Human Computation (2019) 6:1:98-112

ISSN: 2330-8001, DOI: 10.15346/hc.v6i1.6

MetaCrowd: Crowdsourcing Biomedical Metadata Quality Assessment

AMRAPALI ZAVERI, MAASTRICHT UNIVERSITY

WEI HU, NANJING UNIVERSITY

MICHEL DUMONTIER, MAASTRICHT UNIVERSITY

ABSTRACT

To reuse the enormous amounts of biomedical data available on the Web, there is an urgent need for good quality metadata. This is extremely important to ensure that data is maximally Findable, Accessible, Interoperable and Reusable. The Gene Expression Omnibus (GEO) allows users to specify metadata in the form of textual key: value pairs (e.g. *sex: female*). However, since there is no structured vocabulary or format available, the 44,000,000+ *key: value* pairs suffer from numerous quality issues. Using domain experts for the curation is not only time consuming but also does not scale. Thus, in our approach, MetaCrowd, we apply crowdsourcing as a means for GEO metadata quality assessment. Our results show that crowdsourcing is a reliable way to identify similar as well as erroneous metadata in GEO. This is extremely useful for data consumers and producers to curate and provide good quality metadata.

1. INTRODUCTION

Advancements in molecular technologies have enabled extensive profiling of biological samples, resulting in massive amounts of data that can be analyzed to better understand living systems. Increasingly, journals, funding agencies, and investigators all realize the value that these data have to reproduce published findings, validate their own results, and generate new and interesting hypotheses (Barrett et al., 2013b). However, being able to find, interpret, evaluate, and reuse relevant datasets remains a substantial challenge. Consider the work from Khatri and colleagues (Khatri et al., 2013), who used publicly available expression data from the Gene Expression Omnibus (GEO) (Edgar et al., 2002; Barrett et al., 2013a) to identify gene signatures that were predictive for tissue graft rejection. Their work required them to search for and tediously curate important sample characteristics (organism, tissue, protocol, etc) for deposited samples. This documentation, otherwise known as *metadata*, helps investigators understand the meaning and provenance of the data (Borgman, 2012). Ambiguous, incomplete, imprecise metadata makes it difficult to find datasets

that meet particular criteria and will make it impossible interpret the data elements or the context by which the data were generated.

As of May 2017, the GEO database contained 84,220 study records (also known as Series) submitted by over 3,000 laboratories, comprising of 2,066,179 sample records (Sample) derived from over 1600 organisms and 17,214 Platforms[1]. A Sample[2] describes the conditions under which a particular biological sample was handled, the manipulations it underwent, and their abundance. A Series groups together related Samples and provides a focal point and description of the whole study. Gene expression profiling data are typically produced on a small scale, in targeted studies that vary in terms of tissue or cell type, disease model, expression assay platform and model organism. Data submitters can submit data to GEO via three ways: (i) spreadsheets, (ii) SOFT format (plain text), (iii) MINiML format[3] (XML). When users submit data to GEO via a spreadsheet, it requires them to fill out a metadata template that follows the guidelines set out by the Minimum Information About a Microarray Experiment (MIAME) guidelines (Brazma et al., 2011). The metadata template includes fields for title, summary, overall design, contributors, the protocols as well as sample characteristics (e.g. sex, organism, tissue, cell type), but does not refer to a standardized vocabulary for these fields. After submission, a curator checks the content and validity of the information provided (Barrett et al., 2011), and will work with the submitter until all issues are resolved. However, the problem is that the submitter-supplied data is heterogeneous in the style, content and level of detail with which the experiments are described (Soboleva et al., 2008).

While domain experts are well suited to address this problem of curating metadata, they are both expensive and in short supply, and cannot easily cope with the growing amount of new knowledge. One approach to this problem is crowdsourcing, in which non-experts are recruited to execute simplied tasks. Human Intelligent Tasks (HITs) are submitted to a crowdsourcing platform (e.g. Amazon Mechanical Turk[4], Figure Eight[5] etc.) for a worker (non-expert) to perform and to obtain a (financial) reward (Howe, 2006). The ability to execute these tasks depend more on basic understanding of what is being asked rather than any specific skills (such as domain knowledge). One advantage of crowdsourcing is the degree of task parallelization such that the work can be divided and completed in shorter periods of time. However, a key part of crowdsourcing is to obtain consensus from multiple provided answers. This makes solving the tasks time as well as cost efficient and also offers a means to cross-check the accuracy of the answers (by assigning each task to more than one person).

Crowdsourcing has been used for entity linking or resolution (Demartini et al., 2012; Guoliang, 2017), quality assurance and resource management (Wang et al., 2012) and for enhancement of ontology alignments (Sarasua et al., 2012). Crowdsourcing has been previously used to assess the quality of Linked Data with domain experts (Zaveri et al., 2013) as well as with workers from Amazon Mechanical Turk (Acosta et al., 2013). Others have used crowdsourcing to annotate and extract gene expression signatures from GEO (Wang et al., 2016), to improve automated mining of biomedical text for annotating diseases (Good et al., 2015), to cureate gene-variant relations (Burger

[1]Derived from https://www.ncbi.nlm.nih.gov/geo/, last accessed May 2017.

[2]https://www.ncbi.nlm.nih.gov/geo/info/overview.html

[3]'MIAME Notation in Markup Language' format

[4]http://mturk.com

[5]https://figure-eight.com

et al., 2014), to identify drug side-effects (Gottlieb et al., 2015), drug indications (Khare et al., 2015), as well microRNA functions (Vergoulis et al., 2015). These efforts are significant as they focus on applying crowdsourcing to biomedically relevant problems, with workers who do not have domain expertise. These studies produce large collections of high-quality datasets that can be further utilized by algorithms that can extract new knowledge from already-published data that require better annotation, cleaning and reprocessing. Another study (Hadley et al., 2017) employed 72 users that used 731 tags to make 40,361 annotations of digital experiments representing 2,835,037 annotations of digital samples. However, the 72 users were biomedical graduate students and it took them over a year to provide the annotations. This makes the annotation difficult to maintain on a large scale and for a long time. Additionally, it can get expensive and time consuming to have them continuously annotating the Samples. Thus, with our project, we aim to propose a more efficient and inexpensive method for curating biomedical metadata. To the best of our knowledge, there have been no efforts in applying paid microtask crowdsourcing to curate the quality of gene expression metadata, which is the main aim of this paper.

