Knowledge Representation and Management: a Linked Data Perspective

M. Barros, F. M. Couto
LaSIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal

Summary

Introduction: Biomedical research is increasingly becoming a data-intensive science in several areas, where prodigious amounts of data is being generated that has to be stored, integrated, shared and analyzed. In an effort to improve the accessibility of data and knowledge, the Linked Data initiative proposed a well-defined set of recommendations for exposing, sharing and integrating data, information and knowledge, using semantic web technologies.
Objective: The main goal of this paper is to identify the current status and future trends of knowledge representation and management in Life and Health Sciences, mostly with regard to linked data technologies.
Methods: We selected three prominent linked data studies, namely Bio2RDF, Open PHACTS and EBI RDF platform, and selected 14 studies published after 2014 (inclusive) that cited any of the three studies. We manually analyzed these 14 papers in relation to how they use linked data techniques.
Results: The analyses show a tendency to use linked data techniques in Life and Health Sciences, and even if some studies do not follow all of the recommendations, many of them already represent and manage their knowledge using RDF and biomedical ontologies.
Conclusion: These insights from RDF and biomedical ontologies are having a strong impact on how knowledge is generated from biomedical data, by making data elements increasingly connected and by providing a better description of their semantics. As health institutes become more data centric, we believe that the adoption of linked data techniques will continue to grow and be an effective solution to knowledge representation and management.

Keywords
RDF, ontologies, medical informatics, information management, information storage and retrieval, common data elements

Yearb Med Inform 2016:178-83
http://dx.doi.org/10.15265/IY-2016-022
Published online November 10, 2016

1 Biomedical Data

Biomedical research is increasingly becoming a data-intensive science in several areas, where prodigious amounts of data is being generated that has to be stored, and most of the times integrated, shared and analyzed. Moreover, biomedical data is highly complex when compared to standard big data projects that, for example, store and analyze short messages and their authors, often collected from a single source and using common formats. In opposition, even a single health institution has to deal with multiple types of data, in heterogeneous formats and from different sources, such as electronic health records, clinical images and reports, or genome sequences.

The challenge of how to store and manage biomedical data in the most precise way possible has a long-standing history, and besides the big technological advances it still remains an open issue. For example, in 1985 The Committee on Models for Biomedical Research proposed a structured and integrated view of biology to cope with the available data [1]. Nowadays, the BioMedBridges [2] initiative aims at constructing the data and service bridges needed to connect the emerging Biomedical Sciences Research Infrastructures (BMSRI), which are on the roadmap of the European Strategy Forum on Research Infrastructures (ESFRI).

Knowledge

Besides all the technological advances that we may deliver to make data easily accessible, researchers need more than raw data, they need a clear and objective characterization of who, what, where, why and how that data was collected. For example, due to the Galileo's strong commitment to the advance of Science, he integrated the direct results of his observations of Jupiter with careful and clear descriptions of how they were performed, which he shared in Sidereus Nuncius [3]. These descriptions enabled other researchers not only to be aware of Galileo's findings but also to understand, analyze and replicate his methodology. We must understand the meaning of data to replicate experiments and their outcomes, otherwise they are just sequences of zeros and ones where we are able to find useless correlations but no causality. For example, knowing the raw sequence of our genome is useless without the knowledge that science gave us after all these years of studying its meaning.

Even if you have easy access to all biomedical data, its real value can only be leveraged through how effectively we can analyze it towards the acquisition of knowledge that needs to be represented and managed. Creating data without producing knowledge is like writing books that are never read, and biomedical data is like erudite books in terms that they normally are not easy to read with their challenging writing styles. Thus, biomedical literature has been the traditional and natural mean for representing knowledge, where all the findings are properly described and their limitations and potentials fully discussed. As a consequence, a large amount of the knowledge acquired in Life and Health Sciences is available through literature. However, representing knowledge as unstructured free text hinders its accessibility and usage, since the retrieval of information from a large collection of texts is a tedious and time-consuming task for humans and a hard and prone to error task for machines [4].

