Scientific Lenses over Linked Data: An approach to support task specific views of the data. A vision.

Christian Brenninkmeijer[1], Chris Evelo[2], Carole Goble[1], Alasdair J G Gray[1], Paul Groth[3], Steve Pettifer[1], Robert Stevens[1], Antony J Williams[4], and Egon L Willighagen[2]

[1] School of Computer Science, University of Manchester, UK.
[2] Department of Bioinformatics - BiGCaT, Maastricht University, The Netherlands.
[3] Department of Computer Science, VU University of Amsterdam, The Netherlands.
[4] Royal Society of Chemistry, ChemSpider, USA

Abstract. Within complex scientific domains such as pharmacology, operational equivalence between two concepts is often context-, user- and task-specific. Existing Linked Data integration procedures and equivalence services do not take the context and task of the user into account. We present a vision for enabling users to control the notion of operational equivalence by applying *scientific lenses* over Linked Data. The scientific lenses vary the links that are activated between the datasets which affects the data returned to the user.

1 Introduction

Developing an integrated view over pharmacological data is challenging due to the complexity of the domain. For example, searches for the chemical *"Fluvastatin"* on ChemSpider[5] and DrugBank[6] return different compounds: although their basic chemical structure matches, the compounds differ in their stereochemistry[7]. Another challenge is that the domain scientists have differing opinions on when data records can be related. For example, when searching for information about *"Protein Kinase C Alpha"*, the scientist may want information returned for that protein as it exists in humans[8], mice[9] or both. Additionally, when connecting across databases one may want to treat these proteins as equal in order to bring back all possible information whereas in other cases (e.g. when the researcher is focused on humans) only information on the particular protein should be retrieved.

Data integration in Linked Data relies on equality links between resources across different datasets. Relevant existing approaches include Bio2RDF [2],

[5] `http://www.chemspider.com/393587` accessed 7 Sept 2012.
[6] `http://www.drugbank.ca/drugs/DB01095` accessed 7 Sept 2012.
[7] For details see `http://www.chemconnector.com/2012/07/29/` accessed 7 Sept 2012.
[8] `http://www.uniprot.org/uniprot/P17252` accessed 7 Sept 2012
[9] `http://www.uniprot.org/uniprot/P20444` accessed 7 Sept 2012

Chem2Bio2RDF [4], and LODD [9]. These present a single "fixed" view over the data which is composed of the links between the data resources. Applications are supported in discovering the links by identity mapping services e.g. sameas.org [5], Identifiers.org [8], and BridgeDB [7]. These identity mapping services do not consider the "meaning" of the link, typically treating everything as being `owl:sameAs` equivalent. The semantics of such links between datasets are often not trivial: as Halpin et al. have shown sameAs, is not always sameAs [6]. Operational equality is often context-specific, especially in the complex domains considered in science where opinion can depend on the interpretation applied to results. Current linked pharmacology datasets and identity mapping services do not consider the context, task and roles of users. However, it is important to have the flexibility of applying operational equality on a per task basis.

We aim to support users in controlling and varying their view of the data by applying a *scientific lens* which govern the notions of equivalence applied to the data. Users will be able to change their lens based on the task and role they are performing rather than having one fixed lens. To support this requirement, we propose an approach that applies context dependent sets of equality links. These links are stored in a stand-off fashion so that they are not intermingled with the datasets. This allows for multiple, context-dependent, linksets that can evolve without impact on the underlying datasets and support differing opinions on the relationships between data instances. This flexibility is in contrast to both Linked Data and traditional data integration approaches. We look at the role personae can play in guiding the nature of relationships between the data resources and the desired affects of applying scientific lenses over Linked Data.

2 Scientific Personae

We present two personae representative of the scientific users of the data integration system being developed in the Open PHACTS project[10] [10].

Chris the Integrative Systems Biologist. Chris is a researcher in a university who investigates processes in mice and compares these with processes in humans. When searching for data he needs to be able to switch between species depending on the kind of experimental (protein and mRNA) data he wants to compare in pathways, i.e. sometimes he will be interested in just mice data, sometimes just human, and at other times the combination of both. However, he always insists that genes and proteins are kept distinct. When evaluating toxicological effects of Fluvastatin he is interested in both the pharmacological active stereoisomer of Fluvastatin and inactive one, since the latter still might have side effects. He therefore allows the link between ChemSpider and DrugBank for Fluvastatin.

Fiona the Bioinformatician. Fiona works for a pharmaceutical company. She searches for potential drug-like compounds by mining existing literature for similar compounds to interact with known targets. For compounds that have been well studied, she would like to apply strict operational equivalence conditions, i.e.

[10] `http://www.openphacts.org/` accessed 7 Sept 2012.

equating chemical compounds only if their stereochemistry matches and keeping genes and proteins distinct, in order to work with the most relevant data. However, for compounds that have not been so widely studied, she is willing to apply more permissive notions of equivalence to increase the volume of data returned.

