Evaluation of Metadata in Research Data Repositories: The Case of the DC.Subject Element

Dimitris Rousidis[1,2](✉), Emmanouel Garoufallou[1,2], Panos Balatsoukas[3],
and Miguel-Angel Sicilia[1]

[1] University of Alcala, Madrid, Spain
drousid@gmail.com, mgarou@libd.teithe.gr, msicilia@uah.es
[2] Alexander Technological Educational Institute of Thessaloniki, Kentriki Makedonia, Greece
[3] University of Manchester, Manchester, UK
panagiotis.balatsoukas@manchester.ac.uk

Abstract. Research Data repositories are growing in terms of volume rapidly and exponentially. Their main goal is to provide scientists the essential mechanism to store, share, and re-use datasets generated at various stages of the research process. Despite the fact that metadata play an important role for research data management in the context of these repositories, several factors - such as the big volume of data and its complex lifecycles, as well as operational constraints related to financial resources and human factors - may impede the effectiveness of several metadata elements. The aim of the research reported in this paper was to perform a descriptive analysis of the DC.Subject metadata element and to identify its data quality problems in the context of the Dryad research data repository. In order to address this aim a total of 4.557 packages and 13.638 data files were analysed following a data-preprocessing method. The findings showed emerging trends about the subject coverage of the repository (e.g. the most popular subjects and the authors that contributed the most for these subjects). Also, quality problems related to the lack of controlled vocabulary and standardisation were very common. This study has implications for the evaluation of metadata and the improvement of the quality of the research data annotation process.

Keywords: Big data · DC.subject · Data quality · Descriptive analysis · Open access repositories · Metadata

1 Introduction

Modern e-Science and e-Research infrastructure has revolutionized the way scientists can store, retrieve, analyse, use, reuse and share data [4]. In this context, research data repositories have become an important predicate of the scientific workflow and a vital tool for research collaboration. To date, several studies have been conducted in order to examine the use, reuse, interoperability, sustainability, dissemination and long-term preservation of data repositories [3], [6], [7], [8]. Yet, there is little known about the use of metadata for research data repositories and in particular, the challenges and

E. Garoufallou et al. (Eds.): MTSR 2015, CCIS 544, pp. 203–213, 2015.
DOI: 10.1007/978-3-319-24129-6_18

problems associated with metadata application [11], [24]. Understanding the use of metadata in research data repositories is important for improving the quality of metadata for data re-use; and analyzing the growth and characteristics of this type of repositories for audit and policy making.

The aim of the research reported in this paper was to perform a descriptive analysis of the use of the DC.Subject metadata element and to identify the data quality issues associated with the specific element in the context of the Dryad repository. This work extends a previous study by [11] and [24] who performed a preliminary analysis of three metadata elements of the Dryad repository. These were: the DC.Creator, DC.Date and DC.Type metadata elements. The decision to focus our analysis on the DC.Subject was made for two reasons. First, there is a consensus among metadata specialist that subject metadata (e.g. keywords or controlled vocabularies) are frequently prone to bias and a lack of adherence to some form of standardization [5]; second, there is no previous work investigating the subject coverage of a mainstream research data repository, like the Dryad.

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

2 Background

2.1 The DRYAD Repository

Dryad is an open-access international repository hosting peer-reviewed scientific, medical and evolutionary biology literature; and a membership organization administered by journals, publishers, scientific societies, and other interested parties [12], [14]. The repository's developers followed a two-pronged approach in order to create a long-term, sustainable system that will support academia's immediate needs. Dryad's metadata requirements are simplicity, interoperability and Semantic web compatibility [16]. Data are deposited as files with permanent identifiers (DOIs) and metadata.

The repository's development allows collections of related files and datasets to be grouped into data packages with metadata describing a combined set of files. By May 2015, the repository contained: approximately 8.700 data packages (an increase of 90% since the beginning of 2014 when the data used for the study was collected); 27.450 data files (100% increase) deposited by 21.360 authors (90% increase) associated with scholarly articles published in almost 410 international journals (42% increase) [11], [13], [24].

