Using Semantic Technologies to Enhance Metadata Submissions to Public Repositories in Biomedicine

Attila L. Egyedi, Martin J. O'Connor, Marcos Martínez-Romero, Debra Willrett, Josef Hardi, John Graybeal, and Mark A. Musen

Stanford Center for Biomedical Informatics Research
Stanford University, Stanford, CA 94305, USA
attila.egyedi@stanford.edu

Abstract. The emergence of the FAIR principles is driving renewed efforts in the biomedical community to produce high-quality metadata that describe datasets submitted to public repositories. A variety of organizations are now involved in developing submission pipelines that place a strong emphasis on accompanying submissions with highly descriptive metadata. However, these pipelines have highly variable requirements, which range from using ontology-based metadata in existing submission pipelines to supporting end-to-end metadata management in new pipelines. There is a lack of tools for integrating metadata support when building these pipelines. In this paper we describe a system called CEDAR that aims to address this challenge. The described tools provide a flexible, highly configurable solution for producing submission workflows with semantically rich metadata support. We outline how we have used these tools to deliver robust metadata submission pipelines for several communities, including the Adaptive Immune Receptor Repertoire (AIRR), the NIH Cloud Credits Model Pilot (CCP), and the Library of Integrated Network-based Cellular Signatures (LINCS).

Keywords: Metadata, Metadata Management, Ontologies

1 Introduction

A large number of data repositories have been developed in the biomedical community over the past few decades. These repositories are usually provided by large government-funded institutions for general scientific use or may be developed by specialized communities for domain-specific purposes. For example, the U.S. National Center for Biotechnology Information (NCBI) provides an array of repositories, including GenBank [1], which holds DNA sequence metadata, and Sequence Read Archive (SRA) [2], which contains descriptions of biological sequence data. Scientists submitting their datasets to these repositories are required to accompany these submissions with metadata describing the associated experiments. The ability to discover datasets and reproduce experiments is highly dependent on the quality of the metadata in these repositories.

CEUR-WS.org/Vol-2275/paper1.pdf

The submission interfaces provided by many repositories are often notoriously difficult to use. Often the submission process is spread over multiple stages, with metadata upload requiring a mixture of Web-based acquisition forms and population of spreadsheets. If associated data files are large, they may need to be submitted separately—or submitters may be required to assemble metadata and data files into submission packages for batch upload. Users are frequently responsible for ensuring the consistency of internal references in their metadata (e.g., to sample identifiers) and may also need to ensure that those references align with the associated data files. In some cases, references may be required to identify previous submissions. Validation and error reporting processes are often poor. Frequently, validation occurs post-submission, and users are informed of failures via email. Intervention by repository staff to manually repair submissions is not uncommon. As a result, generating conforming metadata for many repositories can require significant effort and often involves a degree of trial and error. Additionally, current submission interfaces typically lack any standard way of semantically annotating the data. Despite the availability of a large number of controlled terminologies in biomedicine, submission systems have weak or nonexistent mechanisms for linking terms from these terminologies to the metadata for the submissions.

There is evidence that this combination of limited semantic enforcement and onerous submission processes negatively affects the quality of metadata in repositories [3]. However, there are no general-purpose tools that can support the diverse requirements when developing metadata-submission interfaces and associated submission pipelines. Several tools address individual parts of the submission process. Some of these tools focus on improving the spreadsheet entry part of the submission process. One of the most popular is ISA Tools [4], which is a desktop application that allows curators to create spreadsheet-based submissions. A later evolution of this tool called Linked ISA [5] provided a means of annotating spreadsheets with controlled terms. RightField [6], an Excel-based plugin, also focuses on semantic annotation, allowing users to embed ontology terms in spreadsheets. A variety of custom tools have also been developed to improve the submission processes to existing repositories. A desktop-based tool called Annotare [7] supports metadata submission for functional genomics experiments to the ArrayExpress repository, replacing its previous spreadsheet-based submission mechanism. NCBI's Sequence Read Archive (SRA) repository has spawned the development of a variety of custom submission tools that replace its spreadsheet-based upload. Projects including BaseSpace[1], mothur[2], and CyVerse[3] have developed submission systems that semi-automatically upload metadata to SRA.

