Article published online: 2023-12-26

Knowledge Representation and Management 2022: Findings in Ontology Development and Applications

Jean Charlet[1,2], Licong Cui[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] AP-HP, DRCI, Paris, France

[3] McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

Summary

Objectives: To select, present, and summarize the best papers in 2022 for the Knowledge Representation and Management (KRM) section of the International Medical Informatics Association (IMIA) Yearbook.

Methods: We conducted PubMed queries and followed the IMIA Yearbook guidelines for performing biomedical informatics literature review to select the best papers in KRM published in 2022.

Results: We retrieved 1,847 publications from PubMed. We nominated 15 candidate best papers, and two of them were finally selected as the best papers in the KRM section. The topics covered by the candidate papers include ontology and knowledge graph creation, ontology applications, ontology quality assurance, ontology mapping standard, and conceptual model.

Conclusions: In the KRM best paper selection for 2022, the candidate best papers encompassed a broad range of topics, with ontology and knowledge graph creation remaining a considerable research focus.

Keywords

Knowledge representation and management; ontology; knowledge graph; International Medical Informatics Association

Yearb Med Inform 2023:225-9
http://dx.doi.org/10.1055/s-0043-1768747

1 Introduction

The year 2022 has yielded an abundant number of publications in the field of 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 [1-6]. In this synopsis, we present the best paper selection process for the KRM section of the 2023 International Medical Informatics Association (IMIA) Yearbook and summarize the findings of the nominated candidate best papers.

2 Paper Selection Method

We performed literature search based on PubMed/MEDLINE to identify KRM-related papers in the context of medical informatics published in international peer-reviewed journals and conference proceedings indexed by PubMed. We reused last year's query set with Medical Subject Headings (MeSH) descriptors [6]. We considered original research articles published between 01/01/2022 and 12/31/2022 and excluded those with the following publication types: reviews, editorials, comments, case reports, and letters to the editors.

We followed the standard process [7] commonly used by the IMIA Yearbook sections to select the best papers. The selection process involved three steps. The section editors first conducted an initial screen based on the title, abstract, and publication type, and then performed a second review to select a collection of candidate best papers for further peer-review. Each candidate paper was reviewed and evaluated by two IMIA Yearbook editors, two section editors, and two external reviewers. The evaluation criteria included topic's importance to medical and health informatics, scientific and/or practical impact of the paper to the topic, quality of scientific and/or technical content, originality and innovativeness, coverage of related literature, and organisation and clarity of presentation. The final selection of the best papers was achieved during a meeting of the whole editorial board, based on the result of peer reviews and the report of the section editors.

3 Results

3.1 Best Paper Selection for 2022

We retrieved a total of 1,847 KRM-related papers published in 2022 from PubMed, which is more than last year (1,231). We obtained 387 papers after the section editors' initial screening. The section editors further reviewed these papers jointly and reached a consensus list of 15 papers, which were nominated as the candidate best papers [8-22]. External reviewers, IMIA Yearbook editors and section editors further evaluated these 15 papers and finally selected two best papers (see Table 1).

Table 1 Selection of best papers for the 2023 IMIA Yearbook of Medical Informatics for the Knowledge Representation and Management section. The articles are listed in alphabetical order by the first author's surname.

Section Knowledge Representation and Management
▪ Kaliyaperumal R, Wilkinson MD, Moreno PA, Benis N, Cornet R, dos Santos Vieira B, Dumontier M, Bernabé CH, Jacobsen A, Le Cornec C, Godoy MP. Semantic modelling of common data elements for rare disease registries, and a prototype workflow for their deployment over registry data. J Biomed Semantics 2022;13(1):9. ▪ Matentzoglu N, Goutte-Gattat D, Tan SZ, Balhoff JP, Carbon S, Caron AR, Duncan WD, Flack JE, Haendel M, Harris NL, Hogan WR. Ontology Development Kit: a toolkit for building, maintaining and standardizing biomedical ontologies. Database 2022:baac087.

