JECH Online First, published on November 15, 2012 as 10.1136/jech-2012-201142

On the usefulness of ontologies in epidemiology research and practice

João D Ferreira,[1] Daniela Paolotti,[2] Francisco M Couto,[1] Mário J Silva[1,3]

INTRODUCTION

Epidemiology research is a truly multidisciplinary subject, relying on areas of knowledge as diverse as medicine, biology, statistics, sociology and geography.[1] The creation of large-scale epidemiological models and the development of effective model-based prediction methods can only be achieved if efficient data collection techniques based on reliable policies for data sharing between research communities and health authorities are adopted.[2] As a research domain that so strongly depends on heterogeneous data from diverse origins, epidemiology greatly requires a proper integrative framework to cope with its inherent multidisciplinarity.

One promising way to meet these requirements is the adoption by the epidemiology community of Semantic Web technologies. The Semantic Web is a vision of information management and sharing that promotes intelligent access to data on the world wide web, both by human beings and by computers.[3] The adoption of the Semantic Web is not new in biomedical research: for instance, in molecular biology, it has been applied in the past with intent to create successful applications. One of these is GoPubMed, a platform that enables a deep and structured exploration of PubMed abstracts[4]; another one is a method to identify gene functions associated with specific biological phenomena.[5]

The remainder of this manuscript will illustrate the advantages of adopting this paradigm for epidemiological studies, together with a brief introduction of standard Semantic Web concepts and practices, that could be useful for current and prospective epidemiologists. We also present the Epidemic Marketplace, a case study for storing and describing epidemiological resources following the Semantic Web vision.

THE SEMANTIC WEB VISION

The world wide web is, by itself, an extremely useful content-sharing platform, but the content of its resources is not expressed through a common data format and is mainly directed at human users. To achieve machine-readability, the Semantic Web perceives information as *resources* (datasets, documents, etc), which are characterised with links to other resources. Each of these links, also called *metadata* or *annotations*, can be seen as a description of the information contained in the resource, that is, its metadata (see figure 1 for an illustration). For instance, a resource about a disease can link to the concept of 'Europe' through the property 'occurs-in', while a resource about a person can link to 'Europe' through the property 'born-in'. In the Semantic Web, everything is a resource, so the concept of Europe, used above, can also be described through links to other concepts, like 'Europe contains Portugal'. A description of key Semantic Web concepts can be found in table 1.

IMPACT OF SEMANTIC WEB ON EPIDEMIOLOGY

The benefits of applying Semantic Web principles to epidemiology are multiple. Perhaps the most important is an increase of interoperability observed by users of epidemic resources, that is, an increase in the ability to exchange datasets between researchers without needing to convert the data to any specific format: the adoption of a common set of shared concepts harmonises the description of the resources.[6] Furthermore, machines can automatically trace and process resources based on their semantic relations, since they are described in a standard and formal representation,[7 8] which opens the possibility to exploit today's computational power over the semantic information available on the web.

The ability to search and browse data by concepts rather than by simple text highly facilitates the management and sharing of resources about a given subject. For instance, one can search resources about influenza or about people born in Europe, so long as the *identifiers* for those concepts are known, because this information is expressed in a machine-readable format.[9] Additionally, if the semantic search engine knows that Europe contains Portugal, it can retrieve the resources about people born in Portugal, when asked to find resources about people born in Europe.

The development of epidemiological models can also benefit from Semantic Web technologies. Many models refer to the same concepts: from the disease they model to the location where the first case was identified, or to where it spread; from the vector transmitting the disease to the modes of transmission and so on. Currently, the description of these models relies on text, and it is not always apparent that two models refer to the same disease (due to different spellings, abbreviations, etc), nor is it easy for computers to know that two models refer to two similar diseases. The principles of the Semantic Web recommend the publication of a model as a resource characterised with unambiguous metadata, for instance, 'resource_1 models *influenza*' or 'resource_2 is about *vector-borne transmission*'. Adopting standard concepts and standard properties brings the model into the Semantic Web, where interested people could use existing tools to find exactly the needed data, again provided that the correct identifiers for the standard concepts are known.

