# Knowledge Representation and Management, It's Time to Integrate!

## Findings from the IMIA Yearbook Section on Knowledge Representation and Management

F. Dhombres[1, 2], J. Charlet[1, 3], Section Editors for the IMIA Yearbook Section on Knowledge Representation and Management

[1] INSERM, UMR_S 1142, LIMICS, 75006 Paris, France
Sorbonne Universités, UPMC Univ. Paris 06, UMR_S 1142, LIMICS, 75006 Paris, France
Université Paris 13, Sorbonne Paris Cité, UMR_S 1142, LIMICS, 93430 Villetaneuse, France
[2] Department of Fetal Medicine, Armand Trousseau Hospital, APHP, 75012 Paris, France
[3] AP-HP, Department of Clinical Research and Innovation, Paris, France

### Summary

**Objectives:** To select, present, and summarize the best papers published in 2016 in the field of Knowledge Representation and Management (KRM).

**Methods:** A comprehensive and standardized review of the medical informatics literature was performed based on a PubMed query.

**Results:** Among the 1,421 retrieved papers, the review process resulted in the selection of four best papers focused on the integration of heterogeneous data via the development and the alignment of terminological resources. In the first article, the authors provide a curated and standardized version of the publicly available US FDA Adverse Event Reporting System. Such a resource will improve the quality of the underlying data, and enable standardized analyses using common vocabularies. The second article describes a project developed in order to facilitate heterogeneous data integration in the i2b2 framework. The originality is to allow users integrate the data described in different terminologies and to build a new repository, with a unique model able to support the representation of the various data. The third paper is dedicated to model the association between multiple phenotypic traits described within the Human Phenotype Ontology (HPO) and the corresponding genotype in the specific context of rare diseases (rare variants). Finally, the fourth paper presents solutions to annotation-ontology mapping in genome-scale data. Of particular interest in this work is the Experimental Factor Ontology (EFO) and its generic association model, the Ontology of Biomedical AssociatioN (OBAN).

**Conclusion:** Ontologies have started to show their efficiency to integrate medical data for various tasks in medical informatics: electronic health records data management, clinical research, and knowledge-based systems development.

### Keywords

Knowledge representation (computer); biomedical ontologies; controlled vocabularies; information storage and retrieval; data integration

Yearb Med Inform 2017:148-51
http://dx.doi.org/10.15265/IY-2017-030
Published online August 18, 2017

## Introduction

The year 2016 has produced a large amount of publications related to the field of Knowledge Representation and Management (KRM) in Medicine. KRM focuses on the development of techniques to be used and leveraged in other medical informatics domains. During the last years, we observed a growing interest in integrating medical data to ensure interoperability between heterogeneous sources of data [1-4]. Not surprisingly, the use of knowledge representation models, especially ontologies, has become a significant approach for enabling complex integration tasks.

In this paper, we present the papers published in 2016 in the KRM domain, selected as the best papers because of their impact or the novelty of the approach they provide in the medical knowledge representation and management field.

## Paper Selection Method

We conducted the selection of KRM papers in PubMed/MELDINE based on the query used in the previous edition of the IMIA Yearbook. We followed a generic method, commonly used in all sections of the Yearbook, defined in [5]. As for the last four years, the search was performed on MEDLINE by querying PubMed. Our query includes MeSH descriptors related to the 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 2016 (from 01/01/2016 to 12/31/2016) were considered; we excluded the following publications types: reviews, editorials, comments, letters to the editors. We limited the search on major MeSH descriptors (for example "biomedical ontologies [MAJR]") to avoid a large set of articles and we completed it by non-MeSH terms searched on the titles and abstracts of articles (for example "terminologies [TIAB]").

The selection of best papers was performed in a three-step process on the papers returned by the query. In the first step, the section editors reviewed all the returned papers on the basis of titles, abstracts, and types of publication to establish a short list of 15 candidate best papers. In the second step, five experts (including the section editors) reviewed the candidate best papers using

the IMIA Yearbook quality criteria scoring method. More specifically, the following criteria were evaluated: significance to medical and health informatics, quality of scientific content, originality and innovativeness, coverage of the related literature, organization and quality of the presentation. The final step of the selection was achieved during the editorial board meeting, taking into account the external reviews and the report of section editors.

