Universität Leipzig
Fakultät für Mathematik und Informatik
Institut für Informatik
Abteilung für betriebliche Informationssysteme

Master Thesis

DataID

Semantically Rich Metadata for Complex Datasets

Abstract: <Gegenstand und Resultate der Arbeit. Was ist neu? Warum sollte man die Arbeit lesen?>

Leipzig, January 2017

Markus Freudenberg
freudenberg@informatik.uni-leipzig.de

Betreuender Hochschullehrer:

Dr.-Ing. Sebastian Hellmann and
Dimitris Kontokostas
Fakultät für Mathematik und Informatik
Betriebliche Informationssysteme, Semantic Web

Acknowledgements

Contents

List of Figures

1 Introduction

1.1 Motivation

In 2006, Clive Humby coined the phrase "the new oil" for (digital) data[1], heralding the ever-expanding realm of what is now summarised as Big Data. Attributed with the same transformative and wealth-producing abilities, once connected to crude oil bursting out of the earth, data has become a cornerstone of economical and societal visions. In fact, the amount of data generated around the world has increased dramatically over the last years, begging the question if those visions have already come to pass.

The steep increase in data produced can be ascribed to multiple factors. To name just a few:

- The growth in content and reach of the World Wide Web[2].
- The digitalising of former analogue data (e.g. [3], [4])
- The realisation of what is called the Internet of Things (IoT)[5].
- The shift of classic fields of research and industry to computer-aided processes and digital resource management (e.g. digital humanities[6], industry 4.0[7]).
- Huge data collections about protein sequences or human disease taxonomies are established in the life sciences[8].
- Research areas like Natural Language Processing or Machine Learning are generating and refining data[9].

[1] https://www.theguardian.com/technology/2013/aug/23/tech-giants-data
[2] http://www.internetworldstats.com/emarketing.htm
[3] https://www.loc.gov/programs/national-recording-preservation-board/
[4] https://archive.org
[5] http://siliconangle.com/blog/2015/10/28/page/3\#post-254300
[6] http://www.dh.uni-leipzig.de/wo/
[7] http://www.europarl.europa.eu/RegData/etudes/STUD/2016/570007/IPOL_STU(2016)570007_EN.pdf
[8] https://www.ncbi.nlm.nih.gov/genbank/statistics/
[9] http://archive.ics.uci.edu/ml/

- In addition, open data initiatives like the Open Knowledge Foundation[10] are following the call for 'Raw data, Now!'[11] of Tim Berners-Lee, demanding open data from governments and organisations.

But with Big Data comes a big challenge. The increasing deluge of data is submerging data producers and possible consumers in a wave of unfiltered, unstructured and apparently unmanageable information. As a new discipline, data engineering is dealing with the fallout of this trend, namely with issues of how to extract, aggregate, store, refine, combine and distribute data of different sources in ways which give equal consideration to the four V's of Big Data: Volume, Velocity, Variety and Veracity[12].

Datasets are the building blocks of these endeavours. They are the combination of multiple data points (datums) bundled together by at least one dimension of distinction (such as source, topic or temporal information). When working with these chunks of data, extra data about data (or metadata) is needed. Dataset metadata enables users to discover, understand and (automatically) process the data it holds, as well as providing provenance on how a dataset came into existence. This metadata is often created, maintained and stored in diverse data repositories featuring disparate data models that are often unable to provide the metadata necessary to automatically process the datasets described. In addition, many use cases for dataset metadata call for more specific information than provided by most available metadata vocabularies. Extending existing metadata models to fit these scenarios is a cumbersome process resulting often in non-reusable solutions.

One vocabulary for dataset metadata is breaking this trend. Since its introduction in 2013, the Data Catalog Vocabulary [1], has been widely adopted as a foundation for dataset metadata in research, government and industry[13]. The very general approach adopted by the authors of DCAT allows for portraying any given (digital) object with this ontology. Extending DCAT is very easy and mappings to other metadata formats are not difficult to achieve.

Conversely, the general approach of DCAT is often too imprecise where specificity is needed, resulting in:

- Insufficient provenance information
- Missing relations between Datasets
- Relations to agents are too cursory

[10] https://okfn.org
[11] http://www.wired.co.uk/news/archive/2012-11/09/raw-data
[12] http://www.ibmbigdatahub.com/infographic/four-vs-big-data
[13] https://joinup.ec.europa.eu/sites/default/files/isa_field_path/2016-05-13_dcat-ap_intro_v0.05.pdf

- Technical description of resources on the Web (e.g. API endpoints) is lacking, restricting the accessibility of the data
- General lack of specificity, inviting non-machine-readable expressions of resources

Similar findings were concluded at the W3C/VRE4EIC workshop 'Smart Descriptions & Smarter Vocabularies' (SDSVoc) in 2016 [2].

add dbpedia

As a result of lacking specificity, current representations of datasets with DCAT are often not contributing to the main benefits of publishing data on the Web: *"Reuse, Comprehension, Linkability, Discoverability, Trust, Access, Interoperability and Processability"* [3]. This, in turn, amplifies broader problems with published datasets, especially in the open data community, reflected by the Open Data Strategy[14], defining the following six barriers for "open public data" [15], proposed by the European Commission in 2011:

1. a lack of information that certain data actually exists and is available,
2. a lack of clarity of which public authority holds the data,
3. a lack of clarity about the terms of re-use,
4. data made available in formats that are difficult or expensive to use,
5. complicated licensing procedures or prohibitive fees,
6. exclusive re-use agreements with one commercial actor or re-use restricted to a government-owned company.

Many issues with DCAT itself or their manifestation in reality can be solved by existing ontologies, even when restricted only to W3C recommended ontologies. For example, the PROV Ontology [4], deals with questions on how to record provenance information on a very granular level. While the Open Digital Rights Language [5] provides machine readable descriptions of licenses and other policies. The existence of problems, like those listed above, despite these offered solutions, speaks to a larger problem of missing organisational structures for landscaping of vocabularies (offering recommendations on combining, revising and usage of ontologies). A study of 91 commonly used vocabularies concluded:

explain what that means

> "Our validation detected a total of 6 typos, 14 missing or unavailable ontologies, 73 language level errors, 310 instances of ontology namespace violations and 2 class cycles which we believe to be errors." [6]

[14] http://europa.eu/rapid/press-release_IP-11-1524_en.htm?locale=en
[15] http://europa.eu/rapid/press-release_MEMO-11-891_en.htm

These errors accumulate when strong interdependencies exist between vocabularies, adding logical and practical problems and aggravating unification issues of ontologies.

1.2 Objectives

In this thesis, I will present the metadata model of **DataID**, a multi-layered metadata ecosystem, which, in its core, describes complex datasets and their different manifestations, their relations to other datasets and agents (such as persons or organisations) endowed with rights and responsibilities.

Improving the portrayal of **provenance**, **licensing** and **access**, while maintaining the easy **extensibility** and **interoperability** of DCAT, are the linchpin objectives in my effort to present a comprehensive, extensible and interoperable metadata vocabulary. Multiple well established ontologies (such as PROV-O, VOID and FOAF) are reused for maximum compatibility to establish a uniform and accepted way to describe and deliver dataset metadata for arbitrary datasets and to put existing standards into practice.

The **DataID Ecosystem** is a suite of ontologies comprised of **DataID core** and multiple extension ontologies, clustered around **DataID core**. It is the result of a modularisation process, which was necessary to preserve **extensibility** and **interoperability** of the DCAT vocabulary, on which all ontologies are based.

I want to present my solution for most of the current problems with dataset metadata in general and DCAT in particular, following these objectives:

1. Provide sufficient support for extensive and machine-readable representations for **provenance**, **licensing** and data **access**.
2. Extend DCAT with well-established ontologies to resolve the discussed issues.
3. Show, that by modularising into a landscape of ontologies, **DataID** preserves the general character of DCAT, supporting **extensibility** and **interoperability**.
4. Prove that the resulting ecosystem is capable of serving for complex demands on dataset metadata (proving **extensibility**).
5. Demonstrate the **interoperability** with other metadata formats.

6. Evaluate the universal applicability of **DataID** for datasets against common demands on data publications.

In addition, **DataID** shall support the FAIR Data principles [7] (section 2.1.5) as well as the best practices defined by the Data on the Web Best Practices working group [3] of the W3C (restricted to those practices where metadata is of concern).

DataID was developed under the sponsorship of the H2020 project ALIGNED[16] (GA-644055), following its main goals:

- to be part of a unified software and data engineering process;
- describing the complete data lifecycle and domain model;
- with an emphasis on quality, productivity and agility.

In the context of ALIGNED, **DataID** is part of a shared model of software and data engineering to enable unified governance and coordination between aligned co-evolving software and data lifecycles:

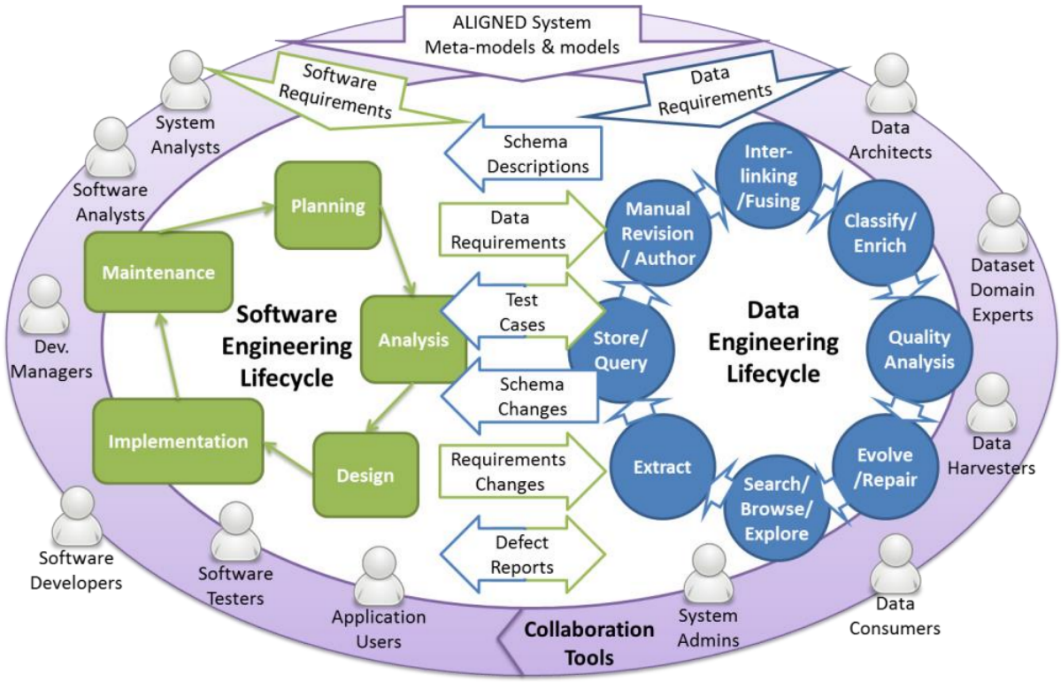

Figure 1.1: ALIGNED Software and Data Engineering Processes

add a last sentence?

[16] http://aligned-project.eu

1.3 Structure

This work is a comprehensive introduction to **DataID**, with a particular focus on the **DataID core** ontology, at the heart of the **DataID Ecosystem**. It is largely based on four publications:

1. Martin Brümmer, Ciro Baron, Ivan Ermilov, Markus Freudenberg, Dimitris Kontokostas and Sebastian Hellmann "DataID: Towards Semantically Rich Metadata for Complex Datasets". In: Proceedings of the 10th International Conference on Semantic Systems. SEM '14. Leipzig, Germany: ACM, 2014, pp. 84–91. [8] An introduction to the first version of **DataID**.
2. Monika Solanki, Bojan Bozic, Markus Freudenberg, Dimitris Kontokostas, Christian Dirschl and Rob Brennan "Enabling Combined Software and Data Engineering at Web-Scale: The ALIGNED Suite of Ontologies". In: The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part II. 2016, pp. 195–203. [9]: An overview of the landscape of ontologies developed by the ALIGNED project.
3. Markus Freudenberg, Martin Brümmer, Jessika Rücknagel, Robert Ulrich, Thomas Eckart, Dimitris Kontokostas and Sebastian Hellmann "The Metadata Ecosystem of DataID". In: Metadata and Semantics Research: 10th International Conference, MTSR 2016, Göttingen, Germany, November 22-25, 2016, Proceedings. Ed. by Emmanouel Garoufallou et al. Cham: Springer International Publishing, 2016, pp. 317–332. I S B N : 978-3-319-49157-8. [10]: An introduction the **DataID Ecosystem**
4. DataID core Ontology (2017): A W3C member submission of the University of Leipzig, under review by the W3C at the time of writing, authored by Martin Brümmer and me.

add ref

After a look at related work (chapter 3) on the subject of dataset metadata, I will present the **DataID Ecosystem** in chapter 4, to introduce the guiding principles of this work. Chapter 5 describes the **DataID core** ontology in detail, containing a running example of a DBpedia language edition. Chapter 6 provides a best practice about publishing data on the Web with **DataID**, followed by an application of those practice to a real example on how to solve complex metadata challenges with the **DataID Ecosystem**, by looking at Data Management Plans (chapter 7). Chapter 8 provides mappings between **DataID** and multiple CMD profiles of the Component MetaData Infrastructure (CMDI). I will evaluate **DataID** in chapter 9 and discuss its development and future work in chapter 10.

what is this?

