SCIRES it
SCIentific RESearch and Information Technology
Ricerca Scientifica e Tecnologie dell'Informazione
Vol 8, Issue 1 (2018), 33-50
e-ISSN 2239-4303, DOI 10.2423/i22394303v8n1p33

AN AGGREGATION FRAMEWORK FOR DIGITAL HUMANITIES INFRASTRUCTURES: THE PARTHENOS EXPERIENCE

Luca Frosini **, Alessia Bardi*, Paolo Manghi*, Pasquale Pagano**

* Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", Italian National Research Council - Pisa, Italy
** Dipartimento di Ingegneria della Informazione, Università di Pisa, Italy

Abstract

Digital Humanities Infrastructures (DHIs) are research infrastructures supporting researchers in the field of humanities by providing ICT tools and facilities for performing their studies and investigation activities. A DHI typically serves either researchers of one specific sector of humanities (e.g. history, archaeology) or focused research groups working on specific research topics (e.g. studies on the holocaust, on a specific manuscript), with little or no re-use of tools, services and data that could be shared and successfully adopted to answer research questions of different research disciplines. This fragmentation often represents a barrier to inter-disciplinary research collaborations. We present a technical framework for the federation of DHIs where tools, data, services, and knowledge available from each DHI are shared in an integrated environment where researchers can collaborate on specific research topics by creating customized Virtual Research Environments.

Keywords

Research infrastructures, e-infrastructures, Hybrid Data Infrastructures, PARTHENOS, Virtual Research Environments, Aggregative Data Infrastructures

1. *Introduction*

Research infrastructures (RIs) are "facilities, resources, and services used by the science community to conduct research and foster innovation"[1]. Researchers' needs for digital services led to the realization of e-Infrastructures, i.e. RIs offering digital technologies for data management, computing and networking. Relevant examples are high speed connectivity infrastructures (e.g. GÈANT), computing infrastructures (e.g. European Infrastructure EGI), scholarly communication infrastructures (e.g. OpenAIRE), data e-infrastructures (e.g. D4Science).

Digital Humanities Infrastructures (DHIs) are e-Infrastructures supporting researchers in the field of Humanities with a digital environment where they can find and use ICT tools and research data for conducting their research activities.

A growing number of DHIs have been realized, most of them targeting a specific sector of the Humanities. Thanks to their discipline-specific feature, DHIs offer specialized services and tools to researchers, who are now demanding support for interdisciplinary research, common solutions for data management, and access to resources traditionally relevant to different sectors (e.g. text-mining algorithms traditionally used by linguistics can also be useful to historians and social scientists).

To close this gap, the European Commission launched the PARTHENOS project (Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies - EC-H2020-RIA grant agreement 654119)[2]. The main goal of the PARTHENOS project is to bridge existing DHIs by forming a federation where researchers of different sectors of the Humanities can collaborate and share data, services and tools in an integrated environment.

Within PARTHENOS, a complete technical framework is produced for the federation of DHIs, enabling transparent access to resources managed by different DHIs and enabling the creation and operation of Virtual Research Environments (Blanke, Candela, Hedges, Priddy, & Simeoni, 2010; Candela, Castelli, & Pagano, 2013).

[1] https://ec.europa.eu/research/infrastructures/index.cfm

[2] http://www.parthenos-project.eu/

According to Candela et al. (Candela, Pagano, Castelli, & Manzi, 2014) "VREs are web-based working environments where groups of scientists, possibly geographically distant from each other, have user friendly, transparent and seamless access to the flexible and shared set of remote resources (data, services and computing capabilities) needed to perform their work collaboratively".

The technical framework supports the realisation of the federation by offering tools for:

- The creation of a homogeneous information space where all resources (data, services, and tools) of the different DHIs are described according to a common data model;
- The discovery of available resources;
- The use of available resources (e.g. for download or processing);
- The dynamic creation of VREs where users can find resources relevant for a research topic, run services, and share the computational results.

Section 2 provides an overview of the DHIs federated in the context of PARTHENOS project. Section 3 describes the technical framework, its configuration for the PARTHENOS use case based on the PARTHENOS Entities Model (section 3.1) and its main components: the Content Cloud Framework (section 3.2) and the Joint Resource Registry (section 3.3). Section 4 discusses related work. Conclusions and next steps are presented in section 5.

2. *The PARTHENOS Federation of Digital Humanities Infrastructures*

Many European and national projects funded the creation of DHIs in different research sectors. Some of them are collaborating in the PARTHENOS context to create a federation allowing researchers in different fields to collaborate, share data, services and tools in an integrated environment (Bardi & Frosini, 2017). In particular, the DHIs federated by PARTHENOS are: ARIADNE[3] (archaeology); CENDARI[4] (history); CLARIN[5] (linguistic studies); CulturaItalia (archaeology, arts and humanities); DARIAH[6] (arts and humanities); EHRI[7] (studies on the holocaust); TGIR Huma-Num[8] (humanities and social sciences) (Aloia, Candela, et al., 2017).

ARIADNE (Aloia, Debole, Felicetti, Galluccio, & Theodoridou, 2017; Meghini et al., 2017) brings together and integrates existing archaeological research data infrastructures so that researchers can use the various distributed datasets and new and powerful technologies as an integral component of the archaeological research methodology. Resources include data, services, language resources. The ARIADNE registry is addressed to cultural institutions, either private or public, that wish to describe their assets in order to make them accessible within the research communities. The registry data model, called ACDM (ARIADNE Catalogue Data Model), extends the Data Catalog Vocabulary (DCAT)[9] and the ISO/IEC 11179 'Specification and Standardization of Data Elements' (management & interchange Technical Committee, 1999). The central notion of the model is the class *ArchaeologicalResource*, specialized as: *DataResource*, which represents the various types of data containers (e.g. databases, GIS, collections, datasets) owned by the ARIADNE partners; *LanguageResource, which can be used to model* vocabularies, metadata schemas, gazetteers and mappings between language resources; and *Services*, used to model services owned by the ARIADNE partners.

CENDARI (Boukhelifa et al., 2018; Gartner & Hedges, 2013) (Collaborative European Digital Archive Infrastructure) is a research infrastructure project aimed at integrating digital archives for medieval and modern European history. The CENDARI Repository holds information on over 1,200 collection holding institutions and over 300,000 records (metadata and some full text) relevant for the study of the medieval period and First World War. Main entities of the model are: *Places/spaces* i.e., geographic locations relevant to research topics or other contextual entities; *Persons/role* i.e., individuals associated with the research topics or other contextual entities, and their associated

[3] ARIADNE: http://www.ariadne-infrastructure.eu/
[4] Collaborative European Digital Archive Infrastructure (CENDARI): http://www.cendari.eu/
[5] Common Language Resources and Technologies infrastructures (CLARIN): https://www.clarin.eu/
[6] Digital Research Infrastructure for Arts and Humanities (DARIAH) http://www.dariah.eu/
[7] European Holocaust Research Infrastructure (EHRI): https://www.ehri-project.eu/
[8] https://www.huma-num.fr/
[9] https://www.w3.org/TR/vocab-dcat/

roles; *Institutions* i.e., organizations associated with the research topics or other contextual entities; *Dates* i.e., specific dates or periods of time associated with specific events, people, or institutions; *Events* i.e., notable events associated with the particular topics, as well as with other contextual entities; *Topics* i.e., subjects associated with the two research areas.

CLARIN (Váradi, Wittenburg, Krauwer, Wynne, & Koskenniemi, 2008) is a research infrastructure initiative that aims at providing a single access point to language resources and language technology to researchers from the Social Sciences and Humanities (SSH) disciplines. In the context of CLARIN, several European initiatives have been carried out in the last decade with the goal of improving and unifying the documentation of language resources. CLARIN provides the Virtual Language Observatory (VLO) (Van Uytvanck, Stehouwer, & Lampen, 2012) faceted browser to explore linguistic resources, services and tools available within CLARIN. All information in the VLO is based on the metadata descriptions of resources as provided by the parties (CLARIN centres[10]) that host the original data. Unfortunately, the unification is not yet completed for the registries of META-SHARE and Language Resource and Evaluation Map (LRE Map). For those, resources are accessible from the original sources. The META-SHARE (Gavrilidou et al., 2012) registry is designed as a network of distributed repositories of language resources, including language data and basic language processing tools (e.g., morphological analysers, PoS taggers, speech recognizers, etc.). The LRE Map initiative issued out of the FLaReNet project (Soria et al., 2012), whose mission was to develop a common vision of the area of language resources and to foster a European strategy for consolidating the sector, thus enhancing competitiveness at EU level and worldwide. Today the LRE Map is a repository of data, documenting language resources using a lightweight metadata scheme (Calzolari et al., 2012).

