RESEARCH

Curami: an Assisted Attribute Curation Tool for the BioSamples Database.

Matthew Green*

Correspondence:
hewgreen@ebi.co.uk
European Bioinformatics Institute,
European Molecular Biology
Laboratory (EMBL), Wellcome
Trust Genome Campus, CB10 1SD
Cambridge, UK
Full list of author information is
available at the end of the article
*Equal contributor

Abstract

First part title: Text for this section.

Second part title: Text for this section.

Keywords: metadata; biosamples; assisted curation; data quality

Background

Architects of biological databases are forced to decide which trade offs to make in order to best suit their data consumers and available resources. Whilst flexible validation and pliable data schema lower the barrier for initial data submission, the resulting poor data quality later incurs a high curation cost or reduced community benefit. Many tools have been built which aim to improve data quality whilst minimising incurred time and financial costs of operation and these can be broadly considered to follow either pre-data submission (validation) or post-data submission (curation) strategies [REF NEEDED]. Within these two strategies, the numerous tactics employed fall somewhere between fully automated or manual. Often increased manual effort correlates with increasing data value which again forces a compromised approach depending on the requirements of the resource [1]. Nevertheless, there are successful applications that leverage the benefits of automated data assessment and combine this with expert manual curation to measurably increase the efficiency of manual curation efforts [2, 3, 4].

The BioSamples database at the European Bioinformatics Institute (EBI) is a metadata repository for all biological samples used to generate datasets across the EBI [5, 6]. The database consists of sample records which contain a unique BioSamples identifier, links to external data, relationships to other samples (derived from, child of etc.) and metadata expressed as key value pairs (note that keys are hereinafter referred to as attributes). As the database provides a single point of entry to metadata from a wide range of scientific domains and technologies, it allows users to create fields that are not predefined to suit their needs. Whilst this increases utility and flexibility for submitters, the dataset incurs redundancy and inconsistency. For example there are 22 different attributes that contain longitudes. Although BioSamples mitigates this problem by providing elastic search with ontology expansion, the samples returned may not necessarily be findable, accessible, interoperable and reusable, the universal targets defined as the FAIR principles [7]. A snapshot of the dataset containing 4,790,415 sample records (taken on the 5th January 2018) was used for initial analysis. These samples contained 29,751 unique attributes used conjointly 45,275,314 times, giving a mean average of 9.45 attributes per sample record. The redundancy and inconsistency within the large number of attributes in the BioSamples database prompted the creation of the assisted-curation tool described herein.

The overarching goal was to provide a tool that facilitates the reduction of redundancy and inconsistency of the attributes in the BioSamples database, ultimately to reduce the total number of attributes in the database without loosing information. In order to achieve this, we aimed to identify pairs of attributes that were semantically similar enough to merge and to further identify the correct polarity of the merging. Although these aims differ from recent work to identify more generally topically related attributes in the Gene Expression Omnibus (GEO) [8, 9], the strategy the authors employ highlighted some key challenges to solving this problem [10]. Whilst GEO encourages data submitters to conform to the Minimum Information About a Microarray Experiment (MIAME) guidelines [11], similarly to BioSamples, the lack of a controlled vocabulary leads to the usage of different terms to represent the same concept. The authors developed a clustering methodology to group attributes that covered similar concepts and in doing so aim to separate curation activities into more manageable chunks. Unfortunately, applying clustering methodologies to data with heavily biased distributions leads to biased results which impedes semantic understanding. Furthermore as is the case of BioSamples, the majority of metadata submissions come from a few pipelines which may impose their own validation, recommendations and guidelines which lead to further bias and render frequency of term usage a less definitive predictive metric.

