Published online: 2020-08-21

Design and Use of Semantic Resources: Findings from the Section on Knowledge Representation and Management of the 2020 International Medical Informatics Association Yearbook

Ferdinand Dhombres[1,2], Jean Charlet[1,3], Section Editors for the IMIA Yearbook Section on Knowledge Representation and Management

[1] Sorbonne Université, Université Paris Nord, INSERM, UMR_S 1142, LIMICS, Paris, France
[2] Médecine Sorbonne Université, Service de Médecine Fœtale, Hôpital Armand Trousseau, Paris, France
[3] AP-HP, DRCI, Paris, France

Summary

Objective: To select, present, and summarize the best papers in the field of Knowledge Representation and Management (KRM) published in 2019.

Methods: A comprehensive and standardized review of the biomedical informatics literature was performed to select the most interesting papers of KRM published in 2019, based on PubMed and ISI Web Of Knowledge queries.

Results: Four best papers were selected among 1,189 publications retrieved, following the usual International Medical Informatics Association Yearbook reviewing process. In 2019, research areas covered by pre-selected papers were represented by the design of semantic resources (methods, visualization, curation) and the application of semantic representations for the integration/enrichment of biomedical data. Besides new ontologies and sound methodological guidance to rethink knowledge bases design, we observed large scale applications, promising results for phenotypes characterization, semantic-aware machine learning solutions for biomedical data analysis, and semantic provenance information representations for scientific reproducibility evaluation.

Conclusion: In the KRM selection for 2019, research on knowledge representation demonstrated significant contributions both in the design and in the application of semantic resources. Semantic representations serve a great variety of applications across many medical domains, with actionable results.

Keywords

Knowledge management; knowledge representation; vocabulary, controlled; information storage and retrieval; semantics

Yearb Med Inform 2020:163-8
http://dx.doi.org/10.1055/s-0040-1702010

1 Introduction

The year 2019 has produced a large amount of publications related to Knowledge Representation and Management (KRM) in Medicine. KRM focuses on the development of resources and techniques to be used and leveraged in other medical informatics domains. In this review, we present a selection of some of the best papers published in 2019 in the KRM domain, based either on their impact or on the novelty of the approach proposed within the medical knowledge representation and management field.

2 Paper Selection Method

We conducted the selection of KRM papers based on the set of queries established in the 2019 edition of the IMIA Yearbook of Medical Informatics [1]. As compared with the previous editions of the IMIA Yearbook in 2017 and 2018 [2, 3], both PubMed/MELDINE and Web of Knowledge were used to search for KRM articles published in 2019. We followed a generic method to select the best papers, commonly used in all sections of the Yearbook since 2013 []. As for last year, the search was performed on MEDLINE by querying PubMed and also on the Institute for Scientific Information (ISI) Web of Knowledge database (WoL). Additionally, the articles of the *Journal of Biomedical Semantics* (JBS) and of the *Journal of Biomedical Informatics* (JBI) were manually analyzed. Our query includes Medical Subject Headings (MeSH) descriptors related to KRM in the context of medical informatics with a restriction to international peer-reviewed journals, including conference proceedings indexed in PubMed. Only original research articles published in 2019 (from 01/01/2019 to 12/31/2019) were considered; publications types such as reviews, editorials, comments, and letters to the editors were excluded.

The selection of best papers was performed among the results of the query process in three steps. At the first step, section editors reviewed all titles, abstracts, and publication types to establish a short list of 15 candidate best papers. At the second step, five expert reviewers (including the section editors) reviewed the candidate papers using the IMIA Yearbook quality criteria scoring method. More specifically, the following aspects of the papers were evaluated: significance, quality of scientific content, originality and innovativeness, coverage of related literature, organization and quality of the presentation. The final step of the selection of best papers was achieved during a meeting of the whole editorial team, based on the reviews of section editors, chief editor of KRM, and external reviewers (at least 4 reviewers per paper).

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2020 in the section 'Knowledge Represenation and Management'. The articles are listed in alphabetical order of the first author's surname.

