Context Aware Recommendation Engine for Metadata Submission

Maryam Panahiazar, Michel Dumontier and Olivier Gevaert

Stanford Center for Biomedical Informatics Research,
School of Medicine, Stanford University

{marypan, michel.dumontier, olivier.gevaert}@stanford.edu

ABSTRACT

While good metadata is essential in finding, interpreting, and reusing data, the authoring of metadata is considered highly tedious and often incomplete. It is imperative to make the authoring of metadata a manageable task. Towards easing the burden of authoring high quality metadata, we have developed a data-driven framework that learns association between data elements to suggest context-sensitive metadata values. We demonstrate our framework in the context of microarray annotations from the Gene Expression Omnibus (GEO).

Keywords

metadata, recommendation engine, data submission

1. INTRODUCTION

Recent years have seen a groundswell of efforts to develop guidelines and checklists for metadata that describe biomedical experiments. Biomedical organizations such as the Global Alliance for Genomics and Health [12], and FORCE11 [6], and more general associations such as the Research Data Alliance [10], work to evangelize the essential role that metadata plays in data sharing. Activities to collect and define community-driven standards [13], such as BioSharing [1], offer important resources not only to biomedical researchers, but also to journal editors and to biomedical curators who seek guidance regarding which standards to use [11].

Despite these initiatives, the authoring of metadata remains an ad-hoc task that produces results that are often incompatible with one another. In an era when data are generated at rates and in quantities never before imaginable, there is an urgent need to understand the structure of datasets, the experimental conditions under which they were produced, and the information that other investigators may need to make sense of the data [3]. The ultimate Big Data challenge lies not in the data, but in the metadata; the machine-readable descriptions that provide data about the data. It is not enough to simply publish the data online; data are not usable until they can be “explained” in a manner that both humans and computers can process[9].

For instance, Gene Expression Omnibus (GEO) is one of the largest, best-known biomedical databases [2]. GEO is a public Functional Genomics data repository supporting *Minimum information about a microarray experiment (MIAME)* [5] compliant data submissions. The current GEO search interface relies upon very broad terms (e.g. ”influenza”) as input and often returns hundreds of datasets.

The current situation requires users to review each dataset to confirm appropriateness which is a time consuming job and limits the re-usability of data collected in public repositories. Moreover, because the task of data submission itself is time consuming, there is a possibility of carelessness regarding adding the metadata during submission. It takes time and effort to create well-specified metadata, and investigators view the task of metadata authoring (or data annotation) to be a burden that may benefit other scientists, but not the team that did the work in the first place. There is no verification for the correctness of metadata during submission. Because of this negligence the valuable information may not be re-usable by other researchers.

The conundrum hereby is that without comprehensive, normalized metadata, datasets are not easily comparable and hence are difficult to retrieve and to suggest. Simply allowing annotators to enter any description of their choice does not alleviate the problem, if the annotators do not follow similar naming conventions. However, imposing a compliance requirement with a controlled vocabulary not only deters annotators from making the extra effort of finding the “correct” annotation, it is also likely that the vocabulary is not comprehensive. Allowing annotator to use ontologies of their choice may simply shift the normalization problem.

Our solution is to let people add descriptions of their choice, but make suggestions based on context and corpus statistics that help the convergence toward a standard vocabulary that can then potentially be more easily mapped to a controlled vocabulary. It has been shown that, at least in the case of collaborative tagging, the community will con-

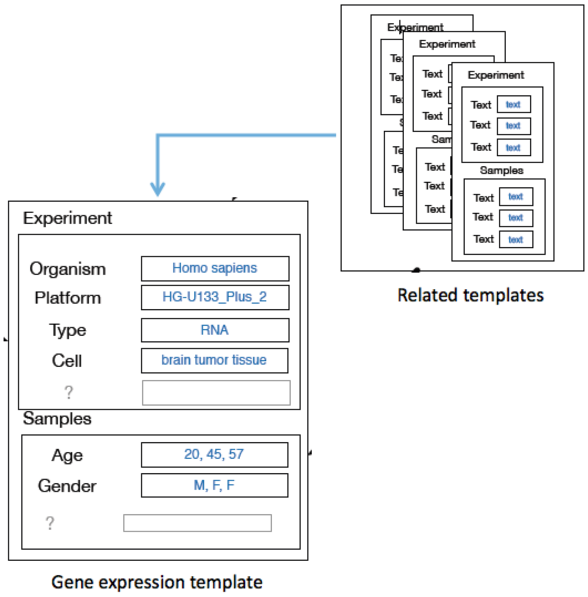


Figure 1: Learning matadata annotation from GEO dataset, we use the existing metadata as a source for value set suggestions.

verge around a power law distribution of tags for a topic [8] and each topic will eventually be well-represented by a small number of highly representative tags.

