A peer-reviewed version of this preprint was published in PeerJ on 16 August 2016.

View the peer-reviewed version (peerj.com/articles/2331), which is the preferred citable publication unless you specifically need to cite this preprint.

Dumontier M, Gray AJG, Marshall MS, Alexiev V, Ansell P, Bader G, Baran J, Bolleman JT, Callahan A, Cruz-Toledo J, Gaudet P, Gombocz EA, Gonzalez-Beltran AN, 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 CJ, Rietveld L, Wimalaratne SM, Yamaguchi A. 2016. The health care and life sciences community profile for dataset descriptions. PeerJ 4:e2331 https://doi.org/10.7717/peerj.2331

The health care and life sciences community profile for dataset descriptions

Michel Dumontier, Alasdair J G Gray, M. Scott Marshall, Vladimir Alexiev, Peter Ansell, Gary Bader, Joachim Baran, Jerven T Bolleman, Alison Callahan, José Cruz-Toledo, Pascale Gaudet, Erich Gombocz, Alejandra N Gonzalez Beltran, Paul Groth, Melissa Melissa Haendel, Maori Ito, Simon Jupp, Nick Juty, Toshiaki Katayama, Norio Kobayashi, Kalpana Krishnaswami, Camille Laibe, Nicolas Le Novère, Simon Lin, James Malone, Michael Miller, Chris Mungall, Laurens Rietveld, Sarala M Wimalaratne, Atsuko Yamaguchi

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 guideline covers elements of description, identification, attribution, versioning, 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, 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 health care and life sciences community profile for dataset descriptions

Michel Dumontier[1], Alasdair J. G. Gray[2], M. Scott Marshall[3], Vladimir Alexiev[4], Peter Ansell[5], Gary D. Bader[6], Joachim Baran[1], Jerven Bolleman[7], Alison Callahan[8], José Cruz-Toledo[8], Pascale Gaudet[7], Erich Gombocz[9], Alejandra Gonzalez-Beltran[10], Paul Groth[11], Melissa Haendel[12], Maori Ito[13], Simon Jupp[14], Nick Juty[14], Toshiaki Katayama[15], Norio Kobayashi[16], Kalpana Krishnaswami[17], Camille Laibe[14], Nicolas Le Novère[18], Simon Lin[19], James Malone[14], Michael Miller[20], Chris Mungall[21], Laurens Rietveld[11], Sarala M. Wimalaratne[14], and Atsuko Yamaguchi[15]

[1]Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA
[2]Computer Science, Heriot-Watt University, Edinburgh, UK
[3]Department of Radiation Oncology (MAASTRO), GROW - School for Oncology and Developmental Biology, The Netherlands
[4]Ontotext Corporation, Bulgaria
[5]CSIRO, Australia
[6]The Donnelly Centre, University of Toronto, Canada
[7]SIB Swiss Institute of Bioinformatics, Switzerland
[8]Carleton University, Canada
[9]IO Informatics, USA
[10]University of Oxford, UK
[11]VU University Amsterdam, The Netherlands
[12]Oregon Health and Science University, USA
[13]NIBIO, Japan
[14]EMBL-EBI, UK
[15]Database Center for Life Sciences, Japan
[16]RIKEN, Japan
[17]Metaome, USA
[18]Babraham Institute, UK
[19]Nationwide Children's Hospital, Ohio, USA
[20]Institute for Systems Biology, USA
[21]Lawrence Berkeley National Laboratory, USA

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 guideline covers elements of description, identification, attribution, versioning, 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, 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.

Keywords: dataset descriptions, data profiling, metadata, provenance

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 (Ohno-Machado et al., 2015). 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 and aggregation of data (Vasilevsky et al., 2013). To overcome these challenges, we have come together as a community to provide guidance for defining essential metadata to accurately describe a dataset, and the manner in which we can express it. The resulting descriptions support the publication of FAIR datasets that are Findable, Accessible, Interoperable, and Reusable (Wilkinson et al., 2016).

For the purposes of this article, we reuse the definition of a dataset from (Maali and Erickson, 2014). 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. Research data are available in research publications and supplemental documents such as the Nucleic Acids Research database issue[1], in literature curated databases such as ChEMBL (Bento et al., 2014), PharmGKB[2] or the CTD[3], or from research repositories such as BioMedCentral-BGI GigaScience[4], Nature Publishing Group's Scientific Data[5], Dryad Digital Repository[6], FigShare[7], and Harvard Dataverse[8]. Cross-repository access is possible through data catalogs such as Neuroscience Information Framework (NIF)[9], BioSharing[10], Identifiers.org Registry[11], Integbio Database Catalog[12], Force11[13], and CKAN's datahub[14].

