Bringing epidemiology into the Semantic Web

João D. Ferreira [1]*, Catia Pesquita [1], Francisco M. Couto [1], Mário J. Silva [1,2]

[1]Lasige, Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa
[2]IST, INESC-ID, Technical University of Lisbon

ABSTRACT

Epidemiology is a domain of knowledge interconnected with many other domains, thus making it a good candidate for reusing existing ontologies that, despite having been created for different purposes, characterize information frequently manipulated by epidemiologists and public health scientists. This paper presents an evaluation of existing ontologies for the semantic annotation of epidemiological resources. We selected a set of ontologies and proposed a Network of Epidemiology-Related Ontologies (NERO), which can form the core of semantic annotation for data-intensive epidemiology-related information systems, such as epidemic forecasting infrastructures. To support this selection, we defined a set of requirements for inclusion of ontologies in NERO, based on good ontology practice, the interdisciplinary nature of the epidemiological domain and support of semantic web technologies. Most of the selected NERO ontologies are current candidates or members of the Open Biological and Biomedical Ontologies initiative.

1 INTRODUCTION

Epidemiology is a truly multidisciplinary subject in the sense that it relies on diverse areas of knowledge, such as medicine, statistics, social sciences and geography. As a result, epidemiological data is both one of the most diverse and also one whose characterization can benefit the most from semantic technologies. In fact, only a framework able to understand heterogeneous and multidisciplinary resources can ultimately deal with all aspects of epidemiology. Ontologies can help address this issue by improving the integration, categorization and sharing of epidemiological resources.

Consider the following example: a research team is building a model for influenza spread and they need to know the number of infections caused by this disease over time to fit the parameters of their model. They are interested in building a model that works in France. By means of an appropriate query, they can try to find data resources about "influenza in France." To effectively find the resources in need using an automated process, these should be correctly annotated in advance. One of the best approaches for creating such annotations is to associate ontological concepts to each resource. The annotations could then be used by a service that would locate the resources satisfying the query and rank them by means of semantic similarity applied to the ontological concepts associated to the query and the annotated resources. Furthermore, by using ontologies, the process of annotating a resource can be facilitated through the identification of concept labels using text mining for analysis of the content of the resource.

To provide the service described in the above scenario, it is paramount to establish a comprehensive and powerful set or inter-related semantic concepts ensuring the consistency of the annotations. By doing so, we stimulate the creation of tools to serve epidemic modelers and epidemiologists in general, thereby contributing to the adoption of semantic technologies in epidemiology.

This paper proposes NERO, a Network of Epidemiology-Related Ontologies, a first step towards the generation of a useful source of epidemiological concepts and the relations among them supporting the characterization of data resources used in epidemiological studies. In this aspect, NERO can be considered an ontology *per se*, but the disparate set of domains in this area suggests that it should, in fact, be composed of a number of distinct inter-related ontologies. To ensure good interoperability between these ontologies, we propose a list of requirements for inclusion in NERO. A survey to the state-of-the-art in epidemiologically relevant ontologies yielded a set of ontologies that partially fulfill the proposed requirements.

2 METHODOLOGY

The ontologies to be included in NERO were selected based on the domains of knowledge that are generally present in epidemiological ontologies (diseases, locations etc.; see Table 4.1 for a full list of NERO domains). However, since its goal is to serve epidemiological research, we also based this selection on the needs of one particular epidemic research platform, the Epidemic Marketplace (EM) (Lopes *et al.*, 2010). This website provides storage capabilities for all kinds of epidemiological resources, as well as a full set of associated services, such as adding comments to the resources and sharing the resources with other parties. Within the EM, NERO concepts can be used to annotate epidemiological resources (see, for an example, http://epimarketplace.net/metadata/example). However, as illustrated in Figure 1, even though NERO is based on our experience of developing the EM, it can be applied outside the

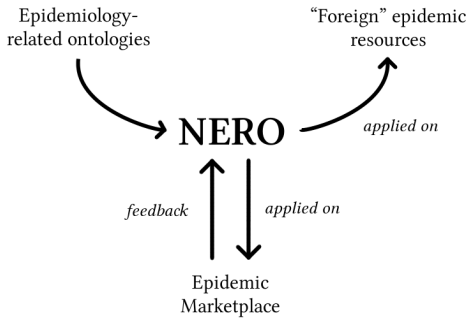

Fig. 1. NERO was created by analysing epidemiology-related ontologies; additionally, NERO and the EM are related to one another since NERO was created based on specific needs for the EM. However, NERO is external to the EM and can be used outside that context by any epidemiologist.

