CEDAR OnDemand: a browser extension to generate ontology-based scientific metadata

Syed Ahmad Chan Bukhari 1*, Marcos Martínez-Romero 4, Martin J. O’Connor 4, Attila L. Egyedi 4, Debra Willrett 4, John Graybeal 4, Mark A. Musen 4, Kei-Hoi Cheung 2,3* and Steven H. Kleinstein 1,2*

Abstract

Background: Public biomedical data repositories often provide web-based interfaces to collect experimental metadata. However, these interfaces typically reflect the ad hoc metadata specification practices of the associated repositories, leading to a lack of standardization in the collected metadata. This lack of standardization limits the ability of the source datasets to be broadly discovered, reused, and integrated with other datasets. To increase reuse, discoverability, and reproducibility of the described experiments, datasets should be appropriately annotated by using agreed-upon terms, ideally from ontologies or other controlled term sources.

Results: This work presents “CEDAR OnDemand”, a browser extension powered by the NCBO (National Center for Biomedical Ontology) BioPortal that enables users to seamlessly enter ontology-based metadata through existing web forms native to individual repositories. CEDAR OnDemand analyzes the web page contents to identify the text input fields and associate them with relevant ontologies which are recommended automatically based upon input fields’ labels (using the NCBO ontology recommender) and a pre-defined list of ontologies. These field-specific ontologies are used for controlling metadata entry. CEDAR OnDemand works for any web form designed in the HTML format. We demonstrate how CEDAR OnDemand works through the NCBI (National Center for Biotechnology Information) BioSample web-based metadata entry.

Conclusion: CEDAR OnDemand helps lower the barrier of incorporating ontologies into standardized metadata entry for public data repositories. CEDAR OnDemand is available freely on the Google Chrome store https://chrome.google.com/webstore/search/CEDAROnDemand

Keywords: Ontology, Metadata, CEDAR, FAIR, BioPortal, NCBI, NCBO

Background

Biomedical data are increasingly being deposited in public repositories accompanied by descriptive metadata. These metadata are crucial for facilitating the discovery of the associated datasets and for reproducing the corresponding experiments. Many public data repositories provide web-based forms for researchers to enter metadata describing their datasets as part of the submission process. However, most repositories make limited use of controlled vocabularies in the metadata entry process and, as a result, metadata are often described using inconsistent terminologies [1]. This lack of standardization makes it difficult to access, find, interoperate, and reuse the datasets, and—crucially—to understand how the associated experiments were performed. Improvements are needed to make these datasets more FAIR (Findable, Accessible, Interoperable, and Reusable) [2]. The use of terms from controlled terminologies and ontologies can provide an important first step for creating FAIR metadata descriptions [3].

A wide array of ontology-based services have been developed in order to promote scientific data interoperability and reusability in biomedicine through the use of
standard terminologies. These include BioPortal [4], the Ontology Lookup Service (OLS) [5], EBI Zooma [6], and NCBO Annotator and Recommender [7, 8]. In addition, data (metadata) standardization efforts have been established by different communities to ensure sufficient amount of information (metadata) be provided for reporting results in a way that facilitates reproducibility such as MIAIME (Minimum Information for Reporting Microarray Experiment) [9], MiAIRR (Minimal Information about Adaptive Immune Receptor Repertoire) [10, 11] and MIBBI (Minimum Information for Biological and Biomedical Investigations) [12]. The Center for Expanded Data Annotation and Retrieval (CEDAR) [13] has leveraged existing data standards and the ontologies available at BioPortal to develop the CEDAR Workbench with the goal of creating semantically rich metadata. A user either can create a new template (web form) or can use existing ontology-controlled templates to author standardized metadata within CEDAR Workbench. An example employing CEDAR Workbench for customized data submission is [14]. Expanding CEDAR’s approach of metadata creation outside of its environment, we have incorporated BioPortal ontologies and web services to develop a decentralized metadata authoring tool called “CEDAR OnDemand”. CEDAR OnDemand is a platform-independent program running as a web browser extension designed to help creating standardized metadata in repository-native web forms. The key advantage of this approach is that it enables users to seamlessly enter ontology-based metadata into existing web forms without requiring the individual repositories to provide these services.

