Journal of Biomedical Semantics

RESEARCH

Open Access

Digital evolution: Novo Nordisk's shift to ontology-based data management

Shawn Zheng Kai Tan[1], Shounak Baksi[1], Thomas Gade Bjerregaard[1], Preethi Elangovan[1], Thrishna Kuttikattu Gopalakrishnan[1], Darko Hric[1], Joffrey Joumaa[1], Beidi Li[1], Kashif Rabbani[1], Santhosh Kannan Venkatesan[1], Joshua Daniel Valdez[1*] and Saritha Vettikunnel Kuriakose[1*]

Abstract

The amount of biomedical data is growing, and managing it is increasingly challenging. While Findable, Accessible, Interoperable and Reusable (FAIR) data principles provide guidance, their adoption has proven difficult, especially in larger enterprises like pharmaceutical companies. In this manuscript, we describe how we leverage an Ontology-Based Data Management (OBDM) strategy for digital transformation in Novo Nordisk Research & Early Development. Here, we include both our technical blueprint and our approach for organizational change management. We further discuss how such an OBDM ecosystem plays a pivotal role in the organization's digital aspirations for data federation and discovery fuelled by artificial intelligence. Our aim for this paper is to share the lessons learned in order to foster dialogue with parties navigating similar waters while collectively advancing the efforts in the fields of data management, semantics and data driven drug discovery.

Keywords Ontology, Data management, Data strategy, FAIR, Pharmaceutical industry

Introduction

The increase in volume, variety, and velocity of biomedical data [1] poses challenges, rendering traditional forms of knowledge management and transfer used in science, like lab notebooks and publications, insufficient. Organisations that can successfully manage their data assets have a significant opportunity in accelerating their drug discovery pipeline [2, 3].

Findable, Accessible, Interoperable and Reusable (FAIR) data principles [4] were introduced in 2016, and have since become pervasive in discussions, policies and implementations across disciplines in scientific research. Organizations have adopted different strategies to ensure FAIR data which can be broadly divided into "FAIR@source" or "born FAIR" [3] and "made FAIR". Most cookbooks and strategies for making data FAIR follow the "made FAIR" strategy where data is "FAIRified" by a dedicated team after it has been generated [5]. In contrast, a FAIR@source strategy empowers data producers to ensure FAIRness of data at point of origin thereby ensuring that the context is accurate and factual rather than inferred. Irrespective of the strategy employed, FAIR data is pivotal in reducing time-to-value and accelerating research [6].

Despite the importance of FAIR principles being recognized widely by the research community, its adoption has proven to be a rather cumbersome task [6]. With the "made FAIR" strategy, lack of access to siloed data and the effort necessary to clean and wrangle data has remained an organizational challenge [7, 8]. Whereas, challenges for the FAIR@source approach stem from the scalability requirements for enterprise-level solutions,

*Correspondence:
Joshua Daniel Valdez
jdnv@novonordisk.com
Saritha Vettikunnel Kuriakose
szvk@novonordisk.com
[1]Novo Nordisk A/S, Novo Nordisk Park 1, Måløv 2760, Denmark

the substantial investment of time and financial resources into infrastructure, and the technical complexities associated with the implementation process.

A corner stone to ensuring FAIRness of data is rich metadata which ensures context, and as a consequence, improves Findability, Interoperability and Reusability. While controlled vocabularies and taxonomies are widely used in data FAIRification, they are not without their limitations. Ontologies, and data systems that are based on them, overcome many of these limitations, and provide a potential solution to much of the difficulties stated above - they are by design FAIR and scalable. However, ontologies are also challenging to develop, maintain, and implement. This is especially so in a Research & Early Development (R&ED) environment, where success hinges on the use of a variety of tools that are ever changing, which inevitably generates a large variety of data, much of which do not have standard structures. This difficulty is compounded in a company like Novo Nordisk, where 100 years of legacy comes with historical data management systems and practices.

In this manuscript, we describe our strategy for digital transformation in Novo Nordisk R&ED using an Ontology-Based Data Management (OBDM) strategy for bridging the gap between data producers and consumers, ensuring that context of experiments are captured in the same terminologies on both producer and consumer side, and metadata is understood regardless of time between data generation to when it is consumed, overall ensuring the best use and reuse of our data assets. We consider this publication to be among the initial discussions addressing a FAIR@source strategy and our hope for this manuscript is to offer valuable insights for FAIR practitioners and developers, shedding light on the challenges encountered by large-scale pharmaceutical companies, and by extension, various data-producing corporations. Furthermore, we aim to share the lessons learned through these experiences, fostering a broader dialogue with parties navigating similar waters while collectively advancing the efforts in the fields of semantics and data management.

Results

An ontology-based data management strategy

In Novo Nordisk R&ED, we are implementing a "FAIR@source and scale" approach to data management. The FAIR@source strategy aims to expedite the response time to scientific queries while ensuring factual context and reducing time-to-value. Simultaneously, our emphasis on FAIR@scale prioritizes scalability and the development of enterprise-level solutions. This strategic emphasis is crucial for the practical utilization of solutions within a large organization, a prerequisite particularly pertinent to global organizations that span multiple geographies. Ensuring the scalability of our data management solutions is integral to their effective implementation and utility across diverse facets of our organization.

To operationalise our "FAIR@source and scale" ambitions, we employ an OBDM strategy. In this strategy, centralized ontologies serve as the Single Source of Truth (SSOT), ensuring consistency and accuracy in data representation. Employing ontology-based structures aids in the alignment and integration of data across various domains, thus streamlining the process of adhering to the FAIR principles from the point of data creation. This approach also ensures scalability within the operational framework of Novo Nordisk. To operationalise our OBDM strategy we use a combination of ontologies, taxonomies, and controlled vocabularies, each aligned to each other, in order to deliver achievable vocabulary management, even to teams with limited or no semantic background, allowing reduction in integration overhead. (Fig. 1 shows an overview, more details of each step can be found in later sections). Enforcing use of ontologies is notoriously difficult. In order to aid in this process, we embed ontologies at source registration systems where we can. We do this through providing drop downs with preferred labels where the system registers ontology URIs in the backend (see "Delivery of controlled vocabularies" in our methods section).

