Let's Put Data to Use: Digital Scholarship for the Next Generation
P. Polydoratou and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-409-1-49

Data Quality Issues and Content Analysis for Research Data Repositories: The Case of Dryad

Dimitris ROUSIDIS[a,1], Emmanouel GAROUFALLOU[b], Panos BALATSOUKAS[c], Miguel-Angel SICILIA[a]
[a] *University of Alcala, Madrid, Spain*
[b] *Alexander Technological Educational Institute of Thessaloniki, Greece*
[c] *University of Manchester, UK*

Abstract. Research Object (RO) repositories extend traditional forms of scholarly communication by providing scientists the means necessary to store, share and re-use datasets generated at various stages of the research process. Yet this shift to digital publication does not guarantee that outputs, results or methods are reusable. Data quality is absolutely vital for the dissemination, reuse and sharing of digital resources. Manual metadata quality control is practically impossible and as a result, many quality criteria, both semantically and structurally get overlooked and digital objects may become problematic. The aim of the research reported on this paper was to identify the data quality problems associated with the Dryad research data repository. In particular, three metadata elements (Creator, Date and Resource Type) were analysed and quality issues associated to these elements were identified. The paper concludes with some recommendations for improving the quality of metadata in research data repositories.

Keywords. Big Data, Data Quality, Descriptive Analysis, Open Access Repositories, Metadata, Research Objects, e-Research

1. Introduction

The parallel growth the availability of scientific data (big data) and the emergence of cloud computing has 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, analyse, use and share datasets generated at various stages of the research process [1]. 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. Although research on the technical and architectural characteristics of research data repositories has progressed (e.g.[2], [3], [4]), there is still a need to measure their growth and analyse their contents. This includes knowledge on the size, composition and growth dynamics of these repositories. Such knowledge would eventually result in insights on the behaviour of researchers and the usability of their research products for reuse, e.g. for experiment repetition.

It is well documented in the literature that measuring the growth and analysing the contents of digital repositories in general is important for the sustainability and

usability of this type of technology (e.g. [5]). Yet, data quality issues (e.g. in terms of metadata) may influence the effectiveness of the analysis of this type of repositories.

The aim of the research reported in this paper was to identify the data quality problems associated with the analysis of the contents of a research data repository, called Dryad. Being this a first attempt to measure research data repositories, the objectives were chiefly exploratory, concretely:

- To perform a descriptive analysis of the contents of the Dryad repository across different variables (metadata), such as the type and format of datasets, the size and submission date of data packages and files; and
- To identify data quality issues and challenges related to the analysis of these metadata elements;

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

2. Previous work

2.1. Quantitative analysis of Repositories

Since the concept of a Research Repository or Research Object is relatively new [6], our knowledge of analysing the contents of research repositories comes primarily from studies conducted in other types of data infrastructures, such as Learning Object Repositories (LORs). A series of seminal studies on the analysis of the contents and growth of digital repositories have been reported by Ochoa and Duval. The goal of these studies [7], [8], [9] was the application of techniques that measured and analysed the processes that create, publish, consume or adapt information in the context of learning object repositories. Several techniques and algorithms were employed in this respect. The purpose of these algorithms was: to apply a set of metrics that would facilitate the assessment of the quality of the learning object metadata within repositories and establish the potential relevance of the learning objects for a given user and situation; to assess the growth of the repositories by analysing the contents of metadata elements, such as the repository's size and growth over time, contributors' characteristics and the number of published material; to examine the relationship between the popularity of an object and its reuse [7], [8], [9].

The findings of the studies conducted by Ochoa and Duval [8], [9] showed some interesting patents regarding the growth, reuse and quality of metadata within repositories. For example, they observed abnormalities on the size distribution of repositories and surprisingly enough a linear growth over time regardless the size and popularity of the repositories. The number and the growth of contributors within repositories varied across repositories due to differences in the size and nature of each individual repository. Regarding the contributor's publication distribution (i.e. the amount of content deposited in the repository by a contributor), the conclusion was that it is relevant to the contributor's engagement with the repository. The issue of reusability of content within repositories was examined by Ochoa in a follow up study in the context of Learning Object repositories [7]. The results of the quantitative analysis were rather discouraging as on average a mere 20% of Learning Objects was reused. A very interesting and rather unexpected result was the lack of correlation

between the popularity of a learning object and its reuse. Finally, in terms of the quality metadata included in the repositories, Ochoa and Duval found the growth of repositories and changes in the nature of information deposited may have an effect on the actual quality of metadata used for the description of the contents of learning object repositories (e.g. inconsistencies in the use of metadata and vocabularies, different levels of completeness within and across repositories).