Linked Data

In an effort to improve the accessibility of data and knowledge without losing too much flexibility, the Linked Data initiative [5] proposed a well-defined set of recommendations for exposing, sharing and integrating data, information and knowledge, using semantic web technologies. This paradigm is more than just a standardized messaging and text communications protocol to avoid data silos, such HL7, Linked Data enables the association and characterization of any kind of data in the form of links reinforcing our tools to represent and manage knowledge. The links are described using Resource Description Framework (RDF) [6] that provides a universal graph-based data model to connect the data between themselves but also to add semantics to them [7]. This model is more flexible than traditional data storage models, but still not as much as unstructured free text. Thus, literature will not be replaced by linked data but more data and knowledge can be easily expressed this way without hindering its accessibility. Besides RDF there are other graph-based models, such as Property Graphs [8], which may prove to be more effective in some specific areas of biomedicine.

One of the earliest well-known attempts of applying Linked Data to biomedical data was Bio2RDF [9], an open access platform that provided access to millions of documents in normalized RDF format with data from hundreds of different organisms. The potential of Bio2RDF was demonstrated in a case where a knowledgebase about Parkinson's disease was successfully built and some specialized questions were efficiently answered.

A few years ago, a public-private partnership between the pharmaceutical industry and the academia, publishers, small and medium sized enterprises initiated the project Pharmacological Concept Triple Store (Open PHACTS) [10]. Its goal was to build an open pharmacological knowledgebase that could overcome with the complexity of data access and licensing hurdles intrinsic to this domain with a solid plan for sustainability, service provision and maintenance in the public domain. Like Bio2RDF, Open PHACTS platform is based on RDF format with a bottom-up perspective of data standards where information from multiple providers is exposed by adaptive integration of the information.

More recently, the European Bioinformatics Institute (EBI), a major provider of bioinformatics data and services, made available to the community the EBI RDF platform [11]. This platform integrated multiple EBI data resources, such as UniProt, Gene Expression Atlas, ChEMBL, BioModels, Reactome and Biosamples, based on the RDF format and accessible through a standard query language interface (SPARQL). EBI RDF platform is the web interface for online access, but besides just providing data in a common format, this platform makes an effort in including as much as possible common vocabularies to describe their semantics and provenance.

Just by adopting the Linked Data paradigm does not mean that we are sharing knowledge. Each human has his own set of links in his mind, and to start communicating we need a common ground. For example, by using spoken English two human can share their knowledge about the world and therefore create more links in their minds. There are specific predicates that should be used for linking datasets, such as *owl:sameAs*, *rdfs:seeAlso*, *skos:exactMatch* and *skos:closeMatch*, and a recent study showed that in Life Sciences the predicate *owl:sameAs* was the most widely used linking predicate (52.17%) [12].

One important aspect of the Linked Data paradigm is the usage of common vocabularies that are expressed using the RDF Schema [13] (RDFS) and the Web Ontology Language [14] (OWL). These vocabularies are used to describe the data elements and their relations by defining classes and their properties. The usage of common vocabularies is incentivized (but not compulsory) to establish a common interpretation of data and by consequence enable knowledge sharing. These vocabularies can vary from simple terminologies to highly complex semantic models of a given domain encoded in the form of ontologies, such as Gene Ontology (GO) [15]. The Linked Data paradigm uses RDF as its data model together with its vocabulary definition languages RDFS and OWL. However, in fact the usage of RDF and ontologies goes beyond the scope of Linked Data, and many biomedical projects exploit them without necessarily following the Linked Data paradigm

Ontologies

Above biomedical data was compared to erudite books with challenging writing styles, but now imagine if each one of them were written in an exclusive language that could not be easily mapped to English, and therefore without any thorough translation available. Reading each book required us to learn a new language to fully understand its message. The knowledge was there but accessible to just a few. Thus, standard classification vocabularies represent a solution that prevents data and knowledge from being stored as silos by enabling data annotations with common terms, which makes data and its meaning more accessible. These vocabularies are instantiated by Knowledge Organization Systems (KOS) [16] in the form of classification systems, thesauri, lexical databases, gazetteers, and taxonomies, and ontologies. The latter can be loosely defined as "a vocabulary of terms and some specification of their meaning" [17, 18]. If an ontology is accepted as a reference by the community then the representation of its domain becomes a standard, and knowledge sharing and management is facilitated.

