HUMBOLDT-UNIVERSITÄT ZU BERLIN
INSTITUT FÜR BIBLIOTHEKS- UND INFORMATIONSWISSENSCHAFT

BERLINER HANDREICHUNGEN
ZUR BIBLIOTHEKS- UND
INFORMATIONSWISSENSCHAFT

HEFT 509

MODELING INSTITUTIONAL RESEARCH DATA REPOSITORIES USING THE DCAT3 DATA CATALOG VOCABULARY

A CASE STUDY ON TUDATALIB

VON
ANDREAS GEIßNER

MODELING INSTITUTIONAL RESEARCH DATA REPOSITORIES USING THE DCAT3 DATA CATALOG VOCABULARY

A CASE STUDY ON TUDATALIB

VON
ANDREAS GEIßNER

Berliner Handreichungen zur
Bibliotheks- und Informationswissenschaft

Begründet von Peter Zahn
Herausgegeben von
Vivien Petras
Humboldt-Universität zu Berlin

Heft 509

Geißner, Andreas

Modeling institutional research data repositories using the DCAT 3 Data Catalog Vocabulary : A case study on TUdatalib / von Andreas Geißner. – Berlin : Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin, 2023. – 99 S. : graph. Darst. – (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft ; 509)

ISSN 14 38-76 62

Abstract:

Semantic Web and Linked Data technologies might solve issues originating from research data being published by independent providers. For maximum benefit from these technologies, metadata should be provided as standardized as possible. The Data Catalog Vocabulary (DCAT) is a W3C recommendation of potential value for Linked Data exposure of research data metadata.

The suitability of DCAT for institutional research data repositories was investigated using the TUdatalib repository as study case. A model for TUdatalib metadata was developed based on the analysis of selected resources and guided by a draft of DCAT 3. The model allowed for providing the essential information about the repository structure and contents indicating suitability of the vocabulary and, conceptually, should permit automated data conversion from the repository system to DCAT 3. A loss of expressiveness comes from the omission of dataset series. Conformance with DCAT 3 class definitions led to a highly complex model, thus creating challenges with actual technical realizations. A comparative study revealed simpler models to be used at two other repositories, but implementation of the TUdatalib or a similar model would have potential to improve alignment to DCAT specifications.

DCAT 3 was observed to be a promising option for Linked Data exposure of institutional research data repository metadata and the TUdatalib model might serve towards developing a general DCAT 3 application profile for institutional and other research data repositories.

Eine Online-Version ist auf dem edoc Publikationsserver der Humboldt-Universität zu Berlin verfügbar.

Acknowledgments

First, I would like to thank Prof. Vivien Petras for acting as my thesis adviser and for all helpful input during our meetings. I also want to thank Laura Rothfritz for agreeing to act as second referee.

I am thankful to my colleagues, in particular Prof. Dr. Thomas Stäcker, Dr. Marc Fuhrmans, Dr. Ashish Karmacharya, Qin Zhao, and Gerald Jagusch, for valuable discussions on the Semantic Web, RDF, DCAT, Linked Data as well as DSpace and TUdatalib.

I want to acknowledge the University and State Library Darmstadt for hosting me during my *Referendariat* and Nora Hölzinger for the great organization.

Thanks to all visitors, especially the usual suspects, of the *M26 Zoom Stammtisch* for making a remote Covid study program feel a bit more like a normal one.

Finally, I am deeply grateful to my family, my roommates and my other friends for all support during the preparation of this thesis.

Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

ADMS	Asset Description Metadata Schema
API	Application programming interface
BIBFRAME	Bibliographic Framework
BIBO	Bibliographic Ontology
DC	Dublin Core
DCAT	Data Catalog Vocabulary
DCAT-AP	DCAT Application profile for data portals in Europe
DCMI	Dublin Core Metadata Initiative
DDC	Dewey Decimal Classification
DFG	Deutsche Forschungsgemeinschaft
DOI	Digital object identifier
DXWG	Dataset Exchange Working Group
FaBiO	FRBR-aligned Bibliographic Ontology
FAIR	Findable, accessible, interoperable, reusable
FOAF	Friend of a Friend
FRBR	Functional Requirements for Bibliographic Records
HTTP	Hypertext Transfer Protocol
OAI-PMH	Open Archives Initiative protocol for metadata harvesting
ORCID	Open Researcher and Contributor ID
OWL	Web Ontology Language
R2RML	RDB to RDF Mapping Language
RDF	Resource Description Framework
RDFS	RDF Schema
RML	RDF Mapping Language
SKOS	Simple Knowledge Organization System
SPDX	Software Package Data Exchange
SWRC	Semantic Web for Research Communities
TU Darmstadt	Technical University of Darmstadt

Turtle	Terse RDF Triple Language
ULB Darmstadt	University and State Library Darmstadt
URI	Uniform Resource Identifier
W3C	World Wide Web Consortium

1 Introduction

Making research data publicly available is one of the fundamental processes of open science (1). This allows, among others, for evaluating research reproducibility and tackling new scientific questions based on already available data (2). However, there can be valid reasons, such as privacy concerns, not to make research data available to anyone (1).

A solution to this issue is the concept of FAIR data recommending data to be made "Findable, Accessible, Interoperable and Reusable" (Wilkinson et al. 2016, ref. 3: section 1). FAIR data can be open data that can be accessed and used by anyone, or restrictions can be assigned to it that are then clearly communicated (4). The FAIR qualities are intrinsically reliant on sufficient metadata to describe the data as well as to having computational mechanisms in place that allow for connecting data providers and potential users (3, 4).

Published research data is usually stored in repositories together with its description in form of metadata (5). There is a large number of providers that offer thematically overlapping content (6). A subset of these repositories are those that specialize in a certain content type, for example by collecting research data from one research field without regard for where the data was produced. However, another major subset is formed by institutional repositories that are set up by research institutions for the task of storing and making accessible research data generated at that institution (7).

Ways have to be established to conform to the FAIR principles in such a congregation of different players. One means to do so is Semantic Web technology, that is "currently a popular solution to the knowledge-sharing problem that also fulfil[s] the requirements of FAIR" (Mons et al. 2017, ref. 4: p. 51). Semantic Web technology aims at allowing "[c]omputers [to] find the meaning of semantic data by following hyperlinks to key terms and rules for reasoning about them logically" (Berners-Lee et al. 2001, ref. 8: p. 36). Especially relevant to the issue of research data are the concepts of Linked Data and Linked Open Data that are a collection of rules, guidelines, technologies, and formats built on top of Semantic Web standards (9: chapter 8). In brief, the general idea behind Linked (Open) Data is that data providers create and expose their datasets with links to external entities in the same way that textual web pages refer to other web pages with additional information. However, in the Linked Data context, all information is provided in a structured way with machine-readable semantics (9: chapter 2, 10: chapter 1).

One prerequisite to efficient linking and search is to describe common concepts with identical terms, meaning the use and reuse of published formal vocabularies that define these terms and their relation to other concepts (9: section 8.3.4). Institutional repositories typically provide access to research data from many different scientific domains (7). A generic vocabulary is needed to cover essential information from all datasets (11). Domain specific extension vocabularies could then be added and linked to already existing entities (12). Data Catalog Vocabulary (DCAT) is a World Wide Web Consortium (W3C) recommendation that has been suggested in the context of research data repositories as it "captures many essen-

tial features of a description of a dataset" (Mendez et al. 2020, ref. 11: p. 23). Version 3 (DCAT 3)[1] is currently in development (13). The vocabulary will be described in more detail in Section 2.1.2. Formal studies to test its suitability in real world cases of institutional research data repositories are lacking.

Thus, this study aims at answering several research questions. Is DCAT 3 a suitable vocabulary to provide information about institutional research data repositories in the context of Linked (Open) Data? Can a DCAT-based model provide all essential metadata that is used to describe the repository structure and the datasets contained within? Can such a model be created in a way so that automated data conversion from the repository system is possible? In order to answer these research questions, a model needs to be developed according to DCAT specifications to describe an existing institutional research data repository. TUdatalib[2], the institutional research data repository of the Technical University of Darmstadt (TU Darmstadt), which is based on the repository software DSpace, was selected as study case here.

The model has to match the guidelines set out in the context of Linked Data as well as by the employed main vocabulary, DCAT 3. These are the Linked Data principles (9: section 8.2.1) and the requirements for conformance with DCAT 3 (14: § 4). Those two rule sets, however, not only concern the underlying model but also the actual technical implementation. For easier assessment, they were condensed and transferred to the following requirements for the model:

- Map all available metadata to DCAT 3 and only leave out information or revert to other vocabularies if no fitting DCAT 3 class or property is available, or in case of conflicts
- Include the different classes of the DSpace hierarchy in the model
- Distinguish clearly between the DCAT 3 concepts of catalogs, datasets, data services, and distributions
- Use all vocabulary terms in conformance with the specifications
- Provide a Uniform Resource Identifier (URI) naming scheme that uses unique, Hypertext Transfer Protocol (HTTP)-dereferenceable URIs
- Include links to external entities

The model was created based on the analysis of selected resources from TUdatalib and compared to DCAT models from other research data repositories. This was followed by an evaluation of the possibilities to implement the model in a productive research data repository system. The insights gained in these steps were then used to discuss the research questions.

[1] In this thesis, the acronym DCAT without a version indicator is used to refer to general concepts or multiple versions of DCAT. A reference to one version is specified by DCAT followed by the version number.

[2] https://tudatalib.ulb.tu-darmstadt.de/

2 Background and Related Literature

2.1 RDF vocabularies for describing datasets

The Semantic Web is built upon the idea of a common language to express statements in a machine-readable fashion (15, 16). This language is named Resource Description Framework (RDF). Statements in RDF are called triples because of their pattern of three terms with a fixed order, a subject followed by a predicate and an object with each of these terms, with few exceptions, being defined by a URI or, for objects, as a literal (17). Further languages, RDF Schema (RDFS) and the Web Ontology Language (OWL), have been designed to provide meaning to the different constituents of these triples and to define relationships between them (18, 19). Using these tools, more complex vocabularies or ontologies can be designed to capture the concepts of a certain domain and model those in a machine-readable fashion (20). The classes and properties in these vocabularies can then be used to provide data in the Semantic Web. Typically, in Linked Data, the URIs where the term definitions of a given vocabulary can be found all start in the same way. This identical URI fragment is called the vocabulary's namespace (21).

If multiple providers offer similar Linked Data datasets, clients should be able to process those in an identical fashion with minimal effort to integrate data of multiple sources (9: section 8.3.4). Thus, these datasets should be provided in a way that supports a common understanding of the contents. At the core of this common understanding is the use of broadly employed, standardized vocabularies: If two data providers refer to the same concept, they should use the same term from the same vocabulary (9: section 8.3.4).

Adding to this is the good practice of reusing terms from existing vocabularies when new vocabularies are created (10: section 4.4.6). This is done when existing vocabularies cover part of the requirements for a new application but leave some necessary concepts undefined (22).

Several vocabularies that are in common use provide classes that may be used to describe research datasets.

2.1.1 Dublin Core vocabularies

The Dublin Core (DC) metadata schema predates the invention of the Semantic Web and was originally a set of 15 metadata elements to provide essential information about electronic objects (23). These terms are called the Dublin Core Metadata Element Set. Qualifiers to those 15 terms were added later to be able to create more detailed descriptions leading to the DCTERMS set (23). DC subsequently became one of the first metadata schemas to be expressed as RDF vocabulary with the original set, called DC Elements, and the qualified DCTERMS becoming separate vocabularies in separate namespaces (23, 24). DC vocabularies are maintained by a non-profit organization called Dublin Core Metadata Initiative (DCMI) (25).

DC lists types of electronic resources in a controlled vocabulary called DCMI Type Vocabulary. Datasets are defined as "[d]ata encoded in a defined structure." (DCMI Usage Board, 2020, ref. 24: section 7). However, DC does not offer specialized terms to describe the characteristics of datasets above the generic terms for all electronic resources and also does not offer functionality to describe the aggregation of datasets into data catalogs.

2.1.2 Data Catalog Vocabulary

DCAT: A W3C recommendation for describing data catalogs

Development of the Data Catalog Vocabulary (DCAT) was started at the Digital Enterprise Research Institute at the National University of Ireland, Galway in the context of open government data portals (26). Its design was based on the analysis of seven such data portals with the aim to create "an RDF Schema vocabulary as an interchange format among data catalogues and as a way of bringing them into the Web of Linked Data" (Maali et al. 2010, ref. 26: p. 339). Interoperability was a main focus of DCAT design, thus, "[c]lasses and properties from existing vocabularies, especially Dublin Core, were re-used whenever possible" (Maali et al. 2010, ref. 26: p. 345). Responsibility for DCAT was subsequently transferred to the W3C, first the eGov Interest Group followed by the Government Linked Data Group, leading to the first W3C recommendation for DCAT being published in 2014 (27).

DCAT 1 comprised four core classes (27). It distinguished between the abstract entity of dcat:Dataset[3], implemented as subclass of dataset from the DCMI type vocabulary, and dcat:Distribution that "[r]epresents a specific available form of a dataset" (Maali et al. 2014, ref. 27: § 5.4). In this regard, it is similar to models from the library community that differentiate between layers of abstraction. However, those models, which were developed for handling any kind of bibliographic resource, feature a higher number of abstraction layers: three for the Bibliographic Framework (BIBFRAME) version 2 (29) and four for the Functional Requirements for Bibliographic Records (FRBR) model (30).

Two classes were provided by DCAT 1 to describe the aggregation of datasets into catalogs. The first, dcat:Catalog, was defined as "a curated collection of metadata about datasets" (Maali et al. 2014, ref. 27: § 5.1). The optional class dcat:CatalogRecord allowed for adding "metadata about the *dataset's entry in the catalog*" (Maali et al. 2014, ref. 27: § 5.2) separately from the dataset metadata. This characterization of entire data catalogs and not only the individual datasets is a feature not provided by most of the vocabularies introduced here, except for DCAT and Schema.org (see Section 2.1.3).

Further development of DCAT was performed in the W3C Dataset Exchange Working Group (DXWG) taking into account new requirements, including "current practice in different communities" (Pullmann et al. 2019, ref. 31: § 1). DCAT 2, the current recommendation published in 2020, introduced major changes to the DCAT class structure (32). Noting that datasets are not the only type of resources that might be entries in a data catalog, dcat:Resource was created, a common parent class for all kinds of cataloged resources. DCAT recommends that no instances of dcat:Resource be created, but only instances of specialized subclasses. DCAT 2 provides two direct subclasses of dcat:Resource, namely dcat:DataService

[3]To refer to specific terms from a vocabulary, the prefix:term notation is used in the text of this thesis as, for example, in the Terse RDF Triple Language (Turtle) (28). The prefix denotes the RDF vocabulary the term is defined in. An overview of employed vocabularies with their prefixes, namespaces, and references to definition documents is shown in Appendix Table A.1.

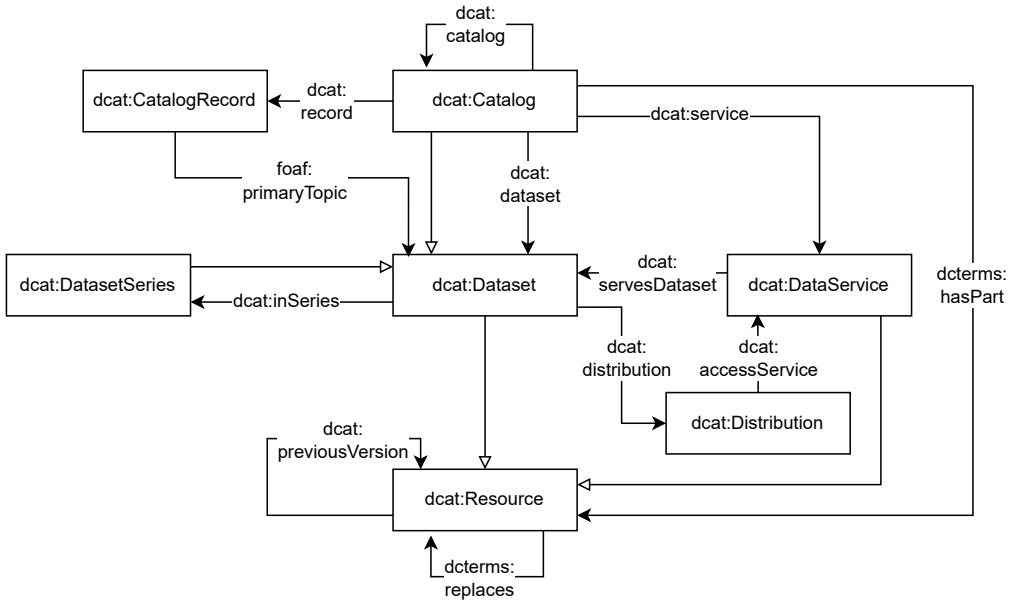

Figure 2.1: DCAT 3 main classes and their relations. The diagram is based on Figure 1 in the third public DCAT 3 draft by Albertoni et al. (14), but restricted to the seven DCAT 3 main classes (without property lists) and their relations. The versioning properties dcterms:replaces and dcat:previousVersion were added. Cardinalities were omitted.

to describe application programming interfaces (APIs) (32: § 5.1) that provide access to dcat:Distributions, and dcat:Dataset with a slightly modified definition from DCAT 1. Thus, since DCAT 2, dcat:Dataset has no longer been a subclass of dataset from the DCMI Type Vocabulary. Furthermore, dcat:Catalog was made a subclass of dcat:Dataset in accordance with the newly introduced scope note that it "represents a catalog, which is a dataset in which each individual item is a metadata record describing some resource; the scope of dcat:Catalog is collections of metadata about **datasets** or **data services**" (Albertoni et at. 2020, ref. 32: § 5.1).

DCAT 3 is currently in development and its third public draft was published in January 2022 (14), followed by a fourth in May 2022 (33) and a fifth in March 2023 (34). The third draft was used as basis for model development here. An overview of the seven main DCAT 3 classes (14: § 5.1) and their relations as defined in that document is depicted in Figure 2.1. DCAT 3 features the implementation of two major additions. With dcat:DatasetSeries, a subclass of dcat:Dataset is introduced to characterize "collection[s] of datasets that are published separately, but share some common characteristics that groups them" (Albertoni et al. 2022, ref. 14: § 6.7). Several terms have also been added to allow for defining relationships between different versions of the same dataset (14: §11). The versioning approach is based on another RDF vocabulary, the Provenance, Authoring and Versioning ontology (35).

DCAT application profiles

Application profiles are vocabularies that are tailor-made for a certain setting by reusing terms from one or multiple other vocabularies and introducing additional constraints on the use of these terms but without creating new terms (36). Several application profiles have been built on top of DCAT, most importantly the DCAT Application profile for data portals in Europe (DCAT-AP) spearheaded by the European Union (37). DCAT-AP 1.0 was published

in 2013 and revised multiple times (37, 38). DCAT-AP 2.0.0 in 2019 aligned the application profile with the new features of DCAT 2 (39). Several other application profiles have been developed based on DCAT-AP (14: § 16).

2.1.3 Other vocabularies

Other RDF vocabularies offer functionalities and terms to describe datasets. However, those vocabularies are usually not tailored for datasets, but include datasets as one possible resource type. The FRBR-aligned Bibliographic Ontology (FaBiO) offers the dataset class as a subclass at the FRBR work abstraction level with a data file as expression (40). Similarly, BIBFRAME 2.1 features datasets as subclass of bf:Work (41). Importantly, while descriptions of datasets are possible with these two vocabularies, they lack the features to describe their aggregation to form data catalogs.

In addition to DCAT, this feature is also provided by the Schema.org vocabulary (42). Schema.org has been designed to annotate web pages in a machine-readable fashion to provide structured information to data processing applications such as search engines (43). The Schema.org representation of datasets and data catalogs was implemented based on the DCAT structure (44) and thus is highly compatible (14: §B). A recommended mapping between DCAT 2 and Schema.org is available (14: § B).

2.2 Linked Data exposure of repository metadata

Libraries, and other cultural heritage institutions, administer large amounts of diverse data with significant informational value when being processed by outside users. Thus, libraries have put effort into converting information to Linked Open Data to make at least part of the data publicly available in a machine-readable format leading to libraries being one of the major applications of Linked Data identified in a systematic review in 2020 (45). The scope of the literature presented in this overview section is limited to studies and concepts that directly concern repositories. A lot of the research cited below has been performed on repositories that handle text-based publications such as books or articles. It is still relevant in the context of data repositories because the same repository systems tend to be used for institutional repositories covering research data even though they are ”clearly optimized for standard research publications; data with different affordances and intended use fit only poorly and with difficulty” (Salo 2010, ref. 46: section 5).

In the context of dissemination of repository metadata, the Open Archives Initiative protocol for metadata harvesting (OAI-PMH) is an important mechanism. OAI-PMH offers a way to exchange metadata via HTTP queries by providing a basic set of commands for selecting records and a standardized way for metadata serialization with XML (47). The XML model allows for embedding of metadata according to different standards (48). Several of the following studies used OAI-PMH as data source for metadata conversion to RDF.

Haslhofer and Schandl (49), in 2008, presented OAI2LOD, a server platform that allowed for creation of RDF triples including assignments of URIs from metadata harvested from an OAI-PMH endpoint, storage in a triple store, and data exposure via request to dereferenceable URIs or SPARQL query. A mechanism to identify outside entities for linking based on metadata field similarity was included. OAI2LOD was a quite inflexible implementation that did not allow mappings between vocabularies and was restricted to one metadata schema

from each OAI-PMH endpoint per OAI2LOD instance. The proof-of-concept implementation delivered DC metadata (49). A follow-up study (47) using a refined version of OAI2LOD demonstrated limitations of employing DC as the use of the same unqualified DC terms for conceptually unrelated information limited the ability to automatically identify by similarity outside entities for linking. Haslhofer and Schandl suggested the use of more semantically rich vocabularies for OAI-PMH and thus, in this implementation, for Linked Open Data (47).

Also in 2008, transformation of qualified DC OAI-PMH metadata to RDF was the approach followed by Koutsomitropoulos et al. (50). Their tool of choice was the OAI-PMH RDFizer (51). Working on the institutional DSpace repository of the University of Patras, they went further by assigning more semantic meaning to the extracted values. They not only converted the metadata to the DCTERMS vocabulary, added datatypes, links to controlled vocabularies and re-used terms from other Semantic Web ontologies like Friend of a Friend (FOAF) (52), they also built their own DC ontology by expanding and refining semantic specifications in the DCTERMS RDF implementation. This ontology was finally expanded to include DSpace-specific concepts and relations. For example, they created new classes for DSpace items, communities, and collections that they implemented as subclass of community (see Section 3.1 for an introduction into the main DSpace classes). Items were being assigned to collections with the dcterms:isPartOf property that links "A related resource in which the described resource is physically or logically included" (DCMI Usage Board 2020, ref. 24: section 2) while the relation between collections and communities was not described in the publication (50).[4] Refinement of this ontology led to a semantic search plugin for early versions of DSpace (53–55).

The reliance on DC metadata in this projects can be explained as this vocabulary is the mandatory minimum for providing metadata via the OAI-PMH interface (47). Dorothea Salo even pointed out that the restraints of the OAI-PMH standard to always provide DC metadata lead to institutional repositories generally storing their metadata using this vocabulary, sometimes with qualifiers (46). This has changed for repositories based on the software Fedora. Fedora has introduced native support of RDF-based metadata with version 4 published in 2016 using the Portland Common Metadata model (56, 57). For others, such as DSpace, qualified Dublin Core continues to be the default metadata schema (58). Additionally, Salo considered the metadata models based on key-value pairs as one of reasons for repository software to be more suited to text-based publications such as books and articles than to research data, specifically adding that with "many, if not most, metadata and exchange standards for research data [using] XML or RDF as a base, this limitation seriously vitiates repositories' ability to manage datasets" (Salo 2010, ref. 46: section 5.2).

Still, in some scenarios, Linked Open Data based on other vocabularies than DC has been created from repositories. Piedra et al. (59) used DCAT 1 to generate Linked Data from the metadata of institutional repositories with a focus on open educational resources. The metadata was harvested from OAI-PMH interfaces and mapped to a model created from multiple vocabularies with DCAT 1 at the center. However, only a partial mapping was presented without, for example, the use of dcat:Distribution[5]. From their own experience of

[4] The publication links to an ontology file which would allow for obtaining this information, but the file appears to be no longer available.

[5] The article provides several links to URIs associated with the model, including a SPARQL endpoint, which could potentially be used to retrieve additional information on the model. However, during writing of this thesis, the respective resources were not available anymore (HTTP 404)

merging data from multiple sources, Piedra et al (59) also outlined key aspects to Linked Open Data publication of repository metadata. These were data source and vocabulary selection, URI design, RDF triple generation from extracted data, and linkage with information made available by other providers.

Latif et al. (60) based their workflow to provide RDF triples from the metadata of the EconStor DSpace repository, specialized in text documents in the field of economics, on extracting metadata directly from the relational database. In their RDF model, URIs and metadata were assigned to authors as well as to the main DSpace classes community, collection, and item. URIs were designed using a pattern based on the entity type and the handle system that is used internally for identification of resources in DSpace. The model used a selection of terms from different vocabularies, but with a focus on DC Elements (60). Information provided for DSpace collections and communities was minimal. Collections were assigned to the class swrc:Collection from the Semantic Web for Research Communities (SWRC) ontology (61) whose definition document is no longer officially available (62, 63). Linked Open Vocabularies recalls swrc:Collection as "book produced from a collection of separate papers" (64). For communities, rdf:type was used to assign the properties dc:publisher and dc:contributor (24: section 3) in a way that is incompatible with the rdfs:Class range of rdf:type (65: section 3.3). A title was provided for communities in the same way as for collections using dc:title and rdfs:label that both allow for naming an entity (24, 52, 60). Latif et al. (60) used dc:publisher, defined as "entity responsible for making the resource available" (DCMI Usage Board 2020, ref. 24: section 3), to connect collections and communities.

Gonzalez-Toral et al. (58) compared workflows to produce RDF triples from qualified DC metadata of the University of Cuenca institutional DSpace repository that specializes in text-based resources. The authors concluded that data extraction from the OAI-PMH interface and subsequent transformations should be preferred to a workflow based on data extraction directly from the relational database. This is mainly due to easier data cleanup and shorter runtimes. For the OAI-PMH approach, the metadata profiles dim and xoai offered by the DSpace module were used that are more expressive than the standard DC. The resulting RDF data used a combination of vocabularies including DCTERMS, the Bibliographic Ontology (BIBO) (66), and FOAF (52). A detailed description of the model was not provided, apart from naming the vocabularies and providing a URI design schema[6]. In this schema, communities and collections as part of the hierarchical structure of DSpace data organization were assigned URIs as well (58).

Dorobăț and Posea (67) proposed a workflow to convert DC XML files, created by DSpace, to RDF graphs according to the specifications of the Europeana Data Model. They stress the importance of creating HTTP URIs to conform with Linked Data principles in addition to having a faithful mapping between the properties of the vocabularies (67, 68).

