The Metadata Ecosystem of DataID

Markus Freudenberg[1], Martin Brümmer[2], Jessika Rücknagel[3], Robert Ulrich[3], Thomas Eckart[4], Dimitris Kontokostas[1], and Sebastian Hellmann[1]

[1] Universität Leipzig, Institut für Angewandte Informatik (InfAI), AKSW/KILT
http://aksw.org/Groups/KILT
{lastname}@informatik.uni-leipzig.de
[2] eccenca GmbH, Hainstr. 8, 04109 Leipzig, Germany, http://eccenca.com
martin.bruemmer@eccenca.com
[3] re3data, http://www.re3data.org
ruecknagel@sub.uni-goettingen.de, robert.ulrich@kit.edu
[4] Universität Leipzig, Abteilung Automatische Sprachverarbeitung
http://asv.informatik.uni-leipzig.de/en
teckart@informatik.uni-leipzig.de

Abstract. The rapid increase of data produced in a data-centric economy emphasises the need for rich metadata descriptions of datasets, covering many domains and scenarios. While there are multiple metadata formats, describing datasets for specific purposes, exchanging metadata between them is often a difficult endeavour. More general approaches for domain-independent descriptions often lack the precision needed in many domain-specific use cases. This paper introduces the multilayer ontology of DataID, providing semantically rich metadata for complex datasets. In particular, we focus on the extensibility of its core model and the interoperability with foreign ontologies and other metadata formats. As a proof of concept, we will present a way to describe *Data Management Plans (DMP)* of research projects alongside the metadata of its datasets, repositories and involved agents.

1 Introduction

In 2006, Clive Humby coined the phrase ”the new oil” for (digital) data[5], 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: (a) The growth in content and reach of the World Wide Web. (b) The digitalising of former analogue data. (c) The realisation of what is called

[5] https://www.theguardian.com/technology/2013/aug/23/tech-giants-data

the Internet of Things (IoT)[6]. (d) The shift of classic fields of research and industry to computer-aided processes and digital resource management (e.g. digital humanities, industry 4.0). (e) Huge data collections about protein sequences or human disease taxonomies are established in the life sciences. (f) Research areas like natural language processing or machine learning are generating and refining data. (g) In addition, open data initiatives like the Open Knowledge Foundation are following the call for 'Raw data, Now!'[7] of Tim Berners-Lee, demanding open data from governments and organisations.

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[8]. Instrumental to all of this, is providing rich metadata descriptions for datasets, thereby enabling users to discover, understand and 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 depending on the circumstances. Extending existing metadata models to fit these scenarios is a cumbersome process resulting often in non-reusable solutions.

In this paper we will present the improved metadata model of DataID (cf. Sections 4 and 5), a multi-layered metadata system, which, in its core, describes datasets and their different manifestations, as well as relations to agents like persons or organisations, in regard to their rights and responsibilities. In a previous version of DataID[1] we already provided a solution for an accessible, compatible and granular best-practice of dataset descriptions for Linked Open Data (LOD).

We want to build on this foundation, presenting improvements in regard to PROVENANCE, LICENSING and ACCESS. In particular, we want to address the aspects EXTENSIBILITY and INTEROPERABILITY of dataset metadata, demonstrating the universal applicability of DataID in any domain or scenario. As a proof of concept for its EXTENSIBILITY we will show how to provide extensive metadata for Data Management Plans (DMP) of research projects (cf. Section 6) by extending the DataID model with properties specific to this scenario. The INTEROPERABILITY with other metadata models is exemplified by the mapping of common CMDI (CLARIN) profiles to DataID in Section 7.

2 Related Work

The Data Catalog Vocabulary (DCAT) is a W3C Recommendation [2] and serves as a foundation for many available dataset vocabularies and application profiles.

[6] http://siliconangle.com/blog/2015/10/28/page/3#post-254300

[7] http://www.wired.co.uk/news/archive/2012-11/09/raw-data

[8] http://www.ibmbigdatahub.com/infographic/four-vs-big-data

In [3] the authors introduce a standardised interchange format for machine-readable representations of government data catalogues. The DCAT vocabulary includes the special class Distribution for the representation of the available materialisations of a dataset (e.g. CSV file, an API or RSS feed). These distributions cannot be described further within DCAT (e.g. the type of data, or access procedures). Applications which utilise the DCAT vocabulary (e.g. datahub.io[9]) provide no standardised means for describing more complex datasets either. Yet, the basic class structure of DCAT (Catalog, CatalogRecord, Dataset, Distribution) has prevailed. Range definitions of properties provided for these classes are general enough to make this vocabulary easy to extend.

DCAT, as opposed to PROV-O, expresses provenance in a limited way using a few basic properties such as `dct:source` or `dct:creator`, thus it does not relate semantically to persons or organisations involved in the publishing, maintenance etc. of the dataset. There is no support or incentive to describe source datasets or conversion activities of transformations responsible for the dataset at hand. This lack is crucial, especially in a scientific contexts, as it omits the processes necessary to replicate a specific dataset, a feature easily obtainable by the use of PROV-O.

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[10](Comprehensive Knowledge Archive Network), which is used as a metadata schema in data portals like datahub.io, partially implements the DCAT vocabulary, but only describes resources associated with a dataset superficially. Additional properties are simple key-value pairs which themselves are linked by `dct:relation` properties. This data model is semantically poor and inadequate for most use cases wanting to automatically consume the data of a dataset.

