PARTHENOS

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

Report on common design requirements

Consiglio Nazionale delle Ricerche – Istituto di Scienza e Tecnologie dell'Informazione

30 January 2019

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

HORIZON 2020 - INFRADEV-4-2014/2015:

Grant Agreement No. 654119

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

REPORT ON COMMON DESIGN REQUIREMENTS

Deliverable Number	D5.8
Dissemination Level	Public
Delivery date	31 January 2019
Status	Final
Author(s)	Carlo Meghini (CNR)

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	D5.8
Document title	Report on common design requirements
Author(s)	Carlo Meghini (CNR)
Dissemination level/distribution	Public

Document History

Version/date	Changes/approval	Author/Approved by
V 0.1 17.12.18	First draft	Carlo Meghini (CNR)
V 1.0 30.01.2019	Reviewed.	S. Bassett (PIN)

Table of Contents

Executive Summary

The present deliverable reports the results of Task 5.5 “Common design requirements”. This Task “identifies and describes common requirements for the design and architecture of new datasets, to optimize their cross-discipline interoperability” and “produces recommendations for optimizing dataset design as regards interoperability”[1].

Data Interoperability is one of the FAIR properties of data, the others being Findability, Accessibility and Re-usability[2]. The PARTHENOS project has paid great attention to ‘data FAIRness’ in the context of Task 3.2 – Quality assessment of digital repositories, data and metadata. In particular, Task 3.2 has identified the FAIR principles as crucial in establishing the quality of data, and has provided guidelines for quality assessment, published in final form in Deliverable D3.2 of the Project.

Needless to say, the outcomes of Task 3.2 are very relevant to the present deliverable. In particular, Section 2 recapitulates the recommendations for data FAIRness presented in D3.2, and for each such recommendation, it discusses whether and how the PARTHENOS Research Infrastructure complies with the recommendation. This is done not only for the recommendations concerning Interoperability, but for all of them.

While discussing each recommendation, the present report also indicates the documents, mostly PARTHENOS Deliverables, that contain the relevant technical information for making data published on or produced within the PARTHENOS Infrastructure FAIR. This technical information indicates to data providers what they need to do when publishing data on the PARTHENOS Infrastructure. In other words, what is recommended for optimizing dataset design as regards findability, accessibility, interoperability and re-use.

Section 3 presents the conclusions.

[1] Both quotations in this sentence are from the PARTHENOS Description of Work, Work Package 6.
[2] https://www.force11.org/group/fairgroup/fairprinciples

Recommendations to FAIRify data management and their implementation in the PARTHENOS Infrastructure

FINDABLE

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

Recommendations for findability

- Each resource must be assigned a permanent and unique identifier which can be used for determining the location of the representation of the original authoritative copy. A suitable standard from the area of language resources is ISO 24619:2011 ("Language resource management -- Persistent identification and sustainable access (PISA)"). The choice of a persistent identifier schema must rely on careful assessment of advantages and disadvantages. Suitable example implementations for these are: handle systems including Digital Object Identifiers (DOI), and URNs.

The original data must be taken care of in the original infrastructure, the PARTHENOS infrastructure does not impose any requirement on this. The metadata that are collected from the original infrastructure are identified in the PARTHENOS infrastructure by means of a Unique Universal Identifier, or UUID, which is an identifier of a well-known kind[3]. That UUID is embedded into a unique Internationalized Resource Identifier, or IRI[4], of the HTTP protocol that resolves to that metadata record within the PARTHENOS Catalogue. The resolver is a service of the D4Science infrastructure at data.d4science.org. The same goes for the datasets that are generated in the PARTHENOS infrastructure as a result of a processing operation. Each of these datasets is uniquely identified by a UUID that is embedded into a unique IRI as described above. The generated identifiers are persistent and resolved as described.

[3] See, for example https://en.wikipedia.org/wiki/Universally_unique_identifier

[4] IRIs are the identifiers of resources used on the Web. See, for example https://en.wikipedia.org/wiki/Internationalized_Resource_Identifier

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

The PARTHENOS infrastructure does not offer a preservation service; this is demanded of the participating RIs. This is the same for metadata. The data generated in the PARTHENOS infrastructure are made available for the period of time during which the user uses the Infrastructure. After that period, which typically spans the lifetime of the project in the context of which the activity takes place, a Service Level Agreement is negotiated with the Infrastructure which may include the long-term preservation of the data.

