Linking model

Introduction defining context and scope

The role of the semantic linking model is to provide a framework for translating between the different standards used for data and process specification in the environmental sciences in the context of the ENVRI reference model. This model should provide a formal basis on which to improve the interoperability of RI services and products, by focusing on the vocabularies used by the ENVRI RIs, feeding into the design of the abstract architecture for interoperable RIs in general. The model also serves to provide the machine-readable formalisation of the ENVRI reference model (or at least its concept model).

Ultimately, based on the relevant task (5.3) of the ENVRI+ project, we will need to:

1. Capture the conceptual vocabulary of the ENVRI reference model and the correspondences between different concepts described by different viewpoints.
2. Define a framework by which existing standards, taxonomies and ontologies can be mapped to the reference model and via the reference model to each other.
3. Provide tool support for defining new mappings between standards, and for searching the semantic space defined by the resulting interlinking.

Thus the purpose of the technology review in ENVRI+ from the linking model perspective is to determine what technologies are available for ontology specification and formal verification, and what technologies exist that could help us to develop new (or adapt existing) tools.

Change history and amendment procedure

The review of this topic will be organised by Paul Martin. He will partition the exploration and gathering of information and collaborate on the analysis and formulation of the initial report. Record details of the major steps in the change history table below. For further details of the complete procedure see item 4 on the Getting Started page.

Note: Do not record editorial / typographical changes. Only record significant changes of content.

Date	Name	Institution	Nature of the information added / changed
21/3/16	Paul Martin	UvA	Initial draft for technology review of T5.1.

Sources of information used

Analysis of state of the art and trends

Combining all environmental domains into one single RI is neither feasible in development nor manageable in operation. During the past several years, interoperability between infrastructures has been extensively studied, with different interoperability solutions proposed for different levels of interoperation: between computing infrastructures (Charalabidis et al. 2012 and Ngan et al. 2011), between middleware (Blair and Grace 2012), and between computational workflows (Zhao et al. 2006). These solutions iteratively build adapters or connectors between two infrastructures and then derive new service standards via focusing community efforts. Such iteration promotes the evolution of services in infrastructures, but cannot fully realize infrastructure interoperability while these solutions only focus on specific layers of the global problem without considering the overall e-science context (Riedel et al. 2009). Meanwhile, White et al. (White et al. 2012) argued the importance of an ontological reference model in the development of interoperable services in infrastructure.

The linking framework for ENVRI+ is being founded on semantic web technologies (Berners-Lee et al. 2001), though the core principles are *techn ology-agnostic*. Key among these technologies is the Resource Description Framework (RDF) that has come to be used as a generic means to describe information implanted in web resources; building upon RDF, the Web Ontology Language (OWL) is a knowledge representation language used to describe ontologies, and is a significant factor in many semantic infrastructure modelling projects (Zhao et al. 2011 and Baldine et al. 2010). Within ENVRI+, the core of the linking framework would be the OIL-E ontologies, which are described in OWL. OWL is well-used in the semantic description domain, but limitations of OWL include the inability to describe integrity constraints or perform closed-world querying (Motik et al. 2006), which might otherwise be useful in (for example) certain well-prescribed areas of the ENVRI reference model. There are also various problems with dealing with diverse schemas, incomplete metadata and the limitations of query interfaces (Gölitz 2007).

The notion of mapping out the topology of standards in environmental science, research practice and e-infrastructure reflects very much the linked open data approach. The linked data approach offers certain advantages, such as ensuring openness, shareability and reusability (Ferris 2014). There is however a lack of good tool support for linked data solutions (Enoksson et al. 2009), which is one of the areas that Task 5.3 is intended to address.

Semantic linking is often investigated in the context of ontology matching, mapping or alignment. The key task is to compare similarity between entities from different semantic models and measure the similarity distances at different layers: the *data* layer, comparing data values and objects; the *ontology* layer, comparing the labels and concepts of entities; and the *context* layer, comparing semantic entities with inclusion of application contexts. We posit that the five viewpoints of the ENVRI reference model are applicable for grouping the different modelling contexts of concern to environmental science research infrastructures.

Different metadata standards have been observed from those RIs that are in operation, including NASA DIF (Miled et al. 2001) and SensorML in EMSO, ISO 19115 (ISO 2014) geospatial metadata in SeaDataNet and ISO 19139 (ISO 2007) geospatial XML in EUROGOOS, and a combination of ISO 19115, INSPIRE and NetCDF-CF based standards in IAGOS (Boulanger et al. 2014). In addition we have observed the use of Dublin Core (ISO 2009), ISO 19156 (ISO 2011), SeaDataNet Cruise Summary Reports metadata, CERIF (Jeffery et al. 2014), and CSMD. These standards can be linked via the information viewpoint of the ENVRI reference model and mapped to functional subsystems of RIs. There is prior work mapping information viewpoint concepts in the reference model to concepts found in those standards (Zhao et al. 2014).

