Registry Metadata Quality Assessment by the Example of re3data.org Schema

Suntae Kim*, Myung-Seok Choi**

ARTICLE INFO

Article history:
Received 29 September 2016
Revised 12 December 2016
Accepted 4 May 2017

Keywords:
Data Repository,
Repository Registry,
Re3data.org

ABSTRACT

Nowadays, research data repositories (RDR) have become progressively widespread all over the world. To expand repository services and build up inbound linking strategy, organizations list their repositories with so called Global Registries. Accordingly, such registries should be carefully described by the related data. In this study, I explore the metadata schema of *re3data.org*. I collect and analyze descriptions from the listed repositories, and come up with some suggestions concerning possible improvements to the metadata schema. To accomplish this, I develop a crawler program, which collects necessary data from the *re3data.org*. Based on the analysis results, I have identified two issues that required elements is missing, one issue that required element value is missing when the corresponding property is applied, five inconsistency issues with re3data controlled vocabulary, six issues with undescribed optional elements, and two inconsistency issues between the elements and their attributes which do not pair with. I believe this discussion can facilitate improvements to the existing *re3data.org* schema and further help researchers who analyze data repository trends.

1. Introduction

Data repositories play increasingly larger role in academic research. Reliable storage and fair re-use of the research data are of paramount importance in terms academic ethics, and thus become an imperative for any research institution. Researchers require infrastructures that ensure a maximum of accessibility, stability and reliability to facilitate working with and sharing of research data. Such infrastructures are being increasingly summarized under the term Research Data Repositories (RDR) (Pampel et al., 2013). Against this background, former Institutional Repositories (IR) are rapidly evolving into Institutional Data Repositories (IDR). On the other hand, global services emerge to help locate individual repositories and assess their content. By way of example, maps.repository66.org service created by Lewis uses OpenDOAR and ROAR data to discover repositories from all around

* Researcher, Korea Institute of Science and Technology Information, Korea (stkim@kisti.re.kr)
** Researcher, Korea Institute of Science and Technology Information, Korea (mschoi@kisti.re.kr)
International Journal of Knowledge Content Development & Technology, 7(2): 41-51, 2017.
http://dx.doi.org/10.5865/IJKCT.2017.7.2.041

the world. re3data.org is a representative example of such specialized registry. As of now, March 2016, it incorporates 1513 research data repositories maintained by 3253 providers: 1349 data providers and 630 service providers, of which 468 registrants provide both data and services. On this account, many institutions opt in re3data.org to promote their content and build up inbound liking. As this takes place, good metadata design becomes the key to data discoverability and sustained click stream. To facilitate the process, re3data.org maintains a Metadata Schema, which is constantly updated to better reflect changes in the landscape of RDRs. Each element and each attribute is scrutinized whether or not they meet real-life needs, and the controlled vocabulary is examined with respect to domain applicability. I will try to address some of these issues by collecting and analyzing metadata from the repositories listed with re3data.org, and based on the obtained results make some suggestions on possible improvements to the existing schema.

2. Previous Research

Pampel et al. (2013) outline the background of the re3data project, and examine its main features and outcomes. Jones (2012) focuses on evolution of the OpenDOAR management ecosystem, and analyzes how it affects data reliability and accessibility. Norris, Oppenheim, and Rowland (2008) compares the relative effectiveness of OpenDOAR, Google, Google Scholar, and OAIster in terms of open access to peer reviewed journal articles. Shafi, Gul, and Shah (2013) studies Web 2.0 interactivity in open access repositories on the ground of OpenDOAR data. Lone , Rather, and Shah (2008) makes a case study of repository status according to the provided data as exemplified in Indian resources. DRI (2013) conducts a detailed study into the outlook of repositories with a view to requirements for constructing nation-wide RDR. The above studies delve mostly in OpenDOAR and ROAR, which have been around before re3data.org. Also, they deal primarily with bulk analysis of the repository metadata, taking a particular interest in open access. There are a quite few research studies on re3data.org, and thus far no study has aimed at improving metadata schema on the strength of the analysis of actual data related to the RDR itself.