- For granularity, there is no clear guideline, but the recommendations from ISO 24619 are good to follow:
- The level of granularity of existing identifier schemes for a type of resources should be retained, for example for books there are ISBNs, so this level would be retained.
- An identifier should be assigned if the resource is associated with the complete content of a digital file.
- An identifier should be assigned if a resource is autonomous and exists outside a larger context, such as a collection of poems by one author being used independently of the collection of all works by the same author, hence the collection of poems is assigned a separate identifier despite the fact that it is also part of the larger unit.
- An identifier should be assigned if a resource is intended to be citable apart from any larger unit. The intention is left vague and can be seen as part of the required negotiations between the depositor and the archive.

The PARTHENOS infrastructure does not discard any identifier defined by the providing RIs, therefore the chosen level of granularity is retained. Concerning the data generated by the infrastructure, any file is identified as a separate resource, whether or not it is included in a larger unit such as a collection or a folder.

Additional guidelines concerning descriptions of resources, necessary to make the resources findable:

- Select an appropriate metadata schema for the type of resource being described. Metadata can have various functions, such as citation metadata, disciplinary metadata, preservation information, provenance, etc. The metadata intended for findability are the type of metadata used for citation and descriptive data in a catalogue. This should be the principal format for maintaining the descriptive metadata. Utilise existing metadata schemas, such as schemas according to ISO 24622-1 (Component Metadata Infrastructure, adjustable to each type of resource), or MARC21 (if appropriate for the type of data). Using only less detailed schemas for describing research data, such as Dublin Core or DataCite MDS, is not recommended.

The PARTHENOS infrastructure has adopted the PARTHENOS Entities Model as a metadata schema for any type of resource being managed by the Infrastructures. The only exception are the data generated in the infrastructure as a result of a processing operation: for the description of the provenance of these, data the PROV-O standard is used.

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

The PARTHENOS infrastructure provides metadata in the PARTHENOS Entities Model as an RDF-XML document, available also from a SPARQL end-point offering different export formats.

- The metadata provided should be high-quality, i.e. as correct and complete as possible, including enough information for later access and comprehensibility.

This is the responsibility of the research infrastructure for the provided data/metadata. For the data generated on the infrastructure as a result of a processing operation the PROV-O description is as accurate as it can be.

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

Any persistent identifier schema is accepted as long as it provides the required functionality.

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

UUID and URI as described above.

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

These reference resources to be used in conjunction with the PARTHENOS Entities Model descriptions are given in Deliverable D5.7.

ACCESSIBLE

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

Recommendations for accessibility

Recommendations for resources that can be retrieved based on their identifiers:

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

The PARTHENOS Infrastructure uses HTTP IRIs.

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

The PARTHENOS Infrastructure provides a resolver for the HTTP IRIs used to identify resources and has put in place the procedures to ensure that this resolution service will continue to function in the long-term.

- Describe the access restrictions and licenses of a resource in the metadata.

The PARTHENOS Entities Model meets this requirement by making the description of the restrictions and licensing an option when crating the description. This has been made only optional for maximum flexibility.

- implement an Authentication, Authorization, and Access Infrastructure (AAAI or AAI).

The PARTHENOS Infrastructure provides an internal AAAI that is federated with several external identity providers (IPs) for maximum usability. Currently, the following IPs are federated:

- EOSC
- Google
- LinkedIn

It is expected that this set will increase in the future.

Recommendations for resources that can be retrieved based on their descriptive metadata:

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

This is not yet implemented on the PARTHENOS infrastructure.

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

The PARTHENOS Infrastructure provides a SPARQL end-point fulfilling this recommendation, at: https://virtuoso.parthenos.d4science.org/sparql.

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

This is not yet implemented in the PARTHENOS infrastructure.

INTEROPERABLE

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

Recommendations for interoperability

- Give an easy to find and detailed overview of accepted (meta)data formats, ideally in a single page that can be referenced directly.