CulturaItalia[11] is the Portal of Italian Culture, managed by the Central Institute for the Union Catalogue of Italian Libraries (ICCU) of Ministry of Cultural Heritage, Activities and Tourism (MiBACT). CulturaItalia is an aggregator playing an important role for the development of European research infrastructures on Cultural Heritage such as ARIADNE, DARIAH and Europeana[12], making available cooperative networks and agreements and coordinating technical activities. The catalogue of CulturaItalia currently manages about three million metadata records from museums, libraries, foundations and institutions, both public and private. CulturaItalia contains a section devoted to open data. It offers access to metadata in CIDOC-CRM[13] (Antoniou, Christophides, Plexousakis, & Doerr, 2005; interoperability Technical Committee, 2014) format via a SPARQL endpoint and it features an OAI-PMH Publisher that makes metadata available as XML or RDF structured according to different schemas: Dublin Core, PICO Application Profile, Europeana Data Model[14], and CIDOC-CRM[15].

DARIAH (Blümm & Schmunk, 2016) is a pan-European infrastructure for arts and humanities scholars working with computational methods. The mission of DARIAH is to enhance and support digitally enabled research across the humanities and arts. It supports digital research as well as the teaching of digital research methods. DARIAH currently connects several hundreds of scholars and dozens of research facilities in 17 European countries. The DARIAH repository contains publications (e.g. journal articles, conference paper, books, posters, patents); documents (e.g. preprint, working papers, report); academic works (e.g. theses, lectures); and research data (e.g. photos, videos, maps, audios) organised in so called "collections". Researchers can search for content in the repository via the DARIAH Collection Registry. A collection description contains information about the services offering access to the collection, its location and the responsible agents (e.g. curating institution). It may also carry collection specific metadata such as spatial and temporal coverages of the contained objects. The DARIAH Collection Registry can be accessed via a web portal, an OAI-PMH Publisher, and a REST web service. Among the 17 national nodes of DARIAH, PARTHENOS federates DARIAH-GR/DYAS and DARIAH-DE. Additional national nodes may join the federation at any time.

EHRI (European Holocaust Research Infrastructure) (Bryant, Reijnhoudt, Speck, Clerice, & Blanke, 2014) seeks to overcome one of the hallmark challenges of Holocaust research: the

[10] https://centres.clarin.eu/
[11] http://dati.culturaitalia.it/
[12] https://www.europeana.eu/
[13] http://www.cidoc-crm.org/
[14] https://pro.europeana.eu/page/edm-documentation
[15] http://erlangen-crm.org/

wide dispersal of the archival source material across Europe and beyond, and the concomitant fragmentation of Holocaust historiography"[16]. It provides users with a range of tools to find, explore, organize and share such information. EHRI makes the available sources accessible via an online platform and offers tools and methods that enable researchers and archivists to collaboratively work with such sources. Apart from providing an online platform, EHRI also facilitates an extensive network of researchers, archivists and other stakeholders to increase cohesion and co-ordination among practitioners and to initiate new transnational and collaborative approaches to the study of the Holocaust.

Très Grande Infrastructure de Recherche (TGIR) Huma-Num[17] is a research infrastructure aimed at facilitating the turning of digital research in the humanities and social sciences. The TGIR Huma-Num offers services dedicated to the production and reuse of scientific data. To do this, Huma-Num supports research teams throughout their digital projects to allow the sharing, reuse and preservation of data thanks to a chain of devices focused on interoperability. The aim is to foster the exchange and dissemination of metadata and data via standardized tools and lasting, open formats. Two of the main services offered by Huma-Num are NAKALA and ISIDORE. NAKALA provides functionality for accessing data, export metadata, and assign persistent identifier to data and metadata. ISIDORE is a platform allowing access to digital data in the Humanities and Social Sciences. Its architecture relies on the languages of the semantic web (RDF/RDFS/OWL) and provides open access to data.

3. *Technical Framework*

In order to create the federation of the above mentioned DHIs (see chapter 2) and to provide Virtual Research Environments, the technical framework has to deal with the following challenges:

- Create a homogeneous information space where all resources (data, services and tools) of the different DHIs are described according to a common data model;
- Provide discovery of all available resources;
- Provide access to available resources (for download or processing);
- Provide support for on-demand creation of VREs where users can find resources relevant for a research topic, run services, and share the results of data analytical studies.

Fig. 1 shows an overview of the main components of the technical framework: the PARTHENOS Content Cloud Framework (CCF) and the Joint Resource Registry (JRR).

The Content Cloud Framework supports the aggregation workflow that starts with the collection of the metadata about resources from the DHIs of the federation and manage the harmonization, curation, collation, and publication of those resources in the homogenous information space. It is composed of the PARTHENOS Aggregator, the 3M Editor and a set of endpoints exposing the aggregated resources according to different (de-facto) standard protocols.

The PARTHENOS Aggregator is realized with the D-NET software toolkit (Manghi et al., 2014),

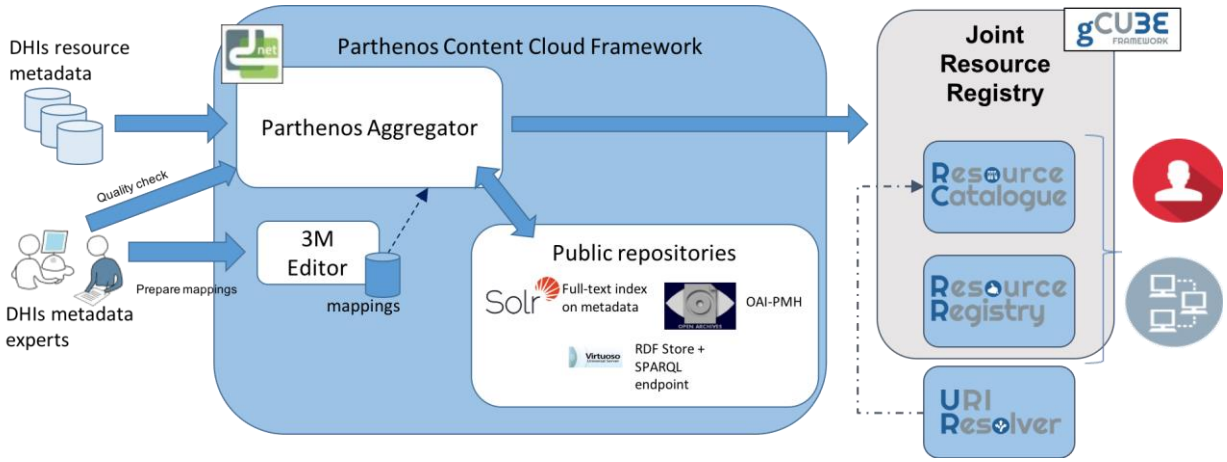

Fig. 1: The PARTHENOS Technical Framework

[16] https://www.ehri-project.eu/

[17] https://www.huma-num.fr/

an enabling framework for the realization of Aggregative Data Infrastructures (ADIs) developed and maintained by CNR-ISTI[18]. D-NET provides functionality for the automatic collection, harmonization, curation and delivery of metadata coming from a dynamic set of heterogeneous data providers. In the context of the PARTHENOS project, D-NET has been configured to collect metadata made available by existing DHIs operated by PARTHENOS partners and harmonize them according to the PARTHENOS Entities Model (PE model) (Bruseker, Doerr, & Theodoridou, 2017) (see model rationale in paragraph 3.1) by applying X3ML[19] (Marketakis et al., 2017) mappings defined by DHIs metadata experts via the 3M Editor. When defining the mappings, DHIs metadata experts can also decide to generate persistent URIs in the PARTHENOS namespace for the resources. The mapping language, editor and execution engine are realized and maintained by FORTH[20]. Aggregated content is then published via different endpoints, supporting a set of (de-facto) standard protocols for metadata search (Solr API, SPARQL) and exchange (OAI-PMH).