2. MAIN CONTRIBUTIONS

We explore the following research questions:

- RQ1: What is the performance, in terms of time, money and accuracy, of the crowd on gene expression quality assessment? We will execute crowdsourcing microtasks for categorization of different gene expression metadata key types. We will evaluate the performance by calculating the amount of time the workers took for the task, the overall cost as well as accuracy, specifically in terms of sensitivity and specificity.
- RQ2: Are there differences in the performance of the crowd on different gene expression metadata? We will select keys from different categories of gene expression metadata for the crowdsourcing tasks. We will compare and evaluate the results based on how accurately the workers are able to identify the category correctly for each key type.

The main contributions of this paper are as follows:

- The development of MetaCrowd, a crowdsourcing platform for metadata quality assessment.
- Empirical analysis of crowdsourcing quality assessment of gene expression metadata.
- Qualitative analysis of a set of gene expression characteristics from the GEO.
- Analysis of the results, lessons learned, and implications of using MetaCrowd for large-scale biomedical metadata quality assessment.

3. METHODOLOGY

In the crowdsourcing experiment, the main aim is to ask the crowd to classify a metadata key into one of the provided categories. In particular, the worker has to choose one of the eight provided categories for a given metadata key. Additional information such as five example values of that key along with definitions of the categories are also provided so as to help the worker choose the right category. This classification would help determine (i) lexical and conceptually similar keys that are grouped into the same category (e.g. age and age(months)), (ii) lexically different but conceptually similar keys that are grouped into the same category (e.g. disease and illness) and (iii) outliers, those that do not fit into any category that are potentially flagged as erroneous (e.g. healthy control).

In this section, we first describe the GEO metadata elements used in our experiments, followed by details of the methodology and design of the crowdsourcing experiment.

3.1. GEO Metadata

In this work, we focused on experimental metadata from GEO, in particular on 'Sample' records. A Sample record describes the conditions under which an individual Sample was handled, the manipulations it underwent, and the abundance measurement of each element derived from it. From the different metadata elements in a Sample[6], we chose the 'Characteristics' metadata, which contains information about the tissue, age, gender, cell type, disease etc. of the study. We chose this field as it is represented in a `key: value` pair format as opposed to the free text format of the other fields. For example, the sample `GSM549326`[7] consists of several `key: value` pairs as depicted in Figure 1. While `GSM549326` contains information of the disease in the form `illness: Still`, one of the other Samples `GSM550400` captures the same disease information such as `healthy control: pSLE`. The term 'healthy control' as a 'key' illustrates a data quality of inconsistency in the representation of health status (healthy, diseased) for a Sample.

The `key: value` pairs are not uniformly captured due to the lack of standard set of terms provided during submitting the metadata. The problems in the keys range from

- spelling errors (e.g. `age at diagonosis (years)`; `genotype/varat, genotype/varation genotype/variataion, genotype/variation`)
- syntactic variance (e.g. `age (years), age(yrs) and age_year`)
- synonyms (e.g. `disease` vs. `illness` vs. `healthy control`)
- multi-category terms (e.g. `disease/cell type`, `tissue/cell line`, `treatment age`)

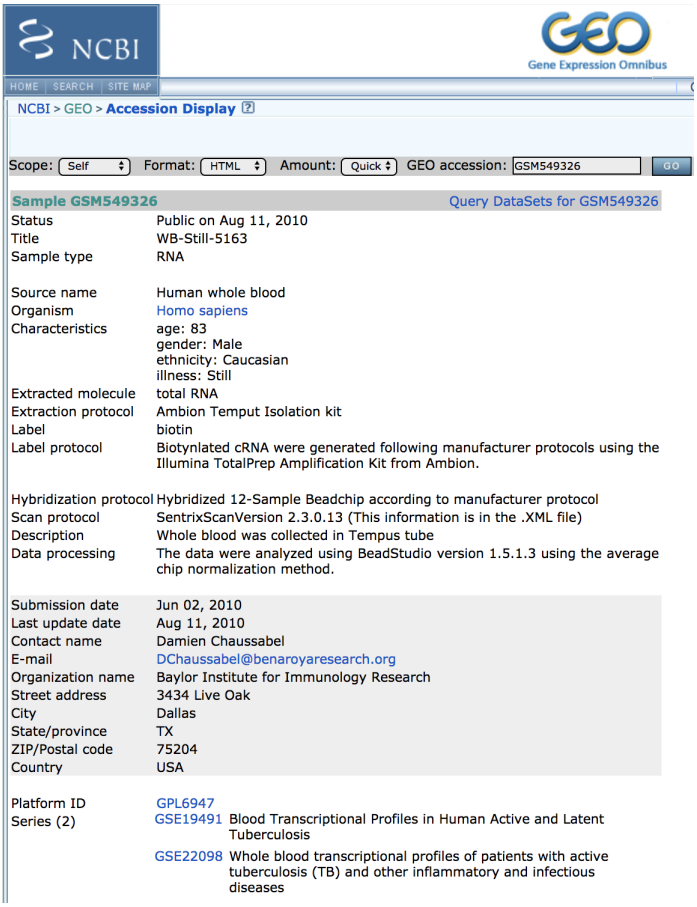
Moreover, the values for these keys are problematic themselves from

- not being consistent (e.g. `time of treatment` key values: after disease establishment, zt24, zt36, 6h, time of grafting - denoting heterogeneous representations of time points)
- also suffering from the minor spelling discrepancies (`genotype` key values: wild type, wt, wild-type, wildtype)
- lacking semantics such as units for the values (e.g. `age at presentation` key values: 72, 50, 70, 69, 66) or (iv) being ambiguous (e.g. `dopamine-agonists treatment` key values: no, -, yes).

