Supporting Ontology-Based Standardization of Biomedical Metadata in the CEDAR Workbench

Marcos Martínez-Romero*, Martin J. O'Connor, Michael Dorf, Jennifer Vendetti, Debra Willrett, Attila L. Egyedi, John Graybeal, and Mark A. Musen
Center for Biomedical Informatics Research, Stanford University, 1265 Welch Rd, Stanford, CA 94305, USA

ABSTRACT

The availability of associated descriptive metadata for scientific datasets is important for discovering and reproducing scientific experiments. The use of ontologies has become a key focus for increasing the quality of these metadata. Despite the wide availability of biomedical ontologies, scientists wishing to use these ontologies when developing metadata descriptions face a number of practical difficulties. A core difficulty is the lack of tools for developing ontology-linked metadata specifications that can be published and shared. Additional difficulties include the lack of support for defining new terms in cases when no existing terms are found and for creating custom term collections to meet domain-specific needs. To address these problems, we developed tools that allow scientists to find terms in ontologies for annotating their data and to dynamically create new terms and value sets. This work has been incorporated into a Web-based platform called the CEDAR Workbench. The resulting integrated environment presents a set of highly interactive interfaces for creating and publishing ontology-rich metadata specifications.

1 INTRODUCTION

In biomedicine, high-quality, standardized metadata are crucial for facilitating the discovery of scientific datasets and reproducibility of the corresponding experiments. In the last few years, the biomedical community has driven the development of metadata standards and guidelines for a variety of experiment types. Scientists use these specifications to inform their annotation of experimental results (Tenenbaum, Sansone, & Haendel, 2014). One of the earliest examples is the MIAME standard (Brazma et al., 2001), which is used to describe metadata about microarray experiments. These standards and guidelines underpin metadata submissions to many public metadata repositories (Edgar, Domrachev, & Lash, 2002). The BioSharing resource (McQuilton et al., 2016) catalogs hundreds of these standardization efforts.

Despite the growing use of standards for defining metadata and the wide availability of biomedical ontologies, metadata submitted to public repositories rarely use standard terms (Bui & Park, 2006). As a result, finding or reusing the metadata is a challenge and understanding the underlying experiments can be extremely hard, often requiring significant post-processing of metadata to extract useful content.

A key problem is that scientists face considerable practical barriers when attempting to link their metadata to ontology terms. Submission mechanisms for biomedical repositories are typically based on spreadsheets, with a variety of *ad hoc* formats that rarely support inclusion of ontology-based annotations. Even in cases where such annotations can be entered, scientists have no easy way to find and use terms from ontologies to include in their metadata submissions. Other difficulties include poor support for on-the-fly term creation when the necessary terms are not found and for creating custom lists of terms to meet domain-specific needs.

A variety of tools have been developed to address the challenge of metadata quality. Foremost among these are the ISA Tools (Rocca-Serra et al., 2010), which allow curators to create spreadsheet-based submissions for metadata repositories. LinkedISA provides a means to interoperate with Linked Open Data, effectively adding controlled term linkage to templates (González-Beltrán, Maguire, Sansone, & Rocca-Serra, 2014). A similar spreadsheet-based tool called RightField (Wolstencroft et al., 2011) provides a mechanism for embedding ontology annotation capabilities in Excel or Open Office spreadsheets using ontologies from the BioPortal repository (Noy et al., 2009). Annotare (Shankar et al., 2010), which is used to submit experimental data to the ArrayExpress metadata repository (Parkinson et al., 2005), also supports ontology-based suggestions. These tools address specific issues of metadata quality but they do not provide an integrated environment that can support the entire metadata specification and submission process for widely used biomedical repositories.

The Center for Expanded Data Annotation and Retrieval (CEDAR)[1] is developing a computational ecosystem to overcome the barriers to creating high-quality metadata in biomedicine (Musen et al., 2015). CEDAR provides a suite of highly sophisticated tools designed to make the authoring of metadata as natural as possible, while also using ontologies to enrich the generated descriptions with standard terms.