While not implementing the DCAT vocabulary, META-SHARE [4] does provide an almost complete mapping to DCAT, providing an extensive description of language resources, based on a XSD schema. In addition it offers an exemplary way of describing licenses and terms of reuse. Yet, META-SHARE is specialised on language resources, thus lacking generality and extensibility for other use cases.

Likewise the Asset Description Metadata Schema[11] (ADMS) is a profile of DCAT, which only describes a specialised class of datasets: so-called Semantic Assets. Highly reusable metadata (e.g. code lists, XML schemata, taxonomies, vocabularies etc.), which is comprised of relatively small text files.

DCAT-AP (DCAT Application Profile for data portals in Europe[12]) is a profile, extending DCAT with some ADMS properties. It has been endorsed by the ISA Committee in January of 2016[13]. Due to the stringent cardinality restrictions, extending DCAT-AP to serve more elaborate purposes will prove difficult. As remarked in section 7 the representation of different agent roles is lacking in the current version of DCAT-AP. Neither DCAT-AP nor ADMS give any consideration to

[9] http://datahub.io/
[10] http://ckan.org/
[11] https://www.w3.org/TR/vocab-adms/
[12] https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-ap-v11
[13] https://joinup.ec.europa.eu/community/semic/news/dcat-ap-v11-endorsed-isa-committee

defining responsibilities of agents, extending provenance or providing thorough machine-readable licensing information.

Similar problems afflicted the previous version of the DataID ontology[1]. Rooted in the Linked Open Data world, it neglected important information or provided properties (e.g. `dataid:graphName`) which are orphans outside this domain. While already importing the PROV-O ontology, it was lacking a specific management of rights and responsibilities.

3 Motivation

In 2011, the European Commission published its *Open Data Strategy* defining the following six barriers[14] for "open public data":

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.

Taking these as a starting point, enriched by requirements of multiple use cases (e.g. section 6) and considering the existing and missing features of related vocabularies described in the previous section, we contrived the following short list of important aspects of dataset metadata:

(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 (as we will see in Section 6), 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

[14] http://europa.eu/rapid/press-release_MEMO-11-891_en.htm

extensions to a vocabulary might encroach on its ability to translate into other useful metadata formats. This notion is corroborated by this document [5]. Note: We (the authors) do not differentiate between EVOLVABILITY and EXTENSIBILITY in the context of this paper. The discrepancies with INTEROPERABILITY are true for both concepts.

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 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 not to jeopardise EXTENSIBILITY and INTEROPERABILITY, we modularised DataID in a core ontology and multiple extensions. The onion-like layer model (cf. Figure 1) illustrates the import restrictions of different ontologies. An ontology of a certain layer shall only import DataID ontologies from layers below their own. 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. While non of them are a mandatory import for use case specific extensions, as opposed to DataID core, in many cases some or all of them will be useful contributions.

DataID core provides the basic description of a dataset (cf. Section 5) and serves as foundation for all extensions to DataID.

Linked Data[15] extends DataID core with the VOID vocabulary[6] 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.

Activities & Plans[16] provides provenance information of activities which generated, changed or used datasets. 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[7].

Statistics will provide the necessary measures to publish multi-dimensional data, such as statistics about datasets, based on the Data Cube Vocabulary[8].

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. Section 8) was declared using SHACL[17].

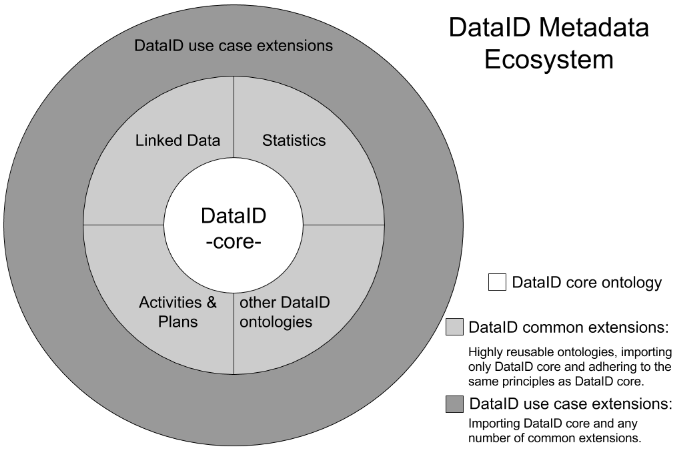
Extending this ecosystem of dataset metadata with domain-specific OWL ontologies adds further opportunities for applications clustered around datasets, as we will showcase in Section 6.

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

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

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

Fig. 1. The Metadata Ecosystem of DataID


DataID Metadata Ecosystem
DataID use case extensions
Linked Data
Statistics
DataID -core-
Activities & Plans
other DataID ontologies
□ DataID core ontology
▨ DataID common extensions:
Highly reusable ontologies, importing only DataID core and adhering to the same principles as DataID core.
■ DataID use case extensions:
Importing DataID core and any number of common extensions.

5 DataID core

This section provides a concise overview of the DataID-core ontology, highlighting important features and improvements to the previously presented version in 2014 [1]. The current version (2.0.0) adheres to the OWL profile OWL2-RL[18]. Figure 2 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.