In 2014, Pascal-Nicolas Becker created a design concept how to use plugins of repository software to make content available as Linked Data (69). The requirements for this concept included observance of Linked Data best practices and observance of the repository design architecture but also ease of use, flexibility and possibility to customize (69: section 4.1). Regarding content, he argued that general solutions should be found for RDF conversion of metadata but conversion of attached files, if required, would have to be handled by additional

[6]The publication refers to two URLs to access the data via SPARQL or an API. This would allow for assessment of the model. However, during writing of this thesis, the respective services could not be reached.

plugins due to heterogeneity (69: sectiona 4.1 and 5.1). He developed a proof-of-concept implementation as a module for DSpace that is still state-of-the-art for DSpace in June 2022 (70). This RDF module allows for generation of URIs and contains an RDF-based configuration to map triples between the DSpace metadata table and RDF vocabularies with regular expressions for string manipulations (69: section 4.3.6).

In the context of this concept, Becker also discussed the use of URIs for information artifacts in repositories. He argued that the approach for real-world objects or non-information resources should be transferred with separate URIs for identification of the entity and for descriptive representations in the form of RDF and HTML documents with the latter being information resources (69: section 3.1.1). On the other hand, he also recommended to reuse the URIs of the HTML representation as identifiers for the RDF resources and to use content negotiation to forward to the RDF serialization which is in contrast with the previous argumentation (69: section 4.2.6). His favorite were the HTTP-URIs associated with digital object identifiers (DOIs) (69: section 5.3).

Neumaier et al. used DCAT 1 in the context of data repositories to assess metadata quality by analyzing what metadata was provided and whether it adhered to expected formats (71). To obtain a homogeneous input for the quality assessment algorithm, metadata harvested from all data portals was converted to DCAT 1. To achieve this, metadata schemas of three data portal software solutions, namely CKAN, Socrata, and OpenDataSoft, were mapped to DCAT 1. For CKAN, Neumaier et al. only made minor changes to an existing mapping provided as part of the CKAN extension ckanext-dcat (71, 72).

This extension was specifically developed to allow CKAN repositories to provide their metadata as Linked Data using DCAT as well as harvest metadata from other repositories (72). As of June 2022, the mapping provided by this extension was still announced to be compatible with DCAT-AP 1.1 (72), even though DCAT 2 and DCAT-AP 2 have been released and DCAT 3 is under development.

The existence of the CKAN DCAT extension and its use by several dozen data portals enabled Neumaier et al. to assess conformance with the DCAT 1 standard as well as identify differences in the data models (73). They pointed to two key issues with the retrieved DCAT RDF data.

They saw differences in how the names and values of custom metadata fields were provided that are not included in the standard mapping of CKAN metadata to DCAT 1 provided by the extension (73). Furthermore, they observed that many datasets had multiple files attached that appeared to contain non-identical information. Those files and the linked metadata were considered to be individual instances of the class dcat:Distribution. The property dcat:distribution was used for the relation between these dcat:Distribution instances and the instances of dcat:Dataset. Neumaier et al. considered this an incorrect use of the concept of distributions (73), citing the DCAT 1 definition of dcat:Distribution that read "Represents a specific available form of a dataset. Each dataset might be available in different forms, these forms might represent different formats of the dataset or different endpoints. Examples of distributions include a downloadable CSV file, an API or an RSS feed" (Maali et al. 2014, ref. 27: section 5.4). The DCAT 2 specifications confirmed their view on distributions, explicitly stating that "all distributions of one dataset should broadly contain the same data" (Albertoni et al. 2020, ref. 32: § 6.7). Neumaier et al. limited their study to datasets and distributions and did not look at the DCAT functions regarding modelling of the entire catalog built around the

class dcat:Catalog (73). Furthermore, the DCAT class structure has undergone significant changes since the study was published (see Section 2.1.2).

The mentioned issue how to model in DCAT the relation between datasets and files belonging to the dataset but providing non-identical information has led to significant discussion. The introduction of a dcat:componentDistribution property along with additional metadata on distribution level has been suggested to be able to express that a dcat:Distribution covers part of the data of a dcat:Dataset (74). However, it was decided not to implement this feature (75). Instead, the concept of a loosely structured catalog was introduced where "There is no distinction made between distribution (representation), and other kinds of relationship (e.g., documentation, schema, supporting documents) from the dataset to each of the files" (Albertoni et al. 2022, ref. 14: § C.1) alongside examples and guidance how to model it in DCAT (14: § C.1, 32: § C.1).

In context of the European Union FAIRsFAIR initiative, Lambert et al. interviewed several data providers and aggregators about their views on the suitability of DCAT 2 and the extensions of DCAT 3 for providing research data metadata (12). The consensus opinion was that DCAT, especially due to its well-fitting class structure, would be a good option to expose metadata from research data repositories with the additions of DCAT 3 offering useful functionality. Significant conversion problems were not expected. No detailed evidence on a model level was provided to substantiate these assessments obtained by interviews (12). It was also suggested but considered challenging to systematically collect already existing mappings between DCAT and other standard vocabularies (12). One such mapping, provided by Milan Ojsteršek in context of the European Open Science Cloud (76), highlights the challenge of transferring the DCAT semantics as the main DCAT classes and the properties connecting those were only mapped to Schema.org or not at all (77). Lambert et al. also pointed towards a recent seminar where specific use cases of data providers were to be investigated for representation in different metadata schemes, including DCAT (12). Investigations of these use cases suggested "potential value in using DCAT" (Lambert et al. 2021, ref. 12: p. 5) but model details do not appear to have been released as of June 2022. A second major conclusion from the interviews was that limited provision of DCAT metadata led to limited incentive for aggregators to harvest this format (12). This low uptake was confirmed by Kazumi Tomoyose both for governmental as well as research data repositories by querying re3data.org for repositories claiming to provide DCAT metadata, identifying only 19, which was less than one percent of those listed (78). Furthermore, directly checking the repositories, only 14 of those 19 could be confirmed to offer metadata in this vocabulary (78).

In summary, the scientific discussion has identified different key challenges related to publishing repository metadata as Linked Data. This includes building workflows and solutions to convert the native metadata of repositories to RDF as well as creating models for Linked (Open) Data publishing using, in general, standard vocabularies and assigning URIs. This study will briefly touch the first challenge but focus on the second one. The center of the model will be DCAT, a W3C standard vocabulary to provide information about datasets and data catalogs of which version 3 is currently in development. For DCAT, incompatibilities between the RDF vocabulary and metadata models of some repositories have been seen, but those observations and the provided solutions should be re-evaluated regularly as DCAT is still under development and addition of new classes and properties might lead to new possibilities or conflicts. Furthermore, studies on research data repositories have so far focused on

datasets in instances of CKAN and should be expanded to repositories using other software solutions.

3 TUdatalib, an Institutional DSpace Research Data Repository

In order to answer the research questions, the research data repository operated by the TU Darmstadt, named TUdatalib, was investigated. The following sections will introduce the DSpace repository software in general and the instance TUdatalib in particular as well as explain the rationale behind choosing this repository as study case.

3.1 DSpace repository software

DSpace is a repository system written in the Java programming language that was started to be developed in the year 2000 and first released in 2002 (79). It is an open source system. The latest major release was DSpace 7 in 2021 (80).

Descriptive metadata for entities in DSpace is represented in a flat fashion as "basically a list of key-value pairs" (Prabhune et al. 2018, ref. 81: p. 173). The standard metadata schema of DSpace to describe digital objects is DC. To provide additional information, the DC Elements properties are refined with qualifiers such as from the DC-Library Application Profile (58, 82). The use of custom metadata fields or alternative vocabularies is possible (58). In addition to this descriptive metadata, there is administrative and structural metadata that is stored in other locations than the descriptive metadata but with partially overlapping information (83).

Content in DSpace repositories is organized in a hierarchical fashion using four classes called community, collection, item, and bitstream (see Figure 3.1 a) (60). Communities and collections form categories into which items, representing datasets, are sorted. Bitstreams are files attached to items. This model leads to a tree-shaped hierarchical structure as depicted in Figure 3.1 b for example entities from TUdatalib[7].

DSpace is considered one of the most commonly used repository solutions (58, 84, 85). This was confirmed by two recent studies that analyzed repositories listed in the OpenDOAR database (86, 87). Both reported that more than 40 % of repositories ran on DSpace. However, whether this high percentage holds true for research data repositories remained unclear as repositories for text-based based publications constituted the majority of repositories in the studies in line with the focus of OpenDOAR (87). A data analysis to interrelate repository content type and repository software was not included in the publications (86, 87). DSpace was also the most often explicitly identified software for research data repositories in a study by Kindling et al. who investigated the repositories listed on re3data.org in 2015 (6). However, DSpace only accounted for 2.6 % of all repositories in the study with most software solutions either "unknown" or "other" (Kindling et al. 2017, ref. 6: section 3.6.2).

[7]Note: Some entities in TUdatalib are named in German while others are named in English. For better readability, German names in text and figures have been translated to English using the name that is provided on the entity's website where such a name could be found. Such translations are not marked.

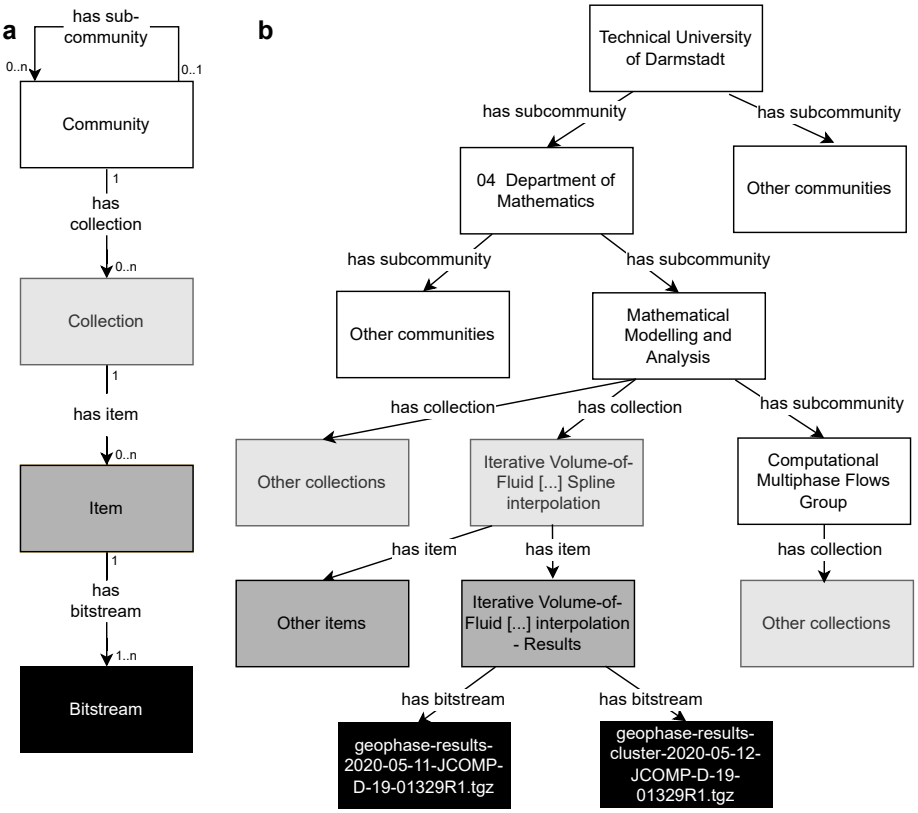

Figure 3.1: Class hierarchy in DSpace and TUdatalib. (a) Abstract diagram showing relations between classes based on Latif et al. (60). As shown by the cardinalities (58), higher level entities may be linked to multiple lower level entities while, in TUdatalib, lower level entities belong to one higher level entity. DSpace would allow more flexibility in this regard (60), but this feature is not currently in use at TUdatalib[8]. (b) Resulting tree structure shown by an excerpt from TUdatalib based on the branch leading to item 2537, one of the investigated items (see Section 4.1).

In Germany, some universities host one common DSpace repository for both text-based and data publications. This includes the eDoc-Server of HU Berlin (88), Refubium of FU Berlin (89), and DepositOnce of TU Berlin (90). Other universities have set up DSpace instances exclusively for research data archiving and publication, such as data_UMR of the University of Marburg (91), DaKS of the University of Kassel (92), and TUdatalib of TU Darmstadt (93). The high popularity of DSpace for repositories in general and its use for institutional research data repositories make a DSpace repository a good choice to study repository modeling for Linked Data.

3.2 TUdatalib

TU Darmstadt offers TUdatalib as a service to its scientists to store, with long-term archiving, data and metadata associated with research at the institution. There is the option but no obligation to publish (94). TU Darmstadt is a university of technology, but research in natural and social sciences as well as humanities is performed besides engineering and related

[8]I confirmed during modeling that the overall number of publicly visible items is the same as the sum of items listed in collections. Thus, every dataset was listed in only one collection.

sciences (95). Thus, TUdatalib was selected as study case being the research data repository of an institution with a broad research interest.

TUdatalib is run by the TUdata team that is a joined effort by the University and State Library Darmstadt (ULB Darmstadt) and the University Computing Centre (94). It was set up for trial operation in 2019 and switched to regular operation in 2022 (96). The service runs on DSpace 6 (97). While TUdata holds central responsibility for the service, the contents as well as the community and collection structures of different subject areas in the repository are self-administered by the scientists who appoint group administrators (96). As such, lower branch levels of the repository tree can look entirely different between subject areas.

The metadata model used in TUdatalib will be looked at in detail in the following chapters for transfer to the DCAT 3 model. In summary, it is mainly DC with few additional TUdatalib-specific fields. The TUdata team offers scientists to add custom, discipline-specific metadata fields (96: section 6.7.2) but that service has not been requested yet[9].

As of June 2022, research data from twelve of the university's thirteen research department has been published on TUdatalib with the Department of Human Studies being the exception (98). Additionally, data published by cooperative research projects and central facilities can be accessed via TUdatalib. Preparations are ongoing for several Hessian universities of applied science to use TUdatalib as research data repository instead of having to host their own repository.

[9]Personal communication with Dr. Marc Fuhrmans, team lead research data services, ULB Darmstadt

4 Methodology

The research presented here was designed as an instrumental case study (99: chapter 1) to perform "*research on a case to gain understanding of something else*" (Stake 1995, ref. 99: p. 171). A model of TUdatalib was created to discern the general suitability of DCAT 3 for institutional research data repositories in the context of Linked Data. As outlined in Chapter 3, the broad research interest of the host institution and the often-employed software solution make TUdatalib a good study case. However, concerning transferability to other repositories, the further the repository setup from the study case used here, the lower the informative value of the model especially when it comes to actual applicability without major changes.

Medina et al. (100), based on Bizer et al. (101), recommended an ordered, four step procedure to publish Linked Data from bibliographic information. These steps are data identification, vocabulary identification, URI assignment, and mapping including links to external entities. This had to be adapted to reflect that the main vocabulary was already decided on and a model was developed instead of actual data dissemination. The workflow used for modeling as such was data identification and collection with rough mapping of properties, mapping of main classes followed by refinement of property mapping and addition of missing classes and external references. Identification of supplementary terms from other vocabularies and proposal of a URI pattern were done last.

Diverse example entities were selected from TUdatalib for metadata analysis. Based on this analysis, the hierarchical structure and metadata model were translated to a graph model specified by RDF triples. Main guidance in the process were the resource definitions and further instructions in the third public DCAT 3 working draft (14). Term definitions of the model originate from this document unless stated otherwise. The entity selection and the modeling steps are described in detail below. As a draft version of the vocabulary was used as modeling reference, the resulting model will have to be validated with the final recommendation once that document is published.

The resulting model was compared to those of two other research data repositories that were selected from the Registry of Research Data Repositories (re3data.org) (102) and whose implementations were analyzed using various tools (see Section 4.4).

4.1 Selection of investigated entities

Entity selection was performed on the level of DSpace items as these are the main entities prepared, equipped with metadata, and submitted by the scientists for storage and publication. The references for the selected items are included in Table 4.1. Entities belonging to the DSpace classes community, collection, and bitstream were investigated based on their link to the selected DSpace items (see below). Only publicly accessible entities were investigated whose metadata is available with a Creative Commons Zero public domain license (103). In

this document, communities, collections, and items are generally referred to via their handle that is their unique identifier in DSpace (104). The landing page of these entities can be accessed via https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/[handle] with [handle] replaced by the respective identifier.

DSpace items for investigation were selected to represent a high diversity with regard to different factors to increase the chance of covering many cases. A maximum of one item was selected per department, but care was taken to include the scientific disciplines humanities and social science as well as natural science and engineering (Table 4.1). Due to the research focus of TU Darmstadt, that resulted in a selection in which engineering and natural sciences were stronger represented than humanities and social sciences. The selection also covered all resource types that scientists can choose for their items (Table 4.2) (105), as well as a range with regard to the number of authors (one to six), the number of resource types chosen for a given item (one to five), and the number of attached bitstreams (one to eight) (Table 4.1).

Table 4.1: References for and characteristics of DSpace items selected for investigation

		Number of		
Handle (ref.)	Department	Authors	Resource Types	Bitstreams
2480 (106)	Law and Economics	1	1	4
2955 (107)	History and Social Sciences	1	2	1
2537 (108)	Mathematics	1	5	1
2662 (109)	Chemistry	3	4	4
2279 (110)	Biology	6	1	2
2416.3 (111)	Architecture	4	3	8
2904 (112)	Mechanical Engineering	3	1	3
2879 (113)	Electrical Engineering and Information Technology	4	1	1
1915.2 (114)	Computer Science	4	3	2

Table 4.2: Resource types associated with investigated DSpace items

Handle	Audiovisual	Dataset	Image	Interactive Resource	Model	Software	Text	Workflow	Other
2480						X			
2955		X					X		
2537		X	X	X		X	X		
2662				X	X	X		X	
2279		X							
2416.3	X	X					X		
2904		X							
2879									X
1915.2		X					X	X	
N	1	6	1	2	1	3	4	2	1

Bitstreams were investigated together with the items they were attached to. Communities and collections were investigated individually if they were located in a branch of the TUdatalib hierarchy leading directly to one of the selected DSpace items. As access to six of the collections was restricted, only data from three out of nine collections was collected.

4.2 Data sources

As the Linked Data graph would be used to present public information, metadata was retrieved from public sources between February 16 and March 11, 2022. Data collected earlier in this period was verified to still be valid between March 8 and March 11, 2022. As such, the data reflects the state of TUdatalib at this point. For items, the main data source was the metadata table included in the full item record located at https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/[handle]?show=full. This was augmented by information found outside the metadata table in the full item record or on the landing page. Furthermore, information was added from the item XML record according to the xoai standard obtained from the OAI-PMH interface of TUdatalib. This XML record was retrieved from https://tudatalib.ulb.tu-darmstadt.de/oai/request?verb=GetRecord&metadataPrefix=xoai&identifier=oai:tudatadatalib.ulb.tu-darmstadt.de:tudatalib/[handle].

Bitstream information was obtained from the same sources as for items and collected alongside those.

Information on communities and collections was obtained from the landing page and from the alphabetical title list belonging to the respective entity with the address https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/[handle]/browse?type=title. Information that was found during data collection on the landing pages of other entities was added.

4.3 Data collection and analysis

A text document was prepared to serve as template for data collection (see Appendix: Data). This document contained in Terse RDF Triple Language (Turtle) (28) style notation the properties of the main DCAT 3 classes (dcat:Catalog, dcat:Dataset, dcat:DatasetSeries, dcat:DataService, dcat:CatalogRecord, and dcat:Distribution) as well as the normative namespace definitions from the third public DCAT 3 working draft (14). This draft was chosen because it was the latest stable document in DCAT 3 development when work on this thesis started. Using this Turtle template, the following analysis workflow was executed for each community, collection, and item including bitstreams. Copies of the files after each step were made to allow for retracing the analysis (see Appendix: Data). All handling of Turtle files was performed using the software Notepad++ v8.3.3[10].

Step 1: Data collection

A copy of this template was created for each investigated entity and information from the data sources as described above was added to fitting properties. Duplicates were allowed both with regard to properties (different information could be added to the same property) as well as with regard to the piece of information (the same information could be added to

[10]Downloaded from https://notepad-plus-plus.org/downloads/v8.3.3/

different properties). More specific sub-properties were preferred to parent properties. Data collection on bitstreams were restricted to those files referred to on the item landing page that were submitted by the user. Bitstreams, usually license documents, that were added to items by the repository system were ignored. Metadata without fitting property was noted as well. If some information in an entity was highly repetitive (e.g. many subcommunities in a community), only part of this information was added to the data with a note to this fact in the respective data file.

Step 2: Data rearrangement and class analysis

The information in the template was distributed to the main classes (see above) so that for each document and class, a list of potential properties was created. These properties were ordered within the class according to how common they were in the DCAT 3 model (e.g. only available for the respective class, inherited from a parent class, or a general property available for all main classes). Based in this distribution of properties to classes as well as definitions in the DCAT 3 draft (14), the assignment between DSpace and DCAT 3 classes was done.

Step 3: Data cleanup and tabulation

Data cleanup was conducted in particular by removing from the documents information made obsolete by the choice of classes. Minor adjustments were made to the syntax to allow for visualization with RDF grapher[11]. Afterwards, properties used for the different instances of each DSpace class and the origin of the respective value was tabulated in Microsoft Excel 2013 (Microsoft, Redmond, USA) to obtain an overview of used properties and metadata fields for each class (see Appendix A.3). Based on this tabulated data, the final decisions on inclusion into the model, creation of additional entities, and inclusion of terms from other vocabularies were made as explained in Section 5.2.

Step 4: Property and entity adjustment and URI assigment

Decisions made in the previous steps on the use of properties or the creation of instances of additional classes to represent multiple bitstream resources or entities such as authors or projects were implemented in the files. A URI pattern was designed and URIs according to this were added to the entities.

Step 5: Merge to create examples

The assigned URIs allowed for merging of item/bitstream files with the collection and community files upward in the hierarchy and cross-linking of entities. These merged files represent examples of what information provided by the model implementation would look like (see Appendix: Data).

4.4 Analysis of data models of other repositories

The Registry of Research Data Repositories (re3data.org) (102) was used to identify repositories that might expose DCAT RDF data. For this, the repository database was filtered by

[11] https://www.ldf.fi/service/rdf-grapher

selecting the criterion "DCAT - Data Catalog Vocabulary" for the field "Metadata standards" in the re3data.org search interface[12] on April 22 and June 11, 2022. The same approach was used in an earlier study by Kazumi Tomoyose to discover data repositories claiming use of DCAT (78). From the resulting list, research data repositories were identified according to the repository descriptions and further examined.

For each of these repositories, at least two English or German datasets were identified arbitrarily (see Appendix A.5) but with a preference to those that had attached multiple, informationally non-equivalent resources if available and not the exception. The landing pages of these datasets were screened for RDF files for download which was also Tomoyose's approach, but who also contacted repositories by e-mail (78). Additionally, the Quick and Dirty (Q&D) RDF browser[13] was used to evaluate whether technical features were used to provide RDF data from the landing page URL. If no DCAT RDF data was obtained using these methods, the following steps were taken: Screen the repository documentation for information regarding DCAT or RDF, or screen the website or the re3data.org repository data for a SPARQL endpoint to directly query for entities of the classes dcat:Catalog or dcat:Dataset.

For all available RDF documents, use of the classes dcat:Catalog and dcat:Dataset was examined. URIs assigned to at least dcat:Dataset were assessed for information exposure upon request with the Q&D RDF browser.

The depositar (metadata licensed with a Creative Commons Zero license (115)) and RDP-CIDAT (investigated datasets licensed with CC-BY 4.0) repositories were evaluated in more detail. The downloaded RDF data of three datasets each (116–121), in case of depositar together with partial catalog data obtained according to the user guide (122), was inspected in Turtle documents and visualized using RDF grapher. Based on the output, a model of the class structure was created. The information included in the RDF data as seen in the visualized graph was compared to the information provided via HTML on the landing page to assess whether all available information was provided via RDF. The data delivered by Linked Data methods that was seen by request with the RDF browser was compared to the information available in the downloaded RDF files.

4.5 Other tools

EasyRdf Converter[14] was used to convert RDF/XML documents to Turtle format. RDFa was analyzed and converted to Turtle using three tools in addition the the Q&D RDF browser whose output was compared: Ruby RDF Distiller[15], W3C RDFa 1.1 Distiller and Parser[16], and RDFa Play[17]. Linked Open Vocabularies[18] (123) was used to search for vocabulary terms outside DCAT. Affinity Designer (Serif Europe Ltd., Nottingham, UK) and draw.io Desktop (JGraph Ltd, Northampton, UK) were used to create figures.

[12] https://www.re3data.org/search
[13] http://graphite.ecs.soton.ac.uk/browser/
[14] https://www.easyrdf.org/converter
[15] http://rdf.greggkellogg.net/distiller?command=serialize
[16] https://www.w3.org/2012/pyRdfa/Overview.html
[17] https://rdfa.info/play/
[18] https://lov.linkeddata.es/dataset/lov

5 Model Development

As described in Chapter 4, the main guidance for modeling was the third draft of DCAT 3 specifications by Albertoni et al. (14). As such, unless otherwise noted, statements on DCAT 3 classes and properties, including those reused from other vocabularies, as well as their relations and use in this chapter are based on that document. It will only be cited if reference to a specific paragraph was deemed essential or advantageous for clarity, which was in particular for references to explanatory, non-normative sections of the document.

Model development was based on data extracted from TUdatalib and collected in text documents in Turtle notation. Metadata from various sources was added to fitting DCAT 3 properties irrespective of the classes that make use of these properties. For some DSpace metadata, this assignment was unambiguous. Other assignments were less clear and will be looked at carefully below, but were treated as correct at this stage. After data collection, for each analyzed entity, properties were distributed to the main DCAT 3 classes according to the DCAT 3 specifications and clustered there depending on whether they were exclusive to this DCAT 3 class.

5.1 Selection of DCAT classes and their relations

5.1.1 Items and bitstreams

First step of modeling was to map DSpace classes (community, collection, item, and bitstream) to DCAT 3 classes. This was attempted at this stage of data analysis.

It was obvious that, if properties that are exclusive to dcat:Distribution were used, those were filled with metadata from bitstreams. Thus, it was decided to map dcat:Distribution to DSpace bitstream. This is in line with the dcat:Distribution definition of "A specific representation of a dataset. A dataset might be available in multiple serializations that may differ in various ways, including natural language, media-type or format, schematic organization, temporal and spatial resolution, level of detail or profiles (which might specify any or all of the above)" (Albertoni et al. 2022, ref. 14: § 6.8).

