Metadata Analyser: measuring metadata quality

Bruno Inácio, João D. Ferreira, and Francisco M. Couto

LaSIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal

Abstract. Scientific research is increasingly dependent on publicly available information and data sharing. So far, the best practices to ensure that data is accessible and shareable has been to deposit it in public repositories. However, these repositories often fail to implement mechanisms that measure data quality, which could lead to improving the discoverability of existing data, and contribute to its future integration. In light of this, we present Metadata Analyser, a tool that measures metadata quality. It assesses the quality of metadata by considering the proportion of terms actually linked to ontology concepts, as well as the specificity of the terms used in the metadata. Metadata Analyser applied to Metabolights, a real-world repository of metabolomics data, and results show that the tool successfully implements the proposed measures, that there is indeed a lack of effort in the annotation task, and that our tool can be used to improve this situation. Metadata Analyser's frontend is available at `http://masterweb-metadataanalyser.rhcloud.com`.

Keywords: Metadata quality, Data sharing, Ontologies, Specificity, Coverage

1 Introduction

A significant portion of scientific research has recently become producer and consumer of large volumes of data, from multiple sources and in various formats [3]. In this scenario, data sharing takes an important role in the success of any scientific endeavour, as it allows scientific advances to "stand on the shoulders" of previous works, either performed by the authors themselves or by other teams [5]. This can only happen if data is properly integrated (i.e. categorized and organized in meaningful groups that reflect the data's similarities and differences), which enables information to be retrieved automatically [1]. However, ensuring data integration is a non-trivial task, sometimes regarded as non-scientific, and costly both in terms of human and time resources. Thus, it tends to be postponed, or even neglected.

The goals of this work are thus threefold: (*i*) to propose two measures of metadata quality, (*ii*) to implement a tool that is able to evaluate these measures in a public repository, and (*iii*) to show that these measures are valid and significant in a real-world scientific repository.

2 Materials and Methods

We propose two measures of metadata quality: (*i*) the proportion of annotations in the metadata file that link to an ontology concept, and (*ii*) the average specificity of those ontology concepts.

The dependence on a notion of ontology is justified because ontologies are regarded by the biomedical community as standard representations of knowledge [6]. An ontology can be thought of as a graph that connects nodes (the relevant concepts) with edges (the relations between the concepts). For example, CHEBI contains statements about small molecules such as "carbon dioxide is-a greenhouse gas" and "glucose is-a carbohydrate". On the one hand, relying on ontologies allows us to base our measures in community-approved knowledge; on the other hand, an ontology concept is unambiguous, traceable, and represents a quantum of information that can be shared between the scientific community without potential for misinterpretation, enabling and enhancing data sharing.

2.1 Term coverage

Usually, metadata files contain a mixture of ontology concepts and natural language terms. Since data sharing relies on the ability to find and retrieve information with automatic tools, ensuring that metadata is expressed as reference to ontology concepts improves its potential for being found in the future.

The first measure of metadata quality, therefore, is **term coverage**. It is the ratio between the number of annotations that refer to ontology concepts and the total number of annotations in the metadata file.

2.2 Semantic specificity

Each ontology concept contains a certain amount of information, which can be measured by its specificity. More specific concepts have a higher information content and thus contribute with more specific knowledge to the metadata file. As such, we propose **semantic specificity**, a measure that reflects the average specificity of the concepts in the metadata file.

For a given concept, we consider the path from itself up to the root of the ontology and all the paths from itself down to the leaves of the tree. Let $T = \{t_1, t_2, \cdots, t_n\}$ be the set of ontology concepts found in a metadata file. For each t in T, its specificity $S_{\text{concept}}(t)$ is computed as

$$S_{\text{concept}}(t) = \frac{A(t)}{A(t) + D(t)} \tag{1}$$

where $A(t)$ is the number of ascendant concepts up from t and $D(t)$ is the average distance between t and all its leaf descendants, measured in number of edges. Concepts with low specificity are located at the top of the tree (near the root). A non-specific annotation contains small amounts of knowledge and is a weak descriptor of the contents of the resource: a more specific descendant concept

would be a better descriptor, since it would provide a more specific semantics to the resource and thus increase its potential for future integration. Concepts with high specificity are located near the leaves of the ontology, and correspond to informative annotations.

In order to determine the semantic specificity of an annotated resource, we average the specificity of the concepts in its metadata.

2.3 Motivation for the measures

The two measures presented reflect the quality of the metadata associated with a resource. On the one hand, high coverage by ontology concepts in a resource's metadata file reflects a greater amount of computationally meaningful knowledge provided about that resource. On the other hand, as demonstrated above, the highest the specificity value of a concept, the better it is in describing the content of the resource. Therefore, high values for these measures enhance the meaning and discoverability of the data to those who wish to use it.

