The ARIADNEplus Knowledge Base: a Linked Open Data set for archaeological research

Alessia Bardi[1,*], Miriam Baglioni[1,†], Michele Artini[1,‡], Andrea Mannocci[1,‡] and Gina Pavone[1,‡]

[1]*Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Italy*

Abstract

The ARIADNE infrastructure provides tools and services for researchers to address archaeological grand challenges that require discovery and analysis of information scattered across different thematic and geographically distributed sources. The ARIADNEplus Knowledge Base (KB) is an archaeological Linked Open Data set modelled according to the ARIADNE ontology, based on CIDOC-CRM, and provided by an international network of organisations leaders in different domains of archaeological sciences. In February 2024, the ARIADNEplus KB features about 4 million archaeological resources. Thanks to the ARIADNE infrastructure, data providers increased the level of fairness of their resources and contributed to a unique asset for the archaeology research community, the European Open Science Cloud and society at large.

Keywords

Knowledge graph, semantic web, e-infrastructure, interoperability

1. Introduction

ARIADNE is the Archaeological Research Infrastructure for Archaeological Data Networking in Europe. Since 2013, ARIADNE establishes a community in archaeological research, gathering together more than 40 organisations in the domain and about 11,000 archaeologists, corresponding to one third of all European archaeologists and probably more than 50% of those using some computer support in their research activities [1]. The ARIADNE infrastructure facilitates open access to Europe's archaeological heritage and proposes technical solutions to overcome the fragmentation of digital repositories, placed in different countries and compiled in different languages [1].

ARIADNE offers resources for digital data analysis, exploration, and research collaboration in line with the Open Science principles. One of those resources is the ARIADNEplus Knowledge Base (KB), an archaeological Linked Open Data set modelled according to the ARIADNE ontology and provided by an international network of organisations in the field. In February

SEBD 2024: 32nd Symposium on Advanced Database Systems, June 23-26, 2024, Villasimius, Sardinia, Italy

[*] Corresponding and main author.

[†] Second author.

[‡] These authors contributed equally.

alessia.bardi@isti.cnr.it (A. Bardi); miriam.baglioni@isti.cnr.it (M. Baglioni); michele.artini@isti.cnr.it (M. Artini); andrea.mannocci@isti.cnr.it (A. Mannocci); gina.pavone@isti.cnr.it (G. Pavone)

0000-0002-1112-1292 (A. Bardi); 0000-0002-2273-9004 (M. Baglioni); 0000-0002-5193-7851 (A. Mannocci); 0000-0003-0087-2151 (G. Pavone)

2024, the ARIADNEplus KB integrates about 4 million archaeological resources including archaeological reports, findings, inscriptions, archaeological sites and monuments from archives and repositories in Europe and beyond (e.g. Argentina, Japan).

The integration of archaeological datasets is realised by a metadata aggregation system based on the D-NET Framework Toolkit [2], which is a framework developed by CNR-ISTI for the realization, maintenance, and operation of (meta)data aggregative infrastructures. D-NET has been successfully applied to various thematic domains such as social history (HOPE [3]), the preservation of film archives (EFG [4]), and ancient epigraphy (EAGLE [5]). It has also been used to build scholarly communication infrastructures (DRIVER [6] for the aggregation of metadata about research publications from Open Access repositories) and it is currently used by OpenAIRE [1] for the construction of a scientific knowledge graph about scholarly communication. The D-NET framework provides developers with 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 exposing information objects to third-party applications through a number of standard access APIs. D-NET features 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.

For the ARIADNE research infrastructure, D-NET was configured to harmonise metadata records according to the ARIADNE Ontology (AO) [7], an extension of the CIDOC-CRM standard. Data providers define the mapping from their model to the AO with the 3M Editor developed by FORTH-ICS [8]. The resulting records are stored in a triple store implemented with a GraphDB[2] server and enriched with information from the Linked Open Datasets of the Getty Art & Architecture Thesaurus (AAT)[3] and PeriodO[4], a public gazetteer of scholarly definitions of historical, art-historical, and archaeological periods. Enriched records form the ARIADNEplus KB, accessible via a SPARQL endpoint, the GraphDB Workbench and the ARIADNE portal [5].

