Global Data Repository Status and Analysis: Based on Korea, China and Japan Data in re3data.org*

Suntae Kim**

ARTICLE INFO

Article history:
Received 30 January 2018
Revised 15 February 2018
Accepted 7 March 2018

Keywords:
Data Repository,
IDR,
Institutional Data Repository,
Data Repository Registry,
Re3data.org

ABSTRACT

We collected and analyzed data from e3data.org, which is a global registry of data repository services. We analyzed data profile for three leading Asian economies–Korea, China, and Japan–against the reference data for other participating countries. In particular, we examined how individual countries contribute to the repository, organizational type, versioning and product quality management, and subject tagging. We come to the conclusion that all three Asian countries still fall short in terms of involvement. As for participating institutions, there are 7 from Korea, 64 from China, and 120 from Japan. Among Chinese organizations, 3 are profit, 61 non-profit, and 37 organizations (which yields 1.8%) are involved in repository building. In Japan, there is 1 is commercial and 119 non-profit organizations, of which 57 (3.0%) are involved in repository building. All 7 organizations from Korea are non-profit, and 6 of them (0.3%) are involved in repository building. As regards versioning and product quality management, Korea, China, and Japan are up to par with other countries. Subject analysis reveals that Korea contributes more to geosciences, Japan to physics and geosciences, while China, unlike Korea and Japan, is more active in life sciences. It is hoped that this study will help planning domestic infrastructure for research data repositories with proper consideration for specific research domains and national characteristics.

1. Research Objectives

Today is the time of fourth-generation research paradigm based on data retrieval instrumentation. Accordingly, long-term storage and open access to research data so that academic community could reproduce them becomes the major challenge. Long-term data storage and access requires data collection, preservation, management, and warehousing facilities. In other words, data repositories are required. More and more universities and research centers are starting to build research data repositories allowing permanent access to data sets in a trustworthy environment. Newly developed

* This research was supported by Korea Institute of Science and Technology Information (KISTI).
** Researcher, Korea Institute of Science and Technology Information, Korea (stkim@kisti.re.kr)
International Journal of Knowledge Content Development & Technology, 8(1): 79-89, 2018.
http://dx.doi.org/10.5865/IJKCT.2018.8.1.089

research data management services register their repositories with re3data.org (Vierkant et al., 2014). re3data.org offers researchers, funding organizations, libraries and publishers an overview of the heterogeneous research data repository landscape (Pampel et al., 2013). re3data.org is a global registry of research data repositories that covers research data repositories from different academic disciplines (Vierkant et al., 2014). The focus of re3data.org is on research data repositories and despite its size the re3data.org registry does not claim global coverage (Klump & Huber, 2017). At the time of this study (February 2018), it contains 2,029 registered data repositories. Among them there are such members as International Mouse Phenotyping Consortium cooperatively managed by 16 organizations from 12 countries. Another example is Datenbank Gesprochenes Deutsch, which is a nation-wide repository managed by a specific country (Germany in this case.) Also, services that allow publishers to access and publish data on a long-term basis are emerging around nongovernmental organizations, developing Data Citation Index and providing repository quality assessment. This shows that new core content of scientific information is quickly emerging. On the one hand, we observe a positive trend of moving towards a shared global environment for research data. On the other hand, we also see that international competition for priority is gaining momentum. This means, it is time to take active consideration on global data repository management and national participation strategy. Accordingly, a close analysis of different country profiles is needed, based on information available from re3data.org, which is a global registry of research data repositories. In this study we examine re3data.org repositories where three leading Asian economies (China, Japan, and Korea) are involved, and compare national profiles. In particular, this article will tell us which organizations contribute to the repository? And what kind of organizational types are? And who do the data versioning and product quality management? And who do the subject tagging service?

2. Previous research

Open DOAR and ROAR are the most used repository registration services. In this study, however, we examine only the repositories registered with re3data.org. Accordingly, we do not consider previous research on OpenDOAR and ROAR. Recently, many studies on re3data.org have been made. Initially, they were mostly introductory in nature. Pampel et al. (2013) describes the heterogeneous RDR landscape and presents a typology of institutional, disciplinary, multidisciplinary, and project-specific RDR, outline the features of re3data.org, and shows how this registry helps to identify appropriate repositories for storage and search of research data. Hayslett (2015), based on data librarian experience, points out the importance of data standardization at all levels, and introduces re3data.org as an example. Elger, Pampel, Vierkant, and Witt (2016) describe how new re3data.org features, such as user-friendly icon system and Application Programming Interface (API), help collect re3data metadata. According to Pampel et al. (2016), publishers like Copernicus Publications, PeerJ, and Nature's Scientific Data make it their editorial policy to recommend researchers re3data.org as a relevant platform for data store. Klump and Huber (2017) examined the impact of persistent identifier system, including how it facilitates repositories registered

withre3data.org. ShaSha, ZHANG. GuoBin, HUANG. Qian and GENG (2017) analyzed four major features (responsibility subjects, platform functions, data resources, and data transmissions) across 247 data publishing platforms, including re3data. Kim and Choi (2017) conducted an examination of metadata registered with re3data.org and carried out product quality assessment. Generally, re3data.org has been studied intensively for many important aspects from simple introduction of the service through features and properties of the registered data to the quality of provided metadata and in-depth analysis of the registered repositories themselves. However, there has been no investigation using re3data.org metadata to deduce how different countries participate in data repositories. Accordingly, this study retrieves and analyzes information for involved organizations related to major Asian economies (Korea, China, and Japan).