3. Material and Methods

The number of RDRs listed on re3data.org (Registry of Research Data Repositories): 3 in 2010, 136 in 2012, 401 in 2013, 582 in 2014, 282 in 2015, and 109 as of March 2016. To collect RDR metadata, I develop a special Crawler program. The collected data (totally 1,513 records) is stored in a relational database and evaluated against the proposed re3data.org schema. I check how elements and attributes are used, thus enabling us to reveal possible issues and suggest potential improvements to the schema. For data collector I modify the source code by Kim and Lee (2014) so that it works with OpenDOAR and ROAR registries. The development environment is outlined in the Table 1 below.

Table 1. Development environment for re3data.org metadata collection

• OS: Windows 7 Professional K, Service Pack 1 • Database Server and Client: MySQL Server 5.5 / MySQL Workbench 6.2 • IDE: Eclipse Java EE IDE for Web Developers / Luna Service Release 1 (4.4.1) / build 20140925-1800 • Programming Language and VM: Java 1.7.0_67 / JavaTM SE Runtime Environment (build 1.7.0_67-b01) /Java HotSpotTM 64-bit Server VM (build 24.65-b04, mixed mode)

Figure 1 shows program flowchart by Kim and Lee (2014) and Figure 2 shows the modified crawler flowchart used in this study. As may be seen, metadata collection procedure has been considerably simplified. The program uses ISO national codes to query OpenDOAR and ROAR data. At the same time, re3data.org provides REST API (http://service.re3data.org/api/v1/repositories) whereby one can retrieve the list of available RDRs. Unfortunately, at the time of writing (March, 2016) no interface for querying individual repositories is supported. Accordingly, I simply download repository XML by the corresponding ID from the following URL: http://service.re3data.org/api/v1/repository/[repository_id]. The Figure3 is showing us the re3data.org Metadata Items for the registry entry.

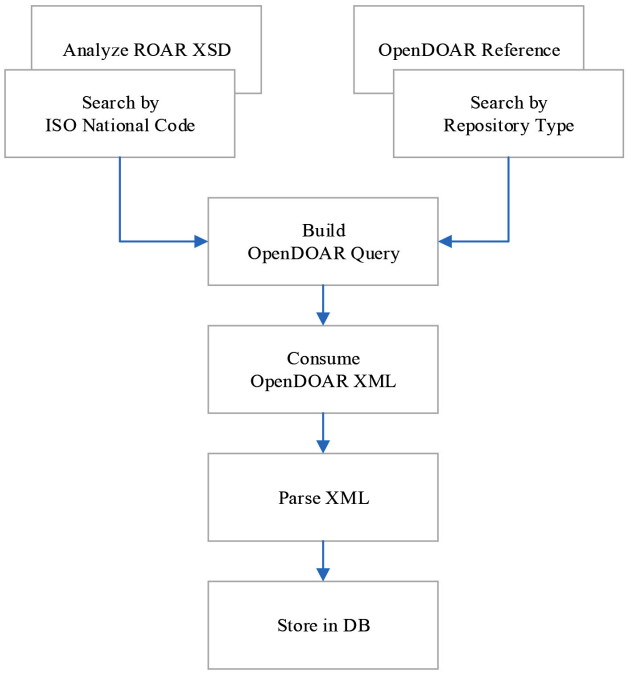

Fig. 1. Flowchart of OpenDOAR and ROAR Crawler (Kim and Lee, 2014)

