Querying the Semantic Web for Concept Identifiers to Annotate Research Datasets

André Langer, Christoph Göpfert and Martin Gaedke

Distributed and Self-organizing Systems Group
Chemnitz University of Technology
Chemnitz, Germany
Email: {andre.langer,christoph.goepfert.martin.gaedke}@informatik.tu-chemnitz.de

***Abstract*—Researchers are encouraged to describe and publish research datasets so that others can find and reuse it. Following a semantic approach, well-known concept identifiers are necessary that can be used as values for meta-data properties to describe relevant characteristics of such a research artifact. Multiple research disciplines, communities or initiatives have already created and published standardized terms as taxonomies or ontologies for that. However, these developments are distributed on the Web. As a consequence, it can be difficult for researchers to become aware of already recommended structured terminologies. Thus, they will further rely on ambiguous, literal annotations. In this paper, we investigate existing data sources in the Semantic Web that contain relevant terms to describe a research dataset in a structured, content-oriented and fine-grained way and how to integrate it in corresponding applications. We therefore analyze both Linked Data services and traditional terminology services on how to retrieve and filter terms for particular research-relevant characteristics. It is shown that a variety of well-structured community-specific terminologies with relevant concepts already exist, but that community-overspanning building blocks are nevertheless missing. Furthermore, filtering and mapping particular concepts is still a challenge to improve interdisciplinary publishing.**

Keywords–Linked Data; Research Data Management; Data Publishing; FAIR; NFDI.

I. INTRODUCTION

The publication of research datasets is increasingly recognized as an essential part of scientific research [1]. Publishing research data in the World Wide Web has various advantages for both the creator and the consumer of the data [2]. It facilitates the reproducibility of research results, raises awareness and allows to discover, reuse and repurpose existing datasets [3]. However, the publication of scientific artifacts also poses challenges, in particular regarding the description and provisioning of a research dataset. In contrast to other types of publications, which can traditionally be classified by librarians, research datasets have to be annotated by the originating researcher or at least domain experts as they are normally not self-descriptive. The *FAIR Guiding Principles for scientific data management and stewardship* [4] address this challenge by defining requirements for publishing research datasets. These principles are intended to make data discoverable to humans and machines. Since their original publication in 2016, the *FAIR Principles* have received broad support, particularly from research journal publishers, including *Springer Nature*, *GigaScience*, or *Gates Open Research*.

Based on these principles, additional information about the dataset should be provided, such as administrative metadata on the creator, the involved institution and publication license, technical metadata on the media type, extent, recording software or device, and also further domain-specific descriptive metadata. Focusing on predicates, a set of established ontologies already exists that can be used to provide a basic metadata description for research datasets, including properties from initiatives, such as the *DataCite [5], Dublin Core Metadata Element Set (DCMES) [6], DCMI Terms [7], DCAT-AP [8], MARC [9], MODS [10], PREMIS [11]*, or projects like *schema.org* [12]. However, approaches on providing structured content and domain related object values are apparently still vague.

Nowadays, the provision of structured descriptive meta information on the content of the research dataset seems often neglected or only done in natural language in the abstract or a separate *ReadMe* description of the dataset [13]. Persistent Digital Object Identifiers (DOIs) are commonly provided to reference the dataset resource itself, but other concepts that describe characteristics of the dataset are provided as a free-text string in many research disciplines, although controlled vocabularies or established identifiers for common concepts would be possible to use as well that also allow semantic linking operations. This hinders the discovery and selective filtering possibilities in established data repository directories and crawling services, especially in an interdisciplinary context.

This situation can be improved, if the meta description of a research dataset does not only rely on predefined, well-understood properties provided by established ontologies, but also makes heavier use of unambiguous identifiers for object values in a Resource Description Framework (RDF) statement. A Linked Data-based approach allows to define type restrictions for the range of these object values and enables inference operations to discover even taxonomically similar concepts with different terms in the description.

Relying on single controlled vocabularies can only partially solve this issue. Ontologies already include a set of predefined persistent identifiers for concepts but they are limited in their expressiveness and focus only on a small scope of characteristics that can be described. In contrary to that, a comprehensive, atomic description of content characteristics requires many more necessary identifiers that have to be provided in a simple fashion. Sometimes, existing identifiers from general-purpose services in the Web, such as *DBpedia* [14], *Wikidata* [15] or *ConceptNet* [16], can be additionally used, but research dataset related concepts are likely to be too specialized in order to be listed there.

 ISBN: 978-1-61208-813-6

Even within the same research area, there can be vastly different types of research data characteristics. National and international research initiatives, such as the *National Research Data Infrastructure (NFDI)* in Germany, have started to work on harmonizing these terminologies into taxonomies, but we nevertheless face a distributed scenario where to query and retrieve existing relevant concept identifiers from.

