Advancing data sharing and reusability for restricted access data on the Web: introducing the DataSet-Variable Ontology

Margherita Martorana
Tobias Kuhn
Ronald Siebes
Jacco van Ossenbruggen
m.martorana@vu.nl
t.kuhn@vu.nl
r.m.siebes@vu.nl
j.r.van.ossenbruggen@vu.nl
m.martorana@vu.nl
Vrije Universiteit
Amsterdam, Netherlands

ABSTRACT

In response to the increasing volume of research data being generated, more and more data portals have been designed to facilitate data findability and accessibility. However, a significant portion of this data remains confidential or restricted due to its sensitive nature, such as patient data or census microdata. While maintaining confidentiality prohibits its public release, the emergence of portals supporting rich metadata can help enable researchers to at least discover the existence of restricted access data, empowering them to assess the suitability of the data before requesting access.

Existing standards, such as CSV on the Web and RDF Data Cube, have been adopted to facilitate data management, integration, and re-use of data on the Web. However, the current landscape still lacks adequate standards not only to effectively describe restricted access data while preserving confidentiality but also to facilitate its discovery. In this work, we investigate the relationship between the structural, statistical, and semantic elements of restricted access tabular data, and we explore how such relationship can be formally modeled in a way that is Findable, Accessible, Interoperable, and Reusable. We introduce the DataSet-Variable Ontology (DSV), that by combining CSV on the Web and RDF Data Cube standards, leveraging semantic technologies and Linked Data principles, and introducing variable-level metadata, aims to capture high-quality metadata to support the management and re-use of restricted access data on the Web. As evaluation, we conducted a case study where we applied DSV to four different datasets from different statistical governmental agencies. We employed a set of competency questions to assess the ontology's ability to support knowledge discovery and data exploration.

By describing high-quality metadata, both at the dataset- and variable levels, while maintaining data privacy, this novel ontology facilitates data interoperability, discovery, and re-use and it empowers researchers to manage, integrate, and analyze complex restricted access data sources.

CCS CONCEPTS

• **Information systems** → **Data structures**; **Document representation**; • **Security and privacy** → *Usability in security and privacy*.

KEYWORDS

Restricted access data, FAIR principles, Variable-level metadata, Privacy-preserving web data, Semantic Web

ACM Reference Format:
Margherita Martorana, Tobias Kuhn, Ronald Siebes, and Jacco van Ossenbruggen. 2023. Advancing data sharing and reusability for restricted access data on the Web: introducing the DataSet-Variable Ontology. In *Knowledge Capture Conference 2023 (K-CAP '23), December 05–07, 2023, Pensacola, FL, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3587259.3627559

1 INTRODUCTION

The ability to explore, analyze, and generate value from complex data sources plays a major role in both the public and the private sectors. The design of new public policies, advancements in intelligent transportation systems, scientific research, and customer needs analysis are just a few applications where intensive data use - and reuse - has become not only beneficial but also necessary [10]. The growing need for finding, accessing, and interpreting data has fueled initiatives to facilitate data-driven research, *Open Science* and *Open Data*. Scientific repositories, museums, and libraries have increasingly been publishing their resources as Open Data [8, 12], and initiatives such as Wikidata [22], have committed to "a world in which every single human being can freely share in the sum of all knowledge" [1]. Initiatives such as the Center for Open Science[2]

K-CAP '23, December 05–07, 2023, Pensacola, FL, USA

ACM ISBN 979-8-4007-0141-2/23/12.
https://doi.org/10.1145/3587259.3627559

[1] https://wikimediafoundation.org/wiki/Vision
[2] https://www.cos.io

and the Linked Open Data Cloud[3], have contributed to the development of Open Data infrastructures, recommender systems, search engines, and web applications to promote data management and reuse [9]. Further, with the increasing popularity of machine learning and artificial intelligence technologies, the need for reliable data for training, testing, and validation becomes a critical component for ensuring high-performance algorithms [18].

To navigate the millions of resources available on the Web, users must be able to search, find, access, manipulate, and analyze datasets. The deployment of Google Dataset Search [7] has shed light on the requirements, constraints, solutions, and problems linked to the domain of *Dataset Search*. However, there is still a gap between user needs and the availability, findability, and trustworthiness of datasets currently on the web [21, 23], particularly for restricted access data like patient or citizen data. Due to its very own nature, sensitive - or otherwise restricted - data, cannot be made *Open*, but it can be made FAIR. Following the FAIR Guiding Principles [25], data is not expected to be free or open but, instead, has to be Findable, Accessible, Interoperable, and Reusable.

This research primary focus is to explore existing models used for publishing structured data on the Web, and assess how comprehensively they can capture in the metadata the structural-, statistical- and semantic- nature of restricted access tabular data. The motivation behind centering this research on the tabular data format is twofold: firstly, tabular data is the most common data format used in real-world applications [20]. Secondly, in recent years tabular data has gained more and more attention due to its application in deep learning technologies [6], and thanks to initiatives, such as the SemTab Challenge[4], which focuses on automatic semantic annotation of tabular data.

To better define the terminology used in this paper, by *structural nature* we refer to the organization and format of a table, e.g. rows and columns, which allows to navigate and understand the relationships between different elements in a table. The *statistical nature* refers to the aspects impacting statistical analysis, like column completeness and data type. Lastly, the *semantic nature* denotes the properties of the individual variables measured in the table, and also the underlying semantic meaning within each column, often taking the form of a label or code. These variables can be domain-specific and are regularly described in codebooks or external vocabularies. In the context of this research, we base our definition of a variable on the one described by [14]: a variable is defined as "WHAT has been observed, measured, simulated, or calculated independently of WHERE (site description, geographical coordinates), HOW (procedure, protocol), and WHEN (measurement time, time resolution) the data acquisition has taken place". Overall, the main research question of this study is:

How can we formally model the relationship between the structural, statistical and semantic elements of restricted access tabular data, in a way that is Findable, Accessible, Interoperable and Reusable?

By addressing this, we aim to identify any gaps or limitations in the existing models and develop a more comprehensive, integrated and unified approach that captures the various aspects of restricted tabular data on the Web. Our contributions include: the DataSet-Variable Ontology, which describes the diverse nature of restricted tabular data; the formulation of competency questions; and a case study to assess the ontology's effectiveness and applicability.

2 RELATED WORK

As follows, we present the relevant background for this research, starting with an overview of the state of the art for publishing data on the web. After, we introduce the FAIR Guiding Principles, which serve as a fundamental framework for our ontology.