The PARTHENOS Infrastructure can cope with any (meta)data format. For metadata the additional requirements are: the XML UTF-8 format and the mapping from the original format to the PARTHENOS Entities Model. The latter requirement guarantees the proper usage of the metadata, for instance for discovery purposes.

- Present the possibilities for interoperability in a finely granulated and well-structured way, making use of up-to-date design and user interface methodology.

Both the PARTHENOS Aggregator and Infrastructure provide a detailed documentation about interoperability, in Deliverables D5.6 Report on mapping (final) and D6.4 Report on services and tools (final), respectively.

- Document and give easy access to the data model or models in use in a repository. Make clear which parts of the data model enable interoperability, and which parts are relevant when connecting datasets between projects.

The PARTHENOS Entities Model is fully described in Deliverable D5.5 Report on the common semantic framework (final), while the reference resources to be used in conjunction are detailed in D5.7 Report on the integration of reference resources.

- Develop, as a joint effort between repositories, scripts and tools for the (automatic) transformation of data in the ingest phase, enabling interoperability at an early stage.

In the PARTHENOS Infrastructure, metadata mappings are defined using the 3M editor, while the PARTHENOS Aggregator provides a sophisticated suite of tools for the application of those mappings and the quality management of the transformation of the data. These tools are described in Deliverable D5.6 Report on mapping (final).

Additional recommendations on actionable metadata:

- Establish quality assurance processes, with a special focus on the data creation phase.

Quality insurance of the ingested metadata are detailed in the previous point. As far as the data generated in the PARTHENOS Infrastructure, the quality is ensured by the accuracy of the PROV-O record that is automatically associated with the data. This is due to the fact that the PARTHENOS Infrastructure is not a data centre but only an infrastructure that federates existing infrastructures for the purpose of making services available across them. Thus, the only data that are generated in the infrastructure are those resulting from processing operations.

- Combine and apply the push of data providers and automatic processes to boost (meta)data quality.

For metadata, the PARTHENOS Infrastructure has in place the following mechanisms to boost quality:

- The metadata are collected periodically so that improvements implemented by the provider side can be acquired automatically whenever available.
- The mappings from the original format can be re-applied any time, so as to reflect, for instance, changes in the data model. This also guarantees that improvements to the mappings can be exploited whenever available.
- The values in the metadata fields can be harmonized according to controlled vocabularies through the Metadata Cleaner service, which is part of the suite of tools in the PARTHENOS Aggregator.
- Invest in tools for cleaning up (meta)data and converting raw data into other, standardized and interoperable, data formats.

See previous point.

- Establish well documented machine-actionable APIs for the (meta)data.

As already pointed out above, a SPARQL end-point is available to access the PARTHENOS data. Moreover, discovery of data and metadata is also supported by offering a Solr access point.

- Give more information on best practices for machine driven automatic data search and reuse.

This information is given in the Appendix II PARTHENOS Entities Minimal Metadata of Deliverable D.5.5.

Additional recommendations on shared vocabularies and/or ontologies for (meta)data formats:

- The description of metadata elements should follow community guidelines that use an open, well defined vocabulary.

The PARTHENOS Entities Model is described in the style of the CIDOC CRM model that has a wide community of users and a long history of usage in prestigious memory institutions such as the German Library and the British Museum.

- Convince researchers to use FAIR compatible vocabularies and ontologies from the very start. Give recommendations on how to do this and how to integrate references in their research data and metadata.

In PARTHENOS, this is demanded of the participating Research Infrastructures. However, the project has indicated clear guidelines how to do this in Deliverable D3.2 Guidelines for common policies implementation (final).

- Give pointers to vocabularies and ontologies that can be used, based on research domain specifics and tangible use cases.

In addition to the above mentioned Deliverables on the PARTHENOS Entities Model and associated Reference Resources, PARTHENOS has dedicated a specific Work Package on Standardization, WP4. This WP has produced Deliverable D4.4 Report on standardization (final) and an associated tool D4.3 Standardization Survival Kit (final) point to vocabularies and ontologies that can be used by researchers in the Humanities.

