CEDAR: Semantic Web Technology to Support Open Science

Mark A. Musen*
Stanford University
Stanford, California, USA
musen@stanford.edu

Susanna-Assunta Sansone
University of Oxford
Oxford, UK
sa.sansone@gmail.com

Kei-Hoi Cheung
Yale University
New Haven, Connecticut, USA
kei.cheung@yale.edu

Steven H. Kleinstein
Yale University
New Haven, Connecticut, USA
steven.kleinstein@yale.edu

Morgan Crafts
Northrop Grumman Corporation
Falls Church, Virginia, USA
morgan.crafts@ngc.com

Stephan C. Schürer
University of Miami
Miami, Florida, USA
sschurer@miami.edu

John Graybeal
Stanford University
Stanford, California, USA
jgraybeal@stanford.edu

ABSTRACT

There is an expectation that scientists will archive their experimental data online in public repositories to enable other investigators to verify their work and to re-explore their data in search of new discoveries. When left to their own devices, however, scientists do a poor job creating the metadata that describe their datasets. A lack of standardization makes it difficult for other investigators to find relevant datasets and to perform secondary analyses. The Center for Expanded Data Annotation and Retrieval (CEDAR) was founded with the goal of enhancing the authoring of experimental metadata to make online datasets more useful to the scientific community. CEDAR technology includes Web-based methods for creating and managing libraries of templates for representing metadata. CEDAR's templates interoperate with a repository of scientific ontologies to standardize the way in which the templates may be filled out. Collaborations with several major research projects are allowing us to explore how CEDAR may ease access to scientific data sets stored in public repositories.

KEYWORDS

Metadata; scientific data; semantic technology

ACM Reference Format:
Mark A. Musen, Susanna-Assunta Sansone, Kei-Hoi Cheung, Steven H. Kleinstein, Morgan Crafts, Stephan C. Schürer, and John Graybeal. 2018. CEDAR: Semantic Web Technology to Support Open Science. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3184558.3186200

*Dr. Mark A. Musen is the principal investigator for this project and the presenter.

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186200

1 INTRODUCTION

The past few years have seen an increasing demand for open science, where investigators make their data available for public access and reuse. 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 [1]. 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 scientists to author high-quality metadata and to archive their datasets in a manner that will assure that the data will be findable, accessible, interoperable, and reusable [7].

We believe that the fundamental challenge of the open-science movement is effective annotation of datasets with metadata that are complete and comprehensive. CEDAR is committed to the development of tools that make it easy for scientists to create high-quality metadata [3].

2 THE CEDAR WORKBENCH

CEDAR is building an open-source suite of tools, known as the CEDAR Workbench, that form a pipeline for authoring experimental metadata [4]. 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.

Metadata Template Repository: We have developed a standardized representation of metadata templates, together with Web services to store, search, and share these templates. Templates created using CEDAR technology are stored in our openly accessible repository. Researchers use the repository to search for appropriate templates to annotate their studies. Web-based interfaces and REST APIs enable access to the metadata templates, as well as to the corresponding metadata collected using those templates [4].

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 ontology respository, template authors can find terms in ontologies 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 metadata elements and suggests to the user context-sensitive metadata values [2]. The system can recommend possible values for metadata elements during the submission process as each field is selected and the user begins to type. The template editor also sorts possible selections in drop-down windows so that the terms 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 allow the authors to do so as quickly and as accurately as possible.

3 EVALUATION AND DISSEMINATION

The CEDAR team includes several groups who are helping to develop and evaluate our current system. For example, the Library of Network-Based Cellular Signatures (LINCS) is a consortium of six Data and Signature Generation Centers (DSGCs) and a Data Coordination and Integration Center (DCIC) in the United States. LINCS scientists use a variety of experimental techniques to study the consequences of disruption of biological pathways. LINCS has developed extensive metadata standards [6], and the consortium is automating the process of data and metadata submission to the DCIC using CEDAR. CEDAR and the LINCS DCIC are collaborating to enable all six of the LINCS DSGCs to use the CEDAR Workbench for metadata management, standardization, and submission.

The Adaptive Immune Receptor Repertoire (AIRR) community began in 2014 as an effort to bring together an international group of investigators utilizing high-throughput methods to study the immune response. The AIRR community has produced its own metadata standard for depositing data across four different data repositories at the U.S. National Center for Biotechnology Information (NCBI): BioProject, BioSample, SRA, and GenBank[5]. Unfortunately, none of the more than 30 databases managed by NCBI provides infrastructure for the curation of standardized metadata. AIRR is working with NCBI and with CEDAR in a project to allow users to submit controlled metadata to the NCBI. The AIRR community is beginning to use CEDAR technology to generate and upload its metadata. The NCBI is monitoring the AIRR activities to evaluate the potential use of CEDAR technology for the submission of metadata to all NCBI data repositories.

Other collaborators are helping us to evaluate the CEDAR Workbench. They include the Protein Data Commons, under development by the U.S. National Cancer Institute; Stanford University Library, which is experimenting with the CEDAR Workbench as a tool for accessioning parts of its collections; and SimTK, a project that manages enormous amount of data related to the biomechanics of human mobility.

4 CONCLUSION

The benefits of open science depend on the availability of acessible, online datasets that are self-describing, using metadata based on standard frameworks and standard ontologies. CEDAR is a large, international collaboration that is working to develop open-source, Web-based tools and services that enable scientists to author and manage comprehensive metadata to annotate such datasets. Deployment of the CEDAR Workbench in several active research settings is enabling us to assess both the usability of our technology and its ability to enhance the metadata that allow experimental datasets to be identified, retrieved, and re-used by other investigators.

ACKNOWLEDGMENTS

CEDAR is supported by NIAID grant U54 AI117925 through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative in the United States. 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, and Debra Willretr have contributed to the development of the CEDAR Workbench. S. Ahmad Chan Bukhari, Daniel Cooper, Alejandra Gonzalez-Beltran, and Phillipe Rocca-Serra have made valuable contributions to the CEDAR project. Additional information about CEDAR is available from the Center's Web site: http://metadatacenter.org.

REFERENCES

[1] Rafael S Gonçalves, Martin J. O'Connor, Marcos Martínez-Romero, et al. 2017. Metadata in the BioSample online repository are impaired by numerous anomalies. In *Proceedings of the First Workshop on Enabling Open Semantic Science (SemSci)*. 39–46. http://ceur-ws.org/Vol-1931/#paper-06

[2] Marcos Martínez-Romero, Martin J O'Connor, R Shankar, et al. 2017. Fast and accurate metadata authoring using ontology-based recommendations. In *Proceedings of the AMIA Annual Symposium*.

[3] Mark A Musen, Carol A Bean, Kei-Hoi Cheung, et al. 2015. The center for expanded data annotation and retrieval. *Journal of the American Medical Informatics Association* 22, 6 (2015), 1148–1152.

[4] Martin J O'Connor, Marcos Martínez-Romero, Attila L Egyedi, et al. 2016. An open repository model for acquiring knowledge about scientific experiments. In *European Knowledge Acquisition Workshop*. Springer, 762–777.

[5] Florian Rubelt, Christian E Busse, Syed Ahmad Chan Bukhari, et al. 2017. Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. *Nature Immunology* 18, 12 (2017), 1274.

[6] Uma D Vempati, Caty Chung, Chris Mader, et al. 2014. Metadata standard and data exchange 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 Biomolecular Screening* 19, 5 (2014), 803–816.

[7] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3 (2016).