The Pistoia Alliance Pharma General Ontology: Experience Using LinkML in Pharma.

Joshua Valdez[1†], Philippe Rocca-Serra[2,8†*], Markus Hartmann[3], Darko Hric[1], Asiyah Yu Lin[4], Birgit Meldal[5], Peter McQuilton[6], Riccardo Mariani[7], Martin Romacker[8], Giovanni Nisato[4] and Ben Gardner[2].

[1] *Novo Nordisk, Copenhagen, Denmark*
[2] *AstraZeneca, 1 Francis Crick Avenue Cambridge Biomedical Campus Cambridge CB2 0AA UK.*
[3] *Merck Group, Germany.*
[4] *Pistoia Alliance, US.*
[5] *Pfizer, Cambridge, UK.*
[6] *Glaxo Smith Kline PLC, Stevenage, UK.*
[7] *Chiesi Farmaceutici, 26/A, Via Palermo, 43122 Parma – Italy*
[8] *Hoffman la Roche, Basel, CH.*
[9] *Oxford e-Research Centre, University of Oxford, OX1 3QG, Oxford, UK.*

Abstract

Pharmaceutical organizations generate vast amounts of data, often fragmented across multiple systems, domains, and silos. This fragmentation is exacerbated by the proliferation of ontologies, each offering its own diverse interpretations of key concepts. The complexity grows further when pharmaceutical companies, contract research organizations (CROs), regulatory bodies, and other stakeholders need to exchange information. Although the FAIR principles (Findable, Accessible, Interoperable, Reusable) promote good data management, achieving wide interoperability at scale is challenging with the current risk of creating "FAIR silos" resources compliant with the FAIR principles, but interoperable only within specific organizations or units.

To overcome these challenges, several members of the Pistoia Alliance initiated the Pharma General Ontology (PGO) project, which will identify, select and recommend a set of core vocabularies for use in relation to describing core pharmaceutical R&D concepts and deliver shared semantics, thereby, supporting interoperability across pharmaceutical domains.

The PGO will supply a set of public URIs for key R&D concepts, establishing a community-aligned, controlled vocabulary. It is hoped this shared controlled vocabulary will enable smoother data exchange and improve understanding across the pharmaceutical sector by creating a community consensus and convergence hub. Initially, the project will focus on R&D during 2024, with plans to broaden its scope to the entire medicinal product lifecycle.

Project link: https://github.com/PistoiaAlliance/Pistoia-Alliance-PGO

Keywords:

Pharmaceutical Industry, interoperability, data integration, active metadata, machine actionability, ontology, knowledge graph, FAIR

[1*] Corresponding author.
[†] These authors contributed equally.
jdnv@novonordisk.com (J.D. Valdez); philippe.rocca-serra@astrazeneca.com (P. Rocca-Serra); markus.hartmann@merckgroup.com (M. Hartmann); drhc@novonordisk.com (D. Hric); asiyah.yulin@pistoiaalliance.org (A.Y. Lin); birgit.meldal@pfizer.com (B. Meldal); pete.x.mcquilton@gsk.com (P. McQuilton); r.mariani@chiesi.it (R. Mariani); martin.romacker@roche.com (M. Romacker); giovanni.nisato@pistoiaalliance.org (G. Nisato); ben.gardner@astrazeneca.com (B. Gardner).
0000-0001-9853-5668 (P. Rocca-Serra); 0000-0003-1266-9230 (D. Hric); 0000-0003-2620-0345 (A.Y. Lin); 0000-0003-4062-6158 (B. Meldal); 0000-0003-2687-1982 (P. McQuilton); 0009-0004-9727-1647 (R. Mariani); 0000-0001-6898-0226 (M. Romacker); 0000-0002-5824-0061 (G. Nisato).