This work is part of The Center for Expanded Data Annotation and Retrieval (CEDAR). CEDAR provides the technology both for making it easier, faster, and less error-prone for scientists to populate metadata templates and for making the metadata themselves computable, reusable, and shareable [9]. With the context-aware recommendation engine we can recommend possible values for the metadata elements during the submission. These value suggestions are based on the more frequent values which have been used in metadata repository. To this end we analyze the metadata of GEO data repository to investigate most common value sets in a large, well-known biomedical database. We adopt a data-driven approach to the development of value set suggestions by using real time statistical techniques over GEO metadata. Figure 1 shows an example of the template authoring that helps to suggest value sets based on existing metadata.

We use this data-driven approach to provide "autocomplete" capabilities to select the most appropriate value for metadata elements from recommended values. These recommendations are dynamically changing the suggestions based on previously entered metadata elements. The contextual recommendation also helps in normalizing the metadata values, because more common values are ranked higher. Therefore, the data submitter can use these suggestions to fill the submission form more effectively. Data seekers, in turn will be able to more effectively explore, query and browse big data for their own use. Data become accessible, verifiable, and reusable. The big winner, of course, is science itself.

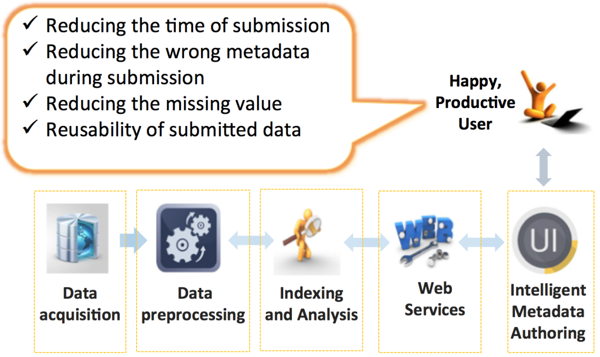

✓ Reducing the time of submission
✓ Reducing the wrong metadata during submission
✓ Reducing the missing value
✓ Reusability of submitted data
Happy, Productive User
Data acquisition
Data preprocessing
Indexing and Analysis
Web Services
Intelligent Metadata Authoring
UI

Figure 2: Components of the System; data acquisition, data preprocessing, data indexing and analysis, web services and user interface.

In the next section we will describe the architecture of the system in terms of its components including data data acquisition, data preprocessing, data indexing and analysis, web services and user interface. Then we discuss the future plan for expert validation and how we use this study as a foundation for the more intelligent recommendation engine. Then we conclude our discussion with the limitations of the approach and the impact of this study for the community.

2. ARCHITECTURE AND TOOL DETAILS

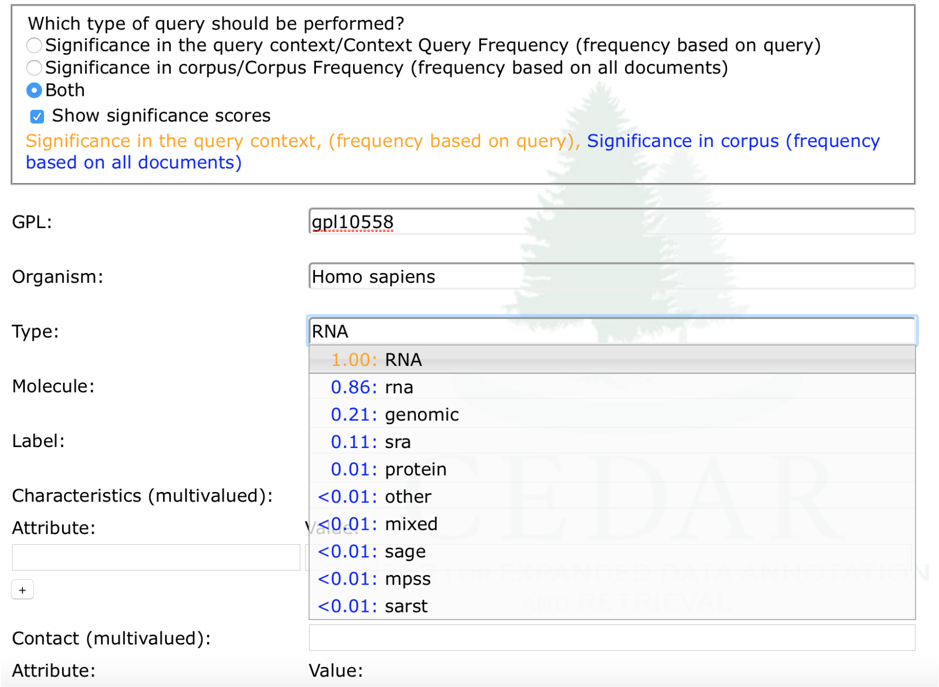
CEDAR's context-aware recommendation engine facilitates metadata submission by suggesting possible value sets for the selected elements based on existing metadata in the repository. Contextual metadata information can be useful for providing better recommendations. The context of information can range from contact information of the data submitter to the type of study. This system is designed around a basic workflow consisting of 5 steps:

1. data acquisition
2. data preprocessing
3. indexing and analysis
4. web service
5. metadata authoring user interface

Figure 2 illustrates the major components of the system. We suggest the possible values for metadata elements in a friendly searchable user interface as an autocomplete-suggestion to make the process of metadata submission faster and less error-prone for the metadata submitter. We will explain the submission workflow as a use case in the user interface section. In the following we explain each individual component of this system in detail.

2.1 Data Acquisition

We downloaded the structured metadata of the GEO repository to have a large well-known biomedical database as a real-world use case for our system [2]. Data submitters to the GEO repositories are required to submit their data in the form of series (GSExxxx), samples (GSMxxxx), and platforms (GPLxxxx). Each series describes the overall study including the samples of the studies and the corresponding



Which type of query should be performed?

○ Significance in the query context/Context Query Frequency (frequency based on query)

○ Significance in corpus/Corpus Frequency (frequency based on all documents)

● Both

☑ Show significance scores

Significance in the query context, (frequency based on query), Significance in corpus (frequency based on all documents)

GPL: gpl10558

Organism: Homo sapiens

Type: RNA

Molecule:

Label:

Characteristics (multivalued):

Attribute: Value:

Contact (multivalued):

Attribute: Value:

Figure 3: Metadata Authoring User Interface

platforms for the samples. We used publicly available data as of May 2015 totalling 55951 series, where each series on average includes 25 samples. The dataset includes 1,366,559 records. We extracted all corresponding metadata elements for all studies including title, sample id, series id, status, submission data, last update, type, sources name, organism, characteristics, molecular, label, treatment, supplementary file link, contact, and more. Here, we discuss some of these elements which are more commonly provided. This includes protocol, organism, type, molecule, label, characteristics and contact information.

To download GEO, we used the *GEOmetadb* package in R [14] The *GEOmetadb* package is an attempt to make querying the metadata describing microarray experiments both easier and more powerful. At the heart of *GEOmetadb* is a SQLite database that stores all the metadata associated with all GEO data types including GEO samples (GSM), GEO platforms (GPL), GEO data series (GSE), and curated GEO datasets (GDS), as well as the relationships between these data types. The entire GEO metadata was accessible with simple SQL-based queries. The database was exported to one single JSON [4] file for the next steps.

2.2 Preprocessing

Some of the fields in the GEO metadata inherently contain semi-structured content, even though they are represented as unstructured data. For example, the *address* field contains information for name, zip, phone number, etc., the *characteristics* field contains attribute-value pairs for *sex*, *age*, *tissue*, etc. The problem is that these useful semistructured data are not just disguised as unstructured data, but the delimiters that separate each attribute-value pair tends to change between different records.

In the preprocessing step we split these unstructured fields into semi-structured key-value pairs. This is made difficult not only by the above mentioned non-uniformity of the delimiters; some annotators only enter the description values without attribute names. Moreover, annotators use different conventions for descriptions creating heterogeneity in the annotations. For example, for the patient's sex, some annotators use "sex : male/female", others use "sex:M/F" or "sex (male-0, female-1):0/1" or even 'sex (male-1, female-2):1/2". Each identified attribute-value pair is added to the JSON object for further processing.

2.3 Indexing and Analysis

Indexing is performed using Elasticsearch [7], Elasticsearch is distributed full-text search engine based on the Lucene Information Retrieval library. It is accessible through a RESTful web interface and communicates using schema-free JSON documents. It is developed in Java and is released as open source under the terms of the Apache License. Indexing is performed by bulk-uploading the converted and preprocessed JSON records to Elasticsearch, which then analyzes the different fields in the data depending on the index description. For example, the *Contact* field is tokenized and stemmed before indexing whereas the *molecule* field is not tokenized and hence treated as a keyword-field.

