Visualising Time-evolving Semantic Biomedical Data

Arnaldo Pereira*, João Rafael Almeida*†, Rui Pedro Lopes‡, and José Luís Oliveira*
*DETI / IEETA, University of Aveiro, Aveiro, Portugal
†Department of Information and Communications Technologies, University of A Coruña, A Coruña, Spain
‡ CeDRI, Polytechnic Institute of Bragança, Bragança, Portugal
Email: arnaldop@ua.pt, joao.rafael.almeida@ua.pt, rlopes@ipb.pt, jlo@ua.pt

2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS) | 978-1-6654-6770-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/CBMS55023.2022.00053

***Abstract*—Today, medical studies enable a deeper understanding of health conditions, diseases and treatments, helping to improve medical care services. In observational studies, an adequate selection of datasets is important, to ensure the study's success and the quality of the results obtained.**

During the feasibility study phase, inclusion and exclusion criteria are defined, together with specific database characteristics to construct the cohort. However, it is not easy to compare database characteristics and their evolution over time during this selection. Data comparisons can be made using the data properties and aggregations, but the inclusion of temporal information becomes more complex due to the continuous evolution of concepts over time.

In this paper, we propose two visualisation methods aiming for a better description of data evolution in clinical registers using biomedical standard vocabularies.

***Index Terms*—Biomedical Data, Temporal Data, Data Visualisation, Evidence-based Medicine, Ontology Evolution**

I. Introduction

When assembling a cohort for an observational study, medical researchers need to choose and access the most suitable databases for their purposes. This task becomes simpler when database catalogues oriented to the researcher's interests are available [1]. It is even more necessary to use these resources in the case of multicentre studies, where the study objectives imply challenges in terms of research protocol development, work management, and harmonised access to data [2].

Biomedical database catalogues are increasingly in everyday use among communities of researchers who share common ground because they can take advantage of third-party data and publish their data for the benefit of all [3]. Data owners use catalogues to publish descriptive information about their databases. They provide database fingerprint information, characteristics of the database content, which include institutional details, access policies and governance rules [4].

Medical data are constantly evolving [5]. Therefore, the time dimension is of interest for the data selection process. It is useful to store information about insertions, deletions, and data changes. Historical data could increase the range of possible options in database selection, considering that it is not uncommon for biomedical data catalogues to have a log system that allows tracking data over time. However, these data are only used to verify the sanity of the solution or carry out data restoration actions in case of system disruptions [6].

Using information visualisation helps improve decision processes. Adding visualisations to biomedical database catalogues allows a better analysis and comparison of data from a single database or assessing the network level of several databases. Incorporating the concepts' temporal evolution into these two visualisation levels can mitigate possible data scarcity.

This paper proposes two types of visualisation that show the temporal evolution of semantic data. The first proposal presents the data at the database level and shows the temporal evolution of a particular selected element. The second proposal aims to visualise the temporal evolution of the data at the semantic network level. The main goal of these visualisations is to improve the use of biomedical data catalogues to help researchers make better data choices for their research studies.

II. Background

Our proposal addresses the temporal visualisation challenges from two points of view. Therefore, we conducted a background analysis, subdividing this issue into 1) standard biomedical vocabularies and 2) time-evolving semantic data.

A. Standardised medical and biomedical vocabularies

When making claims about data resources, a model is defined as a starting point for building a knowledge base. Furthermore, it is necessary to consider mechanisms that allow representing domains of interest, giving semantic meaning to the identifiers of the different resources. New concepts are added by semantic extension [7]. This strategy makes it possible to present specific logical-linguistic constructions to make assertions as unique elements for future use with a precise meaning. The most elementary form of knowledge organisation in this context involves the definition of vocabularies. A vocabulary is a set of identifiers that establish entities and relationships (referred to as terms) used to describe an area of interest.

