Collaborative Research on Academic History using Linked Open Data: A Proposal for the Heloise Common Research Model

Francesco Beretta, Thomas Riechert

Collaborative Research on Academic History using Linked Open Data: A Proposal for the Heloise Common Research Model

Investigación colaborativa sobre la historia académica a través de los datos abiertos enlazados: una propuesta para el modelo de investigación común de Héloïse

Francesco Beretta*
Laboratoire de recherche historique Rhône-Alpes (UMR 5190 LARHRA), CNRS/Université de Lyon, Lyon, France

Thomas Riechert
Hochschule für Technik, Wirtschaft und Kultur (HTWK), University of Applied Sciences, Leipzig , Germany

DOI: http://dx.doi.org/10.20318/cian.2016.3147

Recibido: 05- 05-2016
Aceptado: 23-05-2016

Abstract: The paper presents a proposal for the Heloise Common Research Model (HCRM), to be implemented for the European research network on digital academic history – Heloise. The objective of Heloise is to interlink databases and other digital resources stemming from several research projects in the field of academic history, to provide an integrated database for federated research on the network databases. The HCRM defines three layers: the Repository Layer, the Application Layer and the Research Interface Layer, which are presented in detail. As part of

Resumen: El artículo presenta una propuesta para el modelo de investigación común de Héloïse (HCRM en sus siglas en inglés) para su implementación por la red de investigación europea sobre la historia académica digital – Héloïse. El objetivo de Héloïse es interconectar las bases de datos y otros recursos digitales que pertenecen a varios proyectos de investigación en el campo de la historia académica con el fin de ofrecer una base de datos integrada para la investigación federada sobre las bases de datos de la Red. El HCRM define tres niveles: el nivel del re-

* francesco.beretta@ish-lyon.cnrs.fr – thomas.riechert@htwk-leipzig.de. All websites were accessed on 29 April 2016.

the application and research interface layer, essential concepts are the symogih.org ontology and a Heloise network-specific thesaurus. The concepts have been tested on a sample of Heloise network's datasets as a part of a prototype of the envisaged platform that the authors have started implementing. The paper concludes with future developments to be accomplished within the Heloise network.

Keywords: academic history, domain ontologies, data interoperability, semantic web technologies, linked open data.

positorio, el nivel de la aplicación y el nivel de la interfaz de la investigación que se explican de forma detallada. Como parte del nivel de la interfaz de la investigación, la ontología symogih.org y un diccionario de sinónimos para la red Héloïse constituyen conceptos fundamentales. Los conceptos han sido probados sobre una muestra de datos de la red Héloïse como parte de un prototipo de la plataforma que los autores han empezado a desarrollar. El artículo concluye con propuestas de futuro desarrollo a realizar dentro del marco de la red Héloïse.

Palabras clave: historia académica, ontologías de dominio, interoperabilidad de datos, tecnologías de web semántica, datos abiertos enlazados

1. Motivation

Answering research questions using existing datasets that are available on the Linked Open Data web is of high relevance for future research in the humanities, particularly in the domain of historical research[1]. A constantly increasing amount of archive materials, bibliographic resources, research results and databases is available online. Through standardization efforts, e.g. the use of RDF (Resource Description Framework) and OWL (Web Ontology Language) as description languages for data resources, it is possible to link data sources and apply inference algorithms to use the information in a broader research context.

The authors are active in various research projects in France and Germany within the fields of computer science and history. F. Beretta is participating in the development of the symogih.org project[2], a method and platform in development since 2007 for collectively producing and storing historical data[3]. T. Riechert is working on emergent semantic web technologies at AKSW research group[4]. According to the authors' experience, the issue of providing standardization in the domain of historical research is a difficult

[1] Cf. Meroño Peñuela et al., "Semantic technologies for historical research: A survey". Semantic Web Journal (IOS Press) 6(2015): 539-564.

[2] Système Modulaire de Gestion de l'information Historique: http://symogih.org .

[3] Beretta, F. et al. "Modeling and sharing geo-historical information: the SyMoGIH project." 2012 and Beretta, F. et al. "The SyMoGIH project: publishing and sharing historical data on the semantic web." 2014.

[4] Agile Knowledge Engineering and Semantic Web: http://aksw.org .

one. This is primarily due to the high degree of domain-specific characteristics of existing data and to the significant role of project-specific research questions during manual data acquisition and processing.

Heloise, the European research network on digital academic history[5], was founded in 2012 to interlink databases and other digital resources stemming from several research projects in the field of academic history. Workshops are regularly held in different European countries allowing to present the content of the participating projects' research databases and discuss the issues of interlinking the existing datasets. The aim of the network is to provide a platform for aggregating the available data and publish them in a format that will enable researchers to use them for answering new and more comprehensive research questions. It is also important to connect these data with external resources provided by the National Libraries and other GLAM institutions[6].