Within this paper, we will discuss possibilities and challenges in querying concept identifiers from multiple existing sources in the application domain of research dataset meta descriptions. Our results can be used to build semantic-aware Web applications in the future that can provide structured explicit Linked Data research dataset meta descriptions with an improved user experience. The paper is part of the *PIROL* [17] PhD project about Publishing Interdisciplinary Research over Linked Data and has the following contributions:

1) We systematize existing data sources for concepts relevant for research dataset meta descriptions.
2) We describe a concept on how to query and filter these decentralized knowledge bases for relevant identifiers.
3) We run performance measurements on how to retrieve these identifiers in a Web application for research dataset management

The rest of the paper is structured in the following way: Section II describes the problem domain in detail and defines requirements. Section III provides a systematic mapping of existing knowledge sources for domain-specific research dataset concepts. Section IV discusses a proof-of-concept and different query strategies on how to incorporate these data sources into an application, which is then evaluated in Section V. Section VI contrasts our work to other existing approaches and Section VII summarizes our results and gives an outlook to future work.

II. Interdisciplinary Research Dataset Description

When publishing a research dataset, additional meta information has to be provided that can be used later so that other researchers are able to discover it based on their particular needs. Therefore, the corresponding meta information has to satisfy the following five aspects:

A1 The provided information has to be correct
A2 The provided information has to be machine-readable
A3 The provided information has to be sufficiently extensive
A4 The provided information has to be comprehensive
A5 The provided information has to be usable across multiple user groups

A1 is a necessity, as the provided meta information will be the foundation to discover a particular research dataset.

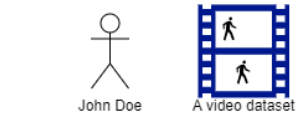
A2 is given, when a separate digital metadata description file is provided. However, this can either be done as a quite unstructured natural-language text, in a semi-structured way with key-value pairs, where the values can again contain descriptive continuous text, or highly structured where both the keys and values contain unambiguous identifiers.

A3 is commonly a trade-off between what can be stated about the data set and what is relevant information to actually discover it. An extensive number of statements can be made to describe the research dataset, but it should focus on filter criteria important for the consumer.

A4 asks for a certain understandability of the provided information, both for humans and machines. The provided terms have to represent a commonly known concept in this knowledge domain.

A5 is important especially in an interdisciplinary context when research data is not only relevant for a particular community but across multiple disciplines. It should therefore be possible to identify and link related or similar concepts.

Discovery operations nowadays commonly apply a keyword-based or fuzzy search on existing metadata descriptions in combination with some kind of natural language processing and named entity recognition. The metadata description itself concentrates on administrative meta information, whereas the description of the dataset content is either based on plain descriptive text or literal keywords. Figure 1 illustrates a scenario, where a research associate publishes, e.g., a research dataset that contains a set of recorded videos of elderly men walking.




```
{
  "@context": "https://schema.org/",
  "@id": "https://doi.org/10.5281/zenodo.12345678",
  "@type": "Dataset",
  "name": "An elaborate data set on human gait of elderly people"
  "description": "<p>This video dataset comprises 65 recordings of
elderly people moving outside in a parc, recorded with a Sony NEX
50EA.</p>",
  "creator": [{
     "@id": "https://orcid.org/0000-0001-8672-0508",
     "@type": "Person",
     "name": Doe, John"
   }],
  "datePublished": "2020-07-01",
  "encodingFormat": "video/mp4",
  "keywords": [
    "gait",
    "motion capture",
    "elderly people",
    "male participants",
    "deterioration"
  ],
"researchMethod": "video-based observation",
"license": "CC0 1.0",
  "version": "0.0.1"
}
```

Figure 1. Example metadata description for a video dataset in JSON-LD.

To improve the discovery and reuse of existing research dataset meta descriptions, such a Linked Data based approach can be valuable. The exemplary description satisfies aspect A1-A4, but we still face challenges when we want to find this research dataset among multiple disciplines based on certain filter criteria. Therefore, it is necessary on one hand to provide structured RDF statements on a research dataset subject, and on the other hand to make use of well-known unambiguous identifiers from controlled vocabularies for predicates and values in these statements. In this research activity, we particularly put focus on Uniform Resource Identifiers (URIs) that are provided as object values in these descriptions to express a concrete concept in an unambiguous way. We follow the hypothesis that this is an important requirement to improve the interdisciplinary discoverability of research data among multiple disciplines with the means of terminology mapping and linking and Linked Data inference capabilities for related concepts and sub-concepts.

In order to identify concept groups of major relevance, our pre-analysis consisted out of three steps:

 ISBN: 978-1-61208-813-6

Examine established vocabularies for attribute groups
The *DataCite / OpenAIRE* metadata schema specification and schema.org/Dataset were reviewed for common attributes and yielded the following reoccuring concept domains: *topic, resource type, (file) format/media type, rights/license, discipline, measurement technique/device, material, audience*.