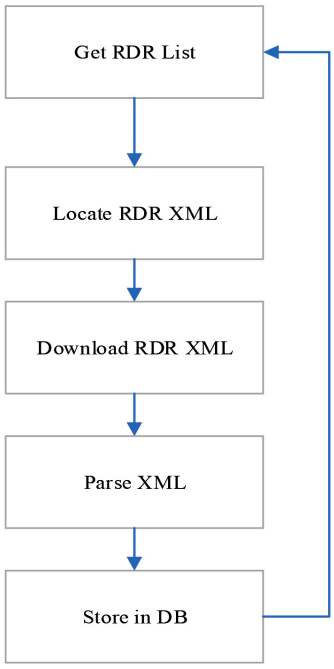

Fig. 2. Flowchart of re3data.org Crawler

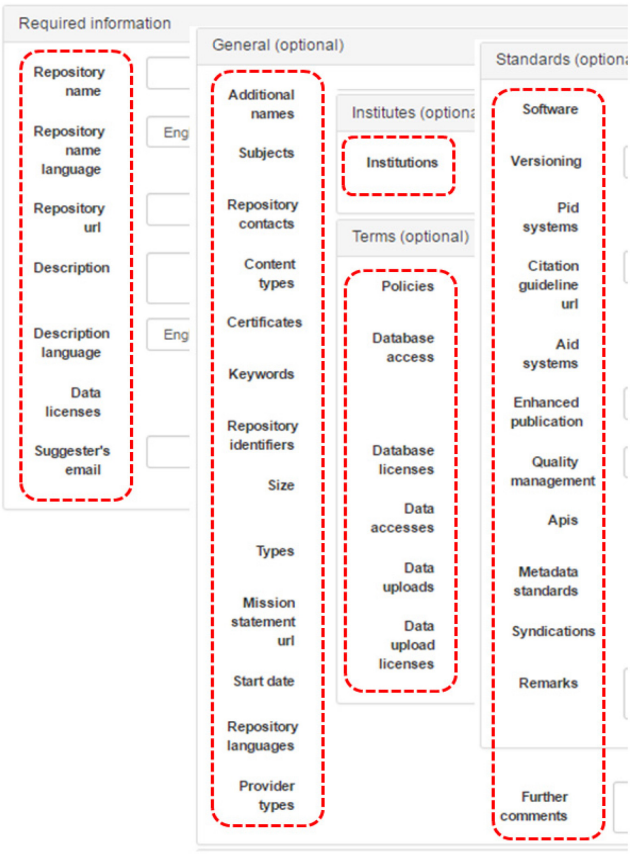

Fig. 3. re3data.org Metadata Items for registry entry

4. Results and Discussion

4.1 Repository Metadata Quality Assessment

4.1.1 Policy Type

The re3data controlled vocabulary supports the following policy types: Access Policy, Collection Policy, Data Policy, Metadata Policy, Preservation Policy, Submission Policy, Terms of Use Policy, and Quality Policy. Most RDRs (1307, 86.4%) have more than one policy in place. Unfortunately, re3data schema fails to specify this property. I believe there should be a required 'policyType' element, and users should be prompted accordingly. This would require changes to the existing schema and RDR listing interface. Also, the controlled vocabulary should be expanded. By way of example, 153 RDRs (10.1%) specify "Privacy Policy," and 41 RDRs (2.7%) specify "Service Policy," none of which is currently supported by the controlled vocabulary.

4.1.2 PID System and Institution Type

PID (persistent identifier) System is distributed as follows: DOI 322 (21.3%), HDL 108 (7.1%), URN and PURL 17 (1.1%) each, ARK 12 (0.8%), and other 73 (4.8%). 990 RDRs (65.4%) do not specify this property at all. Currently, 'pidSystem' element is optional, and available data seems to be inadequate for the analysis. On the other hand, this may indicate that PID System is of less invigoration than research record type (article, report, patent.) At the time of writing, 3267 institutions (with regard to multiple instances 4778) from 65 countries are listed on re3data.org. This means 1511 institutions join in building more than two RDRs. With consideration for multiple instances, 4630 (96.9%) are non-profit and 106 (2.2%) commercial organizations. Quite apparently, non-profit institutions are much more active. On the other hand, 42 institutions (0.9%) — e.g., Polish Geological Institute — do not specify institution type whatsoever. This means, they disregard re3data controlled vocabulary, which may need revision to incorporate missing types.

4.1.3 Metadata Standard

Metadata standards are outlined in Table 2 below.