The Heloise network faces the challenge described above, i.e. connecting data produced according to different data models and research agendas, and transcribing them into a format allowing to query them as a whole without losing the semantic richness of original data. Thus, the interlinking of participating projects' datasets is one of the network's main tasks and objectives. As a result of the preliminary works and exchanges within Heloise workshops from 2013 to 2015, the authors proposed the Heloise Common Research Model (HCRM) as a general methodology to cope with this issue. The HCRM is based on the authors' current research and is under discussion among the Heloise network's partners.

The implementation of HCRM will result in a publication platform to provide interfaces on all three HCRM layers for Heloise network's interlinked repositories. Section 2 of the paper describes the HCRM in detail. A prerequisite to data alignment and sharing in the field of academic history is to provide a domain-specific common vocabulary. Section 3 will provide essential concepts about the solution the authors have devised for developing such a vocabulary using the symogih.org ontology and a Heloise network-specific thesaurus. This solution has been tested on a sample of Heloise network's datasets as a part of a prototype of the platform envisaged[7], which the authors have started implementing by using some of the HCRM modules. Section 4 provides an architecture of the Heloise platform that focuses on using open

[5] European Network on Digital Academic History: http://heloisenetwork.eu/.

[6] Open Knowledge held by Galleries, Libraries, Archives and Museums: http://openglam.org/.

[7] Homepage: http://platform.heloisenetwork.eu.

source software and container virtualization to provide basic infrastructure for supporting cloud technologies. In Section 5, the authors summarize future developments to be accomplished within the Heloise network.

2. Heloise Common Research Model

The Heloise Common Research Model (HCRM) was presented to the Heloise consortium in Madrid in 2015. HCRM presents a layer-oriented concept of a methodology that will be jointly refined by the partners and used to share the projects' datasets to build a common research platform (Figure 1). Layer models are used in different fields of science for modeling the structure of complex and distributed systems. In the field of computer science, the 3-tier architecture is a common structural concept based on layers[8]. Layer models enable the development of independent technical solutions within different abstraction layers. A key feature of a layer is that modules of a layer can provide interfaces that can be used by modules of the overlying layer. These modules are also able to use the interfaces of modules of the immediate underlying layer. The HCRM defines three layers: the Repository Layer, the Application Layer, and the Research Interface Layer, which are presented in detail below. Figure 1 depicts the three layers and the modules they contain. Modules represent either datasets and ontologies or related activities and processes for ontology evolution. All modules depicted within the diagram are proposed for the emerging Heloise network's platform. They can be extended or replaced in the future according to the researchers' and projects' needs. In addition, Figure 1 depicts the Heloise Network Platform and provides applications that are needed to implement modules within the layers (cf. Section 4).

2.1. HCRM Repository Layer

The Repository Layer provides domain-specific databases, partly interlinked. These databases are provided by research repositories of the Heloise network partners and from external institutions (GND, BNF, DBpedia, etc.). So far, the available Heloise datasets are available as relational databases or as documents in XML format. Some of the databases, e.g. Catalogus Professorum

[8] Cf. Sommerville, I. *Software Engineering*. Harlow, England: Addison-Wesley, 2010.

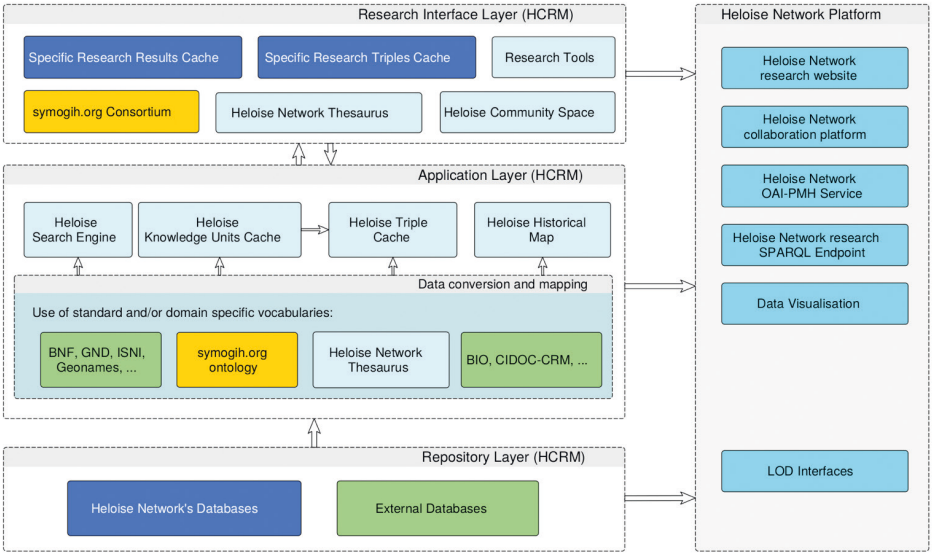

Fig. 1. Heloise Common Research Model - Overview.

