The HCLS Community Profile: Describing Datasets, Versions, and Distributions

Michel Dummontier[1], Alasdair J. G. Gray[2], and M. Scott Marshall[3]

[1] Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, California, USA
[2] Computer Science, Heriot-Watt University, Edinburgh, UK
[3] Department of Radiation Oncology, Netherlands Cancer Institute, Amsterdam, The Netherlands

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 captures all the necessary metadata. This prevents uniform indexing and querying of dataset repositories. Towards providing a practical guide for producing a high quality description of biomedical datasets, the W3C Semantic Web for Health Care and the Life Sciences Interest Group (HCLSIG) identified Resource Description Framework (RDF) vocabularies that could be used to specify common metadata elements and their value sets. The resulting HCLS community profile covers elements of description, identification, attribution, versioning, provenance, and content summarization. The HCLS community profile reuses existing vocabularies, and is intended to meet key functional requirements including indexing, discovery, exchange, query, and retrieval of datasets, thereby enabling the publication of FAIR data. The resulting metadata profile is generic and could be used by other domains with an interest in providing machine readable descriptions of versioned datasets.

The goal of this presentation is to give an overview of the HCLS Community Profile and explain how it extends and builds upon other approaches.

Keywords: Dataset description, Metadata, FAIR Data Principles, Data profile

1 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. The W3C Health Care and Life Sciences (HCLS) Interest Group have developed a community profile [1,2] that defines the required properties to provide high-quality dataset descriptions that support finding, understanding, and reusing scientific data, i.e. making the data FAIR (Findable, Accessible, Interoperable and Reusable) [3].

The HCLS Community Profile reuses the definition of a dataset from [4]. That is, a dataset is defined as

> A collection of data, available for access or download in one or more formats.

For instance, a dataset may be generated as part of some scientific investigation, whether tabulated from observations, generated by an instrument, obtained via analysis, created through a mash-up, or enhanced or changed in some manner.

While several vocabularies are relevant in describing datasets, none are sufficient to completely provide the breadth of requirements identified in Health Care and the Life Sciences. The Dublin Core Metadata Initiative (DCMI) [5] Metadata Terms offers a broad set of types and relations for capturing document metadata. The Data Catalog Vocabulary (DCAT) [4] is used to describe datasets in catalogs, but does not deal with the issue of dataset evolution and versioning. The Provenance Ontology (PROV) [6] can be used to capture information about entities, activities, and people involved in producing or modifying data. The Vocabulary of Interlinked Datasets (VoID) [7] is an RDF Schema (RDFS) [8] vocabulary for expressing metadata about Resource Description Framework (RDF) [9] datasets. Schema.org[4] has a limited proposal for dataset descriptions. Thus, there is a need to combine these vocabularies in a comprehensive manner that meets the needs of data registries, data producers, and data consumers, i.e. to support the publication of FAIR data.

Here we provide a brief overview of the HCLS Community Profile; full details can be found in [1,2]. We will first outline the key requirements that led to the development of the HCLS Community Profile (Section 2) before giving a summary of the HCLS Community Profile (Section 3).

2 Motivation for the HCLS Community Profile

The HCLS Special Interest Group gathered use cases from all members involved with the activity of developing the Community Profile. The use cases were analysed for common usage patterns for metadata in order to identify the metadata properties that would be required. In addition to the common metadata properties found in most dataset description vocabularies, the use cases identified requirements to support:

1. Distinct resolvable identifiers for the metadata about a dataset, its versions, and the distributions of these versions;
2. Descriptions of the identifiers used within a resource;
3. Details of the provenance of the data;
4. Rich statistics about RDF data to support querying.

In order to satisfy these requirements, the HCLS Community Profile was developed that combines properties from existing metadata vocabularies.

[4] `http://schema.org/Dataset` accessed June 2016

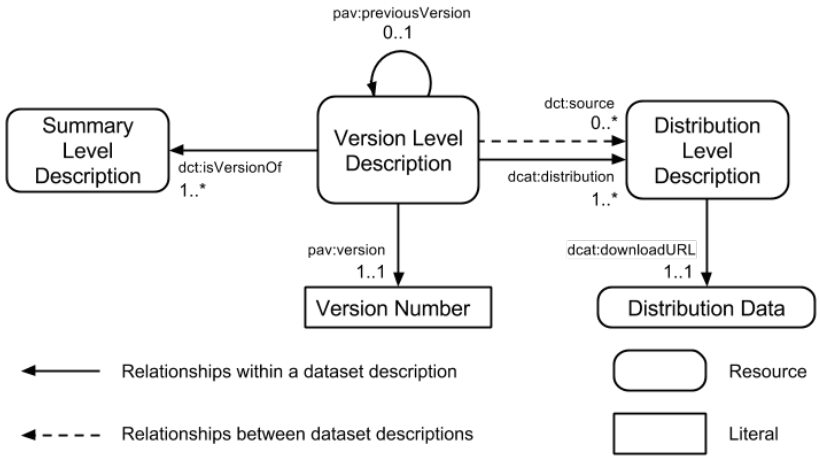

Fig. 1. Overview of the HCLS Community Profile

3 Overview of the HCLS Community Profile

We developed a community profile for the description of a dataset that meets key functional requirements (dataset description, linking, exchange, change, content summary), reuses 18 existing vocabularies, and is expressed in a machine readable format using RDF [9]. The specification covers 61 metadata elements pertaining to data description, identification, licensing, attribution, conformance, versioning, provenance, and content summary. For each metadata element a description and an example of its use is given. Full details of the specification can be found in the W3C Interest Group Note [1].