Table 2. Metadata Standards of re3data.org RDRs

Metadata Standard	Count
Other	80
DDI (Data Documentation Initiative)	49
Dublin Core	38
ISO 19115 (Geographic information - Metadata)	24
FGDC/CSDGM (Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata)	14
RDF Data Cube Vocabulary	12
CF (Climate and Forecast) Metadata Conventions	9
EML (Ecological Metadata Language)	8
DataCite Metadata Schema	4
Darwin Core	3
SDMX - Statistical Data and Metadata Exchange	2
DCAT - Data Catalog Vocabulary	2
ABCD - Access to Biological Collection Data	1
Genome Metadata	1
ISA-Tab	1

Table 2 is showing the all metadata standards enlisted in RDRs. As can be seen from the table, except for the "Other," DDI and Dublin Core are most commonly encountered metadata standards. DDI is of frequent use in humanities. As pointed by Caplan (2003), DDI is particularly well suited for census and survey data, health and election statistics, and is widely used by governmental and

polling organizations, which then make data available for the researches. I believe this to be why DDI is so popular and tops the list. Following DDI, comes Dublin Core developed by DCMI and established by ISO 15836 as a standard for cross-domain resource description. This standard was designed with flexibility in mind, and widely occurs in a varied number of fields. Accordingly, it comes as no surprise that this standard commonly occurs across RDRs. It would be advisable for institutions and groups who aim at designing new RDRs, to take a close look at metadata standards from re3data.org, as this may help avoid lots of metadata-related problems with ease.

4.1.4 Subject and Keywords

Most popular subjects and keywords across re3data.org RDRs (as of the time of writing, March 2016) are outlined in Table 3 below.

Table 3. Top 10 Subjects and Keywords across re3data.org RDRs

Subject	Count	Percent	Keyword	Count
Natural Sciences	778	51.4	Genomics	84
Life Sciences	753	49.8	Bioinformatics	83
Medicine	601	39.7	Multidisciplinary	77
Biology	513	33.9	Biology	77
Geosciences (including Geography)	490	32.4	Biodiversity	72
Humanities and Social Sciences	415	27.4	Health	68
Basic Biological and Medical Research	292	19.3	DNA	64
Atmospheric Science and Oceanography	259	17.1	Genetics	62
Social and Behavioural Sciences	246	16.3	Meteorology	54
Physics	216	14.3	Climate	51

As can be seen from the table, natural (778, 51.4%) and life sciences (753, 49.8%) top the subject list, followed by medicine and biology. Most popular keywords are genomics (84) bioinformatics (83), multidisciplinary (77), biology (77), biodiversity (72), health (68), and DNA (64). This shows well the prevailing domain of the research data. Total percentage is not equal to 100 because an RDR can register many subjects and keywords at a time. According to the proposed schema, the ‘subject’ element is mandatory, while the ‘keyword’ element is optional; at that, both elements can occur multiple times. Even though keywords are optional, all analyzed RDRs (1513) specify this property. As they do so, three RDRs — e.g., “Open Data by Socrata” — skip subject (re3data.org IDs: r3d100011686, r3d100011777, r3d100011900). This suggests that guidelines for registration procedure on re3data.org need to be checked and updated accordingly.

4.1.5 Data Upload, Data License, Database License, Data Access

The mandatory ‘dataUpload’ element can occur multiple times and has two child elements: ‘dataUploadType’ and ‘dataUploadRestriction’. Data Upload Type is distributed as follows: closed

(579), open (54), restricted (871). Nine RDRs — e.g., "Coriolis" — do not specify this element at all. In turn, the 'data Upload Restriction' element is optional, and 632 RDRs omit this value. It seems that re3data schema is deficient here, and the controlled vocabulary needs further revision. Data License, along with its child elements 'dataLicenseName' and 'dataLicenseUrl', should be specified at least once. All analyzed RDRs have this element in place. This clearly indicates that institutions attach much importance to research data licensing. Table 4 below outlines data and databases licenses in use.