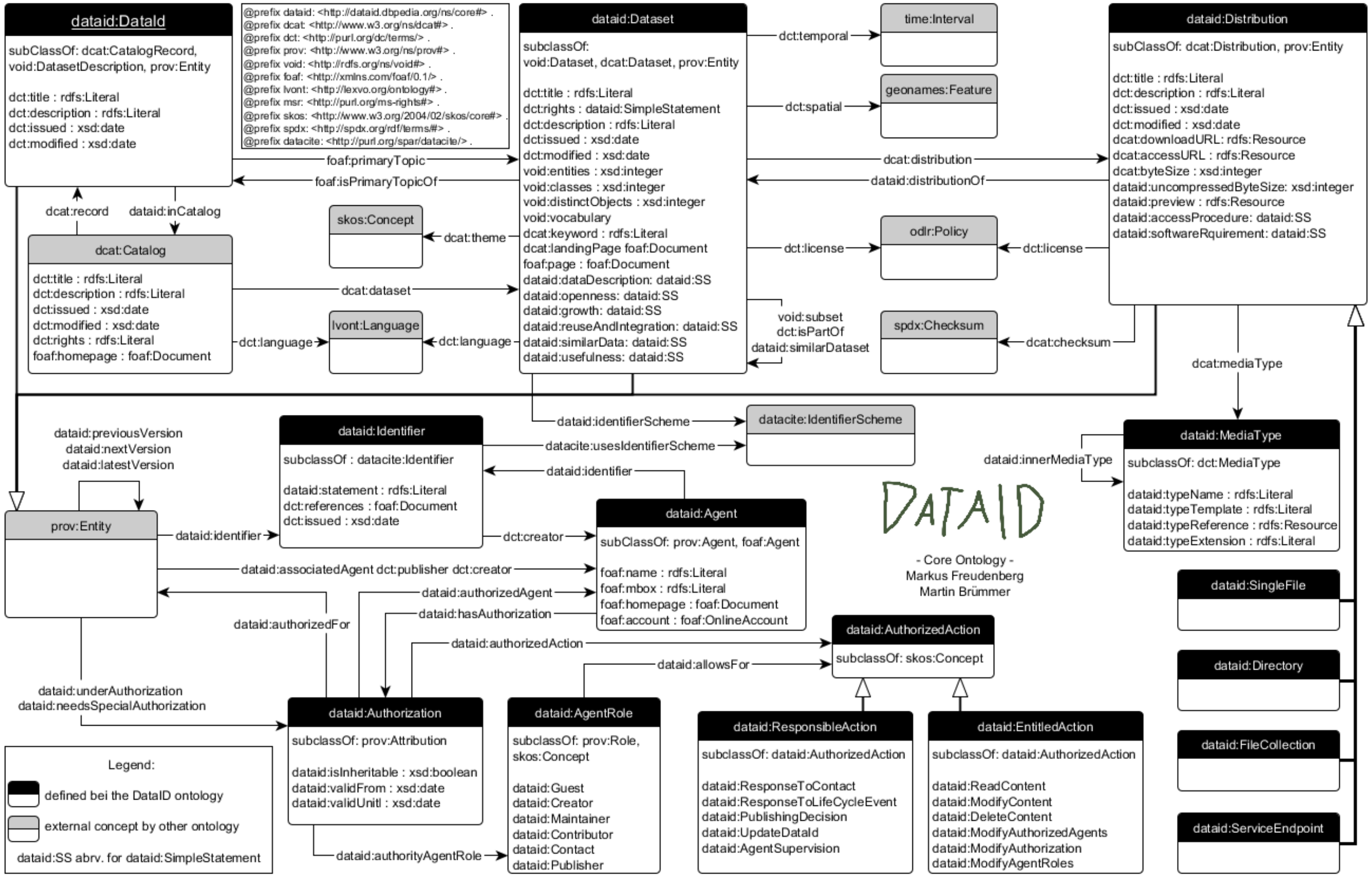
In addition the VOID vocabulary plays a central role, as the dataset concept of both the DCAT and VOID were merged into `dataid:Dataset`, providing useful properties about the content of a dataset from both ontologies. In particular, the property `void:subset` allows for the creation of dataset hierarchies,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

