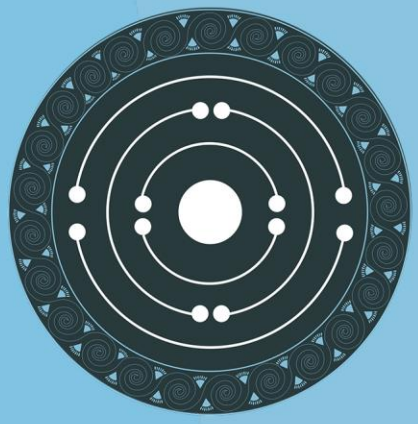
Standardization survival kit (Draft)

Laurent Romary, Emiliano Degl'innocenti, Klaus Illmayer, Adeline Joffres, Emilie Kraikamp, Nicolas Larrousse, Maciej Ogrodniczuk, Marie Puren, Charles Riondet, Dorian Seillier



PARTHENOS

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

STANDARDIZATION SURVIVAL KIT

INRIA (overall coordination)

PIN

CLARIN

KNAW

CNR

CNRS

CSIC

FORTH

OEAW

MIBACT-ICCU

SISMEL

AA

FHP

With contributions from all PARTHENOS partners

31 October 2016

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

STANDARDIZATION SURVIVAL KIT

Deliverable Number	D4.1
Dissemination Level	Public
Delivery date	31 OCTOBER 2016
Status	Draft
Main Author(s)	Laurent Romary (INRIA) Emiliano Degl'Innocenti (CNR-OVI) Klaus Illmayer (OEAW) Adeline Joffres (CNRS) Emilie Kraikamp (DANS-KNAW) Nicolas Larrousse (CNRS) Maciej Ogrodniczuk (CLARIN) Marie Puren (INRIA) Charles Riondet (INRIA) Dorian Seillier (INRIA) 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	D4.1, Standardization Survival Kit
Document title	PARTHENOS-D4.1- Standardization Survival Kit
Author(s)	Laurent Romary (INRIA, overall coordination) Emiliano Degl'Innocenti (CNR-OVI) Klaus Illmayer (OEAW) Adeline Joffres (CNRS) Nicolas Larrousse (CNRS) Maciej Ogrodniczuk (CLARIN) Marie Puren (NRIA) Charles Riondet (INRIA) Dorian Seillier (INRIA) With contributions from all PARTHENOS partners
Dissemination level/distribution	Public

Document History

Version/date	Changes/approval	Author/Approved by
V 0.1 10 October 2016	First draft of chapters	Laurent Romary (INRIA) and collaborators
V 0.2 19 October 2016	Revised complete draft of all chapters	Laurent Romary (INRIA) and collaborators
V 0.3 21 October 2016	Second revised draft with English corrections by Piotr Banski and Reinier de Valk	Laurent Romary (INRIA) and collaborators
V 0.4 24 October 2016	First version submitted to PIN	Laurent Romary (INRIA) and collaborators

V .5 26 October 2016	Reviewed and revised for English	Sheena Bassett (PIN)
Final 31 October 2016	Final version submitted to the European Commission	Laurent Romary (INRIA) and collaborators

Table of content

Executive Summary

The present report provides an overview of the various components of 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 report is organized around five main components:

- A series of use cases, elaborated and complemented from the ones designed within WP2, and articulating the role of standards in concrete research scenarios. New use cases have been identified when they correspond to strong needs in the research community or have the potential for swift development when the standardization agenda has reached a high level of maturity;
- Specific standards related resources (schemas, transforms, documentation) that have been compiled or elaborated within WP4 and that serve as background content for the SSK;
- The actual design concept for the SSK that intends to provide a single entry point for scholars, whether novice or advanced, in the domain of digital methods, so that he/she can have quick and precise access to the information needed for managing digital content or applying the appropriate method in his/her scholarly context;
- A specific awareness package articulated around the “why standards?” leaflet to make scholars understand the essential role of standardized methods and content for the reusability of research assets.
- A set of tools (demonstrator) accessible through a specific VRE, implementing the domain-driven scenarios provided by the research communities represented in PARTHENOS.

The report anticipates a transversal analysis of the various scenarios to identify meta standardization domains around generic objects types in particular (text, 3D, archival information, audio/video) for which we feel it is necessary to provide a comprehensive guidance in the domain of standards. This will serve as a basis for deliverable D4.2, and will be accompanied by a series of focused standardization activities in the second phase of the PARTHENOS project.

Note: a lot of the material presented in this version of the report will be included in the online version of the SSK thus making the next version of the deliverable likely to be a brief update on this one.

1 Introduction: why standards at all?

The various infrastructures represented in PARTHENOS (CLARIN, DARIAH, EHRI and the forthcoming E-RIHS[1]) are all committed to advancing the digital revolution that has captured the Arts and Humanities. As more legacy primary and secondary sources become digital, more digital content is being produced[2] and more digital tools are being deployed, we see emerging a next generation of digitally aware scholars in the Humanities[3]. The role of our infrastructures is to connect these resources, tools and scholars, ensuring that the state-of-the-art in research is sustained and integrated across European countries.

What we need is a data-centred strategy, as echoed in various reports and statements that have been issued recently, in particular "Riding the wave"[4], which has placed the management of scientific data very high on the EU commission's agenda. This report stresses the importance of a long-term strategy concerning the management of scholarly data in all disciplines, which comprises both technical aspects (identification, preservation), editorial aspects (curation, standards) and sociological aspects (openness, scholarly recognition).

In the context of the present report, it is important to understand the actual role that proper data modelling and standards play in making digital content sustainable. Even if it does not seem obvious at first sight that the Arts and Humanities would be fit for taking up the technological prerequisites of standardization, the work carried out in WP4 of the PARTHENOS project intends to show that we can and should integrate standardization issues at the core of our infrastructural work. This work may contribute in turn to a wider understanding of the role of scholars within a digital infrastructure and consequently on how infrastructures could better integrate a variety of research communities in the Arts and Humanities.

[1] http://www.e-rihs.eu/
[2] Martin Hilbert *How much information is there in the "information society?"* Significance, 9 (4) 2012, 8–12.
[3] Bruno Latour **How Better to Register the Agency of Things: Ontology**, 2014. http://www.bruno-latour.fr/node/563
[4] **Riding the wave. How Europe can gain from the rising tide of scientific data**, report of the High Level Expert Group of Scientific Data. A submission to the European Commission, October 2010. http://ec.europa.eu/information_society/newsroom/cf/itemlongdetail.cfm?item_id=6204

We should all contribute to excellence in research by being seminal in the establishment of a large coverage, coherent and accessible data space for the Humanities. Whether acting at the level of standards, education or core information technologies (IT) services, we should keep this vision in mind when setting priorities as to what will impact the sustainability of the future digital ecology of scholars. Above all, such a strategy should directly influence the way we will advocate towards funding or supporting institutions, and also how we will manage our collaboration schemes with other initiatives in Europe and worldwide.

1.1 What is a standard?

The main issue in defining a policy about standards is to understand what they actually are. Standards usually take the form of documents informing about practices, protocols, artefact characteristics or data formats that can be used as reference for two parties working in the same field of activity to be able to produce comparable (or interoperable) results. This will also foster innovative, cross-disciplinary research paths, and eventually contribute to bridge the gap between the different cultures that are represented in the wide landscape of the Humanities and Cultural Heritage studies.

Standards are usually published by standardization organisations (such as ISO, W3C or the TEI Consortium) which ensure that the following three requirements for standards are fulfilled:

- Expression of a consensus: the standard should reflect the expertise of a wide (possibly international) group of experts in the field,
- Publication: the standard should be accessible to anyone who wants to know its content,
- Maintenance: the standard is updated, replaced or deprecated depending on the evolution of the corresponding technical field.

Standards are not regulations. There is no obligation to follow them except when one actually wants to produce results that can be compared with those of a wider community. This is why a standardization policy for an infrastructure in the Arts and Humanities should include recommendations as to what attitude the scholarly communities could or should adopt with regard to specific standards.

The preceding characteristics outlined for standards put a strong emphasis on the role of communities of practice and the corresponding bodies that represent them. Ideally, a good standard reflects the work of the relevant community and is maintained by the appropriate body. This is exactly the case of the TEI (Text Encoding Initiative) with respect to text representation standards and, to a lesser extent, of EAD (Encoded Archival Description), whose maintenance is undertaken by the Library of Congress with support of the Society of American Archivists.

Because there is no obligation to use a given standard, it is essential to provide potential users with a) awareness about the appropriate standards and the interest in adopting them, and b) the cognitive tools to help them identify the optimal use of standards through selection and possibly customisation of a reference portfolio. The experience gained within the various communities and infrastructure represented in PARTHENOS that have been in the need of adopting existing standards, is that there is always an initial phase during which scholars should be made aware of some core standards that are systematically related to the definition of interoperable digital objects. This is basically what has led us to identify the notion of the Standardization Survival Kit (SSK). In Table 1, for instance, we can see a first group of standards without which it is more or less impossible to deal with digital content in a correct manner.

ISO 639 series	Codes for the representation of languages and language families
ISO 15924	Codes for the representation of scripts
ISO 3166	Codes for the representation of country names
IETF BCP 47	Standard for encoding linguistic content, combining ISO 639, ISO 15924 and ISO 3166
ISO 10646, Unicode	Universal encoding of characters
ISO 8601	Representation of dates and times
XML recommendation	Provides the basic technical concept related to XML documents

Outline of a Standardization Survival Kit

As we shall see later in this document, the concept of the SSK goes far beyond these baseline examples and now aims at covering reference digital scenarios in the Arts and Humanities; in this respect we will capitalize on the work done in WP2, to collect and represent the needs of the scholarly communities represented in PARTHENOS. It is also clear that identifying or even documenting these needs is not enough. We need to carry out concrete action, and in particular contribute to educational activities targeted at researchers as planned in collaboration with WP7. We should also act at a more political level by interacting with funding agencies to make them aware that they could make strong recommendations concerning the systematic use of standards in scholarly work.

An important aspect in this dissemination strategy is that elementary projects should be encouraged to refrain from defining their own local formats, and to first demonstrate that their needs are not covered by the wide variety of already existing initiatives in the Digital Humanities landscape. The role of WP4 is thus seen as pro-active in helping communities to participate in standardization activities where they exist. Such a strategy will also contribute to the actual stabilization of existing conceptual and technical knowledge within ongoing projects, and provide a channel for the wider dissemination of the corresponding results.

1.2 The life cycle of standards

The development of standards follows most of the time a three-stage process:
During a preparation phase experts identify a need and gather possible evidence concerning existing practices and/or collect pre-normative documents, for instance when some community reports could be taken as a background for the corresponding work;
When there is a clear standardization project at hand and depending on the actual standardization organisation that takes this under its auspices, the standard is being **elaborated** by a) gathering enough/relevant/significant experts in the corresponding field and b) iterating on the definition of a draft document until a consensus is found that leads to the publication of the standard;
Finally, once the standard has been published, its adoption requires specific efforts to reach out towards community but also to identify and provide the necessary accompanying material that may help end users implement it.

We can mention here the specific case of the TEI Consortium which issues two releases of its **Guidelines** per year. This means that even if the guidelines as such are a strong basis for the representation of various types of textual objects in the Humanities, the identification of new use cases, or ones that are incompletely covered by the current guidelines, offer the possibility to constantly improve on their development. We present a typical example of this in the domain of lexical resources or speech transcription (cf . 2.3.2.2)

As we shall see in this report, the various activities falling in the scope of WP4 of the PARTHENOS project correspond to different stages of the standardization process, ranging from well-established standards to basic concepts that may develop as such.

1.3 Methodological aspect

The work carried out for the SSK covers three types of activities related to the deployment and use of standards in the Humanities and Cultural Heritage scholarship:

- **Documenting** existing standards to provide reference material for scholars who want to know more about their role and content. This relates to the specific provision of bibliographic sources, available documentation, technical introductions to the use of specific standards but also to the provision of prototypical examples which can serve as models for similar works;
- **Supporting** the actual adoption of standards by identifying how they relate to research scenarios and gathering the essential materials for their deployment (e.g. schemas);
- **Communicating** to research communities so that they can be made aware of both the need to apply standards in their digital scholarly practices and also be informed of the essential standards for their own fields.

These principles are reflected in the work carried out in WP4 of PARTHENOS so far, with, on the one hand, the elaboration of the “Why standards?” , accompanied by a simple awareness-raising cartoon and, on the other hand, the design of a comprehensive interface (front-end) to guide scholars through all available resources, on the basis of the reference scenarios that we have identified since the beginning of the project. This is the basis of the SSK as described further in this report.

In this context, the present deliverable provides:

- an overview of the research-oriented scenarios with the identification of the role of standards at this stage;
- the technical settings and the preliminary content of the infrastructure that we have designed for the SSK;
- the "Why standards" and the graphical elements of the strip cartoon.

The work carried out for the SSK, and in particular the precise identification of necessary standards at each stage of the research scenarios, will also be the basis for the preparation of D4.2 where the objective is to provide an overview of important standardization domains where PARTHENOS could optimally contribute to in developing or improving the existing landscape.

2 Documenting

2.1 Overview

2.1.1 Understanding researcher's needs to support the use of standards in Arts and Humanities

Standardization is generally not well perceived among researchers in the Arts and Humanities. It seems to impose time-consuming rules and is difficult to apply successfully in practice, especially by researchers often lacking the technical abilities to implement the standards. Besides these practical obstacles, a part of research in Arts and Humanities is ideologically reluctant to develop standards, their research being characterized by their ability to question societal norms and conventions[5]. Nevertheless, as expressed in the Grant Agreement, *"[...] to* **understand each other, people — including researchers — rely on tacit knowledge and agreement on the meaning of expressions, what is of course not possible if they want to share computer data**[6]". With the growing amount of digital data in the Arts and Humanities, standardization becomes a necessity for researchers, who have to be made aware of the direct benefits arising from this new framework.

Good practices for standardization have indeed to be widely spread among the research communities, before being fully adopted by researchers. Creating "good data" is in fact a key point for a standardized workflow: an accurate use of standards relies on good data modeling at the very beginning of a research project. With this goal in mind, a complete state-of-the-art on transformation stylesheets — from closed and proprietary formats towards XML or other structured formats — could be made available, for example. Developing research scenarios where computer data is fully re-usable for further experiments could also illustrate the high value of well-modelled data.

However, technical abilities vary widely between researchers, depending on their skills and training in computer science. These disparities seem the main sticking point preventing dissemination of good practices for standardization. The Standardization Survival Kit or "SSK" should allow this hindrance to be overcome. In the Grant Agreement, it is defined as

[5] "Standards for Reporting on Humanities-Oriented Research in AERA Publications: **American Educational Research Association**", **Educational Researcher**, August 2009, 38, p.482. DOI:10.3102/0013189X09341833.
[6] European Commission, **Grant Agreement: 654119 - Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies (PARTHENOS), Amendment Reference No AMD-654119**, 25 November 2015, Part B, p.26.

"[...] a minimum standardization package and the supporting tools for its use. This kit **corresponds to the baseline of standards that need to be applied in all kinds of digital activity (character encoding, language codes, country codes, time and dates, etc.)**[7]*"*. The main goals of the SSK can therefore be summarized as follows:

- demonstrating to researchers, with practical examples and specific use cases, that standards could be useful for them;
- encouraging researchers to use standards during their research projects;
- assisting researchers in using standards during their research projects.

Within this framework, the work package 4 is focusing on the user experience (or UX): according to this vision, the SSK will allow researchers in the Arts and Humanities to independently find useful standards for their research. Thus, the Standardization Survival Kit needs to be based on strong user-oriented use cases. This workflow will allow to collect the researchers' needs and constraints regarding standardization, and in the end to respond to these requirements. In collaboration with the work package 2, WP4 members selected use cases, mainly provided by the Report on User Requirements (deliverable 2.1)[8], where:

- a researcher (or a group of researchers) is the primary actor. He created or wants to create digital data.
- an institution is the primary actor, but expresses the same needs as a researcher. In this case, the institution is regarded as a researcher.

After selecting and gathering relevant WP2 use cases — or creating new ones for uncovered fields[9] —, PARTHENOS members who provided and developed the scenarios were asked to adopt a "standards" point of view. Their task consisted of identifying the standards that should be used by the researchers to achieve their projects successfully. With these new inputs, the work package 4 will create an environment where researchers are supported at every step of their research activities. They will find all the information and tools they need and thus they will be able to solve their standards problems on their own.

[7] European Commission, **Grant Agreement: 654119 - Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies (PARTHENOS), Amendment Reference No AMD-654119**, 25 November 2015, p.19.
[8] PARTHENOS, **Reports on user requirements**, 31 January 2016. 318 pages.
[9] Cf.: 2.3. below

2.1.2 Methodology

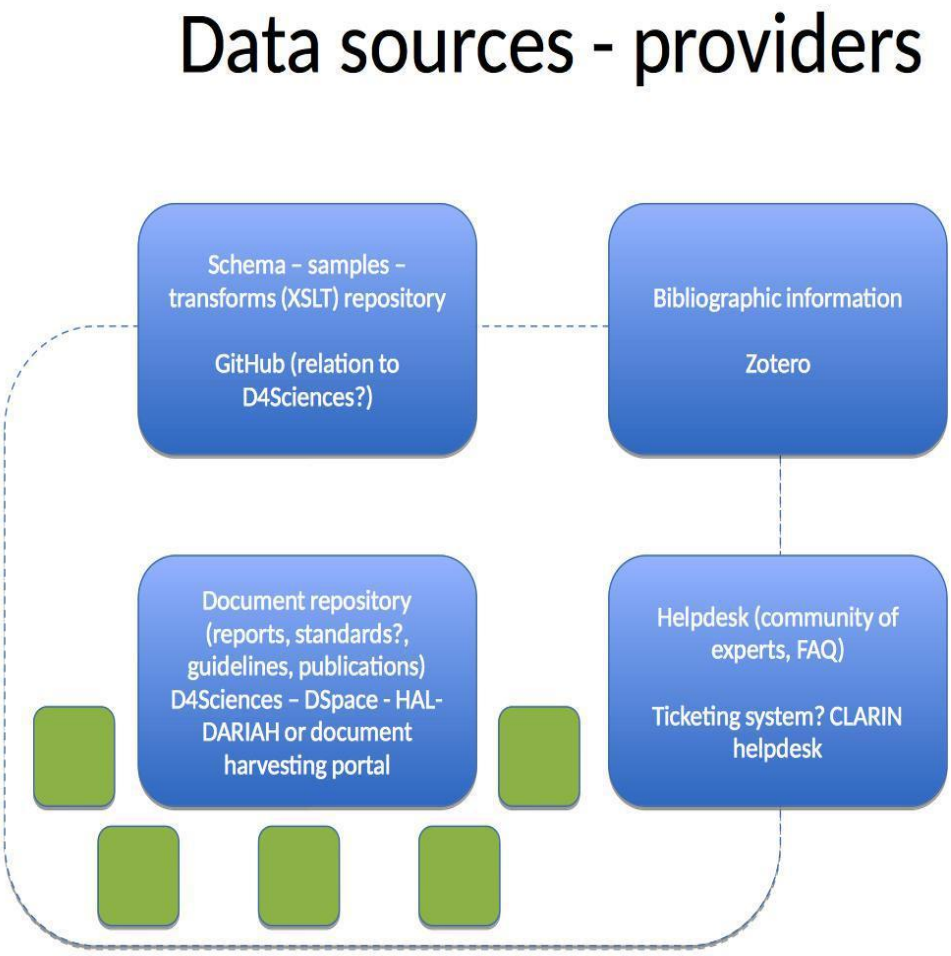
In the deliverable 2.1, the Cockburn simplified description[10] has been used to structure the content of every use case. With the Cockburn description, the main success scenario is a typical and illustrative scenario: it describes each stage of an ideal case, in which researchers solve their problems easily and achieve their objectives without difficulty. Analysing main success scenarios described by each use case puts forward a complete vision of the research process in Arts and Humanities, and offers a better understanding of the role of standards for these disciplines. Nevertheless, some selected scenarios do not allow precise identification of the user requirements regarding standardization. These use cases had to be completed to comply with WP4 goals, especially on three points:

- Explicitly mentioning data models, formats and standards used in the scenarios;
- Imagining or revising scenarios where researchers start from scratch and have no or little experience in Digital Humanities;
- Pointing out, in every relevant steps of the scenarios, when new inputs on standards (from PARTHENOS or another partner) are needed.

These inputs will collect documentation, services, software etc., and then be integrated to the PARTHENOS architecture. Users will be able to find:

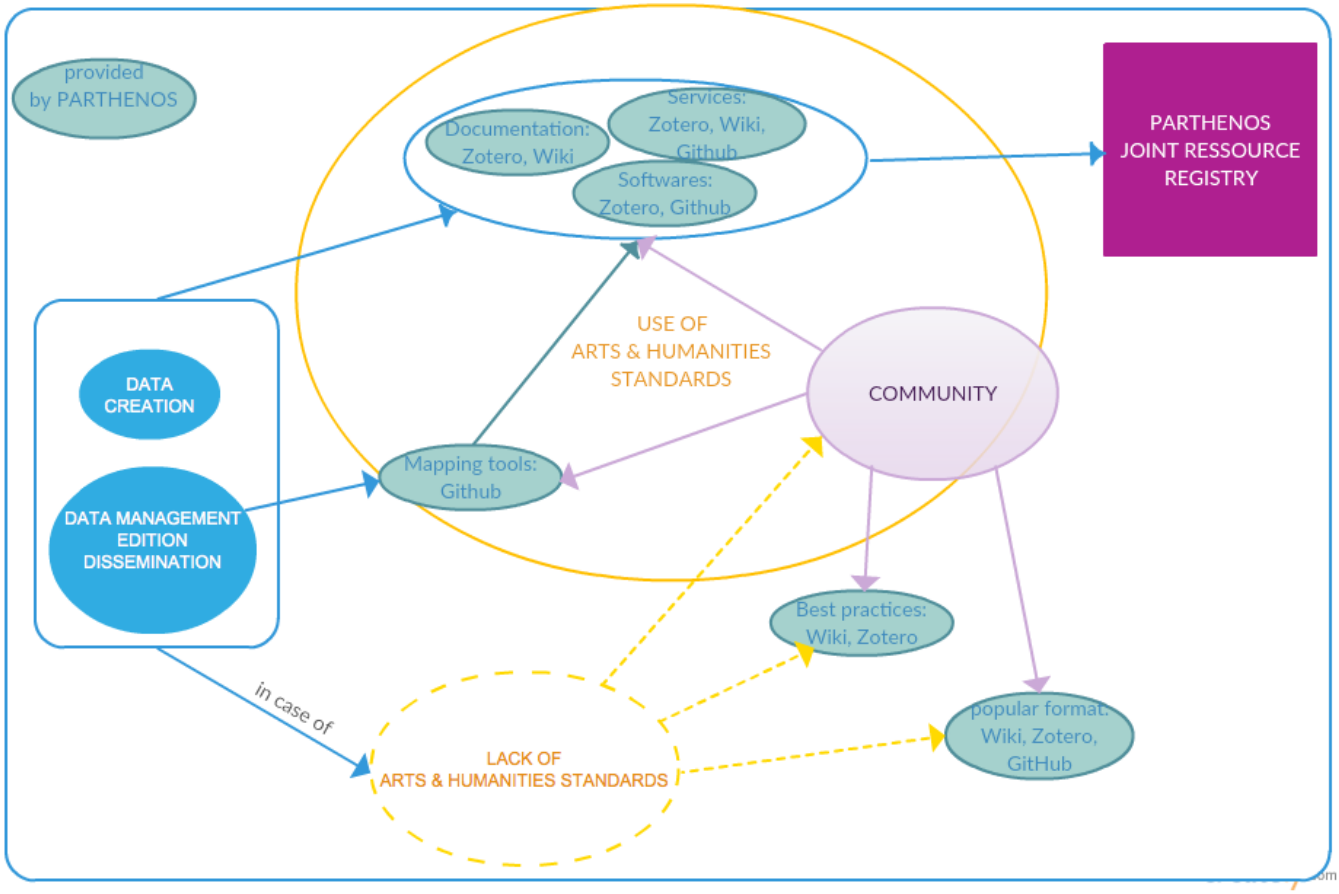
- Documentation and guidelines (Cf. Bibliographic information on WP4 Zotero folder[11]) such as:
- web links to official documentation or tutorials made by PARTHENOS partners or other institutions;
- web links towards tools offering appropriate solutions to researchers.
- Stylesheets or transformation tools (Cf. WP4 Github environment[12])
- Identified user communities who already faced the same type of scenarios, and can provide advice and help (Cf. Helpdesk, listserv, wiki, ...).

[10] Alistair Cockburn, **Writing effective use cases**, pre-publication draft #3, edit date: 21 February 2000, published by Addison-Wesley, 2001. 113 pages. Url: http://alistair.cockburn.us/get/2465
[11] Parthenos WP4, Zotero Group: https://PARTHENOS.zotero.org/groups/parthenos-wp4
[12] Available here: https://github.com/ParthenosWP4



SSK's architecture: first draft

If standardized formats and procedures do not exist, researchers will be able to refer to best practices, and to be put into contact with the appropriate user community. They will also be made aware of popular formats that could be considered as de facto standards.
Analysing main success scenarios allows researchers to spot precisely when they need to use standards, and to understand how PARTHENOS could help them with this:



Overview of data creation process and possible PARTHENOS inputs

In order to identify the inputs that the Standardization Survival Kit could provide, the work package 4 proposed to remodel some use cases selected in the D2.1. It suggested that PARTHENOS members transform the text of these use cases into tables. Each line of the tables corresponds of an identified step of the main success scenario:

- A “Context” column has been added to respond to the wish expressed by some PARTHENOS members, who needed a free space to explain their approach and provide contextual information on their use cases. Some of them chose to add a short paragraph before the table instead of fulfilling the dedicated column.
- The PARTHENOS members must divide their use cases in several steps. For each step, the phase of the scholarly process[13] to which it belongs is indicated in the "Phase" column.
- “Activity” was therefore defined by the work packages 5 and 6 in **PARTHENOS Use Cases from D2.1 as viewed/understood by WP5&6**[14]: “The type of activity to

[13] As defined in Martin Doerr, Carlo Meghini, Kostas Stefanidis, **The PARTHENOS Vision**, joint meeting WP2, WP5, WP6, Rome, 12 October 2015, p.9, url: https://goo.gl/Pk97Rq
[14] **PARTHENOS Use Cases from D2.1 as viewed/understood by WP5&6**, working document, p.1, url: https://goo.gl/DTUItZ

perform / task to accomplish **by the user** within given use case. (indicate as verb)". To fulfill this column, PARTHENOS members were asked to use TaDiRAH, a Taxonomy of Digital Research Activities in the Humanities[15]. It "has been developed for use by community-driven sites and projects that aim to structure information relevant to Digital Humanities and make it more easily discoverable. The taxonomy is expected to be particularly useful to endeavors aiming to collect information on Digital Humanities tools, methods, projects, or readings[16]". This taxonomy is therefore perfectly suitable to describe the activities as they were previously defined by the work packages 5 and 6.

- The last column aims to gather all "Inputs" that could be included in the PARTHENOS architecture, like documentation and guidelines in the WP4 Zotero folder, stylesheets or transformation tools in the WP4 Github environment, or link(s) to user communities.