*To whom correspondence should be addressed:
joao.ferreira@lasige.di.fc.ul.pt

scope of that system, as it is a general network of epidemiological concepts applicable to any epidemiologically-related context.

We started by discussing which requirements should be fulfilled by the ontologies to be included in NERO, in order to create a concise, comprehensive and good quality network of ontological concepts for the epidemiological domain. We then surveyed current ontologies in this domain and evaluated them according to their fulfillment of the requirements. Since it was found that these ontologies generally fail in satisfying many of the requirements, we then considered general-purpose ontologies, like UMLS, and ontologies with a more focused domain, like the Disease Ontology.

3 THE NETWORK OF EPIDEMIOLOGICALLY-RELATED ONTOLOGIES

The ability to reuse concepts from one application in another is one of the premises of the semantic web (Shadbolt *et al.*, 2006). Given that the main purpose of this work is to define a source of concepts to use primarily for the annotation of epidemiological datasets, we want the advantage of not having to deal with maintaining and curating these concepts ourselves, but instead to leverage on existing ontologies for that. Therefore, we propose the Network of Epidemiology-Related Ontologies (NERO), a collection of ontologies that cover the epidemiological domain and enable the annotation of epidemiological resources with the relevant concepts.

In any comprehensive collection with multiple provenances, there must be a set of requirements ensuring and enabling both a good interoperability among those resources and an overall cohesive structure. Besides the specific requirements derived from the particular goals of NERO, some of the requirements presented here include adaptations of (i) principles of the W3C semantic web (Koivunen and Miller, 2001); and (ii) principles of the OBO Foundry (Smith *et al.*, 2007).

This list should be considered as a set of guidelines, not as absolutely mandatory rules. They lead NERO towards an ideal scenario where ontologies have a high level of expressibility and good interoperability between each other, all the while enabling a simple, yet powerful, implementation of semantic web technologies.

We propose ten requirements:

1. **Relevant Domain** The most important requirement for incorporating an ontology in NERO is that it should encode a domain of knowledge that is interesting from the point of view of epidemiology, i.e. it must contain concepts that are relevant to annotate epidemiological resources. Likewise, the full network should cover almost all of the epidemiological domain (diseases, modes of transmission, geographical locations, etc.).
2. **Appropriate granularity** To achieve high domain coverage and improve the semantic characterization of epidemiological resources, an ontology must provide an adequately detailed representation of its domain. Biomedical and geospatial ontologies tend to comply to this requirement quite well, and in fact some contain many thousands or even tens of thousands of concepts spread over many levels of depth, allowing specific annotations such as the exact strain of a virus instead of its family. In contrast, the best ontology to describe a given domain can be too granular for the purpose of epidemiological annotation (e.g. *photon* in ChEBI). In these cases, NERO should disregard the unwanted branches of the ontology.
3. **Expressiveness with tractability** A large number of concepts is an important advantage, but being able to manage the ontologies is also a technical requirement, and as such the ontologies must be well structured and tractable from a computational point of view. Specifically, this means that the properties must be formally defined and that these definitions should be adjusted to the domain in question. For example, in an the anatomy domain, it does not make sense to have only an *is-a* relationship type, since properties like *part-of* and *arterial-supply* are equally relevant.
4. **Cross-references** Ontologies encoding different domains can, nevertheless, be related to one another. For example, symptoms are usually associated with diseases and vice-versa. External references that cross from one ontology to another link together concepts from different domains, enabling semantic web technologies to explore multiple-domain relations.
5. **Textual definitions** Since NERO will be employed primarily to annotate epidemiological resources, it is important that its users understand the meaning of each concept. Ontologies encode theirs concepts' meaning in machine-readable code, but for users to correctly identify the concept they want, concepts should be complemented with textual definitions.
6. **List of synonyms** As synonyms are abundant in natural language, particularly in the biomedical field, it is important that NERO explicitly states these synonyms (e.g., "AIDS" is a synonym of "acquired immune deficiency syndrome" and users expect that both refer to the same concept).
7. **Popularity** NERO ontologies should be well known in the epidemiology community, since familiarity increases the chance that users more easily choose the correct concept. Furthermore, if an ontology is popular, there is a higher probability that its development does not stall in the foreseeable future, ensuring that NERO is kept updated with the most current knowledge.
8. **Publicly available** It is imperative to adopt an open-source ontology rather than one that needs licenses or other form of control over usage. This requirement stems from the fact that we intend NERO to be publicly available to anyone in order to further increase the spread of semantic web technologies into epidemiology. Also important, in this context, is that users have the opportunity to submit corrections, suggestions and other improvements to the ontologies.
9. **Persistent identifiers** Since ontologies constantly change in response to advances in the field, errors found, etc., some concepts may change their definition, ultimately resolving in some annotations becoming wrong. To mitigate this issue, ontologies in NERO should have semantic-free identifiers which are never removed from the ontology.
10. **Distributed access to the ontology** Several languages have been developed to encode ontologies (most notably OWL and OBO format); but other formats exist, from simple tree-like structures described in a text document to tables on a database. Instead of having to cope with all these differences, NERO ontologies should be easily accessible through "the cloud", (web services or equivalent). Moreover, by not having a local copy of the ontology, there is no need to take special actions in order to keep it up-to-date.