In the first paper, Kaliyaperumal *et al.* [8] presented a semantic model based on the Semantic Science Integrated Ontology to represent common data elements (CDEs) for rare disease registries across Europe. This work aimed at addressing a significant challenge faced by the European Platform on Rare Disease Registration to integrate scarce patient data from hundreds of rare disease registries in compliance with the FAIR (Findable, Accessible, Interoperable, and Reusable) principles. To this end, the authors mapped the CDE concepts and their value sets into standardized ontologies including the Orphanet Rare Disease Ontology and the Human Phenotype Ontology (HPO). In addition, the authors built an exemplar Extract/Transform/Load (ETL) pipeline to export data from source registries.

The second article is a contribution by Matentzoglu *et al.* [9], where the authors provided an overview of the Ontology Development Kit (ODK), a Docker-based tool for creating and managing ontologies. ODK consists of a toolbox equipped with diverse tools for ontology editors to build, test, and release ontologies as well as a set of standardized and executable ontology-engineering workflows following the best practices recommended by the Open Biological and Biomedical Ontology (OBO) Foundry. ODK also empowers non-expert users to create and edit ontologies with little training required. ODK has been used for maintaining over 70 ontologies, such as the Cell Ontology, HPO, and Uberon.

Additional content summaries of the two best papers can be found in the appendix of this synopsis.

Considering all the 15 candidate best papers, they can be categorized into five topic areas: ontology and knowledge graph creation, ontology applications, ontology quality assurance, ontology mapping standard, and conceptual model. We also note that many articles of the 15 papers focus on following the FAIR principles. Last but not least, the reuse of reference ontologies or ontologies specific to particular data (PROV-O, BFO, HPO, …etc.) is becoming commonplace in articles, and contributes to the *Interoperability* dimension of the FAIR principles.

3.2 Ontology and Knowledge Graph Creation

Similar to last year [6], ontology and knowledge graph creation continues to receive significant research attention. Seven out of the 15 candidate best papers are with regard to developing ontologies and one is about knowledge graph creation. While the best paper from Matentzoglu *et al.* [9] focuses on tool development in support of creating and managing ontologies for the ontology community, the other six candidate papers contributed to the creation of domain-specific ontologies.

In the candidate paper from Azzi *et al.* [10], the authors have developed a Pneumonia Diagnosis Ontology (PNADO) leveraging clinical practice guidelines and reusing related ontologies from OBO Foundry and BioPortal. The PNADO was the first pneumonia diagnosis ontology to represent different aspects of pneumonia including subtypes, symptoms, and lab tests. Both data-driven evaluation and domain expert evaluation were performed.

Cardoso *et al.* [11] presented the Data Management Plan (DMP) Common Standard Ontology (DCSO), facilitating researchers to systematically manage data and metadata following FAIR principles. The DCSO was proposed to overcome the limitations of the existing machine-actionable DMP specification in the JavaScript Object Notation (JSON) format, created by the Research Data Alliance (Europe) DMP Common Standards working group, which lacks explicit links to relevant data models or ontologies, a standardized approach for describing controlled vocabularies, and a clear mechanism to differentiate between the core specification and its extensions.

Gillespie *et al.* [12] proposed the Neuron Phenotype Ontology (NPO) for naming and representing large quantities of neurons in the nervous system. It offers a FAIR framework (modelling a neuron type as a collection of key phenotypes) to represent the intricate cellular phenotypes produced by neuroscientists engaged in the US Brain Initiative Cell Census Network, Human Cell Atlas, Blue Brain Project, and other individual and large initiatives.

The candidate paper from Fisher *et al.* [13] describes the development of the Xenopus Phenotype Ontology (XPO). Incorporating related information from the Unified Phenotype Ontology, Xenopus Anatomy Ontology, Phenotype and Trait Ontology, and Gene Ontology empowers the XPO with comprehensive phenotypic curation and linkage to phenotype data from other model organisms and human diseases. This exemplifies best practices utilized to address the intrinsic challenges involved in harmonizing phenotype data across diverse species.

The paper from González-Eras *et al.* [14] presented the engineering process of integrating multiple existing COVID-19 related ontologies to construct a new ontology called COVID-19 Pandemic Ontology, comprehensively covering various aspects of the infectious disease. The integration process involved matching, linking, and merging tasks. The resulting ontology was tested through different case studies, demonstrating its ability to derive useful information about the pandemic.

In the paper from Lokala *et al.* [15], the authors reported the process of development, evaluation and application of the Drug Abuse Ontology (DAO) for analyzing web-based data including social media data, web forums, and dark web data. The DAO has been primarily utilized for knowledge extraction from these web-based platforms to inform substance use epidemiology research.