Building an ontology-based data management ecosystem

Having a good approach for ontology consumption is crucial for the development of an OBDM ecosystem. As of January 2024, BioPortal [9] contains 1065 published models, the Ontology Lookup Service [10] by EMBL-EBI and Ontobee [11] contain 246 and 263 ontologies respectively. Given the large corpus of work that already exsist, one should, as far as possible, utilise existing ontologies instead of creating new reference ontologies. In the spirit of FAIR principles, reusing ontologies would lead to greater interoperability. The choices for ontologies have been extensively described in literature [12, 13]. Regardless of the choice of public ontology it is likely that none of them are fully fit for purpose for the organisation. This is not surprising as most public ontologies are built as general-purpose reference ontologies, while organisational requirements tend to be specific. In order to cater to our specific requirements, our OBDM strategy advocates for the development of organisation-specific ontologies derived from public ontologies. This approach allows for the flexibility required to cater to the needs of the organisation, while remaining interoperable with external data. We ensure that concepts specific to our organisation are parented by a public ontology which allows for easier interoperability, a strategy also used by ontology extensions [14]. In order to avoid conflicts and issues arising from redundancies, our domain models are

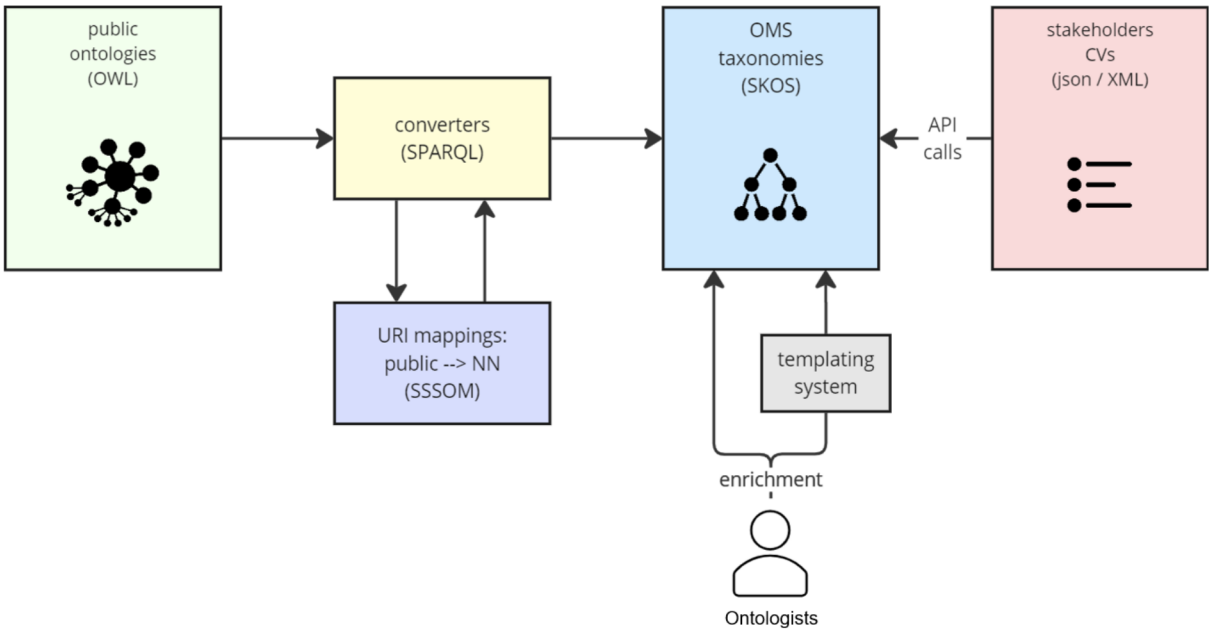

Fig. 1 General workflow diagram representing the ontology-based data management system. This diagram shows how we utilise ontologies, converting them to taxonomies in our Ontology Management Systems (OMS), and serving it up as controlled vocabularies (CVs) through Access Point Interfaces (API). Our conversion process from ontologies to taxonomies also converts public ontology URIs to in-house Novo Nordisk (NN) URIs and maps them using Simple Standard for Sharing Ontological Mappings (SSSOM)

built on selected ontologies (or subsets of ontologies). Other ontologies that are needed are either mapped in or terms in them are brought in as needed in a similar fashion to enrichments. In our strategy we do not allow duplicity of concepts to avoid difficulties in integrating data down the road.

While various strategies exist for organizations to bridge the gap between the scope of public ontologies and organisational needs, our approach was shaped by several key considerations. Through multiple iterations, we have been able to continuously learn and refine our approach. Despite the multitude of ontologies and associated resources, it became apparent that no singular representation could fully satisfy our needs across all domains of interest. With this in mind, we developed what we term "domain models" in which we defined the domains which are of interest, identified relevant ontologies, and created our own internal taxonomies based on them (described below). Another key decision was based on the OBO Foundry principle [12, 15] of orthogonality which asserts that for each domain there should be convergence upon a single ontology. Based on this, we decided against bringing in multiple full ontologies with the same scope since having multiple concepts with the same definitions would lead to difficulties in integration.

Our domain models are based either directly on a public ontology or on a composite of multiple ontologies. The latter are amalgamated to form a cohesive model through a process that involves extracting subgraphs and merging them where appropriate. This process also includes the harmonization and merging of concepts from multiple ontologies within the same domain. To ensure flexibility, we mint new URIs for our domain models. This allows us to modify logical axiomatization of ontologies or append ontology terms to other ontologies. Where it is sensible (e.g. like identifying duplicate terms, refining hierarchies, or incorporating non-proprietary terms such as anatomical parts), we prioritize pushing changes/fixing at source. This has a few strategic benefits compared to fixing in-house including reducing the burden of maintenance and ensuring that our data remains interoperable with data annotated with those ontologies. We use Simple Standard for Sharing Ontological Mappings (SSSOM) [16] systems to maintain interoperability with external sources, allowing us to update our internal ontologies, and keep in sync with public ontologies, avoiding drift. This was important to us as ontologies are not static artefacts, but models that evolve alongside knowledge. Additional benefits of using shared standards include easier utilization of community efforts like biomappings [17], and availability of open source tooling (e.g. sssom-py). From here, tools like the aforementioned biomappings and OXO [18] can help naturalise annotated external data to our ontologies. Additionally, there are semiautomated systems using

named entity recognition (NER), both commercial and open-source, can aid in the annotation of unannotated data. The specific NER tooling used is dependent on use case, team capabilities and preferences, and performance among other things.