The first three requirements (**domain**, **granularity** and **expressiveness**) are scope-related, since they refer specifically to the knowledge encoded in the ontology itself. The others are properties that simplify the tractability of an ontology and improve its usefulness as a scientifically sound source of concepts for annotating epidemiological resources, while ensuring a certain degree of user-friendliness, which is important given that one of the aims of NERO is to facilitate the annotation process.

4 RESULTS

This section presents a survey of the state-of-the-art in ontological representation of the epidemiological domain and an overview of currently existing ontologies that, despite having been created for other purposes, can be used to describe concepts relevant for this field of research, such as diseases, modes of transmission, demographics or geography. A summary of the considered ontologies and the domains they represent is given in Table 4.1.

As we are interested in covering the whole spectrum of epidemiological domains, we have not defined a minimum number of requirements to be fulfilled by NERO ontologies. In fact, not all requirements are equally important. Instead, NERO includes the most well-adjusted ontology for each domain, and for the topics where we could not find ontologies of good quality, it fills the gap with controlled vocabularies, such as taxonomies and dictionaries. It is expectable that relevant ontologies will be developed in the future, and given the modular nature of NERO, these could, when available, easily replace the lower quality terminologies.

4.1 Ontologies specific to the epidemiological domain

There have been two attempts at organizing epidemiological terminologies in a hierarchical manner (Frank *et al.*, 2009; Lynch *et al.*, 2007). These two works have resulted in published material describing the ontologies, but neither points the reader to a place where such ontologies can be downloaded or browsed.

There has also been a number of automatic systems designed to monitor epidemic surges. One example is the BioCaster Global Health Monitor (Collier *et al.*, 2008), a news filter created with the aim of providing "an early warning monitoring station for epidemic and environmental diseases". BioCaster is based on an ontology published in a standard format (OWL), allowing for an easy integration in current semantic web technologies. It contains approximately 2000 entities. While this number may be appropriate for BioCaster purposes (text mining of news articles), it is poor as a source of annotation concepts. For example, only five countries appear in the ontology, and while there are a number of diseases and syndromes, they are shallowly organized (diseases are instances of *Avian Disease*, *Human Disease* or other similar classes, all of which are direct subclasses of the concept *Disease*). Concepts of the therapeutics domain are not well represented (the ontology contains the concept *therapeutic role*, but no other in this area) and there is no concept of vaccination. Overall, we observe that the majority of concepts in this ontology is better represented in other ontologies.