1. Introduction

The current enthusiasm around AI sweeping the industry, beyond the headlines and some success stories, underscores a critical requirement: accurate, trustworthy, and high-quality data essential for effective model training. This necessitates the assembly of well-annotated data corpora, rich in both depth and breadth. Recent advancements and community experiences with Large Language Models (LLM) and building LLM-based agents further highlights the need for capabilities to mitigate the risks associated with spurious output of such tools, commonly referred to as "hallucinations". Among the techniques available, Graph-based Retrieval Augmented Generation (GraphRAG) is gaining popularity [1, 2]. This technique relies on developing and applying knowledge graphs or neuro-symbolic artefacts used as a source of truth and "grounding reference". They therefore define an envelope of legitimate and trusted knowledge, against which the LLM agent output can be evaluated [3].

This growing recognition within business circles for accurate, trustworthy and high-quality data places the FAIR principles of data management [4] at the heart of many corporate strategies, aiming to improve the governance, stewardship, and ultimately the value-driven use of data assets.

The benefits of FAIR for the Pharma industry have been comprehensively outlined by Wise and colleagues [5] and we encourage our readership to consult this work as it is not the purpose of the present manuscript to reiterate it here. Rather, our objective is to document an approach and methodology for addressing the known risk of generating "FAiR" silos, where the lowercase "i" denotes limited, reduced or impaired interoperability. This describes the situation where interoperability is localized and thus partial, restricted to a domain within an organization or limited functional areas. The factors contributing to this "interoperability of limited span" are multifactorial. This limited interoperability can often be traced back to a range of conflicting operational constraints. For instance, regulatory pressures may mandate the use of specific resources as required by authorities, irrespective of their potential for interoperability of other artefacts. Furthermore, distinct licensing terms and usage restrictions, applied to semantic artifacts, affect their diffusion and uptake. Then, despite community initiatives to drive concerted development [6, 7, 8], independent initiatives sprout and develop resources in isolation [9], leading to competing and overlapping knowledge representations within the same domain. The resulting "offer redundancy" may, in the absence of mapping efforts [10], lead to reduced interoperability. This fragmentation of data along distinct semantic lines can significantly hinder integration efforts, creating barriers to seamless data exchange.

These observations put a particular facet of data management in a sharp light. The importance of high-quality reference data as a foundation for producing data corpora suitable for artificial intelligence applications.

Against this backdrop, the goal of the Pharma General Ontology (PGO) initiative is two-fold. First, it seeks to establish consensus on a set of core concepts, or entities relevant to the pharmaceutical R&D domain. Second, it aims to survey semantic artefacts to review and recommend a set of FAIR resources which can reliably be used to provide value-sets for key attributes of the core concepts identified in the first round. An added goal for the project is to communicate both outcomes in an explicit, machine-actionable fashion. We aim to do so using open-source framework and models devised to structure metadata profiles [11, 12, 13, 14], while following community best practices sourced from FAIRsharing such as MIRO or the FAIR cookbook content (FCB:020) [9, 15, 16]. These best practices build on the availability of persistent resolvable identifiers and the availability of infrastructure to ensure resolution and redirection [17,18].

2. Methods

The PGO project, still ongoing at the time of writing this article is governed by the PGO steering group, which defines milestones and deliverables. The PGO governance process was defined and the collaborative work among distributed actors is facilitated by a project manager supported by the Pistoia Alliance through contributions from participating organizations. To support the development, the group elected to use GitHub [19], the popular code sharing platform, taking advantage of issue tracking, wiki, deployment, continuous build features provisioned by the platform. Owing to the Pistoia Alliance bylaws, the main outcomes of the development work are shared via a private repository. Communication of public outcomes was carried out through a dedicated channel, also hosted on GitHub but in the form of a public resource. The group used teleconferencing platforms (Microsoft Teams and Zoom) during regular calls for coordinating day to day operations. Quarterly face-to-face meetings were set up to consolidate progress. The group relied on instruments such as surveys and interviews to collect user requirements and used standard "card sorting" exercises to review and prioritize concepts submitted by contributing members to the working group.