Using standard vocabularies and ontologies, the semantic modelling of shared conceptualisations of knowledge domains allows the establishment of generic entities, concrete objects, and relationships [8]. Standardised vocabularies' reuse permits greater efficiency in semanticising new domains as mapping data to established elements reinforces the interconnection

2372-9198/22/$31.00 ©2022 IEEE
DOI 10.1109/CBMS55023.2022.00053

and takes greater advantage of the opportunities to infer new knowledge.

A dataset uses a vocabulary if a term in that vocabulary appears in the predicate position of a triple or in the object position of a triple whose predicate is `rdf:type` [9]. Regarding life sciences, several examples of standardised vocabularies can be mentioned, such as Orphanet Rare Disease Ontology (ORDO) for rare disease data annotation [10], Gene Ontology (GO) for describing genes [11], and the Human Phenotype Ontology (HPO) vocabulary for phenotypic abnormalities [12].

B. Time-evolving semantic data

Temporal data management allows querying, access and navigating through different data versions to understand their evolution or choose pieces of information from a given moment that are more suited to the user's interests [13]. Within the scope of relational databases, the temporal dimension is considered by creating specialised data structures to optimise accesses. The same principle is valid for pure graph databases. For instance, Khurana and Deshpande [14] proposed a historical graph store for large scale volumes of data integrating a new temporal graph index and a temporal graph analysis framework to perform complex temporal analytical tasks.

Time coding strategy can be divided into copy systems or log systems [15]. With each change, the updated full copy of the data is saved in the copy approach. In the log approach, the first complete version of the data is kept, and changes are recorded in a log. Hybrid systems that consider both approaches can also be adopted.

Querying data that evolves can follow alternative patterns [16]. The first considers a time interval and extracts valid entities for that time interval. Another querying approach takes a time interval and a set of entities to retrieve those entities' temporal evolution. Finally, we can take just a collection of entities and check their entire evolutionary history.

The visualisation of semantic data considering the temporal dimension allows observing patterns and determining when there is a greater concentration of entities of interest. Time-oriented data visualisation techniques can be classified from the arrangement point of view as linear or cyclic and from the time primitives point of view as instant oriented or interval [17].

The discovery and study of patterns are facilitated when using time curves, violating the linearity of the spatial provision of the most usual timelines. Bach *et al.* [18] consider a non-linear time tape curving according to data similarity at each moment, with the advantage of being possible to ascertain the depth of the changes to the concepts.

III. Database-level visualisation

Semantic database descriptions are guided by ontologies established by the community and allow the creation of database catalogues. Semantic data related to each database considered separately allows for studying the adequacy of data used in any particular research. Using data relating to the evolution of concepts over time requires specialised technical solutions. The system must be able to store the data in an acceptable way to take into account the temporal dimension. In addition, it must be possible to visualise the data at the database level, considering the temporal dimension as one more element that allows researchers to make better decisions when choosing data for their studies.

A. Temporal knowledge bases

It is necessary to define a data structure to capture the temporal evolution of the entities and relationships of an ontology. We start by defining the concept of a knowledge graph, and then add the time dimension.

Definition 1. *A Knowledge Base (KB) is an edge labelled multi-digraph $K = (V, E^*)$ that is defined by a node set $V = V_1 \cup V_2$ and a labelled arc set $E^* = \{(v_1, l, v_2) : v_1 \in V_1, v_2 \in V_2, l \in L\}$, l being an element of the label set L.*

Adding the time dimension to this data model, we get the following definition.

Definition 2. *A Temporal Knowledge Base (TKB) is a triple of the form $K = (V, E^*, T)$, with V and E^* as defined before, and a set $T = T_i \times T_f$ of timestamps.*

The ontological concepts and individuals constitute the set of vertices. Two timestamps are associated with each entity. The first timestamp, $t_i \in T_i$, records the moment of inserting the element in the KB. The second timestamp, $t_f \in T_f$ reports the moment of concept evolution (removal or alteration). Thus, a timespan is implicitly defined, useful for applications, namely when discussing visualisations.

