The first group of entities refers to the aspects can be considered as a part of an intellectual production. The second group refers to the entities that are involved in the following process: creation, implementation, distribution and management of the first group. The third group refers to the entities that are the subjects of the works. Moreover, all the entities are associated to some attributes, divided in two main categories: attributes directly related to the entities; attributes outside the entities. Another relevant aspect of FRBR is represented by the relationship. They allow the users, through a search, to identify links between entities and to surf between records.		https://www.ifla.org/publications/functional-requirements-for-bibliographic-records			IFLA			

ISBD		Heritage and applied disciplines	interpretation / Modeling	The ISBD (International Standard Bibliographic Description) is a standard produced by the International Federation of Library Associations and Institutions (IFLA) to make universally and easly available, in a format accepted at international level, bibliographic data about all kind of resource published in each country. The stable version was released in 2011 after a preliminary version produced in 2007. its promotion, revision and update are managed by ISBD revision group. It aim to give uniform criteria for sharing bibliographic information between national bibliographic agency on one hand and the entire library on the other hand. The ISBD establishes a registration format composed by nine areas. This standard was adopted by national cataloguing codes and used by most of national bibliographies. The descriptive are the following: 0: Content form and media type area 1: Title and statement of responsibility area 2: Edition area 3: Material or type of resource specific area 4: Publication, production, distribution, etc., area 5: Material description area 6: Series area 7: Notes area 8: Resource identifier and terms of availability area					IFLA			
OAI-PMH + OAI-ORE	XML	Heritage and applied disciplines	capture/gathering	The Open Archives Initiative was born to make easier the connection between archives that contains documents produced in academic field. Actually, it aim to provide, at a reasonable cost, an instrument able to share information structured in different ways. The first version of OAI-PMH was replaced by the version 2.0 after that the W3C consortium modified the xml standard. The OAI-PMH, in fact, is based on xml and http. The OAI-PM provides a framework of interoperability, independent from the application, based on metadata collection. Thanks to a specific protocol requests, it is possible, for the service provides, to make a series of queries in order to harvest one record or a specific dataset of digital resources. The OAI-ORE (Open Archives Initiative - Object Reuse Exchange) was developed by OAI to implement a standard model for interoperability of aggregated resources. The stable version was released in 2008 after two years of work. This standard provides the description and exchange of digital resources, available online. In fact, one of biggest issue of world wide web was the absence of a standard way to describe the constituents or boundary of an aggregation, and this is what OAI-ORE aims to provide. "The ORE standard is concerned with the description of aggregations of web resources. It defines 4 entities: • the Aggregation itself, identified by its URI but without any corresponding concrete resource. It is a conceptual resource. Being uniquely identified it can enter into relationships with other resources, in particular aggregations of aggregation become possible. • the Aggregated Resource: any resource part of an aggregation, identified by its URI • the Resource Map: a resource describing an aggregation based on a set of assertions. A mandatory assertion indicates which aggregation the Resource Map itself is describing. Other assertions indicate the aggregated resource(s). Certain metadata are mandatory as well, such as the map creator. Dublin Core terms are used for this purpose. • the proxy: a virtual resource acting as a proxy for a certain aggregated resource in the context of a certain aggregation. Its use is optional. A so-called lineage relationship can be established between proxy resources to trace the origin of an aggregated resource from another aggregation. The standard accounts for the possible redundant description of the same aggregation and defines the notion of authoritative Resource Map. It also consider the notions of similar aggregations and of type of aggregated resources. The standard is open in the sense that a Resource Map may include any additional assertions about resources. Finally, the ORE standard builds upon the Cool URIs guideline, which discussed two strategies for not confusing a thing and some representation of it." (www.en.wikipedia.org)		www.openarchives.org/			OAI	2002 (OAI-PM) - 2008 (OAI-ORE)		
CreativeCommons		Heritage and applied disciplines	Dissemination/sharing	Creative commons are probably the most widely used licensing framework. they arise to provide a tool that clearly define the possibilities of reusing data. Although it is currently being used by a heterogeneous community, it is clear that CC was born mainly in the debate on the reuse of digital resources produced in the public domain. It was born on 2001 with the support of the Center for the Public Domain and actually it is managed by a Board of Directors composed by relevant expert in the field of data reuse. This licensing framework provides three possible levels of data reuse: the highest level is Public Domain Dedication and Public Domain Mark licenses. The public domain is license is referred to resources that can be reused in any way and for any purpose. At a lower level, even if they are defined 'free culture', there are licenses that request the attribution (excluding CC-BY-ND, or non-derivative works). In this case, the only limit imposed on users who intend to reuse data is the attribution. Finally, there is a third level, commonly referred to as 'no free culture', which includes all other Creative Commons licenses and which restrict, in various ways, the possibilities for re-use by users.		https://creativecommons.org/						
IIIF (International Image Interoperability Framework)		Heritage and applied disciplines	capture/gathering	The International Image Interoperability Framework (IIIF) is a set of shared application programming interface (API) specifications for interoperable functionality in digital image repositories. The IIIF is comprised of and driven by a community of libraries, museums, archives, software companies, and other organizations working together to create, test, refine, implement and promote the IIIF specifications. Using JSON-LD, linked data, and standard W3C web protocols such as Web Annotation, IIIF makes it easy to parse and share digital image data, migrate across technology systems, and provide enhanced image access for scholars and researchers. In short, IIIF enables better, faster and cheaper image delivery. It lets you leverage interoperability and the fabric of the Web to access new possibilities and new users for your image-based resources, while reducing long term maintenance and technological lock in. IIIF gives users a rich set of baseline functionality for viewing, zooming, and assembling the best mix of resources and tools to view, compare, manipulate and work with images on the Web, an experience made portable–shareable, citable, and embeddable.		http://iiif.io/	http://iiif.io/apps-demos/					

Language-related studies

Name	Standards involved (separated by ";")	Research field(s) (the disciplinary scope of use of the standard)	TaDiRAH Activity(ies)	Overview	Working material	Bibliographical references	Tools, services and samples	Prioritizing of the resources	Creator / Developer of the standard	Last Update	License	Remarks and comments
TEI Lex0 (ENeL)	TEI	Language related studies	Interpretation/modeling	TEI has provided the lexicographic community with diverse alternatives for encoding different kinds of lexical resources. The flexibility that this de-facto standard ensures has engendered an explosion of the TEI schemes and consequently limited exchange and exploitation possibilities by the means of commun Natural Language Processing systems. We do not aim here to specify a mandatory format for the variety of dictionary content that we deal with, but define a baseline encoding (TEI-Lex-0) against which existing dictionaries can be compared, and which could serve as a target transformation format for generic querying or visualization tools. Aggregating such a baseline relies on the restriction of the use of TEI elements the refinement of their definitions, and if necessary, to remove any persistent ambiguity. The outcome of such a customization would be best practice guidelines accompanied by illustrative dictionary samples.	Lexicons	Ide, N. & Suderman, K. Lang Resources & Evaluation (2014) 48: 395. doi:10.1007/s10579-014-9268-1, The Linguistic Annotation Framework: a standard for annotation interchange and merging	https://github.com/ParthenosWP4/standardsLibrary/tree/master/Lexicography/ENeL-WG2		ENeL	March 2017	Both Creative Commons Attribution 3.0 Unported License and a BSD 2-Clause license.	