[18] https://www.w3.org/TR/owl2-profiles/

Fig. 2. DataID core


dataid:DataId
subClassOf: dcat:CatalogRecord,
void:DatasetDescription, prov:Entity
dct:title : rdfs:Literal
dct:description : rdfs:Literal
dct:issued : xsd:date
dct:modified : xsd:date
@prefix dataid: <http://dataid.dbpedia.org/ns/core#> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix void: <http://rdfs.org/ns/void#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix lvont: <http://lexvo.org/ontology#> .
@prefix msr: <http://purl.org/ms-rights#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix spdx: <http://spdx.org/rdf/terms#> .
@prefix datacite: <http://purl.org/spar/datacite/> .
foaf:primaryTopic
foaf:isPrimaryTopicOf
dcat:record
dataid:inCatalog
dcat:Catalog
dct:title : rdfs:Literal
dct:description : rdfs:Literal
dct:issued : xsd:date
dct:modified : xsd:date
dct:rights : rdfs:Literal
foaf:homepage : foaf:Document
skos:Concept
dcat:theme
dcat:dataset
lvont:Language
dct:language
dct:language
dataid:Dataset
subclassOf:
void:Dataset, dcat:Dataset, prov:Entity
dct:title : rdfs:Literal
dct:rights : dataid:SimpleStatement
dct:description : rdfs:Literal
dct:issued : xsd:date
dct:modified : xsd:date
void:entities : xsd:integer
void:classes : xsd:integer
void:distinctObjects : xsd:integer
void:vocabulary
dcat:keyword : rdfs:Literal
dcat:landingPage foaf:Document
foaf:page : foaf:Document
dataid:dataDescription: dataid:SS
dataid:openness: dataid:SS
dataid:growth: dataid:SS
dataid:reuseAndIntegration: dataid:SS
dataid:similarData: dataid:SS
dataid:usefulness: dataid:SS
dct:temporal
time:Interval
dct:spatial
geonames:Feature
dcat:distribution
dataid:distributionOf
dct:license
odrl:Policy
dct:license
void:subset
dct:isPartOf
dataid:similarDataset
spdx:Checksum
dcat:checksum
dataid:Distribution
subClassOf: dcat:Distribution, prov:Entity
dct:title : rdfs:Literal
dct:description : rdfs:Literal
dct:issued : xsd:date
dct:modified : xsd:date
dcat:downloadURL: rdfs:Resource
dcat:accessURL : rdfs:Resource
dcat:byteSize : xsd:integer
dataid:uncompressedByteSize: xsd:integer
dataid:preview : rdfs:Resource
dataid:accessProcedure: dataid:SS
dataid:softwareRqurement: dataid:SS
dcat:mediaType
dataid:MediaType
subclassOf: dct:MediaType
dataid:typeName : rdfs:Literal
dataid:typeTemplate : rdfs:Literal
dataid:typeReference : rdfs:Resource
dataid:typeExtension : rdfs:Literal
dataid:innerMediaType
dataid:SingleFile
dataid:Directory
dataid:FileCollection
dataid:ServiceEndpoint
dataid:previousVersion
dataid:nextVersion
dataid:latestVersion
prov:Entity
dataid:identifier
dataid:Identifier
subclassOf : datacite:Identifier
dataid:statement : rdfs:Literal
dct:references : foaf:Document
dct:issued : xsd:date
dataid:identifierScheme
datacite:IdentifierScheme
datacite:usesIdentifierScheme
dataid:identifier
dct:creator
dataid:Agent
subClassOf: prov:Agent, foaf:Agent
foaf:name : rdfs:Literal
foaf:mbox : rdfs:Literal
foaf:homepage : foaf:Document
foaf:account : foaf:OnlineAccount
dataid:associatedAgent dct:publisher dct:creator
dataid:authorizedAgent
dataid:authorizedFor
dataid:hasAuthorization
dataid:authorizedAction
DATAID
- Core Ontology -
Markus Freudenberg
Martin Brümmer
dataid:AuthorizedAction
subclassOf: skos:Concept
dataid:allowsFor
dataid:underAuthorization
dataid:needsSpecialAuthorization
dataid:Authorization
subclassOf: prov:Attribution
dataid:isInheritable : xsd:boolean
dataid:validFrom : xsd:date
dataid:validUnitl : xsd:date
dataid:authorityAgentRole
dataid:AgentRole
subclassOf: prov:Role,
skos:Concept
dataid:Guest
dataid:Creator
dataid:Maintainer
dataid:Contributor
dataid:Contact
dataid:Publisher
dataid:ResponsibleAction
subclassOf: dataid:AuthorizedAction
dataid:ResponseToContact
dataid:ResponseToLifeCycleEvent
dataid:PublishingDecision
dataid:UpdateDataId
dataid:AgentSupervision
dataid:EntitledAction
subclassOf: dataid:AuthorizedAction
dataid:ReadContent
dataid:ModifyContent
dataid:DeleteContent
dataid:ModifyAuthorizedAgents
dataid:ModifyAuthorization
dataid:ModifyAgentRoles
Legend:
defined bei the DataID ontology
external concept by other ontology
dataid:SS abrv. for dataid:SimpleStatement

An exact description of all classes and properties can be found under the DataID namespace uri http://dataid.dbpedia.org/ns/core including this depiction. The ontology RDF document is also available there: http://dataid.dbpedia.org/ns/core.ttl (.owl)

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.

6 Data Management Plans

Over the last years Data Management Plans (DMP) have become a requirement for project proposals within most major research funding institutions. It states what types of data and metadata are employed,The use case described here will introduce an extension to the DataID ontology to extensively describe a Data Management Plan for digital data in a universal way, laying the foundation for tools helping researchers and funders with the drafting and implementing DMPs. Based on multiple requirements, raised from different DMP guidelines, we will showcase the creation of a DataID extension. We incorporated the re3data ontology to describe repositories and institutions, exemplifying the use of external ontologies.

Requirements of Data Management Plans The following requirements were distilled from an extensive list of DMP guidelines of different research funding bodies, covering most of the non-functional demands raised pertaining to digital datasets. A complete list of funding organisations and their DMP guidelines involved in this analysis is available on the web[19].

1. Describe how data will be shared (incl. repositories and access procedures).
2. Describe the procedures put in place for long-term preservation of the data.
3. Describe the types of data and metadata, as well as identifiers used.
4. Provisioning of copyright and license information, including other possible limitations to the reusability of the data.

[19] http://wiki.dbpedia.org/use-cases/data-management-plan-extension-dataid#Organisation

5. Outline the rights and obligations of all parties as to their roles and responsibilities in the management and retention of research data.
6. Provision for changes in the hierarchy of involved agents and responsibilities (e.g. a Primary Investigator (PI) leaving the project).
7. Include provenance information on how datasets were used, collected or generated in the course of the project. Reference standards and methods applied.
8. Include statements on the usefulness of data for the wider public needs or possible exploitations for the likely purposes of certain parties.
9. Provide assistance for dissemination purposes of (open) data, making it easy to discover it on the web.
10. Is the metadata interoperable allowing data exchange between different meta data formats, researchers and organisations?
11. Project costs associated with implementing the DMP during and after the project. Justify the prognosticated costs.
12. Support the data management life cycle for all data produced.