2.1 Publishing Data on the Web

The RDF Data Cube Vocabulary[5] and the CSV on the Web Primer[6] are two of the most commonly used standards for representing, describing and publishing tabular data on the Web. Data Cube offers a structured way to model multidimensional data, as well as the measures, dimensions, and attributes of datasets. It is designed to primarily capture aggregated statistical data, and it enables interoperability and data exchange through the use of Linked Data principles and the Semantic Web. It provides a flexible way to represent multi-dimensional data, supporting higher-level abstractions that go beyond the traditional tabular data structure. Additionally, the relationship between Data Cube and the Statistical Data and Metadata Exchange (SDMX)[7] plays an important role in facilitating the exchange of statistical data by enhancing semantic representation, and supporting interoperability through the SDMX Glossary.

CSV on the Web, on the other hand, is designed specifically to describe the structure of tabular data. The CSV on the Web toolkit can be used to transform tabular data from a CSV format into JSON-LD, by defining the structural template that the CSV file follows. Such aspect is very useful when translating multiple CSV files that adhere to the same patterns into an RDF format, and it facilitates bridging the gap between tabular data and the Semantic Web. However, CSV on the Web has limitations in representing relationships between columns or the hierarchical structure of data, which limit its expressiveness. Moreover, even though it provides mechanisms for adding metadata annotations, it is less suited to represent domain-specific concepts and variable-level descriptions.

Another available standards for describing datasets is the Vocabulary of Interlinked Datasets [8], which main goal is to express metadata about RDF datasets, and to facilitate communication between the publishers and the users of RDF data. VoID has been extensively used to describe and share RDF data available on the Web, for example from Wikipedia, and also to define access protocols and data integration tasks. Nevertheless, the VoID schema does not focus on structured data and, for that reason, does not provide terms for defining the structural architecture of tabular data.

In addition to these standards, the I-ADOPT Framework [14] has been recently introduced with the aim to support "a common approach to what is observed, measured, calculated, or derived", also referred to as *variable*. I-ADOPT proposes a framework to

[3] https://www.lod-cloud.net
[4] https://www.cs.ox.ac.uk/isg/challenges/sem-tab/
[5] https://www.w3.org/TR/vocab-data-cube/
[6] https://www.w3.org/TR/tabular-data-primer/
[7] https://sdmx.org
[8] https://www.w3.org/TR/void/

describe observable natural phenomena and properties, and their primary goal is to facilitate interoperability of terminology within the biodiversity domain. While I-ADOPT presents rich semantic properties to enrich and describe variables information, it is unclear how it could be applied outside of its original domain. Another example, is the Semantic Sensor Network (SSN) ontology [9], which scope is to describe sensors, their observation, study procedures and samples used. Lastly, the Data Documentation Initiative (DDI) addresses various aspects of the data lifecycle[10], from the collection to the publishing and archiving of data, and it offers both a codesbook and RDF version of its schema[11].

2.2 FAIR Guiding Principles

In the last two decades, academic interest around *Open Science* and *Open Data* has driven research into technologies for more accessible and reusable data. *Open Data* refers to data that is freely accessed, and that conforms to common machine-readable formats. Linked Data (LD), introduced in 2006 by Tim Berners-Lee [2], is defined as structured machine-readable data, based on the Resource Description Framework (RDF) [15]. With a primary goal in Open Science, Linked Open Data (LOD) is set to facilitate the replication of results, discovery, exchange and re-use of secondary data.

Initiatives such as Center for Open Science [12], have been involved in the development of Open Data infrastructures to promote the development, management, storage and reuse of secondary data [9]. Archives, repositories, museums and libraries have increasingly been publishing their resources as Open Data, developing systems to enhance research functionalities across various fields, from biomedicine to social sciences and cultural heritage. EUDAT [24], is an example of an international collaboration that focuses on providing Open Data for the broader European audience, and promotes cross-domain and cross-institutional research through toold like the Collaborative Data Infrastructure (CDI). Additionally, several government offices have started publishing statistical data online to facilitate easy reuse. However, several challenges arise during such data reuse, due to the quality and limited machine-readability, often in PDF, DOC or XLS formats [1][19][5].

Despite the significant interest in LOD, there are still challenges in reusing data containing personally identifiable information (PII), like patient and citizen data. Due to its sensitive nature, such data is rarely publicly shared without extensive data minimization and anonymization. Moreover, confidential data is often not only difficult to access and reuse, but also challenging to find. For example, the Central Bureau for Statistics Netherlands (CBS) has been collecting data since the 19th century, and although they started "riding the wave of digital revolution" in the 1960s [13], navigating through their catalogues is still a complicated and time consuming process. Researchers are limited to browsing PDF codebooks to understand the data, and the search functionalities available in the CBS Microdata Catalogue are limited to exact titles or keywords.

To address these challenges, the FAIR Guiding Principles, introduced in 2016 by Wilkinson et al [25], aim to improve the Findability, Accessibility, Interoperability and Reusability of data, with a special focus on high quality metadata. Several studies have found that applying the FAIR Principles can improve data management and stewardship [4][17], facilitate data reuse and resource citation [13], and ensure transparency, reproducibility and discoverability of secondary data [26]. High-quality metadata is particularly important when describing confidential or restricted access data, like CBS microdata, as standard text indexing methods are insufficient for this unique data format. In a previous systematic review, the authors of this paper investigated common practices used by researchers when dealing with restricted access data within the context of the FAIR Principles, and our findings highlighted the importance of metadata representation and accessibility, suggesting that high quality metadata has a key role in the reuse of restricted access data [16].

3 THE DATASET-VARIABLE ONTOLOGY

In this section, we introduce the main contribution of our work: the DataSet-Variable Ontology (DSV) for describing restricted access tabular data on the Web (Figure: 1). DSV is an OWL[11] based ontology, available at [14]. Our work enhances the representation of restricted access datasets through high quality dataset- and variable-level metadata, while following the FAIR Guiding Principles and maintaining confidentiality.

3.1 The Dataset Layer

While not explicitly defined in our ontology, the Dataset Layer represents those metadata descriptors that are considered essential for understanding and organising data on the Web. Within this layer, we can expect to find information such as the title, description, publisher, temporal and spatial coverage of a dataset. These properties provide the context to understand the data, and they have been widely standardized, adopted and defined by various available standards. Thus, our ontology does not focus on defining such properties and we assume that such information is already available as part of the metadata associated with the dataset. In Figure 1, we show how the dataset layer information can be described in the metadata using schema.org[15] (**schema:**) and DCTerms[16] (**dct:**), but other standards, such as DCAT[17], can also be used.