Section Knowledge Representation and Management
▪ Burek P, Scherf N, Herre H. Ontology patterns for the representation of quality changes of cells in time. J Biomed Semantics 2019;10(1):16.
▪ Denaxas S, Gonzalez-Izquierdo A, Direk K, Fitzpatrick NK Fatemifar G, Banerjee A, Dobson RJB, Howe LJ, Kuan V, Lumbers RT, Pasea L, Patel RS, Shah AD, Hingorani AD, Sudlow C, Hemingway H. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. J Am Med Inform Assoc 2019;26(12):1545-59.
▪ Rector A, Schulz S, Rodrigues J-M, Chute CG, Solbrig H. On beyond Gruber: "Ontologies" in today's biomedical information systems and the limits of OWL. J Biomed Inform: X 2019 Jun 1;2:100002.
▪ Shen F, Zhao Y, Wang L, Mojarad MR, Wang Y, Liu S, Liu H. Rare disease knowledge enrichment through a data-driven approach. BMC Med Inform Decis Mak 2019;19(1):32.

3 Results

For 2019, the KRM query retrieved 1,105 citations from PubMed (JBI and JBS excluded), 18 additional citations from WoL, and 66 manually selected citations from JBI and JBS. The new optimized set of queries introduced last year was stable as compared with the previous query (15% increase of the query in 2018), and had an overall good precision of KRM relevant papers. In contrast, there was a 45% decrease in comparison with the results of the previous query used for 2017 [3]. Section editors achieved a first selection of 148 papers based on titles and abstracts. After a second review of this set of papers, including full text reviews, a selection of 15 candidate best papers was established [4-18]. Five reviewers reviewed these pre-selected papers to best four best papers [4-7].

In direct line with the research presented last year [1], the four best papers published in 2019 demonstrated even further the added-value of ontology-based data integration approaches and that the development of ontology methods is an active area of bioinformatics research.

3.1 Best Papers Selection for 2019

The selection of best papers published in 2019 in the KRM subfield of biomedical informatics are displayed in Table 1. Burek *et al*., [4] investigate the fundamental problem of modeling quality changes over time in biomedical ontologies specified in the Web Ontology Language (OWL). They propose a precise analysis of different design patterns of time representation. This work provides six options to represent time in ontologies, with a supportive description of extensibility, maintainability, Terminological Box/Assertional Box (T-Box/A-Box) complexity, and adequateness for a use case. In the same methodological vein, the article of Rector *et al*., [6] recalls the experiences of commonly constructed ontologies and the development of ontological reasoning. These works are described from an historical perspective, and the paper invites us to rethink about knowledge modeling while describing possible directions for the future development of semantic resources. The limitations of OWL and the consequences of an open-world assumption reasoning process are illustrated, with actionable proposal of alternatives approaches, based on the experience of the design of the 11th revision of the International Classification of Diseases (ICD-11) ontology.

In another selected paper, Denaxas *et al*., [5] describe the CALIBER platform developed for validation and sharing of reproducible phenotypes in national structured Electronic Health Record (EHR) in the United Kingdom (UK). This EHR-based phenomics approach, applied on the data of 15 million individuals, is an important step towards the international use of UK EHR data for health research, with applications for translational research at the population level. Also selected as a best paper for 2019, Shen *et al*., [7] describe a work which aims at enriching available rare disease resources by mining phenotype-disease associations from a 5-year collection of 12.8 million clinical notes from electronic medical records at the Mayo Clinic. Their approach was able to enrich existing rare disease knowledge resources with phenotype-disease associations, with an application to the differential diagnosis across rare and non-rare diseases. The four best papers selected in 2020 are detailed in the appendix.

3.2 Other Pre-selected Papers for 2019

Among the 11 other short-listed papers for 2019, we observed two research directions similar to the distribution of best papers, one with a focus on semantic resource design, and another focused on the use of semantic representations in different applications. The medical domains in semantic resource design, visualization, and curation are represented by Cardiology, Genetics, Pharmacology, Mental Health, and Neurology. The applications of semantic representations are mainly focused on the integration and the enrichment of data, with promising results in the characterization of molecular mechanisms in concomitant phenotypes, biomedical data analysis based on semantic-aware machine learning solutions, flexible phenotypes capture through pre and post-coordination, and scientific reproducibility evaluation based on semantic provenance information representations.