However, existing tools tend to be either highly customized or address only a small part of the metadata submission process. There is a need for general-purpose tools that can both enhance existing pipelines and support end-to-end frameworks for new repositories. In this paper, we describe such a tool set and outline how it has been adopted by several communities to develop and enhance a variety of metadata submission pipelines.

[1] https://basespace.illumina.com/
[2] https://mothur.org/
[3] https://www.cyverse.org/

2 Requirements for Developing Metadata Submission Pipelines

The requirements for developing metadata submission pipelines can be divided into three broad areas: (1) metadata-template specification, which primarily involves formally encoding the structure of anticipated metadata; (2) metadata acquisition, which involves gathering conforming metadata from users; and (3) metadata submission, which targets the final uploading of acquired metadata and associated data to repositories. In general, communities developing submission pipelines target either one or more existing public metadata repositories, or topic-specific community-developed repositories. The tool requirements for these development efforts depend both on the difficulty of satisfying target repository interfaces and the level of automation and user assistance desired in the pipeline.

The first challenge that many of these efforts must tackle is formally encoding a template for the relevant metadata standard. The most common strategy of defining metadata attribute names as entries in a spreadsheet is too imprecise to support rigorous metadata definitions. More structured formats such as XML or JSON add some precision but also suffer from a lack of rigor. Irrespective of format, there is no agreed way to use these specifications, so the developers of metadata pipelines typically define an *ad hoc* metadata template specification approach using their preferred technology choice and formality level. Additionally, there are no standard ways of semantically enhancing these specifications—for example, to restrict acquired values to controlled terminologies—so the eventual metadata templates can be very loosely defined.

Once the metadata template is defined, mechanisms for acquiring template-conformant metadata from users must be created. Again, however, since there are no standard approaches, pipeline developers must create custom metadata acquisition interfaces—or leave users to their own devices when, for example, manually populating spreadsheets or generating XML-based files. The recent focus on enhancing metadata quality has driven the desire for easy-to-use Web-based acquisition interfaces, which can require significant development effort. These interfaces should both reflect the relevant metadata standard and—ideally—enforce strong quality standards.

Converting the acquired metadata to meet repository submission specifications is the next challenge. Because the final submission workflow for many public repositories can be onerous, this step often emphasizes simplifying the submission process. Challenges include validating the metadata to ensure that they conform to repository specification, uploading associated data files (which for many biomedical experiment types can be very large), monitoring the submission, and reporting outcomes.

While not all pipeline developers need to address all requirements in depth, a general solution must satisfy the needs of many different pipelines. A significant number of tools are needed to meet even minimal requirements for a complete core set of metadata tasks. Such tools must be flexible and configurable and must be able to support a variety of integration strategies.

3 The CEDAR Metadata Authoring and Submission Workflow

The Center for Expanded Data Annotation and Retrieval (CEDAR) [8] has developed a system that addresses these diverse requirements. CEDAR supports a workflow for metadata management that is organized around the three main stages of the metadata submission process—namely, metadata-template specification, metadata acquisition, and metadata submission. A driving goal of CEDAR is to provide highly configurable tools that support the creation of metadata-submission pipelines to meet a wide range of deployment scenarios. It is a modular system that provides components that can be integrated into existing workflows to address specific tasks in a metadata submission pipeline or that can be assembled together to provide an end-to-end pipeline. The system—referred to as the CEDAR Workbench [9]—is built around the notion of creating *templates* that define the structure and semantics of metadata specifications. These templates support a metadata-submission workflow that acquires conforming metadata and uploads the resulting metadata to repositories (see Figure 1).