A selection of repository development oriented technologies have been used for the implementation and set up of Dryad like the Singapore framework metadata architecture (a framework created in order to maximize interoperability and reusability [15]) in a DSpace environment via an Extensible Markup Language (XML) schema [14, 15]. This infrastructure allows the automatically generated metadata to inherit characteristics from their original sources by harvesting keywords assigned by authors and controlled vocabularies – ontologies [16].

Finally, the metadata application profile of the Dryad repository is based on the DC Singapore Framework [15].

2.2 Previous Work

Several studies related to the technical and architectural components of Dryad have been published and the most notable papers and presentations can be found at Dryad's wiki [12]. Since the Dryad repository went live on January 2008 [12], the majority of the studies conducted e.g. [7], [15], [16], [18] were focused on the implementation and development of Dryad, its curation workflow, the metadata activities and the analysis of its technologies (mainly DSpace). Practical issues about the repository's further development were discussed in [14].

The phenomenon of metadata re-use and metadata quality in the context of Dryad has received less attention. In [8] the reusability of Dryad's metadata elements was examined and the main findings were that 8 out of the 12 metadata elements (contributor, corresponding author, identifier citation, subject, publication name, description, relation is referenced by and title) had a reuse level of 50% or greater. Also, the authors showed that the metadata reuse was more common for basic bibliographic elements like the author, title and subject. However, re-usability is still limited for more specific and complex scientific metadata elements (e.g. those related to spatial, taxonomic and temporal information).

Finally, [11] and [24] performed a statistical analysis of the Dryad repository and examined the quality issues associated with selected metadata elements of the Dryad's application profile. They found that 50% of the creators contributed two or three objects, 70% of which were datasets. The authors also examined the quality issues associated with selected metadata elements of the Dryad repository. Three metadata elements (Creator, Date and Resource Type) were analysed, quality issues associated with these elements were identified and recommendations for improving metadata quality were made. In particular it was shown that approximately 9% of the names of the Creators had various issues and the distribution of the problems was demonstrated. Problems were identified as well with the date as there were several different formats, while 2% of the dates were invalid. 21,4% of the quality problems associated with the DC.Type element consisted of non-standardised use jargon, blanks and non-relevant input. The work validated the results of a previous study by Sokvitne [19] regarding the DC.Creator element. Sokvitne questioned the suitability of Dublin Core for information retrieval by identifying serious issues with several bibliographic metadata, including the Title, Subject, Creator, Contributor and Publisher metadata elements.

2.3 DC.Subject

Since the research datasets deposited to the Dryad repository are linked to original journal papers published elsewhere, each dataset (data packages and data files) inherits the keywords assigned to the given publication [21]. However, other keywords may be manually applied to datasets. For this purpose additional descriptive attributes have been assigned to the DC.Subject metadata element in order to enhance its specificity. For example, the ‘Field Label’ is an attribute used to represent the Subject Keywords; the ‘Formal Definition’ is the general topic of the resource; and the ‘User Definition’ contains the specific Dataset keywords. The contents of Dryad can be searched using a SOLR[1] interface (a standalone enterprise search server with a REST-like API).

Despite the fact that this is the first study to examine the use of the DC.Subject metadata element in the context of research data repositories, findings from the institutional repository and learning object communities have shown that the subject metadata element was one of the most challenging areas for both metadata creation and resource discovery. In the majority of cases this happened because untrained in metadata authors failed to create proper and unproblematic subject metadata [3]. Evidence presented in several case studies showed that in order to achieve high quality subject metadata, both authors and metadata specialists should provide input collaboratively [5].