To implement these demands in an ontology we can already make the following observations: 1. making further use of PROV-O is necessary to deal with the extensive demands for provenance, 2. a clear specification of involved agents and their responsibilities is needed and, 3. an extensive description of repositories retaining the described data is inescapable.

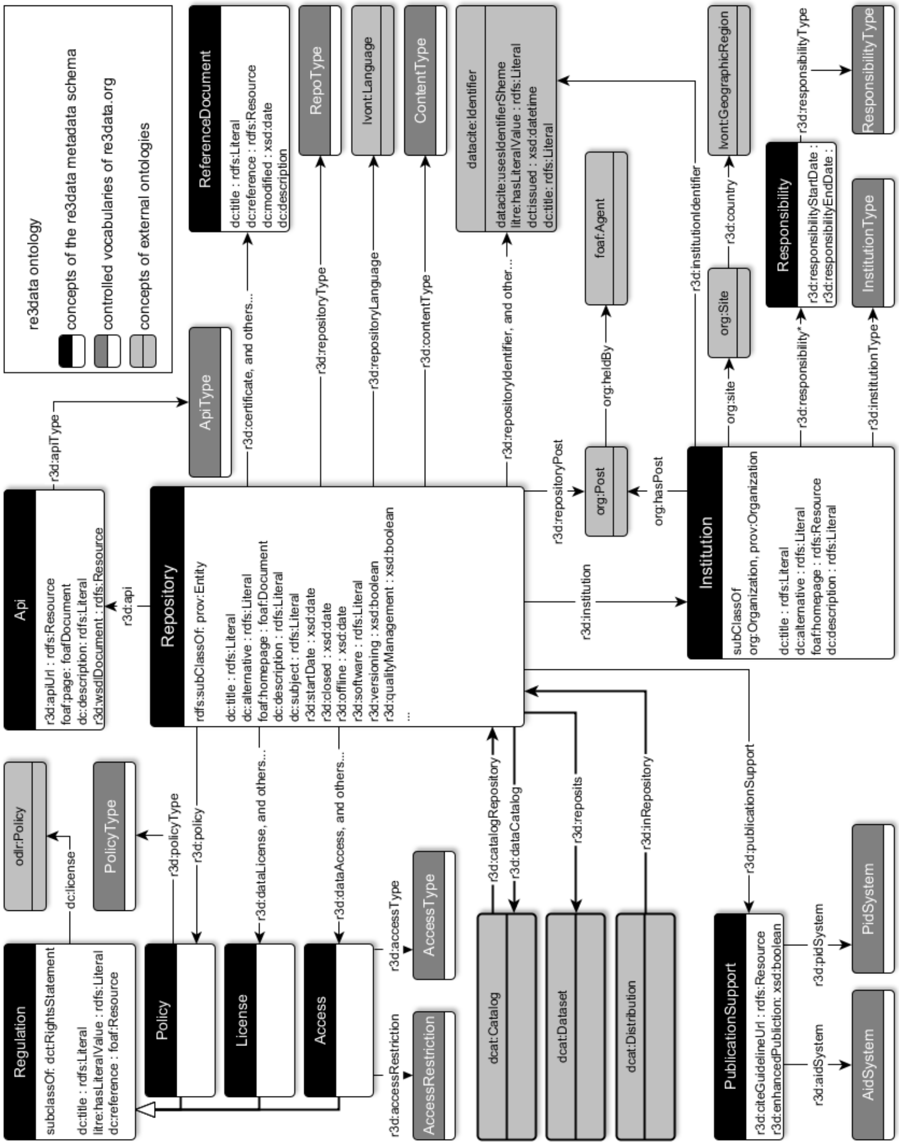
Our goal is to provide aid for researchers in drafting a DMP and implementing it with all requirements in mind: during the proposal phase, while the project is ongoing and the long term implementation of the DMP.

Registry of Research Data Repositories - re3data The re3data[20] registry currently lists over 1.600 research repositories, making it the largest and most comprehensive registry of data repositories available on the web. By providing a detailed metadata description of repositories, the registry helps researchers, funding bodies, publishers and research organisations to find an appropriate data repository for different purposes[9]. Initiated by multiple German research organisations, funded by the German Research Foundation[21] from 2012 until 2015, re3data is now a service of DataCite[22]. In 2014 re3data merged with the DataBib registry for research data repositories into one service[23].

One central goal of re3data is to enhance the visibility of existing research data repositories and to enable all those who are interested in finding a repository to assess a respective information service. This is achieved by an extensive and quality approved metadata description of the listed research data repositories. The basis for this description is the “Metadata Schema for the Description of Research Data Repositories”, having 42 properties in the current version 3.0 [10]. Considering the increasing number of funding bodies demanding a research data

[20] http://www.re3data.org/
[21] http://www.dfg.de/
[22] https://www.datacite.org/
[23] http://www.re3data.org/tag/databib/