2.4 Metadata Analyser Architecture

To automatically evaluate the quality of a metadata file based on the quality measures described previously, we designed an architecture to analyse and evaluate the metadata file contents, Metadata Analyser. This is a modular architecture that can be adapted to other domains. For example, one module is responsible for reading Metabolights metadata files (see "Case Study" below), and another for computing the quality measures. Both modules can be exchanged by other ones, specific to other repositories or designed to compute other measures.

The tool is composed of the following layers:

1. An **interface layer** that interacts with the user by requesting a metadata file, informing the user on the analysis progress, and outputting the result.
2. An **application layer** that analyses the metadata file and evaluates the annotations found therein.
3. A **data layer** that holds the ontologies in local databases.
4. A web **API layer** that connects the interface layer to the application layer, coded in commonly used web technologies.

Source code is available at `https://github.com/lasigeBioTM/MetadataAnalyser`.

3 Case study

To evaluate our work, we applied Metadata Analyser to Metabolights, a database of metabolomics experiments [4, 9]. Metabolomics is the study of the chemical processes that occur in life-related contexts, usually within a cell or in its surroundings. This data often refers to a large number of scientific domains, as it can be cross-species and cross-technique, while covering metabolite structures, biological roles, locations and concentrations, as well as experimental factors.

Metabolights stores metadata associated with the actual data describing the information in each resource. For example, the metadata of the resource called "LCMS analysis of seven apple varieties with a leaking chromatographic column" claims that the data was collected through "liquid chromatography" and "mass spectroscopy", and that the study factors include "Sample type", "Apple number", etc.[1]. These pieces of metadata are collected (by the researcher or the curator) using the ISA-tools software suite [8]; in particular, metadata is saved in the ISA-TAB format, which has the built-in ability to refer to ontology concepts. At the moment of this study, the repository had 161 resources.

This repository has been developed and maintained by the EBI since 2012, and is therefore a relatively recent addition to the panorama of knowledge stores in the biomedical domain. Its use of ontology concepts in the metadata files has been advocated since the beginning, since Metabolights has always recommended its users to prepare and submit the data with that possibility in mind.

Our evaluation consisted of three steps: we first evaluated the measures on all the resources of the Metabolights database, then manually evaluated the results obtained in a selection of resources, and finally performed an evaluation of the metadata quality before and after a curation step performed by a team of metabolomics experts.

3.1 Metadata quality in Metabolights

From the 161 resources, 6 did not contain any ontology annotation, i. e. the semantic specificity was 0.0. The average coverage was 0.25. Only 9 resources show a coverage of 0.40 or higher. The average semantic specificity was 0.81. Histograms of distributions are shown in Fig. 1. From these distributions, we can see that more effort is put into the semantic characterization of the resources than into making sure the terms are actually from a reference ontology.

Discarding the 6 resources with no ontology annotation, a small negative correlation was found between the two measures (see Fig. 2), with a Pearson correlation coefficient of -0.28 (corresponding to a p-value of 5.1×10^{-4}). This trend is only slightly negative, at best, even if statistically significant. Nonetheless, we argue that this may be related to the fact that the tasks of (*i*) looking for the most specific concept to use in the annotation and (*ii*) finding all the locations in the metadata file where an ontology concept can be used take time and thus cannot both be performed perfectly given time constraints.

The most relevant conclusion is that the semantic annotation of metadata describing the Metabolights resources is still far from the desired state of affairs.

3.2 Manual evaluation of the measures

To validate the correctness of the implementation, we randomly selected 6 resources and calculated the two quality measures manually. Results were compared both with a manual verification as well as with a previous work [7]. The

[1] See `www.ebi.ac.uk/metabolights/MTBLS99`

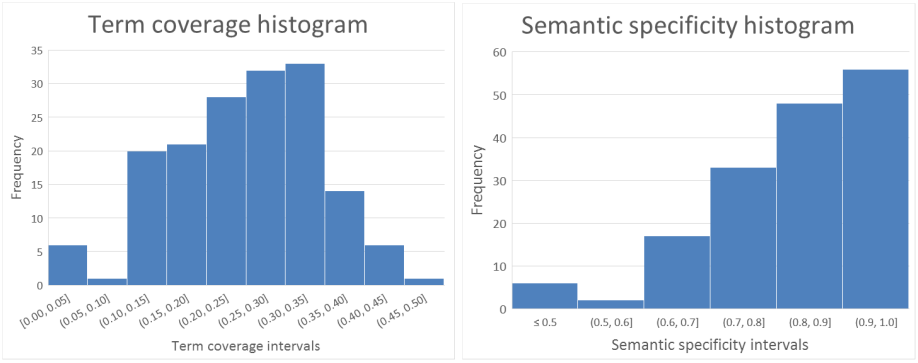

Fig. 1. The histograms of the distribution for the two measures of metadata quality in Metabolights. On the left, the distribution for term coverage; on the right, the distribution for semantic specificity.