2.4 Web Service

We query ElasticSearch using its RESTful API with JSON over HTTP, through a web client. ElasticSearch provides official clients for several languages such as JavaScript, .NET, PHP, Perl, Python, and Ruby [7]. We configured the JavaScript JQuery package to communicate with ElasticSearch.

2.5 Metadata Authoring User Interface

We propose an approach for identifying most appropriate value sets for selected metadata elements. Values for metadata elements are suggested and ranked based on previously entered values, defined as the query context. We use this representation to guide an adaptive user interface for template authoring. Our aim is focused on using an intuitive entry-mechanism that not only makes the authoring as natural as possible, but also uses the existing metadata in the repository to pre-fill much of the metadata, providing auto-complete capabilities, selection from value sets, and dynamic queries.

To help the data submitter, we offer the user different options depending on the type of recommendation support he/she wants. Figure 3 illustrates a demo of the UI explaining our use case. The user has the choice of having the metadata value suggestions scored and ranked by either significance of the result in the query context for the context-based query, by significance of the result or the full corpus, independent of any context or by both recommendation measures. Once the user clicks on a metadata text box, a ranked list of values is displayed. When no context is available, the ranking is based only on corpus statistics. If one or more metadata fields are already filled, the ranking of suggested items will reflect the likelihood that the suggested item occurs together with the already entered metadata items.

As the user starts typing, the context-ranked list changes to only match the characters that were typed. This process will reduce the burden of entering metadata terms and significantly shorten the time that is needed for investigations to enter metadata. We will use the increasing amounts of structured metadata to learn from that. This incorporation of structural knowledge into the learning technology will allow us to infer common metadata patterns and their value sets of the context of an investigator, technology platform, organism, or sample type. Our key goal is to facilitate as much of the metadata submission process as possible, by suggesting possible value sets for the fields in templates based on available data.

3. CONCLUSION

We propose a contextual interface to recommend metadata. The suggestions are given using Information Retrieval (IR) methods, i.e. the contextual significance is computed online and does not rely on precomputed correlation scores and hence does not need to be re-trained when new data is added. Suggesting the value sets from existing metadata will help ensure that template authors do not miss important value sets that appear frequently in the data. Thus, the data submitter will be able to be assisted in finding the terms they need, thereby making it easier to use a common vocabulary and hence improving the quality of metadata. Moreover, when investigators have previously entered data into repositories, investigator-specific learning will allow predicting the metadata entries based on previous behavior.

We can start by confirming that the investigator has collected all the information that he/she will need to populate the metadata, such as protocol, type, organism, contact information, characteristics, and more. Finally, the submitter will be able to specify the type of experiments for each value set if that is valid or not. If we have the information on the type of experiment that a scientist has performed, we will be able to present her with a set of value suggestions that are most likely appropriate to describe to describe her experiment. This process makes metadata submission easier, faster, and less error-prone for scientists to populate metadata templates and for making the metadata themselves comparable, computable, reusable, and sharable.

4. FUTURE WORK

We will adopt an approach to the development of value sets recommendation using item set suggestion to infer metadata patterns using diverse sources of information. In summary we will use the metadata that investigators have already defined the existing data deposited in data repositories in big data platform to infer metadata attributes that are more commonly used and the common values for those attributes.

We will extend the current online contextual recommendation computation to also incorporate more in-depth knowledge of correlations between annotation values in the dataset in order to make better recommendations. Specifically, we will train models that can identify degrees of correlation between metadata items using distributional methods. Specifically, we want to ascertain how high the correlation is between attribute-value pairs in the characteristics field. Not only does this strengthen the positive associations that our current recommendation engine provides, it also allows us to point out possible errors in the annotation. Finally, once our models are sufficiently accurate enough, this technique will also be used to suggest or even automatically fill in missing values.

We will evaluate the effectiveness of our methods in three ways. First, the prediction quality of the model can be computed over the full dataset by leaving an element out of a value set and determining how high the prediction accuracy based on the remaining items is. Second, we will present experts with the suggested value sets for a different set of fields and ask them to evaluate whether or not the suggestions are useful.