The aggregated content is also ingested into the Joint Resource Registry, which exposes an end-user GUI, i.e. the Resource Catalogue, and a machine-oriented API for resource discovery, i.e. the Resource Registry. Data and services registered in the JRR become discoverable by and accessible to users and other services of the federation. Moreover, the JRR provides functionality for infrastructure management. For example, a user can run a CLARIN service for full-text mining on a dataset of medieval full-texts provided by CENDARI because the two resources, dataset and service, would have been provided in the same administrative domain. Computational results can be easily stored and shared with a selection of colleagues or publicly, by publishing them into the JRR. The JRR is based on the gCube enabling technology (Candela & Pagano, 2015), an open-source software toolkit used for building and operating Hybrid Data Infrastructures (Candela, Castelli, & Pagano, 2012) enabling the dynamic deployment of Virtual Research Environments by favouring the realization of reuse oriented policies. gCube is developed and maintained by CNR-ISTI.

Fig. 1 also shows the URI Resolver, which is the component in charge of resolving the persistent URI generated by PARTHENOS mappings to the corresponding URL of the Resource Catalogue. The URI Resolver service allows to decouple the unique and persistent URI associated with the resource from the technology and the service managing the resource.

3.1 PARTHENOS Entities Model

The PARTHENOS Entities Model (PE Model) (Bruseker et al., 2017) aims at capturing and representing the knowledge generation process: which actors and which services are involved in the knowledge creation, resource curation, and management chain.

The PE Model is formalized by using CIDOC-CRM and its extension CRMdig (Doerr & Theodoridou, 2014). The former is able to capture the knowledge of cultural heritage objects, while the latter to describe the provenance of information and the digitization process. The PE Model is available as ontological model in RDFS[21]. The current version (version 2.1) defines a total of 33 classes and 37 properties extending CIDOC-CRM and CRMdig entities.

The PE Model has been defined starting from the analysis of the entities provided by the registries of the federated DHIs described in section 2 and it is based on five main entities:

1) *Dataset* is a set or collection of data, records or information that is kept as a persistent unit of information in the knowledge generation process. This concept occurs in different forms or denominations in the various registries:
   a) The ARIADNE *DataResource*, a class describing resources that are data containers (such as databases, GIS, collections or datasets) and the ARIADNE *LinguisticResources*, a class that has as instances resource of a linguistic nature, whether in natural language (such as a gazetteer) or in a formal language (such as a vocabulary, metadata schema and mapping between schemas). *LinguisticResources* (vocabulary, authority files) are also in the EHRI registry and in the LRE Map registry (ontology).

[18] http://www.isti.cnr.it/
[19] http://www.ics.forth.gr/isl/index_main.php?l=e&c=721
[20] https://www.forth.gr/
[21] The PARTHENOS Entities Model RDFS: http://parthenos.d4science.org/CRMext/CRMpe.rdfs

b) The CENDARI Archival descriptions: Dataset of *Places*, *Topics* (Concepts), *Events* and *Dates* (Timespans).
c) The CLARIN Dataset of concepts (CCR), Dataset of metadata (CMDI), Educational package and textual file of different format (NL Resource List).
d) The Dataset of archival descriptions (EHRI, CulturaItalia), Datasets from museum, library (CulturaItalia), Dataset of collections (DARIAH-GR/DYAS).
e) All TGIR Huma-Num datasets vocabularies and collections.

2) *Actor* is an institution, a team or an individual person that participates in the research infrastructure as partner providing data and/or services. This entity is present in every federated registry with the name: *Person*, *Organization*, *Institution*, or *Agent*.
3) *Service* is defined as the continued, declared willingness and ability of an actor to execute on demand certain activities of specific benefit to the client. Most of the surveyed registries contain such an entity without any specification, while ARIADNE classifies services as *StandAloneService*, *WebService*, *ServiceForHumans*, *InstitutionalService*.
4) *Software* is an artefact that can be executed on a computer to perform specific operations. This entity occurs in the LRE Map registry and it is also present in the ARIADNE registry as a specialization of the Service class with the name of *StandAloneService*.
5) *Knowledge* generation process represents the workflow of the processes used to produce specific datasets.

The PE Model defines the categorical descriptions of those five entities through a minimal metadata set. It is important to note that the mappings and transformations to the PARTHENOS Entities (PE) are by design lossy: the model does not aim at representing all aspects of the source data in the target model, but rather establish the identities of the main entities and relations between them.

3.2 Content Cloud Framework

The Content Cloud Framework supports the realization of a single-entry point to resources available from different DHIs. The framework generates a digital information space where resources of different DHIs are described homogeneously and accessible via a set of (de-facto) standard protocols.

The main components of the Content Cloud Framework are the 3M Editor and the PARTHENOS Aggregator. The 3M Editor is part of the X3ML Toolkit web application suite, whose main functionality is to assist users during the mapping definition process. It provides a human-friendly user interface and a set of sub-components that either suggest or validate the user input. Mappings are specified using the X3ML mapping definition language, an XML-based, declarative, and human readable language that supports the cognitive process of a mapping in such a way that the generated maps can be collaboratively created and discussed by domain experts with little to no IT knowledge (Marketakis et al., 2017).

In the context of PARTHENOS, data experts define mappings from each DHI model to the PE Model. Those mappings are then used by the Aggregator component to transform the input metadata records and generate the information used to populate the uniform information space.

The PARTHENOS Aggregator is realized with the D-NET Software Toolkit (D-NET for brevity) (Manghi et al., 2014), a service-oriented framework specifically designed to support developers at constructing custom aggregative infrastructures in a cost-effective way. D-NET offers data management services capable of providing access to different kinds of external data sources, storing and processing information objects of any data models, converting them into common formats, and exporting information objects to third-party applications through a number of standard access APIs. D-NET services are obtained by encapsulating advanced and state-of-the-art open-source products for data storage, indexing, and processing – such as PostgreSQL, MongoDB, Apache Solr, and Apache HBase – in order to serve broad application needs. Most importantly, D-NET offers infrastructure enabling services that facilitate the construction of domain-specific aggregative infrastructures by selecting and configuring the needed services and easily combining them to form autonomic data processing workflows. The combination of out-of-the box data management services and tools for assembling them into workflows makes the toolkit an appealing starting platform for developers having to face the realization of aggregative infrastructures.

Data management services that are typically combined for the implementation of custom aggregation workflows are:

- *Collection service:* the service embeds modules capable of handling the collection of metadata records via different access protocols. Currently, the service supports the collection from local file system and from remote data sources offering APIs implementing the following protocols: OAI-PMH, FTP(S), SFTP, HTTP(S), RESTful. The service is easily extendable to support additional collection modes.
- *Transformation service:* the service addresses the general problem of transforming metadata records from one metadata data model into records of one output metadata data model by applying a mapping. The service integrates a set of transformation engines and thus supports the execution of different types of mappings: XSLT 1.0/2.0, Groovy rules, D-NET mapping rule scripts (an idiosyncratic mapping engine exploiting the transformation language defined and maintained by University of Bielefeld). Additional transformation engines can be further integrated into the Transformation service to support more types of mappings. For example, during the PARTHENOS project, the X3ML Engine has been integrated into the Transformation Service to support the execution of X3ML mappings.
- *Metadata Cleaner service:* the service harmonises values in metadata records based on a set of thesauri. A D-NET thesaurus consists of a vocabulary that is a list of authoritative terms together with associations between terms and their synonyms. Data curators – typically based on instructions from data providers and domain experts – are provided with user interfaces to create/remove vocabularies and edit them to add/remove new terms and their synonyms. Given a metadata format, the Metadata Cleaner service can be configured to associate the metadata fields to specific vocabularies. The service, provided records conforming to the metadata format, processes the records to clean field values according to the given associations between fields and vocabularies. Specifically, field values are replaced by a vocabulary term only if the value falls in the synonym list for the term. If no match is found, the field is marked as 'invalid'. Curation tools may exploit the "invalid" mark to highlight non-cleaned records and suggest either the update of D-NET vocabularies or the update of the values in the input record. D-NET offers such a curation tool through a web GUI called Metadata Record Inspector (see below).
- *Metadata Record Inspector:* a web GUI integrated into D-NET that provides data curators with an overview of the information space, where they can search and browse records and verify the correctness of the transformation phase (e.g. no mapping mistakes or semantic inconsistencies) and the cleaning phase. Upon positive verification of the records, data curators can inform the PARTHENOS infrastructure administrators that the records can be publicly exported.
- *OAI-PMH Publisher service:* the service offers OAI-PMH interfaces to third-party applications (i.e. harvesters) willing to access metadata objects. The service is implemented on MongoDB and can be configured to offer OAI sets based on the record provenance (i.e. data source from which records have been collected) and other criteria based on the common data model.
- *Index service:* the service guides the feeding of Solr indices and it is also responsible for transforming the aggregated metadata records into Solr documents. The transformation is guided by a configuration that the aggregator administrator can change at runtime.