# Results

For 2016, the KRM query retrieved 1,421 citations from PubMed. The section editors achieved a first selection of 100 papers based on titles and abstracts. After a second review of this set of papers, a selection of 15 candidate best papers was established [6-20]. Five reviewers reviewed these pre-selected papers to select the best four final papers [6-9].

The four best papers of 2016 demonstrated the added-value of ontology-based integration approaches for phenotype-genotype association mining [6], for clinical data from electronic health records (EHRs) integration and analysis [8], for efficient reference dataset production [7, 21] and introduced a new statistical method for phenotyping rare genetic disorders [9]. Table 1 lists the four papers selected as best papers for the section Knowledge Representation and Management. A brief summary of each one can be found in the appendix of this synopsis.

Among the other selected papers, we observed several research directions within the KRM field, mostly focusing on data integration leveraging ontology-based annotation. Interestingly, many other articles published in 2016 leveraged semantic representations in particular in bioinformatics or natural language processing.

## KRM Solutions for EHR Data Integration

In addition to the solutions for i2b2 presented by Bauer et al. [8], three other papers among the candidate best papers had a specific focus on clinical data management. These papers show different ways to implement Knowledge Organization Systems (KOS) and sometimes ontologies to organize EHR and clinical data.

Klann et al. [17] proposed using the i2b2 data warehouse as a hub, to rapidly reconfigure data to meet new analytical requirements without new ETL programming. The originality of this approach is in the generation of a PCORnet Common Data Model (CDM), which acts as a pivot representation.

Hochheiser et al. [13] translated a subset of the Fast Healthcare Interoperability Resources (FHIR) in OWL2 and extended it with terms from the National Cancer Institute (NCI) thesaurus. The resulting model supports cancer phenotype integration in clinical documents.

Johnson et al. [15] demonstrated the use of a data quality ontology to assess the quality of EHR data. Improvement in research based on EHR data might result from shared quality metrics and from an automated quality assessment using a data quality ontology.

Finally, in his survey paper for the KRM section of the 2017 IMIA Yearbook [22], Rosenbloom *et al.* identified the representation of clinical knowledge as a major trend in the past year, with a perspective of wide-scale EHR data integration.

## Knowledge Bases and Integrated Portals

Besides the direct use of clinical data, the KRM community develops integrated semantic resources to support research and clinical practice. In the field of adverse drug event research, Banda *et al.* [6] developed an integrated and open resource supporting the future updates of the FDA (US Food Drug Administration) dataset. Other knowledge bases and information portals of interest in Medicine are described in the four articles below.

Hayman *et al.* [12] developed a curated portal for multispecies gene-disease relationships descriptions centered on the rat genome. This research-oriented resource introduces a dedicated ontology with annotations from the Rat Genome Database and also from ClinVar and OMIM (Online Mendelian Inheritance in Man).

Hoffman *et al.* [14] from the Clinical Pharmacogenetics Implementation Consortium Informatics Working Group established a set of principles for developing efficient knowledge bases in pharmacogenomics to support precision medicine.

Workman *et al.* [20] described Spark, a graphical knowledge discovery application based on Serendipitous Knowledge Discovery studies and data structures known as semantic predications. The source of knowledge for this application is the Semantic MEDLINE database, containing over 70 million predications (stored as triples), extracted from all PubMed citations and abstracts.

Saunders *et al.* [18] proposed a model of BioOntological Relationship Graph database (BORG), which integrates multiple sources of genomic and biomedical knowledge into

**Table 1** Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2017 in the section 'Knowledge Representation and Management'. The articles are listed in alphabetical order of the first author's surname.

| Section<br>Knowledge Representation and Management |
|---|
| ▪ Banda JM, Evans L, Vanguri RS, Tatonetti NP, Ryan PB, Shah NH. A curated and standardized adverse drug event resource to accelerate drug safety research. Sci Data 2016;3:160026. |
| ▪ Bauer CR, Ganslandt T, Baum B, Christoph J, Engel I, Lobe M, Mate S, Staubert S, Drepper J, Prokosch HU, Winter A, Sax U. Integrated Data Repository Toolkit (IDRT). A Suite of Programs to Facilitate Health Analytics on Heterogeneous Medical Data. Methods Inf Med 2016;55(2):125-35. |
| ▪ Greene D, NIHR BioResource, Richardson S, Turro E. Phenotype Similarity Regression for Identifying the Genetic Determinants of Rare Diseases. Am J Hum Genet 2016;98(3):490-9. |
| ▪ Sarntivijai S, Vasant D, Jupp S, Saunders G, Bento AP, Gonzalez D, Betts J, Hasan S, Koscielny G, Dunham I, Parkinson H, Malone J. Linking rare and common disease: mapping clinical disease-phenotypes to ontologies in therapeutic target validation. J Biomed Semantics 2016;7-8. |

