

ASIO: a Research Management System based on Semantic technologies

Jose Emilio Labra Gayo¹[0000-0001-8907-5348], José Barranquero Tolosa²,
Guillermo Facundo Colunga¹[0000-0003-1283-2763], Alejandro González
Hevia¹[0000-0003-1394-5073], Emilio Rubiera Azcona¹[0000-0002-0292-9177],
Daniel Ruiz Santamaría², and Paulino Álvarez de Ron Ondina¹

¹ WESO Research Group, University of Oviedo, Spain

² Izertis S.A., Spain

Abstract. In this paper we describe the architecture of a Research Management System based on Semantic technologies. The system is composed from two main modules: ontological infrastructure and research management system which are communicated through an RDF triple store that integrates all the information. The data model is defined in terms of Shape Expressions which are synchronized with Java entities that define the data model. The shapes also act as core layer that can be used to describe the main entities that will be employed and to validate their ontological definitions with test data. In this way, we propose a test-driven development approach for ontological engineering that improves the quality of both the ontologies defined and the data. The semantic architecture is based on a reactive approach which combines both a clean architecture and a stream-based pattern. This paper describes the architecture of the system and the main quality attributes and design decisions that have been taken into account.

Keywords: research management, semantic technologies, linked data, ontology, stream processing, shape expressions

1 Introduction

There is an increase interest in the development of Research management systems which improve the available services for both researchers, research administrators and citizens in general. However, representing scholarly information is a complex task which requires to take into account the differences between disciplines, the decentralized nature of research advances and collaborations, as well as the increasing amounts of research data that are generated and need to be collected and taken into account. Although several approaches have tackled this problem promoting the use of linked data and semantic web technologies there is still a need to develop research management systems which are adopted by the different institutions and easily integrate their data models.

The HERCULES project³ has been proposed as a University Research Data Semantics system for Spanish universities with the goal of developing new semantic web technologies that gather new information and integrate multiple nodes with heterogeneous ontologies and vocabularies. The University of Murcia (hereinafter UM) signed an agreement with the Spanish Ministry of Economy, Industry and Competitiveness (MINECO) in 2017 backing the HERCULES project with an 80% of cofinancing from the European Regional Development Fund program (ERDF) within the 2014-2020 period. The purpose of this agreement was to establish collaboration amongst MINECO and the UM, directed towards the improvement of public services and business innovation through the Public Procurement of Innovation. Some goals of the system were to create a research management system with semantic capabilities and infrastructures and create support systems for the detection of synergies in R&D between universities. The HERCULES project was divided in three main sub-projects: semantic architecture and ontological infrastructure, research management system; and data enrichment and methods of analysis.

In this paper we present the architecture of the first sub-project called ASIO by its Spanish acronym, which has been proposed by one of the contractor teams formed by Izertis⁴ company and the WESO research group⁵ from the University of Oviedo, Spain. More information about the ASIO project can be obtained at <https://www.um.es/web/hercules/proyectos/asio>.

2 Architecture

The architecture of the system is decomposed in two main building blocks: firstly, the ASIO Semantic architecture offers a number of services to end-users and other applications through a linked data platform. It also has access to external data and offers logging and monitoring services. All important information from this system is stored in a triple store which offers a SPARQL endpoint. Secondly, the Ontological infrastructure building block contains a core ontology and a list of vertical modules defined using OWL by domain experts and ontology engineers. Some resources come from the conversion of external resources to OWL.

2.1 ASIO semantic architecture backend

The ASIO semantic architecture backend module is itself decomposed in two main modules:

- A front-end module which offers a linked data platform API as well as a web publication service. Those systems are offered by the Trellis framework⁶.