Sample records contain the precise metadata by which other biologists can look-up and find studies related to their own experiments, thus enabling re-use. When one attempts to find similar studies by querying the metadata using keywords (as available by the GEO website), all the related studies are not retrieved resulting in loss of important information. Thus, as a first step, we chose the GEO Samples to identify and resolve such quality issues in the *keys* of the millions of GEO Sample records. Resolving quality issues for the values and then the `key: value` pairs is part of the future work.

[6] https://www.ncbi.nlm.nih.gov/geo/info/spreadsheet.html

[7] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM549326, last accessed April 2017



NCBI

Gene Expression Omnibus

HOME | SEARCH | SITE MAP

NCBI > GEO > **Accession Display**

Scope: Self Format: HTML Amount: Quick GEO accession: GSM549326 GO

Sample GSM549326 Query DataSets for GSM549326

Status	Public on Aug 11, 2010
Title	WB-Still-5163
Sample type	RNA
Source name	Human whole blood
Organism	Homo sapiens
Characteristics	age: 83 gender: Male ethnicity: Caucasian illness: Still
Extracted molecule	total RNA
Extraction protocol	Ambion Temput Isolation kit
Label	biotin
Label protocol	Biotynlated cRNA were generated following manufacturer protocols using the Illumina TotalPrep Amplification Kit from Ambion.
Hybridization protocol	Hybridized 12-Sample Beadchip according to manufacturer protocol
Scan protocol	SentrixScanVersion 2.3.0.13 (This information is in the .XML file)
Description	Whole blood was collected in Tempus tube
Data processing	The data were analyzed using BeadStudio version 1.5.1.3 using the average chip normalization method.
Submission date	Jun 02, 2010
Last update date	Aug 11, 2010
Contact name	Damien Chaussabel
E-mail	DChaussabel@benaroyaresearch.org
Organization name	Baylor Institute for Immunology Research
Street address	3434 Live Oak
City	Dallas
State/province	TX
ZIP/Postal code	75204
Country	USA
Platform ID	GPL6947
Series (2)	GSE19491 Blood Transcriptional Profiles in Human Active and Latent Tuberculosis
	GSE22098 Whole blood transcriptional profiles of patients with active tuberculosis (TB) and other inflammatory and infectious diseases

Figure 1. Example of the GEO Sample.

To retrieve the relevant metadata, we executed a SPARQL query over an Resource Description Framework (RDF)[8] transformed dataset of the GEO database[9]. For the conversion, a copy of the SQLite3 GEO database[10] was obtained and converted to the RDF format using Sparqlify[11], a scalable SPARQL-SQL rewriter. All seven tables (GSM, GSE, GPL, GDS Dataset and GDS Subset, GSE_GPL and GSE_GSM[12]) in GEO were converted to RDF[13] by mapping the column name as properties and using the unique Sample ID as the resource IDs. For example, the Listing 1 shows an excerpt of the RDF representation of a single Sample (GSM1272900).

[8]https://www.w3.org/RDF/

[9]Detailed description of the RDF format is out of scope of this article

[10]available at http://gbnci.abcc.ncifcrf.gov/geo/index.php version January 23 2016, 264.5 MB, 07:23:09)

[11]https://github.com/AKSW/Sparqlify

[12]Information about these tables provided at https://www.ncbi.nlm.nih.gov/geo/info/overview.html

[13]Scripts available at https://github.com/amrapalijz/GEO

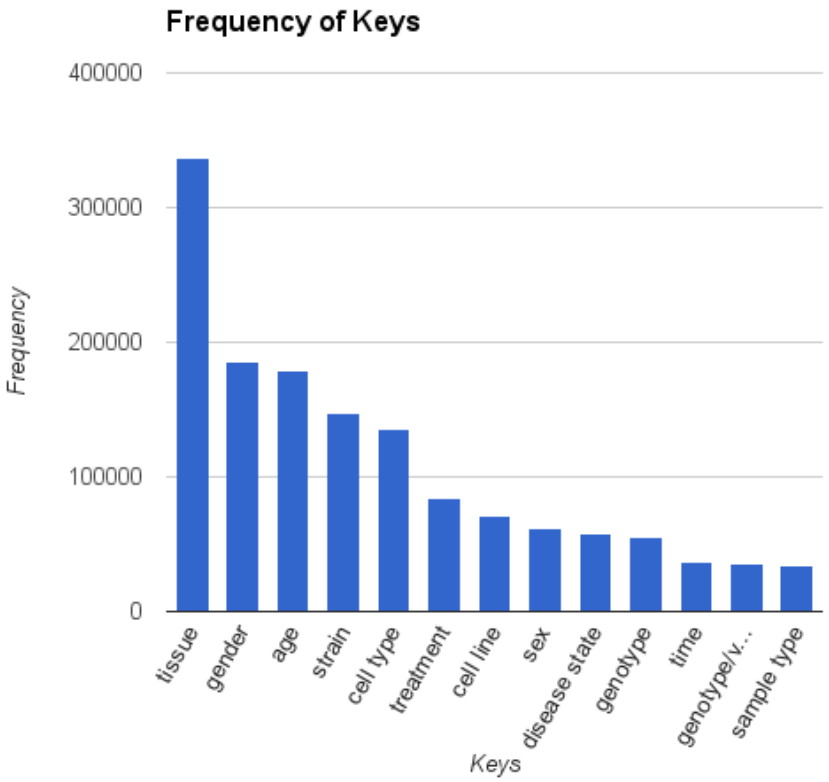


Figure 2. *Top most frequently occurring keys in GEO.*

```
PREFIX geo: <http://bio2rdf.org/geo:>
PREFIX gvoc: <http://bio2rdf.org/geo_vocabulary:>

geo:GSM1272900          a                       geo_vocabulary:Sample;
                        geo_vocabulary:gsmID    "GSM1272900";
                        gvoc:key                geo:GSM1272900/cell%20line,
                                                geo:GSM1272900/strain,
                                                geo:GSM1272900/treatment .

geo:GSM1272900/cell%20line    geo_vocabulary:value    geo:GSM1272900/rn2;
                              rdfs:label              "cell line" .

geo:GSM1272900/rn2            rdfs:label              "rn2" .
```