Morse *et al.* [16] created a Postpartum Depression Ontology (PDO), which encompassed relevant comorbidities, symptoms, treatments, and risk factors associated with postpartum depression (PPD). The PDO incorporated both structured (e.g., International Classification of Diseases versions 9 and 10 codes) and unstructured information (e.g., synonyms of symptoms without standardized codes), aiming to assist in identifying postpartum depression (PPD) patients and supporting PPD research based on electronic health record data.

Wood *et al.* [17] developed a comprehensive biomedical knowledge graph called RTX-KG2 by integrating 70 external knowledge sources including the UMLS, SemMedDB, ChEMBL, DrugBank, UniProtKB and Reactome, and offering a web-based API to query the integrated knowledge graph. RTX-KG2 complies with the standard Biolink model to ensure interoperability and includes provenance information. It has been utilized in the NCATS Biomedical Data Translator project to facilitate computational reasoning in translational science.

3.3 Ontology Applications

Three out of 15 candidate best papers are with regard to ontology applications. We consider the best paper from Kaliyaperumal et al. [8] as an ontology application, since it leverages an ontological model to represent rare disease CDEs and support data integration from distributed data registries.

Another ontology application paper is from Rosenau *et al.* [18], who have demonstrated an automatic approach to generating ontological concepts allowing users to utilize these concepts as criteria to perform federated cohort queries against Fast Healthcare Interoperability Resources (FHIR)-formatted data in the German Corona Consensus Dataset (GECCO), which was developed to address semantic interoperability challenges at a national scale during the COVID-19 pandemic.

The study by Yan *et al.* [19] focused on developing a neural network model for recognizing HPO concepts from free text to support phenotype-based analyses. The model capitalizes on the valuable information present in the ontology (*e.g.*, terms, definitions, and comments) and leverages a pre-trained re-ranking model to enhance overall performance of HPO concept recognition tools. This study sets an example to leverage rich ontological information to empower deep learning models.

3.4 Ontology Quality Assurance

In the candidate paper from Burse *et al.* [20], the authors presented a lexical-based method for quality assurance of SNOMED CT. In particular, the stopwords in concept names were explored to identify potentially missing logical definitions for partially defined concepts (*i.e.*, concepts that are not sufficiently defined). This was achieved by leveraging the logical definitions of those fully defined concepts (*i.e.*, concepts that are sufficiently defined) which are lexically and semantically similar to the partially defined concepts.

3.5 Ontology Mapping Standard

Matentzoglu *et al.* [21] introduced a Simple Standard for Sharing Ontological Mappings (SSSOM) to address the challenge of lacking precise description of mapping metadata (e.g., narrow or broad match between two entities). SSSOM introduces a machine-readable vocabulary to explicitly describe imprecision, inaccuracy, and incompleteness in mappings. It offers a simple table-based format that can be easily integrated into existing data science pipelines in compliance with Linked Data principles. This could serve as a model for the biomedical terminology, database and ontology mapping communities to adopt and follow.

3.6 Conceptual Model

In the paper from Bernasconi *et al.* [22], the authors developed the Ontological Viral Conceptual Model (OntoVCM), based on the Viral Conceptual Model (VCM) that represents virus sequencing, to facilitate semantic interoperability of virology and genomic research data and resources. OntoVCM offers conceptual clarity and ontological grounding from a specific viewpoint concerning viral information encompassing infection, sampling, sequencing, annotations, and depositing.

4 Conclusions

The 15 candidate best papers selected for the KRM section in the year 2022 cover diverse topics including ontology and knowledge graph creation, ontology applications, ontology quality assurance, ontology mapping standard, and conceptual model. Among the two best papers, one is about an ontology development toolkit and the other pertains to ontology-based modelling of CDEs in rare diseases. Seven of the candidate papers were devoted to the development of domain-specific ontologies, namely Pneumonia Diagnosis Ontology, Data Management Plan Common Standard Ontology, Neuron Phenotype Ontology, Xenopus Phenotype Ontology, COVID-19 Pandemic Ontology, Drug Abuse Ontology, and Postpartum Depression Ontology.

