A FAIR Core Semantic Metadata Model for FAIR Multidimensional Tabular Datasets

Cassia Trojahn, Mouna Kamel, Amina Annane, Nathalie Aussenac-Gilles, Bao-Long Nguyen

A FAIR Core Semantic Metadata Model for FAIR Multidimensional Tabular Datasets

Cassia Trojahn[1(✉)], Mouna Kamel[1], Amina Annane[2], Nathalie Aussenac-Gilles[1], and Bao Long Nguyen[1]

[1] IRIT, Université de Toulouse, CNRS, Université Toulouse 2 Jean Jaurès, Toulouse, France
Cassia.Trojahn@irit.fr
[2] Geotrend, Toulouse, France
amina.annane@geotrend.fr

Abstract. Tabular format is a common format in open data. However, the meaning of columns is not always explicit which makes if difficult for non-domain experts to reuse the data. While most efforts in making data FAIR are limited to semantic metadata describing the overall features of datasets, such a description is not enough to ensure data interoperability and reusability. This paper proposes to reduce this weakness thanks to a (FAIR) core semantic model that is able to represent different kinds of metadata, including the data schema and the internal structure of a dataset. This model can then be linked to domain-specific definitions to provide domain understanding to data consumers.

1 Introduction

Tabular format structures data into columns and rows. Each row provides values of properties of what is described by the row. Cells within the same column provide values for the same property. Columns can characterize dimensions within a multidimensional view. According to [2], the tabular format is the most widespread format for publishing data on the Web (37% of the datasets indexed by Google are in CSV or XLS). Data in this format has been made available as open data and datasets on various open data portals. On these portals, however, these datasets are described and presented with properties that are relevant to domain experts but not properly understood and reusable by other communities. For the latter, one of the challenges is to find relevant data among the increasingly large amount of continuously generated data, by moving from the point of view of data producers to the point of view of users and usages.

One way to overcome these weaknesses is to guarantee compliance of data to the FAIR principles: Findability, Accessibility, Interoperability, and Reusability [15]. These principles correspond to a set of 15 recommendations that aims to facilitate data reuse by humans and machines. They are domain-independent and

https://doi.org/10.1007/978-3-031-17105-5_13

may be implemented principally by: (F), assigning unique and persistent identifiers to datasets, and describing them with rich metadata that enable their indexing and discovery; (A), using open and standard protocols for dataset access; (I), using formal languages, and (FAIR) vocabularies to represent (meta)data; and (R), documenting (meta)data with rich metadata about usage license, provenance and data quality. So the first step towards the fulfilment of FAIR principles is to define precise metadata schemes. Indeed, 12 out of the 15 FAIR principles refer to metadata [15]. To go a step further in improving data FAIRness, several authors have shown that metadata schemes should be based on semantic models (i.e., ontologies) for a richer metadata representation [5]. Thanks to their ability to make data types explicit, in a format that can be processed by machines, ontologies are essential to make data FAIR [6].

While most efforts in data FAIRification are limited to specific kinds of metadata, mainly those describing the overall features of datasets, such a description is not enough to fully address all FAIR principles [7], in particular for promoting reuse of their data by other scientific communities. This paper addresses this challenge by proposing a core semantic model capable of representing different types of metadata, including the data schema and the internal structure of a dataset. This core model can be used in different domains and can be linked to domain-specific definitions to provide domain understanding for data consumers. The proposed model relies on existing FAIR vocabularies and ontologies and is itself compliant with the FAIR principles.

The rest of this paper is organised as follows. Section 2 discusses the main related work, followed by the description of the reused vocabularies and ontologies in Sect. 3. Section 4 presents the proposed model and Sect. 5 reports its evaluation. Finally Sect. 6 concludes the paper.