This qualitative evaluation allows us to see if we made predictions that do not match the gold standard but are nevertheless correct. Third, we will monitor the adoption of the suggestions by metadata submitters and record their choice. We will measure the acceptance rate for the suggestions, focusing on how many of the submitters follow the suggestions in their entirely, how many take some elements from the suggested values, and how many expand it. We will use this feedback to improve our system further.

5. ACKNOWLEDGEMENT

Research reported in this publication is supported by the National Institute of Allergy and Infectious Diseases of the National Institute of Health under Award Number U54 AI11792. The content is solely the responsibility of the authors and does not necessary represent the official views of the National Institute of Health.

6. COMPETING INTERESTS

None.

7. REFERENCES

[1] The BioSharing Registry : connecting data policies , standards & databases in life sciences. *Proposed WG to the RDA - Case statement*, pages 1–15, 2014.

[2] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. a. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva. NCBI GEO: archive for functional genomics data sets–update. *Nucleic acids research*, 41(Database issue):D991–5, Jan. 2013.

[3] C. L. Borgman. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6):1059–1078, 2012.

[4] T. Bray. The JavaScript Object Notation (JSON) Data Interchange Format. In *Internet Engineering Task Force RFC7159, ISSN: 2070-1721*, 2014.

[5] A. Brazma, P. Hingamp, and J. Quackenbush. Minimum information about a microarray experiment (MIAME) - toward standards for microarray data. *Nature*, 29(December):365–371, 2001.

[6] T. Clark, R. Dale, A. D. Waard, I. Herman, E. H. Hovy, D. Shotton, A. Birukou, J. A. Blake, P. E. Bourne, S. Buck, L. Chan, O. Chiarcos, P. Ciccarese, T. Clark, R. Dale, A. D. Liddo, D. D. Roure, A. D. Waard, S. Decker, G. Castro, C. Goble, E. Gray, P. Groth, U. Hahn, I. Herman, E. H. Hovy, M. J. Kurtz, F. Murphy, C. Neylon, S. Pettifer, M. W. Rogers, D. Shotton, and J. Siren. Force11 White Paper : Improving The Future of Research Communications and e-Scholarship. 11, 2012.

[7] R. Gheorghe, M. Hinman, and R. Russo. Elasticsearch in Action. *Manning Publications Co*, 2015.

[8] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. *n Proceedings of the 16th international conference on World Wide Web (WWW '07). ACM, New York, NY, USA,*, pages 211–220.

[9] M. A. Musen, C. A. Bean, K.-H. Cheung, M. Dumontier, K. A. Durante, O. Gevaert, A. Gonzalez-Beltran, P. Khatri, S. H. Kleinstein, M. J. O'Connor, Y. Pouliot, P. Rocca-Serra, S.-A. Sansone, and J. A. Wiser. The Center for Expanded Data Annotation and Retrieval. *Journal of the American Medical Informatics Association : JAMIA*, (650):1–6, June 2015.

[10] M. A. Parsons and F. Berman. The Research Data Alliance : Implementing the Technology , Practice and Connections of a Data Infrastructure. pages 33–36, 2013.

[11] S.-A. Sansone, P. Rocca-Serra, D. Field, E. Maguire, C. Taylor, O. Hofmann, H. Fang, S. Neumann, W. Tong, L. Amaral-Zettler, K. Begley, T. Booth, L. Bougueleret, G. Burns, B. Chapman, T. Clark, L.-A. Coleman, J. Copeland, S. Das, A. de Daruvar, P. de Matos, I. Dix, S. Edmunds, C. T. Evelo, M. J. Forster, P. Gaudet, J. Gilbert, C. Goble, J. L. Griffin, D. Jacob, J. Kleinjans, L. Harland, K. Haug, H. Hermjakob, S. J. H. Sui, A. Laederach, S. Liang, S. Marshall, A. McGrath, E. Merrill, D. Reilly, M. Roux, C. E. Shamu, C. A. Shang, C. Steinbeck, A. Trefethen, B. Williams-Jones, K. Wolstencroft, I. Xenarios, and W. Hide. Toward interoperable bioscience data. *Nature Genetics*, 44(2):121–126, 2012.

[12] S. F. Terry. The global alliance for genomics & health. *Genetic testing and molecular biomarkers*, 18(6):375–6, June 2014.

[13] L. Yarmey and K. S. Baker. Towards Standardization: A Participatory Framework for Scientific Standard-Making. *International Journal of Digital Curation*, 8(1):157–172, June 2013.

[14] X. Zhu, H.-I. Suk, and D. Shen. A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis. *NeuroImage*, 100C:91–105, June 2014.