3. Material and Methods

The collected data are stored in a relational database and evaluated against the proposed re3data.org schema. For this purpose, we use data collection crawler developed by Kim and Choi (2017) with some modification of access point and data retrieval protocol. The collected data are converted to the form suitable for further analysis and stored in the database. At the time of this study (February 2018) we collected information for 2,029 data repository entries, 12,126 subject entries, 3,185 policy entries, 2,028 licensing entries, 2,025 data access entries, 2,011 data upload entries, 6,275 institution entries (4,185 with full mapping), 15,038 keyword entries, 606 metadata standardization entries, 2,014 identifier entries, 2,021 repository responsibility entries, and 2,427 repository type entries.

There are, however, some questions with regard to collected data content and quality. First, the lack of identifier. re3data.org does not provide institution identifiers, so we need to use mapping. re3data supports the following entries for institutions: the name (institutionName), country (institutionCountry), responsibility type (responsibilityType), institution type (institutionType), web address (institutionURL), responsibility start date (responsibilityStartDate), responsibility end date (responsibilityEndDate), contact (institutionContact). The problem, however, is that no identifier is provided. In the end, we do not examine institution type and responsibility type. For the purpose of our analysis, we group institutions by name. Later on, more detailed analysis will be required, which would consider full and abbreviated name of the institution to construct the Authority file. Second, some country codes contain inaccurate values. For example, there are 224 instances of "AAA" country code for institutions registered with the repository. In this study, we use country codes as is, without adjustment. Figure 1 shows how the crawler retrieves the data from re3data.org and the database schema for storing the collected data.

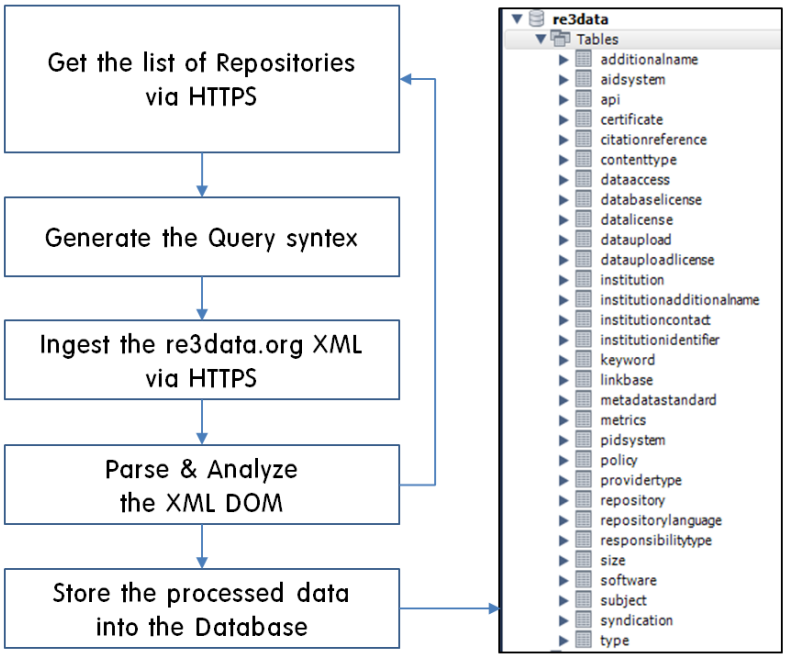

Fig. 1. re3data.org crawling process and database schema

4. re3data.org data analysis

In this study we analyze repository info, institution info, versioning and product management info, as well as subject tags collected from re3data.org.

4.1 Participation by countries

International Mouse Phenotyping Consortium is a repository where the maximum number of countries are involved. In all, there are 12 participating countries, including China, Japan, and Korea, represented by 16 institutions. Next comes Fishbase repository with 10 participating countries represented by 11 organizations. As regards repositories managed by a single state, USA takes the lead with 711 repositories, followed by Germany (202), England (133), Canada (102), and Australia (56). Japan solely manages 31 repositories, China 25, and Korea only 3. Table 1 below shows the number of repositories managed by individual country. Table 2 shows the number of repository involved institutions by country. There are 2,613 institutions from the USA, 766 from Germany, and 561 from England. Korea is represented by 7 institutions, China 64, and Japan 120.