Examine UI of established research dataset repositories
We carefully analyzed the input interface for research dataset meta description of *Zenodo* [18], *Open Science Framework (OSF)* [19] and *Mendeley* 20 and identified similar terminology groups as in the previous step.

Examine meta descriptions of existing research datasets
We verified the results through the result list of the *Google Dataset Search*. Apart from the already identified groups, it was obvious that additional relevant knowledge-domain specific concepts are often mentioned in the content description field text such as *demographic characteristics, examined objects, research and evaluation methods, metrics, measurement characteristics, models* or other applied paradigms.

In the following, we assume the existence of reusable terminologies as several communities have already worked on a standardization of such vocabularies throughout the last decades to represent particular research-related concepts. However, this knowledge is scattered along the entire Web in a decentralized way and can be found in different types of data sources. This complicates the reuse of existing terminology. In the following, we are, therefore, interested in existing data sources that fulfill the following requirements:

REQ1 DOMAIN The data source provides research-relevant terminologies of a specific domain that can be used as object values in the meta description of a research dataset
REQ2 SCHEME The data source provides the information in a semantic data serialization format with a clearly defined meta scheme to group and access similar concepts
REQ3 LABELING The data source provides labeled entities and persistent URIs for each concept
REQ4 API The data source offers a mechanism to access and filter these concepts remotely
REQ5 EXTENT The data source is actively maintained and has a complete or at least sufficient extent of entries

III. Sources for Research Data Concept Identifiers

Resource URIs from *DBpedia*, *Wikidata* or *ConceptNet* are commonly used in the Linked Open Data Cloud (LODC) to provide links to nameable entities. However, they focus on general-purpose data whereas scientific descriptions might need a domain-specialized vocabulary that is not part of *Wikipedia* or similar services. Additionally, the information there might be incomplete or of intermediate data quality.

We therefore conducted a systematic search for alternative sources for research dataset related concepts and mapped them to 4 groups as mentioned in the following sections. Deprecated or unavailable services were excluded from the mapping. We also excluded entity related groups for which appropriate authority services already exist, such as for identifying individual persons (Open Researcher and Contributor ID (ORCID)) [21], organizations (GRID [22], GND [23], LCCN [24], VIAF [25]), geographical information, such as countries and cities, or publications

A. Ontology catalogs

Ontology catalogs are a directory or collection of proposed vocabularies with a certain focus. Within these ontology catalogs, "1) metadata should be stored and handled based on a well defined syntax and semantics, i.e., a documented schema, 2) the catalog software must offer both a user interface and a widely accepted API for access by other software like applications and data portals" [26], as shown in Table I. The focus is set on providing standardized schemes and established ontologies with well-known properties, but these vocabularies might also contain (sub-)class definition or instances with a unique identifier that is appropriate to describe and filter certain meta-data value specific concepts.

TABLE I. COMPARISON OF ONTOLOGY CATALOGUES

Name	DOMAIN	SCHEME	LABELING	API	EXTENT (2019)
NCBO BioPortal	+ (biomed)	+	+	+	+ (792 vocabs)
LOV	+ (various)	+	+	+	+ (682 vocabs)
AberOWL	+ (various)	+	+	+	+ (522 vocabs)
ORR	+ (marine)	+	+	+	+ (499 vocabs)
OLS	+ (biomed)	+	+	o	+ (233 vocabs)
Ontobee	+ (biomed)	+	+	+	+ (201 vocabs)
IBC AgroPortal	+ (agro)	+	+	+	+ (106 vocabs)
Smart City OC	+ (smart city)	+	o	-	+ (70 vocabs)
RDA	+ (various)	-	+	-	+ (60 vocabs)
finto	+ (various)	+	+	o	+ (47 vocabs)
DCC	+ (various)	-	+	-	+ (40 vocabs)
HeTOP	+ (biomed)	+	+	-	+ (36 vocabs)
LinkedData.es	+ (various)	+	+	-	+ (35 vocabs)
Biblioportal	+ (biblio)	+	+	o	+ (31 vocabs)
SIFR BioPortal	+ (biomed)	+	+	+	+ (30 vocabs)
gfbio	+ (biomed)	+	+	o	+ (29 vocabs)
ONS Geography	+ (geography)	+	+	+	o (7 vocabs)

B. Authority services

Several terminology, thesauri and taxonomy services already exist for general or specific application domains, commonly built with the Simple Knowledge Organization System (SKOS) vocabulary as exemplary shown in Table II. Although they are often provided as a searchable Web page or data dump download without any API, they commonly also provide uniform resource identifiers and a hierarchical concept classification.