To formally structure information requirements, the group considered several well-established frameworks and specifications, with JSON schema and LinkML being ultimately selected for evaluation. LinkML (linked data Modeling Language), is a flexible data modeling language, specifically designed from the ground up to specify data schemas and mappings to vocabularies and ontologies. By using YAML to describe models, LinkML is both human-readable and machine-actionable, making it an ideal choice for managing and sharing controlled vocabularies. The formalization of data models in LinkML allows for robust instance data validation. LinkML's suite of libraries facilitates schema definition validation, automatic Python code generation (e.g., Pydantic or dataclass classes), and database schema generation (SQL) to support data persistence. Additionally, its Semantic Web-oriented stack enables near-seamless generation of JSON-LD contexts and RDF serializations, broadening compatibility with other data models. For further details, we direct readers to the LinkML documentation and learning resources available in code notebooks.

2.1. Why LinkML for the PGO?

In pharmaceutical R&D, data interoperability is challenged by the diverse ontologies and terminologies used across organizations and research domains. For the PGO to function as an effective resource, it must facilitate seamless mappings to external vocabularies, accommodating the diverse data needs of the industry. LinkML offers a structured approach to managing these complex mappings, making it a strategic choice for PGO's development. One of LinkML's core strengths is its alignment with FAIR principles, which are essential for enhancing the usability of data across different platforms. By structuring data models in a format that is both human-readable and machine-actionable, LinkML enhances the findability, accessibility, interoperability, and reusability of data resources. Its built-in support for semantic annotations and mappings to external vocabularies ensures that each concept and attribute within the PGO can be clearly defined and linked to a globally recognized identifier, creating a foundation for cross-organizational data sharing. LinkML's semantic web-oriented stack further supports interoperability by enabling automatic generation of JSON-LD contexts and RDF serialization of instance data. This compatibility with Linked Data standards allows PGO's data to be readily integrated into larger semantic networks, broadening its accessibility and applicability across diverse data systems. LinkML's use of YAML makes the language accessible to domain experts without extensive coding expertise, facilitating smoother onboarding and adoption by subject matter experts, and ensuring that PGO's framework can be managed effectively by those closest to the data.

In addition to its accessibility, LinkML allows for defining multiple representations of the same concept, each associated with unique identifiers (URIs) and links to external ontologies. This feature allows the PGO to serve as a unified reference point for concepts, while also capturing alternative

definitions from widely used biomedical ontologies. By providing this flexibility, LinkML empowers the PGO to act as a central hub for interoperable data exchange, supporting both intra-organizational and cross-organizational integration. LinkML's robust tooling also contributes to PGO's sustainability. Automated code generation, validation, and database schema creation, essential for maintaining consistent and high-quality data models while adapting to new industry requirements. This adaptability ensures that PGO can incorporate emerging concepts and align with evolving industry standards, helping to ensure its long-term relevance and scalability.

However, working with LinkML isn't without challenges that the PGO team has carefully managed. Although LinkML reduces some barriers to adoption, its integration with complex corporate infrastructures and existing ontologies requires ongoing support and customization. Additionally, as with any open-source framework, contributing enhancements and documentation back to the LinkML project from a corporate setting involves navigating legal and compliance reviews, which can slow the pace of collaborative improvements. Ultimately, LinkML provides the PGO with a structured, flexible, and FAIR-aligned framework to foster data interoperability, support industry collaboration, and serve as a sustainable resource in the pharmaceutical R&D domain.

2.2. LinkML Schema and Concept Modeling

The PGO defines each core pharmaceutical concept as a distinct LinkML class, allowing each concept to incorporate synonyms that accommodate the range of terminology used by stakeholders. This ensures consistent mappings across terminological variations. Synonyms and alternative definitions are documented with provenance information, enabling users to select definitions suited to their specific contexts.