The etymological encyclopedia (Book IV: Medicine) [19] compiled by Isidore of Seville (c. 560–636) was one of the first attempts to systematize medicine knowledge. In the seventeen century, London bills of Mortality [20] established a classification terminology for registering morbidity and mortality cases that enabled the study of mortality rates and their causalities. However, only in the last decades the biomedical community openly engaged on developing and using ontologies to represent and manage knowledge. Perhaps the most known KOS in medicine nowadays is the vocabulary provided by the International Classification of Diseases ICD [21], a classification system that is being maintained by the World Health Organization (WHO), which originally aimed at providing a statistical analysis tool for disease incidence and mortality. The current release ICD-10 [22] provides a vocabulary containing a list of generic clinical terms mainly arranged and classified according to anatomy or etiology.

Another well-known ontology is the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT), originally created by the College of American Pathologists and currently maintained by the International Health Terminology Standards Development Organization. The SNOMED CT provides a highly comprehensive and detailed set of clinical terms used in many systems to enrich the information in electronic health records. The July 2016 release provided 321.901 active concepts [23]. SNOMED also includes logic-based definitions to represent terminological knowledge, i.e., facts about the meaning of the terms. For example, the term *myocardial infarction* includes the fact that it must involve the *myocardium*, and it must involve an *infarction*. SNOMED CT is available through the Unified Medical Language System (UMLS) [24] maintained by the U.S National Library of Medicine. The UMLS provides a Metathesaurus that integrates more than one hundred miscellaneous vocabularies (e.g. Medical Subject Headings thesaurus (MeSH)), which in the 2015AB release covered more than three million concepts [25].

One of the criticisms of SNOMED CT is the fact that is proprietary. Therefore the Open Biomedical Ontology foundry (OBO) [26] proposed an alternative approach where design patterns and best practices in ontology specification are stimulated on an open usage and collaborative development basis. OBO established a set of principles that the ontologies have to satisfy before becoming part of the project. These principles ensure high quality, formal rigor and also interoperability between OBO member ontologies. One of these principles requests the ontology to be open and available without any constraint other than acknowledging its origin. They also provide an alternative format to OWL to represent ontologies, named OBO format.

GO is one of the most popular OBO ontologies, which has been extensively used to annotate gene-products with terms describing their molecular functions, biological processes and cellular components. In September 2014, GO provided more than 41,775 terms and a total of 53,042,843 gene-products annotated with them [27]. GO relies on a large consortium of collaborators that cooperate in maintaining and updating the ontology which made it widely used and accepted, and thus considered a major example of success for biomedical ontologies. Another OBO ontology is the Disease Ontology (DO) [28] that provides human disease terms, phenotype characteristics and related medical vocabulary disease concepts. In October 2014, DO contained 8,803 terms of which 2,384 are considered obsolete [29]. The Human Phenotype Ontology (HPO) [30] is also an OBO ontology that provides terms for describing phenotypic abnormalities seen in human disease. In 2015, HPO contained more than 11,000 terms, and over 116,000 annotations to over 7,000 rare diseases [31].

2 Methodology

To present an overview of the significant developments in knowledge representation based on linked data approaches over the past one or two years, we started by identifying a representative set of articles by analyzing the citations to the well-known projects described above. Thus, in December 2015 we started by collecting the list of articles on Google Scholar that cited articles presenting Bio2RDF, Open PHACTS and/or EBI RDF platform. At the time they were quoted by 498, 128, and 59 articles, respectively.

The list of articles was then automatically filtered using the following restrictions: i) published after 2014 (inclusive), ii) having the word data in their title, and iii) published in biomedical journals. The first restriction limited the survey to approaches published over the past one or two years. The second one limited the survey to approaches that have a strong focus on data. The third restrictions limited the survey to biomedical studies that are already well-established.

