TOWARDS SEMANTIC APIS FOR RESEARCH DATA SERVICES

by Anna Fensel

Abstract: *Rapid development of Internet and Web technology is changing the state of the art in communication of knowledge, or results of research activities. Semantic technology as well as linked and open data in particular are becoming key enablers for successful and efficient progress in research. At first, I define the research data service (RDS) and discuss typical current and possible future usage scenarios involving RDSs. Furthermore, I discuss the state of the art in the areas of semantic service and data annotation and API construction, as well as infrastructural solutions applicable for RDS realization. Finally, innovative methods of online dissemination, promotion and efficient communication of research are discussed.*

Keywords: *Research Data Service (RDS); Research Data; RDS Metadata; Web API; Semantic Web Service; Semantic Technology; RDS Publication; Research Dissemination*

AUF DEM WEG ZU SEMANTISCHEN APIS FÜR FORSCHUNGSDATENDIENSTE

Zusammenfassung: *Die schnelle Entwicklung der Internet- und Web-Technologie verändert den Stand der Technik in der Kommunikation von Wissen oder Forschungsergebnissen. Insbesondere semantische Technologien sowie verknüpfte und offene Daten werden zu entscheidenden Faktoren für einen erfolgreichen und effizienten Forschungsfortschritt. Zuerst definiere ich den Research Data Service (RDS) und diskutiere typische aktuelle und mögliche zukünftige Nutzungsszenarien dafür. Darüber hinaus bespreche ich den Stand der Technik in den Bereichen semantische Dienstleistung, Datenannotation und API-Konstruktion sowie infrastrukturelle Lösungen, die für die RDS-Realisierung anwendbar sind. Zum Schluss werden noch innovative Methoden der Online-Verbreitung, Förderung und effizienten Kommunikation der Forschung diskutiert.*

Schlüsselwörter: *Forschungsdatendienste (RDS); Forschungsdaten; RDS Metadaten; Web API; Semantic Web Service; Semantische Technologie; RDS Publikation; Verbreitung von Forschung*

Contents

1. Introduction to and Motivation for Research Data Services (RDSs)

Acceptance of the open science principles[1] entails open access not only to research data, but also to tools that allow researchers to perform various types of activities over these data including mining, visualization, and analysis. The data and tools can be called Research Data Services (RDSs), enabling researchers to conduct their research activities efficiently and effectively.

One of the challenges faced by researchers in a globally networked scientific world is to be able to locate RDSs that fulfil their research needs. RDSs should be discoverable, i.e. have a feature at the semantic service description level that enables automatically locating research data services that fulfil a researcher goal. Making a RDS discoverable enables service (re-)use. Research data and service infrastructures are becoming increasingly interlinked, and semantic modeling and linked data are playing an important instrumental role in this process (Thanos, 2016).

Essentially, RDSs should have the following characteristics:

- they are subclasses of Services in a general sense (have a service provider and a service consumer, added value, ...),
- they are data services, part of a data economy,
- they are applicable in scenarios implementing some part of the research process,
- they may be delivered by a program/IT system, but also via other means, e.g. a human.

Wikipedia defines "Research" as "creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of humans, culture and society, and the use of this stock of knowledge to devise new applications." Therefore, RDSs modeled as semantic Application Programming Interfaces (APIs) can increase efficiency of research in a broad sense, including the following tasks in particular:

- discovering new knowledge facilitating the process of research,
- delegation and contracting of research tasks to humans (e.g. in crowdsourcing, interdisciplinary research, etc.) as well as machines,
- combination of data and information coming from heterogeneous sources,
- seamless benchmarking and integration of stand-alone research efforts,
- etc.

According to the RDA Europe project (Thanos and Candela, 2016), a "research data service is a rule of correspondence between two sets", or "a Concrete Research Data Tool on which there exists an Institutional Commitment in the form of a Service-Level Agreement".

