A Metadata Profile to Promote Real-World Data Discoverability

Emiliano Reynares[1], Andrea Splendiani[2] and Hanne van Ballegooijen[3]

[1] *IQVIA, Provença 392, Barcelona, Spain*
[2] *IQVIA, Kirschgartenstrasse 14, Basel, Switzerland*
[3] *IQVIA Solutions B.V., Herikerbergweg 314, Amsterdam, the Netherlands*

Abstract

Real-World Data (RWD) are patient level data routinely collected from a variety of sources other than traditional clinical trials. RWD is increasingly needed for purposes such as making drug discovery more efficient and supporting healthcare systems to become more sustainable. However, the specific characteristics of RWD pose integration and findability-related challenges that arise from the heterogeneity, dynamism, and voluminosity of their sources. While previous work has addressed these issues by defining common data models and frameworks to integrate RWD into specific processes, a machine-actionable representation is needed to promote data discoverability and fit-for-purpose assessment. This paper introduces a Data Catalog Vocabulary (DCAT)-based metadata profile for an RWD-specific metadata vocabulary. The profile aims to tackle the findability-related challenges by enabling querying and processing of metadata in a standardized fashion.

Keywords

Real World Data, DCAT, metadata, findability, machine-actionable, metadata profile

Introduction

RWD are defined as routinely collected data relating to a patient's health status or the delivery of healthcare as opposed to data retrieved from traditional clinical trials. The current landscape of RWD comprises a large variety of types: clinical observations, medication, claims, molecular profiling, family history, mobile health, environmental, patient reported and social media [1, 2].

RWD can provide a holistic view of people's health by capturing information from sources that reflect the experience, behavior, and social and environmental context of patients collected by health records, claims data, registries, patient-reported outcomes, consumer credit-card spendings, geospatial data, and medical devices-harvested data [3, 4].

RWD can make drug discovery more efficient by better identifying unmet medical needs, optimizing the design of clinical studies, informing benefit-risk assessments, and enabling large-scale analytics. It can expand the on-label indications for existing products by showing their effectiveness in different populations and/or conditions [4, 5].

RWD can be used to make the healthcare systems more sustainable by supporting evidence-based decision making, comparative effectiveness research, health technology assessment, and pharmacovigilance. It can support the approval and market access of new drugs by demonstrating their safety, efficacy, and cost-effectiveness in real-world settings. [6].

However, RWD have specific characteristics, which poses challenges for its integration and findability [4, 7]. These challenges arise from the heterogeneity of RWD since it is retrieved from sources that differ by nature and quality, their dynamism and voluminosity as more data are generated and collected over time. Integration of RWD requires harmonization of data formats, standards, and terminologies across sources and domains. Findability of RWD requires effective

Proceedings of SWAT4HCLS 2024: The 15th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences, February 26-29, 2024, Leiden, The Netherlands.

emiliano.reynares@iqvia.com (E. Reynares); andrea.splendiani@iqvia.com (A. Splendiani); hanne.vanballegooijen@iqvia.com (H. van Ballegooijen)

0000-0002-5109-3716 (E. Reynares); 0000-0002-3201-9617 (A. Splendiani)

metadata (i.e., descriptive data that characterize other data creating a clearer understanding of their meaning and suitability for a specific purpose).

While some initiatives have addressed the data integration issue by providing common data models (CDMs) and analytical tools for observational data such as the Observational Medical Outcomes Partnership (OMOP) CDM [8], the Heads of Medicines Agencies (HMA) and the European Medicines Agency (EMA) advance on the findability topic by developing a framework to better integrate RWD into the regulatory decision-making processes [9]. This framework includes a list of metadata for RWD catalogues, a guide of good practices for the use metadata [10, 11]. HMA/EMA have also published *Data Quality Framework for EU medicines regulation* that provides actions, metrics, and a maturity model to support data-driven regulatory decision making [12].