Finally, we manually analyzed the scope of each article and selected only the ones that represented case-studies, repositories or frameworks working with data of Life and Health sciences. Thus, we removed all the articles mainly describing software and tools, statistical analysis of data, opinions, reviews and surveys.

The result of this process was a list of 14 articles that we believe provide a good representative overview, not a comprehensive list, of the significant developments in knowledge representation over the past one or two years.

3 Results

Bio2RDF Related Work

As displayed on Table 1, from the 14 selected articles, i, ix, x, xi, xii, xiii and xiv quote Bio2RDF work, but they do not use this platform. Study i, for example, proposed a Linked Clinical Data Cube, which main goal was to use data from Australian, Imaging, Biomarker and Lifestyle study of Ageing (AIBL) and makes it available as linked data for the research community. The authors quote Bio2RDF as related work, since both studies transform data from databases in RDF and share them in an easy and accessible way, allowing the extraction of relevant information from these databases. Another example is study ix. They created a platform (eXframe platform), that as well as Bio2RDF, makes the information available as RDF. In study xiii, a Web tool (TogoTable) is built. It utilizes the features of RDF to connect several Linked Open Data (LOD) databases, enabling links to Bio2RDF.

Articles iv, vi, vii and viii, not only quote Bio2RDF, but also used it in their methodology. In iv, they presented an approach to integrate pathway data from four different Linked Data repositories using Bio2RDF Kegg's data as the core and Bio2RDF Reactome distribution as an extension. The goal of article vi was to mine linked open data and they used Bio2RDF ontologies to link some entities. Article vii used Bio2RDF biological database applying new standards for LOD necessary to communicate effectively with other reference databases already operating under the scheme or Semantic Web. Finally, article viii, proposes a nanopublication publishing format that uses Bio2RDF since it provides RDF and URIs for different biomedical resources.

Table 1 Results of the 14 selected articles regarding to Bio2RDF, EBI RDF platform and Open PHACTS. Label: x - quote; xx - quote and use.

Nº	Article	Bio2RDF	EBI RDF platform	Open PHACTS
i	Leroux, 2015 [32]	×		
ii	Cases, 2014 [33]			×
iii	Hettne, 2014 [34]			×
iv	Navas¬Delgado, 2015 [35]	××		
v	Davies, 2015 [36]		××	
vi	Personeni, 2014 [37]	××	×	
vii	Bertel¬Paternina, 2014 [38]	××		
viii	Mina, 2015 [39]	××		
ix	Merrill, 2014 [40]	×	×	
x	Ranzinger, 2015 [41]	×	×	
xi	Hoehndhorf, 2015 [42]	×	×	×
xii	Wolstencroft, 2015 [43]	×	×	
xiii	Kawano, 2014 [44]	×	×	
xiv	Rebholz¬Schuhmann, 2014 [45]	×		×

Table 2 Results from the 14 selected articles regarding the type of data they use and create.

Nº	Article	Input	Output
i	Leroux, 2015	RDF	RDF
ii	Cases, 2014	OWL/OBO	New ontology
iii	Hettne, 2014	OWL/OBO	RDF
iv	Navas¬Delgado, 2015	RDF	RDF
v	Davies, 2015	Other	RDF
vi	Personeni, 2014	RDF	RDF
vii	Bertel¬Paternina, 2014	RDF	RDF
viii	Mina, 2015	RDF/OWL	Nanopublications
ix	Merrill, 2014	OWL/OBO	RDF
x	Ranzinger, 2015	OWL	RDF
xi	Hoehndhorf, 2015	RDF/OBO/OWL	OWL
xii	Wolstencroft, 2015	OWL	RDF
xiii	Kawano, 2014	RDF/OBO	RDF
xiv	Rebholz¬Schuhmann, 2014	OWL/OBO	RDF

EBI RDF Platform Related Work

The EBI RDF platform is quoted in 7 of the 14 papers: v, vi, ix, x, xi, xii and xiii. However, only article v work is related with this platform. From the articles that only quote EBI RDF platform, article x compares the efforts from EBI-RDF to provides an innovative approach to queries and explore rich biological data collections, with their own efforts to create an ontology for generating standardized RDF for glycan structures and related data. Another example is article xii. The goal was to create SEEK platform: a suite of tools to support the management, sharing and exploration of data and models in systems biology. SEEK stores metadata in RDF which promotes greater interoperability with other platforms like EBI-RDF.