Table 4. Data and Database Licenses across re3data.org RDRs

Data License	Count	Percent	Database License	Count	Percent
Other	756	50	Copyrights	136	38.4
Copyrights	435	28.8	Other	122	34.5
CC	187	12.4	CC	51	14.4
Public Domain	96	6.3	Apache License 2.0	22	6.2
CC0	16	1.1	BSD	10	2.8
ODC	17	1.1	ODC	7	2
OGL	3	0.2	Public Domain	5	1.4
BSD	2	0.1	CC0	1	0.3
RL	1	0.1			

As can be seen from the table, 767 (50%) RDRs specify data license as "Other" that suggests 'data License Name' element needs revision so that ensure better control. The 'dataAccess' element is mandatory as well, and is specified across all RDRs. Its child element 'dataAccessType' is also mandatory, and according to the proposed schema can have the three following values: open, restricted, and closed. Across the analyzed RDRs these values are distributed as follows: open (758, 50.1%), restricted (590, 39.0%), and closed (105, 6.9%). What is more, 60 RDRs (4.0%) specify data access as "embargoed." This value is not defined by the schema that clearly indicates this attribute should be properly updated.

4.1.6 Content Type, Citation Reference, API

The 'contentType' element is optional and repeatable. It can have 17 type values, among which "Scientific and statistical data formats" (968 RDRs, 64%) is most commonly occurring, followed by "Standard office documents" (885, 58.5%). If I look at "Raw data" and "Images," it may be concluded that they refer to quite different things. This suggests that the proposed values are not granular enough and need further consideration. This is further supported by the fact that 516 RDRs (34.1%) simply specify "Other." Supposedly, PARSE.Insight value defined by re3data schema is not sufficient. The 'citation Reference' and 'api' elements are not controlled. This may imply technical difficulties with repository descriptor or, again, suggest these properties are not needed at all. Be it as it may, this issue deserves further investigation.

5. Metadata Description Quality

Altogether, re3data.org schema specifies 91 properties. Among them, 41 properties are defined as wrapper elements having 50 children. In this study, I assess metadata quality from the point of whether or not mandatory elements are properly used and how accurately their values are specified. On this basis, I suggest the following five items for improvement. First, some properties are simply missing, as with 'policyType'. In such cases, re3data.org schema should be updated, and the corresponding interface provided for RDR registration process. Second, 'institutionType', 'policyType', 'dataUploadRestriction', 'dataAccessType', and 'contentType' controlled vocabulary should be revised for missing and inadequate values. Third, it is essential that properties such as 'subject' and 'data Upload Type' are duly specified. Fourth, 'citation Reference', 'metrics', 'api Type', 'api Url', 'api Documentation', 'institution Identifier', and 'institution Identifier Value' elements need further consideration as to controlled values. Fifth, 'institutionAdditionalName' along with its child element 'language', and 'size' along with its child element 'updated', are missing matching values. Table 5 below summarizes quality indices for each schema element.

Table 5. re3data.org Metadata Quality Indices

ID: re3data.org schema element ID; Property: element name; W/A/C: Wrapper, Attribute, Child; Occ: Occurrence; D/Q: Description Quality