³ <https://www.um.es/web/hercules/>

⁴ <https://www.izertis.com/>

⁵ <http://www.weso.es>

⁶ <https://www.trellisldp.org/>

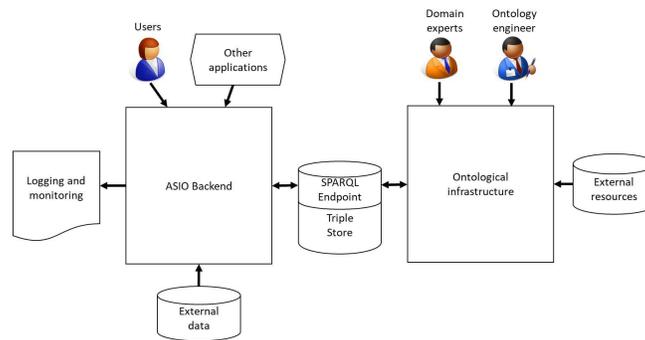


Fig. 1: Main diagram

- The backend is an evolution of the architecture presented at [11] combined with a clean architecture where entities are obtained from the shapes defined by domain experts. The management system stores the events in the event log using Apache Kafka⁷. Following the event sourcing pattern [9], two event processors and storage adapters capture the events and generate different serving data stores, one for Trellis using Apache Jena TDB⁸ and another for a Wikibase instance⁹. The URI Factory component is in charge of normalizing the URIs employed in the system as well as linking between the internal URIs generated by Wikibase and the external URIs exported by the system that follow a specific pattern tailored to the system.

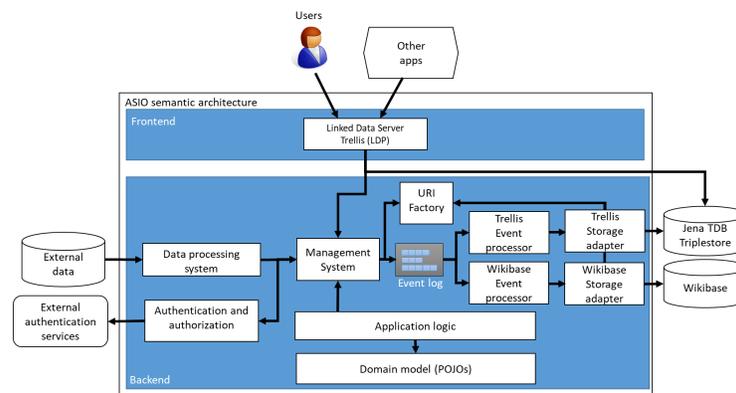


Fig. 2: ASIO Backend

⁷ <https://kafka.apache.org/>

⁸ <https://jena.apache.org/>

⁹ <https://wikiba.se/>

The ASIO architecture follows the clean architecture pattern [20] where the domain model is defined using Plain-old Java Objects (POJOs) and is separated from the application logic. An auto-imposed constraint is that the POJOs which define the domain model are generated from Shapes defined at the ontological infrastructure. An intended advantage of this approach is that the ASIO ontology and the Domain model used by the Java developers will be consistent without requiring the developers to understand the underpinnings of the ontologies or the ontology engineers to delve into Java code.

2.2 Ontological infrastructure

The goal of this module is to allow domain experts and ontology engineers to have an environment where they can define ontologies that form the backbone of system's data. We have decided to employ software engineering best practices like modularity, test-driven development, and continuous integration. With regards to modularity, the ontology has been divided in a core part which defines the generic concepts from the research domain, and several vertical ontologies in more specific domains like: geographical names, human resources, some specific university systems like the Spanish one, the English one, etc. The use of a test-driven development methodology for ontologies [18] is accomplished by the use of shape expressions (ShEx) [23]. We are using a subset of ShEx, called ShEx-Lite¹⁰, which enables domain experts to declare the shapes in a tabular way and at the same time can be used to generate Java code. Indeed some shapes can be also used to generate the POJOs employed as domain entities in the semantic architecture, as can be seen in Fig. 3.

