

Take CARE of your patient data. Clinical And Registry Entries (CARE) Semantic Model

Pablo Alarcón-Moreno¹, Mark Denis Wilkinson¹

¹*Departamento de Biotecnología-Biología Vegetal, Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas, Centro de Biotecnología y Genómica de Plantas. Universidad Politécnica de Madrid (UPM)–Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria-CSIC (INIA-CSIC), Campus Montegancedo 28223 Pozuelo de Alarcón (Madrid), Spain*

Abstract

The Clinical And Registry Entries Semantic Model (CARE-SM) is designed to represent healthcare information stored in patient data registries through the use of Semantic Web technologies, with the objective of facilitating reasoning over federated data sources. Evolving from its origins as the Common Data Element Semantic Model (CDE-SM), CARE-SM improves on this prior art by the standardization and homogenization of its core structure, and also by the addition of a contextual metadata layer with temporal and event-based information. Consistency between data elements' representations allows several implementation improvements, including simplified data transformation and improved data discoverability.

Keywords

CARE-SM, Semantic Web, FAIR, Semantic model, Common Data Elements, Interoperability

1. Introduction

The “Big Data” era provides unprecedented opportunities to undertake large-scale analytics over combined clinical and molecular data. To achieve this, however, there is a need for standardized and interoperable healthcare data models such that federated exploration and analysis can be more easily achieved. There is an increasing number of sensitive registered patient data sources that are intended to be used for research purposes, but the lack of interoperability between data repositories thwarts this goal, causing researchers to invest valuable time finding, preparing, filtering, and combining datasets. [1, 2]


The FAIR Data Principles[1] call for data to be findable, accessible, interoperable, and reusable (FAIR), such that the value of data can be fully realized. Many of the FAIR objectives are realized through a combination of Web and Semantic Web technologies. For example, globally unique identifiers, such as URLs, are a requirement of FAIR, and the use of shared vocabularies (i.e. ontologies) and machine-readable syntaxes such as Resource Description Framework (RDF) [3] are hallmarks of most Semantic Web data architectures. CARE-SM [4] is intended to assist


SWAT4HCLS, February 26-29, 2024, Leiden, NL

✉ pabloalarconmoreno@gmail.com (P. Alarcón-Moreno); mark.wilkisonon@upm.es (M. D. Wilkinson)

🌐 <https://github.com/pabloalarconm> (P. Alarcón-Moreno); <https://github.com/markwilkinson> (M. D. Wilkinson)

🆔 0000-0001-5974-589X (P. Alarcón-Moreno); 0000-0001-6960-357X (M. D. Wilkinson)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

domain experts in achieving “FAIRness” by providing a pre-defined, generic data model that is inherently FAIR, and leverages the more complex features of the Semantic Web such as upper ontologies - specifically, the SemanticScience Integrated Ontology (SIO) [5] - which will help ensure that federated data can be used for logical reasoning.

2. The Genesis of CARE-SM

CARE-SM is an expanded and enriched representation of a prior model, which was primarily drafted to be capable of representing the Common Data Elements (CDE) for Rare Disease Registration [6] from the European Commission. Since that earlier work, the project was faced with the need to expand the number and variety of data elements that should be modeled in a FAIR manner. This included modeling treatments and interventions, imaging, and Patient Reported Outcomes (PROs) [7], and to keep a longitudinal history of patient events. These had a level of complexity that did not exist in the CDEs, and thus necessitated an extensive re-consideration and revision of the earlier CDE semantic model. Nevertheless, the requirement to be able to support semantic reasoning in the future remained, and thus we retained the use of SIO as the “semantic backbone” for CARE-SM.

SIO has a well-defined set of design patterns [8] for modeling scholarly data, which guides the entity-relationships that can exist in a SIO-based data representation. For the CDE model, the core design pattern was: an Identifier identified a Role; a Person played the Role; that Role was materialized in a Process; the Process had an Output; that Output was (generally) the measurement of an Attribute; the Attribute was an attribute of the Person. This core set of entity-relationships needed to be expanded to suit the broader range of observational and molecular data that needed to be modeled by CARE-SM. A full description of these extensions and revisions follows in the next section.