Prediction methods usually require a lot of data to adjust their parameters,[1 10] such as the fraction of people in a given country with a given socioeconomic condition. As shown, the Semantic Web is equipped with technologies to help discover data about both that country and that condition, facilitating the process of creating significant models and more accurate predictions.

The adoption of Semantic Web conventions can help find hidden patterns in epidemiological data. For instance, consider several outbreak datasets annotated to an infectious disease and the location of the event. Even if the datasets do not mention the population of the location, the environmental or the socioeconomic conditions, the data known about the

[1]Department of Informatics, Faculdade de Ciências da Universidade de Lisboa, Lisbon, Portugal
[2]Institute for Scientific Interchange (ISI), Torino, Italy
[3]IST, INESC-ID, Technical University of Lisbon, Lisbon, Portugal

Correspondence to João D Ferreira, Department of Informatics, Faculdade de Ciências da Universidade de Lisboa, Lisbon 1749-016, Portugal; joao.ferreira@lasige.di.fc.ul.pt

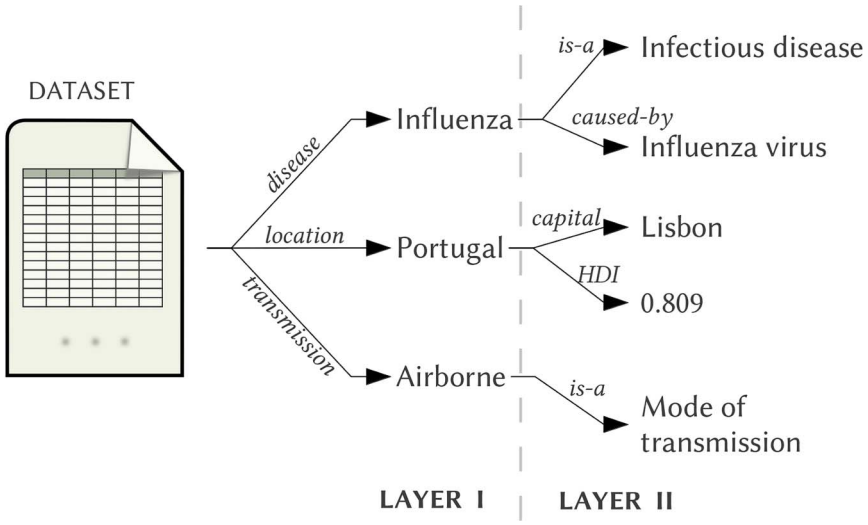

Figure 1 By publishing under the Semantic Web, information becomes more accessible and easier to be shared. Here we illustrate the Semantic Web in action with a resource annotated with its respective metadata. The first-layer concepts represent the direct annotation of the resource (such as 'location' is 'Portugal'), while the second layer includes the properties of these concepts as stated in disease and geospatial ontologies, which can also be accessed in the context of the Semantic Web. This figure is only reproduced in colour in the online version.

locations can help find patterns, such as a correlation between temperature at the time of the outbreak and the fraction of infected people, or between Human Development Index and the severity of the outbreak. The Semantic Web is extremely useful in linking together all this information, as illustrated in figure 1. To be able to do this, it is vital that resources are annotated with standard concepts that are linked to additional related information.

It is worth noting that the Semantic Web is meant to facilitate the access to information, and is not about making the information freely accessible, or about releasing it into the public domain. In the Semantic Web, resources, and the data they contain, can still be protected behind the necessary authentication methods or privacy policies, thus allowing only the right people to access the data,[11] albeit in a more efficient way.

THE ROLE OF ONTOLOGIES

Until now, we have hinted at three particular problems with the Semantic Web: (1) Semantic Web tools require facts about concepts (for instance, pattern finders must have access to facts about the concepts being analysed), (2) an effective interoperability between resources relies on users reusing the same identifiers to represent the same concepts and the same properties and (3) users need to know which identifier maps into which concept. These problems can be addressed by the use of ontologies.