2 Foundations

2.1 Data

Data is an almost intangible term. It is highly ambiguous and touches many fields of interest, stretching from philosophy to digital signal processing. Even in the context of Information Science, Data has multiple possible definitions. Here are some of them:

> "Data is a symbol set that is quantified and/or qualified." (Prof. Aldo de Albuquerque Barreto, Brazilian Institute for Information in Science and Technology, Brazil [11])
>
> "Data are sensory stimuli that we perceive through our senses." (Prof. Shifra Baruchson–Arbib, Bar Ilan University, Israel [11])
>
> "By data, we mean known facts that can be recorded and that have implicit meaning." (Prof. Shamkant Navathe, College of Computing at the Georgia Institute of Technology, USA [12])
>
> "Etymologically, data [...] is the plural of datum, a noun formed from the past participle of the Latin verb dare–to give. Originally, data were things that were given (accepted as "true"). A data element, d, is the smallest thing which can be recognized as a discrete element of that class of things named by a specific attribute, for a given unit of measure with a given precision of measurement." (Prof. Charles H. Davis, Indiana University, USA [11])
>
> "Data are the basic individual items of numeric or other information, garnered through observation; but in themselves, without context, they are devoid of information." (Dr. Quentin L. Burrell, Isle of Man International Business School, Isle of Man [11])

Information and Data seem to be closely linked and are often used interchangeably, yet they are not the same thing:

> "Datum is every thing or every unit that could increase the human knowledge or could allow to enlarge our field of scientific, theoreti-

> cal or practical knowledge, and that can be recorded, on whichever support, or orally handed. Data can arouse information and knowledge in our mind." (Prof. Maria Teresa Biagetti, University of Rome 1, Italy; based on C. S. Peirce, 1931, 1958 [11])

I will take a broader look at the term Data, delineating it from the concepts of Information and Knowledge.

2.1.1 Data, Information, Knowledge

Data is the result of the application of syntactical rules against a set (sequence, string etc.) of signs or signals, out of which a message between a sender and recipient is constructed [13]. Additional schematics might apply, adding structural data. *Information* can be gleaned from data if one can find meaning in it. The result of this semantic expansion (or interpretation) of *Data* can be weighted by its novelty value or exceptionalness in the context of existing information, using Shannon's entropy [14]. *Data* becomes tokens of perceptions, or more commonly used in Information Science: instances of concepts, capable of changing the understanding of a context for the recipient of the original message. Linking *Information* in a context to determine their interrelations and inferring additional information under the presumption of an intellectual goal, are the processes (*Pragmatics*) turning *Information* into *Knowledge*.

The pyramidal structure depicted (fig. 2.1) is often crowned by an additional field called "Wisdom" [15]. Wisdom could be described as a form of evaluated *Knowledge* or understanding, as the pinnacle of human endeavour or enlightenment. But in the era of fake news, pathological mistrust and truthiness[17], this last step does not seem to follow naturally.

check if we need quotation here

Many authors describe this hierarchy or derivations of it. The following common aspects are presented throughout this literature [15]:

- the key elements are data, information and knowledge,
- these key elements are virtually always arranged in the same order, some models offer additional stages, such as wisdom, or enlightenment,
- the higher elements in the hierarchy can be explained in terms of the lower elements by identifying an appropriate transformation process,
- the implicit challenge is to understand and explain how data is transformed into information and information is transformed into knowledge.

[17] https://en.wikipedia.org/wiki/Truthiness

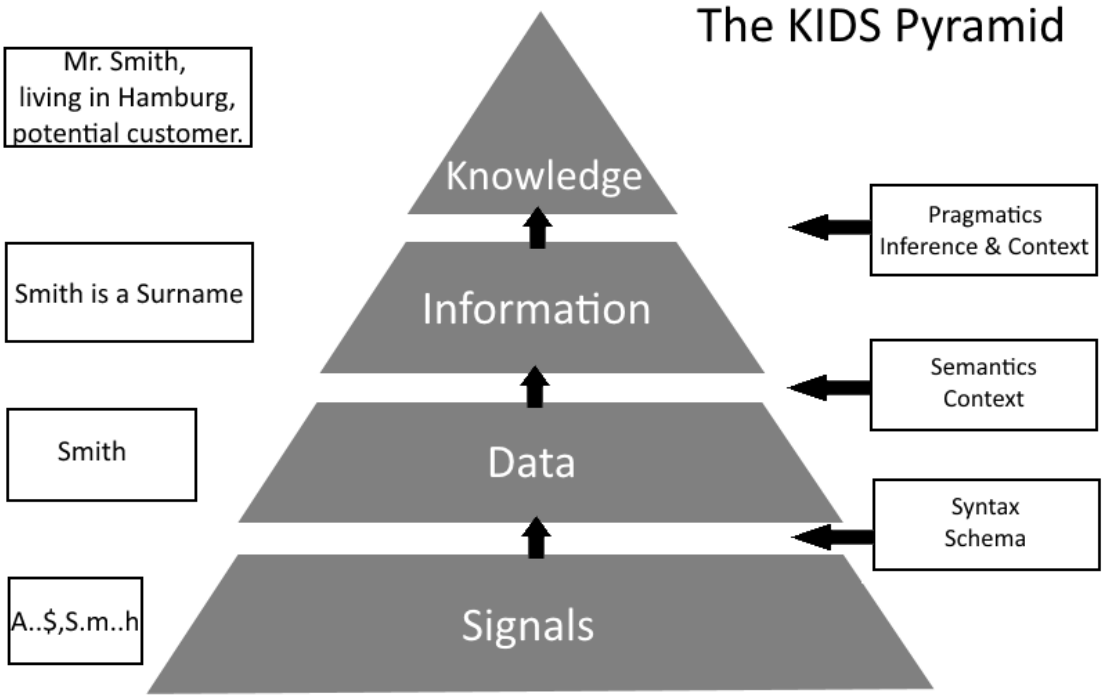

Figure 2.1: The KIDS Pyramid [13]

2.1.2 Digital Data

Digital data is represented using the binary number system of ones (1) and zeros (0). Typically, these are combined into eight of their kind - named Byte. Bytes are used to identify characters of a given alphabet, which, in turn, provide the building stones for any operation, program or data point (datum).

In general, the term Digital Data is used to describes a collection of bytes, representing a digital mapping of an analogue counterpart (e.g. sound waves), or characters of an alphabet understood by humans or programs.

Digital Data is often categorized in structured and unstructured data. Unstructured data does not follow any predefined model and has to be interpreted by the recipient (reader) by its own merit (usually free text). Structured data is strictly adherent to a given data model, facilitating its interpretation by machines and humans alike (such as [please insert]).

2.1.3 Dataset

A dataset[18] is a bundle of data, which have at least one common dimension of distinction. For example, a music album of an artist can be viewed as a dataset, where a single song represents a unit of data. Multiple songs are collected in an album with the common feature (among others) - the artist. Most commonly a dataset corresponds to a collection of structured (digital) data in a single location (e.g. a database table, XML document etc.).

Datasets can manifest in different formats (e.g. in different file types). Therefore, the distinction between dataset as a container for collecting data points of similar content or structure, and its final manifestation on a file system, is advisable.

2.1.4 Metadata

The National Information Standards Organization[19] (NISO), a United States non-profit standards organisation, published a paper in 2004, defining metadata in a widely adopted manner:

> "Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information." [16]

Metadata does not contributes additional content to the original message, but it can ease its transmission, procession and understanding. The meaning of the term metadata and for what kind of data it applies is different, depending on context, disciplines and communities. For example, library catalogue information about a certain book might be understood as metadata in regard to the book itself, while in the context of a Library Management System this type of metadata is considered data.

Specialized types of metadata can be broadly separated into three types [16]:

- **Descriptive metadata** describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords.
- **Structural metadata** indicates how compound objects are put together, for example, how pages are ordered to form chapters.

[18] or 'data set', though this spelling seems to be replaced more and more
[19] `http://www.niso.org/home/`

- **Administrative metadata** provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it. Subsets of administrative metadata include:
  - **Rights management metadata**, which deals with intellectual property rights and licenses.
  - **Preservation metadata**, which contains information needed to archive and preserve a resource.

A metadata record conforms to a given schema, since the use of unstructured data to qualify a different data resource is an exercise in futility. Various metadata schemata or ontologies are available, often describing similar types of data. The most commonly used metadata vocabulary is Dublin Core[20] (DC) by the Dublin Core Metadata Initiative (DCMI):

> "The original objective of the Dublin Core was to define a set of elements that could be used by authors to describe their own Web resources. [...] the goal was to define a few elements and some simple rules that could be applied by noncatalogers [sic]." [16]

The following example illustrates the use of Dublin Core attributes (e.g. `dc:description`) to describe a publication released as an PDF file:

```
dc:title="Metadata Demystified"
dc:creator="Brand, Amy"
dc:creator="Daly, Frank"
dc:creator="Meyers, Barbara"
dc:subject="metadata"
dc:description="Presents an overview of metadata conventions."
dc:publisher="NISO Press"
dc:publisher="The Sheridan Press"
dc:date="2003-07-01"
dc:type="Text"
dc:format="application/pdf"
dc:identifier="http://www.niso.org/standards/resources/metadata.pdf"
dc:language="en"
```

Listing 2.1: Dublin Core example

All Dublin Core attributes are optional, repeatable (non-functional) and present without order. Controlled vocabularies[21] are recommended to be used in connection with some fields (such as `dc:subject`). Since its introduction in 1995,

[20] `http://dublincore.org/documents/dces/`
[21] a set of predefined values, controlled by an institution or a body of experts

the initial list of 15 attributes has been revised and extended, forming the so called DCMI Metadata Terms[22] (DCT). The very general character of this vocabulary provides a useful foundation for more complex schemata reusing DC (for example DCAT).

Metadata can describe any resource in any state or aggregation (single resource, collections, a part of a resource). Any resources, at any abstraction level of a domain model can be substantiated with metadata. The International Federation of Library Associations and Institutions[23] (IFLA) defined the "Functional Requirements for Bibliographic Records" [17], a conceptual model for retrieval and access in online library catalogues and bibliographic databases: **Item** is an exemplar of a **Manifestation**, which embodies an **Expression**, realising a **Work** of an author.

is this a quote?

For example, a metadata record could describe a report, a particular edition of the report, or a specific copy of that edition of the report.

Metadata can be embedded in the described digital object, alongside the object (e.g. in the same directory) or stored separately. HTML documents often keep metadata in the HTML header or completely emerged in the document (see RDFa [18]). While a close coupling between data and metadata is useful when updating them together, a separate approach often simplifies the management of large numbers of records.

The advantages of reliable metadata for digital objects are manifold. These are the more poignant attributions of metadata:

Resource Discovery: Metadata is the foremost source of information which aids agents (persons, software etc.) to discover wanted data.

Organising Electronic Resources: Entities (e.g. books in a library) are organised in catalogues with the help of their digital metadata records.

Interoperability: The interoperability of two digital resources can be determined by comparing their metadata entries, on a syntactical as well as on a semantic level.

Digital Identification: Digital identifiers (such as URIs section 2.2) are stored with metadata records.

Provenance: An extensive record on provenance is a key for trustworthiness, detailing facts like source data, responsible agents or origin activities.

[22] http://dublincore.org/documents/dcmi-terms/
[23] http://www.ifla.org

Quality: The quality of a data source is also established by the quality of its metadata.

Reusability: Increasing discoverability, interoperability, provenance and quality are instrumental requirements for increasing reusability.