Researcher.shex	Researcher.java
<pre> 1 # Prefixes... 2 :Researcher { 3 :name xsd:string ; 4 :surname xsd:string ; 5 :orcid xsd:xstring ; 6 :publications :Publication * 7 ... 8 }</pre>	<pre> 1 // Imports... 2 public class Researcher { 3 private String name; 4 private String surname; 5 private String orcid; 6 private Publication[] publications; 7 ... 8 // Constructor... 9 // Getters and Setters... 10 }</pre>

Fig. 3: Schema modeling a **Researcher** in ShEx syntax to the left. And the generated equivalent Java code to the right.

We have also developed a synchronization system between the ontology source code in RDF and a Wikibase instance. In this way, domain experts can edit the ontology either with a test editor, with an IDE like Protégé¹¹ or in a collaborative system like Wikibase.

¹⁰ <https://www.weso.es/shex-lite/>

¹¹ <https://protege.stanford.edu/>

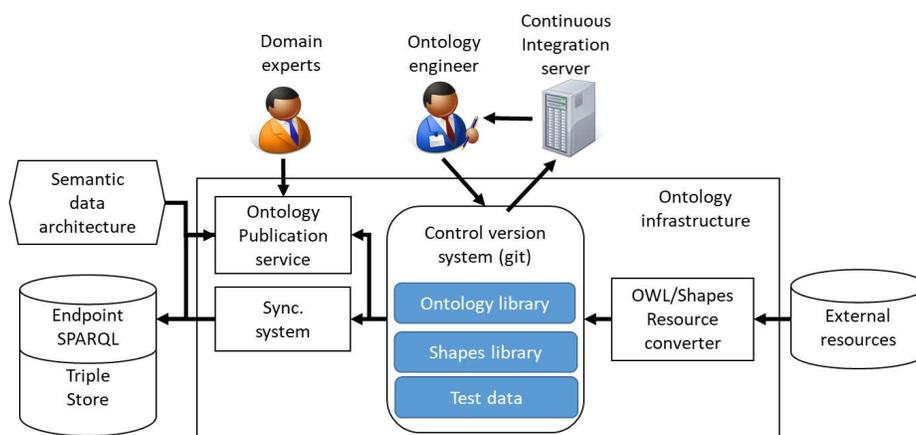


Fig. 4: Ontological infrastructure

This synchronization system is already being used to synchronize the contents of the ASIO ontology¹² with a custom Wikibase instance. This instance is not publicly accessible but we have deployed a public copy at <https://herc-core.wiki.opencura.com/>. In Fig. 5 we can see a screenshot of an example property from the ontology that has been synchronized to the Wikibase.