Implementation

The CEDAR OnDemand script has been developed as a Google chrome browser extension [15] (a browser extension is essentially a small software program that can access contents of a web page, modify it and can enhance the functionality of a web browser). It is powered by the NCBO Annotator [7] and Recommender [8] Web services and facilitates users to suggest entry-time ontology controlled metadata to fill up web forms. After installation, the extension will appear as an icon on the chrome extension bar (upper right side of the browser). It is designed to be manually toggled on upon entry of a web form (it can be toggled off later if needed). Although CEDAR OnDemand can be programmed to be auto-activated, we used the manual activation method to minimize the system memory usage and to protect users from browser-based security attacks [16]. The extension operates in three phases (described below) that are initiated when a user visits a new web-based (metadata) entry form.

Fig. 1 CEDAR OnDemand Workflow. (1) CEDAR OnDemand is installed as a Google Chrome browser extension. When activated by the user (by toggling the icon), the web page in the browser will be analyzed. Users can customize a list of suggested ontologies used by CEDAR OnDemand and any point through a dialogue box with dropdown list (2) An HTML tags detection script identifies the text fields by analyzing the HTML INPUT tags and fetches their labels. Fetched labels are then passed to the NCBO to get the related ontology recommendations. This BioPortal ontology recommendation list is compared to the user suggested ontology list to find the qualified ontologies (3) CEDAR OnDemand associates the qualified ontology list to the detected input fields. Subsequently, NCBO Annotator service is invoked to match field values with the qualified set of ontologies to suggest ontology-based metadata.
Identification of data entry fields
To detect data entry fields, the web page is analyzed to identify text input fields and the associated field labels (Fig. 1, left side). CEDAR OnDemand parses the content of a web page into the document object model (DOM) [17], which defines the content, structure and style of an HTML document (Fig. 1, left panel treeview). The current implementation of CEDAR OnDemand recognizes the standard INPUT fields (HTML5 and previous versions) and their associated labels (HTML5 element). The recognized fields are highlighted with light yellow color. The metadata entry of the detected input fields will be controlled by the list of ontologies chosen by the qualified ontologies.

Ontologies recommendation algorithm
The CEDAR OnDemand ontology recommendation algorithm is designed to recommend ontologies relevant to each input field listed in a webform from the BioPortal [4] ontologies. CEDAR OnDemand takes each field label as input (as shown in Label 2 in Fig. 1) to the NCBO Recommender 2.0 service [8] to get a list of BioPortal ontologies (containing terms matching the field label). Moreover, a user can also define ontologies through a dialogue box which appears by toggling the CEDAR OnDemand extension. The CEDAR OnDemand algorithm takes the intersection of the set of user-defined ontologies and that of ontologies recommended automatically (by the NCBO recommender) to produce the set of qualified ontologies for each field. These field-specific qualified ontologies are then linked to each input field in a web form. If the intersection is an empty set, then the full user-defined list is used for as the qualified ontologies for controlling the field entry. By default, the user-defined list includes six ontologies: ChEBI Ontology [18], Human Disease Ontology (DOID) [19], Gene Ontology (GO) [20], Ontology for Biomedical Investigations (OBI) [21], Phenotypic Quality Ontology [22], Protein Ontology (PR) [23] (Fig. 1, Label 2). Not only do these ontologies cover a broad range of biological domains, but they are also ranked among the top ten by OBO Foundry in terms of their compliance to ontology best practice [24]. The user may change the default ontology list by adding/removing ontologies anytime during the metadata entry process. In its default behavior CEDAR OnDemand works fully automatically and does not require an ontology input from the user. However, customizing the default ontology list may help the user to get domain-specific metadata suggestions.