The typical process for semantic linking involves several iterations of the following steps: 1) preprocessing of features by a small set of excerpts of the overall ontology definition to describe a specific entity; 2) definition of the search space in the ontology for candidate alignment; 3) computation of the similarity between two entities from different ontologies; 4) aggregation of the different similarity results of each entity pair, depending on the algorithms used; and 5) derivation of the final linking between entities using different interpretation mechanisms, including the analysis of human experts.

The linking component of OIL-E glues concepts both *inside* ENVRI-RM and *between* ENVRI-RM and external concepts belonging to outside vocabularies. The ENVRI-RM ontology only contains a limited set of vocabularies derived from common functionality and patterns, so linking ENVRI-RM with external RI-specific concepts will enable RI-specific extensions to the ENVRI-RM vocabulary. Similarly, linking ENVRI-RM with external vocabularies provides bridge between those vocabularies and ENVRI-RM, and indirectly between the vocabularies themselves. Notably, the internal correspondences between different ENVRI-RM viewpoints (enterprise, information, etc.) can potentially be used to indirectly link external vocabularies of quite different foci (data, services, infrastructure, etc.).

Distributed applications and systems can be described using published ontologies, permitting services both internal and external to a system to potentially interact with application components without having had to be explicitly designed to do so, provided that they can process the ontology used to describe the component.

There already exists work on doing this kind of semantic modelling of computing and network infrastructure, however the modelling of applications running on cloud platforms is less well-developed—in (Ortiz 2011), the author articulates some of the challenges facing standardisation of cloud technologies, and the lack of concrete formal models is a major factor. Even outwith the cloud however, information models for modern computing infrastructure are often lacking in some dimension. For example modern infrastructure modelling languages must be able to model virtualisation and management of virtualised resources as well as physical resources.

In (Ghijsen et al. 2013), the authors describe the Infrastructure and Network Description Language (INDL), a product of the Open Grid Forum (OGF) Network Markup Language Working Group (NML-WG). INDL is designed to be extensible, linkable to existing information models, and technology independent. NDL-OWL (Baldine et al. 2010) provides a Semantic Web model for networked cloud orchestration modelling network topologies, layers, utilities and technologies. It extends the Network Description Language upon which INDL is based and uses OWL. Meanwhile (Zhao et al. 2010) presented a workflow planning system called NEtwork aware Workflow QoS planner (NEWQoSPlanner) based on INDL; NEWQoSPlanner is able to select network resources in the context of workflow composition and scheduling.

The longer-term horizon

The generation of formal descriptions for complex entities is essential for the mechanisation of processes involving those entities—this is not in question. What is in question is the extent to which different systems can be integrated within common models with shared vocabularies, and to what extent we must accept the existence of proliferation of alternative models, and thus have to expend effort in bridging between the resulting heterogeneous concept spaces.

Relationships with requirements and use cases

The linking model is strong tied to the reference model, which provides its core vocabulary. The linking model should also itself contribute vocabulary and relations that are useful for the interoperable architecture design task.

Regarding use-cases, any of the use-cases might benefit from a linking of formal descriptions, depending on the extent to which the use-cases cross between domains, or make use of formal descriptions that need linking to the reference model concepts. Particular ENVRI+ cases where linking between different existing standards and vocabularies might be useful include (see https://envriplus.manageprojects.com/projects/wp9-service-validation-and-deployment-1/notebooks/625/pages/324):

- Identifying **trends in the emergence of mosquito born diseases** requires interaction between a number of different data centres and compute providers.
- The **description of a national biodiversity data archive centre** requires a formal model for how data from a national facility is to be delivered to and integrated with Europe-wide data providers.
- **Domain extension of existing thesauri**.

Summary of analysis highlighting implications and issues

The question that underlies the semantic linking task is: how do we make it easier to map between different vocabularies? Autonomous mapping processes are highly error prone, and extremely sensitive to the quality of the underlying taxonomies or ontologies. Manual mapping requires expert oversight, but can be supported by tools.

The base contribution of a linking model in the environmental science research infrastructure domain is the ability to map out the space of existing standards, models and vocabularies being used in different datasets, architecture designs, instrument specifications, service profiles, etc. used by different research communities, and the ability to associate them via the viewpoints of the ENVRI reference model or its successors. This in and of itself would constitute a useful contribution, since as it stands it requires substantial research to truly understand the full current research landscape, and even experts' views are often narrow, focused on a particular domain or a particular geographic region (i.e. the standards produced within their home continent).

Bibliography and references to sources

Ilia Baldine, Yufeng Xin, Anirban Mandal, Chris Heermann Renci, Jeff Chase, Varun Marupadi, Aydan Yumerefendi, and David Irwin. 2010. Networked cloud orchestration: a GENI perspective. In GLOBECOM Workshops (GC Wkshps), 2010 IEEE, pp. 573-578. IEEE.

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. The semantic web. Scientific american 284, no. 5, 28-37.

Blair, Gordon, and Paul Grace. "Emergent middleware: Tackling the interoperability problem." *IEEE Internet Computing* 1 (2012): 78-82.