Long-term archival TEI	TEI	Language related studies	Interpretation/modeling	A lot of digital resources coming from Research Communities, at least from Humanities, are using TEI format. Considering the huge amount of work required to create theses resources, there is a need to think about their preservation in order to make them reusable in the future. There is a great diversity within TEI community, which represents also the different types of described objects they deal with. Anyway, they share a common way of encoding by using the TEI Guidelines both for documentation and definition of their corpora. Surely, it's a good practice, but that's not enough for the digital archival community which main goal is to ensure that the resource should be readable and understandable in the future, say more than 20 years, by someone who was not involved in the creation of this resource. To fulfil this objective, the data archivist requires to verify both technical coherence of the resource and its reusability which means that documentation, taken in an expanded meaning of the term, ensure that you don't need to find a "(TEI) Rosetta Stone" to decipher and understand it. Therefore, the idea is to identify some additional criteria, compared to those commonly used for scholarly research purposes, to reach the goal of long term preservation of TEI corpora in conjunction with the CINES (the French National Digital Archive service - https://www.cines.fr) which will preserve them.	Texts	TEI Guidelines http://www.tei-c.org/Guidelines ODD http://www.tei-c.org/Guidelines/Customization/odds.xml ROMA as a tool to create ODD http://www.tei-c.org/Roma TEI GitHub https://github.com/TEIC OAIS https://en.wikipedia.org/wiki/Open_Archival_Information_System Poster presented during TEI conference in ROMA (2013) http://digilab2.let.uniroma1.it/teiconf2013/program/posters/abstracts-posters#C146 CINES (Centre Informatique National de l'Enseignement Supérieur) General how to archive https://www.cines.fr/en/long-term-preservation/archive-at-cines/ File format https://www.cines.fr/en/long-term-preservation/expertises/file-format/			CINES, CNRS	?	Both Creative Commons Attribution 3.0 Unported License and a BSD 2-Clause license.	
Stand-off annotation	TEI	Language related studies	Enrichment/annotating	Stand-off annotation assumes that the source text in the corpus, ideally kept in an unannotated form and in read-only files, is the root of independent possibly multi-file system of data descriptions (each description focusing on a distinct aspect of the source data). The source text is typically accompanied by a level of primary segmentation, which may be the lowest-level XML layer of annotation. The other files form a possibly multi-leaved and multi-leveled hierarchy referencing either the level of primary segmentation, or higher order levels of description.	Annotations	Bański P., Przepiórkowski A. (2009). Stand-off TEI annotation: the case of the National Corpus of Polish. In Proceedings of the 3rd Linguistic Annotation Workshop (LAW III) at ACL-IJCNLP 2009, pp. 64–67, Singapore, 2009. Bański P. (2010). Why TEI stand-off annotation doesn't quite work: and why you might want to use it nevertheless. In Proceedings of Balisage: The Markup Conference, 2010. 10.4242/BalisageVol5.Banski01. Bański, P., Wójtowicz, B. (2010). The Open-Content Text Corpus project. In V. Arranz., L. van Eerten (eds.) Proceedings of the LREC workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management (LRSLM2010), 23 May 2010, Valletta, Malta, pp. 19–25. Available from http://www.lrec-conf.org/proceedings/lrec2010/workshops/W20.pdf. TEI Consortium (eds.) (2010). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.1.0. Last updated on 15th December 2016. http://www.tei-c.org/Guidelines/P5/			?	?	Both Creative Commons Attribution 3.0 Unported License and a BSD 2-Clause license.	Stand-off annotation is not really a standard. Should it be here? Perhaps expressed in another way?