TABLE II. COMPARISON OF A SELECTED SUBSET OF AUTHORITY SERVICES BASED ON [27]

Name	DOMAIN	SCHEME	LABELING	API	EXTENT (2019)
EU NALs/Eurovoc	+ (general)	-	+	+	+ (150 groups)
Library of Congress	+ (general)	-	+	-	+ (70 groups)
UNESCO	+ (general)	-	+	o	+ (7 groups)

C. Instance datasets

This category basically contains all services from the Linked Open Data Cloud that provide structured meta information on a particular entity. Beside many less relevant concepts for research activities, they are also eligible to describe a research object related concept and provide established resource URIs. Table III focuses on aggregators of instance data sets and most prominent instance data providers.

TABLE III. COMPARISON OF INSTANCE DATASET PROVIDERS

Name	DOMAIN	SCHEME	LABELING	API	EXTENT (2019)
LOD Cache	o (general)	-	o	+	+ (50b stmts.)
LOD-a-lot	o (general)	o	o	-	+ (28b stmts.)
DBpedia	o (general)	+	o	+	+ (9.5b stmts.)
Wikidata	o (general)	+	+	+	+ (7.9b stmts.)
BTC	o (general)	+	o	-	+ (2b stmts.)
YAGO	o (general)	+	+	+	+ (1.4b stmts.)

 ISBN: 978-1-61208-813-6

D. Other concept sources

Beside these ontology, terminology and instance data collections and services, a variety of other data sources exist that might be relevant to retrieve concept identifiers. They are typically provided on separate websites in static text files by services like DataHub [28]. Examples are specifications, such as *CERIF* [29] or *KDSF* [30], use case-related developments, such as from data.gov.uk, or individual recommendations, such as vocabularies for representing data licenses, geographical information or file specific aspects. If these concepts are relevant for research dataset annotation processes or tool development, they can be downloaded and stored as a local data source and are therefore not further considered here.

Dedicated encyclopedic dictionary services exist, such as WordNet [31] and related projects like ConceptNet [16] or BabelNet [32], Wiktionary [33] or OmegaWiki [34]. Applications to annotate research datasets can also benefit from these service as they can also provide APIs, but were not in the particular focus of this research.

We also examined the usage of semantic search engines for concept discovery and retrieval purposes. However, at the point of writing, none of the existing Linked Data search services from the past was publicly available and functional, such as *Swoogle [35], Sindice [36], Falcons, SWSE, LOTUS or IBM Watson*.

E. Discussion

We manually reviewed the mentioned data sources against the relevant concept that we identified in Section II. It became obvious that no data source contained all relevant concepts. The list below shows exemplary data sources:

demographics BioPortal
device AberOWL, OLS, OntoBee
discipline UNESCO and other Authoritative Services
file format Static vocabularies
license Static vocabularies
measurements NCBO BioPortal
research methods LOV

Instead, we face a scattered scenario, where available terminologies and ontologies are provided only by some established aggregation services, or not at all (such as for certain *devices, materials, methods, metrics, models etc.*). In other cases, a researcher needs explicit knowledge on where to find terms for a particular knowledge domain in a decentralized landscape. It may even be misleading, that portals related to biomedical aspects might also identify interdisciplinary relevant concepts.

Characteristics of a research dataset meta description, such as the *topic* or *examined object*, are challenging to systematize at all. In these cases, the usage of established Linked Data entity description services, such as *DBpedia*, *Wikidata* or *ConceptNet*, is considerable to make use of persistent identifiers for a distinguishable concept.

Beside that, the interdisciplinary reuse of existing terms is hindered by the variety of representation formats for the hierarchical grouping of related concepts. Using rdf:type or a categorization is an approach commonly used by instance data sets to state that a concept is an instance of a specific type. Other concepts are represented as subclasses in the Web Ontology Language (OWL) or as a terminological hierarchy in SKOS. Hybrid approaches relying on SKOS and certain RDF Schema (RDFS) and OWL properties do also exist.

IV. AD-HOC TERMINOLOGY QUERYING

In practice, frontend Web applications to describe research datasets contain input interfaces where users have to enter or select a particular concept with a certain domain focus. Text input of literals is still dominating. Auto-suggestion elements can be applied in combination with Linked Data sources [37] so that a user can select the correct concept out of a list of existing concepts which can be solely based on the input literal or restricted to a certain concept type. In order to bridge the gap between existing terminology and ontology services and frontend user interaction, we focus on concept queries that can retrieve RDF statements (description, URI, etc.) for a given concept label/URI or which can retrieve a list of concepts based on a given type or super class via the SPARQL Protocol and RDF Query Language.

