Ontology-Driven Metadata Enrichment for Genomic Datasets

Anna Bernasconi*[0000−0001−8016−5750], Arif Canakoglu*[0000−0003−4528−6586], Andrea Colombo, and Stefano Ceri[0000−0003−0671−2415]

Politecnico di Milano, Via Ponzio 34/5, 20133, Milano, Italy
{anna.bernasconi,arif.canakoglu,stefano.ceri}@polimi.it
andrea55.colombo@mail.polimi.it

Abstract. Data-driven genomic research requires accessing several repositories of genomic datasets, produced by international consortia, which provide open access to extremely valuable and well curated biological content. The associated metadata, describing experimental and biological conditions, are highly heterogeneous; consequently, dataset collection and integration is difficult – it requires data conversions and term matching which needs to be done by humans, with biological expertise.
In this paper, we present a method and tools for ontology-driven metadata enrichment. We select few relevant features which are provided by most repositories, and then we comparatively evaluate several search services providing ontological access, eventually associating each feature with the specific ontologies which are most suited to describe them. We also provide an expert validation of the approach. The method and tools are deployed in a large repository of open data, which will be soon available to the research community.

Keywords: Data Integration · Genomic Datasets · Metadata Annotation · Open Data · Bioinformatics.

1 Introduction

With the growth of diversity and complexity of scientific databases, the role of metadata – describing their content and data production process – is becoming more relevant. In particular, genomic computing often requires collecting datasets from multiple heterogeneous sources; unfortunately, metadata describing datasets across such sources are structured differently, they are often incompatible or incomplete. This raises a huge problem of data integration, which can be solved through ontological mediation, bridging the sources and enabling metadata interoperability.

In this paper, we describe metadata enrichment, which is the process of annotating existing structured metadata with ontological terms, their definitions, synonyms, ancestors, and descendants, to instrument a semantically enriched search of datasets linked to such metadata. Metadata enrichment is performed

* These two authors contributed equally to this work.

at the end of a data integration procedure for data loading, cleaning and mapping that is outside of the scope of this paper.

Metadata are converted to fit the format of a Genomic Conceptual Model (GCM, [2]), gathering the most important properties shared between heterogeneous sources. GCM is centered on the *item* entity (representing an experimental unit stored as a file of genomic regions) and organized as a four-pointed-star whose parts describe connected aspects about biology, technology, extraction, and management of the item.

Among all GCM attributes, we call "ontological" the ones that are present in all sources and require ontological agreement, thus are worthy of enrichment. These are: the *Platform* of items, i.e., the NGS platform used for sequencing, the *Ethnicity* and *Species* of donors, i.e., the individual of the organism from which the biological material is derived; the *Disease*, storing information about the pathology investigated with the sample; the *Tissue* and *CellLine* of samples, which distinguish the kind of biological material used for the experiment; the *Technique* or assay used to produce the genomic experiment (e.g., "Chip-seq", "miRNA-seq", "Genotyping Array"); the specific *Feature* or aspect described by the experiment (e.g., "Copy Number Variation","Histone Modification","Transcription Factor"); and the *Target* gene or protein of experiments (e.g., "CTCF", "MYC").

In this paper, we propose a metadata enrichment system, specific for genomic datasets, with a four-fold contribution: 1. description of the existing services to search ontologies related to biomedical content; 2. scoring and selection of the service and ontologies most relevant for our data; 3. organization of ontological knowledge in a well-structured taxonomy; 4. production of a tool for ontological annotations extraction. As a first integration effort, we include three important data sources used in the genomic research community, namely: Genomic Data Commons (GDC, [9]), with over 310,000 files covering aspects of cancer genomics; the Encyclopedia of DNA Elements (ENCODE, [6]), with almost 420,000 files related to functional DNA sequences and regulatory elements controlling gene expression; Roadmap Epigenomics Project (REP, [13]) containing around 2,000 datasets related to genetic variation.

The paper is structured as follows. Section 2 presents our solution to the problem of selecting appropriate search services and ontologies to annotate metadata. Section 3 describes how the enrichment procedure works and how we validated the process. Section 4 overviews related work, and finally Section 5 concludes the paper.