Another ontology built especially for epidemiological studies is the Epidemiology Ontology, developed by HuGE NET (Khoury and Dorman, 1998; HuGE Net, 2007). It is not as well structured as the BioCaster ontology, as it consists of a single hierarchy of terms related with each other by a single property type, leading to *Person* being described under *Hypothesis Formulation from Descriptive Studies* or *Hospital* under *Notifiable disease*. It contains 791 distinct concepts, some of which also appear in the Dictionary of Epidemiology (Porta, 2008), a dictionary containing a detailed list of concepts of the epidemiological field. Despite the alphabetic organization and the absence of a hierarchy, each entry has a detailed description of its meaning and some form of structure given in the form of references to other entries. However, it is also poor in quantity and ontological structure. Considering all the domains of NERO, and given the low coverage of the Epidemiology Ontology and the Dictionary of Epidemiology in domains such as geography or diagnostic methods, we believe that they have limitations. Just like BioCaster, however, they can help by providing a sense of which domains NERO should represent.

4.2 Other ontologies containing epidemiological concepts

Given the low suitability of those resources for inclusion in NERO, our focus moved to ontologies not built for epidemiology but which, nonetheless, contain relevant epidemiological concepts. The relevant domains were assumed to be the ones in need for the EM metadata model, which agree with the domains of knowledge represented in BioCaster, the Epidemiology ontology and the Dictionary of Epidemiology, namely: demography, diagnostic methods and other clinical methods, diseases, drugs, environment, geographical location, socio-economic conditions, symptoms, taxonomy, transmission modes and vaccination.

Some research has been conducted in epidemiology based on the use of existing ontologies containing epidemiologically relevant concepts, but which were not designed with that specific domain of knowledge in mind. This category includes the Unified Medical Language System (UMLS), a "collection of ontologies and terminologies that promote the creation of more effective and interoperable biomedical information systems and services" (Lindberg *et al.*, 1993), and Medical Subject Headings (MeSH), a controlled vocabulary used to index articles in biomedical sciences (Lipscomb, 2000). These resources can be seen as hierarchies of terms, where a term directly descends from one or more terms, thus creating a graph-like structure.

As an example, the work of Xu *et al.* (2010) uses UMLS to mine for epidemiologically relevant concepts in articles. While it could prove useful in our endeavor, UMLS is a large resource, with over one million concepts; properly scanning through this terminology and determining the relevant concepts is too colossal a task for the typical epidemic modeler.

Additionally, MeSH is relatively unstructured and makes use of a single relation (just like the Epidemiology Ontology presented above). For example, *Axial length* and *Eyebrow* are categorized under *Eye*, but one is a property and the other is a nearby structure. Likewise, *Eye* is both categorized under *Sense Organs* and *Face*, but while it *is* a sense organ, it is *part of* the face. MeSH makes no distinction between these semantic relations, which we consider one of the main drivers for the use of ontologies.

There are other limitations with UMLS and MeSH: since they have a generic and broad domain, the addition of new concepts is non-trivial, and there is a high risk of introducing errors and inconsistencies. In fact, it is known that UMLS houses many inconsistencies (Geller *et al.*, 2009). These two resources are not

Terminology	Domain	Ref.	Fulfills requirement #...										In NERO?
			1	2	3	4	5	6	7	8	9	10	
BioCaster	Epidemiology	(Collier *et al.*, 2008)	Y	N	±	Y	Y	Y	N	Y	N	N	No
Epidemiology Ontology	Epidemiology	(HuGE Net, 2007)	Y	Y	N	N	N	Y	N	Y	N	N	Yes
Dictionary of Epidemiology	Epidemiology	(Porta, 2008)	Y	Y	N	N	Y	±	N	Y	N	N	No
UMLS	General	(Lindberg *et al.*, 1993)	Y	Y-	N	Y	Y	Y	Y	±	Y	N	No
MeSH	General	(Lipscomb, 2000)	Y	Y-	N	N	Y	Y	Y	Y	Y	N	Yes
SNOMED-CT®	General	(Stearns *et al.*, 2001)	Y	Y-	Y	Y	Y	Y	Y	N	N	N	No
GeoPlanet™	Geography	(Yahoo!, 2011)	Y	Y	±	Y	±	Y	Y	±	Y	Y	Yes
GeoNames	Geography	(Geonames.org, 2011)	Y	Y	N	Y	N	Y	Y	Y	Y	Y	No
Geo-Net-PT	Geography	(Lopez-Pellicer *et al.*, 2009)	Y	N	Y	Y	±	Y	N	Y	Y	Y	No
OBO ontologies													
ChEBI	Biochemistry	(de Matos *et al.*, 2010)	Y	Y-	Y	Y	Y	Y	Y	Y	Y	Y	Yes
DOID	Diseases	(Osborne *et al.*, 2009)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Yes
ENVO	Environment	(EnvO developers, 2012)	Y	±	Y	±	Y	Y	N	Y	Y	Y	Yes
HP	Symptoms	(Robinson and Mundlos, 2010)	Y	±	Y	Y	Y	Y	Y	Y	Y	Y	Yes
IDO	Diseases	(Cowell and Smith, 2010)	Y	±	Y	N	Y	N	Y	Y	Y	Y	Yes
NCBI Taxonomy	Taxonomy	(Wheeler *et al.*, 2007)	Y	Y-	Y	N	N	N	Y	Y	Y	Y	Yes
NCI Thesaurus	General	(Sioutos *et al.*, 2007)	Y	Y-	Y	Y	Y	Y	Y	Y	Y	Y	Yes
SYMP	Symptoms	(Schriml *et al.*, 2010)	Y	Y-	Y	Y	Y	Y	Y	Y	Y	Y	Yes
TRANS	Disease transmission	(Schriml *et al.*, 2010)	Y	N	Y	±	Y	N	N	Y	Y	Y	Yes
VO	Vaccines	(Yang *et al.*, 2011)	Y	Y	Y	N	Y	N	N	Y	Y	Y	Yes