ID	Property	W/A/C	Occ	D/Q	ID	Property	W/A/C	Occ	D/Q
1	Identifiers	W	1	**	19	policy	W	0-n	**
1.1	re3data	C	1	100	19.1	policyType	C	1-n	0
1.2	Doi	C	1	*	19.2	policyName	C	1	86.4
2	respositoryName		1	100	19.3	policyUrl	C	1	86.4
2.1	Language	A	Req	100	20	databaseAccess	W	1	**
3	additionalName		0-n	81.2	20.1	databaseAccessType	C	1	***
3.1	Language	A	Req	81.2	20.2	databaseAccessRestriction	C	0-n	***
4	repositoryUrl		1	100	21	databaseLicense	W	0-n	**
5	repositoryIdentifier		0-n	*	21.1	databaseLicenseName	C	1	23.4
5.1	repositoryIdentifierType	C	1	*	21.2	dtabaseLicenseUrl	C	1	23.4
5.2	repositoryIdentifierValue	C	1	*	22	dataAccess	W	1-n	**
6	Description		0-1	99.9	22.1	dataAccessType	C	1	100
6.1	language	A	Req	99.9	22.2	dataAccessRestriction	C	0-n	39.7
7	repositoryContact		0-n	*	23	dataLicense	W	1-n	**
8	type		1-n	99.7	23.1	dataLicenseName	C	1	100
9	size		0-1	39.3	23.2	dataLicenseUrl	C	1	100
9.1	updated	A	Req	39.1	24	dataUpload	W	1-n	**
10	startDate		0-1	76.6	24.1	dataUploadType	C	1	99.4
11	endDate	W	0-1	**	24.2	dataUploadRestriction	C	0-n	57.6
11.1	closed	C	0-1	0	25	dataUploadLicense	W	0-n	**
11.2	offline	C	0-1	0	25.1	dataUploadLicenseName	C	1	27.4
12	repositoryLanguage		1-n	99.9	25.2	dataUploadLicenseUrl	C	1	27.5
13	subject	W	1-n	**	26	software		0-n	83.9

ID	Property	W/A/C	Occ	D/Q	ID	Property	W/A/C	Occ	D/Q
13.1	subjectScheme	A	Req	99.8	27	versioning		1	47.9
13.2	subjectId	C	1	99.8	28	api	W	0-n	**
13.3	subjectName	C	1	99.8	28.1	apiType	C	1	0
14	missionStatementUrl		0-1	71.6	28.2	apiUrl	C	1	0
15	contentType		0-n	99.9	28.3	apiDocumentation	C	1	0
15.1	contentTypeScheme	A	Req	99.9	29	pidSystem		0-n	98.1
16	providerType		1-2	99.9	30	citationReference		0-n	0
17	keyword		0-n	100	31	metrics		0-n	0
18	institution	W	1-n	**	32	citationGuidelineUrl		0-1	57.4
18.1	institutionName	C	1	100	33	aidSystem		0-n	98.1
18.1.1	Language	A	Req	100	34	enhancedPublication		1	67.0
18.2	institutionAdditionalName	C	0-n	99.9	35	qualityManagement		1	89.0
18.2.1	Language	A	Req	100	36	certificate		0-n	15.1
18.3	institutionCountry	C	1	100	37	metadataStandard	W	0-n	**
18.4	responsibilityType	C	0-n	100	37.1	metadataStandardName	C	1	16.4
18.5	institutionType	C	0-1	99.6	37.2	metadataStandardUrl	C	1	16.4
18.6	institutionUrl	C	0-1	***	38	syndication	W	0-n	**
18.7	institutionIdentifier	C	0-n	0	38.1	syndicationType	C	1	33.1
18.7.1	institutionIdentifierType	C	1	0	38.2	syndicationUrl	C	1	33.1
18.7.2	institutionIdentifierValue	C	1	0	39	remarks		0-1	68.3
18.8	responsibilityStartDate	C	0-1	21.5	40	entryDate		1	100
18.9	responsibilityEndDate	C	0-1	6.1	41	lastUpdate		1	100
18.10	institutionContact	C	0-n	***					

* Downloaded XML does not contain the corresponding element.
** Inasmuch the corresponding element is a wrapper, Description Quality is assessed for its child element.
*** Crawler error; analysis deferred.

On the strength of metadata quality assessment, I identify the following problems. “Coriolis” RDR does not specify ‘repositoryLanguage’, whereas the element should appear at least once. The ‘size’ element cannot be repeated; 919 RDRs (60.7%) supply empty string whereas a value is required. With the exception of “Digitale Sammlungen, Goethe-Universität Frankfurt am Main”, all RDRs specify ‘description’ along with its child element ‘language. It seems appropriate to consider making this element mandatory. As of March 2016, re3data.org description XML does not contain citing DOI. On the other hand, citing DOI is available on individual RDR web pages. It seems reasonable to add this information into description XML. (In that event, Crawler should be updated correspondingly.) The schema should be modified so that the ‘updated’ attribute of the ‘size’ element is defined. On the other hand, the ‘size’ element occurs only in 594 RDRs. The ‘updated’ attribute is specified by 592 RDRs. The difference is due to the fact that two RDRs (re3data.org IDs: r3d100011904 and r3d100011908) omit this attribute in violation of the schema. Many RDRs fail to specify ‘institutionIdentifier’. This case needs further investigation, however, as there are reasons to think that institutions simply do not consider it worthwhile.