To identify and select relevant semantic artifacts, the PGO working group conducted a survey of state-of-the-art resources, identifying potential value set sources and metadata profiles for each core concept. Concepts were selected and prioritized based on their relevance to pharmaceutical R&D workflows and, feedback from key stakeholders, ensuring that high-impact terms were addressed early in the development process.

LinkML's structure enables systematic mappings between core concepts and external vocabularies, following a standardized alignment process. While external ontologies can present challenges, particularly when definitions conflict, this approach ensures that mappings are driven by community consensus, minimizing disruption and optimizing interoperability.

To maintain relevance and adaptability, the PGO includes provisions for periodic updates, enabling the integration of new concepts and refinements to existing mappings as industry standards and needs evolve. LinkML's inherent flexibility supports seamless schema updates, allowing the PGO to scale in response to pharmaceutical advancements. These practices aim to sustain the PGO as a robust, FAIR-aligned ontology that addresses the diverse data requirements of pharmaceutical R&D.

3. Results

The Pistoia Alliance PGO project builds upon the foundational work of the FAIR project and its community of experts in 2023. This groundwork included identifying needs, defining the scope, and providing an initial list of core concepts. The first outcome of the PGO group was the identification and refinement of a set of 25 core concepts, deemed essential for the semantic representation of the Pharma R&D domain and its value chain. For each concept, the PGO project team aims to provide the following metadata: a preferred label, a definition, the source of the definition, and at least one persistent, resolvable identifier anchoring and linking the entity to an authoritative reference resource.

The group strives to obtain references from a single resource to ensure consistency and gauge which of the available resources could provide complete or near-complete coverage.

schema.org/Bioschemas, wikidata, HL7 FHIR have been appraised [7, 8, 20, 21]. The rationale for selecting these resources was the use of open-source resources, with no restriction to access to obtain semantic anchoring in resources widely used for specific application such as search engine optimization (schema.org) or data integration (wikidata, HL7 FHIR). Furthermore, for entities lending themselves to univocal instance declaration, a source of identifiers was identified. We will expand on this aspect in the following section.

3.1. Strategy Execution

The PGO's development followed a structured, multi-step approach designed to ensure comprehensive concept identification, schema development, validation, and sustainable governance. Below is a breakdown of the key steps executed by the project team:

Step 1: Concept Identification and Mapping

- **Collaboration with Domain Experts**: The team worked closely with subject matter experts to identify key pharmaceutical R&D concepts, prioritizing those with high relevance to cross-functional workflows. These concepts were mapped to widely adopted public ontologies such as **MeSH**, **ChEBI**, and **OBO Foundry** ontologies or licensed resources (**SNOMED CT),** to maximize compatibility and alignment with established industry standards.
- **Compilation of Definitions and URIs**: For each concept, definitions, URIs, and version information were compiled from selected ontology sources. This foundational reference set forms a comprehensive, consistent baseline that supports cross-referencing, interoperability, and accurate semantic alignment.

Step 2: Schema Development and Population

- **Schema Implementation**: Model each concept within the LinkML schema, allowing for the inclusion of multiple definitions and mappings to external ontologies.
- **Concept Population**: Populate the schema with the core concepts derived from expert collaboration, with attention to accommodating diverse terminology and definitions as used across pharmaceutical R&D sectors.

3.2. Core Concepts

Table 1 below lists the key concepts identified for inclusion in the PGO, each essential to covering the Research & Development domain of the pharmaceutical industry value chain. For each concept, definitions will be sourced from authoritative vocabularies or ontologies, with agreed-upon sets of preferred labels and synonyms established to support interoperability. The alphabetical list corresponds to the current version, which may be expanded in the future to cover domains beyond R&D.

Table 1: list of PGO core concept to cover the Research & Development Domain of the Pharmaceutical Industry value chain.

