Article published online: 2021-09-03

Knowledge Representation and Management: Interest in New Solutions for Ontology Curation

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

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

Summary

Objective: To select, present and summarize some of the best papers in the field of Knowledge Representation and Management (KRM) published in 2020.

Methods: A comprehensive and standardized review of the medical informatics literature was performed to select the most interesting papers of KRM published in 2020, based on PubMed queries. This review was conducted according to the IMIA Yearbook guidelines.

Results: Four best papers were selected among 1,175 publications. In contrast with the papers selected last year, the four best papers of 2020 demonstrated a significant focus on methods and tools for ontology curation and design. The usual KRM application domains (bioinformatics, machine learning, and electronic health records) were also represented.

Conclusion: In 2020, ontology curation emerges as a significant topic of research interest. Bioinformatics, machine learning, and electronics health records remain significant research areas in the KRM community with various applications. Knowledge representations are key to advance machine learning by providing context and to develop novel bioinformatics metrics. As in 2019, representations serve a great variety of applications across many medical domains, with actionable results and now with growing adhesion to the open science initiative.

Keywords

Knowledge representation and management, ontologies, ontology design, ontology curation, IMIA

Yearb Med Inform 2021:185-90
http://dx.doi.org/10.1055/s-0041-1726508

1 Introduction

The year 2020 has produced a large number of publications related to Knowledge Representation and Management (KRM) in Medicine. KRM focuses on the development and application of resources and methods to be used in other medical informatics domains. During the last four years, we have observed significant contributions to data integration, machine learning, bioinformatics and genomics data analysis, and a growing maturity in ontology development.

In this review, we present a selection of some of the best papers published in 2020 in the KRM domain, based either on their impact or the novelty of the approach in the medical KRM field.

2 Paper Selection Method

We conducted the selection of KRM papers, based on the set of queries optimized during the last editions of the International Medical Informatics (IMIA) Yearbook [1-4]. In comparison with the latest editions of the IMIA Yearbook [3, 4], where both PubMed/MELDINE and Web of Science™ were used to search for KRM articles, we did not expand the search to the Web of Science (WoS) database this year. We have observed limited additional contributions in WoS in 2018 (3.4%, n=34/962) and 2019 (1.5%, n=18/1189) and no candidate paper was selected among these additional contributions, mostly because they were more bioinformatics-related than actual KRM papers. We followed a generic method to select the best papers, commonly used in all sections of the Yearbook and have done so since 2013. As for the last years, the search was performed on MEDLINE by querying PubMed and additionally, the articles published in 2020 in the *Journal of Biomedical Semantics* (JBS) and in the *Journal of Biomedical Informatics* (JBI) were manually analyzed.

Our query set 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 2020 (from 01/01/2020 to 12/31/2020) were considered; we excluded the following publication types: reviews, editorials, comments, and letters to the editors.

The selection of the best papers was performed among the results of the query process, in three steps. At the first step, the section editors reviewed all title, abstract and type of publications to establish a short list of 15 candidate papers. At the second step, five expert reviewers (including the section editors and two editors in chief) 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 papers was achieved during a meeting of the whole editorial board, based on the reviews and the report of the section editors.

3 Results

For 2020, the KRM query retrieved 1,140 citations from PubMed (JBI and JBS excluded) and 35 manually selected citations from JBI and JBS. The trend in citations retrieved by our optimized set of queries on PubMed remains stable in comparison with the 1,189 citations reviewed in 2019 [4] and we observed a good precision for KRM salient papers. The section editors achieved a first selection of 226 papers based on the title and abstract. After a second review of this set of papers, including full text reviews, a selection of 15 candidate best papers was established [5-19]. Five reviewers reviewed these pre-selected papers to select the best four final papers [5-8]. In contrast with the papers selected last year, the four best papers of 2020 demonstrated a significant focus on methods and tools for ontology curation and design, and the usual KRM application domains: bioinformatics, machine learning, and electronic health records (EHRs).

3.1 Best Paper Selection for 2020

Among the best papers in 2020, ontology curation research is in the spotlight. Zheng uses the Unified Medical Language System (UMLS) to discover missing is-a relations in standard terminologies [5]. They developed a transformation-based method which uses the rich knowledge provided by the UMLS for auditing and improving the qualities of its source terminologies. Slater *et al.* [6] investigate Open Biomedical Ontologies (OBO) ontologies' inconsistencies and proposes a method to detect hidden unsatisfiabilities in an ontology that arise when combined with other ontologies. They identified a large set of hidden unsatisfiability across a broad range of OBO biomedical ontologies and their results provide elements towards more consistent ontologies. Le presents the UFO [7], a new multifaceted tool for evaluating and analyzing semantic similarity in ontologies in OBO format. This solution integrates one of the most comprehensive set of features: large coverage of existing similarity metrics for term/entity/network analysis and enrichment, and a visualization interface. This tool brings a unified solution for ontology curation and similarity-based research.