Following FAIR principles, removing a concept does not determine its exclusion from the TKB. The TKB arcs indicate the relationships between the different KB entities. As a simplification, we can consider the temporal dimension only for concepts, thus excluding individuals. The relationships that make up the KB arcs are also temporally annotated. Relationships are only marked when they are inserted to avoid inconsistencies. Thus, the evolution of a relationship generates a new relationship that does not affect the triples previously entered.

B. Visualisations of temporal knowledge bases

When viewing temporal databases, navigation between different moments in time must be fluid because users need to perceive the existence of a timeline. The chart has a timeline that becomes visible by choosing each selectable visual element of the visualisation. Each choice of a given moment corresponds to the historical evolution of the node to which that timeline corresponds. Naturally, some concepts are static because they have never changed since insertion into the knowledge base or because there was a gap in the collection of historical data. It is also necessary for each element to show the past and subsequent states closest in time to the displayed highlighted state.

Figure 1 shows a graph-level representation of a single database. Each concept is selectable, opening a highlighted

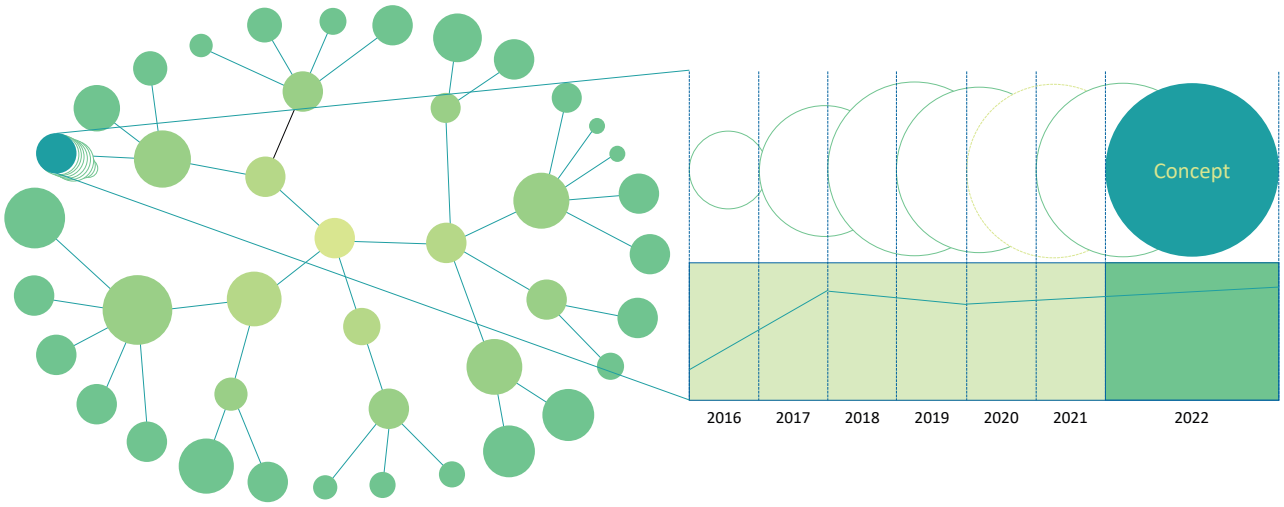

Fig. 1. Database-level visualisation UI mockup proposal.

view where a timeline is displayed. Using this line, we can choose between different moments in time to better study the state of concept in that version of the data. In this way, researchers improve their process of selecting data because the possibilities of choice are greater when considering the different versions stored in the system.

As new concepts are added, modified or removed from the ontology, the different versions that document these changes are saved and serve as a basis for visualisations. Visualisations that do not consider other moments in time aggregate all the information, making the presentation of different concepts confusing. With the proposed visualisation, we can navigate the graph and discover temporal details for each node individually, which helps to have a clearer perception of the data.