6. Conclusion

In this study, I collected and analyzed listed RDR description data from re3data.org, and evaluated them against the proposed schema with the intent of its improvement. To do this, I developed a Crawler, which retrieves necessary data. Based on the analysis results, I identify five basic issues. First, two issues that mandatory elements are missing ('subject', 'dataUploadType'). Second, an issue that mandatory element value is missing ('policyType'). Third, five issues that re3data controlled vocabulary is insufficient or inconsistent ('institutionType', 'policyType', 'dataUploadRestriction', 'dataAccessType', and 'contentType'.) Fourth, some optional elements are not defined by design ('citationReference', 'metrics', 'apiType', 'apiUrl', 'apiDocumentation', 'institutionIdentifier', and 'institutionIdentifierValue'.) Fifth, there are two cases that do not have paired descriptions, such as 'institution Additional Name' along with its child element 'language' and 'size' along with its child element 'updated'. The above issues can be addressed by governing mandatory and optional usage of the corresponding elements, and adjusting the controlled vocabulary. Also, the process of RDR metadata submission can be further controlled by the system. The ethical issues on data quality are very important. Each repository's policy for these issues will be addressed in the future studies. I hope this results will help improve re3data schema, and may be of advantage researchers who analyze data repository trends.

References

Caplan, P. (2003). *Metadata fundamentals for all librarians*. American Library Association.

DRI. (2013). *Caring for Digital Content: Mapping International Approaches*. Royal Irish Academy. ISBN 978-1-908996-25-1

Jones, S. (2012). Curation policies and support services of the main UK research funders. Retrieved from http://www.dcc.ac.uk/sites/default/files/documents/RC%20policy%20overview%20v2.2.pdf

Kim, S., & Lee, W. (2014). Global data repository status and analysis: based on Korea, China and Japan. *Library Hi Tech, 32*(4), 706-722.

Lone, F., Rather, R., & Shah, G. J. (2008). Indian Contribution to Open Access Literature: A Case Study of DOAJ & OpenDOAR. Chinese Librarianship. *International Electronic Journal*, *29*.

Norris, M., Oppenheim, C., & Rowland, F. (2008). Finding open access articles using Google, Google Scholar, OAIster and OpenDOAR. *Online Information Review, 32*(6), 709-715.

Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., Goebelbecker, H., Gundlach, J., Schirmbacher, P., & Dierolf, U. (2013). Making Research Data Repositories Visible: The re3data.org Registry. *PLOS ONE, 8*(11), 1-10. DOI: 10.1371/journal.pone.0078080

Shafi, S. M., Gul, S., & Shah, T. A. (2013). Web 2.0 interactivity in open access repositories. *The Electronic Library, 31*(6), 703-712.

Web References

re3data.org. http://re3data.org/. (accessed 31 March 2016)

[About the authors]

Suntae Kim is a principal research engineer in the division of advanced information at the Korea Institute of Science and Technology Information. He works for the University of Science & Technology as an associate professor. He received his Ph.D. degree in library and information science from Chonbuk National University. Before his current appointment, he worked as a computer program developer at Linksoft which developed the NDSL and NOS. His research interests include research data management, research data platform, research data sharing, semantic web, metadata.

Myung-Seok Choi received his B.S., M.S and Ph.D degree in Dept. of Computer Science from KAIST. He works for Korea Institute of Science and Technology Information(KISTI) as a senior researcher. He has experience in big data processing/analytics and content management. His research interests are in the area of research data management, scientific data sharing and big data analytics.