3.2 The Structural Layer

To model the structural metadata of a dataset identified by some given IRI, we start by assigning this IRI to be of type **`dsv:Dataset`**, where **dsv:** is the prefix for the DataSet-Variable Ontology. To define the column structure of the table, we also mint URIs for each column and make each column an instance of the class **`dsv:Column`**. We collect the set of columns for a dataset explicitly in a named table schema of type **`dsv:DatasetSchema`** using the **`dsv:column`** property to connect the schema with each of the columns. Finally, we connect the original dataset IRI with the schema IRI by using the **`dsv:datasetSchema`** property.

[9] https://www.w3.org/TR/vocab-ssn/
[10] https://ddialliance.org/Specification/DDI-Lifecycle/3.3/XMLSchema/instance.xsd
[11] http://www.ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/
[12] https://www.cos.io
[13] https://www.cbs.nl/en-gb/about-us/organisation/the-statistical-process
[14] https://w3id.org/dsv-ontology
[15] https://schema.org
[16] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[17] https://www.w3.org/TR/vocab-dcat-3/

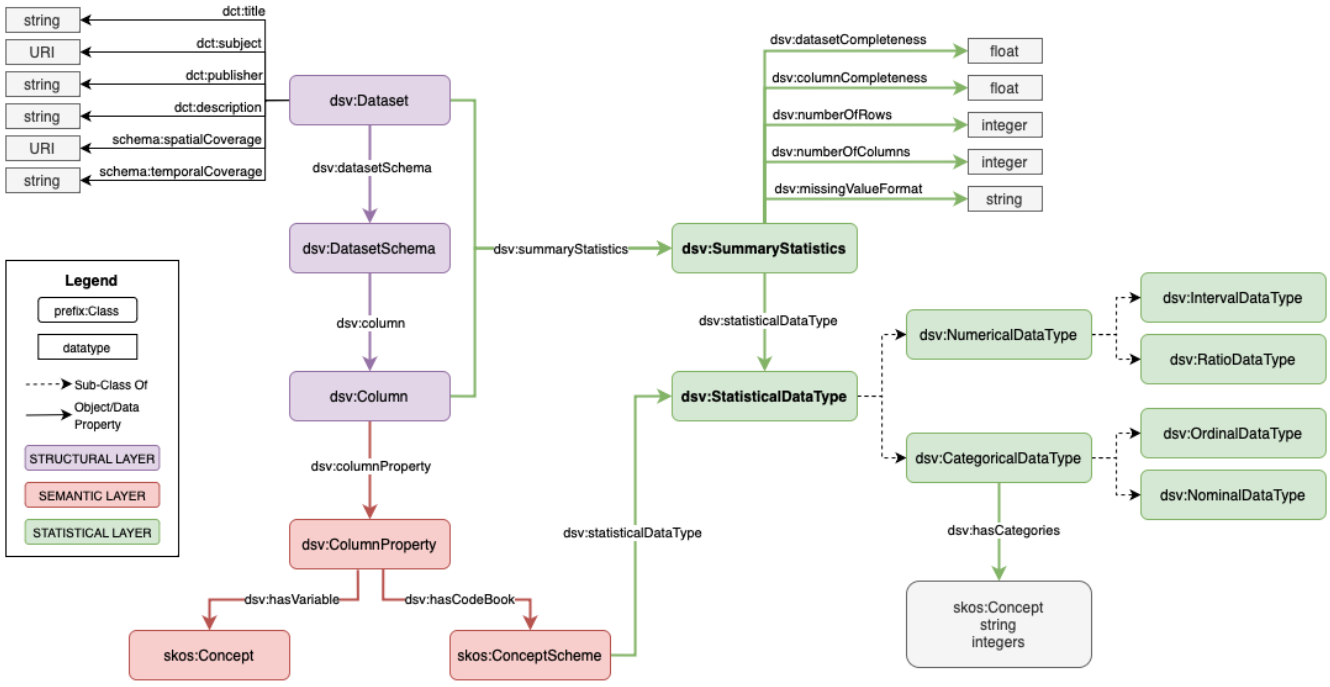

Figure 1: Visualisation of the DataSet-Variable Ontology, highlighting its structural-, semantic- and statistical layers.

3.3 The Statistical Layer

To address the statistical nature of tabular data, we can connect the dataset IRI to the class **`dsv:SummaryStatistics`** through the property **`dsv:summaryStatistics`**. This class can include a set of statistical information about the dataset: the number of columns and rows, the overall completeness, and the format by which the missing values are encoded within the dataset (e.g. with blank spaces, NaN or other placeholders). The properties that we introduce to describe such information are: **dsv:numberOfRows**, **dsv:numberOfColumns**, **dsv:datasetCompleteness** and **dsv:missingValuesFormat**. Moreover, we can also specify the format of the missing values and the completeness of each columns, by connecting the classes of dsv:Column and dsv:SummaryStatistics, using again the property dsv:summaryStatistics. In the latter case, we can use the same property as before for defining the missing value format, but in order to specify the completeness of each column we have introduced the property **dsv:columnCompleteness**. To add extra information about the specific type of data included in each column, we have also introduced the class **dsv:StatisticalDataType**, with further subclasses to define categorical and numerical data.

3.4 The Semantic Layer

In our ontology, we introduce classes to describe the semantics of individual columns. To define the semantic property measured in each column, we connect the class dsv:Column to the class **dsv:ColumnProperty**, which is an instance of rdf:Property, through the property **dsv:columnProperty**. For example, a column called *"Age Group"* connects to the property *example:hasAgeGroup*. We also introduce an approach for describing variable-level metadata. The description of variables can be achieved through the properties **dsv:hasVariable** and **dsv:hasCodeBook**, connecting from the class dsv:ColumnProperty. The first property associates the column with higher level variables, and the second to lower level variables. Continuing on the previous example of the *"Age Group"* column: we can connect this column to a higher level variable of *Age*, such as the WikiData entity *age of a person (Q185836)*[18]. The lower level variable, instead, can connect to a less general term than *age of a person*. This type of terms are usually domain-specific, and they are often found in codebooks or internal vocabularies, where also extra information about the context and hierarchies are present. An example of a lower level variable, in this case, could be the concept of *age of mother*, as it is still related to its higher level variable counterpart (*age of a person)*, but it is indeed more specific about its context and application.

