

Using Ontologies to Drive the Creation of High-Quality Metadata in CEDAR

Rafael S. Gonçalves^[0000-0003-1255-0125], Csongor I. Nyulas, Marcos Martínez-Romero, Martin J. O'Connor, John Graybeal, and Mark A. Musen

Stanford Center for Biomedical Informatics Research
Stanford University, Stanford, CA, USA
rafael.goncalves@stanford.edu

Abstract. The Center for Expanded Data Annotation and Retrieval (CEDAR) developed a suite of tools—the CEDAR Workbench—that allows users to build metadata templates using ontologies to annotate template fields and to constrain the options available to metadata authors for specific fields; to fill in those templates with metadata; to upload data and their metadata to online repositories; and to perform searches over the metadata stored in CEDAR’s metadata repository. The CEDAR Workbench is released under a BSD 2-Clause open-source license, and it is freely available at <https://metadatacenter.org>.

Keywords: Metadata, metadata authoring, metadata repository.

1 Introduction

We present the CEDAR [1] software to produce high-quality, structured, standards-based metadata. The software we have developed—the CEDAR Workbench [2]—is a suite of Web-based tools and APIs that offers users the ability to build highly-modular metadata acquisition forms (templates) that can be annotated with ontology terms, and whose fields can be constrained using terms or branches of terms from ontologies. Rather than having a single monolithic template, CEDAR allows users to recursively construct templates from existing, more granular templates. CEDAR template designers can share templates with individuals or groups—the metadata authors, who fill in the metadata templates, validate field entries, and submit the metadata to online repositories. The metadata produced using CEDAR templates are, by design, adherent to the FAIR data principles [3]. Our goal is ultimately to provide scientists with a robust, end-to-end software solution to author and to manage high-quality FAIR metadata about scientific experiments.

2 The CEDAR Workbench

The CEDAR Workbench is an open-source Web-based platform for the acquisition, storage, search, and reuse of metadata templates and metadata instances. At the core

of the CEDAR technology lies a lightweight, standards-based model [4] designed to provide a common format for describing templates and metadata. All CEDAR resources are represented as JSON-LD documents that conform to our model, which is specified by a JSON Schema. These resources can be viewed and retrieved as RDF documents. Fig. 1 shows an overview of CEDAR.

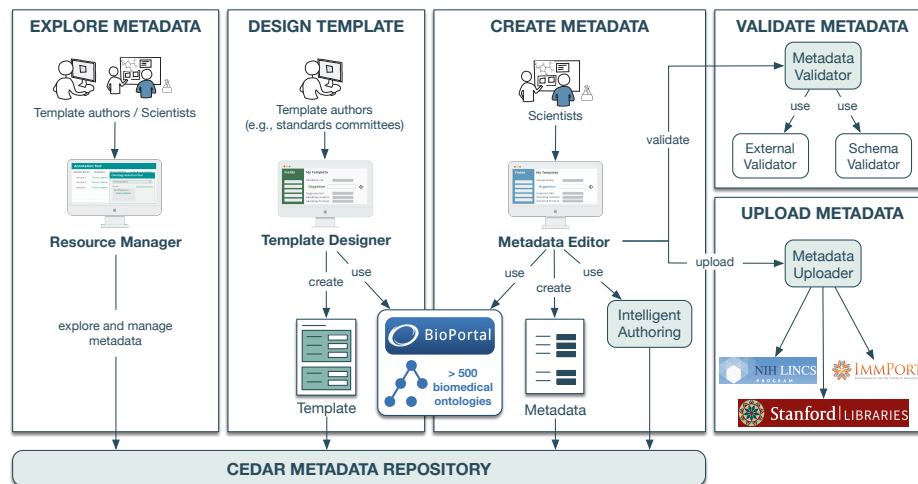


Figure 1. Overview of CEDAR. Users can manage and search for resources; design metadata templates using ontologies from BioPortal; create metadata with support from intelligent authoring features; validate metadata; and upload metadata to external repositories.

The following are the main components of the CEDAR Workbench software.

Resource Manager. Template authors and scientists who use the CEDAR Workbench are initially presented with the Resource Manager tool. The Resource Manager allows users to create and store resources in the CEDAR Metadata Repository; to organize templates and metadata into folders; and to search for these resources. From the Resource Manager, users can define groups composed of their team members for purposes of collaboration. CEDAR users can share resources (with read or write permissions) among users, among groups, or with the general community.

Template Designer. Template authors can build metadata templates using the Template Designer. In the Template Designer, users piece together fields of various types (e.g., text, checkbox, and multiple choice) to form templates. Possible field values can be constrained to terms from ontologies using an interactive look-up service linked to NCBO's BioPortal [5]. With the BioPortal lookup service (Fig. 2), users can interactively create new ontology terms (which can be mapped to terms in other ontologies) and value sets at template design-time for their annotation purposes. The metadata templates and their fields can be annotated using properties from ontologies.