Lipsiensium[9] and the symogih.org project[10], provide Linked Open Data (LOD) in RDF format. An overview of all available datasets within the network is available online[11]. External resources are used to enrich the data produced by the participating research projects with information on publications, geolocations, or general lexicographical knowledge provided by library catalogs, community projects like Open Street Map[12] and Wikipedia[13], as well from external research projects. An overview of such external resources is provided on Datahub.io[14], which is also used to generate the LOD cloud diagram[15].

[9] Catalogus Professorum Lipsiensium: http://catalogus-professorum.org.

[10] Cf. http://symogih.org/?q=rdf-publication.

[11] Heloise partners and online available datasets: http://heloisenetwork.eu/partners.

[12] OSM RDF-Representation: http://linkedgeodata.org. Stadler, C.; Lehmann, J.; Höffner, K. and Auer, S. "LinkedGeoData: A Core for a Web of Spatial Open Data." *Semantic Web Journal* 3 (2012): 333-354.

[13] Wikipedia RDF-Representation: http://dbpedia.org. Lehmann, Jens, Bizer, Chris, Kobilarov, Georgi, Auer, Sören, Becker, Christian, Cyganiak, Richard and Hellmann, Sebastian. "DBpedia - A Crystallization Point for the Web of Data." *Journal of Web Semantics* 7 (2009): 154-165.

[14] Open Knowledge Foundation Datahub: https://datahub.io/dataset.

[15] Cf. Cyganiak, Richard and Jentzsch, Anja. "Linking open data cloud diagram." *LOD Community (http://lod-cloud. net/)* 12(2011).

Processes that can be realized within the repository layer are:

- Publishing Linked Open Data
- Interlinking datasets

LOD can be published using established tools such as those provided within the LOD2 technology stack[16]. The platform will provide applications like OntoWiki[17] for collaboratively collecting data and D2RQ[18] to publish relational databases as LOD. The linking has so far been done within established authoring processes in the database, but may be supported by link discovery algorithms e.g. Silk[19] and Limes[20].

2.2 HCRM Application Layer

Modules within the Application Layer provide common access to data sources described in Section 2.1 by using standardized vocabularies. The access is typically restricted to the expressiveness of the individual vocabulary.

Processes and modules that have to be realized within the application layer are:

- Alignment of specific vocabularies of research databases to standardized vocabularies and upper-level ontologies
- Application of tools for accessing standardized vocabularies
- Caching triples to access federated databases according to the vocabularies

First applications for the Heloise consortium have currently been developed. Among them, the implementing of a search engine for persons on all databases. For these aspects, the BIO vocabulary[21] is used which was de-

[16] LOD 2 Technology Stack: http://stack.lod2.eu .

[17] Frischmuth, P. et al. "OntoWiki - An Authoring, Publication and Visualization Interface for the Data Web". *Semantic Web Journal* (IOS Press) 6 (2015): 215-240.

[18] Bizer, C. and Cyganiak, R. "D2R Server – Publishing Relational Databases on the Semantic Web (Poster)" *Poster at the 5th International Semantic Web Conference*, Athens, USA, 2006.

[19] Volz J. et al., "A Link Discovery Framework for the Web of Data". Madrid 2009.

[20] Ngonga n. et al. "LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data". 2011.

[21] A vocabulary for biographical information: http://vocab.org/bio/0.1/.html.

veloped by Ian Davis, in particular, for displaying biographical information. The necessary vocabulary alignment is carried out using logic rules regarding the relevant classes and properties of the respective ontologies. Also, the consortium will develop a domain-specific vocabulary for academic history resources. The adoption of the symogih.org ontology for constructing such a vocabulary is currently being tested (cf. Section 3). Furthermore, a study group is working on aligning geographical resources with commonly used gazetteers like geonames.org.

2.3. HCRM Research Interface Layer

The Research Interface Layer supports access to the interlinked databases for researchers. This is possible by using the upper-level research vocabularies as provided e.g. by the symogih.org ontology. Furthermore specific research activities are involved including organizational and community activities like the production and administration of a Heloise Network Thesaurus, or the participation in the development of the symogih.org vocabulary.

The process modules that have to be realized within the Research Interface Layer are:

- Collaborative development and evolution of an upper-level research vocabulary that interlinks knowledge across multiple projects within the field of academic history
- Providing an infrastructure for processing research queries, storing results, as well as developing further research tools e.g. to proof research results.

3. Heloise Network Vocabularies

Research data distributed in the information systems of participating projects are not only stored in different formats (relational databases, XML, RDF), a technological issue that can easily be solved, but they are also structured according to the project-specific research agendas and expressed using different project internal semantic categories. This happens in spite of the fact that data produced in this research domain often aim at expressing the same contents: geographical and social origins of university students and teachers, disciplines learned and taught, degrees obtained by them, publications, etc. Another issue

collaborative projects have to cope with is the fact that some information systems provide data in the form of structured data, i.e. all instances in the database (actors, places, universities, disciplines, etc.) are identified and described, but many datasets are available in the form of semi-structured data: properties are stored in fields containing strings (sequences of characters) providing the label of an instance (in the form of a named entity) but without identifying the resource (the institution, the concept, etc.) intended by the researcher.