In another best paper, Robinson *et al.* introduce the LIkelihood Ratio Interpretation of Clinical AbnormaLities [8] to revisit the calculation of Likelihood Ratio (LHR) of phenotypes to support prioritization in candidate genes or diseases in bioinformatics workflows. This algorithm leverages the ontological structure of the Human Phenotype Ontology (HPO) to compute phenotype probabilities used in the calculation of phenotypes LHR.

The four best papers are detailed in the Appendix. Among all candidate papers for 2020, we observed three main directions in research: ontology curation, semantics for machine learning and bioinformatics, and knowledge representation in EHRs.

3.2 Ontology Curation and Design

Besides the three selected best papers focusing on ontology curation [5-7], other candidate papers addressed curation and design of ontologies in their publications. Hier and Brint [9] created a target ontology (NEO) for capturing the neurological examination using 1,100 concepts from the UMLS Metathesaurus. As no UMLS ontology used alone has enough depth capture neurological examination descriptions, the authors combined UMLS terminologies: SNOMED CT, MEDCIN, OMIM, MeSH, and HPO. This article is a good advocate of a multi-terminology coding approach vs. the development of an ad hoc ontology. NEO is validated by a coverage study.

Hou *et al.* [10] propose a large and open knowledge representation model dedicated to precision medicine, the Precision Medicine Ontology (PMO). They developed a semi-automated method, from the collection of terms (4.53 million terms, from 62 biomedical vocabularies) to the evaluation and applications of PMO. This work illustrates how a large terminology can be designed through UMLS integration to cover precision medicine use cases. However, its sustainability over time remains uncertain.

3.3 Knowledge Representation for Machine Learning and Bioinformatics

The contributions of knowledge representations to bioinformatics and machine learning methods are illustrated by the best paper by Robinson *et al.* [8]. Among the candidate papers, three are using machine learning methods enriched by semantic representations [11-13] and the other are open science bioinformatics initiatives enforced by semantic representations [14-16].

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2021 in the special section 'Knowledge Representation and Management'. The articles are listed in alphabetical order of the first author's surname.

Section
Knowledge Representation and Management
▪ Le DH. UFO: A tool for unifying biomedical ontology-based semantic similarity calculation, enrichment analysis and visualization. PLoS One 2020;15(7):e0235670.
▪ Robinson PN, Ravanmehr V, Jacobsen JOB, Danis D, Zhang XA, Carmody LC, Gargano MA, Thaxton CL, Core UNCB, Karlebach G, Reese J, Holtgrewe M, Kohler S, McMurry JA, Haendel MA, Smedley D. Interpretable clinical genomics with a likelihood ratio paradigm. Am J Hum Genet 2020;107(3):403-17.
▪ Slater LT, Gkoutos GV, Hoehndorf R. Towards semantic interoperability: finding and repairing hidden contradictions in biomedical ontologies. BMC Med Inform Decis Mak 2020;20(Suppl 10):311.
▪ Zheng F, Shi J, Yang Y, Zheng WJ, Cui L. A transformation-based method for auditing the IS-A hierarchy of biomedical terminologies in the Unified Medical Language System. J Am Med Inform Assoc 2020;27(10):1568-75.

Koulmanov *et al.* [11] demonstrate how ontologies are increasingly being used to represent background knowledge in similarity-based analysis and machine learning models, with a broad overview of several research topics supported by an impressive list of references. Of particular interest is how semantic similarity measures and ontology embeddings can exploit the background knowledge in ontologies, and how ontologies can provide constraints that improve machine learning models. One use case is provided along with a novel benchmark dataset for the prediction of protein-protein interactions with ontologies. In another machine learning driven method, Sousa *et al.* [12] use genetic programming over a set of semantic similarities derived from different semantic representations, to support supervised learning tasks. They combine symbolic and statistical methods of artificial intelligence with applications in bioinformatics. This approach is validated by a use case of protein-protein interaction prediction. Alag [13] exploits the ClinicalTrials.gov dataset to build a corpus annotated with HPO and MeSH terms and to extract protein mutations and single nucleotide polymorphisms. He developed a repository of reports for each mutation and associated trials with all meta-data available via APIs. One use case demonstrates potential machine learning applications based on these APIs.