2.2. (Meta)data Quality

Data quality is defined as the state of completeness, validity, consistency, timeliness and accuracy that makes data suitable for a specific use [10]. Dekker [11] 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 [11]. Metadata quality is a vital factor for electronic interoperability [9], [12], [13], [14]. 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 [9], [15]. In many cases, the larger the dataset, the greater the probability a problem will emerge [12]. 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. [16])

2.2.1. Quality Issues and Metadata Elements

Sokvitne[17] conducted a research about the effectiveness of the metadata elements of the Dublin Core for retrieval. Sokvtne was focused on the following metadata elements: title, creator, publisher, contributor and subject. The study showed problems with all the above elements. In particular, the DC.title and the DC.subject weren't adding any value for retrieval purposes, while the DC.creator, DC.publisher and DC.contributor presented inconsistent name formats. He 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.

Barton [12] outlined the areas where metadata element problems most commonly arise. These were:

- Spelling, abbreviations and other similar data entry errors and ambiguities.
- Author and other contributor fields. The most common issues are that the same name is entered differently (e.g. inconsistent entry of initial, first-last name ambiguity), a name can change (for instance if one gets married and adopts/adds the spouse's name) and finally synonyms especially in common names.
- Title. Many resources have more than one possible title, while others, particularly non-textual resources, may have no title at all.
- Subject – in the form of keywords and classifications. The main issue with the subject is who should enter the data; the author or the metadata specialist? The author can ensure the entry of the correct terminology but the metadata specialist can ensure the data consistency. The use of taxonomies and subject classification schemes is part of the solution.

- Date. Two main sets of problems are met in this element. Initially the format is the main issue as there are numerous formats that one could use. The format of the date entry should be strict, predefined and unique. The second issue is that date is often ambiguous as it may refer to publication date, submission date, date the record became available, etc.

2.3. 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 [18].

Dryad's primary aim is to facilitate data discovery and reuse, thus guaranteeing the long-preservation of this [19]. Greenberg [3] established as the main two goals of the repository, "the one-stop deposition and shopping for data objects supporting published research" and "the support of the acquisition, preservation, resource discovery, and reuse of heterogeneous digital datasets". One of the strong and appealing characteristics of Dryad according to Peer [20] is that its curatorial team "works to enforce quality control on existing content".

Dryad's developers, by using the Singapore framework metadata architecture in a DSpace environment via an Extensible Markup Language (XML) schema [21], [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[3]. The Singapore Framework for Dublin Core Application Profiles is a framework created in order to maximize interoperability and reusability (Dublin Core Metadata Initiative) by shifting from the "resource-driven legacy approach", representing an information package, to the granular component parts of a resource [22]. Dryad's metadata requirements are simplicity, interoperability and Semantic web compatibility [23].

Greenberg initially [24] and [25] 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 the measurement of the growth of its contents.

3. Methodology

A mechanism that involved the downloading of the metadata elements from the Dryad and their transformation to a proper format was employed. On January 2014, the metadata of the repository were harvested. 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 (135MB). 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 analyse the data from the xml files had to be employed. Due to the descriptive nature of the statistical analysis performed it was decided to analyse the data using Microsoft Excel 2010. It was anticipated that the records per file would be more than 65.536. so using an earlier version of MS Excel would be rather problematic. Therefore the xml to Csv Conversion Tool[3] was used to transform the XML files into CSVs and import these to Excel. It is woth mentioning that importing directly the xml file to Excel provided very frustrating results. The converter provided 19 csv files: 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 sample of metadata elements was analysed. 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.

4. Results

4.1. Creator

The number of contribution per author is depicted on table 1. In total 16.567 authors 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.

Table 1. Amount of objects published by each contributor

Contributions	Amount	Contributions	Amount	Contributions	Amount
1	1422	11	248	21-30	286
2	6131	12	225	31-40	128
3	2282	13	137	41-50	129

1 http://validator.oaipmh.com/
2 https://code.google.com/p/mergex/
3 http://xmltocsv.codeplex.com/

4	1541	14	144	51-60	70
5	1060	15	84	61-70	46
6	773	16	92	71-80	35
7	601	17	100	81-90	25
8	396	18	82	91-100	15
9	362	19	55	>100	2
10	242	20	47	**Total**	**16567**

4.2. Date

This metadata element was assigned to various types of dates like date accessed, date available and date issued. For the purpose of this analysis we gathered the dates corresponding to the date issued (according to the cataloging guidelines of Dyad's[4] wiki, dc.date.issued is the official date of publication, inherited by dataset; the date of the formal issuance of the resource) of the 43.453 objects in the repository. The distribution per year is depicted in Table 2.