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) (DCMI Usage Board, 2012) Metadata Terms offers a broad set of types and relations for capturing document metadata. The Data Catalog Vocabulary (DCAT) (Maali and Erickson, 2014) is used to describe datasets in catalogs, but does not deal with the issue of dataset evolution and versioning. The Provenance Ontology (PROV) (Lebo et al., 2013) can be used to capture information about entities, activities, and people involved in producing or modifying data. The Vocabulary of Interlinked Datasets (VoID) (Alexander et al., 2011) is an RDF Schema (RDFS) (Brickley and Guha, 2014) vocabulary for expressing metadata about Resource Description Framework (RDF) (Cyganiak et al., 2014) datasets. Schema.org[15] 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 describe the results of a multi-stakeholder effort under the auspices of the W3C Semantic Web for Health Care and Life Sciences Interest Group (HCLS[16]) to produce a specification for the description of datasets that meets key functional requirements, uses existing vocabularies, and is expressed using the Resource Description Framework. We discuss elements of data description including provenance and

[1] www.oxfordjournals.org/nar/database/paper.html accessed November 2015
[2] https://www.pharmgkb.org// accessed July 2015
[3] http://ctdbase.org/ accessed July 2015
[4] http://www.gigasciencejournal.com/ accessed July 2015
[5] http://nature.com/sdata accessed July 2015
[6] http://datadryad.org/ accessed July 2015
[7] http://figshare.com/ accessed July 2015
[8] https://dataverse.harvard.edu/ accessed July 2015
[9] http://www.neuinfo.org/ accessed July 2015
[10] http://biosharing.org/ accessed July 2015
[11] http://identifiers.org accessed July 2015
[12] http://integbio.jp/dbcatalog/?lang=en accessed July 2015
[13] https://www.force11.org/catalog accessed July 2015
[14] http://datahub.io/ accessed July 2015
[15] http://schema.org/Dataset accessed July 2015
[16] http://www.w3.org/blog/hcls/ accessed October 2015

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1982v1 | CC-BY 4.0 Open Access | rec: 21 Apr 2016, publ: 21 Apr 2016

versioning, and describe how these can be used for data discovery, exchange, and query (with SPARQL (The W3C SPARQL Working Group, 2013)). This then enables the retrieval and reuse of FAIR data to encourage reproducible science.

The contributions of this paper are:

- A description of the process followed to generate the community profile for dataset descriptions (given in the Methods Section);
- A summary of the community profile (given in the Results Section), a full specification is provided in the W3C Interest Group Note (Gray et al., 2015);
- An analysis of where the community profile fits with existing efforts for providing dataset descriptions (given in the Discussion Section).

METHODS

The Health Care and Life Sciences Community Group

The health care and life sciences community profile for describing datasets was developed as a collaborative effort within the Health Care and Life Sciences Interest Group (HCLSIG[17]) of the World Wide Web Consortium (W3C[18]). This group has a mission to support and develop the use of Semantic Web technology across the health care and life sciences domains. Membership of the interest group is drawn from research and industry organisations from all over the world. A self-selecting subset of these members, i.e. those with an interest in the community project, were actively involved in the discussions for the community profile. Interested individuals could take part in the process via the weekly telephone conferences that were advertised on the HCLSIG mailing list, email discussions on the HCLSIG mailing list, or commenting directly on the working drafts of the various documents or raising issues on the issue tracker. The initial stages of the work saw high levels of engagement with later stages only seeing the core team involved. However, in general it was during the early stages that most of the community agreement was made and the later stages were devoted to writing up the process.

Developing the Profile

The purpose of developing a community profile was to promote the discovery of datasets, enable their reuse, and tracking the provenance of this reuse. An overarching goal in the development of the community profile was to identify and reuse existing vocabulary terms rather than create yet another vocabulary for describing datasets. We believe that this approach will enable greater uptake of the profile due to the existing familiarity with many of these terms.

The development of the community profile was driven by use cases that were collected from the interest group members. These were analysed for common usage patterns for metadata in order to identify the metadata properties that would be required. The identified properties fed into the second activity of the group.

The second strand of activity was to collect existing vocabulary terms for each of the desired properties. This included analysing existing dataset metadata publication practices (see the Discussion Section for details) as well as identifying other potential terms using tools such as the Linked Open Vocabulary repository[19] (Vandenbussche and Vatant, 2014), BioPortal[20] (Whetzel et al., 2011), and Web search engines.