3.2.1 Semantic Resources Design, Visualization, and Curation

Several articles describe ontology designs [9, 10, 18]. First, Brenas *et al*., [9] describe the ins and outs of the Adverse Childhood Experiences (ACEs) Ontology for Mental Health Surveillance. This ontology was created to be used by major actors in the ACEs community with different applications, from the diagnosis of individuals (and the prediction of potential negative outcomes), to the prevention of ACEs in a population (and the design of interventions and policies). Doing-Harris *et*

al., [10] describe the development of a cardiac-centered frailty ontology. This ontology is designed to cover the portions of reality relevant to assess the patient frailty, with a focus on cardiac care decisions. The authors gathered terms using different frailty-measuring instrument findings and physician interviews and they applied realist principles to reconcile clinical texts, medical literature, and existing ontologies. The hierarchical structure is interoperable with the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT), considering only a subset of SNOMED-CT findings. Yu *et al.,* [18] describe the ontology of Drug Adverse Events (ODAE). In order to logically represent the complex relations among drug ingredients and mechanisms of action, adverse events, age, disease, and other factors, an ontology design pattern was developed and applied to generate the ODAE. The result is a community-driven open-source semantic resource that follows the Open Biological and Biomedical Ontology (OBO) Foundry development principles (e.g., openness and collaboration). The use of ODAE is demonstrated with SPARQL Protocol and RDF Query Language (SPARQL) queries, on a knowledge base built under ODAE for 224 neuropathy-inducing drugs, allowing deep investigations of mechanisms of action and clinical contexts.

Other articles deal with ontology visualization and development tools [12,13]. Jackson *et al.*, [12] propose ROBOT, a tool for automating OWL/OBO ontology workflows. This framework is an open-source library with command-line tool for various ontology development and curation tasks. The library can be called from any programming language that runs on the Java Virtual Machine (JVM). This helps ontology developers to efficiently create, maintain, and release high-quality ontologies in OWL and OBO formats, so that they can spend more time focusing on task development. Kuznetsova *et al*., [13] present an open source software, CirGO (Circular Gene Ontology), that provides the visualization of non-redundant two-level hierarchically structured ontology terms from gene expression data in a 2D space. This software displays the most enriched gene ontology terms in an informative, comprehensive, and intuitive format that is achieved by organizing data from the most relevant to the least, as well as the appropriate use of colors and supporting information. CirGO is freely available at https://github.com/IrinaVKuznetsova/CirGO.git.

The rapid accumulation of new biomedical literature not only causes curated knowledge graphs (KGs) to become outdated and incomplete, but also makes manual curation impractical and unsustainable. Hoyt *et al.*, [11] have developed two workflows to address this issue: the first for re-curating KGs to control syntactic and semantic quality, and the second for rationally enriching KGs through the manually revision of automatically extracted relations for the nodes with low information density. They applied these approaches to the KGs of the NeuroMMSig inventory. This KG curation workflow is freely available at https://github.com/bel-enrichment/bel-enrichment.

3.2.2 Semantic Resources Applications: Annotations, Mining, and Enrichment

In their paper, Babbi *et al.,* [8] introduce PhenPath, a new set of resources: PhenPathDB and PhenPathTOOL. PhenPathDB is a database of human genes associated with phenotypes described in the Human Phenotype Ontology (HPO) and in OMIM Clinical Synopses. Phenotypes are then associated to biological functions and pathways by means of NET-GE, a network-based method for functional annotation enrichment of sets of genes. PhenPathTOOL enables the identification of molecular features relevant for investigating diseases characterized by multiple phenotypes. This framework provides a support for the characterization of molecular mechanisms and biological functions underlying the concomitant manifestation of phenotypes. The resource is freely available at http://phenpath.biocomp.unibo.it.

In their paper, Lamurias *et al*., [14] propose a new model to detect and classify relations in text, named BO-LSTM, that takes advantage of domain-specific ontologies, by representing each entity as the sequence of its ancestors in the ontology. The authors implemented BO-LSTM as a recurrent neural network with long short-term memory units integrating open biomedical ontologies, specifically the Chemical Entities of Biological Interest (ChEBI), the Human Phenotype Ontology, and the Gene Ontology. This work demonstrates how domain-specific ontologies can improve deep learning models for classification of biomedical relations.