3. CARE-SM Overview

3.1. Core structure

Compared to the CDE-SM, there was a need to expand the core set of entity-relationships that were captured. Of particular relevance were the following additions to the core model: The process (e.g. a clinical procedure) is now related to several additional entities beyond the process output, including inputs, agents, routes, protocols and targets (see Figure 1). The output in the CDE model is now enhanced with a unit of measurement. Various kinds of observations will use different combinations of these new elements depending on the data element being modeled.

CARE-SM is built upon the Open Biological and Biomedical Ontology (OBO) Foundry [9, 10] to describe domain-specific ontological classes for every data element. The dual combination of SIO and OBO terms have been standardized compared with the previous CDE-SM where the prior used an arbitrary number of ontological classes to annotate each data element subcomponent. This standardization increases the data model consistency for transformation

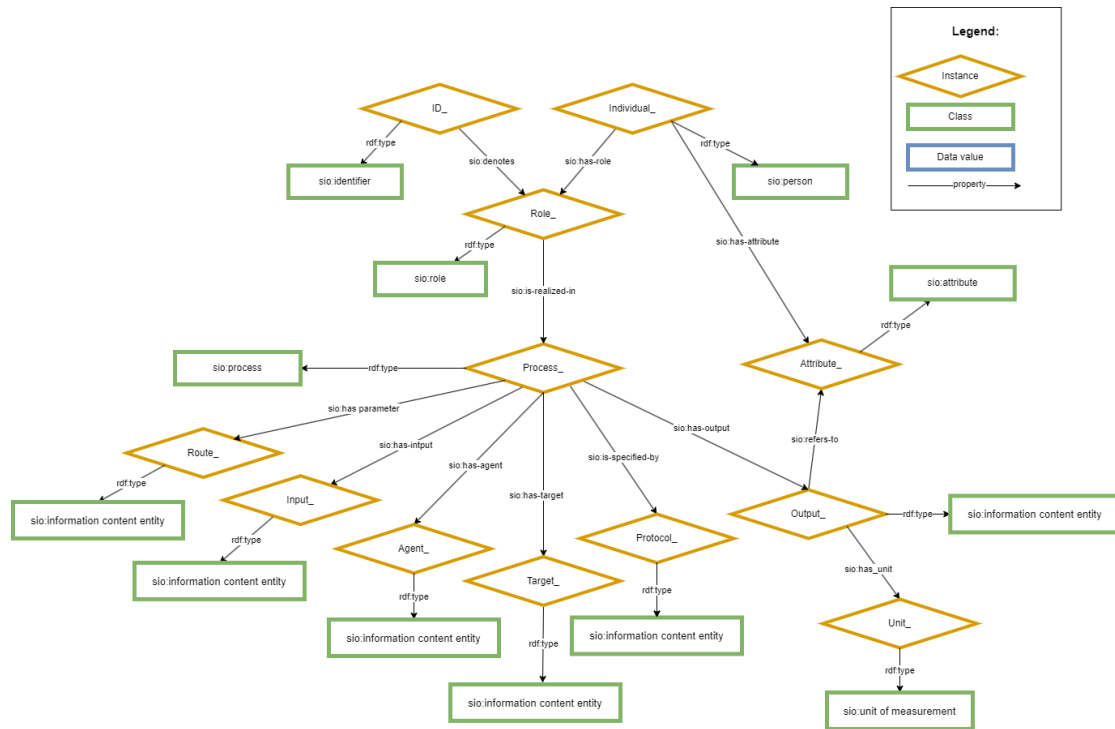


Figure 1: CARE-SM core structure.

and querying. Non-OBO ontologies such as Orphanet Rare Disease Ontology (ORDO) [11] are also present in the CARE-SM to annotate clinical conditions.