This paper describes how the different tools and services for semantic interoperability have been integrated to realise the ARIADNE aggregator and produce the ARIADNEplus Knowledge Base. Section 2 introduces the ARIADNE Ontology (AO) and its main classes. Section 3 describes how archaeological resources are aggregated and enriched. Section 4 provides insights about the ARIADNEplus KB and how the data is organised to support incremental aggregation of resources and to keep provenance information at different levels. Section 5 concludes the paper and outlines future work panned to be carried out in the context of the EC Horizion Europe project ATRIUM and the ARIADNE Research Infrastructure AISBL, the no-profit organisation founded to ensure long-term sustainability of the ARIADNE infrastructure.

[1] OpenAIRE, www.openaire.eu
[2] GraphDB, https://graphdb.ontotext.com/
[3] Getty AAT, https://www.getty.edu/research/tools/vocabularies/aat/about.html
[4] PeriodO Gazetteer, https://perio.do/en/
[5] ARIADNE Portal, https://portal.ariadne-infrastructure.eu/

2. The ARIADNE Ontology

The ARIADNE Ontology (AO) [7] was developed to integrate archaeological data of different type, granularity and geographical scope into a common information space. The CIDOC-CRM, the standard ontology in the cultural heritage domain, is the conceptual backbone of AO. AO specialises CIDOC-CRM to address specific modelling needs of archaeological sub-domains:

- AO-Cat for the representation of cataloguing information. It captures the basic 'What', 'When' and 'Where' information and provides an adequate representation for the discovery of resources relating to archaeological sites, monuments, artefacts, and data from the palaeo-anthropology, environmental, maritime and underwater archeology, and public archaeological finds;
- CRMhs for the representation of scientific data;
- aDNA for the representation of bio-archaeology and ancient DNA.

For other sub-domain we used AO-Cat in combination with existing extensions of CIDOC-CRM: CRMarchaeo for field survey information, CRMba for standing structures, and CRMtex for inscriptions [9, 10].

Figure 1 shows the AO-Cat class taxonomy:

- AO_Entity: the most general class of AO-Cat, all classes being sub-classes of AO_Entity.
- AO_Resource: any digital resource in the ARIADNE research infrastructure. The class hierarchy of AO_Resource is shown in Figure 2. Instances of AO_Resource can be:
  - AO_Service: a digital representation of a service, intended as an offer by some actor of their willingness and ability to execute an activity or series of activities upon request;
  - AO_Data_Resource: represents an archeological data resource at different granularity levels via the subclasses AO_Individual_Data_Resource and AO_Collection. Documents and digital images are sub-classes of AO_Individual_Data_Resource.
- AO_Object: digital representation of a physical object (e.g. an item found during an excavation).
- AO_Concept: terms used to classify entities in terms of type and subject.
- AO_Spatial_Region: represent spatial location identified as points, polygons, bounding boxes or simple place names.
- AO_Temporal_Region: represent time as absolute dates, time intervals expressed as absolute dates or period names. Period names are harmonised with the PeriodO gazetteer service.
- AO_Event and AO_Activity: correspond to CIDOC CRM events and activities.
- AO_Agent: persons or organisations that hold responsibilities for resources or that carry out activities (e.g. publisher, contributor).

For the full description of the model, refer to [7].

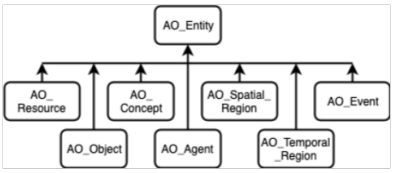

Figure 1: AO-Cat class taxonomy

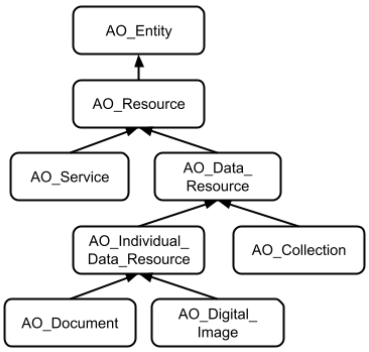

Figure 2: AO_Resource class hierarchy

3. Methodology

The ARIADNEplus Knowledge Base (KB) is populated via the ARIADNE aggregator, a D-NET instance customised for ARIADNE. The ARIADNE aggregator is capable of collecting detailed descriptions of archaeological resources (metadata records) in different formats, transforming the records according to the ARIADNE Ontology, enriching them with subject terms via Getty AAT and dating information via PeriodO and making them available via a SPARQL endpoint and the ARIADNE portal. The ARIADNE portal [6], developed by the Swedish National Data Service, where the resources can be searched and filtered by different criteria (e.g. by location, by historical period, by subject, by contributor).