2 Related Work

A number of vocabularies has been proposed so far to represent metadata in general (Dublin core, VoID, Schema.org, DCAT, DCAT-AP), with extensions for accommodating specific kinds of data, such as geo-spatial data (GeoDCAT-AP) or statistical data (StatDCAT-AP). Several works also proposed specific metadata vocabularies. This is the case of [10] which presents a data model for generating ontology-based semantic metadata for spatial and temporal data, or of the European Open Science Cloud (EOSC) initiative[1] in the context of social sciences and humanities. Concerning observational data, several proposals have used RDF Data Cube (qb) combined to other vocabularies. In [9] the authors combined qb and SOSA to represent 100 years of temperature data in RDF. More recently, metereological data was represented in RDF with a semantic model that reused a network of existing ontologies (SOSA/SSN, Time, QUDT, GeoSPARQL, and qb) [16]. Instead, we propose here a core model that is common to different types of tabular data and that can be extended according to the specifics of the datasets (for instance, using QUDT, SOSA, GeoSPARQL).

[1] https://ddialliance.org/learn/what-is-ddi (accessed on 10th June 2022).

The work done here is an evolution of a previous one presented in [1]. While that first model had a special focus on representing spatio-temporal data (using GeoDCAT-AP and QB4st), here we propose a more generic core semantic model for representing any kind of tabular data for any domain by adopting DCAT and qb. Furthermore, we have introduced new notions such as the notion of a slice that can be also considered as a dataset, and the notion of collection of tabular data, as further detailed in Sect. 4.

3 Reusing Existing Vocabularies

The proposed model was developed following the NeOn methodology scenario 3 "*Reusing ontological resources*" [13]. We introduce here the main existing vocabularies that we relied on to build the core model, without detailing each activity of the methodology. These vocabularies provide metadata describing general features of datasets (DCAT), as well as the internal structure of a dataset (RDF data Cube and CSVW). All these vocabularies are recommended at least by the W3C or FairSharing[2] and thus act as FAIR vocabularies.

DCAT (Data CATalogue vocabulary). DCAT[3] is an RDF vocabulary designed to describe the datasets and data services in a catalog, thus facilitating the aggregation of metadata from multiple catalogs published on the web. It is based on 6 main classes (`Catalog`, `Resource`, `Dataset`, `Distribution`, `DataService` and `CatalogRecord`). It incorporates terms from existing vocabularies, including FOAF (relationships that people maintain with each other), PROV-O (provenance information), Dublin Core (metadata terms including properties, vocabulary encoding schemes, syntax encoding schemes and classes), SKOS (basic structure and content of concept schemes of controlled vocabulary) and vCard (people and organisations). DCAT was standardized in 2014 and has acquired the status of W3C recommendation.

RDF data Cube (qb). qb[4] is a W3C vocabulary dedicated to the representation of multidimensional data or hyper-cubes [14]. It builds upon several existing and recommended RDF vocabularies, such as SKOS, VoID (metadata about RDF datasets, intended to serve as a bridge between publishers and users of RDF data), Dublin Core, SCOVO (representation of statistical data), FOAF and ORG (organizational structures). qb allows the selection (i) of subsets of observations thanks to the notion of `Slice`, and (ii) of subsets of a given slice when the key slice has been fixed. Thus the publisher can identify and label those particular subsets. qb also allows the structure of a dataset to be described using the `DataStructureDefinition` and `ComponentProperty` entities. Each component property can be linked to the concept it represents (modelled as a SKOS concept).

[2] https://fairsharing.org/ (accessed on 10th June 2022).
[3] https://www.w3.org/TR/vocab-dcat-2/ (accessed on 8th June 2022).
[4] https://www.w3.org/TR/eo-qb/ (accessed on 8th June 2022).

CSVW. Resulting from the work of a W3C group on tabular data, CSVW[5] provides metadata at various levels, from table to groups of tables and how they are related to each other. A `Table` can be described with its url, schema, number of columns, foreign keys, transformations in other formats, etc.; each `Column` is then described by its name, title, type, position, whether the value is mandatory, etc. Furthermore, the interdependence between two tables may be represented by linking a column (or a set of columns) of a given table to a column (or a set of columns) of another table, thanks to references to their `ForeignKey`.

4 Proposed Model

We propose the Core Dataset Metadata Ontology (*dmo-core*) as a core semantic model capable of representing various types of metadata, including the data schema and the internal structure of tabular datasets (Fig. 1). It is a domain-independent model that can be enriched and specialized with domain ontologies to describe datasets in that domain. It is based on the FAIR vocabularies presented above. *dmo-core* is available online at https://w3id.org/dmo.