Open PHACTS Related Work

The Open PHACTS is quoted in only 4 of the 14 selected articles (ii, iii, xi and xiv), but none of them use it. The eTOX data-sharing project (ii) is gathering data from public and private domain, being the main goal the development of a common ontology and it quotes Open PHACTS as another initiative to gather chemical related toxicity information. The same description is given in article xiv. This work is related with The Semantic Enrichment of the Scientific Literature (SESL) project. Articles iii and xi shows Open PHACTS as an example of other efforts where Semantic Web technology has been used for the biomedical data integration.

Table 2 shows the data input and output used in the selected articles. Most articles (ii, iii, viii, ix, x, xii and xiv) used Ontology Web Language (OWL) and Open Biomedical Ontologies (OBO) as input or just as complementary data to improve the output. The eTOX project (ii), for example, used data from preclinical studies, extracted from papers and PDF's through data mining, and ontoBrowser ontology to confirm and standardize the data so they could be used to create a new ontology of toxicity. This is useful to create a predictive model for drug development process. In ix, several ontologies such NCBITaxon, EFO, FMA, BTO, CL, NCI Thesaurus and CHEBI were used to annotate data. Ontologies from Bio2RDF platform are particularly employed, as can be read in articles iv, vi, vii, and viii. Gene Ontology was used in articles iii, xi and xiv. Article v do not specify the type of input data, just mentioning that the data were manually extracted and curated from chemistry literature.

RDF data was the input in articles i, iv, xi, vii and xiii. In i, clinical study data was extracted in Clinical Data Interchange Standard Consortium – Operational Data Model (CDISC ODM) format and Data Documentation Initiative RDF (DDI-RDF) vocabulary was used to enrich clinical data based on the CDISC standards. Both RDF and ontologies are used in two particular cases, viii and xi. In the first one, data from Bio2RDF platform is used as well as NIF Standard ontology (NIF-STD), NCI Thesaurus (NCI), Gene Regulation Ontology (GRO), SemanticScience Integrated Ontology (SIO) and Sequence Ontology (SO). The former, a bio-ontology repository, used several ontologies and RDF data to create the platform. Final output for most articles is RDF data. Exceptions are article ii, that created a new ontology of toxicological terms, article viii, which has as output nanopublications, and xi, as already said, intended to create an ontology repository, thus its output is a research ontology platform.

In the selected articles we were not able to find information about dereferencability of the vocabularies/ontologies. However, according to recent statistics, most vocabularies terms (71.73%) are not dereferencable, 19.47% are totally dereferencable and 8.8% are only partially derefenrencable. Particularly in Life Science, 66.67% are not dereferencable, 27.78 are totally dereferencable and 5.56% are partially dereferencable [12].

4 Discussion

There is no doubt that nowadays we have access to more data, more easily and with higher quality than a decade ago. However, is it our knowledge keeping up the pace? As presented above, RDF technologies are having a strong impact on how the Life and Health Sciences community is storing, integrating and sharing data and knowledge. Even if not fully following Linked Data paradigm, the community is now making a large effort in exploiting some of its technologies for connecting the data elements and consequently providing a better description of their semantics. In that sense, ontologies are performing a crucial role in making the semantic annotations consistent and interoperable. Unlike the knowledge concealed in the articles, the knowledge shared through annotations using standard ontologies can be easily processed and analyzed by computational methods. For example, it enables us to search for similar and related entities based on their biomedical meaning, such as similar molecular functions and similar diseases [46].