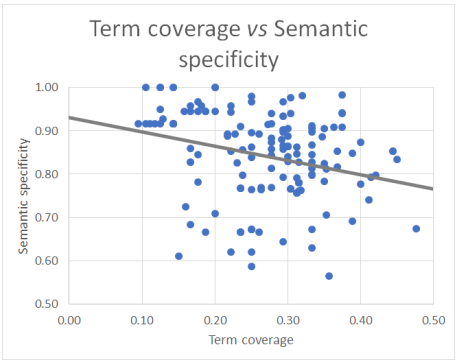

Fig. 2. The correlation between the term coverage and semantic specificity of all the metadata files for all resources in Metabolights.

previous work computes the same results based on a web API that can be used to query biomedical ontologies (BioPortal). The metadata quality measures for the present work, the previous work and the manual validation are presented for the selected resources in Table 1.

These results show that values from the Metadata Analyser are close to the ones obtained from a manual computation. The only significant difference is that in two of the resources the term coverage is lower for our tool. This reduced amount of ontology concepts found in the metadata file leads to an artificial increase in the semantic specificity since the concepts that were exclusively found in the manual validation are non-specific. This limitation in our methodology is due to the fact that not all ontologies used to annotate the resources were included in the local database (e. g. one of the concepts used in MTLBS166 is from MeSH, but since Metadata Analyser did not include it in the database, it failed to compute a semantic specificity for the concept).

Table 1. Results from the manual validation

Resource	Manual results		This work		Previous work	
	Semantic specificity	Term coverage	Semantic specificity	Term coverage	Semantic specificity	Term coverage
MTBLS1	0.89	0.30	0.88	0.30	0.00	0.00
MTBLS36	0.96	0.17	0.96	0.17	0.00	0.00
MTBLS88	0.75	0.31	0.75	0.31	0.69	0.75
MTBLS110	0.84	0.28	0.91	0.14	0.87	0.50
MTBLS137	0.94	0.20	0.94	0.20	0.87	0.37
MTBLS166	0.60	0.23	1.00	0.14	0.00	0.54

Table 2. Results from the pre- and post-curation analysis. **N** is the number of annotations that refer to ontology concepts.

Resource	Pre-curation			Post-curation		
	Semantic specificity	Term coverage	N	Semantic specificity	Term coverage	N
MTBLS286	0.00	0.00	9	0.96	0.25	16
MTBLS287	0.92	0.22	9	0.96	0.25	16
MTBLS288	0.92	0.22	9	0.87	0.25	16

The results from the previous study show a small semantic specificity value compared with the manual validation. They also present higher values of term coverage because that study uses a different algorithm to compute it. Finally, given that the previous work relies on a service over the web and that our methodology uses a local knowledge base, it is unsurprising to notice that Metadata Analyser is faster. In fact, it computes the results, on average, more than 10,000 times faster than the previous work (results not shown).

3.3 Evolution of metadata quality

To study the effect of an expert-driven curation process, we applied our measures of metadata quality to consecutive versions of three resources in the repository. The development team of Metabolights provided the pre- and post-curation versions of the resources MTBLS286, MTBLS287 and MTBLS288. The numbers for these three resources are presented in Table 2.

There are three general differences between the pre- and post-curation process. First, we notice an increase in the number of annotations, from 9 to 16 in each of the three resources. Furthermore, even though there are more annotations, we observe an increase in the amount of annotations that make use of ontology concepts, since the term coverage measure increases from an average of 0.15 to an average of 0.25. Finally, we also observe a mild increase in specificity. These three facts suggest that curators are able to increase the amount

of machine-readable metadata that is available for each resource as well as its information content, measured by the semantic specificity. This experiment suggests that our measures do indeed capture a notion of metadata quality, since both experienced an increase after being handled by curation experts.

It is interesting to notice that the resources MTBLS287 and MTBLS288 already presented high values of semantic specificity prior to curation (higher than the full repository average), which means expert-driven curation could not improve then by much. That the curation process did not significantly alter them suggests that the annotation from the authors was already of high quality.

4 Discussion

There is an increasing usage of linked data techniques in Life and Health Sciences and many of them using biomedical ontologies, however to enhance their impact and value they need to produce high quality semantic descriptions of the data [2].