Technically, RDSs can be based on Web services, as the latter implement a service-oriented architecture in a specific manner, and are essentially a programmatic layer on top of distributed systems. Therefore, RDSs, as defined here earlier:

- May or may not be implemented as Web services.
- In any case have specific characteristics related to research.

This paper is organized as follows: Section 2 discusses the modeling aspects of RDSs, the details of RDSs' technical management and programming in Section 3, and innovative dissemination techniques for RDSs in Section 4. Section 5 concludes the paper.

2. Modeling RDSs

As a typical Web service, an RDS API would have the following types of properties:

- ***Functional*** (contain the formal specification of ***what*** exactly the service can do.),
- ***Behavioral*** (describe the functionality of the service that can be achieved in terms of interaction with the service and in terms of functionality required from other Web services.), and
- ***Non-functional properties*** (capture constraints over the previously mentioned properties.)

Similarly, the tasks one can perform with RDSs would, as in the world of Web services, be as follows (Fensel et al., 2011; Cardoso et al., 2014):

- Discovery: "Find services that match the service requester specifications".
- Selection and Ranking: "Choose the most appropriate services among the available ones".
- Composition: "Assembly of services based in order to achieve a given goal and provide a higher order of functionality".
- Mediation: "Solve mismatches among domain knowledge used to describe the services, protocols used in the communication, data exchanged in the interaction (types used, and meaning of the information) and business models of the different parties".
- Execution: "Invocation of a concrete set of services, arranged in a particular way following programmatic conventions that realize a given task".
- Monitoring: "Supervision of the correct execution of services and dealing with exceptions thrown by composed services or the composition workflow itself".
- Handover: "Replacement of services by equivalent ones, which solely or in combination can realize the same functionality as the replaced one, in case of failure while execution".

As to any linked data services, the semantic web and linked data principles, e.g. on linked data publishing in particular (Heath and Bizer, 2011), would be applicable to RDSs. Scientists and librarians, who comprise the relevant communities here, are de facto more on the early adopter side of semantic web and linked data principles. In the bibliographic domain, semantic mark-up complying with specialized vocabularies such as Dublin Core has already been in use for many years, even decades (Klee, 2013). Many research efforts around using linked data for open science, such as in distribution of educational and research content – especially the adoption of the practices –, are fostered by educational and research institutions, universities in particular (Mouromtsev & d'Aquin, 2016). Semantic formalism has also been applied for decades in some fields, e.g. the life sciences[2]. Another relevant development here is schema.org. Schema.org provides a collection of vocabularies for sharing information relevant in the context of the Web. It was launched in June 2011 by Bing, Google and Yahoo!, further joined by Yandex in November of the same year. Its purpose is to create a common set of schemas for webmasters to mark-up their websites with structured data. It has proven to be a very large success: eventually everything that can be consumed or booked through the Web is semantically annotated with schema.org – there are over 4.8 million schema.org

annotations describing hotels available on the Web, for example (Kärle et al., 2016). However, as with any linked data, data quality is and remains a large issue that needs to be addressed in the near future (Zaveri et al., 2016; Kärle et al., 2016).

The advantages of schema.org are as follows:

- Webmasters can use schema.org to mark up their web pages (creating enriched snippets) in a way that is recognized by major search engines.
- The enriched snippets enable search engines to understand the information on web pages that results in richer and more attractive search results for the users. Hence it is easier for users to find relevant and right information on the web.
- Search engines including Bing, Google, Yahoo! and Yandex rely on this markup to improve the display of search results.
- It helps webmasters achieve higher rankings of their pages in search results.
- This markup has the potential to enhance the CTR (click through ratio) from the search results from anywhere between 10–25%[3].
- Schema.org can be also used for structured data interoperability.
- Its usage can also lead to the development of new tools, for example Google Recipe Search, which may open up other marketing channels if not now, then in the near future.
- Query/Answer based search engines will be improved, particularly in terms of semantic search, by making use of structured data, i.e. the search engine can understand the content of a website and make use of it to give a direct and accurate search result.