In this paper, we describe the main features CEDAR developed to make it possible to easily construct Web-based metadata-acquisition forms, enrich those forms with ontology concepts, and then fill out the forms to create ontology-annotated descriptions of scientific experiments.

[1] https://metadatacenter.org/

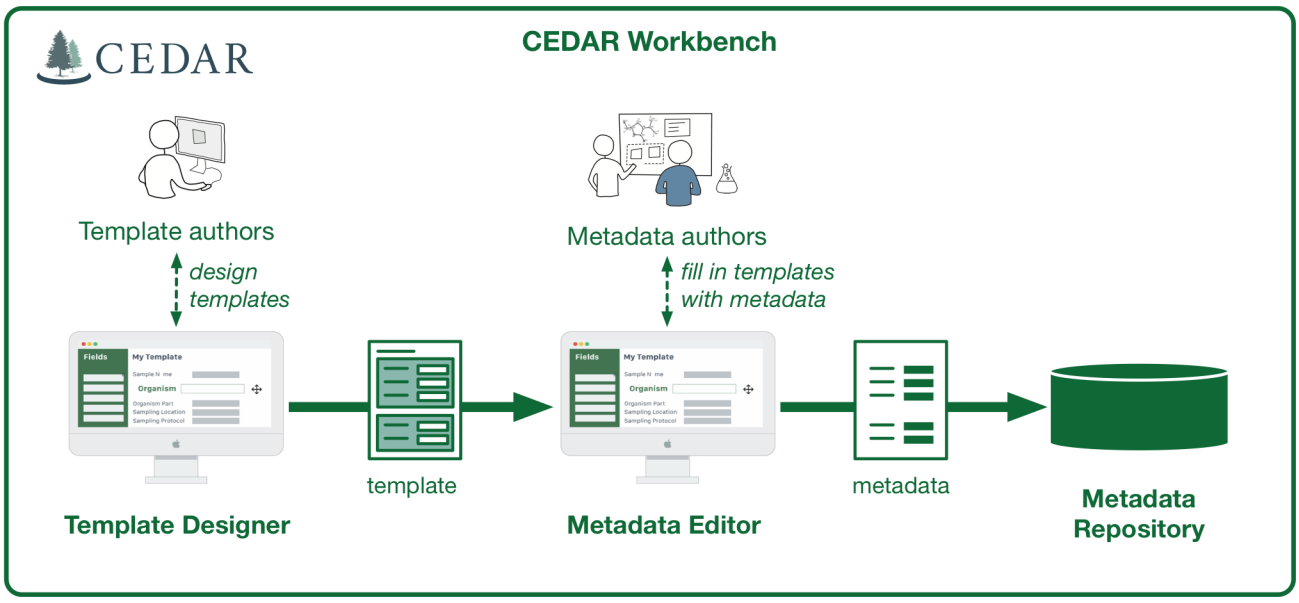

Fig. 1. An overview of CEDAR's metadata authoring workflow. Template authors use the Template Designer tool to create metadata templates. The Metadata Editor uses these templates to generate a graphical interface to acquire metadata from scientists. Acquired metadata are saved in CEDAR's Metadata Repository.

2 BACKGROUND

The CEDAR Workbench[2] is a suite of Web-based tools and REST APIs centered on the use of highly-modular metadata-acquisition forms called *metadata templates* (or simply *templates*). These templates define the data attributes—termed *template fields* or *fields*—needed to describe biomedical experiments. For example, an *experiment* template may have an *organism* field containing the name of the organism being studied by the experiment (e.g., *Homo sapiens*). The templates may specify lists of permissible values for template fields. The central goal when designing a template is to enable the capture of sufficiently precise and complete metadata about experimental data to facilitate data discovery, interpretation, and reuse.