D-NET has been adopted for the realization of several aggregative infrastructures in the Cultural Heritage domain (Bardi, Manghi, & Zoppi, 2014) like the HOPE (Heritage of the People's Europe)[22] infrastructure (Artini et al., 2014), EFG (European Film Gateway)[23], EAGLE[24] (Mannocci, Casarosa, Manghi, & Zoppi, 2015) and in the scholarly communication domain like OpenAIRE[25] (Atzori, Bardi, & Manghi Paolo, 2017; Manghi, Manola, Horstmann, & Peters, 2010), the Spanish national repository aggregator Recolecta[26], and the Latin American aggregator La Referencia[27].

[22] Social history Portal: https://socialhistoryportal.org/about-hope/
[23] European Film Gateway: http://www.europeanfilmgateway.eu/
[24] EAGLE: https://www.eagle-network.eu/
[25] OpenAIRE: https://www.openaire.eu/
[26] Recolecta: https://buscador.recolecta.fecyt.es/
[27] La Referencia: http://www.lareferencia.info/

In the context of PARTHENOS, D-NET has been extended with:

- Specific plugins for the collection of metadata from data sources that are not offering access points compliant to standard exchange protocols;
- The integration of the X3ML engine for the execution of X3ML mappings;
- A new software component for exporting aggregated metadata records in form of RDF resources via a SPARQL endpoint based on the Virtuoso technology;
- A new software component for registering aggregated metadata records into the Joint Resource Registry.

D-NET has also been configured with a custom workflow for the aggregation and export of metadata records, as depicted in Fig. 2. For each federated DHI endpoint, the workflow is configured with proper parameters (e.g. endpoint base URL, endpoint protocol, X3ML mapping to apply) and run autonomously by the PARTHENOS Aggregator.

The workflow comprises two automatic sub-workflows: the aggregation workflow and the provision workflow. The aggregation workflow automatically (i) collects input metadata records and checks if they are well-formed XML records. Non-well-formed records are discarded and stored in a dedicated metadata store where the infrastructure administrator can inspect them to provide a report to the manager of the data source; (ii) transforms input records into RDF/XML files compliant to the PARTHENOS Entities model by applying the specified X3ML mapping. Optionally, before the transformation takes place, records are checked for compliance to a specified XML schema (the XML schema can be automatically detected from each record or provided as a configuration parameter of the workflow); (iii) cleans values in the generated RDF records based on controlled vocabularies selected and/or defined by the PARTHENOS consortium; and (iv) makes the resulting RDF records available to data curators for inspection via the Metadata Record Inspector. Upon positive verification of the records by the data curators, the PARTHENOS infrastructure administrators are notified and the records can be exported to publicly accessible repositories.

The provision workflow is executed in order to export the aggregated records on publicly accessible repositories. The first repository provides access via OAI-PMH to maximise the accessibility of the generated information and the possibility to bulk download the aggregated metadata records in Dublin Core and RDF formats. The PARTHENOS OAI-PMH Publisher has been configured to export one OAI set per research infrastructure. If needed, more OAI sets can be defined. A second repository is based on Virtuoso, a triple store database where all the transformed RDFs of all DHIs are stored. Virtuoso is the first storage where all metadata records of all DHIs are properly integrated. In fact, the storage devices used in the aggregation workflow are silos where only the records collected from the same data source endpoint are available. In Virtuoso, instead, we can find resource descriptions whose provenance may include distinct DHIs, because several DHIs may offer metadata descriptions

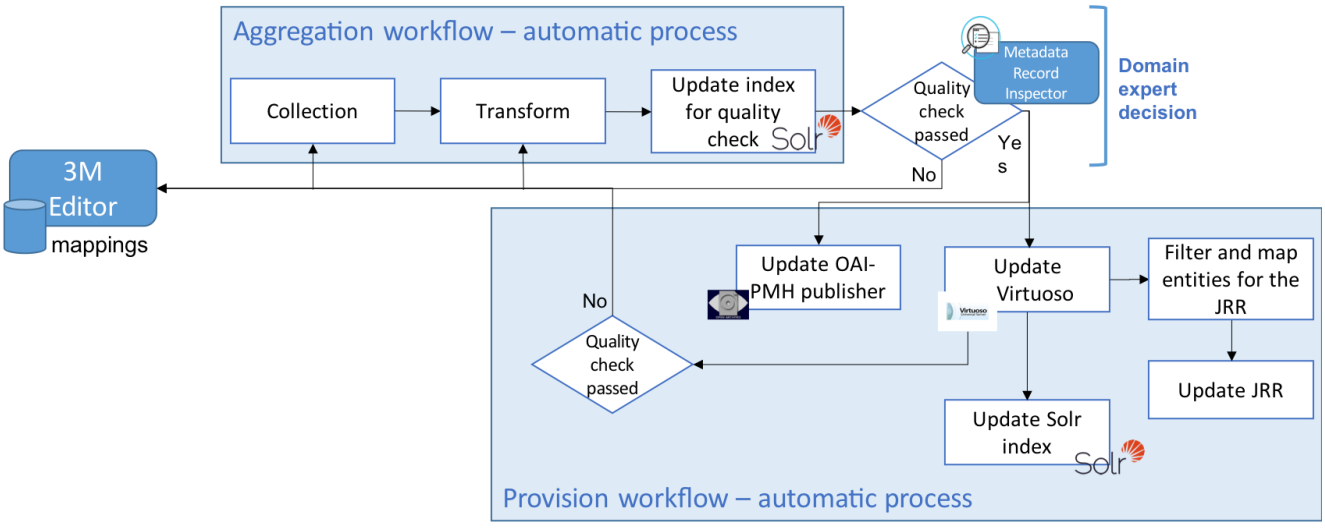

Fig. 2: Aggregation and provision workflows of the PARTHENOS Aggregator

about the same digital or real-world resource. Since the correct integration and cross-references among resources offered by different DHIs is a fundamental feature of the aggregated information space, PARTHENOS data curators are invited to perform an additional step of quality check by running SPARQL queries on Virtuoso. Whenever unexpected results are found, data curators, together with the aggregator operators, perform an investigation to identify the reason of the unexpected results and the actions to be performed to fix them. Then a third repository based on Solr is used to index all the generated information. Finally, the Joint Resource Registry (JRR) is fed with the information to serve web GUI for resource search and discovery. The Solr and JRR repositories are populated starting from the integrated information collected in Virtuoso, maximising the benefits of the integration phase.

The consistency among the different repositories is ensured by the provision workflow that performs all the feeding processes in consistent transactions. Being this workflow completely automatic, tested, and validated it also ensures a repeatable and affordable process. The four different repositories are tailored manifestations of the common information space providing access to the resources according to different standard protocols. This approach was selected to boost the accessibility to the integrated resources, ensure the take up from the different sectors of humanities, and remove all possible barriers for the exploitation of the integrated resources.

3.3 The Joint Resource Registry

The Joint Resource Registry is the PARTHENOS component designed to support the management of the integrated resources. It ensures the correct enforcement of the policies specified by the different DHIs. From the technical point of view, it is composed of two main components: the Resource Registry, providing machine-oriented APIs, and the Resource Catalogue, featuring a human oriented GUI (see Fig. 1).

The Resource Catalogue (RC) is a component of the Human Interaction Framework offered by gCube (Candela & Pagano, 2015). The Resource Registry (RR), instead, is part of gCube Enabling Framework (see Fig. 3).