Blatti *et al.* [14] present the KnowEnG (Knowledge Engine for Genomics) cloud platform. It includes tools such as gene prioritization, sample clustering, gene set analysis, and expression signature analysis. It adheres to the open science principles and complies with all "FAIR" principles. This platform is a good example of a bioinformatics initiative developing standardized tools allowing reproducible research. Beck *et al.* [15] describe the latest evolutions of Genome-wide Association Study (GWAS) Central: a comprehensive resource for the discovery and comparison of genotype and phenotype data from more than 3.8K genome-wide association studies. MeSH and HPO support a precise identification of genetic variants associated with diseases, phenotypes and traits of interest. The Experimental Factor Ontology (EFO), the Disease Ontology Lite (DOLite) and the International Classification of Diseases (ICD-10) are also used to code the phenotype data in GWAS. Shefchek *et al.* [16] describe the status of the Monarch initiative in 2019. Even if this is not a research paper per se, the importance of such initiative is obvious to support research and is a large-scale example of knowledge management across species. The Monarch platform integrates data and allows rich analytics for connecting phenotypes to genotypes. From a KRM perspective, they use a large collection of widely adopted ontologies but also develop (and maintain) additional ontologies and mappings. RDF modelling of the covered biological knowledge supports the data integration and allows efficient data access (with APIs), reports and visualization.

3.4 Knowledge Representation in Electronic Health Records

Knowledge representation in EHRs follows various directions in research: improving granularity of patient description [17], large-scale integration of patient data [18], and formal representation to drive EHR development [19].

Jani *et al.* [17] have designed an ontology of social determinants of health to explore potential improvement in the current EHR coding based on three standardized primary care codes recommended by the National Health Service in England. The 668 codes from the ontology captures social prescriptions with a high granularity, as demonstrated by 5 million instances recorded between January 2011 and December 2019 by a national sentinel network.

Lamer *et al.* [18] evaluate the feasibility of implementing French national EHRs in the observational medical outcome partnership (OMOP) Common Data Model (CDM). Such standardization of medical data allows improved semantic interoperability. The challenges in mapping some of French terminologies and codes (ICD10FR, CIP13, UCD) to the OMOP vocabularies (ICD, RxNorm) are detailed. Besides the limitations of integrating some other French vocabularies, the audit and evaluation of the resulting implementation are promising.

In a mostly manual effort, Colicchio *et al.* [19] develop a collection of concept-relationship-concept tuples to formally represent patients' care context based on 48 semantic relationships and 14 distinct classes. With a set of 82 representative tuples validated on clinical data and reviewed by experts, they build a pragmatic basis for improving the design of EHRs. This work suggests that a limited number of representations is needed for an efficient coverage of possible patient representation in EHRs.

4 Conclusions

In the KRM selection for 2020, we observe significant contributions in research on knowledge representation addressing the challenge of ontology curation. As in previous years, bioinformatics, machine learning and EHRs are still of major interest in the KRM community. Knowledge representation remains key to advance machine learning by providing context and to develop novel bioinformatics metrics. As in 2019, representations serve a great variety of applications across many medical domains, with actionable results and now with growing adhesion to the open science initiative.

Acknowledgements

We would like to acknowledge the support of Lina Soualmia, Fleur Mougin, Adrien Ugon, Martina Hutter, and the whole IMIA Yearbook Editorial Committee as well as the numerous reviewers in the selection process of the KRM best papers.