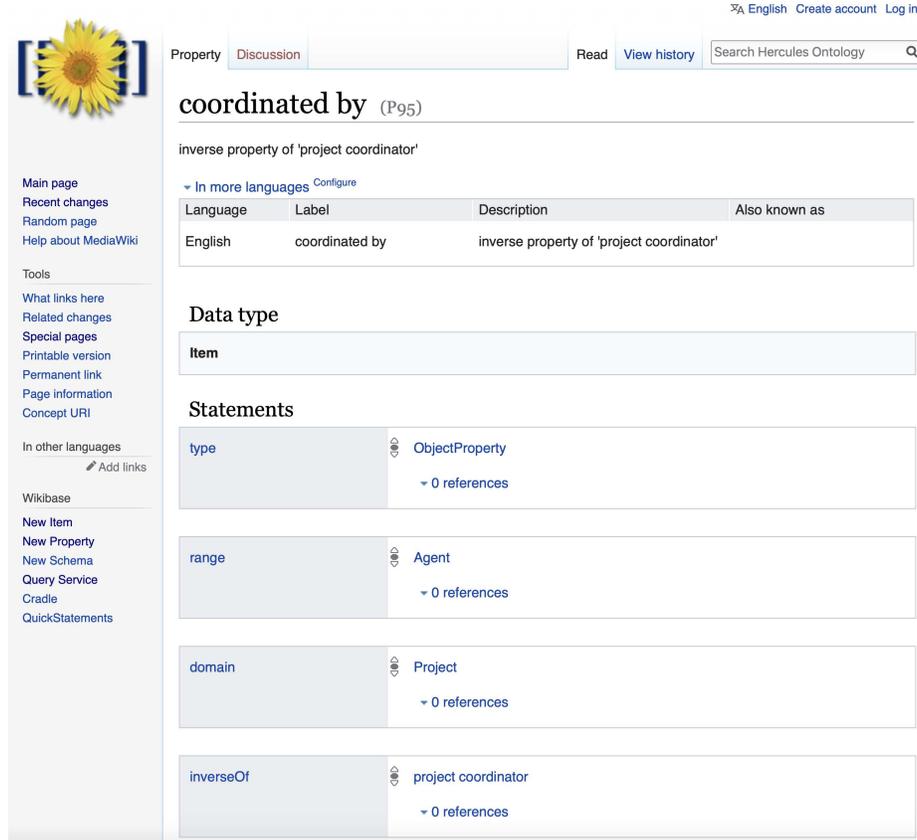
3 ASIO ontology

The ASIO ontology is designed to address the Research Management of the Spanish University System through the particular case of the University of Murcia but applying an encompassing model capable of addressing the rest of Spanish universities and also others from the European level.

The ontology is split into a central and peripheral modules, loosely inspiring ourselves in Fodor [8]. To do so we need to distinguish between two fundamentally different types of information processing (relying upon information architecture and datasets). In any informational system, in this case an ontology, there must be non-modular processing or what we call central processing, to distinguish it from modular processing, which is peripheral (our vertical modules). To say that a part of the ontology is core (i.e. involves central processing actions) is, essentially, to say that it is not informationally encapsulated (as the vertical modules). In principle, any part of the system is relevant to confirming any other and we do not draw boundaries within it.

On one hand, there are highly-specialised information processing datasets aimed at identifying and retrieving data from very specific environments. Informational based datasets in these specific environments involve only a limited type of information. That is why information retrieval tasks having to do with

¹² Available at <https://github.com/weso/hercules-ontology>



Property [Discussion](#) [Read](#) [View history](#)

coordinated by (P95)

inverse property of 'project coordinator'

[In more languages](#) Configure

Language	Label	Description	Also known as
English	coordinated by	inverse property of 'project coordinator'	

Data type

Item

Statements

type	ObjectProperty	0 references
range	Agent	0 references
domain	Project	0 references
inverseOf	project coordinator	0 references

Wikibase

- New Item
- New Property
- New Schema
- Query Service
- Cradle
- QuickStatements

Fig. 5: Example of property *coordinated by* in the Wikibase.

this first type of information are carried out by dedicated parts of the ontology that we call *modules*. These modules are *domain-specific*, that is, they are responsible only for containing data falling in particular domains (geopolitical, scientific, administrative, staffing, etc.).

On the other hand, there are central informational tasks that involve much more complex and wide-ranging inferences and to which an indefinite amount of background information is potentially relevant. The information processing involved in carrying out these tasks is *domain-general* (conversely to domain-specific) and it concerns our main *university domain* (our *core*), because we understand *general* here as our general domain.

On the basis of this distinction, we develop an architecture of the ontological organization as involving both very specialized modules (*vertical modules*) and what we call domain-general, non-modular knowledge (*core ontology*). Two properties of modularity in particular, *information encapsulation* and *domain*

specificity, make it possible to tie together questions of functional architecture with those of knowledge content.

As far as the *vertical modules* are concerned, six modules have been implemented so far, namely:

- geopolitical entities
- administrative entities
- scientific domains
- subject areas
- Spanish universities
- human resources from some national university systems (Spain, Portugal and others)

The aim of this ontological architecture split into core and vertical modules is to facilitate the integration with other possible ontologies from university stakeholders, enabling some customisation through the following procedures:

1. At the ontological core level, some examples of mappings between entities are proposed and illustrated by means of the `owl:equivalentClass` property, which map to equivalent entities in related ontologies, such as VIVO¹³ or CERIF¹⁴.
2. With the use of SKOS¹⁵ in vertical modules and their properties `skos:exactMatch`, `skos:closeMatch`, etc. the possibility of mapping the realities of the different stakeholders is enabled.