Fig. 3. re3data Ontology


re3data ontology
concepts of the re3data metadata schema
controlled vocabularies of re3data.org
concepts of external ontologies
Api
r3d:apiUrl : rdfs:Resource
foaf:page: foafDocument
dc:description: rdfs:Literal
r3d:wsdlDocument : rdfs:Resource
r3d:apiType
ApiType
r3d:api
Repository
rdfs:subClassOf: prov:Entity
dc:title : rdfs:Literal
dc:alternative : rdfs:Literal
foaf:homepage : foaf:Document
dc:description : rdfs:Literal
dc:subject : rdfs:Literal
r3d:startDate : xsd:date
r3d:closed : xsd:date
r3d:offline : xsd:date
r3d:software : rdfs:Literal
r3d:versioning : xsd:boolean
r3d:qualityManagement : xsd:boolean
...
ReferenceDocument
dc:title : rdfs:Literal
dc:reference : rdfs:Resource
dc:modified : xsd:date
dc:description
r3d:certificate, and others...
RepoType
r3d:repositoryType
lvont:Language
r3d:repositoryLanguage
ContentType
r3d:contentType
datacite:Identifier
datacite:usesIdentifierSheme
litre:hasLiteralValue : rdfs:Literal
dct:issued : xsd:datetime
dc:title: rdfs:Literal
r3d:repositoryIdentifier, and other...
r3d:repositoryPost
org:Post
org:heldBy
foaf:Agent
org:hasPost
r3d:institution
Institution
subClassOf
org:Organization, prov:Organization
dc:title : rdfs:Literal
dc:alternative : rdfs:Literal
foaf:homepage : rdfs:Resource
dc:description : rdfs:Literal
r3d:institutionIdentifier
org:site
org:Site
r3d:country
lvont:GeographicRegion
r3d:responsibility*
Responsibility
r3d:responsibilityStartDate :
r3d:responsibilityEndDate :
r3d:responsibilityType
ResponsibilityType
r3d:institutionType
InstitutionType
odrl:Policy
dc:license
PolicyType
r3d:policyType
r3d:policy
Regulation
subclassOf: dct:RightsStatement
dc:title : rdfs:Literal
litre:hasLiteralValue : rdfs:Literal
dc:reference : foaf:Resource
Policy
License
r3d:dataLicense, and others...
Access
r3d:dataAccess, and others...
r3d:accessType
AccessType
r3d:accessRestriction
AccessRestriction
r3d:catalogRepository
r3d:dataCatalog
dcat:Catalog
r3d:reposits
dcat:Dataset
r3d:inRepository
dcat:Distribution
r3d:publicationSupport
PublicationSupport
r3d:citeGuidelineUrl : rdfs:Resource
r3d:enhancedPublic;tion: xsd:boolean
r3d:pidSystem
PidSystem
r3d:aidSystem
AidSystem

Note: This is a reduced version of the ontology omitting some properties and all instances of controlled vocabularies (white font on grey boxes). The re3data ontology has not been finalised by the time of submission. Some minor changes are still being discussed with re3data. The current version can be accessed here: https://github.com/re3data/ontology/blob/master/r3dOntology.ttl.

management plan as an integral part of a grant proposal, information regarding research data repositories is of great importance. The re3data schema does provide a thorough description of repositories and the unique opportunity to incorporate an existing, up-to-date collection of research repositories in future DataID-based applications. To accomplish the integration into the DMP ontology extension, we transformed the current XML-based schema into an OWL-ontology, using established vocabularies like PROV-O and ORG. The schema as well as the data provided by re3data will be available as Linked Data (e.g. via re3data ReSTful-API), thus making it discoverable and more easily accessible for services and applications, reaching a larger circle of users.

Alongside the repository concept, a rudimentary description of institutions which are hosting or funding a repository is needed to ensure long-term sustainability and availability of a repository. The derived re3data ontology supplements `r3d:Repository` and `r3d:Institution` with fitting PROV-O subclasses making them subject to provenance descriptions. The ORG ontology is used to further extend the Institution class, providing organisational descriptions.

Access regulations to the repository and the research data must be clarified, as well as the terms of use. The re3data ontology unifies all license and policy objects under the class `r3d:Regulation`, using the property `dct:license` to point out `odrl:Policy` descriptions of licenses, as used in the DataID ontology.

By linking to `dcat:Catalog` via `r3d:dataCatalog` and `dcat:Dataset` with `r3d:reposits`, we introduced the necessary means to relate descriptions of data stored inside a repository. By providing this interface with the DCAT vocabulary, DataIDs can be used for the description of data in the re3data context.

Implementation The DataID core ontology, the Activities & Plans extension (cf. Section 4) and the re3data ontology are the foundational components of the DMP extension (depiction: Figure 4). On top of which we added additional semantics, solving the requirements listed in Section 6.