Figure 3 shows the workflows defined for the processing of each dataset:

1. *Ingestion of XML records of the provider*. The workflow applies a 3M mapping [8] to each of the input records of a data source and generates RDF/XML records compliant with the ARIADNE Ontology. Transformed records are suitable for ingestion into the ARIADNEplus KB, an instance of GraphDB.
2. *Enrichment with Getty AAT subjects*. To better qualify the archaeological resources, providers are asked to map their local subjects and concepts into the Getty AAT. Getty AAT is a thesaurus of concepts describing different aspects of cultural heritage, such as materials, techniques, cultures (e.g., amphora, oil paint, Buddhism). The mapping between terms used by data providers and Getty AAT terms is done using the Vocabulary Matching Tool developed by University of South-Wales [7] [11]. The subject mapping to Getty AAT is transformed into RDF and fed into the KB. As a result, the KB contains the correspondences between native subjects and terms within the Getty AAT vocabulary.
3. *Enrichment with PeriodO*. For the enrichment with dating information, providers curate an authority file on PeriodO. The authority file specifies the time spans in absolute dates of historical periods that are referred in the metadata records [12]. The authority file is ingested into the knowledge base and used to generate explicit *aocat:has_period* properties.

[6] ARIADNE Portal, https://portal.ariadne-infrastructure.eu/
[7] Vocabulary Matching Tool, https://vmt.ariadne.d4science.org/vmt/vmt-app.html

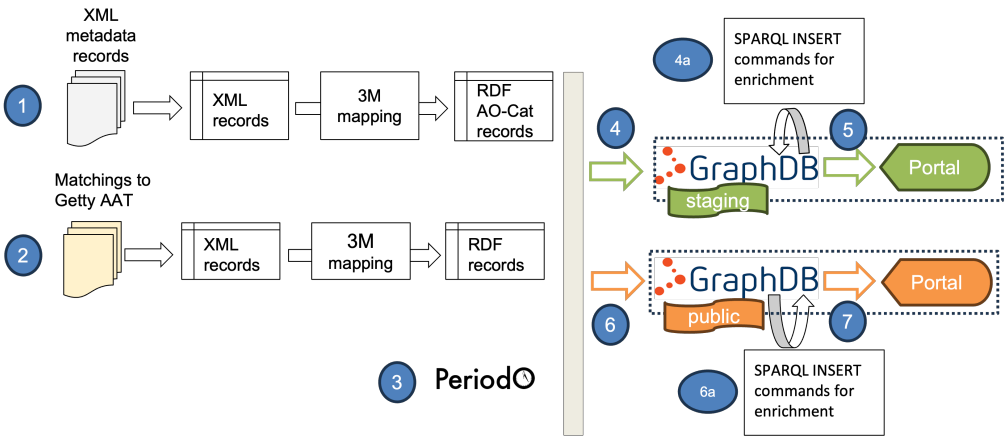

Figure 3: Workflows for the aggregation and enrichment of archaeological resources

4. *Feed the staging knowledge base.* In order to support the providers with checking the content before it is made publicly available, the push on the knowledge base initially targets a "staging" instance. Specific SPARQL INSERT statements are executed (4a) in order to enrich records with AO-Cat properties that are not explicitly available in the mapped records, but that are statically known or can be inferred from other properties or related records (e.g. inheritance of properties of a dataset from its collection). Step 4a also allows the aggregation manager to address possible peculiarities in the data that could complicate the automatic feeding to the portal.
5. *Feed the public knowledge base.* This flow is executed if the provider successfully completed the content checking on the staging knowledge base and portal.
6. *Feed the public portal.* The flow applies the same procedure as step 5, using the public instances of the knowledge base and portal instead of the staging instances.