3 Methodology

A mechanism that involved the downloading of the metadata elements from the Dryad repository and their transformation to a proper format for analysis was employed. In particular, metadata was harvested in January 2014. At that point the Dryad repository contained 4.557 packages and 13.638 data files. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) Validator and data extraction tool was used for the metadata harvesting[2]. A total of 516 .xml files were downloaded (135MB). The XML files were merged into a single file using Mergex, a command line tool for merging xml files[3]. 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 Access (as opposed to the use of more advanced analytics tools, like R). The .xml to .csv Conversion Tool[4] was used to transform the .xml files into .csv and import these in Access. The converter provided 19 .csv files, each corresponding to a metadata element of the Dryad. These are shown in Table 1:

Table 1. CSV files extracted from Dryad Repository

contributor	Format	Record	setSpec
coverage	Header	Relation	***Subject***
creator	***identifier***	Request	***Title***
date	listRecords	responseDate	***Type***
dc	Metadata	resumptionToken	

[1] http://lucene.apache.org/solr/

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

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

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

The .csv files contained the metadata downloaded from the Dryad. In the above table the .csv files in bold are the ones containing data suitable for statistical analysis, whereas the remaining are used as interconnection points between .csv files as they contain tokens, specifications and resumption or response dates. In order to demonstrate the relationship between the different .csv files, a mapping of these files was performed using MS Access (Figure 1).

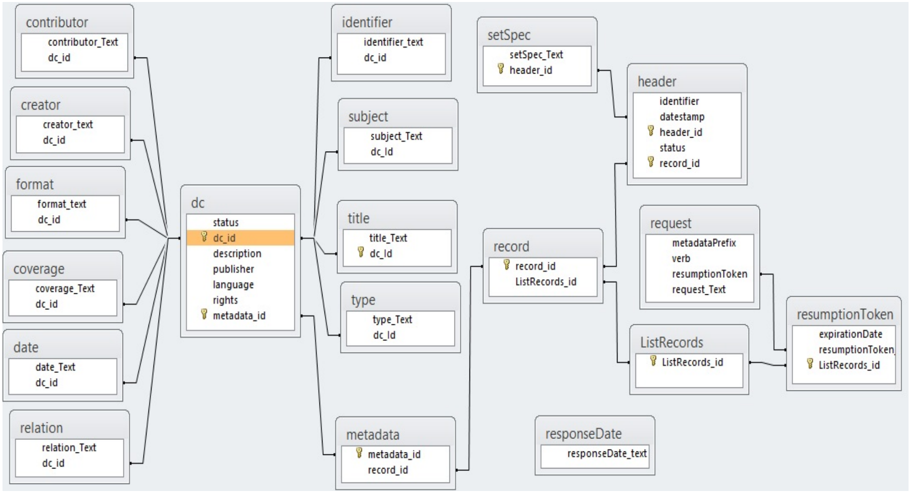

Fig. 1. The Dryad Mapping

3.1 Pre-processing: Preparing the DC.Subject for Analysis

Since the purpose of the paper is to present the results of the analysis of the DC.Subject metadata element, this section summarises the actions made to prepare this metadata element for descriptive analysis. Because we were interested in stratifying the analysis by author (in order to identify the top authors per subject) the analysis involved also the analysis of all author-related metadata elements. These were the Creator and Contributor elements. The workflow used to complete the analysis was split into two phases. Phase 1 involved the downloading of the repository's metadata, while Phase 2 initiated a set of steps needed for preparing metadata for analysis.

Specifically, the Identifier, Creator, Contributor and Subject CSV files were imported in MS Access to create the corresponding tables. These four tables had only two fields: a textual one (i.e. the actual values for the identifier, author, contributor and subject fields) and a numerical one (i.e. the dc identifier, a unique number for each dataset).

The DC.Identifier element was used to identify Dryad’s packages and files. Most types of files can be uploaded (e.g., text, spreadsheets, video, photographs, software code) including compressed archives of multiple files. In order to distinguish the package from its files a ‘/number’ suffix is added to the package’s identifier in order to denote the file (i.e. doi:10.5061/dryad.20 is the package identifier and doi:10.5061/dryad.20/2 is the identifier of the second file of the package).