Table 1. The terminologies found in the survey on epidemiological ontologies, their evaluation based on the requirements of NERO, and whether they were included in NERO. **Legend**: **Y** – terminology fulfils the requirement; ± – terminology partly fulfils the requirement; **N** – terminology does not fulfil the requirement; **Y-** (on requirement 2) – terminology is more granular than required. This last rating is used to suggest that, before inclusion in NERO, the ontology should suffer a previous step of selecting the appropriate branches needed in Epidemiology.

published in a standard semantic web format, meaning that they do not integrate well with semantic web technologies. Moreover, UMLS needs a license to be used.

SNOMED-CT®, a subset of UMLS, was also considered individually, since it is a comprehensive terminology for diseases and other clinical terms (Stearns *et al.*, 2001). Since this is a terminology that needs a license to be used, it does not entirely fit NERO's purposes. Furthermore, its domains are (although sometimes with less detail) represented in other ontologies (like DOID, see Table 4.1).

In face of these issues, we turned to attempts to create and organize more formal ontologies in the biomedical field. OBO should be highlighted, since it aims at providing a suite of interoperable reference ontologies in the biomedical domain (Smith *et al.*, 2007). The OBO Foundry defines a set of principles that must be fulfilled by an ontology before it is included. There are currently eight OBO ontologies, but other candidates are presently working to fulfill the required principles for being endorsed by the OBO Foundry. Given that OBO's set of principles enforces good quality ontologies by promoting good practices in ontology development, and that any one of these ontologies, both supported and candidate, strives to fulfill those principles, we included some of them in NERO (see Table 4.1). Because the ontologies of OBO span over many biological and biomedical domains of knowledge, these domains can thus be well covered in NERO.

Non-biological concepts must be retrieved from other resources. Yahoo! GeoPlanet™ (Yahoo!, 2011) contains a representation of the world geography, and is in fact a good candidate for inclusion in NERO. Other geographical ontologies were considered, such as GeoNames and Geo-Net-PT. GeoNames (Geonames.org, 2011) is a flat dictionary of locations on Earth, lacking an ontological structure. For instance, there is no relation between *Italy* and *Rome* (its capital) or *Italy* and *France* (one of its neighbors). Geo-Net-PT (Lopez-Pellicer *et al.*, 2009) is an ontology of the Portuguese territory and, despite being rich in detail, it covers a small scope of the Earth. However, there are correspondences between Yahoo! Geoplanet™ and Geo-Net-PT (Ferreira *et al.*, 2010); therefore, if a more detailed annotation is required, Geo-Net-PT would be a good complement in the area it covers.

We have been unable to find ontologies that specifically represent demography or social and economic conditions, and suspect that none exist that are publicly available. As such, we will have to rely on resources such as MeSH, the Epidemiology Ontology and NCI Thesaurus for those domains. In this context, it is important to mention that resources like the Dictionary of Epidemiology can be curated in an ontological format (e.g., through triplification (Hitzler and van Harmelen, 2010)), introducing semantics to its entries and allowing the application of semantic web technologies to them.