Steps 4 and 5 support the quality checks of content and can be bypassed once the input format of the records and the 3M mappings of a provider are stable, so that it will be possible to automatically update the knowledge graph and the ARIADNE portal with updated and new records without human intervention. If necessary, steps 4 and 5 might be reactivated for a given source (e.g. because the provider upgraded the information system and wants to perform extensive checks before the data supplied by the new system goes public).

4. The ARIADNEplus Knowledge Base for Archaeological Sciences

As of February 2024, the ARIADNEplus KB integrates about 4 million archaeological resources including archaeological reports, findings, inscriptions, archaeological sites and monuments from archives and repositories in Europe and beyond.

On GraphDB, we count about 490M of triples (subject - predicate - object) describing 13K instances of AO_Collections and 3.9M instances AO_Individual_Data_Resource provided by 59

publishers. The ARIADNEplus KB is stored on a GraphDB server (version 9.8, free edition).

Data on GraphDB is organised in *named graphs* so that the ARIADNE aggregator can incrementally update the KB. GraphDB features one named graph for each data source. All records aggregated from the same data source are stored as triples in the same name graph.

The aggregator is thus able to request the deletion of that specific data source without affecting triples of other data sources. This feature of isolation was a requirement to support continuous aggregation and automatic update of the KB. Every time a dataset is updated (because the input metadata records changed or the 3M mapping changed), the aggregator requests the deletion of the named graph that corresponds to the data source at hand and then proceeds with the feeding of the updated records.

Enhancements to the provided records are added to dedicated named graphs, following a similar logic. As a result, for each data provider, GraphDB features several named graphs:

- One named graph with the triples of the records aggregated from the same data source. If the provider manages different data sources (e.g. two databases), then GraphDB will feature one named graph per data source.
- One named graph with the matches between local subjects and Getty AAT terms as defined by the provider with the Vocabulary Matching Tool.
- One named graph with the PeriodO authority files of the provider.
- One named graph with the triples inferred by intersecting the aggregated data and Getty AAT based on the provided matching.
- One named graph with the triples inferred by intersecting the aggregated data and the PeriodO authority files of the provider.

In addition, the aggregator adds provenance information to a special graph. The provenance graph contains information about when and which endpoints and which data sources have been added to the ARIADNEplus KB. Its triples are compliant with the PAV (Provenance, Authoring and Versioning) ontology [13].

The benefits of such a partition of content on GraphDB target data curators, aggregation managers, and end-users in different ways:

- Machine discoverability of new content and new providers in the KB thanks to the provenance graph;
- Easy identification of what has been aggregated and what has been inferred;
- Easy update of each subset of triples. If there are mistakes in the inference rules, only the graphs with inferred triples can be deleted, the inference rules updated and the relative graphs regenerated;
- Continuous updates of PeriodO terms, Getty AAT matching, and input data do not affect each other and can be run in isolation.

The main drawback is that SPARQL queries have to explicitly target different named graphs to get complete information about a resource. This may be not very convenient, especially for end-users who might not be fully aware of how the data is organised. We addressed the problem by engaging with the users and providing clear and public documentation.

Table 1
Top 5 results of query in Listing 1

Publisher	Subject	Count
Archaeology Data Service	Site/monument	776,056
British Museum	Artefact	476,224
Historic Environment Scotland	Site/monument	334,636
Museum of Cultural History, University of Oslo	Artefact	296,165
Archaeology Data Service	Fieldwork	189,803

In December 2022 we organised a hackathon at the LinkedPasts conference in York, where we engaged with users of the ARIADNE infrastructure and IT people with a knowledge of the archaeological field. User support and feedback is continuously gathered via the ARIADNEplus Lab Virtual Research Environment (VRE)[8]. The VRE offers a social feed where users can post and reply to comments and questions.

Finally, we prepared documentation with examples in the form of a Jupyter Notebook that can be run on the JupyterHub available via the ARIADNEplus Lab VRE. The Jupyter Notebook uses Python and the SPARQL Wrapper library[9] to execute queries that are useful to understand the organisation of the data, its coverage and richness. Listing 1, for example, gets the number of resources grouped by publisher and typology (*aocat:has_ARIADNE_subject*). The first five results are in Table 1.