2 Search Service and Ontology Selection

First, we present the four most used and well-known ontology search services in literature (see Section 2.1), and how we score them (see Section 2.2) in order to select the most appropriate search service for our project (see Section 2.3). Next, we compare the ontologies provided by that search service, and select the specific ontology that is most suitable to annotate values for each ontological attribute (see Section 2.4).

2.1 Ontology Search Services

Ontological access to genomic data is well supported by several search services, which are capable in turn to integrate a high number of ontologies. Therefore, we are initially concerned in choosing the best search service, that will then be used within our system as broker to the underlying ontologies. We consider four different search services, which appear suitable for our purpose.

BioPortal [19] is a repository of biomedical ontologies and terminologies whose access is provided through a Web portal and Web services. We exploit its *term search* service, an endpoint which takes a free text input and provides a result in json format, listing a (configurable) number of annotations to ontological terms, showing different degrees of matching with the free text. These can be considered as possible annotations for the input text. Each term is identified by the pair $\langle ontology, id \rangle$, describing the code which references the ontology inside the BioPortal system and an identification number which references the term inside the ontology. A term also contains a single preferred label and its synonyms. An annotation is composed by a term and a match type: "PREF" if the match with the term is established with the preferred label or "SYN" if the match is with one of the term synonyms.

Ontology Recommender [15] is a BioPortal service that receives a free text or a list of keywords and suggests a set of ontologies appropriate for annotating the indicated terms, considered all together. The structure of annotations is identical to BioPortal's. Additionally, Recommender provides four scores that reflect how well the ontology (set) annotates the input data: *Coverage*, measures with which extent the ontology represents the input; *Acceptance*, indicates how well-known and trusted the ontology is by the biomedical community; *Detail*, shows the level of specification provided by the ontology for the input data; *Specialization*, indicates how specialized the ontology is w.r.t the input data domain.

Ontology Lookup Service (OLS, [12]) provides ontology search, visualization, and ontology-based services. The accepted input is a keyword, the provided result is a list of annotations, similar to the other services but not including a match type. In the API request, a *fieldList* parameter can be used to specify the specific elements to be included in the output along with other formatting preferences.

Zooma[1] is a service from OLS which provides mappings between textual input and a manually curated repository of text-to-ontology-term mappings. If no mappings are found, it uses the basic OLS search. In addition to the usual annotation information, Zooma also returns a confidence label associated to the annotation, ranging from HIGH to LOW.

We exclude other important ontology search portals such as HeTOP [8] and UMLS [3], as they are more focused on multilingual support and medical terminologies, therefore do not include many ontologies that are important to annotate our values. Also the NCBO Annotator [10] is not considered since its functionalities are completely covered by the Ontology Recommender.

[1] https://www.ebi.ac.uk/spot/zooma/

2.2 Scoring

Every search service provides a search API, which is repeatedly used for the score evaluation. For each API call we store: the used service; the attribute from GCM characterizing the values (the "type" of the values); the original *raw* value deriving from the GCM, imported through the mapping phase; possible *parsed* values deriving from a simple syntactic pre-processing of *raw* values (e.g., removal of punctuation, split of long expressions...); the ⟨`ontology`,`ontology_id`⟩ pair, uniquely identifying an ontological term in a service; `pref_label` and `synonym`, respectively the textual expression primarily used for the term and its alternative versions; `score`, textual information regarding the goodness of a match, directly retrieved from the services, if available.

In total, we performed 1,783 API calls to each of the four services, corresponding to 1,299 original values to be enriched; some of these were splitted during a pre-processing phase. As a result, we retrieved 1,783 interesting matches from BioPortal, 885 from Recommender, 1,782 from OLS, and 1,779 from ZOOMA, all of which were used for the following processing after calculating our scores.