LMF Diachrony	ISO-24613:2008.	Language related studies	Interpretation/modeling	The scope of this standard will cover the encoding of all lexical, conceptual and metadata relevant to born digital and retro-digitized etymological datasets. They are as follows: Etymological processes; Dating and sequence; Language information; Lexical forms; orthographic and phonetic Related forms: etymons, roots, cognates Grammatical information Semantic information Bibliographic information Notes: editors notes and other common miscellaneous content Level of confidence External references to ontological or other knowledge sources	Lexicons	Bowers, J., & Romary, L. (2016). Deep encoding of etymological information in TEI. Retrieved from https://hal.inria.fr/hal-01296498/ Salmon-Alt S., L. Romary, E. Buchi (2005). "Modeling Diachrony in Dictionaries". ACH-ALLC 2005, Vancouver, Canada. Salmon-Alt Susanne (2006) "Data structures for etymology: towards an etymological lexical network", BULAG 31 1-12 — http://hal.archives-ouvertes.fr/hal-00110971	https://github.com/anasfkhan81/LMFEty		Technical Committee:ISO/TC 37/SC 4 Language resource management	From 2008, last reviewed in 2012	https://www.iso.org/terms-conditions-licence-agreement.html	
MAF	ISO standard 24611:2012;	Language related studies	Enrichment/annotating	Morphosyntactic Annotation Framework (MAF), an ISO standard 24611:2012, is intended to provide a data model for morphosyntactic annotation of textual data, i.e. grammatical classes (part of speech, e.g. noun, adjective, verb), morphological structure of words and grammatical categories (e.g. number, gender, person). Rather than proposing a single tagset or a family of tagsets the standard offers a generic way to anchor, structure and organize annotations. The standard also describes an XML serialization for morphosyntactic annotations, with equivalences to the guidelines of the TEI (Text Encoding Initiative). Raw original document is accompanied by a set of annotations – word forms covering a set of tokens, identifying non-empty continuous parts of the document. The material corresponding to a token can be embedded inside a token or identified by a pair of document positions (e.g. character offsets, time durations for speech, frames for video etc.) Word forms correspond to tokens (in a many-to-many model), may embed word form subcomponents to represent compound terms and link output of tokenization to some lexicon. Word forms provide morphosyntactic information about a word (POS, lemma, morphology etc.) by means of specifying feature structures conformant to a tagset. Tagset data (types, features, feature values) may be mapped to data categories from ISOCat data category registry and feature structure declarations may be used to identify valid morphosyntactic content. Similarly, feature structure libraries may be used to name the most common morphosyntactic contents. Structural ambiguities are represented by lattices – direct acyclical graphs with single initial and terminal nodes. Lexical ambiguities can be handled by using alternations on word forms while morphological ambiguities by alternations inside feature structures.	Annotations	ISO 24611:2012. Language resource management – Morpho-syntactic annotation framework (MAF). Clémen L., de la Clergerie É. (2005). MAF: a morphosyntactic annotation framework. In Proceedings of the Second Language and Technology Conference, Poznań, Poland. Monachini, M., Calzolari N. (1994). Synopsis and Comparison of Morpho-syntactic Phenomena Encoded in Lexicon and Corpora. A Common Proposal and Applications to European Languages. Internal Document, EAGLES Lexicon Group, ILC, Università Pisa, October 1994. Przepiórkowski A., Bański P. (2011). Which XML standards for multilevel corpus annotation? In Z. Vetulani (ed.) Human Language Technology. Challenges for Computer Science and Linguistics: 4th Language and Technology Conference (LTC 2009), Poznań, Poland, November 6–8, 2009. Revised Selected Papers, vol. 6562 of Lecture Notes in Artificial Intelligence, pp. 400–411, Berlin, 2011. Springer Verlag.			Technical Committee : ISO/TC 37/SC 4 Language resource management	2012	https://www.iso.org/terms-conditions-licence-agreement.html	

SynAF	ISO standard 24615:2010	Language related studies	Enrichment/annotating	Syntax Annotation Framework (SynAF), a multipart ISO standard 24615:2010, is intended to represent the semantic annotation of textual data such as grammatical features, phrase structures and dependency structures. SynAF defines both a meta-model for syntactic annotation (graphs made of nodes and edges) and a set of data categories. Syntactic nodes are either terminal nodes equivalent to MAF word forms, annotated with syntactic data categories according to the word level, or non-terminal nodes annotated with syntactic categories from the phrasal, clausal and sentential level. Relations between syntactic nodes, such as dependency or constituency relations are represented with syntactic edges. Annotations can be applied to nodes and edges. The standard does not propose a specific tagset but only generic classes and specific data categories. Annotation vocabulary should be defined be means of a data category registry, e.g. ISOCat. Several possible serialization formats may be used such as TIGER-XML format or Graph Annotation Format defined in LAF.	Annotations	ISO 24615. Language resource management—Syntactic annotation framework (SynAF). Bunt H., Alexandersson J., Choe J.-W., Fang A. C., Hasida K, Petukhova V., Popescu-Belis A., Traum D. (2012). ISO 24617-2: A semantically-based standard for dialogue annotation. In Proceedings of the 8th International Conference on Language Resources and Evaluation, Istanbul, Turkey, pp. 430–437. European Language Resources Association (ELRA). Bunt H., Prasad R., Joshi A. (2012) First Steps Towards an ISO Standard for Annotating Discourse Relations. In Proceedings of the Joint ISA-7, SRSL-3, and I2MRT LREC 2012 Workshop on Semantic Annotation and the Integration and Interoperability of Multimodal Resources and Tools, Istanbul, Turkey, pp. 60–69. European Language Resources Association (ELRA). Declerck, T. (2006). SynAF: Towards a Standard for Syntactic Annotation. In Proceeedings of LREC 2006, pp. 229–232. European Language Resources Association (ELRA). Pustejovsky J., Lee K., Bunt H., Romary L. (2010). ISO-TimeML: An International Standard for Semantic annotation. In Proceedings of the 7th International Conference on Language Resources and Evaluation, Valletta, Malta, pp. 394–397. European Language Resources Association (ELRA). Romary, L., Zeldes A., Zipser F. (2015). <tiger2/> – Serialising the ISO SynAF Syntactic Object Model. Lang Resources & Evaluation 49: 1. doi:10.1007/s10579-014-9288-x. Stührenberg M. (2012). The TEI and Current Standards for Structuring Linguistic Data: An Overview. Journal of the Text Encoding Initiative (3), pp. 1–14. http://jtei.revues.org/523.			Technical Committee: ISO/TC 37/SC 4 Language resource management	201400,00%	https://www.iso.org/terms-conditions-licence-agreement.html	