Context	Contextual information on use cases (in this column or in a paragraph put aside)
Phase	One verb (or two if necessary[17]) among those used in **The PARTHENOS Vision** to describe the scholarly process[18]: 1. **Collecting** and **organizing** evidence, proactively (i.e., anticipating users' needs) or reactively (i.e., triggered by users' needs) ● Museums, archives, research, private collections, SMRs, corpora, analytical reference databases, geo-data, gazetteers 2. **Connecting** facts ● Search through collections, texts, data, publications ● Create knowledge bases 3. **Interpreting** facts ● Critical revision of reliability, quality and validity of collected facts

[15] TADiRAH, **Taxonomy of Digital Research Activities in the Humanities / TaDiRAH**, url: http://tadirah.dariah.eu/vocab/. Accessed September 23th.
[16] "TaDiRAH - Taxonomy of Digital Research Activities in the Humanities / TaDiRAH", **Ibid.**, url: http://tadirah.dariah.eu/vocab/sobre.php. Accessed September 23th.
[17] Like did the authors of **PARTHENOS Use Cases from D2.1 as viewed/understood by WP5&6**.
[18] Martin Doerr, Carlo Meghini, Kostas Stefanidis, **The PARTHENOS Vision**, p.9.

	● Create knowledge by inference = produce data ● Ask for more evidence 4. **Presenting** results ● Bring facts and interpretation in context ● Describe methods and sources of knowledge ● Publish as the authors' consolidated, justified opinion
Activity	Describing the activity to perform with the controlled vocabulary "TaDiRAH"
Task description	A sentence explaining clearly what the researcher intends to do
Methodological framework	Method(s) used to perform the task
Standards, formats and services	The standards, tools and services used to achieve the task
Inputs	Inputs that could be offered by the SSK: - Documentation, guidelines (Zotero) - Stylesheets (GIthub) - Transformation tools (softwares, websites, Github) - User communities (Helpdesk, listserv, wiki)

2.1.3 The selected use cases

As a basis for the Standardization Survival Kit, twenty-two researcher-oriented use cases were selected by WP4 members. In these use cases:

- Researchers are engaged in a process of data **creation**, **enrichment** or **transformation**
- Researchers need to **use** and **respect** one or more **standard(s)**.
- Researchers belong to one of the **four target user communities** identified by PARTHENOS. In this selection, the "Language-related studies" community is best represented with eight use cases. Six use cases represent the "Studies of the past",

and six use cases were selected for the “Cultural Heritage, applied disciplines, and Archaeology”. But only two use cases have been selected for the “Social sciences” community. These numbers reflect the representation of these communities in the PARTHENOS project. The deliverable 2.1 highlighted indeed that “The Social Sciences were much less strongly represented [in PARTHENOS] [...]. History, Language Studies and Cultural Heritage, applied disciplines, and Archaeology can thus be considered the highest priority for PARTHENOS[19].”

The majority of the twenty-two chosen use cases has been provided by the deliverable 2.1: eighteen are taken from the deliverable, in particular from the chapter 2 “Standardization requirements”. Besides, four of them have been provided by WP4 members: MIBACT-ICCU included a use case for the “Language-related studies”; IESL-FORTH and CSIC added two new use cases to the “Heritage and Applied disciplines” field; and one new use case has been proposed by DANS-KNAW for the “Social sciences” field.

	Use cases	**Provided by**
Studies of the past	WW1 Historian and the transnational/trans-institutional question of the development of the railways	TCD-CENDARI
	Historian wants publish her research data and make it reusable with the DARIAH-DE repository	FHP
	Holocaust Researcher investigates person information and networks	EHRI-NIOD
	Collection Holding Institution publishes data on the EHRI portal[20]	EHRI-NIOD
	Scholars needing standards to represent information (reference tools) about persons, titles of works, documents identifiers	SISMEL

[19] PARTHENOS, **Reports on user requirements**, p.16.
[20] This use case has been provided for the deliverable 2.1, but it has been also developed in 1.3.2. below.

	(shelfmarks…)	
	An historian wants to track the dissemination of a given author's works during the Medieval and Early Modern period	SISMEL-CENDARI
Language related studies	Build a corpus of linguistic data for analysis	OEAW
	Create annotated digital edition	OEAW
	Natural Language Processing Expert wants to test her tool for semantic annotation on an available digital edition of historical texts	CNR
	Interoperability in literature using the TEI	Huma-Num
	Linking original text in literature studies to commentary, translations and external sources	UCPH
	Sustainability and improved viewing of Assyrian text resources	UCPH
	Private Foundation wants to publish the digital collections of its library and museum in the online Public Access Catalogue and in Internet Culturale and CulturaItalia	MIBACT-ICCU
	Librarian that manages a digital library wants to extend the thesaurus, publishing it in SKOS format and mapping it with other SKOS thesauri	MIBACT-ICCU
Heritage and Applied disciplines	Working on 3D formats for archiving and on common metadata	Huma-Num
	Multispectral imaging for the in-situ characterization of painted works of art. The results of the multispectral imaging	IESL-FORTH

	characterization is published in the POLYGNOSIS online knowledge platform[21]	
	Conservation scientist wants to publish information about experimental conditions for Raman analysis of wall painting fragments and report in particular proper experimental measurement conditions for safely detecting and identifying certain types of pigments	IESL-FORTH
	Researcher using lasers in conservation/restoration identifies the necessity of standardized reports of the laser application conditions and the evaluation of the obtained results	CSIC
	Researcher identifies the convenience of standardized Laser-Induced Breakdown Spectroscopy (LIBS) analysis for glass substrate characterization in the field of Cultural Heritage[22]	CSIC
	A dataset for the products used in conservation treatments in order to share information about their application parameters, their effectiveness and their durability in time, related to the type of material and its state of conservation	CNR-ICVB
Social sciences	Platform for inventorying and archiving field surveys in political science (political sociology)	Huma-Num
	A researcher processes raw data on historical and contemporary occupations and performs	DANS-KNAW

[21] This new use case has been added by IESL-FORTH.
[22] This new use case has been added by CSIC.

	statistical analysis and modelling techniques on these data[23]	

Besides, WP4 is currently adding six new use cases[24]. They will allow new objects or partially uncovered fields in the Standardization Survival Kit to be taken into account. The "Language-related studies" community confirms its central position with three new use cases. With a new use case regarding the Music Encoding Initiative (MEI), the "Studies of the past" are reinforced by three new use cases.

	Use cases	**Provided by**
Studies of the past	Collection Holding Institution publishes data on the EHRI portal (new developments)	INRIA-EHRI
	How to easily create and manage a prosopographical database EAC-CPF with Omeka CMS	INRIA
	Music notation and open source software: how to build a musical corpus in MEI standard	Huma-Num
Language related studies	A TEI Format for Layout and Logical Annotation of Historical Corpora[25]	BBAW
	Transcription of speech	INRIA, CLARIN-IDS, Huma-Num
	Towards terser guidelines for the representation of digital dictionaries and the interchange of lexical data	Inria, CLARIN-IDS, CLARIN-BBAW, OEAW, CNR-OVI

[23] This new use case has been added by DANS-KNAW.
[24] Cf. 1.3.
[25] For the Cockburn description, cf. Appendix "Cockburn description of the DTABf use case".

2.2 Use cases from the D2.1 as seen by WP4

These outcomes are the result of a collaborative work, performed by WP4 members and colleagues who are not directly involved in the work package's activities.
The WP4 team worked with twenty-three different people from thirteen institutions. They were generally quick to react, and WP4 had an hundred percent response rate.

2.2.1 Studies of the past[26]

Most of the use cases within the 'Studies of the past' section are connected with the need to provide discoverability for the digital primary sources required by historians in different phases of their research practice, so that they can use various distributed datasets and tools as an integral component of their research methodology. However, the current situation of the historical data ecosystem is characterized by a high degree of fragmentation, preventing resource discovery and access. This situation is due to the fact that, in the process of data creation and management, a number of different actors, including individual researchers, research groups and memory institutions, are playing an active role. As a result, a large part of the available historical digital datasets – spanning different periods, languages and documents types – that are the accumulated results of the research of individuals, teams and institutions, form a vast and fragmented corpus and their potential is constrained by difficult access, lack of interoperability and non-homogenous perspectives.

Relevant data may also be embedded in, or attached to descriptions, records, documentations etc. produced and managed in different research contexts and domains. Furthermore, data and content types are various and comprise, for example, textual descriptions, maps of diverse scales, multimedia objects, grey literature and academic publications. Moreover, as the amount of historical datasets is continuously increasing, a "big data" issue in data management and access is rising, as discussed in the 2007 Big Data report from English Heritage & Archaeology Data Service (ADS) and confirmed by the reference literature in the field.

Historical resources currently available in the digital ecosystem, such as Databases of texts, Digital libraries, Bibliographical and Lexicographical resources — covering different

[26] For this section, the overview has been written by Emiliano Degl'Innocenti.

languages, from greek and latin to vernacular and modern European languages — 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.: TEI, EAD, MODS, METS, DC, EAG, EDM, DM2E, PDF), preventing their full discoverability. In the worst case scenario, valuable resources are made available without using any widely accepted standard to properly represent or encode research data, making those objects virtually inaccessible.

The work of PARTHENOS WP4 (documentation of the standards used in each research community), together with the results of WP2 activities (documenting the interoperability requirements for datasets, services and tools that are relevant for the researchers) should result in the provision of a set of VREs supporting different phases of the research process:

- tracking entities (persons, objects, concepts) through space (i.e.: visualize on a map) and/or time (i.e.: build timelines);
- represent relations connecting those entities (i.e.: networks visualization);
- extract meaningful (often not explicit) information from the digital sources (i.e.: entities extraction, web crawling and data mining).

Other than the availability of entity recognition tools, annotation and data curation tools, linked data and domain vocabularies/thesauri integration, a key issue shared with other communities in the PARTHENOS landscape — in particular language related studies and heritage and applied disciplines — is to bridge the gap between tangible and intangible aspects of the Cultural Heritage Objects (CHO) allowing researchers to follow truly innovative research paths.

2.2.1.1 WW1 Historian and the transnational/trans-institutional question of the development of the railways

TCD-Cendari

Authors: Jennifer Edmond, Vicky Garnett, Francesca Morselli

In this case, A WW1 Historian wants to analyse the change as well as the expansion of the railway system in East Central Europe (starting with Lithuania and Poland) at the end of World War I. What she needs to have at the very beginning are maps of railway lines before the outbreak of World War I, maps of the construction of new railroad tracks under German occupation, and maps of railroad construction plans of Poland and Lithuania after their respective declarations of independence and the setup of traffic/infrastructure ministries. After finding these maps, the challenge is that of bringing them in agreement regarding their scale so that she can create a unified map of track modifications, as well as new constructions in the whole region. Moreover, as she wants to integrate these maps with chronological data about train timetables, she retrieves a great number of timetables from different memory institutions in many east-european countries. Once these have been OCRed she realises that the data collected are heterogenous and need to be cleaned and harmonised in a database that she creates for her project. At this point, the historian has digitised, cleaned and harmonised both geographic and chronological data and can proceed with their unification and consequent visualisation and interrogation thanks to GIS software, such as ARCGIS.

Phase	Activity	Task description	Methodological framework	Standards, formats and services	PARTHENOS inputs
Collect	Discovery	Find maps of railway lines	Online search through Research Infrastructure or database	Google, Repositories of Archives and Libraries, Archival Repositories, Archival research guides, Europeana, EDM, EAD,	Web link

				MARC, D Core	
Collect	Discovery	Find timetables of railway lines	Online search through Research Infrastructure or database	Google, Repositories of Archives and Libraries, Archival research guides, Europeana, EDM, EAD, MARC, D Core	weblink
Connect	Capture/ Imaging	OCR scan timetables	Scan timetables through OCR software	OCR software	Documentation and guidelines / Github
Connect	Interpretation/ Contextualizing	Find equivalent institutions in different states	Research infrastructure, archival guides	EAG, National/ International Registries of Archives and Libraries	Federated directory data/metadata
Interpret	Enrichment / Annotating / Cleanup	Change the scale of maps	Create comparability in data	Measurement standards for mapping	Documentation and guidelines / User community
Present	Analysis / Spatial Analysis	Integrate chronological data into a map	Using GIS software (or similar)	GIS software, such as ArcGIS, or possibly also Google Earth	Training. Documentation and Guidelines. User community.

2.2.1.2 Historian wants publish her research data and make it reusable with the DARIAH-DE repository

FHP

Author: Jenny Oltersdorf

For a long time research data in the humanities and social sciences were stored locally on a hard drive, on a CD or in another non-public place. They could not be found by other researchers and / or were difficult to access. DARIAH-DE has developed a repository for the publication and long term preservation of digital research data in the humanities and social sciences. It is available in an advanced beta version. It is the central component of the DARIAH-DE data federation architecture that allows the sustainable long term preservation and publication of research data including metadata. Every object receives a permanent and unique persistent identifier (PID) to guarantee sustainable referencing and citing. By doing so, research data can easily be reused and researchers can get credits for them. The DARIAH-DE repository is freely available for DARIAH associated research projects as well as for individual researchers. A research, e.g. a historian, can easily select the research data she wants to publish and upload them via the “DARIAH-DE Publikator”.

Phase	Activity	Task description	Methodological framework	Standards, formats and services	PARTHENOS inputs
Collect	Capture/ Gathering	Gather relevant research data to enable verification of published article	• Researcher selects relevant data from own storage	JPG, TXT, spreadsheet	
Connect	Meta-Activities/ Assessing	Check offered repository solutions for publishing research data in the humanities and social sciences and open for several research data formats	• Browse existing inventories of research data • Select relevant repository	Re3data, ZENODO	List of relevant repositories
Interpret	Enrichment/ Annotating	Index research data for further re-use	• Describe data based on relevant / and or provided standard	http://dublincore.org/documents/dcmi-terms/	Overview of relevant meta data standards in the humanities and social sciences, training on how to apply them and why

Collect	Meta-Activities/ Assessing	Determine legal situation of processed research data	• Scrutinize relevant legal status	Creative Commons	Documentation of existing collections and platforms on that topic, examples, best practices
Present	Dissemination/ Publishing	Upload data to DARIAH-DE repository			
Present	Dissemination/ Sharing	Link research data with publication			

2.2.1.3 Holocaust Researcher investigates person information and networks

EHRI-NIOD

Author: Karolien Verbrugge

Within the framework of EHRI, a Holocaust researcher aims to investigate the networks in which European Jews operated during their persecution in the Second World War through prosopography. A prosopography can be defined as an investigation of a historical group linked by a common factor based on the connections between individual members of this group. The leading question is the way these members operated within and upon the social, political, legal, economic, and intellectual institutions of their time.

The prosopographical approach as proposed deals with large quantities of archival source materials and involves mapping out and analysing the various networks represented in those sources by using computational techniques (Natural Language Processing). The

researcher sets up a “Linked Data” information model to capture relationships between entities. The information about the personal entities is structured according to the standard Encoded Archival Context – Corporate Bodies, Persons, and Families (EAC-CPF). An open-source toolkit is used for the creation and enrichment of prosopographical resources, integrating text mining tools and services to automatically tag and disambiguate the mentions of known entities, as well as to discover new entities that need to be added to the knowledge base.

The researcher gathers the archival sources from several collection holding institutions (CoHIs). He identifies this material through the EHRI portal and requests the CoHIs to digitize the materials and to make the content available in a machine readable format, e.g. alto-xml. The researcher analyses the prosopography and answers his research question.

Phase	Activity	Task description	Methodological framework	Standards, formats and services	PARTHENOS inputs
Connect	Capture/ discovering	Holocaust researcher selects relevant sources	• Browse the EHRI portal • Check Privacy protection laws • Creates selection of relevant sources for his/her research	EHRI portal functionality (Watched Items & Notes)	
Connect	Capture/ gathering	Identify relevant sources	• Researcher identifies if the sources are digitized and preferable	EAC-CPF	List

			in machine readable format (metadata) ● Contacts the collection holding institution for digital copies ● Best possible is a CHI that has standardized descriptions of persons in EAC-CPF		
Collect	Capture/ Imaging	Digitize the document	● CHI digitizes the source for the researcher if it is not digital	Standard or popular file formats: Tiff (preservation format) Jpeg (presentation format)	List (Zotero)
Collect	Capture/ Data recognition	OCR of the text	1. The digital copy is being OCR'd and the text is presented in a standardized XML-format (ALTO)	1. ALTO-XML	1. Schema, documentation (Zotero)

Collect	Enrichment/ Cleanup	Correcting OCR errors	(semi-)automatic text recognition & OCR clean-up		
Connect/ interpret	Analysis/ Content Analysis	Extract information on persons, families and/or corporations and their relationship	1. Persons, families or corporations are recognised 2. possible relationships are also recognised and 3. Captured the entities in a standardized way	Text-mining tools (Ontotext) EAC-CPF EAD Named Entity Recognition Tools (Ontotext)	● Tools ● Services ● Controlled EHRI authority files/vocabularies on persons & corporations http://ehri.github.io/ http://ehri.github.io/ http://ehri.github.io/
Connect/ interpret	Analysis/ Network Analysis	Extract network information	1. Analysis of selected entities and their relations 2. Possibly reference against Linked Open	● Named Entity Recognition tools ● Services Relational Analysis ● Controlled EHRI	examples

			Data sets of Persons	authority files/vocabularies on persons & corporations http://ehri.github.io/ http://ehri.github.io/ Linked Open Data sets ● Nodegoat ● Palladio	
connect/ interpret	Interpretation/ Modeling	Model data to capture relationships between entities.	Model the extracted data, based on analysis of entities and their relations	Linked Open Data Relational Analysis tools	
Connect/ interpret	Enrichment/ Annotating	Characterization of relationships between corpus entities	Described the relationships standardized with relationship ontologies	Linked Open Data	Documentation, examples, best practices
Connect/ interpret	Analysis/ Visualization	Data is available/presented in a tagged network	The entities and their relationships are available in a service that enables	Software services, eg. ● Nodegoat ● Palladio	

		or graph structure	the researcher to answer his/her research questions about networks		
Interpret/present	Interpretation/ Theorizing	Researcher analyzes the prosopography	The researcher researches and publishes his/her findings		

2.2.1.4 Collection Holding Institution publishes data on the EHRI portal

EHRI-NIOD

Author: Karolien Verbrugge

EHRI identifies a Collection Holding Institution (CoHI) with Holocaust related sources. The data is considered as relevant for Holocaust scholars. The relevant archives and collections have been created during WWII by persons, organisations or companies, or have been created after the war, for instance by survivors. The sources are usually kept in paper format, and in some cases they have been digitized, in rare instances as Optical Character Recognition (OCR). EHRI aims to integrate descriptions of the identified sources into its portal. EHRI invests time and expertise to support the archive to make digital descriptions of the Holocaust related sources and the CoHI is willing to invest in making digital collection descriptions. EHRI wants the CoHI to make standardised descriptions. EHRI follows international archival standards: EAD-CPF, EAD, EAG, ISAD (G), ISAAR(CPF), ISDIAH. The CoHI makes the descriptions and exports it to EHRI, who publishes the data on the EHRI portal.

Phase	Activity	Task description	Methodological framework	Standards, formats and services	PARTHENOS inputs

Collect	Meta-activities/ project management	Contact is established between CHI and EHRI	• EHRI contacts CHI • EHRI explains purpose EHRI programma	Email Phone	
Collect	Meta-activities/ project management	CHI agrees with EHRI conditions	EHRI and CHI sign agreement about activities	EHRI Project agreement with CHI	
Connect	Capture/ discovery	EHRI or Collection Holding Institution identifies Holocaust Sources	• EHRI/CHI browse collection on relevant sources • EHRI/CHI selects sources	CMS of CHI (eg. selection via keywords) Non digital finding aids	
Collect	Meta-activities/ project management	EHRI and CHI make plan of action	• EHRI and CHI make further plans about standards, activities and planning		
Interpret	Enrichment/ annotating	CHI makes descriptions (metadata)	CHI makes standardized description of Holocaust Sources in EAD using ISAD(G)	EAD ISAD (G)	

Collect	Capture/ imaging	Alternative 1: EHRI creates the sources metadata • Digitize metadata from non digital source	If CHI is not willing to invest in digitization of collection description: • EHRI digitizes the metadata if not digital	EAD ISAD (G) http://PARTHENOS.icacds.org.uk/eng/ISAD(G).pdf	
Collect	Capture/ conversion	Alternative 2: EHRI converts metadata in a standard format	If CHI is not willing to invest in digitization of collection description, EHRI converts digital metadata in standard format.	EAD ISAD (G) http://PARTHENOS.icacds.org.uk/eng/ISAD(G).pdf	
Interpret	Enrichment/ annotating	Alternative 3: EHRI creates the sources metadata • Creates (high level) metadata from scratch	If CHI is not willing to invest in digitization of collection description: • EHRI makes standardized description of Holocaust Sources	EAD ISAD (G)	

			in EAD using ISAD (G)		
Collect	Capture/ gathering	CHI exports metadata from CMS to EHRI	• EHRI registers dataset CHI in EHRI cloud • CHI exports metadata using OAI-PMH protocol	OAI-PMH Resourcesync CSV XML	
Connect	Capture/ conversion	Checking the data provided by the CHI according to EHRI descriptions requirements	A specific EAD schema, with additional rules to comply with EHRI requirements.	This schema is generated by an ODD file, that put together every information needed (Large schema, additional rules and documentation of the whole). This file is implemented in	ODD file and relaxNG schema and HTML documentation (temporary URLs)

				an user interface that gives to the CHI informations on what to change in their descriptions to import them more easily	
Connect	Capture/ conversion	Mapping the data	A mapping tool based on the messages generated by the standard checking. The output format is XML-EAD that strictly respects EHRI guidelines	XSLT transformations, Output format tested with an EHRI_strict relax NG schema	
Interpret	Interpretation/ contextualizing	EHRI describes CHI according to EHRI guidelines	EHRI makes description of CHI in EAG using ISDIAH/EHRI Guidelines	ISDIAH EHRI guidelines EAG	
Present	Dissemination/ publishing	EHRI publishes metadata on EHRI portal		EHRI Portal	
Present	Dissemination/	EHRI publishes		EHRI Portal	

	publishing	organization profile of CHI on EHRI portal		https://portal.ehri-project.eu/	

2.2.1.5 Scholars needing standards to represent information (reference tools) about persons, titles of works, documents identifiers (shelfmarks…)

SISMEL

Authors: *Emiliano Degl'Innocenti, Roberta Giacomi*, Maurizio Sanesi

In medieval studies, it is often needed to define in an unambiguous way information like person names (i.e.: authors), titles of works and documents identifiers (i.e.: shelfmarks). In fact, information can frequently be ambiguous: there are several cases of homonimy; an author can have more than one appellation known; the title of a work can have different forms, and also shelfmarks may have different structures. The quantity of material involved in the research and the time employed to perform it may vary depending on the subject of the research itself. This kind of research can be performed, in a large part, consulting websites, digital platforms of resources and digital libraries. Nevertheless, no standard is actually used by scholars: therefore this use case describes in steps the best practice carried out in order to produce a coherent dataset for research purposes.

Phase	Activity	Task description	Methodological framework	Standards, formats and services	PARTHENOS inputs
Collect	Capture/ Gathering	Formulate a research question	Define the target population and the geographic, chronological	http://digitalhumanist.net/standards/index.php/Prosopography	

			and thematic boundaries		
Collect	Capture/ Gathering	Survey the source material and the general historical and theoretical literature	Check all available prosopographical resources	http://digitalhumanist.net/standards/index.php/Prosopography	Provide access to the resources, maybe with a faceted search
Connect	Interpretation/ Modeling	Build a systematic and uniform prosopographical database	Build the database using primary sources and literature, including both background studies and other relevant studies (e.g.: historical and theoretical – sociological, anthropological etc.). The reliability of selected sources should always be checked	http://digitalhumanist.net/standards/index.php/Prosopography	Provide tools in the VRE for integration and managing of dataset built on personal research
Connect	Analysis	Decide the working	Choose an approach (e.g.:qualitative or	http://digitalhumanist.net/standards/index.php/Prosopography	

		methodology	quantitative), adopt methods of analysis (e.g.: software/tools to use; or mixed approach), etc.		
Interpret	Analysis/ Content analysis	Analyse the data from the available (digital) resources; Combine/interpret data (sources and literature)	Propose answers to solve research questions (see Task description ”Formulate a research question;”). Different kinds of sources can provide information about a certain population: • Demographic sources; • Economic sources; • Administrative sources; • Religious sources; • Juridical sources.	http://digitalhumanist.net/standards/index.php/Prosopography	Integration of a programming language in the VRE for writing scripts for analysis and domain most used tools (recommend tools extensively used by domain researchers)
Present	Dissemination/ Publishing	Present research	Publish datasets and/or other traditional/enhanced	http://digitalhumanist.net/standards/index.php/Prosopography	Provide tools in the VRE for integration of

		results	publications		the data and the research results into publication platforms

2.2.1.6 Historian wants to track the dissemination of a given author's works during the Medieval and Early Modern period

SISMEL

Authors: ***Emiliano Degl'Innocenti, Roberta Giacomi***

In medieval studies, the dissemination of a work can give important information on life and movements of the author; plus, it can influence local history and everyday activities: i.e. some medical practices were performed just in places where the work carrying information about them was spread. Moreover, it can be useful also to understand if the work was disseminated alone or together with other works, and why: for instance, in medieval universities the essential works for the study of a discipline were often gathered in the same manuscript. Therefore, scholars may have interest in the itinerary made by a given work, and may want to track it. The use case below describes the best practice to be performed in order to obtain information about the dissemination of a work, from the survey of the work to the final publication. In the 'Standards, formats and services' column are listed all the digital catalogues, platforms and libraries that can support the scholar in every phase of his research.

Phase	Activity	Task description	Methodological framework	Standards, formats and services	PARTHENOS inputs
Connect / Collect	Interpretation/ Contextualizing	Survey and collect all existent		Incunabula Short-Title Catalogue (ISTC) and TEXT-	Provide access to the resources,

		editions of Donatus, **Ars minor**		Inc	maybe with a faceted search
Interpret	Capture/ Discovering	Assess the 15th and 16th-century use of the editions	Discover who were the users of the surviving copies	Material Evidence in Incunabula (MEI)	
Interpret	Meta:Assessing	Assess how many medieval and renaissance manuscripts of this work survive today in our libraries	Access a wide number of electronic catalogues of manuscripts	META-OPAC CERL Portal	Provide access to the resources
Interpret	Meta:Assessing	Understand the popularity and circulation of this work in the medieval and	Assess the presence of this work in catalogues of medieval libraries in Europe	Medieval Libraries of Great Britain (MLGB3), Biblissima and TRAME tools	Provide access to the resources

		early modern period			
Interpret	Meta: Assessing	Ensure inclusiveness of data	Ensure that the CERL Thesaurus is running at the back of the above listed tools	CERL Thesaurus	
Collect	Capture/ Gathering	Bundle collected data	The collected data needs to be packaged in one location for a better use	Cloud / Local Network	Provide tools in the VRE for integration and managing of collected data
Connect	Interpretation/ Contextualizing	Link out to secondary literature on this work		Biblissima and TRAME	
Interpret	Interpretation/ Contextualizing	statistical information, metrics, or other methods are used	Services: analysing frameworks (e.g. for statistical analysis), or with self-developed scripts	Integration of tools into the VRE.	Integration of a programming language in the VRE for writing scripts for

		to analyse the corpus			analysis and domain most used tools
Present	Dissemination/ Publishing	Publish data and results of analysis	Publish datasets and/or other traditional/enhanced publications		Provide tools in the VRE for integration of the data and the research results into publication platforms Integration of a tool that provides information about the publication methodology (e.g.: good practices, guidelines) Integration of an arranging data tool for publication

2.2.2 Language-related studies[27]

A core issue for this chapter is a working definition of the scope of Language-related studies (LRS). An inspection of all the use cases shows that many overlaps exist among the research fields of interest to PARTHENOS. Especially the use cases from the Studies of the Past community share many commonalities with the LRS ones. A characteristic feature of the LRS use cases appears to be the strong focus on textual data and on the processing of such data by NLP tools. This seems to be responsible for domain-specific approaches and concerns in all of the phases of the data research cycle. The most important issue is probably that the processing of data for LRS purposes is to a high degree language-dependent. Some languages are better supported by tools than others. This also has an effect on the usage of standards. Despite the varying degrees of tool support for particular languages, we can observe broad use and wide acceptance of the Text Encoding Initiative Guidelines in the LRS community, a fact that makes TEI XML a reference standard for the encoding of textual material. Nearly every collected use case from the LRS field refers to TEI. Even if in some of the use cases there is no specific reference to NLP, the adoption of TEI makes it, in principle, possible to rely on a relatively well-defined bunch of NLP tools. For this reason, it makes sense for PARTHENOS WP4 to adopt and promote TEI as the main standard for encoding textual data. This can be done by providing documentation, best practices and workflows as well as by registering and facilitating resources and tools for LRS researchers to support the adoption of TEI.