Preservation: "Metadata is key to ensuring that resources will survive and continue to be accessible into the future." [16]

In the chain of processes transforming Signals, Data, Information into Knowledge (section 2.1.1), metadata can help with all transformation steps:

- To provide information about schemata to which a set of structured data adheres or the syntax description needed to understand the signs transmitted.
- To advance the interpretation of data by providing context information (e.g. geographical or temporal).
- To point out related (web-) resources to broaden the context of an information (e.g. links to similar datasets or related web site).

2.1.5 The FAIR Data Principles

In 2014, a workshop in Leiden, Netherlands, was held, named "Jointly Designing a Data Fairport". A wide group of academics and representatives of companies and other organisations concluded the workshop by drafting a concise and measureable set of principles to overcome common obstacles, impeding data discovery and reuse of (scientific) data.

The **FAIR** Guiding Principles [7]

F To be Findable:

1 (meta)data are assigned a globally unique and persistent identifier

2 data are described with rich metadata (defined by R1 below)

3 metadata clearly and explicitly include the identifier of the data it describes

4 (meta)data are registered or indexed in a searchable resource

A To be Accessible:

1 (meta)data are retrievable by their identifier using a standardized communications protocol

i. the protocol is open, free, and universally implementable

ii. the protocol allows for an authentication and authorization procedure, where necessary

2 metadata are accessible, even when the data are no longer available

I To be Interoperable:

1 (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

2 (meta)data use vocabularies that follow FAIR principles

3 (meta)data include qualified references to other (meta)data

R To be Reusable:

1 (meta)data are richly described with a plurality of relevant attributes

2 (meta)data are released with a clear and accessible data usage license

3 (meta)data are associated with detailed provenance

4 (meta)data meet domain-relevant community standards

2.2 Semantic Web

I presented our common understanding of how data can herald information and knowledge in section 2.1.1. I refrained from specifying which step or state might be restricted to humans or machines. None of those restrictions would probably be correct. While I don't want to elaborate on the question of: "Can machines have Knowledge?", I do state that machines can glean information from data. To interpret a message and derive meaning is not limited to the human mind. All what is needed is context and an understanding of the concepts a domain is constituted of (semantics).

In 2001, Tim Berners-Lee, Hendler, and Lassila layed out their expectation of how the World Wide Web will eventually extend to become a Semantic Web [19]. The simple extension of web resources with structured, well defined data (or metadata) would give meaning to resources, previously only decipherable by humans. To identify these data resources uniquely by Universal Resource Identifiers (URI) and provide links to other resources, would be the first steps in the direction towards a "web of data that can be processed directly and indirectly by machines". quote?

> "The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning. better enabling computers and people to work in cooperation." [19]

Tim Berners-Lee is the director of the World Wide Web Consortium[24] (W3C), which is responsiblefor the development of standards for the World Wide Web, rewrite and by extension for the Semantic Web. At its core, the Semantic Web is defined by a collection of standards, which are *Recommendations* by the W3C. "Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data." [25].

2.2.1 Resource Description Framework (RDF)

This foundational technology of the Semantic Web and recommendation of the W3C [20] is used to describe any resource, using URIs as identifiers. Resources are defined by set of characteristics which are expressed as attributes or relations to other resources. Statements in the form of "subject, predicate, object" (named "triples") are used to convey such characteristics. The resource described is uniquely identified by the URI of the subject. The object corresponds to the

[24] https://w3.org
[25] https://w3.org/standards/semanticweb/

content or reference of the statement. Literals (strings) are used to serialize content (or values), URIs of resources provide the target of a reference. The predicate is the semantic link between the subject and the object and defines the meaning of this statement. This simple linguistic construct makes RDF data understandable for humans and machines alike:

```
<http://dbpedia.org> publisher "DBpedia Association".
```

This triple describes the resource http://dbpedia.org (identified by the URI of the subject). Its object is the literal "DBpedia Association" connected to the subject with the predicate: publisher. Without the predicate, no meaning could be ascribed to the datum "DBpedia Association", as the type of relation between subject and object would be unknown. Thus, the predicate is what lends meaning to a statement.

The same information about the publisher of a website could be expressed differently in RDF data model. Since "DBpedia Association" represents an organisation, it could be introduced and described as an instance of a concept named "Organisation". This instance is capable of providing multiple attributes, not only its name:

```
<http://dbpedia.org> publisher <http://dbpedia.org/DBpediaAssociation>.
<http://dbpedia.org/DBpediaAssociation> type "Organisation".
<http://dbpedia.org/DBpediaAssociation> name "DBpedia Association".
<http://dbpedia.org/DBpediaAssociation> headquarter "Leipzig".
```

It is obvious that the instantiation of the objects "Organisation" and "Leipzig" would bear the same benefits. By extending this list of statements, describing additional resources and their characteristics, a directed graph is constructed. The data of this graph is highly interlinked and has the ability to relate to resource descriptions outside this graph as well. A URI identifies a resource in the world without ambiguity (when carefully constructed based on existing domain names). This allows linking of data objects without regard to resource locations (datasets, service endpoints etc.) and institutions, creating, what is called, "Linked Data" (section 2.2.3).

While labels like "publisher" have meaning for humans, they are ambiguous and could be misinterpreted. Especially machines cannot resolve this ambiguity and could not infer meaning from such statements. To address this issue, predicates are also identified by URIs that can be looked up for further information. Predicates are called properties in the RDF data model.

```
<http://dbpedia.org> <http://purl.org/dc/terms/publisher> <http://dbpedia.org/
    DBpediaAssociation>.
<http://dbpedia.org/DBpediaAssociation> <http://purl.org/dc/terms/name> "DBpedia
    Association".
```

Sets of properties can be defined and documented by institutions, like DCMI (section 2.1.4). They can then be reused by others, increasing interoperability and reusability. These sets of properties together with associated concepts and annotations are called ontologies.

2.2.2 Web Ontology Language (OWL)

The Web Ontology Language is a W3C recommended standard [21] based on the RDF data model, which could be summarized as an ontology for defining ontologies. An ontology is a set of concepts, properties and logical axioms, with which to model a domain of knowledge or discourse. This conceptualization of a given domain or idea, allows for a formal representation with RDF as well as its automatic interpretation by machines.

Concepts are classes under which all objects of a domain can be classified, dividing up the domain in abstract objects. Properties provide meaning for the links to instances of concepts or literal objects (values), which, in turn, lends meaning to the concepts themselves. Additionally, subclass relations and restrictions on properties (e.g. domain and range definitions) help to specify more complex relations of a domain. This semantic layer between the domain knowledge on one side and its representation in RDF on the other is free of restrictions imposed by underlying technologies, distinguishing it from other data models (e.g. database schemata). A wide range of ontologies are available for any type of domain. Upper ontologies are general use vocabularies (such as Dublin Core section 2.1.4) which can be reused together with other ontologies. High reusability is another difference to many other domain description languages.

Multiple langauge profiles are available [22], introducing different logical regimes to OWL, such as Description Logic (OWL-DL) or based on Rule Languages (OWL2-RL). Profiles allow for reasoning over RDF data, adhering to ontologies under such regimes. The profile OWL Full is an extension of the RDF Schema (RDFS [23]), which provides basic elements for the description of ontologies (allowing for class hierarchies and basic relations).

Description Logic is a fragment of first-order predicate logic [24] and a formalism for representing knowledge. OWL is heavily influenced by Description Logic (projects like DAML+OIL [25]) in order to achieve a beneficial trade-off between language expressiveness and computational complexity of reasoning.

2.2.3 Linked Data

Linked Data is the idea of how data is freed from an islands of unconnected data, so it can be used and referenced all over the world referencing simply their URIs. Links between data objects from different sources are not bound to restrictions, authorisation procedures, licensing or any technical obstacles, they simply state that: "There is a data object (published in a well defined manner - e.g. RDF) which is related and it is identifiable on the Web with this URI". Machines and humans can follow up these links and explore the "Web of Data", expanding the contextual information.

needs link

> "Technically, Linked Data refers to data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external data sets, and can in turn be linked to from external data sets." [26]

Additionally, a set of best practices were contrived by Tim Berners-Lee to identify the necessary steps needed to publish and link structured data on the Web, since "a surprising amount of data isn't linked in 2006, because of problems with one or more of the steps" [27].

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
4. Include links to other URIs. so that they can discover more things.

Linked Data extends the demands of RDF as a data model by specifying HTTP as the access layer of choice and requiring the openness of the resource to be regarded of high quality. The result is a Web of Data, a machine-readable, semantic network of structured data, opposed to the HTML based web of documents (from humans, for humans).

quote

3 Related Work

3.1 Dataset vocabularies

This section is dedicated to dataset metadata vocabularies and application profiles [26], to compare them and list their (dis-) advantages.

Based on the FAIR Data Principles (section 2.1.5) and the list of important aspects of dataset metadata (section 1.1), I contrived the following list of aspects, against which I want to evaluate each vocabulary.

A **The vocabulary encourages the use of richly described and machine-readable resources.** Concepts are defined as exhaustive as necessary to describe all relevant aspects, avoiding free text properties in general. For example, replacing a literal with a well structured instance of `foaf:Agent`.

B **The vocabulary assigns globally unique URIs to metadata resources.** Demanding URIs as identifiers, independent of the chosen data representation (even for non RDF or XML metadata).

C **The vocabulary can describe data access related properties and restrictions, enabling access for humans and machines alike.** Sufficient effort has been made to describe this important aspect in detail, considering all possible formats of a dataset.

D **The vocabulary can portray provenance information extensively.** Dataset provenance can be described extensively, including other datasets (e.g. sources), activities (e.g. data generation activities) and agents (e.g. publisher) as well as inter-relational properties between these concepts.

E **The vocabulary provides for detailed descriptions of rights and licenses.** Machine readable licenses are of utmost importance.

F **The vocabulary provides properties to cite identifiers of the data described.** The possibility to reference the data described directly (by identifier) in the metadata is available.

[26] a set of metadata elements defined for a particular application or other limited purpose, often based on a broader schema (like an ontology)

G **The vocabulary provides for qualified references between resources.** Relations between instances of dataset metadata can be qualified by roles (specifying the type of relations), time and other restrictions.

H **The vocabulary is easy to extend, to fit any given use case.** The vocabulary is general enough to fit any use case, it can easily be extended and no unnecessary restrictions, like restrictive cardinalities, are in place.

I **The vocabulary is unambiguous and easy to map to other metadata vocabularies.** The vocabulary is general enough to be able to match other metadata formats. Properties are defined clearly without overlapping the purpose of others (so users know which property to use).

J **The vocabulary offers additional properties to aid dissemination and discovery.** Extra properties are in place to provide keywords, genres, taxonomy concepts and general statements explaining the use of a dataset.

I will assign one of the following ratings to every item: **(2)** The requirement is supported in full. **(1)** The requirement is partially met. **(0)** The vocabulary does not support this requirement. While this list is helpful for evaluation and comparison purposes, the quality of a dataset vocabulary is also dependent on the intended domain of use and other factors, which I will mention as well.

3.1.1 The Data Catalog Vocabulary (DCAT)

In [28] the authors introduce a standardised interchange format for machine-readable representations of government data catalogues. The Data Catalog Vocabulary (DCAT) is a W3C recommendation [1] and serves as a foundation for many available dataset vocabularies and application profiles. Vocabulary terms for DCAT are inferred from the survey on seven data catalogues from Europe, US, New Zealand and Australia.

> "By using DCAT to describe datasets in data catalogs, publishers increase discoverability and enable applications easily to consume metadata from multiple catalogs. It further enables decentralized publishing of catalogs and facilitates federated dataset search across sites." [1]

DCAT defines three levels of abstraction, based on the following distinctions: A dataset describes a "collection of data, published or curated by a single agent, and available for access or download in one or more formats"[1], and represents the commonalities and varieties of the data held within (the 'idea' or intellectual content of that dataset). A Dataset is part of a data catalogue, representing

multiple datasets (e.g. of an organisation). Datasets manifest themselves (are available) in different forms (such as files, service endpoints, feeds etc.), expressed with the class `dcat:Distribution`. A dataset might be available for download at two different locations on the Web and available to be queried through an API endpoint. This scenario can be described by using three different `dcat:Distribution` instances.