Smaili *et al.,* [17] propose to use formal axioms in biomedical ontologies to improve the analysis and interpretation of biomedical data. The general principle is to consider each axiom of the ontology as a sentence processed by an algorithm similar to Word2vec. They use ontology-based machine learning methods to evaluate the contribution of formal axioms and ontology metadata to evaluate the prediction of protein–protein interactions and gene–disease associations. They find that the background knowledge provided by the Gene Ontology (and other ontologies) significantly improves the performance of prediction models through the provision of domain-specific background knowledge. Their results have implications on the further development of knowledge bases and ontologies, especially since machine learning methods are more frequently applied across the life sciences.

In a short paper, Siegele *et al*., [16] illustrate the task of phenotype annotation with the ontology of microbial phenotypes (OMP). They describe an OMP-based annotation framework that supports the representation of a wide range of phenotypes and provides flexibility for different levels of detail. This framework can support research by the capture of phenotypes from the experimental literature for a variety of microbes, with pre and post-coordination methods.

Sahoo *et al*., [15] introduce the ProvCare platform for mining semantic provenance information in the biomedical literature, with the aim of evaluating scientific reproducibility. This platform relies on a S3 model and a formal ontology. A provenance-focused text processing workflow generates provenance triples consisting of subject, predicate, and object, using metadata extracted from articles. The resulting ProvCaRe knowledge repository (available at https://provcare.case.edu/) supports "provenance-aware" hypothesis-driven search queries. This repository is one of the largest provenance resources for biomedical research studies that combines intuitive search functionality with a new provenance-based ranking feature to sort the retrieved articles.

4 Conclusions

In the KRM selection for 2019, research on knowledge representations demonstrated significant contributions both in the design and in the application of semantic resources. Novel ontologies have emerged and new tools are available for the scientific community. Some methodological foundations have been revisited and large scale applications have been deployed. Semantic representations serve a great variety of applications across many medical domains, with actionable results. As in 2018, we observed promising research combining knowledge representations and machine learning techniques.

Acknowledgements
We would like to thank Adrien Ugon for his support and the reviewers for their participation in the selection process of the IMIA Yearbook best papers.