Many information retrieval systems, such as Google, use similarity measures to calculate the similarity between a query and a document that takes in account its relevance to the user. For instance, if we try to look for physiology models annotated with *Scaphoid* we may be interested in receiving the models annotated with *Wrist* as well, but probably not all the models annotated with other *Upper limb* segments. This relevance can be captured by semantic similarity measures that return a numerical value reflecting the closeness in meaning between semantic annotations [47]. These semantic similarity measures have been successfully developed and applied to biomedical ontologies, particularly to the Gene Ontology, where they are mainly used to compare genes or proteins based on the similarity of their functions. Another popular technique is enrichment analysis that exploits the semantic annotations to identify clinical and biological characteristics that may better describe the outcome of a group of patients with a common disease. For example, recently this technique was effectively applied to improve the disease prognosis of the hypertrophic cardiomyopathy [48].

In a near future health institutes would be data centric, where each situation is analyzed according to previous situations by comparing similar patient profiles with similar phenotypes. For example, screening processes that are crucial to detect life-threatening situations in a short period of time would benefit from having a large knowledgebase together with advanced information retrieval systems that could provide these alerts in real time. Due to privacy issues these knowledgebase are normally restricted to local data that hinders their effectiveness, but for sharing data we do not need open data. For example, we can use a remote similarity service from an external knowledgebase and if we get a hit, we may automatically send a request to access that matching information. If permissions are granted, we may access the information in an anonymized and controlled way, i.e., in case of any leak we know who, how, why and what was granted and accessed. Thus, by dealing with sensitive data does not mean that we cannot share metadata and services following a linked data perspective, by the contrary it is one (if not the) of the best approaches to represent and manage knowledge in such a setting.

Linked Data offers an effective solution to break down data silos, however, the systematic usage of these technologies requires a strong commitment from the research community. Creating linked data resources with sound and comprehensive characterizations of their meaning and using semantic annotations to common ontologies is a complex and subjective process, which can be supported by automatic methods, such as text mining [49], but in most cases it requires a lot of specialized human intervention. So recognition and reward mechanisms besides bibliometric indicators will be essential to avoid the creation of raw data silos that cannot be reused by others, or even by the owners themselves [50]. This incentive is currently so low that sometimes even authors cannot recover the data associated with their own publications. Public funding agencies and journals may enforce data-sharing policies, but adherence is most of the times inconsistent and scarce [51]. The problem, therefore, is obtaining a proactive involvement of the community in integrating and sharing data. To support these, we have to go beyond technological advances, and create motivation mechanisms that encourage data owners to share their data in a meaningful way [52].

Linked Data is not free or open data and is not sound data, it can have access restrictions, be incomplete, have errors, but the technological advances and the successful use cases in the Life and Health Sciences shown above are a promising sign that linked data may in near future be omnipresent in our daily lives as the Internet is today.

Acknowledgements

We would like thank Gonçalo Figueiró, Joana Teixeira and Marta Antunes for helping us in identifying some of the details from the articles reviewed. This work was supported by FCT through funding of the LaSIGE Research Unit, ref. UID/CEC/00408/2013.

References

1. Morowitz H. Models for biomedical research: A new perspective. Washington DC: National Academy of Sciences Press; 1985.
2. Duarte AM, Psomopoulos FE, Blanchet C, Bonvin AM, Corpas M, Franc A, et al. Future opportunities and trends for e-infrastructures and life sciences: going beyond the grid to enable life science data analysis. Front Genet 2015 Jun 23;6:197.
3. Galilei G, Van Helden A. Sidereus Nuncius, or The Sidereal Messenger. University of Chicago Press; 1989 Apr 15.
4. Pradhan S, Elhadad N, Chapman W, Manandhar S,

Savova G. Semeval-2014 task 7: Analysis of clinical text. SemEval 2014. 2014 Aug 23;199(99):54.