The Identifier .csv file had a series of issues and irrelevant data. In particular, there was no clear distinction made between packages and files as information about these was contained in the same metadata element and not in different elements as one would expect. Also, data from other repositories was found within the downloaded metadata files that actually contaminated Dryad's metadata. The repositories that are obviously collaborating with Dryad are the Knowledge Network for Biocomplexity (KNB)[5] - an international repository intended to facilitate ecological and environmental research, and the Long-Term Ecological Research (LTER) Network program[6] - a network that seeks to inform the broader scientific community by providing open access to well designed and well documented databases via a Network-wide information system.

Based on the Dryad's DOI identifier, the data was "cleaned" and the correct packages with the corresponding identifiers were retrieved. The technique for cleaning the data was based on SQL queries: i) Records containing the 'doi' string were retrieved and, ii) Records containing as last characters a forward slash and one or two numbers were firstly identified, catalogued and saved in new tables and then removed. Using the correct data and via a SQL query, the number of each unique Keyword was calculated. The Creator and Contributor tables were merged into a single table called 'Author'. This decision was made because after inspection of the data we observed that both metadata elements were used for the same purpose (i.e to denote the authors and co-authors of a given dataset). Then via the relationship (the common dc_id field) of the new 'Authors' table and the 'Subject' table the total contributions per subject for each author was calculated.

4 Results

4.1 General Results for Packages and Files

The initial dc.identifier file was consisted of 127.853 records which as mentioned earlier included the identifiers for packages and files from Dryad along with data from the KNB and LTER repositories. With a series of queries the identifiers for the 4.557 packages (a 100% success) were retrieved. The number of files per packages was calculated and Figure 2 provides a depiction of the findings.

As it is shown in Figure 2, approximately half of the packages contained one file (49%), while two files were included in 810 packages (17,7%). By multiplying the number of packages by the number of files per package we managed to calculate the total of the files that were uploaded to Dryad. According to our calculations the total number of files in the repository was 13.633 - just five less than the ones referred to the Dryad's site. This means that each package contained on average three files.

[5] https://knb.ecoinformatics.org/

[6] http://www.lternet.edu/

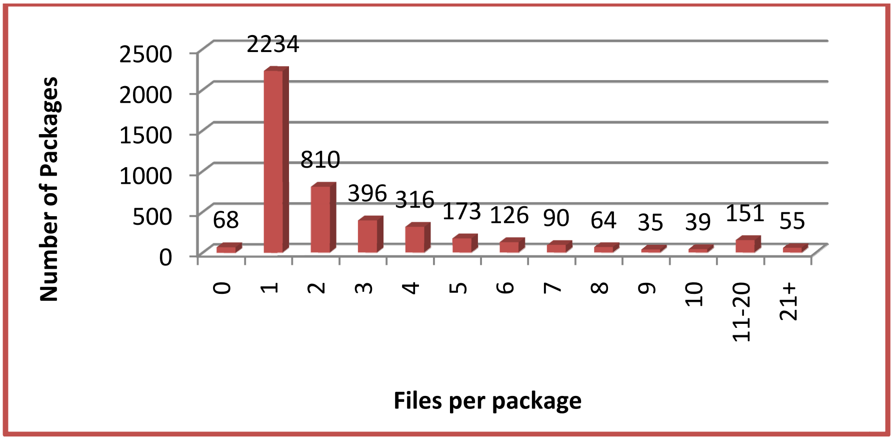

Fig. 2. Files per package diagram

4.2 Subject

The DC.Subject provides the keywords of each object of the repository. According to [21], initially only-explicitly stated keywords were meant to be catalogued and the final goal of Dryad's developers was to perform an automatic keyword insertion.