The CEDAR Workbench provides three core components that form a metadata construction pipeline (Fig. 1): (1) a Template Designer, which supports interactive template creation; (2) a Metadata Editor, which allows end-users to fill in templates with metadata; and (3) a Metadata Repository for storing both templates and the metadata created using those templates. The CEDAR Workbench also allows scientists to upload the metadata created to public biomedical repositories.

2.1 Template Designer and Metadata Editor

In the Template Designer, template authors assemble templates from one or more input fields. There are numerous field types available to template authors (e.g., text, paragraph, e-mail, numeric, and date). Users can also define reusable groups of fields, called *elements*. For example, the fields that describe a publication (e.g., *authors*, *title*, *year*, *publication type*, etc.) could be grouped together to form a *publication* element, which can then be reused in multiple templates. After a template is created, the Metadata Editor can be used to automatically generate a forms-based acquisition interface for entering metadata for that template. Scientists entering metadata using the Metadata Editor are prompted in real time with drop-down lists, auto-completion suggestions, and verification hints, significantly reducing their error rate while speeding metadata entry and repair. These prompts are driven by the value constraints specified in templates.

2.2 Metadata Repository

Templates and metadata produced by the Workbench are stored in CEDAR's metadata repository. CEDAR incorporates a standardized model of templates and metadata, together with Web-based services to store, search, and share these resources (O'Connor et al., 2016). This model is based on the JSON Schema and JSON-LD specifications. It allows users to publish their metadata as both JSON-LD and RDF, thus facilitating interoperation with Linked Open Data.

2.3 Support for ontology-based metadata

The CEDAR tools provide mechanisms for structurally describing templates and publishing metadata created using those templates in an open format. To increase the metadata quality further, we offer the ability to enrich these descriptions with controlled terms from ontologies. We extended the Template Designer and Metadata Editor to let users specify semantic content for templates and to easily enter semantically precise terms in their metadata. These extensions, can help to improve metadata adherence to the FAIR data principles (Wilkinson et al., 2016) and interoperability with Linked Open Data.