Another key issue is communicating with the LRS community. Even though only some of the use cases refer to CLARIN directly, it needs to be highlighted that CLARIN is the main reference for the LRS community. This includes not only data management and visibility but also stable recommendations for standards and tools. As CLARIN is a partner in PARTHENOS, the cooperation already works well. But it should be stressed that the dissemination of the LRS relevant results of PARTHENOS WP4 should also be integrated in the infrastructure of CLARIN. PARTHENOS WP4 could be therefore seen as a broker between researchers and both the TEI and CLARIN communities. As it is mentioned in some use cases, a worthwhile effort could be to work on guidelines that help homogenize different TEI formats/schemas. This would allow more interoperability between otherwise enclosed projects. Thereby reducing the costs of transforming and adapting data from different sources.

[27] For this section, the overview has been written by Klaus Illmayer.

Although we already have a lot of use cases from the LRS domain, we still lack some more fields that are frequently used. That would include use cases covering different approaches of corpus based processing (e.g. comparative corpus generation, text machine learning) and such ones considering multilingual topics. Also, many of the use cases challenge the process of transforming heterogeneous material for the needs of their research approach. It would be good to focus more on this process, because it seems that there is a lack of widely accepted best practices, guidelines and tools. This raises the need to gather more use cases from the PARTHENOS partners. Besides, it can be observed that there is a strong bias on digital edition work that needs to be more balanced with other LRS fields in the next deliverables. Another point to be mentioned is that most use cases cover only one specific project. It will be a follow-up task to introduce more abstract views on LRS topics and to focus on enabling (stronger) connections between projects with the help of standards.

One important observation from the use cases already collected is that they give insights where LRS can profit from other research fields, and vice versa. This could lead to recommendations for researchers from other communities to enrich their data with language and NLP relevant standards that have the potential to boost interdisciplinary approaches. The LRS use cases also show readiness to strengthen the interlocking of projects based on standards. PARTHENOS WP4 will support this by offering in the SSK not only resources but also metascenarios to point out how these resources can be linked together. This will allow projects to implement best practices for a better interoperability and interchange of data and research results.

2.2.2.1 Build a corpus of linguistic data for analysis

OEAW

Author: Klaus Illmayer

This use case is not based on a specific project but instead it documents best practice from the field. To access a corpus is a requirement for linguistic data analysis. It is therefore important to apply standards and document workflows when building up a corpus. This is not only because standards allow other scholars to adapt their linguistic data analysis on the corpus but also to make the analysis reproducible. The research infrastructure CLARIN is a beacon for all kinds of digital language resources and tools in Europe. By building up and registering a corpus under the premise of CLARIN, the visibility and quality of the data is guaranteed. The aim of this use case is to show how a corpus can be build up by using resources from CLARIN.

Phase	Activity	Task description	Methodological framework	Standards, formats and services	PARTHENOS inputs
Collect	Capture/ Discovering	Receive machine-readable data from repositories	According to research question find relevant data sets by - using search infrastructures - having in-house data - consulting research	Services: In case of language-related studies researchers (LRSR) they can use either CLARIN-FCS (Federated Content Search) or CLARIN-VLO (Virtual Language Observatory), based on CMDI (Component Metadata Infrastructure)	Integration of possible resources for search in the Joint Resource Registry (JRR, see Task 6.4). Discovering sources with the help of Resource Discovery Tools (see Task 6.5)

			community Use keywords, domain knowledge, etc. to discover useful data sets e.g. minimal/maximal size of data set, required types of annotation, etc. Precondition: Availability of data	But there are a lot more possibilities, depending on the research question, e.g. simple websites, Wikipedia, data from digital edition projects, etc. They can have a lot of different standards and formats in use. Formats/Standards: In general the received data should be already available in the favoured format for analysis and presentation (for LRSR in many cases TEI).	There should be also some way to recommend the integration of resources in the JRR (besides an automatic data/resource crawling) - maybe a report workflow for missing resources. For the Resource Discovery a faceted search would be useful (especially to limit research results in regard to formats, standards, and availability).
Collect	Creation/ Translation Interpretation/ Modeling	Convert into TEI	If data is in different format convert it to TEI (or at least to a text format). A TEI schema is needed or if there are more than one TEI schema is involved, a	Data is quite often available in different formats and standards. With appropriate tools some of this data can be converted to TEI. But in many cases it will be only possible to convert it to a (plain) text format.	Explain process of converting into TEI. Give list of tools to convert (in case of a VRE this tools could be integrated).

			mapping has to be done.		
Collect	Capture/ Gathering	Bundle received data	The discovered data needs to be bundled in one location for better handling of the following steps/activities.	Services: Often the local computer or a local infrastructure (e.g. network shares, the cloud of the research institution or of a commercial company) is used.	A Virtual Research Environment (VRE) that allows to bundle data from different sources.
Collect/Connect	Interpretation/ Modeling Enrichment/ Cleanup Enrichment/ Annotating Analysis/ Structural Analysis	Compile data into a corpus	On base of the research question a model for interpretation of the data is created. For TEI this will be a schema. The bundled data needs to be merged regarding the model. A cleanup of the data is necessary. Probably before and after the annotation. The annotation will be	Services: NLP tools are in many cases language sensitive. There exist different tools for different languages. For PoS taggers you need a tagset, e.g. in German you could use the Stuttgart-Tübingen Tagset (STTS). There are a lot of NLP tools available. But you can not recommend in general a specific one, as the research question should guide you on choosing suitable tools.	An overview on tools and their domain could help to find suitable solutions. This would also need search possibilities. updates on new versions and functions, reference projects, tutorials, and so on. Something that the DiRT Directory is already doing (could it be integrated into PARTHENOS JRR?). NLP tools could be

			done as much as possible automatically with natural language processing (NLP) tools, e.g. stemmers, tokenizer, part-of-speech (PoS) taggers, chunker, sentence/word segmentation, etc. But there will be also the need to manually annotate, therefore a text editor is used.	Standards: For the annotation use XML TEI P5. Named entities should be linked to controlled vocabularies (e.g. VIAF - Virtual International Authority File or http://geonames.org) Standards/Formats: For the different steps in NLP there is a recommendation of CLARIN what standards and formats to use: https://PARTHENOS.clarin.eu/faq/what-standards-are-recommended-clarin	integrated into the VRE so that the bundled data can be compiled to a corpus within the VRE. The availability of an XML text editor in the VRE is necessary for manual annotation. Collect and create tutorials for using NLP tools (a shared space where this information is collected and disseminated). List of recommended controlled vocabularies (JRR?). Recommendations on the use of standards and formats for compilation, annotation, analysis, and

					preparation of the data (WP4 helpdesk, SSK, tools to check compliance with Standards, in case of a VRE there could be a guidance: “Your data seems not to take Standards in account. We recommend to use this tool/tutorial to convert into TEI”, etc.)
Connect	Storage/ Identifying Enrichment/ Annotating Storage/ Organizing	Prepare corpus for further investigation	The different parts of the corpus should get metadata information and (persistent) identifiers. This helps to analyse, re-organise, and investigate the data. Also it is a preparation for (long-term) storage. It will also help to use the data of the corpus for other projects. To	Services/Standards: For persistent identifiers (PID) use a PID service. CLARIN has a PID policy on this: https://PARTHENOS.clarin.eu/content/pid-policy-summary For electronic documents a digital object identifier (DOI) could be acquired (but that is probably more useful for research results).	The VRE should establish an uniform PID service. There should be the possibility to organize data in different ways: Allow to assign metadata, structure data hierarchically (e.g. folders), associate data (tagging).

			organise the corpus in such way helps to find data and to review the research results.		
Interpret	Depending on research question everything from Analysis could be useful Interpretation/ Contextualizing	Analysis of corpus	The enriched corpus is analysed using computation of statistical information, metrics, or other methods depending on the research question.	Services: This will either be done with ready-to-use tools, with analysing frameworks (e.g. for statistical analysis R could be used), or with self-developed scripts	Integration of tools into the VRE. Integration of a programming language in the VRE for writing scripts for analysis (for a seamless experience of the knowledge creation process). Alternatively, export functions or API in the VRE to get relevant data out of the corpus for usage in self-developed scripts.
Present	Storage/ Preservation	Prepare results of analysis for further usage in	Ideally this is long-term storage of the research data (but also of the	Service: Provider of long-term storage should be officially recognized, e.g. with the Data	WP3 will give recommendations on data management and data

	Storage/ Archiving	the research community	input data, if the legal situation allows it)	Seal of Approval or as a Certified center of CLARIN Standards: For metadata of the stored resources and for harvesting the provider should deliver Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)	research life cycle. List of providers where to long-term store data (JRR?). It would be great if it is possible with “one click” to process all the data (input, bundle, research results) into the long-term storage space of a provider.
Present	Dissemination/ Commenting Dissemination/ Publishing Dissemination/ Sharing Creation/ Designing	Publish data and results of analysis	There are many possibilities, usually the research results are commented (that could be text and/or visualizations) and a paper is published. An increase would be the sharing of the research results (and of the data) e.g. in social	Services: Web environments for publishing the data e.g. Content Management Systems (CMS) or tailored repositories with the possibility to exploit the data (APIs, Export functions, etc.)	Provide tools in the VRE for integration of the data and the research results into social media, blogs, websites. Best practice recommendations for publishing data and results of analysis. List of tools that help to

	Creation/ Web development		media, blog entry, etc. An effort would be the development of a website for inspection/exploration of the data for the public.		prepare data for presentation.

2.2.2.2 Create annotated digital edition

OEAW

Author: Klaus Illmayer

This is a somewhat generic description of workflows that lead to an annotated digital edition. It is used for projects in the OEAW department and it is based on best practices in the field of digital editions. Not every step needs to be done, but covering most of them will result in better quality of data. A digital edition can have a lot of starting points: The legacy of a writer, specific documents to a research question, even one single document. Not only the provenance but also the material itself can be of a homogenous type but more likely it will have a heterogeneous aspect (e.g. handwritten documents, typescripts, letters, etc.). What all projects have in common is a corpus of text material with a research interest in language-technology based scholarship. As a de-facto standard (as it is widely used in the community) TEI comes into play. The different contexts of the projects are handled with tailored TEI, specific enrichments, and selected NLP analysis. While there is a common agreement on TEI, the project specific part is handled in very different ways.

Phase	Activity	Task description	Methodological framework	Standards, formats and services	PARTHENOS inputs
Collect	Capture/ Gathering	Prepare and bundle documents for the digital edition	All of the documents for the digital edition should be collected at one place. As the material can be very heterogeneous, it makes sense to handle documents in regard to the type of material (handwritten text, manuscripts, letters, etc.). The main concern for the bundling is to establish a structure in accordance with the targeted result.	Services: Usually this is done on a local computer, if there are more than one people involved in a shared space (e.g. cloud service of an University). Often the structuring is done by creating different folders and/or by putting information on the type of document into the filenames. There are also tools that can help you, but no standard tool/format for this comes into my mind (examples for helpful tools: ATLAS.ti, MAXQDA — but both are not very useful for the following steps of the creation of a digital edition)	Technical framework: A Virtual Research Environment (VRE) to gather the documents and to structure them (tag documents, group documents in different perspectives, etc.). WP4: Best practices on structuring heterogeneous documents (e.g. how to handle letters, manuscripts, etc.). Raise awareness that a well thought structure will simplify the later work.
Collect	Interpretation/	Choose a	This is a harmonizing	Services: Use a text editor for	Technical framework:

	Modeling Creation/ Translation	TEI schema for encoding. Apply TEI schema on documents	process. If there are documents that are encoded in a different TEI schema, they need to be converted.	this. Ideally the text editor supports creating XML e.g. Oxygen XML Editor For conversion: Manually or tools e.g. OxGarage or self developed tools/command-line scripts (using XSL stylesheets).	• integrated XML Editor in the VRE • Finding resources and editors in the Joint Resource Registry (JRR) WP4: • list of XML TEI compatible editors • best practices/highlighted project for TEI encoding and references to TEI Guidelines, TEI mailing list, etc.
Connect	Enrichment/ Annotating	Tokenization and lemmatization of the	Tokenization usually on word and sentence level. Lemmatization may also	It is language dependent how good esp. lemmatization tools work.	Technical framework: • Integration of Tools in the VRE + finding tools via

		texts	include Part-of-speech (PoS) tagging.	Services: (some examples of tools) ● Apache OpenNLP ● Stanford Tokenizer + other tools from the Stanford NLP Group ● TreeTagger ● ... Formats: XML TEI P5 should be the result of the Tokenization and lemmatization. Something that the TTLab Preprocessor or the MorphAdorner does. Something that is missing is a Token editor (e.g. for this we have at our department a self developed tool).	JRR ● Results of Resource Discovery should point out if resources already tokenized/lemmatized ● Integration of XML files in the VRE, possibility to edit and to annotate this files on base of tokens, lemmas, etc. WP4: ● list of tools and experiences with them ● comparison list of tools would be

					interesting (using research domain specific interests e.g. supported languages) • Tutorials, reading list for beginners, discussion forum, examples, ...
Connect	Enrichment/ Annotating Enrichment/ Cleanup	Perform Named entity recognition (NER)	Enrich the tokens with references, this should be done as far as possible automatically, but it will need a lot of manually work (especially clean-ups of not correct references e.g. to a different geographic entity)	Services: A lot of tools but also a lot of controlled vocabularies that can/should be used for referencing (and could be expanded with the results, as for domain specific data it is possible that no standardized vocabulary exist). The tools are used to (semi-)automatically annotate the data. To get better results you combine tools like Apache Stanbol or Babelfy. A problem is the	Technical framework: • Integration of NER tools and controlled vocabularies in the VRE + finding tools and vocabularies via JRR • Results of Resource Discovery should point out if resources are NER

				consolidation process. For this we are using a self developed web tool. Afterwards you will have to clean-up/correct/complement the annotations manually. Formats: XML TEI as result of the tokenization and lemmatization will be enriched considering the defined model for TEI annotations	performed and if so, what are the controlled vocabularies the NER refers WP4: ● List of NER tools + experiences, tutorial material, etc. ● List of controlled vocabularies + info how to resolve results and to link to data
Present	Dissemination/ Commenting Dissemination/ Publishing	Publish edition	There are different levels of publishing: It could be that the main goal of publishing is a printed edition where with	Services: ● For the print process you will transform the XML data to the needs of the print publishing process ● A website will often be	Technical framework: ● Registering digital editions in the Joint Resource Registry (JRR) ● Finding digital

	Creation/ Designing Creation/ Web development		the help of the digital edition the publishing and commenting is enriched. Ideally the digital edition is published as a website (for this you will need a design and the development of a website) or with the help of a web environment.	developed individually, this will also need a transformation either to HTML or into a database. ● An alternative would be the use of web environments that are tailored to Digital editions (e.g. TextGrid) or integration into Content Management Systems (CMS) like Drupal, Wordpress, etc. Formats: For the transformation use XSLT and TEI XSL Stylesheets	editions via the resource discovery ● Manage digital edition in the VRE and allow deployment on a web environment where Digital editions are published (something like TextGrid), alternatively allow exchange of data between PARTHENOS and web environments (APIs, Registration, Mapping, etc.) WP4:

					• List of tools and material for transforming, publishing, web development (e.g. how to publish a digital edition with Drupal, Wordpress, ...) • Manuals and sharing of experiences on using web environments for digital editions (e.g. how to connect pictures of facsimile to the text in the Digital edition) • Insights into the creation process of

					best practice projects: How did you do it?
Present	Storage/ Preservation Storage/ Archiving	Long-term preservation of digital edition	At least the result should be stored in a long-term preservation manner. Ideally every part of the process should also be stored (e.g. raw text > enriched text > published text). Enable hooks for analysing the digital edition for the research community	Identical with the section “Prepare results of analysis for further usage in the research community” in the use case “Build a corpus of linguistic data for analysis”	Identical with the section “Prepare results of analysis for further usage in the research community” in the use case “Build a corpus of linguistic data for analysis”

2.2.2.3 Natural Language Processing Expert wants to test her tool for semantic annotation on an available digital edition of historical texts

CNR

Authors: Francesca Frontini, Monica Monachini

Within the Language Technologies (LT) community, strong interest is building up in the potential for testing text analysis tools on corpora other than newspaper articles. Using CLARIN resource repositories, a language technology provider identifies a set of corpora that a particular community of scholars have made available. It may be a philologically curated electronic edition of a historical text, for instance the Nuova Cronica, a history of Florence by the medieval merchant Giovanni Villani. The tool the expert wants to test performs some type of semantic annotation, for example Named Entity recognition, in particular of persons and places. This could be via linking to DBpedia.

LT experts would like to test their tools on this kind of data, but unfortunately they face a series of issues concerning input and output formats. More specifically, it is often the case that the tool that a LT expert or experts have developed takes plain text as input, whereas an electronic edition is – in the best scenario – in TEI/XML, or – in the worst scenario – a HTML page. As a consequence, some code needs to be written in order to extract plain text from the TEI/XML or HTML. Even in the best scenario this may be complicated, as the details of the structure of the TEI schema are not well known among a wider community of LT experts.

Moreover it is often the case that LT experts would like to re-inject the automatic annotation in the original TEI, so as to send it back to the editors for validation, as they too might find it useful to have an enriched version of their text. But the tool only outputs data in a plain, one token per line, tab-separated format that is commonly used by many LT applications. Building a wrapper that converts this in the right format is costly and may require collaboration with the editors of the TEI text.

Phase	Activity	Task description	Methodological framework	Standards, formats and services	PARTHENOS inputs
Collect	Discovering	Search and find corpus in catalogue		Catalogue with resources described with standard metadata format	PARTHENOS registry
Collect	Meta-Assessing	Check the legal situation of corpus		Creative commons	Documentation (Zotero)
Collect	Meta-Assessing	Assess the encoding format of the corpus		• Character encoding: ISO -10646 UNICODE; UTF-8 • TEI Guidelines	• Documentation (Zotero) • List of TEI schemas for particular uses (GitHub)
Collect	Data Recognition	Extract plain text from TEI	XML parsing	TEI Guidelines, XPATH	
Connect / interpret	Enrichment/ Annotating	Annotating plain text with linguistic tools	Natural language processing, morpho-syntactic tagging,	• Feature structure representation: ISO 24610-1:2006 • Morpho-syntactic annotation:	Documentation, samples (Zotero, Github)

			Named entity recognition and linking	MAF (Morpho-syntactic Annotation Framework), ISO/DIS 24611 • Syntactic annotation: SynAF (Syntactic Annotation Framework), ISO/CD 24615 • Lexical annotation: LMF (Lexical Markup Framework), ISO 24613:2008 • Linguistic annotation: LAF (Linguistic Annotation Framework), ISO/DIS 24612 • RDF	
Connect / interpret	Analysis/ Content Analysis Annotating	Enrichment of original TEI text with Named Entity annotation	1. Named entity recognition 2. Linked data	TEI Guidelines (to re-inject NE-annotation and linking into TEI)	Documentation on NER tools (Zotero), services Reference datasets (Services on Github, documentation)
Connect /	Analysis/ Network	Network analysis (ego-	Graph theory		Transformation stylesheet from TEI to

interpret	Analysis	network) based on the TEI elements: <person>, <place>, <date>, <event>, …			network visualization formats (Mapping tools, Github)

2.2.2.4 Interoperability in literature using the TEI

Huma-Num

Authors: Adeline Joffres, Nicolas Larrousse

Huma-Num's Consortium "Authors of Corpora for the Humanities: Computerization, Edit, Search" (CAHIER) which aims at bringing together the various existing or planned initiatives in France in the fields of "Authors' Corpora", providing coordination, sharing experience and promoting access to data. In that perspective, Consortium CAHIER agree on a common use of TEI format and create a collaborative platform to gather and disseminate corpora.

Phase	Activity	Task description	Methodological framework	Standards, formats and services	PARTHENOS inputs
Collect	Meta activities/ Community building	Initiating and coordinating a "grand dialogue" between the partners of the	- Meetings (virtual and face-to-face), - participating in		Enlarge users community

		consortium from various disciplines (literature, linguistic, philosophy, history, political science, etc.) on metadata and data	conferences and organising research events (including specific ones for TEI community members) - funding several projects of corpora's digitalization,		
Collect	Creation/ Programming	Building a unique core format based on TEI (Text Encoding Initiative) to describe digital objects (above all corpora) derived from various sources and different formats.	- Choosing a common type of description - networking	TEI	Documentation, guidelines, samples (Zotero, Github)
Connect/ Interpret	Creation/ Programming	Building a tool in order to publish and share in an interoperable way data shaped in this format	WEB-OAI tool provides a virtual research environment in order to describe in a normalized way all the literature's objects considered in CAHIER's consortium and	TEI, OAI-PMH	Contribution to enhance tools (Github).

			give (human) access to the catalogue		
Interpret/ Present	Enrichment/ Editing	Publish TEI header's normalized metadata through an OAI-PMH repository in order to be harvested with rich metadata vocabulary (dcterms…)	- Computing programmation - Organising training session on tools,	TEI, OAI-PMH, metadata models	Documentation, guidelines, samples (Zotero, Github)

2.2.2.5 Linking original text in literature studies to commentary, translations and external sources

UCPH

Author: Lene Offersgaard

Researchers working with old Latin and Greek texts at UCPH express the need for linking information in commentary, translations and other sources to different versions of the original texts. This linking of information can facilitate publishing texts with commentary in two major uses: i) in a simple reading system that can easily display needed and interesting information based on the reading skills of the user, and ii) in a more advanced system that can support research, development of new commentaries to students, and other material. The ability to link in a standardized way should also enable researchers to easily extend the information in the commentary and the linking to other resources in collaboration with other researchers. The primary challenges now is that current standards are available on different sub areas of this setup, but a single researcher do not have the time resources, the knowledge of many standards, and the overview of how to combine the right standards and formats when creating a commentary or a translation of a source text.This case

includes refining TEI formats that can facilitate the formatting, linking and annotation of the texts. Annotation tools are also needed to enrich the texts, and finally a user interface is needed to enable browsing of the material.

Phase	Activity	Task description	Methodological framework	Standards, formats and services	PARTHENOS inputs
Collect	Capture/ Discovering	Collect/select the texts and translations to be linked	Texts are selected based on research interest or focus in teaching	Texts are available in digital files in the formats DOC, HTML, XML, and plain text, but metadata has to be defined and determined for the text.	
Collect	Enrichment/ Annotation	Create metadata for collected files	Create almost minimalistic metadata. A schema for the minimal and preferred TEI header should be created.	Create a TEI-header: specify enough metadata to be compliant with information needed in web-interface, including acknowledgments or references to sources/ authors of the material.	
Convert	Capture/ Conversion	Convert files to TEI-format	If data is in non-TEI format convert it to TEI (use specific TEI format).	Convert data to TEI, and merge with header. Use a special TEI-format with xml-id's attached to all div's and with options to specify notes about the text. Xslt is used to convert for each individual xml format.	Example of converted files with teiHeader: a text and a translation snippet. A plain tool/description of converting the fields

					into TEI can be given. TEI-schema and XSLT stylesheets available on Github.
Interpret	Analysis/ elational Analysis	Align text and translation	Automatic alignment. Researcher checks out problematic cases/errors. The complexity of the alignment task is depending heavily on the translators “loyalty” in the translation phase of the original text, e.g. the focus on adding extra information to the reader.	Align text and translation using proprietary format, that resembles TEI. Suggest to use “de-facto” xml standard used for parallel corpora released from EU or UN. Format should allow for aligning a set of texts e.g. more translations from different locations or date.	Example of alignment file. Documentation of xml-format, and xml-schema on Github
Interpret	Dissemination/ Commenting	Create commentary file. (Commentary should be	The researcher decides what to comment on, and refers to XML-ids in the text or/and the translation file. Commentary should	Commentary also to be formatted in TEI-format.	Example of commentary file and TEI schema.

		extendable later)	be able to refer to word, sentence, and sections of different sizes.		
Interpret	Enrichment/ Annotation	Mark up of NER (persons and places)	Automatic annotation of NER using NLP tool	Extend TEI elements: <person>, <place>, and <event>	Example of file with mark-up of NERs. Tutorial for using NLP tools, and Readme-file to showing how example file is created.
Interpret	Enrichment/ Cleanup	Inspection/Correctiong of mark-up of NERs.	Web page with view mode of annotation allowing for easy edition of mark-up.	Mark-up can be labelled added, deprecated, wrong or “to be validated/checked” in an attribute of the annotation of the NER.	
Interpret	Interpretation/ Contextualizing	Linking to other resources e.g. dbPedia, online dictionaries, thesauri.	Specify how to link to other resources in a way that is easy for a web-interface to handle. Nice to see them by “mouse over”, or to follow by opening in new browser window.	Determine format, e.g. use something like the link to the commentary, links has to be marked to be external.	

Present	Dissemination/ Publishing	Create or extend web-application. Documentation including examples of files and standards.	Implement web-interface with use of xmlDB as knowledge source. Write user and maintenance documentation for the xmldB and the web-interface	All text, files, and info are placed in xml db, e.g. existDB. Web interface query the xml db for info.	A small pilot implemtation can be seen here: http://PARTHENOS.locus-classicus.org Source code in github or PARTHENOS chosen sw repository. Documentation placed in Zotero or D4Science
Collect	Capture/ Gathering	Upload text in web-application			
Connect	Meta Activities/ Assessing	Testing of upload, web interface, links to external sources	Testing of interface and external linking facilities. Prioritizing changes, and implementation of bug fixes.		Test log placed in Zotero or D4Science.

Present	Dissemination/ Publishing	Publish web-interface for teaching	Announcing the service for use in teaching	Using channels defined in PARTHENOS; support for organizing workshops, seminars, and in the user group created in Denmark.	

2.2.2.6 Sustainability and improved viewing of Assyrian text resources

UCPH

Author: Claus Povlsen

The Old Assyrian corpus consists of roughly 23,000 individual documents of which approximately 6,000 have been published (in print). Most of the corpus originates from merchants and represents commercial letters, legal documents etc. A collaboration between researchers at the University of Copenhagen has resulted in a smaller controlled archive of about 1,200 texts (being stored in separate files). The texts are transliterations of the clay tablets and their corresponding English translations. The texts are stored in separate files. The archive also contains digital photos of most of the clay tables and text-specific metadata such as dates and contextual notes. In short, the task besides securing sustainability, is to make it possible for the users to make queries for relevant text collections by exploiting the metadata assigned to them. Finally, the user has expressed the need to get the three representations of a given text segment displayed in parallel.

