Ontology-based Approaches to Medical Data Integration

Ines Obradovic[1], Mario Milicevic[1], Boris Vrdoljak[2], Krunoslav Zubrinic[1]

[1] Department of Electrical Engineering and Computing

University of Dubrovnik

Dubrovnik, Croatia

[2] Faculty of Electrical Engineering and Computing

University of Zagreb,

Zagreb, Croatia

ABSTRACT

Medical data come in a variety of forms and are stored in different types of databases. Some are structured relational database stores, while others are semi-structured and unstructured, such as the emerging NoSQL data stores. A lot of valuable data can be found in the form of unstructured text, such as clinical notes and discharge letters. To analyze and discover hidden patterns and extract knowledge from the data, they should be integrated. In the field of medicine, many ontologies have been created to provide a common basis for information exchange and to improve semantic interoperability. In this paper, we provide an overview of ontology-based integration approaches for various sources of medical data. We also identify current challenges and provide directions for future research.

Keywords: Medical data, Ontologies, Ontology-based data integration

INTRODUCTION

With the widespread adoption of electronic health records (EHRs) (Caceres, 2013) in hospitals and other medical facilities, a tremendous volume of medical data has already been produced and will continue to grow. EHRs are intended to exchange information between health care providers and organizations, though, the detailed clinical data they contain are often stored in proprietary formats with nonstandard codes and structures.

Data in electronic health records can be entered in a variety of formats. Structured data typically includes basic information, such as patient demographics; numeric values, such as height, weight, blood pressure; categorical values, such as blood type; and health data with a standardized code system, such as SNOMED and ICD-10 codes. However, information such as those in clinical notes, pathology and radiology reports, admission and discharge summaries is commonly written in the form of unstructured or semi-structured text. Additionally, medical data are stored in different types of databases. A large portion of the data is stored in traditional structured relational databases, although, medical data can also be found in semi-structured and unstructured stores, such as the emerging NoSQL data stores which are created to meet the demands of the rapidly growing data volumes and to accommodate data for specific use cases where relational databases have proven inadequate (Chen and Lee, 2019).

Although traditional analytics typically uses structured data that are easily accessible, a lot of valuable insights can be found in unstructured free text. Accessing and understanding the unstructured data is not as straightforward. Health systems must use sophisticated technologies, such as natural language processing (NLP), to derive value from a large amount of everyday language. To comprehensively analyze the unstructured and structured data, discover hidden patterns, and extract knowledge from them, they need to be integrated. Three types of heterogeneity should be considered in the integration process. In addition to structural (schematic) heterogeneity, medical data from multiple sources are also characterized by syntactic (format) and semantic (meaning) heterogeneity. Conventional approaches to heterogeneous database integration are not able to fully solve the integration of multiple sources because they cannot efficiently deal with problems of semantic heterogeneity (Asfand-E-Yar and Ali, 2020).

Ontologies, as formal models of knowledge representation with explicitly defined concepts and named relationships linking them (Gruber, 1993), can be used to overcome the problem of semantic heterogeneity between data sources (Wache et al., 2001). In this paper, we present the basics of ontology-based integration approaches for various sources of medical data.

The paper is organized as follows. In Section 2, we introduce medical ontologies and present repositories of medical and biomedical ontologies. In Section 3, we give a

high-level description of the use of ontologies in data integration and the approaches to ontology-based integration of possible sources of medical data. Section 4 discusses the issues and challenges faced by these approaches. We conclude in Section 5.

MEDICAL ONTOLOGIES AND ONTOLOGY REPOSITORIES

In the field of medicine, many ontologies have been created to standardize terminology, provide access to domain knowledge, verify data consistency, and facilitate integrative analysis over heterogeneous data (Hoehndorf, Dumontier and Gkoutos, 2013). The list of ontologies is constantly expanding and many of them are independently developed by many different groups and institutions. Most of these ontologies have been created for a specific area of healthcare, such as human disease (e.g., Human Disease Ontology (Schriml et al., 2019)), drug development (e.g., Drug Discovery Investigations (Qi et al., 2010)), and rehabilitation (e.g., Physical Medicine and Rehabilitation (Subirats and Ceccaroni, 2011)).

Since medical ontologies are provided to users in heterogeneous formats and interoperability between them is limited, several repositories have been established to provide some degree of semantic interoperability and to facilitate ontology discovery and access, as well as to support ontology reuse (Fung and Bodenreider, 2012). Such repositories provide access to integrated ontologies through powerful graphical and programming interfaces. Table 1 shows the largest repositories for medical and biomedical ontologies.