Metadata Editor. Scientists generate metadata instances by filling in metadata templates using the Metadata Editor. This tool builds a metadata-acquisition form interface from template specifications built in the Template Designer. We implemented a computer-assisted value recommender [6] in the Metadata Editor that provides

context-sensitive suggestions for field values during metadata submission. The value recommender learns associations between field values in previous metadata entries using rule mining, and ranks their applicability to specific fields. The goal of the value recommender is to ease the burden of authoring high-quality metadata. Metadata generated through CEDAR templates can be submitted to external repositories, such as the NCBI BioSample [7] and SRA [8] repositories, or the ImmPort repository for immunology-related datasets [9].

Find terms in BioPortal or [Create New Terms](#) to constrain the values of the 'Organism' field [Start Over](#)

Search in BioPortal
Organism ⚙️ 🔍

497 results for the query 'Organism'. Click on a term below to select it

TERM	DEFINITION	TYPE	SOURCE	ID
Organism	-	Class	SNMI	T001
organism	A material entity that is an individual living system, such as animal, plant, bacteria or virus, that is capable of...	Class	OBI	OBI_0100026 Hide Details
organism	A material entity that is an individual living system, such as animal, plant, bacteria or virus, that is capable of...	Class	CLO	OBI_0100026
Organism	-	Class	NIFSTD	birnlex_2
organism	A material entity that is an individual living system, such as animal, plant, bacteria or	Class	NIFSTD	OBI_0100026

OBI classes:

- Organism
 - Eukaryota
 - Viruses
 - Genetically Modified Organism
 - Bacteria
 - Infectious Agent
 - Archaea
 - Host Of Immune Response
 - Chimera
 - Selectively Maintained Organism
 - Immaterial Entity
 - Generically Dependent Continuant
 - Specifically Dependent Continuant

TERM DETAILS | **ONTOLOGY DETAILS**

Name organism

Id http://purl.obolibrary.org/obo/OBI_0100026

Definition A material entity that is an individual living system, such as animal, plant, bacteria or virus, that is capable of replicating or reproducing, growth and maintenance in the right environment. An organism may be unicellular or made up, like humans, of many billions of cells divided into specialized tissues and organs.

TERM | **BRANCH** | ONTOLOGY

Figure 2. The BioPortal lookup service allows users to search for ontology terms in BioPortal, browse ontologies, and select terms or branches of terms to constrain the possible options available when filling-in specific fields in CEDAR metadata templates. It also allows users to create new terms, and to create value sets made up of existing terms in BioPortal.

The CEDAR Workbench can be used through the Web-based components described above, or using the CEDAR API—a collection of REST-based services that provide comprehensive access to the CEDAR ecosystem. The API allows creating, reading, updating, and deleting CEDAR resources programmatically. With this API, users can also export templates or metadata to other repositories or applications. All our software is distributed and versioned on GitHub, at <https://github.com/metadatascenter>.

3 Summary

The CEDAR Workbench provides a comprehensive solution for authoring, validating, searching, and (re)using metadata. The goal behind CEDAR is to significantly improve the way scientists work with metadata, and the quality and interoperability of the metadata that they create. We meet this goal by equipping the community with a collaborative platform to build standards-based metadata templates that use ontologies as sources for standard terms, and to author and submit high-quality metadata to online repositories. CEDAR's metadata repository gives scientists a means to search for and to use metadata templates developed by the community, and to build new ones from scratch or based on existing templates. CEDAR allows its users to submit their metadata to external repositories, such as NCBI databases. We are working to allow our users to submit metadata to an increasing number of external repositories.

Acknowledgements

CEDAR is supported by NIAID grant U54 AI117925 through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative. The NCBO BioPortal has been supported by the NIH Common Fund under grant U54HG004028.

References

- [1] M. A. Musen *et al.*, "The center for expanded data annotation and retrieval," *J. Am. Med. Informatics Assoc.*, vol. 22, no. 6, pp. 1148–52, Jun. 2015.
- [2] R. S. Gonçalves *et al.*, "The CEDAR Workbench: An Ontology-Assisted Environment for Authoring Metadata that Describe Scientific Experiments," in *Proc. of International Semantic Web Conference (ISWC)*, 2017.
- [3] M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship.," *Sci. data*, vol. 3, p. 160018, 2016.
- [4] M. J. O'Connor, M. Martínez-Romero, A. L. Egyedi, D. Willrett, J. Graybeal, and M. A. Musen, "An Open Repository Model for Acquiring Knowledge About Scientific Experiments," in *Proc. of International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, 2016.
- [5] N. F. Noy *et al.*, "BioPortal: ontologies and integrated data resources at the click of a mouse," *Nucleic Acids Res.*, vol. 37, pp. W170–W173, 2009.
- [6] M. Martínez-Romero *et al.*, "Fast and Accurate Metadata Authoring Using Ontology-Based Recommendations," in *Proc. of AMIA Annual Symposium*, 2017.
- [7] T. Barrett *et al.*, "BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata," *Nucleic Acids Res.*, vol. 40, pp. D57–D63, 2012.
- [8] R. Leinonen, H. Sugawara, and M. Shumway, "The Sequence Read Archive," *Nucleic Acids Res.*, vol. 39, no. Database, pp. D19–D21, Jan. 2011.
- [9] S. Bhattacharya *et al.*, "ImmPort: disseminating data to the public for the future of immunology.," *Immunol. Res.*, vol. 58, no. 2–3, pp. 234–9, 2014.