CIDOC-CRM	ISO 21127:2014	Language related studies	Interpretation/modeling	CIDOC-CRM has been designed and is maintained by the International Committee for Documentation at ICOM - the International Council of Museums - to help Cultural Heritage Organizations develop adequate documentation. Started as an effort to create a general data model for museums, it eventually shifted from the Entity Relation model, used by traditional databases - to adopt an object oriented approach and become a Conceptual Reference Model enabling information interchange and integration also beyond the museum community. After a transition period (2000), it eventually became an official ISO Standard ISO 21120:2006, revised as ISO 21127:2014. The reason behind CIDOC-CRM is to provide compatibility to data and information produced by different institutions using different data models, workflows, and terminologies. Rather that trying to fix this gap by providing yet another set of custom transformation rules, or by oversimplificating the complexity of original data, concentrating on a limited sub set of 'core' descriptors, the CIDOC reference model aims to overcome these limitations by providing a semantic reference point which will enable Cultural Heritage Organizations to render their information resources mutually compatible without sacrificing detail and precision. The CIDOC-CRM is a standard for domain experts in cultural heritage and related domains, providing a common and extensible semantic framework, with definitions and a formal structure to describe the implicit and explicit concepts and relationships used in cultural heritage documentation, map and describe relevant information on cultural heritage objects, formulate requirements for information systems. In this way, it can provide the "semantic glue" needed to mediate between different sources of cultural heritage information participating in PARTHENOS. Together with the PARTHENOS Entity Model - an application profile of CIDOC-CRM developed to manage the descriptions of the PARTHENOS Entities (digital objects available in the PARTHENOS Dataspace as well as services available for the users via the PARTHENOS VREs) - CIDOC-CRM is the format used to encode all the data produced and managed by the project. FORTH developed a specific component - already integrated with the D4Science Platform - to manage and support the mapping process from specific formats (EAD,TEI etc.) to CIDOC and vice versa.	Ontologies	Data repository http://www.cidoc-crm.org/ http://old.cidoc-crm.org/ Bibliography (Zotero) http://lists.ics.forth.gr/pipermail/crm-sig/ Blog entries http://old.cidoc-crm.org/press.htm	Github Mapper X3ML https://github.com/delving/x3ml		http://www.cidoc-crm.org/	November 2011	https://www.iso.org/terms-conditions-licence-agreement.html	We think CIDOC-CRM should be considered a cross-disciplinary standard in PARTHENOS, and not related to language studies. Maybe, when we'll proceed to the writing of the chapters, we should think about a general section to be filled in with standards used in all the research domains (as CIDOC-CRM). (Roberta Giacomi and Maurizio Sanesi, SISMEL)
Multilingual Thesaurus Building		Language related studies	Interpretation/modeling	Information resources may be of very different kinds: books, chapters in books, papers in periodicals and conference volumes, newspapers, case records, data tables, graphs, images, maps, music sheets, etc. The contents may be in different languages. These resources may be available in their conventional physical document forms and/or in digital form. Directories, indexes, lists, catalogues and such other tools are used to know contents and retrieve information. KOTs (Knowledge Organising Tools) are useful to manage the vocabulary/terminology of these tools. The KOTs include ontologies, taxonomies, lexicons, dictionaries, schemes for subject classifications, thesauri, wordnets, semantic nets, self-organising systems, etc. These tools are useful in order to standardise and manage vocabularies in indexes. In a multilingual indexing thesaurus both the terms and the relationships are represented in more than one language. Since the drawing up of the Guidelines for the Establishment and Development of Multilingual Thesauri in 1976, the multilingual access to information has followed two main developments: the building of nonsymmetrical thesauri and the linking of two or more thesauri and/or controlled vocabularies.	Thesauri	- https://www.ifla.org/publications/ifla-professional-reports-115 - UNESCO, Guidelines for the Establishment and Development of Multilingual Thesauri, Paris 1976. - https://www.ifla.org/best-practice-for-national-bibliographic-agencies-in-a-digital-age/node/9041 - https://www.iso.org/standard/53657.html - IFLA Working Group on Guidelines for Multilingual Thesauri. 2009 Guidelines for multilingual thesauri. The Hague. International Federation of Library Associations and Institutions. Available at: http://archive.ifla.org/VII/s29/pubs/Profrep115.pdf (06 April 2017). - International Organization for Standardization 2011 ISO 25964-1:2011, information and documentation. Thesauri and interoperability with other vocabularies. Part 1: thesauri for information retrieval. Geneva. International Organization for Standardization. - International Organization for Standardization 2013 ISO 25964-2:2013, information and documentation. Thesauri and interoperability with other vocabularies. Part 2: interoperability with other vocabularies. Geneva. International Organization for Standardization			International Federation of Library Associations and Institutions (IFLA)	.	.	Is this a standard?
BackBone Thesaurus (BBT)		Language related studies	Interpretation/modeling	The aim of the BackBone Thesaurus is to develop a model and a proposal of how existing thesauri and ontologies will become interoperable and can be maintained in a sustainable and scalable way. This has been undertaken by the Thesaurus Maintenance WG which was established in 2014 in the framework of DARIAH EU: The Digital Research Infrastructure for the Arts and Humanities- a research infrastructure. This Research Infrastructure aims at enhancing and supporting digitally-enabled research and teaching across the arts and humanities. The Thesaurus Maintenance WG aims at designing and establishing a coherent overarching thesaurus for the humanities, a "backbone" or "metathesaurus", under which all the vocabularies and terminologies in use in the domain can be aligned. Therefore, this work focuses on identifying the top-level-concepts (facets and hierarchies) that will become its common basis, meeting the demands for intersubjective and interdisciplinary validity. The approach is nevertheless bottom-up – rather than by theoretical argument; top-level concepts are developed by adequate abstraction from existing terminological systems. This requires an effective methodology in order to decide, if a more generic concept has the power to effectively subsume enough narrower terms from different thesauri and to determine whether it is comprehensible enough in its definition to allow experts from different sub-disciplines to align their terms by themselves under these concepts. This alignment has the ambition to provide a comprehensive first-level integration of terminologies in DARIAH and possibly beyond, and to foster a shared good practice of terminology definition. One of the major advantages of this kind of classification is the potential of a sustainable and manageable expansion of the thesauri into new areas of knowledge, in which it continues to be effective and efficient, without forcing the experts to abandon their terminology. Furthermore, it enables collaboration, crossdisciplinary resource discovery, and detection of common principles and ensures compatibility with other thesauri that are restricted to particular areas of knowledge. Following this methodology, the Working Group decided to define an initial set of top-level concepts based on evidence from vocabularies the Group had so far access to. These concepts constitute a first operational draft, which on one side demonstrates the feasibility of the methods and illustrates it for didactic purposes, and on the other side allows its intended operation for terminology integration. This first draft will be adapted and extended as the integration of more terminologies will	Thesauri	Introduction to BBT http://83.212.168.219/DariahCrete/en/bbt_intro_en BBT documentation http://83.212.168.219/DariahCrete/en/documents BBT releases http://83.212.168.219/DariahCrete/en/bbt_releases	THEMAS (username:reader, no password) http://139.91.183.44:8090/THEMAS/			September 2016	CC BY NC SA	can this be characterized as a standard - or is it something under development?