A graphical summary of this survey is shown on Table 4.1, which details the requirements fulfilled by each of the ontologies found, and which of them were included in NERO.

5 CONCLUSION

This paper proposes the Network of Epidemiology-Related Ontologies (NERO), which aims at providing a core of relevant concepts to semantically characterize epidemiological resources and therefore enable the application of semantic technologies in the epidemiological domain. NERO is being developed as part of the

Epidemic Marketplace platform (EM) to annotate its resources, but can be further used by any epidemiologist or public health scientist.

The choice of a network of ontologies was motivated by the fact that existing epidemiology ontologies (BioCaster and the Epidemiology Ontology) are not comprehensive enough to suitably model the epidemiology domain. Nevertheless, they offer an insight into what an epidemiological network of ontologies should cover and how it should be organized, since they contain the branches of knowledge required in epidemiology (diseases, modes of transmission, locations, social conditions, etc.). By crossing this information with the needs of the EM, we concluded that selected ontologies from the OBO project complemented with branches of MeSH and the Epidemiology Ontology are the most appropriate terminologies for the biomedical portion of epidemiology, since together they span over a large amount of the biomedical domain on this area. Other resources, such as UMLS or SNOMED-CT®, are not as suitable due to their licensing, complexity, and difficulty in integrating into a fully semantic web approach.

For geographical information, we included the geographical ontology of Yahoo! GeoPlanet^{TM}, based on its higher quality when compared to the other candidates. For the domains of demography and social and economical conditions, the lack of any specific ontologies prompted us to consider branches of MeSH and NCI Thesaurus. Additionally, the Epidemiology Ontology has also concepts in these domains.

Once epidemiological resources are annotated with NERO, it will be possible to exploit these annotations to perform complex semantic analysis on diverse data processing tasks tasks, such as information retrieval, integration and extraction. These tasks will provide epidemiologists, particularly epidemiology modelers, with tools that enable an easy discovery of models and the parameters to use in them. The Epidemic Marketplace is currently being developed with the intention of serving as a starting point for this semantic analysis.

Finally, it is worth noticing that, at the moment, NERO is little more than a list of requirements and the identified set of ontologies. As future work, we plan to completely integrate these ontologies in a more tangible network. For instance, semi-automatic alignments between the ontologies should produce pairs of equivalent concepts that can be merged. The mapping of these ontologies into upper-domain ones (such as BFO) should also contribute to a better integration of all the ontologies. Eventually, we plan to introduce other semantic web technologies that will make use of NERO, such as RDF triple stores and SPARQL endpoints.

ACKNOWLEDGEMENTS

The authors want to thank the European Commission for the financial support of the EPIWORK project under the Seventh Framework Programme (Grant #231807), and the Portuguese Fundação para a Ciência e Tecnologia through the financial support of the SOMER project (PTDC/EIA-EIA/119119/2010), the PhD grants SFRH/BD/42481/2007 and SFRH/BD/69345/2010, and the PIDDAC Program funds (INESCID multi annual funding) and through the LASIGE multi annual support.

REFERENCES

Collier, N., Doan, S., Kawazoe, A., *et al.* (2008). Biocaster: detecting public health rumors with a web-based text mining system. *Bioinformatics*, **24**(24), 2940–2941.

Cowell, L. G. and Smith, B. (2010). Infectious disease ontology. In V. Sintchenko, editor, *Infectious Disease Informatics*, pages 373–395. Springer New York.

de Matos, P., Alcántara, R., Dekker, A., *et al.* (2010). Chemical Entities of Biological Interest: an update. *Nucl. Acids Res.*, **38**, D249–D254.

EnvO developers (2012). The envo project. http://gensc.org/gc_wiki/index.php/EnvO_Project. Accessed Jan 5, 2012.

Ferreira, J. D., Batista, D. S., Couto, F. M., and Silva, M. J. (2010). The geo-net-pt/yahoo! geoplanet (tm) concordance. Technical Report 2010;5, Universidade de Lisboa, Faculdade de Ciências. http://hdl.handle.net/10455/6677.