Over the course of several months, the community group discussed each property for its inclusion in the community profile. An important consideration in this process was the set of possible semantic consequences for each choice, e.g. usage of the Vocabulary of Interlinked Dataset properties would entail that the dataset was available as RDF which is not true of all health care and life sciences datasets. A vote was conducted for each property to choose the appropriate term for the property and its level of requirement. The requirement level was specified with terms such as MUST, SHOULD, or MAY, as defined by RFC 2119 (Bradner, 1997).

[17] `http://www.w3.org/blog/hcls/` accessed July 2015
[18] `http://w3.org` accessed July 2015
[19] `http://lov.okfn.org/` accessed October 2015
[20] `http://bioportal.bioontology.org/` accessed October 2015

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1982v1 | CC-BY 4.0 Open Access | rec: 21 Apr 2016, publ: 21 Apr 2016

Process Management

During the development of the community profile, a variety of document management and discussion approaches were used. The reason for using different approaches was due to the affordances that they offered.

During the requirement capture phase it was important to support as many people contributing their use cases and vocabulary terms as possible. We decided that there should be as low an entry barrier to participation as possible. This phase of the development was conducted through a shared and open Google document[21] and spreadsheet[22] respectively. These enabled multiple participants to concurrently edit the document using a Web browser in an interface similar to the corresponding desktop application. It is possible to comment on sections of text and discuss issues through threading of comments in the context of the document, as well as to have live chat sessions in the margins of the document Web page while editing. The majority of the community profile was developed within the shared Google document.

In preparation for publishing the community profile as a W3C Interest Group Note, the content of the Google document was transformed into a HTML document and stored in an open GitHub repository[23]. At this point the Google document was made read-only, to avoid missing edits, and a link to the new location inserted at the top. The GitHub repository allowed for the continued collaborative editing of the document, although not through a WYSIWIG online editor or in the same interactive collaborative way. However, the GitHub repository enabled better management of the HTML version as well as tracking and responding to issues relating to the document. A real-time preview of the editors' draft version of the interest group note was available using the GitHub HTML preview feature[24]. It was this preview location that was circulated on the mailing list and used in discussions during the telephone conferences.

In accordance with W3C procedures, once the community profile was finalised, the preview version was circulated via the W3C HCLSIG mailing list with a link to the issue tracker for generating new issues. Once these final issues were resolved, the HCLSIG formally voted to accept the community profile during a telephone conference in April 2015. The note was then published on the W3C pages in May 2015 once styling issues had been resolved.

RESULTS

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 (Cyganiak et al., 2014). 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 (Gray et al., 2015). Here, we will summarise the features of the community profile.

The community profile extends the DCAT model (Maali and Erickson, 2014) 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.

[21] `https://docs.google.com/document/d/1zGQJ9bO_dSc8taINTNHdnjYEzUyYkbjglrcuUPuoITw/edit?usp=sharing` accessed July 2015

[22] `https://docs.google.com/spreadsheets/d/1bhbw1HAp5I_c9JvAxyURKW0uEGJ8jV5hlzf07ggWDxc/edit?usp=sharing` accessed July 2015

[23] `https://github.com/W3C-HCLSIG/HCLSDatasetDescriptions` accessed August 2015

[24] See `http://htmlpreview.github.io/?https://github.com/W3C-HCLSIG/HCLSDatasetDescriptions/blob/master/Overview.html` for an example (accessed July 2015).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1982v1 | CC-BY 4.0 Open Access | rec: 21 Apr 2016, publ: 21 Apr 2016

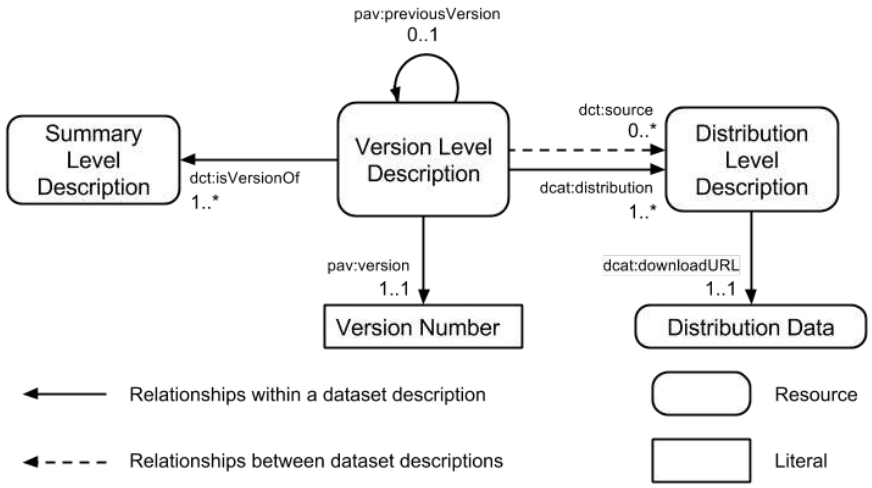

Figure 1. Three component model for dataset description.