In conclusion, schema.org is also obviously relevant for RDSs, as research is being done on a variety of objects, most of these existing in real life and already annotated with schema.org.

Metadata needed for the description of the RDS should contain an "input set" (domain) and the "output set" (co-domain). Whereas the service itself is typically a process, which is not really possible to describe formally. Examples of details that can be annotated as input and output of the service are shown in Figure 1.

Ideally, we need to specify the syntax and semantics of the elements of the domain and co-domain. Establishment and spread of such specifications is also a realistic development path as the amount of research data services increase and the research steering service economy is becoming more interdisciplinary and it becomes more difficult to identify and find relevant services.

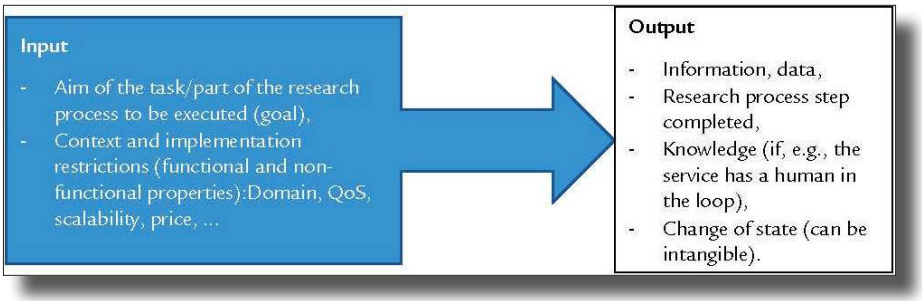

Fig. 1: Inputs and outputs of RDSs

In the real world, however, simpler things and models spread better. For example, despite many developments in semantic web services, the largest service registries, like ProgrammableWeb[4], still do not have semantic descriptions for service APIs. Adding these clearly constitutes progressive usage potential, and is also a likely to happen here first since the research community tends to be an early adopter of new technology.

An RDS instance definitely has the dynamicity aspect as well, and changes in time in the following ways:

- Its quality may alter, its non-functional properties change,
- Its context and usage may vary,
- It may appear and disappear,
- Its implementation may change, etc.

Modeling such state changes programmatically is difficult, and making them widely used is even more difficult. Therefore, RDSs should be designed as stateless.

In order to appropriately describe its functionality, the RDS profile description should include:

- Aimed dataset or service,
- Scientific discipline,
- Scientific method,
- Domain,
- Information about quantity, quality, availability, creator(s)/provider(s),
- Access and license policies,
- Origin and annotation of reused/subcontracted sources (if applicable).

Some or even all of the above attributes can be optional. Examples for a service can include:

- "compare performance of my semantic repository according to criteria X",
- "find datasets with energy consumption of fridges in Vienna".

Scientific workflows are appropriate for describing the process model of a RDS, for example in conjunction with research method description. Existing models, such as BPEL[5], USDL[6] and Linked USDL (Cardoso et al., 2014) can be used to interlink business models, service systems, service models, service instances and service descriptions (Cardoso et al., 2014).

3. Managing and programming RDSs

There are existing service description frameworks like OWL-S[7] and WSMO (Fensel et al., 2011), but in most practical cases they are too complex, except where the related technologies are already being built on RDF infrastructures. In genetics, for example, some data annotations are made using OWL, and building a service on top of it would be a natural extension. But in many cases it would not be the best choice, as most data are not shared on the Web with OWL, and its semantic and syntactic complexity and expressivity are higher than necessary for developers.

Using real data in research is essential, so the frameworks containing the real data are the most important and should therefore comprise and rely on linked (open) data and schema.org annotations. For the modelling level, a feasible approach could be applying the Linked Service System (LSS) model structure, defining a human-friendly way to model a service by representing key aspects of the service by answering the essential questions about it (Who/Role, Why/Goal, What/Resource, How/Process, When/Time, Where/Location), and linking to data represented in existing formalisms like USDL and Linked USDL (Cardoso et al., 2014).