As our upper ontology, we use BFO [19] which allows us to use reasoner-based coherency checks in the future. Deciding on an implementation for a middle/unifying ontology is a bit more complicated. For example, in the biomedical area, the OBO foundry has created an experimental unifying middle ontology, Core Ontology for Biology and Biomedicine (COB)(https://github.com/OBOFoundry/COB), that aims to bring together key terms from OBO ontologies. Work is also underway by Pistoia Alliance to develop a similar middle/upper ontology to unify high level concepts in the pharma space (termed Pharma General Ontology (PGO)) and we are actively contributing to the thought leadership underlying its construction. Federated solutions such as mapping of terms is an alternative to having unifying ontologies, and community efforts like biomappings [17] are already ongoing. However, since these unified solutions are in their infancy, we decided to take an approach interoperable with either of them. As of April 2024, we do not map to any middle ontology, but instead directly to BFO, with the knowledge that mappings can be made in the future.

Securing interoperable scientific metadata

Our FAIR@source and scale strategy relies on metadata in all applications being interoperable and served from an SSOT. We use taxonomies derived from ontologies to act as the SSOT for scientific metadata. Since ontologies are complex and difficult to maintain, we build SKOS taxonomies based on public ontologies to maintain enrichments (terms specific to Novo Nordisk) to domain models. These taxonomies only maintain annotations, hierarchical structures, and minimal relationships between concepts (as opposed to ontologies, which contain richer and more expressive relationships). This allows us a quick turnaround required in an R&ED environment. As SKOS is not as expressive as OWL, we enforce conventions in conversion where we treat skos: narrower to be equivalent to rdfs: subClassOf (an agreement among the team rather than a logical assertion). This allows back conversion to OWL/RDFS when needed for use cases that require it such as building semantically controlled knowledge graphs described in the next section. In cases where modelling has to done using individuals rather than classes, we use rdf: type to be equivalent to skos: narrower. Relationships like part_of that can be conceptualised as narrower in a taxonomy are instead left as associative relationships if they are needed to be brought in. More details and links to snippets can be found in the methods section "Conversion to SKOS taxonomies". Given that our enrichments are always parented by a concept that is derived from a public ontology, it affords us flexibility in our system. For example, if we decide to maintain OWL ontologies at a later date, our enrichments are already parented by concepts modelled as such, and conversion of enrichment to owl objects can be as rich or shallow as we choose. We however do acknowledge that utilising CVs comes with a risk of drift that may lead to issues down the road. For example, Roche utilised an internal CV which is mapped to a set of terms from the Gene Ontology (GO) for gene enrichment analysis. However, when converting them to OWL class expression, incomplete mappings of CV terms and unmappable CV terms to GO were found [20]. To mitigate this, we ensure that conversions are done in automated pipelines which allow us to easily update to new versions of ontologies– something we do on a regular basis, and as mentioned above, we ensure enrichments are parented by concepts from public ontologies. Delivery of these taxonomies to stakeholders takes the form of ontology governed controlled vocabularies (a structured list of concepts derived from the taxonomies) delivered in any form the stakeholders prefer– mostly APIs. Where changes in public ontologies that will affect users (e.g. deprecation of terms), we follow conventions of OBO ontologies (e.g. bringing in 'term replaced by' annotations) and inform our downstream users appropriately.

Integration across data silos using knowledge graphs

Given the legacy of a century-old organisation, we have diverse data sources originating from legacy and federated systems. Ensuring semantic interoperability across these silos gives us the opportunity to better leverage insights across the data landscape. One way to enable semantic interoperability is through the use of a Knowledge Graph (KG). However, the ability to integrate disparate data sources into a KG can present a formidable challenge. An effective solution to this issue can be found in the implementation of a Virtual Knowledge Graph (VKG) or Ontology-Based Data Access (OBDA) approach. In this approach, data sources such as databases are mapped to an ontology, thereby presenting a unified KG [21, 22]. Compared to materialisation, this methodology offers significant advantages, including the ability to leverage existing security measures and access controls, as the data remains in its original location, eliminating the need for duplicating and storing large volumes of data which in turn reduces storage costs and minimizes data redundancy. Additionally, this approach provides a real-time, on-demand view of the underlying data, which better supports the compliance and governance frameworks that are could be crucial in pharma. A VKG approach also provides a more scalable method for data ingestion. Maintaining scalable mappings, as

opposed to the resource-intensive processes of data materialization and constant reindexing, results in a more efficient and sustainable system. This is highly crucial in a R&ED environment is decidedly dynamic, with frequent updates and revisions, and a materialised graph can quickly become outdated or inconsistent. In essence, virtualization in the context of a semantic KG offers a flexible approach to data integration, allowing for the addition or removal of data sources with minimal impact on the overall structure of the KG and has benefits over a materialised graph in compliance, maintaining data consistency, and managing security. A challenge to having a VKG approach is that it but can struggle with high query complexity and reasoning overhead at scale. There are however several techniques to mitigate this like using Large Language Model (LLM) enabled advanced query rewriting to transform high-level semantic queries into optimized database queries [23], caching of inference results to minimize redundant computations, and partitioning complex reasoning tasks across multiple nodes. While we acknowledge that scalability remains a key challenge in handling highly complex queries, we believe that VKGs are well suited to address our primary focus is on establishing a robust conceptual framework that can integrate siloed data with fine-grained access control and the reason why this method is increasingly considered a viable alternative to traditional data integration techniques [21, 22].

Developing a KG in the biomedical domain takes a lot of time, resources, and commitment due to the variety and heterogeneity of biomedical data sources [24, 25]. Legacy and disparate sources of data, for example, require huge curation effort. We therefore have adopted the strategy of incremental improvements based on concrete use cases with stakeholders that can champion it. Our KG is built with scalability, useability, and flexibility in mind. Given that we link our data through our own internal URI, any rewiring needed can be done easily. Changing ontologies can be done simply by mapping our internal URIs to the new ontology concepts' URIs. Furthermore, if we decide to eventually maintain our own internal owl ontologies, switching can be done in very similar ways. Public ontologies are also brought in as imports in a modular fashion, this would mean that if we require slices of the ontology/KG for specific future applications, it can easily be done. All this points to a generalisable, flexible, and scalable system that fits our strategy of incremental improvements.

We take a semantic approach to our KG construction with public ontologies as its underlying structure. Our semantic KG utilises our domain models' links to public ontologies to establish a bridge between internal URIs and those of public ontologies using equivalence assertions. Figure 2 shows a diagrammatic representation of how we built our application-specific ontology. In this example, the aim was to query the graph on any "experiment" that "involves" a "chemical entity" that has_role "anti-obesity agent". The solid lines show explicit relationships between entities in our graph, while the dotted lines show inferred relationships. The inferred response to the query above is highlighted in red.