4 Related work

The vivo platform was proposed as a mechanism to enable collaboration between scientists [5]. It was based on semantic web technologies with the VIVO-ISF (Integrated Semantic Framework) providing a set of ontologies for the system. OpenVIVO [16] was later proposed as an demonstration of the VIVO project where anyone with an ORCID account can join. VIVO has been adopted by a large number of institutions, specially in the United States, but also in other countries around the world. On the other hand, euroCRIS¹⁶, is an international organization of research information systems whose mission is to promote cooperation and share research information through the CERIF, the Common European Research Information Format. Although CERIF data model is based on XML, an initial ontology was developed for CERIF. The need to connect research management systems using linked open data was already proposed by [17, 25].

The OpenAIRE (Open Access Initiative for Research in Europe)¹⁷ initiative started as a project to promote Open Access and gradually moved to Open

¹³ <http://vivoweb.org/ontology/core>

¹⁴ <http://www.eurocris.org/ontologies/cerif>

¹⁵ <https://www.w3.org/2004/02/skos/>

¹⁶ <https://eurocris.org/>

¹⁷ <https://www.openaire.eu/>

Science offering a research graph of linked data that relates publications, funders, research infrastructures, etc. [1].

The FAIR principles [26] promote the adoption of research data management systems that provide findable, accessible, interoperable and reusable data. The FAIR principles have been related to the linked open data guidelines [13] and the intersection between research data management, FAIR and open data has also been proposed by [14]. In a report developed by European Commission Expert Group on FAIR data, wikidata was mentioned as one of the technologies enabling FAIR data [15]. More recently the TRUST mnemonic has been proposed to promote trustworthy digital repositories which offer transparency, responsibility, user focus, sustainability and technological capabilities [19]. The ASIO model presented in this paper starts from the Spanish University system with a more modest goal but trying to offer a flexible system that can be later adopted by other research institutions, adopting FAIR and linked data principles and a scalable architecture. From a semantic point of view, the ASIO ontology maps its concepts to the VIVO and CERIF ontologies enabling future interoperability with those models. The reproducibility crisis of research has fostered the appearance of new initiatives like research objects [2–4, 12, 22]. The ASIO technology will give first class support to publish research objects following FAIR principles.

Combining ontological representations of science with research data platforms leads to the development of science knowledge graphs. ResearchSpace is a platform designed at the British Museum to help collaborations between researchers [21]. The Science Knowledge Graph Ontologies is a suite of ontologies that model research findings in various fields of modern science and which enable the development of a knowledge graph for science [7].

Shape Expressions (ShEx) have been proposed as a concise, human-friendly language to describe and validate RDF data [23]. There have been multiple applications of ShEx to different domains already [24], although as far as we know, the use of ShEx schemas to define data models and generate domain models in Java is new. The software architecture proposed is based on the event sourcing and CQRS patterns proposed by Young [27] and Dahan [6] for high performance applications. The pattern has already been employed at [11]. The ontological infrastructure usage of SKOS to map concepts between ontologies has also been proposed in [10]. We have adopted a test-driven approach for ontology development inspired by [18] but validating the ontologies using test data plus shapes.

5 Conclusions

We have presented the ASIO semantic architecture for research data management. The main quality attributes that we aim to fulfil are interoperability, scalability and reusability. Following the linked data platform principles we also aim to offer a decentralized solution were the different universities can gradually join the HERCULES project. A prototype system is being developed where the main components are glued together.

Acknowledgements. The HÉRCULES Semantic University Research Data Project is backed by the Ministry of Economy, Industry and Competitiveness with a budget of 5.462.600,00 euros with an 80% of cofinancing from the 2014-2020 ERDF Program. It has also been partially funded by the Spanish Ministry of Economy and Competitiveness (Society challenges:TIN2017-88877-R).