3.5 Compatibility with Data Cube and CSVW

The DataSet-Variable Ontology is based on the established standards of CSV on the Web (prefix **csvw:)** and RDF Data Cube (prefix **qb:)**. Our classes and properties are designed as sub-classes and sub-properties of those found in such standards to ensure compatibility and integration with already available datasets that are implemented using CSV on the Web or RDF Data Cube. In Figure 2, we show the sub-class and sub-property relationships between the DSV ontology and RDF Data Cube and CSV on the Web.

[18] https://www.wikidata.org/wiki/Q185836

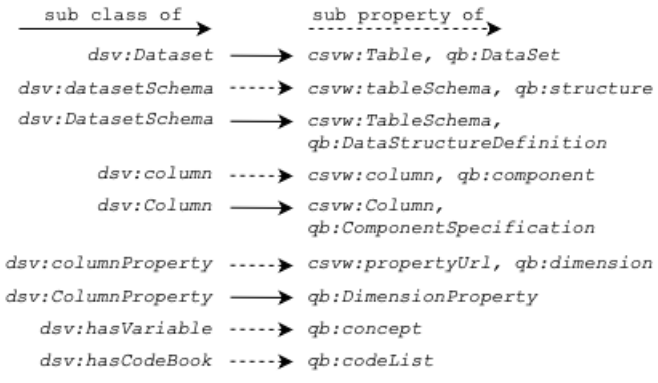

Figure 2: Illustration of the sub-class and sub-property relationships between the DataSet-Variable Ontology (dsv:), CSV on the Web (csvw:) and RDF Data Cube (qb:).

4 CASE STUDY

Our aim is to investigate how our model can be used to represent restricted access datasets, such as microdata, by allowing expressivity in the *structural*, *statistical* and *semantic* natures of the data. Because the intended data for our ontology is by nature restricted, we have conducted the case on example datasets obtained from four different national statistical organizations. This case study aims to address the quality of the model by answering the Competency Questions (CQs) below. The competency questions are intended as typical questions researchers would answer by inspecting the underlying data in case the metadata is insufficient. We followed the guidelines for CQ development introduced by [3]. Since our approach was intended to work for restricted tabular data, it is crucial that these CQs can be answered by using only the metadata:

1 What are the topics and variables represented in a dataset?
2 What other datasets represent the same topics or variables?
3 Are there any conceptual relationship between columns?
4 Given a variable, which columns are used to represent measurements of this variable in the various datasets?
5 What quantitative insights, such as number of columns and dataset completeness, can be derived from a dataset?
6 How many entries does the dataset have?
7 If there is any missing data, how is it encoded?
8 And how is the missing data distributed across columns?
9 Are there any potential sources of bias in the dataset?
10 Does the dataset have any primary keys? If yes, are the keys shared with other datasets?
11 Can two datasets with the same primary keys be merged?
12 How would the metadata description of the merged datasets look like?

4.1 Study Design

For this case study, we selected datasets from national statistical organizations: U.S. Government's Open Data[19], U.K. Open Data[20], Canada Open Government[21] and Central Bureau for Statistics Netherlands (CBS)[22]. From each data portal we selected one dataset related to *"fertility"*. This choice aimed for a balance between a general subject and diverse sources. With the CBS dataset, we specifically chose a subset of columns to align with the other datasets, as the original dataset contained more extensive population statistics beyond the scope of this research.

We used the DSV ontology to model the metadata for these four datasets, generating RDF files in turtle format. Additionally, we calculated summary statistical for each dataset using a Python script, available at [23]. To evaluate the model, we uploaded the turtle files into a GraphDB[24] instance, and translated the CQs into SPARQL queries, which can be found at [25]. The queries and study results are discussed in the next section.

5 EVALUATION

We present the case study results, starting from a descriptive analysis of the datasets, followed by an evaluation of DSV through competency questions implemented as SPARQL queries.

We used four datasets, in CSV format, from different countries: the USA (4 columns, 110 rows), the UK (10 columns, 962,760 rows), Canada (5 columns, 40,208 rows), and the Netherlands (56 columns, 72 rows). Due to the NL dataset's large number of columns, we selected a subset of 12 columns that better aligned with the other datasets. The application of the DSV ontology resulted in a total of 1095 triples, manually generated. The number of triples varied based on the dataset's column count. These triples describe metadata at the dataset and variable levels, but not at the cell level. We also represented in RDF hypothetical codebook as supplementary information, as illustrative examples. The Canada dataset had the most triples, with 147 for the core dataset and 276 for the codebook. The Netherlands dataset had 241 for the core dataset and 62 for the codebook. The UK dataset had 215 for the core dataset and 61 for the codebook. The USA dataset didn't require a code-book, being a very simple and straightforward dataset, and 93 triples were generate for the core dataset. All RDF files in turtle format are available at [26].

5.1 Competency Questions as SPARQL Queries

Hereby, we present the SPARQL queries addressing the competency questions (CQs) introduced in 4. Due to space constraints, the SPARQL queries shown here are provided in a shortened format. However, the complete set of queries can be access through this GitHub repository [27]. Overall, the DSV ontology allows to connect the structural, statistical and variable layer of restricted tabular data through the combination of DCT Terms (**@prefix dct:**), RDF Data Cube (**@prefix qb:**) and CSV on the Web (**@prefix csvw:**).