Starting from the retrieved information, we calculate the `match_score` as a measure of how well a term matches a value, by using a scoring system that is specifically designed for the task, which is next described. The general formula returning the `match_score` value, shown in Eq. 1, subtracts from an initial maximum number (10, when there is a perfect match with a `pref_label`, 9 with a `synonym`) a penalty measuring how the raw value differs from the label retrieved from the services:

$$match\_score(raw, label) = \{10, 9\} - distance(raw, label) \tag{1}$$

To compute the distance, we use a modified version of Needleman-Wunsch algorithm [16], a protein and nucleotide sequence alignment algorithm which is widely used in bioinformatics. In the original algorithm, the input is represented by two strings whose letters need to be aligned. The letters may have a "match", a "mismatch" or an "indel" (i.e., adding a gap in one of the strings). In our modified version, we define each word as a distinct letter of the original algorithm and we add another type of *mismatch*, i.e., the *swap*. All in all, the total distance is calculated as a sum of distances between words:

- **Match:** Two words are the same, then their distance is 0
- **Mismatch:** Two words are different, then their distance is 2.5
- **Swap:** Two consecutive words traded places, then their distance is 0.5
- **Delete:** One word is deleted from the *raw*, then their distance is 2
- **Insert:** A new word is added to the *raw* then their distance distance is 1

The indicated distance values are chosen in such a way that the number of deletions is minimized (i.e., we penalize a *label* which does not include a word present in *raw*) and the swap is preferred to indel and mismatch. For example, for the *raw* "breast invasive carcinoma", the *label* "invasive breast carcinoma" (i.e., one *swap*) is considered better than "breast carcinoma" (i.e., one *deletion*).

Additional calculated scores are: `onto_suitability`, a measure of how much an ontology is adequate for a given attribute, calculated as the average

`match_score` over all *raw* values for that attribute; `onto_acceptance`, a measure of how well-known and trusted the ontology is by the biomedical community, computed through Recommender Web Services [15][2]; the `overall_score`, obtained by multiplying each *raw* value `match_score` by a weighted average of the two measures typical of the ontology.

2.3 Service Evaluation

Table 1 describes the obtained results. The "Service Properties" part contains an overview of service properties. BioPortal and Recommender provide a *match_type* (MT) in their APIs response, which means that they specify if the input text is more similar to the preferred label rather than to one of the synonyms associated to a term. Recommender offers the additional function of searching for multiple key-words at the same time (MK) and consequently suggests a minimal set of ontologies suitable for annotating the maximum possible number of key-words. This function is also offered by ZOOMA which, however, in practice just performs multiple single key-word requests and lists all results at the same time. Only Recommender executes a good attempt of annotating free texts (FT). BioPortal's set of ontologies is much broader than OLS' since minor efforts are also included. ZOOMA exploits search results from OLS but also provides results coming from previous manual curation works as an additional service to the user.

Table 1. Summary of Ontology Search Services as of October 1, 2018

		BioPortal	Recommender	OLS	ZOOMA
Service Properties	Search properties	MT	MT,MK,FT	-	MK
	Num. of ontologies	728	728	214	214
	Previous curation	no	no	no	yes
Example scoring of "cervical adenocarcinoma"	1st best match	ncit_c4029	ncit_c4029	ncit_c4029	efo_0001416
	2nd best match	efo_0001416	None	efo_0001416	None
	3rd best match	doid_3702	None	ncit_c136651	None
	Occurrence score	1	0.5	1	0.5
	Coverage score	1	1	1	1
Aggregated scores	Occurrence	83.17%	46.97%	**90.54%**	75.96%
	Coverage	100.00%	49.88%	**99.94%**	99.78%

The "Example scoring" part contains an example of how services are rewarded based on the matching terms they find. To evaluate the match, we use the `overall_score` described in Section 2.2. When the disease-related text "cervical adenocarcinoma" is searched, BioPortal suggests, on top of others, the three terms "ncit_c4029", "efo_0001416", and "doid_3702", while Recommender just

[2] It is derived from the number of visits to the ontology page in BioPortal and the presence or absence of the ontology in UMLS [3].

provides one result, "ncit_c4029". Our algorithm for *Occurrence* computes the set of terms which occur the highest amount of times in the top three matches of the services (in this case ["ncit_c4029","efo_0001416"]) and assigns a weighted reward (1 if the set only contains one entry, 0.5 if it contains 2, and so on) to the services which include that term in the top results. Indeed BioPortal scores 1 since it contains both top results, while Recommender scores 0.5 since it contains just one. *Coverage* is 1 when the service provides at least one result, 0 otherwise.

We use as scores for service selection the average *Occurrence* and *Coverage* over all the searched raw values. On this basis, OLS is selected as the best suited search service to pursue the enrichment annotations in our system.