TBX in TEi	TEI; ISO standard 320042 (TBX — TermBase eXchange) ; ISO standard 16642 (TMF — Terminology Markup Framework)	Language related studies	Interpretation/modeling	TEI offers a plethora of means for modeling lexical data, nevertheless those means are rooted in a semasiological approach, in which the lemma is the basis of the representation. Contrasting and complementing this view, an onomasiological approach puts the respective concepts of lexical units at its center, i.e. all synonymous words - and in particular spanning over various languages - as associated with their concept. Such models are the basis for thesauri, synonym dictionaries, and terminological dictionaries which are commonly used in translation work, language learning, and technical writing as well as in software environments that include indexing, documentary system, or machine translation capabilities. The present work is an adaptation of ISO standard 320042 (TBX — TermBase eXchange) and optimises the re-use of TEI constructs in combination with TBX elements. TBX is itself an application of ISO standard 16642 (TMF — Terminology Markup Framework) which provides a meta-model for the description of terminologies and other onomasiological structures. Historically, TMF has its roots in the TEI but following its fork was not able to profit from a large body of work done in the context of TEI and vice versa, the TEI lack a native model for conceptually structured lexical data. The present work is trying to bridge this gap.	Terminologies	Data repository: https://github.com/ParthenosWP4/standardsLibrary/tree/master/terminology/use_cases Bibliography (Zotero): https://www.zotero.org/groups/parthenos-wp4/items/collectionKey/5IQ9TPWS	Github: https://github.com/ParthenosWP4/standardsLibrary/tree/master/terminology		?	?	https://www.iso.org/terms-conditions-licence-agreement.html	
CMDI	ISO standard ISO-24622	Language related studies	Interpretation/modeling	Component Metadata Infrastructure (CMDI), an ISO standard ISO-24622, is one of the technical pillars of CLARIN's infrastructure. It features a (meta-)model to define/create and (re)use metadata schemas and at the same time a technical infrastructure to create and share these schemas as well as to create, collect and distribute actual resource descriptions (metadata records) adhering to (one of) these schemas. Thus CMDI is specifically: NOT one (single) format. There is a schema expressing the metamodel CMDI spec and there are currently around 200 profiles or schemas defined for different types of resources and different contexts. NOT a (single) tool. It is a set of software components forming an integrated technical infrastructure. The whole infrastructure is supported by a number of recommended components, guidelines and best practices, tools for validation and benchmarking, etc.	Metadata	CLARIN ERIC, Frequently Asked Questions - Metadata in CLARIN: basics, https://www.clarin.eu/faq-page/273 (last accessed April 2017). CMDI Task Force 2016, CMDI 1.2 specification, https://office.clarin.eu/v/CE-2016-0880-CMDI_12_specification.pdf (last accessed April 2017). Goosen, T, Windhouwer, M, Ohren, O, Herold, A, Eckart, T, Durco, M & Schonefeld, O 2015, CMDI 1.2: Improvements in the CLARIN Component Metadata Infrastructure. in J Odijk (ed.), Selected Papers from the CLARIN 2014 Conference, October 24-25, 2014, Soesterberg, The Netherlands., 116:004, Linköping Electronic Conference Proceedings, Linköping University Electronic Press, Linköpings universitet, Linköping, pp. 36-53, https://pure.knaw.nl/portal/en/publications/cmdi-12-improvements-in-the-clarin-component-metadata-infrastructure%2891536b93-31cb-4f4a-8125-56f4fe0a1881%29.html (last accessed April 2017). Wittenburg, P, van Uytvanck, D 2012: The Component Metadata Initiative (CMDI), in: CLARIN-D AP 5, CLARIN-D User Guide, https://media.dwds.de/clarin/userguide/text/metadata_CMDI.xhtml (last accessed April 2017).	CMDI toolkit		CLARIN	July 2016	The specification is in some ways the core of the standard and is distributed as CC BY-SA (https://www.clarin.eu/cmdi1.2-specification). The CMDI toolkit is the implementation and is GPLv3 (https://github.com/clarin-eric/cmdi-toolkit)	
Syntax Annotation Framework (SynAF)	multipart ISO standard 24615:2010; TEI; TIGER-XML; Graph Annotation Format	Language related studies	Enrichment/annotating	Syntax Annotation Framework (SynAF), a multipart ISO standard 24615:2010, is intended to represent the semantic annotation of textual data such as grammatical features, phrase structures and dependency structures. SynAF defines both a meta-model for syntactic annotation (graphs made of nodes and edges) and a set of data categories. Syntactic nodes are either terminal nodes equivalent to MAF word forms, annotated with syntactic data categories according to the word level, or non-terminal nodes annotated with syntactic categories from the phrasal, clausal and sentential level. Relations between syntactic nodes, such as dependency or constituency relations are represented with syntactic edges. Annotations can be applied to nodes and edges. The standard does not propose a specific tagset but only generic classes and specific data categories. Annotation vocabulary should be defined be means of a data category registry, e.g. ISOCat. Several possible serialization formats may be used such as TIGER-XML format or Graph Annotation Format defined in LAF.	Annotations	ISO 24615. Language resource management—Syntactic annotation framework (SynAF). Bunt H., Alexandersson J., Choe J.-W., Fang A. C., Hasida K, Petukhova V., Popescu-Belis A., Traum D. (2012). ISO 24617-2: A semantically-based standard for dialogue annotation. In Proceedings of the 8th International Conference on Language Resources and Evaluation, Istanbul, Turkey, pp. 430–437. European Language Resources Association (ELRA). Bunt H., Prasad R., Joshi A. (2012) First Steps Towards an ISO Standard for Annotating Discourse Relations. In Proceedings of the Joint ISA-7, SRSL-3, and I2MRT LREC 2012 Workshop on Semantic Annotation and the Integration and Interoperability of Multimodal Resources and Tools, Istanbul, Turkey, pp. 60–69. European Language Resources Association (ELRA). Declerck, T. (2006). SynAF: Towards a Standard for Syntactic Annotation. In Proceedings of LREC 2006, pp. 229–232. European Language Resources Association (ELRA). Pustejovsky J., Lee K., Bunt H., Romary L. (2010). ISO-TimeML: An International Standard for Semantic annotation. In Proceedings of the 7th International Conference on Language Resources and Evaluation, Valletta, Malta, pp. 394–397. European Language Resources Association (ELRA). Romary, L., Zeldes A., Zipser F. (2015). <tiger2/> – Serialising the ISO SynAF Syntactic Object Model. Lang Resources & Evaluation 49: 1. doi:10.1007/s10579-014-9288-x. Stührenberg M. (2012). The TEI and Current Standards for Structuring Linguistic Data: An Overview. Journal of the Text Encoding Initiative (3), pp. 1–14. http://jtei.revues.org/523.			Technical Committee: ISO/TC 37/SC 4 Language resource management	2014		