3.1. A Collaborative Heloise Network Thesaurus

Therefore, two different tasks are to be realized. On the one hand, the HCRM needs a domain-specific vocabulary to provide identification for the resources the researchers are interested in and collecting information about (actors, institutions, places, concepts, etc.). This vocabulary will provide some functionalities like the handling of taxonomies and the interlinking with external well-known identifiers of resources (GND[22], BNF[23], VIAF, etc.) that will allow identification and organization of the resources that are relevant in the academic history research domain and provide the identifiers in the form of URIs (Uniform Resource Identifiers) for interlinking the projects' resources among them and providing alignment with external reference datasets. The semi-structured data in present databases will be progressively transformed into structured data in order to allow advanced domain-specific queries and inference, possibly supported by an algorithm comparing the quality of available data. This task will be realized by creating a collaborative Heloise Network Thesaurus which will be instantiated and discussed by all interested participating scholars.

3.2. The Choice of an Ontology for Historical Data Modeling

On the other hand, the HCRM needs to define a common ontology for expressing the semantic richness of the projects' databases and information systems in a uniform and flexible way. One approach to cope with this issue can be to adopt an existing vocabulary, e.g. the CIDOC-CRM vocabulary devoted to the domain of cultural heritage or the BIO vocabulary for biographical information,

[22] Integrated Authority File, managed by the German National Library: http://www.dnb.de/EN/Standardisierung/GND/gnd_node.html.

[23] http://www.bnf.fr/fr/professionnels/autorites_bnf/s.autorite_bnf_presentation_statistiques.html.

and extract all semantic contents that can be expressed using these vocabularies from the participating projects' information systems. These vocabularies were not developed to model the richness of information which is typically provided by historical research. This process can be useful in the first step of data exploration, but will inevitably lead to a simplification and flattening of the semantic contents of the data provided by the projects. Furthermore, this will cause information loss and reduce the value of data for historical research.

These vocabularies have not been conceived to express the issues of sourcing, dating, handling imprecision and contradictory sources that should be provided to qualify and compare the available data. Moreover, they have only modeled a small portion of the contents that are relevant to academic history and a large amount of participating projects' information present in data could not be expressed. Of course, one could extend these vocabularies by adding new classes and properties or add classes from other ontologies, to express concepts like dating, sourcing, etc. However, this would mean producing a patchwork of classes and properties conceived in different vocabularies and missing the methodological coherence and ontological commitment which is an important element of scientific research in the domain of humanities and social sciences.

Therefore, the authors propose to adopt the symogih.org ontology, an upper-level flexible vocabulary developed with the aim of collaboratively modeling historical knowledge, for handling the issue of aligning the semantic structure of the Heloise network's datasets and for creating a common research repository. The symogih.org project, which is currently housing 16 ongoing projects including several on a European scale (France, Belgium, Italy, Germany and Switzerland) with about 50 active users representing different approaches (economic, political, religious, cultural, science and academic history), is evolving towards the creation of an international consortium for producing and sharing historical data[24]. Interested members of the Heloise network could participate in this process which mainly consists of collaborative modeling of historical data and building up open research data repositories.

3.3. The symogih.org Ontology

The symogih.org ontology was developed in 2007 as a method for modeling knowledge in the form of structured data produced by historians by applying

[24] Beretta, F. et al. "The symogih.org Project: Towards an International Consortium." 2016.

the critical method[25]. In a first step, generic classes for grouping objects are defined as subclasses of an *Object* abstract class: actors, collective actors, abstract objects (concepts), places, etc. (cf. Figure 2). Similar classes in standard ontologies, like CIDOC-CRM, are often equivalent classes or subclasses of the symogih.org objects' classes. They are defined with the highest level of abstraction and objectivity possible so that they can be used independently from the research agenda of the historians. Instances of these classes are created progressively by researchers providing a preferred name for an object, alternatives names, (approximate) dating and a definition. These instances can then be used by all participants in the symogih.org project by referring to the resource's URI. The Heloise Network Thesaurus mentioned above should be built using the symogih.org objects' classes and should support the collaborative definition and discussion of concepts like disciplines, degrees, time-related specificities of institutions, etc. in the domain of academic history.