Table 2. Amount of Objects issued per year

Date	Amount	Date	Amount	Date	Amount
1995	1	2002	10	2009	416
1996	10	2003	11	2010	3172
1997	10	2004	13	2011	25411
1998	59	2005	12	2012	5035
1999	50	2006	13	2013	8005
2000	17	2007	27	1/1/2014-9/1/2014	176
2001	67	2008	97	Invalid input	841

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.

4.3. Type

A total of 53.598 records were retrieved for the DC.Type element and their distribution is shown in Table 3. 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.

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

Table 3. Type distribution of objects

Type	Amount	Percentage %	Type	Amount	Percentage %
Activity	4	0,007	Image	62	0,116
Article	4451	8,304	Map	1	0,002
Book	3	0,006	none	4086	7,623
Blank	4	0,007	oneyear	830	1,549
custom	109	0,203	protocol	11	0,021
Dataset	36708	70,167	untilArticleAppears	6429	11,995

As shown in Table 3, 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 dc.type 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 are inserted. If we clean the data and leave only the suitable type files, then 42.129 records remain and the percentages change: 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 are dataset and nearly 10% are articles.

4.4. Data quality problems

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 analyse 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. The most common quality issues are summarised below.

4.4.1. Creator

The highest variety of issues was identified in this element. Out of 16568 records, a total of 1443 (8,71%) demonstrated the following issues:

- Additional names: Many authors were input with just their first name. The problem emerged in 614 (42,55%) cases when the authors' additional name were added as a different record and also by including additional ones (e.g. Aradhya, Mallikarjuna K. and Aradhya, Mallikarjuna).
- Using initials: Another serious issue was the use of initials instead of the whole name (11,64%). For instance Schim van der Loeff, M. F. and Schim van der Loeff, Maarten Franciscus.
- Differentiation of languages: A percentage of 12,06% occurred with this issue. There are numerous variations for writing a name in non-English language. Trying to convert a name by the English alphabet may be problematic as there are many symbols that do not exist. For instance, accent aigu or accent grave in French, umlaut in German, etc. make an error when writing a name very

possible. 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 (e.g. Zach instead of Zachariah).

- Miswritten: With a percentage of 2,56% many errors due to typos were indemnified (e.g. Philipp instead of Phillip). In this category errors like when a first name was missing or when the name was inserted at the surname field were also counted.
- Dots and commas. The second most frequent mistakes (23,08%) were the absence of dots or the use of commas at the end of initials.
- Spacing: Different 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%).
- Ambiguous: There were around 71 cases (4,92%) were there was serious doubt whether different writings of a creator were belonging to the same person, mainly because they were very common (e.g. Gold, John and Gold, J. or Edwards, Mary and Edwards, M.).

It should be noted that in one occasion the names of a certain creator (that we will not write his surname) were input with six different ways (A Rus. – A. Rus – A. Russel – Alan R. – Alan Rus – Rus).The problems appeared in this element were also identified in the Contributor element; although an analysis was not performed, a rough review validated the same symptoms.

4.4.2. Date

Serious issues were also met at the DC.Date element. There was absolutely no consistency in the format and no control for the insertion of dates. As it is mentioned in section4.2 there were dates with invalid format: 4 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. Table 4 depicts the second main issue; the inconsistency in the date format. The length of the date varies from being blank up to 20 characters.

Table 4. Length of issued date

Length	Count	Example
4	156	2009
6 to 7	163	2009-03
8 to 10	42590	2009-09-07
20	503	2009-10-01T10:19:28Z
Various	41	Blanks, unacceptable format

4.4.3. Type

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. In addition to the reusability of research objects, 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 this objective all the metadata from the Dryad repository were harvested and analysed. 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.

We propose a set of ideas that might enhance the quality of Dryad's metadata. For example, a solid format of the names should be specified. Each creator and contributor should be assigned with a unique ID that would hold their full name. When requesting an entry of the full name at the repository this unique ID should be inserted. To avoid any complications, the ID might be interlinked with an email. Possible synonymies can be resolved by the use of unique full names (e.g. different writing of first names, the use of initials, or the use of a father name should be implemented). 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). Finally, in the case of the type metadata element, inconsistencies can be fixed through the use of pre-defined list of values for authors to select from.

Based on the belief that "metadata solutions will become common-place for accomplishing various tasks" [26], our future work will be focused on Dryad Repository and the rest of its metadata elements. More elaborate statistical analysis by using R will be employed and data mining and text mining techniques will be applied in order to provide a better understanding of the repository's data, to identify any associations, clusters or hidden patterns and provide a visualization of these results.