- Assay (Method)
- Assay (Biological)
- Biomarker
- Biospecimen
- Cell Line
- Cell Type
- clinical study
- Compound
- Equipment
- Disease
- Drug
- Gene
- Indication
- Molecular Target
- Product
- Program
- Project
- Protein
- Site
- Species
- Subject-Person
- Substance
- Target
- Unit of Measure
- Vocabulary

Each concept will be documented by instantiating a custom-built LinkML schema that serves as the backbone for representing each core concept. This schema is designed to accommodate multiple definitions and mappings to trusted external ontologies, ensuring flexibility and alignment with diverse standards. By structuring each concept within the PGO, we aim to establish a unified reference point that supports precise and consistent data exchange across pharmaceutical organizations.

In the second phase of development, each concept will be assessed to identify at least one authoritative source of information that supports the creation of value sets for either the resource itself or for key facets defining the entity.

To illustrate this point, for the simplest and least controversial case, PGO experts selected **Uniprot** [22] as the reference source for identifiers to fully qualify values instantiating an entity that would be cast as **PGO:Protein** in a document.

More complex concepts require a multi-source approach. For instance, the "Clinical Study" concept will utilize ClinicalTrials.gov [23] to provide stable, regulator-approved trial identifiers, while the CDISC glossary [24] will supply value sets for attributes like "clinical trial type" and "clinical trial phase." This multi-source approach allows the PGO to incorporate reliable data across the lifecycle of clinical studies, balancing specificity with broad applicability.

To optimize the reusability potential, PGO developers strived to not reinvent whenever possible and considered efforts such as the Biolink model [14]. Biolink modeled a number of essential bio domain entities to facilitate data collection and data integration projects. Therefore, schemas developed by PGO exploit specific LinkML elements, `close_mapping` and `exact_mapping` to provide bridges to Biolink entities when they apply. Then, PGO developers relied on the linkML 'id_prefixes', which is a mechanism that allows to list explicitly the resources that instances of this

class ought to have as part of their CURIE. This feature meets precisely the requirements defined in the PGO use cases and extensively used in the Biolink approach.

Beyond selecting sources, the team established metadata profiles to fully qualify each resource and unambiguously communicate its essential characteristics. By consulting the FAIRSharing catalog, which indexes standards, policies, vocabularies, and databases, the PGO working group identified the relevant metadata profiles. Collaboration with the Pistoia Alliance's FAIR Ontologies project further supported the selection and structuring of these metadata resources [15,16].
To summarize, two distinct LinkML schemas have been defined and one more is under development:

- A LinkML schema to describe each semantic artifact selected by the expert group, adhering to a minimal information profile. This schema is available at [public GitHub repository link].
- A LinkML schema to describe each PGO concept, enabling structured, machine-actionable definitions. This schema is also accessible at [public GitHub repository link].
- A LinkML schema for data dictionaries (currently under development), which will provide an additional layer of structure for mapping data fields and terms to PGO concepts.

This structured approach ensures that each concept in the PGO is clearly defined, annotated, and linked to authoritative sources, promoting interoperability across pharmaceutical data systems.

3.3. Experience using the LinkML library

LinkML isn't just a schema definition language; It encompasses a suite of specialized libraries, purpose-built to deliver machine actionable metadata. Machine actionability refers to the ability of metadata descriptors to be mobilized more effectively but not supplying "just strings" but "things" [25] via the association of labels with resolvable identifiers, in accordance to the FAIR principles. The framework is meant to ensure that every entity and attribute is semantically annotated. This means that in theory, instance data meeting the requirements defined by a semantically anchored schemas, can readily be serialized to RDF thus enabling data integration through the provision of a semantic graph of instance data.