References

1. Dhombres F, Charlet J. Formal Medical Knowledge Representation Supports Deep Learning Algorithms, Bioinformatics Pipelines, Genomics Data Analysis, and Big Data Processes. Yearb Med Inform 2019;28(1):152-5.
2. Dhombres F, Charlet J. As Ontologies Reach Maturity, Artificial Intelligence Starts Being Fully Efficient: Findings from the Section on Knowledge Representation and Management for the Yearbook 2018. Yearb Med Inform 2018;27(1):140-5.
3. Dhombres, F, Charlet J. Knowledge Representation and Management, It's Time to Integrate! Yearb Med Inform 2017;26(1):148-51.
4. Burek P, Scherf N, Herre H. Ontology patterns for the representation of quality changes of cells in time. J Biomed Semantics 2019;10(1):16.
5. Denaxas S, Gonzalez-Izquierdo A, Direk K, Fitzpatrick NK, Fatemifar G, Banerjee A, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. J Am Med Inform Assoc 2019;26(12):1545-59.
6. Rector AL, Schulz S, Rodrigues JM, Chute CG, Solbrig H. On beyond Gruber: "Ontologies" in today's biomedical information systems and the limits of OWL. J Biomed Informatics: X 2019;2:100002.
7. Shen F, Zhao Y, Wang L, Mojarad MR, Wang Y, Liu S, et al. Rare disease knowledge enrichment through a data-driven approach. BMC Med Inform Decis Mak 2019;19(1):32.
8. Babbi G, Martelli PL, Casadio R. PhenPath: a tool for characterizing biological functions underlying different phenotypes. BMC Genomics 2019;20(Suppl 8):548.
9. Brenas JH, Shin EK, Shaban-Nejad A. Adverse Childhood Experiences Ontology for Mental Health Surveillance, Research, and Evaluation: Advanced Knowledge Representation and Semantic Web Techniques. JMIR Ment Health 2019;6(5):e13498.
10. Doing-Harris K, Bray BE, Thackeray A, Shah RU, Shao Y, Cheng Y, et al. Development of a cardiac-centered frailty ontology. J Biomed Semantics 2019;10(1):3.
11. Hoyt CT, Domingo-Fernandez D, Aldisi R, Xu L, Kolpeja K, Spalek S, et al. Re-curation and rational enrichment of knowledge graphs in Biological Expression Language. Database (Oxford). 2019;2019:baz068.
12. Jackson RC, Balhoff JP, Douglass E, Harris NL, Mungall CJ, Overton JA. ROBOT: A Tool for Automating Ontology Workflows. BMC Bioinformatics 2019;20(1):407.
13. Kuznetsova I, Lugmayr A, Siira SJ, Rackham O, Filipovska A. CirGO: an alternative circular way of visualising gene ontology terms. BMC Bioinformatics 2019;20(1):84.
14. Lamurias A, Sousa D, Clarke LA, Couto FM. BO-LSTM: classifying relations via long short-term memory networks along biomedical ontologies. BMC Bioinformatics 2019;20(1):10.
15. Sahoo SS, Valdez J, Kim M, Rueschman M, Redline S. ProvCaRe: Characterizing scientific reproducibility of biomedical research studies using semantic provenance metadata. Int J Med Inform 2019;121:10-8.
16. Siegele DA, LaBonte SA, Wu PI, Chibucos MC, Nandendla S, Giglio MG, et al. Phenotype annotation with the ontology of microbial phenotypes (OMP). J Biomed Semantics 2019;10(1):13.
17. Smaili FZ, Gao X, Hoehndorf R. Formal axioms in biomedical ontologies improve analysis and interpretation of associated data. Bioinformatics 2020;36(7):2229-36.
18. Yu H, Nysak S, Garg N, Ong E, Ye X, Zhang X, et al. ODAE: Ontology-based systematic representation and analysis of drug adverse events and its usage in study of adverse events given different patient age and disease conditions. BMC Bioinformatics 2019;20(Suppl 7):199.

Correspondence to:
Dr. Ferdinand Dhombres
Médecine Sorbonne Université, INSERM and APHP
Hôpital Universitaire Armand Trousseau
service de médecine foetale
26 rue du Dr Arnold Netter
75012 Paris, France
E-mail: ferdinand.dhombres@inserm.fr

Appendix: Summaries of Selected Best Papers for the 2020 IMIA Yearbook, Section Knowledge Representation and Management

Burek P, Scherf N, Herre H

Ontology patterns for the representation of quality changes of cells in time

J Biomed Semantics 2019;10(1):16

In this paper, the authors investigate the fundamental problem of modeling changes of quality occurring over time in biomedical ontologies specified in OWL. The paper is the result of the lessons learned during the development of an ontology for the annotation of cell tracking experiments. They present, discuss, and evaluate six representation patterns for specifying cell changes in time. In particular, they discuss two patterns of temporally changing information: n-ary relation reification and 4d fluents. They analyze the performance of each pattern with respect to standard criteria used in software engineering and data modeling, *i.e.*, simplicity, scalability, extensibility, and adequacy. There is no ideal solution and the patterns behave differently depending on the temporal distribution of the information modeled.

Finally, modeling quality value change is not limited to cell tracking experiments. It is a common and non-trivial task across many biomedical domains. The presented patterns are domain-independent. Since a change of quality values is common to many biomedical domains, this work has possible applications to represent the evolution of a patient's condition according to his/her treatments.

Denaxas S, Gonzalez-Izquierdo A, Direk K, Fitzpatrick NK Fatemifar G, Banerjee A, Dobson RJB, Howe LJ, Kuan V, Lumbers RT, Pasea L, Patel RS, Shah AD, Hingorani AD, Sudlow C, Hemingway H

UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER

J Am Med Inform Assoc 2019;26(12):1545-59

In this paper, the authors present the CALIBER EHR platform for developing, validating, and sharing reproducible phenotypes from national structured EHR in United Kingdom with applications to translational research.