When relevant databases, such as listed in Section III, exist and the requirements from Section II are satisfied, it is possible to query concepts of a particular characteristic, as shown in Figure 2, by either

- importing relevant terminologies in a centralized data base
- running follow-up queries along relevant data sources
- using federated query approaches along multiple endpoints.

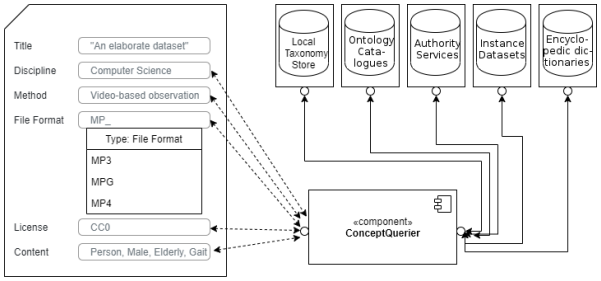


Figure 2. Conceptual architecture of a Concept Query Component.

The *ConceptQuerier* component provides a WebAPI that accepts requests with parameters stating the data the user has already entered in a text input field together with optional filters that describe the scope of the concepts that shall be retrieved. Such a component might analyze these parameters in advance and then query a set of appropriate services for existing concepts. This can either be done until the first Web service is able to satisfy the scope and returns corresponding concept data or in a parallel fashion, where the *ConceptQuerier* aggregates the results of multiple Web service responses.

Querying a remote service for a label or entity URI is considered as an already-understood trivial task. However, restricting existing resources based on filters requiring a particular class is more challenging as the type and hierarchy of an entity has to be identified additionally and the name of this type can either be filtered based on a keyword or based on a qualified identifier as conceptually shown in Figure 3.

```
SELECT DISTINCT ?concepturi ?conceptlabel
WHERE {
    ?concepturi  rdfs:label ?conceptlabel.
    ?concepturi rdf:type ?type.
    ?type rdfs:label ?typelabel.
    FILTER(CONTAINS(lcase(str(?typelabel)),lcase(str(?query)))).
    # ... - other additional concept scope filter patterns -
}
```

Figure 3. Conceptual SPARQL query for concepts of a particular type, where ?query contains the type restriction of the Web application text input field.

V. EVALUATION

We have implemented a software prototype [38] as a proof of concept of such a *ConceptQuerier* in a *NodeJS* based Web application. It offers a simple Web form with multiple text input fields which have an auto-suggest extension that provides concepts of a corresponding scope. After entering a keyword in such a text field, an AJAX request is created and sent to local a */suggest* REST API endpoint of our demo application, containing the entered literal string and a list of filter expressions defined in advance by a developer based on the domain scope of the input field in a JSON string. The *ConceptQuerier* implements a simple query strategy to SPARQL endpoints investigated in this paper to the Wikidata and DBpedia SPARQL endpoint. It first retrieves a list of matching concept URIs and then executes a DESCRIBE on each entity URI to get additional meta information for each of these concepts.

We used this application to measure a selection of indicators for service quality and data quality metrics of the identified concept data sources in order to assess to which extent they are appropriate in practice to retrieve Linked Data identifiers for concepts of a particular knowledge domain. We therefore focused on a PopulationCompletenessMetric, RelevanceMetric and LatencyMetric calculated on the *extent* (Table IV) and *processing time* (Table V) for the retrieved result list for four exemplary concept groups: *gender* (for a structural interdisciplinary demographic characteristic), *license* (for a data-related, interdisciplinary aspect), *file format*(for a computer domain-specific characteristic) and *research method* (for a research-concept oriented characteristic).

Querying via a SPARQL endpoint concrete concept labels or concept URIs was a trivial task. Retrieving concepts which are an instance or sub concept of a certain class was also straight-forward and yielded results in less than *1.0* second as long as caching strategies were established (**) or the entity URI of the super concept is known. However, this is typically not the case and includes a tedious manual lookup activity. And these URIs differ in practice between multiple data sources as long as no linking/inference operation is executed in the background. We therefore focused on a keyword-based search for appropriate super classes and retrieved a list of concepts based on these classes. This aspect and the measured latency times make these concept queries inappropriate for federated SPARQL approaches.

We evaluated at least one appropriate representative for each of the identified data source groups. For ontology catalogs, candidates were the *BioPortal* and *LOV*. Querying the *BioPortal* Ontology Catalog had to be done over the REST API and included the manual retrieval of subclasses from identified ontologies (*), as the provided SPARQL interface was only in beta status and limited to ontology meta information. For authoritative terminology services, we focused on *EuroVoc* and additionally provided *EU Named Authority Lists*. For instance data collections, we selected *Wikidata* and the *LODCache*. But the SPARQL endpoint of *LODCache* always ended with a timeout without text search index optimizations. Instead, we therefore considered *DBpedia*.