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).

Modular Approach

The community profile is split into five thematic modules, each focusing on a different aspect of the metadata. However, this is simply to ease the presentation and understanding of the properties within the specification. The properties covered in each module must be supplied to provide a conformant description of a dataset. The modules and their focus are as follows.

Core Metadata: captures generic metadata about the dataset, e.g. its title, description, and publisher.

Identifiers: describes the patterns used for identifiers within the dataset and for the URI namespaces for RDF datasets.

Provenance and Change: describes the version of the dataset and its relationship with other versions of the same dataset and related datasets, e.g. an external dataset that is used as a source of information.

Availability/Distributions: provides details of the distribution files, including their formats, in which the dataset is made available for reuse.

Statistics: used to summarise the content of the dataset.

Note that in the current specification the statistics presented only make sense for RDF distributions of the data. However, it is in this case that these play the most important role since they provide a summary that will enable others to more effectively assess the contents of the data, as well as query the data. Indeed, the statistical summary matches the results of the queries used to explore a new SPARQL endpoint, e.g. providing the total number of triples, the number of distinct subjects and objects, and the number of relationships between each subject and object pair. Another motivation for providing these statistics in the dataset description is due to the fact that many of the required queries are computationally demanding on the data provider. Thus it is better to execute them once and publish the results in the distribution level description. For other distributions, such as a relational database dump, the data generally provides a schema to describe the relationships in the data.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1982v1 | CC-BY 4.0 Open Access | rec: 21 Apr 2016, publ: 21 Apr 2016

Vocabulary Reuse

As stated in the Methods Section, a goal of the community profile was to reuse existing vocabulary terms. This was broadly possible except in three cases, which we will now discuss and along with our chosen solutions.

To capture the link between a Summary Level Description and the current version of a dataset, a new vocabulary term was required. No existing term could be found to provide this information. Other versioning related properties such as the version number or the relationship to the previous version were supplied using the Provenance, Authoring and Versioning Vocabulary (PAV) (Ciccarese et al., 2013). We engaged the creators of the PAV ontology to have them create the property `pav:hasCurrentVersion` as we felt that this was an omission from the original vocabulary and would be beneficial to the wider community. After discussing the use cases where this was required, the PAV authors added the term to their vocabulary.

The second case involved the description of identifiers and their relationship to namespaces used in the dataset. Identifiers.org (Juty et al., 2012) is a key resource that provides metadata about life science databases and their identifier schemes. We worked with the Identifiers.org team to develop the idot vocabulary to include terms to define primary short names (`idot:preferredPrefix`), alternate short names (`idot:alternatePrefix`), core identifier patterns (`idot:identifierPattern`) and example identifiers (`idot:exampleIdentifier`). We leverage the VoID vocabulary to specify full URI templates (`void:uriRegexPattern`) and URI identifiers (`void:exampleResource`).

The final case was the ability to capture the relationship between a resource in an RDF dataset and the description of the distributions of the dataset that contain it. On the surface, the `void:inDataset` property would seem to convey this information. However, the domain of this property is `foaf:Document`, not the resource. This is due to the linked data assumptions inherent in VoID where triples are collected and served as documents on the Web. Since VoID is no longer actively developed or maintained, it was decided that a new term should be created in the Semanticscience Interlinked Ontology (Dumontier et al., 2014). The new terms, `sio:has-data-item` and `sio:is-data-item-in`, can be used to capture the relationship between a resource and the RDF distributions that it is available in.

Implementations

A worked example using the ChEMBL dataset (Bento et al., 2014) as an exemplar is provided in the Interest Group Note (Gray et al., 2015), and included in the supplementary material for this article (hcls-chembl-example.ttl.txt); a summary of the ChEMBL example is given in Table 1. The provided description is not intended as an accurate description of the ChEMBL dataset, but an illustration of how to provide the different description levels of the community profile and the various properties that should be used at each of the descriptions levels. In particular, the example provides a sample usage, at each description level, of each of the 61 properties that may be included. However, these are not complete, e.g. the ChEMBL dataset has many authors but only one has been declared in the example.