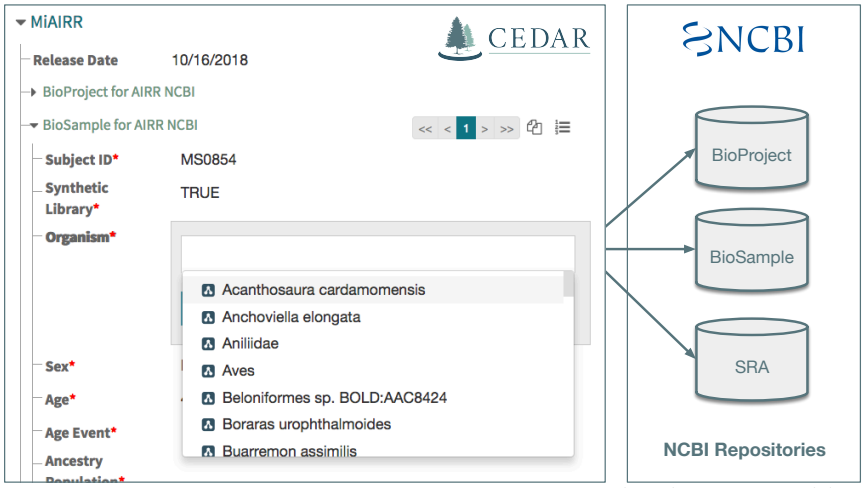

Figure 1. Example of CEDAR metadata submission. A metadata form generated by the CEDAR Workbench collects semantically annotated metadata and submits the metadata to three NCBI repositories: BioProject, BioSample, and SRA. Some of the BioSample attributes are visible in the above form; here the user is filling out the *organism* attribute.

CEDAR's overall metadata workflow (Figure 2) comprises the following three steps: (1) Template authors use a CEDAR tool called Template Designer to create templates describing metadata, typically following discipline-specific standards or minimum information models based on the type of experimental data to be annotated. Authors can define their templates using controlled terms, ontologies, and standard datasets supplied by the BioPortal ontology portal [10, 11]. (2) When a scientist or other metadata provider chooses to populate a template, a CEDAR tool called the Metadata Editor automatically generates a form-based interface from the template; the scientist then uses the Metadata Editor interface to enter the descriptive metadata. When users

populate these forms, the semantic annotations specified by the associated template are used to present ontology-controlled suggestions to users and ensure that collected metadata conform to the published specification. (3) Once the metadata have been entered, scientists can use the CEDAR Submission Service to upload metadata and associated experimental data to a target repository.

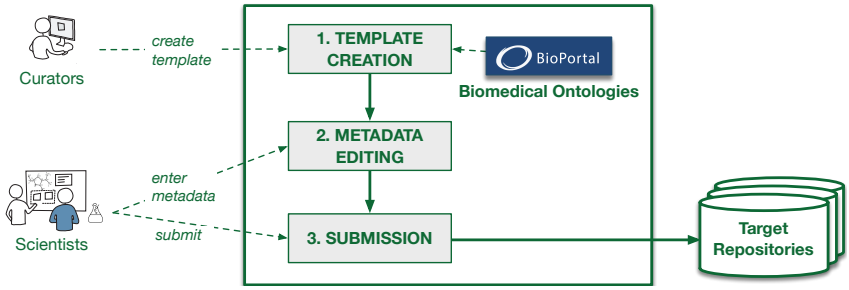

Figure 2. CEDAR's Template-Based Metadata Submission Workflow. (1) A template author creates a template to define the structure and content of a particular metadata specification; (2) CEDAR generates a graphical interface with which a scientist can enter metadata; (3) the metadata and optionally the accompanying data can be submitted to an external repository.

4 Using CEDAR to Develop Metadata Submission Pipelines

A variety of communities use CEDAR to satisfy varied metadata submission requirements. We describe three communities' deployments that illustrate CEDAR techniques to operationalize metadata submission pipelines. The three communities are: (1) the Adaptive Immune Receptor Repertoire (AIRR) Community [12], which studies human immune response; (2) the National Institutes of Health (NIH) Cloud Credits Model Pilot (CCP) [13], which was established to evaluate approaches for allocating scientific computational resources in the NIH Data Commons; and (3) the Library of Integrated Network-based Cellular Signatures (LINCS) [14], a consortium studying cell signaling to learn how cells respond to various genetic and environmental stressors.