References

1. Dhombres F, Charlet J. Knowledge Representation and Management, It's Time to Integrate! Yearb Med Inform 2017;26(1):148-51.
2. Dhombres F, Charlet J, Section Editors for the IYSoKR, Management. 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, Section Editors for the IMIA Yearbook Section on Knowledge Representation and Management. 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.
4. Dhombres F, Charlet J, Section Editors for the IYSoKR, Management. Design and Use of Semantic Resources: Findings from the Section on Knowledge Representation and Management of the 2020 International Medical Informatics Association Yearbook. Yearb Med Inform 2020;29(1):163-8.
5. Zheng F, Shi J, Yang Y, Zheng WJ, Cui L. A transformation-based method for auditing the IS-A hierarchy of biomedical terminologies in the Unified Medical Language System. J Am Med Inform Assoc 2020;27(10):1568-75.
6. Slater LT, Gkoutos GV, Hoehndorf R. Towards semantic interoperability: finding and repairing hidden contradictions in biomedical ontologies. BMC Med Inform Decis Mak 2020;20(Suppl 10):311.
7. Le DH. UFO: A tool for unifying biomedical ontology-based semantic similarity calculation, enrichment analysis and visualization. PLoS One 2020;15(7):e0235670.
8. Robinson PN, Ravanmehr V, Jacobsen JOB, Danis D, Zhang XA, Carmody LC, et al. Interpretable Clinical Genomics with a Likelihood Ratio Paradigm. Am J Hum Genet 2020;107(3):403-17.
9. Hier DB, Brint SU. A Neuro-ontology for the neurological examination. BMC Med Inform Decis Mak 2020;20(1):47.
10. Hou L, Wu M, Kang HY, Zheng S, Shen L, Qian Q, et al. PMO: A knowledge representation model towards precision medicine. Math Biosci Eng 2020;17(4):4098-114.
11. Kulmanov M, Smaili FZ, Gao X, Hoehndorf R. Semantic similarity and machine learning with ontologies. Brief Bioinform 2020;bbaa199.
12. Sousa RT, Silva S, Pesquita C. Evolving knowledge graph similarity for supervised learning in complex biomedical domains. BMC Bioinformatics 2020;21(1):6.
13. Alag S. Unique insights from ClinicalTrials.gov by mining protein mutations and RSids in addition to applying the Human Phenotype Ontology. PLoS One 2020;15(5):e0233438.
14. Blatti C, 3rd, Emad A, Berry MJ, Gatzke L, Epstein M, Lanier D, et al. Knowledge-guided analysis of "omics" data using the KnowEnG cloud platform. PLoS Biol 2020;18(1):e3000583.
15. Beck T, Shorter T, Brookes AJ. GWAS Central: a comprehensive resource for the discovery and comparison of genotype and phenotype data from genome-wide association studies. Nucleic Acids Res 2020;48(D1):D933-D40.
16. Shefchek KA, Harris NL, Gargano M, Matentzoglu N, Unni D, Brush M, et al. The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res 2020;48(D1):D704-D15.
17. Jani A, Liyanage H, Okusi C, Sherlock J, Hoang U, Ferreira F, et al. Using an Ontology to Facilitate More Accurate Coding of Social Prescriptions Addressing Social Determinants of Health: Feasibility Study. J Med Internet Res 2020;22(12):e23721.
18. Lamer A, Depas N, Doutreligne M, Parrot A, Verloop D, Defebvre MM, et al. Transforming French Electronic Health Records into the Observational Medical Outcome Partnership's Common Data Model: A Feasibility Study. Appl Clin Inform 2020;11(1):13-22.
19. Colicchio TK, Dissanayake PI, Cimino JJ. Formal representation of patients' care context data: the path to improving the electronic health record. J Am Med Inform Assoc 2020;27(11):1648-57.

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: Content Summaries of Selected Best Papers for the IMIA Yearbook 2021, Section Knowledge Representation and Management

Le DH

UFO: A tool for unifying biomedical ontology-based semantic similarity calculation, enrichment analysis and visualization

PLoS One 2020;15(7):e0235670

In this article, Le presents a unified tool to support semantic similarity-based research called UFO. Ontology-based similarity has become a routine approach in many applications (ontology curation, enrichment, decision support, gene association studies…). Similarity measurements in ontologies can be performed using many metrics and methods, and current solutions are scattered over different tools and platforms.

UFO is implemented as an app for the Cytoscape platform (open-source software for visualizing complex networks and integrating data in molecular and systems biology, genomics, and proteomics) and supports the OBO format. This article (and supplementary material) describes all UFO features and refers to case studies of relevant implementations: human disease phenotype similarity based on Human Phenotype Ontology (HPO), prediction of disease-associated genes and protein complexes based on gene and protein complex similarity networks using Gene ontology, prediction of disease-associated genes and long non-coding RNAs based on disease similarity network using HPO and Disease ontology, and enrichment analysis with HPO.

The main detailed functions are similarity calculation, enrichment analysis, and visualization. The similarity matrices can be calculated between terms (with 11 metrics, node and/or edge-based), between annotated entities (pairwise or groupwise) and between two sets of entities. Statistical tests (binomial or Fischer's exact) can be applied to a set of entities to search for additional salient terms to enrich the set. Graph visualization facilitates the understanding of the relationships among selected terms, their ancestors (e.g., shared ancestors) and descendants, or among entities (similarity networks).

