Help me describe my data: A demonstration of the Open PHACTS VoID Editor

Carole Goble[1], Alasdair J G Gray[2], and Eleftherios Tatakis[1]

[1] School of Computer Science, University of Manchester, Manchester, UK
[2] Department of Computer Science, Heriot-Watt University, Edinburgh, UK

Abstract. The Open PHACTS VoID Editor helps non-Semantic Web experts to create machine interpretable descriptions for their datasets. The web app guides the user, an expert in the domain of the data, through a series of questions to capture details of their dataset and then generates a VoID dataset description. The generated dataset description conforms to the Open PHACTS dataset description guidelines that ensure suitable provenance information is available about the dataset to enable its discovery and reuse.
The VoID Editor is available at `http://voideditor.cs.man.ac.uk`. The source code can be found at `https://github.com/openphacts/Void-Editor2`.

Keywords: Dataset descriptions, VoID, Provenance, Metadata

1 Motivating Problem

Users of systems such as the Open PHACTS Discovery Platform[3] [1,2] need to know which datasets have been integrated. In the scientific domain they particularly need to know which version of a dataset is loaded in order to correctly interpret the results returned by the platform. To satisify this need, the provenance of the datasets loaded into the Open PHACTS Discovery Platform are needed. This provenance information is then available for any data returned by the platform's API. Within the Open PHACTS project we have identified a minimal set of metadata that should be provided to aid understanding and reuse of the data [3]. Additionally, we recommend that the metadata is provided using the VoID vocabulary [4] so that the data is self-describing and machine processable.

Open PHACTS does not publish its own datasets; it integrates existing publicly available domain data. Typically the publishers of these scientific data sets are experts in their scientific domain, viz. chemistry or biology, but not in the semantic web. They need to be supported in the creation of VoID descriptions of their datasets which may have been published in a database and converted into RDF. A tool which hides the underlying details of the semantic web but enables the creation of descriptions understandable to a domain expert is thus needed.

[3] `https://dev.openphacts.org/` accessed July 2014

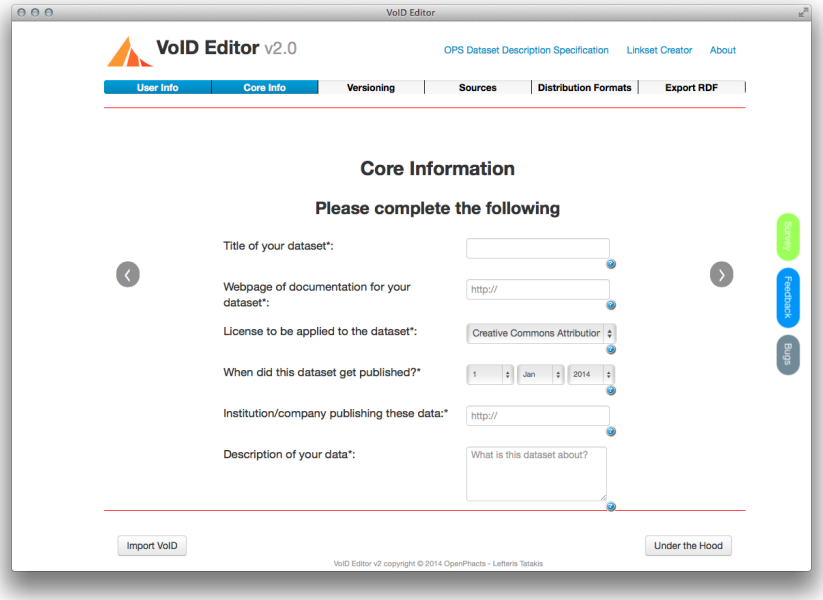

Fig. 1: Screenshot of the VoID Editor

2 VoID Editor

The aim of the VoID Editor (see screenshot in Figure 1) is to allow a data publisher to create validated dataset descriptions within 30 minutes. In particular, the data publisher does not need to read and understand the Open PHACTS dataset descriptions guidelines [3] which provide a checklist of the RDF properties that MUST and SHOULD be provided. There is also no need for the data publisher to understand RDF or the VoID vocabulary.

The VoID Editor is a web application that guides the data provider through a series of questions to acquire the required metadata properties. The user is first asked for details about themselves and other individuals involved in the authoring of the data. Core publishing metadata such as the publishing organisation and the license are then gathered. The user is then asked for versioning information and the expected update frequency of the data. The Sources tab helps the user to provide details of source datasets from which their data is derived. They can either select from the datasets already known to the Open PHACTS Discovery Platform or enter the details manually. The list of known datasets is populated by a call to the Open PHACTS API. The Distribution Formats tab allows the user to describe the distributions in which the data is provided, e.g. RDF, database dump, or CSV. The final screen allows the user to export the RDF of their dataset description as well as providing a summary of any validation errors, e.g. not supplying a license which is a required field, such errors