Frank, G., Wheaton, W., Bakalov, V., Cooley, P., and Wagener, D. (2009). An ontology for designing models of epidemics. In *Proceedings of ICBO*, volume 2009, pages 47–50.

Geller, J., Morrey, C., Xu, J., *et al.* (2009). Comparing inconsistent relationship configurations indicating umls errors. In *AMIA Annual Symposium Proceedings*, volume 2009, page 193. American Medical Informatics Association.

Geonames.org (2011). Geonames. http://www.geonames.org/. Accessed Dec 9, 2011.

Hitzler, P. and van Harmelen, F. (2010). A reasonable semantic web. *Semantic Web*, **1**(1), 39–44.

HuGE Net (2007). Guidelines for the epidemiological ontology. http://www.hugenet.org.uk/resources/informatics/Ontology_Version_1.pdf. Accessed Dec 16, 2011.

Khoury, M. J. and Dorman, J. S. (1998). The human genome epidemiology network. *American journal of epidemiology*, **148**(1), 1–3.

Koivunen, M.-R. and Miller, E. (2001). W3c semantic web activity. *Semantic Web KickOff in Finland*, pages 27–41.

Lindberg, D. A., Humphreys, B. L., and McCray, A. T. (1993). The unified medical language system. *Methods of information in Medicine*, **32**(4), 281.

Lipscomb, C. E. (2000). Medical subject headings (mesh). *Bulletin of the Medical Library Association*, **88**(3), 265.

Lopes, L., Silva, F., Couto, F., *et al.* (2010). Epidemic marketplace: an information management system for epidemiological data. *Information Technology in Bio-and Medical Informatics, ITBAM 2010*, pages 31–44.

Lopez-Pellicer, F. J., Chaves, M., Rodrigues, C., and J., S. M. (2009). Geographic ontologies production in grease-ii. Technical Report 09-18, Universidade de Lisboa, Faculdade de Ciências. http://hdl.handle.net/10455/3256.

Lynch, C. O., Cunningham, C., Schripsema, E., Morris, T., and Rhodes, B. (2007). A biosurveillance platform for biosense message analysis using integrated reference ontologies and intelligent agents. In *AAAI Fall Symposium*, volume 2007.

Osborne, J., Flatow, J., Holko, M., *et al.* (2009). Annotating the human genome with disease ontology. *BMC genomics*, **10**(Suppl 1), S6.

Porta, M., editor (2008). *A Dictionary of Epidemiology*. Oxford University Press, USA, 5^{th} edition edition.

Robinson, P. N. and Mundlos, S. (2010). The human phenotype ontology. *Clinical genetics*, **77**(6), 525–534.

Schriml, L. M., Arze, C., Nadendla, S., *et al.* (2010). Gemina, genomic metadata for infectious agents, a geospatial surveillance pathogen database. *Nucleic Acids Research*, **38**(suppl 1), D754.

Shadbolt, N., Hall, W., and Berners-Lee, T. (2006). The semantic web revisited. *Intelligent Systems, IEEE*, **21**(3), 96–101.

Sioutos, N., Coronado, S., Haber, M. W., *et al.* (2007). Nci thesaurus: a semantic model integrating cancer-related clinical and molecular information. *Journal of biomedical informatics*, **40**(1), 30–43.

Smith, B., Ashburner, M., Rosse, C., *et al.* (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, **25**(11), 1251–1255.

Stearns, M. Q., Price, C., Spackman, K. A., and Wang, A. Y. (2001). Snomed clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*, page 662. American Medical Informatics Association.

Wheeler, D. L., Barrett, T., Benson, D. A., *et al.* (2007). Database resources of the national center for biotechnology information. *Nucleic acids research*, **35**(suppl 1), D5–D12.

Xu, H., Lu, Y., Jiang, M., *et al.* (2010). Mining biomedical literature for terms related to epidemiologic exposures. In *AMIA Annual Symposium Proceedings*, volume 2010, page 897. American Medical Informatics Association.

Yahoo! (2011). Yahoo! geoplanet^{TM}. http://developer.yahoo.com/geo/geoplanet/. Accessed Dec 16, 2011.

Yang, B., Sayers, S., Xiang, Z., and He, Y. (2011). Protegen: a web-based protective antigen database and analysis system. *Nucleic Acids Research*, **39**(suppl 1), D1073.