IV. Network-level visualisation

In collaborative scenarios, where multiple entities are responsible for storing data in their own facilities, the information is distributed among those involved. In some scenarios this is a forced practice due to data sensitivity, namely when dealing with medical records. However, as shown in other work [2], [19]–[21], it is essential to adopt distributed data to conduct more impactful studies.

In multicentre studies, one crucial step for the study's success is the dataset selection. Another issue when performing a multicentre study is correct definition of the vocabularies used to define the inclusion and exclusion criteria. Although there are several standard vocabularies for characterising medical concepts, as described previously, different institutions may adopt different vocabularies or different versions of the same vocabulary. Therefore, for situations in which it is important to have an overview of the data stored in all databases, we proposed a network-level visualisation with comparative and selective features.

Figure 2 shows the network view, where researchers can access temporal information when selecting a particular concept. By doing so, the time data view already explained for the database-level visualisation is displayed.

A. Concept network coverage

When conducting medical studies, the amount of data affects the quality of the results. A simple visualisation to learn the number of database records is to use a table, as seen on the left of Figure 2. In multicentre studies, it is essential to understand the study's feasibility before recruiting the datasets to support it. This view can easily provide an overview of the number of records in each database, even when some of them do not have the concepts defined for inclusion criteria in the study.

Using a tabular view, researchers can inspect row by row or column by column to see the greater or lesser density of the data, that is, to understand how many records of each concept there are in each database. However, this form of visualisation is not the simplest way to make comparisons between different databases. Depending on the study, sometimes it is necessary to identify one database to be used for analysis and others for validation. An example of these cases is patient-level prediction studies, in which one database is usually chosen to train machine learning models. These are tested and validated using other databases. Knowing the number of samples for the concepts in the study helps determine which databases should be used to train the models.

The comparison of admissibility of one database in contrast to others must be checked considering all the concepts. The tabular view puts all data at the same level and does not allow for a hierarchical view facilitating comparisons. This problem is solved by complementing this view with another presenting data in a hierarchical network, as described in the following subsection.

Concept	Database A	Database B	Database C	Database D
3671655	2.4k			
4553810	50k	30k		
3671560	3.6k			
1411491			304k	
3708323		12.7k		
4561632		23.6k		55k
4561635			271.1k	
1411500			43k	
1029193			398.6k	
3708331				152.9k
3296466				73.4k

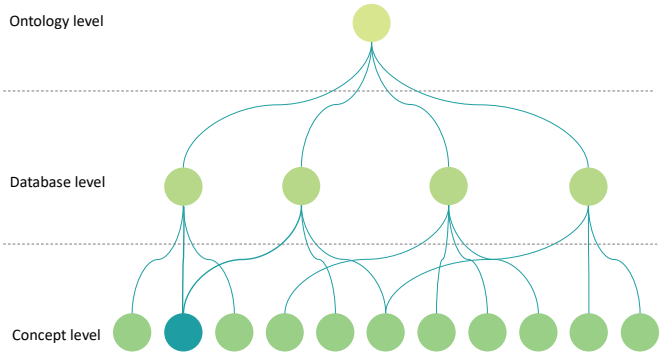

Fig. 2. Network-level visualisation UI mockup proposal.

B. Browsing at the concept level

The network view shown on the right of Figure 2 allows navigating the concepts of the different vocabularies. Selecting a particular node causes it to be highlighted. Also highlighted are its connections with other nodes. This navigation is important to simplify the database selection process. With the huge number of vocabularies and the complexity involved due to the existence of many connections between concepts and databases, navigating the tree can help researchers to identify which concepts are under a specific domain and which ones exist in the databases of the network.

For simplicity, Figure 2 presents the nodes spread over three core layers of data types, but many more levels could have been considered. The root is the level of the ontology from which all other terms derive. At the database level, we find entries for more detailed information on each of them. Finally, at the level of the leaves, we have the concepts themselves. Arcs model the relationships between the different nodes. For example, a quick inspection of the dendrogram connections shows that concept 4553810 exists in databases A and B.