2.4 Ontology Selection

Based on the `overall_score` described in Section 2.2, we also aggregate results over specific attributes and ontologies. This calculation produces, as a result, one top ontology for each attribute. Since most of the times only one ontology does not provide an acceptable coverage for all the values belonging to that attribute, we use an algorithm to compute a small set of ontologies to annotate values from an attribute. Such algorithm first tries to match values only with the first ontology, then tries to match only the ones left unmatched with the following ontologies, until a fixed point for coverage is found. If the computational costs become too high, the algorithm can be stopped at a predefined threshold coverage, considered acceptable. In our case we set the threshold equal to 95%.

The resulting choice of ontolgies sets is: OBI for *Platform*, NCIT for *Ethnicity*, NCBITaxon for *Species*, NCIT for *Disease*, UBERON for *Tissue*, {EFO,CL} for *CellLine*, NCIT for *Feature*, and OGG for *Target*. All the above choices meet the set threshold. Our best choice for the attribute *Technique* is the set {OBI,EFO}, but for this attribute we are not able to achieve the coverage threshold, as we reach a best coverage of 85.7%.

3 Metadata Enrichment

After selecting such sets, we proceed with the enrichment of the values contained in the ontological attributes of the GCM. Section 3.1 presents the relational schema which supports this phase. Section 3.2 describes the enrichment process and Section 3.3 shows how the automatic annotation is aided by curators intervention. Finally, Section 3.4 overviews the expert validation.

3.1 Relational Schema

Figure 1 describes the logic schema of the relational database. The Genomic Conceptual Model frame contains the tables from the GCM (of which we only show in detail the ones which have ontological attributes). The Local Knowledge Base (LKB) frame stores all the information retrieved from OLS services and relevant to annotate our values. The main tables are: `vocabulary` (storing the reference term ids), `synonym` (containing synonyms of the preferred label in

the vocabulary), `reference` (identifiers of equivalent terms in alternative ontologies), `ontology` (dimension table for used ontologies), and `relationship` (representing links between terms in the ontology). The Expert Support frame contains the tables used to contain information for expert users. Each GCM ontological attribute X is equipped with a companion-attribute X_tid, which references the ontological term in the vocabulary table (e.g., *Platform* with value "Illumina Human Methylation 450" is associated to $Platform\_tid = 10$, representing the vocabulary object OBI_0001870, taken from the Ontology of Biomedical Investigations [1]). The Vocabulary table is the central entity of the LKB schema. The *tid* column is the primary key which is referenced by all other tables in LKB and from the tables in GCM. Also tables from the LKB and from the Expert tables are linked using *tid*s.

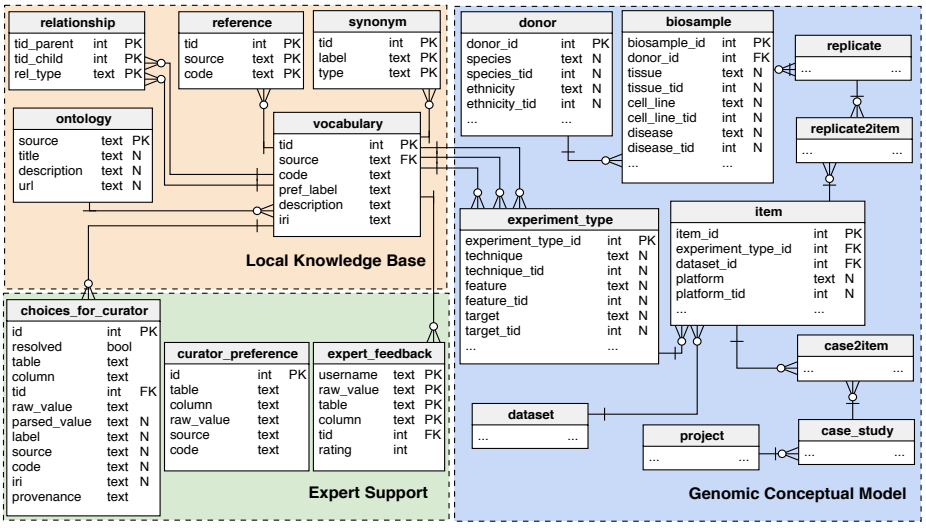

Fig. 1. Relational schema for tables of the GCM, LKB and user feedback routines.