It is also worth noting that schema.org has Actions as a part of its model, which makes modeling typical actions relevant in the Web context possible. Using it of course also ensures direct compatibility with the multitude of real data available in schema.org.

Given that many service representation languages are still evolving at present, it is important to note the main characteristics and capabilities of a knowledge representation language appropriate for the description of the functionality of a data service as well as for effectively supporting reasoning in the matchmaking process. These include the following parts:

- Presence of the ways to annotate the functionality, domain, ...
- Assumption of "incorrectness and incompleteness", or the need to combine heterogeneous techniques in reasoning (Fensel & Van Harmelen, 2007) – given that working with RDS would have an open world assumption as well as any reasoning on the Web, the principles of reasoning as performed before in "closed" infrastructures (databases) would become irrelevant.
- Much of matchmaking and reasoning should be moved to the applications – but: semantics can support maintenance of community-generated reusable mapping (e.g. stating that two service parts are the same).

Discipline-specific classification of data services (classes of data services) supported by discipline-specific ontologies would be considered state of the art. We particularly need it, because:

- Data sets vary from domain to domain, and often the research is domain-specific.
- It facilitates discoverability.

This would also be in agreement with other trends in the services area, e.g. microservices they have a very narrow focus – typically very domain-specific ones, but with semantic APIs that enable easy integration into more complex services that, in turn, can be modified, re-created and re-published by the developers as well as the end users (Davies et al., 2011).

The role of registries, directories, and catalogs of services – or essentially, infrastructures, is very important as they provide a) a single point to make services discoverable, b) meta-Research Data Services in themselves, as a collection of services. Ideologically they may be constructed architecturally similarly to UDDI from the past, or like ProgrammableWeb for Web APIs now.

Once semantically annotated, RDSs can be found by various features e.g. "stateless/state-based", "type of input data: discrete data/vectors/functions/streaming data", "types of output data", etc. – by data, by domain, by functionality, ... The classifications do not have to be created a priori, but could be created ad hoc once the annotations are there in order to avoid restricting usage. De facto, RDSs are classified by provider platforms, e.g. in the areas the platforms are operating in mainly: like Linked Open Vocabularies (LOV) or datahub.io for structured data, or in the library domain, where numerous repositories for publications are provided by various publishers, such as Zenodo or Google Scholar.

Last but not least, being able to identify and reference RDSs is essential, and "citation" is instrumental in making research data services discoverable. Efforts in bridging such data to semantic formats are already ongoing, e.g. opening up bibliographic data (Peroni et al., 2015). Of course, eventually identification and citation data will become a necessary accompanying part of research data and RDSs as well.

Citation information will be used in processes, e.g. in ranking. As in other cases where citation numbers influence ranking (e.g. in Google Scholar– most cited publications are displayed on top), this is not optimal, as searches according to these criteria may overlook data with closer matches, and the output of such ranking processes itself also impacts citations as the items that are ranked higher are more visible and cited more. To ameliorate this situation, several ways for choosing semantic annotations have been suggested. These include taking into consideration the provenance specifics (e.g. author reputation) of newly appearing annotations (Stavrakantonakis et al., 2016) as well as inclusion of end-user perspectives in the ranking process (Klan & König-Ries, 2014).

4. Disseminating RDSs

Nowadays, research is disseminated not only via classical channels like digital libraries, but also in a multi-channel manner, e.g. via social media, collaborative infrastructures, and a multitude of other diverse communication and dissemination channels (Fensel et al., 2014). Arguably, social media are used by younger generations more commonly than email, and social media networks like Facebook, LinkedIn, Xing, Twitter and Google+ have become a mainstream mode of communication in general. Channels focusing on research communication exist as well, such as ResearchGate, Academia.edu, Google Scholar and SlideShare, to name a few. They are increasingly used by researchers, and research can be directly followed by interested parties, e.g. by following accounts of relevant groups and relevant researchers.