TABLE IV. RETRIEVED INSTANCES PER REQUESTED CLASS LABEL

Concept Group	LOV	BioPortal	EuroVoc	Wikidata	DBpedia
Gender	27	37*	4	34	28
License	11	42*	41	435	108
File Format	128	51*	172	4201	432
Research Method	16	149*	0	16	5

We used existing fulltext index query extensions of the services, where possible. Retrieving concepts based on a keyword search in associated class labels had the advantage that also concepts from different but similar groups could be retrieved (e.g., a query via *Wikidata* for instances containing the string *"license"* also returned *435* relevant concepts from groups, such as *"software license"*, *"free license"* or *"data copyright license"*, in comparison to a URI based constraint *wd:Q207621* with only *47* results). However, this also resulted in extended processing times which were ten times higher in our experiment than in the explicit case, and might also lead to false-positive results (the search for concepts related to *"gender"* in *Wikidata*, e.g., returned *546* results, where the majority instantiated the group *"tennis tournament edition by gender"*). Additionally, the terms used for describing a certain concept class differed between the services (*"License"* vs. *"Licence"*, *"Media Type"* vs. *"File Format"*, or *"Research Method"* vs. *"Scientific Method"*).

TABLE V. PROCESSING TIME PER REQUESTED CONCEPT LABEL IN SECONDS

Concept Group	LOV	BioPortal	EuroVoc	Wikidata	DBpedia
Gender	1.5s	1.5s*	1.0s	2.7s	0.2s**
License	1.5s	1.5s	1.4s	5.3s	0.2s**
Media Type	1.8s	2.9s	1.0s	5.8s	0.5s**
Research Method	1.5s	3.9s	1.0s	13.2s	0.2s**

False-positive results also originated from the data basis of the data provider itself. Queries for *"File Format"*, e.g., in *Wikidata* and *DBpedia* returned many concepts with multiple literal duplicates representing the same concept with additional appendices in the label, or no *file format* at all. Despite the high number of results from instance data providers for this use case, a high-quality population completeness was not given as some concepts were still missing. But using this kind of data sources for retrieving other specialized concept groups (such as research objects, devices, material) was still a valid strategy in comparison to approaches based on general taxonomies or ontology catalogs, where none of these concepts might be provided in a controlled fashion at all.

Searching for other, research-specific entities, such as a *research method*, revealed actual weaknesses of the tested data sources. Surprisingly, 3 out of 4 tested data sources returned some results for such a concept class. However, the obtained concept results were limited and also contained inappropriate concepts, e.g., from *DBpedia*. Services providing research-oriented, domain specific taxonomies or ontologies are a better choice in such a case as they commonly provide controlled terms and vocabularies.

From a technical point of view, it is demonstrated that a *ConceptQuerier* with a homogeneous interface to query multiple Linked Data concept sources was feasible to implement. However, separate queries had to be carefully designed for each data provider as the underlying data model differed on how concepts are classified into groups, based on rdf:type, rdfs:Class/subClass relationships, skos:broader or even skos:inScheme.

 ISBN: 978-1-61208-813-6

VI. Related Work

Using standardized identifiers to classify publications is already common for decades in a librarian environment [39–41]. Authoritative services exist there to represent entities, such as *authors*, *disciplines*, *keywords*, *publications* and *publishers* [42]. In this context, the usage of Linked Data in a librarian environment was discussed and applied multiple times [43]. However, this topic also became increasingly important for the description and discovery of other scientific publication artifacts. Especially the publication of research datasets requires expert insights where only the originating researcher can precisely provide a meta description of the provided content. Embedded librarians [44] might help to reuse existing classification systems, but interdisciplinary data exchange requires atomic research concepts [45] from established terminologies to support Quality-Driven Information Filtering among different disciplines [46]. The Semantic Web community has already presented concepts on federated SPARQL engines [47], and how to execute SPARQL queries over the Web of Data [48] and how to establish links between similar concepts from multiple ontologies [49]. Beyond that, science put emphasis on the development of ontologies, such as *DataCite*, *SWAP*, *LinkedScience*, *SciData* or *ModSci]*, for modelling relationships between scientific branches and scientific entities with a focus on established predicates. Querying and proxying decentralized data sources, such as NCBO, was discussed for single examples, such as the *BioPortal* [50] or *ONKI* [51]. Beside that, general-purpose encyclopedia and thesaurus-based terminology-providing services exist [52, 53]. Dedicated semantic terminology services providing concrete interdisciplinary concepts are still rare and limited to discipline-specific approaches, such as [54]. In both cases, relying on a single API to query for particular concepts will fail if these terminologies are very specific and not present in the knowledge base of the addressed service. Research dataset related concepts might be such an example, where an approach to query specific data sources as presented in this paper can provide better results.

