Digital preservation of epidemic resources: coupling metadata and ontologies

João D. Ferreira
LASIGE, University of Lisbon
Portugal
joao.ferreira@lasige.di.fc.ul.pt

Catia Pesquita
LASIGE, University of Lisbon
Portugal
cpesquita@di.fc.ul.pt

Francisco M. Couto
LASIGE, University of Lisbon
Portugal
fcouto@di.fc.ul.pt

Mário J. Silva
INESC-ID, University of Lisbon
Portugal
mjs@inesc-id.pt

ABSTRACT

The preservation of epidemiological information is challenging in several aspects, since this is both a data-intensive and multidisciplinary subject, with large amounts of data spanning several domains of knowledge. We present, as a case study, the Epidemic Marketplace (EM), a platform dedicated to the preservation of epidemiological resources. To ensure integrity of the data, the EM uses a metadata model coupled with the Network of Epidemiology-Related Ontologies (NERO), a compilation of ontologies covering several domains of epidemiology. This enables users to quickly annotate their resources with concepts from those ontologies, increasing their visibility. Additionally, the ontologies of NERO offer support for future development, guaranteeing longevity of the metadata. This ensures that the information about the resources, such as its authors, is preserved and can be searched even in the absence of the data itself.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries; H.3.5 [**Online Information Services**]: Data Sharing

Keywords

Data-intensive research, Digital Curation, Ontologies, Data sharing, Epidemiology

1. INTRODUCTION

Epidemiology research is a truly multidisciplinary subject as it relies on areas of knowledge as diverse biology, medicine, statistics, social sciences and geography. It requires, for instance, computational methods to predict the spread of a disease, realistic large scale models and automatic data collection. Only a framework able to accommodate these methodologies can ultimately deal with epidemiology.

Epidemiology is also highly data-intensive, making it a direct case supporting the fourth paradigm of science [9], which addresses the challenges raised from the need to validate, analyze, visualize, store, and curate the large amounts of generated data. For example, models of the spread of an epidemic disease rely on large amounts of information. This information can be very general and easily located (*e.g.* the size of the population being infected), or specific to the population and to the disease (*e.g.* the rate of contact between people). Finding this information in literature can be difficult, if not impossible, in the time frame of utility of those models. In fact, while some diseases have periodic surges, like the flu, with an expected number of peaks per year, other surges are more unpredictable, and modeling them in real time, as the disease progresses, requires the quick collection of the necessary information. The *E. coli* outbreak in Europe in 2011 is an example of such a situation.

Given these characteristics, it is particularly crucial to guarantee the preservation of epidemiological data, to ensure that it remains available, reliable and usable for the future [3]. If past data can be easily retrieved and explored, then the probability of it being reused increases, which is especially relevant in a complex domain such as epidemiology, where combining data across different locations, diseases or even time can lead to new insights and new knowledge.

2. THE EPIDEMIC MARKETPLACE

To answer to this need in information preservation, we have created the Epidemic Marketplace (EM), a platform for epidemic research that enables and encourages epidemiological data sharing, enabling the community to perform data-intensive research [11]. The EM was developed as part of the European Epiwork project, which aims at creating the appropriate framework of tools and knowledge needed for the design of epidemic forecast infrastructures to be used by epidemiologists and public health scientists [13]. It emphasizes the urge to share the information within the epidemiology community, and directly demonstrates the advantages of doing so. In fact the collaboration with other partners of the Epiwork project has shown the need to standardize the sharing of the digital resources.

3. THE EM METADATA MODEL

Metadata is an essential component of digital sharing and preservation, since it ensures that the data can be uniquely identified and accurately described to support future retrieval and reuse. As such, the EM establishes a metadata model to annotate its resources, based on the Dublin Core (DC). DC was chosen to be the base of this metadata model because it is an interoperable metadata standard, it provides a semantic vocabulary with many of the elements needed to manage an online datastore, and it enables a straightforward extension.

The metadata model defines the Network of Epidemiology-Related Ontologies, which contains concepts that are relevant for characterization of epidemiological resources. Our approach has the benefit of increasing interoperability with external services and, by restricting annotation to ontology-based controlled vocabularies, we move closer to the idea of a Web of Knowledge instead of a Web of Text [1].