Ontology association and auto-completion of metadata
To associate ontology terms (e.g., “myasthenia gravis” from DOID) with the field entry (e.g., disease), CEDAR OnDemand matches the term(s) entered by the user with the terms defined in the qualified ontologies (Fig. 1, Label 3). This is done by invoking the NCBO Annotator web service [8] through AJAX (asynchronous JavaScript and XML) call [25]. AJAX communicates with NCBO BioPortal server [26] asynchronously (in the background) through XMLHttpRequest Object to send and retrieve the data. This asynchronous communication model of CEDAR OnDemand enables entry-time suggestion for ontology controlled metadata entry. The NCBO Annotator returns a ranked list of ontology term matches for the user to choose.

Results
We tested CEDAR OnDemand by entering metadata using the NCBI human BioSample web form” [27]. In this use case, we first extended the user defined ontology list by adding several field-specific ontologies identified through NCBO recommender: Cell ontology (CL) [28], Cell Line Ontology (CLO) [29], NCBI thesaurus NCIT [30], NCBI Taxonomy ontology NCBITAXON [31], and Uber Anatomy Ontology (UBERON) [32]. The NCBI human BioSample web form contains twenty-one text input fields. CEDAR OnDemand suggested eight ontologies based on the input fields in the NCBI human BioSample web form. After intersection with the user defined ontologies (extended list), the final ontology list recommended by the CEDAR OnDemand includes: NCIT thesaurus NCIT [30], Cell Ontology [28], Cell Line Ontology [33], (UBERON) [32], Human disease Ontology [19], Gene Ontology (GO) [20] and OBI [21] (See Table 1). Controlled vocabularies do not make sense for some text fields, such as “Sample Name”, “Age” and “isolate”. Therefore, CEDAR OnDemand allows the user to override ontology suggestions for all fields with the user-defined entries. CEDAR OnDemand provides the field’s specific metadata suggestion controlled by ontologies. Thus, users are no longer entering free text but they are instead using standardized ontology terms. An auto-completion feature is provided at runtime through a drop-down list. As an example (Fig. 1, Label 3), CEDAR OnDemand suggests “myasthenia gravis” as controlled term (defined in DOID) for the disease field.

Discussion
Although many public repositories, such as those run by the NCBI, provide easy-to-use tools and interfaces for entering and querying metadata, scientists who upload their datasets are generally not constrained to use standard terminologies when they define the necessary metadata. As a result, metadata are often described using inconsistent terminologies, limiting scientists’ ability to access, find, interoperate and reuse the datasets and to understand how the experiments were performed. Scientific data analysis or mining [34] often requires multiple datasets to be integrated within a single repository or
Table 1 CEDAR OnDemand Qualified Ontologies for each NCBI BioSample Field