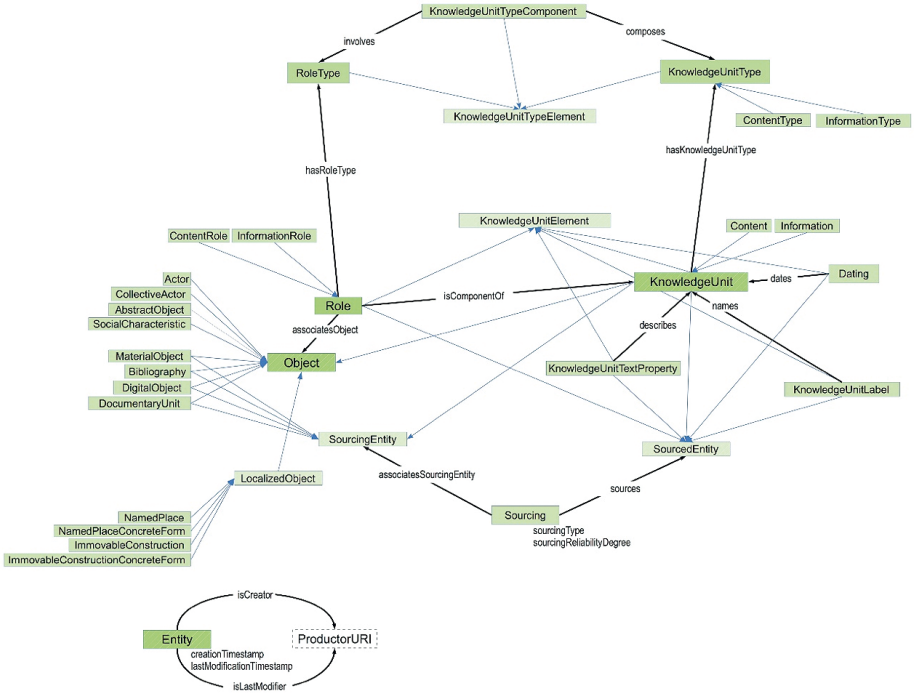

Fig. 2. The symogih.org ontology, 0.2.0 version, October 17, 2015.

[25] Beretta, F. and Vernus, P. "Le projet SyMoGIH et la modélisation de l'information: une opération scientifique au service de l'histoire". *Les Carnets du LARHRA* 1(2012): 81-107.

In a second step, the symogih.org ontology provides a method for modeling objects' properties and relationships among them. These are modeled in the form of a *KnowledgeUnit* class defined by a type, a situation in a time span, optionally by a label, and by providing one or more information sources. The principle of atomization expressed by the concept of 'knowledge unit' (i.e. split wherever possible knowledge into simple propositions) ensures reusability of data for new research agendas. Knowledge units are produced by applying the historical method: sourcing, criticism, and deduction are insured. Therefore, knowledge units can be thought of as assertions stated by the historian about object's properties and relationships among objects. The participation of an object in a knowledge unit is expressed by a role, specified by a role type that indicates what kind of participation the object has: for instance, in the case of a letter, an actor can be the sender or the receiver, which is expressed by his role.

Knowledge units are divided into two subclasses: An instance of the *Information* class expresses knowledge as it is constructed by the researcher using historical criticism and extracting it from different sources, whereas a instance of the *Content* class reproduces knowledge as it was meant by one and only one source, even if we know the source is wrong about an event date or circumstances. In both cases, the sourcing class provides the origin of each knowledge unit. This distinction is very important in academic history because some projects have a source-driven approach, producing data exactly as they appear in sources, and other projects compare different sources to find out, with a degree of probability defined by historical criticism, how reality was. The status of knowledge is not the same in both cases and one should be aware of that in comparing different projects' datasets.

Specific types of knowledge units and roles appear as instances of the *KnowledgeUnitType* and *RoleType* classes: they are progressively and collaboratively created by historians insofar as they need to treat new kinds of topics. This means that the ontology is versatile, can be gradually expanded and virtually handles any type of knowledge. Instances of the *KnowledgeUnitType* and *RoleType* classes make explicit and document the meaning of the structured data produced using them. They are therefore published on the symogih.org project website[26].

[26] http://symogih.org/?q=type-of-knowledge-unit-classes-tree.

3.4. Upper-level Ontologies and Data Interoperability

The symogih.org generic data model has an ontological structure which is equivalent to the DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering) upper-level ontology[27] and to the upper classes of CIDOC-CRM[28]: They model human discourse about objects' properties, and present and past events. The symogih.org *Object* class is equivalent to the *Endurant* class in DOLCE and *Persistent Item* class in CIDOC-CRM; the *KnowledgeUnit* class is equivalent respectively to the *Perdurant* and *Temporal Entity* class. Therefore, the symogih.org ontology, like CIDOC-CRM, is an instantiated upper-level ontology in the sense that "its specific domain concepts instantiate broader top-level notions"[29]. The symogih.org ontology has no hierarchical structure like CIDOC-CRM but a flat one: All knowledge units types are on the same level, which increases the ontology flexibility and the capacity of being collaboratively enlarged by the user's community.

This ontological structure easily allows for rewriting and expressing any kind of information. For instance, the birth of the astronomer Johannes Kepler is expressed by the DBpedia ontology in the following way:

```
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX dbpedia-owl: <http://live.dbpedia.org/ontology/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

dbpedia:Johannes_Kepler dbpedia-owl:birthDate "1571-12-27"^^xsd:date ;
                        dbpedia-owl:birthPlace dbpedia:Weil_der_Stadt .
```

The same information can be rewritten using the symogih.org ontology:

```
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX sym: <http://symogih.org/ontology/>
PREFIX syr: <http://symogih.org/resource/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
```

[27] http://cordis.europa.eu/result/rcn/41438_en.html – http://en.wikipedia.org/wiki/Upper_ontology#DOLCE_and_DnS.