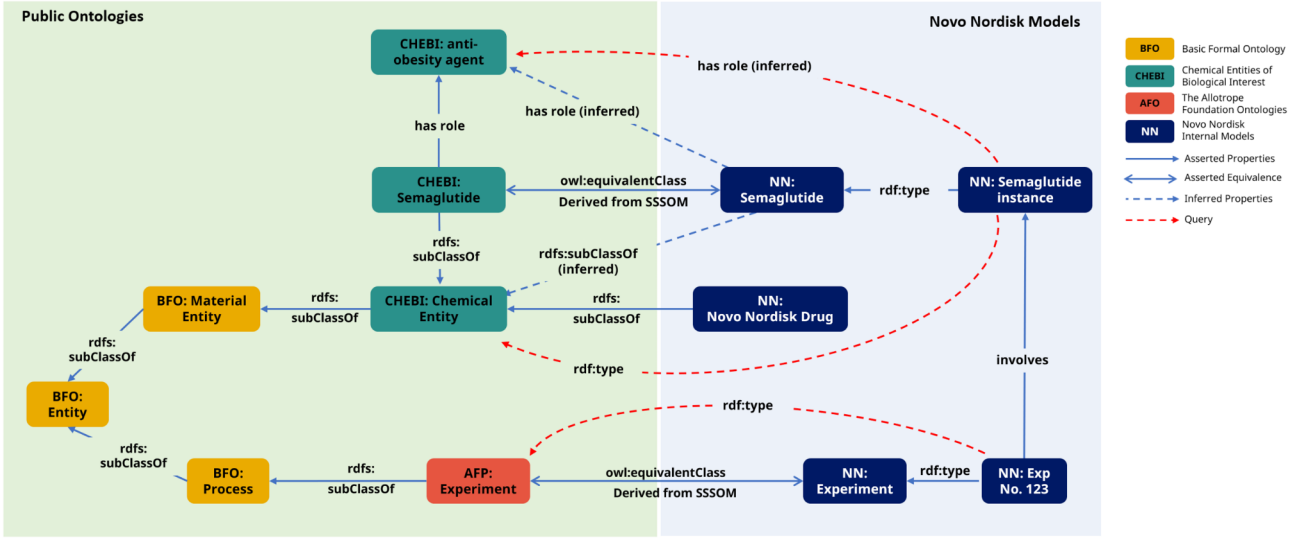

Fig. 2 Diagrammatic representation of our knowledge graph approach which illustrates the response path to the query: any "experiment" that "involves" a "chemical entity" that has_role "anti-obesity agent". The left side represents public ontologies with BFO (yellow) as the upper ontology, and CHEBI (green) and AFO (red) as ontologies of interest. Novo Nordisk specific nodes (blue) are linked either by owl: equivalentClass or rdfs: subClassOf. The solid lines show explicit relationships between entities in our graph, while the dotted lines show inferred relationships. The inferred response to the query is highlighted in red

An additional benefit of utilizing common public ontologies to underly our semantic KG is the ability to bring easily align with public initiatives. For example, the Monarch Initiative Graph integrates phenotypes, genes, and diseases [26]– something that is highly useful to work done in R&ED. Bringing in such a graph would be relatively simple given that we already utilize the ontologies underlying it, something we are already scoping. In order to optimally harness such alignment, we utilize public data models where possible. For example, for our single cell RNA sequencing data, we utilize CellXGene standards [27] and have harmonized our legacy data accordingly.

Metrics and evaluation

In order to measure the impact of the OBDM strategy on enabling FAIR@source in applications and understand the landscape of the impact, we measured uptake of CVs by the spread of departments, mapped to phases of the pharma value chain, which use our APIs (Fig. 3A) and the number of fields (headers) and values under them that we provide to applications (Fig. 3B). The spread of areas shows diversity in stakeholders and the increase from 2023 (when we first developed domain models) to 2024 shows an increase in uptake. In order to understand how our models have developed, we analyzed the number of concepts present in each of our models and broken down into what came from public ontologies, and what was enriched (Fig. 3C). We do note that these metrics are a snapshot in a single point in time (when this paper was written) and our models and stakeholders do continually evolve. These metrics should hence be taken only as a glimpse of our current landscape.