Although we omitted the filters in this representative figure, researchers need to search for the concepts or domains they want to use in the study. The dendrogram is then updated based on the filter applied. Several sub-layers of relations may appear between the database and concept level layers, depending on the concepts searched for. Only some nodes at the concept level are presented to avoid overloading the visualisation. It is possible to focus attention on particular concepts, choosing a node and activating it to navigate to related nodes, namely in cases where the same concept has different identifiers in different vocabularies.

There are concepts with an enormous diversity of child concepts, so it is possible to obtain more details visually using a view like these. For example, in the SNOMED vocabulary, the concept “Aspirin” has more than 700 child concepts when considering the relation “Specific active ingredient of”. The table supports the dendrogram because the profusion of child concepts makes the visualisation more challenging to use due to the number of leaf nodes.

Comparing databases enables informed choices about the data to use in medical studies. The possibility of visually comparing the contents of different data sources is a bonus. When using tables, as is currently done in several database catalogues, it is challenging to identify the databases of interest. Considering graphs to represent semantic data, researchers can make comparisons and observe hierarchies easily and intuitively.

V. Discussion

Visualisations of semantic data with temporal information allow evaluating the evolution of these concepts and orient database selection to carry out medical research. In this section, we present examples of the application of our visualisation proposals and point out some open challenges.

A. Research application

Observational Health Data Sciences and Informatics (OHDSI)[1] initiative [19] is an international, interdisciplinary, multi-stakeholder project to develop applications to access and analyse large-scale observational health data. The core of this project lies in adopting a common data model for the treatment of health data. Solutions to extract, transform, and load data from different sources in the proposed standard format are available, as well as other tools related to the data modelling process.

Observational Medical Outcomes Partnership (OMOP) promotes the proper use of observational healthcare databases [22]. OMOP Common Data Model (CDM)[2]has been proposed as an open relational data model standard designed to establish the structure and content of observational health data. OMOP CDM allows the creation of relational databases to load transformed data from other sources of information. In this data schema, a set of tables was defined to store the standard vocabularies in an interoperable structure. These tables can represent each vocabulary and all the information associated with it. This is essential to ensure database interoperability in multicentre studies when using institutions that adopt different vocabularies in the original data. The

[1]https://www.ohdsi.org/
[2]https://ohdsi.github.io/CommonDataModel/cdm54.html

tables are defined in the collection denominated "Standardised Vocabularies".

ATHENA[3] (which stands for "Automated Terminology Harmonisation, Extraction and Normalisation for Analytics) is a standard vocabulary repository based on an automated building process. ATHENA allows keyword searching for terms using filters to select the application domain (drugs, conditions, procedures, devices, observations, and measurements), type of concepts (for classification, standard or not), class, vocabulary and validity. The search results are presented in a tabular format, and it is possible to browse the terms shown for those lying inside a hierarchy.

Inspired by the principles of aggregating multiple vocabularies on a centralised platform and due to the existence of several medical vocabularies based on semantic ontologies, we proposed these two visualisations to simplify data exploration and analysis. The first allows visualising the temporal evolution of vocabulary or ontology concepts, enabling researchers to see the impact of concept insertion, modification and removal operations. The second view allows comparing different databases by observing the number of records for each concept.

Visualisations at the database level allows assessing which concepts are present and understanding how they are related. Analysing the evolution of a given concept can enable researchers to simplify the process of choosing data. The main limitation of this type of visualisation is that it does not allow comparisons between different databases. Another limitation relates to concepts that have changed but for which historical data have not been correctly recorded. We can mention a couple of applications for this type of visualisation. As an example of an application, Esteban-Gil *et al.* [23] carried out a study that could have benefited from this visualisation to select data from a repository on cancer patients containing temporal information.