[2] https://cedar.metadatacenter.net

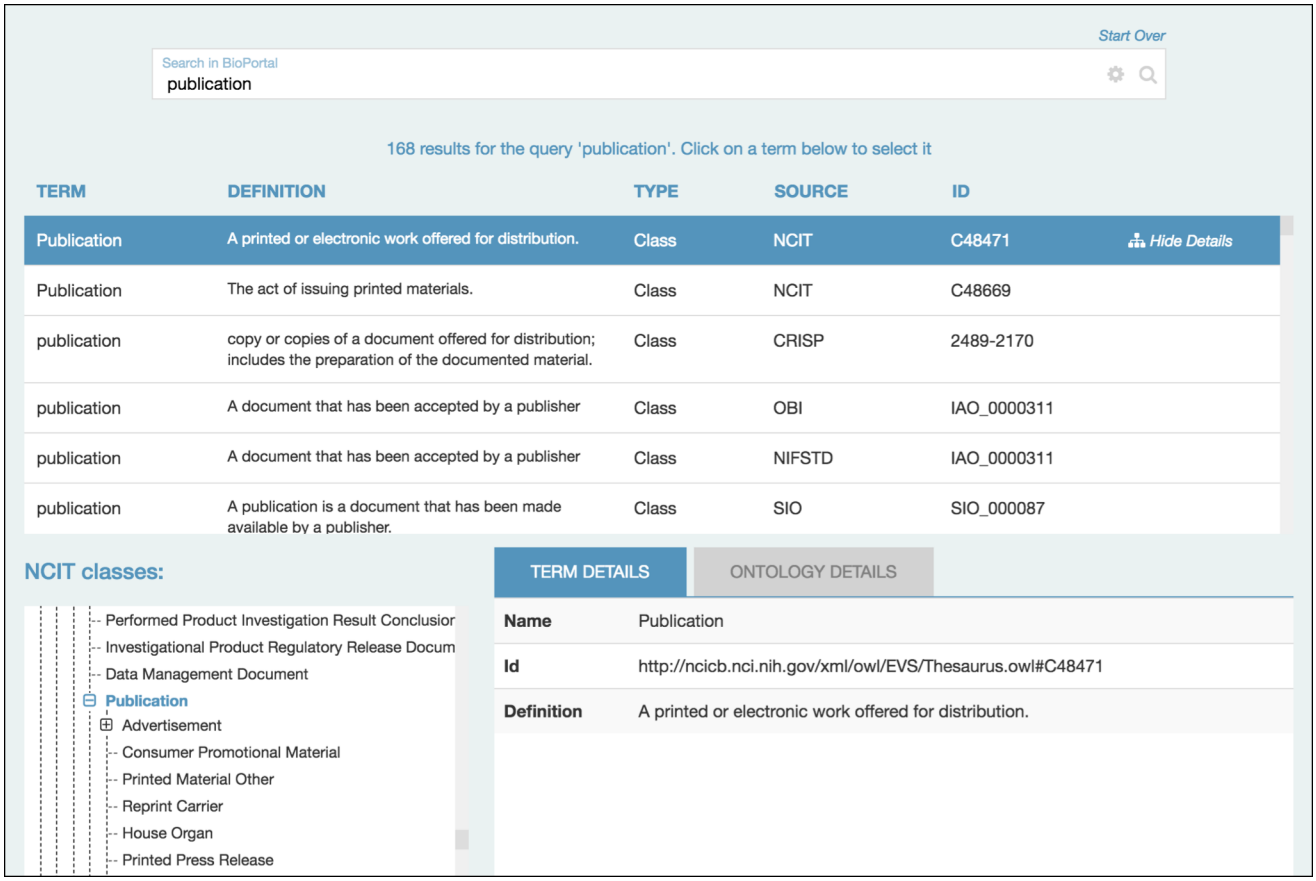

Fig. 2. Screenshot of CEDAR Template Designer's ontology lookup interface. Here, the user entered the search term *publication* and selected the class *Publication* from the National Cancer Institute Thesaurus (NCIT). The location of the selected class in the class tree is presented, as well as class and ontology details.

3 IMPLEMENTATION

We have enhanced the CEDAR Workbench to provide the ability to link ontology terms selected from BioPortal to biomedical metadata. BioPortal, developed by the National Center for Biomedical Ontology (NCBO) (Musen et al., 2012), is a popular platform for hosting and sharing biomedical ontologies. It provides access to more than 550 ontologies, and contains over 8 million classes and 64,000 properties. The BioPortal API provides a rich set of operations to access and use ontologies. We extended this API to provide the fine grained, highly interactive class and property lookup features needed by CEDAR's term search and selection features. To facilitate general use of these features, we encapsulated BioPortal's API as a CEDAR service and made it available as a public REST endpoint.[3] We now describe these extensions.

3.1 Class and Property Search

CEDAR allows template authors to search for ontology terms to annotate their templates, that is, to add type and property assertions to template elements and fields using ontology classes and properties. Classes and object, data, and annotation properties for performing these annotations can be selected from terms supplied by BioPortal. Fig. 2 shows a screenshot of the ontology lookup user interface of the CEDAR Workbench. In the example shown, the template author entered the search term *publication* and then selected the *Publication* class from the National Cancer Institute Thesaurus (NCIT). The interface shows detailed information both for the selected class and the associated ontology, as well as for the position of the class in the class tree of NCIT.

3.2 Value Set creation

A value set is a list of possible values for a specific purpose. In the CEDAR Workbench, value sets are a useful mechanism to define pick lists of permissible values for template fields. CEDAR works in conjunction with BioPortal to allow template authors to dynamically create value sets containing the terms in these pick lists. Value sets can contain

[3] The REST endpoints that provide ontology-based services to the CEDAR Workbench are documented at https://terminology.metadatacenter.net/api.