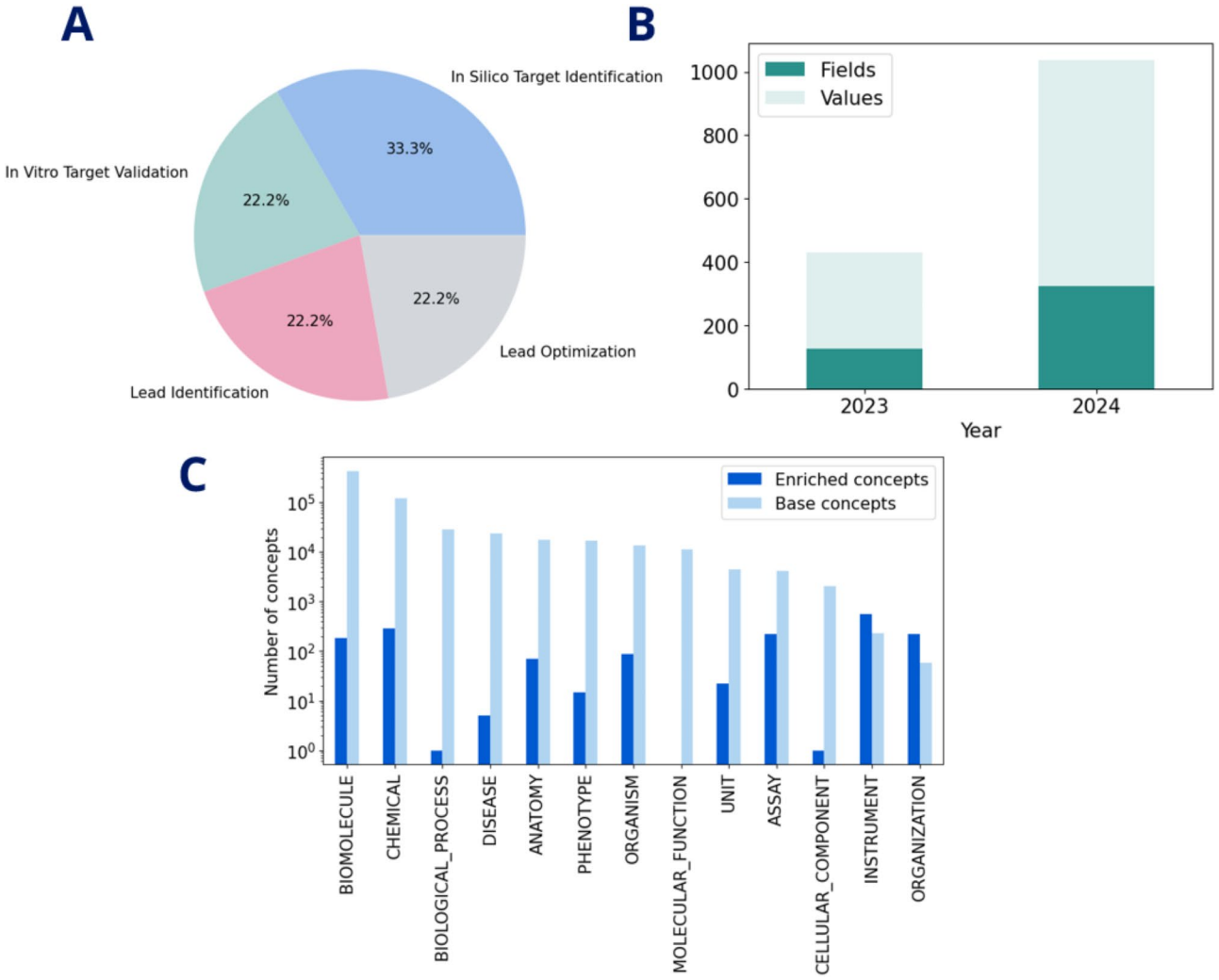

Fig. 3 Snapshot of metrics on use of controlled vocabulary and number of concepts in domain models in Novo Nordisk R&ED. (**A**) shows the spread of applications across phases of the pharma value chain. (**B**) shows the number of fields (headers) and values under them by year. (**C**) shows the number of concepts present in each of our models broken down into what came from public ontologies (base concepts), and what was enriched

Discussion

In this manuscript, we have described our strategy for building a FAIR@source and scale ecosystem based on OBDM. Although such a strategy has its merits, initiating it is notoriously challenging. Our four-year journey has underscored the importance of addressing the immediate needs of stakeholders while also planning for future use cases. Key components of an OBDM strategy include the models themselves and the capacity of downstream applications to utilize these models from a centralized source. Thus, the complexity of implementing this strategy is multifaceted and heavily reliant on collaboration across various sectors of a large organization. Inevitably, this raises the question of the value gained from investing in the development of such intricate systems, especially given the complexity, time, and costs involved. Developing an OBDM ecosystem is a strategic investment that can yield significant benefits, particularly in the realms of decentralized data architecture, semantic interoperability, and consequently, Artificial Intelligence (AI). Specifically, the integration of ontologies into modern AI architectures can considerably enhance data interpretation and utilization.

Ontologies provide a semantic context that enhances the capabilities of AI technologies, such as LLMs, machine learning algorithms, and knowledge graph-based systems. Consider the application of an LLM within the context of a drug discovery workflow. Without a systematized semantic context, the LLM may encounter difficulties in interpreting the intricate relationships between various biological entities or in disambiguating between similar entities (e.g. diseases, rare diseases, and symptoms or signs [28]). The integration of biomedical ontologies, which also function as curated knowledge bases, such as the Gene Ontology (GO) [29, 30] or the Chemical Entities of Biological Interest (ChEBI) [31], can substantially mitigate these challenges. These ontologies provide a semantic framework that equips the LLM with the necessary tools to accurately interpret complex biomedical data, thereby enhancing its accuracy and reliability.

One of the biggest challenges with the use of LLMs has been hallucinations, a common and dangerous occurrence related to the way these models operate. This highlights the need for rigorous validation processes to address these issues [25–27]. Ontologies, and semantic knowledge graphs developed from them, can function as a 'reality check' for LLMs, ensuring that outputs are both semantically and contextually accurate. An example of how this can be done is through the use of the Common Coordinate Framework (CCF) validation tool [32], which utilizes ubergraph to validate structured expert-curated tables and atlases. Such tools can be utilized to validate the accuracy of the information generated by LLMs. Additionally, the incorporation of KGs into LLMs enhances their performance by enabling Retrieval Augmented Generation (RAG) architectures, affording the ability to leverage search for information from federated sources while responding to user queries [30]. This integration, exemplified by efforts such as the Monarch Initiative [31], has demonstrated improved language model performance within the biomedical domain [33]. Generative AI can further be used to democratize the use of KGs by enabling the generation of querying languages such as SPARQL using natural language, enhancing the accessibility and usability of KGs [34]. Our OBDM strategy enables us to utilize the aforementioned benefits through a GraphRAG approach [33] which transforms natural language queries into precise semantic queries extracting not only relevant individual entities but also the interconnected relationships among them, which are then used to enrich the inputs provided to LLMs, enabling the generation process to be better informed by structured and semantically precise information. In essence, this approach preserves data provenance and traceability, ensuring that every piece of augmented information can be directly linked back to its original source. Overall, ontologies and semantic knowledge graphs play a pivotal role in reducing hallucinations and enhancing the performance of mixed generative retrieval strategies.

Beyond the advantages that an OBDM strategy brings to the AI capabilities of an organization, perhaps most notably, this approach plays a crucial role in the implementation of enterprise search. The pharmaceutical industry generates a vast array of complex and heterogeneous data, spanning from chemical and biological data to clinical trial data and patient records. Often, this data is stored in disparate systems and in various formats, posing a challenge when it comes to efficiently searching and retrieving relevant information. An OBDM strategy addresses this challenge by providing a unified view of the data, facilitating the integration of data from different sources and in different formats, thereby breaking down the silos and enhancing the accessibility and interoperability of the data. In the context of enterprise search, this means that users can search for information across different systems using a common set of terms and receive results that are contextually relevant and comprehensive. For instance, a researcher looking for information on a specific drug compound can use the same search terms to retrieve information from chemical databases, clinical trial databases, and patient records within the confines of our organization's access policies. Furthermore, the inherent structural semantics of ontologies can improve the specificity and relevance of search results. This is accomplished by leveraging the axiomatic and hierarchical relationship context between various data elements, moving beyond the limits of string-matching techniques.

By providing a unified and semantically rich view of the data, an OBDM approach supports more informed decision-making.