This work proposes two measures of metadata quality: (*i*) semantic specificity, which measures the average specificity of the ontology concepts referred to in the metadata and (*ii*) term coverage, which measures the proportion of annotations associated with actual ontology concepts. Based on them, we developed Metadata Analyser, an application that assesses metadata quality. It was evaluated by comparing its results both with a manual evaluation and a previous tool: results suggest that our measure corresponds to the expectations for metadata quality, as they increase after an expert-driven curation process. The tool is also significantly faster than the previously presented one and more accurate.

The major conclusion is that the two proposed measures can effectively measure the effort put into the semantic annotation of digital resources. This includes the annotation of a resource's metadata with explicit references to concepts from ontologies accepted by the community as machine-readable, standard representations of a domain of knowledge.

The results obtained from the Metabolights case study confirm the problem that motivated the creation of this tool, as we observe a weak term coverage (average of 0.25) and we hope it can be applied in existing repositories as a way to provide users feedback on their metadata quality, as well as motivating the general scientific community to increase their annotation efforts, so that we can, as a whole, spend more effort in ensuring proper data integration.

4.1 Helping with scarce semantic integration

One possible cause behind the poor state of affairs in semantic annotation is of social nature, rather than technical [3]: metadata files are usually compiled by the authors of the data, who (*i*) may not know the ontologies that contain the concepts they need, (*ii*) do not fully know the structure of the ontologies in order to perform annotation with the appropriate specific terms, (*iii*) lack the proper skills to carry on the annotation process because of the technical difficulties associated with this task, (*iv*) do not consider data sharing to be relevant, or (*v*) consider that the cost of ensuring proper semantic integration outweighs the benefits.

The apparent correlation between specificity and coverage shown in Fig. 2 shows that a general effort exists to ensure specific concepts are used in the metadata, but not to ensure that ontology concepts are used throughout the metadata files, which suggests that indeed the perceived benefits may not significantly counterbalance the time costs of doing so. Without mandatory high quality metadata publication, it becomes difficult or even impossible to create automatic information retrieval mechanisms that can handle these author-created metadata files. While the short-term solution is to leverage on curators to help increase metadata quality, in a long-term scenario we wish to empower data creators with a means to measure the quality of their metadata, who would then use this feedback to improve metadata quality and thus the integration potential of the data.

Acknowledgments This work was supported by FCT through funding of the LaSIGE Research Unit, ref. UID/CEC/00408/2013. We thank the EBI team in charge of the development and maintenance of Metabolights for their support in this study.

References

1. Baker, M.: Quantitative data: learning to share. nature methods 9(1), 39 (2012)
2. Barros, M., Couto, F.M., et al.: Knowledge representation and management: a linked data perspective. IMIA Yearbook pp. 178–183 (2016)
3. Couto, F.M.: Rating, recognizing and rewarding metadata integration and sharing on the semantic web. In: Proceedings of the 10th International Conference on Uncertainty Reasoning for the Semantic Web-Volume 1259. pp. 67–72 (2014)
4. Haug, K., Salek, R.M., Conesa, P., Hastings, J., de Matos, P., Rijnbeek, M., Mahendraker, T., Williams, M., Neumann, S., Rocca-Serra, P., Maguire, E., González Beltrán, A., Sansone, S., Griffin, J.L., Steinbeck, C.: Metabolights - an open-access general-purpose repository for metabolomics studies and associated metadata. Nucleic Acids Research 41(Database-Issue), 781–786 (2013), http://dx.doi.org/10.1093/nar/gks1004
5. Innovative Medicine Initiatives: IMI2: 9th Call for proposals. http://www.imi.europa.eu/sites/default/files/uploads/documents/IMI2Call9/IMI2_Call9_TopicsText.pdf (April 2016), accessed April 2016
6. Noy, N.F., McGuinness, D.L.: Ontology development 101: A guide to creating your first ontology (2001)
7. Ramos, C., Louro, M., Santos, M., Couto, F.M.: Knowledge ratings in metabolights. arXiv preprint arXiv:1604.07997 (2016)
8. Rocca-Serra, P., Brandizi, M., Maguire, E., Sklyar, N., Taylor, C., Begley, K., Field, D., Harris, S., Hide, W., Hofmann, O., et al.: ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. Bioinformatics 26(18), 2354–2356 (2010)
9. Salek, R.M., Haug, K., Conesa, P., Hastings, J., Williams, M., Mahendraker, T., Maguire, E., González Beltrán, A., Rocca-Serra, P., Sansone, S., Steinbeck, C.: The metabolights repository: curation challenges in metabolomics. Database 2013 (2013), http://dx.doi.org/10.1093/database/bat029