Acknowledgements

We would like to acknowledge the support of Fleur Mougin, Adrien Ugon, Lina Soualmia, Kate Fultz Hollis, and the whole IMIA Yearbook editorial team as well as the reviewers in the selection process of the 2022 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. doi: 10.15265/IY-2017-030.
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. doi: 10.1055/s-0038-1667078.

3. 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. doi: 10.1055/s-0039-1677933.
4. Dhombres F, Charlet J. 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. doi: 10.1055/s-0040-1702010.
5. Dhombres F, Charlet J. Knowledge Representation and Management: Interest in New Solutions for Ontology Curation. Yearb Med Inform 2021;30(01):185-90. doi: 10.1055/s-0041-1726508.
6. Cui L, Dhombres F, Charlet J. Knowledge Representation and Management: Notable Contributions in 2021. Yearb Med Inform 2022;31(01):236-40. doi: 10.1055/s-0042-1742523.
7. Lamy JB, Séroussi B, Griffon N, Kerdelhué G, Jaulent MC, Bouaud J. Toward a formalization of the process to select IMIA Yearbook best papers. Methods Inf Med 2015;54(02):135-44. doi: 10.3414/ME14-01-0031.
8. Kaliyaperumal R, Wilkinson MD, Moreno PA, Benis N, Cornet R, dos Santos Vieira B, et al. Semantic modelling of common data elements for rare disease registries, and a prototype workflow for their deployment over registry data. J Biomed Semantics 2022;13(1):9. doi: 10.1186/s13326-022-00264-6.
9. Matentzoglu N, Goutte-Gattat D, Tan SZ, Balhoff JP, Carbon S, Caron AR, et al. Ontology Development Kit: a toolkit for building, maintaining and standardizing biomedical ontologies. Database 2022:baac087. doi: 10.1093/database/baac087.
10. Azzi S, Michalowski W, Iglewski M. Developing a pneumonia diagnosis ontology from multiple knowledge sources. Health Informatics J 2022;28(2). doi: 10.1177/14604582221083850.
11. Cardoso J, Castro LJ, Ekaputra FJ, Jacquemot MC, Suchánek M, Miksa T, et al. DCSO: towards an ontology for machine-actionable data management plans. J Biomed Semantics 2022;13(1):21. doi: 10.1186/s13326-022-00274-4.
12. Gillespie TH, Tripathy SJ, Sy MF, Martone ME, Hill SL. The Neuron Phenotype Ontology: a FAIR approach to proposing and classifying neuronal types. Neuroinformatics 2022;20(3):793-809. doi: 10.1007/s12021-022-09566-7.
13. Fisher ME, Segerdell E, Matentzoglu N, Nenni MJ, Fortriede JD, Chu S, et al. The Xenopus phenotype ontology: bridging model organism phenotype data to human health and development. BMC Bioinformatics 2022;23(1):99. doi: 10.1186/s12859-022-04636-8.
14. González-Eras A, Dos Santos R, Aguilar J, Lopez A. Ontological engineering for the definition of a COVID-19 pandemic ontology. Inform Med Unlocked 2022;28:100816. doi: 10.1016/j.imu.2021.100816.
15. Lokala U, Lamy F, Daniulaityte R, Gaur M, Gyrard A, Thirunarayan K, et al. Drug abuse ontology to harness web-based data for substance use epidemiology research: ontology development study. JMIR Public Health Surveill 2022;8(12):e24938. doi: 10.2196/24938.
16. Morse RB, Bretzin AC, Canelón SP, D'Alonzo BA, Schneider AL, Boland MR. Design and Evaluation of a Postpartum Depression Ontology. Appl Clin Inform 2022;13(01):287-300. doi: 10.1055/s-0042-1743240.
17. Wood EC, Glen AK, Kvarfordt LG, Womack F, Acevedo L, Yoon TS, et al. RTX-KG2: a system for building a semantically standardized knowledge graph for translational biomedicine. BMC Bioinformatics 2022;23(1):400. doi: 10.1186/s12859-022-04932-3.
18. Rosenau L, Majeed RW, Ingenerf J, Kiel A, Kroll B, Köhler T, et al. Generation of a Fast Healthcare Interoperability Resources (FHIR)-based ontology for federated feasibility queries in the context of COVID-19: feasibility study. JMIR Med Inform 2022;10(4):e35789. doi: 10.2196/35789.
19. Yan S, Luo L, Lai PT, Veltri D, Oler AJ, Xirasagar S, et al. PhenoRerank: A re-ranking model for phenotypic concept recognition pre-trained on human phenotype ontology. J Biomed Inform 2022;129:104059. doi: 10.1016/j.jbi.2022.104059.
20. Burse R, McArdle G, Bertolotto M. Targeting stopwords for quality assurance of SNOMED-CT. Int J Med Inform 2022;167:104870. doi: 10.1016/j.ijmedinf.2022.104870.
21. Matentzoglu N, Balhoff JP, Bello SM, Bizon C, Brush M, Callahan TJ, et al. A simple standard for sharing ontological mappings (SSSOM). Database 2022:baac035. doi: 10.1093/database/baac035.
22. Bernasconi A, Guizzardi G, Pastor O, Storey VC. Semantic interoperability: ontological unpacking of a viral conceptual model. BMC bioinformatics 2022;23(Suppl 11):491. doi: 10.1186/s12859-022-05022-0.