There are even solutions to make the multi-channel communication of research more efficient and less time-consuming. For example, ONLIM[8] is an online tool based on semantic technologies that aims to facilitate managing various social media platforms by means of publishing posts and tracking the feedback given by other users. ONLIM supports several social media platforms such as Facebook, Twitter, Youtube, LinkedIn, Xing and Flickr. It also allows users to schedule their posts to enable more effec-

tive social media management and marketing. In contrast to other similar tools, ONLIM also has an automatic post generation feature that creates posts for publication in social media from external sources. RDSs that are annotated with schema.org can be converted to social media posts by ONLIM automatically, for example. In addition, these posts can be automatically forwarded to all appropriate channels to disseminate the research. While they are already being used in mainstream marketing (Fensel et al., 2014; Fensel et al., 2016), such developments are becoming a game changer in RDS discoverability and in how research is being disseminated and accessed.

With content and data, there are a few particularities hampering their potential (re-)use in the research data value chain. Data licensing in particular is still complicated, and semantic formats for licensing data, as well as the tools processing data license annotations are under-defined or nonexistent. Semantic standards for licenses are being developed right now, and include efforts such as ODRL[9] and RightsML (Ermilov and Pellegrini, 2015). Another example of an ongoing Austrian project focusing on semantic languages and tools for data licensing is DALICC[10].

5. Conclusion

RDSs are data services specifically addressing the production processes needed for research and education, and as such belong to the data service value chain. In the long run, they would be accessible via and manageable with semantically annotated APIs, over a network of RDS repositories and e-Infrastructures.

Promising semantic languages and technologies exist that can be applied to solutions of RDS modeling and discovery problems, e.g. linked services, linked data and schema.org. These solutions would facilitate the use of semantic data already available in terms of research, discoverability and applicability.

Efficient dissemination of research is very important. Dissemination also needs to be multi-channel now, and new kinds of channels appear, e.g. social media and collaborative environments. Eventually, the Web of RDSs would be accessible over a semantic API, so that the most relevant RDSs can be delivered or activated by the user merely as a result of a query. Such functionality, as well as RDS delivery and activation across different communication and dissemination channels, would facilitate conduction of interdisciplinary research.

Relevant data value chain languages, techniques and tools are in development, e.g. on (semantic) data licensing. These solutions would facilitate implementation of the data value chain in research. In particular, data contributions from professional researchers as well as unprofessional researchers (e.g. data generated via crowdsourcing) would be clearly annotated in terms of rights, permissions and obligations associated with its usage.

Ass. Prof.in Dr.in Anna Fensel
ORCID: http://orcid.org/0000-0002-1391-7104
University of Innsbruck, Semantic Technology Institute (STI)
E-Mail: anna.fensel@sti2.at

References

Cardoso, J., Lopes, R., & Poels, G. (2014). Service systems: concepts, modeling, and programming (pp. 1–91). Springer. DOI: https://dx.doi.org/10.1007/978-3-319-10813-1.

Davies, M., Carrez, F., Heinilä, J., Fensel, A., Narganes, M., & Carlos dos Santos Danado, J. (2011). m: Ciudad: enabling end-user mobile service creation. International Journal of Pervasive Computing and Communications, 7(4), pp. 384–414. DOI: https://doi.org/10.1108/17427371111189683.

Ermilov, I., & Pellegrini, T. (2015, September). Data licensing on the cloud: empirical insights and implications for linked data. In Proceedings of the 11th International Conference on Semantic Systems (pp. 153–156). ACM. DOI: https://doi.org/10.1145/2814864.2814878.

Fensel, A., Akbar, Z., Toma, I., & Fensel, D. (2016). Bringing Online Visibility to Hotels with Schema.org and Multi-channel Communication. In Information and Communication Technologies in Tourism 2016 (pp. 3–16). Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-28231-2_1.

Fensel, A., Toma, I., García, J. M., Stavrakantonakis, I., & Fensel, D. (2014). Enabling customers engagement and collaboration for small and medium-sized enterprises in ubiquitous multi-channel ecosystems. Computers in Industry, 65(5), pp. 891–904. DOI: https://doi.org/10.1016/j.compind.2014.02.001.