[28] www.cidoc-crm.org/official_release_cidoc.html. CIDOC-CRM is since 2006 a ISO norm: ISO21127.

[29] Domingue, J., Fensel, D. and Hendler, J. A. (eds.). *Handbook of semantic web technologies* (Berlin, Heidelberg: Springer-Verlag, 2011): 523.

```
_:info_1 rdf:type sym:Information ;
          sym:hasKnowledgeUnitType syr:TyIn14 . # birth
          sym:hasCreator viaf:14907585 . # Francesco Beretta
          sym:hasCreationTimestamp "2015-02-02T14:55:33"^^
             xsd:dateTime.

_:d1 rdf:type sym:Dating ;
          sym:dates _:info_1 ;
          sym:dateTime "1571-12-27"^^xsd:date ;
          sym:datingType syr:AbOb246 ; # unique date
          sym:datingCertainty 3 . # certain
          sym:hasCreator [...]

_:s1 rdf:type sym:Sourcing ;
          sym:sources _:info_1 ;
          sym:associatesSourcingEntity <http://dbpedia.org/page/
             Johannes_Kepler> ;
          sym:sourcingType 3 ; # literal
          sym:sourcingReliabilityDegree 1 . # uncertain
          sym:hasCreator [...]

_:r1 rdf:type sym:Role ;
          sym:isComponentOf _:info_1 ;
          sym:associatesObject dbpedia:Johannes_Kepler ;
          sym:hasRoleType syr:TyRo40 ; # being born
          sym:associatedObjectIdentificationCertainty 3 ; # certain
          sym:hasCreator [...]

_:r2 rdf:type sym:Role ;
          sym:isComponentOf _:info_1 ;
          sym:associatesObject dbpedia:Weil_der_Stadt ;
          sym:hasRoleType syr:TyRo8 ; # localize
          sym:associatedObjectIdentificationCertainty 3 ; # certain
          sym:hasCreator [...]
```

This rewritten knowledge unit not only provides the information itself, but also the metadata the historian needs to evaluate its reliability for historical research. Given the upper-level structure of the symogih.org ontology, any information can be rewritten and provided with metadata. If the appropriate instance of the *KnowledgeUnitType* class is missing, it can be created and its meaning publicly documented. The symogih.org ontology is, therefore, suitable for providing interoperability among structured and semi-structured datasets that were produced according to different data

models[30]. In addition, the PROV ontology[31] can be used to represent provenance from the different projects' datasets. A test of this rewriting process was realized with data provided by four different Heloise network's projects and the results are available on the symogih.org public SPARQL-endpoint[32] and in tabular form on a web page publishing a portion of the RDF data[33].

4. Heloise Research and Publication Platform

This section presents basic technologies to implement modules on the different HCRM layers for the Heloise network, and will present some interfaces, that are available at the Heloise Network Platform. A particular challenge is to manage the heterogeneity of the available software application. Accordingly, to the linked data lifecycle[34] these applications implements extraction, storage, authoring, interlinking, enrichment, quality analysis, evolution and exploration of linked data sources. Unlike typical integration scenarios where infrastructure is centralized, Heloise applications will be run distributed on infrastructure at all network partners. Although in the initial phase of implementation, modules and related services will be offered centrally for all partners.

The technical objective is to implement an infrastructure that provides individual modules in a virtualized environment. This will allow to duplicate and distribute modules and applications in a productive state on nodes within the web. As part of the prototypical implementation of the first modules, a container-based operating system virtualization will be realized by Docker. The Docker project was started in 2013 and experienced a rapid increase in its popularity[35]. Container-based virtualization provides an isolated runtime environment for multiple individual applications sharing the complete

[30] Cf. Beretta, F. "Publishing and sharing historical data on the semantic web: the *symogih.org* project", 2015 and Beretta, F. "L'interopérabilité des données historiques et la question du modèle: l'ontologie du projet SyMoGIH". In *Quels enjeux numériques pour les médiations scientifique et culturelle*, Jean-Luc Minel (ed.) (Paris: Presses universitaires de Paris Ouest, 2016).

[31] https://www.w3.org/TR/prov-o/.

[32] Home page: http://symogih.org/graph/heloise.

[33] Cf. e.g. http://heloise.ish-lyon.cnrs.fr/actors/birth-origin-list.

[34] Cf. Auer, S. et al. "Introduction to Linked Data and Its Lifecycle on the Web" In *Proceedings of the 9th international conference on Reasoning Web: semantic technologies for intelligent data access* (RW'13), 2013.

[35] Cf. Merkel, D. "Docker: Lightweight linux containers for consistent development and deployment." Linux Journal, 239 (2014).

hardware and the operating system components of the host system. [36] The Dockerizing Linked Data[37] project provides an orchestration and integration platform for Linked Data applications. Triple stores i.e. OpenLink Virtuoso[38] and Fuseki SPARQL server for Apache Jena[39] are ready-to-use and provide SPARQL and Linked Data endpoints for semantic web applications[40]. Within the Heloise network, a collection of available applications is provided on Heloise Application Stack[41].