Table 2. Subject Top-25

Subject	Count	Percentage (%)	Subject	Count	Percentage (%)
Adaptation	366	1,68	Quantitative Genetics	115	0,53
Population Genetics – Empirical	304	1,39	Molecular Evolution	110	0,50
Speciation	258	1,18	Biogeography	104	0,48
Phylogeography	235	1,08	Microsatellites	98	0,45
Ecological Genetics	197	0,90	Birds	96	0,44
Hybridization	184	0,84	Phylogenetics	95	0,44
Conservation Genetics	177	0,81	climate change	93	0,43
Insects	143	0,66	Life History Evolution	90	0,41
Population Genetics	141	0,65	Mammals	90	0,41
Phenotypic Plasticity	132	0,61	Sexual Selection	87	0,40
Phylogeny	129	0,59	Invasive Species	85	0,39
Fish	123	0,56	natural selection	84	0.39
Gene Flow	119	0,55			

In total, 21.809 subjects were identified as keywords for the packages and the unique ones were 8.149. Therefore, approximately five keywords (4,79) were used on average per package. Table 2 shows Dryad's 25 most popular subjects. The most popular keywords were in accordance with the Dryad's subject coverage, i.e. focused on medical and evolutionary biology topics.

4.3 Subject distribution by author

The most frequent authors per subject area were also identified. With the aid of a query that used the *Authors* Table (which was created by merging the creator and the contributor tables) and the Subject Table, the count of subjects per author was requested. The query provided data for a new table with 3 fields: Subject, Author, and Author's Number of Contributions per Subject. Table 3 shows the top 10 of the subjects that were most frequently used from a unique author.

Table 3. Most frequent keyword from a unique author

Subject	Author	# of Contributions
Phylogeny	Douzery, Emmanuel J. P.	10
Fish	Bernatchez, Louis	10
Speciation	Bernatchez, Louis	8
Molecular dating	Douzery, Emmanuel J. P.	8
Supermatrix	Douzery, Emmanuel J. P.	7
Speciation	Rieseberg, Loren H.	7
Phylogeny	Delsuc, Frédéric	6
Conservation Genetics	Narum, Shawn R.	6
DNA metabarcoding	Taberlet, Pierre	6
sexual selection	Rundle, Howard D.	6

The analysis can identify also if an Author has contributed heavily to a specific subject. For instance 10 out of the 129 contributions (7,75%) and 10 out of the 123 contributions (8,13%) of the 'Phylogeny' and the 'Fish' subject respectively, come from specific authors.

For the top 25 most popular subjects we managed to identify the most contributive authors per subject. In table 4 only the top 10 subjects are shown along with the authors with more than 3 contributions per subject. For the 'speciation' subject there are 5 additional Authors with 4 contributions. An additional analysis of this table can provide associations between Authors and also between group of Authors and subjects.

Table 4. Most contributive Authors for the Top-10 Subjects

Subject	Author	# of Con.	Author	# of Con.	Author	# of Con.
Adaptation	Laurila, Anssi	4	Seehausen, Ole	4	Butlin, Roger K.	4
	Sota, Teiji	4	Merilä, Juha	4		
Population Genetics	Bernatchez, Louis	4				
Speciation*	Bernatchez, Louis	8	Rieseberg, Loren H.	7	Rieseberg, Loren H.	5
Phylogeography	Moritz, Craig	5	Schönswetter, Peter	4	Searle, Jeremy B	4
Ecological Genetics	Narum, Shawn R.	4	Bernatchez, Louis	4	Gagnaire, Pierre-Alexandre	4
	Kempenaers, Bart	4				
Hybridization	Moritz, Craig	4	Rieseberg, Loren H	4	Bernatchez, Louis	4
Conservation Genetics	Narum, Shawn R.	6	Campbell, Nathan R	4		
Insects	Foitzik, Susanne	4	Traugott, Michael	4		
Population Genetics	Bernatchez, Louis	4				
Phenotypic plasticity	Simmons, Leigh W.	4				