While evaluating LinkML, the PGO working group identified some limitations which prevented the full execution of the RDF conversion of instance data. These restrictions have been traced not to LinkML specifications *per-se* but to limitations of library the stack relies on. Specifically, a discrepancy between features allowed by JSON-LD 1.1 specifications [26] and the implementation of said specifications by the RDFlib [27] 7.0 library. This gap restricts the flexibility needed for associating multiple URIs with a single class, impacting use cases that require broad interoperability

As we outlined in section 3.1, each core PGO entity can be tied to a *class_uri* attribute in a LinkML schema. However, the current specification does not allow for more than one CURIE per entity to be provided.

We therefore aim to provide class equivalence mapping by relying on the SSSOM mapping syntax described by Matentzoglu and colleagues (7), enabling the provision of class equivalence mappings between PGO concepts and external ontologies.

The PGO group operated as a good citizen of open science, reporting issues and documenting possible fixes, for instance https://github.com/linkml/schema-automator/issues/149.

The PGO group has plans to actively contribute to open science by documenting issues and suggesting fixes within the LinkML ecosystem, including raising issues (e.g., Schema-Automator Issue #149) and participating in discussions to advance community tools. The different industrial partners, each working with the stack, have each identified issues or suggestions to improve the framework. However, contributing back to open science projects from a corporate environment presents unique challenges. Legal and compliance barriers often require review and approval before contributions can be made publicly, adding time and procedural complexity to the process. Additionally, corporate firewalls can impede direct interaction with open-source repositories and collaboration platforms,

complicating the group's efforts to work seamlessly with the open science community. The intent is nevertheless to proactively assist the community.

To address documentation gaps and contribute further to LinkML, the PGO team plans to develop a series of Jupyter notebooks demonstrating programmatic schema generation. These notebooks will clarify method invocation, offering practical guidance for other organizations and projects and improving LinkML's accessibility and usability. By sharing these best practices and workflows, the PGO team aims to streamline onboarding for new users and provide a resource to address common challenges in pharmaceutical and life sciences data management, ultimately enhancing LinkML's utility across the industry.

4. Conclusion

The Pistoia Alliance PGO working group is keen to evaluate the approach "*in the wild*", for instance through collaboration and creation of a community of practice exploring frameworks, such as LinkML, to deliver "FAIR by Design", "FAIR at the first mile" metadata definitions and the definition of data dictionaries. With this early release and preliminary work, the PGO working group, under the Pistoia Alliance, aims to trigger the interest of major pharma players and associated service providers. This phase is an opportunity to engage with subject matter experts to review and validate content to ensure consistency and accuracy across different ontologies. It also enables interoperability testing by evaluating the resources against real-world use-cases where different systems and organizations need to exchange data to ensure interoperability and integration is achieved.

Finally, the group aims to develop a governance framework to manage updates and revisions to the ontology mappings and definitions over time, ensuring the PGO remains a living resource. This should help define a long-term sustainability plan for the PGO by establishing a process for updating mappings as external ontologies evolve.

We invite all parties to evaluate, test this initial output and provide comments and suggestions to refine or expand the list of FAIR ontologies to align against when considering data exchange, data interoperability and data integration scenarios between organizations operating in the domain, from big Pharma to service providers.

Acknowledgements

The authors thank Dr Chris Mungall for the useful discussions on LinkML. We thank Alexandra Grebe de Barron, Andrea Splendiani, Berenice Wulbrecht, Christian Senger, Erwin Weiler, Gabriel Backiananthan, Irina Filitovich, Jane Lomax, Jim Rynker, Marius Michaelis, Pablo Porras Millan, Peter Winstanley, Ping Li , Steve Penn, Thomas Liener, Tom Plasterer, Umesh Bhatt, Wendy Zimmermann for their inputs as well as all experts contributing to the Pistoia Alliance PGO program.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] Wang, X., Kapanipathi, P., Musa, R., Yu, M., Talamadupula, K., Abdelaziz, I., Chang, M., Fokoue, A., Makni, B., Mattei, N. and Witbrock, M., 2019, July. Improving natural language inference using external knowledge in the science questions domain. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 7208-7215).

