Community-Based Guidelines for Describing Biomedical Datasets

Community-Based Guidelines for Describing Biomedical Datasets

Michel Dumontier[*1], *M. Scott Marshall*[2], *Alasdair JG Gray*[3], *and the W3C HCLS Interest Group*

[1]*Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA,* [2]*MAASTRO Clinic, The Netherlands,* [3]*Heriot-Watt University, UK*

1 ABSTRACT

Access to consistent, high-quality metadata is critical to finding, understanding, and reusing scientific data. However, while there are many relevant vocabularies for the annotation of a dataset, none sufficiently capture all the necessary metadata. This prevents uniform querying of dataset repositories. Towards providing guidance for producing a high quality description of biomedical datasets, the W3C Semantic Web for Health Care and the Life Sciences Interest Group (HCLSIG) identified published RDF vocabularies that could be used to specify common metadata elements and their value sets. The resulting guideline covers elements of description, identification, attribution, versioning, attribution, provenance, and content summarization. This guideline reuses existing vocabularies, and is intended to meet key functional requirements including indexing, discovery, exchange, query, and retrieval of datasets.

2 INTRODUCTION

Big Data presents an exciting opportunity to pursue large-scale analyses over collections of data in order to uncover valuable insights across a myriad of fields and disciplines. Yet, as more and more data is made available, researchers are finding it increasingly difficult to discover and reuse these data. One problem is that data are insufficiently described to understand what they are or how they were produced. A second issue is that no single vocabulary provides all key metadata fields required to support basic scientific use cases. For instance, the Data Catalog Vocabulary (DCAT) (Maali, 2014) is used to describe datasets in catalogs, but does not deal with the issue of dataset evolution and versioning. A third issue is that data catalogs and data repositories all use different metadata standards, if they use any standard at all, and this prevents easy search, aggregation, and exchange of data descriptions. Thus, there is need to combine these vocabularies in a comprehensive manner that meets the needs of data registries, data producers, and data consumers.

3 RESULTS

We developed a specification for the description of a dataset that that meets key functional requirements (dataset description, linking, exchange, change, content summary), reuses

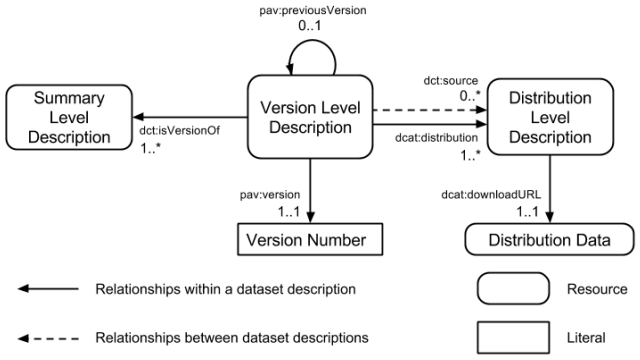

Figure 1. Three component model for dataset description

18 existing vocabularies, and is expressed using the Resource Description Framework (RDF). The specification covers 61 metadata elements pertaining to data description, identification, licensing, attribution, conformance, versioning, provenance, and content summary. Each metadata element includes a description and an example of use. The specification extends DCAT with versioning through a three component model (Figure 1). The summary level description focuses on file independent metadata that is often captured by dataset registries; the distribution level description focuses on specific data files, their formats and downloadable location; and the version level description links summary descriptions with distribution descriptions with a version number. Each description level is bound to a different set of metadata requirements – mandatory, recommended, optional. A full worked example using the ChEMBL dataset is provided. The group is currently evaluating the specification with implementations for dataset registries such as Identifiers.org and Linked Data repositories such as Bio2RDF (Callahan, 2013). The specification is currently available at http://www.w3.org/TR/hcls-dataset/

REFERENCES

Maali, F and Erickson J. (2014) *Data Catalog Vocabulary*. W3C Recommendation. http://www.w3.org/TR/vocab-dcat/

Callahan,A, Cruz-Toledo, J, Ansell, P and Dumontier, M. (2013) Bio2RDF Release 2: Improved Coverage, Interoperability and Provenance of Life Science Linked Data. *LNCS*, **7882**, 200–12.

Williams AJ, et al. (2012) OpenPHACTS:semantic interoperability for drug discovery. Drug Dis. Tod. **17**, 1188-1198.