The EM metadata model provides elements for (i) *technical information* (the uploader, an internal identifier and the date of submission); (ii) *general information* related to the digital resource, (*e.g.* title, creator – which need not be the same as the uploader); and (iii) *content-specific information*, such as the subject, the sources used by the resource or even the epidemiological information that makes up the resource.

The terms offered by the DC can already handle many of the requirements of the EM, especially in the *technical* and *general* information areas. However, epidemiology relies on multiple domains of knowledge. Accordingly, the metadata model devised for this purpose must extend the core elements of DC with tags appropriate for these domains of epidemiology. For example, many epidemiological resources deal with one or more diseases, a concept absent from DC; as such, the EM metadata model contains a specific element, `<em:disease>`, suitable for annotating a resource with the diseases it refers to. This property roughly translates to "the resource refers to disease X". Using metadata in this fashion ensures that the resource is searchable not only based on the general information provided by the DC but also based on its epidemiology-specific contents.

Furthermore, we extended some of the DC elements with new epidemiological elements. For example, the content-specific element `<em:biologicalInformation>` is refined by a number of biologically relevant elements, such as the previously mentioned `<em:disease>`.

Additionally, the metadata model specifies the expected values that can be used to fill each element. Some expect literal values, such as `<em:title>`, which expect a string. Most of the *content-specific* information must be selected from ontologies of an appropriate domain. For example, to fill the `<em:disease>` element, instead of the literal "flu", the URI `http://purl.obolibrary.org/obo/DOID_8469` should be used. This is the identifier of the concept named "Influenza" in the Human Disease Ontology. Several ontologies have been collected in a network of relevant ontologies, which have been integrated in the EM so that users of the platform can search them and correctly annotate their resources.

4. NERO

Most of the *content-specific* elements of the EM metadata model are filled with concepts from ontologies. To properly encourage users to annotate their data and ensure preservation, we integrated into the EM a number of ontologies that provide appropriate concepts that assist users during the annotation process. These were collected into a Network of Epidemiology-Related Ontologies (NERO) [8].

NERO directly contributes to the preservation of epidemiological resources in at least three ways:

1. its ontologies were selected in order to ensure both availability and longevity;
2. the meaning of the concepts is guaranteed to remain unchanged; if some modification happens, the concepts are marked as deprecated and a pointer to the new concept is made. This means that there will always be opportunity to update a deprecated annotation with the new term or to reconsider it;
3. as with any ontology, NERO allows the full spectrum of semantic web technologies to be used to search resources: it enables performing simple but powerful queries on the EM, or to draw inferences based on the semantics of these annotations [2], ensuring that pertinent data can be more easily retrieved and subsequently used, and thus fulfilling one of the main goals of digital preservation [3].

The ontologies contained in NERO were selected based on a number of requirements, some of which are related to the preservation of epidemiological resources. For example, these ontologies are required to provide textual definitions for their concepts, to be popular among the communities that use them and to be publicly available. All these properties contribute to the preservation of metadata integrity.

In our search, we found ontologies that already try to model the epidemiological domain. Given their low coverage and granularity, they were deemed inappropriate for inclusion in NERO. However, they provided a sense of the concepts that should be modeled in an epidemiological resource. Some general-purpose ontologies contain concepts of epidemiological interest. From a preservation point of view, these ontologies are adequate for annotation. However, properly scanning through these large terminologies and determining which of their concepts are relevant would be too colossal a task for the typical epidemiologist.

In face of these issues, we ended up selecting mainly single-domain ontologies for NERO. The OBO Foundry [14] defines a set of principles that must be fulfilled by an ontology before it can be included, enforcing good quality by promoting good practices in ontology development. In particular, its ontologies are public domain and must guarantee versioning, documentation, etc., which contribute to manageable preservation of their contents. Given their association to a high profile initiative, these ontologies are more likely to be kept available and up-to-date in the future.

5. INTEGRATION OF NERO IN THE EM

Annotation of resources with metadata is only effective if the users are encouraged to create this annotations. For this reason, there are two mechanisms in NERO that facilitate this process.

When users upload a resource to the EM, they are required to fill-in a minimal set of mandatory metadata elements, which include `<em:title>` and `<em:description>`. For the annotation of epidemiological information, the EM provides an autocomplete function that, based on user input, retrieves concepts from NERO which are appropriate for the metadata field in question. This effectively hides the technical details of the ontologies from the regular users, letting them focus on semantic annotation.

