Automatic Integration of Metadata into the Web of Linked Data

Olaf Hartig[1], Jun Zhao[2], and Hannes Mühleisen[1]

[1] Humboldt-Universität zu Berlin
(hartig|muehleis)@informatik.hu-berlin.de
[2] University of Oxford
jun.zhao@zoo.ox.ac.uk

Abstract. In order to enable a reliable and approved consumption and processing of Linked Data in applications it requires various information about the consumed data (e.g. licensing and provenance). Since a large amount of this information is available to the publishers of the data it should become a practice to provide this information as metadata. We demonstrate components that provide an easy-to-use mechanism for adding metadata to RDF graphs served by Linked Data publishing tools.

1 Introduction

With the adoption of the Linked Data principles by a significant amount of data publishers we foresee the emergence of novel applications that enable users to benefit from a virtually unbound set of data sources. During their execution these applications may require certain kinds of information about the consumed data. For instance, an application requires information about the licensing terms for different data items in order to check whether the data can be used as intended [1]. Furthermore, an application may require information about the provenance of data to decide about its quality and trustworthiness [2]. Another example is statistical information or data summaries about a dataset that could be used for source selection algorithms in, e.g., query distribution scenarios [3]. It is impossible to collect and use information like this beforehand for all data that might be discovered and used by an application. Therefore, applications have to collect this information during runtime as part of the data consumption process. Such an on-the-fly collection relies on the availability of metadata about the consumed data. We are convinced that a successful discovery and utilization of such metadata can only be possible if the metadata is an integral part of the Web of Linked Data. This means, the publication of this metadata should adhere to the same principles that are used for the data itself. To initiate such a practice with a minimal effort for data publishers we extended several Linked Data publishing tools with corresponding metadata components. In the remainder of this demonstration proposal we give a brief overview of our metadata components and we outline how we present these components during the demonstration session.

2 Metadata Components for Publishing Tools

The integration of metadata into the Web of Linked Data can only be successful if generating and publishing this metadata is possible with minimal effort for data providers. For this reason, we developed metadata components for widely used Linked Data publishing tools, including Triplify[1], Pubby[2] and D2R server[3]. These extensions provide a mechanism for adding metadata to the RDF graphs served by the publishing tools. Hence, with our metadata components the tools publish metadata together with the data itself. Additionally, our extensions augment the HTML templates so that the provided metadata is also visible in the human-readable output of the publishing tools (cf. Fig. 1 in the Appendix).

The original aim of developing the metadata components was the automatic publication of provenance-related metadata. For this reason, our components include a default template that –out of the box– adds various pieces of provenance information to the published data. Furthermore, data providers can easily enrich the generated metadata by simply configuring a few parameters, such as the name and URI identifying the provider or the URI of the dataset. Hence, with data providers upgrading their Linked Data servers to the latest release of the aforementioned publishing tools the amount of provenance information added to the Web of Linked Data will increase significantly.

The metadata provided by our components is generated on-the-fly from a template RDF graph by replacing placeholders in the template. These placeholders refer to i) run-time data, ii) typical configuration options, and iii) to additional configuration variables that are very easy to define. In addition to the definition of configuration variables referred to in a given template, the operator of a Linked Data server can also adjust the template graph itself. This customization is as easy as writing an RDF graph in a human-friendly format. Hence, our components provide a flexible metadata generation framework that can be customized for every data source with minimal effort.

3 The Demonstration

During the demonstration session we present the latest Pubby release as an exemplar of publication tools that include our metadata components. We set up a sample Pubby instance to showcase what provenance-related metadata Pubby provides out of the box. Furthermore, participants learn how Pubby can be configured to generate and provide additional metadata.