Mapping DSpace items was less unambiguous. The distribution of the properties pointed to one of the four subclasses of dcat:Resource as many properties were those belonging to these classes. Items not being APIs (14: § 5.1) and dcat:DataService having the least number of fitting properties, this class was ruled out. The class dcat:Catalog had two fitting, class-specific properties. However, the information assigned to foaf:homepage could be delivered using other properties as well and linking to a classification system using dcat:themeTaxonomy would be unnecessary if done at a higher hierarchy level. The respective terms in the classification system assigned for specific items were linked via the property dcat:theme that was available for all dcat:Resources. Importantly, the heterogeneous data described did not fit the dcat:Catalog definition of "A curated collection of metadata about resources." (Albertoni et al. 2022, ref. 14: § 6.3). A choice between the two remaining classes, dcat:Dataset and the

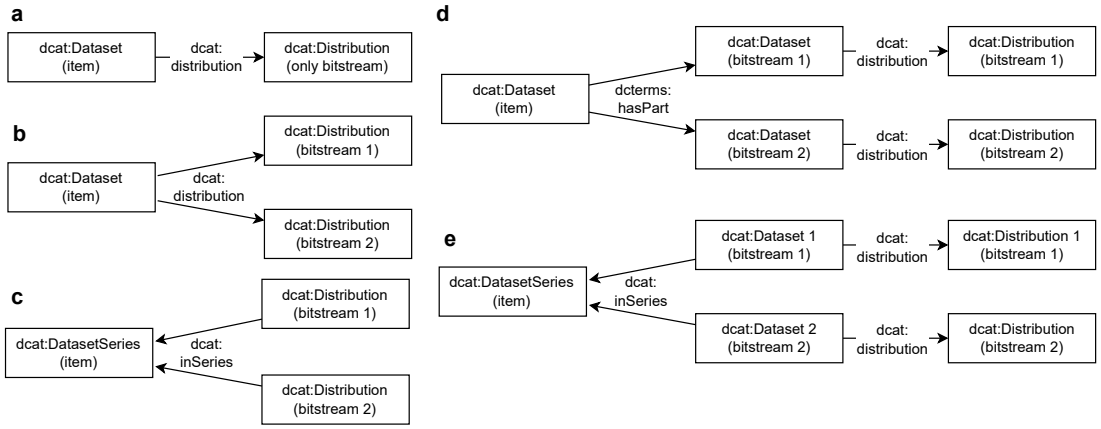

Figure 5.1: Possibilities for describing the DSpace item-bitstream relation using classes and properties from DCAT 3. (a) Relation for an item with a single bitstream. (b-e) Relation for an item with two bitstreams attached representing items with more than one bitstream in general. Details are described in the text.

DCAT 3-introduced dcat:DatasetSeries was not possible on the basis of class properties. This was due to dcat:DatasetSeries being a subclass of dcat:Dataset without additional properties.

Thus, the decision was made according to the relation between the items and their attached bitstreams for which possibilities are shown in Figure 5.1. For three of the nine items, the choice was simple as there was only one bitstream to serve as dcat:Distribution pointing to the relation depicted in panel a that shows the one dcat:Dataset with one dcat:Distribution relationship as in the DCAT 3 basic example (14: § 5.3). Transferring this model to multiple bitstreams in panel b would need each bitstream to be "A specific representation of [the] dataset" (Albertoni et al. 2022, ref. 14: § 6.8) with all of them "broadly contain[ing] the same data" (Albertoni et al. 2022, ref. 14: § 6.8). This was not the case for the analyzed items with multiple bitstreams. For example, those included item 2480 with documentation, software, and data as different bitstreams or item 2416.3 with videos taken at different timepoints from different positions. Thus, in DCAT logic, these bitstreams were distributions of datasets not explicitly seen in DSpace (124).

This left the choice between the models represented in Figure 5.1 panels c, d, and e. Panel d shows the model of a loosely structured catalog employing dcterms:hasPart to connect dcat:Datasets (14: § C.1, 125, 126). In contrast, dcat:DatasetSeries for panel e, modeled based on DCAT 3 examples 37 and 38 (14: § 12.1), is defined as "A collection of datasets that are published separately, but share some common characteristics that groups them" (Albertoni et al. 2022, ref. 14: § 6.7). Panel c represents a short version of panel e that allows for modeling of a dcat:DatasetSeries without having to explicitly show the containing dcat:Datasets (127). This was possible in the DCAT 3 draft used here, but is controversial and probably will be dropped in the final DCAT 3 recommendation (33, 127, 128).

The data suggests that, in TUdatalib, it cannot be generally assumed that the files attached to an item meet the criterion of common characteristics. They did, for example, in case of item 2416.3 that had eight videos as bitstreams representing different times and locations in the same study. Item 2279 had two bitstreams named SingleCellData_CellCycle.zip and SingleCellData_TimeSeries.zip without bitstream description and only a short item description reading "Time series data as well as time of S-phase entry and cell division for individual cells under the different experimental conditions used in the published study" (Benary et al. 2020,

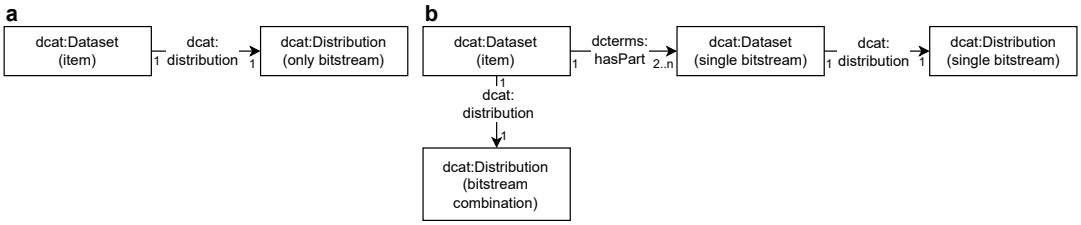

Figure 5.2: Models for representing DSpace items and bitstreams in DCAT 3. (a) Model for items with one bitstream. (b) Model for items with two or more bitstreams.

ref. 110). That kind of item needs someone familiar with the field as well as clear rules for the repository what is considered to have common characteristics. Item 2480 clearly belonged to the case of loosely structured catalog with the different bitstreams containing documentation, software, and data. Items 1915.2, 2662 and 2904 were also seen to be mixtures of different file types with different intrinsic characteristics delivering partial content. Thus, to avoid having to make manual, case-specific decisions, the approach in Figure 5.1 d that is based on unknown relations between item and bitstreams was generally assumed for TUdatalib.

dcat:Datasets of items with multiple bitstreams of different content have dcat:Distributions as well, namely those that represent the combination of all bitstreams (129). Those may even have a direct download link in form of a package (14: § C.5), which was the case for four of the six multi-bitstream items investigated as the repository offered the download as a zip file. Otherwise, the link to a web page that lists all bitstreams is possible using the property dcat:accessURL (130). As DSpace landing pages list all bitstreams attached to an item, these could be linked in this context.

In summary, DSpace item was mapped to dcat:Dataset and DSpace bitstream to dcat:Distribution with the two cases of one or multiple bitstreams handled differently. These are depicted in Figure 5.2 panel a and b, respectively.

5.1.2 Communities and collections

Data analysis suggested two alternatives for DCAT 3 classes to represent DSpace communities and collections, namely dcat:DatasetSeries and dcat:Catalog.

When collecting data from DSpace items, the property dcat:inSeries was used as the only property of DCAT 3 to refer to a higher level entity apart from the generic dcterms:relation. This property is available for the class dcat:Dataset that items were mapped to. Its range would require communities and/or collections to be mapped to the class dcat:DatasetSeries.

A major argument against the use of dcat:DatasetSeries is the issue of diversity, similar to the reason multi-bitstream items were not mapped to this class, seen for the three publicly-accessible collections. Collections 1914 and 2476 included software as well as data from different analyses. Collection 2840 was a container for "measurement data used for publications which are related to the project AMOS [...] sorted by publications" (DFG Project AMOS, ref. 131). It cannot be assumed that data relating to different publications meets the dataset series criterion of shared characteristics.

During the discussion of introducing dcat:DatasetSeries, it was stressed that "a dataset series is *conceptually* a single dataset, but composed of several subsets, which share most of their properties" (Cox 2021, ref. 132). This is a different concept than a collection that serves as a container for various items.

Datasets included in communities at different levels in the DSpace hierarchical tree only got more diverse as multiple collections and subcommunities are joined further up the hierarchy.

The alternative to dcat:DatasetSeries was dcat:Catalog that was suggested by the data analysis of communities and collections themselves. Several dcat:Catalog-exclusive properties were used. This included foaf:homepage whose information could also be provided by other properties as well as dcat:themeTaxonomy to refer to the classification systems used in TUdatalib. Importantly, among the properties were also those that could link to entities lower in the DSpace hierarchy such as dcat:dataset that would connect the dcat:Catalog to the dcat:Datasets representing DSpace items.

The definition of dcat:Catalog reads "A curated collection of metadata about resources" (Albertoni et al. 2022, ref. 14: § 6.3). This fits well to the concept of collections and communities as they organize the repository structure and basically provide an index of the items that are linked to the respective collection or, for communities, linked to any collection downward in the DSpace hierarchy.

That left the challenge of modeling the hierarchy between communities, their subcommunities, and collections. An property had to be found that allowed for representing this relationship in DCAT 3. DCAT 3 provided dcat:catalog with domain and range dcat:Catalog for relations between dcat:Catalogs. However, its description reads "A catalog that is listed in this catalog" (Albertoni et al. 2022, ref. 14: § 6.3.6). This is very similar to dcat:dataset (domain: dcat:Catalog; range: dcat:Dataset) being defined as "A collection of data that is listed in the catalog" (Albertoni et al. 2022, ref. 14: § 6.3.4) and intended to illustrate that a given dcat:Dataset is included in a dcat:Catalog as an entry. Thus, dcat:catalog means that a dcat:Catalog lists another dcat:Catalog as entry, not that the resources listed in the second dcat:Catalog are a subset of the resources listed in the first dcat:Catalog. Importantly, it would mean that the second dcat:Catalog is a resource that the first dcat:Catalog is supposed to provide metadata on (14: § 5.1), not that it is an entity that organizes repository content.

The scope note of DCAT describes dcat:Catalog as "a dataset in which each individual item is a metadata record describing some resource" (Albertoni et al. 2022, ref. 14: § 5.1). Viewing the entries in the catalogs as the data of dcat:Catalog, a subclass of dcat:Dataset, a model similar to the one for multi-bitstream datasets (see Section 5.1.1) should be possible. There, a dcterms:hasPart relation, which allows for breaking datasets into parts (133), was used to illustrate the relationship between the main dcat:Dataset and a second dcat:Dataset representing the part of the data encoded in a specific bitstream. Based on this modeling logic regarding dcat:Dataset, it should be possible to connect a dcat:Catalog with a subcatalog that only contains part of the data, meaning part of the listed entries, with dcterms:hasPart.

However, for dcat:Catalog, dcterms:hasPart has been repurposed to also demonstrate listing items in a dcat:Catalog, namely for any dcat:Resource, reading "An item that is listed in the catalog" (Albertoni et al. 2022, ref. 14: § 6.3.3). As dcat:Catalog, via dcat:Dataset, is a subclass of dcat:Resource, using dcterms:hasPart to connect two instances of dcat:Catalog would mean one dcat:Catalog is listed in the other like for dcat:catalog, not that one dcat:Catalog lists a subset of the entries of the other as would be the dcat:Dataset interpretation.

As part of the modeling process for this thesis and using the argumentation of the previous paragraphs, this difference of interpretation between dcat:Catalog and its parent class dcat:Dataset and the resulting conflict when modeling the DSpace hierarchical category struc-

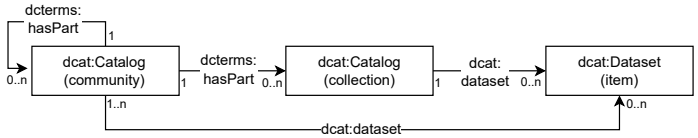

Figure 5.3: Model for representing the DSpace community, collection, and item hierarchy in DCAT 3. The property dcterms:hasPart is used to represent both the relations of a subcommunity within a community and the collection belonging to a community. The cardinalities are chosen to represent the same case as in Figure 3.1 of one lower level entity belonging to exactly one entity of the next higher level as seen for all entities analyzed here. The overall model would be flexible enough to accommodate a more complex hierarchy that DSpace also supports (60). As DSpace items are listed on every level, dcat:dataset is used to connect both communities and collections to the listed items.

ture was raised by me as a GitHub issue with W3C DXWG (134). While leaving it unclear if this was seen as a conflict or only ambiguity (135–137), referring to this post as the only example for problems with the use of dcterms:hasPart in a new GitHub issue specifically created to make a change (135), dcat:resource was introduced as a new object property for dcat:Catalog, defined as "A resource that is listed in the catalog" (Albertoni et al. 2022, ref. 33: § 6.3.3). This property replaced dcterms:hasPart in the context of dcat:Catalog and dcterms:hasPart was made a general property of any dcat:Resource with no specialization for dcat:Catalog (135).

This change allowed for use of dcterms:hasPart to transfer the DSpace hierarchical structure to a model of nested dcat:Catalogs as shown in Figure 5.3 with the dcat:Catalog property dcat:dataset to connect to dcat:Datasets representing DSpace items.

5.2 TUdatalib metadata and DCAT properties

Having decided on the mapping of the main DSpace classes to DCAT 3, the Turtle documents were edited to remove properties not belonging to the chosen DCAT classes or made obsolete by the selected relations as well as to add the relations resulting from the model as developed so far. Following this, it was possible to visualize RDF graphs of single entities. As depicted in Figure 5.4 for item 2904, this visualization led to identification of further issues that needed to be looked at, and changed if problematic, to finish the model. The issues included further properties of the main DCAT 3 classes as well as classes to represent other entities. These will be looked at first before the systematic approach to assigning URIs in Section 5.4.

5.2.1 Item metadata

To obtain a systematic overview of the properties and metadata included in the Turtle documents for DSpace items at this point, the contents were summarized in tabular form. See Appendix Table A.2 for the uncondensed table and Table 5.1 for a condensed form. DCAT 3 provides guidance on whether properties should be used as object property, i.e. refer to another entity, or as data property with a literal in the object position. This information was included in Table 5.1. Additionally, unused properties and unused item metadata were summarized in Table 5.2.

Table 5.1: Assignment of properties of dcat:Dataset to DSpace item metadata fields tabulated from the Turtle documents of item analysis. The uncondensed table is shown in Appendix Table A.2. The third column shows the number of items where the respective metadata field was used (out of the nine analyzed). LP in the second column refers to information available on the landing page. Gray text flags information removed from the model as described in the main text. Other properties were added or specific entities were created or linked. See Appendix Table A.9 for final assignment.

DCAT property	Metadata field/object	N Items	Property type
dcat:landingPage	dc.identifier.uri	9	Object property
dcterms:relation	dc.relation (+subproperty)	3	Object property
	tud.tubiblio	1	Object property
dcat:distribution	Link: distribution	9	Object property
dcterms:creator	dc.contributor.author	9	Object property
dcterms:description	dc.description	9	Data property
dcterms:title	dc.title	9	Data property
dcterms:issued	dc.date.accessioned	9	Data property
	dc.date.available	9	
	dc.date.issued	9	
dcterms:modified	xoai: lastModifyDate	9	DataProperty
dcterms:language	dc.language.iso	4	Object property
dcterms:publisher	tud.unit	4	Object property
dcterms:identifier	dc.identifier.uri	9	Variable
dcat:theme	dc.subject.ddc	9	Object property
	dc.subject.classification	8	
dcterms:type	dc.type	9	Object property
dcat:qualifiedRelation	dc.relation (+subproperty)	3	Object property
	tud.tubiblio	1	
dcat:keyword	dc.subject	5	Data property
dcterms:license	dc.rights.uri	9	Object property
dcat:previousVersion	LP Version history: Item	2	Object property
dcterms:replaces	LP Version history: Item	2	Object property
dcat:version	LP Version history: Version	2	Data property
	dc.description.version	2	
adms:versionNotes	LP Version history: Summary	1	Data property
dcterms:temporal	dc.date.accessioned	9	Data property
	dc.date.available	9	
	dc.date.issued	9	
	dc.subject	1	
prov:wasGeneratedBy	tud.project	3	Object property
	tud.unit	4	
dcterms:hasPart	Link: distribution dataset	6	Object property

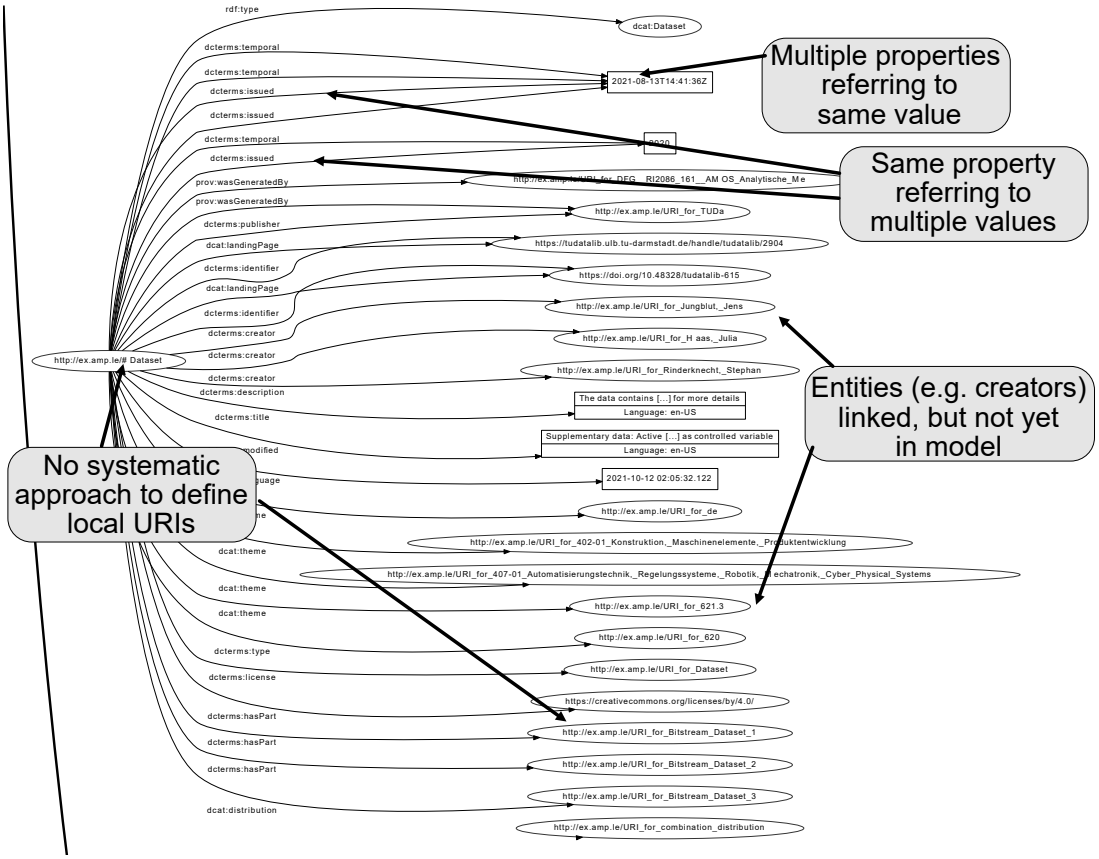

Figure 5.4: Visualization of the RDF graph created by the Turtle document of item 2904 without bitstream information. Boxes highlight open issues of modeling.

Table 5.2: Unused dcat:Dataset properties and item metadata values. Xoai refers to information found in the item's xoai XML representation from the OAI-PMH interface.

Unused property		Unused metadata (N items)
dcat:inSeries	dcat:hasCurrentVersion	dc.rights (9)
dcat:contactPoint	prov:qualifiedAttribution	tud.history.classification (1)
dcterms:rights	dcterms:isReferencedBy	LP creator ORCID link (6)
odrl:hasPolicy	dcterms:accessRights	xoai creator local
adms:status	dcterms:conformsTo	authority key (9)
dcat:first	dcat:accrualPeriodicity	xoai creator local
dcat:last	dcterms:spatial	authority confidence (9)
dcat:prev	dcat:spatialResolutionInMeters	
dcat:hasVersion	dcat:temporalResolution	

Consistently used, unambiguous properties

With dcterms:title, dcterms:description, and dcterms:modified there were three data properties that were used consistently throughout all items. Keywords, while seen for less items, could be mapped to the data property dcat:keyword. Additionally, there were several object properties that were used by many if not all items, but whose linked entities needed to be looked at.

This included the item creators. The corresponding property is dcterms:creator with the range foaf:Agent (14: § 6.4.4). Thus, the scientists as creators should be represented as foaf:Person (14: § 6.12). The metadata provided three pieces of information on the creators: name, local authority key, and – for some authors – a link to an Open Researcher and Contributor ID (ORCID) record (see Tables 5.1 and 5.2). Thus, a minimal local instance of foaf:Person could be created for all authors, as well as an external link to ORCID for a subset as ORCID provides RDF data with authors as foaf:Person instances (138). Direct linking to the ORCID RDF data without creating local foaf:Person instance for scientists with known ORCID would have one disadvantage. A local foaf:Person instance with a URI might be created before the ORCID is added to the profile, resulting in direct linking after RDF update and loss of the locally created instance. Therefore, the owl:sameAs strategy for linking (9: section 8.3.3) was selected to be used in a way as shown in the following example for a creator of item 2904.

```
<#URI> a foaf:Person ;
   foaf:name "Jens Jungblut" ;
   owl:sameAs <https://orcid.org/0000-0002-8056-4345> .
```

TUdatalib uses DSpace Authority Control (139) and, ideally, should only have one record per scientist even if they authored multiple items. This authority information could be used to only model one foaf:Person entity per scientist. The so-far unused DSpace authority key could be included in the URIs for these entities (see Section 5.4).

Another consistently used object property was dcat:theme (14: § 6.4.12) to refer to classes in classification systems. TUdatalib was seen to use two classification systems, the Dewey Decimal Classification (DDC) and the system of Deutsche Forschungsgemeinschaft (DFG). As these are also in use elsewhere, ideally, central instances should be referred to by all providers. This should be possible in case of the DFG classification as an ontology has been designed for it (140). In contrast, the Linked Data service (141) that had been available for the DDC was terminated (142). One strategy here might be to set up, if legally possible, local URIs for the DDC classes that provide the essential information and refer via owl:sameAs (9: section 8.3.3) to the former URIs of Dewey Linked Data (as other services might still do) and possibly in the future to a new central service.

The property dcterms:type (14: § 6.4.13) was used to refer to the resource types, which are, for TUdatalib, a subset of the DataCite resource types (105, 143). The DataCite schema belongs to the ones recommended by DCAT 3 in the context of resource types (14: § 6.4.13). The Linked Data Registry had created RDF entities of those resource types for linking (144), but the service appears to no longer be available.

The value of dc.rights.uri used by all items fit the scope of dcterms:license delivering official URIs of license documents (14: § 9). However, only the linked Creative Commons documents appeared to have RDF descriptions. Official RDF descriptions of the other licenses may

become available in the future and for now the link to the HTML page was used. The DSpace metadata field dc.rights only delivered the name of the license and did not fit any DCAT 3 property. In contrast, the property dcterms:accessRights was used to refer to an external controlled vocabulary stating whether access is public, which was the case for all but one item where access to one bitstream was restricted, or restricted (14: § 9).

Only four of the nine items had associated language information. This information can be provided as linked data by referring to Library of Congress resources as recommended by DCAT 3 (14: § 6.4.9).

Landing page and identifiers

Two different identifier systems were used by the items with the metadata field dc.identifier.uri. In addition to the handle system that DSpace uses internally, four items had a DOI assigned. Both values were used for dcterms:identifier in their dereferenceable form as suggested (14: § 8). Additionally, the handle URIs are identical to the URL of the HTML landing page describing the item and thus were also used with the property dcat:landingPage (14: § 6.4.17). DOIs, in contrast, were seen as identifiers for the abstract entities of a DSpace item that forward to the URLs of the landing pages, represented by handles if HTML is requested, or to other formats by content negotiation (145). Thus, all values of dc.identifier.uri would be used with dcterms:identifier, but only the handle ones with dcat:landingPage and selection would have to be carried out, for example by pattern recognition, upon RDF triple generation.

Date and time

In addition to dcterms:modified, usage of two time related DCAT 3 properties was seen, namely dcterms:issued and dcterms:temporal. The first one relates to the publication date. For dc.date.issued, TUdatalib asks about the date the respective dataset was created (96: section 6.7.2). As this information is distinct from the publication date in dcterms:issued, one of the other two values that are automatically generated upon submission should be referred to here. Officially, dc.date.available should refer to publication date in the repository (146) and therefore be used for dcterms:issued, even though there are still technical problems with the implementation of the value of this field (147).

The property dcterms:temporal indicates "[t]he temporal period that the dataset covers" (Albertoni et al. 2022, ref. 14: § 6.6.6). In certain cases, this would be identical to the dataset creation time saved in dc.date.issued, for example for the videos in item 2416.3. For others, such as the list of Middle High German words in item 2955, it would be different. As such, reliable values were not available for this property, but it might be a candidate for future information collected upon item submission. While DCAT 3 did not provide a fitting mapping for the TUdatalib value of dc.date.issued, dcterms:created to provide the "[d]ate of creation of the resource" (24: section 2), was an easy outside addition to the model and potentially would be to DCAT itself.

Relations

There were three items (1915.2, 2662, and 2879) with user-submitted relations. Overall, there were five relations defined to external resources as well as one to TUbiblio[19], the bibliography of TU Darmstadt. The types of three external relations (references, is version of, is part of) belonged those available in the DCTERMS vocabulary (24: section 2) and those terms should be used as per DCAT recommendation (14: § 6.4.14). The TUbiblio relation could be modeled using the term dcterms:references that connects to "A related resource that is referenced, cited, or otherwise pointed to by the described resource" (24: section 2). Ideally, this would be to an RDF description, but this is not available yet. The remaining two external relations (is described by, is supplement to) were of a type from the DataCite relations as are all relations available for selection in the submission form of TUdatalib. Thus, for all relations that cannot directly be mapped to DCTERMS, use of dcat:qualifiedRelation would be possible (14: § 6.4.15). This is a property that allows for use of relation types outside the standard set from DCAT or DCTERMS. The DataCite relations belong to those recommended for use (14: § 15).

Publisher and projects

The dataset publisher is supposed to be referred to with the object property dcterms:publisher with entities of foaf:Agent as recommended values (14: § 6.4.10). In connection with the expansion of TUdatalib to several universities of applied science, tud.unit was established as a metadata field that automatically refers to the institution in whose catalog the item was published[20]. Thus, the value "TUDa" seen for items 2662, 2879, 2904, and 2955 corresponds to creating an RDF representation of TU Darmstadt as instance of foaf:Organization and referring to it via dcterms:publisher. This instance should also be linked in legacy items that had not yet been assigned a value in the tud.unit field. The property owl:sameAs (9: section 8.3.3) could be used to refer from that entity to external ones also describing the institution, such as wikidata[21] or dbpedia[22].

Another metadata field that had information about the dataset origin for three items was tud.project that referred to third party funded projects. This kind of information fits the scope of prov:wasGeneratedBy with the range of prov:Activity that is included in DCAT 3 (14: § 6.6.8) reused from the PROV Ontology (148).

Versioning

Two of the analyzed items were newer versions of previously published datasets (items 1915.2 and 2416.3). Properties newly introduced in DCAT 3 can be used to refer to the URI of the dcat:Dataset representing the previous (dcat:previousVersion) and any version (dcterms:replaces) in the item's history (14: § 11). Older versions of items in TUdatalib can only be referred to via the new versions or the direct link and are not listed in collections or communities (96: section 6.8.5). This was done accordingly in the TUdatalib model by referring to old versions from later ones via the properties introduced here but not via the property dcat:dataset from catalogs.