Phase	Activity	Task description	Methodological framework	Standards, formats and services	PARTHENOS inputs

Collect	Meta:Assessing	The IPR issue is solved, i.e. the permission to make the corpora publicly available is given.			Specific licence templates
Collect	Capture/ Imaging	The documents are digitized		The documents are stored in plain text UTF8	See above in the Context section
Collect	Capture/ Data recognition Capture/ Transcription	Transliteration of the clay tablets was done manually.	Manual transcription	A text editor (Word) was used in the transliteration process	
Connect	Interpretation/ Modeling	Use of a core set of standards needed in all kind of activities		Mandatory core set standards: • Character encoding: ISO -10646 UNICODE; UTF-8 • Country codes: ISO 3166; • Language codes: ISO 639-1 and 639-3	Documentation on ISO standards

Connect	Interpretation/ Modeling Enrichment/ Modeling	Regarding the header part of TEI schema for annotating, an obvious choice would be the one that can be found here: http://catalog.clarin.eu/ds/ComponentRegistry?item=clarin.eu:cr1:p_1380106710826	1. Reuse of an existing schema 2. Customization of an existing schema 3. Creation of a particular schema	TEI Guidelines TEI Roma TEI subset Customized TEI subset Extended TEI subset	Reference in Zotero (Hansen, D.PARTHENOS., Offersgaard, L., Olsen, S. (2014). Using TEI, CMDI and ISOcat in CLARIN-DK. In **Proceedings of the Ninth International Conference on Language Resources and Evaluation** (LREC'14). Reykjavik, Island.)
Connect/ interpret	Capture/ Conversion	Extension: Only a minor part of the corpora are embedded in a TEI			Transformation stylesheet from TEI to TEI

		format			(Mapping tools, Github)
Connect/ interpret	Enrichment/ Annotating	Enrichment of text Annotating the text for linguistic research	1. Named entity recognition 2. Linked data	• Feature structure representation: ISO 24610-1:2006 • Morpho-syntactic annotation: MAF (Morpho-syntactic Annotation Framework), ISO/DIS 24611 • Syntactic annotation: SynAF (Syntactic Annotation Framework), ISO/CD 24615 • Lexical annotation: LMF (Lexical Markup Framework), ISO 24613:2008 • Linguistic annotation: LAF (Linguistic	Documentation, samples on NER tools (Zotero), services Reference datasets (Services on Github, documentation) (Could be relevant at a later stage of this search and curation project.)

				Annotation Framework), ISO/DIS 24612	
Connect/ interpret	Enrichment/ Annotating	Characterization of relationships between corpus entities		• Bio ontology, FOAF • relationship ontology • The organization ontology	Documentation, samples (Zotero, Github)
Connect/ interpret	Analysis/ Structural Analysis	Statistical analysis	1. Textometry 2. Stylometry	Softwares, services	
Connect/ interpret	Analysis/ Network Analysis	Network analysis (ego-network) based on the TEI elements: <person>, <place>, <date>, <event>, …	NLP processing?	Softwares, services	Transformation stylesheet from TEI to network visualization formats (Mapping tools, Github)
Connect/ interpret	Analysis/ Visualization	Network visualization		Softwares, services • Gephi	Documentation, samples

				• Nodegoat • Vistorian	(Zotero, Github)
Connect/ interpret	Enrichment/ Editing	Connect pictures to text		• ALTO-XML • TEI • International Image Interoperability Framework (IIIF)	Github (samples, schema), services, documentation
Present	Dissemination/ Publishing	Publish data: something similar to CLARIN.DK		• TEI Boilerplate • Nakalona • Ortolang • ...	Transformation stylesheets from TEI to publication formats (Mapping tools, Github)

2.2.2.7 Private Foundation wants to publish the digital collections of its library and museum in the online Public Access Catalogue and in Internet Culturale and Culturaltalia

ICCU

Authors: Sara di Giorgio, Antonio Davide Madonna

A cultural institution holds library and museum heritage and wants to digitize, catalogue and share it online, including through portals like Internet Culturale and CulturaItalia for a wider dissemination. CulturaItalia is the Italian National Aggregator that gathers metadata from a large partners network, like InternetCulturale (the portal of Italian Libraries) and sends metadata via OAI-PMH protocol to European Infrastructures such as EUROPEANA and ARIADNE. For digitazing and catalouging the librarian and museums heritage, the curator of the Foundation follows the guidelines and the standards provided by ICCU. The digital collections of the libraries are hosted by Internet Culturale and the ones of the museums by MuseiD-Italia that is a digital library integrated into the CulturaItalia platform devoted to museums collections. MuseiDsItalia is based on Fedora Commons framework and supports METS standard for the digital objects. The curator investigates the intellectual property rights of the objects and its digital reproductions and associates the correct licence. For the metadata level description the Creative Common Public Domain Dedication is chosen. The metadata in the repository of InternetCulturale and in the repository of MuseiD-Italia are harvested by CulturaItalia and available for Europeana and other similar initiatives.

Phase	Activity	Task description	Methodological framework	Standards, formats and services	PARTHENOS inputs
Collect	Meta:Assessing	1. Identification of Intellectual Property Rights conditions of the content 2. Choose the right license for publishing the digital content (metadata and digital objects)	Legal framework about IPR and open data at national and international level	Creative Commons licenses Europeana Licensing Framework The Rights Statements	PARTHENOS Guidelines (D4Science – Zotero)

Collect	Capture/ Imaging	Identification of guidelines to digitize physical or analogue originals such as photographic material, maps, proclamations, broadsides, single-sheet publications, museums' collections and archive contents	Adherence to internationally recognized standards and best practices	Standards or popular file formats	List (Zotero)
Connect	Dissemination/ Publishing	Dissemination and sharing data	Publication of a OAI-PMH repository according to specific application profiles (PICO, CIDOC_CRM, EDM)	OAI-PMH repository	Software and documentation (GitHub - Zotero)
LIBRARIES					
Collect	Conceptualize / model	Identification of cataloguing rules, metadata standards and thesauri in use in	Adherence to national and internationally recognized cataloguing standards and best	1. Reicat, Catalographic National Code 2. Guideline for cataloguing modern material in SBN	List (Zotero)

		the librarian domain	practices	(Sistema Bibliotecario Nazionale – Italian National Librarian System) 3. Thesaurus of National Library in Florence 4. ISBD: International Standard Bibliographic Description (trad. it) 5. FRANAR: Functional Requirements and Numbering of Authority Records 6. FRBR 7. SKOS	
Connect	Interpretation/ Modeling	Use Metadata Standards for libraries to ensure interoperability with other systems	Adherence to national and internationally recognized cataloguing standards and best practices	1. Dublin Core 2. UNIMARC 3. MAG STANDARD 4. MAG USER MANUAL 5. METS 6. CIDOC-CRM	List (Zotero)

MUSEUMS					
Collect	Conceptualize / model	Identification of cataloguing rules, metadata standards and thesauri in use in the museums domain	Adherence to national and internationally recognized cataloguing standards and best practices	1. ICCD Standards (Italian) 2. VRA Core for the description of works of visual culture (International) 3. SKOS	List (Zotero)
Connect	Interpretation/ Modeling	Use metadata standards to ensure the interoperability with other system	Adherence to national and internationally recognized cataloguing standards and best practices	1. Dublin Core 2. METS 3. PICO AP 4. EDM 5. CIDOC-CRM	List (Zotero) Best practices
Connect / interpret	Convert / transform	Data transformation	Transformation of catalogued data and technical information of digital objects in METS format	METS Generator	Software and documentation (GitHub - Zotero)

Checking data	Convert / transform Validation	Data quality checking	Automatic validation of xml records	METS Validator	Software and documentation (GitHub - Zotero)

2.2.2.8 Librarian that manages a digital library wants to extent the thesaurus, publishing it in SKOS format and mapping it with other SKOS thesauri

ICCU

Authors: Sara di Giorgio, Antonio Davide Madonna

A cultural institution holds a digital collection of art objects that are identified by a controlled vocabulary. To guarantee a wider interoperability and multilingualism, the curator of the digital collection wants to SKOS-ify its vocabulary and map it with the AAT Thesaurus. A web application could upload the local vocabulary in XML format and the AAT Thesaurus in SKOS. The application allows the mapping between the concepts and the publication in the SKOS format of the enriched vocabulary.

Phase	Activity	Task description	Methodological framework	Standards, formats and services	PARTHENOS inputs
Collect	Meta:Assessing	Identification and acquisition of a controlled vocabularies in order to	Reuse of an existing schema		List (Zotero)

		define thesaurus terms.			
Collect	Meta:Assessing	Analysis of existing thesauri at national and international level and creation of a thesaurus for a Digital Library in Skos format via specific tool (thesaurus manager)	Creation of a particular thesaurus	Format: Skos Standard: Thesaurus schema: ISO 25964 Thesaurus Categories: ISO19115/19139 Register Format:ISO19135	Software and documentation (GitHub - Zotero)
Connect / interpret	Enrichment	Creation of relations between thesaurus terms	Specification of SKOS URI (skos:broadmatch; skos:exactmatch; skos:closematch; skos:narrowmatch)		Software and documentation (GitHub - Zotero)
Connect / interpret	Enrichment	mapping thesaurus with other international thesauri for implementing a multilingual extension.			List (Zotero)
Connect / interpret	Enrichment/ Editing	After a check, thesaurus must be updated and / or	Check thesaurus		Software and documentation

		amended			(GitHub - Zotero)
Present	Dissemination: Publishing	Publishing thesaurus	Thesaurus publication	Format: html, xml	Software and documentation (GitHub - Zotero)

2.2.2.9 A TEI Format for Layout and Logical Annotation of Historical Corpora//Berlin-Brandenburgische Akademie der Wissenschaften

BBAW

Author: Susanne Haaf

A literary studies scholar is interested in the way Goethe and Schiller might have influenced one another in their writing. To examine this question, the researcher plans to compare their writings from the period before their contact to the ones created during their friendship. His interest covers very different fields, from linguistic features via the style of writing to the appearance of the texts.
The analysis is to be partially performed by help of computer-based methods. There are already digitized full-texts of most of the writings of these authors scattered throughout the internet — often several digital versions per work. Thus, the best digital version of a text has to be selected and to be converted to a final common format.

Phase	Activity	Task description	Methodological framework	Standards, formats and services	Parthenos inputs
Collect	Capture/ Discovering	Selection of works of primary interest	Qualitative Analysis of Goethe's and Schiller's		

			works wrt the given research topic		
Interpret	Meta-Activities/ Assessing	Definition of the principal digitization guidelines and quality standards the researched texts should fulfill	Research on guidelines of other projects; Evaluation of the tools to use on the texts wrt their requirements on data quality; Evaluation of the intended subsequent use of the project's data and resulting requirements	Project guidelines and documentation of similar projects	DTABf documentation (DTA 'Base format', format for the homogeneous annotation of historical texts in a large corpus)
Connect	Capture/Discovering	Selection of an unambiguous, TEI-based annotation format	Research on the TEI formats (provided by TEI Technical Council and TEI community) which might meet the project's annotation requirements; Where necessary adjustment of a chosen format according to the project's needs; Selection of a suitable metadata	TEI; TEI formats; ODD; Metadata formats (TEIHeader, CMDI, ...)	DTABf

			format		
Connect	Meta-Activities/ Teaching / Learning	Training in usage of the TEI guidelines	Participation in workshops for training in TEI and adjacent formats or technologies	DH Summer Schools (e.g. ESU Leipzig, DHOxSS Oxford, DHSI Victoria)	workshops and training tutorials on TEI and XML technologies
Connect	Capture/ Discovering	Search for digitized versions of the works of interest	Research in online text archives	Metadata and standardized metadata formats	for text research: CLARIN's VLO, DTA
Interpret	Meta: Assessing	Evaluation of the licenses of works found	Gathering of license information; Research on licenses and their restrictions; Exclusion of texts that are not reusable because of their license	Creative Commons	CLARIN's legal helpdesk
Interpret	Meta: Assessing	(if applicable) Comparison of different digitized versions of one work in terms of their proximity to/distance	Selection of the most suitable digitized text version per work	TEI, ODD, RNG	RNG scheme to evaluate proximity of a TEI text to the DTABf

		from the target format			
Collect	Capture: Data recognition	(if required) Text recognition for works of the research corpus where no suitable digitized version existed	OCR + post-correction; manual transcription (if applicable, Double Keying)	Tools for OCR and (semi-automatic post-correction), e.g. Transkribus	CLARIN's PoCoTo; existing data recognition workflows in CLARIN (e.g. DTA workflow)
Connect/Interpret	Enrichment/Annotating | Cleanup	Conversion of the final corpus texts into the target format	Conversion task is performed partially automatically, partially with the help of a tool to support manual annotation, correction, and homogenization of text	oXygen XML Editor; (if applicable) XSLT, Regular Expressions, scripting languages for automatic conversion	DTABf oXygen Framework
Interpret	Analysis/ Content Analysis/ Structural Analysis/ Network	Research: Analysis of the mutual influence of Goethe and Schiller on their respective works during the time of	Analysis of the source texts with different NLP tools and comparison of the respective outcomes	Standard tagsets (e.g. STTS), Standard text corpus annotation format (e.g. CLARIN's TCF)	CLARIN's WebLicht tool chain, DTA's DiaCollo

	Analysis	their friendship			
Present	Dissemination/ Publishing/ Sharing	Provision of text corpus to the public	Publication of the corpus data	Text corpus: TEI (XML) Metadata: TEI; CMDI	Repositories of CLARIN service centers (e.g. at BBAW Berlin); CLARIN's VLO

2.2.3 Heritage and applied disciplines and Archeology[28]

The objective of PARTHENOS is to strengthen the cohesion of research on Heritage from a large variety of applied disciplines including architecture, archaeology and art conservation. In those fields, researchers, conservator scientists and engineers have to collaborate closely to succeed the long-term sustainability of Heritage. There is significant diversity of scientific and technical expertise between these groups, therefore the extended use of standards will support their work to preserve, protect and maintain and, moreover, to enhance the significance of tangible cultural heritage.

To overcome the limited number of standards and, furthermore, to embrace the epistemological questions that arise, PARTHENOS has elaborated six use cases. Two of these are linked to the "standardization of analytical techniques used in the characterization of cultural heritage objects" and are related to data collection, analysis and interpretation of the actual cultural heritage material/artifact. The remaining four are focused on the data documentation, preservation, dissemination and dissemination aspects of the materials/artifacts through questioning "standardized documentation of cultural heritage data and metadata".

With respect to ISO standards related to technical applications in the cultural heritage domain, the draft of a standard mentioned in the CSIC use case of laser cleaning has been recently approved and announced: EN 16782:2016 Conservation of cultural heritage - Cleaning of porous inorganic materials - Laser cleaning techniques for cultural heritage.
Regarding the other standards mentioned in the following use cases, they represent a first approach to be completed later with various actions. For example, as far as the practices of 3D in the Humanities and social sciences are concerned, a PARTHENOS workshop organised by CNR, CNRS & INRIA will be held in Bordeaux from November 30 to December 2. Indeed, as few ISO standards are known or probably used in the field, one of the objectives of the selected use cases on Heritage and applied disciplines is to identify good practices, to guide researchers in their use, and to promote the more useful and valuable ones for the field.
This kind of work on best practices and standards have also been studied in other H2020 projects in the field of heritage like ARIADNE or IPERION-CH, from which PARTHENOS

[28] For this section, the overview has been writenn has been written by Adeline Joffres and Nicolas Larrousse.

will extract some inputs, such as use cases. Other projects will eventually also benefit from this work, like E-RIHS.

2.2.3.1 Standardization of analytical techniques used in the characterization of cultural heritage objects

The following use cases exemplify the necessity of developing standards in the use of analytical techniques for CH objects characterization.

- Multispectral imaging for the in-situ characterization of painted works of art. The results of the multispectral imaging characterization is published in the POLYGNOSIS online knowledge platform

IESL-FORTH

Authors: Demetrios Anglos, Panayiotis Siozos

Multispectral imaging is a method used to produce several images of an object using selected ranges of wavelengths in the electromagnetic spectrum that include and extend beyond the capabilities of the human eye. Due to the simplicity of the method and the low cost of the equipment multispectral images is extensively used from for pigment characterization on painted artworks.

The main objective of the present case is to establish standardized procedures for multispectral imaging analysis in order to accomplish qualitative results and furthermore, to increase the availability and secure sustainability of the data, using the POLYGNOSIS online knowledge platform.

Phase	Activity	Task description	Methodological framework	Standards, formats and services	PARTHENOS inputs
Collect	Capture/ Gathering	Research in digital resources concerning information of the	Simple searching and browsing	Abstract and citation databases (WoS, Scopus etc), scientific journal	Web link to documentation by PARTHENOS

		painting (pigments, drawing etc)		websites, project websites, institutions repositories	partners, other institutions or bibliographic sources, Polygnosis, Documentation (Zotero)
Collect	Capture/ Imaging	Spectral Image acquisition	• focusing adjustments, • reference image acquisition, • intensity adjustment, • background image acquisition	BMP and PNG image file	Images stored in D4science, Diagnosis documentation system
Interpret	Enrichment/ Cleanup	Post processing of images	• background correction, • reference correction • image registration	• BMP and PNG image file format • Image processing software (ImageJ, MATLAB, etc)	Images stored in D4science, Diagnosis documentation system
Interpret	Interpretation/ Contextualizing	Further post processing and analysis,	• production of diffuse reflectance spectrum of a specific area • visual inspection of	• BMP and PNG image file format • Image processing software (ImageJ,	Data stored in D4science, Diagnosis documentation system

			the corrected spectral images	MATLAB, etc) • UNICODE, UTF-8, ASCII	
Interpret/ Present	Storage/ Organizing	Organizing data to prepare a document with the results and the conclusion of the multispectral imaging analysis		• Image processing software (ImageJ, etc) • XML format	Data stored in D4science, Diagnosis documentation system
Present	Dissemination/ Publishing	Publish the results of the multispectral imaging characterization of the paint art in the POLYGNOSIS online knowledge platform		• XML format	Best practices Documentation Guidelines and Stylesheets (D4Science – Zotero) Helpdesk, Polygnosis

- Researcher identifies the convenience of standardized Laser-Induced Breakdown Spectroscopy (LIBS) analysis for glass substrate characterization in the field of Cultural Heritage

CSIC

Authors: Marta Castillejo, Esther Carrasco

A researcher is interested in accessing systematic works related to LIBS analysis of glass substrates and its application to cultural heritage glass artefacts given the technique strengths (speed, absence of sampling and sample preparation, microscopic resolution and in-situ characterization). With that purpose, the researcher first needs to select information for specific analytical studies of glass characterization by LIBS. Then, he/she needs to apply the technique for own research work. The application of LIBS involves the acquisition of their own spectra and their analysis (evaluation of their quality, corrections applications, identification of relevant elements and concentration estimations). A successful output of this work would potentially allow for the glass composition determination, its classification, provenance determination, cromophores and opacifiers detection as well as the evaluation of possible alterations or grade of degradation. All the obtained data from the LIBS analysis are important for the preservation and restoration of glass artefacts in the CH field, if all the relevant information of the characterization is reported systematically by the researcher and evaluated in comparison with the analysis of the same objects performed with other techniques. The usefulness of that research study for CH is linked to the possibility of an effective dissemination of the elaborated report, its access in a sustainable repository and its reuse by other researchers. Finally, the researcher would like to establish contact with people who might be interested in standardization of glass characterization by LIBS.

Phase	Activity	Task description	Methodological framework	Standards, formats and services	PARTHENOS inputs
Collect	Capture/ Discovering	Search and finding info about procedures related to experimental glass characterization by LIBS, analysis of	Simple searching and browsing	Scientific journal websites, project websites, institutions repositories, …	Web link to documentation by PARTHENOS partners, other institutions or bibliographic sources (Zotero)

		obtained data and reporting			
Collect/ Connect	Capture/ Gathering	Collect and select info for the specific analytical case (glass object, experimental set-up, methodology, major or trace glass components identification, stratigraphic analysis, calibration methods, quantitative analysis, …)	Gathering LIBS analysis and glass characterization procedures by keywords	Scientific publications (journal and conference papers, books, …), technical reports, industrial application notes, …	Documentation (Zotero)
Collect	Capture/ Imaging	Spectra acquisition	Laser parameters (wavelength, pulse duration, fluence, repetition rate, …), focusing optics, sample environment, spectrometer	- Data format (ASCII files) - Standard glasses (NIST, DGG, Corning, Breitländer, BAM) for	

			resolution and detector operation conditions (gate delay and integration times) selection	calibration purposes	
Interpret	Analysis/ Visualization	Spectra display	Evaluation of the reproducibility of acquired spectra, detection limit, signal-to-background and signal-to-noise to optimize spectra acquisition if required	Display and processing software from equipment manufacturers	
Interpret	Enrichment/ Cleanup	Pre-processing of spectra after acquisition	Background correction, normalization, internal standard selection	Processing software (commercial and homemade)	
Connect/ Interpret	Interpretation/ Contextualizing	Qualitative identification of elemental	Selected emission line assignments based on reference	NIST atomic spectra database Kurucz database	Links to the Databases, PARTHENOS joint resource registry

		components of glass substrate	data	US Army Laboratory Applied Photonics	
Interpret	Analysis/ Content analysis	Quantitative analysis (composition from major to minor and trace elements) of glass substrate under study	Calibration and calibration free methods application, matrix effects evaluation, uncertainty determination	- Univariate analysis - Multivariate analysis (Partial Least Squares, Principal Component Analysis, Support Vector Regression) - Statistical analysis software of preprocessed data	Guidelines, best practices (D4Science - Zotero)
Interpret	Interpretation/ Theorizing	Glass characterization, origin, conservation	Analysis of own research data in combination with		Guidelines, best practices (D4Science - Zotero)

		state of glass artefact	related published works to extract relevant info and conclusions		Helpdesk
Connect/ Interpret	Interpretation/ Contextualizing	Compare the obtained results by LIBS with results from other analytical techniques employed in the characterization of the same glass object or other glasses of expected similar characteristics	Gathering and evaluating results of glass characterization by combination of LIBS technique with: -Fluorescence (LIF) -Plasma techniques (ICP-OES, ICP-MS) -X-ray techniques (XRF, EDX) -Ion-beam techniques (PIXE) -Raman spectroscopy	Scientific publications, technical reports, industrial application notes, collaborative works, …	Web link to documentation by PARTHENOS partners, other institutions, projects or bibliographic sources (Zotero)
Interpret/ Present	Creation/ Writing	Elaborate a systematic report			Guidelines and Stylesheets (D4Science

		which include the results of the characterization together with all the relevant parameters, procedures and followed methodology.			– Zotero) Helpdesk
Present	Dissemination/ Publishing	Publish the elaborated report.			Support/Feedback from PARTHENOS/Iperion CH community Best practices, documentation
Present	Storage/ Archiving	Keep access to the characterization report.			Feedback from PARTHENOS/Iperion CH community Best practices, documentation
Present	Storage/ Preservation	Facilitate the reuse of the research report.			Feedback from PARTHENOS/Iperion CH community

Interpret/ Present	Meta-Activities/ Community building	Build an interdisciplinary community around standardization of glass characterization by LIBS.			Participation in workshops, seminars, conferences

2.2.3.2 Standardized documentation of cultural heritage data and metadata

The present use cases refer to cultural heritage data presentation, gathering and archiving.

- Conservation scientist wants to publish information about experimental conditions for Raman analysis of wall painting fragments and report in particular proper experimental measurement conditions for safely detecting and identifying certain types of pigments

IESL-FORTH

Authors: Demetrios Anglos, Panayiotis Siozos

Raman spectroscopy is widely used for material characterization in cultural heritage. The results of the characterization process are typically published in scientific journals and textbooks. However, the information that can be included in these manuscripts is limited, the data reuse is inadequate and furthermore the overall process is time demanding.

The present case aims to demonstrate a standardized procedure for reporting and updating proper experimental measurement conditions for safely detecting and identifying certain types of pigments using Raman.

Phase	Activity	Task description	Methodological framework	Standards, formats and services	PARTHENOS inputs
Collect	Capture/ Discovering	Research in digital resources to determine the experimental measurement conditions for Raman analysis of wall painting fragments	Simple searching and browsing	Abstract and citation databases (WoS, Scopus etc), scientific journal websites, project websites, institutions repositories	Web link to documentation by PARTHENOS partners, other institutions or bibliographic sources Documentation
Collect	Capture/ Gathering	Obtain the experimental conditions for Raman analysis from a digital library	Obtain experimental conditions by keywords (laser wavelength, laser power, detector integration time, number of accumulations, etc)	Scientific publications (journal and conference papers, books), technical reports, surveys by institutions or museums magazines	Documentation (Zotero)
Connect	Creation/	Apply the			

	Translation	experimental conditions for Raman analysis of the material			
Connect	Capture/ Imaging	Characterization of the material using Raman spectroscopy.	Perform the analysis on the materials using the Raman spectroscopic instrument		
Collect	Storage/ Identifying	Raman spectra and images are stored and prepared for interpretation		Standard image file formats (tiff, jpg, bmp etc) Data formats from Raman manufacturer (e.g HORIBA Jobin Yvon IBH format). UNICODE, UTF-8, ASCII	Best practices for image file formats Survey and documentation on data formats
Interpret	Interpretation/ Contextualizing	Detection of weak discoloration of the pigment. Clarification of			

		the discoloration.			
Interpret/ Present	Storage/ Organizing	Organizing data to prepare a report about experimental conditions creating the discoloration		Numerical analysis software (MATLAB, GNU Octave). Software for graph preparation (Origin, MATLAB, GNUPLOT, etc) Image processing software (ImageJ, etc)	
Present	Dissemination/ Communicating	Prepare and deliver the report to the authors by using the platform of the digital library		Standard document file formats	Best practices Documentation Guidelines and Stylesheets (D4Science – Zotero) Helpdesk
Present	Dissemination/ Publishing	The report is publicly available from the digital library		Scientific journal websites, project websites, institutions repositories	Feedback from PARTHENOS/Iperion CH community

- Working on 3D formats for archiving and on common metadata

Huma-Num

Authors: Adeline Joffres, Nicolas Larrousse

The general context is the creation of 3D Huma-num consortium in 2013. It aims at sharing good practices and developing common standards and tool for creating, processing and archiving 3D objects.