The most significant advantage of network-level data visualisation is the possibility of making comparisons between different databases. Given that the same concept has various records depending on the database, the proposed visualisation means researchers can speed up the database selection process. The most significant limitation of this visualisation lies in handling large volumes of data, which requires adequate filters to avoid visual display saturation. Pointing to a use case, we can refer to the work of Reps *et al.* [24]. They reported a multicentre study based on data from multiple health care databases that could have benefited from network-level visualisations. In the case of this study, the objective was to reproduce patient-level prediction results for type 2 diabetes and dementia.

B. Challenges and opportunities

The proposed time-evolving semantic data charts present advantages for researchers carrying out medical studies that depend on careful database selection. However, there are some challenges in implementing and adopting the proposed visualisations.

Concept evolution characterisation is challenging because it implies keeping a succession of states to be able to trace this evolution. It also means the need to compare different versions of the same ontology. We are faced with this scenario when evaluating data at the level of a single database. At the database level, it would be desirable, for two versions of the ontology, to see the operations of adding and deleting semantic entities. Visualisation of these two basic operations becomes complex when we overlap all the ontology elements. Cardoso *et al.* [25] solve this problem by building a historical knowledge graph that collects data related to all critical semantic operations: add, delete, split, move concepts, relationships or attributes. However, this problem still lacks an adequate solution, and its resolution would allow more informative visualisations.

Network-level data visualisation suffers from problems when the number of entities considered is very high. Visualising a considerable number of concepts and relationships is messy and uninformative. This happens when a node is connected to many concepts at a lower level in a hierarchical structure. In this case, it is necessary to have filtering mechanisms that highlight the most relevant information and hide or place the non-relevant information with less prominence. An approach that promises good results in solving this problem involves using supervised machine learning techniques considering node embeddings [26]. Another method is to consider layering and data separation to highlight the relevant information [27].

The same concept can be coded in different ways in different vocabularies. There is no single authoritative form or universal formulation accepted by all researchers. Also, within the same vocabulary, concepts can evolve, considering their different versions. Regarding semantic heterogeneity, several concept alignment techniques have been investigated, such as establishing new metrics for calculating semantic distances or new disambiguation approaches using generic knowledge bases [28]. Another way to tackle this problem is to consider common data models and mapping rules for extracting, transforming and loading as proposed by OHDSI partners.

Choosing medical and biomedical databases is crucial in conducting multicentre medical studies. Given that there is not always a sufficient amount of data to guarantee the quality of studies, it is essential to design strategies that increase the possibilities of choice. When using database catalogues, it is possible to access a varied set of historical data to help decide which data to choose and include in the study. Visualisations support decision-making processes and are valuable for selecting databases from catalogues. Our proposal allows visualisations at the database and data network levels, taking the time dimension into account to observe the evolution of vocabulary concepts. With this proposal, it is hoped that researchers can streamline their choices of databases.

[3]https://athena.ohdsi.org/

VI. Conclusion

Data scarcity is a drawback when conducting observational medical studies. Considering historical data from the concept evolution of biomedical vocabularies can expand the range of data choices when using database catalogues. Information visualisation mechanisms are needed to facilitate decision-making, allowing for a more detailed view of the evolution of concepts in a database. It is also essential to compare different databases using the most appropriate data depictions.

The objective of this work was to propose two visualisation strategies for time-evolving semantic data. The first view allows studying the evolution of individual database concepts. The second offers a solution to explore the entire data network that considers all the databases described. The proposed visualisations were considered within the scope of a catalogue of standard vocabularies, with the purpose of using them in choosing databases for medical studies.

Acknowledgements

This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the Project Scope: DSAIPA/AI/0088/2020. Arnaldo Pereira and João Rafael Almeida are funded by the FCT – Fundação para a Ciência e Tecnologia (national funds) under the grants PD/BD/142877/2018 and SFRH/BD/147837/2019 respectively.