Fensel, D., Facca, F. M., Simperl, E., & Toma, I. (2011). Semantic web services. Springer Science & Business Media. DOI: https://doi.org/10.1007/978-3-642-19193-0.

Fensel, D., & van Harmelen, F. (2007). Unifying Reasoning and Search to Web Scale. IEEE Internet Computing, 2(11), pp. 96–95. DOI: https://doi.org/10.1109/MIC.2007.51.

Heath, T., & Bizer, C. (2011). Linked data: Evolving the web into a global data space. Synthesis lectures on the semantic web: theory and technology, 1(1), pp. 1–136. DOI: https://doi.org/10.2200/S00334ED1V01Y201102WBE001.

Kärle, E., Fensel, A., Toma, I., & Fensel, D. (2016). Why Are There More Hotels in Tyrol than in Austria? Analyzing Schema.org Usage in the Hotel Domain. In Information and Communication Technologies in Tourism 2016 (pp. 99–112). Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-28231-2_8.

Klan, F., & König-Ries, B. (2014, June). A user-centered methodology for the evaluation of (semantic) web service discovery and selection. In Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14). ACM. DOI: https://doi.org/10.1145/2611040.2611069.

Klee, C. (2013). Vokabulare für bibliographische Daten. (Open) Linked Data in Bibliotheken, S. 45–63. DOI: https://doi.org/10.1515/9783110278736.45.

Mouromtsev, D., & d'Aquin, M. (Eds.). (2016). Open Data for Education: Linked, Shared, and Reusable Data for Teaching and Learning (Vol. 9500). Springer. DOI: https://doi.org/10.1007/978-3-319-30493-9.

Peroni, S., Dutton, A., Gray, T., Shotton, D. (2015). Setting our bibliographic references free: towards open citation data. Journal of Documentation, 71 (2): pp. 253–277. DOI: https://doi.org/10.1108/JD-12-2013-0166.

Stavrakantonakis, I., Fensel, A., & Fensel, D. (2016, September). Linked Open Vocabulary ranking and terms discovery. In Proceedings of the 12th International Conference on Semantic Systems (pp. 1–8). ACM. DOI: https://doi.org/10.1145/2993318.2993338.

Thanos, C. (2016). A Vision for Open Cyber-Scholarly Infrastructures. Publications, 4(2), 13. DOI: https://doi.org/10.3390/publications4020013.

Thanos, C., Klan, F., Kritikos, K., & Candela, L. (2016). White Paper on Research Data Service Discoverability. Publications, 5(1), 1. DOI: https://doi.org/10.3390/publications5010001.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for linked data: A survey. Semantic Web, 7(1), pp. 63–93. DOI: https://doi.org/10.3233/SW-150175.

Zhdanova, A. V. (2008, May). Community-driven ontology evolution: Gene ontology case study. In International Conference on Business

Information Systems (pp. 118-129). Springer Berlin Heidelberg. DOI: https://doi.org/10.1007/978-3-540-79396-0_11.

1 Open Science – Digital Single Market: https://ec.europa.eu/digital-single-market/en/open-science.
2 See, for example, an analysis of the community-driven development of Gene Ontology (Zhdanova, A. V, 2008).
3 See, for example, an analysis of a hotel web presence annotated with schema.org (Fensel et al., 2016).
4 ProgrammableWeb – APIs, Mashups and the Web as Platform: http://www.programmableweb.com.
5 BPEL: https://www.oasis-open.org/committees/wsbpel/.
6 USDL: https://www.w3.org/2005/Incubator/usdl/XGR-usdl-20111027/.
7 OWL-S: https://www.w3.org/Submission/OWL-S/.
8 ONLIM – Social Media Management Tool: www.onlim.com.
9 ODRL – Open Digital Rights Language: https://www.w3.org/ns/odrl/2/.
10 DALICC – Data Licenses Clearance Center: https://dalicc.net.