The Heloise platform[42] integrate basic modules within the application layer and the research interface layer. The platform is based on the content management system Drupal[43] hosted at LARHRA. It provides collaborative functionalities within a discussion forum and collaborative document editor. Furthermore, it integrates other hosted distributed modules[44]. Modules are services that are typically realized by interlinked applications (cf. Section 2). The following basic modules to be implemented have been proposed by the authors to the Heloise consortium (cf. Figure 1): Heloise Search Engine, Heloise Knowledge Units Cache, Heloise Triple Cache and the Heloise Historical Map.

The Heloise Search Engine[45] is the entry point for researchers to look for existing resources within the network. It is implemented by an Elastic Search Engine and individual parser plugins for each online available dataset. All partners that provide a URL of their database index (list of all individuals) can be added to the search index that is restricted to individuals.

The implementation of an Heloise Knowledge Units Cache[46] is based on the symogih.org ontology and is the integrated development of the Heloise Thesaurus (cf. Section 3). Caching in this case means that all data within the cache is recoverable at any time. The module provides a SPARQL interface (provided by an underlying Virtuoso triple store) to apply integrated queries

[36] Scheepers, M. J. "Virtualization and containerization of application infrastructure: A comparison." 21st Twente Student Conference on IT, Twente, Nederland's, June 2014.

[37] Dockerizing Linked Data: https://dockerizing.github.io/.

[38] Virtuoso universal server: http://virtuoso.openlinksw.com/.

[39] Fuseki SPARQL server: https://jena.apache.org/documentation/serving_data/.

[40] Arndt, N. et al. "Knowledge Base Shipping to the Linked Open Data Cloud" In Proceedings of the 11th International Conference on Semantic Systems (SEMANTICS '15) 2015.

[41] Heloise Applications Stack: http://apps.heloisenetwork.eu.

[42] Heloise Platform: http://platform.heloisenetwork.eu.

[43] Drupal CMS: https://www.drupal.org/.

[44] List of available applications: http://platform.heloisenetwork.eu/apps.

[45] Heloise Search Engine: http://search.module.heloisenetwork.eu.

[46] Heloise Knowledge Facts: http://knowledge-units.module.heloisenetwork.eu.

on all federated datasets. Closely related to this module are further Heloise Triple Cache modules that align the datasets to alternative vocabularies e.g. BIO.

The Heloise Thesaurus Module[47] is located in the Research Interface Layer as part of the scientific discussion on the description of resources. The module is implemented by the collaborative knowledge authoring tool OntoWiki. OntoWiki provides a generic user interface for authoring ontologies including discussion and rating functionality on the resources level.

5. Future Work

This paper presented the main concepts of the HCRM. The technical implementation of the Heloise Research and Publication Platform is shown by the first applications and the alignment of database samples using the simogih.org ontology. On the basis of these achievements, more Heloise databases will be aligned and will be available through the Heloise Search Engine and the Heloise Knowledge Units Cache. The modules and applications of the core platform have to be co-developed with an increasing participation of the Heloise partners in platform-supported research processes, e.g. developing the Heloise Network Thesaurus.

To cope with the large amount of datasets, more applications must be made available within the platform in the future. Particularly, applications to support the publishing of structured and semi-structured data as Linked Open Data. The use of a link discovery algorithm between datasets is envisaged, and also achievements in the evolution of the thesaurus. Furthermore, the Heloise partners should discuss an alignment of data licenses to make the data integration possible. The authors advise using the Creative Commons Attribution-ShareAlike 4.0 International License[48] accordingly to LOD.

Acknowledgements

We want to thank our colleagues within the Heloise Network for their helpful comments and inspiring discussions during the Heloise Workshops, as well for the providing their repositories as open data. In addition, we would like to thank our colleagues and students from Laboratoire de Recherche Historique Rhône-Alpes (LARHRA), University of Leipzig and Hochschule für Technik,

[47] Heloise Network Thesaurus: http://thesaurus.module.heloisenetwork.eu/.
[48] CC4.0-BY-SA License: http://creativecommons.org/licenses/by-sa/4.0/.

Wirtschaft und Kultur (HTWK), who contributed to the application stack, especially Djamel Ferhod, Natanael Arndt, Roy Meissner, Simeon Ackermann, Georgie Alkhouri, Tobias Hahn and Tom Neumann. This work was supported by LARHRA, University of Leipzig, HTWK and Leipzig Institute of Applied Informatics (InfAI) by providing basic infrastructure and IT services.

Bibliography

Arndt, Nathanael, Ackermann, M., Brümmer, M. and Riechert, Thomas. "Knowledge Base Shipping to the Linked Open Data Cloud". 11th International Conference on Semantic Systems Proceedings, September, Vienna, Austria 2015: 73-80.