Listing 1: Get the number of resources grouped by publisher and typology

```
PREFIX aocat: <https://www.ariadne-infrastructure.eu/resource/
    ao/cat/1.1/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT (count(?resource) AS ?cnt) ?publisherName ?asl WHERE {
    ?resource aocat:has_publisher ?publisher .
    ?publisher aocat:has_name ?publisherName .
    ?resource aocat:has_ARIADNE_subject ?as .
    ?as rdfs:label ?asl
}
GROUP BY ?publisherName ?asl
ORDER BY DESC(?cnt)
```

5. Conclusion and future work

To our knowledge, the ARIADNEplus KB provides access to the largest international archaeological dataset available online, with about 4 millions resources maintained by 59 publishers in Europe and beyond.

[8]ARIADNE Lab Virtual Research Environment by D4Science.org, https://ariadne.d4science.org/group/ariadneplus_lab
[9]SPARQL Wrapper Library, https://github.com/RDFLib/sparqlwrapper

Thanks to the ARIADNE Ontology, based on the CIDOC-CRM standard, heterogeneous data can be harmonised and offered via a single-entry point.

The harmonisation process is managed by the ARIADNE aggregator, a system based on the D-NET framework toolkit, the 3M Editor and the Vocabulary Matching Tool. The combination of the three services proved to be effective to deal with very different use cases and to manage interoperability challenges due to idiosyncratic exchange protocols, metadata models and formats.

Organizations providing content to the KB improved the level of FAIRness of their data. Each resource in the KB is described according to the ARIADNE Ontology, which was defined together with the research community to address the needs of researchers in archaeology ad its many sub-domains. Each resource is assigned a unique and persistent URL that resolves either on its landing page on the ARIADNE portal or to its RDF/XML representation (based on content negotiation). Descriptions of the resources are enriched with properties and links to standard vocabularies and gazetters (e.g. PeriodO and Getty AAT). The KB is a Linked Open Dataset, compliant with the Resource Description Framework and queriable via the standard SPARQL protocol.

The re-usability of the KB and of the software of the portal[10] is demonstrated by the Unpath'd Waters Portal[11] launched in April 2023. The Archaeology Data Service adapted the ARIADNE portal to provide a discovery portal for resources about maritime heritage of UK coastal waters available in the ARIADNEplus KB. As highlighted also in [10], the same approach could be easily adopted by other projects or initiatives willing to provide thematic or national portal without the burden and costs of maintaining a dedicated aggregation system.

In November 2022, ARIADNE has become a not-for-profit association registered under Belgian law, but operating internationally, named ARIADNE Research Infrastructure AISBL. As of February 2024, ARIADNE RI AISBL has 29 organisational members from 20 countries, including Italy, United Kingdom, and Japan.[12] The setup of the association was fundamental to ensure the long-term sustainability of the ARIADNE infrastructure.

Thanks to the participation in the project ATRIUM (Advancing FronTier Research In the Arts and hUManities), the ARIADNE infrastructure will further grow its community and the coverage of its knowledge base. The project is funded by the European Commission under the Horizon Europe Framework programme. It started in January 2024 and will last for 4 years. During the project the ARIADNEplus KB will be enriched with additional content like reports on primary fieldwork, standing building surveys, images, and an improved management of geo-spatial data. ARIADNE services will also be registered in the SSH Open Marketplace[13], contributing to the European Open Science Cloud for Social Sciences and Humanities.

[10] ARIADNE portal software, https://github.com/ariadne-infrastructure
[11] Unpath'd Waters Portal, https://unpathd.ads.ac.uk/
[12] ARIADNE Research Infrastructure AISBL members, https://www.ariadne-research-infrastructure.eu/partners/
[13] SSH Open Marketplace, https://sshopencloud.eu/ssh-open-marketplace

6. Data availability statement

The ARIADNEplus KB is accessible via the ARIADNEplus Lab Virtual Research Environment (VRE) hosted by the D4Science infrastructure at https://ariadne.d4science.org/group/ariadneplus_lab/.

The Jupyter Notebook with instructions and sample queries is available at https://data.d4science.net/YuUq and can be run on the JupyterHub available in the VRE linked above.

The ARIADNE portal is accessible at https://portal.ariadne-infrastructure.eu/. Its source code is at https://github.com/ariadne-infrastructure.