An ontology is a description of a domain of knowledge consisting of the domain's concepts and their relationships, and can be used as a standard specification of such domain, for instance, diseases or geographical locations.[12 13] A relationship between concepts stands for facts about those concepts: for instance, 'Lisbon is the capital of Portugal' and 'influenza is an infectious disease'. Figure 2 shows a snippet of three ontologies, one of the geospatial domain, and the other from the disease domain. Consider these: (1) the capital of a country is part of that country, (2) a part of A is a part of all the things that contain A, (3) Lisbon is the capital of Portugal and (4) Portugal is part of Europe. A computer equipped with an inference engine can be taught these facts and then derive that 'Lisbon is part of Europe'. With this newly inferred knowledge, queries about events in Europe also return resources describing events in Lisbon. This kind of inference mechanism also enables powerful pattern recognition. This means that authors can annotate their resources with as much detail as they want, while users interested in finding resources can be as general as they want, while still maintaining compatibility between the two needs.

Table 1 Key concepts in the semantic web

Key concept	Description
Resource	Anything published under the Semantic Web vision, that is, with links to other resources. For example, a dataset on influenza or the concept of Europe
Concept	A kind of resource that represents a single idea. For example, the concept of influenza
Identifier	A unique string of characters that refers to exactly one concept. For instance, 'http://purl.obolibrary.org/obo/DOID_8469', which identifies the concept of influenza
Ontology	A collection of concepts and the relationships between them. These relationships provide concepts with a machine-readable meaning that can be explored for pattern recognition, knowledge discovery or any other computational analysis. Ontologies are the main source of concepts used in the metadata of Semantic Web resources.
Metadata	A machine-readable description of the contents of a resource made through linking the resource to the concepts that describe it. For instance, a dataset links to the concept of 'influenza' because it contains data concerning that disease. Figure 1 illustrates the metadata of a resource
Annotation	The process of enriching a resource with information about itself (metadata) by means of semantically defined properties pointing to other resources, especially properties pointing to concepts from ontologies
Property	A relationship between two resources, such as the ones that are used to annotate a resource with its metadata. Properties are also often called *relations*.
Semantic search engine	A software that searches for annotated resources in the Semantic Web based on a query. Inference can be used to find the appropriate results; additionally, results can be ranked according to similarity and relevance to the query. See the text for an example
Pattern recognition	The application of ontology-stored knowledge and inference mechanisms in the pursuit of unapparent correlations found in data. By using the semantics of the resources, information that is not explicit in the data but can be derived from ontologies can lead to more comprehensive results (see the text for an example).