[2] Matsumoto N, Moran J, Choi H, Hernandez ME, Venkatesan M, Wang P, Moore JH. KRAGEN: a knowledge graph-enhanced RAG framework for biomedical problem solving using large language models. Bioinformatics. 2024 Jun 3;40(6): btae353. doi: 10.1093/bioinformatics/btae353.
[3] Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. Nature. 2024 Jun;630(8017):625-630. doi: 10.1038/s41586-024-07421-0. Epub 2024 Jun 19.
[4] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018. doi: 10.1038/sdata.2016.18. Erratum in: Sci Data. 2019 Mar 19;6(1):6. doi: 10.1038/s41597-019-0009-6.
[5] Wise J, de Barron AG, Splendiani A, Balali-Mood B, Vasant D, Little E, Mellino G, Harrow I, Smith I, Taubert J, van Bochove K, Romacker M, Walgemoed P, Jimenez RC, Winnenburg R, Plasterer T, Gupta V, Hedley V. Implementation and relevance of FAIR data principles in biopharmaceutical R&D. Drug Discov Today. 2019 Apr;24(4):933-938. doi: 10.1016/j.drudis.2019.01.008.
[6] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ; OBI Consortium; Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007 Nov;25(11):1251-5. doi: 10.1038/nbt1346
[7] Bioschemas: URL: https://bioschemas.org/
[8] Schema.org: URL: https://schema.org/
[9] Sansone SA, McQuilton P, Rocca-Serra P, Gonzalez-Beltran A, Izzo M, Lister AL, Thurston M; FAIRsharing Community. FAIRsharing as a community approach to standards, repositories and policies. Nat Biotechnol. 2019 Apr;37(4):358-367. doi: 10.1038/s41587-019-0080-8.
[10] Matentzoglu N, Balhoff JP, Bello SM, Bizon C, Brush M, Callahan TJ, Chute CG, Duncan WD, Evelo CT, Gabriel D, Graybeal J, Gray A, Gyori BM, Haendel M, Harmse H, Harris NL, Harrow I, Hegde HB, Hoyt AL, Hoyt CT, Jiao D, Jiménez-Ruiz E, Jupp S, Kim H, Koehler S, Liener T, Long Q, Malone J, McLaughlin JA, McMurry JA, Moxon S, Munoz-Torres MC, Osumi-Sutherland D, Overton JA, Peters B, Putman T, Queralt-Rosinach N, Shefchek K, Solbrig H, Thessen A, Tudorache T, Vasilevsky N, Wagner AH, Mungall CJ. A Simple Standard for Sharing Ontological Mappings (SSSOM). Database (Oxford). 2022 May 25;2022:baac035. doi: 10.1093/database/baac035
[11] Musen MA, O'Connor MJ, Schultes E, Martínez-Romero M, Hardi J, Graybeal J. Modeling community standards for metadata as templates makes data FAIR. Sci Data. 2022 Nov 12;9(1):696. doi: 10.1038/s41597-022-01815-3.
[12] Batista D, Gonzalez-Beltran A, Sansone SA, Rocca-Serra P. Machine actionable metadata models. Sci Data. 2022 Sep 30;9(1):592. doi: 10.1038/s41597-022-01707-6.
[13] LinkML, URL: LinkML: https://linkml.io/linkml/
[14] Unni DR, Moxon SAT, Bada M, Brush M, Bruskiewich R, Caufield JH, Clemons PA, Dancik V, Dumontier M, Fecho K, Glusman G, Hadlock JJ, Harris NL, Joshi A, Putman T, Qin G, Ramsey SA, Shefchek KA, Solbrig H, Soman K, Thessen AE, Haendel MA, Bizon C, Mungall CJ; Biomedical Data Translator Consortium. Biolink Model: A universal schema for knowledge graphs in clinical, biomedical, and translational science. Clin Transl Sci. 2022 Aug;15(8):1848-1855. doi: 10.1111/cts.13302. Epub 2022 Jun 6. PMID: 36125173; PMCID: PMC9372416.
[15] Matentzoglu N, Malone J, Mungall C, Stevens R. MIRO: guidelines for minimum information for the reporting of an ontology. J Biomed Semantics. 2018 Jan 18;9(1):6. doi: 10.1186/s13326-017-0172-7.
[16] Rocca-Serra P, Gu W, Ioannidis V, Abbassi-Daloii T, Capella-Gutierrez S, Chandramouliswaran I, Splendiani A, Burdett T, Giessmann RT, Henderson D, Batista D, Emam I, Gadiya Y, Giovanni L, Willighagen E, Evelo C, Gray AJG, Gribbon P, Juty N, Welter D, Quast K, Peeters P, Plasterer