Table 1 shows that only a few properties are required at each description level. We note that many of these are repeated from one level to another to enable each level to be as self-contained as possible. For example, various core metadata properties such as publisher, license, and rights are kept the same on all levels of the description. The RDF distribution description is significantly larger due to the inclusion of an example of each of the types of statistical information provided. However, these are automatically generated using the SPARQL queries given in the Interest Group Note. Thus, they can be generated as part of the data publishing pipeline used to create the dataset and its metadata description.

We note that separate distribution level descriptions need to be provided for each RDF serialisation. However, since we expect that the dataset descriptions are generated automatically as part of the data publishing pipeline, there is no additional human effort required. The advantage is that the provenance of dataset reuse can be captured at the level of the distribution format used and thus problems identified in the data can be more effectively tracked. For example, the discrepancy could be a result of a problem in the transformation script that generates the N-triples serialisation of a dataset.

The HCLSIG is currently evaluating the specification with implementations for dataset registries such as Identifiers.org (Juty et al., 2012) and Riken MetaDatabase[25], and Linked Data repositories such as Bio2RDF (Callahan et al., 2013) and the EBI RDF Platform (Jupp et al., 2014).

[25] `http://metadb.riken.jp/metadb/front` accessed July 2015

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1982v1 | CC-BY 4.0 Open Access | rec: 21 Apr 2016, publ: 21 Apr 2016

Resource	*Description*	*Number of triples*
`:chembl`	Summary level description of the ChEMBL dataset	23
`:chembl17`	Version level description corresponding to version 17 of the ChEMBL dataset	42
`:chembl17db`	Distribution level description corresponding to an SQL dump of the ChEMBL17 database	48
`:chembl17rdf`	Distribution level description corresponding to an RDF release of the ChEMBL17 database in the turtle serialisation	107

Table 1. Summary of the resources in the ChEMBL example dataset description.

Validating Dataset Descriptions

To help encourage uptake of the community profile, we have developed an online validation tool to identify when a dataset description conforms to the community profile (Baungard Hansen et al., 2015). The online version of the tool enables the user to paste their dataset description in a variety of RDF serialisations. It then analyses the RDF, identifies the resources that provide dataset descriptions and validates the properties expressed against those expected for that level of description. The user can override the suggested description level, e.g. if the tool suggests that a resource is a summary level description but the user wants to validate it as a version level description they can specify that. The validator will either report a successful validation with a green tick or provide contextualised details of where errors have occurred. This allows dataset publishers to verify their descriptions.

The validation tool is also available for download[26] and using the node.js framework[27] can be incorporated into data publishing pipelines. Thus, validation of the dataset description can become an integral part of data publishing.

DISCUSSION

Existing Vocabularies

A number of existing vocabularies have been developed for describing datasets. However it was only by using a combination of multiple vocabularies that we were able to satisfy the use cases identified within the HCLSIG. We will now discuss why some prominent vocabularies in the area were insufficient to meet the needs of the HCLS community.

The Dublin Core Metadata Initiative (DCMI) publish the Dublin Core Terms and Dublin Core Types ontologies (DCMI Usage Board, 2012). These are widely used for providing metadata about web resources. They provide a core set of metadata properties such as `dct:title` for the title, `dct:license` for declaring the license, and `dct:publisher` for the publisher. However, there is no prescribed usage of Dublin Core terms, i.e. each and every resource is free to pick and choose which properties to use, nor does it cover all of the properties deemed necessary by the HCLSIG. In order to support the reuse of datasets, it was deemed important that there was a prescribed set of properties that would appear. The community profile recommends using 18 Dublin Core Terms in conjunction with properties drawn from other vocabularies to augment its coverage. This reuse is due to their suitability and existing wide-spread usage.

The Dataset Catalog Vocabulary (DCAT) (Maali and Erickson, 2014) is a W3C Recommendation that was developed in the eGovernment Interest Group[28] and turned into a recommendation by the Government Linked Data Working Group[29]. The goal of the vocabulary is to support the exchange of metadata records between data catalogs. The DCAT specification prescribes the properties that should appear, with many of these drawn from the Dublin Core ontologies. DCAT also created new terms which are used to distinguish between the catalog record of a dataset and the distribution files of the dataset. However, it does not distinguish between different versions of a dataset. Thus, the HCLS community profile extends this two

[26] `https://github.com/HW-SWeL/ShEx-validator` accessed November 2015
[27] `https://nodejs.org/` accessed July 2015
[28] `http://www.w3.org/egov/` accessed July 2015
[29] `http://www.w3.org/2011/gld/wiki/Main_Page` accessed July 2015

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1982v1 | CC-BY 4.0 Open Access | rec: 21 Apr 2016, publ: 21 Apr 2016

tier model into a three tier model where the versions of a dataset can be distinguished. The extra terms for describing the versioning information in the HCLS community profile are drawn from the Provenance, Authoring and Versioning Vocabulary (PAV) (Ciccarese et al., 2013).