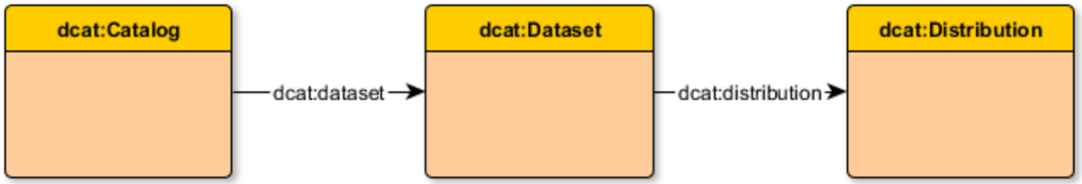

Figure 3.1: The basic idea of DCAT

This basic idea of differentiating between catalogue, dataset and distribution, has prevailed throughout the metadata domain for digital resources (not only datasets), and has become an quasi-requirement for metadata representations of web resources[27]. In fact, the general approach the authors of DCAT took, makes it possible to describe any digital object. This is a highly desirable feature, supporting **`extensibility`** and **`interoperability`** of the vocabulary.

As stated before (section 1.1), a downside to this approach is the possible unspecificity of resources, especially in regard to machine-readability, where uncertainty about formats is problematic.

Insufficient provenance information:

- DCAT expresses provenance in a limited way using a few basic properties such as `dct:source` or `dct:creator`, which can not be further qualified.
- No possibility to specify activities involved in the creation of datasets.
- There is no support or incentive to describe source datasets, related publications or conversion activities of transformations responsible for the dataset. This lack is crucial, especially in a scientific contexts, as it omits the processes necessary to replicate a specific dataset.
- Insufficient portrayal of context information (e.g. licenses, geography etc.)

Missing relations between Datasets:

- in general: Referencing related datasets is only possible on a very generic scale (e.g. `dct:relation`).

[27] `https://www.w3.org/TR/dwbp/#context`

- hierarchical: No inherent portrayal of dataset hierarchies is possible.
- evolutionary: No versioning pointers between dataset representations.

Relations to agents are too cursory:

- A very restricted number of properties pointing out agents, without any further qualification (e.g. `dct:publisher`, `dcat:contactPoint`), other related entities, like software, projects, funding etc., are neglected all together.
- No agent role concept to define new relations.

Technical description of distributions is lacking, restricting the accessibility of the data:

- Only superficial attributes for describing the technical characteristics of a distribution are available.
- No access information are available, such as access restrictions or service description, needed to describe service endpoints.
- No specificity when describing serialisation or media type of a distribution (e.g. file format).

General lack of specificity, inviting non-machine-readable expressions of resources:

- Insufficient specificity of property ranges (e.g.: `dct:license`, `dct:temporal`, `dct:spatial`, `dct:language`, `dcat:mediaType`), thereby neglecting exactness of relevant metadata resources, such as licenses.
- A lack of referential and functional integrity due to missing role-based qualifications of properties (such as `dct:maintainer`) [38].

While this seems to be an extensive list of shortcomings, most of the points listed above are due to the general approach of DCAT. The list merely indicates that portraying dataset metadata with DCAT alone, might not be sufficient for most domains and use cases.

In turn, extending DCAT as upper ontology provides a sufficient basis for any metadata descriptions, without regard to domain or use case. Adopting not only its basic ideas, but the aspects of easy **`extensibility`** and **`interoperability`**, should prove beneficial. Addressing the issues with **`provenance`**, **`licensing`** and **`access`** as well as domain specific demands for metadata is the central task I set out to complete.

The evaluation table is reused in the eval section. The numbers within have to be revisited...

Requirement	A	B	C	D	E	F	G	H	I	J	Sum
Evaluation of DCAT	2	2	0	0	0	1	0	2	2	2	11

3.1.2 Vocabulary of Interlinked Datasets (VoID)

The Vocabulary of Interlinked Datasets (VOID) [29] is widely accepted and used within the Semantic Web community, for instance in projects such as: OpenLink Virtuoso[28] , LODStats[29] , World Bank[30] and others. VOID can be used to express general metadata, access metadata, structural metadata and links between Linked Data datasets. Tools to create VOID metadata are described in [30] where authors also presents techniques of reduction in order to create descriptions for Web-scale datasets. In the same paper the importance of VOID is well established but there is still a lack of important metadata which is not described, for example license and provenance. For simple datasets VOID performs well, which is supported by the fact of wide acceptance of the vocabulary. However, in a case of complex datasets VOID is not expressive enough. In particular, access metadata includes the `void:dataDump` property, which points to the data files of the particular dataset. This property should link directly to dump files as described in W3C Interest Group Note: Describing Linked Datasets with the VOID Vocabulary[31]. Thus, additional semantic information about the data files and the structure of the dataset can not be expressed using VOID. Moreover, VOID does not provide a distribution concept depriving the vocabulary of an important level of abstraction. Yet, this ontology is useful especially for Linked Data datasets, offering many useful statistical properties (such as `void:triples`) and the very handy concept of `void:LinkSet` describing the particular relation between datasets, where one holds links to instances of the other.

Requirement	A	B	C	D	E	F	G	H	I	J	Sum
Evaluation of VOID	2	2	1	1	0	0	1	1	1	1	10

3.1.3 Comprehensive Kerbal Archive Network (CKAN)

Metadata models vary and most of them do not offer enough granularity to sufficiently describe complex datasets in a semantically rich way. For example, CKAN[32] (Comprehensive Knowledge Archive Network), a data management

[28] http://virtuoso.openlinksw.com/
[29] http://stats.lod2.eu/
[30] http://worldbank.270a.info/
[31] http://www.w3.org/TR/void/
[32] http://ckan.org/

system used widely in (Open) Data Portals (such as datahub.io[33]) to provide web representations for datasets, operates on a JSON based schema[34] developed by the Open Knowledge Foundation[35]. CKAN allows simple access to a whole range of functions related to the management of datasets (such as search and faceting of data-sources) accessible via a REST interface. Its metadata schema has some similarities with DCAT: Datasets are collected under organisation objects, which are used as s primitive stand in for catalogues. Datasets have 'Resources' which assume a similar role as a distribution in the DCAT vocabulary.

Alas, there is no clear definition of the resource-object within CKAN documentation, nor are there any noteworthy restrictions. This has led to a medley of different use cases for this resource, where it assumes the role of a `dcat:Distribution` in one dataset (containing all data of the dataset) and providing different slices of the data in another example [31]. Furthermore, the extensive use of key-value pairs for additional data led to a shier host of (semi-) structured data, with only marginal agreement on key names between different Data Portals [31] adding to the general unclarity of this metadata format, complicating the mapping to other vocabularies.

This data model is semantically poor and inadequate for most applications consuming data automatically. I strongly discourage any organisation from adopting the CKAN format for their dataset metadata. Since it is used so frequently in Data Portals, I feel obliged to point out that there are mapping tools for most vocabularies to CKAN (as I provided for DataID in [8]).

Requirement	A	B	C	D	E	F	G	H	I	J	Sum
Evaluation of CKAN	2	1	0	0	0	0	0	0	1	1	5

3.1.4 Metashare

The META-SHARE ontology[32] is the offspring of a prior, XSD[36] based "metadata schema that allows aspects of [language resources] accounting for their whole lifecycle from their production to their usage to be described"[32]. META-SHARE differentiates between language resources (basically datasets with a language related purpose - text, audio etc.), technologies (e.g., tools, services) used for their processing and additional entities like reference documents, agents, projects or licenses. This allows for the portrayal of provenance in the

[33] `http://datahub.io/`

[34] `https://github.com/KSP-CKAN/CKAN/blob/master/CKAN.schema`

[35] `https://okfn.org`

[36] XML Schema Definition: a W3C recommendation for how to formally describe the elements in an XML document (https://www.w3.org/TR/xmlschema11-1/)

domain of Natural Language Processing. In addition it offers an exemplary way of describing licenses and terms of reuse[37]. Yet, META-SHARE is highly specialised for language resources, thus lacking generality and extensibility for other use cases. While not implementing the DCAT vocabulary, META-SHARE does provide an almost complete mapping to DCAT. Mappings to other ontologies might prove difficult, due to the large size of the vocabulary and the often employed (and ample) controlled vocabularies. The related META-SHARE XSD schema has been implemented in the META-SHARE web portal[38], providing many NLP related datasets for download.

Requirement	A	B	C	D	E	F	G	H	I	J	Sum
Evaluation of META-SHARE	2	2	1	1	2	0	0	0	1	1	10

3.1.5 Asset Description Metadata Schema (ADMS)

The Asset Description Metadata Schema[39] (ADMS) is a profile of DCAT, which is specialised to describe "Semantic Assets". Assets (as subclass of `dcat:Dataset`) are highly reusable metadata (e.g. code lists, XML schemata, taxonomies, vocabularies etc.) expressing the intellectual content of the data, which is represented (in most cases) in relatively small files.

ADMS adopts the DCAT structure and provides a well defined way of versioning between entities. Its specialised nature makes it unsuited for a broader approach of portraying datasets (as intended by the authors), but it can still contribute useful properties to DCAT based vocabularies (e.g. section 3.1.6). Since ADMS does not impose any restrictions it can be extended to DCAT without any consequences for DCAT based metadata documents. The evaluation below, therefore does not differ from DCAT.

Requirement	A	B	C	D	E	F	G	H	I	J	Sum
Evaluation of ADMS	2	2	0	0	0	1	0	2	2	2	11

[37] http://www.cosasbuenas.es/static/ms-rights/
[38] http://www.meta-share.org
[39] https://www.w3.org/TR/vocab-adms/

3.1.6 DCAT Application Profile for data portals in Europe (DCAT-AP)

The DCAT Application Profile for data portals in Europe[40] (**DCAT-AP**) is a specification based on **DCAT** (extended with **ADMS** properties) for describing public sector datasets in Europe. It was developed by a working group under the auspices of the European Commission. Its basic use case is "to enable cross-data portal search for datasets and to make public sector data better searchable across borders and sectors" [33]. This can be achieved by the exchange of descriptions of datasets among data portals.

Traits of the resulting profile[41] (version 1.1) released in October 2015 [34]:

- It proposes mandatory, recommended or optional classes and properties to be used for a particular application;
- It identifies requirements to control vocabularies for this application;
- It gathers other elements to be considered as priorities or requirements for an application such as conformance statement, agent roles or cardinalities.

DCAT-AP has been endorsed by the Standards Committee of ISA2[42] in January of 2016[43] for the use in data portals. Further, it has been implemented by over 15 open data portals in the European Union, including the European Data Portal[44].

In general, while some recommendation are in place (e.g. using **ODRL** license documents - section 3.2.2), **DCAT-AP** can not propose concrete improvements in extending **DCAT** to advance **`provenance`**, **`licensing`** or **`access`**. As remarked in section 7 of its specification[45], the representation of different agent roles is lacking in the current version of **DCAT-AP**. In my opinion, the second solution proposed within, using **PROV-O** (Provenance Ontology - see section 3.2.1), is the most comprehensive way of resolving this issue. Due to some cardinality restrictions (e.g. those on `dcat:accessURL`) and its specialisation for data portals, extending **DCAT-AP** to serve more elaborate purposes, can pose challenges.

Requirement	A	B	C	D	E	F	G	H	I	J	Sum
Evaluation of **DCAT-AP**	2	2	1	0	0	1	0	1	2	1	10

[40] https://joinup.ec.europa.eu/asset/dcat_application_profile/home
[41] https://joinup.ec.europa.eu/catalogue/distribution/dcat-ap-version-11
[42] https://ec.europa.eu/isa/isa2/index_en.htm
[43] https://joinup.ec.europa.eu/community/semic/news/dcat-ap-v11-endorsed-isa-committee
[44] https://www.europeandataportal.eu/
[45] https://joinup.ec.europa.eu/catalogue/distribution/dcat-ap-version-11

3.1.7 The HCLS Community Profile

The W3C interest group Semantic Web for Health Care and Life Sciences (HCLS) represents many stakeholders of the Life Sciences, seeking to "develop, advocate for, and support the use of Semantic Web technologies across health care, life sciences, clinical research and translational medicine" [35].

Their community profile (an ongoing effort by this W3C interest group[46]), extends DCAT with versioning and detailed summary statistics, through a three component model. This model introduces an additional abstraction level between dataset and distribution, the so called 'Version Level Description', which contains version specific properties (e.g. `dct:isVersionOf`). The profile is structured in multiple modules, dealing with different levels of specificity [36]:

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.

The HCLS profile reuses 18 vocabularies with 61 properties [36], covering many goals of my evaluation. The chosen approach of this profile is sound and achieves qualitatively good metadata, with an emphasis on FAIR Data principles.

One problem is the large number of reused vocabularies with overlapping purposes, which cause difficulties when mapping to other vocabularies. Its cardinality restrictions can pose problems when extending the profile to use cases, especially outside the health care domain. Although, efforts were made to cover provenance in general, problems like qualifying otherwise static relations to agents or datasets can not be solved by the incorporated vocabularies. The approach to portray licenses does not improve DCAT.