The notion of catalog is represented by `dcat:Catalog`, as a curated collection of metadata about resources (e.g., datasets and data services in the context of a data catalog). A catalog associates `dcat:Dataset`, which can be described with several types of metadata [4] using DCAT or DCT properties: descriptive metadata (`dct:description`, `dct:title`, `dcat:keywords`, etc.), quality (`dct:conformsTo`, etc.), provenance (`dct:publisher`, `dct:creator`), access rights (`dct:accessRights`, etc.) and versioning (`dct:hasVersion`, etc.). A `dcat:Dataset` may have different distributions `dcat:Distribution` (described by `dct:format`, `dcat:accessURL`, etc.), some of which may be in a tabular format. A table (`csvw:Table`) is described by its schema (`csvw:Schema`) which specifies the various columns (`csvw:Column`) it contains, as well as foreign keys (`csvw:ForeignKey`). A dialect description associated with a table provides hints to parsers on how to parse the distribution file (`csvw:delimiter`, `csvw:encoding`, etc.). The concept `csvw:TableGroup` represents a collection of datasets that share the same structure, what allows for defining the schema of these datasets for reuse. A `qb:Dataset` is associated to its structure metadata (`qb:DataStructure Definition`) as a set of measures (`qb:MeasureProperty`) organized along a group of dimensions (`qb:DimensionProperty`), together with associated metadata (`qb:AttributeProperty`). A dataset may be split into several subsets called slices (`qb:Slice`). A slice is characterized by a `qb:SliceKey` that specifies which dimensions are fixed (at least one). The `qb:concept` property allows to associate a `qb:ComponentProperty` (i.e., measure, dimension or attribute) to a concept to make its semantics explicit using domain ontologies.

The integration of these vocabularies relies on the definition in *dmo-core* of new concepts and properties (shown in orange in Fig. 1). A `dmoc:Dataset` is both a `dcat:Dataset` and a `qb:Dataset`, which allows a dataset to be

[5] https://www.w3.org/ns/csvw (accessed on 10th June 2022).