The gCube Enabling Framework is composed of three main systems: the Resource Management System, the Security System, and the Information System. These are complex ICT systems that exploit tailored persistence technologies managed via web services. The Resource Management System supports the creation of Virtual Research Environments (VREs) and the exploitation of the resources dynamically assigned to it. The Security System ensures the correct exploitation, auditing, and accounting of the resources under the policies defined at registration time and customised at VRE definition time. It is orthogonal to all services operating in the infrastructure and its components

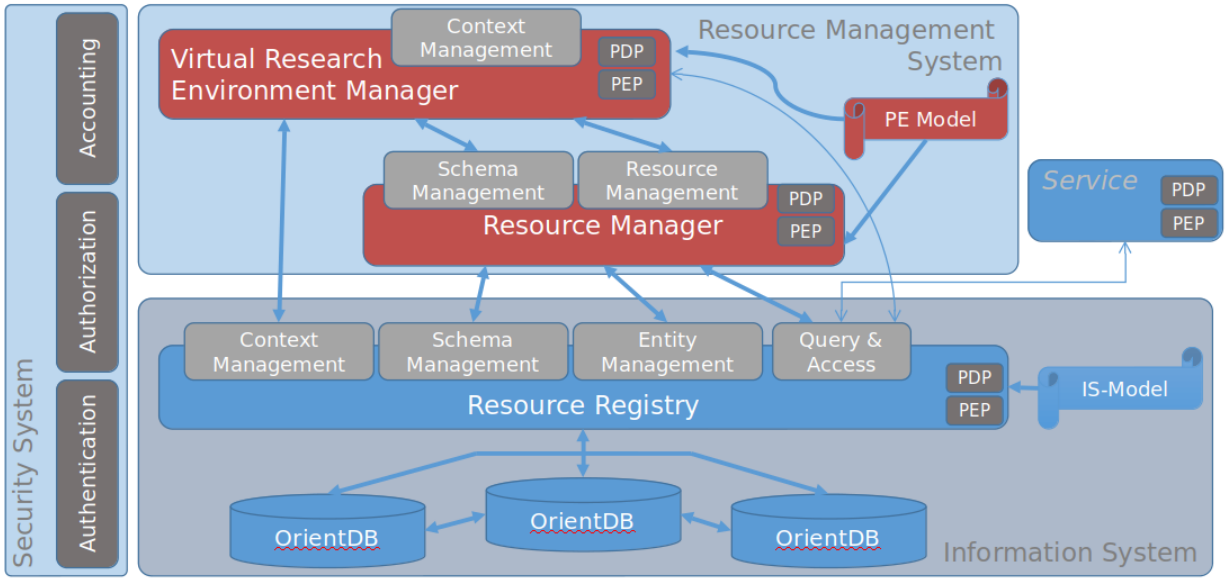

Fig. 3: gCube Enabling Framework

are deployed on all computing nodes. The Information System supports the registration, discovery, and access to resources described in terms of its internal data model, the Information System Model (see 3.3.1). The IS Model defines basic resource types and relationships that have to be specialised by configuring it to meet domain-specific requirements.

In the context of PARTHENOS infrastructure, the gCube Enabling Framework has been configured to support the PARTHENOS Entities Model (see section 3.1). The configured gCube Enabling Framework promotes the optimal exploitation of the resources available in the PARTHENOS Cloud Infrastructure and the integration of technology operated and maintained by external resource providers (i.e. the federated DHIs). It insulates, as much as possible, the management of the infrastructure from the data and the data management services that are hosted by or accessible through the infrastructure itself.

3.3.1 IS Model

The Information System Model (IS Model) is conceived to provide the building blocks needed to develop an information system suitable for data e-infrastructures (Frosini & Pagano, 2018). The model is based on a graph model having Entities as nodes and Relations as edges.

As depicted in Fig. 4, two typologies of Entities are envisaged:

1. *Resources*: entities representing a description of a "thing" to be managed;
2. *Facets*: entities contributing to "build" a description of a Resource.

A Resource is characterised by a number of Facets. A Facet, once attached to a Resource, captures a specific aspect/characterisation of the resource. Every facet is characterised by a number of properties.

Fig. 4: Inheritance model of entities and relations

Two typologies of Relations are envisaged (see Fig. 4):

1. *isRelatedTo*: a relation linking any two Resources;
2. *consistsOf*: a relation connecting a Resource with one of its Facets.

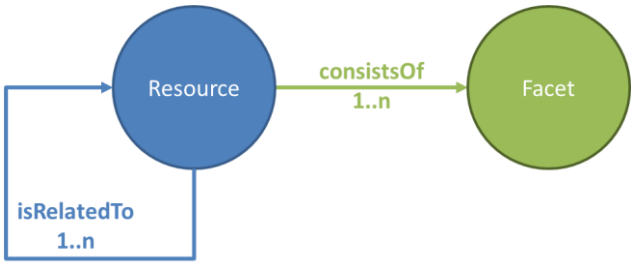

Fig. 5: Conceptual graph model of defined entities and relations

A Relation has a direction, i.e. a "source" and a "target". However, relations can be navigated in both directions when inspecting the graph (e.g. at query time), i.e. from source to target and from target to source. Facet describes a characteristic of a Resource. For such a reason, it is not permitted to define a Relation having a Facet as "source". In other words: it is neither permitted to define a Relation connecting a Facet with another one nor a Relation connecting a Facet with a Resource (as target) (see Fig. 5)

Each Entity and Relation has a header that is automatically generated for the sake of identification and provenance of the specific information and can be specialised with custom properties. The header is composed by the following properties: *uuid* i.e., used to unequivocally identify the Entity or the Relation; *creator* i.e., the actor has created the Entity or Relation; *modifiedBy* i.e., the actor made the last update to the Entity or Relation; *creationTime* i.e., creation time instant in milliseconds; *lastUpdateTime* i.e., last update time instant in milliseconds. Facet and Relation instances can have an arbitrary number of properties.

3.3.2 The Resource Registry

The Resource Registry is the core subsystem of the Information System as it connects producers and consumers of resources. It acts as a registry of the infrastructure by offering global and partial views of its resources, their current status and notification tools. The Resource Registry supports the Resource Manager system at offering dynamic allocation of resources to the VRE.

The Resource Registry offers Java and REST APIs and uses OrientDB as persistence layer (see Fig. 3). OrientDB is a graph database (but it can be used also as document store and key-value store) supporting Apache TinkerPop™ standard. Apache TinkerPop™ is a graph-computing framework for both graph databases (OLTP) and graph analytic systems (OLAP). When a data system is TinkerPop-enabled, its users are able to model their domain as a graph and analyse that graph using the Gremlin graph traversal language[28].

The Resource Registry is a stateless web service, it can be replicated horizontally and provides scalability capability. It provides four REST service endpoints for:

1. *Context Management*: managing hierarchical Context. A Virtual Research Environment is a typical context managed by the Resource Registry;
2. *Schema Management*: registering and defining entities and relations schema. This choice allows easy extension and support modification to the resource model and this is a key factor for the sustainability of the service and the Cloud infrastructure.
3. *Entity Management*: managing entities and relations instances compliant with registered schema.
4. *Query and Access*: supporting discovery and access of instances of registered entities and schema of registered types.

Schema Management validates the registered type according the following inheritance and type compatibility rules. For entity schema definition, the following rules apply: (*i*) multiple inheritance is permitted; (*ii*) any new type can have only one ancestor between Resource and Facet. For relation schema definition, the following rules apply: (*i*) multiple inheritance is permitted; (*ii*) any new type can have only one ancestor between *isRelatedTo* and *consistsOf*; (*iii*) relations definition must indicate "source" and "target" entity classes;