VII. Conclusion

In this paper, we presented an analysis of data sources that provide labels and persistent identifiers for concepts that can be used as values in meta descriptions of research datasets and other interdisciplinary relevant scientific publications. We have identified four groups of potentially relevant services (ontology catalogs, authoritative services, instance dataset collections, static independent vocabularies). We provided an implementation of a Web-based prototype that is capable of querying these remote concept sources based on a particular concept scope represented by a concrete type or class label. In an evaluation, we showed a varying service and data quality of existing data sources. Response times, especially for a keyword-based class search, are still too high to consider remote services for ad-hoc queries in real-time user interaction. Apart from that, different underlying data models require adapted query patterns for each data service which make federated query approaches difficult in practice. From a content-perspective, we still face a scattered distributed scenario, as none of the data sources provided a set of discipline-overspanning, research-focusing, interdisciplinary-usable concepts in a single point of access. To improve the interdisciplinary discovery and reuse of research datasets, additional research in the future is needed.

Acknowledgment

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 416228727 – SFB 1410

References

[1] C. C. Austin *et al.*, “Research data repositories: Review of current features, gap analysis, and recommendations for minimum requirements,” *IASSIST Quarterly*, vol. 39, no. 4, p. 24, Jun. 2016.

[2] C. Steinhof, “Success criteria of research data repositories and their relevance for different stakeholder groups,” masterthesis, Fachhochschule Potsdam, 2018.

[3] D. S. Sayogo and T. A. Pardo, “Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data,” *Gov. Inf. Q.*, vol. 30, pp. 19–31, 2013.

[4] M. D. Wilkinson *et al.*, “The FAIR Guiding Principles for scientific data management and stewardship,” *Scientific Data*, vol. 3, no. 1, 2016.

[5] Datacite. Accessed: 2020-07-10. [Online]. Available: https://schema.datacite.org/

[6] Dublin core metadata element set (dcmes). Accessed: 2020-07-10. [Online]. Available: https://www.dublincore.org/specifications/dublin-core/dces/

[7] Dcmi terms. Accessed: 2020-07-10. [Online]. Available: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

[8] Dcat-ap. Accessed: 2020-07-10. [Online]. Available: https://www.dcat-ap.de/

[9] Marc. Accessed: 2020-07-10. [Online]. Available: http://www.loc.gov/marc/

[10] Mods. Accessed: 2020-07-10. [Online]. Available: https://rd-alliance.github.io/metadata-directory/standards/mods.html

[11] Premis. Accessed: 2020-07-10. [Online]. Available: https://www.loc.gov/standards/premis/

[12] schema.org. Accessed: 2020-07-10. [Online]. Available: https://schema.org/docs/documents.html

[13] M. Assante, L. Candela, D. Castelli, and A. Tani, “Are Scientific Data Repositories Coping with Research Data Publishing?” *Data Science Journal*, no. 15, 2016.

[14] Dbpedia. Accessed: 2020-07-10. [Online]. Available: https://wiki.dbpedia.org/

[15] Wikidata. Accessed: 2020-07-10. [Online]. Available: https://www.wikidata.org/wiki/Wikidata:Main_Page

[16] Conceptnet. Accessed: 2020-07-10. [Online]. Available: https://conceptnet.io/

[17] A. Langer, “PIROL : Cross-domain Research Data Publishing with Linked Data technologies,” in *Proceedings of the Doctoral Consortium Papers Presented at the 31st CAiSE 2019*, M. La Rosa, P. Plebani, and M. Reichert, Eds. Rome: CEUR, 2019, pp. 43–51.

[18] Zenodo. Accessed: 2020-07-10. [Online]. Available: https://zenodo.org/

[19] Osf. Accessed: 2020-07-10. [Online]. Available: https://osf.io/

[20] Mendeley. Accessed: 2020-07-10. [Online]. Available: https://data.mendeley.com/