Figure 2 provides an example of the application of the core semantic model to a specific type of data - in this case, a tumor resection surgery of a patient: A person that has the role of a patient, denoted by a patient identifier, is participating in a tumor resection process. Several entities are associated with this process, such as the intervention protocol and anatomic structure that targets the surgery (defined as lung tissue in this example). Furthermore, the administration of a drug during the intervention (denoted by a drug identifier), followed by its route of administration. Intervention comments can be also added to the clinical process to enrich the contextual information in a human-readable way.

3.2. An added layer of metadata

One of the most consequential changes to the overall model when comparing CARE-SM to its predecessor is the introduction of a metadata layer that imparts context on each data element. Semantically, the contextual metadata layer groups every instance, class or property used to describe each data element. As shown in Figure 3, temporal information, in the form of time points or time intervals, is an example of the use of this layer, allowing the definition of a timeline of patient clinical encounters. The use of an encounter identifier can be added to the

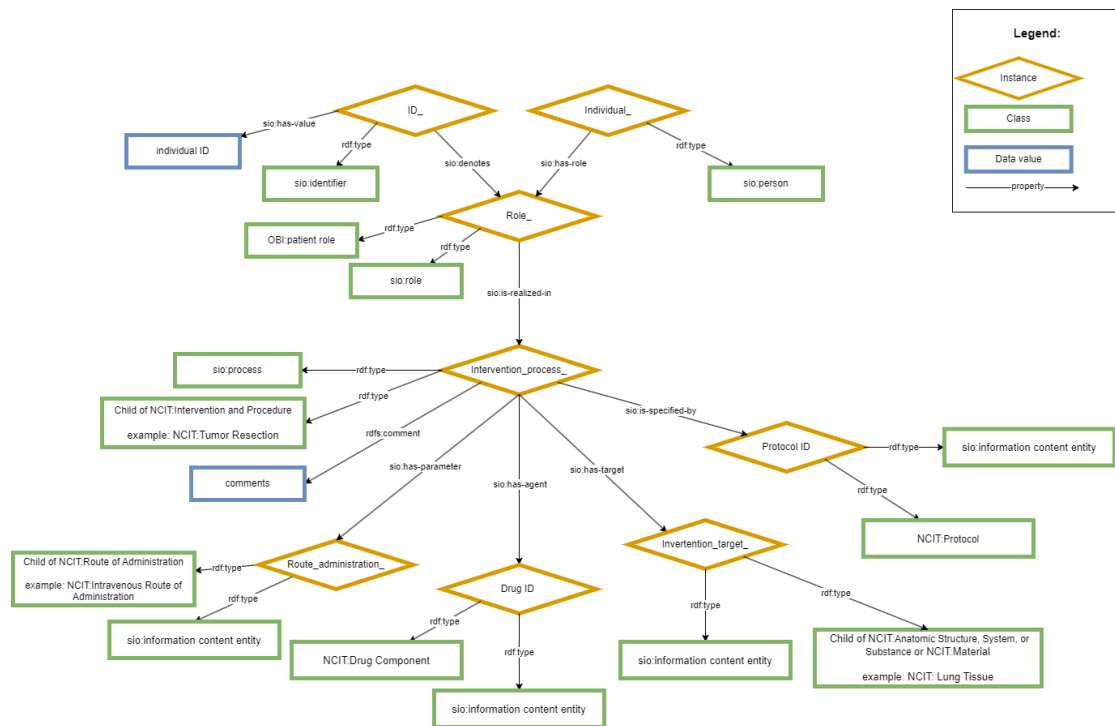


Figure 2: Exemplar tumoral resection surgery.

model to further relate several of these data elements under the same clinical episode or event, for example, a treatment regimen.

As is typical for RDF data, context is modeled using RDF-Quads [12] - that is, a fourth URI element accompanies every RDF triple. This context URI can then be used as the subject for additional triples in order to, for example, add temporal or administrative information about that data element, or to group sets of triples into other higher-level structures.