4.1 AIRR

The AIRR Community[4] uses advanced DNA sequencing technologies to study the human immune response. AIRR researchers identified the lack of standards to describe their datasets as a bottleneck to their progress and created a working group to establish formal community-driven guidelines for metadata. Metadata conforming to those guidelines are targeted for submission to several repositories provided by the National Center for Biotechnology Information (NCBI). One of the first formal standards produced by this group is called MiAIRR [15]. MiAIRR is a metadata standard for capturing the minimal information, or principal characteristics, of experiment types collectively referred to as repertoire sequencing. Metadata described by this standard, along

[4] https://www.antibodysociety.org/the-airr-community/

with the corresponding datasets, are submitted to NCBI's BioProject, BioSample, and SRA repositories. These repositories have quite complicated submission interfaces, particularly when multi-repository upload is required. The AIRR community wanted to provide a unified, user-friendly submission interface that reflected the MiAIRR standard. They also wanted an interface that enforced the strong semantic restrictions placed on field values by the standard and that could handle the submission of the large sequencing files associated with AIRR studies.

In collaboration with members of the AIRR community, we operationalized an end-to-end submission pipeline [16] (Figure 3). First, members of the AIRR Community used the Template Designer to create a template that captured the structural and semantic requirements of the MiAIRR standard. Submitting scientists can now use the Web-based form generated from the MiAIRR template by the Metadata Editor to enter ontology-controlled metadata associated with their AIRR studies, and to upload metadata and associated sequencing data to the three target NCBI repositories. We extended CEDAR's Submission Manager to transform the entered metadata into a form compatible with NCBI's BioProject, BioSample, and SRA repositories, and configured the Metadata Editor tool to allow submission of MiAIRR-based metadata through the Submission Manager. We also implemented a file upload mechanism that could submit large sequencing files using the NCBI FTP-based file upload service. This submission pipeline was released in September 2018 and is in routine use by the AIRR community.

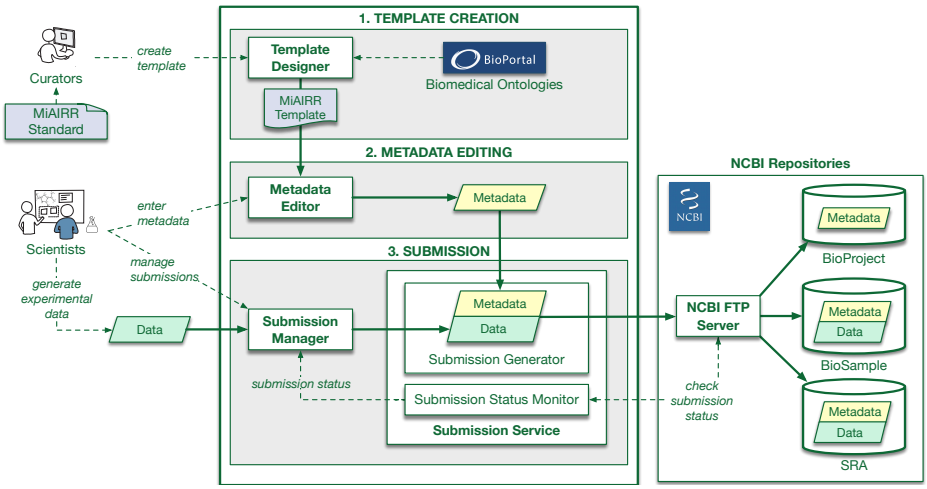

Figure 3. AIRR Metadata Submission Workflow. (1) *Create Template.* AIRR Community members defined their MiAIRR specification and used the Template Designer to encode it. (2) *Enter Metadata.* A scientist with AIRR data fills out descriptive metadata using a single form, rather than the three web forms provided by NCBI, and can validate the form's BioSample information against an NCBI submission validator. (3) *Submit Metadata and Data.* The scientist associates the metadata form with the data sets that it describes, and the Submission Server then submits the relevant metadata and the accompanying data to the three NCBI repositories.