Listing 1. *Excerpt of the RDF representation of a single Sample (GSM1272900).*

The 'Characteristics' column in the GSM Samples table, however, was poorly formatted with several metadata keys and values in the same column. The standard representation of `key1: value1; key2: value2` was not always followed (e.g. "Gender: unknown; Tissue: liver; Tumor stage: carcinoma (GSM341738) vs. Age (yrs),38.63,PMI (hrs),8,gender,m (GSM341548)). Thus, with a

direct mapping of the cell to the object value of a triple lead to poor representation of the metadata. Thus, we first converted the data to JSON by separating out the key: value pairs using different delimiters and then converted it to RDF[14]. The RDF data was then loaded in a Graph Database[15] to query and retrieve specific parts of the dataset available at http://graphdb.dumontierlab.com[16]. Listing 2 shows an example of a SPARQL query to retrieve all the values belonging to the key 'age' and sorting them in descending order.

```
PREFIX gvoc: <http://bio2rdf.org/geo_vocabulary:>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT DISTINCT ?valuelabel COUNT(?valuelabel)
WHERE
{
?x      a              gvoc:Key .
?x      rdfs:label     "age"^^<http://www.w3.org/2001/XMLSchema#string>.
?x      gvoc:value     ?value.
?value  rdfs:label     ?valuelabel.
} ORDER BY DESC(COUNT(?valuelabel))
```

Listing 2. SPARQL query to retrieve all the values belonging to the key 'age' sorted in descending order.

Figure 2 shows the top most frequently occurring keys in GEO sample data.

3.2. Crowdsourcing Methodology

For the crowdsourcing experiments, we selected total of 1643 keys associated with eight (top most frequently occurring) categories: (i) `cell line`, (ii) `disease`, (iii) `gender`, (iv) `genotype`, (v) `strain`, (vi) `time`, (vii) `tissue` and (viii) `treatment`. We retrieved the corresponding top 5 most frequently occurring values for each of the keys and, using the SPARQL query shown in Listing 1 to display to the crowd. Definitions for each of these categories were obtained the Semanticscience Integrated Ontology (SIO[17]) (Dumontier et al., 2014) and Medical Subject Headings ontology[18].

Candidate syntactic variants of each key were obtained through regular expressions using SPARQL FILTER clause. For example, by querying for 'disease', all keys with this keyword were retrieved such as 'disease state', 'disease specific survival years', 'disease onset' etc. Two independent researchers manually categorized 60 of the total set of keys, which we used as our gold standard. Then, we generated microtasks as depicted in Figure 3. For each microtask, the worker is provided with a 'Term' (i.e. the key) along with five of the most frequently occurring example values. A list of the eight key categories are provided along with a definition for each. The worker's task is to analyze the given 'Term' and five values, and choose the category that the 'Term' best belongs to. For example, if the 'Term' is 'disease specific survival (years)' and the values are '8.22, 17.66, 4.51, 0.89, 12.19', the *correct* category is 'time', but not 'disease'.

[14]https://github.com/yamalight/gsmCharacteristics

[15]http://graphdb.ontotext.com/

[16]Toggle for the option 'GEO' on the upper right hand corner

[17]Available at https://bioportal.bioontology.org/ontologies/SIO

[18]http://bioportal.bioontology.org/ontologies/MESH, last accessed April 2017

An additional option of 'Don't know/I cannot tell' was also provided. When the worker clicks on this option, five reasons are presented (also shown in Figure 3):

– Does not fit into any category
– The term is ambiguous.
– There is not enough information provided to choose the right category.
– I do not understand the examples.
– I am not sure.

The worker must choose one of these five reasons when they are unsure which category that term can belong to. This information can then be used to understand the rationale for disagreement for particular keys.

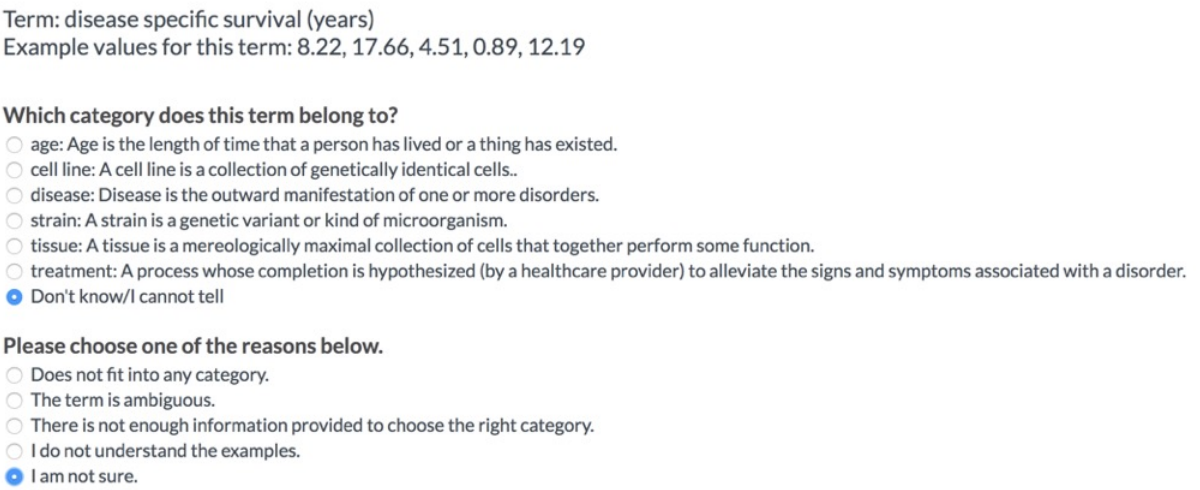


***Figure 3.** Example of the microtasks provided to the workers on CrowdFlower.*

Task Settings:

– Workers were recruited using the Figure Eight platform. We chose 'Level 2' workers, which are a small group of more experienced, higher accuracy contributors[19]. The workers were provided with an overview of the task, instructions for the steps to follow, rules and tips (e.g. 'If you do not know which category the term can belong to, please choose "Don't know/I cannot tell"'.) and positive and negatives examples of the answers to guide them.
– We initially assigned 3 workers per microtask and enabled the 'Dynamic Judgment' setting, which automatically requests additional judgments if contributors disagree on an answer. We

[19]In CrowdFlower, there is an option to choose 'Level 1' workers, which provides the fastest throughput (but not necessarily high quality judgments), 'Level 3' workers, which are the smallest group of experienced, highest accuracy contributors and 'Level 2' workers, which are a small group of more experienced, higher accuracy contributors. Thus, since we wanted a larger set of workers as well as those with high accuracy, we chose 'Level 2' workers.

set the maximum number of judgments to be 7 or when the minimum confidence of 0.80 is reached. The minimum accuracy for the test questions was set to 80%. Majority consensus was used to select the final answer and to test the workers' reliability based on their agreement with other workers.
– We paid 5 cents per judgment and grouped 10 rows (each row corresponding to a metadata key) per page and set the minimum time a worker spends on each page to be 12 seconds.
– For quality control, there were 60 test questions in total and each page contained 9 rows and 1 test question. These test questions ranged from simple (e.g. keys belonging to one category) to complex ones (e.g. keys that can belong to more than one category).

The datasets analyzed during the current study are available in the GEO repository available at http://gbnci.abcc.ncifcrf.gov/geo/index.php. All the input GEO metadata and the crowdsourcing results are available on our website at http://ws.nju.edu.cn/geo-clustering/.

4. EVALUATION AND RESULTS

We evaluated our approach for 1,643 GEO metadata keys belonging to eight key categories: (i) `cell line`, (ii) `disease`, (iii) `gender`, (iv) `genotype`, (v) `strain`, (vi) `time`, (vii) `tissue` and (viii) `treatment`. Table 1 summarizes the overall results. The approach resulted in an overall accuracy[20] of 0.93 on all the tasks (including gold standard questions). Furthermore, we calculated the specificity and sensitivity and observed that crowd workers obtained slightly higher values of specificity than sensitivity (0.95 vs. 0.91), which suggests that workers perform better when correctly detecting true negatives.

A total of 145 workers performed the tasks and a total of 7835 judgments were provided (including test questions). The total cost of the experiment was $470 (7835 judgments * $0.05 paid per judgment = $391.75 + 20% transaction fee $78.35) and total time was ~1 hour. The IQM (Interquartile Mean) task time by trusted contributors was 3m 29s and that of untrusted contributors was 7m 43s. The average accuracy for gold standard questions only for trusted contributions was 90% and that of untrusted contributions was 62%[21].

Table 2 reports the number of keys and accuracy achieved with 3 workers as well as overall accuracy with all workers for each of the eight key categories. The highest overall accuracy was achieved for the 'genotype' category at 0.98 and the comparatively lowest accuracy was for the 'gender' category at 0.90. Accuracy for the other categories ranged from 0.908 to 0.96. From the total of 1643 keys, 1454 achieved consensus with 3 workers, 66 keys with 5 workers and 131 keys with 7 workers. The option 'Don't know/I cannot tell' was chosen 100 times. From the 5 reasons, 42 workers chose R1, 29 chose R3, 16 chose R2, 7 chose R4 and 6 chose R5.

For the 'cell line' category (including the 'cell type' keys), 98 keys of 109 key were correctly classified by 3 workers per key. Keys that belonged to the 'time' category such as 'cell line source age' and 'cell line initiation date' were assigned to 7 workers. 5 of the workers correctly classified these keys since the values indicated a time point. The workers chose the reasons R1 and R3, R4 for

[20] Accuracy is the percentage of the correct answers on the total.

[21] https://success.crowdflower.com/hc/en-us/articles/202703305-Glossary-of-Terms#trust_score

Table 1. Results of the crowdsourcing experiment

No. of microtasks (keys)	1643 rows
Total no. of workers	145
Total no. of judgments	7835
Overall accuracy	0.93
Sensitivity	0.91
Specificity	0.95
No. of gold standard questions	60
Accuracy on gold standard questions	0.93
Agreement (%)	94.42
Average confidence for workers	0.91
Total cost	470$
Total time	1 hour
Inter-quartile mean task time by trusted and untrusted contributors	3m 29s, 7m 43s

the keys 'targeted cell type' and 'pancreatic cell type'. However, these workers had comparatively lower trust scores of 0.80.

In case of the 'disease' category, from the 85 keys, 61 keys were correctly classified by 3 workers for each key. Keys that, in particular, belong to the 'time' category as indicated by the values such as 'disease free survival (month)', 'duration of disease (month)', 'mean disease duration', 'disease duration (yrs)', 'disease-free survival (dfs)', 'disease phase' were assigned and correctly classified by 7 workers for each key. However, there was low consensus for the keys 'code disease-specific survival', 'stage of disease', 'trg disease state', 'disease progression (event)', 'diseasestatus', 'disease/treatment status', 'disease_state'. One of the workers chose the reason R1 for the 'diseasestatus' key since it's values '1, 2, nafld, nash, normal control' were inconsistent and non-informative to choose the best fit.