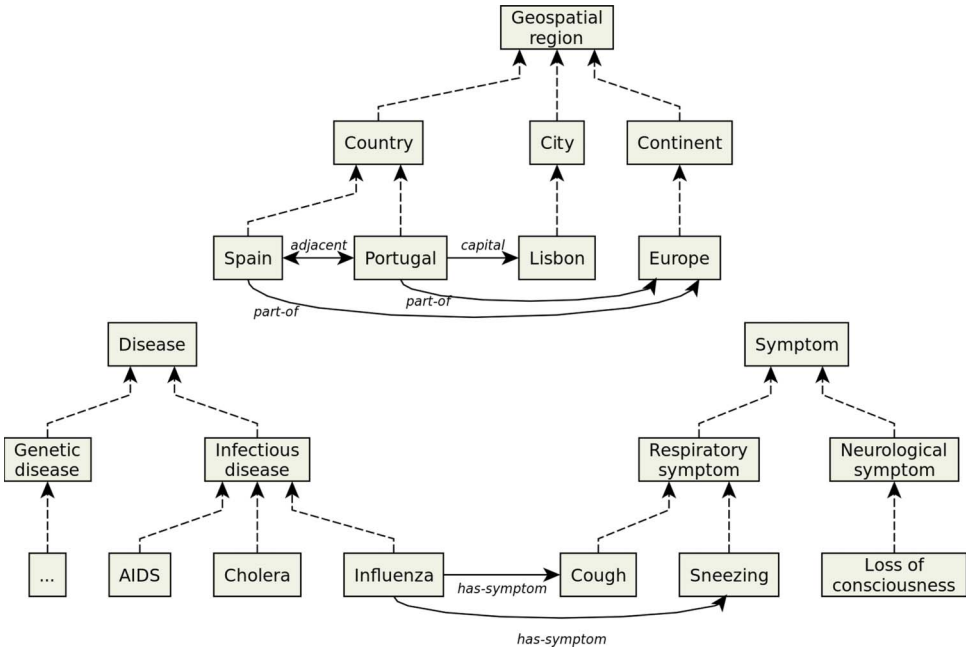

Figure 2 A snippet of three ontologies showing some concepts and some of the relationships between them. Most relationships are simple class-subclass ones, where a concept is a specialisation of the other, such as 'influenza' being a kind of 'infectious disease'; other relationships include, for instance, the borders of a geospatial region or the symptoms of a disease. This figure is only reproduced in colour in the online version.

Ontologies are also used to map identifiers to concepts, since ontologies contain concepts and identify them using URIs (Universal Resource Identifiers).[14 15] For example, the Human Disease Ontology[12] assigns the URI <http://purl.obolibrary.org/obo/DOID_8469> to the concept of 'influenza', and uses it to indicate facts about that disease (for instance, that it is an infectious disease). Any other person can use the same URI to associate its resources with the concept of 'influenza'. This removes ambiguity, since any single concept identifier is totally context independent and can be used by everyone, allowing Semantic Web tools to gather more information on a concept, if needed. For instance, other names of 'influenza' (like 'flu' or even names in other languages) can be associated with this identifier. Additionally, existing ontology-matching software tools can find equivalent concepts in different ontologies,[16] allowing ontology developers to merge ontologies from different domains in a single, more manageable knowledge base, which further increases the amount of knowledge available in the Semantic Web.

Web portals can be used as an intermediary in this process to assist the process of annotating an epidemiological resource; an example of such an intermediary is BioPortal,[17] a collection of biological and biomedical ontologies that can be used to access the information about the concepts on those ontologies. In the next section, we focus on the Epidemic Marketplace, a platform designed to store epidemiological resources that also includes an intermediary between names and URIs.

Furthermore, ontologies can be used to explore the degree of similarity between two concepts or two resources annotated with them. For instance, it is reasonable to assume that the concepts 'cough' and 'sneezing' are more related to each other than 'cough' and 'loss of consciousness'. This similarity between concepts can be computed based on the knowledge stored in the ontology[18 19]; if a user searches for data on 'diseases whose symptoms include cough' and finds no results, then they may be satisfied with data on diseases associated with sneezing, but not with diseases associated with loss of consciousness. Additionally, this similarity can be used to sort results according to how relevant they are to the query. Semantic similarity can also be used when comparing data collected using different methodologies. For example, data collected with an x-ray scan is more related to data collected with a CT scan than with a blood screening process. The Semantic Web can be made aware of this fact, thus enabling the comparison of heterogeneous datasets.

A CASE STUDY

The Semantic Web offers increased interoperability, more efficient search engines, and other useful tools for ontology creation, ontology maintenance and data analysis. To bring this potential into the hands of epidemiological data users, they must have access to a comprehensive collection of resources annotated with concepts from epidemiology-related ontologies. This can be bootstrapped by annotating existing epidemiological datasets using Semantic Web principles.

Following this vision, a platform for epidemiological resources named Epidemic Marketplace[20 21] is being developed at http://www.epimarketplace.net/ under the umbrella of the EPIWORK project,[22] a European project sponsored by the Seventh Framework Programme. Its primary goal is to serve as a repository of epidemiologically relevant information, in the form of resources. Each resource is given a unique identifier, and can be annotated by its author or any authorised Epidemic Marketplace user, with several ontological concepts in order to give them machine-readable semantics.