Phase	Activity	Task description	Methodological framework	Standards, formats and services	PARTHENOS inputs
Collect	Capture/ Imaging	Acquisition of 3D object (by scanning)	Scanning system: -Triangulation based artefact scanners -Terrestrial time of flight laser scanners -Airborne laser scanning -Mobile Mapping systems Methodological framework: -Data acquisition -Register scan -Georeference scan -Creation polygonal mesh -Optimization/Decimation mesh	Proprietary formats	Providing a survey of existing formats and documentation on them

Collect	Capture/ Imaging	Acquisition of 3D object (by photogrammetry)	Photogrammetry device and related software Photogrammetry device: -Digital SLR camera -Bridge camera -Compact Camera -Mobile phone -UAV photogrammetry Related software -Agisoft photoscan -Accute 3D -MICMAC -Reality Capture -Visual SFM -123 catch Methodological framework photogrammetry -Data acquisition (overlapping photographs)	Formats: OBJ, PLY, 3DS, VRML, COLLADA, DXF, JPEG, TIFF, PNG, BMP	Providing a survey of existing formats and documentation/information on them

			-Calibration -Align photo -Point cloud creation -Mesh generation -Texture generation		
Connect/ Interpret	Interpretation/ Modeling	Migration to target format	Specific software: -3ds max -Autocad -Google sketchup -Blender -Maya	OBJ, DAE, DXF, DWG,VRML	Analyse and improvement on tools
Interpret	Interpretation/ Modeling	Verification of quality data	Human and technical verification on content and structure	PLY and COLLADA formats	Improvement of documentation and verification tools
Present	Dissemination/ Publishing	Visualisation	Viewer tools, specific software: -A360 viewer -CC viewer -Potree -Sketchfab -3D Hop		Feedback from PARTHENOS 3D community and registry of existing viewers

			-Cryengine -Unity		
Present	Storage/ Archiving	Find a common and interoperable metadata format	Work based on existing Europeana format for metadata -3D coform -3D icons	Enhanced CARARE (Europeana)	Feedback from PARTHENOS 3D community
Present	Storage/ Preservation	Agree on a "pivot" format with "acceptable" information losses of archiving 3D objects	Collective dialogue within different disciplines of 3D Huma-Num consortium	PLY and COLLADA formats	Feedback from PARTHENOS 3D community
Interpret/ Present	Meta-Activities/ Community building	Building an interdisciplinary community around 3D topics	3D Open source tools -Meshlab -3Dhop -Blender -Cloud Compare -Google sketchup -123 catch -MicMAC -open flipper		Participation in workshops, seminars, conferences and any initiative, meeting or general reflexion initiated by the 3D HN consortium / Scientific collaboration

			-CC viewer		

- A dataset for the products used in conservation treatments in order to share information about their application parameters, their effectiveness and their durability in time, related to the type of material and its state of conservation.

CNR-ICVB

Authors: Rachele Manganelli, Marco Realini

Among the goals pursued by research institutions and professionals involved in the protection and conservation of Cultural Heritage (CH), there is the efficacy and the durability in time of the conservation treatments, carried out for buildings or objects of artistic and historical importance. In literature it is possible to find many examples to verify the efficacy of the treatments carried out on buildings, artifacts or laboratory tests on the durability of products subjected to accelerated aging cycles, but no tools exist that are able to "correlate" the performance of protective or reinforcement treatments and their durability in time, with different kinds of materials or substrates, decay and climatic conditions to which they are subjected.

In other words, it is not possible to obtain in a direct and simple way, but still based on scientific data, the most appropriate products for a particular material, exposed to specific environmental conditions.

The aim is to create:

- a dataset containing information about a large series of treatments carried out on site and in the laboratory, from which the benefits provided by various products in different situations can be deduced.
- a system that is able to describe for each material, the investigations performed, the obtained results, and its conservative history (present and past works).

- A large collection of data about conservation treatments in order to advise planners and restorers to identify the most suitable products for the conservation of artifacts under certain conditions (environment, decay), and choose the best treatment methodology (application technique, application time, concentration).

The diagnostic techniques shows are just one example. There are many techniques to be applied **in situ** and in laboratory to determine the characteristics of the material and their state of Conservation.

Phase	Activity	Task description	Methodological framework	Standards, formats and services	PARTHENOS inputs
Collect	Capture/ Discovering	Searching and finding data about conservation treatments	Searching and browsing	• Scientific journals, publications, conference papers, books, technical reports, surveys... • Project websites • Institutional repositories	Web links to documentation provided by PARTHENOS partners and other institution Bibliographic sources (Zotero) Community (Experts)
Collect/ Connect	Capture/ Gathering	Collect and select information for the specific research case	Gathering information on conservation treatments and products for a particular material	Scientific journals, publications, conference papers, books, technical reports, surveys...	Documentation (Zotero)
Collect/	Analysis/	Physical, chemical,	Portable Fiber Optic	ASCII	Documentation

Connect	Content analysis	minerological and petrographic characterization of materials and definition of its state of conservation	Reflectance Spectroscopy (FORS)		(Zotero) Guidelines, best practices (Zotero, D4S) Helpdesk
			Portable X-Ray Fluorescence Spectroscopy (XRF)	ASCII	
			FTIR and Raman Benchtop and Portable Spectroscopy	ASCII	
			Benchtop X-Ray Diffractometry	ASCII	
			Benchtop Polarized Light and Reflected Light Microscopy	2D images: JPG	
			Benchtop Mercury Intrusion Porosimetry	ASCII	
Collect/ Connect	Interpretation/ Theorizing	Evaluation of the environmental parameters that have most influence on the	• Data registered from direct sensors on site (thermo-hygrometric	ASCII	Weblinks to databases available online

		degradation of constitutive material	variations, concentration of pollutants, rain, wind, solar radiation) • Data gathered from databases avalaible on line		
Collect/ Connect	Analysis/ Content analysis	Determine the characteristics of the surface and of the material before treatment	Portable Optical Microscopy	2D images: JPG	Documentation Best practices, guidelines
			Portable colorimetry/Spectrophotometry	CSV	
			Benchtop Mercury Intrusion Porosimetry	ASCII	
			Portable Contact Sponge Method	CSV	
			Portable Peeling Test Device	ASCII	

			Portable Ultra-close range photogrammetry	ASCII	
Interpret	Meta-Activities	Monitoring the behavior of the applied product to determine its durability and its effectiveness			Documentation Best practices, guidelines
Interpret	Interpretation/ Theorizing	Choosing the best product and treatment methodology			Best practices, guidelines
Present	Storage/ Archiving	Make available datasets containing a large series of treatments carried out on site and in laboratory	Enter new data in a repository/ a database		Repository provided by PARTHENOS Weblinks to databases provided by partners Documentation

- Researcher using lasers in conservation/restoration identifies the necessity of standardized reports of the laser application conditions and the evaluation of the obtained results

CSIC

Authors: Marta Castillejo, Esther Carrasco

A researcher is interested in having access to procedures related to the use of lasers in conservation and restoration of artworks, including the possibility of effectively selecting information for specific cases of laser cleaning in conservation/restoration. He/she needs to have the capacity to obtain the relevant parameters of laser-material interaction in the case of interest and to assess the effects of the laser treatment (physicochemical modifications, undesired side effects or collateral induced damage) based on systematically documented published works and guidelines or best practices. These capacities are the first objective. The data will facilitate the elaboration of a standardized report of his/her own research study, which is the second aim. The usefulness of this report in the CH field is linked to the possibility of its effective dissemination, its access in a sustainable repository and its reuse by other researchers. Finally, the researcher would like to establish contact with people who might be interested in standardization of laser use reporting in cultural heritage.

Phase	Activity	Task description	Methodological framework	Standards, formats and services	PARTHENOS inputs
Collect	Capture/ Discovering	Searching and finding info about procedures that use lasers for conservation and restoration of artworks and heritage objects and substrates	Simple searching and browsing	Scientific journal websites, project websites, institutional repositories, …	Web link to documentation by PARTHENOS partners, other institutions or bibliographic sources (Zotero)
Collect/ Connect	Capture/ Gathering	Collect and select info for the specific research case (object, substrate, …)	Gathering cleaning procedures by keywords (materials, laser types, techniques of analysis, induced effects, …)	Scientific publications (journal and conference papers, books, …), technical reports, surveys by institutions or museums	Documentation (Zotero)

				magazines	
Connect/ Interpret	Analysis/ Content analysis	Obtain the relevant parameters related to the conditions of the laser-material interaction in the CH conservation/ restoration area of interest	Work can be based on published studies where parameters of laser conditions are identified and listed	Tables, graphs, schema and images obtained from the bibliographic study	Documentation (Zotero) Helpdesk
Interpret	Interpretation/ Theorizing	Assess the physicochemical effects of the laser irradiation, the undesired side effects or collateral induced damage and their systematic documentation for own research work.	Work can be based on published studies where laser effects are systematically documented		Guidelines, best practices (D4Science - Zotero) Helpdesk
Interpret/ Present	Creation/ Writing	Elaborate a standardized report		CEN/TC 346. FprEN 16782	Guidelines and Stylesheets

		according to the previous phases and applied to own research study.		Conservation of cultural heritage - Cleaning of porous inorganic materials - Laser cleaning techniques for cultural heritage (standard under approval)	(D4Science – Zotero) Helpdesk Community: CEN/TC
Present	Dissemination/ Publishing	Publish the elaborated standardized report.			Support/Feedback from PARTHENOS/Iperion CH community Best practices, documentation
Present	Storage/ Archiving	Keep access to the created research report.			Feedback from PARTHENOS/Iperion CH community Best practices, documentation

Present	Storage/ Preservation	Facilitate the reuse of the created research report.			Feedback from PARTHENOS/Iperion CH community
Interpret/ Present	Meta-Activities/ Community building	Build an interdisciplinary community around standardization of laser use reporting in cultural heritage.			Participation in workshops, seminars, conferences

2.2.4 Social sciences[29]

The field of the social sciences is inherently focused on people and their relation to society. As such, data is often gathered on individuals. This can be done in numerous ways, of which one prominent approach is to conduct a survey. Inventorying and archiving the survey is an important aspect in studies that employ surveys, and one in which PARTHENOS can provide valuable input. How this can be done is presented in the first use case by Huma-Num. Often, a next step in the research process is to (re)code, and refine the data gathered and perform statistical analysis and modelling techniques on the data. Several tools and standards are commonly used in this step, of which some well-known standards and tools are presented in the second use case, by DANS.

[29] The Social Sciences use cases have been presented and analyzed by Emilie Kraikamp (DANS-KNAW).

2.2.4.1 Platform for inventorying and archiving field surveys in political science (political sociology)

Huma-Num

Authors: Adeline Joffres, Nicolas Larrousse

The archiPolis Huma-Num consortium was labeled as such in 2012.

The main mission of this consortium is to develop a collective strategy for inventorying, collecting, preserving — through digitization, for older surveys, which have not been entered in digital format — and define common metadata of field surveys conducted by political scientists, sociologists and other social scientists interested in the political subject. This is to make these investigations intelligible through documentation and commissioning a consistent context. The objective is indeed to avoid depletion or even abuse of the research work that could lead to the conservation and availability of data out of context.

Phase	Activity	Task description	Methodological framework	Standards, formats and services	PARTHENOS inputs
Collect	Meta activities/ Community building	Initiating and coordinating a "grand dialogue" between the partners of the consortium to agree on a common description on various surveys	- Meetings (virtual and face -to-face), - participating in conferences and organising research events	DDI, METS	Documentation, best practices, guidelines
Collect	Creation/ Programming	Work on an adaptation of DDI format for qualitative	- Meetings (virtual and face -to-face)	DDI	Documentation, best practices, guidelines

		surveys in political science and sociology	- Partnership with BE Quali		(Zotero)
Interpret/ Present	Dissemination/ Sharing	Gathering, editing and normalising social science surveys collected and transferring them in Be Quali repository.	Setting up a processing chain	DDI, METS, TEI	Best practices and guidelines
Interpret/ Present	Storage/ Preservation	Work with Huma-Num on the creation of packets for archiving following the OAIS recommendations.	- Work based on CINES recommendations - Setting up a processing chain for a submission to CINES, the French Archiving Centre	DDI, METS, TEI , Dublin Core	Documentation, best practices, guidelines (Zotero)

- A researcher processes raw data on historical and contemporary occupations and performs statistical analysis and modelling techniques on these data

DANS-KNAW

Author: Emilie Kraikamp

Given a large dataset with several raw variables, a social science researcher needs to (re)code some of the data in order to properly conduct statistical analysis and modelling techniques. To allow comparison across studies, researchers often develop and use certain standards for recoding some aspects of individuals, such as recoding contemporary occupations into the ISCO. After (re)coding variables, scales are often constructed by using standard methods such as factor and reliability analysis. Several software packages exist, such as SPSS Statistics, that are widely used for these purposes across social science researchers. Statistical analyses and modelling also have certain software standards. Finally, reporting the results is done within a consistent style, such as APA Style. Some of these coding, software and reporting standards are covered in this use case.

Phase	Activity	Task description	Methodological framework	Standards, formats and services	PARTHENOS inputs
Interpret	Interpretation/ Contextualizing	Recode occupation variable into a class and status score for occupations in the 19th and 20th century, and for contemporary occupations.		HISCO and Leeuwen, M. V., Maas, I., & Miles, A. (2002). **HISCO: Historical international standard classification of occupations**. Leuven: Leuven University Press. HISCAM	

				Lambert, P. S., Zijdeman, R. L., Van Leeuwen, M. PARTHENOS., Maas, I., & Prandy, K. (2013). The construction of HISCAM: A stratification scale based on social interactions for historical comparative research. **Historical Methods: A Journal of Quantitative and Interdisciplinary History**, **46**(2), 77-89. for the 19th and 20th century. (standards) ISCO The International Standard Classification of Occupations 2008 (ISCO-08). (2012). International Labour Organisation, Geneva. for contemporary occupations (standard)	

Interpret	Interpretation/ Contextualizing	Combine variables into meaningful scales	Exploratory Factor Analysis (EFA)	SPSS Statistics standard (tool) http://PARTHENOS-03.ibm.com/software/products/nl/spss-stats-standard	
Interpret	Meta:Assessing	Testing for reliability of the scales	Reliability Analysis	SPSS Statistics standard (tool) http://PARTHENOS-03.ibm.com/software/products/nl/spss-stats-standard	
Interpret	Analysis/ Structural Analysis	Statistical modelling of the relationship between variables	Structural Equation Modelling (SEM)	SPSS AMOS (tool) http://PARTHENOS-03.ibm.com/software/products/nl/spss-amos	
Interpret	Analysis/ Structural Analysis	Statistical analysis	Calculate p-values for significance, by using Analysis of Variance (ANOVA) or regression-	SPSS Statistics standard http://PARTHENOS-03.ibm.com/software/products/nl/spss-stats-standard	

			analyses.		
Present	Creation/ Writing	Reporting on the research		APA referencing (standard) American Psychological Association. (2010). Publication manual of the American Psychological Association (6th ed.). Washington, DC: Author. EndNote (tool) http://endnote.com/	

A promising standard for describing and documenting social science data is the **Data Documentation Initiative (DDI)**, which complies with the widely used metadata standards of Dublin Core. The DDI standard is freely available and can be used to describe observational (e.g. survey) data from several fields in the social sciences. Using this standard fosters high-quality metadata and improves the understanding of the data by both humans and computers. Work is being done to broaden the scope of the DDI format for qualitative surveys in political science and sociology. **METS (Metadata Encoding & Transmission Standard)** is another standard for encoding metadata of objects in digital libraries. Similarly, the **TEI (Text Encoding Initiative)** has a set of widely used guidelines that specify encoding methods for machine-readable texts and is used in multiple fields including social sciences.

However, for optimal (re)use of data not only the metadata, but also the data itself and methods of working with the data can benefit from standardization. One prominent example of a standard in the recoding of data is the recoding of occupations into the **International Standard Classification of Occupations (ISCO)**. The use of such standard classification schemes allows the conducting of international comparative research on social stratification.

Further data refinement and analysis can be done via statistical software. **SPSS** is the leading statistical software for social sciences. Due to its intuitive graphical user interface it is accessible for a large variety of researchers, while also allowing for more command-driven ways of working via syntax.

After data analysis, results and conclusions need to be reported, often in a research paper. Several reporting and referencing styles exist for such papers, of which the **APA (American Psychological Association)** style is a prominent standard. Reference managers assist in multiple ways for writing these papers. **EndNote** is widely-used reference manager that is compatible with many word processors.

2.3 New use cases developed by WP4

2.3.1 Studies of the past

2.3.1.1 Interoperability between archival description formatted in XML-EAD (EHRI)

The European Holocaust Research Infrastructure (EHRI) support Holocaust researchers by providing online access to information about dispersed sources relating to the Holocaust through its online portal (http://portal.ehri-project.eu) This portal has put together descriptions from more than 1,800 institutions. The search and discovery aids come in a variety of formats, not necessarily in EAD or even in XML. There is a need to map all these data before their ingestion in the portal database. This mapping must use XML-EAD as its target format, because this is the format that the EHRI database is able to process. However, the EHRI database has its own constraints. This use case proposes a workflow to ease the maintenance and interoperability of archival description expressed in XML-EAD.

We created an ODD instance of the XML-EAD guidelines and schema. ODD ("One document does it all") is the TEI documentation and schema-definition format that "includes the schema fragments, prose documentation, and reference documentation (...) in a single document"[30]. It can generate schemas (DTD, Relax NG, XML Schema), HTML documentation and, if necessary, Schematron rules. This file is available from the PARTHENOS WP4 GitHub repository[31].

More importantly, this ODD file can be used as a basis for the customization of the XML-EAD format, by adding additional constraints on the structure and the content of the resulting representations. With this method, we transformed the EHRI guidelines in a specific flavour of EAD, an EHRI-EAD schema, 100% compliant with the EAD specifications, with additional rules expressed in the Schematron format. The two types of rules are the ones related to the presence or the absence of elements, and the ones

[30] http://www.tei-c.org/Guidelines/Customization/odds.xml, last seen September 23rd 2016
[31] https://github.com/PARTHENOSWP4/standardsLibrary/blob/master/archivalDescription/EAD/odd/EADSpec.xml

related to the value of the elements (regular expressions). This schema is used in EHRI's mapping tool.

The ODD customization, as well as the corresponding Relax NG schema and HTML documentation can be found here: https://github.com/charlesriondet/data-validations/tree/master/ODD-RelaxNG

2.3.1.2 How to easily create and manage a prosopographical database EAC-CPF with the CMS Omeka

The increasing amount of digital data in the Arts and Humanities has introduced new challenges to simplify the access, the management and the interoperability of prosopographical data. In particular, the fact that data is shared among different types of authorities: user, researcher, teams, etc., having different skills and backgrounds, makes the problem more complex. The proposed solution is a file management tool, with publishing and interoperable capabilities that can be handled without a steep learning curve.

The goal is to deploy a dataflow and a storage system able to ingest and produce authority files in different markup formats (XML, HTML, CSV, etc) supporting different standards (Dublin Core, FOAF, TEI, EAC-CPF, etc.) without requiring any special operation on the part of users. The work focused on a very specific use case in the context of historical studies. The authority records were 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[32].

[32] The EAC-CPF standard is maintained by the Society of American Archivists in partnership with the Berlin State Library http://eac.staatsbibliothek-berlin.de/

2.3.1.3 Music notation and open source software: how to build a musical corpus in MEI standard

Huma-Num

Authors: Adeline Joffres, Nicolas Larrousse

Huma-Num's Consortium Musica aims to coordinate the debate on digital musicology in France, and to develop new technologies for representing and analyzing music**.** Following MEI standards, Consortium Musica's tool system (mainly based on the CMS Omeka) permit to realize online and open source musical editions.

Phase	Activity	Task description	Methodological framework	Standards, formats and services	Parthenos inputs
Collect	Transcription/ Editing	Creating a corpus of musical compositions on Sibelius	• Transcriptions from original musical sources, • adding critical editorial signs, • normalizing, where applicable, ancient poetic texts to modern usage	Sibelius	
Collect	Annotating	Collecting and organizing metadata according to DublinCore standards	• Creating an Omeka database about people, places, sources, etc. related to the musical edition	DublinCore, Omeka	
Connect	Annotating	Enrichment of musical edition with critical apparatus metadata	• Adding to music edition one staff for each source attesting variants or person suggesting emendation	Sibelius + 'Wagner files' protocol	

			• encoding variants/emendations only in the measures where interventions occur		
Interpret	Conversion	Format change of Sibelius files into MEI files	• Converting the Sibelius files into MEI files through the plugin SibMei	Sibelius, Plugin SibMei	
Interpret	Cleanup/ Conversion	Enrichment of MEI files	• Correcting errors occurred during the conversion process, • enriching through JavaScript libraries • converting MEI files into drawing instructions for the VexFlow online music notation rendering API (MEItoVexFlow)	MEI Massaging, MeiView, MEItoVexFlow	To be discussed
Present	Publishing	Present in an open source digital CMS the result of the musical edition	• Publishing on Omeka + VexFlow (if the published musical corpus is a critical edition), • publishing on Omeka + Verovio (in all other cases)	Omeka, Verovio, VexFlow	

2.3.2 Language-related studies

2.3.2.1 A new TEI format: the “DTA Basis-format” or DTABf (FHP)

The Deutsches Textarchiv project (German Text Archive; DTA)[33] has developed a specific TEI format, the “DTA Basisformat” (DTA 'Base format'; DTABf)[34] for the annotation of historical corpora. It has been created on the basis of the Deutsches Textarchiv corpus (16th-19th century) and applied to more than 3,000 corpus texts of various genres, disciplines, authors and publishers, originally digitized by the DTA or curated from various external digitization or edition projects. The goal is to allow for homogeneous text annotation and thus ensure interoperability of all DTABf annotated texts. This is achieved by reducing the extensive TEI P5 tagset to a much smaller subset with determined attribute and value sets. The DTABf not only comprises a tagset and guidelines for text annotation but also for metadata recording.

The format currently consists of four components:

- A TEI ODD[35] and an RNG schema[36] derived from that ODD
- A set of accompanying Schematron constraints[37]
- Detailed documentation with various examples from the DTA corpus[38]
- A large number of DTABf texts within the DTA corpora, freely licensed and available for download via the DTA website[39]

The DTA 'Base format' has been recognized as a reference format by the Deutsche Forschungsgemeinschaft (German Research Foundation; DFG)[40] and by CLARIN[41], and the DTA hopes it will also be recognized at an European level. In collaboration with WP4, the Deutsches Textarchiv will create a TEI-ODD which will evaluate the distance of a TEI document from the DTABf. It will also provide an extensive English documentation for the DTABf.

[33] http://www.deutschestextarchiv.de
[34] http://www.deutschestextarchiv.de/doku/basisformat_en; https://jtei.revues.org/1114
[35] http://www.deutschestextarchiv.de/basisformat.odd
[36] http://www.deutschestextarchiv.de/basisformat.rng
[37] http://www.deutschestextarchiv.de/basisformat.sch
[38] http://www.deutschestextarchiv.de/doku/basisformat_en; https://jtei.revues.org/1114
[39] http://www.deutschestextarchiv.de/dtag/book; http://www.deutschestextarchiv.de/list
[40] http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/foerderkriterien_editionen_literaturwissenschaft.pdf
[41] http://de.clarin.eu/en/help/user-handbook, II.6.Text Corpora

The use case describes a typical usage scenario for the DTABf as it is and for the further developments planned in PARTHENOS[42].

2.3.2.2 Transcription of speech

For quite some time the speech transcription data landscape was determined by the various tools[43] available for researchers to transcribe their audio or video resources. Each tool, even when offering similar functionalities and data models as its competitors, had its own storage format for storing the transcriptions and annotations created with it. Even though since an early version (e.g. TEI P3), the TEI Guidelines had contained a specific chapter dedicated to the transcription of speech[44], the fact that it was not prescriptive enough or that at times it even failed to offer such relevant constructs as those needed for interlinear annotation prevented it from getting widely acknowledged and adopted.

This is the context in which a new project was started within ISO committee TC 37/SC 4, under the leadership of Thomas Schmidt (Institut für Deutsche Sprache) to design a reference customisation of the TEI guidelines to identify the reference features needed for transcribing spoken resources that are anchored on a single reference timeline, but also integrating mechanisms to encompass most usual transcription conventions. The standard[45], published in 2016, is already implement in EXMARaLDA[46] and is already recognized as a major step forward for the field.

In the sense of the general standardization workflow that we have identified for the Arts and Humanities, we are in a typical post-publication situation where the main emphasis should be put on providing support for the adoption of the standard as well as communicating widely about its availability.

Within PARTHENOS, the work planned for this use case will cover the following activities:

- Gathering information about existing corpora and identify reference samples that could be wider disseminated as best practice;

[42] See the appendix below: “Cockburn description of the DTABf use case”.
[43] See the seminal paper by T. Schmidt (2011): http://jtei.revues.org/142
[44] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TS.html
[45] ISO 24624:2016 Language resource management -- Transcription of spoken language
[46] http://www.exmaralda.org

- Work towards the provision of additional tools for the conversion of existing legacy formats, the connection to automatic annotation mechanisms and the (possibly lossless) Import to existing tools;
- Identify or develop the necessary components to enable better query, visualisation or browsing of ISO compliant resources, as well as facilitated software developments (schemas, XSLT libraries, API).

2.3.2.3 Towards terser guidelines for the representation of digital dictionaries and the interchange of lexical data

The proper management of lexical information is an essential aspect in various domains in the Humanities. It is the basis for most natural language processing tasks but also takes a strong role in tackling various linguistic, literary, philological or historical tasks. Lexical data comes mainly in two forms:

- Semasiological (word to sense), when lexical forms are associated to the various senses they can have in a given language. This is the traditional form of human readable dictionaries;
- Onomasiological (concept to term), when, for a given concept, the various possible expressions used to refer to it in several languages are recorded. This is the basis for terminological work but also for standard dialectology work.

Both domains have been well covered in the recent years by standardization work:

- The ISO 24613 (ISO LMF) standard provides a model for semasiological data and the TEI Guidelines have included for many years a specific chapter for “Dictionaries”. Still, both standards are not so well aligned and lack at times a good coverage of essential components such as etymology;
- Two ISO standards (16642 and 30042) provide a background for the modelling and serialisation of terminological information with the potential to cover wider use cases. Still, no mirror chapter has been available in the TEI guidelines since the nineties (when the ancestor of ISO 30046 was actually initiated) and there is a lack of actual guidance for scholars in the Humanities and of an easy off-the-shelf onomasiological format.

Recent discussions among WP4 partners, but also within the several related groups (COST ENe-L[47], CLARIN Standards Committee, DARIAH WG Lexical Resources, TEI Linguistic SIG), have shown that the appropriate time has come to work on a strong improvement of the lexical standards landscape at the service of scholarship and with the underlying vision of a) a better convergence between ISO activities and the TEI guidelines and b) eliciting more precise constraints related to the use of the TEI guidelines for lexical information.

This can be further articulated into the following action plan which will be further implemented in the forthcoming phase of the PARTHENOS project:

- Participation in the ongoing revision of the ISO LMF standard, with a particular focus on the provision of a TEI serialisation and the definition of a new part dedicated to diachrony and etymology;
- In collaboration with the COST action ENeL, defining some precise guidelines for prototypical constructs in dictionary entries that would serve as target practices for optimising interoperability across lexical data bases;
- Explore the possibility for machine learning technique to automatically extract structured information from legacy print dictionary, with the setting up of a PhD on the subject;
- Finalizing a proposal to incorporate an onomasiological data model in the TEI guidelines, inspired by ISO 30046.

This is the context in which a workshop[48] will be jointly organized between ENe-L and PARTHENOS with a close liaison with the above-mentioned group to finalize precise recommendations as to the optimal representation of (semasiological) lexical data using the TEI guidelines.

[47] The Working Group “Retro-digitized dictionaries” of the COST Action “European Network of e-Lexicography” or ENe-L aims at setting up guidelines and standards for turning paper dictionaries into digital format.

[48] “Toward Best Practice Guidelines for Encoding Legacy Dictionaries” to be held at the Berlin-Brandenburg Academy of Sciences (BBAW), November 17-19, 2016.
http://www.elexicography.eu/events/workshops/wg2-berlin-2016/

2.4 Cross-use case analysis: data types and associated standardization strategies

2.4.1 Generic resources and tools

2.4.1.1 ISO codes

In any project, the use of standardized ways to represent countries, languages or scripts is recommended, if it's not mandatory. Most of the time, the ISO has adopted a coding format and has entrusted an institution to compile and maintain the list of these codes. However, it's sometimes difficult to get an up-to-date version of the lists that can be used and reused in sustainable way. It is the goal of the SSK to provide such a sustainable resource. Until now, PARTHENOS WP4 has created and maintains three resources related to ISO codes.