A popular vocabulary in the Semantic Web community for describing datasets is the Vocabulary of Interlinked Datasets (VoID) (Alexander et al., 2011). This vocabulary also adopts many terms from the Dublin Core ontologies, although as with Dublin Core there are no prescribed properties to provide, thus making the use of VoID dataset descriptions more challenging. It was not possible to use VoID as the basis for the HCLSIG community profile as it assumes that all datasets are available as RDF. While the community profile uses RDF to enable dataset descriptions to be machine processable, it does not assume that the dataset that is being described is published in RDF; there are many datasets in the HCLS domain that do not publish their data in RDF, hence the need for projects like Bio2RDF (Callahan et al., 2013).

Existing Community Approaches

BioDBCore is a community driven checklist designed to provide core attributes for describing biological databases (Gaudet et al., 2011). The BioSharing catalog[30] is a curated and searchable web portal of interrelated data standards, databases, and data policies in the life, environmental, and biomedical sciences; databases are described using the BioDBCore attributes[31]. The databases catalog is progressively populated via in-house curation, assisted by community contributions via two routes: 1) through a collaboration with the Oxford University Press, where information is obtained from the annual Nucleic Acids Research Database issue and Database journal, and 2) via database developers and maintainers, who register their databases. This centralised and curated approach differs from the idea of the community profile described here. We anticipate the data publishers will publish the HCLS metadata descriptions together with their data and where possible embedded within the data. Registries such as BioDBCore can then harvest these descriptions and BioSharing are working to support automatic submission and updates based on HCLS dataset descriptions, supporting the idea that you write the metadata once and reuse it many times. We note that many of the properties covered in the BioDBCore are included in the HCLSIG community profile, although contact information such as email addresses are not included in the community profile. We anticipate that the usage of ORCID identifiers for individuals (Haak et al., 2012) and the ability to dereference these as RDF will eliminate the maintenance problem of keeping contact details up-to-date in many different registries.

The Big Data to Knowledge (BD2K) NIH-funded biomedical and healthCAre Data Discovery Index Ecosystem (bioCADDIE[32]) consortium (Ohno-Machado et al., 2015) works to develop the BD2K Data Discovery Index (DDI), to support data discovery complementing PubMed for the biomedical literature. The bioCADDIE Metadata Working Group has produced a metadata specification (WG3 Members, 2015) by converging requirements from a set of competency questions collected from the community and a set of existing models, formats and vocabularies used for describing data (including generic ones such as HCLSIG community profile and DataCite, and domain-specific ones such as ISA and SRA-xml). Thus, datasets described using the HCLSIG community profile can be easily included to the bioCADDIE index and work is ongoing at developing a HCLS profile data ingester.

The Open PHACTS Discovery Platform (Gray et al., 2014) provides an integrated view over several datasets with rich provenance information provided to identify where each data property originated. To enable this rich provenance, the Open PHACTS project developed their own standard for describing datasets based on a checklist of properties to provide (Gray, 2013). The Open PHACTS use case and their standard for dataset descriptions served as useful inputs to the HCLSIG community profile. The HCLSIG profile extends the Open PHACTS approach to support a larger number of use cases.

The Bio2RDF project provides scripts for converting biological datasets from their native format to an RDF representation together with SPARQL endpoints for interrogating and integrating the resulting data (Callahan et al., 2013). The conversion process from the original data source to the resulting RDF representation is captured by a provenance record that supplies core metadata about the source and resulting data[33] and is subsumed by the HCLSIG community profile. A key contribution to the HCLSIG community profile from the Bio2RDF project has been the need for providing rich statistics about the

[30] `https://www.biosharing.org/` accessed November 2015

[31] `https://biosharing.org/biodbcore/` accessed November 2015

[32] `https://biocaddie.org` accessed November 2015

[33] `https://github.com/bio2rdf/bio2rdf-scripts/wiki/Bio2RDF-Dataset-Provenance` accessed July 2015

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1982v1 | CC-BY 4.0 Open Access | rec: 21 Apr 2016, publ: 21 Apr 2016

RDF data[34], the purpose of which are to support understanding of the dataset without needing to pose common queries – some of which can be expensive to compute. These recommendations have been included in the distribution level descriptions of the HCLSIG community profile.