Linguistic Annotation Framework (LAF)	ISO 24612	Language related studies	Enrichment/annotating	ISO 24612:2012 specifies a linguistic annotation framework (LAF) for representing linguistic annotations of language data such as corpora, speech signal and video. The framework includes an abstract data model and an XML serialization of that model for representing annotations of primary data. The serialization serves as a pivot format to allow annotations expressed in one representation format to be mapped onto another.	Annotations				Technical Committee: ISO/TC 37/SC 4 Language resource management	2012	https://www.iso.org/terms-conditions-licence-agreement.html	
RDF	RDF Schema; RDF XML Syntax	Language related studies	Interpretation/modeling	RDF Schema: RDF Schema is a semantic extension of RDF. It provides mechanisms for describing groups of related resources and the relationships between these resources. RDF Schema is written in RDF using the terms described in this document. These resources are used to determine characteristics of other resources, such as the domainsand ranges of properties. "The Resource Description Framework (RDF) is a general-purpose language for representing information in the Web. RDF XML Syntax:This standard defines an XML syntax for RDF called RDF/XML in terms of Namespaces in XML, the XML Information Set and XML Base. The formal grammar for the syntax is annotated with actions generating triples of the RDF graph as defined in RDF Concepts and Abstract Syntax. The triples are written using the N-Triples RDF graph serializing format which enables more precise recording of the mapping in a machine processable form. The mappings are recorded as tests cases, gathered and published in RDF Test Cases."	Texts, ontologies, annotations				RDF working group. https://www.w3.org/standards/techs/rdf#w3c_all	2014		
Language codes	ISO 639-1 and 639-3	Language related studies	Enrichment/annotating	ISO 639-1 contains two letter codes, one per language for each ISO macrolanguage. ISO 639-3 contains 3 letter codes, and they are distinct codes for each variety of an ISO macrolanguage	metadata, annotations				Technical Committee: ISO/TC 37/SC 2 Terminographical and lexicographical working methods	639-1: 2002. 639-3:2007.	https://www.iso.org/terms-conditions-licence-agreement.html	
Unicode	The Unicode Standard, Version 9.0.0	Language related studies	Capture/conversion	The Unicode Standard is a character coding system designed to support the worldwide interchange, processing, and display of the written texts of the diverse languages and technical disciplines of the modern world. In addition, it supports classical and historical texts of many written languages.	Texts				The Unicode Consortium	2016	http://www.unicode.org/copyright.html	
Dublin Core	Metadata Element Set	Language related studies	Interpretation/modeling	The Dublin Core Metadata Element Set is a vocabulary of fifteen properties for use in resource description.	texts, annotations				The Dublin Core Metadata Initiative (DCMI)	2012	DCMI documents are licensed under a Creative Commons Attribution 3.0 Unported License	
Eurovoc	EU's Multilingual Thesaurus	Language related studies	Interpretation/modeling	EuroVoc is a multilingual, multidisciplinary thesaurus covering the activities of the EU. It contains terms in 23 EU languages (Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish and Swedish), plus in three languages of countries which are candidate for EU accession: македонски (mk), shqip (sq) and српски (sr).					EU	2016	http://open-data.europa.eu/kos/licence/EuropeanCommission	

OLAC	Metadata annotation	Language related studies	enrichment/annotating	Open Language Archives Community Metadata (OLAC) was developed by Open Archives Intiative (OAI) and is useful for representing specific metadata of language resources and NLP tools for harvesting metadata in archives. OLAC is implemented in XML and can be used for the interchange of metadata descriptions among archives.	texts				OLAC	2008		
Topic Maps	ISO/IEC 13250 Knowledge representation	Language related studies	Interpretation/modeling	opic Maps is an ISO Standard and describes the information management and interchange. It allows to build the abstract data model for knowledge collection, to connect them to relevant information resources and to apply the relation between the encoded knowledges. Topic Maps is suitable for knowledge representation in wide range of domains such as persons, locations, things, events etc.					ISO	A multipart standard with several updates		
Guidelines on the Production and Preservation of Digital Audio Objects	production and preservation of audio	Language related studies			Speech and sound	IASA Technical Committee, Guidelines on the Production and Preservation of Digital Audio Objects, ed. by Kevin Bradley., Second edition 2009. (= Standards, Recommended Practices and Strategies, IASA-TC 04). www.iasa-web.org/tc04/audio-preservation			IASA			
OWL	Knowledge representation	Language related studies	Interpretation/modeling	RDF-based standard for specifying ontologies which are compatible with the World Wide Web					W3C OWL working group	2012	https://www.w3.org/Consortium/Legal/2015/copyright-software-and-document	
IMDI	Metadata annotation	Language related studies	enrichment/annotating	- The IMDI metadatas set are particularly suitable for annotating multi modal corpora. It is very comprehensive and conceptual, and includes other widely known metadata sets like Dublin Core or OLAC.	multi-modal corpora	D. Broeder and P. Wittenburg, "The IMDI metadata framework, its current application and future direction," International Journal of Metadata, Semantics and Ontologies, vol. 1, no. 2, pp. 119–32, 2006. P. Withers, "Metadata Management with Arbil", in Proceedings of the Eight International Conference on Language Resources and Evaluation, Istanbul, Turkey, 2012.				2010		

Social Science