[19] http://tubiblio.ulb.tu-darmstadt.de/
[20] personal communication with Qin Zhao, TUdatalib administrator, ULB Darmstadt
[21] http://www.wikidata.org/entity/Q310695
[22] http://dbpedia.org/resource/Technische_Universität_Darmstadt

DSpace allows to enter a version summary that, in the cases seen here, fit the scope of the property adms:versionNotes, reused from the Asset Description Metadata Schema (ADMS) vocabulary (149), that should be used to outline changes between versions (14: § 6.4.28). Furthermore, the property dcat:version refers to version numbers or identifiers (14: § 6.4.27). Two different identifiers were seen. The metadata field dc.description.version was used in item 1915.2 as well as in item 2537 that in TUdatalib only had one version. Additionally, DSpace counted new versions by itself. It was decided to use dcat:version for the uniform version numbers that TUdatalib assigns automatically when generating a version sequence as dc.description.version clearly did not reflect versioning within the repository. The other information could potentially be delivered with a fitting property from another vocabulary. The ADMS vocabulary (149) that is closely related to DCAT recommends owl:versionInfo for such an identifier, however, DCAT 3 itself warns that this property is supposed to be used in context of resources that are ontologies (14: § 11).

Other versioning properties, dcat:hasVersion and dcat:hasCurrentVersion, remained unused. These properties are intended to refer from constantly changing resources to snapshots (14: § 6.4). As items in TUdatalib are supposed to be stable for permanent referencing, the properties was not considered relevant here.

Unused properties

Thus, almost all available information on the investigated DSpace items could be transferred to DCAT 3. In contrast, several dcat:Dataset properties remained unused. For a subset of those, such as dcat:contactPoint essentially describing corresponding authors (14: § 6.4.3), it might be considered collecting metadata in TUdatalib in the future.

5.2.2 Metadata of bitstreams

According to the general model outlined in Section 5.1, DSpace bitstreams were mapped to dcat:Distributions. In case of multi-bitstream items, an additional dcat:Dataset was necessary to represent the part of the item that is delivered in a given bitstream. The properties of the classes dcat:Distribution and dcat:Dataset needed in this context will be looked at separately in this section.

Metadata of dcat:Distribution

The same approach as for items of tabulating DCAT properties and DSpace metadata fields was used to obtain an overview of alignments for bitstreams. This was done on the level of items. That means for multi-bitstream items that if a dcat:Distribution property was assigned to a metadata field of at least one bitstream of that item, it was included in the table. The full data is displayed in Appendix Table A.3. Table 5.3 shows a condensed version. For conciseness, where choices and argumentation were identical to items, these choices have already been heeded when Table 5.3 was assembled.

The assignment of several properties was unambiguous. However, file names cannot be seen as an ideal choice for the distribution title as they do not necessarily provide information about the file content. In absence of an alternative and, as such speaking file names were generally seen in the analyzed items, the mapping was still performed like that, but requesting submission of titles for files might be considered in the future. Another exception

Table 5.3: Assignment of properties of dcat:Distribution to DSpace bitstream metadata fields tabulated from the Turtle documents of item analysis. The uncondensed table is shown in Appendix Table A.3. The third column shows the number of items where the respective metadata field was used (out of the nine analyzed). Xoai in the second column refers to information found in the item's xoai XML representation from the OAI-PMH interface. Gray text flags information removed from the final model as described in the main text. Further changes were made due to creation of specialized entities. See Appendix Table A.11 for final assignment.

DCAT property	Metadata field/object	N Items	Property type
dcterms:accessRights	BITSTREAM Access	1	Object property
dcterms:description	BITSTREAM Description	4	Data property
dcterms:title	BITSTREAM NAME	9	Data property
dcterms:issued	dc.date.available	9	Data property
dcterms:modified	xoai LastModifyDate	9	Data property
dcterms:license	dc.rights.uri	9	Object property
dcat:accessURL	dc.identifier.uri	9	Object property
dcat:downloadURL	BITSTREAM URL	9	Object property
	ZIP Package URL	4	
dcat:byteSize	xoai BITSTREAM SIZE	9	Data property
dcat:mediaType	xoai BITSTREAM Format	9	Object property
	ZIP Package Format	4	
dcat:compressFormat	BITSTREAM Format	5	Object property
	ZIP Package Format	4	
dcat:packageFormat	ZIP Package Format	4	Object property
spdx:checksum	xoai checksum	9	Object property
	xoai checksum algorithm	9	

was dcterms:modified (14: § 6.8.4) that was removed from the list as the modification date of the whole item does not necessarily reflect the modification date of an attached bitstream. Another exception was the issue of file types. DCAT offers the property dcat:mediaType for the type of any bitstream (14: § 6.8.16), but special properties for compressed and packaged distributions (14: § 6.8.18, § 6.8.19, and § C.5). dcat:mediaType is an object property referring to IANA media types (14: § 8.16) that were also seen to be used in TUdatalib. In the model created here, the use of these properties was made based on the context of the respective file to discriminate between actions of scientists and the TUdatalib system. The property dcat:mediaType was reserved for the type of any attached bitstream as uploaded by the scientist, even if this is a compressed package containing other file types such as for items 1915.2 and 2279 (110, 114). Ideally, dcat:mediaType should refer to the file type within the package (14: § C.5), but this might again be a mix as in case of 1915.2 (114). Thus, this was handled pragmatically. The properties dcat:compressFormat and dcat:packageFormat, in contrast, were assigned to the packaged download of all bitstreams from one distribution, meaning to packages created by TUdatalib to make available single distributions of multi-bitstream items (see Section 5.1.1).

One of the bitstreams from item 1915.2 was access restricted. Same as for items, this information can be delivered using the property dcterms:accessRights and a controlled vocabulary including terms for public and restricted access as suggested by DCAT 3 (14: § 9).

Table 5.3 points to one more class needed in the model. Checksums are not a direct property of dcat:distribution, but this class refers to spdx:Checksum via the property spdx:check-

sum (14: § 6.8.20), reused in DCAT 3 from the Software Package Data Exchange (SPDX) vocabulary (150). There, the properties spdx:checksumValue and spdx:checksumAlgorithm are available for the respective values from TUdatalib (14: § 6.17).

Metadata of dcat:Dataset for multi-bitstram items

The previous sections showed that essentially all metadata for DSpace items and bitstreams could be assigned to dcat:Dataset and dcat:Distribution, respectively. As such, doubling of information was necessary to describe the dcat:Datasets representing the part of an item served in a dcat:Distribution. Taking metadata from both item and bitstream would be possible, however, for information from the item it had to be taken care that only metadata was used that is sure to apply to all bitstreams as well. The chosen metadata is listed in Table 5.4. The bitstream sequence identifier (104: section 4.2) was not added as an identifier to the bitstream dcat:Dataset as its meaning depends on the identifier of another entity, the handle in the item dcat:Dataset.

Table 5.4: Properties used to describe part dcat:Datasets for bitstreams in multi-bitstream items

Property	Metadata field
dcat:landingPage	dc.identifier.uri
dcterms:issued	dc.date.available
dcterms:license	dc.rights.uri
dcat:theme	dc.subject.classification + dc.subject.ddc
prov:wasGeneratedBy	Link: project entity (tud.project)
dcterms:accessRights	BITSTREAM Access
dcterms:title	BITSTREAM Name
dcterms:description	BITSTREAM Description
dcat:distribution	Link: Distribution

5.2.3 Metadata of communities and collections

The same approach of tabulating properties from the Turtle files was used for communities and collections. The full information is shown in Appendix Tables A.4 and A.5, and the condensed form in Table 5.5. In general, a quite uniform appearance was seen both within each class and between the classes that serve similar purposes.

One difference was the information delivered about communities in the higher level communities and about collections in communities they belonged to or their listed items. While collections had a text description and that information was therefore assigned to dcterms:description as a second value beside the full description in the collection itself, communities had internal entity identifiers of TU Darmstadt that were assigned to dcterms:identifier. If such a pattern switched from identifiers to descriptions in lower level communities as well, mechanical pattern analysis might be an option to discriminate between the use of those properties.

Furthermore, even though the landing pages of communities and collections only referred to the DFG classification, the property dcat:themeTaxonomy (14: § 6.3.2) should be used to refer to the DDC as well as its use was seen for items (see Section 5.1.1).

In item 2840, a news text was included to describe the addition of datasets to the collection over time. Even though news is a sort of description about a resource, it is a specialized one.

Table 5.5: Assignment of properties of dcat:Catalog to DSpace community and collection metadata fields tabulated from the Turtle documents. Gray text flags information removed from the final model as described in the main text. The third and fourth columns show the number of communities (out of 24) and collections (out of 3), respectively, where the metadata field was used. The uncondensed tables are shown in Appendix A.3.3.

Property	Metadata field	N	
		comm.	coll.
foaf:homepage	LP URL	24	3
dcat:landingPage	LP URL	24	3
dcat:themeTaxonomy	LP DFG classification	24	3
dcterms:hasPart	LP Subcommunities	16	0
	LP Collections	9	0
dcat:dataset	Title List	24	3
dcterms:description	LP Description	1	3
	External Information: Description in item/community	0	3
	LP News	0	1
dcterms:title	LP Title	24	3
dcterms:identifier	LP URL	24	3
	External Information: Identifier in higher level community	20	0
adms:versionNotes	LP News	0	1

Thus, a more specialized property than dcterms:description should be used. With adms:versionNotes (14: § 6.4.28), the DCAT 3 vocabulary offers one property to describe changes between versions. However, with collection being a non-versioned, potentially continually changing entity, properties relating to describing specific versions should be avoided. A fitting property from another vocabulary for a use case as seen here is skos:historyNote (151: section 7.2) from the Simple Knowledge Organization System (SKOS) vocabulary. If use of the feature were found to be highly diverse, it might be prudent to map it to a generic property such as rdfs:comment (65: section 3.7).

One community, 2212 of the Computational Physical Chemistry Group, had an associated logo depicting a logo of the research group. DCAT 3 does not offer specific terms to include logos. They might be modeled like other DSpace bitstreams, but they do not represent research data and there is insufficient metadata. Thus, using terms from other vocabularies was preferred. Of three fitting properties from widely employed vocabularies, I decided on using vcard:hasLogo (152: section 4.2) as it has the least associated conditions being an owl:ObjectProperty without domain or range definitions. The term schema:logo (153) comes with domain specifications and foaf:logo (52) is an inverse functional property potentially giving issues with logo reuse.

Catalog distributions

Appendix Tables A.4 and A.5 also pointed to properties of dcat:Catalog to which no values were assigned. Many of those were inherited from dcat:Dataset and deemed irrelevant for dcat:Catalog in this model. One notable exception was dcat:distribution. Seeing dcat:Catalog as dcat:Dataset whose content is resource metadata as in the scope note of DCAT 3 (14: § 5.1), a file delivering this metadata would be a dcat:Distribution. In DSpace and other

repository systems, this kind of information is provided in XML format via the OAI-PMH interface. As such, these XML documents could be seen as dcat:Distributions of a dcat:Catalog. A direct link to a specific XML document with a given metadata schema containing all metadata of items in a community or collection would then be the value delivered with the dcat:Distribution property dcat:downloadURL. Furthermore, OAI-PMH would constitute a dcat:DataService to access metadata included in the individual catalogs of communities and collections. Indeed, OAI-PMH had already been identified as a potential use case for dcat:DataService earlier (154), but there appears to have been no follow-up on this. The idea of catalog distributions was also included in my GitHub post (134), with Andreas Perego replying that "the OAI-PMH notion of 'set' can actually be mapped to dcat:Catalog" (Perego 2022, ref. 137). This agrees with this concept as the different OAI-PMH XML formats would be the files/distributions delivered by the dcat:Catalog mapped to set.

OAI-PMH is a widely employed standard for metadata transfer (155). Analyzing this standard and creating a best-practice model for its description in DCAT 3 is outside the scope of this thesis. Thus, the information provided about the OAI-PMH interface in this model is minimal.

Other unused properties

Apart from dcat:distribution, other properties were seen not to contain values in Appendix Tables A.4 and A.5. However, information could be added for several of those. The property dcterms:publisher should again refer to the RDF description of TU Darmstadt. Furthermore, all metadata in TUdatalib is published under the public domain Creative Commons CC0 license (103). Thus, the URI of this license was referred to via dcterms:license (14: § 6.4.19). Public metadata was flagged as such using the property dcterms:accessRights (14: § 6.4.1).

5.2.4 The class dcat:CatalogRecord

The class dcat:CatalogRecord (14: § 6.5) is an optional class that does not describe a resource but resource's catalog entry. In the analyzed data, several properties would have been fitting for dcat:CatalogRecord, but this would have exclusively been redundant information. As such, inclusion of this class based on the available data was deemed unnecessary. In certain cases use of this class might be considered, such as for harvested metadata when date of dataset publication and entry into the catalog might differ.

5.3 Model overview

In the previous sections, diverse examples of the main entities of DSpace in the context of TUdatalib have been analyzed. Based on this analysis. a model to translate this information to the classes and properties of DCAT 3 was created. The core model containing the major classes and relation properties is depicted in Figure 5.5. Tables listing the properties for each class are located in Appendix A.4.

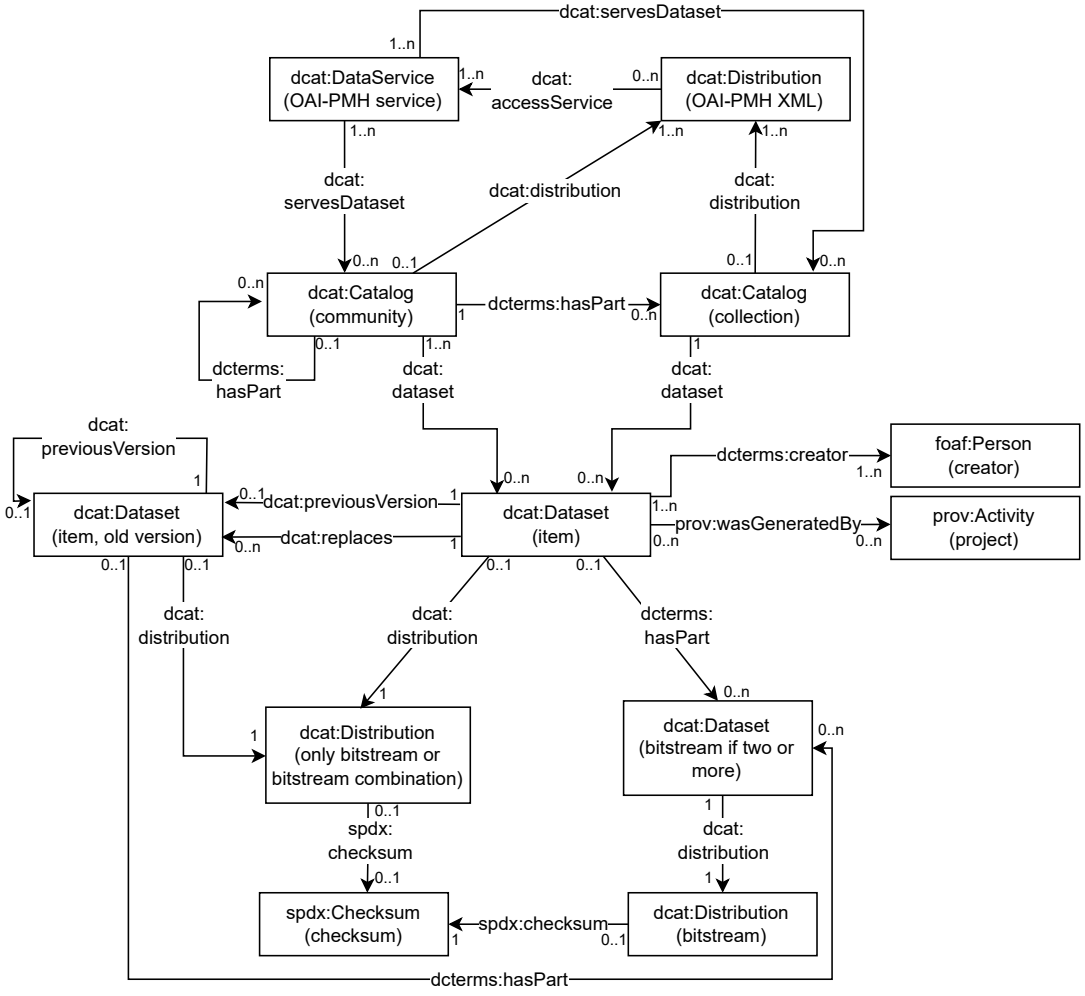

Figure 5.5: Core model for describing TUdatalib using the DCAT 3 vocabulary. Cardinalities differ from Figure 5.2 due to merging of the single and multiple bitstream cases into one diagram as well as inclusion of old item versions. For clarity, only classes based on TUdatalib entities or data have been included in the diagram. For all relations, including external ones, see Appendix A.4. Also for clarity, relations between old versions of datasets and creators/projects were excluded from the diagram as was the class foaf:Organization that would have a dcterms:publisher relation to six other entities.

Table 5.6: Pattern to assign URIs to instances of classes in the model. [handle] denotes to the handle of the respective community, collection, or item, [sid] the DSpace bitstream sequence identifier (104: section 4.2), and [authkey] the authority key of a creator. For OAI-PMH, [context] and [ms] stand for OAI-PMH instances according to different harvester guidelines and the available metadata schemas, respectively (156). [root] abbreviates the beginning of the dereferenceable URI. For TUdatalib, https://tudatalib.ulb.tu-darmstadt.de/rdf as in the standard setting of the DSpace RDF module (70) would be a possibility for [root].

Class	Context	URI pattern
dcat:Catalog	Community and collection	[root]/Catalog/[handle]
dcat:Distribution	Community and collection	[root]/OAI/[context]/[ms]/[handle]
dcat:DataService	Community and collection	[root]/OAI/[context]/Service
dcat:Dataset	Item	[root]/Dataset/[handle]
dcat:Dataset	Bitstream	[root]/Dataset/[handle]/[sid]
dcat:Distribution	Only bitstream or bitstream combination	[root]/Distribution/[handle]
dcat:Distribution	Single bitstream of multi-bitstream item	[root]/Distribution/[handle]/[sid]
foaf:Person	Creator	[root]/Creator/[authkey]

5.4 URI assignment system

According to the Linked Data principles (10: section 2.1), entities should be equipped with permanent HTTP URIs that allow for provision of information about the entity upon request. Pascal-Nicolas Becker (69: section 3.1.1) recommends handling repository content like abstract-entities or non-information resources.

For abstract entities, two general approaches that Aidan Hoagan termed hash recipe (9: section 8.2.2.2) and slash recipe (9: section 8.2.2.3) exist to generating URIs that differ from those of the resources providing information about the entities while still referring to those resources. Briefly, in the hash recipe, the URI of another entity can be extended to create a new URI, but the HTTP request will be directed to the original URI. In the slash recipe, completely independent URIs are designed from which the HTTP request is redirected to another resource providing the information (9: section 8.2.2).

The slash recipe was chosen to represent the abstract entities in the model using a method of URI design similar to Latif et al. (60) who included in their URIs the entity type, the handle, and whether an information or non-information resource was referred to. A general pattern (see Table 5.6) was designed to represent the different entities in the core model without overlapping with the URIs of DSpace HTML pages that start with https://tudatalib.ulb.tu-darmstadt.de/handle/. When designing the pattern, identifiers provided in the DSpace data model were reused allowing easy cross-reference and avoiding the need to design a second set of identifiers. From this URI, technical measures have to ensure referring the client to the appropriate information resource that are HTML pages or serializations of the RDF descriptions in a format that can be processed by the client (10: section 2.3). All URIs would be located in a namespace that is under the control of the RDF provider as recommended (9: section 8.3.1).

DCAT 3 RDF guidelines discourage the use of blank nodes with the DCAT 3 main classes (14: § 5.2). Otherwise, blank nodes might have been used for the dcat:Dataset representing bitstreams that were added for formal correctness of the model but add little informational value. Still, in the case of TUdatalib, there were reasons to use blank nodes instead of URIs for instances of two classes from the core model depicted in Figure 5.5. The first was spdx:Checksum as there was no need seen to refer to file checksums from a context not involving the respective bitstream. The second one was prov:Activity for projects where data was still preliminary as seen by the incomplete information for item 2904 and the recording system might change[23].

Assignment of URIs allowed for cross-linking of all entities related to an item and the communities and collection upwards in the repository hierarchy. Due to their large size, these final documents of the investigated entities were not included in the print appendix but are available alongside all precursors in the research data (see Appendix A.1).

[23] Personal communication with Gerald Jagusch, Head of Information Technology, Research and Development, ULB Darmstadt

6 Comparison to Other DCAT Implementations

A comparison of the TUdatalib model to other repositories that use the DCAT vocabulary to provide metadata was intended. This was to evaluate whether other current implementations differ significantly from the model developed here and and whether adaptation of the new model might improve metadata delivery outside TUdatalib.

The repository database re3data.org was queried to identify potential repositories for this comparison. Filtering for DCAT as metadata standard as done before by Kazumi Tomoyose (78) led to a list of 24 repositories with varying scope out of 2864 listed repositories overall (102). No DSpace repositories were found on the filtered list. A use case not too dissimilar to TUdatalib was preferred for the comparison, thus, government data repositories as well as the musical source database RISM was removed from the list. The remaining thirteen repositories were looked at in more detail (see Appendix A.5). In brief, only a subset (nine of thirteen with another one of the thirteen not publicly accessible but unlikely according to its FAQ page (157)) provided DCAT metadata that could be used for further analysis. Nine (six positives) of those were already investigated in the earlier study to identify repositories exposing DCAT metadata by Tomoyose (78) with identical result. Tomoyose did not perform a detailed analysis of the models (78). Of the ones looked at here, the depositar[24] repository was seen as closest in scope to TUdatalib being a repository provided by an academic institution, Academia Sinica, Taiwan, for mainly academic research data without a focus on a particular field of study (158). Furthermore, depositar clearly stated that its DCAT implementation was supposed to be compatible with DCAT 2, a statement not seen for any other repository, even though the implementation was still considered to be in beta phase (122). In addition to this CKAN installation, a second repository for academic research data but running on another software package was selected. This was the RDPCIDAT[25] DKAN repository of the Ruhr Universität Bochum Research Department Plasmas with Complex Interactions.

For each repository, three datasets of the most complex case, those composed of several files with non-overlapping content, were looked at (116–121). For one dataset, Lin et al. (116), for depositar and for all datasets for RDPCIDAT, file types and contents were, in my opinion, clearly incompatible with the shared characteristics requirement of dataset series (14: § 6.7). An additional resource describing the repository content with the class dcat:Catalog was available for depositar but did not appear to exist for the RDPCIDAT repository. Figure 6.1 shows the core structure of the data models as derived from the downloaded RDF data.

Two major differences to the TUdatalib model could be seen for depositar, one concerning dcat:Catalog and one the relation between dcat:Dataset and dcat:Distribution. Only one instance of dcat:Catalog without subcatalogs for different categories in the repository was included in the model. Furthermore, every file was considered a dcat:Distribution of the

[24] https://data.depositar.io/en/
[25] https://rdpcidat.rub.de/

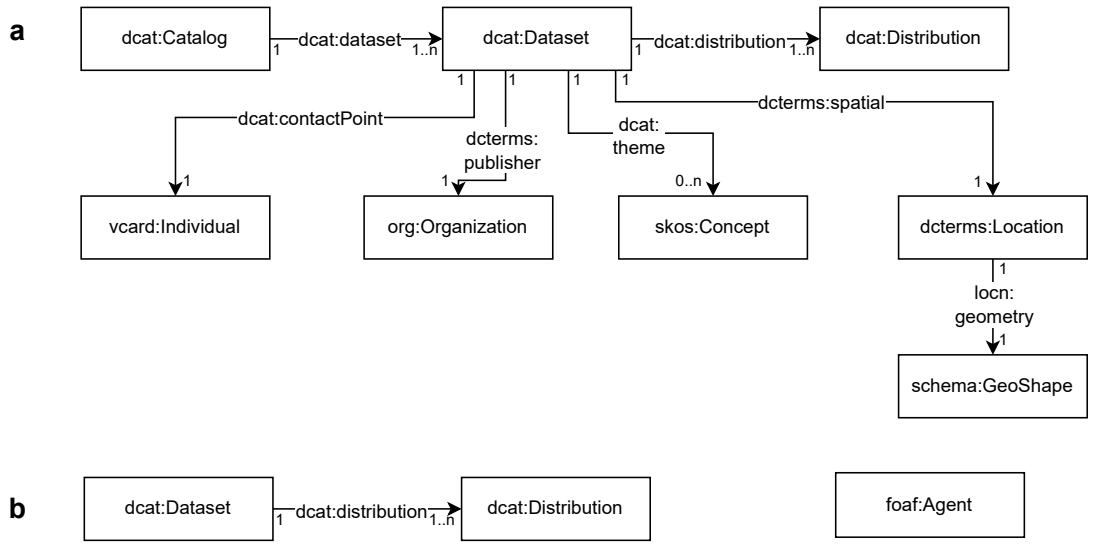

Figure 6.1: Models of depositar (a) and RDPCIDAT (b) DCAT implementations. Cardinalities were included as seen for the analyzed examples and might differ if larger graphs were looked at. Note: the foaf:Agent in panel b was not connected to the rest of the graph, but with the URI https://rdpcidat.rub.de/publisher/n0 supposedly should represent the publisher. A literal was provided for publisher as well.

main dcat:Dataset even if it only served partial, non-overlapping data in contrast to DCAT specification also for DCAT 2 (32: § 6.8). These two differences let the depositar model appear significantly more compact than the TUdatalib model.

Several other classes were used in the model to provide information in a semantic way. For metadata not available in TUdatalib in a similar fashion, geolocations were typed as dcterms:Location and self-created themes as skos:Concept, the second in depositar in addition to links to wikidata themes. The presentation of a corresponding author via the property dcat:contactPoint had been suggested for TUdatalib in section 5.2.1 and was realized here with the class vcard:Individual. Interestingly, creators were provided as one string with the dc:creator property from DC Elements (24) and not modeled as instances of foaf:Person connected via dcterms:creator as per DCAT 2 instructions (32: § 6.4.4).

The RDPCIDAT model was seen to be a minimal implementation. It consisted of the classes dcat:Dataset and dcat:Distribution and provided any additional metadata, except for access and download links, as plain literals. This was sometimes against the range definition of the respective property. As seen for depositar, multiple files with non-overlapping content were each considered a dcat:Distribution of the main dcat:Dataset in RDPCIDAT.

In a similar way as the implementation of the TUdatalib model would be expected to do, the depositar DCAT RDF provided almost all information that was also seen on the landing page. In contrast, for RDPCIDAT, even information on the dataset creators was missing. Other lacking metadata for RDPCIDAT included information for the scientific domain of plasma research that would not be covered by DCAT.

Not shown in Figure 6.1 are differences in the use of properties to provide metadata that was not modeled as classes. The depositar repository used certain properties, also from schema.org, that were not included in the TUdatalib model. On the other hand, several properties included in the TUdatalib model were not observed in the analyzed depositar RDF data, which might be partly due to the investigated datasets not using all available

metadata fields. For clarity, the analysis of differences was restricted to the core model involving the most important entities and properties and analysis of differences in single additional properties was out of scope for this work.