Annotation can be a laborious task, since annotators should be familiar with the properties and concepts to be used, and biomedical and geospatial ontologies tend to be large and very detailed. To assist users in correctly characterising their resources, the Epidemic Marketplace offers an interactive interface for resource annotation that provides a list of selected properties that cover most of the data characterisation needs required by

consumers and curators of epidemiological resources (eg, 'location', 'diagnostic method', 'drug'). Furthermore, to speed up the process of creating machine-readable metadata, the platform provides a user tool that converts free text (such as 'influenza') into the proper URI of the concept (http://purl.obolibrary.org/obo/DOID_8469). To characterise an epidemiological resource, the annotator has to fill the values of these properties with concepts extracted from a set of selected ontologies, Network of Epidemiology-Related Ontologies (NERO). These ontologies include concepts related to chemical compounds, diseases, environment, symptoms, taxonomy, modes of disease transmission, vaccines and geospatial information.[23]

The set of properties together with the concepts from NERO can facilitate the annotation process in the Epidemic Marketplace. Consider this illustration: in the Epidemic Marketplace, a dataset on possible treatments for AIDS can be easily annotated with: (1) the concept of AIDS, (2) the kinds of drugs that are mentioned in the dataset, such as antiretrovirals, (3) the locations where the data were collected, such as a continent, a country or a city, (4) the time where this collection occurred, (5) the socioeconomic conditions of the people included in the dataset, and so on. The Epidemic Marketplace is also already being used by epidemic modellers and statisticians for the storage and management of a Europe-wide surveillance data collection carried on by the consortium of Influenzanet,[24 25] a network of web-based monitoring platforms. Flu incidence data in the Influenzanet participating countries are already available in the Marketplace and can be searched through their metadata, which include, for instance, the pathogen, disease and host species.

CONCLUSION

The Semantic Web is a powerful paradigm that leverages on authors annotating their resources with meaningful, machine-readable metadata to perform tasks, such as information sharing, effective data search and pattern recognition. Ontologies play a very important role, since they provide both the concepts to be used in annotation and the facts about these concepts. Therefore, they allow automatic inference mechanisms underlying the Semantic Web tasks.

A key requirement for the dissemination of Semantic Web technologies among the epidemiology community is the implementation of tools for annotating epidemic resources in an efficient way, such as the ones provided by the Epidemic Marketplace, where an author is guided in the process of choosing the right concepts for use as metadata. Other forms of assistance, such as tools for automatic characterisation of resources, could also be used to bootstrap the availability of a large pool of semantically annotated resources of relevance to epidemiology.

For instance, to improve collaboration and interoperability, most journals on molecular biology encourage authors to submit their data to public databases, and use the respective accession numbers to mention the entities in the text (eg, genes, proteins, diseases). Therefore, the availability of tools, such as Epidemic Marketplace, may represent a starting point to foster authors to submit their resources, and then use the given URIs to mention them in the epidemiological journals. Note, that this does automatically make data publicly available, given data access restrictions that may exist; it means, however, that data will be consistently stored and coherently mentioned, so that both authorised persons and computers can efficiently find this information.

Ultimately, we expect that the implementation of this vision of information management into epidemiology pushes the field into a more self-integrated body of knowledge, with information and data easily flowing among researchers.

Acknowledgments The authors wish to thank the European Commission for the financial support of the EPIWORK project under the Seventh Framework Programme (Grant #231807), and the FCT for the financial support of the PhD grant SFRH/BD/69345/2010, the SOMER project (Grant PTDC/EIA-EIA/119119/2010) and the Multiannual Funding Programme.

Contributors JDF prepared the first draft, after which JDF, DP, FMC and MJS reviewed and commented on the manuscript.

Funding This work was supported by Seventh Framework Programme, grant number: EPIWORK project (Grant #231807). The funder had no role in the preparation/submission of the manuscript.

Competing interests None.

Provenance and peer review Commissioned; externally peer reviewed.

