Bring Your Own Data workshops: a mechanism to aid data owners to comply with Linked Data best practices

Marco Roos[1,*], Alasdair J. G. Gray[2,*], Andra Waagmeester[3], Mark Thompson[1],
Rajaram Kaliyaperumal[1], Eelke van der Horst[1], Barend Mons[1,4],
Mark D. Wilkinson[5]

1) Leiden University Medical Center, Leiden, The Netherlands, m.roos@lumc.nl
2) Heriot-Watt University, Edinburgh, The United Kingdom, A.J.G.Gray@hw.ac.uk
3) Ontoforce, Antwerp, Belgium
4) Dutch Techcentre for Life Science, The Netherlands
5) Universidad Politécnica de Madrid, Spain
*) corresponding author

Abstract. In the context of stimulating data owners to make data 'FAIR' (Findable, Accessible, Interoperable, Reusable), we have developed the concept of the 'Bring Your Own Data workshop' (BYOD) for FAIR data. In a BYOD that focuses particularly on a Linked Data (LD) approach, engineers or bioinformaticians who are experts on their own data sources work with LD experts to generate Linked Data on a preselected set of data from their own resource. We report here on the organisation of a BYOD and on our experience from two BYODs, one that was held June 24-25, 2014 in Leiden, the Netherlands, and one that will be held in Rome, Italy, on November 26-27, 2014[1], following the EpiRare workshop[2].

1   Introduction

The aim of a LD BYOD workshop is to guide the participants in transforming their own data to Linked Data by selecting a set of example cross-domain questions, and by doing so to allow rapid integration with other Linked Data resources. To drive this process, and as a secondary goal, Data Owners and LD experts collaboratively make selected data interoperable for a demonstration. At the end of a BYOD, Data Owners should understand the basic principles of making data interoperable, such that they can begin to follow

---

[1] http://rd-connect.eu/BYOD-1
[2] http://www.epirare.eu/

this strategy. Data owners and LD experts should have created a minimal resource from selected data to answer a driving question.

To achieve our aims we organise two days that are a combination of a hackathon and a hands-on tutorial. In advance, at least two webinars are organised to introduce the basic principles to data owners, and to start the discussion on driving user questions as the starting point for a BYOD. In our experience, the desired ratio between Data Owners and LD experts is 1:1. Ideally, the Data Owners are represented by local engineers and one or two domain experts to aid with the use cases.

The basic workflow that we follow in a BYOD is as follows: (i) define a driving user question and demonstrator, (ii) find appropriate ontologies and ontology concepts, (iii) find appropriate source data (data owner data + existing Linked Data), (iv) create Linked Data, (v) create a Linked Data Cache, (vi) perform a SPARQL query that represents the user question. In the first BYOD, driving user questions were chosen on the first day. In the second BYOD defining driving questions and selecting data starts with the preparatory webinars. Participants share information in a Google Document.

The first BYOD brought together six engineers of the Human Protein Atlas (HPA) and Mycobank and nine LD experts, including two facilitators. HPA is a rich resource of millions of images showing the spatial distribution of proteins in human tissues, cancer types, and human cell lines (http://www.proteinatlas.org/). Mycobank is an on-line database that serves the mycology community by documenting novel mycological nomenclature and associated data (http://www.mycobank.org/). The second BYOD will bring together owners of Rare Disease patient registries and biobanks, and LD experts. The experience of the first BYOD has been used to improve the second BYOD.

2   Experience

The first day of the workshop was mainly devoted to introducing the ideas of publishing data as interoperable RDF and understanding the datasets represented by the data providers. Before the afternoon was over we had split into two teams to work up the showcase studies using either Mycobank or HPA and the experts with knowledge of related datasets.

The next day started with the teams feverishly working up their ideas. There was a general buzz around the room with the experts calling on each other's knowledge across the teams to bring together working demonstrations. The day closed with a show and tell of what had been accomplished.

The MycoBase showcase focused on discovering which compounds fermented. About 10,000 fungal strains in MycoBase were represented in RDF resulting in 2.5 million triples. These were linked to the ChEMBL database by exploiting the Open PHACTS Discovery Platform API to resolve chemical names present in MycoBase to their ChEMBL URI. This formed the key linkage to integrate the two data sets, pulling in key facts (e.g. molecular weight, log p value and hydrogen bound count) from the ChEMBL database.

The HPA team worked up two possible showcases. The first involved discovering for a given HPA protein the pathways, sourced from wikipathways, in which the protein occurs. The second involved linking with the genes present in FANTOM5 and included a resolution step involving the Bio2RDF version of Entrez Gene. These connections were possible due to the lengthy modelling discussions and the development of an RDF generating script that converted part of the HPA relational database into an RDF representation.

3   Conclusion

Overall the workshop was a great success. The data providers felt they had learnt about RDF and were happy with the progress that had been made. While it was recognised that modelling the data in RDF was hard, the interoperability possibilities were a great incentive. The trainers were pleased with the ad hoc training approach, although they had some key suggestions for training material for the next BYOD workshop. The facilitators also played a key role in ensuring that there was an appropriate amount of tutorial time. Both teams left vowing to continue working up their showcases to completion and aiming to produce a paper about their work.

In our experience, we addressed two bottlenecks in the process to create Linked Data 'at the source': (i) driving user questions that aid LD experts to selecting data and choosing appropriate ontologies, (ii) choosing ontologies and ontology concepts by data owners. A secondary bottleneck is choosing the most appropriate conversion tools and Linked Data store. We find that although there is a wealth of ontologies available in the biomedical domain,

it is virtually impossible for novices in Linked Data to make informed decisions. Through BYODs we are beginning to identify the consensus practice that most LD experts follow. This can form the basis of defining clear practical guidelines for novices in Linked Data. At this time, in our opinion, creating Linked Data requires tight collaboration between Linked Data/Ontology experts and domain experts.

The key measure of success for the workshop will be if the data providers are now able to find their way into the world of semantic data publishing without further workshops. A positive sign is that HPA engineers recently included a beta release of RDF-based nanopublications for part of their data[3].

4   Acknowledgements

---

[3]   http://www.proteinatlas.org/about/download