an on-disk semantic network where human genes and their orthologs in mouse and rat are central concepts mapped to ontology terms. This graph was used and analyzed to screen all human genes for potential links to tendinopathy and finally to propose four candidate genes.

## New Ontologies and Solutions for Ontology Development

The KRM community also produced a wide range of methods, models, and tools. The selected article by Greene et al. [9] presents a novel approach, involving knowledge representations and statistical methods to investigate the relationship between genetic variants and an ontological representation of phenotypes. Sarntivijai *et al.* [6] introduced a generic OWL (Web Ontology Language) representation of the disease-phenotype association. Among the other papers reviewed in the KRM section, we identified four papers presenting valuable methods and tools for ontology development. Interestingly, in the development of ontologies, the reusing of other available terminological resources has become a common and shared approach.

Blank *et al.* [10] used up to date methods to develop an ontology for prokaryotic taxonomic descriptions (MicrO). This resource is promising for text mining support and for modeling genotype-phenotype associations in prokaryotes. Their methods correspond to the state-of-the-art of ontology design.

Detwiler *et al.* [11] presented the official conversion of the Foundational Model of Anatomy from Frames to OWL2. The methods for the conversion and for the post conversion clean-up are presented in detail, providing a clear documentation for researchers.

Jupp *et al.* [16] presented a solution (Webulous) to OWL ontology development based on ontology design patterns. It is a Google add-on which is used for the development of few ontologies at the European Bioinformatics Institute.

Thanintorn *et al.* [19] proposed the method of "sketch map" in order to reduce knowledge complexity for precision medicine analytics. The authors tend to demonstrate that the description of pathways with a new KEGG (Kyoto Encyclopedia Genes and Genomes) ontology and their method would be invaluable for hypothesis generation in different domains as precision diagnostics.

# Conclusions

In 2016, the integration of heterogeneous data emerges as a major trend in Knowledge Representation and Management in Medicine. Accordingly, ontology- and terminology-based annotations appear as a fruitful solution to support this integration. Consequently, the construction of ontologies is refined and precise, taking into account existing domain ontologies and terminologies by reuse or alignment. Additionally, these ontologies are integrated into knowledge-based systems (KBS) mainly based on semantic web languages (e.g. Resource Description Framework – RDF) which are successfully developed for different tasks: for the generation of hypotheses of links between diseases and phenotypes, genes or biological pathways, for clinical research, and for the assessment of the quality of EHR data.

**Acknowledgements**

We would like to thank Adrien Hugon and Martina Hutter for her support, the external reviewers for their participation in the selection process of the IMIA Yearbook, and Paul Landais who urgently reviewed articles of the selection.