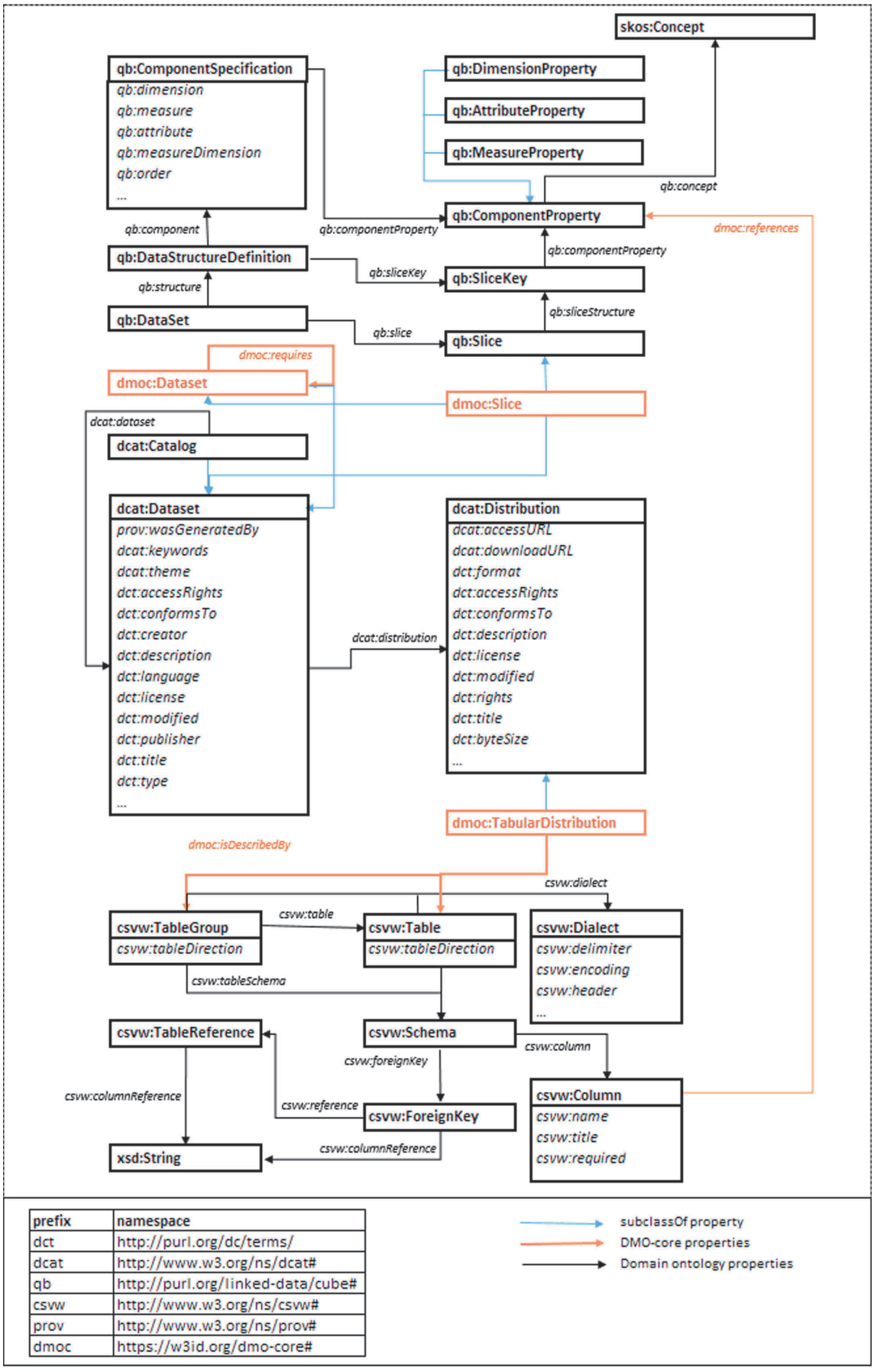

Fig. 1. The *dmo-core* model: main reused concepts and properties.

described in terms of global metadata and structure respectively. We also consider that a slice (a subset of a dataset) is a dataset, and can thus benefit from all properties of a `dmoc:Dataset`. This is why we introduce the `dmoc:Slice` concept. The concept `dmoc:TabularDistribution` is a specification of the `dcat:Distribution` to be able to describe tabular data. It is linked to a table (`csvw:Table`) or to a group of tables (`csvw:TableGroup`) with the `dmoc:isDescribedBy` property. To make the relationship between structural components (i.e., columns) and data schema components (i.e., measures and dimensions) explicit, we introduce the `dmoc:references` property between a `csvw:Column` and a `qb:ComponentProperty` (i.e., `qb:MeasureProperty` or a `qb:DimensionProperty`). For a finer semantics, we propose to link a `qb:ComponentProperty` to a concept of a domain ontology using the `qb:concept` property the range of which is `skos:Concept`. Finally, the `dmoc:requires` property aims to represent the dependency between datasets.

5 Evaluation

Several metrics, such as OntoMetrics [8], and tools such as OntOlogy Pitfall Scanner! (OOPS!) [12] can be used to evaluate ontology quality. As *dmo-core* highly relies on existing (reference) models, the quality measure here rather relies on the consistency when putting together these existing models. *dmo-core* was implemented in OWL2 and its consistency was checked thanks to different reasoners (Hermit, ELK, and Pellet) available in the Protégé[6] ontology editor. In terms of compliance to the FAIR principles, few online tools are available. One of this tools is FOOPS! [3], which takes as input an OWL ontology and generates a global FAIRness score [11]. It runs 24 different checks distributed across the 4 FAIR dimensions: 9 checks on **F** (unique, persistent and resolvable URI and version IRI, minimum descriptive metadata, namespace and prefix found in external registries); 3 checks on **A** (content negotiation, serialization in RDF, open URI protocol); 3 checks on **I** (references to pre-existing vocabularies); and 9 checks on **R** (human-readable documentation, provenance metadata, license, ontology terms properly described with labels and definitions). A score of 79% of FAIRness in FOOPS! is obtained for *dmo-core*. This score can be further improved by indexing the model in a searchable online catalog (LOV, for instance).

6 Conclusion and Future Work

This paper presented a FAIR core semantic model for descriptive and structural metadata of multidimensional tabular datasets. It was used to semantically represent several large collections of metereology datasets from the Météo-France catalog. We have now to evaluate whether the FAIRness of these metereology datasets actually helps non domain experts to reuse them.