They implemented a rule-based phenotyping framework with a systematic validation based on up to six different approaches. This framework was applied on 15 million individuals based on national EHR data collection from four data sources: UK primary care EHR data, hospital care billing data, disease registry data, and national death records data. Five standard terminologies were used to record information: a subset of the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT Read terms) for primary care clinical data, a derivative of the NHS Dictionary of Medicines and Devices (dm+d) for prescription codes, the International Classification of Diseases (9th and 10th revisions) for secondary care diagnoses and causes of mortality, and the Office of Population Censuses and Surveys Classification of Surgical Operations and Procedures (4th revision) for hospital surgical procedures. The authors created algorithms for 51 diseases, syndromes, biomarkers, and lifestyle risk factors, using the CALIBER phenotyping framework. Three are detailed in the article: Heart failure, Myocardial infarction, and Bleeding. Validation evidence was established on cross-EHR source concordance, clinical note review, etiology, prognosis, genetic associations, and external population.

The open-access CALIBER data platform has been used by 40 national and international research groups in 60 peer-reviewed publications. This EHR-based phenomics approach within the CALIBER platform is an important step towards the international use of UK EHR data for health research.

Rector A, Schulz S, Rodrigues J-M, Chute CG, Solbrig H

On beyond Gruber: "Ontologies" in today's biomedical information systems and the limits of OWL

J Biomed Inform: X 2019 Jun 1;2:100002

In this paper, the authors start with the precise definition of the term "ontology", the experience of commonly constructed ontologies, and the development of ontological reasoning. Since the introduction of ontologies, open-world reasoning systems based on description logics have been developed, OWL has become a standard, and philosophical issues have been raised.

The article highlights what OWL-DL statements mean and lists all the potential pitfalls of OWL-DL representations with examples. It illustrates in particular that a reasoning process is not an ontological classification task (open vs. closed world), and that a paradigm shift is needed for reasoning. These developments are described from a historical perspective that shows the Information Technology (IT) context present at the time the ontologies were created.

The authors discuss the confusion that has emerged from the evolution of research in this field with two broad usages for the word "ontology" in the biomedical informatics literature: i) its original usage as a general term for the "background knowledge base"; and ii) as a term for some subset of the background knowledge base that is considered fundamental, on logical and/or philosophical grounds. They advocate for using the invariants as the denomination for traditional ontological statements, and for developing symbolic representation systems as hybrid model combining these invariants and reasoning capabilities based on frames.

In this framework, the authors also state that the International Classification of Diseases, 11th version (ICD-11) is philosophically close to SNOMED CT with an hybrid architecture: an ontology at the base and transformation/linearization in classification for certain uses (e.g. coding).

This paper was rated as a best paper because it is a very interesting and well-documented article that encourages us to rethink about knowledge modeling and the future systems that we will be able to build.

Shen F, Zhao Y, Wang L, Mojarad MR, Wang Y, Liu S, Liu H

Rare disease knowledge enrichment through a data-driven approach

BMC Med Inform Decis Mak 2019;19(1):32

In this paper, the authors applied a data-driven approach to enrich existing rare disease resources by mining phenotype-disease associations from a 5-year collection of 12.8 million clinical notes from electronic medical records (EMRs) at the Mayo Clinic.

They used association rule mining algorithms on EMRs to extract significant phenotype-disease associations and enriched existing rare disease resources: Human Phenotype Ontology (HPO) and Orphanet. They generated three phenotype-disease bipartite graphs: the HPO-Orphanet graph, the EMR graph, and the enriched knowledge base HPO-Orphanet + graph. They conducted a case study on Hodgkin lymphoma to compare performance on differential diagnosis among these three graphs.

The disease-disease similarity generated by the eRAM (an existing encyclopedia of rare disease annotations mined from 10 million scientific publications and medical records) was used as gold standard to compare the three graphs with a sensitivity and specificity of 0.17, 0.36, 0.46 and 0.52, 0.47, 0.51, for the three graphs, respectively.

They also compared the top 15 diseases generated by the enriched knowledge HPO-Orphanet + graph with eRAM and another clinical diagnostic tool, the Phenomizer. The proposed approach was able to significantly enrich existing rare disease knowledge resources with phenotype-disease associations from EMRs. This work provides a solution for differential diagnosis across rare and non-rare diseases.