References

[1] M. Sequeira, J. R. Almeida, and J. L. Oliveira, "A comparative analysis of data platforms for rare diseases," in *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, 2021, pp. 366–371.

[2] J. R. Almeida, L. B. Silva, I. Bos, P. J. Visser, and J. L. Oliveira, "A methodology for cohort harmonisation in multicentre clinical research," *Informatics in Medicine Unlocked*, vol. 27, pp. 1–9, 2021.

[3] L. B. Silva, A. Trifan, and J. L. Oliveira, "MONTRA: An agile architecture for data publishing and discovery," *Computer methods and programs in biomedicine*, vol. 160, pp. 33–42, 2018.

[4] J. L. Oliveira, A. Trifan, and L. B. Silva, "EMIF Catalogue: a collaborative platform for sharing and reusing biomedical data," *International journal of medical informatics*, vol. 126, pp. 35–45, 2019.

[5] E. L. Siegler, "The evolving medical record," *Annals of Internal Medicine*, vol. 153, no. 10, pp. 671–677, 2010.

[6] T.-c. Chiueh and D. Pilania, "Design, implementation, and evaluation of a repairable database management system," in *21st International Conference on Data Engineering (ICDE'05)*, 2005, pp. 1024–1035.

[7] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowledge Acquisition*, vol. 5, no. 2, pp. 199–220, 1993.

[8] R. d. A. Falbo, G. Guizzardi, and K. C. Duarte, "An ontological approach to domain engineering," in *Proceedings of the 14th International Conference on Software Engineering and Knowledge Engineering*, 2002, p. 351–358.

[9] M. Schmachtenberg, C. Bizer, and H. Paulheim, "Adoption of the linked data best practices in different topical domains," in *The Semantic Web - ISWC 2014*, 2014, pp. 245–260.

[10] S. Weinreich, R. Mangon, J. Sikkens, M. Teeuw, and M. Cornel, "Orphanet: A european database for rare diseases," *Nederlands Tijdschrift voor Geneeskunde*, vol. 152, no. 9, pp. 518–519, 2008.

[11] The Gene Ontology Consortium, "Expansion of the gene ontology knowledgebase and resources," *Nucleic Acids Research*, vol. 45, no. D1, pp. D331–D338, 2016.

[12] S. Köhler, N. A. Vasilevsky, M. Engelstad, E. Foster, J. McMurry, S. Aymé, G. Baynam, S. M. Bello, C. F. Boerkoel, K. M. Boycott, M. Brudno, O. J. Buske, P. F. Chinnery, V. Cipriani, L. E. Connell, H. J. Dawkins, L. E. DeMare, A. D. Devereau, B. de Vries, H. V. Firth, K. Freson, D. Greene, A. Hamosh, I. Helbig, C. Hum, J. A. Jähn, R. James, R. Krause, S. J. F. Laulederkind, H. Lochmüller, G. J. Lyon, S. Ogishima, A. Olry, W. H. Ouwehand, N. Pontikos, A. Rath, F. Schaefer, R. H. Scott, M. Segal, P. I. Sergouniotis, R. Sever, C. L. Smith, V. Straub, R. Thompson, C. Turner, E. Turro, M. W. Veltman, T. Vulliamy, J. Yu, J. von Ziegenweidt, A. Zankl, S. Züchner, T. Zemojtel, J. O. Jacobsen, T. Groza, D. Smedley, C. J. Mungall, M. Haendel, and P. N. Robinson, "The human phenotype ontology in 2017," *Nucleic Acids Research*, vol. 45, no. D1, pp. D865–D876, 2016.

[13] M. Kaufmann, A. A. Manjili, P. Vagenas, P. M. Fischer, D. Kossmann, F. Färber, and N. May, "Timeline index: A unified data structure for processing queries on temporal data in sap hana," in *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, 2013, p. 1173–1184.