Finally, due to the diverse nature of data systems, the presence of a decentralized data architecture becomes essential for ensuring appropriate data ownership and governance. This approach fosters a scalable and flexible method for managing data, enabling individual teams to effectively utilize and oversee their specific data assets. A Data Mesh paradigm [34] fulfils such a need through the distribution of data across distinct domains, each characterized by its unique data product and product owner. This distribution, while advantageous in certain aspects, can introduce complexities in data integration and interoperability due to the inherent heterogeneity of data across the domains. An OBDM strategy effectively navigates these complexities by providing a unified semantic framework enabling seamless data integration across the different domains within the Data Mesh. We see the OBDM enabling the federation of data in Novo Nordisk by serving as a semantic mediator that harmonizes disparate data sources. In this context, ontology plays a pivotal role in providing a shared and common understanding of the data domain. This integration ensures that data from disparate domains can be coherently understood and leveraged for various applications in a unified manner, thereby overcoming many of the challenges posed by a distributed system. Moreover, this approach actualizes the full potential of a decentralized data architecture by enhancing data accessibility, interoperability, and usability. The OBDM strategy employs a global-as-view (GAV) strategy, where the data sources are mapped to ontologies. This ensures that data queries can be executed across multiple databases in a semantically consistent manner, thereby improving the efficiency and accuracy of data retrieval. Furthermore, the semantic relationships encoded in the ontology can be leveraged to infer new knowledge from federated data. This not only enriches the data exploration and discovery process but also enhances the expressivity and reasoning capabilities of the data federation system. The OBDM's ability to handle implicit semantics and ontological inconsistencies further strengthens its role in data federation. By resolving semantic conflicts and ambiguities, OBDM ensures the integrity and reliability of the federated data.

In conclusion, while FAIR principles have been codified, their execution has proven difficult, especially at an enterprise level in a heterogenous fast paced environment like pharma R&ED. To address this, we implemented an OBDM strategy, as discussed in this manuscript, which not only addresses our need for scale but also ensures that the data is FAIR@source. An added advantage from such an ecosystem is that it reduces time to value and accelerates data driven research and decision making. In this communication we focussed on the technical implementation of our preferred approach, but the effort required to make this happen both from a change management perspective and resource required to bring scientific knowledge to ontologists should not be discounted. Stewardship is an essential component of this playbook facilitating coordination between research scientists, data scientists, and semantic experts. Additionally, an organisational commitment at different levels starting from leadership commitment to state of the art infrastructure to bench scientists willing to collaborate to keep our knowledge bases up to date is a must. Along our four-year transformation journey, we were helped in our ambitions by the progress in technology, specifically LLMs with the consequent increased attention to semantics. The rapid progress served to underscore the need for a sustainable, scalable, and flexible data foundation which we believe is addressed by our OBDM strategy. We do however acknowledge that there are limitations in how we have implemented our OBDM strategy– practicalities of implementation, digital readiness of stakeholders, and emerging new technologies have huge influence on how we our strategy has evolved and will continue to direct the refinement of our strategy as we learn from our lessons moving forward. We hope that by sharing our journey and technical blueprint, we can foster dialogue, exchange learnings and address potential pain points. Despite the challenges we encountered along the way, we believe that the value that an OBDM ecosystem brings outweighs the challenges, and therefore is worth investing in.

Methods

Consuming public ontologies

As of April 2024, we consume 13 ontologies, either in whole or in part, to develop our 13 domain models: uberon [35], cell ontology (CL) [36], Cellosaurus (CVCL) [37], bioassay ontology (BAO) [38], ontology for biomedical investigations (OBI) [39], allotrope foundation ontology (AFO) [40], gene ontology (GO) [29, 30], protein ontology (PR) [41], Chemical Entities of Biological Interest (CHEBI) [31], Mondo Disease Ontology (MONDO) [42], human phenotype ontology (HPO) [43], NCBITaxon [44], and Quantities, Units, Dimensions and Data Types Ontologies (QUDT) [45]. These ontologies were selected based on a combination of need/use cases, and guidelines from Pistoia Alliance [46] and OBO Foundry [12]. Figure 4 shows a diagram of the public ontologies we use and the domain models they fuel.

Conversion to SKOS taxonomies

In order to convert OWL ontologies to SKOS taxonomies, we developed custom scripts for each ontology. This allows us the flexibility to ensure coherency