| Field names         | Qualified ontologies                                                                 |
|---------------------|-------------------------------------------------------------------------------------|
| Sample Name         | Ontology for Biomedical Investigations (OBI), National Cancer Institute Thesaurus (NCIT) |
| Organism            | National Cancer Institute Thesaurus (NCIT)                                          |
| Isolate             | National Cancer Institute Thesaurus (NCIT)                                          |
| Age                 | National Cancer Institute Thesaurus (NCIT)                                          |
| Biomaterial Provider| National Cancer Institute Thesaurus (NCIT)                                          |
| Tissue              | Uber Anatomy Ontology (UBERON), Ontology for Biomedical Investigations (OBI), National Cancer Institute Thesaurus (NCIT), Cell Ontology (CL), Cell Line Ontology (CLO) |
| Cell line           | Cell Line Ontology (CLO), Ontology for Biomedical Investigations (OBI), National Cancer Institute Thesaurus (NCIT) |
| Cell subtype        | Cell Ontology (CL), Gene Ontology (GO), National Cancer Institute Thesaurus (NCIT) |
| Cell type           | Cell Ontology (CL), Cell Line Ontology (CLO), National Cancer Institute Thesaurus (NCIT) |
| Culture Collection  | National Cancer Institute Thesaurus (NCIT)                                          |
| Development Stage   | Gene Ontology (GO), National Cancer Institute Thesaurus (NCIT)                      |
| Disease             | Human Disease Ontology (DOID), Cell Line Ontology (CLO), Ontology for Biomedical Investigations (OBI), National Cancer Institute Thesaurus (NCIT) |
| Disease Stage       | Human Disease Ontology (DOID), Cell Line Ontology (CLO), Ontology for Biomedical Investigations (OBI), National Cancer Institute Thesaurus (NCIT) |
| Ethnicity           | National Cancer Institute Thesaurus (NCIT)                                          |
| Health state        | National Cancer Institute Thesaurus (NCIT)                                          |
| Karyotype           | National Cancer Institute Thesaurus (NCIT)                                          |
| Phenotype           | Ontology for Biomedical Investigations (OBI), National Cancer Institute Thesaurus (NCIT) |
| Population          | Ontology for Biomedical Investigations (OBI)                                        |
| Race                | National Cancer Institute Thesaurus (NCIT)                                          |
| Sample type         | National Cancer Institute Thesaurus (NCIT)                                          |
| Treatment           | Ontology for Biomedical Investigations (OBI), National Cancer Institute Thesaurus (NCIT) |

Field Names column lists the Human Sample attributes of NCBI BioSample. Qualified Ontologies are the ontologies which CEDAR OnDemand algorithm recommends.
operating systems. We are exploring the possibility of supporting other types of browsers (e.g., Firefox and Microsoft Edge).

Conclusions
CEDAR OnDemand is a chrome browser extension that enables users to seamlessly enter ontology-controlled metadata using existing web-based submission forms provided by metadata repositories. The use of controlled vocabularies for entering metadata can help improve the quality of metadata submitted to repositories and ultimately contributes to the creation of FAIR data.

Availability and requirements
Availability: https://chrome.google.com/webstore/search/CEDAROnDemand
Code Availability: https://github.com/ahmadchan/CEDAROnDemand
Project name: CEDAR OnDemand.
Operating system(s): Operating system independent works within web browser.
Programming language: Javascript.
License: GPL.
Any restrictions to use by non-academics: none.

Endnote
1https://submit.ncbi.nlm.nih.gov/subs/biosample/

Acknowledgements
We acknowledge the BioPortal and CEDAR team for their valuable suggestions during this research work.

Funding
This work was supported by grant USA AI117925 awarded by the National Institute of Allergy and Infectious Diseases through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (https://commonfund.nih.gov/bd2k).

Authors’ contributions
Study conception and design: SACB, KHC, SHK, JB, MAM. Code Implementation: SACB. Validated and interpreted the results: SACB, JB, MOC, DB, ALE. Drafting of manuscript: SACB, SHK, KHC. Critical revision: SACB, MMR. Implementation: SACB. Validated and interpreted the results: SACB, JB, MOC, SH. Study conception and design: SACB, KHC, SHK, JB, MAM. Code Availability: https://github.com/ahmadchan/CEDAROnDemand