Acknowledgments

This work has been supported by ARIADNEplus EC H2020 Grant 823914 and ATRIUM EC HE Grant 101132163. The ARIADNE aggregator is operated by CNR-ISTI on the D4Science infrastructure (https://www.d4science.org/). We thank the members of the ARIADNEplus project, during which the aggregation system was designed and developed, especially the members of the aggregation task force, with their strong commitment and passion: Ceri Binding, Achille Felicetti, Carlo Meghini, Enrico Ottonello, Julian Richards, and Maria Theodoridou.

References

[1] F. Niccolucci, J. Richards, The ARIADNE Impact, ARCHAEOLINGUA FOUNDATION, 2020. URL: https://doi.org/10.5281/zenodo.4319058. doi:10.5281/zenodo.4319058.

[2] P. Manghi, et al., The D-NET software toolkit A framework for the realization, maintenance, and operation of aggregative infrastructures, Program (Lond., 1966) 48 (2014) 322–354. doi:10.1108/prog-08-2013-0045, publisher: Emerald, Bradford, Regno Unito.

[3] A. Bardi, P. Manghi, F. Zoppi, Coping with interoperability and sustainability in cultural heritage aggregative data infrastructures, International Journal of Metadata, Semantics and Ontologies 9 (2014) 138. URL: http://dx.doi.org/10.1504/IJMSO.2014.060341. doi:10.1504/ijmso.2014.060341.

[4] M. Artini, A. Bardi, F. Biagini, F. Debole, S. La Bruzzo, P. Manghi, M. Mikulicic, P. Savino, F. Zoppi, Data interoperability and curation: The european film gateway experience, in: M. Agosti, F. Esposito, S. Ferilli, N. Ferro (Eds.), Digital Libraries and Archives, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 33–44.

[5] A. Mannocci, V. Casarosa, P. Manghi, F. Zoppi, The eagle europeana network of ancient greek and latin epigraphy: A technical perspective, in: D. Calvanese, D. De Nart, C. Tasso (Eds.), Digital Libraries on the Move, Springer International Publishing, Cham, 2016, pp. 75–78.

[6] P. Manghi, M. Mikulicic, L. Candela, D. Castelli, P. Pagano, Realizing and maintaining aggregative digital library systems: D-net software toolkit and oaister system, D-Lib Magazine 16 (2010). URL: http://dx.doi.org/10.1045/march2010-manghi. doi:10.1045/march2010-manghi.

[7] A. Felicetti, C. Meghini, J. Richards, M. Theodoridou, The ao-cat ontology, 2023. URL: https://doi.org/10.5281/zenodo.7818375. doi:10.5281/zenodo.7818375.
[8] Y. Marketakis, N. Minadakis, H. Kondylakis, K. Konsolaki, G. Samaritakis, M. Theodoridou, G. Flouris, M. Doerr, X3ml mapping framework for information integration in cultural heritage and beyond, International Journal on Digital Libraries 18 (2016) 301–319. URL: http://dx.doi.org/10.1007/s00799-016-0179-1. doi:10.1007/s00799-016-0179-1.
[9] J. Richards, A. Felicetti, C. Meghini, M. Theodoridou, D4.4 – final report on ontology implementation, 2023. URL: https://doi.org/10.5281/zenodo.7636720. doi:10.5281/zenodo.7636720.
[10] J. D. Richards, U. of York, Joined up thinking: Aggregating archaeological datasets at an international scale, Internet archaeology (2023). URL: http://dx.doi.org/10.11141/ia.64.3. doi:10.11141/ia.64.3.
[11] C. Binding, D. Tudhope, Improving interoperability using vocabulary linked data, International Journal on Digital Libraries 17 (2015) 5–21. URL: http://dx.doi.org/10.1007/s00799-015-0166-y. doi:10.1007/s00799-015-0166-y.
[12] C. Binding, Implementing archaeological time periods using cidoc crm and skos, in: L. Aroyo, G. Antoniou, E. Hyvönen, A. ten Teije, H. Stuckenschmidt, L. Cabral, T. Tudorache (Eds.), The Semantic Web: Research and Applications, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 273–287.
[13] P. Ciccarese, S. Soiland-Reyes, K. Belhajjame, A. J. G. Gray, C. Goble, T. Clark, PAV ontology: provenance, authoring and versioning, J. Biomed. Semantics 4 (2013) 37.