Ontologies are used extensively in data integration systems because they provide an explicit and machine-understandable description of the semantics of the data source (Wache et al., 2001). Moreover, medical and biomedical ontologies have been used in wide range of biomedical applications such as search and query of heterogeneous biomedical data, data exchange among applications, natural language processing, representation of encyclopedic knowledge and computer reasoning with data (Rubin, Shah and Noy, 2007; Hoehndorf, Schofield and Gkoutos, 2015).

Table 1: The largest repositories for medical and biomedical ontologies and their main features

Repository	**Main features**
BioPortal	BioPortal (https://bioportal.bioontology.org/) (Salvadores et al., 2013) is the most comprehensive repository of biomedical ontologies with more than 800 ontologies to date. It includes ontologies developed in formats such as OWL and OBO, as well as many medical terminologies in US National Library of Medicine's proprietary format. BioPortal also

	contains the metadata about the ontologies, and the mappings between terms in different ontologies. It can be accessed through a SPARQL endpoint.
OBO Library	The Open Biological and Biomedical Ontologies (OBO) library (http://obofoundry.org/) (Smith et al., 2007) consists of a collection of ontologies developed according to a set of agreed-upon principles, including open use, complementarity and collaborative development.
Ontobee	Ontobee (www.ontobee.org/) (Ong et al., 2017) is an ontology repository in which ontologies are presented as Linked Data. Ontobee provides information about the classes and relations used by the OBO project. The repository also includes Ontobeep, the program for ontology alignment, comparison, and result visualization.

ONTOLOGY -BASED INTEGRATION OF MEDICAL DATA SOURCES

Ontologies in Data Integration

Ontologies in data integration tasks are employed in several ways. Wache et al. (2001) distinguish three directions: single ontology approaches, multiple ontology approaches, and hybrid ontology approaches. Figure 1 shows the possible ways to use ontologies in data integration.

In single ontology approaches, all data source schemas are related to a global ontology that provides a shared vocabulary for specifying the semantics. Single ontology approaches assume that all data sources provide nearly the same view of the domain. Moreover, single ontology approaches are vulnerable to changes in the data sources which may entail the changes in the global ontology and in the mappings to the other data sources.

In multiple ontology approaches, each data source is described by its own ontology. Local ontologies are developed without regard to other sources and their ontologies, so there is no need for a common ontology and agreement between data sources, which simplifies the handling of changes in data sources. However, the lack of a common vocabulary makes it difficult to compare local ontologies, so an additional representation formalism is needed to define mappings between ontologies.

Hybrid approaches were created to overcome the drawbacks of the previous two approaches. For each data source schema, a local ontology is created and mapped to a global shared ontology that contains the basic concepts of a domain. New data sources can be easily added without changing the existing mappings.

Since valuable medical data are stored in diverse local data sources and formats, all of these storage types must be considered for mapping to ontologies in integration tasks.

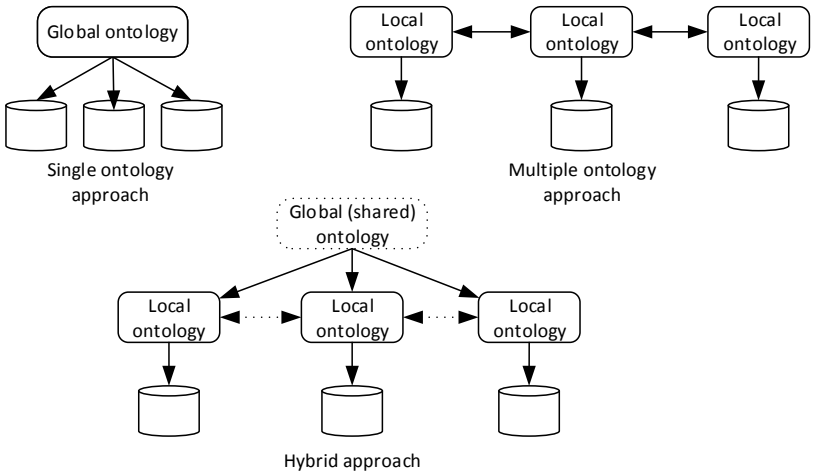

Figure 1. Three ways in which ontologies have been used in data integration systems

Structured Data Sources

Structured data are typically stored in a relational database. The process of mapping a relational database to an ontology follows certain mapping rules, which are described in detail in the literature (Haw and May, 2017; Asfand-E-Yar and Ali, 2020). The mapping rules define how the components in the relational databases can be transformed into ontology components such as classes, properties, instances, etc. In recent years, several tools for mapping relational databases to ontologies have been developed and are discussed and compared by Haw and May (2017). The main problem with ontology mapping is automation, which is not adequately supported. Most approaches still provide at best a semi-automatic mapping (Haw and May, 2017).