References

1. Alexiou, G., Vahdati, S., Lange, C., Papastefanatos, G., Lohmann, S.: OpenAIRE Lod Services: Scholarly Communication Data As Linked Data (2017). <https://doi.org/10.5281/zenodo.293836>
2. Bechhofer, S., Buchan, I., Roure, D.D., Missier, P., Ainsworth, J., Bhagat, J., Couch, P., Cruickshank, D., Delderfield, M., Dunlop, I., Gamble, M., Michaelides, D., Owen, S., Newman, D., Sufi, S., Goble, C.: Why linked data is not enough for scientists. *Future Generation Computer Systems* **29**(2), 599–611 (feb 2013). <https://doi.org/10.1016/j.future.2011.08.004>
3. Belhajjame, K., Zhao, J., Garijo, D., Gamble, M., Hettne, K., Palma, R., Mina, E., Corcho, O., Gómez-Pérez, J.M., Bechhofer, S., Klyne, G., Goble, C.: Using a suite of ontologies for preserving workflow-centric research objects. *Journal of Web Semantics* **32**, 16–42 (may 2015). <https://doi.org/10.1016/j.websem.2015.01.003>
4. Brinckman, A., Chard, K., Gaffney, N., Hategan, M., Jones, M.B., Kowalik, K., Kulasekaran, S., Ludscher, B., Mecum, B.D., Nabrzyski, J., Stodden, V., Taylor, I.J., Turk, M.J., Turner, K.: Computing environments for reproducibility: Capturing the “whole tale”. *Future Generation Computer Systems* **94**, 854–867 (may 2019). <https://doi.org/10.1016/j.future.2017.12.029>
5. Corson-Rikert, J., Cramer, E.J.: Vivo: Enabling national networking of scientists. In: IASSIST (2010)
6. Dahan, U.: Clarified CQRS, <https://udidahan.com/2009/12/0/>
7. Fathalla, S., Auer, S., Lange, C.: Towards the semantic formalization of science. In: Proceedings of the 35th Annual ACM Symposium on Applied Computing. ACM (mar 2020). <https://doi.org/10.1145/3341105.3374132>
8. Fodor, J.A.: *The Modularity of Mind*. MIT Press (1983)
9. Fowler, M.: Event sourcing (Dec 2005), <https://www.martinfowler.com/eaDev/EventSourcing.html>
10. Frosterus, M., Tuominen, J., Pessala, S., Hyvnen, E.: Linked open ontology cloud: managing a system of interlinked cross-domain lightweight ontologies. *International Journal of Metadata, Semantics and Ontologies* **10**(3), 189 (2015). <https://doi.org/10.1504/ijmso.2015.073879>
11. García-González, H., Fernández-Álvarez, D., Labra-Gayo, J.E., de Pablos, P.O.: Applying big data and stream processing to the real estate domain. *Behaviour & Information Technology* **38**(9), 950–958 (may 2019). <https://doi.org/10.1080/0144929x.2019.1620858>
12. Giraldo, O., García, A., López, F., Corcho, O.: Using semantics for representing experimental protocols. *Journal of Biomedical Semantics* **8**(1) (nov 2017). <https://doi.org/10.1186/s13326-017-0160-y>
13. Hasnain, A., Rebholz-Schuhmann, D.: Assessing FAIR Data Principles Against the 5-Star Open Data Principles. In: *Lecture Notes in Computer Science*, pp. 469–477. Springer International Publishing (2018). <https://doi.org/10.1007/978-3-319-98192-5-60>