4. CARE-SM in Action

4.1. The CARE-SM implementation

Although CARE-SM only specifies a generic data model, we have generated a set of tools and guidelines to assist with the implementation of this model over patient data. The European Joint Project on Rare Diseases (EJP-RD) [13] has implemented an automated workflow for transforming tabular data into an RDF representation, which has been adapted to the requirements of the CARE-SM models in a variety of ways since its initial use with the CDE-SM. The CDE-SM workflow consumed data-element-specific CSV tables, where the CSV columns were referenced in data-element-specific templates structured using the YARRRML specification [14]. These YARRRML templates were transformed into the RML mapping

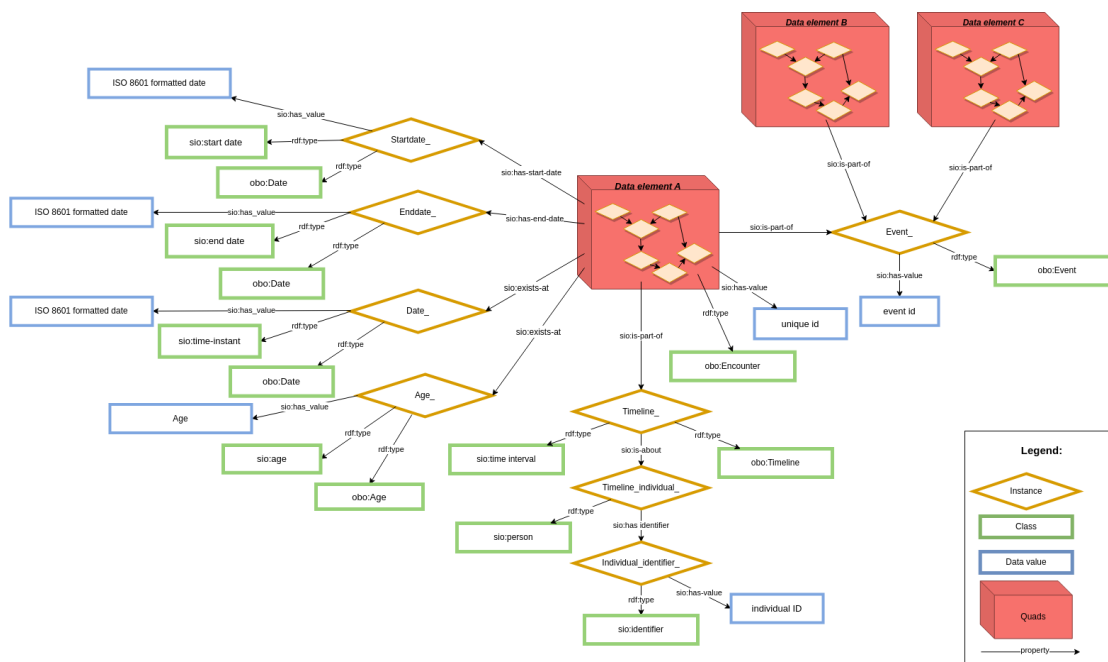


Figure 3: Contextual metadata layer.

language [3] by a YARRRML parser, and this mapping was applied to the CSV to generate the final RDF representation. This same “backbone” still remains in the CARE-SM implementation. However, since all data elements now conform to a single overarching model, every data type can now be represented using a common CSV template, with a common YARRRML. The only remaining data-element specificity is the set of columns that are required/optional for each data type, and these requirements are documented on the project’s GitHub [15]. The flexibility of allowing more optional data facets necessitated the addition of some additional complexity in the transformation templates - in particular, the use of “conditionals” (if/then) within the YARRRML to decide when an RDF statement should be generated. In addition, the transition from RDF Triples to RDF Quads required the addition of a new element (“graph”) in the YARRRML templates. Finally, a toolkit has been created in order to perform quality control, data manipulations, and other pre-processing steps to reduce the burden of accurate CSV generation by the users. This toolkit reorganizes the user-provided CSV template into its final form, compatible with the YARRRML template, prior to the RDF transformation step.

All of these components have been linked into a larger data transformation and publication workflow called FAIR-in-a-Box (FiaB) [16], which utilizes a custom daemon to sequentially execute each transformation step within the confines of a docker network, minimizing the exposure of any individual component to the internet, and finally loads the CARE-SM data into a GraphDB-based Triplestore. Thus, the users of CARE-SM within FiaB need only generate a CSV file in order to become FAIR data publishers.