For the 'gender' category, out of the 72 keys, 63 keys were correctly classified by 3 workers for each key. There were several keys that potentially belong to another category such as 'cell sex', 'sex/age at diagnosis (years)', 'strain c57bl/6 gender', 'cell line source gender', 'genotyped sex', 'gender and age'. 7 workers were assigned for each of these keys who, in majority, classified the keys correctly. However, there were keys with no consensus (e.g. 'w sex', e.g. 'sex chromosome complement') and the workers chose either of the following reasons: R1, R2, R3 or R5 indicating that they either did not understand the term or the examples.

From the 112 keys in the 'genotype' category, 106 achieved consensus by assigning 3 workers for each key. There were disagreements for keys such as 'agenotype' where the worker incorrectly chose 'time' (1 out of the 7 workers) and 'tissue genotype/variation' where the worker incorrectly chose 'tissue' (4 out of the 7 workers). However, for the key 'strain genotype', the workers correctly chose 'strain' (5 out of the 7 times).

Table 2. ***Accuracy for each key of the eight key categories.***

Key	No. of keys	Correct classification by 3 workers	Accuracy with 3 workers	Overall accuracy
Cell line	109	98	0.89	0.95
Disease	85	60	0.70	0.93
Gender	72	61	0.84	0.90
Genotype	112	106	0.94	0.98
Strain	181	154	0.85	0.96
Time	698	658	0.94	0.90
Tissue	145	123	0.84	0.94
Treatment	242	192	0.79	0.94

For the 'strain' category, from the 181 keys, 154 keys were correctly classified by assigning 3 workers for each key. However, for several keys, the workers chose different reasons for not choosing a category: 'strain by4741 ptc3' (R1), 'strain (genetic background)' (R2), 'organ-derived dc strain' (R2 and R5), 'strain sr1187' (R4), 'strain id' (R4) and for the key 'strain fgsc number' (R1 and R3). This indicated that the workers were unsure about which category the key fit best as either they found the term ambiguous or did not understand the examples. Moreover, for the keys 'strain/cell line background' and 'strain/genotype variation', there were disagreements between the workers in choosing the 'strain', 'cell line' and 'genotype' categories. The values for these keys were ambiguous which led to the perplexity.

In case of the 'time' category with 698 keys, 662 keys were correctly classified by assigning 3 workers to each key. Keys for which the workers either chose 'time' or 'treatment' were, for example, 'treatment time', 'time to treatment (days)', 'age at start of treatment', 'time of treatment'. However, the majority consensus was towards 'time' since the values indicated a time point. For the key 'age/disease timepoint', 1 out of 7 worker incorrectly chose 'disease'. For the keys 'sampling age' and 'age at time of surgery', two workers each chose the reason R1 since the values did not indicate any unit, which made it difficult for them to determine the best category. For the '8 weeks. tissue' pair, the workers correctly chose 'tissue' as the category even though the key itself indicated a time point. On the other hand, for the key 'age at rc', the workers chose the R2 and R3 reasons indicating that they found the term ambiguous with not enough information to choose the right category since it had only one value '51.33' with no units. The key 'tissue/development stage' was classified into the 'tissue' category by the workers (only one chose 'time'). The value for this key 'inflorescences including floral stages 1-13', however, does not provide enough information to classify it correctly.

For the 'tissue' category, out of the 145 keys, 126 were correctly classified by assigning 3 workers for each key. Keys for which 7 workers were assigned were 'age and tissue', 'day of tissue dissection' and were correctly classified into the 'time' category. The workers chose the reason (i) R3 for the key 'tissue & age' with values such as 'fiber 2 dpa' and 'npcs of neocortex from 3 littermates wild-type mouse embryos at e14'; (ii) R1 for the key 'tissue ph' with the values '6.46' and '6.43'.; (iii) R2 for the key 'tissue derivation' with the values 'adenoid cystic carinoma of the parotid gland' and 'pancreatic tumor' indicating that they found the term ambiguous or needed more information

to choose the best fit. However, the key 'tissue/treatment id' with values '4', '2' etc, is ambiguous and does not fit into either of the 'tissue' or 'treatment' category correctly.

From the 242 keys in the 'treatment' category, 192 keys were correctly classified by assigning 3 workers for each key. The keys 'duration of il-6 treatment', 'treatment duration', 'days of treatment', 'tnfa treatment time point', 'length of treatment' were correctly classified to the 'time' category by assigning them to 7 workers each. However, there was disagreement regarding the best fit for the key 'small molecule treatment' as the values are 'repsox at 24 hours' and 'repsox at 48 hours' indicated 'time' but the workers chose 'treatment' 4 out of the 5 times.

5. DISCUSSION

Lessons Learned The results of the experiments lead to the following conclusions and lessons learned:

- **RQ1**: From the results achieved in the MetaCrowd experiment, we conclude that the performance of the crowd is sufficient to *partly* perform biomedical metadata quality assessment. Partly because there is still need for experts to identify domain specific quality issues. In terms of the time taken to complete the task, which was within 1 hour for c.a. 1600 tasks, this is efficient as opposed to one person completing all the tasks. The tasks were costly since we added up to 7 workers. However, we can optimize the task-worker assignment using our statistical method, CrowdED (Zaveri et al., 2018), which a-priori estimates the optimal number of worker and task assignment to obtain maximum accuracy. Although the overall accuracy was high with 93%, the sensitivity and specificity values for the tasks gave more insight into the performance of the workers. The lower performance of the workers in terms of sensitivity (as compared to specificity) was not surprising, since this particular task requires certain domain knowledge about biological data and experiments.
- **RQ2**: Overall, the workers could correctly identify the category for those keys that contained the category phrase in the key name. In effect, this showed that non-expert workers have the necessary meta-cognitive skills to assess which keys belong to which category and more importantly, which ones do not belong to any category. However, they were unable to correctly identify categories for certain types of keys. Even though the five top-most frequently occurring values were provided to the workers, it was interesting to note that they were not able to choose the relevant category. This was due to either (i) lack of semantically annotated values, (ii) ambiguous nomenclature of keys as well as the values, (iii) values indicating that keys belong to more than one cluster or (iv) inconsistent usage of the particular metadata key. In particular, there was higher accuracy for the key categories 'genotype' and 'strain' (0.98 and 0.96 respectively) but lower accuracy for the key categories 'gender' and 'time' (0.90 and 0.90 respectively). This was because for the former categories, the keys contained the respective names in the key itself (e.g. 'mouse strain', 'virus strain', 'genotype p53', 'plant genotype') whereas for the latter categories, there were variants in the nomenclature leading to confusions. Additionally, using the MetaCrowd approach only for the top frequently occurring keys gave us an understanding of the strengths and limitations of the crowd. This information will guide us for designing crowdsourcing experiments for the entire dataset.