5. http://linkeddata.org/
6. Klyne G, Carroll JJ. Resource description framework (RDF): Concepts and abstract syntax. W3C recommendation, W3C, February 2004. URL: http://www. w3. org/TR/rdf-concepts.
7. Bizer C, Heath T, Berners-Lee T. Linked data-the story so far. Semantic Services, Interoperability and Web Applications: Emerging Concepts. 2009:205-27.
8. Alocci D, Mariethoz J, Horlacher O, Bolleman JT, Campbell MP, Lisacek F. Property Graph vs RDF Triple Store: A Comparison on Glycan Substructure Search. PloS One 2015 Dec 14;10(12):e0144578.
9. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J Biomed Inform 2008 Oct 31;41(5):706-16.
10. Williams AJ, Harland L, Groth P, Pettifer S, Chichester C, Willighagen EL, et al. Open PHACTS: semantic interoperability for drug discovery. Drug Discov Today 2012 Nov 30;17(21-22):1188-98.
11. Jupp S, Malone J, Bolleman J, Brandizi M, Davies M, Garcia L, et al. The EBI RDF Platform: linked open data for the life sciences. Bioinformatics 2014,30(9):1338–9.
12. Schmachtenberg M, Bizer C, Paulheim H. Adoption of the linked data best practices in different topical domains. In: International Semantic Web Conference 2014 Oct 19. Springer International Publishing; 2014. p. 245-60.
13. Brickley D, Guha RV. RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation, February 2004.
14. McGuinness DL, Van Harmelen F. OWL web ontology language overview. W3C recommendation. 2004 Feb 10;10(10):2004.
15. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet 2000 May;25(1):25-9.
16. Mayr P, Tudhope D, Clarke SD, Zeng ML, Lin X. Recent applications of Knowledge Organization Systems: introduction to a special issue. International Journal on Digital Libraries 2016 Mar 1;17(1):1-4.
17. Jasper R, Uschold M. A framework for understanding and classifying ontology applications. In: Proceedings 12th Int. Workshop on Knowledge Acquisition, Modelling, and Management KAW 1999 Oct. 1999;99:16-21.
18. Uschold M, Gruninger M. Ontologies: Principles, methods and applications. The knowledge engineering review 1996 Jun 1;11(02):93-136.
19. Lindsay WM. The editing of isidore etymologiae. The Classical Quarterly 1911 Jan 1;5(01):42-53.
20. Graunt J. Natural and Political Observations Made Upon the Bills of Mortality; London; 1662.
21. http://www.who.int/classifications/icd/en/
22. de Keizer NF, Abu-Hanna A, Zwetsloot-Schonk JH. Understanding terminological systems I: terminology and typology. Methods Inf Med 2000 Mar 1;39(1):16-21.
23. http://www.ihtsdo.org/snomed-ct/snomed-ct-worldwide
24. https://www.nlm.nih.gov/research/umls
25. https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html
26. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007 Nov 1;25(11):1251-5.
27. Gene Ontology Consortium. Gene ontology consortium: going forward. Nucleic Acids Res 2015 Jan 28;43(D1):D1049-56.
28. Schriml LM, Arze C, Nadendla S, Chang YW, Mazaitis M, Felix V, et al. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Res 2012 Jan 1;40(D1):D940-6.
29. Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, et al. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res 2014 Oct 27:gku1011.
30. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. The Am J Hum Genet 2008 Nov 17;83(5):610-5.
31. Groza T, Köhler S, Moldenhauer D, Vasilevsky N, Baynam G, Zemojtel T, et al. The human phenotype ontology: semantic unification of common and rare disease. Am J Hum Genet 2015 Jul 2;97(1):111-24.
32. Leroux H, Lefort L. Semantic enrichment of longitudinal clinical study data using the CDISC standards and the semantic statistics vocabularies. J Biomed Semantics 2015 Apr 9 6:16.
33. Cases M, Briggs K, Steger-Hartmann T, Pognan F, Marc P, Kleinöder T, et al. The eTOX Data-Sharing Project to Advance in Silico Drug-Induced Toxicity Prediction. Int J Mol Sci 2014;15(11):21136-54.