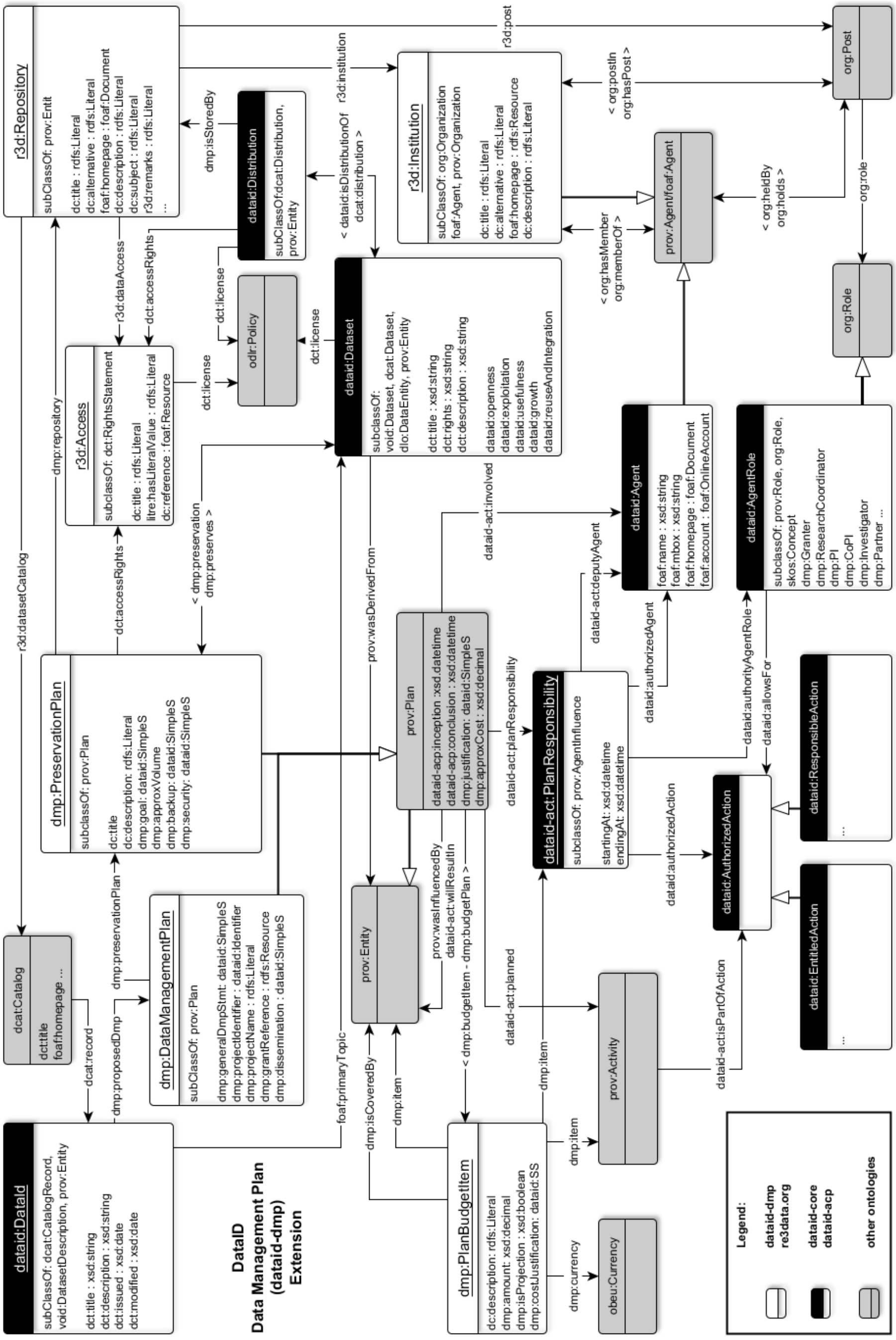
Extensive use of the PROV-O ontology and the concepts and properties introduced by the Activities & Plans extension is key to DMP, providing the means for describing sources and origin activities of datasets **(R7)**.

In the same vein, using the `dataid:Authorization` concept, augmented with a DMP specific set of `dataid:AgentRole` and `dataid:AuthorizedAction`, adds necessary provenance and satisfies requirement **(R5)** and **(R6)**.

A description of repositories involved in a DMP is provided by the concept `r3d:Repository`, including exact documentation of APIs and access procedures **(R1)**. More detailed information on the type of data or additional software necessary to access the data, was introduced with `dataid:Distribution`.

As in DataID core, information about licenses and other limitations are provided via `dct:license` and `dct:rights` **(R4)**, or the complementary properties of the re3data ontology concerning access and other policies. Helpful information on usefulness, reusability and other subjects for possible users of the portrayed datasets are added to the `dataid:Dataset` concept: `dataid:usefulness`, `dataid:reuseAndIntegration`, `dataid:exploitation` etc. **(R8)**.

Fig. 4. Data Management Ontology



This ontology is accessible here: https://github.com/dbpedia/DataId-Ontology/blob/DataManagementPlanExtension/dmp/dataManagementPlanExt.ttl

Requirement **(R3)** is intrinsic to DataID and needs no further representation, while **(R10)** is exemplified by the next section.

Several functional requirements raised by the guidelines of research funding bodies (which are not included in the requirements of this section) will be covered by the DataID service (cf. Section 8). It will provide a versioning system for DataIDs (based on properties like `dataid:nextVersion`), enabling features like tracking changes to a DataID over time. Thereby, the full data management life cycle of datasets is supported **(R12)**, which spans all phases of a Data Management Plan, but this is outside of the scope of this document.

The heart of the DMP extension are two subclasses of `prov:Plan`: The `dmp:DataManagementPlan` provides the most general level of textual statements about the DMP itself or the planned dissemination process **(R9)**, as well as the necessary references to pertaining projects. While `dmp:PreservationPlan` entities can describe different approaches for preservation of different datasets **(R2)** or provide temporal scaling (e.g. regarding embargo periods). Besides textual statements about general goals and provisions for security and backup, using the `dataid-acp:planned` property to point out specific tasks, put in place to preserve data long term, is one of the more notable provenance information.

The concept `dmp:BudgetItem` is an optional tool to list costs pertaining to activities, responsibilities (consequently costs of agents) and any entity involved in a plan like `dmp:PreservationPlan`. Together with `dmp:approxCost` and `dmp:justification` it satisfies requirement **(R11)**.