## References

1. Griffon N, Charlet J, Darmoni S. Knowledge representation and management: towards an integration of a semantic web in daily health practice. Yearb Med Inform 2013;8:155-8.
2. Griffon N, Charlet J, Darmoni SJ. Managing free text for secondary use of health data. Yearb Med Inform. 2014;9:167-9.
3. Charlet J, Darmoni SJ. Knowledge Representation and Management. From Ontology to Annotation. Findings from the Yearbook 2015 Section on Knowledge Representation and Management. Yearb Med Inform. 2015;10(1):134-6.
4. Soualmia LF, Charlet J. Efficient Results in Semantic Interoperability for Health Care. Findings from the Section on Knowledge Representation and Management. Yearb Med Inform 2016(1):184-7.
5. Lamy J-B, Séroussi B, Griffon N, Kerdelhué G, Jaulent M-C, Bouaud J. Toward a formalization of the process to select IMIA Yearbook best papers. Methods Inf Med 2015;54(2):135-44.
6. Sarntivijai S, Vasant D, Jupp S, Saunders G, Bento AP, Gonzalez D, et al. Linking rare and common disease: mapping clinical disease-phenotypes to ontologies in therapeutic target validation. J Biomed Semantics 2016;7:8.
7. Banda JM, Evans L, Vanguri RS, Tatonetti NP, Ryan PB, Shah NH. A curated and standardized adverse drug event resource to accelerate drug safety research. Sci Data 2016;3:160026.
8. Bauer CR, Ganslandt T, Baum B, Christoph J, Engel I, Lobe M, et al. Integrated Data Repository Toolkit (IDRT). A Suite of Programs to Facilitate Health Analytics on Heterogeneous Medical Data. Methods Inf Med 2016;55(2):125-35.
9. Greene D, NIHR BioResource, Richardson S, Turro E. Phenotype Similarity Regression for Identifying the Genetic Determinants of Rare Diseases. Am J Hum Genet 2016;98(3):490-9.
10. Blank CE, Cui H, Moore LR, Walls RL. MicrO: an ontology of phenotypic and metabolic characters, assays, and culture media found in prokaryotic taxonomic descriptions. J Biomed Semantics 2016;7:18.
11. Detwiler LT, Mejino JL, Brinkley JF. From frames to OWL2: Converting the Foundational Model of Anatomy. Artif Intell Med 2016;69:12-21.
12. Hayman GT, Laulederkind SJ, Smith JR, Wang SJ, Petri V, Nigam R, et al. The Disease Portals, disease-gene annotation and the RGD disease ontology at the Rat Genome Database. Database (Oxford) 2016;2016.
13. Hochheiser H, Castine M, Harris D, Savova G, Jacobson RS. An information model for computable cancer phenotypes. BMC Med Inform Decis Mak 2016;16(1):121.
14. Hoffman JM, Dunnenberger HM, Kevin Hicks J, Caudle KE, Whirl Carrillo M, Freimuth RR, et al. Developing knowledge resources to support precision medicine: principles from the Clinical Pharmacogenetics Implementation Consortium (CPIC). J Am Med Inform Assoc 2016;23(4):796-801.
15. Johnson SG, Speedie S, Simon G, Kumar V, Westra BL. Application of An Ontology for Characterizing Data Quality For a Secondary Use of EHR Data. Appl Clin Inform 2016;7(1):69-88.
16. Jupp S, Burdett T, Welter D, Sarntivijai S, Parkinson H, Malone J. Webulous and the Webulous Google Add-On--a web service and application for ontology building from templates. J Biomed Semantics 2016;7:17.
17. Klann JG, Abend A, Raghavan VA, Mandl KD, Murphy SN. Data interchange using i2b2. J Am Med Inform Assoc. 2016;23(5):909-15.
18. Saunders CJ, Jalali Sefid Dashti M, Gamieldien J. Semantic interrogation of a multi knowledge domain ontological model of tendinopathy identifies four strong candidate risk genes. Sci Rep 2016;6:19820.
19. Thanintorn N, Wang J, Ersoy I, Al-Taie Z, Jiang Y, Wang D, et al. Rdf Sketch Maps - Knowledge Complexity Reduction for Precision Medicine Analytics. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing. 2016;21:417-28.

20. Workman TE, Fiszman M, Cairelli MJ, Nahl D, Rindflesch TC. Spark, an application based on Serendipitous Knowledge Discovery. J Biomed Inform 2016;60:23-37.
21. Banda JM, Evans L, Vanguri RS, Tatonetti NP, Ryan PB, Shah NH. Data from: A curated and standardized adverse drug event resource to accelerate drug safety research. Dryad Data Repository; 2016. http://dx.doi.org/10.5061/dryad.8q0s4
22. Rosenbloom ST, Carroll RJ, Warner JL, Matheny ME, Denny JC. Representing Knowledge Consistently Across Health Systems. Yearb Med Inform 2017:139-47.