Analysing from a Linked Data perspective, it was observed that depositar provided dcat:Datasets with URIs that were distinct from the landing page URL and that provided upon request with an RDF browser identical data to the RDF files for download on the landing page. In contrast, the URI for the dcat:Catalog was https://data.depositar.io, which was identical to the catalog homepage provided via foaf:homepage (32: § 6.3.1). Links to external resources were provided in different contexts, for example to deliver information on language, dataset type, file type, and licenses.

At RDPCIDAT, as mentioned above, most of the metadata was provided as literals. As such, no links to external resources were seen in the analyzed RDF data. The URI for the dcat:Dataset was identical to the URL of the dataset landing page in the repository. When this URI was accessed with an RDF browser, different, non-compatible information was returned than in the RDF file for download on the landing page for an entity with an identical URI. In particular, the class foaf:Document was used for typing. This information was encoded via RDFa (159) inside the HTML page, albeit interpreted differently by different parsers including formally incorrect use of terms suggesting issues with the serialization. The RDFa triples inside the landing page included a link via rdfs:seeAlso to the DCAT RDF/XML file available for download on the landing page.

7 Technical Implementation

The TUdatalib DCAT model (see Figure 5.5) turned out to be highly complex with instances of different classes that have to be created during the conversion of single DSpace entities. These instances have to be connected to provide a semantic description of the entities and the overall repository contents. Therefore, a technical system to implement such a model must provide advanced functionality and ability for configuration.

The number one choice for a technical implementation of the model in TUdatalib would be the DSpace RDF module that adds Linked Data functionality to a repository if desired (70). This module was created as proof-of-concept realization of the general repository Linked Data concept developed by Pascal-Nicolas Becker (69). Briefly, the module consists of three parts that handle RDF triple generation, storage in a triple store, and triple exposure upon request.

The RDF module generates RDF triples from the repository content using plugins (69: section 4.2.3). However, not all metadata is available via those plugins and only limited capabilities for configuration are built in. For the plugin that handles the descriptive metadata stored in the metadata table, an advanced RDF-based configuration has been integrated, suggesting that these values should be available as needed (69: section 4.3.6). However, there are only very limited capabilities to configure the relations between the different levels of hierarchy handled by another plugin (69: section 4.3.5). For the TUdatalib model, this is especially an issue for bitstreams as instances of two classes are necessary in case of multi-bitstream items. There appears to be no way to encode the relations between item versions. The third and last plugin only serves to add constant triples to every entity (69: section 4.3.5).

Furthermore, bitstreams are handled as part of items without the possibility to create URIs for entities representing bitstreams (70, 160). In general, URI generation is only possible for the repository itself, communities, collections, and items (160). DSpace e-persons representing registered users would be possible as well, but are being warned against (70) and would only cover part of the authors for whom URI generation would also be necessary. Finally, URIs would have to be generated for entities representing OAI-PMH functionalities (see Table 5.6).

These points reveal that, in its current form, the DSpace RDF module lacks essential functionality to create RDF triples covering the complex DCAT 3 RDF model developed for TUdatalib. There might be additional limitations in the module's components to store and expose triples, but these were not evaluated seeing the insufficient triple generation functionality.

Thus, to use the module with the TUdatalib DCAT model, further development would be necessary. At least metadata conversion would have to be improved by creating either more sophisticated plugins whose configuration would allow a highly complex model as suggested for TUdatalib or bespoke plugins for that model.

Alternatively, external software solutions could be used that access the DSpace data but that are not part of the DSpace software system. These solutions would have to fulfill the same tasks that would be handled by the DSpace RDF module otherwise. At least, they would

need to convert existing repository data to RDF and expose the data upon user request. In a similar way to the DSpace RDF module workflow, the converted RDF triples could be stored for recall after conversion. Alternatively, conversion of repository data could be done on-the-fly upon user request without triple storage (9: section 8.5.4.1).

As seen in chapter 2.2, previous efforts on linked data exposure of repository metadata mostly relied on two data sources, OAI-PMH (49, 50, 58, 59) or direct access to the repository database (58, 60). In principle, these two options exist for TUdatalib as well.

Using OAI-PMH would solve the challenge identified by Becker (69: section 4.2.3) of triple conversion that is access management to not expose any non-public information. Essentially, nothing that would not be available via OAI-PMH could be included in the RDF data. The main issue is information completeness. OAI-PMH allows for querying with six different verbs, among them listing all sets inside a repository (58). OAI-PMH does not allow for retrieving significant descriptive metadata about these sets, just name and assigned content in form of record metadata. In case of DSpace, communities and collections are exposed as OAI-PMH sets using the respective entity title as name (161). Thus, if OAI-PMH were to be used as primary data source, at least some data would have to be added from another source.

The complete metadata would be available in the repository databases. DSpace stores most of its metadata in a relational database, but with exceptions for authority data where a SOLR cache is used (162). SOLR typically returns data as XML upon request (163).

These findings indicate that there is no single data source that will provide all necessary data. Thus, a method will be required that can merge data from different sources during triple generation. One possibility is use of the RDF Mapping Language (RML) (164, 165) that is an unofficial extension of the W3C recommendation RDB to RDF Mapping Language (R2RML) (166). Interpreters for RML have been published and are under continuous development (167, 168). Of course, in the context of using an external tool, mechanisms have to be in place to keep the RDF data in sync with the continuous changes of the repository data as also discussed by Becker (69: section 4.2.2).

Next, the generated RDF triples need to be either directly delivered to the requesting user or stored on the server for later exposure. An on-the-fly triple generation and delivery system surveying the model URIs without a caching mechanism would solve the issue of an updating mechanism as always the most up-to-date version of the repository content would be used, but at the cost of potentially higher response times to requests (9: section 8.5.4.1). Storage of pre-generated triples could be done in a dedicated triple store as for the DSpace RDF module (69: section 4.2) and a suitable tool could be used to load the correct information upon request to one of the URIs if that functionality were not provided by the triple store (10: section 5.2.5). Alternatively, text files could be used to store serialized RDF triples. Scripts would then have to be employed to direct from the URIs used to name entities in the model to the appropriate text file (10: section 5.2.3).

Ideally, users that only know the URL of the DSpace landing page should also be made aware of the RDF content describing the respective entity. This could be solved by referring to the RDF data in a similar way as for RDPCIDAT (see chapter 6) using RDFa with the property rdfs:seeAlso embedded in the landing page HTML, or using an HTML alternate link as suggested by Heath and Bizer (10: section 5.2.1.3).

8 Discussion

A model of TUdatalib in RDF based on the DCAT 3 vocabulary was developed in order to address the research questions. During model creation, different modeling layers had to be approached successively. These layers were mapping of classes and their relations, for each class the mapping of DSpace metadata fields with DCAT 3 properties, and assignment of URIs. The challenges posed by each layer turned out to be quite different.

For mapping classes (see section 5.1), one major challenge was posed by different views of DSpace and DCAT on how the content of datasets is delivered. DSpace, and consequently TUdatalib, follow an organization of items to which one or more potentially very different bitstreams are attached that in sum contain the whole dataset content. This turned out to be incompatible with the DCAT idea of having distributions that, allowing for exceptions such as limitations of file formats, cover the whole dataset content (14: § 6.8). These different views had been noted before, for example for CKAN repositories (73), and a strategy had been developed for DCAT 2 to address this challenge of loosely structured catalogs (32: § C.1). This strategy is inherited by DCAT 3 in a slightly modified form (14: § C.1). Alternatively, DCAT 3 also introduces the concept of a dataset series (14: § 12), which is another approach to splitting dataset information but which has higher requirements for structural uniformity between constituent parts than the loosely structured catalog.

Data analysis showed that only a subset of TUdatalib multi-bitstream items matched the requirements for a dataset series with one item (2416.3), out of six multi-bitstream items, appearing to me like a clear candidate and another one (2279) borderline. Thus, two options existed to model multi-bitstream items of TUdatalib in DCAT 3. The loosely structured catalog approach could be used for all multi-bitstream items or items representing dataset series could be identified and modeled as such and the remainder be modeled as loosely structured catalog. I decided for the first option for two reasons. Manual dataset series identification by experts would impair automated data conversion (discussed below) and addition of another DCAT 3 class would also mean the use of additional properties to describe relations making the whole model significantly more complicated. The decision comes at the cost of expressiveness as the existence of dataset series is not encoded in the model and thus this information cannot be delivered to a client. How bad this cost is cannot be estimated based on the low number of investigated items and because the items were selected to be diverse, not representative. A follow-up study should investigate on a representative set of items, with clear pre-formulated rules on the TUdatalib interpretation of dataset series requirements, how common dataset series are in TUdatalib. If they were highly common, a way would have to be found to integrate them into the model, potentially at the cost of automation.

Earlier attempts to describe DSpace repositories in RDF were impacted by missing vocabulary to transfer the concepts of communities and collections. Koutsomitropoulos et al. (50) defined their own classes while Latif et al. (60), in a questionable fashion, relied on the

SWRC ontology class swrc:Collection, intended for books (64), for collection and broke the range definition of rdf:type (65: section 3.3) to use properties from DC Elements instead of classes to type communities. In contrast, DCAT 3 provides the class dcat:Catalog (14: § 6.3) that suited very well to describe these entities within on the DSpace hierarchy. However, a fitting property to describe the relation between communities and their subcommunities as well as communities and their collections could not be identified. This was solved following a GitHub issue (134) I created to make the DCAT editors aware of a potential interpretation conflict with dcterms:hasPart coming from its specialization in the context of dcat:Catalog (14: § 6.3.3). While it was left unclear whether the DXWG agreed that there really was a conflict, or just ambiguity meaning this property's use would have been possible from the start (135–137), it was decided to introduce a new specialized property, dcat:resource, for listing any kind of dcat:Resource in a dcat:Catalog and thereby making dcterms:hasPart available for nesting catalogs or at least reducing ambiguity (135).

DCAT-AP had already allowed catalog nesting using dcterms:hasPart (39: section 4.1.3, 136) while this property's specialization in DCAT 2 either prevented this or introduced ambiguity for the main vocabulary. If the first was the the case, the use of this property was misaligned between the main vocabulary and a major application profile.

In summary, communities and collection were easily mapped to DCAT 3 after some changes to this RDF vocabulary while items and bitstreams required a complex model whose selection also comes with disadvantages.

The next level (see section 5.2) was finalizing a mapping between the metadata fields of the different DSpace entities and the properties of the DCAT 3 classes. This mapping was straightforward for many DSpace metadata fields. To a large degree, this was because DCAT was designed to reuse many properties from the DCTERMS vocabulary (26) that aligned well to the qualified DC Elements metadata fields used in TUdatalib. For lack of respective fields in TUdatlib, domain-specific metadata could not be included in this study.

In certain cases, especially concerning dates (see section 5.2.1), care had to be taken to choose the most appropriate value for the selection offered by the repository. Even this value was a pragmatic decision that cannot always be guaranteed to exactly fit the DCAT specifications until an issue with DSpace itself (147) is fixed in the future. Such decisions might have to be made on an individual basis for each repository depending on the specific instructions for submission. Another pragmatic decision was the handling of media types of bitstreams and compressed combination distributions of multi-bitstream items (see section 5.2.2). Noting such pragmatic decisions and re-evaluating them in other research data repositories might lead to arguments for vocabulary improvements, or at least to formulating guidelines, for example in form of an application profile, for making the decisions equally in similar repositories.

Very little information could not be mapped to DCAT 3, including dataset creation date and user-provided version identifiers (see section 5.2.1). In the latter case, it was decided that the well-defined version identifier provided for multi-version items by the repository fits conceptually better than the user-provided identifier that is a string without a pre-defined structure. Multiple use of properties would be okay in some cases, for example DCAT-AP allows for multiple use of dcterms:description (39: section 4.1.3). For version identifiers, in contrast, having different values with the same property would create ambiguity. However, identification of a well-fitting one from another vocabulary was not possible. A similar case

was the news field of collections where a broader study would be needed to confirm that the suggested outside property, skos:versionNotes (151: section 7.2), is a good fit.

On the other hand it was seen that DCAT 3 properties exist that remained unused in the model. Several of those might be of interest to scientists looking for data relevant to their studies or, as for dcat:contactPoint (14: § 6.4.3), for reaching out to colleagues. As such, it might be useful to add fields corresponding to so-far unused DCAT 3 properties to TUdatalib, and potentially other repositories, to improve dataset descriptions. Having a common vocabulary for guidance could also improve interoperability between repositories.

It was observed that the dcat:Datasets introduced to represent single bitstreams in the loosely structured catalog model had little metadata assigned. All of it was redundant as it was included in the description of the dcat:Dataset of the item or of the dcat:Distribution of the bitstream (see section 5.2.2). A property, dcat:componentDistribution, had been suggested before to directly relate a dcat:Distribution to a dcat:Dataset even if it were not a full representation (74). This would make the dcat:Datasets that represent bitstreams unnecessary, simplifying the structure of the TUdatalib DCAT 3 model. However, this low availability of information was also an argument against introducing the possibility of modeling files as partial distributions of datasets seeing that this might lead to discouraging people from making an effort to sufficiently describe data in files (169).

Looking at it from the opposite side, having dcat:Dataset instances for individual files gives the opportunity to add deeper descriptions using dcat:Dataset-specific properties of individual bitstreams for future items if the respective metadata were provided upon submission. This, of course, would mean that TUdatalib would have to handle this information and the submitting users were willing to put in the work to provide it.

DCAT 3 introduces two major new features, namely dataset series and dataset versioning (14). Dataset series have not been included in the model as discussed above. In contrast, the introduction of dataset versioning turned out to be of high importance and the new properties were used to reflect the ability of DSpace and TUdatalib to create different versions of items. Furthermore, checksums for files were also added in DCAT 3 and used in the TUdatalib model (14).

To summarize this level, almost all metadata of TUdatalib could be mapped to DCAT 3. For these mappings, case-specific, manual decisions could be avoided supporting automated metadata conversion. In contrast, unused DCAT 3 properties may guide extension of collected metadata in TUdatalib in the future. For the few missing mappings, terms from external vocabularies might be employed but those terms have not yet been identified in a satisfactory manner in all cases. New features of DCAT 3 were included as an important part of the model.

The third level (see section 5.4) was development of a strategy to assign dereferenceable URIs to entities. An important point here was to distinguish between the classes of the TUdatalib model, which are abstract and real-world entities, and documents providing information about these entities. The reason for this distinction for objects in repositories, like for other abstract and real-world entities (170: section 3.1), was argued in detail by Pascal-Nicolas Becker (69: section 3.1.1). Finding a system to create these turned out to be quite straightforward in an approach inspired by Latif et al. (60) as DSpace provided sufficient identifiers for the different entities.

The slash recipe (9: section 8.2.2.3) was used mainly because it creates independence from the documents of the repository system. This might, in the future, be necessary if the repository were to be ported to another software solution. It also allows for handling request without having to rely on DSpace functionalities as turned out to be of advantage when moving towards a technical implementation (see chapter 7 and discussion below). All suggested URIs would stem from a namespace under the control of ULB Darmstadt in line with the recommendation to use local URIs (9: section 8.3.1). Thus, the suggestion to use permanent identifiers, especially DOIs, formulated by Becker (69: section 5.3) was not followed due to several reasons. Having the URIs under control of another institution might mean limited technical features or changes to policies like DataCite has done for content negotiation in the past (171). Moreover, not all items in TUdatalib have DOIs. Finally, the model developed here relies on several entities with different URIs for the same DSpace class. One URI provided by a permanent identifier would not be enough for naming all entities in such a complex model.

The discussion so far has already shown that the developed model meets most of the requirements outlined in the introduction (chapter 1). Care has been taken to conform to vocabulary specification, even though pragmatic decisions were necessary for dates of issuance and for media types. Only the last item from the requirement list has not yet been looked at. This item concerns links to external entities. Such links were included if the information was provided in some way upon submission, for example by provision of ORCIDs, notations in classifications, or even manually added external relations. Also standard terms such as languages and media types were easily converted to external links. It should be possible to expand reference to outside entities in the future, maybe by mechanical entity identification such as proposed by Haslhofer and Schandl (47, 49), for example to central authority records like the *Gemeinsame Normdatei* (172) or to wikidata. Such possibilities will have to be evaluated separately in the future. For this model, it was considered satisfactory to create the external links already included in the metadata, either implicitly or explicitly. Thus, overall, the model requirements have been met sufficiently allowing for the gained insights to be discussed in the context of the research questions.

Expert interviews on the anticipated suitability of DCAT 2 for research data catalogs conducted by the FAIRsFAIR initiative with two domain-focused data providers and two data aggregators resulted in an optimistic assessment, stating that "there was universal agreement that the model is a very suitable one and would map well to current practice with no major obstacles foreseen. The high-level concepts such as Catalog, CatalogRecord, Resource and Dataset were regarded as natural and in some cases useful extensions to the models currently used. DCAT was regarded as a very rich generic approach" (Lambert et al. 2021, ref. 12: p. 15). Asked about the changes of DCAT 3, the same experts regarded "[t]he enhancements concerning versioning and dataset series envisaged in DCAT 3 [...] as valuable" (Lambert et al. 2021, ref. 12: p. 15). For institutional repositories, this study on TUdatalib let to the same general conclusion of suitability of DCAT 3 to describe institutional research data repositories in RDF. The vast majority of available metadata could be included in a DCAT 3-based model leading to a comprehensive description of repository structure and content. This model turned out to be highly complex, however, which might prove to be a hurdle for implementation as discussed below. This complexity might be different for other data providers including other institutional repositories based on different software solutions.

The analyzed datasets that model development was based on were selected to be diverse (see Chapter 4). Still, the focus of a university of technology was visible and results of such a case study might be different for a university with a focus on social sciences or humanities. It would also be important to conduct similar case studies on institutional research data repositories running different software solutions. Furthermore, the model will have to be re-evaluated against the final DCAT 3 recommendation once available.

Conceptually, the model for TUdatalib supports automated data conversion to RDF. This is due to the finding that the mappings of the classes and metadata fields in TUdatalib to the classes and properties of DCAT 3 exhibited very low ambiguity. However, at one point, this unambiguity, as well as the avoidance of an even more complex model, was bought with a loss of information. Dataset series would be modeled just like any other item and not identified as such even though DCAT 3 would have the classes and semantics to allow this (14: § 12). If identification of dataset series were found to be necessary, as still has to be evaluated for TUdatalib (see above), the data provider would have to deal with the even more complex model and find a way to flag dataset series, probably with an additional metadata field that would be occupied by manual curation or machine learning technologies.

The final prove that automatic data conversion would be possible can only be delivered by actually implementing the model in a live system as unforeseen challenges might occur. This implementation will be a formidable challenge because of the model's high structural complexity. A short technical evaluation of the Linked Data module provided by DSpace (see chapter 7) found that its current version offers insufficient functionality to incorporate the TUdatalib model. Importantly, this module does not support URIs for bitstreams in general (160). As such, the DCAT requirement to distinguish between datasets and distributions (14: § 4) is incompatible with the functionalities of this module and the impossibility of implementation with the module's current version is not a specific issue with the application of DCAT as it is presented here. A rough outline of alternatives to using the the current module was provided in chapter 7 as well. These include further development of the DSpace Linked Data module and using external tools with different data sources to bypass DSpace functionality. All these alternatives come with their own challenges. The evaluation of advantages and disadvantages of those approaches, or maybe development of another alternative, will have to be carried out in a future study. Seeing these challenges, it was not attempted to set up a running system.

Two research data repositories, depositar and RDPCIDAT, which already provide their metadata using DCAT were analyzed and their models compared to the one for TUdatalib. They were seen to be quite different, also from each other, in their use of classes and properties. The class dcat:Catalog was only used by depositar. RDPCIDAT also did not link to external resources.

Use of dcat:Catalog for depositar was restricted to a single instance. This probably reflects the software behind the system as I also mentioned in the GitHub issue (134). The discovery system-type organization with facets of the depositar repository seen when selecting datasets is in contrast to the hierarchical one for TUdatalib and not as suited for subcatalogs. The DCAT 3 specifications usage note for dcat:Catalog reads that "A Web-based data catalog is typically represented as a single instance of this class" (Albertoni et al. 2022, ref. 14: § 6.3). However, it can be argued that a dcat:Catalog that has multiple parts is still one catalog, but providing additional structural information (134).

A major difference for both repositories to TUdatalib was how informationally non-equivalent files, modeled as dcat:Distribution, were connected to the dcat:Dataset that represented the combination of all files. A direct dcat:distribution relation was used, in contrast to the much more complex model suggested for TUdatalib. Use of this direct relation with informationally non-equivalent files is known to be in use also for governmental catalogs (173–175) even though already considered to be in conflict with specifications for DCAT 1 by Neumaier et al. (73). A possible reason for this common pattern of modeling is that DCAT-AP suggested this approach to represent dataset series (176).

DCAT 3 allows for this way of modeling dataset series, as well as those to be modeled like loosely structured catalogs, for a transition period, stating that "These options are not formally incompatible with DCAT, so they can cohexist [sic] with dcat:DatasetSeries during the upgrade to DCAT 3" (Albertoni et al. 2022, ref. 14: § 12.3). This statement is controversial with regard to dataset series being modeled as dcat:Datasets with multiple informationally non-equivalent dcat:Distributions for being "in direct conflict with the definition of distributions" (Palmer 2021, ref. 177), which I agree with.

It would be up to those in charge of the repository to enforce that all datasets conform to the uniformity requirements of dcat:DatasetSeries if such a model were intended. Otherwise, it would be good to switch approaches and reflect the diversity of dataset types by implementing a model similar to the TUdatalib suggestion. As the main intention of the analysis here was to compare the current underlying models, a systematic study of conformance with dataset series requirements was not performed using only three datasets each that were selected arbitrarily. One for depositar and all for RDPCIDAT were, even allowing broad interpretation of dataset series requirements (14: § 6.7), in my opinion not dataset series with the possibility of those being exceptions in these repositories. Such studies into file diversity on several repositories along with the reflection of how the results require the respective models to be designed would be incredible valuable on the way towards a general approach of modeling research data repositories with DCAT.

There were additional research data repositories whose DCAT implementations were only briefly looked upon (see Appendix A.5). Those probably have their own strengths and weaknesses and a detailed analysis of their models in comparison to the one developed here, beyond the scope of this thesis, might give inspiration for improvements on the TUdatalib model, or even DCAT itself. Overall, the uptake of DCAT in research data repositories was low with only a minor subset of re3data.org-listed repositories asserting to have implemented DCAT. Not even all of those made metadata in this standard publicly available. This discrepancy and the low uptake was noted before by Kazumi Tomoyose who, in 2021, used the same approach of querying re3data.org to obtain a list of 19 governmental and research data repositories, overlapping with the ones investigated here, claiming to have implemented DCAT (78). When checking metadata availability in DCAT on the repositories themselves, confirmation was only possible for fourteen of the repositories (78). The low uptake of DCAT is also in agreement with the FAIRsFAIR interviews with B2FIND and OpenAIRE that conclude that there is low incentive to harvest DCAT due to limited provision of metadata in this vocabulary by repositories (12).

Ideally, there should be efforts towards standardization of how institutional research data repositories provide their metadata in DCAT-based RDF as the vocabulary turned out to be promising. A standardized approach would not only increase interoperability, but clear

guidelines could also lower the barriers for repositories to implement. This could be done in the form of a DCAT application profile similar to others that have been designed for various applications (14: § 16). The model provided here can help to work towards this, but many more use cases and studies of different repositories will be needed.

9 Conclusions and Outlook

In this thesis, I investigated the suitability of DCAT 3 to describe institutional research data repositories using the TUdatalib DSpace repository as study case. Overall, the vocabulary was found to contain fitting classes and properties to confer the essential metadata about the repository structure and contents for provision as Linked Data. However, the model I developed to transfer repository information to DCAT 3 turned out to be highly complex, mainly to be in conformance with the DCAT definition of distributions and their relation to the datasets they represent. In contrast, the modeling of the general repository structure required a change of DCAT 3 property definitions but then was conceptually simple.

Basing the model on DCAT 3 instead of DCAT 2 was beneficial as one of the new features, namely dataset versioning, turned out to be valuable to mirror the functionalities of DSpace to create different versions of the same items. In contrast, another new feature of DCAT 3 that are dataset series were not included in the model as their manual identification would interfere with automated data conversion. The loss of information due to that omission could not be estimated as the low number of and the selection procedure for the investigated TUdatlib items was unsuitable judge dataset series commonness in the repository. This should be investigated in a follow-up study and if shown to be high, a way would have to be found to include them in the model, potentially at the cost of automation.

The high complexity of the TUdatalib model also became obvious when comparing it to those used in two repositories that currently expose DCAT metadata. However, it could be shown that parts of those simpler models are only in agreement with DCAT definitions for legacy reasons if at all. As such, use of the TUdatalib model, potentially in an adapted fashion, might improve metadata delivery by those repositories. To confirm this, further investigations into the dataset series issues will also be necessary for those repositories.

Conceptually, the developed model should allow for the automated metadata conversion from the repository system to a DCAT Linked Data RDF representation. However, the complexity of the DCAT 3 class model and the even higher complexity of its application in TUdatalib entails significant effort for implementation. Among other, the functionality of the DSpace RDF module would not be sufficient and other solutions come with their own drawbacks. Thus, it was not tried to set up a running system during the course of this work and the ultimate prove of automated metadata exposure using the TUdatalib model remains to be given.

For maximum interoperability, repositories with a similar focus should provide their metadata in the same model. For institutional and other research data repositories, an application profile based on DCAT 3 would constitute a promising approach. While the model presented here is an important step, more repositories will have to be analyzed and the results combined to access such a standardized representation.

Bibliography

(1) Jean-Claude Burgelman, Corina Pascu, Katarzyna Szkuta, Rene von Schomberg, Athanasios Karalopoulos, Konstantinos Repanas, and Michel Schouppe. "Open Science, Open Data, and Open Scholarship: European Policies to Make Science Fit for the Twenty-First Century". In: *Frontiers in big data* 2 (2019), p. 43.

(2) Christine L. Borgman. "The conundrum of sharing research data". In: *Journal of the American Society for Information Science and Technology* 63.6 (2012), pp. 1059–1078.

(3) Mark D. Wilkinson, Michel Dumontier, I. Jsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3 (2016), p. 160018.

(4) Barend Mons, Cameron Neylon, Jan Velterop, Michel Dumontier, Luiz Olavo Bonino Da Silva Santos, and Mark D. Wilkinson. "Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud". In: *Information Services & Use* 37.1 (2017), pp. 49–56.

(5) Jane Greenberg, Hollie C. White, Sarah Carrier, and Ryan Scherle. "A Metadata Best Practice for a Scientific Data Repository". In: *Journal of Library Metadata* 9.3-4 (2009), pp. 194–212.

(6) Maxi Kindling, Heinz Pampel, Stephanie van de Sandt, Jessika Rücknagel, Paul Vierkant, Gabriele Kloska, Michael Witt, Peter Schirmbacher, Roland Bertelmann, and Frank Scholze. "The Landscape of Research Data Repositories in 2015: A re3data Analysis". In: *D-Lib Magazine* 23.3/4 (2017).