A specific problem is the use of the property `dcat:accessURL`, which, according to the profile, could be used on the dataset abstraction level ('Summary Level Description'). This clearly violates the specification of DCAT. In general the

[46] `https://www.w3.org/blog/hcls/`

boarder between `dcat:Dataset` and `dcat:Distribution` has to be defined more carefully when adding an additional layer in between.

Requirement	A	B	C	D	E	F	G	H	I	J	Sum
Evaluation of HCLS Profile	2	2	1	1	0	2	0	1	1	1	11

3.1.8 CERIF

The metadata format CERIF[47] (Common European Research Information Format) is a metadata development, which began in the early 1990s. The shortcomings of its first approach were addressed by CERIF2000, which became a EU recommendation for its member states for research data. Since 2002 CERIF is further developed by the European Current Research Information Systems[48].

CERIF provides generalized base concepts and relations between them, based on an entity-relationship[49] approach [37]. Relations (or 'linking entities') are qualified with roles, temporal and spacial statements, supplemented with provenance and versioning information. In contrast to Dublin Core based metadata standards (e.a. DCAT), CERIF was developed with a particular focus on referential and functional integrity of resources, avoiding ambiguity in interpretation [38]. CERIF provides much more than just metadata for datasets, it addresses metadata needs for the whole science community, describing projects, funding, facilities, organisations and other.

This paper lines out the main characteristics of CERIF metadata [38]:,

- it separates clearly base entities from relationships between them and thus represents the more flexible fully-connected graph rather than a hierarchy;
- it has generalised base entities with instances specialised by role (e.g. <person> rather than <author>), in the linking entities;
- it handles multilinguality by design and temporal information (so representing versions) to the appropriate attribute treated as an entity (example <title> linked to <publication>);
- temporal information is in the link entities not the base entities (e.g. employment between two dates is in the linking relation between <person> and <organisation> and not an attribute of either of the base entities);

[47] `http://www.eurocris.org/cerif/main-features-cerif`

[48] `http://www.eurocris.org`

[49] Entity–relationship model: describes inter-related things of interest in a specific domain of knowledge. An ER model is composed of entity types (which classify the things of interest) and specifies relationships that can exist between instances of those entity types.

- the temporal information in linking entities provides provenance and versioning recording (e.g. versions of datasets and – in the associated role attribute – the method of update or change);
- CERIF separates the semantics into a special 'layer' which is referenced from CERIF instances. The sematic layer includes permissible values for roles in any linking entity and for controlled values of attributes in base entities (e.g. ISO country codes). Thus semantic terms are stored once and referenced many times (preserving integrity).

CERIF offers a comprehensive approach for solving a host of metadata demands in a sophisticated manner. The chosen abstraction levels (layers) are appropriate and the adherence to an entity-relationship approach is arguably a working solution for qualifying relations. The main problem with CERIF is the complexity of its ontology[50] [51] together with the unique approach to metadata (unlike the DCAT based understanding of metadata). The imposes hurdles when studying, mapping and extending this ontology. Furthermore, the ontology proves to be not specific enough when dealing with information on access and licenses.

Requirement	A	B	C	D	E	F	G	H	I	J	Sum
Evaluation of CERIF	2	2	0	2	0	2	2	1	1	1	13

3.1.9 DataID version 1.0.0

The common shortcomings of dataset vocabularies revealed in this section were also afflicting the previous version (1.0.0) of the **DataID** ontology [8]. Rooted in the Linked Data world, it neglected important information or provided properties (e.g. `dataid:graphName`) which are orphans outside this domain.

While it already imported the important Provenance Ontology (PROV-O-section 3.2.1), to cover the general issues with **provenance**, it was lacking in regard to specificity of **access** and **licensing**. The narrow definition of datasets (i.e. restricted to Linked Data datasets) was inadequate for use cases outside this domain and so inhibited **extensibility**.

Requirement	A	B	C	D	E	F	G	H	I	J	Sum
Evaluation of DataID 1.0.0	2	2	0	2	1	0	0	1	2	1	11

[50] http://eurocris.org/ontologies/cerif/1.3/
[51] http://eurocris.org/ontologies/semcerif/1.3/

3.2 Secondary Literature

This section proffers a collection of associated literature, not directly touching on the subject of dataset metadata. Many subjects, such as representation of licenses and data quality, are relevant for providing metadata of datasets.

3.2.1 The Provenance Ontology (PROV-O)

> "Provenance is defined as a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing. In particular, the provenance of information is crucial in deciding whether information is to be trusted, how it should be integrated with other diverse information sources, and how to give credit to its originators when reusing it. In an open and inclusive environment such as the Web, where users find information that is often contradictory or questionable, provenance can help those users to make trust judgements." [39]

The Provenance Ontology[4] (PROV-O) is a widely adopted W3C recommended standard and serves as a lightweight way to express the provenance and interactions between activities, agents and entities (e.g. datasets).

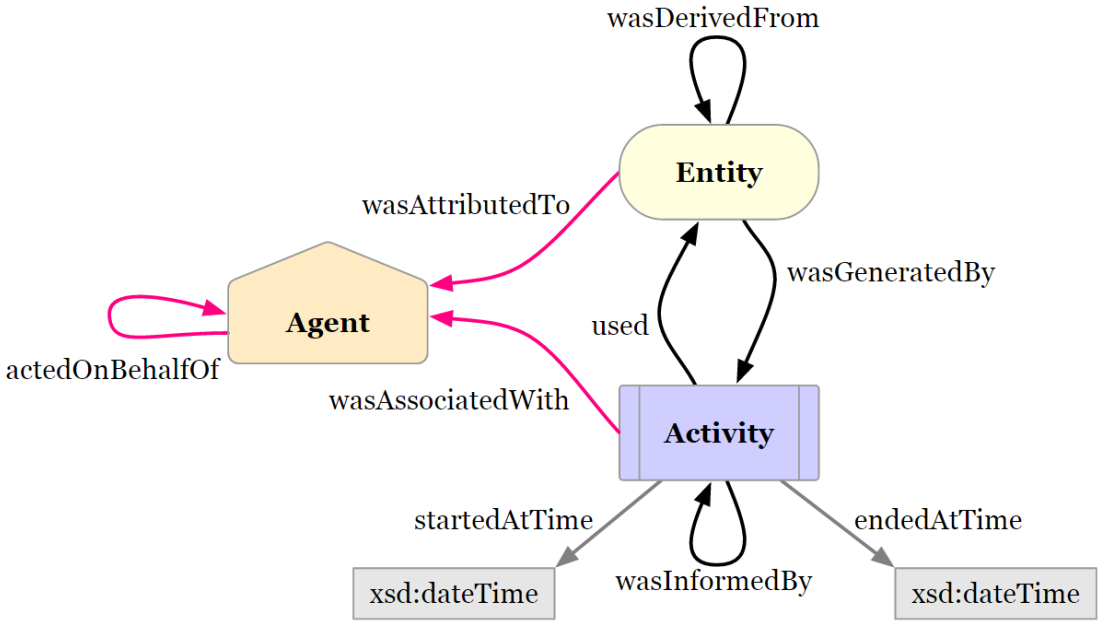

Figure 3.2: Linchpin of the Provenance Ontology: Entities, Agents, Activities [4]

In the context of datasets, provenance information on the activities which helped to create a dataset, which agents were involved in these processes and who to contact about it as well as source datasets or other entities involved are of interest, especially when trying to determain the trustworthiness of data. This ontology is "[...] the foundation to implement provenance applications in different domains that can represent, exchange, and integrate provenance information generated in different systems and under different contexts." [4]

Each of the relations depicted (fig. 3.2) have suitable qualification classes (e.g. `prov:Association`), so that a property (`prov:wasAssociatedWith`) can be inferred from the qualified property path (`prov:qualifiedAssociation/prov:agent`). Qualification classes provide qualification via properties such as `prov:role`.

In this example (from the official specification of PROV-O [4]), an association provides additional description about the :illustrationActivity that an agent named 'Derek' influenced:

```
@prefix prov:                   <http://www.w3.org/ns/prov#> .
@prefix :                       <http://example.org#> .

:illustrationActivity
   a prov:Activity;                  ## this illustration activity was
   prov:wasAssociatedWith :derek;    ##   associated with Derek in some way.
.

:derek a prov:Agent .

:illustrationActivity
   prov:qualifiedAssociation [       ## Qualify how the :illustrationActivity
      a prov:Association;            ##   was associated with the Agent Derek.
      prov:agent   :derek

      prov:hadRole :illustrationist; ## Qualification: The role that Derek served
      prov:hadPlan :tutorial_blog;   ## Qualification: The plan (or instructions)
                                     ##   that Derek followed when creating
   ].                                ##   the graphical chart.

:tutorial_blog   a prov:Plan, prov:Entity .
:illustrationist a prov:Role .
```

PROV-O will have a central role in the creation of DataID. Further reading on the key requirements, guiding principles, and design decisions which influenced the PROV Family of Documents[52] is advised (see [40]).

[52] https://www.w3.org/TR/2012/WD-prov-overview-20121211/

3.2.2 Open Digital Rights Language (ODRL)

The Open Digital Rights Language (ODRL)[53] is an initiative of the W3C community group with the same name[54], aiming to develop an open standard for policy expressions. The ODRL version 2.0 core model defines licensing policies in regard to their permissions granted, duties and constraints associated with these permissions as well as involved legal parties. Thus, an ODRL description allows to specify, in a machine-readable way, if data can be edited, integrated or redistributed.

3.2.3 Lexvo.org

Lexvo.org[55] is a service publishing language related information as Linked Data on the Web. The data published is conform to the Lexvo Ontology, providing unique identifiers for human languages in the context of geography, language families, words and word senses, scripts and characters [41]. All in all, the Lexvo dataset consists of over 8000 languages with a broad spectrum of language-related information that is extensively used by many data publishers and communities.

3.2.4 DataCite Ontology

DataCite is a "global non-profit organisation that provides persistent identifiers (DOIs) for research data [with the goal] to help the research community locate, identify, and cite research data with confidence"[56].

revise citation

DataCite published an XML schema for describing and citing research data[57], which is elaborate and at times novel approach for providing dataset metadata. Yet, due to its rigid XML structure with many cardinality restrictions, it does not feature in my collection of dataset metadata in this chapter.

Of more interest, is the DataCite ontology which was published as an OWL ontology[58] and has a particular focus on representing identifiers (fig. 3.3). The Identifier class is divided into ResourceIdentifier and AgentIdentifier. In addition, an IdentifierScheme defines the format of the literal which represents the

[53] https://www.w3.org/ns/odrl/2/ODRL21
[54] https://www.w3.org/community/odrl/
[55] http://www.lexvo.org
[56] https://www.datacite.org/mission.html
[57] http://schema.datacite.org/meta/kernel-4.0/
[58] http://www.sparontologies.net/ontologies/datacite/source.html

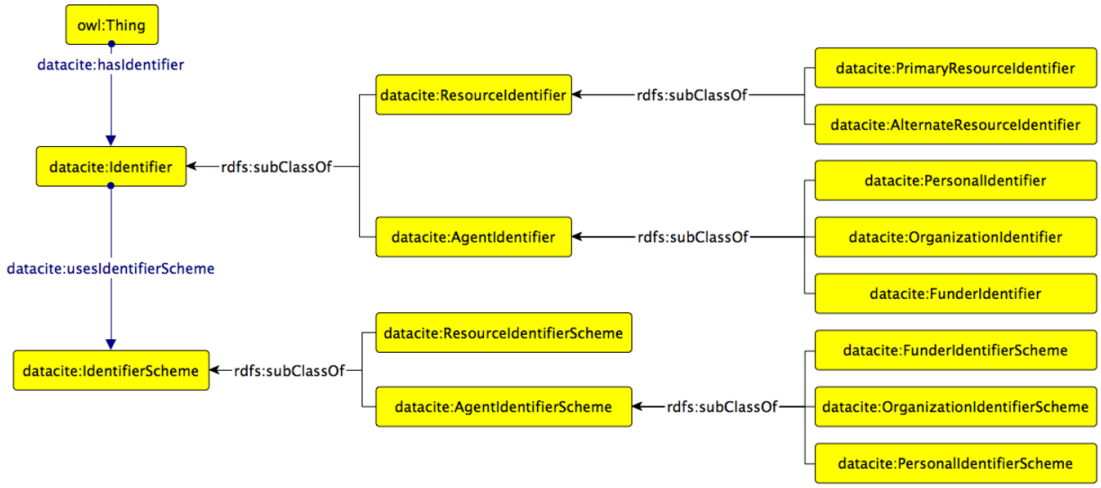

Figure 3.3: DataCite Identifier and IdentifierScheme

identifier. As opposed to an approach with data-types, this ontology allows for adding new schemes without altering the vocabulary itself and adding additional qualifications to the scheme entity [42]. Furthermore, the DataCite ontology contains a multitude of predefined IdentifierScheme instances, ready to be used.