REUSABLE

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

Recommendations for reusability

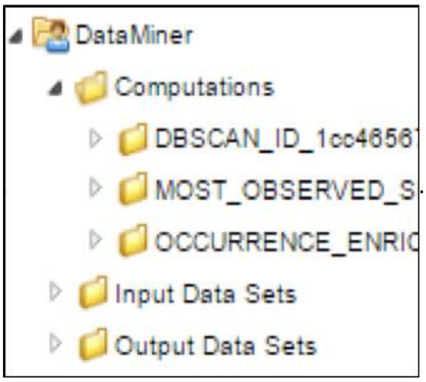
- Document data systematically. To make clear what can and what cannot be expected in a dataset or repository, data should be systematically documented. Being transparent about what's in the data and what isn't facilitates trust and, consequently, data reuse.

This is the responsibility of the participating Infrastructures for all the research data collected by the PARTHENOS Infrastructure. For the data generated in the PARTHENOS Infrastructure via processing operation, the PROV-O record is always attached. Moreover, users can add their own metadata to that record.

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

As for the previous point, this is the responsibility of the participating Infrastructures for all the research data collected by the PARTHENOS Infrastructure. For the data generated in the PARTHENOS Infrastructure via processing operations, file naming is performed automatically following a simple naming convention that would enable users to retrieve, access and understand data objects. In particular (see figure below), the data are placed into a folder that is created for this purpose and is named with the identifier of the process that creates them. This folder is placed into the folder DataMiner/Computations in the

Workspace of the user. The data produced by an execution of the process are placed in a file into the process folder; the name of file is obtained by concatenating:


DataMiner
Computations
DBSCAN_ID_1cc4656
MOST_OBSERVED_S
OCCURRENCE_ENRIC
Input Data Sets
Output Data Sets

- the name of the process
- the string “ID”
- the UUID of the process execution

For instance, *DBSCAN_ID_1cc4564-36385fjs-2873…*

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

As for the previous point, this is the responsibility of the participating Infrastructures for all the research data collected by the PARTHENOS Infrastructure. For the data generated in the PARTHENOS Infrastructure via processing operations, the format of the file depends on the specific processing software. The export format of the provenance metadata is XML, which is a standardised format, widely used across the communities.

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

The Storage component of the PARTHENOS Infrastructure checks the integrity of the data by performing a standardised control.

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

For the data provided by the participating Infrastructures, the PARTHENOS Entities Model allows specifying who the data rights holder is and what licence applies. For the data generated in the PARTHENOS Infrastructure via processing operations, it is the owner of the Workspace who decides the visibility and accessibility of the data. The chosen solution ranges from total accessibility of the generated data, with publication of the description in the Joint Resource Registry, to no accessibility. The project is also working on connecting the PARTHENOS Infrastructure with other Infrastructures, such as B2Share.

Conclusions

The present report provides recommendations to the PARTHENOS Community for optimizing dataset design as regards findability, accessibility, interoperability and re-use.

It achieves this goal by reviewing the recommendations for data FAIRness presented in the PARTHENOS Deliverable D3.2, and by discussing, for each such recommendations, whether and how the PARTHENOS Research Infrastructure complies with the recommendation.

As it turns out, the PARTHENOS Infrastructure meets all recommendations for FAIRness given in D3.2, except two concerning the resources that no longer exists. These two recommendations concern accessibility, therefore from the interoperability point of view the PARTHENOS Infrastructure is fully compliant with the recommendations given in D3.2. Nevertheless, the D4Science management has been informed about these two recommendations and is considering implementing them.

The present report also indicates the documents, mostly PARTHENOS Deliverables, that contain the relevant technical information for making data published on or produced within the PARTHENOS Infrastructure FAIR. This technical information indicates to data providers what they need to do when designing datasets to be published in the PARTHENOS Infrastructure.

Acknowledgements

The author thanks Hella Hollander, Francesca Morselli and Femmy Admiraal at KNAW DANS for their support on the FAIR principles and for making earlier version of Deliverable D3.2 available. The author also thanks his colleagues Alessia Bardi and Massimiliano Assante at the NeMIS Laboratory of CNR ISTI for their contributions on the technical features of the PARTHENOS Infrastructure.