4.2 CCP

The Cloud Credits Model Pilot[5] (CCP) was established to evaluate approaches for allocating scientific computational resources in the NIH Data Commons. It explores a credits allocation model to encourage the sharing of various types of digital objects resulting from NIH research in the Cloud. Researchers funded by the CCP are required to upload metadata describing their experiments' digital objects to the DataMed [17] data discover index. The digital objects generated by these researchers are uploaded to Cloud-based platforms and are referenced by uploaded metadata cached in DataMed. While DataMed's internal metadata is described by a model called Data Tag Suite (DATS) [18], it had no interfaces for external users to submit conforming metadata.

To satisfy the needs of the CCP, the DataMed team worked with the CEDAR system to provide a Web-based submission interface (Figure 4). The DataMed team first developed template representing the DATS model using the CEDAR Template Designer. Metadata submitters use a Web-based acquisition form generated from this template to enter their metadata. While these templates are being filled out, CEDAR's Metadata Editor ensures that the relevant semantic restrictions are enforced. Upon completing the metadata, the submitter sets an attribute in the metadata to indicate that DataMed should index the metadata. DataMed monitors the submissions nightly and indexes any submissions that are marked as ready, updating any previously indexed submissions that have changed. The submission process involves only metadata, since CCP users separately upload their digital objects to Cloud-based platforms. This pipeline went into production in July 2018, and it is regularly accepting submissions from CCP users.

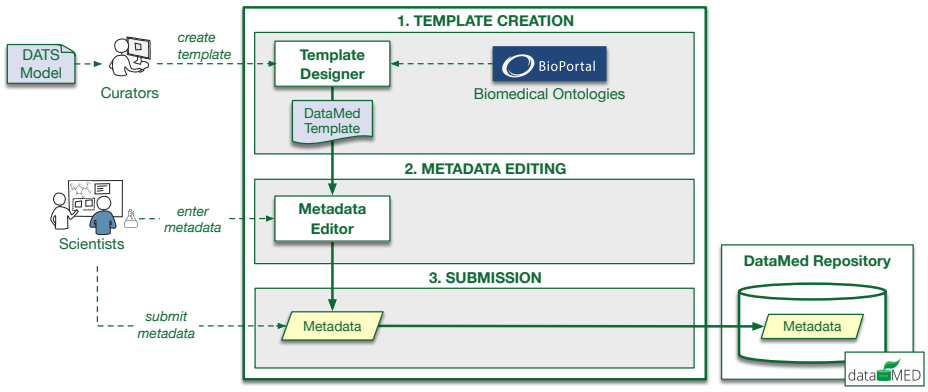

Figure 4. NIH Cloud Credits Model Pilot Metadata Submission Workflow. (1) *Create Template.* The DataMed team created a template describing a DATS-compatible submission. (2) *Enter Metadata.* Using this template, CCP users fill out descriptive metadata about their digital objects. Multiple objects may be described in a single instance, or by entering metadata separately for each object. (3) *Submit Metadata.* When users have finished entering metadata they indicate that the relevant metadata are ready, whereupon DataMed detects and retrieves these new submissions nightly. CCP users separately upload the referenced digital objects to Cloud-based platforms.

[5] https://www.commons-credit-portal.org/

4.3 LINCS

The Library of Integrated Network-Based Cellular Signatures (LINCS) [14] is a consortium of biologists studying cell signaling to learn how cells respond to various genetic and environmental stressors. The LINCS Data Coordination and Integration Center (LINCS-DCIC) created an Integrated Knowledge Environment for managing LINCS-related dataset submissions from Consortium members. The LINCS-DCIC specified templates to represent metadata about the various biological entities involved in the relevant experiment types. LINCS originally obtained metadata through an online platform, the LINCS Dataset Submission Tool (DST), using spreadsheet-based representations of these templates. As the templates became more complex, populating the spreadsheets became more difficult for users. Additionally, LINCS-DCIC had increased the use of controlled terms in the template specifications and wished to enforce these restrictions at data-acquisition time. LINCS desired a more robust, user-friendly acquisition process that could be easily extended to support new templates.