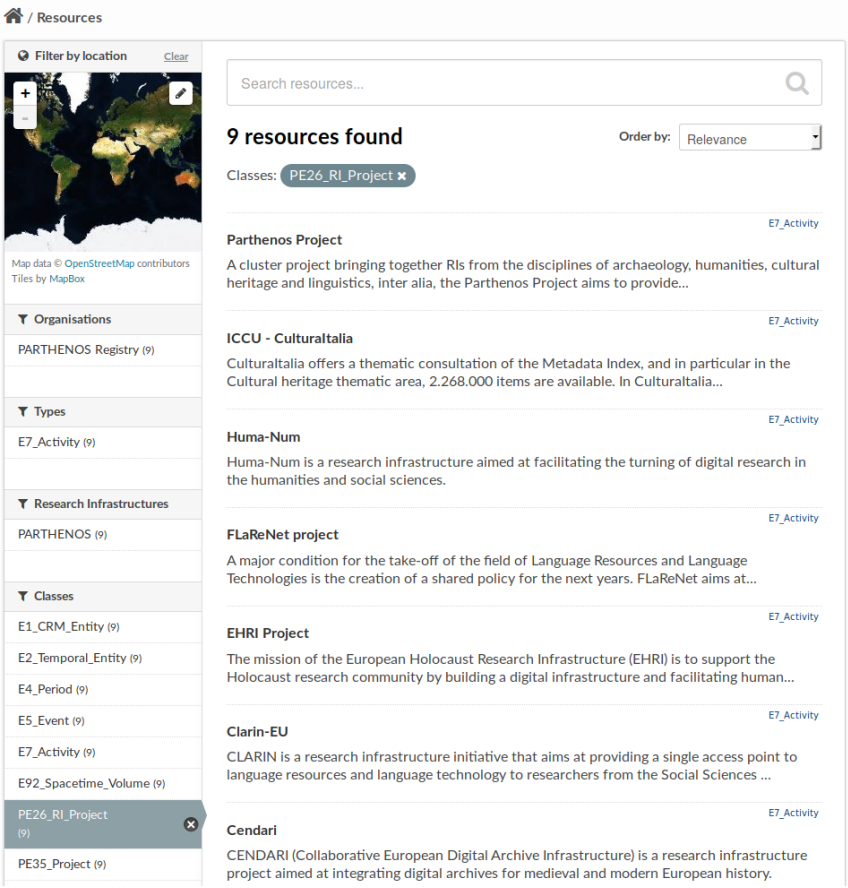

Fig. 6: Example of Resources Browsing and Filtering

[28] Gremlin graph traversal language: http://tinkerpop.apache.org/gremlin.html

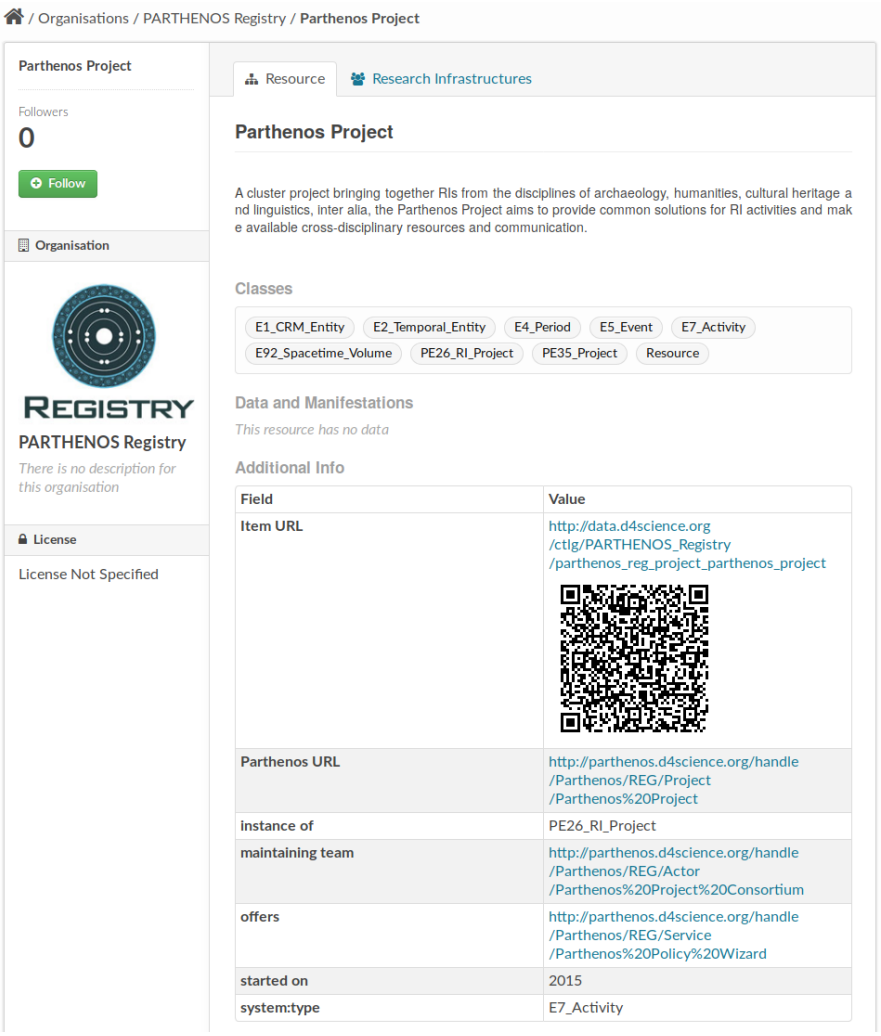

Fig. 7: Example of Resource View

(*iv*) any specialised relation must be between two entities having coherent source and target ancestors with respect to the one defined in the parent relation.

Entity Management validates instances according the following rules: (*i*) any entity and relation instance must belong to one of registered types; (*ii*) for a relation, the "source" and target entity instances must be an instance of (or an instance of a subclass) of the source and target entity types defined for the relation type.

The PARTHENOS Entities Model (PE Model) has been registered in the Resource Registry using the Schema Management REST service. The PE Model has been implemented as model of the Resource Registry by extending the IS Model with the following rationale[29]:

- Entities of the PE Model have been mapped as resources of IS Model;
- Properties of the PE Model have been mapped as *isRelatedTo* relations;
- A number of facets containing metadata describing the resources has been defined. Each facet groups related metadata (e.g. for identification, provenance, access rights information).

3.3.3 *The Resource Catalogue*

The Resource Catalogue service is based on CKAN[30], whose end-user web GUI has been configured to provide search, browse and filtering facilities using the concepts of PARTHENOS, e.g., Research Infrastructures, PARTHENOS Entity (PE)

[29] The complete definition of PE Model on IS Model can be found in Aloia et al. 2017.

[30] https://ckan.org/

types. Users of the Resource Catalogue can perform textual searches, browse and filter resources by selecting the interested PE class and/or the source research infrastructure (see Fig. 6).

When the user selects a specific resource, all details of the resource are displayed (see Fig. 7), such as name, UUID, description, URL on the catalogue and the PARTHENOS persistent URI (which is generated by the mapping applied by the PARTHENOS Aggregator and resolved by the URI Resolver to the URL on the catalogue i.e., the page the user is visiting). A QR code for each URL is proposed to allow acquisition with a smartphone camera. Users can also filter resources based on their geographic locations, by selecting a geographic area on a map. Resources can be geo-located by using a single point indicating latitude and longitude or through the use of a number of point which define a polygon. It is important to remark that the geo-location, if any, is provided by the original sources as resource metadata.

The Resource Catalogue is available by accessing a PARTHENOS dedicated VRE through the PARTHENOS Gateway[31], i.e., a D4Science infrastructure (Candela, Castelli, & Pagano, 2009; Candela et al., 2014) dedicated portal.

4. *Related Work*

Several initiatives in the past years intercepted the need of researchers of having a single-entry point to resources scattered across different and isolated data sources and to support them with tools for conducting their research in digital settings. In this section, we briefly present some of the main initiatives and highlight similarities and dissimilarities with respect to the approach of the PARTHENOS infrastructure.

Europeana (Purday, 2009) is a European initiative providing a single-entry point to more than 50 million digitised items – books, music, artworks and more – coming from thousands of European archives, libraries and museums willing to share cultural heritage for enjoyment, education and research. Items are discoverable via the Europeana portal, which offer advanced search and browsing facilities, and via the Europeana thematic collections on art, fashion, music, photography and World War I. Those thematic collections include galleries, blogs and exhibitions to inform and inspire. In order to support such advanced end-user functionality, Europeana feature a rich data model (Doerr et al., 2010) whose goal is to describe digital and digitised resources with enough details for offering advanced search and browse facilities.

Although also PARTHENOS realises and maintains an aggregation of metadata descriptions, the PARTHENOS approach differs from Europeana's mainly for its final goal: PARTHENOS aims at identifying the resources, the actors that maintain them, and the knowledge generation processes that have been applied to generate them. Because of that, the model does not require extensive descriptive metadata information, but rather focuses on the provenance of data and the provenance relationships with other entities. In addition, PARTHENOS also offers a framework where DHIs can integrate and share their existing services, while Europeana focuses only on metadata search and discovery.