Additionally, each item in the list of suggestions is associated with its description, which users can read to help them choose the concept that better describes the resource. Whenever a given characteristic of the resource cannot be accurately described by any of the available ontology concepts, the user can easily assign a more general concept, which is supported by the inherent hierarchical nature of ontologies.

The second mechanism is the exploration of the actual resource provided by the user with text-mining to suggest back annotations that the user might think are relevant. This functionality uses NCBO BioPortal's Annotator service [10] to read the content of the resource, where possible, and preloads the annotation form with the NERO concepts it finds.

Given the variable nature of epidemiological resources, not all resources will need to be annotated in all metadata fields. For instance, a resource focused on tracking the geographical spread of a disease probably won't refer to any drugs, or if it focuses on the treatment of a disease, it might not include information on diagnostic method. In a recent analysis we conducted of semantically annotating over 100 Epidemiological resources in the EM, and found that all resources mentioned at least one disease and one geographical location, about 80% included information about the diagnostic method and the pathogen involved, but only about 30% mentioned any drugs or vaccines.

One crucial feature of the EM and NERO integration is the ability to assign multiple concepts to the same metadata field, since many resources mention multiple diseases, symptoms, drugs, etc., mirroring the wide scope of epidemiology. This effectively enables crossing information from different resources referring the same or similar entities, such as diseases or drugs, to support broader studies.

The adoption of a metadata model to support the semantic annotation of epidemiological resources, ensures a more structured annotation process, effectively guiding the annotation itself. Furthermore, by coupling the metadata model with NERO, the annotation process is further simplified, since terms to fill-in metadata fields are retrieved from a controlled vocabulary which is backed by the rich properties of ontologies such as hierarchical structure, definitions and other properties and relations.

6. BENEFITS FOR EPIDEMIOLOGY

Once epidemiological resources are annotated with NERO, metadata can be used in complex semantic analysis as part of diverse tasks, such as information retrieval and information extraction. These tasks will provide epidemiologists, particularly the modelers, with tools that enable an easy discovery of models and the data needed to parametrize them.

There are two main challenges in accomplishing this goal. The first is to define a way to effectively compare resources that are annotated using different sets of ontologies, *i.e.* how to compare a resource annotated with disease and pathogen, to another annotated with symptom and treating drug. This problem is relevant in the context of NERO, since different resources have different domains, and are, as such, annotated using different ontologies; and also because resources annotated with NERO may, at some point, be compared with resources annotated with other ontologies.

Semantic similarity [4,7,12] can address this issue. This will improve information retrieval by allowing a user to find resources that are similar to an input query. For instance, a user can be interested in finding all resources related to viral diseases. The system can retrieve resources similar to this query. Alternatively, the user can use as input a resource and find related ones. Semantic similarity across multiple ontologies exploits correspondences between the concepts, but such correspondences can be unavailable; ontology matching techniques can then be used to automatically create them, increasing the accuracy of similarity and, as such, the performance of information retrieval, and the field of ontology matching [5,6].

A second challenge resides in providing a contingency plan for handling cases where few or no annotations exist, which translates to how to generate annotations in an automated fashion for a given resource. Although we expect this situation to become increasingly less frequent as EM gains momentum, it will always remain a necessity to complement manual annotation.

Text mining is already used to handle this by extracting relevant information from the contents of EM resources, and then creating new annotations. This is particularly relevant in poorly annotated resources, since usefulness to the community directly depends on the ability to easily retrieve them. One of the main goals of such techniques is to facilitate the process of annotation to users. By analyzing the content of the files being uploaded, this techniques mine the data to find, for example, diseases or geographical places. These automatic annotations are suggested to the user, who can accept or reject them, improving the quantity and quality of annotations and contributing to a better performance in information retrieval.

When NERO ontologies do not have a sufficient degree of specificity, new concepts can be added using semi-automatic ontology extension, which is capable of automatically suggesting new concepts and relations. New concept suggestions can be derived from text or external ontologies and resources, or more interestingly from the free text annotations made by EM users.

The integration of these techniques into the EM will undoubtedly result in a full-fledged system for semantic web based information retrieval and extraction over its resources.

A final advantage of the EM is in the area of privacy. Since epidemiological data is generally sensitive, the EM manages data access in a fashion where, although the metadata is accessible by all users, access to the data itself can be protected and restricted to authorized users, ensuring that data can remain private, while some of the knowledge about the dataset is still shared, cataloged and found using automatic systems, contributing to its preservation and reuse.

