View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Crossref

Information Services & Use 34 (2014) 279–286
DOI 10.3233/ISU-140746
IOS Press

Metadata for Big Data: A preliminary investigation of metadata quality issues in research data repositories

Dimitris Rousidis [a], Emmanouel Garoufallou [b,*], Panos Balatsoukas [c] and Miguel-Angel Sicilia [a]

[a] *University of Alcala, Madrid, Spain*
[b] *Alexander Technological Educational Institute of Thessaloniki, Thessaloniki, Greece*
[c] *University of Manchester, Manchester, UK*

Abstract. Data-driven approaches to scientific research have generated new types of repositories that provide scientists the means necessary to store, share and re-use big data-sets generated at various stages of the research process. As the number and heterogeneity of research data repositories increase, it becomes critical for scientists to solve data quality problems associated to the data-sets stored in these repositories. To date, several authors have been focused on the data quality issues associated to the data-sets stored in the repositories, yet there is little knowledge about the quality problems of the metadata used to describe these data-sets. Metadata is important for the long-term sustainability of research data repositories and data re-use. The aim of the research reported in this paper was to identify the data quality problems associated with the metadata used in the Dryad data repository. The paper concludes with some recommendations for improving the quality of metadata in research data repositories.

Keywords: Big Data, data quality, descriptive analysis, metadata, research objects, e-Research

1. Introduction

The availability of scientific data (big data) and the emergence of cloud computing have radically changed research activities. eScience and eResearch applications have extended traditional forms of scholarly e-infrastructure (such as institutional repositories and digital libraries) and enabled scientists to store, access, analyze, use and share datasets generated at various stages of the research process [5]. Given the big volume and diversity of scientific data, research repositories are becoming integral part of the communication and collaboration process between scientists and research groups. Yet problems related to data quality may impede the process of analysing, integrating and re-using heterogeneous datasets.

Although several researchers have been focused on the development of new methods to improve the quality of the data stored in research data repositories, e.g. [19,21], there is little research on the data quality issues of the metadata used to describe and annotate datasets in this type of repositories. The use of complete and accurate metadata is important for several processes, including the re-use and sharing

* Corresponding author. E-mail: mgarou@libd.teithe.gr.

0167-5265/14/$27.50 © 2014 – IOS Press and the authors.
This article is published online with Open Access and distributed under the terms of the Creative Commons Attribution Non-Commercial License.

of research datasets among scientists; the application of digital curation and data provenance strategies; and the analysis of the contents of research data repositories.

The aim of the research reported in this paper was to identify the data quality problems associated with the metadata used in a research data repository, called Dryad. Being this a first attempt to analyse the metadata used in research data repositories [17], the objectives were chiefly exploratory, concretely:

- To perform a descriptive analysis of the metadata elements used in the Dryad repository; and
- To identify the main metadata quality issues.

This paper is structured as follows: First, a literature review on previous work is presented and the Dryad repository is described. Then the methodology and results of this study are presented. Finally, conclusions and suggestions for further research are reported on the last section of the paper.

2. Background

Data quality is defined as the state of completeness, validity, consistency, timeliness and accuracy that makes data suitable for a specific use [6]. Dekkers et al. [3] states that data is of high quality "if they are fit for their intended uses in operations, decision making and planning". There is no distinction between the data and metadata quality considerations [3]. The growth, proliferation and evolution of digital objects are accompanied by an analogous transformation of their metadata which causes a consistency issue affecting at the same time their quality [12,13]. In many cases, the larger the dataset, the greater the probability a problem will emerge [2]. Also, research has shown that there are effects of discipline of the quality of metadata, thus suggesting a cultural dimension on data quality (e.g. [1]).