Boulanger, Damien, Benoit Gautron, Valérie Thouret, Martin Schultz, Peter van Velthoven, Bjoern Broetz, Armin Rauthe-Schöch, and Guillaume Brissebrat. "Latest developments for the IAGOS database: Interoperability and metadata." In *EGU General Assembly Conference Abstracts*, vol. 16, p. 6510. 2014.

Charalabidis, Yannis, Marijn Janssen, and Olivier Glassey. "Introduction to cloud infrastructures and interoperability minitrack." In *2012 45th Hawaii International Conference on System Sciences*, p. 2177. IEEE, 2012.

Enoksson, Fredrik, Matthias Palmér, and Ambjörn Naeve. "An RDF modification protocol, based on the needs of editing Tools." *Metadata and Semantics*. Springer US, 2009. 191- 199.

Ferris, Virginia. “Beyond “Showing What We Have”: Exploring Linked Data for Archival Description”, School of Information and Library Science of the University of North Carolina at Chapel Hill, 2014.

Ghijsen, Mattijs, Jeroen Van Der Ham, Paola Grosso, Cosmin Dumitru, Hao Zhu, Zhiming Zhao, and Cees De Laat. 2013. A semantic-web approach for modeling computing infrastructures. Computers & Electrical Engineering 39, no. 8, 2553-2565.

Gölitz, Olaf. “Distributed Query Processing for Federated RDF Data Management”, PhD thesis, Universitat Koblenz-Landau, 2007.

ISO. 2007. Geographic information—Metadata—XML schema implementation. ISO 19139:2007.

ISO. 2009. Information and documentation—The Dublin Core metadata element set. ISO 15836:2009.

ISO. 2011. Geographic information—Observations and measurements. ISO 19156:2011.

ISO. 2014. Geographic information—Metadata. ISO 19115:2014.

Jeffery, K., Houssos, N., Jörg, B., & Asserson, A. (2014). Research information management: the CERIF approach. *International Journal of Metadata, Semantics and Ontologies*, *9*(1), 5-14.

Miled, Zina Ben, Srinivasan Sikkupparbathyam, Omran Bukhres, Kishan Nagendra, Eric Lynch, Marcelo Areal, Lola Olsen et al. "Global change master directory: object-oriented active asynchronous transaction management in a federated environment using data agents." In *Proceedings of the 2001 ACM symposium on Applied computing*, pp. 207-214. ACM, 2001.

Motik, B, I. Horrocks, R. Rosati, and U. Sattler, “Can OWL and Logic Programming Live Together Happily Ever After?”, Proceedings 5th International Semantic Web Conference, 2006.

Ngan, Le Duy, Yuzhang Feng, Seungmin Rho, and Rajaraman Kanagasabai. "Enabling interoperability across heterogeneous semantic web services with OWL-S based mediation." In *Services Computing Conference (APSCC), 2011 IEEE Asia-Pacific*, pp. 471-476. IEEE, 2011.

Ortiz, Sixto. 2011. The problem with cloud-computing standardization. Computer 7, 13-16.

Riedel, Morris, Erwin Laure, Th Soddemann, Laurence Field, John-Paul Navarro, James Casey, Maarten Litmaath et al. "Interoperation of world wide production eScience infrastructures." *Concurrency and Computation: Practice and Experience* 21, no. 8 (2009): 961-990.

Tawfiq Khalil, Ching-Seh (Mike) Wu, “Link Patterns in the World Wide Web”, International Journal of Information Technology & Management Information System (IJITMIS), Volume 4, Issue 3, 2013.

White, Laura, Norman Wilde, Thomas Reichherzer, Eman El-Sheikh, George Goehring, Arthur Baskin, Ben Hartmann, and Mircea Manea. "Understanding interoperable systems: Challenges for the maintenance of SOA applications." In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pp. 2199-2206. IEEE, 2012.

Zhao, Zhiming, Suresh Booms, Adam Belloum, Cees de Laat, and Bob Hertzberger. "Vle-wfbus: a scientific workflow bus for multi e-science domains." In *e-Science and Grid Computing, 2006. e-Science'06. Second IEEE International Conference on*, pp. 11-11. IEEE, 2006.

Zhao, Zhiming, Paola Grosso, Ralph Koning, Jeroen Van Der Ham, and Cees De Laat. 2010. An agent based planner for including network QoS in scientific workflows. In Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on, pp. 231-238. IEEE.

Zhao, Zhiming, Paola Grosso, Jeroen van der Ham, Ralph Koning, and Cees de Laat. 2011. An agent based network resource planner for workflow applications. Multiagent and Grid Systems 7, no. 6, 187-202.

Zhao, Z, P. Grosso, C. de Laat, B. Magagna, H. Schentz, Y. Chen, A. Hardisty, P. Martin, and M. Atkinson. (2014) Interoperability framework for linked computational, network and storage infrastructures, version 2. Accessed: 2015-07-21. [Online]. Available: http://envri.eu/.