As a summary; we created 3 classes and 17 properties, which, together with the concepts and properties introduced by the re3data ontology, can describe Data Management Plans as demanded by the requirements of Section 6. An example of a DataID with DMP extension has been created by the ALIGNED H2020 project (e.g. the English DBpedia dataset[24]).

7 CMDI – Component MetaData Infrastructure

The Component MetaData Infrastructure (CMDI) is a component-based framework for the creation and utilisation of metadata schemata[11]. It allows the distributed development of metadata components (defined as sets of related elements) and their combination to profiles in any level of detail, forming the basis for the creation of resource-specific XML Schemata and around one million publicly available metadata files. CMDI is a flexible metadata framework, which can be applied to resources from any scientific field of interest. It is especially relevant in the context of the European research infrastructure CLARIN[12] where it is used to describe resources with a focus on the humanities and social sciences.

The very flexible and open approach of the CMDI which allows for its wide applicability, may lead in parts to problems regarding consistency and INTEROPERABILITY. Despite being rich in descriptive metadata, some CMD profiles lack consistent information of the kind stated in Section 6. This includes the explicit specification of involved persons, descriptions of authoritative structures

[24] http://downloads.dbpedia.org/2015-10/core-i18n/en/2015-10_dataid_en.ttl

Table 1. Most popular CMD profiles and their completeness regarding DataID classes

CMD profile	CMD instances (in % of all)	Supported properties of dataid:Dataset	Supported dataid:AgentRoles
OLAC-DcmiTerms	156.210 (17,4%)	13	3
Song	155.403 (17,3%)	9	1
imdi-session	100.423 (11,2%)	9	2
teiHeader	87.533 (9,7%)	10	2

as well as technical details and actual download locations. Earlier work on the conversion of CMD profiles into RDF/RDFS[13] reflects the complete bandwidth of CMDI-based metadata, but also some idiosyncrasies that may constrain its usage in other contexts. It is expected that a transformation of relevant data to a uniform, DataID-based vocabulary will enhance visibility and exploitation of CMDI resources in new communities. We created explicit mappings for CMD profiles, accountable for 56% of all publicly available metadata files, matching the appropriate DataID classes and applied them on all respective instance files via XSPARQL[25]. An overview of created mappings can be found on Github[26].

The creation and further adaptation of these mappings showed that the support of data considered essential in DataID differs between all profiles. The summary table 1 demonstrates this effect for primary properties of `dataid:Dataset` and the support of different agent roles specified in `dataid:Agent`. Apparently there is a varying degree of conformance of both approaches, indicating possible shortcomings in specific CMD profiles. An example for such a potential deficit is the fine-grained modelling of involved persons or organisations via DataID's Agent concept that is only partially supported in most profiles.

8 Lessons Learned and Future Work

We modularised the DataID ontology into a multilayer composition arranged around a single core ontology. This was necessary to preserve EXTENSIBILITY and INTEROPERABILITY, as the vocabulary was growing due to a plethora of requirements of different use cases. An example of multiple DataIDs already in use can be found with the latest version of DBpedia (2015-10), we stored alongside the datasets (e.g. for the English DBpedia[27]).

We have shown that by extending DataID core with existing addendums and even external ontologies, we could satisfy complex metadata requirements like those of Data Management Plans, while keeping the ability to inter-operate with other metadata vocabularies (like CMDI) in turn. In the wake of this process we incorporated the re3data XML schema into our metadata system, resulting in homogenised metadata. This holds not only for merging external repositories, but also for the identification of potential shortcomings within the same repository as has been shown by converting CMD profiles. The conversion process especially helps to uncover data quality issues and schema gaps.

[25] https://www.w3.org/Submission/xsparql-language-specification/

[26] https://github.com/dbpedia/Cmdi-DataID-mappings

[27] http://downloads.dbpedia.org/2015-10/core-i18n/en/2015-10_dataid_en.ttl

We are in the process of implementing a DataID service and website to simplify and automate the creation, validation and dissemination of DataIDs, supporting humans in creating DataIDs manually, as well as automation tasks with a service endpoint. Additional work has to be done with DataID extensions, to offer additional dataset description options. Integrating DataID fully into the processes and tools defined by the ALIGNED project is another outstanding task. DataID core is planned to be published as a W3C member submission.

Acknowledgements This research has received funding by grants from the H2020 EU projects ALIGNED (GA 644055) and FREME (GA-644771) as well as the Smart Data Web (GA-01MD15010B).

9 References

[1] 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.

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

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

[4] 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.

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

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

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

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

[9] Heinz Pampel et al. "Making Research Data Repositories Visible: The re3data.org Registry". In: *PLoS ONE* 8(11).e78080 (2013).

[10] Jessika Rücknagel et al. "Metadata Schema for the Description of Research Data Repositories". In: GFZ Germans Research Center for Geosciences.

[11] Daan Broeder et al. "A Data Category Registry- and Component-based Metadata Framework". In: *Proceedings of LREC 2010*. European Language Resources Association, 2010. ISBN: 2-9517408-6-7.

[12] Erhard Hinrichs and Steven Krauwer. "The CLARIN Research Infrastructure: Resources and Tools for e-Humanities Scholars". In: *Proceedings of LREC 2014*. European Language Resources Association (ELRA), 2014.

[13] Matej Durco and Menzo Windhouwer. "From CLARIN Component Metadata to Linked Open Data". In: *LDL 2014, LREC Workshop*. 2014.