2.1. Metadata quality issues in repositories

In an early study, Sokvitne [18] examined the effectiveness of the metadata elements of the Dublin Core for information retrieval. The study showed several problems especially with popular elements. In particular, the authors found that the DC.title and the DC.subject elements did not add any value for retrieval purposes, while the DC.creator, DC.publisher and DC.contributor elements presented inconsistent name formats [18] concluded the study by questioning the suitability of the Dublin Core for information retrieval unless various problematic issues were resolved. The main issues were that the elements should be populated and used correctly, while precise instructions, descriptions and rules should be set. In addition to general metadata standards, like Dublin Core, researchers have examined metadata quality in the context of specialized repositories, such as architectural repositories [15]; digital libraries [20]; agricultural collections [23]; health databases [16]; and learning object repositories [14]. Despite the heterogeneity of metadata and repositories examined in these studies, there is a common set of metadata issues that appears to influence quality. For example, Barton et al. [2] outlined the areas where metadata element problems most commonly occur. These were: Spelling, abbreviations and other similar data entry errors and ambiguities; Inconsistencies with the Author and other contributor/creators metadata elements; Use of multiple Title elements; Use of correct and standardised terminology (in the case of the Subject metadata element); Inconsistencies with the format of the Date metadata element.

While the aforementioned studies provided some evidence about the type of metadata quality issues that apply in the context of information-driven repositories (such as digital libraries and repositories of information resources), there is little known about research data repositories.

2.2. The DRYAD repository

Dryad is an open access repository that permits scientists – in pure sciences and medicine – to store, search, retrieve and re-use research data associated to their scholarly publications. Data are deposited as files with permanent identifiers (DOIs) and metadata. Collections of related files may be grouped into data packages with metadata describing a combined set of files. Currently the repository contains approximately 4500 data packages associated with scholarly articles published in almost 300 international journals [4].

Dryad's developers, by using the Singapore framework metadata architecture in a DSpace environment via an Extensible Markup Language (XML) schema [11,22] and HIVE (Helping Interdisciplinary Vocabulary Engineering), implemented the infrastructure so that the automatically generated metadata inherit characteristics from their original sources by harvesting keywords assigned by authors and controlled vocabularies – ontologies [7].

Greenberg et al. initially [10] and [9] performed quantitative studies which were focused on the reusability of the repository's metadata. The main findings of the studies, based on the study of two Dryad workflows, were that 8 out of 12 metadata elements (contributor, corresponding author, identifier citation, subject, publication name, description, relation is referenced by, title) had a reuse at 50% or greater. The researchers concluded that reuse was more common in the case of traditional bibliographic elements; and the generation of more accurate metadata earlier in the metadata workflow is necessary. As opposed to the studies conducted by Greenberg and colleagues on the re-usability of metadata, the research reported in this paper is focused on the identification of the main quality issues related to analysis of the metadata elements of the Dryad repository and how these may affect re-use and the analysis of the contents of research data repositories.

3. Methodology

A mechanism that involved the downloading of the metadata elements from the Dryad and their transformation to a proper format for analysis was employed. Metadata was harvested in January 2014. At this point the Dryad was holding 4,557 packages, 13,638 data files, 287 journals, 16,595 authors and 751,658 times an instance of the repository was downloaded. The *Open Archives Initiative Protocol for Metadata Harvesting* (*OAI-PMH*) *Validator & Data Extraction Tool* was used for the metadata harvesting.[1] A total of 516 XML files were downloaded (135 MB). The XML files were merged into a single file using *Mergex*, a command line tool for merging XML files.[2] Finally, a method to use and analyze the data from the XML files had to be employed. Due to the descriptive nature of the statistical analysis performed it was decided to analyze the data using Microsoft Excel 2010 (as opposed to the use of more data analysis tools, like R). Therefore the XML to CSV Conversion Tool[3] was used to transform the XML files into CSVs and import these to Excel. It is worth mentioning that importing directly the XML file to Excel provided very frustrating results. The converter provided 19 CSV files, each corresponding to a different metadata element: (i) contributor, (ii) coverage, (iii) creator, (iv) date, (v) dc, (vi) format, (vii) header, (viii) identifier, (ix) listRecords, (x) metadata, (xi) record, (xii) relation, (xiii) request, (xiv) responseDate, (xv) resumptionToken, (xvi) setSpec, (xvii) subject, (xviii) title and (xix) type. A selected

[1] http://validator.oaipmh.com/.

[2] https://code.google.com/p/mergex/.