34. Hettne KM, Dharuri H, Zhao J, Wolstencroft K, Belhajjame K, Soiland-Reyes S, et al. Structuring research methods and data with the research object model: genomics workflows as a case study. J Biomed Semantics 2014;5(1):41.
35. Navas-Delgado I, Garcí a-Godoy MJ, López-Camacho E, Rybinski M, Reyes-Palomares A, Medina MA, et al. kpath: integration of metabolic pathway linked data. Database (Oxford) 2015 Jun 8;2015:bav053.
36. Davies M, Nowotka M, Papadatos G, Dedman N, Gaulton A, Atkinson F, et al. ChEMBL web services: streamlining access to drug discovery data and utilities. Nucleic Acids Res 2015;43(W1):W612-20.
37. Personeni G, Daget S, Bonnet C, Jonveaux P, Devignes MD, Smaí l-Tabbone M, et al. Mining Linked Open Data: a Case Study with Genes Responsible for Intellectual Disability. Data Integration in the Life Sciences - 10th International Conference, DILS 2014, Jul 2014, Lisbon, Portugal. Lecture Notes in Computer Science 2014;8574:16 – 31.
38. Bertel-Paternina L, Castillo LF, Isaza G, Gaitán-Bustamente A. Towards a Linked Open Data Model for Coffee Functional Relationships. Advances in Intelligent Systems and Computing 2014;:232.
39. Mina E, Thompson M, Kaliyaperumal R, Zhao J, van der Horst E, Tatum Z, et al. Nanopublications for exposing experimental data in the life-sciences: a Huntington's Disease case study. J Biomed Semantics 2015;6:5.
40. Merrill E, Corlosquet S, Ciccarese P, Clark T, Das S. Semantic Web repositories for genomics data using the eXframe platform. J Biomed Semantics 2014;5(Suppl 1):S3.
41. Ranzinger R, Aoki-Kinoshita KF, Campbell MP, Kawano S, Lütteke T, et al. GlycoRDF: An ontology to standardize Glycomics data in RDF. Bioinformatics 2015 Mar 15;31(6):919-25.
42. Hoehndorf R, Slater L, Schofield PN, Gkoutos GV. Aber-OWL: a framework for ontology-based data access in biology. BMC Bioinformatics 2015;16:26.
43. Wolstencroft K, Owen S, Krebs O, Nguyen Q, Stanford NJ, Golebiewski M, et al. SEEK: a systems biology data and model management platform. BMC Syst Biol 2015;9:33.
44. Kawano S, Watanabe T, Mizuguchi S, Araki N, Katayama T, Yamaguchi A. TogoTable: cross-database annotation system using the Resource Description Framework (RDF) data model. Nucleic Acids Res 2014;42(Web Server issue):W442-8.
45. Rebholz-Schuhmann D, ller CG, Kavaliauskas S, Croset S, Woollard P, Backofen R, et al. A case study: semantic integration of gene–disease associations for type 2 diabetes mellitus from literature and biomedical data resources. Drug Discov Today 2014; 19(7):882-9.
46. Ferreira JD, Paolotti D, Couto FM, Silva MJ. On the usefulness of ontologies in epidemiology research and practice. J Epidemiol Community Health 2013 May;67(5):385-8.
47. Couto FM, Pinto HS. The next generation of similarity measures that fully explore the semantics in biomedical ontologies. J Bioinform Comput Biol 2013 Oct;11(5):1371001.
48. Machado CM, Freitas AT, Couto FM. Enrichment analysis applied to disease prognosis. J. Biomed Semantics 2013 Oct 8;4:21.
49. Groza T, Köhler S, Doelken S, Collier N, Oellrich A, Smedley D, et al. Automatic concept recognition using the Human Phenotype Ontology reference and test suite corpora. Database (Oxford) 2015 Feb 27;2015.
50. Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, Crosas M, et al. Ten simple rules for the care and feeding of scientific data. PLoS Comput Biol 2014 Apr 24;10(4):e1003542.
51. Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, Ioannidis JP. Public availability of published research data in high-impact journals. PloS One 2011 Jan;6(9):e24357.
52. Couto FM. Rating, recognizing and rewarding metadata integration and sharing on the semantic web. In10th International Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2014) 2014 Sep. p. 67.

Correspondence to:
Francisco M. Couto
LaSIGE
Departamento de Informática
Faculdade de Ciências
Universidade de Lisboa
1749-016 Lisboa, Portugal
E-mail: fjcouto@ciencias.ulisboa.pt