6. CONCLUSIONS, LIMITATIONS AND FUTURE WORK

Improvements in the quality of biomedical metadata aids in (i) understanding the nature of a dataset, (ii) improves the ability to query for studies involving particular characteristics, and (iii) augments reuse of existing data beyond what the original investigators envisioned to uncover novel insights from the data. This would have a huge impact for data consumers as well as producers for curating and providing good quality metadata in order to re-use the data. This is extremely important when it comes to the FAIR principles of data as well as metadata in order to ensure that data that is already available is maximally Findable, Accessible, Interoperable and Reusable (Wilkinson et al., 2016). Ultimately, improving the quality and quantity of metadata for biomedical datasets is crucial to drive the next paradigm shift in data reuse.

In this paper, we describe MetaCrowd, a crowdsourcing solution to curate and assess the quality of gene expression metadata from selected samples in the GEO database. Through the crowdsourcing experiment on Figure Eight and empirical analysis, we found that the crowds input is cost-effective and efficient to identify correct and incorrect biomedical metadata terms. In that, however, there were differences in assessing specific types of metadata keys. In general, those keys which contained the category phrase in the key name were easier to categorize and those keys which contained phrases from two different categories were harder. Moreover, the poor quality of the values for the keys added to the difficulty of choosing the right category for the key.

The proposed approach and the lessons learned can be used to assess other datasets which have (similar) metadata quality issues (e.g. BioSamples[22]). As future work, we aim to combine automated methods such as machine learning with crowdsourcing in an iterative and adaptive cycle such that the automated methods "learn" from human input to ultimately identify erroneous metadata accurately. This combination will facilitate reducing time and costs and maximize the quality of results to perform large-scale metadata quality assessment.

7. ACKNOLWEDGEMENTS

Support for this work was provided by NCATS, through the Biomedical Data Translator program (NIH awards OT3TR002027 [Red]). Any opinions expressed in this document are those of the Translator community writ large and do not necessarily reflect the views of NCATS, individual Translator team members, or affiliated organizations and institutions. Wei Hu is partially supported by the National Natural Science Foundation of China (No. 61872172).

8. REFERENCES

Acosta, M, Zaveri, A, Simperl, E, Kontokostas, D, Auer, S, and Lehmann, J. (2013). Crowdsourcing Linked Data quality assessment. In *Proceedings of the 12th International Semantic Web Conference, (ISWC)*, Vol. 8219. Springer Berlin Heidelberg, 260–276.

Barrett, T, Troup, B, Wilhite, S, Ledoux, P, Evangelista, C, Kim, I, Tomashevsky, M, Marshall, K, Phillippy, K, Sherman, P, Muertter, R, Holko, M, Ayanbule, O, Yefanov, A, and Soboleva, A. (2011). NCBI GEO: archive for functional genomics data sets — 10 years on. *Nucleic Acids Research* 39 (2011), 991 – 995.

Barrett, T, Wilhite, S, Ledoux, P, Evangelista, C, Kim, I, Tomashevsky, M, Marshall, K, Phillippy, K, Sherman, P, Holko, M, Yefanov, A, Lee, H, Zhang, N, Robertson, C, Serova, N, Davis, S, and A, S. (2013)a. NCBI GEO: archive for functional genomics data sets – update. *Nucleic Acids Research* 41 (2013), 991–995.

[22] https://www.ebi.ac.uk/biosamples/

Barrett, T, Wilhite, S, Ledoux, P, Evangelista, C, Kim, I, Tomashevsky, M, Marshall, K, Phillippy, K, Sherman, P, Holko, M, Yefanov, A, Lee, H, Zhang, N, Robertson, C, Serova, N, Davis, S, and Soboleva, A. (2013)b. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Research* 41 (2013), 991 – 995.

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

Brazma, A, Hingamp, P, Quackenbush, J, Sherlock, G, Spellman, P, Stoeckert, C, Aach, J, Ansorge, W, Ball, C, Causton, H, Gaasterland, T, Glenisson, P, Holstege, F, Kim, I, Markowitz, V, Matese, J, Parkinson, H, Robinson, A, Sarkans, U, Schulze-Kremer, S, Stewart, J, Taylor, R, Vilo, J, and M., V. (2011). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics* 29 (2011), 365 – 371. Issue 4.

Burger, J, Doughty, E, Khare, R, Wei, C, Mishra, R, Aberdeen, H, Tresner-Kirsch, D, Wellner, B, Kann, M, Lu, Z, and Hirschman, L. (2014). Hybrid curation of gene–mutation relations combining automated extraction and crowdsourcing. *Database: The Journal of Biological Databases and Curation* (2014).

Demartini, G, Difallah, D, and Cudré-Mauroux, P. (2012). ZenCrowd: Leveraging Probabilistic Reasoning and Crowdsourcing Techniques for Large-Scale Entity Linking. In *21st International Conference on World Wide Web WWW 2012*. 469 – 478.