* there are five additional authors with four contributions

4.4 Data Quality Issues

Identifier

The main issues with the Identifier metadata are the repetitive data and most important of all the irrelevant to Dryad data. As it was mentioned in the methodology section, data from other repositories were included in the downloaded xml files. The main repercussion of such unwanted data is that researchers are led to biased and erroneous results. The downloaded data need first to be cleaned and corrected, via the procedure described in the methodology section. It seems that no data quality mechanisms are in place in the case of the metadata annotation process of the Dryad repository. An obvious solution that could lead to error-free data is the blocking of data that do not contain a DOI. Finally, an implementation of separate metadata identifiers for the packages and the files would aid the analysis of the repository.

Subject

Several quality issues are met in this element. First of all, the manual cataloguing of data entails typos and the input of irrelevant information. This can lead to multiple records for the same subject. Another serious problem was the extreme diversity of similar notions (synonyms). It is apparent that the subjects were not entered through the use of a controlled vocabulary that would obviously restrict and minimize mistakes. For instance, the 'Fertilization' subject has 21 similar entries: fertilization, fertilized, fertilizer, fertilizers and various forms of fertilization such as bias, success, plot, plots, Fertilization nitrogen and Fertilization phosphorus are a few examples that confirm the lack of standardisation.Similar problems were encountered in the case of the ‘Population’ subject where 144 similar/diverse entries were recorded. The inconsistent use of singular and plural, adjectives, synonym terms and misspelled words failed the quality criteria check during the data pre-processing phase and made difficult the analysis of the subject metadata element.

5 Conclusions

The goal of this research was to perform a descriptive analysis of the DC.Subject metadata element used in the Dryad repository. Following this analysis a series of quality problems associated to the specific metadata element and the process implemented to analyse it were identified.

Despite the fact that several metadata quality issues have been documented in the literature during the past few years e.g. [5], [6], [10], [11], [19], yet many of these issues are still present in the case of the Dryad repository. It appears that there is a need for more manual control over the metadata input, since the automatic or semi-automatic method of populating the DC.Subject element with values is prone to quality errors. Improving the quality of the subject metadata in Dryad could also streamline the process of analyzing its contents. Therefore, establishing a coherent pre-processing method for cleaning the metadata is important for strengthening the validity of the analysis process. In this present paper we demonstrated a method for pre-processing specifically for the DC.Subject metadata element. This involved the mapping of the different metadata elements and their relationship (Figure 1); and the use of the DC.Identifier element as a means of identifying unique instances of packages and files for subject analysis.

Future work will be focused on applying data mining and text mining techniques to the DC.Subject metadata element in order to provide a better understanding of the repository’s data; to identify associations, clusters or hidden patterns for this data; and to develop novel visualisations for displaying the contents of the Dryad repository [22].

References

1. Gargouri, Y., Hajjem, C., Lariviere, V., Gingras, Y., Brody, T., Carr, L., Harnad, S.: Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research. PLOS ONE **5**(10) (2010). http://www.plosone.org/article/info:doi/10.1371/journal.pone.0013636 (July 13, 2014)
2. Mabe, M., Amin, M.: Growth dynamics of scholarly and scientific journals. Scientometrics **51**, 147–162 (2001). doi:10.1023/A:1010520913124
3. Hess, C., Ostrom, E.: A Framework for Analyzing the Knowledge Commons : A Chapter from Understanding Knowledge as a Commons: from Theory to Practice (2005). http://surface.syr.edu/cgi/viewcontent.cgi?article=1020&context=sul
4. Garoufallou, E., Papatheodorou, C.: A critical introduction to metadata for e science and e-research, special issue on metadata for e-science and e-research. International Journal of Metadata Semantics and Ontologies (IJMSO) **9**(1), 1–4 (2014)
5. Currier, S., Barton, J., O’Beirne, R., Ryan, B.: Quality assurance for digital learning object repositories: issues for the metadata creation process. ALT-J, Research in Learning Technology **12**(1), 5–20 (2004)