(7) Heinz Pampel, Paul Vierkant, Frank Scholze, Roland Bertelmann, Maxi Kindling, Jens Klump, Hans-Jürgen Goebelbecker, Jens Gundlach, Peter Schirmbacher, and Uwe Dierolf. "Making research data repositories visible: the re3data.org Registry". In: *PloS one* 8.11 (2013), e78080.

(8) Tim Berners-Lee, James Hendler, and Ora Lassila. "The Semantic Web". In: *Scientific American* 284.5 (2001), pp. 34–43. URL: https://www.jstor.org/stable/26059207?seq=3 (visited on 06/12/2022).

(9) Aidan Hogan. *The Web of Data*. Cham: Springer International Publishing, 2020.

(10) Tom Heath and Christian Bizer. "Linked Data: Evolving the Web into a Global Data Space". In: *Synthesis Lectures on the Semantic Web: Theory and Technology* 1.1 (2011), pp. 1–136.

(11) Eva Mendez, Tony Hernandez, Angus Whyte, and Joy Davidson. *D3.6 Proposal on integration of metadata catalogues to support cross-disciplinary FAIR uptake*. Ed. by Fostering FAIR Data Practices in Europe. 2020. URL: https://doi.org/10.5281/zenodo.5357560 (visited on 06/12/2022).

(12) Simon Lambert, Ricarda Braukmann, Eva Méndez, Marina Sánchez, and Joy Davidson. *D3.7 Report on integration of metadata catalogues*. Ed. by Fostering FAIR Data Practices in Europe. 2021. URL: https://doi.org/10.5281/zenodo.5744913 (visited on 06/12/2022).

(13) Riccardo Albertoni, David Browning, Simon Cox, Alejandra González-Beltrán, Andrea Perego, Peter Winstanley, and Makx Dekkers. *Data Catalog Vocabulary (DCAT) - Version 3: W3C Editor's Draft 03 June 2022*. 2022. URL: https://w3c.github.io/dxwg/dcat/ (visited on 06/12/2022).

(14) Riccardo Albertoni, David Browning, Simon Cox, Alejandra González-Beltrán, Andrea Perego, Peter Winstanley, and Makx Dekkers. *Data Catalog Vocabulary (DCAT) - Version 3: W3C Working Draft 11 January 2022*. 2022. URL: https://www.w3.org/TR/2022/WD-vocab-dcat-3-20220111/ (visited on 06/10/2022).

(15) Ioana Robu, Valentin Robu, and Benoit Thirion. "An introduction to the Semantic Web for health sciences librarians". In: *Journal of the Medical Library Association : JMLA* 94.2 (2006), pp. 198–205.

(16) Richard Cyganiak, David Wood, and Markus Lanthaler. *RDF 1.1 Concepts and Abstract Syntax: W3C Recommendation 25 February 2014*. 2014. URL: http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/ (visited on 03/26/2022).

(17) Thomas D. Steele. "What comes next: understanding BIBFRAME". In: *Library Hi Tech* 37.3 (2019), pp. 513–524.

(18) Brian McBride. "The Resource Description Framework (RDF) and its Vocabulary Description Language RDFS". In: *Handbook on Ontologies*. Ed. by Steffen Staab and Rudi Studer. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 51–65.

(19) Grigoris Antoniou and Frank van Harmelen. "Web Ontology Language: OWL". In: *Handbook on Ontologies*. Ed. by Steffen Staab and Rudi Studer. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 67–92.

(20) Javed Ahmad Khan and Suresh Kumar. "Deep analysis for development of RDF, RDFS and OWL ontologies with protege". In: *Proceedings of 3rd International Conference on Reliability, Infocom Technologies and Optimization*. IEEE, 2014, pp. 1–6.

(21) Linked Conservation Data. *What is a namespace?* URL: https://www.ligatus.org.uk/lcd/faq/211 (visited on 05/31/2022).

(22) M. Cristina Pattuelli, Alexandra Provo, and Hilary Thorsen. "Ontology Building for Linked Open Data: A Pragmatic Perspective". In: *Journal of Library Metadata* 15.3-4 (2015), pp. 265–294.

(23) Thomas Baker. "Libraries, languages of description, and linked data: a Dublin Core perspective". In: *Library Hi Tech* 30.1 (2012), pp. 116–133.

(24) DCMI Usage Board. *DCMI Metadata Terms*. 2020. URL: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/2020-01-20/ (visited on 06/12/2022).

(25) Dublin Core Metadata Initiative. *About DCMI*. URL: https://www.dublincore.org/about/ (visited on 03/25/2022).

(26) Fadi Maali, Richard Cyganiak, and Vassilios Peristeras. "Enabling Interoperability of Government Data Catalogues". In: *Electronic Government*. Ed. by Maria A. Wimmer, Jean-Loup Chappelet, Marijn Janssen, and Hans J. Scholl. Vol. 6228. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 339–350.

(27) Fadi Maali, John Erickson, and Phil Archer. *Data Catalog Vocabulary (DCAT): W3C Recommendation 16 January 2014*. 2014. URL: https://www.w3.org/TR/2014/REC-vocab-dcat-20140116/ (visited on 06/12/2022).

(28) David Beckett, Tim Berners-Lee, Berners-Lee, Prud'hommeaux, Eric, and Gavin Carothers. *Terse RDF Triple Language: W3C Recommendation 25 February 2014*. 2014. URL: https://www.w3.org/TR/2014/REC-turtle-20140225/ (visited on 04/06/2022).

(29) Shoichi Taniguchi. "Examining BIBFRAME 2.0 from the Viewpoint of RDA Metadata Schema". In: *Cataloging & Classification Quarterly* 55.6 (2017), pp. 387–412.

(30) Olivia M. A. Madison. "The IFLA Functional Requirements for Bibliographic Records". In: *Library Resources & Technical Services* 44.3 (2000), pp. 153–159.

(31) Jaroslav Pullmann, Rob Atkinson, Antoine Isaac, and Ixchel Faniel. *Dataset Exchange Use Cases and Requirements: W3C Working Group Note 17 January 2019*. 2019. URL: https://www.w3.org/TR/2019/NOTE-dcat-ucr-20190117/ (visited on 06/10/2022).

(32) Riccardo Albertoni, David Browning, Simon Cox, Alejandra González-Beltrán, Andrea Perego, Peter Winstanley, and Makx Dekkers. *Data Catalog Vocabulary (DCAT) - Version 2: W3C Recommendation 04 February 2020*. 2020. URL: https://www.w3.org/TR/2020/REC-vocab-dcat-2-20200204/ (visited on 06/10/2022).

(33) Riccardo Albertoni, David Browning, Simon Cox, Alejandra González-Beltrán, Andrea Perego, Peter Winstanley, and Makx Dekkers. *Data Catalog Vocabulary (DCAT) - Version 3: W3C Working Draft 10 May 2022*. 2022. URL: https://www.w3.org/TR/2022/WD-vocab-dcat-3-20220510/ (visited on 06/10/2022).

(34) Riccardo Albertoni, David Browning, Simon Cox, Alejandra González-Beltrán, Andrea Perego, Peter Winstanley, and Makx Dekkers. *Data Catalog Vocabulary (DCAT) - Version 3: W3C Working Draft 7 March 2023*. 2022. URL: https://www.w3.org/TR/2023/WD-vocab-dcat-3-20230307/ (visited on 05/22/2023).

(35) Paolo Ciccarese, Stian Soiland-Reyes, Khalid Belhajjame, Alasdair Jg Gray, Carole Goble, and Tim Clark. "PAV ontology: provenance, authoring and versioning". In: *Journal of biomedical semantics* 4.1 (2013), p. 37. (Visited on 06/01/2022).

(36) Rachel Heery and Manjula Patel. "Application Profiles: Mixing and Matching Metadata Schemas". In: *Ariadne* 25 (2000). URL: http://www.ariadne.ac.uk/issue/25/app-profiles/ (visited on 03/29/2022).

(37) Fabian Kirstein, Benjamin Dittwald, Simon Dutkowski, Yury Glikman, Sonja Schimmler, and Manfred Hauswirth. "Linked Data in the European Data Portal: A Comprehensive Platform for Applying DCAT-AP". In: *Electronic Government*. Ed. by Ida Lindgren, Marijn Janssen, Habin Lee, Andrea Polini, Manuel Pedro Rodríguez Bolívar, Hans Jochen Scholl, and Efthimios Tambouris. Vol. 11685. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 192–204.

(38) European Commission. *DCAT Application Profile for data portals in Europe: Final version*. Ed. by Makx Dekkers. 2013. URL: https://joinup.ec.europa.eu/sites/default/files/distribution/2014-05/DCAT-AP_Final_v1.01.pdf (visited on 03/30/2022).

(39) Bert van Nuffelen. *DCAT Application Profile for data portals in Europe Version 2.0.0*. 2019. URL: https://joinup.ec.europa.eu/sites/default/files/distribution/access_url/2019-12/12f0dc1d-50b6-43e4-90c2-0afe213ac2be/DCAT_AP_2.0.0.pdf (visited on 03/30/2022).

(40) David Shotton and Silvio Peroni. *FaBiO, the FRBR-aligned Bibliographic Ontology*. 2019. URL: http://purl.org/spar/fabio/2019-02-19 (visited on 03/30/2022).

(41) The Library of Congress. *BIBFRAME 2 List View*. URL: http://id.loc.gov/ontologies/bibframe-2-1-0/ (visited on 03/30/2022).

(42) Schema.org. *Documentation*. URL: https://schema.org/docs/documents.html (visited on 06/01/2022).

(43) Nuno Freire, Valentine Charles, and Antoine Isaac. "Evaluation of Schema.org for Aggregation of Cultural Heritage Metadata". In: *The Semantic Web*. Ed. by Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam. Vol. 10843. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 225–239.

(44) Schema.org. *DataCatalog: A Schema.org Type*. URL: https://schema.org/DataCatalog (visited on 03/30/2022).

(45) Cecilia Avila-Garzon. "Applications, methodologies, and technologies for linked open data: a systematic literature review". In: *International Journal on Semantic Web and Information Systems (IJSWIS)* 16.3 (2020), pp. 53–69.

(46) Dorothea Salo. "Retooling Libraries for the Data Challenge". In: *Ariadne* 64 (2010). URL: http://www.ariadne.ac.uk/issue/64/salo/ (visited on 03/18/2022).

(47) Bernhard Haslhofer and Bernhard Schandl. "Interweaving OAI-PMH data sources with the linked data cloud". In: *International Journal of Metadata, Semantics and Ontologies* 5.1 (2010), p. 17.

(48) Philip Hunter and Marieke Guy. "Metadata for harvesting: the Open Archives Initiative, and how to find things on the Web". In: *The Electronic Library* 22.2 (2004), pp. 168–174.

(49) Bernhard Haslhofer and Bernhard Schandl. "The OAI2LOD Server: Exposing OAI-PMH Metadata as Linked Data". In: *International Workshop on Linked Data on the Web (LDOW2008), co-located with WWW 2008*. Beijing, 2008. URL: http://eprints.cs.univie.ac.at/284/ (visited on 03/17/2022).

(50) Dimitrios A. Koutsomitropoulos, Georgia D. Solomou, and Theodore S. Papatheodorou. "Semantic interoperability of dublin core metadata in digital repositories". In: *2008 International Conference on Innovations in Information Technology*. IEEE, 2008, pp. 233–237.

(51) Stefano Mazzocchi. *OAI-PMH RDFizer*. URL: https://github.com/santteegt/oai2rdf (visited on 03/30/2022).

(52) Dan Brickley and Libby Miller. *FOAF Vocabulary Specification 0.99: Namespace Document 14 January 2014 - Paddington Edition*. 2014. URL: http://xmlns.com/foaf/spec/ (visited on 03/30/2022).

(53) Dimitrios J Koutsomitropoulos, Georgia D Solomou, Andreas D Alexopoulos, and Theodore S Papatheodorou. "Digital Repositories and the Semantic Web: Semantic Search and Navigation for DSpace". In: *4th International Conference on Open Repositories*. 2009. URL: http://hdl.handle.net/1853/28484 (visited on 03/18/2022).

(54) Dimitrios A Koutsomitropoulos, Georgia D Solomou, and Ricardo Borillo Domenech. "DSpace Semantic Search v2.0: What's New and Current Status". In: *Proceeding of the 7th International Conference on Open Repositories (OR 2012), 9–13 July, Edinburgh*. 2012.

(55) Dimitrios A. Koutsomitropoulos, Georgia D. Solomou, and Theodore S. Papatheodorou. "Semantic query answering in digital repositories: Semantic Search v2 for DSpace". In: *International Journal of Metadata, Semantics and Ontologies* 8.1 (2013), p. 46.

(56) David Tenenholtz. *A Complex Web: Upgrading to Linked Data in Digital Repositories*. 2017. URL: https://doi.org/10.17615/4s14-3f06 (visited on 03/31/2022).

(57) David Wilcox. "Stewarding Research Data with Fedora". In: *IFLA WLIC 2017*. 2017. URL: http://library.ifla.org/id/eprint/1796/ (visited on 03/31/2022).

(58) Santiago Gonzalez-Toral, Mauricio Espinoza-Mejia, and Victor Saquicela. "Digital Repositories and Linked Data: Lessons Learned and Challenges". In: *Knowledge Graphs and Semantic Web*. Ed. by Boris Villazón-Terrazas and Yusniel Hidalgo-Delgado. Vol. 1029. Communications in Computer and Information Science. Cham: Springer International Publishing, 2019, pp. 41–55.

(59) Nelson Piedra, Janneth Chicaiza, Jorge Lopez-Vargas, and Edmundo Tovar Caro. "Guidelines to producing structured interoperable data from Open Access Repositories". In: *2016 IEEE Frontiers in Education Conference (FIE)*. IEEE, 2016, pp. 1–9.

(60) Atif Latif, Timo Borst, and Klaus Tochtermann. "Exposing Data From an Open Access Repository for Economics As Linked Data". In: *D-Lib Magazine* 20.9/10 (2014).

(61) David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, et al., eds. *Progress in Artificial Intelligence*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005.

(62) URL: http://ontoware.org/swrc/swrc/SWRCOWL/swrc_updated_v0.7.1.owl (visited on 06/11/2022).

(63) Linked Open Vocabularies. *Semantic Web for Research Communities (swrc)*. URL: https://lov.linkeddata.es/dataset/lov/vocabs/swrc (visited on 06/11/2022).

(64) *Linked Open Vocabularies Search for collection in swrc*. URL: https://lov.linkeddata.es/dataset/lov/terms?q=collection&vocab=swrc (visited on 06/11/2022).

(65) Dan Brickley and R. V. Guha. *RDF Schema 1.1: W3C Recommendation 25 February 2014*. 2014. URL: http://www.w3.org/TR/2014/REC-rdf-schema-20140225/ (visited on 06/01/2022).

(66) Bruce D'Arcus and Frédérick Giasson. *Bibliographic Ontology (BIBO) in RDF*. URL: https://www.dublincore.org/specifications/bibo/bibo/ (visited on 04/06/2022).

(67) Ilie Cristian Dorobăț and Vlad Posea. "Evolving the DSpace Storage into Linked Data". In: *2020 12th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*. IEEE, 2020, pp. 1–5.

(68) Ilie Cristian Dorobăț and Vlad Posea. "Enriching the Cultural Heritage Metadata Using Historical Events: A Graph-Based Representation". In: *Digital Libraries for Open Knowledge*. Ed. by Antoine Doucet, Antoine Isaac, Koraljka Golub, Trond Aalberg, and Adam Jatowt. Vol. 11799. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 344–347.

(69) Pascal-Nicolas Becker. "Repositorien und das Semantic Web". Diploma Thesis. Berlin: Technische Universität Berlin, 2014. URL: https://doi.org/10.14279/depositonce-5015 (visited on 03/31/2022).

(70) Tim Donohue. *DSpace 7.x Documentation: Linked (Open) Data*. URL: https://wiki.lyrasis.org/pages/viewpage.action?pageId=104566644 (visited on 06/03/2022).

(71) Sebastian Neumaier, Jürgen Umbrich, and Axel Polleres. "Automated Quality Assessment of Metadata across Open Data Portals". In: *Journal of Data and Information Quality* 8.1 (2016), pp. 1–29.

(72) Open Knowledge Foundation. *ckanext-dcat README*. URL: https://github.com/ckan/ckanext-dcat/blob/56eb2459293754a57a47bb2b7c0273d2d7263ae0/README.md (visited on 06/03/2022).

(73) 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, 2016. URL: https://www.w3.org/2016/11/sdsvoc/SDSVoc16_paper_16 (visited on 04/01/2022).

(74) Simon Cox. *Distributions, services and implementation-resources #411 [Response 5)*. 2018. URL: https://github.com/w3c/dxwg/issues/411#issuecomment-432851965 (visited on 06/12/2022).

(75) David Browning. *Distributions, services and implementation-resources #411 [Response 35]*. 2019. URL: https://github.com/w3c/dxwg/issues/411#issuecomment-468195225 (visited on 06/12/2022).

(76) European Commission, Directorate-General for Research and Innovation, O. Corcho, M. Eriksson, K. Kurowski, Milan Ojsteršek, C. Choirat, M. Sanden, and F. Coppens. *EOSC interoperability framework : report from the EOSC Executive Board Working Groups FAIR and Architecture*. Publications Office, 2021.

(77) Milan Ojsteršek. *Crosswalk of most used metadata schemes and guidelines for metadata interoperability*. 2021. URL: https://doi.org/10.5281/zenodo.4420116 (visited on 06/12/2022).

(78) Kazumi Tomoyose. "O Data Catalog Vocabulary (DCAT) para a publicação de dados de pesquisa nos princípios Linked Data". Master thesis. 2021. URL: https://repositorio.ufscar.br/handle/ufscar/14116 (visited on 06/12/2022).

(79) MacKenzie Smith, Mary Barton, Margret Branschofsky, Greg McClellan, Julie Harford Walker, Mick Bass, Dave Stuve, and Robert Tansley. "DSpace". In: *D-Lib Magazine* 9.1 (2003).

(80) Tim Donohue. *DSpace Wiki: Releases*. URL: https://wiki.lyrasis.org/pages/viewpage.action?pageId=69010426 (visited on 04/04/2022).

(81) Ajinkya Prabhune, Rainer Stotzka, Vaibhav Sakharkar, Jürgen Hesser, and Michael Gertz. "MetaStore: an adaptive metadata management framework for heterogeneous metadata models". In: *Distributed and Parallel Databases* 36.1 (2018), pp. 153–194.

(82) Robina Clayphan and Rebecca Guenther. *DC-Libraries - Library Application Profile - Draft*. 2004. URL: https://www.dublincore.org/specifications/dublin-core/library-application-profile/ (visited on 04/04/2022).

(83) A. Bollini, C. Cortese, E. Groppo, and S. Mornati. "Extending DSpace to fulfil the requirements of digital libraries for cultural heritage management". In: *IRCDL 2017 Conference. Modena: IRCDL*. 2017. URL: https://www.academia.edu/download/51570645/IRCDL2017_rev.pdf (visited on 04/05/2022).

(84) Michel Castagné. *Institutional repository software comparison: DSpace, EPrints, Digital Commons, Islandora and Hydra*. 2013. URL: https://doi.org/10.14288/1.0075768 (visited on 06/12/2022).

(85) Sinjini Mukherjee and Rajesh Das. "Integration of Domain-Specific Metadata Schema for Cultural Heritage Resources to DSpace: A Prototype Design". In: *Journal of Library Metadata* 20.2-3 (2020), pp. 155–178.

(86) Mufazil Ali, Fayaz Ahmad Loan, and Rabiya Mushatq. "Open Access Scientific Digital Repositories : An Analytical Study of the Open DOAR". In: *2018 5th International Symposium on Emerging Trends and Technologies in Libraries and Information Services (ETTLIS)*. IEEE, 2018, pp. 213–216.

(87) Asma Bashir, Aasif Ahmad Mir, and Zahoor Ahmad Sofi. "Global Landscape of Open Access Repositories". In: *Library Philosophy and Practice* (2019), pp. 1–21. URL: https://www.proquest.com/scholarly-journals/global-landscape-open-access-repositories/docview/2234441884/se-2 (visited on 04/02/2022).

(88) re3data. *edoc-Server - Forschungsdaten: Repository details*. URL: https://www.re3data.org/repository/r3d100012889 (visited on 04/05/2022).

(89) re3data. *Refubium: Repository details*. URL: https://www.re3data.org/repository/r3d100013519 (visited on 04/05/2022).

(90) re3data. *DepositOnce: Repository details*. URL: https://www.re3data.org/repository/r3d100011091 (visited on 04/05/2022).

(91) re3data. *data_ UMR: Repository details*. URL: https://www.re3data.org/repository/r3d100013377 (visited on 04/05/2022).

(92) re3data. *DaKS: Repository details*. URL: https://www.re3data.org/repository/r3d100013608 (visited on 04/05/2022).

(93) re3data. *TUdatalib: Repository details*. URL: https://www.re3data.org/repository/r3d100013029 (visited on 04/05/2022).

(94) TUdata. *About TUdatalib Repository*. URL: https://tudatalib.ulb.tu-darmstadt.de/docs/en/ (visited on 04/04/2022).

(95) TU Darmstadt. *Our mission*. URL: https://www.tu-darmstadt.de/universitaet/index.en.jsp (visited on 06/12/2022).

(96) TUdata. *Guide for TUdatalib administrators*. URL: https://tudatalib.ulb.tu-darmstadt.de/docs/en/leitfaden/ (visited on 06/04/2022).

(97) DuraSpace. *DuraSpace Registry: TUdatalib*. URL: https://duraspace.org/registry/entry/7514/ (visited on 04/04/2022).

(98) TUdata. *TUdatalib*. URL: https://tudatalib.ulb.tu-darmstadt.de/ (visited on 06/06/2022).

(99) Robert E. Stake. *The art of case study research*. 4th print. Thousand Oaks: Sage, 1995.

(100) Maria Auxilio Medina, J. Alfredo Sanchez, Ofelia Cervantes, Antonio Benitez, and Jorge de La Calleja. "LOD4AIR: A strategy to produce and consume linked open data from OAI-PMH repositories". In: *2017 International Conference on Electronics, Communications and Computers (CONIELECOMP)*. IEEE, 2017, pp. 1–8.

(101) Christian Bizer, Tom Heath, and Tim Berners-Lee. "Linked Data - The Story So Far". In: *International Journal on Semantic Web and Information Systems* 5.3 (2009), pp. 1–22.

(102) *re3data.org - Registry of Research Data Repositories*. URL: https://doi.org/10.17616/R3D (visited on 06/11/2022).

(103) TUdata. *Vereinbarung zur Archivierung bzw. Veröffentlichung von Forschungsdaten mit TUdatalib (Nutzungsvereinbarung)*. 2020. URL: https://tudatalib.ulb.tu-darmstadt.de/docs/nutzungsvereinbarung/ (visited on 05/03/2022).

(104) Tim Donohue. *DSpace 7.x Documentation: Functional Overview*. URL: https://wiki.lyrasis.org/pages/viewpage.action?pageId=104566616 (visited on 06/01/2022).

(105) TUdata. *TUdatalib: Browsing by Type*. URL: https://tudatalib.ulb.tu-darmstadt.de/browse?type=type (visited on 04/07/2022).

(106) Jan-Karl Knigge. *Virtual Reality in Manual Order Picking - Software*. 2020. URL: https://doi.org/10.25534/tudatalib-317 (visited on 06/10/2022).

(107) Luise Borek. *Pferdetypen*. 2017. URL: https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2955 (visited on 06/10/2022).

(108) Tomislav Maric. *Iterative Volume-of-Fluid interface positioning in general polyhedrons with Consecutive Cubic Spline interpolation: Singularity container*. 2020. URL: https://doi.org/10.25534/tudatalib-379 (visited on 06/10/2022).

(109) Marvin Bernhardt, Martin Hanke, and Nico van der Vegt. *Data of the publication: Iterative integral equation methods for structural coarse-graining*. 2021. URL: https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2662 (visited on 06/10/2022).

(110) Manuela Benary, Stefan Bohn, Mareen Lüthen, Ilias Nolis, Nils Blüthgen, and Alexander Loewer. *Single cell data*. 2020. URL: https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2279.

(111) Lakshya Pandit, Gladys Vásquez Fauggier, Lanqing Gu, and Martin Knöll. *Time-lapse videos of pre-and post traffic calming scenarios on Mainkai*. 2020. URL: https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2416.3.

(112) Jens Jungblut, Julia Haas, and Stephan Rinderknecht. *Supplementary data: Active vibration control of an elastic rotor by using its deformation as controlled variable*. 2020. URL: https://doi.org/10.48328/tudatalib-615 (visited on 06/10/2022).

(113) Christoph Reich, Tim Prangemeier, Heinz Koeppl, and Christian Wildner. *Multi-StyleGAN weights*. 2021. URL: https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2879 (visited on 06/10/2022).

(114) Darina Benikova, Margot Mieskes, Christian M. Meyer, and Iryna Gurevych. *DBS Corpus*. 2017. URL: https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/1915.2 (visited on 06/10/2022).

(115) depositar. *Terms of Use*. 2021. (Visited on 06/09/2022).

(116) Tzu-Hao Lin, Chong Chen, Hiromi Kayama Watanabe, Tomonari Akamatsu, and Shinsuke Kawagucci. *Deep-sea soundscapes of Japan*. 2021. URL: https://data.depositar.io/en/dataset/deep-sea-soundscapes-of-japan (visited on 06/08/2022).

(117) Andrea Wei-Ching Huang. *Reused Datasets from Our World in Data Statistics*. 2020. URL: https://data.depositar.io/en/dataset/our-world-in-data-statistics (visited on 06/08/2022).

(118) Yu-De Pei, Nathan William Price, Joseph Heard, Chieh-Hsuan Lee, Hsuan Tsang, and Colin Kuo-Chang Wen. *Data for Terpios paper*. 2021. URL: https://data.depositar.io/en/dataset/data-for-terpios-paper (visited on 06/08/2022).

(119) J. Golda, F. Kogelheide, P. Awakowicz, and V. Schulz-von der Gathen. *Dissipated electrical power and electron density in an RF atmsopheric pressure helium plasma jet*. 2021. URL: https://rdpcidat.rub.de/dataset/dissipated-electrical-power-and-electron-density-rf-atmsopheric-pressure-helium-plasma-jet (visited on 05/31/2022).

(120) K. Grosse, M. Falke, and A. von Keudell. *Ignition and propagation of nanosecond pulsed plasmas in distilled water - negative vs. positive polarity applied to a pin electrode*. 2021. URL: https://rdpcidat.rub.de/dataset/ignition-and-propagation-nanosecond-pulsed-plasmas-distilled-water-negative-vs-positive (visited on 05/31/2022).

(121) J. Held, S. Tiemann-Monje, A. von Keudell, and Vvon Schulz-von der Gathen. *Velocity distribution of metal ions in the target region of HiPIMS: the role of Coulomb collisions*. 2020. URL: https://rdpcidat.rub.de/dataset/velocity-distribution-metal-ions-target-region-hipims-role-coulomb-collisions (visited on 05/31/2022).