2.4.1.2 ISO 639 International Standard for language codes

The IANA, Internet Assigned numbers Authority, is responsible for maintaining many of the codes and numbers contained in a variety of Internet protocols, including a Language Subtag Registry, respecting the ISO 639 standard. The SSK provides a XSL stylesheet that transforms the Language Subtag Registry (accessible with the URL http://www.iana.org/assignments/language-subtag-registry/language-subtag-registry) in XML-TEI, and the corresponding XML-TEI output file. The TEI output must be regenerated every once in awhile. It is recommended to keep the name ISO639_TEI.xml.

2.4.1.3 ISO 3166 International Standard for country codes

The XML file ISO3166_TEI.xml is the expression in TEI of the country codes list. It needs to be updated each time a new code is created.

2.4.1.4 ISO 15924 Codes for the representation of names of scripts

The SSK provides an XSL stylesheet that transforms the Alphabetical list of four-letter script names (provided by the Unicode consortium, ISO 15924 Registration Authority, and accessible at the URL http://www.unicode.org/iso15924/iso15924-text.html) in XML-TEI, and the corresponding XML-TEI output file. The TEI output must be regenerated every once in a while. It's recommended to keep the name ISO15924_TEI.xml.
These resources are located on Github:
https://github.com/PARTHENOSWP4/standardsLibrary/tree/master/ISO

2.4.2 Text mining

Text mining refers to a set of techniques used to derive high-quality information from text. It also involves the process of text structuration.

2.4.2.1 GROBID (GeneRation Of BIbliographic Data)

GROBID (https://github.com/kermitt2/grobid) is a machine learning library for extracting, parsing and re-structuring raw documents such as PDF into structured TEI-encoded documents. The input documents are preferably related to technical and scientific domains: scholarly documents, technical manuals and patents. But text with layout information (PDF) or raw text can be processed as well. GROBID also offers normalization of metadata, by the exploitation of external bibliographical databases for correcting/completing results based on extraction results. The extraction covers the usual bibliographical information (e.g. title, abstract, authors, affiliations, keywords, etc.), the references (footnotes), names, affiliation and address blocks, dates, full text structures (section title, paragraph, figures, …).

GROBID is used by Research institutions such as NASA and CERN, as well as scientific papers repositories, Mendeley and Researchgate. On the latter platform, everyday, thousands of PDF are loaded either by Researchgate users or by crawlers on Open Access archives. The “acquisition” document workflow integrates GROBID for citation extraction. 300K PDF documents are processed every months on a cluster of 16

machines. Extracted citations are matched against an internal bibliographical database. This service offers citation notifications for researchers and relevance ranking in search.

GROBID is a very powerful tool that can be used in many contexts, including research in Humanities where a retro digitization of scientific documents is needed. The annotations that GROBID generates are also fully compatible with the TEI, and can be part of the reflexion on standardized representation of scientific documents.

2.4.2.2 NERD (Named Entity Recognition and Disambiguation)

"In data mining, a named entity is a word or a phrase that clearly identifies one item from a set of other items that have similar attributes. In the expression **named entity**, the word **named** restricts the scope of entities that have one or many rigid designators that stands for a referent. Usually, Rigid designators include proper names, but it depends on domain of interest that may refer the reference word for object in domain as named entities[49]."
Disambiguation is the process of resolving a Named Entity mention by identifying the correct entity the mention refers to, using existing knowledge bases. The typical example is Francis Bacon, a mention that refers to two entities, the philosopher and the painter.

The NERD technology is a useful tool for many scenarios. In the scope of WP4, the NERD has to have two essential features. The first is the ability to be called in a standardized way, with an API (or similar services) accepting many standardized formats as input. The second is fundamental. The output text of a NERD service must respect a standardized annotation format. In that respect, a step forward is currently made with the incoming implementation in the NERD (service developed by Science Miner[50]) of the TEI Stand Off annotation model (currently under discussion within the TEI technical council), based on the Open annotation standard developed by the W3C.

[49] Rahul Sharnagat, **Named Entity Recognition: A Literature Survey**, 2014 (http://www.cfilt.iitb.ac.in/resources/surveys/rahul-ner-survey.pdf, accessed October 13th 2016)
[50] http://cloud.science-miner.com/nerd/, accessed October 14th 2016

2.5 Towards a roadmap for standardization activities in the humanities and cultural heritage research – metascenarios

The work carried out on the precise identification of relevant standards and data formats within research processes in the Humanities shows that there are some strongholds that PARTHENOS should further explore in order to provide better guidance and support. The various summaries of the previous sections allow us to put forward the following elements that should be at the centre of our future activities:

- Basic guidance and training documents for simple metadata formats such as METS and MODS, used in various digital library settings (in particular in audio and video repositories) and that scholar are likely to come across when dealing with such material;
- The provision of concise guidelines for the use and publication of archival data (EAD in particular) so that scholarly work is part of a continuous workflow from digital archival collection descriptions to enriched archival content;
- Increase the support of the TEI guidelines for computational linguistic scenarios and, in particular, in the domain of basic sources transcription (comprising spoken sources), stand-off annotation of primary documents and also all domains related to lexical information;
- Accelerating the agenda for the definition of proper standards for cultural heritage material, ranging from identifying a good compromise for the representation of 3D objects to the launch of specific initiatives for the standardization of material analysis procedures.

This overview provides the premises of a real roadmap of action for the second phase of PARTHENOS, but also in the long term for future standardization developments within CLARIN, DARIAH and the future E-RIHS. For each domain, and taking into account the forces and thematics associated to the the subtasks 2, 3 and 4 of WP4, targeted actions will be implemented to push the agenda forward, leading in turn to the elaboration of the future deliverable D4.2.

3 Supporting: the web interface of the Standardization Survival Kit

3.1 General workflow

The work realized on the D2.1 use cases by WP4 members is a first step in the elaboration of the Standardization Survival Kit or SSK. The SSK aims to be a useful tool for researchers, helping them in using standards in their projects. Based on WP4's work on the D2.1 use cases, it will take the form of a website that will:

- Teach researchers to use standards;
- Answer their questions on standards;
- Guide them in using standards;
- Provide them information on standards.

Before diving into design aspects, it is essential to get a clear overview of the future users of the SSK: who they are? what would motivate them to browse the SSK? what are their needs? how do they work and how can we adjust to their methods while improving them by adding standard practices? The process of addressing these questions is called User Research. The purpose of UX or user-centered design, consists in designing to meet the needs of the end user, and User Research allows us to know who that person is, in what context he/she will use the tool or service, and what he/she needs to achieve his/her goal(s).

To have more information on the researchers' needs, the work package used the work done by WP2 in the deliverable 2.1 on users' requirements. The WP4 team focused more specifically on standards that can and should be used by researchers in the Arts and Humanities, and tried to answer the following question: how can researchers work with standards more effectively thanks to the new inputs provided by PARTHENOS? This work highlighted the necessity to focus on the researchers' activities within a project, seen as successive steps requiring particular standard inputs. The activities described in the use cases are based both on the PARTHENOS Vision and on the TaDiRAH. Along with the kind of data used and analysed by the researcher (texts, datasets, artefacts), and the discipline the research project is in, it was also important to focus on what is done in concrete terms, what are the specific tasks carried out and if they can be gathered inside a

general category such as the ones proposed in TaDiRAH. These three concepts (research data, research domain, research activity) should be the main entry points of the SSK's interface, or more precisely, the facets for the navigation.

3.2 Building the information architecture

3.2.1 The taxonomy

The information architecture of a website depends greatly on the taxonomy. In a narrow sense, a taxonomy is a hierarchical classification or categorization system. In the broadest sense, more meaningful for the SSK building task, it is an organization structure that people interact with through an information architecture to achieve their goals. It is the information structure behind the user experience (UX). In order to build a general taxonomy for the SSK, WP4 led a design exercise called "Card sorting[51]". It consists in writing down all the content that should be made available on the future virtual environment, using small pieces of paper (the actual "cards"), and asking potential users to sort them, building categories in the process:

[51] https://www.usability.gov/how-to-and-tools/methods/card-sorting.html

Card sorting for the SSK

This activity helped a lot in figuring out what the topics were that needed to be addressed in the SSK, how they should be organized and categorized. The picture above has been taken during the process, in which it is possible to see:

- The white cards, on which a particular topic was written initially. Testers were asked to reflect upon those topics, with the possibility to change the label or to discard it if considered unnecessary. The ones retained had to be sorted into one column.
- The yellow cards, made by the testers themselves during the process, on which the title of the category (on the top of each column) was written.

This method helped the team to understand what the users' expectations and understanding of the chosen topics were. Even if the participants were members of the WP4 team, they are Arts and Humanities researchers themselves so they are representative of the SSK's targeted audience. The insights gathered thanks to this method will be useful to:

- Build the structure of the SSK
- Decide what to put on the homepage
- Label categories and navigation.

3.2.2 Best practices for populating the Knowledge Base

The SSK is, in essence, a special purpose Knowledge Base (KB) acting as a reference tool for standards related to Humanities and Cultural Heritage. There are many ways to manage the content creation process and update of such a KB. Below, WP4 presents some best practices, acting as guidance to build a valuable knowledge base, which form the process to be adopted for the PARTHENOS SSK.

3.2.2.1 Assigning responsibility

The first essential step towards forming a best practice process is to identify an owner for the knowledge base. The KB owner is the contact for the team and can ensure that content is created and is consistent. The KB owner should regularly monitor and check the list of issues that need further KB documentation (e.g. new standards to populate the KB, additional documentation, new references, new links etc.). This will ensure that new content is created in a timely manner and that existing content is always up-to-date. Depending on how the team is identifying issues or content that needs to be added or updated the KB owner is responsible to prioritize, schedule, and assign content processing accordingly. Having a KB owner is a good way to make sure that the KB content meets the necessary quality standards and is consistent and thorough.

3.2.2.2 Establishing a process for flagging issues for the knowledge base

Identifying issues and new content to update the KB is a standard part of each operational workflow. The PARTHENOS team will make sure that such requests will come from within the consortium but also from external members of the scientific community. For this reason a communication framework will be established to gather the necessary feedback from external stakeholders. The KB manager should always search for existing documentation to avoid creating multiple versions of the same issue. And when there is existing content,

the KB owner or any other person in charge for specific standards documentation should check to see if it needs updates or improvements, and flag it if so. After a while this will become standard support process as part of the daily operational maintenance of the PARTHENOS SSK.

3.2.2.3 Determining content authors for the knowledge base

Creating KB content has to be a priority and part of the regular responsibilities for specific people or a group. In PARTHENOS there are specific teams responsible for writing and updating the respective KB items following the standards that they are closely monitoring. By assigning content authors PARTHENOS can make sure that all articles in the KB are being managed effectively by subject matter experts who are familiar with the standardization topics.

3.2.2.4 Establishing standards for authoring quality knowledge base content

Regardless of who is creating the content, it is important that documentation is clear, concise, and consistent. If users do not easily find answers in the KB, they may get frustrated and ignore it, reducing thus the usability and impact of the PARTHENOS SSK. Below we present some general tips as best practices for authoring KB content:

- **Develop a template for the articles.** A template makes it faster and easier to create content. PARTHENOS authoring team has developed a template with designated sections to fill in, so that authors include the right information. A template also ensures that articles are consistent and users know what to expect.
- **Keep articles, short if possible, and divide content into sections.** Articles should be short enough for users to quickly scan to see if the information they need is there. It is always unnecessary to overwhelm users with too much information in one article.
- **Use of clear, action-based titles.** Users tend to look for articles in the KB when they want to accomplish a task. It is important that articles are clearly titled with the action or task that is documented. Vague and general titles make it hard to users to know if the task they need is covered and what standard is most relevant for their research.

- **Use of bullets and numbered lists.** List items and steps are much easier to scan and follow when they are broken into bullets or numbered lists.
- **Define terms.** PARTHENOS SSK will make sure that all terms are defined appropriately in the articles, and the necessary links to resources will be present defining key terms for the standard sought. Especially for advanced concepts, linking to an explanatory article will be considered. Pointing to reusable content that provides more information is a best practice for KBs.
- **Link articles for showing relationships.** Linking to related articles is a must for a consistent KB. This will help users find all the information they need to solve their problem and it may help them answer some questions they didn't know they had.

3.2.2.5 Performing reviews before content is published

Users trust that they are getting reliable information in a KB but if they find too many errors or inconsistencies, they will not trust the information resulting thus in a poor utilization and usability of the KB. So, it is important to have content reviewed before it is published. KB authors should schedule a subject matter expert (SME) to review articles for accuracy, integrity and thoroughness. Multiple review cycles or multiple reviewers might be necessary for complex topics. This is also a good way to share knowledge. It is always a good practice to assign one final person — someone responsible for publishing having as focus the final quality control. This person must review and approve all articles before publishing them in the internal KB.

3.3 Designing the SSK

To design the interface, the WP4 decided to use Axure[52], a widely popular design software for creating drafts, wireframes and prototypes of projects. Axure was favoured because:

- The UX Designer in the WP4 possessed a license and had experience with the tool and the majority of its features ;
- It is designed to produce two kinds of deliverables, wireframes and prototypes, which makes it flexible.

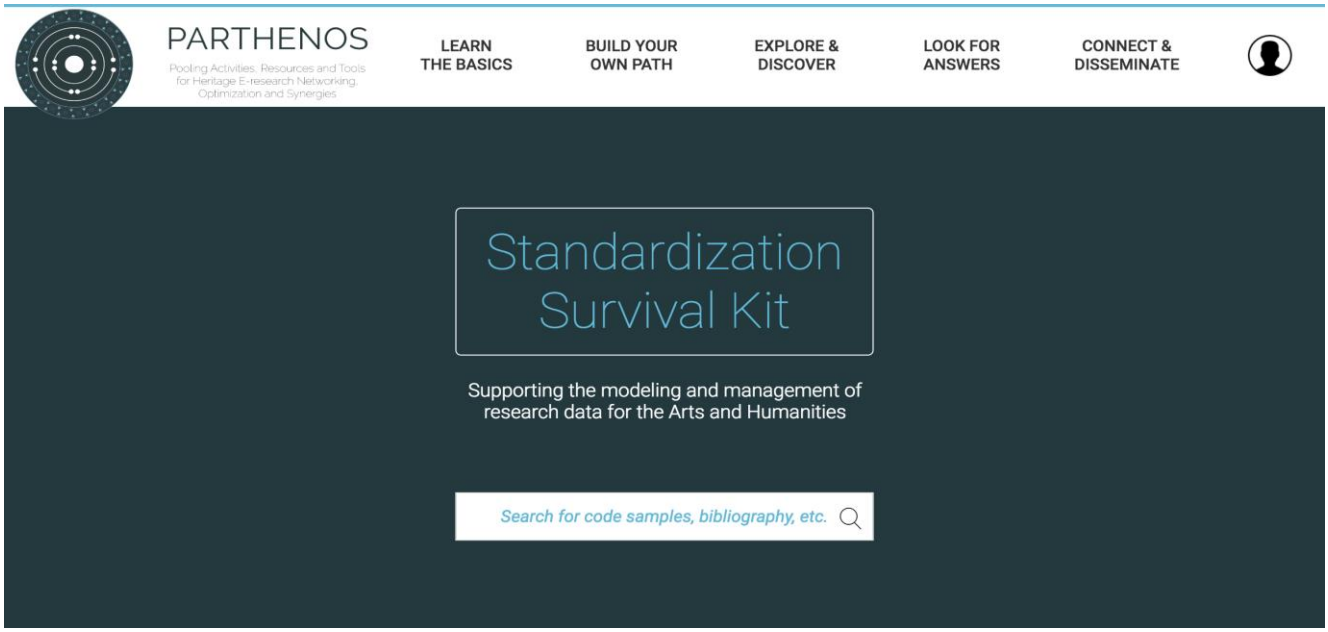
[52] http://www.axure.com/

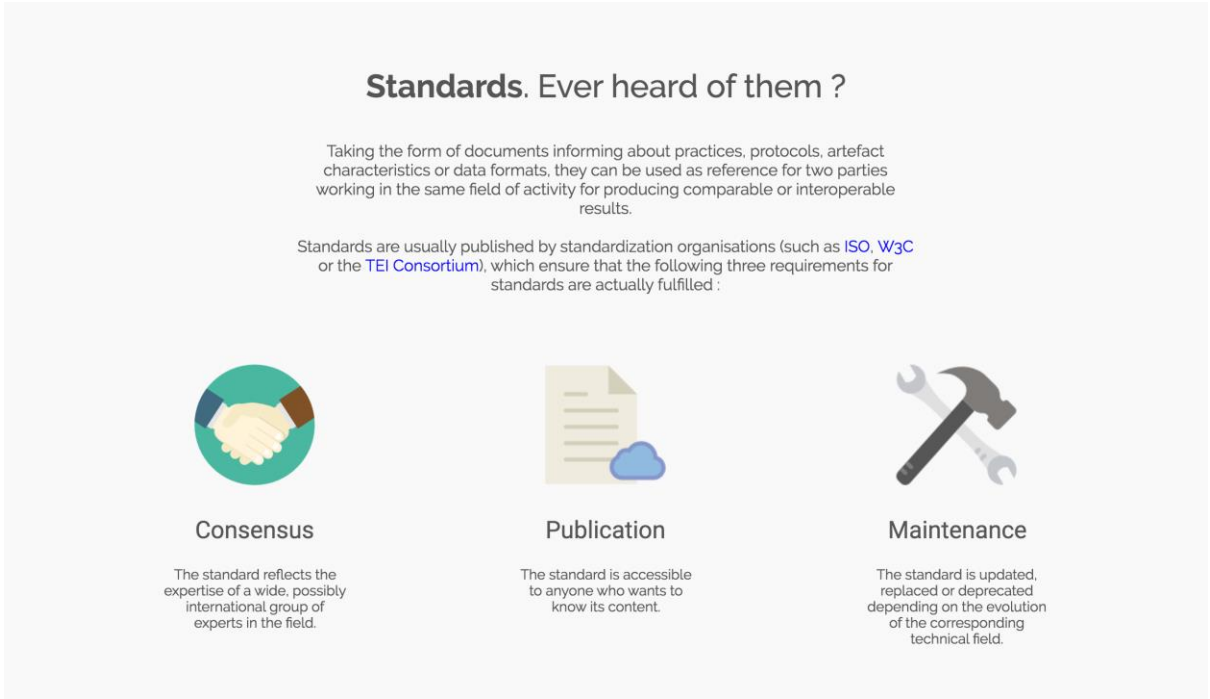
All design deliverables do not address the same requirements and goals. There are three kinds of deliverables for designing a user interface[53]:

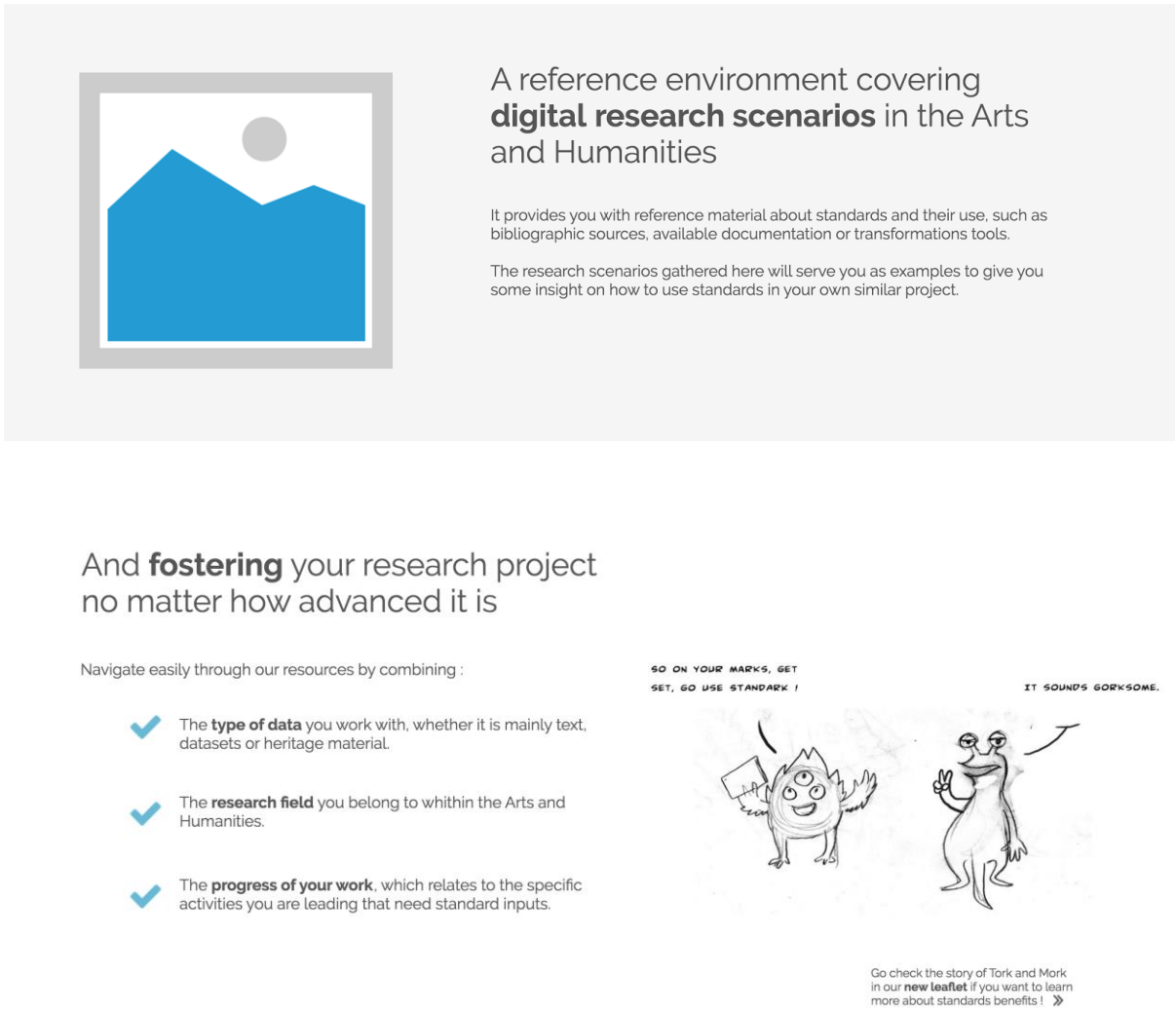
1. A **Wireframe** is a low-fidelity representation of a design. It should clearly show the main groups of content (the what?), the structure of information (the where?) and a basic description / visualisation of the user interface interaction. Wireframes are typically used as the documentation of the project. Since they are static and fix interaction with an interface at a specific point in time, they should be accompanied by the text (from short notes explaining interaction to complex technical documentation). But they can also be used in a less formal way. They are quick and simple, so they serve well as clear sketches for inner communication in the team.
2. A **Mockup** is a middle to high fidelity but static design representation. Very often a mockup is a visual design draft, or even the actual visual design. It is supposed to represent the structure of information, by enabling to visualise the content and showing the basic features in a static way. It also encourages people to actually review the visual side of the project.
3. A **Prototype** is often confused with a wireframe. It is a middle to high fidelity representation of the final product, which simulates user interface interaction. It allows the user to test content and interactions with the interface in a way similar to the final product. Interactions should be modelled with care and have significant resemblance to the final experience. It is not static but a dynamic representation of the interface which can be used to its full potential in user testing before the development begins.

The work done for the SSK until now is closer to wireframes than it is to mockup or prototype, since it is generally not too time-consuming and achieves the first stage of a user-interface design: to see how content is organized on the page, how the information is architectured. But the wireframes designed for now have better fidelity than usual: for better communication and understanding with WP4 partners (and users), the WP4 team clearly needed to follow some visual guidelines taken from the PARTHENOS Graphical Charters, by re-using fonts, logos and colours employed in the PARTHENOS project.

[53] https://www.interaction-design.org/literature/topics/user-research


PARTHENOS
Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies
LEARN THE BASICS
BUILD YOUR OWN PATH
EXPLORE & DISCOVER
LOOK FOR ANSWERS
CONNECT & DISSEMINATE
Standardization Survival Kit
Supporting the modeling and management of research data for the Arts and Humanities
Search for code samples, bibliography, etc.


Standards. Ever heard of them ?
Taking the form of documents informing about practices, protocols, artefact characteristics or data formats, they can be used as reference for two parties working in the same field of activity for producing comparable or interoperable results.
Standards are usually published by standardization organisations (such as ISO, W3C or the TEI Consortium), which ensure that the following three requirements for standards are actually fulfilled :
Consensus
The standard reflects the expertise of a wide, possibly international group of experts in the field.
Publication
The standard is accessible to anyone who wants to know its content.
Maintenance
The standard is updated, replaced or deprecated depending on the evolution of the corresponding technical field.



Homepage of the SSK

This design follows the PARTHENOS graphical charter guidelines in terms of colours, fonts and logo. The main page is divided into five blocks:

1. **The title and navigation menu**. The work done during the card sorting activity helped to identify the main topics and categories that should be present in the SSK. The menu names focus on a specific action allowed by the users. The search bar below the title has to be immediately visible to users who want to start browsing, especially those already familiar with the environment.

2. The “**Standards**” section. Researchers not too familiar with this concept need to be informed right away what standards are and what are they used for in scholarly activities.
3. The “**Reference environment**” section. This is where the question of the SSK’s role and of the service offered with the platform is addressed.
4. The “**Fostering your research project**” section. It clarifies how the users are able to navigate the website by highlighting the ways they can retrieve information.
5. The **footer**. This is the same footer as used for the main PARTHENOS website.

3.4 Further developments

WP4 plans to use unfold some of the most representative use cases to build an entire user-flow, starting from the researcher landing on the homepage to the information retrieval that he/she was looking for. By doing so, it will be possible to organize the links between the visited pages and how the information is organized within them.
Furthermore, the ongoing discussions with WP3, WP5/6 and WP7 on the necessity to coordinate the different activities lead to consider a stronger link between the interactive guides and resource registries developed by PARTHENOS.

4 Communicating: the “Why standards?” leaflet[54]

4.1 The goal of the leaflet

The leaflet is intended to communicate the necessity of standardization in the academic and scientific world, as well as the availability of standard solutions for different applications. It is targeted at scholars with limited technical background. The leaflet is also designed to be the point of entry to the PARTHENOS website, the helpdesk and the Standardization Survival Kit.

4.2 Our approach: a collaborative work

The work on the leaflet was carried out by working groups consisting of representatives of the humanities, standardization experts and graphic designers in several sessions between May and October 2016. Initiated in 2015 within the DARIAH Guidelines and standards (GiST) working group, it featured in-depth analysis consisting of several stages:

1. Gathering ideas for the longer leaflet on standards in a brainstorming manner, resulting in an extensive document describing the landscape of the world of standards, metaphorical examples, drafts of scenarios for going from unstructured to structured data, etc.
2. Reorientation of the initial approach to a short introductory publication aimed at scholars with limited technical background, and intended to raise their interest in standardization without immediately delving into details and solutions.
3. Preparation of the draft of the six-section leaflet containing the catching title page plus concise information on:
    a. what the leaflet is about (motivation to read it)
    b. the role of standards and why to use them
    c. the world without standards (short descriptions of problems resulting from not using standards)
    d. the world with standards (benefits from adopting standards)

[54] List of contributors to the "Why Standards?" brochure/leaflet (members of PARTHENOS WP4 and DARIAH WG GiST): Jack Bowers, Esther Carrasco, Jakob Epler, Agathe Gastaldi, Valentijn Gillssen, Klaus Illmayer, Nicolas LarrousSe, Adeline Joffres, Karlheinz Mörth, Maciej Ogrodniczuk, Marie Puren, Charles Riondet, Laurent Romary, Dorian Seillier, Panayiotis Siozos, Reinier de Valk.

    e. link to PARTHENOS website and helpdesk (to enable action)
4. Working on the ideas to clarify the message and make it attractive:
    a. using a comic as the primary means of communication, with text bubbles integrating the initial text
    b. adopting the sci-fi universe for the leaflet (aliens instead of humans, thus avoiding any cultural or historical rooting)
    c. using professional graphic design
    d. maintaining the PARTHENOS graphic charter (colors and font)
5. Preparing short scripts for future comics in the leaflet
6. Contacting the cartoonist and a graphic designer
7. Creating the visualisations of characters, houses and cities (the latter two both organized vs. unorganized)

The material gathered throughout the process and not used in the leaflet will be reused in further communicating activities and other project deliverables.