[21] Orcid. Accessed: 2020-07-10. [Online]. Available: https://orcid.org/

 ISBN: 978-1-61208-813-6

[22] Grid. Accessed: 2020-07-10. [Online]. Available: https://www.grid.ac/
[23] Gnd. Accessed: 2020-07-10. [Online]. Available: https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html
[24] Lccn. Accessed: 2020-07-10. [Online]. Available: https://www.loc.gov/marc/lccn_structure.html
[25] Viaf. Accessed: 2020-07-10. [Online]. Available: https://www.viaf.org/
[26] E. Lapi, N. Tcholtchev, L. Bassbouss, F. Marienfeld, and I. Schieferdecker, "Identification and utilization of components for a linked open data platform," in *2012 IEEE 36th Annual Computer Software and Applications Conference Workshops*, 2012, pp. 112–115.
[27] Skos datasets. Accessed: 2020-07-10. [Online]. Available: https://www.w3.org/2001/sw/wiki/SKOS/Datasets
[28] Datahub. Accessed: 2020-07-10. [Online]. Available: https://www.datahub.io/
[29] Cerif. Accessed: 2020-07-10. [Online]. Available: https://eurocris.org/services/main-features-cerif
[30] Kdsf. Accessed: 2020-07-10. [Online]. Available: https://kerndatensatz-forschung.de/index.php?id=version1
[31] Wordnet. Accessed: 2020-07-10. [Online]. Available: https://wordnet.princeton.edu/
[32] Babelnet. Accessed: 2020-07-10. [Online]. Available: https://babelnet.org/
[33] Wiktionary. Accessed: 2020-07-10. [Online]. Available: https://www.wiktionary.org/
[34] Omegawiki. Accessed: 2020-07-10. [Online]. Available: http://www.omegawiki.org/
[35] Swoogle. Accessed: 2020-07-10. [Online]. Available: http://swoogle.umbc.edu/
[36] Sindice. Accessed: 2020-07-10. [Online]. Available: https://www.dataversity.net/end-support-sindice-com-search-engine-history-lessons-learned-legacy-guest-post/
[37] A. Langer, C. Göpfert, and M. Gaedke, "URI-aware user input interfaces for the unobtrusive reference to Linked Data," *IADIS International Journal on Computer Science and Information Systems*, vol. 13, no. 2, pp. 62–75, 2018.
[38] A. Langer, C. Göpfert, and M. Gaedke. Conceptquerier prototypical implementation. Accessed: 2020-07-10. [Online]. Available: https://gitlab.hrz.tu-chemnitz.de/vsr/researchinputform
[39] H. Albrechtsen and E. K. Jacob, "The dynamics of classification systems as boundary objects for cooperation in the electronic library," *Library Trends*, vol. 47, no. 2, pp. 293–312, 1998.
[40] P. Rafferty, "The representation of knowledge in library classification schemes," *Knowledge Organization*, vol. 28, no. 4, pp. 180–191, 2001.
[41] M. Satija, "Library classification : An essay in terminology," *Knowledge organization*, vol. 27, no. 4, pp. 221–229, 2000.
[42] E. T. Mitchell 2, *Library Linked Data: Early Activity and Development.*, 2016, vol. 52, no. 1.
[43] M. Hallo, S. Luján-Mora, A. Maté, and J. Trujillo, "Current state of Linked Data in digital libraries," *Journal of Information Science*, vol. 42, no. 2, pp. 117–127, 2016.
[44] D. Shumaker, *Embedded librarian: innovative strategies for taking knowledge where it's needed*. Information Today, 2012.
[45] A. Pfeifer, "More efficient research with Atomic Research," 2018, accessed: 2020-07-19. [Online]. Available: https://usertimes.io/2018/06/20/effizienter-forschen-mit-atomic-research/
[46] C. Bizer, "Quality-driven information filtering in the context of web-based information systems," Ph.D. dissertation, 2007, accessed: 2020-09-10. [Online]. Available: http://dx.doi.org/10.17169/refubium-14260
[47] N. A. Rakhmawati, J. Umbrich, M. Karnstedt, A. Hasnain, and M. Hausenblas, "Querying over federated sparql endpoints —a state of the art survey," 2013.
[48] O. Hartig, C. Bizer, and J.-C. Freytag, "Executing SPARQL Queries over the Web of Linked Data," accessed: 2020-09-10. [Online]. Available: http://olafhartig.de/files/HartigEtAl_QueryTheWeb_ISWC09_Preprint.pdf
[49] R. Parundekar, C. A. Knoblock, and J. L. Ambite, "Discovering concept coverings in ontologies of linked data sources," in *The Semantic Web – ISWC 2012*, P. Cudré-Mauroux *et al.*, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 427–443.
[50] M. Salvadores *et al.*, "Using SPARQL to query bioportal ontologies and metadata," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7650 LNCS, no. PART 2, pp. 180–195, 2012.
[51] K. Viljanen, J. Tuominen, E. Mäkelä, and E. Hyvönen, "Normalized access to ontology repositories," *Proceedings - IEEE 6th International Conference on Semantic Computing, ICSC 2012*, pp. 109–116, 2012.
[52] R. Navigli and S. P. Ponzetto, "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," *Artificial Intelligence*, vol. 193, pp. 217–250, 2012.
[53] R. Speer and C. Havasi, "Representing general relational knowledge in conceptnet 5." in *LREC*, 2012, pp. 3679–3686.
[54] N. Karam *et al.*, "A Terminology Service Supporting Semantic Annotation, Integration, Discovery and Analysis of Interdisciplinary Research Data," *Datenbank-Spektrum*, vol. 16, no. 3, pp. 195–205, 2016.

 ISBN: 978-1-61208-813-6