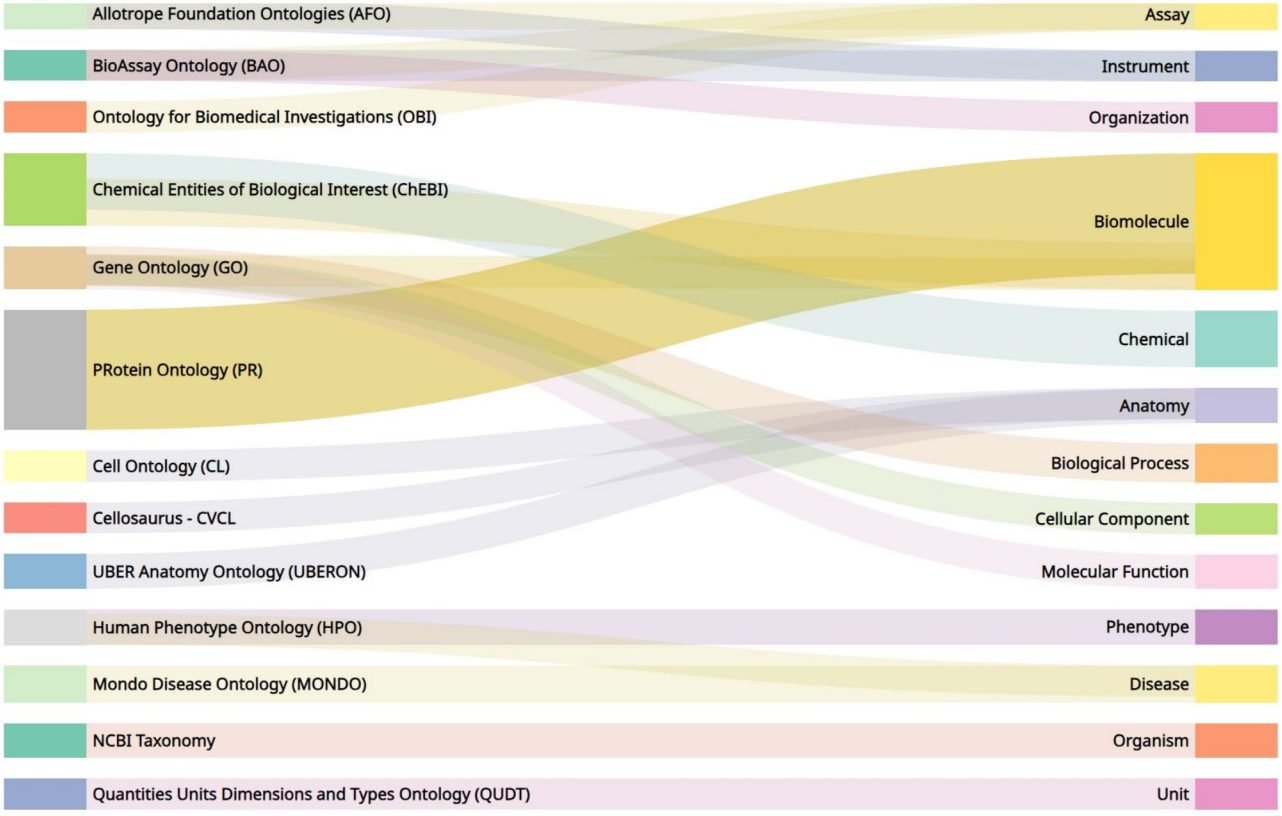

Fig. 4 Sankey diagram of the public ontologies we use (on the left) and the domain models (on the right) they are used by. While some public ontologies directly contribute to our domain models, most of our domain models are composites of different public models or parts thereof

between taxonomies converted from different ontologies. The input files to these convertors vary depending on the ontology. Snippets of codes can be found in our git repository (https://github.com/novonordisk-research/OBDM-manuscript). The convertor scripts are a series of SPARQL queries which convert owl: Class to skos: Concept (example snippet: construct-concepts.ru) and rdfs: subClassOf assertions to skos: broader (example snippet: build-heirachy.ru) and pull over selected annotations (e.g. labels, textual definition, etc.) (example snippet: mapping-metdata.ru). Owl EquivalenceAxioms are relaxed and treated similarly to rdfs: subClassOf using ROBOT [47]. During the conversion, new URIs are also minted, and the convertor outputs an SSSOM file with skos: exactMatch as the predicate_id between the new URIs and the public URIs (see https://github.com/novonordisk-research/OBDM-manuscript/blob/main/modules/replace_URIs.py).

Enrichment to taxonomies

Enrichments to taxonomies are built in a separate module from the underlying SKOS taxonomies described in the previous section, and added as skos: narrower concepts of concepts derived from a public ontology. This is done through a vendor bought centralized taxonomy management system either manually or through a proprietary templating system provided by the vendor that functions very similarly to ROBOT template [47]. Regardless, quality of all enrichments is ensured through a human internal review process. Automated SHACL validation is however in our development roadmap.

Delivery of controlled vocabularies

Controlled vocabularies are delivered to stakeholders in the form of curated dropdowns that are built in collaboration with users. We make them available via APIs from our taxonomy management system. In order to obtain terms in a given dropdown, our stakeholder's system issues pre-defined API call to our taxonomy management system which returns the terms in the requested dropdown. This eliminates the need for manual importing or updating by stakeholders, saving time and minimizing errors.

Building a knowledge graph application specific ontology

In order to construct the ontology that underpins our KGs, we leverage the links in our taxonomies to public ontologies. These links are maintained according to SSSOM standards. In order to generate equivalence class axioms between the concepts, we utilise SSSOM-py

convert function to generate skos: exactMatch annotations, and a SPARQL query to insert corresponding owl: equivalentClass axioms (example snippet: exactmatch-to-equiv.ru). To build an application ontology that underlies the KG, we perform induced subsetting from public ontologies using ROBOT [47], only bringing in the concepts we need using a Syntactic Locality Module Extractor (SLME) method. We link the subset public ontologies using BFO (denoted in yellow nodes in Fig. 1). The above described steps are designed as a scalable workflow which mimics the dynamic import system of the ontology development kit (ODK) [48] but removes the need for Docker.

Figure generation

Figures relating to metrics were generated using matplotlib [49] and source code and data can be found in our git repository (https://github.com/novonordisk-research/OBDM-manuscript).

Abbreviations

AI	Artficial Intelligence
FAIR	Findable, Accessible, Interoperable and Reusable
GAV	Global-As-View
KG	Knowledge Graph
LLMs	Large Language Models
NER	Named Entity Recognition
OBDA	Ontology-Based Data Access
OBDM	Ontology-Based Data Management
ODK	Ontology Development Kit
RAG	Retrieval Augmented Generation
R&ED	Research & Early Development
SLME	Syntactic Locality Module Extractor
SSOT	Single Source of Truth
SSSOM	Simple Standard for Sharing Ontological Mappings
VKG	Virtual Knowledge Graph

Author contributions

SZKT, JDV, and SVK conceptualised and wrote the manuscript. All authors contributed to developing the infrastructure described in the manuscript. All authors reviewed, edited, and approved the manuscript.

Funding

All authors are employees of Novo Nordisk.

Data availability

All code and data utilized in this manuscript can be found in our git repository (https://github.com/novonordisk-research/OBDM-manuscript). All other code for tools that were used are open sourced and are referenced in text. SKOS conversion scripts are highly specific, and reuse is inadvisable, however, the methodology is described in text. Given the non-reusability of such code, placing full code in public repositories would be inappropriate, hence full code will only be made available on request, within the confines of Novo Nordisk confidentiality regulations, to ensure that proper context is provided. However, snippets of key codes are provided in our git repository (https://github.com/novonordisk-research/OBDM-manuscript/tree/main/snippets).

Declarations

Ethical approval

Not applicable.

Competing interests

All authors are full time employees of Novo Nordisk. SZKT is part of the OBO Foundry Operations Committee. JDV and SVK are steering committee members for the PGO project in the Pistoia Alliance.