(122) depositar. *User guide*. URL: https://docs.depositar.io/en/6.5.2/user-guide.html (visited on 06/09/2022).

(123) Pierre-Yves Vandenbussche, Ghislain A. Atemezing, María Poveda-Villalón, and Bernard Vatant. "Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web". In: *Semantic Web* 8.3 (2016), pp. 437–452.

(124) Makx Dekkers. *Best practice for a loosely-structured catalog #253 [Response 18)*. 2018. URL: https://github.com/w3c/dxwg/issues/253#issuecomment-405219461 (visited on 06/11/2022).

(125) Simon Cox. *Best practice for a loosely-structured catalog #253 [Response 14]*. 2018. URL: https://github.com/w3c/dxwg/issues/253#issuecomment-404710185 (visited on 06/11/2022).

(126) Simon Cox. *Best practice for a loosely-structured catalog #253*. 2018. URL: https://github.com/w3c/dxwg/issues/253#issuecomment-404709391 (visited on 06/11/2022).

(127) Riccardo Albertoni. *Model Series of Data as Distributions of a single Dataset #1429 [Response 4]*. 2021. URL: https://github.com/w3c/dxwg/issues/1429#issuecomment-980316895 (visited on 06/10/2022).

(128) Andrea Perego. *Review comment in DXWG GitHub Pull Request #1458: PR to refine the message about dataset series' member (issue 1429)*. 2022. URL: https://github.com/w3c/dxwg/pull/1458#discussion_r815159348 (visited on 04/09/2022).

(129) Simon Cox. *Distribution composed of more than one file, but not packaged #482*. 2018. URL: https://github.com/w3c/dxwg/issues/482#issue-372413103 (visited on 06/09/2022).

(130) Makx Dekkers. *Distribution composed of more than one file, but not packaged #482 [Response 6]*. 2018. URL: https://github.com/w3c/dxwg/issues/482#issuecomment-436755100 (visited on 06/09/2022).

(131) *DFG Project AMOS – 435227428*. URL: https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2840 (visited on 04/10/2022).

(132) Simon Cox. *A separate class for DatasetSeries (?) #1272 [Response 6]*. 2021. URL: https://github.com/w3c/dxwg/issues/1272#issuecomment-784014802 (visited on 06/12/2022).

(133) Simon Cox. *Issue with dct:hasPart as described in DCAT #1205 [Response 2]*. 2020. URL: https://github.com/w3c/dxwg/issues/1205#issuecomment-652662240 (visited on 06/12/2022).

(134) Andreas Geißner. *dcterms:hasPart in the context of nested Catalogs #1454*. 2022. URL: https://github.com/w3c/dxwg/issues/1454#issue-1142593956 (visited on 06/09/2022).

(135) Andrea Perego. *Replace dcterms:hasPart with a specific property #1469:* 2022. URL: https://github.com/w3c/dxwg/issues/1469#issue-1153406743 (visited on 06/09/2022).

(136) Andrea Perego. *dcterms:hasPart in the context of nested Catalogs #1454 [Response 5]*. 2022. URL: https://github.com/w3c/dxwg/issues/1454#issuecomment-1054653629 (visited on 06/09/2022).

(137) Andrea Perego. *dcterms:hasPart in the context of nested Catalogs #1454 [Response 3]*. 2022. URL: https://github.com/w3c/dxwg/issues/1454#issuecomment-1051046026 (visited on 06/09/2022).

(138) Steve Baskauf. *RDF for talking about people*. 2016. URL: https://baskauf.blogspot.com/2016/02/rdf-for-talking-about-people.html?m=0 (visited on 04/13/2022).

(139) *DSpace Wiki: Authority Control of Metadata Values*. URL: https://wiki.lyrasis.org/pages/viewpage.action?pageId=19006062 (visited on 05/03/2022).

(140) André Castro, Deutsche Forschungsgemeinschaft, and Susanne Arndt. *DFG Classification of Subject Areas Ontology*. 2021. URL: https://github.com/tibonto/DFG-Fachsystematik-Ontology/blob/main/dfgfo.ttl (visited on 04/12/2022).

(141) Joan S. Mitchell and Michael Panzer. "Dewey linked data: Making connections with old friends and new acquaintances". In: *JLIS.it* 4.1 (2013), p. 177. URL: https://www.proquest.com/scholarly-journals/dewey-linked-data-making-connections-with-old/docview/1270767450/se-2.

(142) Alice Sneary. *Change to Dewey Web Services*. 2015. URL: https://web.archive.org/web/20180523011138/https://www.oclc.org/developer/news/2015/dewey-down.en.html (visited on 04/12/2022).

(143) DataCite Metadata Working Group. *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.4*. 2021. URL: https://doi.org/10.14454/3w3z-sa82 (visited on 06/12/2022).

(144) Simon Cox. *Dataset type [RDST] #64 [Response 13]*. 2018. URL: https://github.com/w3c/dxwg/issues/64#issuecomment-406128677 (visited on 06/10/2022).

(145) DataCite. *DataCite Content Negotiation*. URL: https://support.datacite.org/docs/datacite-content-resolver (visited on 06/03/2022).

(146) Tim Donohue. *Question about dc.date.accessioned and dc.date. available [Response 1]*. 2016. URL: https://groups.google.com/g/dspace-tech/c/k7qCSJVhRrU/m/_jmmpGwkOAAJ (visited on 06/10/2022).

(147) Ben Heartland. *Question about dc.date.accessioned and dc.date. available [Response 7]*. 2021. URL: https://groups.google.com/g/dspace-tech/c/k7qCSJVhRrU/m/oHHbxTg7BgAJ (visited on 06/10/2022).

(148) Timothy Lebo, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. *PROV-O: The PROV Ontology: W3C Recommendation 30 April 2013*. 2013. URL: http://www.w3.org/TR/2013/REC-prov-o-20130430/ (visited on 04/13/2022).

(149) Phil Archer, Gofran Shukair, and Makx Dekkers. *Asset Description Metadata Schema (ADMS): W3C Working Group Note 01 August 2013*. 2013. URL: http://www.w3.org/TR/2013/NOTE-vocab-adms-20130801/ (visited on 04/14/2022).

(150) Linux Foundation. *The Software Package Data Exchange® (SPDX®) Specification Version 2.2.2*. URL: https://spdx.github.io/spdx-spec/ (visited on 06/01/2022).

(151) Alistair Miles and Sean Bechhofer. *SKOS Simple Knowledge Organization System Reference: W3C Recommendation 18 August 2009*. 2009. URL: http://www.w3.org/TR/2009/REC-skos-reference-20090818/ (visited on 06/10/2022).

(152) Renato Iannella and James McKinney. *vCard Ontology - for describing People and Organizations: W3C Interest Group Note 22 May 2014*. 2014. URL: http://www.w3.org/TR/2014/NOTE-vcard-rdf-20140522/ (visited on 06/12/2022).

(153) Schema.org. *logo: A Schema.org Property*. URL: https://schema.org/logo (visited on 05/27/2022).

(154) Andrea Perego. *DataService and DataDistributionService #432 [Response 6]*. 2019. URL: https://github.com/w3c/dxwg/issues/432#issuecomment-453757666 (visited on 06/10/2022).

(155) Herbert van de Sompel, Michael L. Nelson, Carl Lagoze, and Simeon Warner. "Resource Harvesting within the OAI-PMH Framework". In: *D-Lib Magazine* 10.12 (2004).

(156) Tim Donohue and Mark H. Wood. *DSpace Wiki: OAI 2.0 Server*. URL: https://wiki.lyrasis.org/pages/viewpage.action?pageId=104566650 (visited on 04/20/2022).

(157) Fairdata.fi. *FAQ*. URL: https://www.fairdata.fi/en/faq/#5 (visited on 06/09/2022).

(158) depositar. *What is depositar?* URL: https://data.depositar.io/en/about (visited on 06/08/2022).

(159) Ben Adida, Mark Birbeck, Shane McCarron, and Ivan Herman. *RDFa Core 1.1 - Third Edition: Syntax and processing rules for embedding RDF through attributes. W3C Recommendation 17 March 2015.* 2015. URL: http://www.w3.org/TR/2015/REC-rdfa-core-20150317/ (visited on 04/26/2022).

(160) DSpace. *URIGenerator.java*. URL: https://github.com/DSpace/DSpace/blob/365827ce92e6eb6763d2a59525dc201fb753789b/dspace-api/src/main/java/org/dspace/rdf/storage/URIGenerator.java (visited on 04/25/2022).

(161) Tim Donohue. *DSpace 7.x Documentation: OAI-PMH Data Provider 2.0 (Internals)*. URL: https://wiki.lyrasis.org/pages/viewpage.action?pageId=104566648 (visited on 04/27/2022).

(162) Tim Donohue. *DSpace 7.x Documentation: ORCID Integration*. URL: https://wiki.lyrasis.org/pages/viewpage.action?pageId=104566745 (visited on 04/27/2022).

(163) *Solr Ref Guide 6.6: Response Writers*. URL: https://solr.apache.org/guide/6_6/response-writers.html (visited on 04/28/2022).

(164) Ben de Meester, Pieter Heyvaert, and Thomas Delva. *RDF Mapping Language (RML): Unofficial Draft 06 October 2020*. URL: https://rml.io/specs/rml/ (visited on 04/28/2022).

(165) Anastasia Dimou, Miel Vander Sande, Pieter Colpaert, Ruben Verborgh, Erik Mannens, and Rik van de Walle. "RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data". In: *Proceedings of the 7th Workshop on Linked Data on the Web*. Ed. by Christian Bizer, Tom Heath, Sören Auer, and Tim Berners-Lee. Vol. 1184. CEUR Workshop Proceedings. 2014. URL: http://ceur-ws.org/Vol-1184/ldow2014_paper_01.pdf.

(166) Souripriya Das, Seema Sundara, and Richard Cyganiak. *R2RML: RDB to RDF Mapping Language: W3C Recommendation 27 September 2012*. 2012. URL: http://www.w3.org/TR/2012/REC-r2rml-20120927/ (visited on 04/28/2022).

(167) Pieter Heyvaert, David Chaves-Fraga, Freddy Priyatna, Oscar Corcho, Erik Mannens, Ruben Verborgh, and Anastasia Dimou. "Conformance Test Cases for the RDF Mapping Language (RML)". In: *Knowledge Graphs and Semantic Web*. Ed. by Boris Villazón-Terrazas and Yusniel Hidalgo-Delgado. Vol. 1029. Communications in Computer and Information Science. Cham: Springer International Publishing, 2019, pp. 162–173.

(168) Pieter Heyvaert, Anastasia Dimou, and David Chaves-Fraga. *RML Implementation Report: Unofficial Draft 17 February 2022*. URL: https://rml.io/implementation-report/ (visited on 04/28/2022).

(169) Jakub Klímek. *Best practice for a loosely-structured catalog #253 [Response 12]*. 2018. URL: https://github.com/w3c/dxwg/issues/253#issuecomment-404440434 (visited on 06/11/2022).

(170) Leo Sauermann, Richard Cyganiak, Danny Ayers, and Max Völkel. *Cool URIs for the Semantic Web: 3C Interest Group Note 03 December 2008*. 2008. URL: http://www.w3.org/TR/2008/NOTE-cooluris-20081203/ (visited on 05/06/2022).

(171) Martin Fenner. *Upcoming Changes to DOI Content Negotiation*. 2019. URL: https://doi.org/10.5438/nz0m-rb06 (visited on 05/09/2022).

(172) Renate Behrens-Neumann and Barbara Pfeifer. "Die Gemeinsame Normdatei - ein Kooperationsprojekt". In: *Dialog mit Bibliotheken* (2011), pp. 37–40.

(173) Sabine Maennel. *Model Series of Data as Distributions of a single Dataset #1429*. 2021. URL: https://github.com/w3c/dxwg/issues/1429#issue-1059496244 (visited on 06/10/2022).

(174) Makx Dekkers. *Model Series of Data as Distributions of a single Dataset #1429 [Response 1]*. 2021. URL: https://github.com/w3c/dxwg/issues/1429#issuecomment-975294525 (visited on 06/10/2022).

(175) Anna Odgaard Ingram. *Need for a common approach to modeling dataset series in DCAT-AP #155 [Response 6]*. 2020. URL: https://github.com/SEMICeu/DCAT-AP/issues/155#issuecomment-701261376 (visited on 06/10/2022).

(176) *DCAT-AP: How to model Dataset series?* URL: https://joinup.ec.europa.eu/release/dcat-ap-how-model-dataset-series (visited on 04/27/2022).

(177) Matthias Palmer. *Model Series of Data as Distributions of a single Dataset #1429 [Response 13]*. 2021. URL: https://github.com/w3c/dxwg/issues/1429#issuecomment-987387421 (visited on 06/10/2022).

(178) Paolo Ciccarese, Silvio Peroni, and David Shotton. *Essential FRBR in OWL2 DL*. 2018. URL: http://purl.org/spar/frbr/2018-03-29 (visited on 06/01/2022).

(179) Andrea Perego and Michael Lutz. *ISA Programme Location Core Vocabulary: Second version in w3.org/ns space - 2015-03-23*. 2015. URL: https://www.w3.org/ns/locn (visited on 06/10/2022).

(180) Dave Reynolds. *The Organization Ontology: W3C Recommendation 16 January 2014*. 2014. URL: https://www.w3.org/TR/vocab-org/ (visited on 06/10/2022).

(181) Michael Schneider, Jeremy Carroll, Ivan Herman, and Peter F. Patel-Schneider. *OWL 2 Web Ontology Language RDF-Based Semantics (Second Edition): W3C Recommendation 11 December 2012*. 2012. URL: http://www.w3.org/TR/2012/REC-owl2-rdf-based-semantics-20121211/ (visited on 06/01/2022).

(182) Paolo Ciccarese, Stian Soiland-Reyes, Marco Ocana, Khalid Belhajjame, Alasdair J. G. Gray, and Simon Jupp. *PAV - Provenance, Authoring and Versioning*. 2014.

A Appendix

The Appendix contains the following sections:

A.1 Data publication

Data from TUdatalib encompassing all different modeling steps has been published on TUdatalib at the following DOI: https://doi.org/10.48328/tudatalib-1154

A.2 Vocabularies and namespaces

Table A.1: Used or mentioned RDF vocabularies with their prefixes, namespaces, and references to the documentation pages. Table assembled using information from Albertoni et al. (14: §3.1 and §3.2), Hogan (9: section 8.4.3), Linked Open Vocabularies at https://lov.linkeddata.es/dataset/lov/vocabs/ and the individual references for each vocabulary. If more than one common prefix exists for a vocabulary, one was selected.

Prefix	Name Namespace	Reference
adms	Asset Description Metadata Schema http://www.w3.org/ns/adms#	(149)
bf	Bibliographic Framework (BIBFRAME) http://id.loc.gov/ontologies/bibframe/	(41)
bibo	Bibliographic Ontology http://purl.org/ontology/bibo/	(66)
dc	Dublin Core Metadata Element Set http://purl.org/dc/elements/1.1/	(24)
dcat	Data Catalog Vocabulary http://www.w3.org/ns/dcat#	(14)
dcterms	DCMI Metadata Terms, /terms/ namespace http://purl.org/dc/terms/	(24)
dctype	DCMI Type Vocabulary http://purl.org/dc/dcmitype/	(24)
fabio	FRBR-aligned Bibliographic Ontology http://purl.org/spar/fabio	(40)
foaf	Friend of a Friend vocabulary http://xmlns.com/foaf/0.1/	(52)
frbr	Functional Requirements for Bibliographic Records http://purl.org/vocab/frbr/core#	(178)
locn	ISA Programme Location Core Vocabulary http://www.w3.org/ns/locn#	(179)
org	The Organization Ontology http://www.w3.org/ns/org#	(180)
owl	Web Ontology Language http://www.w3.org/2002/07/owl	(181)
pav	Provenance, Authoring and Versioning http://purl.org/pav/	(182)
prov	The PROV Ontology http://www.w3.org/ns/prov#	(148)
rdf	Resource Description Framework http://www.w3.org/1999/02/22-rdf-syntax-ns#	(16)
rdfs	RDF Schema http://www.w3.org/2000/01/rdf-schema#	(65)

Table A.1 continued

schema	Schema.org https://schema.org/	(42)
skos	Simple Knowledge Organization System http://www.w3.org/2004/02/skos/core	(151)
spdx	Software Package Data Exchange http://spdx.org/rdf/terms#	(150)
swrc	Semantic Web for Research Communities http://swrc.ontoware.org/ontology#	N/A (see (63))
vcard	vCard Ontology http://www.w3.org/2006/vcard/ns#	(152)

A.3 Tables of property use tabulated from Turtle documents

A.3.1 Item

Table A.2: Table of dcat:Dataset properties tabulated from item information in item Turtle documents

Item	1915.2	2279	2416.3
dcat:landingPage	dc.identifier.uri	dc.identifier.uri	dc.identifier.uri
dcterms:relation	dc.relation (+subproperty) tud.tubiblio		
dcat:distribution	[Link: distribution]	[Link: distribution]	[Link: distribution]
dcat:inSeries			
dcterms:accessRights			
dcterms:conformsTo			
dcat:contactPoint			
dcterms:creator	dc.contributor.author	dc.contributor.author	dc.contributor.author
dcterms:description	dc.description	dc.description	dc.description
dcterms:title	dc.title	dc.title	dc.title
dcterms:issued	dc.date.accessioned	dc.date.accessioned	dc.date.accessioned
	dc.date.available	dc.date.available	dc.date.available
	dc.date.issued	dc.date.issued	dc.date.issued
dcterms:modified	{xoai: lastModifyDate}	{xoai: lastModifyDate}	{xoai: lastModifyDate}
dcterms:language	dc.language.iso		dc.language.iso
dcterms:publisher			
dcterms:identifier	dc.identifier.uri	dc.identifier.uri	dc.identifier.uri
dcat:theme	dc.subject.ddc	dc.subject.ddc	dc.subject.ddc
		dc.subject.classification	dc.subject.classification
dcterms:type	dc.type	dc.type	dc.type
dcat:qualifiedRelation	dc.relation (+subproperty) tud.tubiblio		
prov:qualifiedAttribution			
dcat:keyword	dc.subject		dc.subject
dcterms:license	dc.rights.uri	dc.rights.uri	dc.rights.uri
dcterms:rights			
odrl:hasPolicy			
dcterms:isReferencedBy			
dcat:previousVersion	{LP Version history: Item}		{LP Version History: Item 2}
dcat:hasVersion			
dcat:hasCurrentVersion			
dcterms:replaces	{LP Version history: Item}		{LP Version History: Item all}
dcat:version	{LP Version history: Version} dc.description.version		{LP Version History: Version}
adms:versionNotes	{LP Version history: Summary}		
adms:status			
dcat:first			
dcat:last			
dcat:prev			
dcat:accrualPeriodicity			
dcterms:spatial			
dcat:spatialResolutionInMeters			
dcterms:temporal	dc.date.accessioned	dc.date.accessioned	dc.date.accessioned
	dc.date.available	dc.date.available	dc.date.available
	dc.date.issued	dc.date.issued	dc.date.issued
dcat:temporalResolution			
prov:wasGeneratedBy			
dcterms:hasPart	[Link: distribution dataset]	[Link: distribution dataset]	[Link: distribution dataset]

Table A.2, continued

Item	2480	2537	2662
dcat:landingPage	dc.identifier.uri	dc.identifier.uri	dc.identifier.uri
dcterms:relation			dc.relation (+subproperty)
dcat:distribution	[Link: distribution]	[Link: distribution]	[Link: distribution]
dcat:inSeries			
dcterms:accessRights			
dcterms:conformsTo			
dcat:contactPoint			
dcterms:creator	dc.contributor.author	dc.contributor.author	dc.contributor.author
dcterms:description	dc.description	dc.description	dc.description
dcterms:title	dc.title	dc.title	dc.title
dcterms:issued	dc.date.accessioned	dc.date.accessioned	dc.date.accessioned
	dc.date.available	dc.date.available	dc.date.available
	dc.date.issued	dc.date.issued	dc.date.issued
dcterms:modified	{xoai: lastModifyDate}	{xoai: lastModifyDate}	{xoai: lastModifyDate}
dcterms:language		dc.language.iso	
dcterms:publisher			tud.unit
dcterms:identifier	dc.identifier.uri	dc.identifier.uri	dc.identifier.uri
dcat:theme	dc.subject.ddc	dc.subject.ddc	dc.subject.ddc
	dc.subject.classification	dc.subject.classification	dc.subject.classification
dcterms:type	dc.type	dc.type	dc.type
dcat:qualifiedRelation			dc.relation (+subproperty)
prov:qualifiedAttribution			
dcat:keyword		dc.subject	dc.subject
dcterms:license	dc.rights.uri	dc.rights.uri	dc.rights.uri
dcterms:rights			
odrl:hasPolicy			
dcterms:isReferencedBy			
dcat:previousVersion			
dcat:hasVersion			
dcat:hasCurrentVersion			
dcterms:replaces			
dcat:version			
		dc.description.version	
adms:versionNotes			
adms:status			
dcat:first			
dcat:last			
dcat:prev			
dcat:accrualPeriodicity			
dcterms:spatial			
dcat:spatialResolutionInMeters			
dcterms:temporal	dc.date.accessioned	dc.date.accessioned	dc.date.accessioned
	dc.date.available	dc.date.available	dc.date.available
	dc.date.issued	dc.date.issued	dc.date.issued
dcat:temporalResolution			
prov:wasGeneratedBy		tud.project	tud.project
			tud.unit
dcterms:hasPart	[Link: distribution dataset]		[Link: distribution dataset]

Table A.2, continued

Item	2879	2904	2955
dcat:landingPage	dc.identifier.uri	dc.identifier.uri	dc.identifier.uri
dcterms:relation	dc.relation (+subproperty)		
dcat:distribution	[Link: distribution]	[Link: distribution]	[Link: distribution]
dcat:inSeries			
dcterms:accessRights			
dcterms:conformsTo			
dcat:contactPoint			
dcterms:creator	dc.contributor.author	dc.contributor.author	dc.contributor.author
dcterms:description	dc.description	dc.description	dc.description
dcterms:title	dc.title	dc.title	dc.title
dcterms:issued	dc.date.accessioned	dc.date.accessioned	dc.date.accessioned
	dc.date.available	dc.date.available	dc.date.available
	dc.date.issued	dc.date.issued	dc.date.issued
dcterms:modified	{xoai: lastModifyDate}	{xoai: lastModifyDate}	{xoai: lastModifyDate}
dcterms:language		dc.language.iso	
dcterms:publisher	tud.unit	tud.unit	tud.unit
dcterms:identifier	dc.identifier.uri	dc.identifier.uri	dc.identifier.uri
dcat:theme	dc.subject.ddc	dc.subject.ddc	dc.subject.ddc
	dc.subject.classification	dc.subject.classification	dc.subject.classification
dcterms:type	dc.type	dc.type	dc.type
dcat:qualifiedRelation	dc.relation (+subproperty)		
prov:qualifiedAttribution			
dcat:keyword			dc.subject
dcterms:license	dc.rights.uri	dc.rights.uri	dc.rights.uri
dcterms:rights			
odrl:hasPolicy			
dcterms:isReferencedBy			
dcat:previousVersion			
dcat:hasVersion			
dcat:hasCurrentVersion			
dcterms:replaces			
dcat:version			
adms:versionNotes			
adms:status			
dcat:first			
dcat:last			
dcat:prev			
dcat:accrualPeriodicity			
dcterms:spatial			
dcat:spatialResolutionInMeters			
dcterms:temporal	dc.date.accessioned	dc.date.accessioned	dc.date.accessioned
	dc.date.available	dc.date.available	dc.date.available
	dc.date.issued	dc.date.issued	dc.date.issued
			dc.subject
dcat:temporalResolution			
prov:wasGeneratedBy		tud.project	
	tud.unit	tud.unit	tud.unit
dcterms:hasPart		[Link: distribution dataset]	

A.3.2 Bitstream

Table A.3: Table of dcat:Distribution properties tabulated from bitstream information in item Turtle documents

Bitstreams of Item	1915.2	2279, 2662	2416.3
dcterms:accessRights	BITSTREAM Access		
dcterms:conformsTo			
dcterms:description	BITSTREAM Description		BITSTREAM Description
dcterms:title	BITSTREAM NAME	BITSTREAM NAME	BITSTREAM NAME
dcterms:issued	dc.date.available	dc.date.available	dc.date.available
dcterms:modified	xoai LastModifyDate	xoai LastModifyDate	xoai LastModifyDate
dcterms:license	dc.rights.uri	dc.rights.uri	dc.rights.uri
dcterms:rights			
odrl:hasPolicy			
dcat:spatialResolutionInMeters			
dcat:temporalResolution			
dcat:accessURL	dc.identifier.uri	dc.identifier.uri	dc.identifier.uri
dcat:accessService			
dcat:downloadURL	BITSTREAM URL	BITSTREAM URL	BITSTREAM URL
		ZIP Package URL	ZIP Package URL
dcat:byteSize	xoai BITSTREAM SIZE	xoai BITSTREAM SIZE	xoai BITSTREAM SIZE
dcterms:format			
dcat:mediaType	xoai BITSTREAM Format	xoai BITSTREAM Format	xoai BITSTREAM Format
		ZIP Package Format	ZIP Package Format
dcat:compressFormat	BITSTREAM Format	BITSTREAM Format	
		ZIP Package Format	ZIP Package Format
dcat:packageFormat		ZIP Package Format	ZIP Package Format
spdx:checksum	xoai checksum	xoai checksum	xoai checksum
	checksum algorithm	checksum algorithm	checksum algorithm

Bitstreams of Item	2480	2537; 2879, 2955	2904
dcterms:accessRights			
dcterms:conformsTo			
dcterms:description	BITSTREAM Description		BITSTREAM DESCRIPTION
dcterms:title	BITSTREAM NAME	BITSTREAM NAME	BITSTREAM NAME
dcterms:issued	dc.date.available	dc.date.available	dc.date.available
dcterms:modified	xoai LastModifyDate	xoai LastModifyDate	xoai LastModifyDate
dcterms:license	dc.rights.uri	dc.rights.uri	dc.rights.uri
dcterms:rights			
odrl:hasPolicy			
dcat:spatialResolutionInMeters			
dcat:temporalResolution			
dcat:accessURL	dc.identifier.uri	dc.identifier.uri	dc.identifier.uri
dcat:accessService			
dcat:downloadURL	BITSTREAM URL	BITSTREAM URL	BITSTREAM URL
	ZIP Package URL		
dcat:byteSize	xoai BITSTREAM SIZE	xoai BITSTREAM SIZE	xoai BITSTREAM SIZE
dcterms:format			
dcat:mediaType	xoai BITSTREAM Format	xoai BITSTREAM Format	xoai BITSTREAM Format
	ZIP Package Format		
dcat:compressFormat	BITSTREAM Format		BITSTREAM Format
	ZIP Package Format		
dcat:packageFormat	ZIP Package Format		
spdx:checksum	xoai checksum	xoai checksum	xoai checksum
	checksum algorithm	checksum algorithm	checksum algorithm