6. Heery, R., Anderson, S.: Digital repositories review. Other. Joint Information Systems Committee (2005). http://www.jisc.ac.uk/uploaded_documents/digital-repositories-review-2005.pdf
7. Greenberg, J., Vision, T.: The Dryad Repository: A New Path for Data Publication in Scholarly Communication. OCLC, Dublin, Ohio (2011). https://www.oclc.org/content/dam/oclc/community/presentations/guests/greenberg-20110425.pdf (January 22, 2015)
8. Greenberg, J, Swauger, S, Feinstein, E.M.: Metadata capital in a data repository. In: Proceedings of the International Conference on Dublin Core and Metadata Applications, pp. 140–150 (2013)
9. Beagrie, N., Eakin-Richards, L., Vision, T.: Business Models and Cost Estimation: Dryad Repository Case Study, iPRES2010 Vienna (2010)
10. Palavitsinis, N., Manouselis, N., Sanchez-Alonso, S.: 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), 1202–1216 (2014)
11. Rousidis, D., Garoufallou, E., Balatsoukas, P., Sicilia, M.A.: 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, June 19–20, 2014, Thessaloniki, Greece (2014). http://elpub.scix.net/data/works/att/106_elpub2014.content.pdf
12. Dryad Digital Repository Wiki. Main Page, April 29, 2015. http://wiki.datadryad.org/Main_Page
13. Dryad Digital Repository. Frequently Asked Questions, April 29, 2015. http://datadryad.org/pages/faq
14. White, H., Carrier, S., Thompson, A., Greenberg, J., Scherle, R.: The Dryad data repository: a Singapore framework metadata architecture in a DSpace environment. In: The 2008 International Conference on Dublin Core and Metadata Applications, Berlin (2008)
15. Greenberg, J., White, H.C., Carrier, S., Scherle, R.: A metadata best practice for a scientific data repository. Journal of Library Metadata **9**(3), 194–212 (2009). http://dx.doi.org/10.1080/19386380903405090 (February 15, 2014)
16. Greenberg, J.: 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), 380–402 (2009)
17. Peer, L.: The Role of Data Repositories in Reproducible Research. Yale (2013). http://isps.yale.edu/news/blog/2013/07/the-role-of-data-repositories-in-reproducible-research#.UzINafmSxyM
18. Greenberg, J.: Linking and Hiving Data in the Dryad Repository. The Semantic Web: Fact or Myth. CENDI, FLICC, and NFAIS Workshop. National Archives, Washington, DC, November 17, 2009 (2009b)
19. Sokvitne, L.: An Evaluation of the Effectiveness of current Dublin Core Metadata for Retrieval. Proceedings of VALA 2000. Victorian Association for Library Automation: Melbourne (2000)
20. Beagrie, N., Eakin-Richards, L., Vision, T.: Business Models and Cost Estimation: Dryad Repository Case Study, iPRES2010 Vienna (2010)
21. Dryad Digital Repository Wiki. Cataloging Guidelines (2009). http://wiki.datadryad.org/Cataloging_Guidelines_2009 (April 12, 2015)
22. Greenberg, J., Garoufallou, E.: Change and a future for metadata. In: Garoufallou, E., Greenberg, J. (eds.) MTSR 2013. CCIS, vol. 390, pp. 1–5. Springer, Heidelberg (2013)
23. Integrating Manuscript Processing with the Dryad Digital Repository, April 10, 2015. http://wiki.datadryad.org/images/c/c6/DryadIntegrationOverview.pdf
24. Rousidis, D., Garoufallou, E., Balatsoukas, P., Sicilia, M.A.: Metadata for big data: a preliminary investigation of metadata quality issues in research data repositories. Information Services and Use **34**(3), 279–286 (2014)