[3] http://xmltocsv.codeplex.com/.

sample of metadata elements was analyzed. These were: contributor, creator, date, subject, type, relation, coverage, dc, identifier and title. However, since the focus of this paper is on the presentation of the data quality issues, rather than a detailed description of the contents of the Dryad repository, a small subset of three metadata elements is presented: Creator, Type and Date. These elements represent typical cases where data quality issues can impede the quantitative and qualitative analysis of the Dryad repository, as well as the re-use of the data stored in the repository.

4. Results

A significant number of major data problems were identified in the case of the Creator, Date and Type metadata elements. The methodology for the conversion and analysis of data was quite problematic. The noise accumulation and the incorrect assignment of the records to the proper fields were the main problems with the conversion. Data irrelevant to the fields and data misplaced made the initial files difficult to analyze and manipulate making a manual intervention essential. Furthermore, the quality of the data, an issue completely irrelevant with the conversion procedure, was not the anticipated one taking into account Dryad's development.

4.1. Creator

The number of contribution per author is depicted on Table 1. In total 16,567 authors, just 28 less than the number of authors referred at the Dryad's webpage, contributed 86,087 objects. As it is shown in Table 1, the majority of creators (i.e. authors of the research objects) contributed between one to five research objects in the repository.

Out of 16,568 records, a total of 1,443 (8.71%) demonstrated the following issues:

- Additional names: Many authors were entered with just their first name. The problem emerged in 614 (42.55%) cases. Also, this percentage included cases where an author's additional names or surname were added as a different record.
- Use of initials: Another major issue was the use of initials instead of the whole name (11.64%).
- Different languages: Almost twelve percent (12.06%) of problems occurred with this quality issue. There are numerous variations for writing a name in non-English language. Trying to convert a name

Table 1
Amount of objects published by each contributor

Number of contributions	Number of creators	Number of contributions	Number of creators	Number of contributions	Number of creators
1	1,422	11	248	21–30	286
2	6,131	12	225	31–40	128
3	2,282	13	137	41–50	59
4	1,541	14	144	51–60	24
5	1,060	15	84	61–70	11
6	773	16	92	71–80	10
7	601	17	100	81–90	10
8	396	18	82	91–100	13
9	362	19	55	>100	2
10	242	20	47	**Total**	**16,567**

by the English alphabet may be problematic as there are many symbols that do not exist (for example due to different accents). The most frequent mistakes were made in French, Spanish, Scandinavian, German, Chinese, Balkan and East Europe names. The use of short names and diminutives were also included in this category.
- Invalid input: A 2.56% of errors occurred due to typos. Typical examples of errors in this category occurred when a first name was missing or when the first name was inserted at the surname field.
- Dots and commas. The second most frequent type of an error (23.08%) involved the absence of dots or the use of commas at the end of initials.
- Spacing: Invalid creator entries existed as in a few cases (2.36%) no or too many spaces were inserted during the name input.
- Miscellaneous: Issues like using irrelevant text (e.g. et al., PhD, status, code, etc.) were grouped in this category (0.83%).

4.2. Date

This metadata element was assigned to various types of a date like date accessed, date available and date issued. For the purpose of this analysis we gathered the dates corresponding to the date issued of the 43,453 objects in the repository. According to the cataloging guidelines of Dyad's wiki,[4] the DC.date.issued is the official date of publication, inherited by the dataset; i.e. the date of the formal issuance of the resource. The distribution per year is depicted in Table 2.

The growth of the Dryad Repository over time based on the objects' issued date is shown in Fig. 1. It should be noted that there are two abnormalities in the flow of the records within the repository. On October 2010 2,572 publications where entered when the previous month the amount was a few dozens and on April 2011 the number was skyrocketed to around 23,000, more than half (52.67%) of the total publications of the repository. Since it is highly unlike that on a single month half of the input of the repository was published it seems that there is mix-up with date issued and the date input in Dryad.

Major quality issues occurred also in the case of the DC.date element. Most of these included:

- Lack of consistency in the format. For example, four dates from 1900–1904, 321 dates after the date that the metadata was harvested, 476 dates equal to 1/1/9999 and 40 dates that were blank or with text; and
- Lack of standardization of the date format. Table 3 shows the inconsistency in the length of this element's values which varies from being blank to 20 characters long.