However, it is not enough to have a list of abstract metadata to promote data discoverability and fit-for-purpose assessment. This should be accompanied by an RWD-specific metadata vocabulary. A machine-actionable representation of metadata able to be queried and processed in a standardized fashion would be ideal but requires a model for the metadata vocabulary. The Data Catalog Vocabulary (DCAT) and schema.org can be used as the foundation of such a model [13, 14]. For instance, the Google Dataset Search service [15] relies on the structured description of datasets published on the web using both DCAT and schema.org for dataset discovery [16]. These models are domain-agnostic i.e.: they describe cross-domain entities in a high-level fashion and provide mechanisms to extend them for applications in domain-specific settings.

An extension of these models would enable the definition of the semantics and syntactic aspects of the domain-specific metadata elements, as well as the relationships and constraints among them. It would also facilitate the validation and quality assessment of the metadata and allows the mapping and alignment with other metadata standards. A RWD domain-specific extensions is still missing. It would lead to a better description of the RWD assets through standardized metadata models and vocabularies, which would overcome the findability issues.

Schema.org and DCAT

Schema.org is a structured markup vocabulary to describe resources published on the web by defining *types* of information and their corresponding *properties*. It is increasingly being applied to web pages as it improves its findability and enhances the search experience for end users. Schema.org also provides the mechanisms to extend the vocabulary by defining additional types, properties, and constraints [17]. Bioschemas is one of those extensions aimed at improving the exposure of structured data in life sciences by adding data types related to biology and defines a set of minimum metadata properties [18, 19].

DCAT is a structured vocabulary developed by W3C and designed to enable a publisher to describe datasets and data services in a catalog using a standard model. It comprises a minimal set of classes, properties, and constraints specified through the Resource Description Framework (RDF) data model [20]. DCAT also reuses terms from other vocabularies as proposed by the Dublin Core Metadata Terms Recommendation [21]. DCAT can be extended by adding constraints as cardinality restrictions, classes and properties, controlled vocabularies, and requirements for access mechanisms. A DCAT extension is known as a DCAT profile. There are profiles for data portals in Europe and geospatial, statistical, and national variations of it [22].

Despite the similarities, schema.org and DCAT show differences in their technical design and the use cases they address. Schema.org is aimed at enhancing discovery and indexing of online resources via search engines while DCAT is meant to represent detailed information on datasets and data catalogues. Therefore, schema.org has a broader scope since it covers a variety of resources like web pages, articles, books, and events. On the technical side, schema.org has a simpler and flatter structure as it defines fewer classes and properties, does not use nested or complex objects, and allows multiple types for the same resource [23].

A metadata profile for RWD

This paper introduces a DCAT-based metadata profile for an RWD-specific metadata vocabulary. It is aimed to describe the specific details of the variety of types of data and sources comprised in RWD catalogs. It enables the standardized definition of the essential aspects of the RWD datasets such as therapeutic area, patient data, healthcare setting, legal and regulatory, and governance details.

The proposed profile described below extends the core six DCAT classes. A *dataset* is a collection of published data available for access or download in one or more serializations. Each serialization is represented by a *distribution* that can be provided via *data services*. The description of the datasets, distributions, and data services are registered in a *catalog*. The *catalog records* are used to describe such a registration by providing, for instance, provenance details. The *dataset series* allows to represent a collection of datasets that are published separately but share some characteristics that group them.

A catalog compliant with the profile must fulfill the following additional constraints. The catalog must have at least one catalog record. A catalog record must (a) describe the registration of exactly one dataset or data service, (b) have the date of formal recording of the asset, and (c) have the most recent date the catalog entry was updated. A dataset must (a) have an identifier, (b) have a title, (c) have a textual description, and (d) be serialized through at least one distribution.