Methods

Curami's source code and documantation is available for inspection and reuse at *https://github.com/EBIBioSamples/curami.*

Overview of Data and Work Flow

Figure 1

Calculated Data Features

Different setting options

Actual Curation Work (including some observations of problems and potential issues)

Pair Sorting Settings

Results

Discussion

Curami is designed to be used from two perspectives. As an assisted curation application it can be used to navigate the user through their domains of interest within the dataset and provide auxiliary information to the curator which may increase their confidence and allow them to make a curation decision. From a second perspective, this decision capture is incredibly useful as a training set to later discover in which circumstances the calculated features become predictive indicators of a required curation operation. Although this work has not explored analysis of the captured features, this second utility was a major contributing factor to the design of Curami. Displaying the right pairs to the curator and the information required to make decisions will aid the curation process [REF NEEDED] but the potential of later leveraging each of these decisions through machine learning will further increase a curators impact. However, the current lack of training data dictated that as a first step, Curami had to provide immediate utility whilst generating training data for later correlation analysis.

Conclusion

Competing interests

The authors declare that they have no competing interests.

Author's contributions

Text for this section . . .

Acknowledgements
Text for this section . . .

References

1. Goble, C., Stevens, R., Hull, D., Wolstencroft, K., Lopez, R.: Data curation+ process curation= data integration+ science. Briefings in bioinformatics **9**(6), 506–517 (2008)
2. Salgado, D., Krallinger, M., Depaule, M., Drula, E., Tendulkar, A.V., Leitner, F., Valencia, A., Marcelle, C.: Myminer: a web application for computer-assisted biocuration and text annotation. Bioinformatics **28**(17), 2285–2287 (2012)
3. Salimi, N., Vita, R.: The biocurator: connecting and enhancing scientific data. PLoS computational biology **2**(10), 125 (2006)
4. Szostak, J., Ansari, S., Madan, S., Fluck, J., Talikka, M., Iskandar, A., De Leon, H., Hofmann-Apitius, M., Peitsch, M.C., Hoeng, J.: Construction of biological networks from unstructured information based on a semi-automated curation workflow. Database **2015** (2015)
5. Gostev, M., Faulconbridge, A., Brandizi, M., Fernandez-Banet, J., Sarkans, U., Brazma, A., Parkinson, H.: The biosample database (biosd) at the european bioinformatics institute. Nucleic acids research **40**(D1), 64–70 (2011)
6. Faulconbridge, A., Burdett, T., Brandizi, M., Gostev, M., Pereira, R., Vasant, D., Sarkans, U., Brazma, A., Parkinson, H.: Updates to biosamples database at european bioinformatics institute. Nucleic acids research **42**(D1), 50–52 (2013)
7. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., et al.: The fair guiding principles for scientific data management and stewardship. Scientific data **3** (2016)
8. Edgar, R., Domrachev, M., Lash, A.E.: Gene expression omnibus: Ncbi gene expression and hybridization array data repository. Nucleic acids research **30**(1), 207–210 (2002)
9. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., *et al.*: Ncbi geo: archive for functional genomics data setsóupdate. Nucleic acids research **41**(D1), 991–995 (2012)
10. Hu, W., Zaveri, A., Qiu, H., Dumontier, M.: Cleaning by clustering: methodology for addressing data quality issues in biomedical metadata. BMC bioinformatics **18**(1), 415 (2017)
11. Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., *et al.*: Minimum information about a microarray experiment (miame)ótoward standards for microarray data. Nature genetics **29**(4), 365 (2001)

Figures

Figure 1 Data and Work Flow Overview of the modular design of Curami showing separation of scripts and general dataflow. Initialization and analysis scripts 1-5 should be executed in series to pull information from the BioSamples API to generate input files, pair lexically similar attributes, extract pairwise and individual features and store the results in the Neo4J database. The Data Analysis Store has three node types User, Pair and Attribute with corresponding relationships as shown. Edges *r1* represent curation decisions made by the user about a pair of attributes and can be one of four types; merge, merge with reversed polarity, don't merge and skip. Edged *r2* has only one type, 'pair contains', and indicated which individual attributes are in the pair. The Curation Interface allows curators to view the attribute pairs sorted by applying one of the available settings described in section . Curation decisions are then stored in the Neo4J database which can then be directly converted into curation objects, a format that can be directly submitted to the BioSamples API.