[19]https://data.gov
[20]https://www.data.gov.uk
[21]https://open.canada.ca/en
[22]https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS
[23]https://github.com/ritamargherita/DataSet-Variable-Ontology/blob/main/scripts/summary-statistics-generator.ipynb
[24]https://graphdb.ontotext.com
[25]https://github.com/ritamargherita/DataSet-Variable-Ontology/blob/main/case-study/queries.rq
[26]https://github.com/ritamargherita/DataSet-Variable-Ontology/tree/main/case-study
[27]https://github.com/ritamargherita/DataSet-Variable-Ontology/blob/main/case-study/queries.rq

5.1.1 CQ 1: What are the topics and variables represented in a dataset? The query below, retrieves the topics and variables of a given dataset. As an example, we show below an excerpt of the results we get for the Netherlands dataset and we can see its description, topics (Demographic and Social Statistics, Health), and variables (Reference Period, Age, Natality, Fertility Rate). It is important to note here that in this example query we assume that the objects of the property *dct:subject* are defined in some controlled vocabularies, and they are not just string values.

```
SELECT DISTINCT ?description
(GROUP_CONCAT(DISTINCT ?subjectLabel;SEPARATOR='\n') AS ?topics)
(GROUP_CONCAT(DISTINCT ?variableLabel;SEPARATOR='\n') AS ?variables)
WHERE {
 ?dataset a dsv:Dataset ;
    dct:description ?description ;
    dct:subject ?subject ;
    dsv:datasetSchema [ dsv:column ?column ] .
 ?subject skos:prefLabel ?subjectLabel .
 ?column dsv:columnProperty ?property .
 ?property dsv:hasVariable ?variable .
 ?variable rdfs:label | skos:prefLabel ?variableLabel .
 VALUES ?dataset { <..... > }} GROUP BY ?description
```

	description	topics	variables
1	"Number of teen pregnancies and rates per 1,000 females, by pregnancy outcome (live births, induced abortions, or fetal loss), by age group (under 20 years, 20 to 24 years, 25 to 29 years, 30 to 34 years, 35 to 39 years, or 40 years and over), 1974 to 2005."@en	"Demographic and social statistics, Regional and small area statistics, Health"	"Reference Period, Reference Area, Age, NATALITY"

5.1.2 CQ 2: What other datasets represent the same topics or variables? This SPARQL query is designed to retrieve pairs of related datasets and their title, given the condition that they have the same topics and variables. More specifically, given a dataset X (which identifier can be added to "`VALUES ?dataset { < dataset-uri > }`"), what other datasets share the same topics - objects of property dct:subject - and the same variables - object of property qb:concept. The example result below shows what is retrieved by this query when the Canada dataset URI has been specified as VALUES, and we can see that all the other three datasets share at least one topic and one variable. It is important to mention that in this instance we do not take in consideration that some other datasets might partially be related, for example by only matching the variables or the topics. Moreover, we also do not consider textual similarities in the description, as this task would require specific techniques (such as Natural Language Processing) that go beyond the capabilities of SPARQL. Therefore, this query has only partly answered the CQ 2, as at this moment we can only retrieve datasets that completely match the topics, categories or variables of a given dataset.

```
SELECT DISTINCT ?relatedDataset ?title
WHERE {
 ?relatedDataset a dsv:Dataset ;
    dct:title ?title ;
    dct:subject ?subject ;
    dsv:datasetSchema [ dsv:column ?otherColumn ] .
 ?otherColumn dsv:columnProperty ?otherProperty .
 ?otherProperty dsv:hasVariable ?variable .
 FILTER(?relatedDataset != ?dataset)
 { SELECT DISTINCT ?dataset ?subject ?variable
   WHERE {
    ?dataset a dsv:Dataset ;
        dct:subject ?subject ;
        dsv:datasetSchema [ dsv:column ?column ] .
    ?column dsv:columnProperty ?property .
    ?property dsv:hasVariable ?variable .
    VALUES ?dataset { <..... > }}}}
```

	relatedDataset	title
1	http://www.uk-open-data.org/example-dataset	"Annual births by age of mother"@en
2	http://www.usa-open-data.org/example-dataset	"NCHS - Births and General Fertility Rates: United States"@en
3	http://www.nl-open-data.org/example-dataset	"Geboorte: kerncijfers"@nl

5.1.3 CQ 3: Are there any conceptual relationship between columns? Similarly to the previous two questions, the combination of DCT Terms and DataCube allow us to leverage information about variables through the SPARQL query shown below. In this question, by conceptual relationship we mean whether multiple columns are linked to the same variable. From the results we can see that there are 3 columns linked to the SDMX variable *age*, and 1 column linked to the SDMX variable *refPeriod*.

```
SELECT DISTINCT ?column ?variable
WHERE {
 ?dataset a dsv:Dataset ;
    dct:subject ?subject ;
    dsv:datasetSchema [ dsv:column ?column ] .
 ?column dsv:columnProperty ?property .
 ?property dsv:hasVariable ?variable .
 VALUES ?dataset { <..... > }}
```

	column	variable
1	http://www.nl-open-data.org/example-dataset/columnv/perioden	sdmx-concept:refPeriod
2	http://www.nl-open-data.org/example-dataset/columnv/jongerDan20Jaar_9	sdmx-concept:age
3	http://www.nl-open-data.org/example-dataset/columnv/k_20Tot25Jaar_10	sdmx-concept:age
4	http://www.nl-open-data.org/example-dataset/columnv/k_25Tot30Jaar_11	sdmx-concept:age

5.1.4 CQ 4: Given a variable, which columns are used to represent measurements of this variable in the various datasets? Expanding from the previous question, we can query how columns across datasets are represented, which are linked to a certain variable, in this case the SDMX concept for *age*. The SPARQL query below retrieves the URI of the dataset where a certain column appears, together with human-readable information about such column: its label and its description. In doing so, we allow the user to easily identify the column of interest within a CSV dataset, even when the label of the column is not expressive enough to easily identify its relationship with a certain variable. For example, in the result below we can see that in the Canada dataset the column linked to the variable *Age* is called "Age Group"; in the UK dataset, instead, it is called "age_of_mother"; and in the Netherlands dataset is called "jongerDan20Jaar_9".

```
SELECT DISTINCT ?dataset ?columnLabel ?columnDescription
WHERE {
 ?dataset a dsv:Dataset ;
    dct:subject ?subject ;
    dsv:datasetSchema [ dsv:column ?column ] .
 ?column dsv:columnProperty ?property ;
    rdfs:label ?columnLabel ;
    rdfs:description ?columnDescription .
 ?property dsv:hasVariable sdmx-concept:age . }
```

	dataset	columnLabel	columnDescription
1	http://www.canada-open-data.org/example-dataset	"Age group"@en	"Age group of mother."@en
2	http://www.uk-open-data.org/example-dataset	"age_of_mother"@en	"Age group of the mother."@en
3	http://www.nl-open-data.org/example-dataset	"jongerDan20Jaar_9"@nl	"Mothers younger than 20 years old."@en