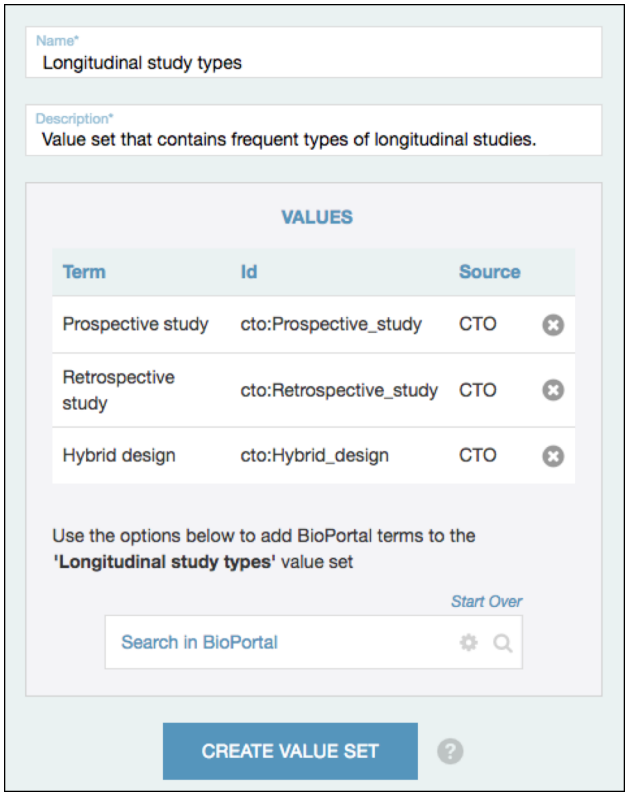

Fig. 3. Screenshot showing an example of value set creation. The user is building a *Longitudinal study types* value set with terms from the Clinical Trials Ontology (shown as CTO).

classes from any combination of BioPortal ontologies. Upon creation, a value set is immediately assigned a unique, provisional IRI. The CEDAR Workbench supports the creation, retrieval, update, and deletion of these value sets.

For example, suppose that the template author wishes to constrain the values of a *Study type* field to three specific types of longitudinal studies (*prospective study*, *retrospective study*, and *hybrid study*). The Clinical Trials Ontology (CTO) is a good source of these types since it contains 375 study type classes (represented as descendants of the *Study type* class). Instead of selecting all these types, the template author can create a value set containing only the desired types. Fig. 3 shows a screenshot of value set creation features for this example presented in the Template Designer. Here, the user creates a value set named *Longitudinal study types* with three terms selected from CTO.

3.3 Class creation

Despite the vast number of classes and properties available in biomedical repositories, ontologies often do not contain the exact term a user requires. To address this problem, CEDAR allows users to dynamically define new classes and immediately to use them. When generating a new class, users can optionally link it to one or several existing classes by means of the RDFS *subclassOf* relationship and SKOS relationships (*closeMatch*, *exactMatch*, *broadMatch*, *narrowMatch*, *relatedMatch*). Upon creation, a class is immediately assigned a unique, provisional IRI.

For example, suppose that a user needs to use the anatomical term *adductor dorsalis*. This term is not available in any BioPortal ontology, though the *adductor muscle* class in the UBERON ontology is a close conceptual match. In this case, the user decides to create an *adductor dorsalis* class via the CEDAR Workbench and indicate that the new term is a subclass of the *adductor muscle* UBERON class. Fig. 4 shows the class creation interface for this example. The *adductor dorsalis* class is stored in BioPortal as a CEDAR provisional class and is immediately available to all CEDAR users. Eventually, maintainers of UBERON may decide to incorporate the *adductor dorsalis* class to the ontology or may decide to reject it. If the class is added to UBERON, the permanent identifier for the class will be stored as part of the information of the provisional class. If *adductor dorsalis* is not included in the next version of the ontology, the *subclassOf* link will removed, but the class will still be valid in CEDAR.