Received: 13 June 2024 / Accepted: 10 March 2025
Published online: 22 March 2025

References

1. Margolis R, et al. The National institutes of health's big data to knowledge (BD2K) initiative: capitalizing on biomedical big data. J Am Med Inf Assoc. 2014;21:957–8.
2. Gadiya Y et al. FAIR data management: what does it mean for drug discovery? Front Drug Discov. (Lausanne). 2023;3:1226727.
3. Harrow I, Balakrishnan R, Küçük McGinty H, Plasterer T, Romacker M. Maximizing data value for biopharma through FAIR and quality implementation: FAIR plus Q. Drug Discov Today. 2022;27:1441–7.
4. Wilkinson MD, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016;3:160018.
5. Rocca-Serra P, et al. The FAIR Cookbook - the essential resource for and by FAIR doers. Sci Data. 2023;10:292.
6. Wise J, et al. Implementation and relevance of FAIR data principles in biopharmaceutical R&D. Drug Discov Today. 2019;24:933–8.
7. Makarov VA, Stouch T, Allgood B, Willis CD, Lynch N. Best practices for artificial intelligence in life sciences research. Drug Discov Today. 2021;26:1107–10.
8. Jeliazkova N, et al. Towards FAIR nanosafety data. Nat Nanotechnol. 2021;16:644–54.
9. Noy NF, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. 2009;37:W170–3.
10. Jupp S, Burdett T, Leroy C, Parkinson HE. A new ontology lookup service at EMBL-EBI. SWAT4LS. 2015;2:118–9.
11. He Y, et al. The eXtensible ontology development (XOD) principles and tool implementation to support ontology interoperability. J Biomed Semant. 2018;9:3.
12. Smith B, et al. The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007;25:1251–5.
13. Malone J, et al. Ten simple rules for selecting a bio-ontology. PLoS Comput Biol. 2016;12:e1004743.
14. Tan SZK et al. Brain data Standards - A method for Building data-driven cell-type ontologies. Sci Data. 2023;10:50.
15. Smith B. Ontology (Sci). Derm Helv. 2008. https://doi.org/10.1038/npre.2008.2027.1
16. Matentzoglu N et al. A Simple Standard for Sharing Ontological Mappings (SSSOM). *Database* 2022.
17. Hoyt CT, Hoyt AL, Gyori BM. Prediction and curation of missing biomedical identifier mappings with biomappings. Bioinformatics. 2023;39(4):btad130.
18. Jupp S et al. OxO - A Gravy of Ontology Mapping Extracts. *International Conference on Biomedical Ontology*. 2017.
19. Arp R, Smith B, Spear AD. Building ontologies with basic formal ontology. London, England: MIT Press. 2015.
20. Osumi-Sutherland DJ, Ponta E, Courtot M, Parkinson H, Badi L. Using OWL reasoning to support the generation of novel gene sets for enrichment analysis. J Biomed Semant. 2018;9.
21. Xiao G, Ding L, Cogrel B, Calvanese D. Virtual knowledge graphs: an overview of systems and use cases. Data Intell. 2019;1:201–23.
22. Xiao G et al. Springer International Publishing, Cham,. The virtual knowledge graph system ontop. in *Lecture Notes in Computer Science*. 2020;259–277.
23. Dharwada S, Devrani H, Haritsa J, Doraiswamy H. Query Rewriting via LLMs. *arXiv [cs.DB]*. 2025.
24. Silva MC, Eugénio P, Faria D, Pesquita C. Ontologies and knowledge graphs in oncology research. Cancers (Basel). 2022;14:1906.
25. Seneviratne O et al. Knowledge integration for disease characterization: A breast cancer example. *arXiv [cs.AI]*. 2018.
26. *The Monarch Initiative in: An Analytic Platform Integrating Phenotypes, Genes and Diseases across Species*. 2024.
27. CZI Cell Science Program. CZ cellxgene discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. Nucleic Acids Res. 2025;53:D886–900.
28. Shyr C, Hu Y, Harris PA, Xu H. Identifying and extracting rare disease phenotypes with large language models. *arXiv [cs.CL]*. 2023.

29. Ashburner M, et al. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet. 2000;25:25–9.
30. Gene Ontology Consortium. The gene ontology knowledgebase in 2023. Genetics. 2023;224 (1):iyad031.
31. Hastings J et al. ChEBI in. Improved services and an expanding collection of metabolites. *Nucleic Acids Res.* 2016;44:D1214-9.
32. Caron AR, Puig-Barbe A, Quardokus EM, Balhoff JP, Belfiore J, Chipampe NJ, Hardi J, Herr BW, Kir H, Roncaglia P, Musen MA. A general strategy for generating expert-guided, simplified views of ontologies. bioRxiv. 2024. https://doi.org/10.1101/2024.12.13.628309.
33. Edge D et al. From local to global: A graph RAG approach to query-focused summarization. *arXiv [cs.CL]*. 2024.
34. Dehghani Z. Data mesh: delivering Data-Driven value at scale. O'Reilly Media. 2022.
35. Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. Uberon, an integrative multi-species anatomy ontology. Genome Biol. 2012;13:R5.
36. Bard J, Rhee SY, Ashburner M. An ontology for cell types. Genome Biol. 2005;6:R21.
37. Bairoch A. The Cellosaurus, a cell-line knowledge resource. J Biomol Tech. 2018;29:25–38.
38. Visser U, et al. BioAssay ontology (BAO): a semantic description of bioassays and high-throughput screening results. BMC Bioinformatics. 2011;12:257.
39. Bandrowski A, et al. The ontology for biomedical investigations. PLoS ONE. 2016;11:e0154556.
40. Millecam T, Jarrett AJ, Young N, Vanderwall DE. & Della Corte, D. Coming of age of Allotrope: Proceedings from the Fall 2020 Allotrope Connect. *Drug Discov. Today*. 2021;26:1922–1928.
41. Natale DA, et al. The protein ontology: a structured representation of protein forms and complexes. Nucleic Acids Res. 2011;39:D539–45.
42. Vasilevsky NA et al. Mondo: Unifying diseases for the world, by the world. *bioRxiv*. 2022 https://doi.org/10.1101/2022.04.13.22273750
43. Köhler S, et al. The human phenotype ontology in 2021. Nucleic Acids Res. 2021;49:D1207–17.
44. Schoch CL et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)*. 2020.
45. FAIRsharing Team. FAIRsharing record for: Quantities, Units, Dimensions and Types, FAIRsharing. https://doi.org/10.25504/FAIRSHARING.D3PQW7 2015.
46. Harrow I, et al. Ontology mapping for semantically enabled applications. Drug Discov Today. 2019;24:2068–75.
47. Jackson RC, et al. ROBOT: A tool for automating ontology workflows. BMC Bioinformatics. 2019;20:407.
48. Matentzoglu N et al. Ontology Development Kit: a toolkit for building, maintaining and standardizing biomedical ontologies. *Database*. baac087. 2022.
49. Hunter JD, Matplotlib. A 2D graphics environment. Comput Sci Eng. 2007;9:90–5.

Publisher's note