CONCLUSIONS

The Health Care and Life Sciences Community Profile for describing datasets has been developed as a community effort to support application needs. The development process has been inclusive allowing a wide varity of individuals to provide use cases, and to extensively discuss the choices of vocabulary terms made. This effort was enabled through the use of collaborative writing tools.

The resulting community profile enables the description of datasets and the different resources they make available: the dataset, its versions, and the distribution files of these versions. These descriptions can be used to publish datasets compliant with the FAIR data principles (Wilkinson et al., 2016) because they i) can be uniquely identified and retrieved using HTTP-based URIs, ii) are made available using RDF - a standardized, machine accessible knowledge representation language, iii) feature a rich set of metadata elements including licensing, provenance, and conformance to existing community-based vocabularies. The community profile has extended existing best practice for describing datasets in several key areas; most notably enabling the versions and distributions of a dataset to be distinguished and by providing rich statistics about the dataset. While the profile has been developed by the W3C HCLSIG, there are no aspects that are specific to this domain.

The community profile is currently undergoing validation by being implemented by several different projects. We anticipate that there will need to be a consolidation period where new use cases are added and improvements made to the existing community profile.

ACKNOWLEDGMENTS

We would like to acknowledge the contributions made by all those involved in the development of the W3C Health Care and Life Sciences community profile. In particular we would like to acknowledge Eric Prud'hommeaux the W3C liaison for the Health Care and Life Sciences Interest Group for his contributions in finalising the formatting of the W3C Interest Group Note.

REFERENCES

Alexander, K., Cyganiak, R., Hausenblas, M., and Zhao, J. (2011). Describing linked datasets with the VoID vocabulary. Interest group note, W3C. http://www.w3.org/TR/void/.

Baungard Hansen, J., Beveridge, A., Farmer, R., Gehrmann, L., Gray, A. J., Khutan, S., Robertson, T., and Val, J. (2015). Validata: An online tool for testing RDF data conformance. In *Semantic Web Applications and Tools for Life Sciences (SWAT4LS2015)*.

Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., Krüger, F. A., Light, Y., Mak, L., McGlinchey, S., Nowotka, M., Papadatos, G., Santos, R., and Overington, J. P. (2014). The ChEMBL bioactivity database: an update. *Nucleic acids research*, 42(Database issue):D1083–1090.

Bradner, S. (1997). Key words for use in RFCs to indicate requirement levels. Best current practice. http://www.ietf.org/rfc/rfc2119.txt.

Brickley, D. and Guha, R. (2014). RDF schema 1.1. Recommendation, W3C. http://www.w3.org/TR/rdf-schema/.

Callahan, A., Cruz-Toledo, J., Ansell, P., and Dumontier, M. (2013). Bio2RDF Release 2: Improved Coverage Interoperability and Provenance of Life Science Linked Data. In *The Semantic Web: Semantics and Big Data*, pages 200–212. Springer.

Ciccarese, P., Soiland-Reyes, S., Belhajjame, K., Gray, A. J. G., Goble, C., and Clark, T. (2013). PAV ontology: Provenance, Authoring and Versioning. *Journal of biomedical semantics*, 4(37).

Cyganiak, R., Wood, D., and Lanthaler, M. (2014). RDF 1.1 concepts and abstract syntax. Recommendation, W3C. http://www.w3.org/TR/rdf11-concepts/.

DCMI Usage Board (2012). DCMI metadata terms. Recommendation, DCMI. http://dublincore.org/documents/dcmi-terms/.

[34] https://github.com/bio2rdf/bio2rdf-scripts/wiki/Bio2RDF-Dataset-Metrics accessed July 2015

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1982v1 | CC-BY 4.0 Open Access | rec: 21 Apr 2016, publ: 21 Apr 2016

Dumontier, M., Baker, C. J., Baran, J., Callahan, A., Chepelev, L., Cruz-Toledo, J., Del Rio, N. R., Duck, G., Furlong, L. I., Keath, N., Klassen, D., McCusker, J. P., Queralt-Rosinach, N., Samwald, M., Villanueva-Rosales, N., Wilkinson, M. D., and Hoehndorf, R. (2014). The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. *Journal of biomedical semantics*, 5(1):14.