Table 2
Contributions per creator

Date	Amount of contributions	Date	Amount of contributions	Date	Amount of contributions
1995	1	2002	10	2009	416
1996	10	2003	11	2010	3,172
1997	10	2004	13	2011	25,411
1998	59	2005	12	2012	5,035
1999	50	2006	13	2013	8,005
2000	17	2007	27	1/1/2014–9/1/2014	176
2001	67	2008	97	Invalid input	841

[4]http://wiki.datadryad.org/Cataloging_Guidelines_2009.

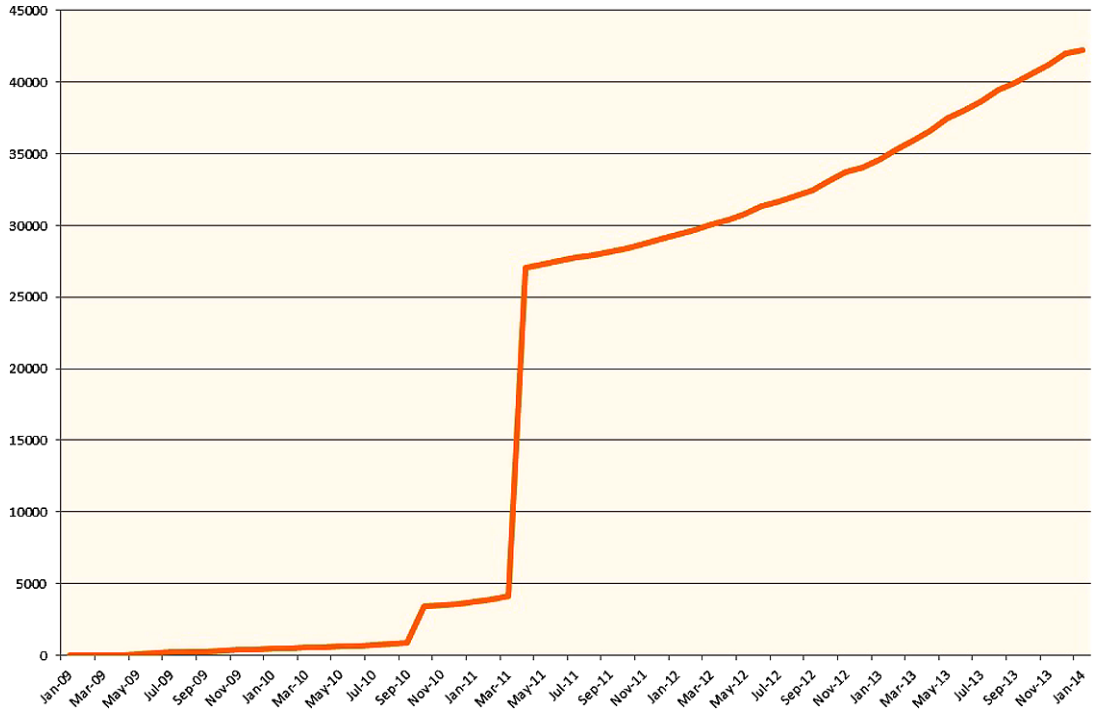

Fig. 1. Growth over time. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/ISU-140746.)

Table 3
Length of issued date

Length	Count	Example
4	156	2009
6–7	163	2009-03
8–10	42,590	2009-09-07
20	503	2009-10-01T10:19:28Z
Various	41	Blanks, unacceptable format

4.3. Type

A total of 53,598 records were retrieved for the DC.type element and their distribution is shown in Table 4. In the type field is shown the exact text that was found in the type field, except from blank were actually there was nothing inserted.

As shown in Table 4, the Dataset type holds the vast majority of the DC.type element with 70.17%, followed by the Article with 8.30%. However, it is apparent that there are types in the table that should not appear in a first place like custom, blanks, none, oneyear, protocol and untilArticleAppears. According to the Dryad's Cataloging Guidelines[5] the DC.type element is the "Code indicating the type of file. This is automatically detected by DSpace, but can be modified manually". Obviously there are issues with the automatic detection and irrelevant/unrelated with the DC.type entries were inserted. If

[5]http://wiki.datadryad.org/Cataloging_Guidelines_2009.