[6] https://protege.stanford.edu/ (accessed on 10th June 2022).

Acknowledgement. This work is funded by the ANR (French National Research Agency) Semantics4FAIR project, contract ANR-19-DATA-0014-01.

References

1. Annane, A., Kamel, M., Trojahn, C., Aussenac-Gilles, N., Comparot, C., Baehr, C.: Towards the fairification of meteorological data: a meteorological semantic model. In: Garoufallou, E., Ovalle-Perandones, M.-A., Vlachidis, A. (eds.) Metadata and Semantic Research. pp, pp. 81–93. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-98876-0_7
2. Benjelloun, O., Chen, S., Noy, N.F.: Google dataset search by the numbers. In: Proceedings of the 19th International Semantic Web Conference, pp. 667–682 (2020)
3. Garijo, D., Corcho, Ó., Poveda-Villalón, M.: Foops!: an ontology pitfall scanner for the FAIR principles. In: Seneviratne, O., Pesquita, C., Sequeda, J., Etcheverry, L. (eds.) Proceedings of the ISWC 2021 Posters, Demos and Industry Tracks: From Novel Ideas to Industrial Practice Co-located with 20th International Semantic Web Conference (ISWC 2021), CEUR Workshop Proceedings, vol. 2980. CEUR-WS.org (2021)
4. Greiner, A., Isaac, A., Iglesias, C.: Data on the web best practices. Technical report, W3C (2017). Accessed 30 Sept 2021
5. Guizzardi, G.: Ontology, Ontologies and the "I" of FAIR. Data Intell. **2**(1-2), 181–191 (2020)
6. Jacobsen, A., et al.: FAIR principles: interpretations and implementation considerations. Data Intell. **2**(1–2), 10–29 (2020)
7. Koesten, L., Simperl, E., Blount, T., Kacprzak, E., Tennison, J.: Everything you always wanted to know about a dataset: studies in data summarisation. Int. J. Hum. Comput. Stud. **135**, 102367 (2020)
8. Lantow, B.: Ontometrics: putting metrics into use for ontology evaluation. In: Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - KEOD, (IC3K 2016), pp. 186–191. INSTICC, SciTePress (2016)
9. Lefort, L., Bobruk, J., Haller, A., Taylor, K., Woolf, A.: A linked sensor data cube for a 100 year homogenised daily temperature dataset. In: Proceedings of the 5th International Workshop on Semantic Sensor Networks, vol. 904, pp. 1–16 (2012)
10. Parekh, V., Gwo, J., Finin, T.W.: Ontology based semantic metadata for geoscience data. In: Arabnia, H.R. (ed.) Proceedings of the International Conference on Information and Knowledge Engineering. IKE 2004, 21–24 June 2004, Las Vegas, Nevada, USA, pp. 485–490. CSREA Press (2004)
11. Poveda-Villalón, M., Espinoza-Arias, P., Garijo, D., Corcho, O.: Coming to terms with FAIR ontologies. In: Keet, C.M., Dumontier, M. (eds.) EKAW 2020. LNCS (LNAI), vol. 12387, pp. 255–270. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-61244-3_18
12. Poveda-Villalón, M., Gómez-Pérez, A., Suárez-Figueroa, M.C.: OOPS! (OntOlogy Pitfall Scanner!): an on-line tool for ontology evaluation. Int. J. Semant. Web Inf. Syst. (IJSWIS) **10**(2), 7–34 (2014)
13. Suárez-Figueroa, M.C., Gómez-Pérez, A., Fernández-López, M.: The neon methodology framework: a scenario-based methodology for ontology development. Appl. Ontol. **10**(2), 107–145 (2015)
14. van den Brink, L., et al.: Best practices for publishing, retrieving, and using spatial data on the web. Semant. Web **10**(1), 95–114 (2019)

15. Wilkinson, M., Dumontier, M., et al.: The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data **3**(1), 1–9 (2016)
16. Yacoubi, N., Faron, C., Michel, F., Gandon, F., Corby, O.: A model for meteorological knowledge graphs: application to Météo-France observational data. In: 22nd International Conference on Web Engineering, ICWE 2022, Bari, Italy (2022)