Correspondence to:
Ferdinand Dhombres, MD, PhD
LIMICS - INSERM U1142
Campus des Cordeliers
5, rue de l'école de médecine
75006 Paris, France
E-mail: ferdinand.dhombres@inserm.fr

# Appendix: Content Summaries of Selected Best Papers for the 2017 IMIA Yearbook, Section Knowledge Representation and Management

### Banda JM, Evans L, Vanguri RS, Tatonetti NP, Ryan PB, Shah NH

### A curated and standardized adverse drug event resource to accelerate drug safety research

### Sci Data 2016;3:160026

This open science paper introduces a large, curated, and publicly available resource for adverse drug event (ADE) research: it includes an ADE dataset with the source code and documentation to be used by the research community. This resource (AEOLUS) derives from the publicly available US Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS) dataset, covering spontaneous reports of adverse drug events since 2012. The resource also integrates a legacy dataset covering 2004-2012. The final dataset results from a pipeline consolidating all relevant data for ADE research, normalizing different term usage, de-duplicating cases, mapping drugs to RxNorm, mapping drug indications and reactions to MedDRA, and generating drug-outcome pairs with associated statistics. The provided documentation and code is expected to support the evolution of this resource when updates of the FAERS dataset are released.

### Bauer CR, Ganslandt T, Baum B, Christoph J, Engel I, Lobe M, Mate S, Staubert S, Drepper J, Prokosch HU, Winter A, Sax U

### Integrated Data Repository Toolkit (IDRT). A Suite of Programs to Facilitate Health Analytics on Heterogeneous Medical Data

### Methods Inf Med 2016;55(2):125-35

This paper presents a set of tools developed within the Integrated Data Repository Toolkit (IDRT) German project in order to facilitate heterogeneous data integration in the i2b2 framework. Among various applications, this toolset efficiently allows researchers to design their own analyses. For example, the Mapping Editor of the IRDT Import and Mapping Tool helps to import different formats of data into the current i2b2 terminology. The originality at this step is to allow users to integrate data described in different terminologies (e.g., ICD-10-GM, MedDRA, LOINC, ICP-O, and others) and to build a new repository, with a unique model able to support the representation of these data. The new target ontology is a dedicated model of the data of the project which can be saved for new analyses – with a prior translation of the data from the original terminological model to a new target hierarchy. As a result, IDRT appears as a step forward to the semantization of i2b2.

### Greene D, NIHR BioResource, Richardson S, Turro E

### Phenotype Similarity Regression for Identifying the Genetic Determinants of Rare Diseases

### Am J Hum Genet 2016;98(3):490-9

This paper is dedicated to model the association between multiple phenotypic traits described with the Human Phenotype Ontology (HPO) and the corresponding genotype, in the specific context of rare disease (rare variants). HPO allows composite phenotypes to be represented systematically but association methods accounting for the ontological relationship between HPO terms do not exist. The authors propose a Bayesian method to model the association between HPO-coded phenotypes and genotypes. The method uncovers associations between rare genotypes and the similarities between patients' phenotypes and a latent characteristic phenotype. The effectiveness of the approach is demonstrated on a simulation study and on a real dataset from the BRIDGE project.

### Sarntivijai S, Vasant D, Jupp S, Saunders G, Bento AP, Gonzalez D, Betts J, Hasan S, Koscielny G, Dunham I, Parkinson H, Malone J

### Linking rare and common disease: mapping clinical disease-phenotypes to ontologies in therapeutic target validation

### J Biomed Semantics 2016;7-8

This paper proposes solutions to annotation-ontology mapping in genome-scale data. Of particular interest in this work is the Experimental Factor Ontology (EFO) and its generic association model, the Ontology of Biomedical AssociatioN (OBAN). EFO is a well-founded ontology, reusing ontologies from the Open Biomedical Ontologies (OBO) community (and other necessary models) for a comprehensive description of the domain, with the Minimum Information to Reference an External Ontology Term (MIREOT) strategy: Chemical Entities of Biological Interest Ontology (ChEBI), the Phenotypic And Trait Ontology (PATO), the Orphanet Rare Disease Ontology (ORDO), the BRENDA Tissue Ontology (BTO), the Uber Anatomy Ontology (Uberon), and the Gene Ontology (GO). OBAN is a means to represent diseases and phenotypes associations and the source of evidence for these associations. This was applied to the use case of linking rare to common diseases at the Centre for Therapeutic Target Validation. Based on these models, this work demonstrates the feasibility of rare and common diseases integration, using shared phenotypes. This paper offers a convincing example of the industrialization of integration. The EFO ontology is updated monthly and allows to propose regularly new associations.