Table 4
Type distribution of objects

Type	Amount	Percentage %	Type	Amount	Percentage %
Activity	4	0.007	Image	62	0.116
Article	4,451	8.304	Map	1	0.002
Book	3	0.006	None	4,086	7.623
Blank	4	0.007	Oneyear	830	1.549
Custom	109	0.203	Protocol	11	0.021
Dataset	36,708	70.167	UntilArticleAppears	6,429	11.995

we cleaned the data and left only the suitable type files, then 42,129 records would remain and the percentages would change as follows: Activity 0.009%; Article 10.565%; Book 0.007%; Dataset 89.269%; Image 0.147%; and Map 0.002%. Consequently, nearly 90% of the stored files were datasets and nearly 10% were articles.

Almost twenty percent (21.4%) of the records in the DC.type metadata element was jargon or blank or completely irrelevant to the element. The absence of data control and quality was more than obvious. As with the other elements a mechanism that will allow only correct data entry has to be employed.

5. Conclusions

The purpose of this study was to illustrate some of the main data quality issues associated with the use of metadata in the Dryad Repository. Our study validated this assumption as major issues were identified in the case of the DC.creator, DC.Date and DC.type metadata elements. In addition to the reusability of research data, addressing issues related to data quality of metadata in the Dryad repository is important for the accurate analysis and monitoring of the growth of the repository. In order to address the aim of this study all metadata from the Dryad repository were harvested and analyzed. A plethora of data misuse issues were identified; issues that constitute the data inappropriate for text mining or data mining purposes. A mechanism that secures the metadata input from the issues that we identified needs to be employed. Data control would make repositories far more appealing and sustainable. For example, a solid format for the creators' names should be specified. Each creator and contributor should be assigned with a unique ID that would hold their full name (e.g. http://orcid.org/). When requesting an entry of the full name at the repository this unique ID should be inserted. If for any reason the creator wishes to change the name, then all of the records related with the name should be updated automatically, through the unique ID. In the case of dates, input should follow the same format (e.g. dd-mm-yyyy). Validation rules must be applied when each date is entered (e.g. it is more than obvious that a date cannot be posterior than the current date or prior than the creator's birthday). In the case of the Type metadata element, inconsistencies can be fixed through the use of pre-defined lists of values for authors to select from. Finally, we validated the fact that poor quality metadata have drastic impact on the results of the quantitative analysis of the repository's metadata elements.

Our future work will be focused on the analysis of the remaining metadata elements of the Dryad repository. More elaborate statistical analysis by using R will be employed and data mining and text mining techniques will be applied to provide a better understanding of the repository's data, to identify associations, clusters or hidden patterns and to develop novel visualisations for displaying the contents of research data repositories based on the analysis of their metadata [8].

References

[1] P. Balatsoukas, A. O'Brien and A. Morris, The effects of discipline on the application of learning object metadata in UK Higher Education: the case of the JORUM repository, *Information Research* **16**(3) (2011).

[2] J. Barton, S. Currier and J.M.N. Hey, Building quality assurance into metadata creation: an analysis based on the learning objects and e-prints communities of practice, in: *Proceeding of 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice – Metadata Research and Applications*, 2013, pp. 39–48.

[3] M. Dekkers, N. Loutas, M. De Keyzer and S. Goedertier, Open data and metadata quality, 2013, available at: https://joinup.ec.europa.eu/sites/default/files/D2.1.1%20Training%20Module%202.2%20Open%20Data%20Quality_v0.09_EN.pdf [25/01/2014].

[4] Dryad Digital Repository, Frequently asked questions, available at: http://datadryad.org/pages/faq [29/12/2013].

[5] E. Garoufallou and C. Papatheodorou, Editorial – Special issue on Metadata for e-Science and e-Research, *International Journal of Metadata Semantics and Ontologies* **9**(1) (2014), 1–4.

[6] K. Gordon, *Principles of Data Management*, Facilitating Information Sharing, 2007.