3.4 Value Constraints

With the above functionality, the system can limit the possible values of a template field to a predefined sets of ontology terms or value sets. Some template authors may need to define value constraints that go beyond predefined term

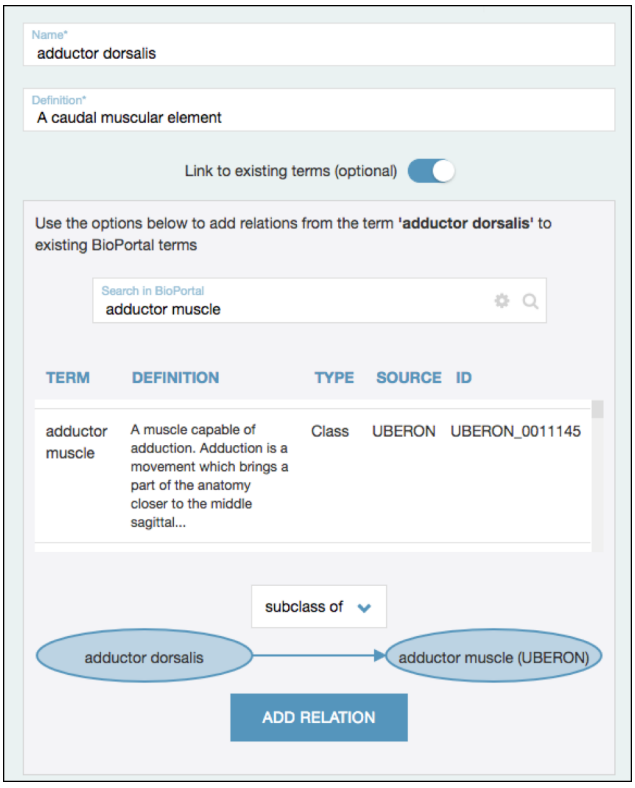

Fig. 4. Example of class creation in the CEDAR Workbench. The user creates the *adductor dorsalis* class and links it to the *adductor muscle* class in the UBERON ontology via the subclassOf relation.

lists. For example, a user may wish to constrain the values of a *disorder* template field to all subclasses in three specific branches of the DOID ontology rooted at the terms *cognitive disorder*, *sleep disorder*, and *dissociative disorder*.

To deal with use cases such as this one, the system effectively allows template designers to constrain field values to any combination of (1) all classes in an ontology branch, (2) all classes from a specific ontology, (3) new or existing classes, and (4) new or existing value sets. Multiple constraint types can be specified for the same field.

Users populating templates using the Metadata Editor are presented in real time with a list of choices driven by these value constraints. Fig. 5 shows an example of choices presented for a *disorder* field that has had its values constrained to come from the three DOID ontology disease branches described in the earlier example. All terms from these three branches are combined in real time and presented as a single list.

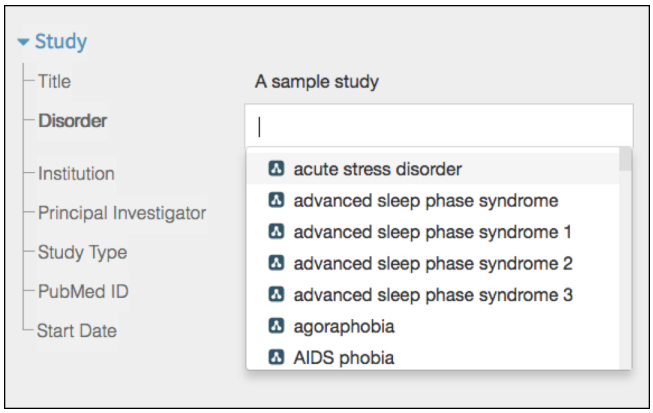

Fig. 5. Screenshot of the Metadata Editor that shows the possible values of a *Disorder* field in a *Study* template. This field has been constrained to accept values from the branches of the DOID ontology with roots *cognitive disorder*, *sleep disorder*, and *dissociative disorder*.