3 Scientific Lenses and Linked Data

Within scientific datasets it is common to find links to the "equivalent" record in another dataset. However, there is no declaration of the form of the relationship. There is a great deal of variation in the notion of equivalence implied by the links both within a dataset's usage and particularly across datasets, which degrades the quality of the data. The scientific user personae have very different needs about the notion of equivalence that should be applied between datasets. The users need a simple mechanism by which they can change the operational equivalence applied between datasets. We propose the use of scientific lenses.

A *scientific lens* is defined in terms of the scientific notions upon which operational equivalence can vary, e.g. stereochemistry, gene and protein equivalence, and cross species matching. Our goal is that the scientists should be able to focus on their science and not on the links between datasets. We do not propose to define an ontology of relationships; Halpin *et al* [6] have already proposed a similarity ontology which captures the nuances of equivalence at a logical level. However, they also showed that it is difficult to accurately identify the relationship predicate to use in a given situation. Instead, we propose to capture the context in which the link holds, and specifically a justification for the link, e.g. stereochemistry. The metadata and the links are captured as a VoID linkset [1]. We have specified the predicates expected in the VoID header in order to support scientific lenses in the Open PHACTS platform [3]. We propose to extend BridgeDB [7] to interpret the metadata about a linkset and decide under which scientific lenses it is active. BridgeDB can then be used by applications to discover operationally equivalent records under different scientific lenses. It is intended that a lens will specify all of the operational equivalence criteria that it applies, i.e. there will be several lenses which match compounds based on their stereochemistry but which vary in some other category of equivalence.

Consider again the scientific personae. Chris would have two lenses that he could switch between: one that keeps species independent of each other and one that matches proteins across species. Fiona would also have two lenses: one that applies strict levels of operational equivalence and a more relaxed one. She could of course define intermediary ones to allow her to slide the level of strictness.

4 Conclusions

Equivalence of different, but for some purposes comparable content from, data resources depends upon a user's opinion, task, and context. We have proposed applying scientific lenses which vary the notion of operational equivalence between datasets. A scientific lens varies the links that are active. Thus, a query

over the Linked Data would see a difference in the data returned when applying lenses with different notions of operational equivalence. Note that although we are developing our lenses in the context of pharmacology, the same principles could be applied in other domains. Our ongoing work is concentrating on two strands. First we are developing mechanisms to define scientific lenses in terms of scientific notions which can then be used to control which links are active. Second we are developing a query infrastructure to support the contextualised stand-off mappings used by our scientific lenses approach.

Acknowledgements

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement number 115191, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007- 2013) and EFPIA companies' in kind contribution.

References

1. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing linked datasets with the VoID vocabulary. Note, W3C (Mar 2011), http://www.w3.org/TR/void/
2. Belleau, F., Nolin, M.A., Tourigny, N., Rigault, P., Morissette, J.: Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. Journal of Biomedical Informatics 41(5), 706 – 716 (2008)
3. Brenninkmeijer, C., Evelo, C., Goble, C., Gray, A.J.G., Waagmeester, A., Willighagen, E.: Dataset descriptions for the open pharmacological space. Working draft, Open PHACTS (August 2012), http://www.openphacts.org/specs/datadesc/
4. Chen, B., Dong, X., Jiao, D., Wang, H., Zhu, Q., Ding, Y., Wild, D.: Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. BMC Bioinformatics 11(1), 255 (2010)
5. Correndo, G., Salvadores, M., Millard, I., Glaser, H., Shadbolt, N.: SPARQL query rewriting for implementing data integration over linked data. In: EDBT/ICDT Workshops (2010)
6. Halpin, H., Hayes, P.J., McCusker, J.P., McGuinness, D.L., Thompson, H.S.: When owl: sameas isn't the same: An analysis of identity in linked data. In: International Semantic Web Conference (1). pp. 305–320 (2010)
7. van Iersel, M.P., Pico, A.R., Kelder, T., Gao, J., Ho, I., Hanspers, K., Conklin, B.R., Evelo, C.T.A.: The BridgeDB framework: standardized access to gene, protein and metabolite identifier mapping services. BMC Bioinformatics 11, 5 (2010)
8. Juty, N., Le Novere, N., Laibe, C.: Identifiers.org and MIRIAM registry: community resources to provide persistent identification. Nucleic Acids Research 40(D1), D580–D586 (2012)
9. Samwald, M., Jentzsch, A., Bouton, C., Kallesoe, C., Willighagen, E., Hajagos, J., Marshall, M., Prud'hommeaux, E., Hassanzadeh, O., Pichler, E., Stephens, S.: Linked open drug data for pharmaceutical research and development. Journal of Cheminformatics 3(1), 19+ (2011)
10. Williams, A.J., Harland, L., Groth, P., Pettifer, S., Chichester, C., Willighagen, E.L., Evelo, C.T., Blomberg, N., Ecker, G., Goble, C., Mons, B.: Open PHACTS: Semantic interoperability for drug discovery. Drug Discovery Today (2012)