A set of previously proposed Linked Data modeling patterns were applied to design the profile [24]. The *Custom Datatype* pattern to further describe the structure of literal values, *Label Everything* to ensure that every resource has a human-readable name, *Link Not Label* to ensure the complex entities are modeled as resources instead of labels, *Multi Lingual Literal* to allow for internationalized text, *Qualified Relation* to improve the modeling of complex relationships, *Repeated Property* to allow for multi-valued properties, *Topic Relation* to states the linkage between content-related resources, and *Typed Literal* to constrain the properties allowed values.

The profile defines a design pattern to assert the presence or absence of the features on the datasets [11]. This pattern comprises (1) the *is described by* property to link a dataset with the *Qualified Existential* class, (2) the *Qualified Existential* class to asserts the occurrence of a dataset feature through the *is data available* and *is about dataset feature* properties, and (3) the *Dataset Feature* vocabulary of features.

Usage of this pattern provides a mechanism to extend the profile by defining new dataset features on the *Dataset Feature* vocabulary. This allows us to avoid an explosion in the number of properties that would hamper its maintainability.

The metadata profile also defines additional classes and a variety of properties. The *Patient Group* class and its descendants – the *Patient Gender Group* and *Patient Age Group* classes – allow to describe the population aspect of the datasets.

An overview of the profile core classes, and the additional properties is shown in Figure 1. The diagram should be interpreted under the RDF open-world assumptions about the sub-class-of semantics in contrast to the object-oriented generalization meaning.

Conclusions and future work

This paper introduced a DCAT-based metadata profile for an RWD-specific metadata vocabulary. The model is aimed at tackling the findability-related challenges by enabling querying and processing of metadata in a standardized fashion. It extends the core DCAT classes through the definition of additional constraints, classes, and properties. A design pattern to specify the data elements applicable to the description of the RWD sources is also proposed.

As a proposal in its early stages, it opens several lines of further development. This involves a comprehensive analysis of the metadata vocabulary defined by the HMA/EMA and the assessment of a potential alignment with the proposed model.