[14] U. Khurana and A. Deshpande, "Storing and analyzing historical graph data at scale," in *Proceedings of the 19th International Conference on Extending Database Technology*, 2016, pp. 65–76.

[15] M. H. Böhlen, A. Dignös, J. Gamper, and C. S. Jensen, "Temporal data management - an overview," in *Proceedings of the 7th European Summer School on Business Intelligence and Big Data (eBISS)*, 2017, p. 51–83.

[16] B. Salzberg and V. J. Tsotras, "Comparison of access methods for time-evolving data," *ACM Computing Surveys*, vol. 31, no. 2, p. 158–221, 1999.

[17] W. Aigner, S. Miksch, H. Schumann, and C. Tominski, *Survey of Visualization Techniques*, ser. Human–Computer Interaction Series. London: Springer, 2011, ch. 7, pp. 147–254.

[18] B. Bach, C. Shi, N. Heulot, T. Madhyastha, T. Grabowski, and P. Dragicevic, "Time curves: Folding time to visualize patterns of temporal evolution in data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 559–568, 2016.

[19] G. Hripcsak, J. D. Duke, N. H. Shah, C. G. Reich, V. Huser, M. J. Schuemie, M. A. Suchard, R. W. Park, I. C. K. Wong, P. R. Rijnbeek, J. v. d. Lei, N. Pratt, G. N. Norén, Y.-C. Li, P. E. Stang, D. Madigan, and P. B. Ryan, "Observational Health Data Sciences and Informatics (OHDSI): Opportunities for observational researchers," *Studies in Health Technology and Informatics*, vol. 216, pp. 574–578, 2015.

[20] G. Hripcsak, P. B. Ryan, J. D. Duke, N. H. Shah, R. W. Park, V. Huser, M. A. Suchard, M. J. Schuemie, F. J. DeFalco, A. Perotte *et al.*, "Characterizing treatment pathways at scale using the OHDSI network," *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7329–7336, 2016.

[21] J. R. Almeida, O. Fajarda, A. Pereira, and J. L. Oliveira, "Strategies to access patient clinical data from distributed databases," in *HEALTHINF*, 2019, pp. 466–473.

[22] P. E. Stang, P. B. Ryan, J. A. Racoosin, J. M. Overhage, A. G. Hartzema, C. Reich, E. Welebob, T. Scarnecchia, and J. Woodcock, "Advancing the science for active surveillance: Rationale and design for the observational medical outcomes partnership," *Annals of Internal Medicine*, vol. 153, no. 9, pp. 600–606, 2010.

[23] A. Esteban-Gil, J. Fernandez-Breis, and M. Boeker, "Analysis and visualization of disease courses in a semantic enabled cancer registry," *Journal of Biomedical Semantics*, vol. 8, pp. 1–16, 2017.

[24] J. Reps, P. Ryan, P. Rijnbeek, and M. Schuemie, "Design matters in patient-level prediction: evaluation of a cohort vs. case-control design when developing predictive models in observational healthcare datasets," *Journal of Big Data*, vol. 8, pp. 1–18, 08 2021.

[25] S. D. Cardoso, M. Silveira, and C. Pruski, "Construction and exploitation of an historical knowledge graph to deal with the evolution of ontologies," *Knowledge-Based Systems*, vol. 194, p. 105508, 2020.

[26] M. Kulmanov, F. Z. Smaili, X. Gao, and R. Hoehndorf, "Semantic similarity and machine learning with ontologies," *Briefings in Bioinformatics*, vol. 22, no. 4, pp. 1–18, 2020.

[27] G. E. Marai, B. Pinaud, K. Bühler, A. Lex, and J. H. Morris, "Ten simple rules to create biological network figures for communication," *PLOS Computational Biology*, vol. 15, no. 9, pp. 1–16, 2019.

[28] P. Shvaiko and J. Euzenat, "Ontology matching: State of the art and future challenges," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 1, pp. 158–176, 2013.