4 EVALUATION

CEDAR is working with several biomedical communities to perform an initial evaluation of our ontology-based annotation functionality. This evaluation is being carried out in the context of using CEDAR to develop metadata submission pipelines for three biomedical groups. These groups are (1) the LINCS Consortium,[4] which is developing a catalog of cellular signatures; (2) ImmPort,[5] a portal for immunology-related datasets; and (3) the AIRR Community,[6] which is developing standards for describing datasets acquired using advanced sequencing technologies. In all three cases, the workflow is: (1) design metadata templates for each group's relevant datasets; (2) enhance these templates with ontology-based annotations; (3) scientists populate the templates with metadata describing their experiments; and (4) submit the generated metadata to the appropriate repositories.

Working together with the LINCS, ImmPort, and AIRR teams we first used the Template Designer tool to develop a basic version of the templates required by each group. We then annotated those templates using ontologies. Each project required a slightly different annotation workflow.

To annotate ImmPort data, members of the Human Immunology Project Consortium (HIPC)[7] performed an analysis of all fields and value constraints in the ImmPort system to identify appropriate controlled-term linkages. They used the Template Designer to comprehensively annotate the ImmPort templates with the controlled terms identified. They also specified value constraints for controlled-value fields to ensure that the generated acquisition interfaces restricted the acquisition of metadata to appropriate terms. In the cases where custom value sets were required for fields, CEDAR used BioPortal's value set features to let users define these resources. The process for the AIRR community was slightly different, since that community already incorporated ontology-based annotations as an integral part of their metadata-specification process. All these annotations were available in spreadsheet format and the only required step was to formalize them using the Template Designer. Finally, the LINCS team identified and encoded controlled term linkage for an initial subset of their templates.

The system successfully represented all required controlled-term annotations for the three groups. We are now completing the metadata submission pipeline for each group. For the LINCS and ImmPort projects, we are submitting the generated metadata into their community domain repositories. The AIRR submission process involves submitting the generated metadata to the public NCBI BioSample repository.[8] We have completed prototype LINCS and NCBI pipeline submissions and will evaluate the speed, reliability, and completeness of the submission process before releasing each submission pipelines for public use.

5 DISCUSSION

Despite the growing number of ontologies in biomedicine, scientists rarely select standard terms for describing their experiments. Consequently, finding scientific datasets and understanding the corresponding experiments can be extremely hard and time-consuming, and often requires considerable post-processing of metadata to extract relevant content. A fundamental problem is the lack of convenient and openly available tools for linking metadata to ontologies. It takes time and effort to create well-specified metadata and scientists often view the task of metadata authoring as a burden that does not bring them any direct benefit.

[4] http://www.lincsproject.org
[5] http://www.immport.org
[6] http://airr-community.org
[7] https://www.immuneprofiling.org/hipc
[8] https://www.ncbi.nlm.nih.gov/biosample

The CEDAR Workbench allows template authors to make extensive use of ontologies from BioPortal to add type and property assertions to template fields and to constrain the values of fields to ontology terms. Once those templates are created, metadata authors can easily use them to generate rich metadata without needing any understanding of ontology structures. The features described in this paper represent a major step toward overcoming the barriers to the creation of high-quality metadata in biomedicine. Through our approach, we hope to make it easier, and even fun, for scientists to annotate their experimental data in ways that ensure their value to the scientific community.

We are studying a variety of technologies to further ease the work of entering metadata. We developed a recommendation service that identifies common patterns in the metadata repository and that generates real-time suggestions for filling out templates (Martínez-Romero et al., 2017). This service is the first of a planned set of intelligent authoring components that will also include the extraction and semantic annotation of templates and metadata from semi-structured sources, such as spreadsheets, scientific articles, and Web pages.

We also plan to develop an ontology enrichment pipeline in which ontology owners receive term requests based on the new classes created from CEDAR, which could be used to refine and extend their ontologies. The TermGenie (Dietze et al., 2014) tool for requesting new Gene Ontology classes provides a model for the planned functionality.