4.2.1 A longer brochure

The initial document was intended to contain a short introduction and a list of domain-specific approaches to be supplemented by scenario specialists. These approaches were to offer ready-to-use solutions to problems caused by lack of standardization in the humanities.

The working group also investigated the possibilities of using metaphorical scenarios such as electricity outlet formats or train track sizes and dimensions, as well as other examples such as data from obsolete or undocumented measurement systems or non-Western writing systems. Possibilities to use cartoon characters (cf. Digiman cartoon series of WePreserve: https://www.youtube.com/channel/UCPbQQpwOWIuxR6AzJnFvqFw) resulted in the first ideas for using cartoons to tell the story about standards.

4.2.2 The draft 6-page leaflet

The initial version of the leaflet was created to gather the most important content in a user-friendly format:

The role of standards and why to use them

Standards have been around for ages and have come to play an ever important part in human society. Without, for instance, referring to particular weights and measures shopping and making appointments would be much more cumbersome. Even if we don't see them, we are dealing with them everywhere: in communications, media, transport, healthcare, food, construction, energy, education and research.

Standards are relevant to the humanities as well, as there are numerous uses of standards that apply to literature, linguistics, lexicography, archaeology, history, etc.

Moreover, standards exist for, and apply to the use of: images, text (character encoding and file formatting), dates, media files (video, voice, etc.), references to places, people, languages, archives, controlled vocabularies and more.

Without standards

Regardless of your particular academic field, you may have already encountered:

- tools not working as expected
- difficulties in reuse and exchange of research data
- compatibility problems with tools and data
- losing access to your data
- having to invest time and/or money in conversion
- not getting funded

All of the above could be avoided by using standards.

With standards

In the humanities and sciences, digital standards have gained even more importance in dealing with formatting and referencing issues. Thus, we need standards to:

- be independent of particular software
- collect, structure, and re-use data
- apply qualitative and quantitative (research) methods to digital data
- reference and find references to common entities like places, people, time, languages, etc.
- make tools interoperable.

That is why politics, funding organisations, infrastructure projects, and many other stakeholders are pushing the use of standards.

Motivation for this leaflet:

What do we offer:

Standardization Survival Kit (SSK): It gives you an overview to commonly used standards and helps you in deciding which standard to use for your project.

Helpdesk: If you have further questions on standards that are not covered by the SSK, please contact our helpdesk. The input we get there will be integrated into the SSK.

Join our efforts on using standards!

Tell us about your experiences and your approach: [Link]

How can we help you to use standards? Learn more and get involved:

[Link to a website with more information]

(IMPRINT)
This leaflet is a conjoined effort by

PARTHENOS stands for "Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies" and is funded by Horizon 2020 of the European Commission. The project started on 1 May 2015 and runs for 48 months.
www.parthenos-project.eu
and
DARIAH Working Group Guidelines and Standards
(Logo + Contact)

[more partners/institutions to add]

[is there a mailing list or something else where people can get information? If so, refer to this list / or at least to the helpdesk/SSK]

Why standards?

An argumentation aid for using standards in the humanities.

Illustration / Logo

Leaflet: first mockup

The text was further adapted to tell a story showing how the use of standards will help the research community.

4.2.3 The draft storyline

This version, introducing alien creatures, was created by Dorian Seillier:

On the planet Digitus, Tork pays his friend Mork a visit, who just started to build his Digit-House.

1)
Tork: « Hi Mork ! How is the house coming on ? »
Mork: « Not bad, apart from a few minor things.

2)
Mork: « My tools keep breaking, I can make no sense of the plans I drew up, and I've wasted half of my €uro-rks buying stuff I thought I'd need. Then finding I didn't need it after all. »

3)
Mork: « I thought it would be easy to do it on my own.
Tork: « Hmm.. »

4)
(House wreckage)
Mork: « MY GORK IT'S FALLING APART ! »
Tork: « Ouch. »

5)
Tork: « On our planet Humanitus, we use STANDARK for our buildings. STANDARK is the most used tool in the Humanitus community. Whatever resources you have, it's usable, reusable, solid and sustainable.

6)
Tork: « Here, let me show you. »
(Tork puts off some kind of tablet and turns it on.)

7)

Tork: « As you can see, with STANDARK you can share resources and building techniques. It's called interoperability. »

Mork: « By Malork's tentacles ! »

8)

Tork: « What's more, you can use it to do a lot of repair work and maintenance — even if you didn't build the house yourself. By producing and using STANDARK, we are sharing the same resources and the associated know-how with all Humanitusians wherever they are.

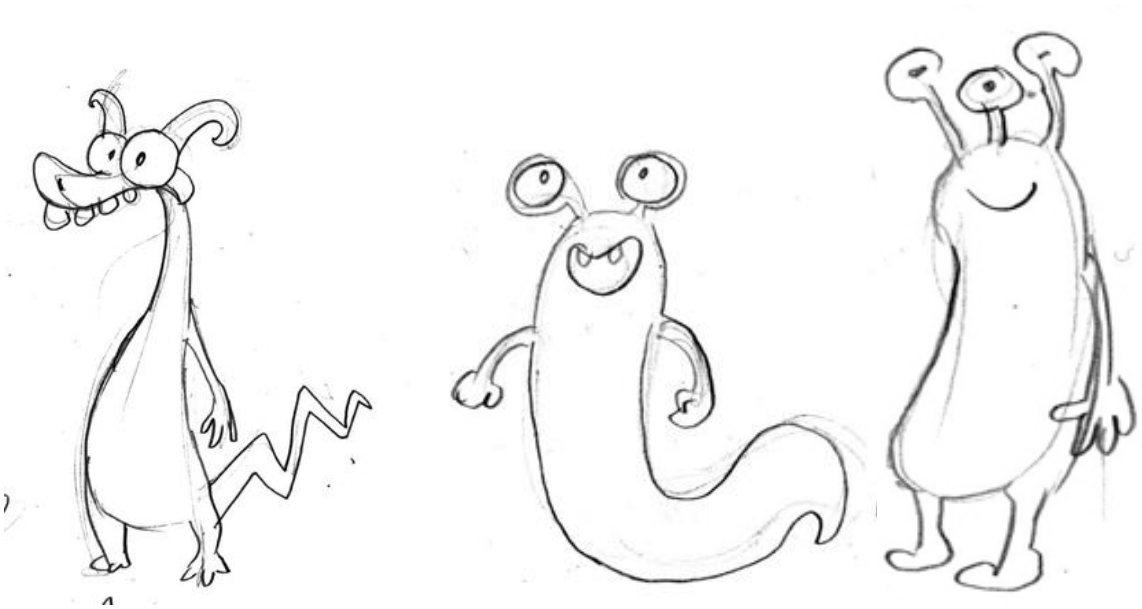
9)

Tork: « So on your marks, get set, go use STANDARK ! »

Mork: « Sounds gorksome. »

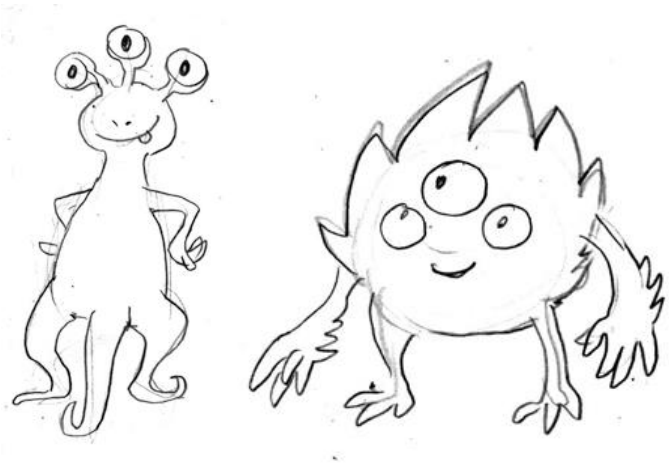
4.2.4 Ideas from the graphic designer and selected characters

Cartoonist Agathe Gastaldi suggested several variants of alien creatures and their houses to be used in the comic.



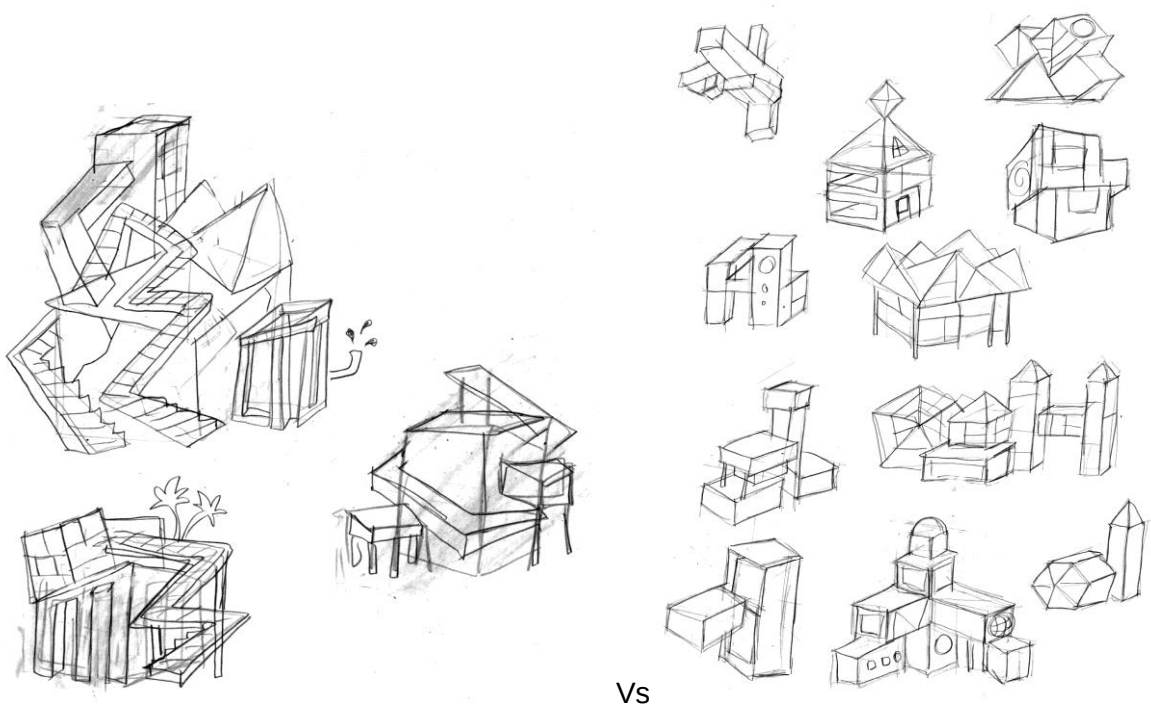
Characters: first proposals (June 2016)

In September the working group selected the initial characters to be used in the final version of the comic:



Selected characters (September 2016)

Also, the concepts of ‘non-standard’ vs. ‘standard’ houses (representing disintegrating and stable constructions, respectively) were visualized:



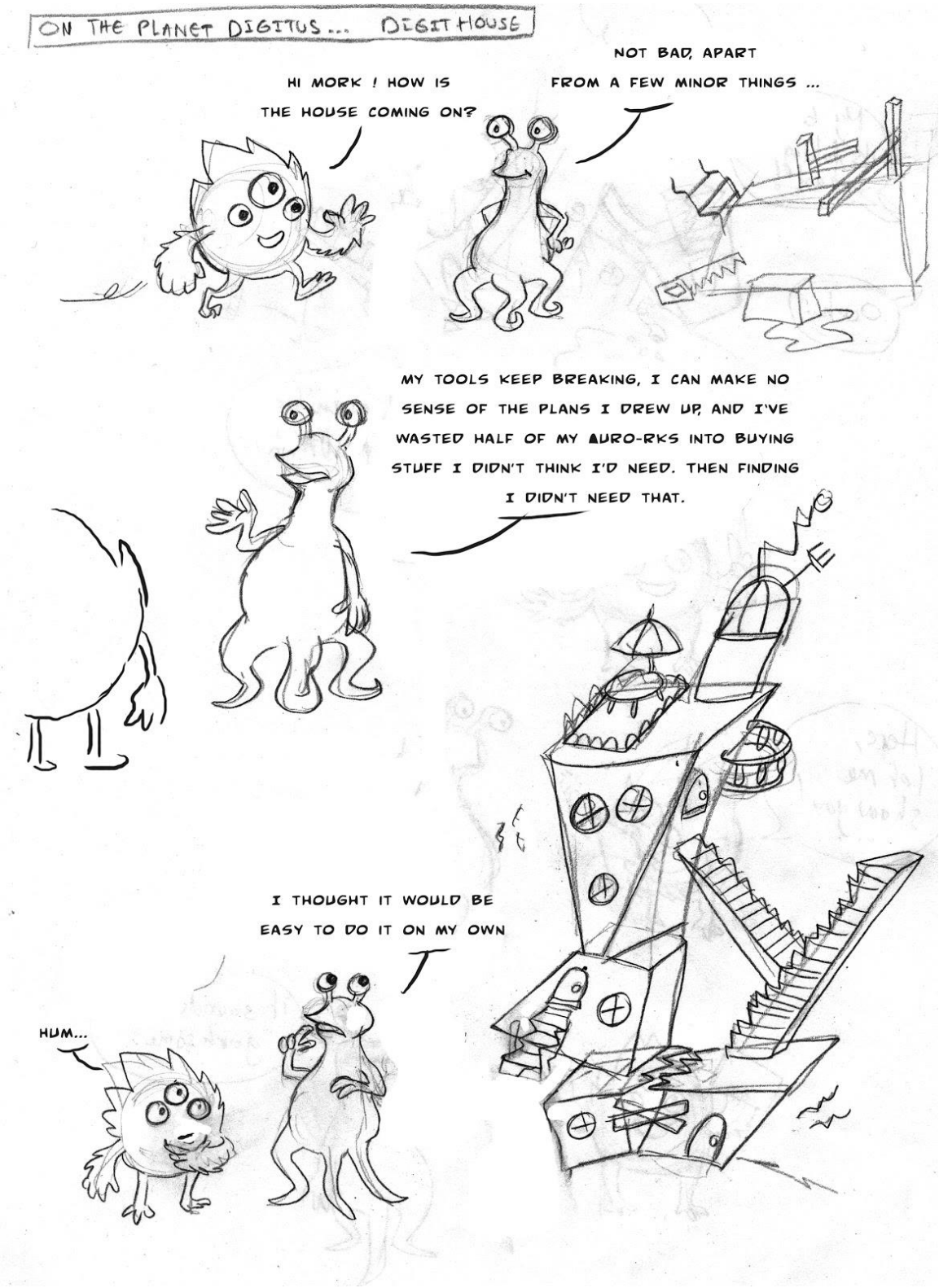
Vs

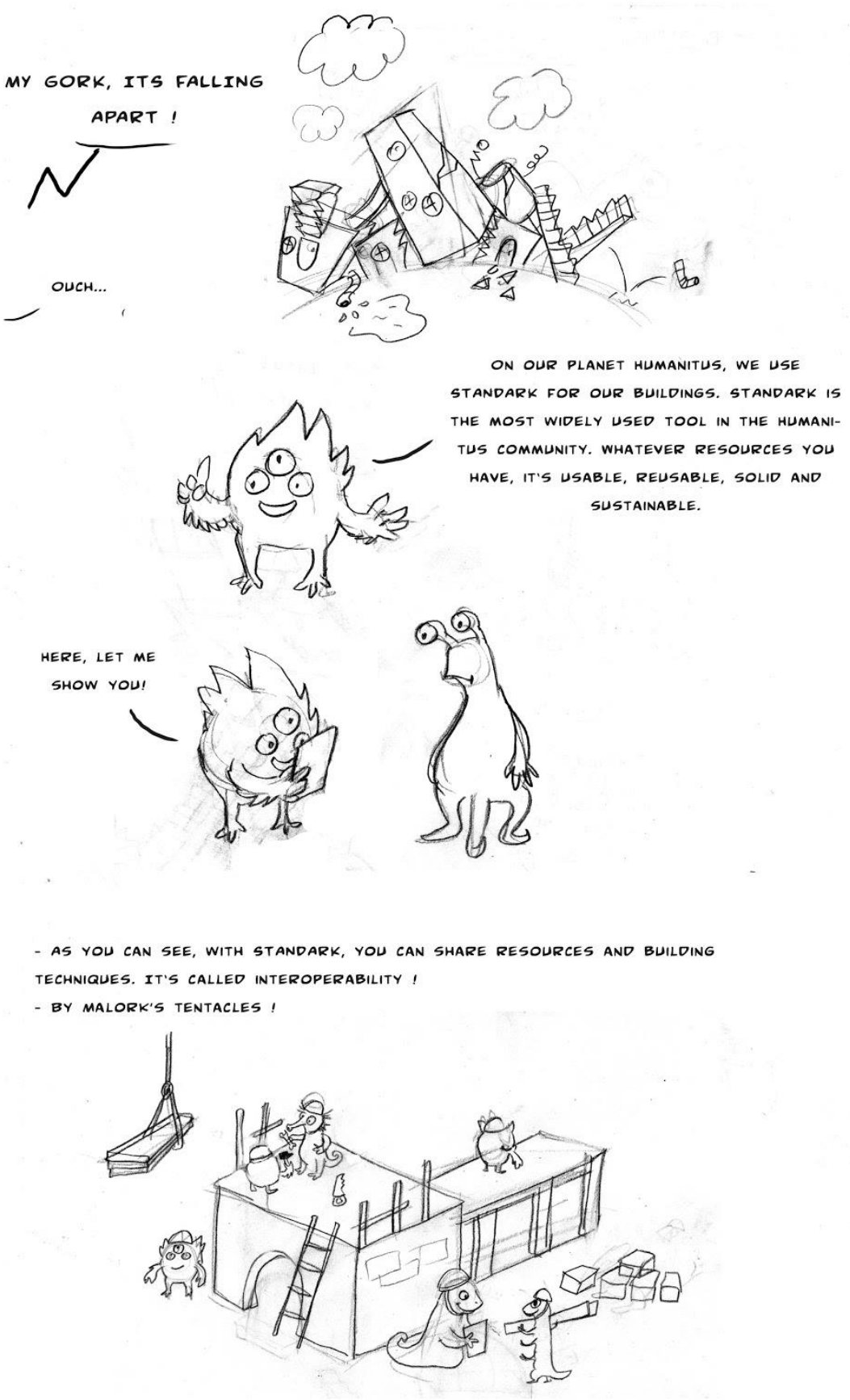
‘Non-standard’ vs. ‘standard’ houses

The concepts were further developed by the graphic designer.

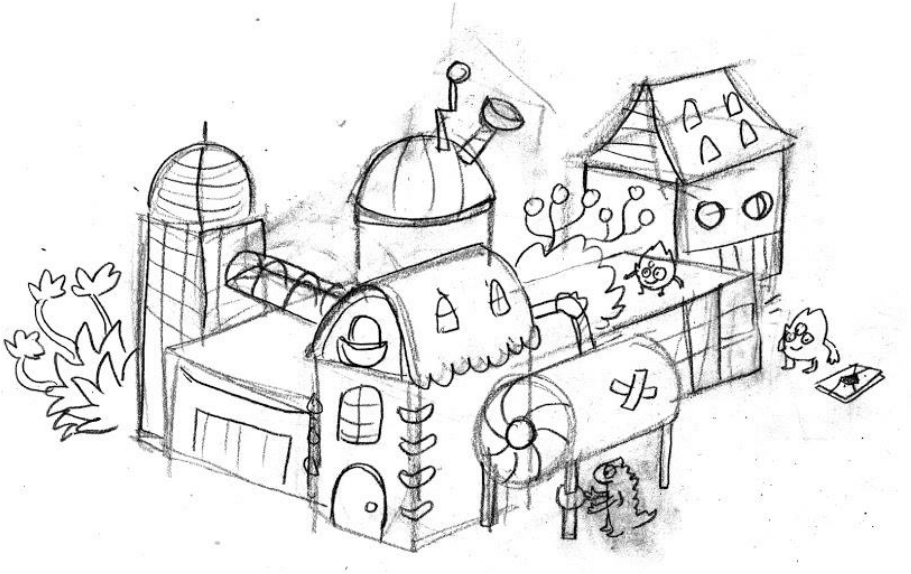
4.3 The draft version of the comic

The draft version of the comic was created by Agathe Gastaldi:




MY GORK, ITS FALLING APART !
OUCH...
ON OUR PLANET HUMANITUS, WE USE STANDARK FOR OUR BUILDINGS. STANDARK IS THE MOST WIDELY USED TOOL IN THE HUMANITUS COMMUNITY. WHATEVER RESOURCES YOU HAVE, IT'S USABLE, REUSABLE, SOLID AND SUSTAINABLE.
HERE, LET ME SHOW YOU!
- AS YOU CAN SEE, WITH STANDARK, YOU CAN SHARE RESOURCES AND BUILDING TECHNIQUES. IT'S CALLED INTEROPERABILITY !
- BY MALORK'S TENTACLES !

WHAT'S MORE, YOU CAN USE IT TO DO A LOT OF REPAIR WORK AND MAINTENANCE - EVEN IF YOU DIDN'T BUILD THE HOUSE YOURSELF. BY PRODUCING AND USING STANDARK, WE ARE SHARING THE SAME RESOURCES AND THE ASSOCIATED KNOW-HOW WITH ALL HUMANITUSIANS WHEREVER THEY ARE.



SO ON YOUR MARKS, GET SET, GO USE STANDARK !

IT SOUNDS GORKSOME.

Storyboard

4.4 The accompanying text

The final version of the leaflet will integrate the comic, the PARTHENOS logo title page and a short text extracted from the longer leaflet text:

You may have already encountered:
– difficulties in reuse and exchange of research data
– compatibility problems with tools and data
– losing access to your data

And there are many more frustrating experiences when working with digital data and tools.

This is where standards come into play. They help you in:
– dealing with formatting and referencing issues
– making tools interoperable
– collecting, structuring, and re-using data

Furthermore, there are a lot of other issues where the use of standards will help you and your research community.

Additionally, standards will help you in your research in applying qualitative and quantitative research methods to digital data, disseminating the results, and in collaborating with others.

Using standards you invest not only in the quality of your data, but you also enable others to benefit from your research — like the people of Humanitus when they apply standards in their community.

4.5 The strategy for dissemination of the leaflet

We anticipate disseminating the leaflet and the comics both in paper and electronic formats. The paper version will serve as communication tool at events organized by PARTHENOS, but also within CLARIN and DARIAH and major related conferences to

point to the more comprehensive content available from the SSK. The two main characters of the comic will become reference characters in the SSK to represent scholars asking themselves questions as a basis for providing hints, answers and solutions to the reader.

It is planned that the leaflet will serve as a reference communication object about standards in both CLARIN (through their standardization committee) and DARIAH (through the GiST working group). In particular, the paper version will be widely disseminated among the various members of the two infrastructures, so that it is also becomes available to young scholars within our high education institutions.

5 Conclusion

The PARTHENOS project is the opportunity to embrace a whole range of concrete actions to foster the use and development of standards for research in the Humanities. The present report gathers a wide range of such activities which have allowed us to acquire a rather comprehensive picture of the corresponding landscape. Our objective now is to reuse most of the material described or elicited here as basic components that will be integrated in the actual SSK as it will be put in place, and complement this with the more precise results we will obtain from the focused actions described above touching the meta-data, archival, language and cultural heritage domain.

In parallel, we will carry out groundwork with respect to the following aspects:

1. Continue the development of the use cases, with more in-depth description of how the identified standards take a role in data creation, enrichment or dissemination;
2. Building on this use case development, develop straightforward documentation to encourage research communities to integrate more and more the use of standards in their practices.
3. Implement a "How-to" guide to present the path research communities have to follow to turn their good practices into an internationally recognized standard;
4. Support the development of emerging standards and the dissemination of solid standards among broader research communities.

6 Abbreviations

BBAW	Berlin-Brandenburgische Akademie der Wissenchaften
CENDARI	Collaborative European Digital Archive Infrastructure
CLARIN	Common Language Resources and Technology Infrastructure
CNR	Consiglio Nazionale delle Ricerche
CSIC	Consejo Superior de Investigaciones Científicas
DANS	Data Archiving and Networked Services
DARIAH	Digital Research Infrastructure for Arts and Humanities
EHRI	European Holocaust Research Infrastructure
E-RIHS	European Research Infrastructure for Heritage Science
FHP	Fachhochschule Potsdam
FORTH	Foundation for Research and Technology – Hellas
ICVBC	Istituto per la Conservazione e la Valorizzazione dei Beni Culturali
IDS	Institut für deutsche Sprache
IESL	Institute of Electronic Structure and Laser
INRIA	Institut National de Recherche en Informatique et en Automatique
ISO	International Organization for Standardization
KNAW	Koninklijke Nederlandse Akademie van Wetenschappen
MEI	Music Encoding Initiative
NIOD	Nederlands Instituut voor Oorlogs-, Holocaust- en Genocidestudies
OEAW	Österreichische Akademie der Wissenschaften
SISMEL	Società Internazionale per lo Studio del Medioevo Latino
SSK	Standardization Survival Kit
TaDiRAH	Taxonomy of Digital Research Activities in the Humanities
TEI	Text Encoding Initiative
TCD	Trinity College Dublin
UCPH	University of Copenhagen
W3C	World Wide Web Consortium

7 Appendix

7.1 Cockburn description of the DTABf use case[55]

User Story

A literary studies scholar is interested in the way Goethe and Schiller might have influenced one another in their writing. To examine this question, the researcher plans to compare their writings from the period before their contact to the ones created during their friendship. His interest covers very different fields, from linguistic features via the style of writing to the appearance of the texts.

The analysis is to be partially performed by help of computer-based methods. Thus, the researcher needs a digital corpus of his source material. Since there has been great interest in the writings of these two authors from various scholars, there are already digitized full-texts of most of their writings scattered throughout the internet — often several digital versions per work. Therefore, in the current case, it seems not to be necessary to digitize the works under consideration from scratch. However, the existing digitized texts come in various formats: different TEI formats as well as various other, non-TEI formats. They have thus to be homogenized with regard to their text transcription and annotation guidelines, quality, and annotation format.

The corpus consists of prose, lyrical, and dramatic texts. Since TEI-XML covers these genres for structural annotation and represents a widespread format within the community, the researcher chooses to resort to the TEI Guidelines for text annotation despite a lack of experience on his side with these guidelines. For his quantitative research, it is important that the corpus is truly interoperable in itself so that results gained from the corpus are reliable. For the qualitative questions he needs easy-to-use, rendered reading versions of the texts.

In addition, in order to save time, in case there are several digital versions of one text, the researcher wants to know which one of them is closest to his final TEI format and thus most convenient to convert.