4.2. Mapping activities using CARE-SM

CARE-SM, and its predecessor, have been used in mapping activities against other standardized data models in the clinical data semantic community. Early initiatives had the objective of schema integration and harmonization between CDE-SM and both RDF and non-RDF-based schemas. One of these initiatives, in collaboration with the Critical Path Institute, led to the creation of common SPARQL [17] queries that could map both CDE-SM and Critical Path Institute's semantic schema by leveraging a "Rosetta Stone" of shared Biolink schema concepts [18]. Other initiatives are currently under development, such as the creation of Extract, Transform, Load (ETL) workflow from CARE-SM to Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) v5.3 [19], which we hope will prove capable of transforming any encounter-based data representation represented in CARE-SM, to an equivalent Observational Health Data Sciences and Informatics (OHDSI) [20] representation.

Other initiatives focus on the creation of federated query tools. A Beacon-v2-compatible API [21] has been created for data discoverability and federation, parsing JSON-based Beacon requests into SPARQL queries executed over Triplestores. The generic adaptation of CARE-SM to the Beacon API was possible due to having a common, predictable semantic data pattern for every Beacon data filter.

5. Conclusions

Compared with its predecessor, CARE-SM simplifies many aspects of FAIR Data publishing and reuse in the clinical space. Having a single CSV template means the data provider does not have to create multiple export routines for each data element, reducing the time required to generate the data extraction layer. Moreover, this allowed the consolidation of the numerous CDE-SM YARRRML templates into a single template, enabling easier maintenance and evolution. The data model consistency achieved by reusing a single design pattern simplifies query, where the primary difference between data elements are the ontological classes that define the various sub-elements of a data type. Thus through minor adjustments to an overall SPARQL query template, any of the CARE-SM data elements can be explored in the same way. This harmonization assists the creation of toolkits and APIs around the model, for example, the implementation of the Beacon API capable of transforming non-semantic JSON calls into a set of templated SPARQL queries.

CARE-SM allows grouping, through the "context" node of RDF-Quads, of arbitrary data elements, producing linkages between multiple data models, for example, the multiple data elements that arise from a single patient encounter with the healthcare system. Few resources seem to be taking advantage (in the rare disease space) of this RDF-Quad technology, despite it being a well-documented and official W3 standard for RDF representation for about a decade.

References

- [1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, *Sci Data* 3 (2016) 160018. URL: <https://www.nature.com/articles/sdata201618>. doi:10.1038/sdata.2016.18, number: 1 Publisher: Nature Publishing Group.
- [2] P. v. Damme, P. A. Moreno, C. H. Bernabé, A. C. Ballesteros, C. M. A. L. Cornec, B. D. S. Vieira, K. J. v. d. Velde, S. Zhang, C. Carta, R. Cornet, P. A. C. Hoen, A. Jacobsen, M. A. Swertz, M. Roos, N. Benis, A Resource for Guiding Data Stewards to Make European Rare Disease Patient Registries FAIR 22 (2023) 12. URL: <https://datascience.codata.org/articles/10.5334/dsj-2023-012>. doi:10.5334/dsj-2023-012, number: 1 Publisher: Ubiquity Press.
- [3] RDF 1.2 Concepts and Abstract Syntax, ??? URL: <https://www.w3.org/TR/rdf12-concepts/>.
- [4] Clinical And Registry Entries (CARE) Semantic Model, 2023. URL: <https://github.com/CARE-SM/CARE-Semantic-Model>, original-date: 2023-10-05T12:51:43Z.
- [5] M. Dumontier, C. J. Baker, J. Baran, A. Callahan, L. Chepelev, J. Cruz-Toledo, N. R. Del Rio, G. Duck, L. I. Furlong, N. Keath, D. Klassen, J. P. McCusker, N. Queralt-Rosinach, M. Samwald, N. Villanueva-Rosales, M. D. Wilkinson, R. Hoehndorf, The Semantic-science Integrated Ontology (SIO) for biomedical research and knowledge discovery, *Journal of Biomedical Semantics* 5 (2014) 14. URL: <https://doi.org/10.1186/2041-1480-5-14>. doi:10.1186/2041-1480-5-14.
- [6] R. Kaliyaperumal, M. D. Wilkinson, P. A. Moreno, N. Benis, R. Cornet, B. dos Santos Vieira, M. Dumontier, C. H. Bernabé, A. Jacobsen, C. M. A. Le Cornec, M. P. Godoy, N. Queralt-Rosinach, L. J. Schultze Kool, M. A. Swertz, P. van Damme, K. J. van der Velde, N. Lalout, S. Zhang, M. Roos, Semantic modelling of common data elements for rare disease registries, and a prototype workflow for their deployment over registry data, *Journal of Biomedical Semantics* 13 (2022) 9. URL: <https://doi.org/10.1186/s13326-022-00264-6>. doi:10.1186/s13326-022-00264-6.
- [7] T. Weldring, S. M. Smith, Article Commentary: Patient-Reported Outcomes (PROs) and Patient-Reported Outcome Measures (PROMs), *Health Serv Insights* 6 (2013) HSI.S11093. URL: <https://doi.org/10.4137/HSI.S11093>. doi:10.4137/HSI.S11093, publisher: SAGE Publications Ltd STM.
- [8] Design Patterns · MaastrichtU-IDS/semanticscience Wiki, ??? URL: <https://github.com/MaastrichtU-IDS/semanticscience/wiki/Design-Patterns>.
- [9] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, S. Lewis, The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, *Nat Biotechnol* 25 (2007) 1251–1255. URL:

- <https://www.nature.com/articles/nbt1346>. doi:10.1038/nbt1346, number: 11 Publisher: Nature Publishing Group.
- [10] OBO Foundry, ????. URL: <https://obofoundry.org/>.
- [11] Orphanet Rare Disease Ontology - Summary | NCBO BioPortal, ????. URL: <https://bioportal.bioontology.org/ontologies/ORDO>.
- [12] RDF 1.2 N-Quads, ????. URL: <https://www.w3.org/TR/rdf12-n-quads/>.
- [13] EJP RD – European Joint Programme on Rare Diseases, ????. URL: <https://www.ejprarediseases.org/>.
- [14] YARRRML, ????. URL: <https://rml.io/yarrml/spec/>.
- [15] CARE Semantic Model Implementation, 2023. URL: <https://github.com/CARE-SM/CARE-SM-Implementation>, original-date: 2023-10-09T15:57:30Z.
- [16] FiaB: FAIR-in-a-box, 2022. URL: <https://github.com/ejp-rd-vp/FiaB>, original-date: 2022-12-12T10:34:05Z.
- [17] SPARQL 1.1 Overview, ????. URL: <https://www.w3.org/TR/sparql11-overview/>.
- [18] P. Alarcon, I. Braun, E. Hartley, D. Olson, N. Benis, R. Cornet, M. Wilkinson, R. L. Walls, Leveraging Biolink as a “Rosetta Stone” Between C-Path and EJP-RD Semantic Models Provides Emergent Interoperability, *Journal of the Society for Clinical Data Management* 3 (2023). URL: <https://www.jscdm.org/article/id/130/>. doi:10.47912/jscdm.130, number: 1 Publisher: Society for Clinical Data Management.
- [19] OMOP CDM v5.3, ????. URL: <https://ohdsi.github.io/CommonDataModel/cdm53.html>.
- [20] OHDSI – Observational Health Data Sciences and Informatics, ????. URL: <https://www.ohdsi.org/>.
- [21] J. Rambla, M. Baudis, R. Ariosa, T. Beck, L. A. Fromont, A. Navarro, R. Paloots, M. Rueda, G. Saunders, B. Singh, J. D. Spalding, J. Törnroos, C. Vasallo, C. D. Veal, A. J. Brookes, Beacon v2 and Beacon networks: A “lingua franca” for federated data discovery in biomedical genomics, and beyond, *Human Mutation* 43 (2022) 791–799. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/humu.24369>. doi:10.1002/humu.24369, _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/humu.24369>.