Auer, Sören, Lehmann, Jens, Ngomo, Axel-Cyrille Ngonga and Zaveri, Amrapali. "Introduction to Linked Data and Its Lifecycle on the Web". In *Proceedings of the 9th international conference on Reasoning Web: semantic technologies for intelligent data access* (RW'13), 2013.

Beretta, Francesco. "Publishing and sharing historical data on the semantic web: the symogih.org project." Paper presented at the workshop: Semantic Web Applications in the Humanities, Göttingen, Germany, March 2015. https://halshs.archives-ouvertes.fr/halshs-01136533 .

Beretta, Francesco. "L'interopérabilité des données historiques et la question du modèle: l'ontologie du projet SyMoGIH". In Quels enjeux numériques pour les médiations scientifique et culturelle, Jean-Luc Minel (ed.), Paris: Presses universitaires de Paris Ouest, 2016.

Beretta, Francesco and Vernus, Pierre. "Le projet SyMoGIH et la modélisation de l'information : une opération scientifique au service de l'histoire". Les Carnets du LARHRA 1(2012): 81-107. http://halshs.archives-ouvertes.fr/halshs-00677658

Beretta, Francesco, Hours, Bernard, Butez, Charlotte and Vernus, Pierre. "Modeling and sharing geo-historical information : the SyMoGIH project." Poster presented at the conference Digital Humanities 2012, Hamburg, Germany http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/le-systeme-modulaire-de-gestion.1.html.

Beretta, Francesco, Ferhod, Djamel, Gedzelman, Séverine and Vernus, Pierre. "The SyMoGIH project: publishing and sharing historical data on the semantic web." Poster presented at the conference Digital Humanities 2014, Lausanne, Switzerland. http://halshs.archives-ouvertes.fr/halshs-01097399 .

Beretta, Francesco, Alamercery, Vincent and Ferhod, Djamel. "The symogih.org Project: Towards an International Consortium." Accepted paper for the conference Digital Humanities 2016, Krakow, Poland. https://halshs.archives-ouvertes.fr/halshs-01310700 .

Bizer, Christian and Cyganiak, Richard. "D2R Server – Publishing Relational Databases on the Semantic Web (Poster)" *Poster at the 5th International Semantic Web Conference*, Athens, USA, 2006.

Cyganiak, Richard and Jentzsch, Anja. "Linking open data cloud diagram." LOD Community (http://lod-cloud. net/) 12(2011).

Domingue, John, Fensel, Dieter, and Hendler, James A. (eds.). Handbook of semantic web technologies. Berlin, Heidelberg: Springer-Verlag, 2011.

Frischmuth, Philipp, Martin, Michael, Tramp, Sebastian, Riechert, Thomas and Auer, Sören. "OntoWiki - An Authoring, Publication and Visualization Interface for the Data Web." Semantic Web Journal (IOS Press), 6 (2015): 215-240. http://www.semantic-web-journal.net/content/ontowiki-authoring-publication-and-visualization-interface-data-web.

Kurtz, D., Parker, G., Shotton, D., Klyne, G., Schroff, F., Zisserman, A., & Wilks, Y. (2009, December). "Claros-bringing classical art to a global public." In e-Science, 2009. e-Science'09. Fifth IEEE International Conference on (pp. 20-27). IEEE.

Lehmann, Jens, Bizer, Chris, Kobilarov, Georgi, Auer, Sören, Becker, Christian, Cyganiak, Richard and Hellmann, Sebastian. "DBpedia - A Crystallization Point for the Web of Data." Journal of Web Semantics 7 (2009): 154-165.

Merkel, Dirk. "Docker: Lightweight linux containers for consistent development and deployment." Linux Journal, 2014. http://www.linuxjournal.com/content/docker-lightweight-linux-containers-consistent-development-and-deployment .

Meroño-Peñuela, Albert, Ashkpour, Ashkan, van Erp, Marieke, Mandemakers, Kees and Breure, Leen. "Semantic technologies for historical research: A survey." Semantic Web Journal (IOS Press) 6(2015): 539-564. http://www.semantic-web-journal.net/content/semantic-technologies-historical-research-survey-0.

Ngonga Ngomo, Axel-Cyrille and Auer, Sören. "LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data." Paper presented at the meeting of the Proceedings of IJCAI, 2011.

Sommerville, Ian. Software Engineering. Harlow, England: Addison-Wesley, 2010.

Scheepers, Mathijs Jeroen. "Virtualization and containerization of application infrastructure: A comparison." 21st Twente Student Conference on IT, Twente, Nederland's, June 2014.

Stadler, Claus, Lehmann, Jens, Höffner, Konrad and Auer, Sören. "LinkedGeoData: A Core for a Web of Spatial Open Data." Semantic Web Journal 3 (2012): 333-354.

Volz, Julius, Bizer, Christian, Gaedke, Martin and Kobilarov, Georgi. "Silk – A Link Discovery Framework for the Web of Data." 2nd Workshop about Linked Data on the Web (LDOW2009), Madrid, Spain, April 2009.