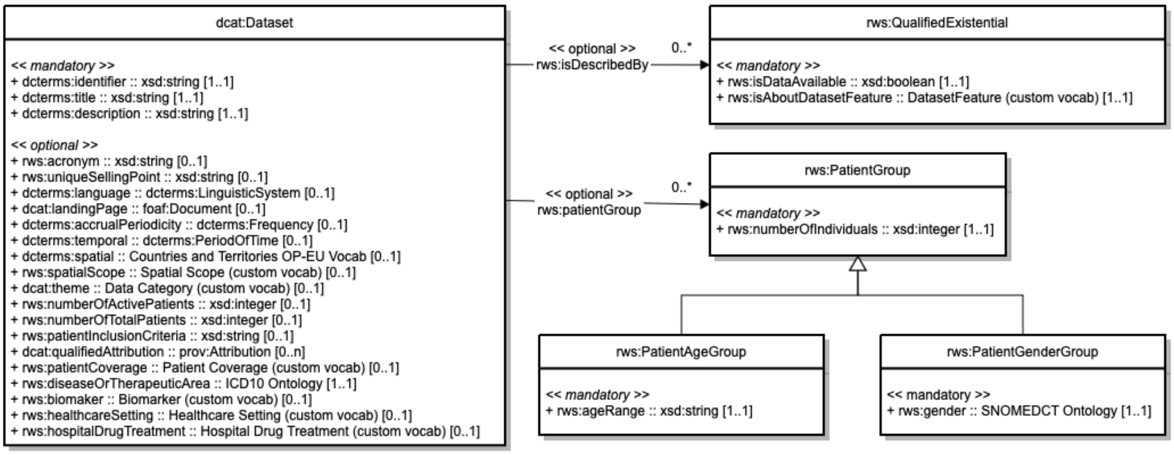
dcat:Dataset
<< mandatory >>
+ dcterms:identifier :: xsd:string [1..1]
+ dcterms:title :: xsd:string [1..1]
+ dcterms:description :: xsd:string [1..1]
<< optional >>
+ rws:acronym :: xsd:string [0..1]
+ rws:uniqueSellingPoint :: xsd:string [0..1]
+ dcterms:language :: dcterms:LinguisticSystem [0..1]
+ dcat:landingPage :: foaf:Document [0..1]
+ dcterms:accrualPeriodicity :: dcterms:Frequency [0..1]
+ dcterms:temporal :: dcterms:PeriodOfTime [0..1]
+ dcterms:spatial :: Countries and Territories OP-EU Vocab [0..1]
+ rws:spatialScope :: Spatial Scope (custom vocab) [0..1]
+ dcat:theme :: Data Category (custom vocab) [0..1]
+ rws:numberOfActivePatients :: xsd:integer [0..1]
+ rws:numberOfTotalPatients :: xsd:integer [0..1]
+ rws:patientInclusionCriteria :: xsd:string [0..1]
+ dcat:qualifiedAttribution :: prov:Attribution [0..n]
+ rws:patientCoverage :: Patient Coverage (custom vocab) [0..1]
+ rws:diseaseOrTherapeuticArea :: ICD10 Ontology [1..1]
+ rws:biomaker :: Biomarker (custom vocab) [0..1]
+ rws:healthcareSetting :: Healthcare Setting (custom vocab) [0..1]
+ rws:hospitalDrugTreatment :: Hospital Drug Treatment (custom vocab) [0..1]
<< optional >>
rws:isDescribedBy
0..*
rws:QualifiedExistential
<< mandatory >>
+ rws:isDataAvailable :: xsd:boolean [1..1]
+ rws:isAboutDatasetFeature :: DatasetFeature (custom vocab) [1..1]
<< optional >>
rws:patientGroup
0..*
rws:PatientGroup
<< mandatory >>
+ rws:numberOfIndividuals :: xsd:integer [1..1]
rws:PatientAgeGroup
<< mandatory >>
+ rws:ageRange :: xsd:string [1..1]
rws:PatientGenderGroup
<< mandatory >>
+ rws:gender :: SNOMEDCT Ontology [1..1]

Figure 1: A DCAT metadata profile overview. The *dcterms:* prefix refers to Dublin Core terms, *dcat:* to the DCAT terms, *foaf:* to the FOAF Ontology, *prov:* to the W3C PROV Ontology, and *xsd:* to the XML Schemas data types. The *rws:* prefix and *custom vocabs* annotated elements refers to items proposed by this profile.

The documentation of the data and metadata quality aspects could be enhanced by reusing specific frameworks such as the *Data Quality Framework for EU medicines regulation* published by the HMA/EMA.

The extension of the schema.org vocabulary by adding RWD-specific terms is a topic to further explore. The experience of the Bioschemas community in the life sciences domain, and the set of mappings to schema.org recommended by the DCAT specification can leverage this line of work.

This article describes work-in-progress for which the authors are looking for collaborations, and kindly ask interested parties to get in touch for further details. Given this scenario, the licensing issue is under discussion. This makes it unfeasible to openly release a fully-fledged RDF serialization of the proposed DCAT profile.

Acknowledgements

Special thanks go to Michèle Arnoe and Marie Sophie Schwalm (IQVIA) for the useful comments on the overall process of cataloguing and publishing real-world data, and for the feedback on this manuscript.

References

[1] FDA, Real-World Evidence, 2023. URL: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence.

[2] B. Swift, L. Jain, C. White, V. Chandrasekaran, A. Bhandari, D. A. Hughes, P. R. Jadhav, Innovation at the Intersection of Clinical Trials and Real-World Data Science to Advance Patient Care. Clinical And Translational Science, 11 (2018): 450-460. doi:10.1111/cts.12559.

[3] McKinsey & Company, Real-world data quality: What are the opportunities and challenges?, 2023. URL: https://www.mckinsey.com/industries/life-sciences/our-insights/real-world-data-quality-what-are-the-opportunities-and-challenges.