3.2.5 DBpedia

DBpedia is a community effort to extract structured information from Wikipedia and to make this information available on the Web, where it is used as a common point of reference for interlinking and enriching most of the structured data on the web today, establishing it as the center of the so called 'LOD cloud'.

this section needs polishing and references!

The main focus of the DBpedia extraction lies in mapping of info boxes, templates and other easily identified structured data found in Wikipedia articles to properties of an ontology. The DBpedia ontology is reflecting the relations and classifications found within Wikipedia. Together with the mappings to the Wikipedia XML templates, this ontology is curated by a community of interested people around the world to update schematical changes in time for the bi-yearly DBpedia releases.

Each release consists of around one hundred datasets for each of its 130 (2016) language editions (reflecting most of the languages of Wikipedia). Datasets are often created around particular properties (such as `rdf:type`) or other distinguishing features. To reflect this large corpus of structured data, the need for specific demands on its metadata are self-evident.

probably gets another field on the data quality vocabulary...

4 The DataID Ecosystem

4.1 Problem Statement

The inadequacies of current metadata vocabularies are manifold and diverse (section 3.1). As already introduced (section 1.1), there are some issues which protrude from the rest, due to their ubiquitousness in use cases or their import on aspects like interoperability, trustworthiness and governance of data.

This list of important aspects of metadata reflects these issues and explains them in detail:

(A1) **`provenance`**: a crucial aspect of data, required to assess correctness and completeness of data conversion, as well as the basis for trustworthiness of the data source (no trust without provenance).

(A2) **`licensing`**: machine-readable licensing information provides the possibility to automatically publish, distribute and consume only data that explicitly allows these actions.

(A3) **`access`**: publishing and maintaining this kind of metadata together with the data itself serves as documentation benefiting the potential user of the data as well as the creator by making it discoverable and crawlable.

(A4) **`extensibility`**: extending a given core metadata model in an easy and reusable way, while leaving the original model uncompromised expands its application possibilities fitting many different use cases.

(A5) **`interoperability`**: the interoperability with other metadata models is a hallmark for a widely usable and reusable dataset metadata model.

When regarding aspects **(A4)** and **(A5)**, taking into account the intricate requirements of many use cases (such as in **??**), **`extensibility`** and **`interoperability`** seem contradictory when leaving the more general levels of a domain description. A vocabulary capable of interacting with other metadata vocabularies might be too general to fit certain scenarios of use. Restrictive extensions to a vocabulary might encroach on its ability to translate into other useful metadata formats. This notion is corroborated by this document [43].

update ref

Note: I do not differentiate between **evolvability** and **extensibility** (as done in this source) in the context of this thesis. The discrepancies with **interoperability** are true for both concepts. Letting features 'die out' over time does not impact, in my understanding, the aspect of **extensibility**.

We conclude, not only is there a gap between existing dataset metadata vocabularies and requirements thereof, but it seems unlikely that we are able to solve all these diverse problems with just one, monolithic ontology.

4.2 The multi-layer ontology of DataID

While trying to solve the different aspects, which we discussed in the previous section, and tending to the needs of different usage scenarios, the DataID ontology grew in size and complexity.

In order to keep the **DataID** ontology reasonable in size and complexity as well as not to jeopardise **extensibility** and **interoperability**, we modularised **DataID** in a core ontology and multiple extensions. The onion-like layer model (fig. 4.1) illustrates the import dependencies between the ontologies:

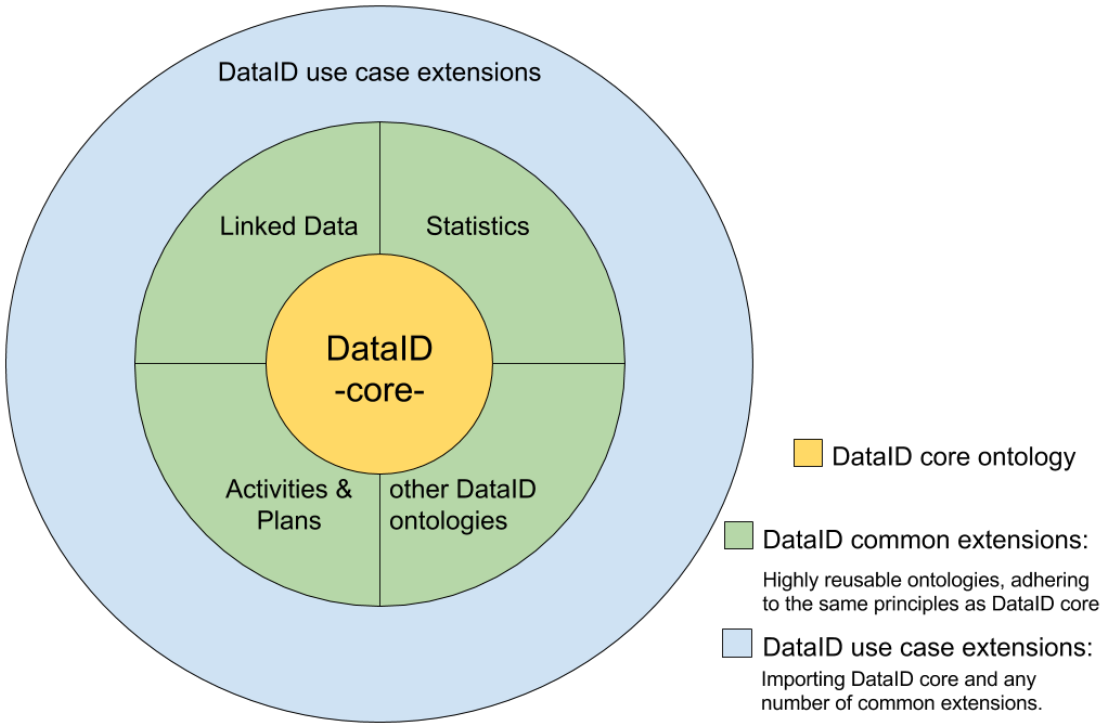

Figure 4.1: The Metadata Ecosystem of DataID

The scaling approach used to modularize the original **DataID** ontology adopts principles of the modular programming technique, separating concepts and properties of a large ontology into independent, interchangeable modules, specialised to fit common use cases, dependent only on **DataID core**. Thus, any vocabulary in this sphere must import **DataID core**, with the exception of **DataID core** itself. The mid-layer (or common extensions) of this model is comprised of highly reusable ontologies, extending **DataID core** to cover additional aspects of dataset metadata. Non of these are mandatory imports for any ontology, yet, in many cases some or all of them will be useful contributions. While I will not (and can not) impose any restrictions as to which ontologies not to import, ideally, ontologies of this layer should only import **DataID core**, to minimize discrepancies between mid-layer ontologies of different authors with overlapping purposes. The outermost layer of this sphere represents all vocabularies importing **DataID core** and any number if mid-layer ontologies and adding additional semantics to portray domain or use case specific demands for metadata.

The **DataID Ecosystem** is a suite of ontologies comprised of **DataID core** and its extensions, which were created to satisfy different use cases in a reusable manner:

DataID core provides the basic description of a dataset (cf. **??**) and serves as foundation for all extensions to DataID.

Linked Data extends DataID core with the VOID vocabulary[44] and some additional properties specific to LOD datasets. Many VOID and Linked Data references from the previous version of DataID were outsourced into this ontology[59].

Activities & Plans provides provenance information of activities which generated, changed or used datasets[60]. The goal is to record all activities needed to replicate a dataset as described by a DataID. Plans can describe which steps (activities, precautionary measures) are put in place to reach a certain goal. This extension relies heavily on the PROV-O ontology[4].

Statistics will provide the necessary measures to publish multi-dimensional data, such as statistics about datasets, based on the Data Cube Vocabulary[45].

Other common extensions of similar general character as the ontologies of that layer, which could be useful in multiple use cases.

[59] https://github.com/dbpedia/DataId-Ontology/tree/master/ld

[60] https://github.com/dbpedia/DataId-Ontology/tree/DataManagementPlanExtension/acp

Ontologies under the DataID multilayer concept do not offer cardinality restrictions, making them easy to extend and adhere to OWL profiles. An application profile for the DataID service (cf. **??**) was declared using SHACL[61].

Extending this ecosystem of dataset metadata with domain-specific OWL ontologies adds further opportunities for applications clustered around datasets, as we will showcase in **??**.

ref

4.3 The interplay of ontologies

title will be revised

Multiple requirements are planned to be enforced for the adoption of new ontologies in the common (or mid) layer of the DataID Ecosystem. They might contain (while not being restricted to):

- Authors must provide information about the reason for the new extension (and why the expected result is not achievable with existing extensions).
- Authors must document the Interoperability with other extensions of this layer (and where problems are to be expected).
- Authors must inform about conformity with OWL profiles.

Deciding on which combination of DataID ontologies to use for a Dataset description is a domain and problem dependent process. It may be necessary to add additional properties on top of the provided metadata properties.

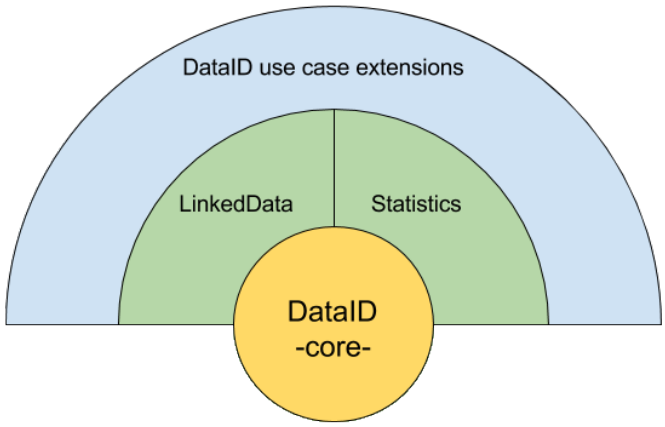

Figure 4.2: Example of combining multiple DataID ontologies

[61] http://w3c.github.io/data-shapes/shacl/

For example, a DataID based ontology for LOD Datasets dealing with multi-dimensional data, may look schematically like this: (importing DataID core and the extensions for Linked Data and Statistics, as well as some additional properties only used in the use case at hand.)

5 DataID core Ontology

5.1 Fundamentals

This section will provide an overview of the **DataID core** ontology and some background on basic design decisions made. Most of the vocabularies reused by this ontology have already been presented in chapter 3 and need no further introduction.

DataID core developed out of a previous version of **DataID** by following the aspiration to modularise **DataID** into a core-ontology, surrounded by its dependents. Many design decisions of **DataID core** were already introduced in its former version, which was designed to make DCAT, combined with the VOID vocabulary, fit the requirements of the DBpedia use case for a hierarchy of Linked Data datasets (see also section 3.2.5):

> "The DBpedia dataset, with its different versions and languages, multiple SPARQL endpoints and thousands of dump files with various content serves as one example of the complexity metadata models need to be able to express. We argue that the DCAT vocabulary as well as the established VoID vocabulary only provide a basic interoperability layer to discover data. In their current state, they still have to be expanded to fully describe datasets as complex as DBpedia [...]." [8]

DataID core is founded on two pillars: the DCAT and PROV-O ontologies. To incorporate DCAT as the basis of **DataID**, to further extend DCAT with PROV-O, introducing extensive provenance records in the process, and adding properties specific to the DBpedia use case was the original premise of this endeavour. In addition, the VOID vocabulary was adopted to cover Linked Data specific semantics and provide more general properties, such as `void:subset` for establishing dataset hierarchies. The wide application of these vocabularies in the context of the Semantic Web was the rational behind these decisions, furthering our goal of **interoperability**.