We worked with the LINCS-DCIC team to develop a Web-based submission pipeline that is integrated into the DST's existing submission workflow (Figure 5). Members of the LINCS community began by defining CEDAR templates to represent all current LINCS templates. CEDAR's Metadata Editor reads these templates to generate corresponding Web-based metadata-acquisition forms. The LINCS-DCIC development team integrated CEDAR's form-based acquisition process with the Dataset Submission Tool, so that, when DST users choose to populate metadata for a particular type of submission, they are presented with the relevant CEDAR-generated form. The DST monitors CEDAR every minute for new forms, importing any new metadata as they are entered, so DST users can quickly see their metadata entry results confirmed in the DST control panel. The LINCS-DCIC pipeline was released in June 2018.

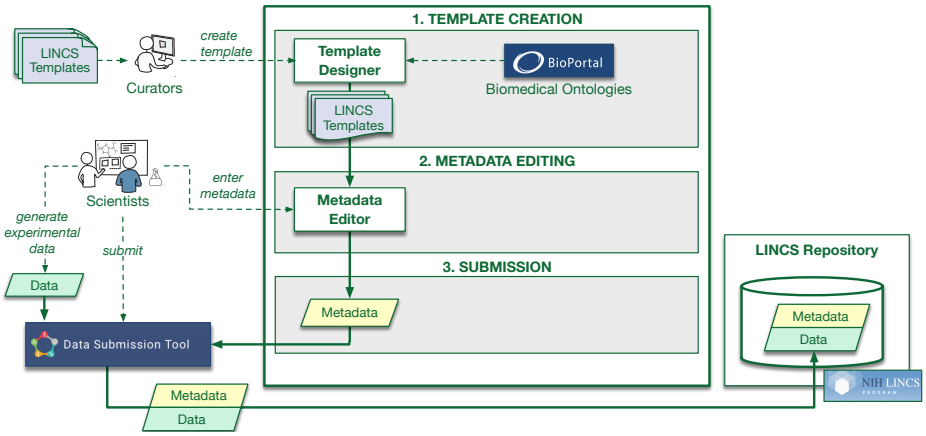

Figure 5. LINCS Metadata Submission Workflow. (1) *Create Template.* The LINCS-DCIC team created all the CEDAR templates describing the objects described by the LINCS community. (2) *Enter Metadata.* After the LINCS user clicks on a button in the LINCS-DCIC Data Submission Tool, the user redirected to a CEDAR page to enter their metadata. (3) *Submit Metadata.* The LINCS user's metadata are automatically harvested by the LINCS DST, which also manages the data submission performed by the user.

5 Discussion

There has been an emerging consensus that scientific data should be archived in open repositories, and that the data should be Findable, Accessible, Interoperable, and Reusable. To make experimental datasets FAIR, they must be accompanied by metadata that can explain what the data are about, under what conditions the data were collected, how the data are formatted, and the provenance of the data. Most online repositories are notorious for containing bad metadata, largely because these repositories allow their users the "freedom" to upload metadata that include arbitrary fields filled with arbitrary values—including missing values and typos.

There is a pressing need for solutions to help investigators to author more complete, more comprehensive, and more standardized metadata. The biomedical community is already making some progress in this direction. Domain-specific data-management tools are able to offer bespoke user interfaces that greatly ease the acquisition of high-quality metadata and that facilitate data exploration and analysis. While useful, these tools tend to be highly targeted to specific repositories and are not easily reusable. There is a need for a comprehensive technological approach to improve the authoring and management of metadata. This approach must target easy-to-use solutions that are generic (that is, not bespoke) to provide guidance over the entire life cycle of metadata—streamlining metadata creation as well as supporting metadata publication to third-party repositories.

In this paper, we outlined such a technological approach. The technology—called CEDAR—offers an example of such an all-purpose, end-to-end solution. CEDAR is a general-purpose system that assists the authoring of metadata to annotate experimental datasets and aims to simplify the submission of the datasets to online repositories. We explained how three major national activities used CEDAR's principled approach to develop community-specific pipelines for submitting biomedical metadata. We reviewed the specialized requirements that must be addressed when developing high-quality metadata submission pipelines, and described how CEDAR's flexible deployment options supported each project, enhancing particular needs and providing end-to-end metadata pipelines. In each case, CEDAR enabled intuitive metadata entry, while adding semantic precision and real-time validation. CEDAR metadata tools offer a rigorous and flexible choice for organizations that may not want to devote development time to providing custom-tailored metadata solutions.