3.2 Enrichment Process

To enrich the values contained in the ontological attributes of the GCM, we iterate over all values associated to a *_tid* column. For each value we call OLS services with the ontologies sets indicated in Section 2.4. When a best `match_score`, calculated as in Eq. 1, is found and is above the threshold 5.0, we select the corresponding term and proceed with the annotation, otherwise the decision is delegated to data curators (see Section 3.3).

Once the term has been selected, we populate the tables of the LKB with all the information derived from OLS regarding the term: description, iri, synonyms, xrefs, hypernyms and hyponyms (both of IS_A and PART_OF kinds). The depths

of ancestors and descendants retrieved from the ontology are configurable by constant specification. The automatic enrichment process currently annotates about 83% of the total raw values, while the remaining are handled using a manual curation procedure.

3.3 Biologists Support

We propose two procedures which allow experts curators to support the annotation algorithm; we assume them to be knowledgeable about biological data management and to be expert in genomic data curation.

In the first procedure, a curator can examine all cases in which the algorithm is not able to provide a high quality match (i.e., the service provides either partial matches with low score or no result). The low scores matches are proposed as suggestions so that the curator may select one of them. In any case, a manual annotation can always be provided. The procedure can be configured so that it also shows the cases with the same score.

The second procedure is started when a pre-existing annotation is not adequate (i.e., a *_tid* column has been filled with a wrong vocabulary term). In this case, the curator can invalidate the annotation and provide an alternative.

3.4 Expert Validation

We conducted a validation by engaging six experts with good biological knowledge. For each considered attribute, we presented to them a set of annotations (i.e., matches between a *raw* and an ontological term, equipped with its descriptions) automatically produced by the enrichment procedure. We asked them to rate the associations according to how accurate they are w.r.t their knowledge.

Table 2. Expert Validation results

Attribute	*Platform*	*Ethnicity*	*Species*	*Disease*	*Tissue*	*CellLine*	*Technique*	*Feature*	*Target*
#annot/total	3/4	20/33	4/4	76/97	82/121	191/282	10/14	9/22	738/787
EXACT	5.56%	67.50%	100.00%	64.17%	88.33%	84.17%	73.33%	61.11%	100.00%
GOOD	27.78%	6.67%	-	6.67%	6.67%	4.17%	8.33%	9.26%	-
ACCEPTABLE	50.00%	19.17%	-	10.83%	2.50%	4.17%	6.67%	12.96%	-
WRONG	16.67%	5.83%	-	17.50%	1.67%	6.67%	6.67%	14.81%	-
DO NOT KNOW	-	0.83%	-	0.83%	0.83%	0.83%	5.00%	1.85%	-

The questionnaire contains up to 20 matches for each attribute (or less in the case of *Platform*, *Species*, *Technique*, and *Feature*, which contain less found matches), selected randomly from their value pools, therefore considered representative of the sets. The test allows five choices: 1. EXACT, 2. GOOD, 3. ACCEPTABLE, 4. WRONG, 5. DO NOT KNOW.

In Table 2, in the first row we indicate, for each attribute, the ratio between the number of automatically annotated values and the number of their total distinct values. Then, we show in detail the results from the attributes presented

to experts. The averaged results highlight that in 83.06% of cases the experts marked as exact or good the examined matches, in 8.81% they rated them as acceptable, and only the remaining 7.05% were marked as wrong. In the 1.08% of cases the experts declared they were not able to evaluate the match.

4 Related Works

Many works in the literature consider the problem of recognizing ontological concepts to perform semantic annotation of data. For example: Bodenreider [4] proposes a (dated) survey on the use of ontologies in biomedical data management and integration; the works [11, 18] debate solutions devoted to data integration; Giles et al. [7] focus on concept extraction from datasets of a specific source; the works [14, 5] consider the problem of metadata authoring by using BioPortal ontology-based recommendations, with a focus on metadata manual creation and preparation. A number of articles have addressed the problem of choosing ontologies for semantic enrichment. Among these: Wilkinson et al. [20] present the *FAIR* principles, which define a set of characteristics that data resources and infrastructures should exhibit; [17] identify key search factors for biomedical ontologies to help biomedical experts in selecting the best-suited ones in their search cases. In Section 2.1 we presented BioPortal [19], Ontology Recommender [15], Ontology Lookup Service [12] and Zooma, since we believe UMLS [3], HeTop [8], and Annotator [10] were not suited for our purpose.