Approaches similar to the one proposed in PARTHENOS has been adopted in completely different research fields by the agINFRA initiative and the LifeWatch Greece Directory.

agINFRA[32] is a European hub providing access to data for research on agriculture, food and the environment (Drakos, Protonotarios, & Manouselis, 2015). Like PARTHENOS, agINFRA is interested in the actors that are offering research tools and facilities to researchers in their field. A catalogue of organisations, initiatives, services and facilities has been created and it is currently supported and maintained by the eROSA (e-infrastructure Roadmap for Open Science in Agriculture) project[33] and AGINFRA+ project[34].

The LifeWatch Greece Directory (Minadakis et al., 2016) is a registry that supports the discovery of resources, especially datasets, within the biodiversity domain. End-users can perform queries to obtain information about the location of a dataset and, possibly, its creators, curators and contact points. Similarly to PE Model, the schema of LifeWatch Greece Directory is based on CIDOC-CRM and its extensions (CRMdig, CRMsci (Doerr, Kritsotaki, Rousakis, Hiebel, & Theodoridou, 2014)).

Other initiatives share the intents of PARTHENOS but focuses on the availability and accessibility of specific types of resources like Linghub (John P McCrae & Cimiano, 2015) for

[31] https://parthenos.d4science.org/
[32] http://www.aginfra.eu/
[33] http://www.erosa.aginfra.eu/
[34] http://www.plus.aginfra.eu/

language resources, the Open Metadata Registry[35] for metadata schemas, controlled vocabularies and application profiles and TERESAH (Tools E-Registry for E-Social science, Arts and Humanities)[36] for software tools for Social Science and Humanities studies.

Linghub[37] (John P McCrae & Cimiano, 2015; John Philip McCrae et al., 2015) is an initiative launched within the EC FP7 project LIDER[38] for the aggregation of metadata information about language resources. Collected metadata is harmonised by applying mappings to standard RDF vocabularies in order to enable queries and faceted search in a uniform environment. Harmonisation focuses on properties relevant for fostering the re-use of data: the resource URLs availability and accessibility, the terms and conditions of re-use and basic metadata information to support browse and discovery by type, language and intended usage of the resource.

The Open Metadata Registry[39] provides a means to identify, declare and publish through registration metadata schemas (element/ property sets), schemes (controlled vocabularies) and application profiles. The registry extends the open-source Dublin Core Metadata Initiative (DCMI) Registry[40] (Weibel & Koch, 2000) to support: (i) the automated creation and maintenance of schemas and application profiles; (ii) the submission of schemas and schemes to a registry workflow for review and publication. All of the development work leverages the latest knowledge and standards for networked knowledge organisation systems, schema and application profile declaration, and registry development. The Open Metadata Registry project was funded by the National Science Foundation for its first three years. It is currently managed by Metadata Management Associates, a consulting partnership committed to maintain the Registry as an open system.

TERESAH (Tools E-Registry for E-Social science, Arts and Humanities)[41] is a cross-community tools knowledge registry aimed at researchers in the Social Sciences and Humanities. It provides an authoritative listing of software tools, services, methodologies, and standards relevant in those domains. The registry features search and browse facilities, API for metadata access and import/export functions in RDF and JSON formats. The metadata format used to describe the tools is the Dublin Core. TERESAH has been developed as part of the Data Service Infrastructure for the Social Sciences and Humanities (DASISH)[42] European project. DASISH collaborates with the five European Strategy Forum on Research Infrastructures (ESFRI)[43] Infrastructures in the field of Social Science and Humanities (CESSDA[44], CLARIN, DARIAH, ESS[45], and SHARE[46]).

5. *Conclusion*

We presented a complete framework enabling the federation of DHIs and supporting the creation and operation of Virtual Research Environments (VREs) where researchers with different backgrounds can collaborate on specific research topics, sharing digital tools and data.

The Content Cloud Framework creates a homogeneous information space where metadata about resources (data, services and tools) of different DHIs are described according to a common data model. The aggregated metadata are made accessible to humans and machines via different service endpoints and exchange protocols: a Solr index (for search and browse), an OAI-PMH publisher (for bulk download), a Virtuoso server (for SPARQL queries). They are also registered into the Joint Resource Registry (JRR), which offers the required support for the discovery of federated resources. In particular, the presented technical framework enables the creation on-demand of Virtual Research Environments, which can be configured to provide a view of services, tools, and data registered in the JRR tailored to the research topic researchers want to collaborate on. The VRE becomes a collaboration environment where researchers can find resources relevant for a research topic, run services, and share the computational results.

The framework has been validated in the context of the PARTHENOS project with the collaboration of the research infrastructures and partners of the consortium.

35 http://metadataregistry.org/
36 http://teresah.dariah.eu/about
37 http://linghub.lider-project.eu/
38 http://www.lider-project.eu/
39 http://metadataregistry.org/
40 http://dublincore.org/groups/registry/
41 http://teresah.dariah.eu/about
42 http://dasish.eu/
43 http://www.esfri.eu/
44 https://www.cessda.eu/
45 http://www.europeansocialsurvey.org/
46 http://www.share-project.org/

The framework has been deployed in the D4Science infrastructure (Candela et al., 2009, 2014) and, as of May 2018, the Content Cloud Framework aggregated more than 200,000 (and growing) metadata records from CLARIN, Huma-Num, Metashare, LRE Map, CulturaItalia, Ariadne, DARIAH-GR, and EHRI. When stored in Virtuoso the aggregated metadata results into about 100 million triples. Three VREs have been created and configured to support researchers with tools and services for named entity recognition, natural language processing, and the integration of ontologies and thesauri.

The project will end in April 2019 and in the last year of the project we expect to complete the aggregation for a total of more than two million metadata records, integrate additional services and configure additional VREs where researchers can perform their research activities and share their results. Those will in fact further grow the PARTHENOS information space, which will then become a "live" knowledge base.

In order to increase the visibility of the tools, services and data, PARTHENOS is collaborating with OpenAIRE[47] to openly publish those products, making them discoverable by different stakeholders of the scholarly communication and the research community at large.

6. *Acknowledgements*

A special thanks to Nicola Aloia for his precious contributions.

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the PARTHENOS project (Grant agreement No. 654119).

[47] OpenAIRE: https://www.openaire.eu/

REFERENCES

Aloia, N., Candela, L., Debole, F., Frosini, L., Lorenzini, M., & Pagano, P. (2017). *Report on the Design of the Joint Resource Registry*. Retrieved from http://www.parthenos-project.eu/Download/Deliverables/D5.2_Report_on_design_Joint_Resource_Registry.pdf

Aloia, N., Debole, F., Felicetti, A., Galluccio, I., & Theodoridou, M. (2017). Mapping the ARIADNE Catalog Data Model to CIDOC CRM: Bridging Resource Discovery and Item-Level Access. *SCIRES-IT - SCIentific RESearch and Information Technology*, *7*(1), 1–8. https://doi.org/10.2423/I22394303V7N1P1

Antoniou, G., Christophides, V., Plexousakis, D., & Doerr, M. (2005). Semantic web fundamentals. In *Encyclopedia of Information Science and Technology, First Edition* (pp. 2464–2468). IGI Global.

Artini, M., Atzori, C., Bardi, A., La Bruzzo, S., Manghi, P., Mikulicic, M., & Zoppi, F. (2014). The {Heritage of the People's Europe} Project: An Aggregative Data Infrastructure for Cultural Heritage. In T. Catarci, N. Ferro, & A. Poggi (Eds.), *Bridging Between Cultural Heritage Institutions* (Vol. 385, pp. 77–80). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-54347-0_9

Atzori, C., Bardi, A., & Manghi Paolo, M. A. (2017). The OpenAIRE workflows for data management. In *Digital Libraries and Archives* (Vol. 733, pp. 95–107). Springer. Retrieved from https://link.springer.com/chapter/10.1007/978-3-319-68130-6_8

Bardi, A., & Frosini, L. (2017). Building a Federation of Digital Humanities Infrastructures. *{ERCIM} News*, (111), 28–29. Retrieved from https://ercim-news.ercim.eu/en111/special/building-a-federation-of-digital-humanities-infrastructures

Bardi, A., Manghi, P., & Zoppi, F. (2014). Coping with Interoperability and Sustainability in Cultural Heritage Aggregative Data Infrastructures. *International Journal of Metadata, Semantics and Ontologies*, *9*(2), 138–154. https://doi.org/10.1504/IJMSO.2014.060341

Blanke, T., Candela, L., Hedges, M., Priddy, M., & Simeoni, F. (2010). Deploying general-purpose virtual research environments for humanities research. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *368*(1925), 3813–3828.