References
1. Gonçalves RS, O’Connor MJ, Martínez-Romero M, Graybeal J, Musen MA: Metadata in the BioSample Online Repository are Impaired by Numerous Anomalies. arXiv [cs.DB] 2017.
2. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bouma PE, Bouwman J, Brooks AJ, Clark T, Crossas M, Dillo I, Dumont O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AIG, Groth P, Goble C, Grethe JS, Heringa J, ’t Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Penison B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Straw G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B: The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016, 3:160018.
3. Shadbolt N, Berners-Lee T, Hall W: The semantic web revisited. IEEE Intell Syst. 2006;21:96–101.
4. Whetzel PL, NCBO Team. NCBO Technology: Powering semantically aware applications. J Biomed Semantics. 2013;4(Suppl 1):S8.
5. Ison J, Kalai M, Jonassen I, Botser D, Uldlag M, McWilliam H, Malone J, Lopez B, Pettifer S, Rice P: EDM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. Bioinformatics. 2013;29:1325–32.
6. ZOOMA text annotations tool. http://www.ebi.ac.uk/spot/zooma/.
7. Jonquet C, Shah NH, Youn CH, Gallendar C, Storey M-A, Musen MA. NCBO annotator: semantic annotation of biomedical data. International Semantic Web Conference, Poster and Demo session. 2009. https://pdfs.semanticscholar.org/9956/99824d17db4b87374931085ca4b6e0d0518a99fd.pdf.
8. Martínez-Romero M, Jonquet C, O’Connor MJ, Graybeal J, Pazzos A, Musen MA. NCBO Ontology Recommender 2.0: an enhanced approach for biomedical ontology recommendation. J Biomed Semantics. 2017;8:21.
9. Brazma A: Minimum information about a microarray experiment (MIAME)–successes, failures, challenges. Sci World J. 2009;9:4020–3.
10. Rubelt F, Busse CE, Bukhari SAC, Büchert J-P, Malottti-Ferrandiz E, Cowell LG, Watson CT, Mathandana N, Faison WI, Hershberg U, Laserson U, Corrie BD, Davis MM, Peters B, Lefranc MP, Scott JK, Breden F: AIRR community. Luning Prak ET, Kleinstein SH: adaptive immune repertoire receptor community recommendations for sharing immune-repertoire sequencing data. Nat Immunol. 2017;18:1274–8.
11. Breden F, Luning Prak ET, Peters B, Rubelt F, Schrarmm CA, Busse C, Vander Heiden JA, Christley S, Bukhari SAC, Thorogood A, Matisen F, Wine Y, Laserson U, Katsuzuma D, Douek D, Lefranc MP, Collins AM, Bubela T, Kleinstein S, Watson CT, Cowell LG, Scott JK, Kepler TB: Perspective: Reproducibility and Use of Adaptive Immune Repertoire Receptor Data. Front Immunol. 2017;8:333.
12. Kartner C, Field D, Sansone S-A, Taylor C, Aerts J, Binns N, Blake A, Britten CM, de Marco A, Fostel J, Gaudet P, Gonzalez-Beltran A, Hardy N, Hettmans J, Herrjakobi J, Huty N, Leebens-Mack J, Maguire E, Neumann S, Orchard S, Parkinson H, Piel W, Ranganathan S, Rocca-Serra P, Santasiero A, Shotton D, Stierk P, Untergasser A, Whetzel PL: Meeting report from the second “minimum information for biological and biomedical investigations” (MIBBI) workshop. Stand Genomic Sci. 2010;3:259–66.
13. Musen MA, Bean CA, Cheung K-H, Dumontier M, Durante KA, Gevaert O, Gonzalez-Beltran A, Khati P, Kleinstein SH, O’Connor MJ, Pouliot Y, Rocca-Serra P, Sansone S-A, Wiser JA, Yokley J, Zeller K: CEDAR team: the center for expanded data annotation and retrieval. J Am Med Inform Assoc. 2015;22:1148–52.
14. Bukhari SAC, O’Connor MJ, Graybeal J, Musen MA, Cheung K-H, Kleinstein SH: Leveraging the CEDAR Workbench for Ontology-linked Submission of Adaptive Immune Repertoire Receptor Data to the Sequence Read Archive (SRA). https://doi.org/10.5281/zenodo.4244126
15. Mehta P: Introduction to Google Chrome Extensions. In: Creating Google Chrome Extensions: Apress. New Delhi: Springer; 2016. p. 1–33. https://link.springer.com/content/pdf/10.1007/978-1-4842-1775-7.pdf.
16. Shital P: Web browser security: different attacks detection and prevention techniques. IJCAI. 2017;170:33–41.