14. Higman, R., Bangert, D., Jones, S.: Three camps, one destination: the intersections of research data management, FAIR and open. *Insights the UKSG journal* **32** (2019). <https://doi.org/10.1629/uksg.468>
15. Hodson, S., Jones, S., Collins, S., Genova, F., Harrower, N., Laaksonen, L., Mietchen, D., Petrauskait, R., Wittenburg, P.: Turning fair data into reality: interim report from the european commission expert group on fair data (2018). <https://doi.org/10.5281/ZENODO.1285272>
16. Ilik, V., Conlon, M., Triggs, G., White, M., Javed, M., Brush, M., Gutzman, K., Essaid, S., Friedman, P., Porter, S., Szomszor, M., Haendel, M.A., Eichmann, D., Holmes, K.L.: OpenVIVO: Transparency in scholarship. *Frontiers in Research Metrics and Analytics* **2** (mar 2018). <https://doi.org/10.3389/frma.2017.00012>
17. Joerg, B., Ruiz-Rube, I., Sicilia, M.A., Dvořák, J., Jeffery, K., Hoellrigl, T., Rasmussen, H.S., Engfer, A., Vestdam, T., Barriocanal, E.G.: Connecting closed world research information systems through the linked open data web. *International Journal of Software Engineering and Knowledge Engineering* **22**(03), 345–364 (may 2012). <https://doi.org/10.1142/s0218194012400074>
18. Keet, C.M., Lawrynowicz, A.: Test-driven development of ontologies. In: *Proceedings of the 13th International Conference on The Semantic Web. Latest Advances and New Domains - Volume 9678*. pp. 642–657. Springer-Verlag, Berlin, Heidelberg (2016). https://doi.org/10.1007/978-3-319-34129-3_39
19. Lin, D., Crabtree, J., Dillo, I., Downs, R.R., Edmunds, R., Giaretta, D., Giusti, M.D., L’Hours, H., Hugo, W., Jenkyns, R., Khodiyar, V., Martone, M.E., Mokrane, M., Navale, V., Petters, J., Sierman, B., Sokolova, D.V., Stockhause, M., Westbrook, J.: The TRUST principles for digital repositories. *Scientific Data* **7**(1) (may 2020). <https://doi.org/10.1038/s41597-020-0486-7>
20. Martin, R.: *Clean architecture*. Prentice Hall (2017)
21. Oldman, D., Tanase, D.: Reshaping the knowledge graph by connecting researchers, data and practices in ResearchSpace. In: *Lecture Notes in Computer Science*, pp. 325–340. Springer International Publishing (2018). <https://doi.org/10.1007/978-3-030-00668-6-20>
22. Pasmín, O.X.G., Corcho, O., Castro, A.G.: SMART Protocols: seMAnTic representation for experimental protocols (2014), <http://oa.upm.es/36778/>
23. Prud’hommeaux, E., Gayo, J.E.L., Solbrig, H.: Shape Expressions: An RDF Validation and Transformation Language. In: *Proceedings of the 10th International Conference on Semantic Systems* (2014). <https://doi.org/10.1145/2660517.2660523>
24. Thornton, K., Solbrig, H., Stupp, G.S., Gayo, J.E.L., Mietchen, D., Prud’hommeaux, E., Waagmeester, A.: Using shape expressions (ShEx) to share RDF data models and to guide curation with rigorous validation. In: *The Semantic Web*, pp. 606–620. Springer International Publishing (2019). <https://doi.org/10.1007/978-3-030-21348-0-39>
25. Wiljes, C., amd Florian Lier, N.J., Paul-Stueve, T., Vompras, J., Pietsch, C., Cimiano, P.: Towards linked research data: An institutional approach. In: Castro, A.G., Lange, C., Lord, P., Stevens, R. (eds.) *3rd Workshop on Semantic Publishing (SePublica)*. CEUR Workshop Proceedings, vol. 994, pp. 27–38 (2013)
26. Wilkinson, M.D., Dumontier, M., Aalbersberg, I., Appleton, G., Al., E.: The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* (2016). <https://doi.org/doi:10.1038/sdata.2016.18>
27. Young, G.: CQRS and Event Sourcing, <http://codebetter.com/gregyoung/2010/02/13/cqrs-and-event-sourcing/>