Semi-structured Data Sources

Data that do not fit into a formal structure such as a relational database or different models, but still contain markers or other elements to separate different data structures, are called semi-structured. Since different data models are considered semi-structured, the mapping process depends on the particular model.

Previous research led to the development of correspondence rules and solutions for mapping XML documents to ontologies, as presented in (Hacherouf, Bahloul and Cruz, 2015). Still, mapping NoSQL data sources to ontologies is still an active

research area. Since most NoSQL data sources are generally schemaless, the schema has to be re-engineered from heterogeneous data. However, certain database models, such as document stores can be considered to have a schema embedded and mixed with the data, with concepts implicitly defined in relations among collections and documents, so the schema must be extracted, which is also an active research question (Ptiček, Vrdoljak and Gulić, 2019).

Unstructured Data

Unstructured data in the form of free text, found in EHRs and elsewhere, contain the deeper, more complex information that often remains unexplored as they cannot be readily processed by a computer. One method for making free text machine-processable is annotation, which is the mapping of free text phrases to ontology concepts that express the meaning of the phrases. NLP can be used for semi-automatic processing of free text. In the field of medicine, NLP algorithms have been widely adopted and implemented (Jovanović and Bagheri, 2017; Kreimeyer et al., 2017; Kersloot et al., 2020). A survey of the current state of development of NLP algorithms that map medical text to ontology is given by Kersloot et al. (2020), along with their evaluation and a list of recommendations for algorithm evaluation.

Ontology Alignment

Once data sources have been mapped to (local) ontologies, their integration becomes a problem of ontology matching and alignment. Alignment of medical ontologies has been actively researched over the last decade, with a variety of approaches that differ both in the number of ontologies covered and the level of automation achieved. Existing approaches to medical ontology alignment are described by Dimitrieski et al. (2016). Although much work has been done in this area, no universal solution has emerged that enables automatic integration of medical ontologies, as they are usually limited to a specific domain of healthcare and used for specific use cases.

Related Work

Mate et al. (2015) proposed an ontology-based approach to organize and describe the medical concepts of both source and target systems to integrate the data across different clinical and research systems. They concentrate on structured but uncoded data, define declarative transformation rules within ontologies and illustrate how these constructs can then be used to automatically generate SQL code to perform the transformations.

Gocheva, Eminova and Batchkova (2016) proposed an ontology-based approach for biomedical data integration based on the concept of Linked Open Data (LOD). Different biomedical data are represented in RDF and integrated applying ontology alignment. Applicability of the approach is demonstrated through the integration of patients personal, medical, and billing records, stored in relational database with the

OWL version of schema.org. This approach considers LOD and relational databases as data sources.

An ontology-based data integration approach for multi-level integrative data analysis of cancer survival is proposed by Zhang et al. (2018). A new ontology is constructed for a global ontology, but many entities are reused from ontologies in BioPortal. This approach also considers only structured data from relational databases. Table 2 outlines the main characteristics of each approach.

Table 2: Main features of ontology-based approaches

	Objective	**Data Sources**	**Characteristics**
Mate et al. (2015)	Integration of data across different clinical and research systems	RDB	• Ontologies used to organize and describe medical concepts • Semi-structured and unstructured sources are not considered
Gocheva, Eminova and Batchkova (2016)	Integration of biomedical data and information using LOD vocabularies and a D2RQ-mapped database	RDB LOD	• RDBs are represented in RDF • Integration through ontology alignment • Unstructured data (free text) are not considered
Zhang et al. (2018)	Semantic data integration framework to support integrative data analysis of cancer survival	RDB	• Global ontology as a common vocabulary • Ontology reuse • Semi-structured and unstructured sources are not considered

CHALLENGES

Data Privacy and Data Unavailability

Medical data are sensitive personal data that need to be protected from unauthorized access and inadvertent disclosure, therefore there are only a limited number of publicly available medical datasets. Before the data can be released for research purposes, they must be carefully de-identified. Manual de-identification of large amounts of data is too costly and algorithmic methods to perform de-identification automatically have yet to be established. Several approaches for de-identifying data

in English have been published (Kushida et al., 2012), but research on data in other languages requires a lot of costly manual work due to the missing language resources.