Blümm, M., & Schmunk, S. (2016). Digital Research Infrastructures: DARIAH. In *3D Research Challenges in Cultural Heritage II* (pp. 62–73). Springer.

Boukhelifa, N., Bryant, M., Bulatović, N., Čukić, I., Fekete, J.-D., Knežević, M., … Thiel, C. (2018). The CENDARI Infrastructure. *J. Comput. Cult. Herit.*, *11*(2), 8:1–8:20. https://doi.org/10.1145/3092906

Bruseker, G., Doerr, M., & Theodoridou, M. (2017). *Report on the Common Semantic Framework*. Retrieved from http://www.parthenos-project.eu/Download/Deliverables/D5.1_Common_Semantic_Framework_Appendices.pdf

Bryant, M., Reijnhoudt, L., Speck, R., Clerice, T., & Blanke, T. (2014). The EHRI project-virtual collections revisited. In *International Conference on Social Informatics* (pp. 294–303).

Calzolari, N., Del Gratta, R., Francopoulo, G., Mariani, J., Rubino, F., Russo, I., & Soria, C. (2012). The LRE Map. Harmonising Community Descriptions of Resources. In *8th International Conference on Language Resources and Evaluation (LREC 2012)* (pp. 1084–1089).

Candela, L., Castelli, D., & Pagano, P. (2009). D4Science: an e-Infrastructure for Supporting Virtual Research Environments. In *IRCDL* (pp. 166–169).

Candela, L., Castelli, D., & Pagano, P. (2012). Managing Big Data through Hybrid Data Infrastructures. *{ERCIM} News*, (89). Retrieved from http://ercim-news.ercim.eu/en89/special/managing-big-data-through-hybrid-data-infrastructures

Candela, L., Castelli, D., & Pagano, P. (2013). Virtual research environments: an overview and a research agenda. *Data Science Journal*, *12*, GRDI75--GRDI81. https://doi.org/10.2481/dsj.GRDI-013

Candela, L., & Pagano, P. (2015). Cross-disciplinary Data Sharing and Reuse via gCube. *{ERCIM} News*, (100). Retrieved from http://ercim-news.ercim.eu/en100/special/cross-disciplinary-data-sharing-and-reuse-via-gcube

Candela, L., Pagano, P., Castelli, D., & Manzi, A. (2014). Realising Virtual Research Environments by Hybrid Data Infrastructures: the D4Science Experience. In *Proceedings of the symposium" International Symposium on Grids and Clouds (ISGC) 2014"(ISGC2014). 23-28 March, 2014. Academia Sinica, Taipei, Taiwan*. Retrieved from http://pos.sissa.it/cgi-bin/reader/conf.cgi?confid=210,id.22

Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C., & van de Sompel, H. (2010). The europeana data model (edm). In *World Library and Information Congress: 76th IFLA general conference and assembly* (pp. 10–15).

Doerr, M., Kritsotaki, A., Rousakis, Y., Hiebel, G., & Theodoridou, M. (2014). *CRMsci: the Scientific Observation Model*.

Doerr, M., & Theodoridou, M. (2014). *CRMdig an extension of CIDOC-CRM to support provenance metadata*.

Drakos, A., Protonotarios, V., & Manouselis, N. (2015). agINFRA: a research data hub for agriculture, food and the environment. *F1000Research*, *4*.

Frosini, L., & Pagano, P. (2018). A facet-based open and extensible resource model for research data infrastructures. *Grey Journal*, *2017–Octob*(2), 69–76.

Gartner, R., & Hedges, M. (2013). CENDARI: Establishing a digital ecosystem for historical research. In *2013 7th IEEE International Conference on Digital Ecosystems and Technologies (DEST)* (pp. 61–65). IEEE. https://doi.org/10.1109/DEST.2013.6611330

Gavrilidou, M., Labropoulou, P., Desipri, E., Piperidis, S., Papageorgiou, H., Monachini, M., … others. (2012). The META-SHARE Metadata Schema for the Description of Language Resources. In *8th International Conference on Language Resources and Evaluation (LREC 2012)* (pp. 1090–1097).

interoperability Technical Committee, T. (2014). *ISO 21127:2014 - A reference ontology for the interchange of cultural heritage information*. Geneva, CH.

management, D., & interchange Technical Committee. (1999). *ISO 11179-1:1999 Specification and standardization of data elements*. Geneva, CH.

Manghi, P., Artini, M., Atzori, C., Bardi, A., Mannocci, A., La Bruzzo, S., … Pagano, P. (2014). The D-NET software toolkit: A framework for the realization, maintenance, and operation of aggregative infrastructures. *Program*, *48*(4), 322–354. https://doi.org/10.1108/PROG-08-2013-0045

Manghi, P., Manola, N., Horstmann, W., & Peters, D. (2010). An infrastructure for managing EC funded research output-The OpenAIRE Project. *The Grey Journal (TGJ): An International Journal on Grey Literature*, *6*(1).

Mannocci, A., Casarosa, V., Manghi, P., & Zoppi, F. (2015). The EAGLE Europeana network of Ancient Greek and Latin Epigraphy: a technical perspective. In *Italian Research Conference on Digital Libraries* (pp. 75–78).

Marketakis, Y., Minadakis, N., Kondylakis, H., Konsolaki, K., Samaritakis, G., Theodoridou, M., … Doerr, M. (2017). X3ML mapping framework for information integration in cultural heritage and beyond. *International Journal on Digital Libraries*, *18*(4), 301–319. https://doi.org/10.1007/s00799-016-0179-1

McCrae, J. P., & Cimiano, P. (2015). Linghub: a Linked Data based portal supporting the discovery of language resources. In *11th International Conference on Semantic Systems* (Vol. 1481, pp. 88–91).

McCrae, J. P., Cimiano, P., Rodriguez-Doncel, V., Suero, D. V., Gracia, J., Matteis, L., … Buitelaar, P. (2015). Reconciling heterogeneous descriptions of language resources. In *4th Workshop on Linked Data in Linguistics: Resources and Applications* (pp. 39–48).

Meghini, C., Scopigno, R., Richards, J., Wright, H., Geser, G., Cuy, S., … Vlachidis, A. (2017). ARIADNE: A Research Infrastructure for Archaeology. *J. Comput. Cult. Herit.*, *10*(3), 18:1--18:27. https://doi.org/10.1145/3064527

Merlitti, D., Procaccia, M., Parrini, U. , & Masci, M. E. (2012). La digital library "SHOAH" dell'Archivio Centrale dello Stato: Un progetto di recupero e di digital preservation di documenti audiovisivi di storia orale. *SCIRES-IT - SCIentific RESearch and Information Technology*, 2(2), 1–16. https://dx.doi.org/10.2423/i22394303v2n2p1

Minadakis, N., Marketakis, Y., Doerr, M., Bekiari, C., Papadakos, P., Gougousis, A., … Arvanitidis, C. (2016). LifeWatch Greece data-services: Discovering Biodiversity Data using Semantic Web Technologies. *Biodiversity Data Journal*, *4*. https://doi.org/10.3897/BDJ.4.e8443

Purday, J. (2009). Think culture: Europeana.eu from concept to construction. *The Electronic Library*, *27*(6), 919–937. https://doi.org/10.1108/02640470911004039

Soria, C., Bel, N., Choukri, K., Mariani, J., Monachini, M., Odijk, J., … Calzolari, N. (2012). The flarenet strategic language resource agenda. In *8th International Conference on Language Resources and Evaluation (LREC 2012)* (pp. 1379–1386).

Van Uytvanck, D., Stehouwer, H., & Lampen, L. (2012). Semantic metadata mapping in practice: the Virtual Language Observatory. In *8th International Conference on Language Resources and Evaluation (LREC 2012)* (pp. 1029–1034).

Váradi, T., Wittenburg, P., Krauwer, S., Wynne, M., & Koskenniemi, K. (2008). CLARIN: Common language resources and technology infrastructure. In *6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Weibel, S. L., & Koch, T. (2000). The Dublin core metadata initiative. *D-Lib Magazine*, *6*(12), 1082–9873.