Name	Standards involved (separated by ";")	Research field(s) (the disciplinary scope of use of the standard)	TaDiRAH Activity(ies)	Overview	Working material	Bibliographical references	Tools, services and samples	Prioritizing of the resources	Creator / Developer of the standard	Last Update	License	Remarks and comments
Data Documentation Initiative (DDI)		Social sciences	- meta activities - community building - creation - programming - dissemination - sharing - storage -preservation	The Data Documentation Initiative (DDI) is an international standard for describing the data produced by surveys and other observational methods in the social, behavioral, economic, and health sciences. DDI is a free standard that can document and manage different stages in the research data lifecycle, such as conceptualization, collection, processing, distribution, discovery, and archiving.	social and behavioral science data	- DDI Working Paper Series (ISSN 2153-8247) http://www.ddialliance.org/publications/working-papers - http://www.ddialliance.org/publications/formal-papers - http://www.ddialliance.org/publications/conferences-list	https://www.ddialliance.org/resources/tools		DDI Alliance	March 2014 (DDI 3.2)	GNU Lesser General Public License (DDI 3.2 schema) / Creative Commons licenses (other DDI documents)	
Metadata Encoding and Transmission Standard (METS)		Social sciences	- meta activities - community building - dissemination - sharing - storage - preservation	The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the XML schema language of the World Wide Web Consortium.	metadata	http://www.loc.gov/standards/mets/news100306.html	http://www.loc.gov/standards/mets/mets-tools.html		Digital Library Federation (DLF)		Creative Commons CC0 1.0 Universal Public Domain Dedication	
Text Encoding Initiative (TEI)		Social sciences	- Dissemination - Sharing - Storage - Preservation	The TEI is a standard for encoding machine-readable texts in the humanities and social sciences.	transcriptions of speech	http://www.tei-c.org/Support/Learn/tei_bibliography.xml	http://www.tei-c.org/Tools/			December 2016	Creative Commons Attribution 3.0 Unported License and a BSD 2-Clause license	
Dublin Core	ISO 15836:2009	Social sciences	- Storage - Preservation	The Dublin Core metadata standard is a simple yet effective element set for describing a wide range of networked resources. The Dublin Core standard includes two levels: Simple and Qualified.	metadata	http://dublincore.org/documents/2000/07/16/usageguide/references.shtml			Dublin Core Metadata Initiative		Creative Commons Attribution 4.0 International License (CC BY 4.0)	
International Standard Classification of Occupations (ISCO)	ISCO-08	Social sciences	- Interpretation - Contextualizing	The ISCO is a classification structure for organizing jobs into a clearly defined set of groups according to the tasks and duties undertaken in the job. It is intended for use in statistical applications and in a variety of client oriented applications. Client oriented applications include the matching of job seekers with job vacancies, the management of short or long term migration of workers between countries and the development of vocational training programmes and guidance. The ISCO is the basis for many national occupation classifications as well as applications in specific domains such as reporting of teaching, agricultural and healthcare workforce information.	information on labour and jobs	http://www.ilo.org/public/english/bureau/stat/isco/docs/publication08.pdf			International Labour Organization (ILO)	2008		
Statistical Data and Metadata eXchange (SDMX) https://sdmx.org/	ISO 17369:2013 (https://www.iso.org/standard/52500.html)	Social sciences		SDMX, which stands for Statistical Data and Metadata eXchange, is an ISO standard designed to describe statistical data and metadata, normalise their exchange, and improve their efficient sharing across statistical and similar organisations. It provides an integrated approach to facilitating statistical data and metadata exchange, enabling interoperable implementations within and between systems concerned with the exchange, reporting and dissemination of statistical data and their related meta-information. It consists of: technical standards (including the Information Model) statistical guidelines an IT architecture and tools (https://sdmx.org/?page_id=2555/)	Statistics metadata	https://sdmx.org/?page_id=5008	https://sdmx.org/?page_id=4500		International initiative sponsored by BIS (Bank for International Settlements), ECB (European Central Bank), EUROSTAT (Statistical Office of the European Union), IMF (International Monetary Fund), OECD (Organisation for Economic Co-operation and Development), UN (United Nations), and the World Bank	2013		
DataCite Metadata Schema https://schema.datacite.org/		Social sciences		The DataCite Metadata Schema is a list of core metadata properties chosen for an accurate and consistent identification of a resource for citation and retrieval purposes, along with recommended use instructions. (https://schema.datacite.org/)	Bibliographic data	https://schema.datacite.org/meta/kernel-4.0/	https://schema.datacite.org/meta/kernel-4.0/		DataCite (https://www.datacite.org/)	2016		
International Standard Classification of Education (ISCED) http://uis.unesco.org/en/topic/international-standard-classification-education-isced		Social sciences	- Interpretation - Contextualizing	The International Standard Classification of Education (ISCED 2011) provides a comprehensive framework for organising education programmes and qualification by applying uniform and internationally agreed definitions to facilitate comparisons of education systems across countries. (http://uis.unesco.org/en/topic/international-standard-classification-education-isced)		http://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf			UNESCO	2013		
da|ra Metadata Schema (https://www.da-ra.de/)		Social sciences	- Storage - Preservation - Dissemination	da|ra operates as the registration agency for social science and economic data jointly run by GESIS (http://www.gesis.org) and ZBW (http://www.zbw.eu). da|ra pursues the goal of long-term, persistent identification and availability of research data via allocation of DOI names. (https://www.da-ra.de/fileadmin/media/da-ra.de/PDFs/TechnicalReport_2014-17.pdf)	Research data	https://www.da-ra.de/fileadmin/media/da-ra.de/PDFs/TechnicalReport_2014-17.pdf	https://datorium.gesis.org/xmlui/		gesis. Leibniz Institute for the Social Sciences (http://www.gesis.org)	2014		Not sure, to which extent it is used outside Germany.