References

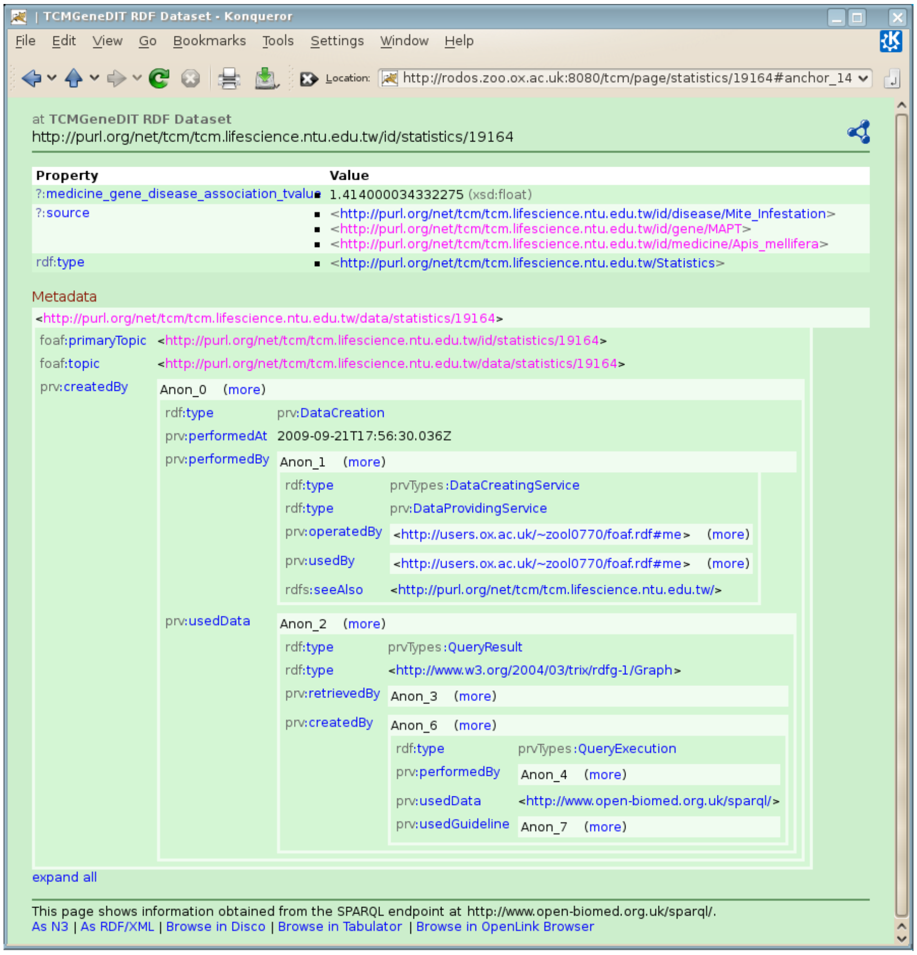
1. Miller, P., Styles, R., Heath, T.: Open Data Commons, a License for Open Data. In: Workshop on Linked Data on the Web (LDOW) at WWW. (2008)
2. Hartig, O., Zhao, J.: Using Web Data Provenance for Quality Assessment. In: Role of Semantic Web in Provenance Management Workshop (SWPM) at ISWC. (2009)
3. Langegger, A., Wöß, W.: RDFStats – An Extensible RDF Statistics Generator and Library. In: 20th International Workshop on Database and Expert Systems Applications (DEXA). (2009)

[1] http://triplify.org/

[2] http://www4.wiwiss.fu-berlin.de/pubby/

[3] http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/

Appendix



http://rodos.zoo.ox.ac.uk:8080/tcm/page/statistics/19164#anchor_14

at **TCMGeneDIT RDF Dataset**
http://purl.org/net/tcm/tcm.lifescience.ntu.edu.tw/id/statistics/19164

Property	Value
?:medicine_gene_disease_association_tvalue	1.414000034332275 (xsd:float)
?:source	▪ <http://purl.org/net/tcm/tcm.lifescience.ntu.edu.tw/id/disease/Mite_Infestation> ▪ <http://purl.org/net/tcm/tcm.lifescience.ntu.edu.tw/id/gene/MAPT> ▪ <http://purl.org/net/tcm/tcm.lifescience.ntu.edu.tw/id/medicine/Apis_mellifera>
rdf:type	▪ <http://purl.org/net/tcm/tcm.lifescience.ntu.edu.tw/Statistics>

Metadata

```
<http://purl.org/net/tcm/tcm.lifescience.ntu.edu.tw/data/statistics/19164>
  foaf:primaryTopic  <http://purl.org/net/tcm/tcm.lifescience.ntu.edu.tw/id/statistics/19164>
  foaf:topic         <http://purl.org/net/tcm/tcm.lifescience.ntu.edu.tw/data/statistics/19164>
  prv:createdBy      Anon_0   (more)
                     rdf:type         prv:DataCreation
                     prv:performedAt  2009-09-21T17:56:30.036Z
                     prv:performedBy  Anon_1   (more)
                                      rdf:type        prvTypes:DataCreatingService
                                      rdf:type        prv:DataProvidingService
                                      prv:operatedBy  <http://users.ox.ac.uk/~zool0770/foaf.rdf#me>   (more)
                                      prv:usedBy      <http://users.ox.ac.uk/~zool0770/foaf.rdf#me>   (more)
                                      rdfs:seeAlso    <http://purl.org/net/tcm/tcm.lifescience.ntu.edu.tw/>
                     prv:usedData     Anon_2   (more)
                                      rdf:type         prvTypes:QueryResult
                                      rdf:type         <http://www.w3.org/2004/03/trix/rdfg-1/Graph>
                                      prv:retrievedBy  Anon_3   (more)
                                      prv:createdBy    Anon_6   (more)
                                                       rdf:type          prvTypes:QueryExecution
                                                       prv:performedBy   Anon_4   (more)
                                                       prv:usedData      <http://www.open-biomed.org.uk/sparql/>
                                                       prv:usedGuideline Anon_7   (more)
```

expand all

This page shows information obtained from the SPARQL endpoint at http://www.open-biomed.org.uk/sparql/.
As N3 | As RDF/XML | Browse in Disco | Browse in Tabulator | Browse in OpenLink Browser

Fig. 1. HTML representation of data about a gene as provided by a Pubby service, including a visualization of the provenance metadata that our metadata component for Pubby generates automatically.