The community profile extends the DCAT model [4] with versioning through a three component model (Figure 1), and detailed summary statistics. The three components of the dataset description model are:

Summary Level Description: provides a description of the dataset that is independent of file formats or versions of the dataset. For example, this level will capture the title of the dataset which is not expected to change from one version of the dataset to another, but will not contain details of the version number. This is akin to the information that would be captured in a dataset registry.

Version Level Description: provides a description of the dataset that is independent of the file formats but tied to the specific release version of a dataset. For example, this level will capture the release date and version number of a specific version of the dataset but will not contain details of where the data files can be obtained.

Distribution Level Description: provides a description of the files through which a specific version of a dataset is made available. Examples of the

types of metadata captured are the file format, the location from which it is made available, and summary statistics about the data model (e.g. number of triples in the RDF distribution).

Each description component has a different set of metadata properties specified at the appropriate requirement level – mandatory (MUST), recommended (SHOULD), and optional (MAY).

The metadata elements are grouped into five modules covering Core Metadata (e.g. dataset title), Identifiers (e.g. patterns for data item identifiers), Provenance and Change (e.g. relationship to source datasets), Availability/Distributions (e.g. released data files), and Statistics (e.g. number of triples).

4 Conclusions

The HCLS Community Profile for dataset descriptions was developed to meet an identified need not satisfied by other dataset description standards. The HCLS Community Profile builds upon many existing metadata standards and vocabularies. While it was developed within the W3C Health Care and Life Sciences Interest Group, it does not contain any domain specific properties. The HCLS Community Profile is beginning to see uptake within the Health Care and Life Sciences community, e.g. Riken MetaDatabase[5], Bio2RDF[6], and PHI-Base[7] [10].

References

1. Gray, A.J.G., Baran, J., Marshall, M.S., Dumontier, M.: Dataset descriptions: HCLS community profile. Interest group note, W3C (May 2015) `http://www.w3.org/TR/hcls-dataset/`.
2. Dumontier, M., Gray, A.J.G., Marshall, M.S., Alexiev, V., Ansell, P., Bader, G., Baran, J., Bolleman, J.T., Callahan, A., Cruz-Toledo, J., Gaudet, P., Gombocz, E.A., Gonzalez-Beltran, A.N., Groth, P., Haendel, M., Ito, M., Jupp, S., Juty, N., Katayama, T., Kobayashi, N., Krishnaswami, K., Laibe, C., Le Novère, N., Lin, S., Malone, J., Miller, M., Mungall, C.J., Rietveld, L., Wimalaratne, S.M., Yamaguchi, A.: The health care and life sciences community profile for dataset descriptions. PeerJ **4** (August 2016) e2331
3. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.w., Bonino da Silva Santos, L., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C.A., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons,

[5] `http://metadb.riken.jp/metadb/front` accessed September 2016
[6] `https://github.com/bio2rdf/bio2rdf-scripts/wiki` accessed September 2016
[7] `http://linkeddata.systems/SemanticPHIBase/Metadata` accessed September 2016

B.: The FAIR Guiding Principles for scientific data management and stewardship Authors. Nature Scientific Data **3** (2016) doi:10.1038/sdata.2016.18.
4. Maali, F., Erickson, J.: Data catalog vocabulary (DCAT). Recommendation, W3C (January 2014) `http://www.w3.org/TR/vocab-dcat/`.
5. DCMI Usage Board: DCMI metadata terms. Recommendation, DCMI (June 2012) `http://dublincore.org/documents/dcmi-terms/`.
6. Lebo, T., Sahoo, S., McGuinness, D.: PROV-O: the PROV ontology. Recommendation, W3C (April 2013) `http://www.w3.org/TR/prov-o/`.
7. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing linked datasets with the VoID vocabulary. Interest group note, W3C (2011) `http://www.w3.org/TR/void/`.
8. Brickley, D., Guha, R.: RDF schema 1.1. Recommendation, W3C (February 2014) `http://www.w3.org/TR/rdf-schema/`.
9. Cyganiak, R., Wood, D., Lanthaler, M.: RDF 1.1 concepts and abstract syntax. Recommendation, W3C (February 2014) `http://www.w3.org/TR/rdf11-concepts/`.
10. Rodrguez Iglesias, A., Rodrguez Gonzlez, A., Irvine, A.G., Sesma, A., Urban, M., Hammond-Kosack, K.E., Wilkinson, M.D.: Publishing fair data: an exemplar methodology utilizing phi-base. Frontiers in Plant Science **7**(641) (2016)