[55] For more information, please read the section 2.3.1.4 “A new TEI format: the “DTA Basis-format” or DTABf (FHP)”

Sustainability is a very important issue to the project. Thus, not only is it planned to re-use existing data and resources within the project, but also, in turn, to provide project resources to the scholarly community. Since possible research scenarios based on these frequently studied texts are manifold, it is planned to provide not only the raw TEI texts but also rendered versions of all corpus texts and their associated stylesheets.

Goal

A researcher wants to gather a corpus based on digitized texts from scattered sources, to homogenize and analyze his resources and to provide his findings and the underlying corpus (as XML texts as well as rendered versions) to the public for further research.

Scope

Corpus-based comparative research on how two authors might have influenced each other's writings.

Preconditions

- Digitized versions of the works to be considered exist and may be reused, i.e. they come in readable and processable formats and with free licencing
- Truly interoperable TEI data for the source texts: similar quality, a uniform annotation format, similar annotation depth, etc.
- A TEI format which prevents ambiguity in tagging; usage of the format is easy and clear due to: a restrictive schema, extensive documentation, and tools to support the application of the format
- A schema to evaluate the closeness of a text to the target schema
- A tool to support the annotation, correction, and homogenization task
- Stylesheets based on the respective TEI format that allow for the rendering of all texts

Success End Condition

The researcher was able to make subsequent use of existing data, to easily convert all data into a homogeneous TEI format and thus create truly interoperable texts. He could then perform his research with reliable results. The research data could be provided to the research community in a sustainable manner.

Failed End Condition

The idea of re-use didn't work out. There were not enough digital text versions to find which were freely licenced and came in non-proprietary formats. Therefore, a significant

amount of texts had to be digitized from scratch. The researcher didn't find a format to re-use for annotation which would cover his annotation needs and support unambiguous, homogeneous text annotation. Such format had to be created from scratch, as well. Conversion then involved lots of manual work directly in the XML texts. Hence, corpus creation was very time-consuming so that there didn't remain enough project resources for the actual research within the project.

Primary Actor

Humanity's Researcher with basic (but limited) text processing skills.

Trigger

Research question, existing text and other resources which may be reused here

Main Success Scenario

1. Definition of the research question
2. Definition of digitization guidelines and quality standards for the researched texts
3. Selection of an unambiguous, TEI-based annotation format
4. Selection of works of primary interest
5. Search for digitized versions of the respective works
6. Comparison of digitized versions of similar works in terms of their proximity to/distance from the target format and selection of final corpus of digitized texts
7. Conversion of the final text corpus into the target format
8. Research: Analysis of the mutual influence of Goethe and Schiller on their respective works during the time of their friendship.
9. Provision of text corpus to the public

Sub-Variations

3'. Such TEI format, which furthermore covers all annotation needs of the researcher, couldn't be found, but had to be created within the project

5'. There are no digital versions of some of the texts. Those texts have to be digitized first.

5'a. The digital text versions are not based on the primary but on later editions of the works. Hence, the problem that changes to the original texts in later editions might distort the results has to be factored in.

5'a. Text selection is based on the accessibility of digital text versions. Hence, points 4. and 5. are interchanged.

7’. Texts are too heterogeneous in annotation and quality, so that they can’t be homogenized with justifiable effort. Therefore texts have to be digitized from scratch in a homogeneous way.

8’. There are no given stylesheets for rendering. Those have to be created by the project.

Extensions

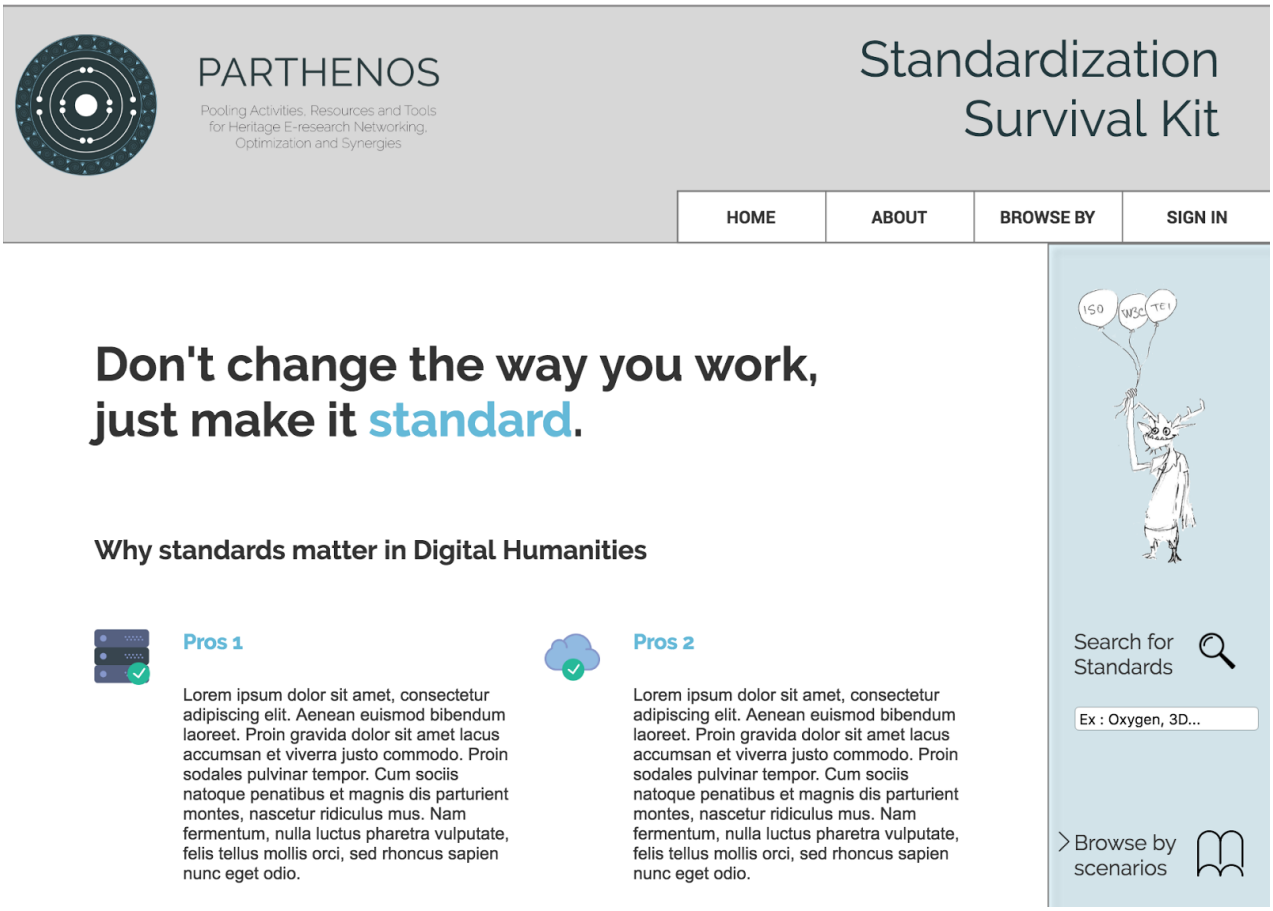
7a. There are tools existing to support the research when performing the annotation, correction and homogenization task (e.g. a WYSIWYG editor creating readable text versions from the XML)

8a. There are linguistic tools and services existing which can be reused for the comparison task.

9a. The corpus can be provided via a research infrastructure.

7.2 First draft of the web interface (July 2016)

7.2.1 Homepage



Pros 3

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean euismod bibendum laoreet. Proin gravida dolor sit amet lacus accumsan et viverra justo commodo. Proin sodales pulvinar tempor. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Nam fermentum, nulla luctus pharetra vulputate, felis tellus mollis orci, sed rhoncus sapien nunc eget odio.

Pros 4

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean euismod bibendum laoreet. Proin gravida dolor sit amet lacus accumsan et viverra justo commodo. Proin sodales pulvinar tempor. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Nam fermentum, nulla luctus pharetra vulputate, felis tellus mollis orci, sed rhoncus sapien nunc eget odio.

Go check our *Why Standards ?* leaflet to learn more about it >>

Digital Edition

Dataset Mangament

Physical Objetcs Analysis

News & Updates

The Standardization Survival Kit : a reference environment for researchers in Digital Humanities.

Our motives

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean euismod bibendum laoreet. Proin gravida dolor sit amet lacus accumsan et viverra justo commodo. Proin sodales pulvinar tempor. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Nam fermentum, nulla luctus pharetra vulputate, felis tellus mollis orci, sed rhoncus sapien nunc eget odio.

Parthenos vision

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean euismod bibendum laoreet. Proin gravida dolor sit amet lacus accumsan et viverra justo commodo. Proin sodales pulvinar tempor. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Nam fermentum, nulla luctus pharetra vulputate, felis tellus mollis orci, sed rhoncus sapien nunc eget odio.

Continue reading >>

TaDiRAH activities

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean euismod bibendum laoreet. Proin gravida dolor sit amet lacus accumsan et viverra justo commodo. Proin sodales pulvinar tempor. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Nam fermentum, nulla luctus pharetra vulputate, felis tellus mollis orci, sed rhoncus sapien nunc eget odio.

Continue reading >>

Multiple resources

GitHub

Zotero

Documentation, blogs

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean euismod bibendum laoreet. Proin gravida dolor sit amet lacus accumsan et viverra justo commodo. Proin sodales pulvinar tempor. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Nam fermentum, nulla luctus pharetra vulputate, felis tellus mollis orci, sed rhoncus sapien nunc eget odio.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean euismod bibendum laoreet. Proin gravida dolor sit amet lacus accumsan et viverra justo commodo. Proin sodales pulvinar tempor. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Nam fermentum, nulla luctus pharetra vulputate, felis tellus mollis orci, sed rhoncus sapien nunc eget odio.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean euismod bibendum laoreet. Proin gravida dolor sit amet lacus accumsan et viverra justo commodo. Proin sodales pulvinar tempor. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Nam fermentum, nulla luctus pharetra vulputate, felis tellus mollis orci, sed rhoncus sapien nunc eget odio.

Start to explore and discover new tools, methods, samples...

By choosing your field of interest

Digital Edition

Dataset management

Physical Objects Data-Analysis

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean euismod bibendum laoreet. Proin gravida dolor sit amet lacus accumsan et viverra justo commodo. Proin sodales pulvinar tempor. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Nam fermentum, nulla luctus pharetra vulputate, felis tellus mollis orci, sed rhoncus sapien nunc eget odio.

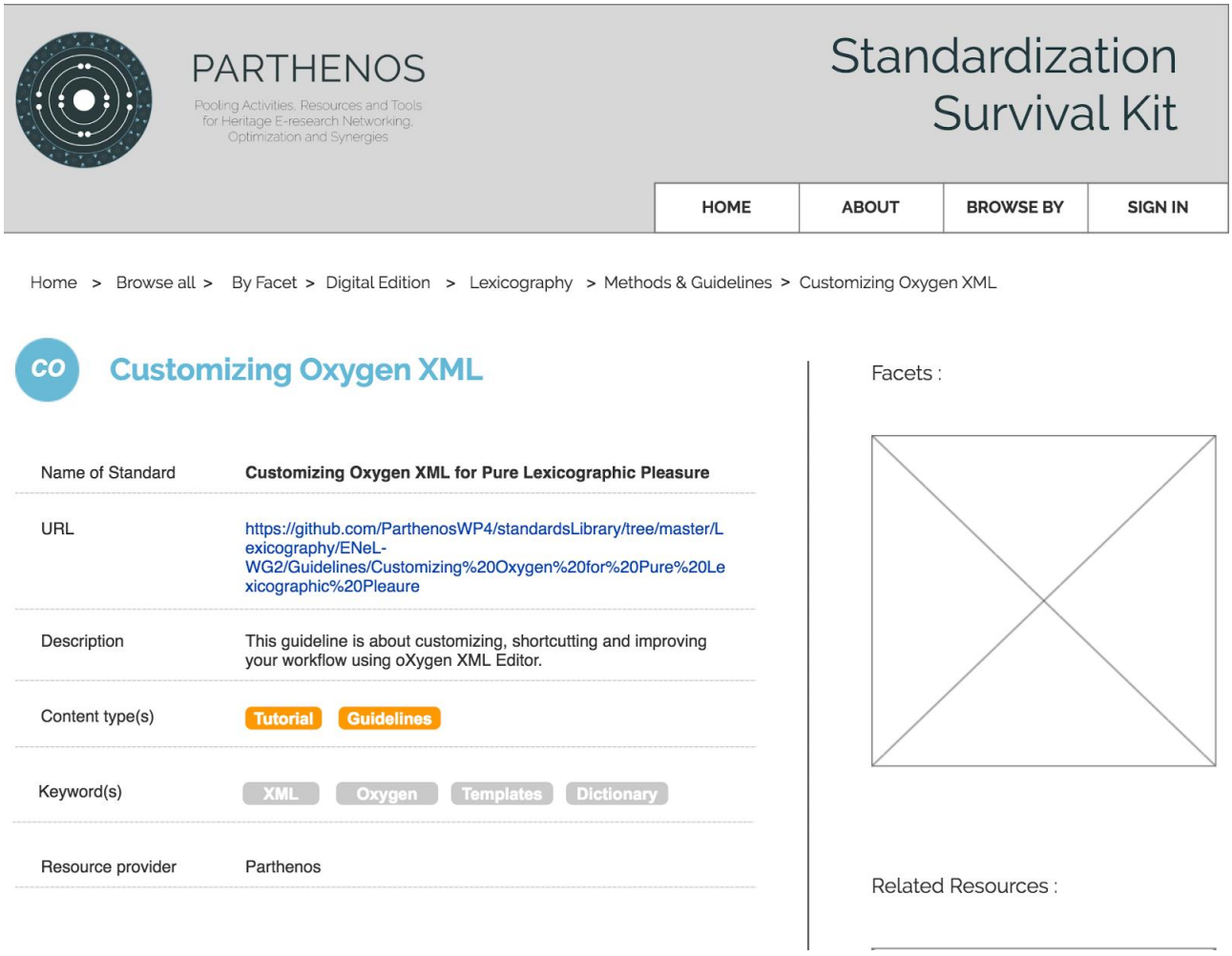
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean euismod bibendum laoreet. Proin gravida dolor sit amet lacus accumsan et viverra justo commodo. Proin sodales pulvinar tempor. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Nam fermentum, nulla luctus pharetra vulputate, felis tellus mollis orci, sed rhoncus sapien nunc eget odio.

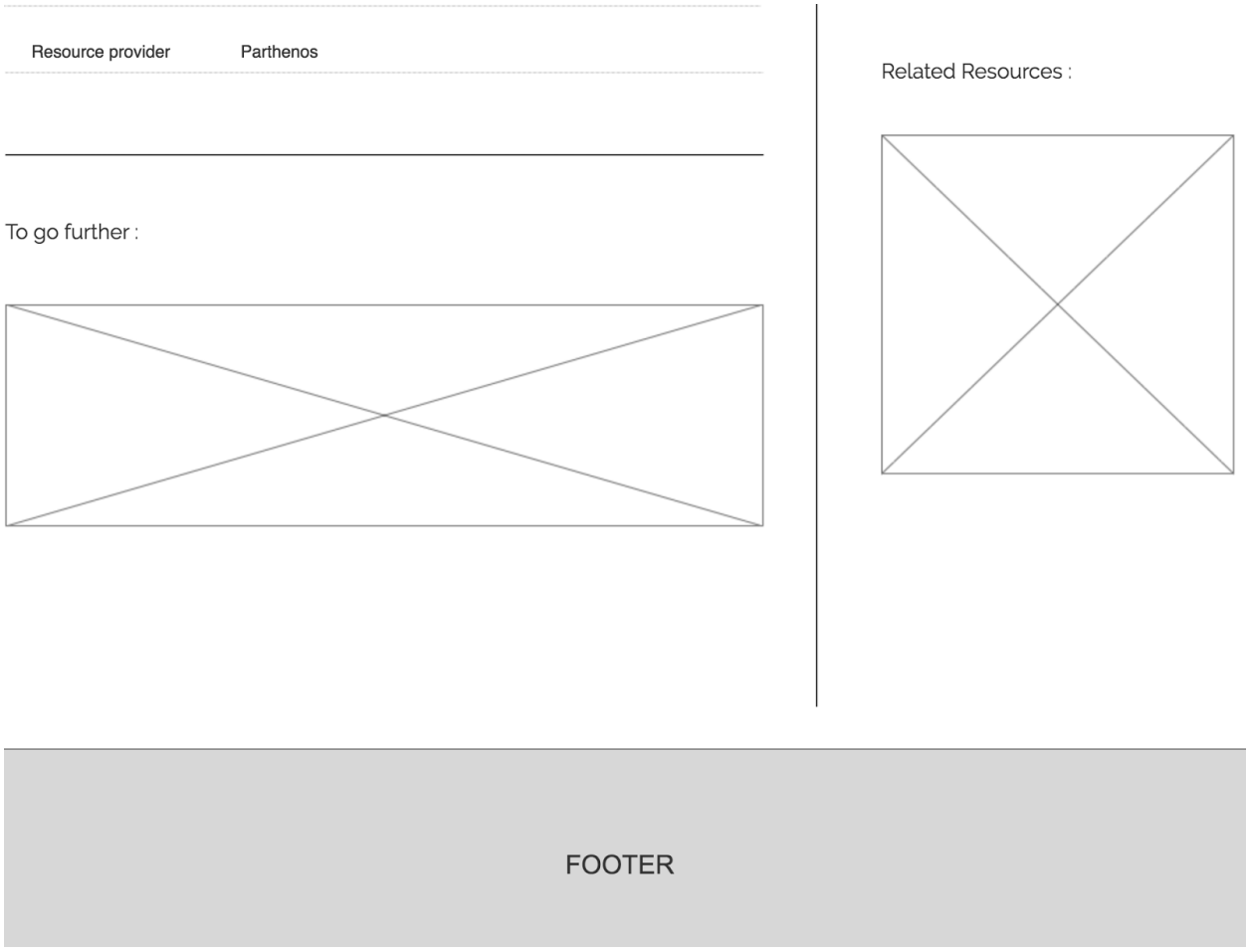
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean euismod bibendum laoreet. Proin gravida dolor sit amet lacus accumsan et viverra justo commodo. Proin sodales pulvinar tempor. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Nam fermentum, nulla luctus pharetra vulputate, felis tellus mollis orci, sed rhoncus sapien nunc eget odio.

FOOTER

Home page of the web interface – SSK

7.2.2 Result page





A result page – SSK

7.3 The “Why standards?” leaflet

7.3.1 The first version of the leaflet (draft)

7.3.1.1 The role of standards and why to use them

Standards have been around for ages and have come to play an ever-important part in human society. Without, for instance, referring to particular weights and measures shopping and making appointments would be much more cumbersome. Even if we don’t see them, we are dealing with them everywhere: in communications, media, transport, healthcare, food, construction, energy, education and research.

In the digital realm, standards have gained even more importance as only explicitly identified units can be dealt with. We need standards

- to collect and integrate data
- to re-use data
- to be independent of particular software
- to create structure in data
- to apply qualitative and quantitative (research) methods to digital data

7.3.1.2 Counter-examples: (e. impact of not using data standards)

Without standards

- tools might not work as expected
- tools and data will hardly be reusable for other purposes
- tools and data will be incompatible with other tools and new data
- you are stuck with your programmer, with a particular company
- you might lose access to your data

Not making use of standards

- you will have to pay for data modelling
- you will probably not be funded
- you will have to invest in conversion

(A couple of failure examples of why standards are needed. Illustrating the concept of re-use, interoperability or sustainability without mentioning the term)

7.3.1.3 Negative questions

Why did I not get my research grant?
Why can't I use particular tools for my data?
Why can't I publish my data directly on the web?
Why can't they reproduce my research?
Why can't I use other peoples' data?
Why can't others use my data?

7.3.1.4 Scenarios

- Historian transcribing a source in MS Word and is now looking for ways to use textometric tools, in conjunction with similar documents transcribed by some of his colleagues
- Field linguist with proprietary software has gathered data for years and cannot see how to deliver it to a wider community of lexicographers, or make it available for use by the speakers of the language they have been studying
- Researcher in literary studies wanting to show an author's correspondence online: what should she be doing, what should she ask the IT specialist

7.3.1.5 (Standards Survival Kit:) The landscape

While many fields need new standards, certain areas are already well covered. In these areas, we simply can take what is there.

There is no need to develop new **image** formats:

Make use of .tiff (Tagged Image File Format) image file formats as they are unproprietary and supported by many different applications. Because, unlike many other file types, they contain no hidden data or links, they not vulnerable to viruses or hidden bugs, thus making them a great option for long term storage/archiving.

Text plays a crucial role in most research projects:

Character Encoding:

Mixing all the different writing systems from all over the world in digital documents is no longer a problem, provided you make use of Unicode[56] which defines almost everything needed to represent letters and characters, regardless of the language, the program or the platform.

File Formatting:

Above the character level, there exist a thousand different ways to create text, most of which output documents which have proprietary encoding that may not easily be

[56] http://unicode.org/standard/WhatIsUnicode.html

searched, shared, converted of even readable several years down the line. To avoid these problems, there is XML (EXtensible Markup Language)[57] which allows document contents to be readable by both humans and machines, safely stored with no risk of corruption in the near or long term, and can very easily be searched, selectively extracted, displayed and converted to other formats.

Encoding Standards:

Additionally, in order to both save time for those creating and working with XML documents, there are well established standards for text markup which are designed to both meet the needs of specific disciplines, but also, where possible, share common practices for information which is not discipline-specific. The most widely used standard within the humanities is the TEI (Text Encoding Initiative) which is made up and maintained by an international consortium of researchers, scholars and academic institutions and is used for representation of early books transcription and annotation of ancient manuscripts and inscriptions, historical archives, anthologies, critical editions, literary and cultural materials, linguistic corpora, dictionaries amongst others.

No matter in which field you work, at some point you will want to give information about **time and date**:

While there are a thousand ways of notating a particular day (1.3.2015, 1/3/2015, 2015-3-1, 3-1-2015, 3/1/2015, first of March 2015, etc) and while all of them might appear intuitive in their contexts, to make this particular day retrievable digitally you should encode it in ISO 8601, i.e. 2015-03-01.

Places:

For attributing data of some kind to a specific (present day) country and/or their subdivisions, there is the ISO 3166 standard in which each is given a unique language independent code (http://www.iso.org/iso/country_codes).

The geographical database GeoNames (http://www.geonames.org/) provides a user friendly open sourced resource which contains multilingual information pertaining to landscape (elevation), geo-coordinates (lat/long), population demographics, covering much of the populated world. The data sources for GeoNames are a

[57] An introductory tutorial on XML can be found at: http://www.w3schools.com/xml/

number of both official governmental agencies, survey/census bureaus, international bodies and research institutions and a number of private comercial sources.

People:

When referring to or searching for people, there are ofter many potential ways in which others may refer contemporarily or historically to the same person. The causes for such ambiguity can be simply due to language differences, usage context, time of reference, etc.,. Then of course there are simply some people who have the same name. The ability to search, discover and organize information about individuals in a way that can accurately group an disabiguate is important for researchers looking for information, as well as archivists, or librarians storing it, and should be of concern to researchers and authors creating new materials if their work is to be maximally discoverable. For this purpose, there exist the Virtual International File Authority (https://viaf.org/), the GND (Integrate Authority File; German:Gemeinsame Normdatei, http://www.dnb.de/EN/Standardisierung/GND/gnd.html) which are open source, linked, and continuously maintained databases of person names (amongst other topics).

Languages:

As the vast majority of the content we produce and work with in the humanities is of course expressed in human language. Thus, it is useful to have a simple and common way to declare and discover the language, or languages used within a given resource. For most uses, the ISO 639 series of standard language codes are the most widely used: (http://www-01.sil.org/iso639-3/). Additionally, there is the Glottolog standard, which is perhaps more linguistically and anthropologically more dynamic than ISO, but is less used and for which there is less technological support (http://glottolog.org/). While making use of standards in the fields of linguistics and lexicography is of particular importance when creating new resources in a given language, the utility of language codes is by no means limited to these fields, as behind the scenes, it is the use of language tags which are central to the functionality of: spelling and grammar checkers, automatic translation programs (such as Google Translate), and the discoverability of content and resources.

GLAM:

Refer to CIDOC-CRM as an standardized ontology to exchange cultural heritage information (it will also be used in PARTHENOS)

7.3.1.6 Detailed scenarios

Use geonames.org (basic skill, introduction into standards)

You are having research data either in unstructured form like continuous text or in structured form like Excel sheets. Part of this research data are references to geographical entities be it street, city, or country names. You now want to share this research data with some colleague. After consulting your colleague you find out that you have different views how to name the geographical entities because of language or historical reasons. Also you discover that some of the names are not unique and they are used for different places (e.g. for your research it is clear that with Vienna you mean the Austrian capital whereas your colleague refers to a town in New York, United States). You both discuss how to come to a sustainable solutions because you want to combine your research data but you want to keep up your different views on the naming of the geographical entities. So you decide to use geographic information systems (GIS) referring to latitude and longitude data for your geographical entities. To make it easier you decide to use the website http://geonames.org where you enter the geographic name and then decide on the correct place. You then take the link to this place from geonames and insert it in brackets behind all of your geographic data (e.g. for Vienna, Austria:
http://www.geonames.org/2761369/vienna.html
or for Vienna, NY:
http://www.geonames.org/4833322/vienna.html) .

Scenario: From unstructured to structured data

So you are using a text editor for writing your papers and you wonder how to reduce periodic work like typing every time the same bibliographic references. You decide to create a document where you collect your bibliographic data. From there you copy and paste your references in your papers avoiding to type this information again and again. That's already one step in turning unstructured text into structured data. Now try to think in bigger terms. You know that you share bibliographic references with your research community colleagues. Why not work together and create a common reference document so that on the one hand you can discover new literature and on the other hand evade recurrent work. Sounds great! But you will run into problems if you and your colleagues don't decide on commonalities in describing the bibliographic data. You will need an understanding on notation and on the necessary data to collect. And here standards joins your effort for the first time. You will need to decide on a common bibliographic style.

Questions/thoughts on scenarios:

- Would it be useful to refer to Research Data Life Cycles? Because we could show how Standards are helping in different parts of research activity (and in connecting them).

- Another thing to keep in mind are guidelines, they could be de facto standards for some communities. This could also mean to list DH infrastructures and projects and try to give access points to them in terms of standards.
- We could also think in terms of tools: Pick tools that solve the use of standards in excellent manner. Name them and describe them in view of why they solve the use of standards so good.

7.3.1.7 Open Access

Probably we will need some words on commonly used digital practices that do not fit in our view of standards, e.g. what about Wikipedia/Social Media/Microsoft Word/GoogleDoc/etc.? As they are a strong part of the scholarly landscape in terms of everyday usage. Should we present some recommendations e.g. arguments for not using Word and instead using an XML editor?

Should we involve a list of resources for community standards? e.g.:

- CLARIN Standard recommendations for LRT: http://www.clarin.eu/content/standard-recommendations (whereupon this is a document demonstrating a good way to frighten people because it is hard to read and to understand)
- CLARIN Standards List: http://dev.clarin.nl/clarin-standards2-list-fs (a long list, probably not helpful if you are a newbie)
- TEI website: http://www.tei-c.org/About/Archive_new/ETE/Preview/ (not limited to the TEI, a list of helpful papers, articles, and guidelines addressing a wide array of issues relevant to text encoding for projects in a number of different fields and use cases)
- ACH (Association for Computers and the Humanities): http://digitalhumanities.org/answers/ (a cross discipline question and answer board for people working on, or interested in working on projects in digital humanities).

7.3.2 Sketches by Agathe Gastaldi: first proposals (June 2016)