Making Online Datasets More Searchable and Accessible: The CEDAR project

Mark A. Musen
Stanford Center for Biomedical Informatics Research
1265 Welch Road, Room X-215, Stanford, California 94305-5479, USA
musen@stanford.edu

Abstract

Scientists increasingly are archiving their data in online repositories to promote open science and data reuse. The ability to find and access datasets that are stored in these repositories depends on the quality of the associated metadata. There is a growing set of community - developed standards for defining such metadata often in the form of metadata templates. The practical difficulties of working with these templates are tremendous, however. The Center for Expanded Data Annotation and Retrieval (CEDAR) is developing technologies to assist in the management of biomedical metadata. By discovering patterns in existing metadata and by linking templates to biomedical ontologies, CEDAR is assisting the authoring of new, high-quality metadata. The availability of comprehensive and expressive metadata will facilitate data discovery, interoperability, and reuse.

1 Introduction

The past few years have seen an increasing call for "open science," where investigators make their data available for public access and reuse (Nosek, B.A., et al., 2015). There are obvious opportunities to make new discoveries by examining, integrating, and analyzing data provided by other scientists. Funding organizations and journal editors are increasingly insisting that investigators place their experimental data in public repositories for the benefit of the scientific community. The problem, however, is that submitting data to a public repository can be an onerous task that most investigators would like to avoid. Online datasets need to be supplemented by metadata data about the data that describe the subjects of the experiment, the conditions under which the data were collected, and the major steps that the investigators followed to perform their study. Good metadata are needed for other scientists to be able to search for relevant datasets, to make sense of the data, and to know how to reanalyze the data. The problem is that most datasets are annotated with very poor metadata (Gonalves, R.S., et al., 2017). Metadata authors are burdened by cumbersome requirements, they receive too little guidance, and the result is that metadata are often riddled with typographical errors and they often fail to incorporate standard ontological terms when required. There is a clear need for methods to make it easier for scientist to author high-quality metadata and to archive their datasets in a manner that will assure that the data will be findable, accessible, interpretable, and reusable (FAIR (Wilkinson, M.D., et al., 2016)). We believe that the fundamental challenge of the open-science movement is effective annotation of datasets with metadata that are complete and comprehensive. to use. CEDAR is committed to the development of tools that make it easy for scientists to create high-quality metadata (Musen, M.A., et al., 2015).

2 The CEDAR Workbench

CEDAR is building a suite of tools, known as the CEDAR Workbench, that form a pipeline for authoring experimental metadata (O'Connor, M.J., et al., 2016). We are working in the area of biomedical science, where there is already a trend for different scientific communities to specify standardized templates that capture the minimal requirements for metadata related to different classes of experiments (Taylor, C.F., Field, D., and Sansone, S.A., 2008).

Metadata Template Repository: We have developed a standardized representation of metadata templates together with Web-based services to store, search, and share these templates. Templates created using CEDAR technology are stored in our openly accessible community repository. Researchers access the repository to search for appropriate templates to annotate their studies. Web-based interfaces and REST APIs enable access to all metadata templates, as well as to all the metadata collected using those templates (O'Connor, M.J., et al., 2016).

Metadata Template Creator and Template Editor: Two highly interactive Web-based tools simplify the process of authoring metadata templates. The Template Creator allows users to create, search, and author metadata templates. Using interactive look-up services linked to the NCBO BioPortal, template authors can find terms in ontologies to annotate their templates and to restrict the values of template fields. The Template Creator automatically produces a user interface specification as it builds a template. The Metadata Editor uses this specification to generate a forms-based acquisition interface for acquiring individual metadata components.

Intelligent Authoring: To ease the burden of authoring high quality metadata, a recommender framework learns associations between data elements and suggests to the user context-sensitive metadata values (Martínez-Romero, M., et al., 2017). The system can recommend possible values for metadata elements during the submission process as each blank is selected and the user begins to type. The template editor also sorts possible selections in drop-down windows so that the terns that occur in the database with the greatest frequency in the context of the other entries that have already been made into the template appear at the top of the drop-down list. The goal is to make it as simple as possible for metadata authors to fill in the templates, using as many entries from standard ontologies as they can, and to do allow the authors to do so as quickly and as accurately as possible.

3 Deployment and Evaluation

The CEDAR team includes several community - based groups who are helping to develop and evaluate our current system. These collaborators include (1) the BioSharing initiative, which catalogs metadata standards for describing biomedical experiments (McQuilton, P., et al., 2016), (2) ImmPort, a data warehouse of immunology-related datasets (Bhattacharya, et al., 2014), and (3) the Human Immunology Project Consortium Standards Working Group, which designs new metadata templates and channels experimental datasets to the ImmPort repository. We successfully have represented metadata from several hundred studies provided by these groups within the CEDAR workbench. We also are working with the LINCS project to develop a more robust metadata management pipeline that supports the authoring of metadata for a wide range of studies (Vempati, U.D., et al., 2014). Collaborations with other scientific consortia are in the planning stage, with the long-term goal of making all scientific data easier to find, access, integrate, and reuse.

Acknowledgments

CEDAR is supported by NIAID grant U54 AI117925 through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative. CEDAR includes participation from groups at Stanford University, Yale University, the University of Oxford, and Northrop Grumman corporation. Martin J. O'Connor, Marcos Martínez-Romero, Attila L. Egyedi, Debra Willrett, and John Graybeal have contributed to the development of the CEDAR Workbench. Additional information about CEDAR is available from the Center's Web site: `http://metadatacenter.org`.

References

Bhattacharya, S., Andorf, S., Gomes, L., et al. 2014. *Imm-Port: disseminating data to the public for the future of immunology.* Immunologic Research 58(23):234239.

Gonalves, R.S., OConnor, M.J., Martnez-Romero, M., et al. 2017. *Metadata in the BioSample online repository are impaired by numerous anomalies. Procedings of SemSci: Enabling Open Semantic Science.* International Semantic Web Conference. Vienna, Austria.

Martínez-Romero, M., OConnor, M.J., Shankar, R., et al. 2017. *Fast and accurate metadata authoring using ontology-based recommendations.* Proceedings of the American Medical Informatics Association Annual Symposium. Washington, DC.

McQuilton, P., Gonzalez-Beltran, A., Rocca-Serra, P., et al. 2016. *Biosharing: curated and crowd-sourced metadata standards, databases, and data policies in the life sciences.* Database 2016, doi: 10.1093/database/baw075.

Musen, M.A., Bean, C.A., Cheung, K.-H., et al. 2015. *The Center for Expanded Data Annotation and Re-*

trieval. Journal of the American Medical Informatics Association 22(6):11481152.

Nosek, B.A., Alter, G., Banks, G.C., et al. 2015. *Promoting an open research culture.* Science 348(6242):14221424.

O'Connor, M.J., Martnez-Romero, M., Egyedi, A.L., et al. 2016. *An open repository model for acquiring knowledge about scientific experiments.* Proceedings of the 20th International Conference on Knowledge Engineering and Knowledge Management. Bologna, Italy.

Taylor, C.F., Field, D., and Sansone, S.A. 2008. *Promoting coherent minimum reporting guidelines for biological and biomedical investigaitons: the MIBBI project.* Nature Biotechnology 26:889896.

Vempati, U.D., Chung, C., Mader, C., et al. 2014. *Specifications to describe, model, and integrate complex and diverse high-throughput screening data from the Library of Integrated Network-based Cellular Signatures (LINCS).* Journal of Bio-molecular Screening 19(5):803816.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., et al. 2016. *The FAIR guiding principles for scientific data management and stewardship.* Nature Scientific Data 3:160018.