Language Resources

Lack of language resources for non-English languages also complicates ontology mappings from medical texts. Precise NLP solutions are highly language and domain dependent, but tools and appropriate corpora exist almost exclusively for English.

Data Quality and Data Standardization

Data quality is an important factor in the successful integration and interpretation of medical data. Poor data quality can occur along several dimensions (Yeh and Puri, 2010), such as accuracy, consistency, validity, and completeness. In addition, there is no uniformity in healthcare classification and coding. Most medical institutions use their own terminology and coding systems. Several terminology standards have been developed to standardize the storage, retrieval, and exchange of medical data, such as SNOMED CT, LOINC, and ICD-10, as well as corresponding upper-level ontologies (e.g., SCTO (El-Sappagh et al., 2018)), but additional efforts are needed to apply these standards to existing data.

Ontology Integration

The number and variety of available medical ontologies makes their integration a challenging task. A large part of the ontologies that are in the aforementioned repositories have many of their concepts mapped to concepts in other ontologies. It remains a problem to find the right way to deal with the available medical ontologies and provide different views on a given domain. In addition, while OBO and OWL are popular formalisms for representing ontologies, many ontologies are only available in proprietary formats. As a result, the same entity often exists under different identifiers in multiple ontologies, making integration difficult. The Yosemite Project (2019) proposes a two-step approach to integrating medical ontologies: 1) converting any ontology format to OWL/RDF and 2) creating an integration algorithm for two OWL/RDF ontologies.

CONCLUSION AND FUTURE WORK

In this paper, we have presented basics of ontology-based integration of heterogeneous sources of medical data. Much work has been done in this area, but most of the existing approaches do not consider diverse data structures.
In our view, the most promising approach for medical data integration would be to develop a comprehensive semi-automated method that integrates data from structured, semi-structured and unstructured sources and utilizes existing medical ontologies that are rich in background knowledge and reside in ontology repositories.

Integrated medical data would be invaluable for various tasks such as data analysis, knowledge extraction, prediction of treatment outcomes, and decision support; however, there are still some unresolved issues that need to be addressed for successful data integration.

REFERENCES

Asfand-E-Yar, M. and Ali, R. (2020), Semantic Integration of Heterogeneous Databases of Same Domain Using Ontology. IEEE Access, 8, 77903–77919.

Caceres, S. (2013), Electronic health records: beyond the digitization of medical files. Clinics, 68 (8), 1077–1078.

Chen, J.-K. and Lee, W.-Z. (2019), An Introduction of NoSQL Databases Based on Their Categories and Application Industries. Algorithms, 12 (5), 106.

Dimitrieski, V., Petrović, G., Kovačević, A., Luković, I., and Fujita, H. (2016), A Survey on Ontologies and Ontology Alignment Approaches in Healthcare. In: H. Fujita, M. Ali, A. Selamat, J. Sasaki, and M. Kurematsu, eds. Trends in Applied Knowledge-Based Systems and Data Science. Cham: Springer International Publishing, 373–385.

El-Sappagh, S., Franda, F., Ali, F., and Kwak, K.-S. (2018), SNOMED CT standard ontology based on the ontology for general medical science. BMC Medical Informatics and Decision Making, 18 (1), 76.

Fung, K.W. and Bodenreider, O. (2012), Knowledge Representation and Ontologies. In: R.L. Richesson and J.E. Andrews, eds. Clinical Research Informatics. London: Springer London, 255–275.

Gocheva D., G., Eminova H., M., and Batchkova I., A. (2016), Ontology based data and information integration in biomedical domain. Machines. Technologies. Materials., 10 (2), 35–38.

Gruber, T.R., (1993), A translation approach to portable ontology specifications. Knowledge Acquisition, 5 (2), 199–220.

Hacherouf, M., Bahloul, S.N., and Cruz, C. (2015), Transforming XML documents to OWL ontologies: A survey. Journal of Information Science, 41 (2), 242–259.

Haw, S.-C., May, J.W., and Subramaniam, S. (2017), Mapping Relational Databases to Ontology Representation: A Review. In: Proceedings of the International Conference on Digital Technology in Education - ICDTE '17. Presented at the the International Conference, Taipei, Taiwan: ACM Press, 54–58.