T, Wood C, van der Horst E, Reilly D, van Vlijmen H, Scollen S, Lister A, Thurston M, Granell R; FAIR Cookbook Contributors; Sansone SA. The FAIR Cookbook - the essential resource for and by FAIR doers. Sci Data. 2023 May 19;10(1):292. doi: 10.1038/s41597-023-02166-3.

[17] McMurry JA, Juty N, Blomberg N, Burdett T, Conlin T, Conte N, Courtot M, Deck J, Dumontier M, Fellows DK, Gonzalez-Beltran A, Gormanns P, Grethe J, Hastings J, Hériché JK, Hermjakob H, Ison JC, Jimenez RC, Jupp S, Kunze J, Laibe C, Le Novère N, Malone J, Martin MJ, McEntyre JR, Morris C, Muilu J, Müller W, Rocca-Serra P, Sansone SA, Sariyar M, Snoep JL, Soiland-Reyes S, Stanford NJ, Swainston N, Washington N, Williams AR, Wimalaratne SM, Winfree LM, Wolstencroft K, Goble C, Mungall CJ, Haendel MA, Parkinson H. Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. PLoS Biol. 2017 Jun 29;15(6):e2001414. doi: 10.1371/journal.pbio.2001414.

[18] Bernal-Llinares M, Ferrer-Gómez J, Juty N, Goble C, Wimalaratne SM, Hermjakob H. Identifiers.org: Compact Identifier services in the cloud. Bioinformatics. 2021 Jul 19;37(12):1781-1782. doi: 10.1093/bioinformatics/btaa864.

[19] GitHub: URL: https://github.com

[20] Waagmeester A, Stupp G, Burgstaller-Muehlbacher S, Good BM, Griffith M, Griffith OL, Hanspers K, Hermjakob H, Hudson TS, Hybiske K, Keating SM, Manske M, Mayers M, Mietchen D, Mitraka E, Pico AR, Putman T, Riutta A, Queralt-Rosinach N, Schriml LM, Shafee T, Slenter D, Stephan R, Thornton K, Tsueng G, Tu R, Ul-Hasan S, Willighagen E, Wu C, Su AI. Wikidata as a knowledge graph for the life sciences. Elife. 2020 Mar 17;9:e52614. doi: 10.7554/eLife.52614. PMID: 32180547; PMCID: PMC7077981.

[21] HL7 FHIR. URL: https://www.hl7.org/fhir/

[22] UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 2023 Jan 6;51(D1):D523-D531. doi: 10.1093/nar/gkac1052. PMID: 36408920; PMCID: PMC9825514.

[23] Gillen JE, Tse T, Ide NC, McCray AT. Design, implementation and management of a web-based data entry system for ClinicalTrials.gov. Stud Health Technol Inform. 2004;107(Pt 2):1466-70. PMID: 15361058.

[24] CDISC glossary. URL: https://www.cdisc.org/standards/glossary

[25] https://blog.google/products/search/introducing-knowledge-graph-things-not/

[26] JSON LD 1.1. URL: https://www.w3.org/TR/json-ld11/

[27] RDFlib. URL: https://github.com/RDFLib/rdflib