Dumontier, M, Baker, C. J, Baran, J, Callahan, A, Chepelev, L, Cruz-Toledo, J, Del Rio, N. R, Duck, G, Furlong, L. I, Keath, N, Klassen, D, McCusker, J. P, Queralt-Rosinach, N, Samwald, M, Villanueva-Rosales, N, Wilkinson, M. D, and Hoehndorf, R. (2014). The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. *Journal of Biomedical Semantics* 5, 1 (06 Mar 2014), 14. DOI:http://dx.doi.org/10.1186/2041-1480-5-14

Edgar, R, Domrachev, M, and Lash, A. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* 30 (2002), 207–210. Issue 1.

Good, B, Nanis, M, Wu, C, and Su, A. (2015). Microtask crowdsourcing for disease mention annotation in PubMed abstracts. *Pac Symp Biocomput.* (2015), 282–293.

Gottlieb, A, Hoehndorf, R, Dumontier, M, and Altman, B. R. (2015). Ranking Adverse Drug Reactions With Crowdsourcing. *J Med Internet Res* 17, 3 (23 Mar 2015), e80. http://www.ncbi.nlm.nih.gov/pubmed/25800813

Guoliang, L. (2017). Human-in-the-loop Data Integration. *Proc. VLDB Endow.* 10, 12 (Aug. 2017), 2006–2017. DOI:http://dx.doi.org/10.14778/3137765.3137833

Hadley, D, Pan, J, El-Sayed, O, Aljabban, J, Aljabban, I, Azad, T. D, Hadied, M. O, Raza, S, Rayikanti, B. A, Chen, B, Paik, H, Aran, D, Spatz, J, Himmelstein, D, Panahiazar, M, Bhattacharya, S, Sirota M., M. M. A, and Butte, A. J. (2017). Precision annotation of digital samples in NCBI's gene expression omnibus. *Scientific Data* 4, 170125 (2017). DOI:http://dx.doi.org/doi:10.1038/sdata.2017.125

Howe, J. (2006). The Rise of Crowdsourcing. *Wired Magazine* 14, 6 (06 2006). http://www.wired.com/wired/archive/14.06/crowds.html

Khare, R, Burger, J. D, Aberdeen, J. S, Tresner-Kirsch, D. W, Corrales, T. J, Hirchman, L, and Lu, Z. (2015). Scaling drug indication curation through crowdsourcing. *Database: The Journal of Biological Databases and Curation* (2015).

Khatri, P, Roedder, S, Kimura, N, De Vusser, K, Morgan, A. A, Gong, Y, Fischbein, M. P, Robbins, R. C, Naesens, M, Butte, A. J, and Sarwal, M. M. (2013). A common rejection module (CRM) for acute rejection across multiple organs identifies novel therapeutics for organ transplantation. *Journal of Experimental Medicine* 210, 11 (2013), 2205–2221. DOI:http://dx.doi.org/10.1084/jem.20122709

Sarasua, C, Simperl, E, and Noy, N. F. (2012). CrowdMap: Crowdsourcing Ontology Alignment with Microtasks. In *The Semantic Web – ISWC 2012 (Lecture Notes in Computer Science)*. Springer Berlin Heidelberg, 525–541.

Soboleva, A, Evangelista, C, Troup, D. B, Rudnev, D, Kim, I. F, Phillippy, K. H, Marshall, K. A, Tomashevsky, M, Sherman, P. M, Ledoux, P, Muertter, R. N, Edgar, R, Wilhite, S. E, and Barrett, T. (2008). NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Research* 37 (10 2008), D885–D890. DOI:http://dx.doi.org/10.1093/nar/gkn764

Vergoulis, T, Kanellos, I, Kostoulas, N, Georgakilas, G, Sellis, T, Hatzigeorgiou, A, and Dalamagas, T. (2015). mirPub: a database for searching microRNA publications. *Bioinformatics* 31 (2015), 1502–1504. Issue 9.

Wang, J, Kraska, T, Franklin, M. J, and Feng, J. (2012). CrowdER: crowdsourcing entity resolution. *Proc. VLDB Endow.* 5 (July 2012), 1483–1494.

Wang, Z, Monteiro, C, Jagodnik, K, Fernandez, N, Gundersen, G, and Rouillard, A. e. a. (2016). Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. *Nat Communications* 26, 7 (2016).

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, Bouwman, J, Brookes, A. J, Clark, T, Crosas, M, Dillo, I, Dumon, O, Edmunds, S, Evelo, C. T, Finkers, R, Gonzalez-Beltran, A, Gray, A. J, Groth, P, Goble, C, Grethe, J. S, Heringa, J, 't Hoen, P. A, Hooft, R, Kuhn, T, Kok, R, Kok, J, Lusher, S. J, Martone, M. E, Mons, A, Packer, A. L, Persson, B, Rocca-Serra, P, Roos, M, van Schaik, R, Sansone, S. A, Schultes, E, Sengstag, T, Slater, T, Strawn, G, Swertz, M. A, Thompson, M, van der Lei, J, van Mulligen, E, Velterop, J, Waagmeester, A, Wittenburg, P, Wolstencroft, K, Zhao, J, and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship.

Scientific Data 3 (2016). DOI:http://dx.doi.org/10.1038/sdata.2016.18

Zaveri, A, Kontokostas, D, Sherif, M, Bühmann, L, Morsey, M, Auer, S, and Lehmann, S. (2013). User-driven Quality Evaluation of DBpedia. In *Proceedings of the 9th International Conference on Semantic Systems (ICSS)*. ACM, 97–104.

Zaveri, A, Serrano, P, Desai, M, and Dumontier, M. (2018). CrowdED: Guideline for optimal Crowdsourcing Experimental Design. In *Proceedings of the workshop on Augmenting Intelligence with Humans-in-the-Loop co-located with TheWebConf (WWW2018)*.