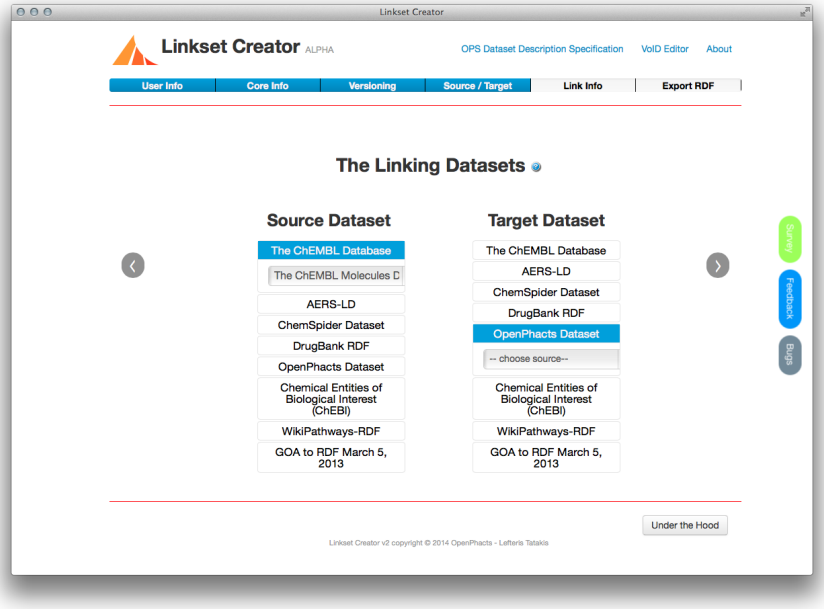

Fig. 2: Screenshots of the Linkset Editor

will already have been indicated by a red bar at the top of the screen containing an error message. Note that the 'Export RDF' button is only activated when a valid dataset description can be created, i.e. all required fields have been filled in.

At any stage, the generated RDF dataset description may be inspected by clicking the 'Under the Hood' button. This button can also be used to save a partially generated description that can later be imported into the editor through the 'Import VoID' button. The 'Under the Hood' feature is also useful for semantic web experts to see what is being generated at any stage.

3 Linkset Editor

In companion with the VoID Editor, a Linkset Editor (see screenshot in Figure 2) has been developed. The Linkset Editor allows for the creation of descriptions of the links between two datasets. The same interface design and framework is used.

The Linkset Editor reuses the first three tabs of the VoID Editor to capture details of the authors, core publishing information, and details about versioning. The Source/Target tab allows the user to select the pair of datasets that are connected by the linkset. Again, the list of possible datasets is generated by a call to the Open PHACTS API. The Link Info tab asks the user to declare

the link predicate used in the linkset and provide some justification to capture the nature of the equality relationship encoded in the links. (For details about linkset justifications, please see Section 5 of [3].)

4 Implementation

The VoID and Linkset Editors have been implemented using AngularJS as a Javascript framework for the web client with a server implementation using Jena libraries. A user-centric approach was followed for the design and development of the VoID Editor. A small number of data providers were consulted about the type of tool they required with regular interviews and feedback on prototype versions. A larger number of potential users were involved in an evaluation of the VoID Editor. Full details can be found in [5].

In the future we plan to investigate how the VoID Editor can genearate template descriptions that can be populated as part of the data publishing pipeline. We also plan to look at how the editor could be adapted to other dataset description guidelines, e.g. DCAT[4] or the W3C HCLS community profile[5]. However, this is not a straightforward process since considerable care and attention is paid to the phrasing and grouping of questions to ensure a pleasant user experience.

Acknowledgements

The research 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. Gray, A.J.G., Groth, P., Loizou, A., Askjaer, S., Brenninkmeijer, C.Y.A., Burger, K., Chichester, C., Evelo, C.T., Goble, C.A., Harland, L., Pettifer, S., Thompson, M., Waagmeester, A., Williams, A.J.: Applying linked data approaches to pharmacology: Architectural decisions and implementation. Semantic Web **5**(2) (2014) 101–113 doi:10.3233/SW-2012-0088.
2. Groth, P., Loizou, A., Gray, A.J.G., Goble, C., Harland, L., Pettifer, S.: API-centric Linked Data Integration: The Open PHACTS Discovery Platform Case Study. Journal of Web Semantics (2014) In press. doi:10.1016/j.websem.2014.03.003.
3. Gray, A.J.G.: Dataset descriptions for the Open Pharmacological Space. Working draft, Open PHACTS (September 2013)
4. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing Linked Datasets with the VoID Vocabulary. Note, W3C (March 2011)
5. Tatakis, E.: VoID Editor v2. Undergraduate dissertation, School of Computer Science, University of Manchester, Manchester, UK (April 2014)

[4] `http://www.w3.org/TR/vocab-dcat/` accessed July 2014
[5] `http://www.w3.org/2001/sw/hcls/notes/hclsdataset/` access July 2014