7. CONCLUSIONS

In this paper we present the EM's metadata model, an extension to the Dublin Core coupled with a Network of Epidemiology-Related Ontologies (NERO). It was created with the aim of preserving digital epidemic resources. NERO was compiled based on a set of requirements that ensure, among other qualities, good preservation of the metadata. The EM metadata model supports the annotation of epidemic resources and the application of semantic web technologies over them. The integration of NERO with the EM made the annotation process easier and more complete, giving users a standard set of concepts to choose from. This has already resulted in a corpus of annotated epidemiological resources, over which the information retrieval and extraction system can operate.

By providing better tools to annotate the EM resources, these will be more easily preserved, guaranteeing an easier sharing of epidemic resources for the foreseeable future. In fact, NERO is able to serve all the epidemiology community, since it is not bound to the EM but can subsist on its own. For example, research teams working on developing approaches to identify and quantify modularity in spatially structured and heterogeneous meta-populations and contact networks can also benefit from using NERO, both as an annotation standard and as a way to search for other resources. The geospatial information that NERO encodes can be of great interest here. The collection of validated data through ICT applications can also benefit from NERO: semantically annotating these data is a major step in its analysis, and NERO can serve as the source of concepts for this annotation. The establishment of this network of ontologies contributes, therefore, to an improvement for all the community, particularly on the topics of preservation, sharing and reusing epidemiological data.

8. ACKNOWLEDGMENTS

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 FCT through the financial support of the SOMER project (PTDC/EIA-EIA/119119/2010) and the PhD grant SFRH/BD/69345/2010, and the PIDDAC Program funds for INESCID (Pest-OE/EEI/LA0021/2013) and LASIGE multi annual support.

9. REFERENCES

[1] T. Berners-Lee and J. Hendler. Publishing on the semantic web. *Nature*, 410(6832):1023–1024, 2001.

[2] C. Bizer, T. Heath, and T. Berners-Lee. Linked data-the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22, 2009.

[3] E. Conway, D. Giaretta, S. Lambert, and B. Matthews. Curating scientific research data for the long term: a preservation analysis method in context. *Work*, 6(2):38–52, 2011.

[4] F. M. Couto, M. J. Silva, et al. Disjunctive shared information between ontology concepts: application to Gene Ontology. *J Biomed Semantics*, 2(5), 2011.

[5] D. Faria, C. Pesquita, E. Santos, F. M. Couto, C. Stroe, and I. F. Cruz. Testing the AgreementMaker System in the Anatomy Task of OAEI 2012. *arXiv preprint arXiv:1212.1625*, 2012.

[6] J. D. Ferreira, D. S. Batista, F. M. Couto, and M. J. Silva. The Geo-Net-PT/Yahoo! GeoPlanet (TM) concordance. *Technical Report. TR 10-05, Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa (doi:10455/6677)*, 2010.

[7] J. D. Ferreira and F. M. Couto. Semantic similarity for automatic classification of chemical compounds. *PLoS computational biology*, 6(9):e1000937, 2010.

[8] J. D. Ferreira, C. Pesquita, F. M. Couto, and M. J. Silva. Bringing epidemiology into the Semantic Web. In *Proceedings of the International Conference on Biomedical Ontologies*, 2012.

[9] J. Gray. Jim Gray on eScience: a transformed scientific method. *The fourth paradigm: Data-intensive scientific discovery*, 2009.

[10] C. Jonquet, N. Shah, C. Youn, C. Callendar, M. Storey, and M. Musen. Ncbo annotator: semantic annotation of biomedical data. In *IntâĂŹl Sem Web Conf (ISWC)*, 2009.

[11] L. Lyon, A. Ball, M. Duke, and M. Day. Developing a Community Capability Model Framework for data-intensive research. In *iPres 2012-9th International Conference on Preservation of Digital Objects*, pages 9–16, 2012.

[12] C. Pesquita, D. Faria, H. Bastos, A. Ferreira, A. Falcao, and F. Couto. Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC bioinformatics*, 9(Suppl 5):S4, 2008.

[13] M. J. Silva, F. Silva, L. F. Lopes, and F. M. Couto. Building a digital library for epidemic modelling. In *Proceedings of ICDL*, pages 23–27, 2010.

[14] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255, 2007.