5.1.5 CQ 5: What quantitative insights, such as number of columns and dataset completeness, can be derived from a dataset? The SPARQL query below addresses this question by retrieving all triples linked to the property **dsv:summaryStatistics**, a novel property introduced in our ontology. In the results we can see a variety of quantitative information about the dataset, that exemplifies DSV's capability to describe quantitative insights of restricted access datasets without exposing confidential information.

```
SELECT ?p ?o
WHERE {
 ?dataset a dsv:Dataset ;
    dsv:summaryStatistics [ ?p ?o ] .
 VALUES ?dataset { <..... > }}
```

	p	o
1	dsv:numberOfColumns	"5"^^xsd:integer
2	dsv:numberOfRows	"40208"^^xsd:integer
3	dsv:datasetCompleteness	"0.99"^^xsd:float
4	dsv:missingValuesFormat	"NA"

5.1.6 CQ 6: How many entries does the dataset have? The SPARQL query below addresses this question by retrieving the number of rows of a given dataset, through the newly introduced property **dsv:numberOfRows**. For example, the result below shows that the Canada dataset have 40208 entries, information that in our ontology is available in the metadata. This might seem trivial, but it is an important piece of information that with our approach can be retrieved from the metadata without having access to the full dataset.

```
SELECT ?numberOfRows
WHERE {
 ?dataset a dsv:Dataset ;
    dsv:summaryStatistics [ dsv:numberOfRows ?numberOfRows ] .
 VALUES ?dataset { <......... > }}
```

	numberOfRows
1	"40208"^^xsd:integer

5.1.7 CQ 7: If there is any missing data, how is it encoded? The corresponding SPARQL query for this competency question, below, retrieves the triples linked to the properties **dsv:datasetCompleteness** and **dsv:missingValueFormat**. Usually, information about missing data and how it is encoded within the dataset are derived from the data itself. With the DSV ontology, instead, we show that such information can be pre-process and made available in the metadata, enhancing efficiency but also transparency. In fact, we can see in the result below that in the case of the Canada dataset there is 0.99 data completeness (where 1.00 corresponds to no missing data), and that the format of the missing data is the actual string "NA".

```
SELECT ?datasetCompleteness ?missingValuesFormat
WHERE {
 ?dataset a dsv:Dataset ;
    dsv:summaryStatistics [
        dsv:datasetCompleteness ?datasetCompleteness ;
        dsv:missingValuesFormat ?missingValuesFormat ] .
 VALUES ?dataset { <......... > }}
```

	datasetCompleteness	missingValuesFormat
1	"0.99"^^xsd:float	"NA"

5.1.8 CQ 8: And how is the missing data distributed across columns? Related to the previous CQ, the SPARQL query for CQ 8 retrieves the URI of each column, as well as human-readable information, such as label and description of the column, as well as the column completeness. Example results for the Canada dataset can be seen below, where the first 4 columns are fully complete and the last one, with label "VALUE", has a completeness of 0.97. By being able to address this query, DSV contributes to identify which columns could require extra attention from the user during analysis, as well as giving insights in data distribution and potential bias.

```
SELECT ?column ?columnLabel ?columnDescription ?columnCompleteness
WHERE {
 ?dataset a dsv:Dataset ;
    dsv:datasetSchema [ dsv:column ?column ] .
 ?column rdfs:label ?columnLabel ;
    rdfs:description ?columnDescription ;
    dsv:summaryStatistics [ dsv:columnCompleteness ?columnCompleteness ] .
 VALUES ?dataset { <......... > }}
```

	column	columnLabel	columnDescription	columnCompleteness
1	http://www.canada-open-data.org/example-dataset/column/refDate	"REF_DATE"@en	"Year of observation."@en	"1.00"^^xsd:float
2	http://www.canada-open-data.org/example-dataset/column/geo	"GEO"@en	"Place of residence of mother."@en	"1.00"^^xsd:float
3	http://www.canada-open-data.org/example-dataset/column/ageGroup	"Age group"@en	"Age group of mother."@en	"1.00"^^xsd:float
4	http://www.canada-open-data.org/example-dataset/column/pregnancyOutcomes	"Pregnancy outcomes"@en	"Pregnancy outcome."@en	"1.00"^^xsd:float
5	http://www.canada-open-data.org/example-dataset/column/value	"VALUE"@en	"Number of births."@en	"0.97"^^xsd:float

5.1.9 CQ 9: Are there any potential sources of bias in the dataset? This competency question cannot be answered by a simple SPARQL query, as bias can can manifest in many different ways across different steps of data collection, representation or analysis. For instance, in the query shown below, we retrieve the narrower concepts of a the concept *Gender* from the hypothetical UK codes-book that we manually generated. We can see from the results that there are only 2 narrower concepts: *Female Gender* and *Male Gender*. This limited conceptual representation of gender might be inadequate to capture the complexity and diversity of real-worl data, and thus be a cause of bias. Our ontology cannot, at the present time, be used to identify bias in data, but it supports the use of external vocabulary and codes-book, which can be a first step for bias investigation.

```
SELECT *
WHERE {
    <http://example.org/ns#uk-codebook/Gender> skos:narrower ?narrowerConcept .
    ?narrowerConcept rdfs:label ?label }
```

	narrowerConcept	label
1	eg-uk:GenderFemale	"Female gender"@en
2	eg-uk:GenderMale	"Male gender"@en

5.1.10 CQ 10: Does the dataset have any primary keys? If yes, are the keys shared with other datasets? The SPARQL query designed to answer this question and the result of it can be seen below. For the first part of this question, we can identify the column used as primary key through the property **csvw:primaryKey**, where the object of this property is the URI of the column - or columns - used as primary key. However, to identify whether that primary key is shared among datasets would require that multiple columns across different datasets would have the same URI, which is not recommended. With the DSV ontology, we can overcome this obstacle by searching among datasets for columns used as primary keys that have the same variable (through the property **qb:concept**). In the results below we can see that multiple datasets have primary keys related to the concept of "Year of observation", which have been retrieved because the columns have been linked to the external **sdmx-concept:refPeriod** concept as variable. Doing so, we do not claim that the primary keys of these datasets are, indeed, the same, but the user has now both machine and human-readable information to decide whether the primary keys are related.

```
SELECT ?dataset ?primaryKey ?label ?description
WHERE {
 ?dataset csvw:primaryKey ?primaryKey .
 ?primaryKey rdfs:label ?label ;
    rdfs:description ?description ;
    dsv:columnProperty [ dsv:hasVariable sdmx-concept:refPeriod ] . }
```

	dataset	primaryKey	label	description
1	http://www.canada-open-data.org/example-dataset	http://www.canada-open-data.org/example-dataset/column/refDate	"REF_DATE"@en	"Year of observation."@en
2	http://www.uk-open-data.org/example-dataset	http://www.uk-open-data.org/example-dataset/column/year	"year"@en	"Year of observation."@en
3	http://www.usa-open-data.org/example-dataset	http://www.usa-open-data.org/example-dataset/column/Year	"Year"@en	"Year of observation."@en
4	http://www.nl-open-data.org/example-dataset	http://www.nl-open-data.org/example-dataset/column/perioden	"perioden"@nl	"Year of observation."@en
5	http://www.nl-open-data.org/example-dataset	http://www.nl-open-data.org/example-dataset/column/perioden	"periods"@en	"Year of observation."@en