Acknowledgments

CEDAR is supported by the National Institutes of Health through the NIH Big Data to Knowledge program under grant 1U54AI117925. NCBO is supported by the NIH Common Fund under grant U54HG004028. All software described in this paper is open source and available on GitHub (https://github.com/metadatacenter). We released a public version of CEDAR (https://cedar.metadatacenter.org) in April 2017.

References

1. Benson, D.A., Cavanaugh, M., et al.: GenBank. Nucleic Acids Res. 41, (2013).
2. Leinonen, R., Sugawara, H., Shumway, M.: The Sequence Read Archive. Nucleic Acids Res. 39, D19–D21 (2011).
3. Gonçalves, R.S., O'Connor, M.J., Martínez-Romero, M., et al.: Metadata in the BioSample online repository are impaired by numerous anomalies. In: Proceedings of 1st International Workshop on Enabling Open Semantic Science (SemSci 2017), co-located with ISWC 2017. pp. 39–46 (2017).
4. Rocca-Serra, P., Brandizi, M., Maguire, E., et al.: ISA software suite: Supporting standards-compliant experimental annotation and enabling curation at the community level. Bioinformatics. 26, 2354 (2010).
5. González-Beltrán, A., Maguire, E., Sansone, S.A., et al.: linkedISA: Semantic representation of ISA-Tab experimental metadata. BMC Bioinformatics. 15, (2014).
6. Wolstencroft, K., Owen, S., Horridge, M., et al.: RightField: Embedding ontology annotation in spreadsheets. Bioinformatics. 27, 2021–2022 (2011).
7. Shankar, R., Parkinson, H., et al.: Annotare - a tool for annotating high-throughput biomedical investigations and resulting data. Bioinformatics. 26, 2470–2471 (2010).
8. Musen, M.A., Bean, C.A., Cheung, K.H., et al.: The Center for Expanded Data Annotation and Retrieval. J. Am. Med. Informatics Assoc. 22, 1148–1152 (2015).
9. Egyedi, A.L., O'Connor, M.J., Martínez-Romero, M., et al.: Embracing Semantic Technology for Better Metadata Authoring in Biomedicine. In: Proceedings of the 10th International SWAT4HCLS Conference, Semantic Web Applications and Tools for Health Care and Life Sciences. pp. 1–10 (2018).
10. Martínez-Romero, M., O'Connor, M.J., Dorf, M., et al.: Supporting ontology-based standardization of biomedical metadata in the CEDAR Workbench. In: Proceedings of the Int Conf Biom Ont (ICBO) (2017).
11. Noy, N.F., Shah, N.H., Whetzel, P.L., et al.: BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. 37, W170–W173 (2009).
12. Rubelt, F., et al.: Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. Nat. Immunol. 18, 1274–1278 (2017).
13. National Institutes of Health: The Big Data to Knowledge Cloud Credits Model, https://commonfund.nih.gov/bd2k/cloudcredits.
14. Keenan, A.B., Jenkins, S.L., Jagodnik, K.M., et al.: The Library of Integrated Network-Based Cellular Signatures NIH Program: System-Level Cataloging of Human Cells Response to Perturbations. Cell Syst. 6, 13–24 (2018).
15. Breden, F., Luning Prak, E.T., Peters, B., et al.: Reproducibility and reuse of adaptive immune receptor repertoire data. Front. Immunol. 8, (2017).
16. Bukhari, S.A.C., O'Connor, M.J., Martínez-Romero, M., et al.: The CAIRR pipeline for submitting standards-compliant B and T cell receptor repertoire sequencing studies to the NCBI. Front. Immunol. 9, 1877 (2018).
17. Ohno-Machado, L., Sansone, S.A., Alter, G., et al.: Finding useful data across multiple biomedical data repositories using DataMed. Nat. Genet. 49, 816–819 (2017).
18. Sansone, S., Gonzalez-beltran, A., Rocca-serra, P., et al.: DATS : the data tag suite to enable discoverability of datasets. 1–11 (2017).