A.3.3 Community and Collection

Table A.4: Table of dcat:Catalog properties tabulated from community Turtle documents

Community	1074, 1119, 1256, 1332, 1403, 1436, 1568, 1585, 1606, 1630, 1638, 1959	1079
foaf:homepage	LP URL	LP URL
dcat:landingPage	LP URL	LP URL
dcat:themeTaxonomy	LP DFG classification	LP DFG classification
dcterms:relation		
dcterms:hasPart	LP Subcommunities	LP Subcommunities LP Collections
dcat:dataset	Title List	Title List
dcat:service		
dcat:catalog		
dcat:distribution		
dcat:inSeries		
dcat:record		
dcterms:accessRights		
dcterms:conformsTo		
dcat:contactPoint		
dcterms:creator		
dcterms:description		
dcterms:title	LP Title	LP Title
dcterms:issued		
dcterms:modified		
dcterms:language		
dcterms:publisher		
dcterms:identifier	LP URL External Information: Identifier in higher level community	LP URL External Information: Identifier in higher level community
dcat:theme		
dcterms:type		
dcat:qualifiedRelation		
prov:qualifiedAttribution		
dcat:keyword		
dcterms:license		
dcterms:rights		
odrl:hasPolicy		
dcterms:isReferencedBy		
dcat:previousVersion		
dcat:hasVersion		
dcat:hasCurrentVersion		
dcterms:replaces		
dcat:version		
adms:versionNotes		
adms:status		
dcat:first		
dcat:last		
dcat:prev		
dcat:accrualPeriodicity		
dcterms:spatial		
dcat:spatialResolutionInMeters		
dcterms:temporal		
dcat:temporalResolution		
prov:wasGeneratedBy		

Table A.4, continued

Community	1132, 1359, 1414, 1448, 1586, 1636, 1977	1409, 2200, 2544	2212
foaf:homepage	LP URL	LP URL	LP URL
dcat:landingPage	LP URL	LP URL	LP URL
dcat:themeTaxonomy	LP DFG classification	LP DFG classification	LP DFG classification
dcterms:relation			
dcterms:hasPart		LP Subcommunities	
	LP Collections		LP Collections
dcat:dataset	Title List	Title List	Title List
dcat:service			
dcat:catalog			
dcat:distribution			
dcat:inSeries			
dcat:record			
dcterms:accessRights			
dcterms:conformsTo			
dcat:contactPoint			
dcterms:creator			
dcterms:description			LP Description
dcterms:title	LP Title	LP Title	LP Title
dcterms:issued			
dcterms:modified			
dcterms:language			
dcterms:publisher			
dcterms:identifier	LP URL External Information: Identifier in higher level comm.	LP URL	LP URL
dcat:theme			
dcterms:type			
dcat:qualifiedRelation			
prov:qualifiedAttribution			
dcat:keyword			
dcterms:license			
dcterms:rights			
odrl:hasPolicy			
dcterms:isReferencedBy			
dcat:previousVersion			
dcat:hasVersion			
dcat:hasCurrentVersion			
dcterms:replaces			
dcat:version			
adms:versionNotes			
adms:status			
dcat:first			
dcat:last			
dcat:prev			
dcat:accrualPeriodicity			
dcterms:spatial			
dcat:spatialResolutionInMeters			
dcterms:temporal			
dcat:temporalResolution			
prov:wasGeneratedBy			

Table A.5: Table of dcat:Catalog properties tabulated from collection Turtle documents

Collection	1914, 2476	2840
foaf:homepage	LP URL	LP URL
dcat:landingPage	LP URL	LP URL
dcat:themeTaxonomy	LP DFG classification	LP DFG classification
dcterms:relation		
dcterms:hasPart		
dcat:dataset	Title list	Title list
dcat:service		
dcat:catalog		
dcat:distribution		
dcat:inSeries		
dcat:record		
dcterms:accessRights		
dcterms:conformsTo		
dcat:contactPoint		
dcterms:creator		
dcterms:description	LP Description	LP Description
	External Information:	External Information:
	Description In item/community	Description In item/community
		LP News
dcterms:title	LP Title	LP Title
dcterms:issued		
dcterms:modified		
dcterms:language		
dcterms:publisher		
dcterms:identifier	LP URL	LP URL
dcat:theme		
dcterms:type		
dcat:qualifiedRelation		
prov:qualifiedAttribution		
dcat:keyword		
dcterms:license		
dcterms:rights		
odrl:hasPolicy		
dcterms:isReferencedBy		
dcat:previousVersion		
dcat:hasVersion		
dcat:hasCurrentVersion		
dcterms:replaces		
dcat:version		
adms:versionNotes		LP News
adms:status		
dcat:first		
dcat:last		
dcat:prev		
dcat:accrualPeriodicity		
dcterms:spatial		
dcat:spatialResolutionInMeters		
dcterms:temporal		
dcat:temporalResolution		
prov:wasGeneratedBy		

A.4 Property tables for the different classes in the final model

Table A.6: Table of dcat:Catalog property use for communities and collections

Property	Metadata field	Comment
foaf:homepage	LP URL	
dcat:landingPage	LP URL	
dcat:themeTaxonomy	LP DFG classification	External Link
dcterms:hasPart	Link: subcommunity catalog	
	Link: collection catalog	
dcat:dataset	Link: dataset	
dcterms:description	LP Description	
	External Information: Description in item/community	
dcterms:title	LP Title	
dcterms:identifier	LP URL	
	External Information: Identifier in higher level community	
dcat:distribution	Link: Distribution for OAI-PMH XML	
(skos:historyNote)	LP News	
dcterms:publisher	Link: publisher	
vcard:hasLogo	LP Logo URL	

Table A.7: Table of dcat:Distribution property use for dcat:Catalog OAI-PMH distributions

Property	Comment
dcterms:accessRights	Constant external link to "public" in controlled vocabulary
dcterms:description	String assembled to provide information on handle, metadata format, etc.
dcterms:title	String assembled to provide information on handle, metadata format, etc.
dcterms:license	Constant external link to CC0
dcat:downloadURL	Link: XML serialization URL
dcat:mediaType	Constant external link to MIME type text/xml
dcat:accessService	Link: dcat:Service for OAI-PMH

Table A.8: Table of dcat:DataService property use for OAI-PMH services

Property	Comment
dcterms:accessRights	Constant external link to "public" in controlled vocabulary
dcterms:description	String assembled to provide information on context
dcterms:title	String assembled to provide information on context
dcterms:license	Constant external link to CC0 License
dcat:servesDataset	Links to all public dcat:Catalogs
dcat:endpointURL	URL of the OAI-PMH service
dcterms:publisher	Link: publisher

Table A.9: Table of dcat:Dataset property use for items

Property	Metadata field	Comment
dcat:landingPage	dc.identifier.uri	Local handle URL only
dcterms:relation	dc.relation (+subproperty)	External link
	tud.tubiblio	
dcat:distribution	Link: distribution	
dcterms:creator	Link: creator	
dcterms:description	dc.description	
dcterms:title	dc.title	
dcterms:issued	dc.date.available	
dcterms:modified	xoai: lastModifyDate	
dcterms:language	dc.language.iso	External link
dcterms:publisher	Link: publisher	
dcterms:identifier	dc.identifier.uri	External link for DOI
dcat:theme	dc.subject.ddc	Potential external link, minimal local instance if not available and legally possible
	dc.subject.classification	External link
dcterms:type	dc.type	Potential external link, minimal local instance if not available
dcat:qualifiedRelation	dc.relation (+subproperty)	External link
dcat:keyword	dc.subject	
dcterms:license	dc.rights.uri	External link
dcat:previousVersion	LP Version history: Item	
dcterms:replaces	LP Version history: Item	
dcat:version	LP Version history: Version	
adms:versionNotes	LP Version history: Summary	
prov:wasGeneratedBy	Link: project	
dcterms:hasPart	Link: distribution dataset	
dcterms:accessRights	Information: public/restricted	External link
dcterms:created	dc.date.issued	
(owl:versionInfo)	dc.description.version	Recommended in ADMS vocabulary, but not ideal choice

Table A.10: Table of dcat:Dataset property use for bitstream part-datasets

Property	Metadata field	Comment
dcat:landingPage	dc.identifier.uri	Local handle URL only
dcterms:issued	dc.date.available	
dcterms:license	dc.rights.uri	External link
dcat:theme	dc.subject-classification	External link
	dc.subject.ddc	Potential external link, minimal local instance if not available and legally possible
prov:wasGeneratedBy	Link: project	
dcterms:accessRights	BITSTREAM Access	External link
dcterms:title	BITSTREAM Name	
dcterms:description	BITSTREAM Description	
dcat:distribution	Link: Distribution	
dcterms:publisher	Link: publisher	

Table A.11: Table of dcat:Distribution property use

Property	Metadata field	Comment
dcterms:accessRights	BITSTREAM Access	External link
dcterms:description	BITSTREAM Description	
dcterms:title	BITSTREAM NAME	
dcterms:issued	dc.date.available	
dcterms:license	dc.rights.uri	External link
dcat:accessURL	dc.identifier.uri	Local handle URL only
dcat:downloadURL	BITSTREAM URL	For distributions representing one bitstream
	ZIP Package URL	For combination distributions with zip download link
dcat:byteSize	xoai BITSTREAM SIZE	
dcat:mediaType	xoai BITSTREAM Format	External link
dcat:compressFormat	ZIP Package Format	External link
dcat:packageFormat	ZIP Package Format	External link
spdx:checksum	Link: checksum	

Table A.12: Table of foaf:Person property use for creators

Property	Metadata field	Comment
foaf:name	dc.contributor.author	
owl:sameAs	LP creator ORCID	External link

Table A.13: Table of foaf:Organization property use for publishers

Property	Metadata field	Comment
foaf:name	Institution name (based on tud.unit)	
owl.sameAs	Link to external ressources describing same institution (e.g. wikidata, GND)	External link
foaf:homepage	Institution homepage	
others	information depending on institution self-description	

Table A.14: Table of spdx:Checksum property use

Property	Metadata field	Comment
spdx:algorithm	xoai checksum	
spdx:checksumValue	xoai checksum algorithm	External link

Table A.15: Table of prov:Activity property use for projects

Property	Metadata field
prov:label	tud.project

A.5 Overview of research data repositories using DCAT according to re3data.org

The following repositories were evaluated before deciding to analyze depositar and RDPCIDAT for comparison with the model developed here.

RDPCIDAT

URL: https://rdpcidat.rub.de/

No catalog found, but data in Turtle format containing descriptions of datasets could be downloaded. Accessing URIs with the Q&D RDF browser resulted in exposure of RDF data, but invalid DCAT (no dcat:Datasets, and XML literals as object for dcat:distribution). Links to the XML documents that contained the valid descriptions were included in the returned RDF data. When checking the issue with the XML literal as dcat:distribution object by using Ruby RDF Distiller, W3C RDFa 1.1 Distiller and Parser, and RDFa Play to convert the data to Turtle, different outcomes, sometimes incorrect use of terms (such as use of dcat:Distribution as property) and occasionally errors were seen suggesting some non-conformance with standards.

See main text references Golda et al. (119), Grosse et al. (120), and Held et al. (121) for investigated datasets.
Overall: positive (in agreement with Tomoyose (78: section 6.17))

TERN Data Discovery Portal

URL: https://portal.tern.org.au/

Only a subset of data listed in the portal provided RDF files for download. These files only included correct RDF serialization for the class dcat:Catalog. No correct information in RDF on the respective dataset was included.

Investigated datasets (all licensed CC-BY 4.0):
Department of Environment and Science, Queensland Government. *Airborne Hyperspectral and LiDAR data - Australian field sites*. 2013. URL: https://geonetwork.tern.org.au/geonetwork/srv/eng/catalog.search#/metadata/4ff0b4c9-cfa0-4d09-9520-b5402adc583f (accessed May 31, 2022)

Department of Environment and Science, Queensland Government. *Seasonal ground cover statistics - Landsat, JRSRP algorithm, Queensland coverage*. 2015. https://geonetwork.tern.org.au/geonetwork/srv/eng/catalog.search#/metadata/86c19d64-7cfe-4557-a874-479d024ac1b5 (accessed May 31, 2022)

Terrestrial Ecosystem Research Network. *AEKOS Poaceae Extraction 2014*. 2014. URL: https://geonetwork.tern.org.au/geonetwork/srv/ger/catalog.search#/metadata/87d6406b-6dbf-48a9-81d6-b42eadbc9ddf (accessed May 31, 2022)

Department of Environment and Science, Queensland Government. *Seasonal fractional cover - Landsat, JRSRP algorithm, Australia coverage*. 2013. https://geonetwork.tern.org.au/geonetwork/srv/eng/catalog.search#/metadata/f0c32576-9ad7-4c9c-9aa9-22787867e28b (accessed May 31, 2022)

Trying to access the listed URLs with the Q&D RDF browser also did not return RDF with DCAT vocabulary, only very general RDF information.
The SPARQL endpoint listed on r3data.org (https://graphdb.tern.org.au/sparql) was queried on April 22, 2022 for instances of dcat:Dataset with the following query, but only returned two datasets for the test-long-format repository.

```
PREFIX dcat: <http://www.w3.org/ns/dcat#>
  select ?dataset where {
    ?dataset a dcat:Dataset .
  } limit 100
```

Because of the use of dcat:Catalog, a positive result was given to this repository, even though it was a borderline case.

depositar

URL: https://data.depositar.io/en/

Datasets from depositar were described in valid RDF using the DCAT classes dcat:Dataset and dcat:Distribution. A valid repository description was also available using dcat:Catalog with the instructions located in the user guide (122). The RDF was designed to be compatible with DCAT 2, but still considered in beta status.

See main text references Lin et al. (116), Huang (117), Pei et al. (118) for investigated datasets.
Additionally, Turtle files describing the catalog were downloaded from https://data.depositar.io/catalog.ttl?page=3 and https://data.depositar.io/catalog.ttl?page=9 as per the user guide instructions on April 22, 2022.
Accessing URIs with the Q&D RDF browser returned valid RDF with DCAT descriptions and same URIs as downloaded RDF files.

Overall: positive

Biological and Chemical Oceanography Data Management Office

URL: https://www.bco-dmo.org/

RDF, in general, was a mix of different vocabularies with classes and properties of DCAT used alongside the PROV ontology and schema.org. The class dcat:Distribution only appeared to be used for data available in textual formats, not for other files for download.

Investigated datasets (all licensed CC-BY 4.0):
Joshua Kohut, George Cutter and Christian Reiss. *Dataset: SWARM AMLR moorings - acoustic data.* 2022. URL: https://www.bco-dmo.org/dataset/872729 (accessed May 31, 2022)

Christopher House and Leah Brandt. *Dataset: 16S rRNA gene from DNA.* 2019. URL: https://www.bco-dmo.org/dataset/780926 (accessed May 31, 2022)

Andrew R. Babbin and Jarek Kwiecinski (2021) *Dataset: Depth-gridded and Density-gridded ODZs.* 2021. https://www.bco-dmo.org/dataset/865316 (accessed May 31, 2022)

The service returned valid RDF when querying URIs with the Q&D RDF browser.
The use of the class dcat:Catalog was confirmed using the following SPARQL query on the endpoint at https://lod.bco-dmo.org/sparql on May 31, 2022.

```
PREFIX dcat: <http://www.w3.org/ns/dcat#>
select ?catalog where {
    ?catalog a dcat:Catalog .
} limit 100
```

Overall: positive (in agreement with Tomoyose (78: section 6.3))

Norwegian Polar Data Centre

URL: https://data.npolar.no/home/

RDF files were delivered for single datasets using the classes dcat:Catalog, dcat:Dataset, and dcat:Distribution.

Investigated datasets (both licensed CC-BY 4.0):
Mikko Vihtakariemail, Jorg Welcker, Børge Moe, Olivier Chastel, Sabrina Tartu, Haakon Hop, Claus Bech, Sébastien Descamps and Geir Wing Gabrielsen. *Black-legged kittiwake diet data from Kongsfjorden 1982-2016.* 2017. URL: https://data.npolar.no/dataset/26dbd004-158b-4909-a67d-4d3b12223842 (accessed June 09, 2022).

Anja Rösel, Dmitry Divine, Jennifer A. King, Marcel Nicolaus, Gunnar Spreen, Polona Itkin, et al. *N-ICE2015 total (snow and ice) thickness data from EM31 [1.0].* 2016. URL: https://data.npolar.no/dataset/70352512-fed8-4f1d-8b9c-30e6a764f5c2 (accessed June 09, 2022).

Content was delivered in JSON-LD format that cannot be processed by the Q&D RDF browser. Encoding errors were seen when, instead, trying to retrieve the data from the RDF file dcat:Dataset URIs with EasyRDF converter on June 10, 2022.

Overall: positive (in agreement with Tomoyose (78: section 6.13))

Health Data Research Innovation Gateway

URL: https://www.healthdatagateway.org/

According to the specification document[26], DCAT terms are reused in this catalog for the internal metadata model. There did not appear to be RDF documents for download on the dataset landing pages.

Investigated datasets (restrictive license according to Terms and Conditions[27]):
NHS Digital. *GPES Data for Pandemic Planning and Research (COVID-19).* 2021. URL: https://web.www.healthdatagateway.org/dataset/696cfc9f-090d-4328-94ac-140760a77c73 (accessed June 09, 2022)

NHS Digital. *Improving Access to Psychological Therapies Data Set.* URL: https://web.www.healthdatagateway.org/dataset/bcf6e5ce-986d-4b84-9c9c-69de966e8bbd (accessed June 09, 2022)

The RDF returned when accessing the links with the Q&D RDF browser only provided very general descriptions without use of the DCAT vocabulary.

With missing public documents containing DCAT classes or properties, this repository was considered negative for providing DCAT-based metadata.

IOS Regensburg Research Data Respository

URL: https://lambda.ios-regensburg.de/

No catalog found. No RDF documents seemed to be available for download for the datasets.

Investigated datasets (licensed CC0 as are metdadata in this repositorium[28])

[26] A. Milward and D. Milward. "Descriptive Metadata Specification V2. 12.8.2020". 2020. URL: https://github.com/HDRUK/schemata/blob/master/docs/dataset/2.0.0/distribution/Descriptive%20Metadata%20Specification%20v2.0.0%2012.8.2020%20.pdf(accessed June 09, 2022)

[27] *Health Data Research UK Innovation Gateway Terms and Conditions.* URL: https://www.healthdatagateway.org/about/terms-and-conditions (accessed June 09, 2022)

[28] *Research Data Policy.* URL: https://lambda.ios-regensburg.de/research-data-policy (accessed June 09, 2022)

Holm Sundhaussen. *Historische Statistik Serbiens: Berufsstatus der städtischen und ländlichen Erwerbstätigen nach Geschlechtern 1895 und 1900.* URL: https://lambda.ios-regensburg.de/dataset/soa_87-44 (accessed June 09, 2022)

Holm Sundhaussen. *Historische Statistik Serbiens: Bewegung der Eheschließungen, Geburten und Sterbefälle im europäischen Vergleich 1841/50 - 1901/10.* URL:https://lambda.ios-regensburg.de/dataset/soa_87-26 (accessed June 09, 2022)

In the same way as for RDPCIDAT, DCAT RDF descriptions were available but partially invalid when accessing via the Q&D RDF browser. The RDF included links to RDF XML documents that were not available at the given address (HTTP 404).

Due to the inclusion of DCAT terms in the linked data representation, this repository was still seen as positive for providing DCAT metadata.

COEMS Open Data

URL: http://dkan.isp.uni-luebeck.de/

Like RDPCIDAT that uses the same respository system, downloaded metadata contained dcat:Dataset and dcat:Distribution, but not dcat:Catalog. No catalog was found at other locations. The same issue with the Q&D RDF browser was observed as for RDPCIDAT.

Investigated datasets (both licensed CC-BY-NC-ND 4.0):
Høgskulen på Vestlandet. *CRV-2014 Offline Trace Data.* 2017. URL: http://dkan.isp.uni-luebeck.de/dataset/crv-2014-offline-trace-data (accessed June 09, 2022)

Høgskulen på Vestlandet. *LOR data race example.* 2017. URL: http://dkan.isp.uni-luebeck.de/dataset/lor-data-race-example (accessed June 09, 2022)

Overall: positive (in agreement with Tomoyose (78: section 6.4))

Fairdata IDA and Etsin Research Data Finder

URLs: https://ida.fairdata.fi/login and https://etsin.fairdata.fi/

Fairdata IDA, the content management system, needs access with a user account. No RDF files appeared to be available for download on Fairdata Etsin. According to the fairdata.fi FAQ Page (157), DCAT is used for internal metadata storage, but users cannot directly access this data.

Investigated datasets for Etsin (licensed CC-BY 4.0):
Tuomas Puttonen and Jukka Kuva. *Aalto 3D prints preview.* 2021. URL: https://etsin.fairdata.fi/dataset/2bd639f6-5738-441e-a28f-3f815844fce3 (accessed June 09, 2022)
Petr Stepanek. *Dataset: Low-concentration measurements of nuclear spin-induced optical rotation using SABRE.* 2020. URL: https://etsin.fairdata.fi/dataset/60946ecf-926d-4882-9d0d-77417bce3533 (accessed June 09, 2022)

No DCAT RDF was returned upon request to these links via Q&D RDF browser.
Fairdata Etsin was considered negative and Fairdata Etsin unlikely to provide DCAT metadata. Tomoyose also regarded Etsin as negative (78: section 6.6) and treated IDA as undetermined (78: section 6.9).

JRC Data Catalogue

URL: https://data.jrc.ec.europa.eu/

The repository About page[29] gave instructions on how to access RDF data for a given dataset or collection. The classes dcat:Catalog were used for collections in these files, and the classes dcat:Dataset and dcat:Distribution in the context of datasets.

[29]Joint Research Centre. *Data Catalogue.* URL: https://data.jrc.ec.europa.eu/about (accessed June 09, 2022)

Investigated datasets:
Silke Haarich, Stephanie Kirchmayr-Novak, Javier Sanchez Lopez, Maria Teresa Borzacchiello, Marios Avraamides. *Regional bioeconomy strategies in the EU.* 2022. URL: https://data.jrc.ec.europa.eu/dataset/a89482ff-83af-4c82-96ef-39b0a59eb345 (accessed June 09, 2022); dataset partially published under a no limitation reuse and partially under a reuse with attribution license)
Carlo Lavalle, Ana Barbosa. *LF433 - Built-up area per inhabitant (LUISA Platform REF2014).* 2015. URL: https://data.jrc.ec.europa.eu/dataset/jrc-luisa-lf433-built-up-area-per-inhabitant-ref-2014 (accessed June 09, 2022); dataset published under a no limitation reuse license

Investigated collections: *Copernicus Sentinel2 L1C cloud-free annual composites.* URL: https://data.jrc.ec.europa.eu/collection/id-00299 (accessed June 09, 2022)
Land Use and Coverage Area frame Survey. URL: https://data.jrc.ec.europa.eu/collection/id-00334 (accessed June 09, 2022)

Only minimal information not using the DCAT vocabulary was returned upon request of URIs with the Q&D RDF browser.

Overall: positive (in agreement with Tomoyose (78: section 6.12))

ICOS Carbon Portal

URL: https://www.icos-cp.eu/

The repository provided RDF documents for download, but these did not use the DCAT vocabulary.

Investigated dataset (both licensed CC-BY 4.0):
Denis Loustau, Christelle Aluome, Christophe Chipeaux, Jean-Luc Denou, Alain Kruszewski, Sebastien Lafont. *ETC NRT Meteosens, Bilos.* 2022. URL: https://meta.icos-cp.eu/objects/hB_mXvOtp0MMLc6nwuzRjb-D (accessed June 09, 2022)
Benjamin Dumont, Gaëtan Bogaerts, Henri Chopin, Anne De Ligne, Loïc Demoulin, Ariane Faurès, Bernard Heinesch, Bernard Longdoz, Tanguy Manise, Ayche Orgun. 2022. URL: https://meta.icos-cp.eu/objects/2PXnPpuPJ2M72zNqOwH04u2O (accessed June 09, 2022)

Requests to the URLs via Q&D RDF browser returned the same information as provided in the RDF files for download.
A SPARQL request for dcat:Catalogs at https://meta.icos-cp.eu/sparqlclient/ on June 09, 2022 revealed a single one. More information was obtained using the following SPARQL request and by accessing its URI[30] by browser. In both cases, the dcterms:description read "ICOS Carbon Portal example dataset metadata export to DCAT vocabulary". Trying to

[30]https://meta.icos-cp.eu/resources/cpmeta/icosL2objects (accessed on June 09, 2022)

access the URIs of several linked dcat:Datasets[31] on June 09, 2022 gave the message "The requested resource could not be found.".

```
PREFIX dcat: <http://www.w3.org/ns/dcat#>
select ?p ?o where {
   <http://meta.icos-cp.eu/resources/cpmeta/icosL2objects> ?p ?o .
}
```

Overall, this repository was seen as not providing DCAT-based metadata as the example did not appear to be created to publicly exhibit DCAT metadata. This repository was also one of those considered negative by Tomoyose (78: section 6.11)

Arctic Permafrost Geospatial Centre

URL: https://apgc.awi.de/

RDF documents using the DCAT vocabulary were available for download. The classes dcat:Dataset and dcat:Distribution were used. There did not appear to be any use of the class dcat:Catalog to describe the repository.

Investigated datasets (both licensed CC-BY 3.0):
Prajna R. Lindgren, Guido Grosse, Vladimir E. Romanovsky. *Lake Database from Landsat TM and ETM+,1970s, 2000s, 2013-2014, Western Alaska (US).* 2015. URL: https://apgc.awi.de/dataset/lake-db-ls-1970s-2000s-2013-2014-w-alaska (accessed June 09, 2022)

Daniel Sabel, Sang-Eun Park, Annett Bartsch, Stefan Schlaffer, Jean-Pierre Klein, Wolfgang Wagner. *Surface Soil Moisture from ENVISAT ASAR GM, 2005-2011, Ob Estuary and Yamal Peninsula, Western Siberia (RU).* 2012. URL: https://apgc.awi.de/dataset/due-rssm-asagm-2005-2011-005 (accessed June 09, 2022)

Access to the URLs with the Q&D RDF browser led to descriptions of the documents including links to the DCAT RDF files that could be downloaded. Accessing the URIs used for dcat:Datasets (different form the landing page URLs) did not directly return the DCAT triples, but information referring to the same RDF files available for download on the landing page.
Overall: positive (in agreement with Tomoyose (78: section 6.2))

[31] https://meta.icos-cp.eu/dcat/objects/fQBJ8LG5gxCq-5GbdOkd55ML
https://meta.icos-cp.eu/dcat/objects/ClUboWUi6uMXfI650Vctkkh3
https://meta.icos-cp.eu/dcat/objects/nbpHuOzH_xbM6DuZ1Pur-CNV
https://meta.icos-cp.eu/dcat/objects/dYO0Mp9d1R91AXVrRHNNfa16
https://meta.icos-cp.eu/dcat/objects/Dl5TgWRP-07Zz8gkb_Jf4Ptf
https://meta.icos-cp.eu/dcat/objects/jt4eCj8bfCoD0pgmEqV7BrQZ