This tool brings to the KRM community a unified solution for semantic similarity research. Now, only OBO format ontologies are supported (other ontology formats will need conversion) and the tool is tailored for a molecular and systems biology platform. However, this solution supports any application domain.

Robinson PN, Ravanmehr V, Jacobsen JOB, Danis D, Zhang XA, Carmody LC, Gargano MA, Thaxton CL, Core UNCB, Karlebach G, Reese J, Holtgrewe M, Kohler S, McMurry JA, Haendel MA, Smedley D

Interpretable clinical genomics with a likelihood ratio paradigm

Am J Hum Genet 2020;107(3):403-17

In this paper, Robinson *et al.* address the phenotype-driven prioritization of variants with a metric providing robust estimates of the strength of the predictions of candidate genes or diseases, beyond the usual placement in a ranked list.

They present a novel algorithm, the LIkelihood Ratio Interpretation of Clinical AbnormaLities (LIRICAL), that calculates the likelihood ratio (LHR) of each observed or excluded phenotypic abnormality. For each candidate diagnosis, LIRICAL calculates the extent to which each phenotypic abnormality (and if available genotype) is consistent with the diagnosis. Phenotypic abnormalities are represented by Human Phenotype Ontology (HPO) terms and the LHR calculations are derived from the subsumption hierarchies in the HPO. In the methods section, the algorithm is entirely described.

This work illustrates how the structure of a knowledge representation (HPO) can contribute to a bioinformatics workflow. The performances of LIRICAL are demonstrated to be state-of-the-art, on simulated data from 384 published case reports and data from 116 solved cases from the 100,000 Genomes Project.

LIRICAL is available for academic use for free, and source code can be downloaded on a GitHub repository.

Slater LT, Gkoutos GV, Hoehndorf R

Towards semantic interoperability: finding and repairing hidden contradictions in biomedical ontologies

BMC Med Inform Decis Mak 2020;20(Suppl 10):311

In this paper, the authors present a method to identify hidden unsatisfiabilities in an ontology that arise when combined with other ontologies. They identified a large set of inconsistencies across a broad range of biomedical ontologies.

One way to combine ontologies is to use the MIREOT (Minimum Information to Reference an External Ontology Term) guidelines which were originally developed to support inclusion of classes from biomedical ontologies. MIREOT has become a standard for term re-use and inclusion throughout the biomedical ontology community. The authors advocate that, while this method allows ontologies to reuse classes in a scalable and efficient manner, the inclusion of external classes without the context of the external ontology's axioms means that contradictions may arise. These contradictions cannot be detected using an automated reasoner that evaluates only the target ontology.

The authors use automated reasoning to determine whether unsatisfiable classes are present. In addition, they designed a novel algorithm that suggest justifications for contradictions across large and complex ontologies. Their experiments identify contradictions that lead to unsatisfiable classes in the OBO ontologies and highlight the axioms that can be removed to solve most cases of unsatisfiability.

Such a work is important since researchers often import pieces of ontologies without considering the associated axioms. It is also important because it shows the challenge of maintaining a coherent group of ontologies in a large repository like the OBO ontologies.

Zheng F, Shi J, Yang Y, Zheng WJ, Cui L

A transformation-based method for

auditing the IS-A hierarchy of biomedical terminologies in the Unified Medical Language System

J Am Med Inform Assoc 2020;27(10):1568-75

In this paper, Zheng *et al.* use the rich knowledge provided by the Unified Medical Language System (UMLS) for auditing and improving the quality of its source terminologies. Given a concept name in the UMLS, they first identify its base and secondary noun chunks. For each identified noun chunk, they generate replacement candidates that are more general than the noun chunk. Then, they replace the noun chunks with their replacement candidates to generate new potential concept names that may serve as supertypes of the original concept. If a newly generated name is an existing concept name in the same source terminology as the original concept, then a potentially missing IS-A relation between the original and the new concept is identified.

This method gives very good results during the tests: a total of 39,359 potentially missing IS-A relations were detected in 13 source terminologies. Domain experts evaluated a random sample of 200 potentially missing IS-A relations identified in SNOMED CT and 100 in Gene ontology. A total of 173 of 200, and 63 of 100 potentially missing IS-A relations were confirmed by domain experts, indicating that the method achieved a precision of 86.5% and 63% for SNOMED CT and for Gene ontology, respectively.

This method is very suitable for large ontologies that are assemblages of thoughts and teams over time and for which it is difficult to audit the whole resource. It would be interesting to test this method on smaller ontologies.