ACKNOWLEDGMENTS

CEDAR is supported by the National Institutes of Health through an NIH Big Data to Knowledge program under grant 1U54AI117925. NCBO is supported by the NIH Common Fund under grant U54HG004028. The CEDAR Workbench is available at https://cedar.metadatacenter.net, and on GitHub (https://github.com/metadatacenter).

REFERENCES

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., et al. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet*, *29*, 365–371.

Bui, Y., & Park, J.-R. (2006). An assessment of metadata quality: A case study of the National Science Digital Library Metadata Repository. *Proceedings of CAIS/ACSI 2006*.

Dietze, H., Berardini, T. Z., Foulger, R. E., Hill, D. P., Lomax, J., Osumi-Sutherland, D., Roncaglia, P., et al. (2014). TermGenie – a web-application for pattern-based ontology class generation. *Journal of Biomedical Semantics*, *5*(1), 48.

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, *30*(1), 207–210.

González-Beltrán, A., Maguire, E., Sansone, S.-A., & Rocca-Serra, P. (2014). linkedISA: semantic representation of ISA-Tab experimental metadata. *BMC bioinformatics*, *15 Suppl 1*, S4.

Martínez-Romero, M., O'Connor, M. J., Shankar, R., Panahiazar, M., Willrett, D., Egyedi, A. L., Gevaert, O., et al. (2017). Fast and accurate metadata authoring using ontology-based recommendations. *Proceedings of AMIA 2017 Annual Symposium (to appear)*.

McQuilton, P., Gonzalez-Beltran, A., Rocca-Serra, P., Thurston, M., Lister, A., Maguire, E., & Sansone, S.-A. (2016). BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. *Database: the journal of biological databases and curation*, *2016*.

Musen, M. A., Bean, C. A., Cheung, K. H., Dumontier, M., Durante, K. A., Gevaert, O., Gonzalez-Beltran, A., et al. (2015). The Center for Expanded Data Annotation and Retrieval. *Journal of the American Medical Informatics Association*, *22*(6), 1148–1152.

Musen, M. A., Noy, N. F., Shah, N. H., Whetzel, P. L., Chute, C. G., Story, M.-A., & Smith, B. (2012). The National Center for Biomedical Ontology. *Journal of the American Medical Informatics Association*.

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., et al. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, *37*(Web Server issue), W170-3.

O'Connor, M. J., Martinez-Romero, M., Egyedi, A. L., Willrett, D., Graybeal, J., & Musen, M. A. (2016). An open repository model for acquiring knowledge about scientific experiments. *Proceedings of the 20th International Conference on Knowledge Engineering and Knowledge Management (EKAW2016)*. (Vol. 10024, pp. 762–777).

Parkinson, H., Sarkans, U., Shojatalab, M., Abeygunawardena, N., Contrino, S., Coulson, R., Farne, A., et al. (2005). ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic acids research*, *33*(Database issue), D553–D555.

Rocca-Serra, P., Brandizi, M., Maguire, E., Sklyar, N., Taylor, C., Begley, K., Field, D., et al. (2010). ISA software suite: Supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics*, *26*(18), 2354.

Shankar, R., Parkinson, H., Burdett, T., Hastings, E., Liu, J., Miller, M., Srinivasa, R., et al. (2010). Annotare-a tool for annotating high-throughput biomedical investigations and resulting data. *Bioinformatics*, *26*(19), 2470–2471.

Tenenbaum, J. D., Sansone, S.-A., & Haendel, M. (2014). A sea of standards for omics data: sink or swim? *Journal of the American Medical Informatics Association*, *21*(2), 200–203.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018.

Wolstencroft, K., Owen, S., Horridge, M., Krebs, O., Mueller, W., Snoep, J. L., du Preez, F., et al. (2011). RightField: Embedding ontology annotation in spreadsheets. *Bioinformatics*, *27*(14), 2021–2022.