[4] P. Naidoo, C. Bouharati, V. Rambiritch, N. Jose, S. Karamchand, R. Chilton, R. Leisegang, Real-world evidence and product development: Opportunities, challenges and risk mitigation. Wien Klin Wochenschr, 133 (2021): 840-846. doi:10.1007/s00508-021-01851-w.

[5] BMC Series blog, Opportunities and Challenges for Evidence-based making in Health Sciences using Real-World Data, 2022. URL: https://blogs.biomedcentral.com/bmcseriesblog/2022/04/12/opportunities-and-challenges-for-evidence-based-decision-making-in-health-sciences-using-real-world-data/.

[6] SDG group, From Real World Data to Real World Evidence, 2023. URL: https://www.sdggroup.com/en-US/insights-room/real-world-data-real-world-evidence.

[7] ICMRA, ICMRA statement on international collaboration to enable real-world evidence (RWE) for regulatory decision-making, 2022. URL: https://www.icmra.info/drupal/sites/default/files/2022-07/icmra_statement_on_rwe.pdf.

[8] OHDSI, Standardized Data: The OMOP Common Data Model. Accessed in December 2023. URL: https://www.ohdsi.org/data-standardization/.

[9] HMA/EMA, Real-world evidence framework to support EU regulatory decision-making, 2023. URL: https://www.ema.europa.eu/en/documents/report/real-world-evidence-framework-support-eu-regulatory-decision-making-report-experience-gained_en.pdf.

[10] HMA/EMA, List of metadata for Real World Data catalogues, 2022. URL: https://www.ema.europa.eu/en/documents/other/list-metadata-real-world-data-catalogues_en.pdf.

[11] HMA/EMA, Good Practice Guide for the use of the Metadata Catalogue of Real-World Data Sources, 2022. URL: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/good-practice-guide-use-metadata-catalogue-real-world-data-sources_en.pdf.

[12] EMA, Data Quality Framework for EU medicines regulation, 2022. URL: https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/data-quality-framework-eu-medicines-regulation_en.pdf.

[13] W3C Working Draft, Data Catalog Vocabulary (DCAT) - Version 3, 2023. URL: https://www.w3.org/TR/vocab-dcat-3/.

[14] Schema.org. Accessed in December 2023. URL: https://schema.org.

[15] Google Dataset Search. Accessed in December 2023. URL: https://datasetsearch.research.google.com.

[16] Google Dataset structured data. Accessed in December 2023. URL: https://developers.google.com/search/docs/appearance/structured-data/dataset.

[17] R. V. Guha, D. Brickley, S. Macbeth, Schema.org: evolution of structured data on the web. Communications of the ACM, 59 (2016): 44-51. doi:10.1145/2844544.

[18] Bioschemas. Accessed in December 2023. URL: https://bioschemas.org.

[19] A. J. G. Gray, C. A. Goble, R. Jimenez, Bioschemas: From Potato Salad to Protein Annotation. In ISWC (Posters, Demos & Industry Tracks), 2017. URL: https://ceur-ws.org/Vol-1963/paper579.pdf.

[20] W3C Working Group. RDF 1.1 Primer, 2014. URL: https://www.w3.org/TR/rdf11-primer/.

[21] Dublin Core, DCMI Metadata Terms Specification. Accessed in December 2023. URL: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

[22] SEMIC EU, DCAT-AP 3.0.0 Release, 2023. URL: https://joinup.ec.europa.eu/collection/semic-support-centre/solution/dcat-application-profile-data-portals-europe/release/300.

[23] W3C, ISO 19115 – DCAT – Schema.org mapping, 2018. URL: https://www.w3.org/2015/spatial/wiki/ISO_19115_-_DCAT_-_Schema.org_mapping.

[24] L. D. I. Davis, Linked Data Patterns, 2022. URL: https://patterns.dataincubator.org.