Semi-automated Annotation of Epidemiological Resources

Catia Pesquita[1], João D. Ferreira[1], Francisco M. Couto[1], Mário J. Silva[2]
[1]LASIGE, University of Lisbon
[2]INESC-ID, University of Lisbon

Abstract

Epidemiology is by nature a data-intensive research field, combining results from empirical, analytical and simulation studies. These would benefit from the new scientific methodology designated as the *fourth paradigm*, which addresses the challenges raised from our need to validate, analyze, visualize, store, and curate the large amounts of generated data [1]. In this context, we developed the Epidemic Marketplace (EM) [2] (available at www.epimarketplace.net) , a platform that enables the sharing of resources and knowledge within the Epidemiology community with a strong focus on the semantic annotation of resources.

To support the browsing and management of the resources stored in the EM, each one is described with a set of metadata elements providing biological information (e.g.: disease, symptom, host, vaccine, vector), geographical information, environment and socio-economic conditions, demographics and the associated time frame. To promote a precise and consistent characterisation, these metadata elements are described with well-defined terms from NERO (Network of Epidemiology Related Ontologies), a collection of existing ontologies and vocabularies with the purpose of covering the epidemiological domain [3, 4]. On uploading their resources to the EM, users can provide an accurate semantic annotation based on the metadata and NERO ontologies. However, this process can be time consuming, since a single epidemiological resource can refer to several diseases, symptoms, locations, etc.

In order to require minimal human intervention, we will develop a semi-automated annotation module for the EM, which automatically identifies terms of NERO in a given text-based resource and suggests them as a default characterization of the resource. This module will be based in the recognition of epidemiologically relevant terms in the text and their resolution, i.e., mapping, to NERO terms. Our proposal is to adapt the machine learning approach and semantic similarity techniques described by Grego et al. [5] to perform the recognition and resolution tasks. The machine learning approach requires a training corpus which we will create from the over 100 fully annotated resources already available in the EM. Semantic similarity techniques are normally used with single domain ontologies, but the multidisciplinary nature of NERO will require novel techniques which we intend to adapt from previous work [6].

The integration of this module in the EM will be able to automatically generate candidate semantic annotations to be later validated by the users. This will cut the time and effort needed to provide a complete semantic annotation for research papers and other text-based epidemiological resources, effectively encouraging users to contribute with more resources and provide richer annotations.

References

[1] Hey, A. J. (2009). The fourth paradigm: data-intensive scientific discovery.

[2] Couto, F. M., Ferreira, J. D., Zamite, J., Santos, C., Posse, T., Graça, P., Domingos, D., & Silva, M. J. (2012). The Epidemic Marketplace Platform: towards semantic characterization of epidemiological resources using biomedical ontologies. In Proceedings of the International Conference on Biomedical Ontologies.

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

[4] Ferreira, J. D., Paolotti, D., Couto, F. M., & Silva, M. J. (2012). On the usefulness of ontologies in epidemiology research and practice. Journal of Epidemiology and Community Health.

[5] Grego, T., Pinto, F. R., & Couto, F. M. (2012). Identifying Chemical Entities based on ChEBI. In Proceedings of the International Conference on Biomedical Ontologies.

[6] Ferreira, J. D., & Couto, F. M. (2011). Generic semantic relatedness measure for biomedical ontologies In Proceedings of the International Conference on Biomedical Ontologies.