5.1.11 CQ 11: Can two datasets with the same primary keys be merged? Similarly to CQ 8, this question cannot be answered by a simple SPARQL query, but our ontology can help users deciding

whether two (or multiple) datasets could potentially be merged. As seen in the query below, the user can extract different kind of information from the metadata that may help deciding whether two or multiple datasets could be merged. For instance, we can check if there are related primary keys between datasets, as shown also in CQ 10, and we can retrieve human-readable information to better understand the content of the datasets, such as description, title and variable labels. Nevertheless, all this information can only indicate whether a potential merge is possible, but not confirm it. A possible solution to overcome this could be an extended version of this model, with explicit metadata information about whether two datasets are merge-able, and by what features (such as primary keys) such datasets can be linked.

```
SELECT DISTINCT *
WHERE {
 ?otherDataset a dsv:Dataset ;
    csvw:primaryKey ?otherPrimaryKey .
 ?otherPrimaryKey rdfs:label ?otherLabel ;
    rdfs:description ?otherDescription ;
    dsv:columnProperty [ dsv:hasVariable ?otherVariable ] .
 FILTER(?otherDataset != ?dataset)
 FILTER(?otherVariable = ?variable)
 { SELECT ?dataset ?primaryKey ?dimension ?label ?description ?variable
   WHERE {
    ?dataset a dsv:Dataset ;
     csvw:primaryKey ?primaryKey .
    ?primaryKey rdfs:label ?label ;
     rdfs:description ?description ;
     dsv:columnProperty [ dsv:hasVariable ?variable ] . }}}
```

5.1.12 CQ 12: How would the metadata description of the merged datasets look like? Addressing CQ 12 involves envisioning how the metadata of merged datasets would look like. While this task may not be directly answered through a SPARQL query, we can consider the metadata attributes and features of the DSV ontology that would characterize the metadata of merged datasets. For example, information about the description, temporal and spatial coverage, topics, and variables from the original dataset could be integrated into the merged one. However, the merging strategy (e.g., left or right joins) introduces differences like data inclusion/exclusion, which should also be documented in the merged dataset's metadata. For example, information about missing data from both datasets was handled, or whether duplicates where removed are some examples of important features that the merged metadata should have.

The provided SPARQL query below illustrates using the "UNION" operator to retrieve spatial/temporal coverage, topic labels, and variable labels from merged datasets. The example below using the Canada and UK datasets, shows that the merged spatial coverage corresponds now to two WikiData entities (Q16 for the Canada and Q145 for the UK). Also the temporal coverage, topics and variables now incolve multiple entries. This example highlights how the merging strategy impacts the metadata, such as whether it exclusively shows overlapping temporal coverage or spans the entire timeline. While this query provides insight into addressing the metadata of merged datasets, a more comprehensive solution requires detailed considerations of merging strategies.

```
SELECT DISTINCT
(GROUP_CONCAT(DISTINCT ?spatialCoverage;SEPARATOR=',\n') AS ?mergedSpatialCoverage)
(GROUP_CONCAT(DISTINCT ?temporalCoverage;SEPARATOR=',\n') AS ?mergedTemporalCoverage)
(GROUP_CONCAT(DISTINCT ?topicLabel;SEPARATOR=',\n') AS ?mergedTopicLabel)
(GROUP_CONCAT(DISTINCT ?variableLabel;SEPARATOR=',\n') AS ?mergedVariable)
WHERE { {
 ?dataset1 schema:spatialCoverage ?spatialCoverage ;
    schema:temporalCoverage ?temporalCoverage ;
    dct:subject [ skos:prefLabel ?topicLabel ] ;
    dsv:datasetSchema [ dsv:column [
        dsv:columnProperty [ dsv:hasVariable [
            rdfs:label ?variableLabel ] ] ] ] . }
UNION {
 ?dataset2 schema:spatialCoverage ?spatialCoverage ;
    schema:temporalCoverage ?temporalCoverage ;
    dct:subject [ skos:prefLabel ?topicLabel ] ;
    dsv:datasetSchema [ dsv:column [
        dsv:columnProperty [ dsv:hasVariable [
            rdfs:label ?variableLabel ] ] ] ] . }
    VALUES ?dataset1 { <..... >}
    VALUES ?dataset2 { <..... >} }
```

	mergedSpatialCoverage	mergedTemporalCoverage	mergedTopicLabel	mergedVariable
1	"http://www.wikidata.org/entity/Q16, http://www.wikidata.org/entity/Q145"	"1974/2005, 2000-07-01/2020-06-30"	"Demographic and social statistics, Regional and small area statistics, Health"	"Reference Period, Reference Area, Age, Sex"

6 CONCLUSION

In this study, we have introduced DSV, an ontology that leverages RDF Data Cube and CSV on the Web to facilitate the findability, interoperability and reusability of restricted access datasets. We have successfully addressed a number of competency questions, each focusing on a different aspect of sharing data on the Web. We have showed how we can investigate the main topics and variables of a given dataset (CQ1), how to identify related datasets (CQ2) and conceptual relationships between columns (CQ3). We have provided insights into the retrieval of statistical information (CQ5-6), the encoding of missing data (CQ7) and the detection of bias (CQ8). Moreover, we have discussed how DSV can facilitate the discovery and integration of merge-able datasets through shared primary keys (CQ10-11), and we also considered and analysed how the metadata of a merged dataset could be presented (CQ12).

The proposed novel DataSet-Variable Ontology, achieves a dual benefit: the ability to encapsulate the intricate conceptual relationships of restricted access tabular data, such as topics and variables, while also maintaining a clear definition of the data structure and statistical elements. We showcase that this ontology not only promotes compatibility between datasets, but also contributes to transparency and understanding. The importance of rich, informative and high-quality metadata is essential, especially when handling confidential data. By moving information into the metadata, we enable users to get insights, understand potential usefulness, and make informed decisions even when the data itself is not accessible.