Correspondence to:
Jean Charlet
DRCI/AP-HP & LIMICS UMR_S 1142
Paris, France
E-mail: jean.charlet@sorbonne-universite.fr

Appendix: Content Summaries of Selected Best Papers Published in 2022 for the IMIA Yearbook, Section Knowledge Representation and Management

Kaliyaperumal R, Wilkinson MD, Moreno PA, Benis N, Cornet R, dos Santos Vieira B, Dumontier M, Bernabé CH, Jacobsen A, Le Cornec C, Godoy MP

Semantic modelling of common data elements for rare disease registries, and a prototype workflow for their deployment over registry data

J Biomed Semantics 2022;13(1):9. doi: 10.1186/s13326-022-00264-6

In this article, the members of the EU Platform on Rare Disease Registration put forward a working procedure to ensure the interoperability and FAIRification of data from the common data representation model, the CDE. The motivation is precise and industrial: the 16 CDEs must be implemented in all EU Rare Disease registries. To do this, they created semantically grounded models to represent each of the CDEs, using the SemanticScience Integrated Ontology as the core framework for representing the entities and their relationships. Within that framework, they mapped the concepts represented in the CDEs, and their possible values, to domain ontologies such as the Orphanet Rare Disease Ontology, Human Phenotype Ontology and National Cancer Institute Thesaurus. Finally, they created an exemplar, reusable ETL pipeline that they will be deploying over non-coordinating data repositories to assist them in creating model-compliant FAIR data without requiring site-specific coding, nor expertise in Linked Data or FAIR. This ETL refers to alignment description languages and execution models with YAML and YARRRML. With the aforementioned ontologies, the authors describe an industrial process at the knowledge level, perfectly operational, and in total respect of FAIR principles.

Matentzoglu N, Goutte-Gattat D, Tan SZ, Balhoff JP, Carbon S, Caron AR, Duncan WD, Flack JE, Haendel M, Harris NL, Hogan WR

Ontology Development Kit: a toolkit for building, maintaining and standardizing biomedical ontologies

Database 2022:baac087. doi: 10.1093/database/baac087

In this paper, the authors provided an overview of the Ontology Development Kit (ODK), a Docker-based tool for creating and managing ontologies in the biomedical domain. ODK consists of a toolbox equipped with diverse tools for ontology editors to build, test, and release ontologies as well as a set of standardized and executable ontology-engineering workflows following the best practices recommended by the Open Biological and Biomedical Ontology (OBO) Foundry. Moreover, the authors attempt to highlight how ODK stimulates standardization efforts in the Knowledge Representation (KR) community. The authors have already observed significantly lower error rates in many of the ontologies that use the ODK, thanks to the ability of the automated testing system provided by the ODK to catch errors early on. Lastly, they seek to harmonize the representation of ontology release files through the use of standard release workflows, which result in standard release serializations and metadata to make ontologies more FAIR and interoperable. ODK is not an ontology editor. It lets ontology creators use an editor like Protégé. ODK supports a templating system such as ROBOT or others. This is not the first time that researchers in the KR domain have proposed an ontology development environment: NEON Toolkit or WebODE are proposals that have been around for over 10 years. ODK is more recent, takes into account modular approaches that are now important, and respects the de facto standards of ontology engineering. Time will tell whether it will become the reference tool. In any case, we need ontology development methodologies and softwares to implement them.