OPEN ACCESS

J Epidemiol Community Health 2012;**0**:1–4. doi:10.1136/jech-2012-201142

REFERENCES

1. **Porta MS.** *A dictionary of epidemiology 2008*. USA: Oxford University Press.
2. **Salathé M,** Bengtsson L, Bodnar T, *et al*. Digital epidemiology. *PLoS Comput Biol* 2012;**8**:e1002616.
3. **Shadbolt N,** Hall W, Berners-Lee T. The semantic web revisited. *Int Syst, IEEE* 2006;**3**:96–101.
4. **Doms A,** Schroeder M. GoPubMed: exploring PubMed with the gene ontology. *Nucleic Acids Res* 2005;**33**:W783–6.
5. **Bastos HP,** Tavares B, Pesquita C, *et al*. Application of gene ontology to gene identification. *Methods Mol Biol* 2011;**760**:141–57.
6. **Michalowski M,** Ambite JL, Thakkar S, *et al*. Retrieving and semantically integrating heterogeneous data from the web. *Int Syst, IEEE* 2004;**19**:72–9.
7. **Bizer C,** Heath T, Berners-Lee T. Linked data-the story so far. *Int J Semantic Web Info Syst* 2009;**5**:1–22.
8. **Sheth A,** Ramakrishnan C, Thomas C. Semantics for the semantic web: the implicit, the formal and the powerful. *Int J Semantic Web Info Syst* 2005;**1**:1–18.
9. **Auer S,** Bizer C, Kobilarov G, *et al*. Dbpedia: a nucleus for a web of open data. *Semantic Web* 2007;722–35.
10. **Lowery JC.** Getting started in simulation in healthcare. *Simulation Conf Proc* 1998;**1**:31–5.
11. **Samarati P,** de Vimercati S. Access control: policies, models, and mechanisms. *Foundations Sec Anal Des* 2001:137–96.
12. **Osborne J,** Flatow J, Holko M, *et al*. Annotating the human genome with disease ontology. *BMC Genomics* 2009;**10**:S6.
13. **Yahoo! Geoplanet™**. Accessed May 2012. http://developer.yahoo.com/geo/geoplanet/
14. **McGuinness DL,** Van Harmelen F. OWL web ontology language overview. *W3C Recomm* 2004;**10**:1–22.
15. **Motik B,** Patel-Schneider PF, Parsia B, *et al*. OWL 2 web ontology language: structural specification and functional-style syntax. *W3C Recomm* 2009;**27**. http://www.w3.org/TR/2009/REC-owl2-syntax-20091027/
16. **Cruz I,** Stroe C, Caimi F, *et al*. *Using AgreementMaker to align ontologies for OAEI 2011*. The Sixth International Workshop on Ontology Matching (OM-2011), 2011.
17. **Noy NF,** Shah NH, Whetzel PL, *et al*. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res* 2009;**37**:W170–3.
18. **Pesquita C,** Faria D, Falcão AO, *et al*. Semantic similarity in biomedical ontologies. *PLoS Comput Biol* 2009;**5**:e1000443.
19. **Ferreira JD,** Couto FM. Generic semantic relatedness measure for biomedical ontologies. *Proc Int Conf Biomed Ontologies* 2011.
20. **Silva M,** Silva F, Lopes L, *et al*. Building a digital library for epidemic modelling. *ICDL 2010—The International Conference on Digital Libraries* 2010:447–59.
21. **Lopes LF,** Silva FAB, Couto FM, *et al*. Epidemic marketplace: an information management system for epidemiological data. *Proceedings of the Information Technology in Bio- and Medical Informatics* 2010:31–44.
22. **Silva F,** Silva M, Couto F. Epidemic Marketplace: an e-Science Platform for Epidemic Modelling and Analysis. ERCIM News 82—Special Theme: Computational Biology, 2010.
23. **Ferreira JD,** Pesquita C, Couto FM, *et al*. Bringing epidemiology into the Semantic Web. *Proceedings of the International Conference on Biomedical Ontologies* 2012.
24. http://www.influenzanet.eu
25. http://www.epiwork.eu/resources/wp5-ict-monitoring-and-reporting-systems/