Hoehndorf, R., Dumontier, M., and Gkoutos, G.V. (2013), Evaluation of research in biomedical ontologies. Briefings in Bioinformatics, 14 (6), 696–712.

Hoehndorf, R., Schofield, P.N., and Gkoutos, G.V. (2015), The role of ontologies in biological and biomedical research: a functional perspective. Briefings in Bioinformatics, 16 (6), 1069–1080.

Jovanović, J. and Bagheri, E. (2017), Semantic annotation in biomedicine: the current landscape. Journal of Biomedical Semantics, 8 (1), 44.

Kersloot, M.G., van Putten, F.J.P., Abu-Hanna, A., Cornet, R., and Arts, D.L. (2020), Natural language processing algorithms for mapping clinical text fragments onto ontology concepts: a systematic review and recommendations for future studies. Journal of Biomedical Semantics, 11 (1), 14.

Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S.F., Forshee, R., Walderhaug, M., and Botsis, T. (2017), Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. Journal of Biomedical Informatics, 73, 14–29.

Kushida, C.A., Nichols, D.A., Jadrnicek, R., Miller, R., Walsh, J.K., and Griffin, K. (2012), Strategies for De-identification and Anonymization of Electronic Health Record Data for Use in Multicenter Research Studies. Medical Care, 50, S82–S101.

Mate, S., Köpcke, F., Toddenroth, D., Martin, M., Prokosch, H.-U., Bürkle, T., and Ganslandt, T. (2015), Ontology-Based Data Integration between Clinical and Research Systems. PLOS ONE, 10 (1), e0116656.

Ong, E., Xiang, Z., Zhao, B., Liu, Y., Lin, Y., Zheng, J., Mungall, C., Courtot, M., Ruttenberg, A., and He, Y. (2017), Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration. Nucleic Acids Research, 45 (D1), D347–D352.

Ptiček, M., Vrdoljak, B., and Gulić, M. (2019), The potential of semantic paradigm in warehousing of big data. Automatika, 60 (4), 393–403.

Qi, D., King, R.D., Hopkins, A.L., Bickerton, G.R.J., and Soldatova, L.N. (2010), An Ontology for Description of Drug Discovery Investigations. Journal of Integrative Bioinformatics, 7 (3).

Rubin, D.L., Shah, N.H., and Noy, N.F. (2007), Biomedical ontologies: a functional perspective. Briefings in Bioinformatics, 9 (1), 75–90.

Salvadores, M., Alexander, P.R., Musen, M.A., and Noy, N.F. (2013), BioPortal as a Dataset of Linked Biomedical Ontologies and Terminologies in RDF. Semantic web, 4 (3), 277–284.

Schriml, L.M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., Felix, V., Jeng, L., Bearer, C., Lichenstein, R., Bisordi, K., Campion, N., Hyman, B., Kurland, D., Oates, C.P., Kibbey, S., Sreekumar, P., Le, C., Giglio, M., and Greene, C. (2019), Human Disease Ontology 2018 update: classification, content and workflow expansion. Nucleic Acids Research, 47 (D1), D955–D962.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R.H., Shah, N., Whetzel, P.L., and Lewis, S. (2007), The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology, 25 (11), 1251–1255.

Subirats, L. and Ceccaroni, L. (2011), An Ontology for Computer-Based Decision Support in Rehabilitation. In: I. Batyrshin and G. Sidorov, eds. Advances in Artificial Intelligence. Berlin, Heidelberg: Springer Berlin Heidelberg, 549–559.

Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., and Hübner, S. (2001), Ontology-Based Integration of Information - A Survey of Existing Approaches. In: OIS@IJCAI.

Yeh, P.Z. and Puri, C.A. (2010), An Efficient and Robust Approach for Discovering Data Quality Rules. In: 2010 22nd IEEE International Conference on Tools with Artificial Intelligence. Presented at the 2010 22nd IEEE International Conference on Tools with Artificial Intelligence, 248–255.

Yosemite Project [online] (2019), GitHub. Available from: https://github.com/yosemiteproject [Accessed 5 May 2021].

Zhang, H., Guo, Y., Li, Q., George, T.J., Shenkman, E., Modave, F., and Bian, J. (2018), An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival. BMC Medical Informatics and Decision Making, 18 (S2), 41.