References

[1] Garoufallou, E. and 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) (2014), 1 – 4.

[2]Bernard, J. et al. A visual digital library approach for time-oriented scientific primary data.*International journal on Digital Libraries*, **11(2),** (2010), 111-123.

[3] 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)** (2009), 380-402.

[4] Heery, R. Digital Repositories Roadmap review: towards a vision for research and learning in 2013. *JISC*. Available: http://kennison.name/files/zopestore/uploads/libraries/documents/reproadmapreviewfinal.pdf [29/12/2013]

[5] Kelly, B et al. Open metrics for open repositories. *In: OR2012: the 7th International Conference Conference on Open Repositories,* (2012).Available: http://opus.bath.ac.uk/30226/ [29/12/2013].

[6] Bechhofer, S., De Roure, D., Gamble, M., Goble, C., Buchan, I. Research Objects: towards exchange and reuse of digital knowledge. *FWCS2010*, (2010), Available: http://eprints.soton.ac.uk/268555/1/fwcs-ros-submitted-2010-02-15.pdf [10/03/2014]

[7] Ochoa X.Learnometrics: metrics for learning objects. *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (LAK '11). ACM, New York, NY, USA, (2010, 1-8. http://doi.acm.org/10.1145/2090116.2090117

[8] Ochoa, Xavier and Duval, Erik. Quantitative Analysis of Learning Object Repositories. *IEEE Transactions on Learning Technologies*, **2(3)**, (2009), 226-238.

[9] Ochoa, X. and Duval, E. Automatic evaluation of metadata quality in digital repositories. *International Journal on Digital Libraries*, **10(2)**, (2009), 67-91.doi:10.1007/s00799-009-0054-4

[10] Gordon, K. Principles of Data Management. Facilitating Information sharing, 2007, ISBN 978-1-902505-84-8. Available: http://www.bcs.org/upload/pdf/data-management-chapter1.pdf [13/03/2014]

[11]Dekkers, M., Loutas, N., De Keyzer M., and Goedertier, S. Open data and metadata quality. (2013). Available: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]

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

[13] Park, J. Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging & Classification Quarterly*, **47(3)**, (2008), 213-228. doi:10.1080/01639370902737240

[14] Palavitsinis, N., Manouselis, N., Sanchex-Alonso, S. Metadata quality in digital repositories: empirical results from the cross-domain transfer of a quality assurance process. *Journal of the Association for Information Science and Technology*, (In press)

[15] Lee, D. Practical maintenance of evolving metadata for digital preservation: Algorithmic solution and system support. *International Journal on Digital Libraries*, **6(4)**, (2007), 313-326. doi:10.1007/s00799-007-0014-9

[16] Balatsoukas, P., O'Brien, A., and Morris, A. 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). Available: http://www.informationr.net/ir/16-3/paper481.html[1/2/2014]

[17] 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).

[18] Dryad Digital Repository. Frequently Asked Questions.Available: http://datadryad.org/pages/faq[29/12/2013]

[19]Beagrie, N., Eakin-Richards, L., and Vision, T. Business Models and Cost Estimation: Dryad Repository Case Study*, iPRES2010,* (2010), Vienna.

[20]Peer, L. The Role of Data Repositories in Reproducible Research.Yale, (2013). Available:http://isps.yale.edu/news/blog/2013/07/the-role-of-data-repositories-in-reproducible-research#.UzINafmSxyM [12/0/2014]

[21] White, H. Carrier, S., Thompson, A., Greenberg, J., &Scherle, R. The Dryad Data Repository: A Singapore Framework Metadata Architecture in a DSpace Environment. *DC 2008, the 2008 International Conference on Dublin Core and Metadata Applications*, (2008), Berlin.

[22] Greenberg, J., White, H., C, Carrier, S. and Scherle, R.A Metadata Best Practice for a Scientific Data Repository.*Journal of Library Metadata*, **9 (3)**, (2009) 194-212. Available: http://dx.doi.org/10.1080/19386380903405090 [15/2/2014]

[23] 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, (2009). Available: http://www.cendi.gov/presentations/11-17-09_cendi_nfais_Greenberg_UNC.pdf [9/1/2014]

[24] Greenberg, J. and Vision, T. The Dryad Repository: A New Path for Data Publication in Scholarly Communication. *OCLC*, Dublin, Ohio, (2011). Available http://www.oclc.org/research/news/2011-03-24.htm [22/1/2014]

[25] Greenberg J, Swauger S, Feinstein E.M. Metadata Capital in a Data Repository. *Proceedings of the International Conference on Dublin Core and Metadata Applications*(2013), 140-150

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