[7] J. Greenberg, Theoretical considerations of lifecycle modeling: an analysis of the Dryad repository demonstrating automatic metadata propagation, inheritance, and value system adoption, *Cataloguing & Classification Quarterly* **47**(3,4) (2009), 380–402.

[8] J. Greenberg and E. Garoufallou, Change and a future for metadata, in: *Metadata and Semantic Research: 7th Research Conference, MTSR 2013*, Thessaloniki, Greece, November 19–22, 2013, Proceedings. Communications in Computer and Information Science (CCIS), E. Garoufallou and J. Greenberg, eds, Vol. 390, 2013, pp. 1–5.

[9] J. Greenberg, S. Swauger and E.M. Feinstein, Metadata capital in a data repository, in: *Proceedings of the International Conference on Dublin Core and Metadata Applications*, 2013, pp. 140–150.

[10] J. Greenberg and T. Vision, The Dryad repository: A new path for data publication in scholarly communication, OCLC, Dublin, Ohio, 2011, available at: http://www.oclc.org/research/news/2011-03-24.htm [22/1/2014].

[11] J. Greenberg, H.C. White, S. Carrier and R. Scherle, A metadata best practice for a scientific data repository, *Journal of Library Metadata* **9**(3) (2009), 194–212.

[12] D. Lee, Practical maintenance of evolving metadata for digital preservation: Algorithmic solution and system support, *International Journal on Digital Libraries* **6**(4) (2007), 313–326.

[13] X. Ochoa and E. Duval, Automatic evaluation of metadata quality in digital repositories, *International Journal on Digital Libraries* **10**(2) (2009), 67–91.

[14] N. Palavitsinis, N. Manouselis and S. Sanchez-Alonso, Metadata quality in digital repositories: empirical results from the cross-domain transfer of a quality assurance process, *Journal of the Association of Information Science and Technology* **65**(6) (2014), 1202–1216.

[15] E. Park, Building interoperable Canadian architecture collections: Initial metadata assessment, *The Electronic Library* **25**(2) (2007), 207–218.

[16] D. Rousidis, E. Garoufallou, P. Balatsoukas, K. Paraskevopoulos, S. Asderi and D. Koutsomicha, Metadata requirements for repositories in health informatics research: evidence from the analysis of social media citations, in: *MTSR 2013*, E. Garoufallou and J. Greenberg, eds, 2013, pp. 246–257.

[17] D. Rousidis, E. Garoufallou, P. Balatsoukas and M.-A. Sicilia, Data quality issues and content analysis for research data repositories: The case of Dryad, ELPUB2014. Let's put data to use: digital scholarship for the next generation, in: *18th International Conference on Electronic Publishing*, Thessaloniki, Greece, 19–20 June 2014, available at: http://elpub.scix.net/cgi-bin/works/Show?106_elpub2014.

[18] L. Sokvitne, An evaluation of the effectiveness of current Dublin core metadata for retrieval, in: *Proceedings of VALA 2000*, Victorian Association for Library Automation, Melbourne, 2000.

[19] B. Stvilia, C. Hinnant, S. Wu, A. Worall, D.J. Lee, K. Burnett, G. Burnett, M. Kazmer and P. Marty, Research project tasks, data and perceptions of data quality in a condensed matter physics community, *Journal of the Association for Information Science and Technology*, in press.

[20] M.H. Vinagre, L.G. Pinto and P. Ochôa, Revisiting digital libraries quality: a multiple-item scale approach, *Performance Measurement and Metrics* **12**(3) (2011), 214–236.

[21] N.G. Weiskopf and C. Weng, Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research, *JAMIA* **20**(1), 144–151.

[22] H. White, S. Carrier, A. Thompson, J. Greenberg and R. Scherle, The Dryad data repository: A Singapore framework metadata architecture in a DSpace environment, in: *DC 2008, The 2008 International Conference on Dublin Core and Metadata Applications*, Berlin, 2008.

[23] T. Zechocke and J. Beniest, Adapting a quality assurance framework for creating educational metadata in an agricultural learning repository, *The Electronic Library* **29**(2) (2011), 181–199.