As DCAT, **DataID core** is centred around the Dataset and Distribution concepts which were imported from DCAT. We introduced the class

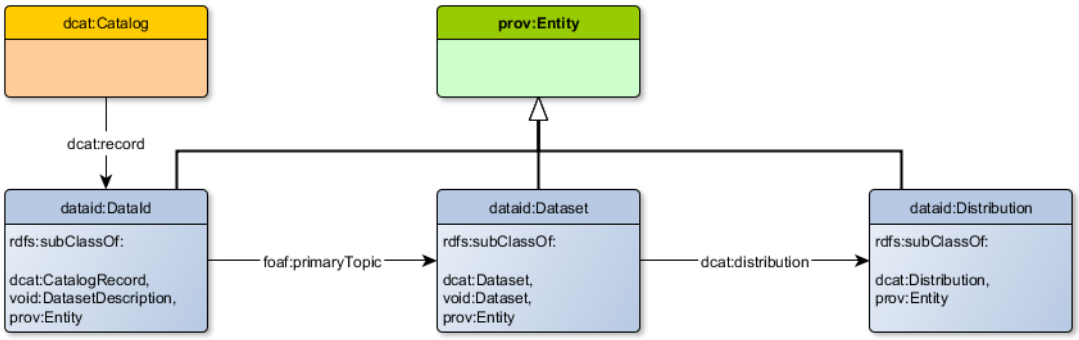

Figure 5.1: Foundations: Combining DCAT, VOID and PROV-O

`dataid:DataId` (see section 5.2.1), merging both `dcat:CatalogRecord` and `void:DatasetDescription` into one, and providing an additional level of abstraction between `dataid:Dataset` and `dcat:Catalog`. Instances of this class can be compared to root elements in the XML world, since all subsequent instances, describing a dataset, are hierarchicaly below a DataId instance. Though, this comparison is of course misleading, since a graph has no orientation. Yet, most of DataID documents[62] will have structural similarities to XML documents.

All dataset related classes of **`DataID core`** (`dataid:DataId`, `dataid:Dataset` and `dataid:Distribution`) are sub classes of `prov:Entity`, which is the interface needed to harness the possibilities of the Provenance Ontology. In the context of this description of **`DataID core`**, the word 'Entities' summarises all possible instances of `prov:Entity` (ergo: instances of these three classes).

With PROV-O, describing any provenance issue is possible (section 3.2.1). Central to this, is the concept of qualification classes, providing qualifications for more general properties (e.g. for `prov:wasAttributedTo`). Describing the interrelations between Entities (such as a particular dataset or just a single distribution of it) and Agents (e.a. a person or an organisation) are salient requirements in most environments of datasets and their metadata. Thus, **`DataID core`** has singled out these relations to further qualify them and provide much needed referential integrity.

explain

The property `dataid:associatedAgent` (which is a sub property of `prov:wasAttributedTo`) is a universal relation between an Entity and an Agent. It can (and should) be qualified by an instance of `dataid:Authorization`, a sub class of `prov:Attribution`. An Authorization adds qualifications and restrictions to the original property, such as an agent role (defining the role the agent has in regard to the Entities involved - see section 5.2.6). Furthermore,

[62] DataID graph serialised as a metadata document

DataID core allows for assigning Actions to an Authorization, which specify what an Agent can do and for which tasks he/she/it is responsible for.

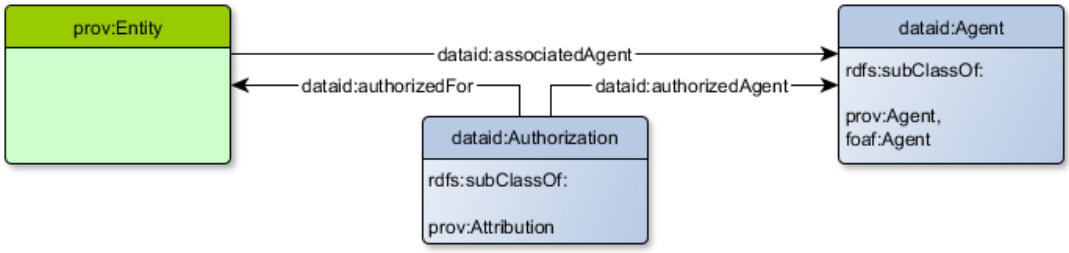

Figure 5.2: Foundations: Using PROV-O to qualify properties

In its current version 2.0.0, **DataID core** offers a general approach for describing dataset metadata, incorporating important ontologies to extend DCAT with **provenance**, qualifying relations and their intended range classes (e.g. to refine **licensing**), hierarchical dataset structures, management of rights and responsibilities of agents and exhaustive descriptions for data **access**.

In general, **DataID core** waived the use of cardinality restrictions[63], helping us with our goal to adhere to the OWL 2 RL[64] profile and maintaining the benefits of easy **extensibility** and **interoperability** of DCAT. In turn, **DataID core** restricts some of the very general property ranges of Dublin Core and DCAT properties (such as `dct:license` or `dcat:mediaType`), to reduce impreciseness and increase machine-readability.

loosing the ability to extend easily This section provides a concise overview of the DataID-core ontology, highlighting important features and improvements to the previously presented version in 2014 [8]. The current version (2.0.0) adheres to the OWL profile OWL2-RL[65]. Figure 5.3 supplies a depiction of this ontology. DCTERMS is used for most general metadata of any concept.

DataID is founded on two pillars: the DCAT and PROV-O ontologies. The class `dataid:DataId` subsumes `dcat:CatalogRecord`, which describes a dataset entry in a `dcat:Catalog`. It does not represent a dataset, but provenance information about dataset entries in a catalog. It is the root entity in any DataID description.

In addition the VOID vocabulary plays a central role, as the dataset concept of both the DCAT and VOID were merged into `dataid:Dataset`, providing useful properties about the content of a dataset from both ontologies. In particular,

[63] this is the purpose of an application profile, not of an ontology
[64] `https://www.w3.org/TR/owl2-profiles/#OWL_2_RL`
[65] `https://www.w3.org/TR/owl2-profiles/`

the property void:subset allows for the creation of dataset hierarchies, while dcat:distribution points out the distributions of a dataset.

The class dcat:Distribution is the technical description of the data itself, as well as documentation of how to access the data described (dcat:accessURL / dcat:downloadURL). This concept is crucial to be able to automatically retrieve and use the data described in the DataID, simplifying, for example, data analysis. We introduced additional subclasses (e.g. dataid:ServiceEndpoint), to further distinguish how the data is available on the Web.

DCAT does not offer an intrinsic way of specifying the exact format of the content described by a distribution. While the property dcat:mediaType does exist, its expected range dct:MediaTypeOrExtend is an empty class (without any further definitions). Therefore, we created dataid:MediaType to remedy this matter. With the property dataid:innerMediaType we can even describe nested formats (e.g. .xml.bz2), useful in pipeline processing.

The most important change to the previous version of DataID is the possible expression of which role an agent can take in regard to metadata entities (e.g. the whole DataID and all datasets, a single distribution etc.). This is achieved by the class dataid:Authorization, which is a subclass of prov:Attribution, a qualification of the property prov:wasAttributedTo. Basically it states, which role(s) (dataid:authorityAgentRole) an agent (dataid:authorizedAgent) has regarding a certain collection of entities (dataid:authorizedFor). This mediator is further qualified by an optional period of time for which it is valid and authoritative restrictions by the entities themselves, allowing only specific instances of dataid:Authorization to exert influence over them (dataid:needsSpecialAuthorization).

The role an agent can take (dataid:AgentRole) has only one property, pointing out actions it entails. A dataid:AuthorizedAction shall either be a dataid:EntitledAction, representing all actions an agent could take, as well as the actions an agent has to take (dataid:ResponsibleAction). Actions and roles defined in this ontology (e.g. dataid:Publisher) are only examples of possible implementations and can be replaced to fit a use case. Hierarchical structures of agent roles or actions can provide additional semantics.

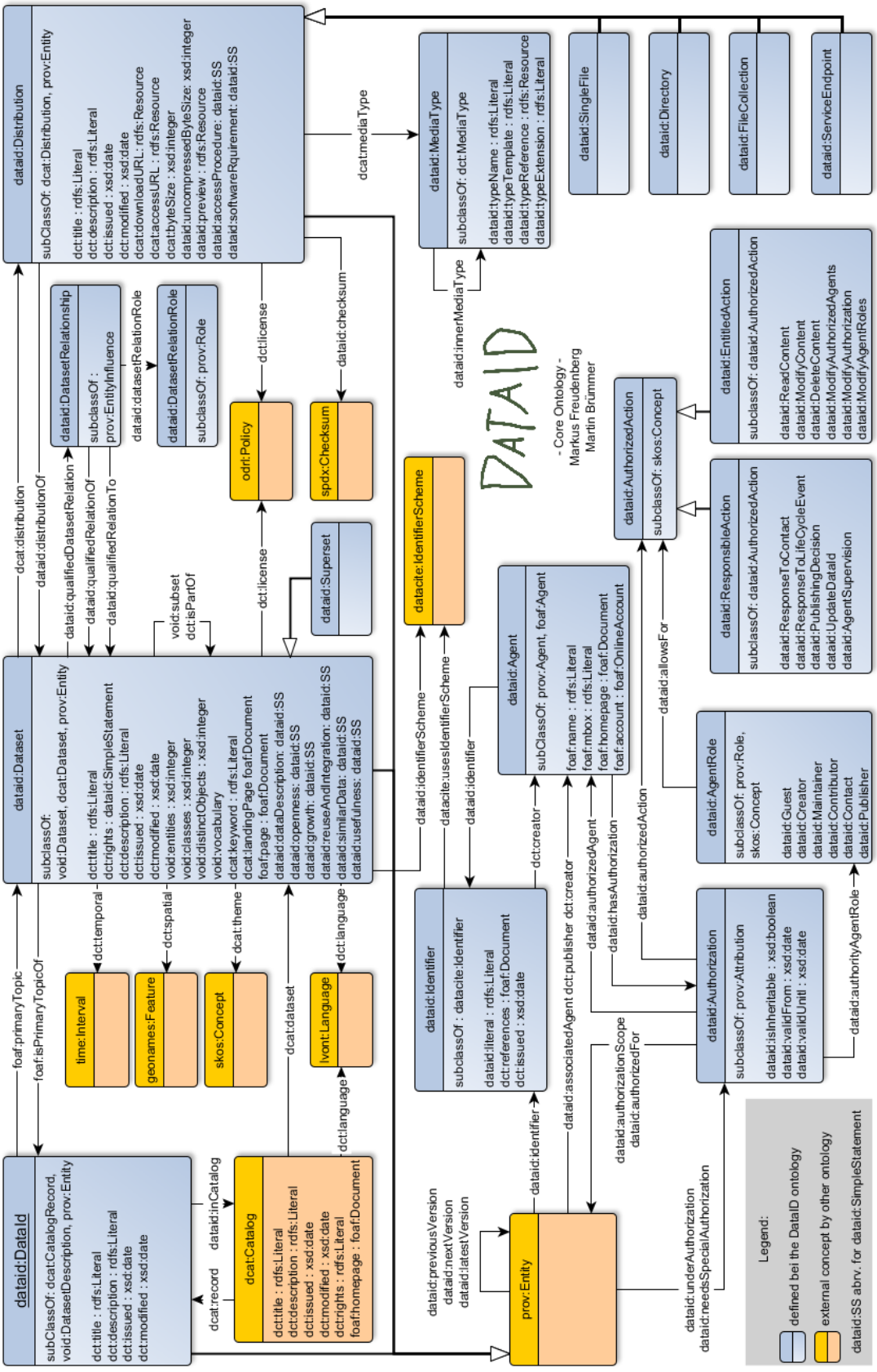

Figure 5.3: DataID core

5.2 Classes

5.2.1 DataId

5.2.2 Dataset

5.2.3 Distribution

5.2.4 MediaType

5.2.5 Agent

5.2.6 Authorization

5.2.7 AuthorizedAction & AgentRole

5.2.8 DatasetRelationship

5.2.9 Identifier

5.2.10 SimpleStatement

6 Publishing Data on the Web with DataID

7 Application: Data Management Plans

In diesem Kapitel folgt die Auswertung der Arbeit. Hier werden Messerreihen wie z.B. die Latenzzeit, die Performanz, der Speicherverbrauch, die Erkennungsrate etc. aufgeführt. Ferner werden die Methoden erläutert, die Ergebnisse interpretiert und diskutiert.

Auch hier gibt es keinen Richtwert, das Kapitel sollte umfassend wie nötig sein.

8 DataID and the Component MetaData Infrastructure (CMDI)

9 Evaluation

10 Conclusion and Future Work

Am Ende der Arbeit steht die Zusammenfassung, die alle wichtigen Punkte und Ergebnisse der Arbeit in einfachen Worten wiedergibt. Anschließend folgt ein Ausblick auf anschließende Arbeiten und Themenvorschläge.