Further implementations of the DSV ontology could assess how to further generalize its application by expanding to other data formats beyond tabular data. Moreover, introducing more structured ways for validating both the ontology and the construction of the RDF metadata, for example through Shapes Constraint Language (SHACL) [28] rules, could promote even further data integration and reusability.

ACKNOWLEDGMENTS

This work is funded by the Netherlands Organisation of Scientific Research (NWO), ODISSEI Roadmap project: 184.035.014.

[28] https://www.w3.org/TR/shacl/

REFERENCES

[1] Fernando Benitez-Paez, Auriol Degbelo, Sergio Trilles, and Joaquin Huerta. 2017. Roadblocks hindering the reuse of open geodata in Colombia and Spain: A data user's perspective. *ISPRS International Journal of Geo-Information* 7, 1 (2017), 6.

[2] Tim Berners-Lee. 2006. Linked data. https://www.w3.org/DesignIssues/LinkedData.html

[3] Camila Bezerra, Fred Freitas, and Filipe Santana. 2013. Evaluating Ontologies with Competency Questions. In *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) - Volume 03 (WI-IAT '13)*. IEEE Computer Society, USA, 284–285. https://doi.org/10.1109/WI-IAT.2013.199

[4] Martin Boeckhout, Gerhard A Zielhuis, and Annelien L Bredenoord. 2018. The FAIR guiding principles for data stewardship: fair enough? *European journal of human genetics* 26, 7 (2018), 931–936.

[5] Sanja Bogdanović-Dinić, Nataša Veljković, and Leonid Stoimenov. 2014. How open are public government data? An assessment of seven open data portals. In *Measuring E-government efficiency*. Springer, New York, NY, 25–44.

[6] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. 2022. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[7] Dan Brickley, Matthew Burgess, and Natasha Noy. 2019. Google Dataset Search: Building a Search Engine for Datasets in an Open Web Ecosystem. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1365–1375. https://doi.org/10.1145/3308558.3313685

[8] Victor de Boer, Jan Wielemaker, Judith van Gent, Marijke Oosterbroek, Michiel Hildebrand, Antoine Isaac, Jacco van Ossenbruggen, and Guus Schreiber. 2013. Amsterdam museum linked open data. *Semantic Web* 4, 3 (2013), 237–243.

[9] Erin D Foster and Ariel Deardorff. 2017. Open science framework (OSF). *Journal of the Medical Library Association: JMLA* 105, 2 (2017), 203.

[10] Moneeb Gohar, Muhammad Muzammal, and Arif Ur Rahman. 2018. SMART TSS: Defining transportation system behavior using big data analytics in smart cities. *Sustainable cities and society* 41 (2018), 114–119.

[11] Pascal Hitzler, Markus Krötzsch, Bijan Parsia, Peter F Patel-Schneider, Sebastian Rudolph, et al. 2009. OWL 2 web ontology language primer. *W3C recommendation* 27, 1 (2009), 123.

[12] Gary King. 2007. An introduction to the dataverse network as an infrastructure for data sharing. , 173–199 pages.

[13] Anna-Lena Lamprecht, Leyla Garcia, Mateusz Kuzak, Carlos Martinez, Ricardo Arcila, Eva Martin Del Pico, Victoria Dominguez Del Angel, Stephanie Van De Sandt, Jon Ison, Paula Andrea Martinez, et al. 2020. Towards FAIR principles for research software. *Data Science* 3, 1 (2020), 37–59.

[14] Barbara Magagna, Ilaria Rosati, Maria Stoica, Sirko Schindler, Gwenaelle Moncoiffe, Anusuriya Devaraju, Johannes Peterseil, and Robert Huber. 2021. The I-ADOPT Interoperability Framework for FAIRer data descriptions of biodiversity. *arXiv preprint arXiv:2107.06547* (2021).

[15] Frank Manola, Eric Miller, Brian McBride, et al. 2004. RDF primer. *W3C recommendation* 10, 1-107 (2004), 6.

[16] Margherita Martorana, Tobias Kuhn, Ronald Siebes, and Jacco van Ossenbruggen. 2022. Aligning restricted access data with FAIR: a systematic review. *PeerJ Computer Science* 8 (2022), e1038.

[17] Barend Mons. 2018. *Data stewardship for open science: Implementing FAIR principles*. Chapman and Hall/CRC.

[18] Yuji Roh, Geon Heo, and Steven Euijong Whang. 2019. A survey on data collection for machine learning: a big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering* 33, 4 (2019), 1328–1347.

[19] Erna Ruijer, Stephan Grimmelikhuijsen, Michael Hogan, Sem Enzerink, Adegboyega Ojo, and Albert Meijer. 2017. Connecting societal issues, users and data. Scenario-based design of open data platforms. *Government Information Quarterly* 34, 3 (2017), 470–480.

[20] Ravid Shwartz-Ziv and Amitai Armon. 2022. Tabular data: Deep learning is not all you need. *Information Fusion* 81 (2022), 84–90.

[21] Michael Stonebraker, Ihab F Ilyas, et al. 2018. Data Integration: The Current Status and the Way Forward. *IEEE Data Eng. Bull.* 41, 2 (2018), 3–9.

[22] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* 57, 10 (09 2014), 78–85. https://doi.org/10.1145/2629489

[23] Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae-Gil Lee. 2023. Data collection and quality challenges in deep learning: A data-centric ai perspective. *The VLDB Journal* (2023), 1–23.

[24] Heinrich Widmann and Hannes Thiemann. 2016. EUDAT B2FIND : A Cross-Discipline Metadata Service and Discovery Portal. In *EGU General Assembly Conference Abstracts (EGU General Assembly Conference Abstracts)*. Article EPSC2016-8562, EPSC2016-8562 pages.

[25] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 1 (2016), 160018. https://doi.org/10.1038/sdata.2016.18

[26] Mark D Wilkinson, Ruben Verborgh, Luiz Olavo Bonino da Silva Santos, Tim Clark, Morris A Swertz, Fleur DL Kelpin, Alasdair JG Gray, Erik A Schultes, Erik M van Mulligen, Paolo Ciccarese, et al. 2017. Interoperability and FAIRness through a novel combination of Web technologies. *PeerJ Computer Science* 3 (2017), e110.