Gaudet, P., Bairoch, A., Field, D., Sansone, S.-A., Taylor, C., Attwood, T. K., Bateman, A., Blake, J. A., Bult, C. J., Cherry, J. M., Chisholm, R. L., Cochrane, G., Cook, C. E., Eppig, J. T., Galperin, M. Y., Gentleman, R., Goble, C. A., Gojobori, T., Hancock, J. M., Howe, D. G., Imanishi, T., Kelso, J., Landsman, D., Lewis, S. E., Karsch Mizrachi, I., Orchard, S., Ouellette, B. F. F., Ranganathan, S., Richardson, L., Rocca-Serra, P., Schofield, P. N., Smedley, D., Southan, C., Tan, T. W., Tatusova, T., Whetzel, P. L., White, O., and Yamasaki, C. (2011). Towards BioDBcore: a community-defined information specification for biological databases. *Database*, 2011:baq027–baq027.

Gray, A. J. G. (2013). Dataset descriptions for the open pharmacological space. Working draft, Open PHACTS. `www.openphacts.org/specs/datadesc/`.

Gray, A. J. G., Baran, J., Marshall, M. S., and Dumontier, M. (2015). Dataset descriptions: HCLS community profile. Interest group note, W3C. `http://www.w3.org/TR/hcls-dataset/`.

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., and Williams, A. J. (2014). Applying linked data approaches to pharmacology: Architectural decisions and implementation. *Semantic Web*, 5(2):101–113.

Haak, L. L., Fenner, M., Paglione, L., Pentz, E., and Ratner, H. (2012). ORCID: a system to uniquely identify researchers. *Learned Publishing*, 25(4):259–264.

Jupp, S., Malone, J., Bolleman, J., Brandizi, M., Davies, M., Garcia, L., Gaulton, A., Gehant, S., Laibe, C., Redaschi, N., Wimalaratne, S. M., Martin, M., Le Novère, N., Parkinson, H., Birney, E., and Jenkinson, A. M. (2014). The EBI RDF platform: linked open data for the life sciences. *Bioinformatics (Oxford, England)*, 30(9):1338–9.

Juty, N., Le Novère, N., and Laibe, C. (2012). Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. *Nucleic Acids Res*, 40:D580–6.

Lebo, T., Sahoo, S., and McGuinness, D. (2013). PROV-O: the PROV ontology. Recommendation, W3C. `http://www.w3.org/TR/prov-o/`.

Maali, F. and Erickson, J. (2014). Data catalog vocabulary (DCAT). Recommendation, W3C. `http://www.w3.org/TR/vocab-dcat/`.

Ohno-Machado, L., Alter, G., Fore, I., Martone, M., Sansone, S.-A., and Xu, H. (2015). Biocaddie white paper - data discovery index. White paper, BioCADDIE. `http://dx.doi.org/10.6084/m9.figshare.1362572`.

The W3C SPARQL Working Group (2013). SPARQL 1.1 overview. Recommendation, W3C]. `http://www.w3.org/TR/sparql11-overview/`.

Vandenbussche, P.-Y. and Vatant, B. (2014). Linked Open Vocabularies. *ERCIM News*, 96.

Vasilevsky, N. A., Brush, M. H., Paddock, H., Ponting, L., Tripathy, S. J., LaRocca, G. M., and Haendel, M. A. (2013). On the reproducibility of science: unique identification of research resources in the biomedical literature. *PeerJ*, 1:e148.

WG3 Members (2015). WG3-MetadataSpecifications: NIH BD2K bioCADDIE Data Discovery Index WG3 Metadata Specification v1. Technical report. `http://dx.doi.org/10.5281/zenodo.28019`.

Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C., Tudorache, T., and Musen, M. A. (2011). BioPortal: Enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Research*, 39(SUPPL. 2).

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., 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,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1982v1 | CC-BY 4.0 Open Access | rec: 21 Apr 2016, publ: 21 Apr 2016

438 K., Zhao, J., and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and
439 stewardship Authors. *Nature Scientific Data*, 3. doi:10.1038/sdata.2016.18.

440 FUNDING STATEMENT

441 Funding for Michel Dumontier was provided in part by grant U54 HG008033-01 awarded by NIAID
442 through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative.
443 Alasdair J G Gray was partly funded by the Open PHACTS project an Innovative Medicines Initiative
444 Joint Undertaking under grant agreement number 115191, resources of which are composed of financial
445 contribution from the European Union's Seventh Framework Programme (FP7/2007- 2013) and EFPIA
446 companies' in kind contribution.
447 M. Scott Marshall was funded by the European Commission through the EURECA (FP7-ICT-2012-6-
448 270253) project.
449 Jerven Bollenman: Swiss-Prot group activities are supported by the Swiss Federal Government
450 through the State Secretariat for Education, Research and Innovation