Das Kapitel sollte zwischen einer und drei Seiten umfassen.

add prefix table

Glossar

Vocabulary On the Semantic Web, vocabularies define the concepts and relationships (also referred to as "terms") used to describe and represent an area of concern. Vocabularies are used to classify the terms that can be used in a particular application, characterize possible relationships, and define possible constraints on using those terms. In practice, vocabularies can be very complex (with several thousands of terms) or very simple (describing one or two concepts only)[46].

Ontology There is no clear division between what is referred to as "vocabularies" (see Vocabulary) and "ontologies". The trend is to use the word "ontology" for more complex, and possibly quite formal collection of terms, whereas "vocabulary" is used when such strict formalism is not necessarily used or only in a very loose sense. Vocabularies are the basic building blocks for inference techniques on the Semantic Web[46].

Prefix Gloassar

dcat DCAT

prov Provenance Ontology

dct Dublin Core Terms

References

[1] Fadi Maali, DERI, and NUI Galway. *Data Catalog Vocabulary (DCAT)*. W3C Recommendation. URL: https://www.w3.org/TR/vocab-dcat/.

[2] Phil Archer and Keith Jeffery. *Smart Descriptions & Smarter Vocabularies Report*. W3C Workshop report. URL: https://www.w3.org/2016/11/sdsvoc/report.html.

[3] Bernadette Farias Loscio, Caroline Burle, and Newton Calegari. *Data on the Web Best Practices*. W3C Proposed Recommendation. W3C, Dec. 2016. URL: https://www.w3.org/TR/2016/PR-dwbp-20161215/.

[4] Deborah McGuinness, Timothy Lebo, and Satya Sahoo. *The PROV Ontology*. W3C Recommendation. URL: http://www.w3.org/TR/prov-o/.

[5] Mo McRoberts and Victor Rodriguez Doncel. *ODRL Version 2.1 Ontology*. W3C Community Group Specification. http://www.w3.org/ns/odrl/2/ODRL21. W3C, Mar. 2015. URL: https://www.w3.org/ns/odrl/2/.

[6] Kevin Feeney, Gavin Mendel-Gleason, and Rob Brennan. "Linked data schemata: fixing unsound foundations". In: *Semantic Web Journal-Special Issue on Quality Management of Semantic Web Assets* (2016).

[7] Mark D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3 (Mar. 2016), pp. 160018+. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18. URL: http://dx.doi.org/10.1038/sdata.2016.18.

[8] Martin Brümmer et al. "DataID: Towards Semantically Rich Metadata for Complex Datasets". In: *Proceedings of the 10th International Conference on Semantic Systems*. SEM '14. Leipzig, Germany: ACM, 2014, pp. 84–91.

[9] Monika Solanki et al. "Enabling Combined Software and Data Engineering at Web-Scale: The ALIGNED Suite of Ontologies". In: *The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part II*. 2016, pp. 195–203. DOI: 10.1007/978-3-319-46547-0_21. URL: http://dx.doi.org/10.1007/978-3-319-46547-0_21.

[10] Markus Freudenberg et al. "The Metadata Ecosystem of DataID". In: *Metadata and Semantics Research: 10th International Conference, MTSR 2016, Göttingen, Germany, November 22-25, 2016, Proceedings*. Ed. by Emmanouel Garoufallou et al. Cham: Springer International Publishing, 2016, pp. 317–332. ISBN: 978-3-319-49157-8. DOI: 10.1007/978-3-319-49157-8_28. URL: http://dx.doi.org/10.1007/978-3-319-49157-8_28.

[11] Chaim Zins. "Conceptual Approaches for Defining Data, Information, and Knowledge: Research Articles". In: *J. Am. Soc. Inf. Sci. Technol.* 58.4 (Feb. 2007), pp. 479–493. ISSN: 1532-2882. DOI: 10.1002/asi.v58:4. URL: http://dx.doi.org/10.1002/asi.v58:4.

[12] Ramez Elmasri and Shamkant B. Navathe. *Fundamentals of Database Systems (5th Edition)*. Addison Wesley, Mar. 2006. ISBN: 0321369572. URL: http://www.amazon.ca/exec/obidos/redirect?tag=citeulike04-20%7B%5C&%7Dpath=ASIN/0321369572.

[13] F. Bodendorf. *Daten- und Wissensmanagement*. Springer-Lehrbuch. Springer, 2003. ISBN: 9783540001027. URL: https://books.google.de/books?id=f1wjPAAACAAJ.

[14] Claude Shannon. "A Mathematical Theory of Communication". In: *Bell System Technical Journal* 27 (July 1948), pp. 379–423, 623–656. URL: http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf.

[15] Jennifer Rowley. "The Wisdom Hierarchy: Representations of the DIKW Hierarchy". In: *J. Inf. Sci.* 33.2 (Apr. 2007), pp. 163–180. ISSN: 0165-5515. DOI: 10.1177/0165551506070706. URL: http://dx.doi.org/10.1177/0165551506070706.

[16] NISO. *Understanding metadata*. ISBN 1-880124-62-9. National Information Standards Organization. 2004. URL: http://www.niso.org/standards/resources/UnderstandingMetadata.pdf.

[17] IFLA. *Functional Requirements for Bibliographic Records: Final Report*. K. G. Saur, 1998. URL: http://www.amazon.com/Functional-Requirements-Bibliographic-Records-Publications/dp/359811382X.

[18] Shane McCarron et al. *RDFa Core 1.1 - Third Edition*. W3C Recommendation. http://www.w3.org/TR/2015/REC-rdfa-core-20150317/. W3C, Mar. 2015.

[19] Tim Berners-Lee, James Hendler, and Ora Lassila. "The Semantic Web A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities". In: (2001). URL: http://www.sciam.com/article.cfm?id=the-semantic-web%5C&%5C#38;print=true.

[20] Markus Lanthaler, David Wood, and Richard Cyganiak. *RDF 1.1 Concepts and Abstract Syntax*. W3C Recommendation. http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/. W3C, Feb. 2014. URL: http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/;print=true.

[21] Sebastian Rudolph et al. *OWL 2 Web Ontology Language Primer (Second Edition)*. W3C Recommendation. http://www.w3.org/TR/2012/REC-owl2-primer-20121211/. W3C, Dec. 2012. URL: https://www.w3.org/TR/owl2-primer/;print=true.

[22] Achille Fokoue et al. *OWL 2 Web Ontology Language Profiles (Second Edition)*. W3C Recommendation. http://www.w3.org/TR/2012/REC-owl2-profiles-20121211/. W3C, Dec. 2012. URL: https://www.w3.org/TR/2012/REC-owl2-profiles-20121211/;print=true.

[23] Ramanathan Guha and Dan Brickley. *RDF Schema 1.1*. W3C Recommendation. http://www.w3.org/TR/2014/REC-rdf-schema-20140225/. W3C, Feb. 2014. URL: http://www.w3.org/TR/2014/REC-rdf-schema-20140225/;print=true.

[24] Jon Barwise. "An Introduction to First-Order Logic". In: *HANDBOOK OF MATHEMATICAL LOGIC*. Elsevier BV, 1977, pp. 5–46. DOI: 10.1016/s0049-237x(08)71097-8. URL: http://dx.doi.org/10.1016/S0049-237X(08)71097-8.

[25] Ian Horrocks et al. "OWL: a Description Logic Based Ontology Language for the Semantic Web". In: *The Description Logic Handbook: Theory, Implementation, and Applications (2nd Edition)*. Ed. by Franz Baader et al. Cambridge University Press, 2007. Chap. 14. URL: download/2003/HPMW07.pdf.

[26] C. Bizer, T. Heath, and T. Berners-Lee. "Linked data - the story so far". In: *Int. J. Semantic Web Inf. Syst.* 5.3 (2009), 1?22.

[27] Tim Berners-Lee. *Linked Data*. 2006. URL: https://www.w3.org/DesignIssues/LinkedData.html (visited on 12/15/2016).

[28] Fadi Maali et al. "Enabling Interoperability of Government Data Catalogues." In: *EGOV*. Ed. by Maria Wimmer et al. LNCS. Springer, 2010.

[29] Keith Alexander and Michael Hausenblas. "Describing linked datasets - on the design and usage of void, the ?vocabulary of interlinked datasets". In: *In Linked Data on the Web Workshop (LDOW 09), in conjunction with 18th International World Wide Web Conference (WWW 09*. 2009.

[30] Christoph Böhm, Johannes Lorey, and Felix Naumann. "Creating voiD descriptions for Web-scale data". In: *J. Web Sem.* 9.3 (2011), pp. 339–345. DOI: 10.1016/j.websem.2011.06.001. URL: http://dx.doi.org/10.1016/j.websem.2011.06.001.

[31] Sebastian Neumaier, Jürgen Umbrich, and Axel Polleres. "Challenges of mapping current CKAN metadata to DCAT". In: *W3C Workshop on Data and Services Integration*. Amsterdam, the Netherlands, Nov. 2016. URL: https://www.w3.org/2016/11/sdsvoc/SDSVoc16_paper_16.

[32] John P. McCrae et al. "One ontology to bind them all: The META-SHARE OWL ontology for the interoperability of linguistic datasets on the Web". In: *Proc. of 12th Extended Semantic Web Conference (ESWC 2015) Satellite Events, Portorož, Slovenia*. Vol. 9341. June 2015, pp. 271–282.

[33] DCAT Application profile working group. *DCAT-AP v1.1*. 2016. URL: https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-ap-v11 (visited on 12/15/2016).

[34] Brecht Wyns et al. "DCAT Application Profile for Data Portals in Europe". In: 2016. URL: %5Curl%7Bhttps://www.w3.org/2016/11/sdsvoc/SDSVoc16_paper_30%7D.

[35] SEMANTIC WEB HEALTH CARE and LIFE SCIENCES (HCLS) INTEREST GROUP. *HCLS mission statement*. 2016. URL: https://www.w3.org/blog/hcls/ (visited on 12/15/2016).

[36] Michel Dumontier et al. "The health care and life sciences community profile for dataset descriptions". In: *PeerJ* (2016). DOI: 10.7717/peerj.2331. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4991880/.

[37] Keith Jeffery and Anne Asserson. "CERIF-CRIS for the European e-Infrastructure". In: *Data Science Journal* 9 (2010), CRIS1–CRIS6. DOI: 10.2481/dsj.CRIS1. URL: http://dx.doi.org/10.2481/dsj.CRIS1.

[38] Keith G Jeffery and Anne Asserson. "Position Paper: Why CERIF?" In: 2016. URL: %5Curl%7Bhttps://www.w3.org/2016/11/sdsvoc/SDSVoc16_paper_15%7D.

[39] Luc Moreau and Paolo Missier. *PROV-DM: The PROV Data Model*. W3C Recommendation. http://www.w3.org/TR/2013/REC-prov-dm-20130430/. W3C, Apr. 2013.

[40] Luc Moreau et al. "The Rationale of PROV". In: *Web Semant.* 35.P4 (Dec. 2015), pp. 235–257. ISSN: 1570-8268. DOI: 10.1016/j.websem.2015.04.001. URL: http://dx.doi.org/10.1016/j.websem.2015.04.001.

[41] Gerard de Melo. "Lexvo.org: Language-Related Information for the Linguistic Linked Data Cloud". In: *Semantic Web* 6.4 (Aug. 2015), pp. 393–400.

[42] Silvio Peroni et al. "DataCite2RDF: Mapping DataCite Metadata Schema 3.1 Terms to RDF". In: (Feb. 2016). DOI: 10.6084/m9.figshare.2075356.v1. URL: https://figshare.com/articles/DataCite2RDF_Mapping_DataCite_Metadata_Schema_3_1_Terms_to_RDF/2075356.

[43] Henrik Frystyk Nielsen. *Interoperability and Evolvability*. https://www.w3.org/Protocols/Design/Interevol.html.

[44] Keith Alexander et al. *Describing Linked Datasets with the VoID Vocabulary*. W3C Interest Group Note. URL: https://www.w3.org/TR/void/.

[45] Richard Cyganiak et al. *The RDF Data Cube Vocabulary*. W3C Recommendation. URL: https://www.w3.org/TR/vocab-data-cube/.

[46] W3C. *VOCABULARIES*. https://www.w3.org/standards/semanticweb/ontology. [Online; accessed 01-December-2016]. 2015.

Erklärung

"'Ich versichere, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe, insbesondere sind wörtliche oder sinngemäße Zitate als solche gekennzeichnet. Mir ist bekannt, dass Zuwiderhandlung auch nachträglich zur Aberkennung des Abschlusses führen kann"'.

Ort Datum Unterschrift