5 Conclusion and Future Work

Annotating metadata with terms from ontologies and providing an expansion to hypernyms and hyponyms allows for easier and semantically flexible dataset search. We provide selection criteria for choosing among search services and ontologies, and a user-friendly process for assisting biologists in checking that suggested terms are indeed acceptable. We also provided an internal validation of annotations produced by our process. As future work, we intend to improve the matching algorithm by exploiting the ontology structures and information. We plan to integrate more sources and test our method on a comprehensive database.

The implementation of the metadata enrichment system described in this paper is is available at: `https://github.com/DEIB-GECO/Metadata-Enricher`. It is used in the broader context of a genomic repository, developed within the GeCo Project[3], which will be available for use in the near future. Enriched metadata help users in locating datasets for genomic data extraction and analysis, either on their original sources or within our repository.

Acknowledgment

This research is funded by the ERC Advanced Grant 693174 GeCo (data-driven Genomic Computing).

[3] http://www.bioinformatics.deib.polimi.it/geco/

References

1. Bandrowski, A., et al.: The ontology for biomedical investigations. PloS one **11**(4), e0154556 (2016)
2. Bernasconi, A., et al.: Conceptual modeling for genomics: Building an integrated repository of open data. In: Mayr, H.C., Guizzardi, G., Ma, H., Pastor, O. (eds.) Conceptual Modeling. pp. 325–339. Springer International Publishing, Cham (2017)
3. Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), D267–D270 (2004)
4. Bodenreider, O.: Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearbook of Medical Informatics p. 67 (2008)
5. Egyedi, A.L., et al.: Embracing semantic technology for better metadata authoring in biomedicine. In: Proceedings of SWAT4LS International Conference 2017 (2017)
6. Consortium ENCODE: An integrated encyclopedia of DNA elements in the human genome. Nature **489**(7414), 57–74 (2012)
7. Giles, C.B., et al.: Ale: automated label extraction from GEO metadata. BMC Bioinformatics **18**(14), 509 (2017)
8. Grosjean, J., et al.: Health multi-terminology portal: a semantic added-value for patient safety. Studies in health technology and informatics **166**, 129 (2011)
9. Jensen, M.A., et al.: The NCI Genomic Data Commons as an engine for precision medicine. Blood **130**(4), 453–459 (2017)
10. Jonquet, C., Shah, N., Musen, M.: The open biomedical annotator. In: AMIA Summit on Translational Bioinformatics. pp. 56–60 (2009)
11. Jonquet, C., et al.: A system for ontology-based annotation of biomedical data. In: International Workshop on Data Integration in The Life Sciences. pp. 144–152. Springer (2008)
12. Jupp, S., et al.: A new Ontology Lookup Service at EMBL-EBI. In: Malone, J., et al. (eds.) Proceedings of SWAT4LS International Conference 2015 (2015)
13. Kundaje, A., et al.: Integrative analysis of 111 reference human epigenomes. Nature **518**(7539), 317–330 (2015)
14. Martínez-Romero, M., et al.: Fast and accurate metadata authoring using ontology-based recommendations. In: AMIA Annual Symposium Proceedings. vol. 2017, p. 1272. American Medical Informatics Association (2017)
15. Martínez-Romero, M., et al.: NCBO Ontology Recommender 2.0: an enhanced approach for biomedical ontology recommendation. Journal of Biomedical Semantics **8**(1), 21 (2017)
16. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology **48**(3), 443–453 (1970)
17. Oliveira, D., et al.: Where to search top-k biomedical ontologies? Briefings in Bioinformatics p. bby015 (2018)
18. Shah, N.H., et al.: Ontology-driven indexing of public datasets for translational bioinformatics. BMC Bioinformatics **10**(2), S1 (2009)
19. Whetzel, P.L., et al.: Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. Nucleic Acids Research **39**(suppl_2), W541–W545 (2011)
20. Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. Scientific Data **3** (2016)