Market, opinion and social research -- Vocabulary and service requirements	ISO 20252:2012	Social sciences	- Interpretation - Contextualizing	This International Standard establishes terms and definitions and service requirements for organizations and professionals conducting market, opinion and social research. (https://www.iso.org/obp/ui/#iso:std:iso:20252:ed-2:v1:en)		https://www.iso.org/standard/53439.html			ISO/TC 225	2012		
SPSS Portable		Social sciences	- Analysis - Interpretation	This is a tool by IBM that is widely used in social studies (sociology, political studies). It is kind of a de-facto standard, as the usage of SPSS is often part of curricula	Statistics	https://www.ibm.com/analytics/de/de/technology/spss/			IBM	in active development	proprietary format	This is a commercial tool and it it produces a proprietary file format. But it is widely used and some repositories in social studies accept it for deposition.
R		Social sciences	- Analysis - Interpretation	R is a language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible.		- https://cran.r-project.org/manuals.html - https://journal.r-project.org/ - https://www.r-project.org/doc/bib/R-books.html			Initially written by Robert Gentleman and Ross Ihaka of the Statistics Department of the University of Auckland, the current R is the result of a collaborative effort with contributions from all over the world.	June 2017	GNU General Public License	Free software

PSPP		Social sciences	- Analysis - Interpretation	PSPP is a free software application for analysis of sampled data, intended as a free alternative for SPSS. It can perform descriptive statistics, T-tests, anova, linear and logistic regression, measures of association, cluster analysis, reliability and factor analysis, non-parametric tests and more. Its backend is designed to perform its analyses as fast as possible, regardless of the size of the input data.You can use PSPP with its graphical interface or the more traditional syntax commands.		https://www.gnu.org/software/pspp/manual/			GNU Project	July 2016	GNU General Public License	Free software
JASP		Social sciences	- Analysis - Interpretation	JASP is a free and open-source graphical program for statistical analysis, designed to be easy to use, and familiar to users of SPSS. Additionally, JASP provides many Bayesian statistical methods.			https://github.com/jasp-stats/jasp-desktop		JASP team	May 2017	GNU Affero General Public License, Version 3	
STATA		Social sciences	- Analysis - Interpretation	Stata is a complete, integrated statistical software package that provides everything you need for data analysis, data management, and graphics. (http://www.stata.com/why-use-stata/)	Statistics	http://www.stata.com/			StataCorp	in active development	proprietary format	This is a commercial tool and it it produces a proprietary file format. But it is widely used and some repositories in social studies accept it for deposition.

4. Annex: Abbreviations

API	Application Programming Interface
CC0	Creative Commons “No Rights Reserved”
CCS	CENDARI Collection Schema
CESSDA	Consortium of European Social Science Data Archives
CHO	Cultural Heritage Objects
CIDOC-CRM	CIDOC Conceptual Reference Model
CLARIN	Common Language Resources and Technology Infrastructure
CMDI	Component Metadata Infrastructure
CSV	Comma-separated Values
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DC	Dublin Core
DDI	Data Documentation Initiative
DEP	The user may distribute derivative works via CLARIN (in CLARIN licence)
DM2E	Data Model 2 Europeana
DMP	Data Management Plan
EAD	Encoded Archival Description
EAG	Encoded Archival Guide
EDM	Europeana Data Model
EGI	European Grid Infrastructure

ESE	Europeana Semantic Elements
EUDAT	European Data Infrastructure
FAIR	findable, accessible, interoperable, reusable
FRBR	Functional Requirements for Bibliographic Records
IASA	International Association of Sound and Audiovisual Archives
ICIP	Indigenous Cultural and Intellectual Property
IETF	Internet Engineering Task Force
IFLA	International Federation of Library Associations and Institutions
IIIF	International Image Interoperability Framework
ILO	International Labour Organization
IMDI	ISLE Meta Data Initiative
IPR	Intellectual Property Rights
ISBD	International Standard Bibliographic Description
ISCED	International Standard Classification of Education
ISCO	International Standard Classification of Occupations
ISO	International Organization for Standardization
JASP	Jeffreys's Amazing Statistics Program
JSON	JavaScript Object Notation
LIDO	Lightweight Information Describing Objects
LOD	Linked open data
LRT	Language Resources and Technology

METS	Metadata Encoding & Transmission Standard
MODS	Metadata Object Description Schema
NORED	The user is not permitted to redistribute the resource (in CLARIN Licence)
OA	Open Access
OAI-ORE	Open Archives Initiative Object Reuse and Exchange
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
ODbL	Open Database License
ODC	Open Data Commons
OCLC	Online Computer Library Center
OLAC	Open Language Archives Community
OWL	Web Ontology Language
PARTHENOS	Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies
PDDL	Public Domain Dedication and License
PDF	Portable Document Format
PSI	Public Sector Information
RDF	Resource Description Framework
SKOS	Simple Knowledge Organization System
SPSS	Statistical Package for the Social Sciences
SSK	Standardization Survival Kit
TEI	Text Encoding Initiative
TK	Traditional Knowledge

XML	Extensible Markup Language
W3C	World Wide Web Consortium