Table 1. Number of repositories managed solely by country (as of February 2018)

Country code	Number of Repositories	Country code	Number of Repositories	Country code	Number of Repositories
USA	711	CAN	102	JPN	31
DEU	202	AUS	56	CHN	25
GBR	133	FRA	33	EEC	24

Table 2. Number of participating institutions by country (as of February 2018)

Number	Country code	Number of Institutions	Number	Country code	Number of Institutions
1	**USA**	**2613**	8	AUS	222
2	DEU	766	**9**	**JPN**	**120**
3	GBR	561	10	CHE	112
4	CAN	351	11	NLD	92
5	EEC	242	**12**	**CHN**	**64**
6	FRA	235	13	IND	61
7	N/A	224	**37**	**KOR**	**7**

Korea participates in 6 global data repositories, China in 37, and Japan in 57. Table 3 shows the number of participating institutions for the aforementioned repositories in descending order. As to International Mouse Phenotyping Consortium and International Ocean Discovery Program, Korea, China, and Japan are collectively participating in these repositories.

Table 3. Number of participating institutions for repositories where Korea, China, and Japan are involved (as of February 2018)

Country code	Repository Name	Access Type	Version Management	Product Management	Participating Countries	Participating Institutions
Republic of Korea	International Mouse Phenotyping Consortium	open	yes	yes	12	16
Republic of Korea	International Ocean Discovery Program	open	yes	yes	9	12
Republic of Korea	WDC Sunspot Index and Long-term Solar Observations	open		yes	3	4
China	International Mouse Phenotyping Consortium	open	yes	yes	12	16
China	International Ocean Discovery Program	open	yes	yes	9	12
China	Fishbase	open	yes	N/A	10	11
Japan	International Mouse Phenotyping Consortium	open	yes	yes	12	16
Japan	International Ocean Discovery Program	open	yes	yes	9	12
Japan	International Service of Geomagnetic Indices	open	yes	yes	7	11

4.2 Participating organizations

Table 4 shows institutions that are actively participating in repository building and management by countries. American National Science Foundation (175 repositories) and National Institutes of Health (67 repositories). European CLARIN-ERIC (31 repositories), German Bundesministerium für Bildung und Forschung (30 repositories) and Deutsche Forschungsgemeinschaft (23 repositories), British Wellcome trust (24 repositories) and European Molecular Biology Laboratory, European Bioinformatics Institute (24 repositories), Australian National Data Service(18 repositories) and National Collaborative Research Infrastructure Strategy (8 repositories), Canadian Environment and Climate Change Canada (17 repositories) and Government of Canada (12 repositories). Chinese National Natural Science Foundation of China (3 repositories), Japanese Kyoto University (4 repositories), ICSU World Data System (3 repositories). From Republic of Korea, including Korea Labor Institute and Ministry of Employment and Labor, altogether 7 institutions, among which 6 are non-profit organizations.

Table 4. Representative institutions involved in repository building and management from leading developed countries, Korea, China, and Japan (as of February 2018) (2 institutions randomly selected from the list ordered by countries.)

Institution Name	Country code	Number of Repositories
Korea Institute of Geology, Mining and Materials	KOR	1
National Space Weather Center, Korean Space Weather Center	KOR	1
National Natural Science Foundation of China	CHN	3
Chinese Academy of Sciences, Institute of Microbiology, Information Network Center	CHN	2
Kyoto University	JPN	4
ICSU World Data System	JPN	3
National Science Foundation	USA	175
National Institutes of Health	USA	67
Wellcome trust	GBR	24
European Molecular Biology Laboratory, European Bioinformatics Institute	GBR	24
CLARIN-ERIC	EEC	31
European Commission, Research & Innovation, Seventh Framework Programm - FP7	EEC	21
Environment and Climate Change Canada	CAN	17
Government of Canada	CAN	12
Australian National Data Service	AUS	18
National Collaborative Research Infrastructure Strategy	AUS	8
Bundesministerium für Bildung und Forschung	DEU	30
Deutsche Forschungsgemeinschaft	DEU	23

4.3 Participating institutions by type

Korea, China, and Japan have a minor role in data repository building and management. They participate in 95 international repositories; of which 46 repositories involve more than two countries among them. They also manage 49 national repositories (Korea 3, China 25, Japan 31). Examination of institutions involved in building re3data repositories reveals that 4,059 of them are non-profit organizations, 124 are commercial, and two organizations (National Center for Education Statistics and U.S. Department of Education) have no type specified. Both unspecified organizations, however, are non-profit in fact. Table 5 below shows participation rate for national, international, Korean, Chinese, and Japanese institutions. Korea participates in 7 repositories altogether, of which 6 are non-profit (0.3%), including Korean Labor & Income Panel Study. Also, 36 non-profit organizations from around the world are participating in repositories where Korea is involved. China is represented by 3 commercial and 61 non-profit organizations, including China Forestry Scientific Data Center, participating in 37 repositories altogether. 102 non-profit and 5 commercial organizations from around the world are participating in repositories where China is involved. In Japan, there is 1 is commercial and 119 non-profit organizations, including JEDI, of which 57 (3.0%) are involved in repository development. 1 non-profit and 196 commercial organizations from around the world are participating in repositories where Japan is involved.

Table 5. Global data repository participation rate (national, international, Korea, China, Japan)

International Participation Rate	National Participation Rate	Korea Participation Rate	China Participation Rate	Japan Participation Rate
74.8	25.2	0.3	1.8	3

4.4 Version management

Whether or not a repository supports version management is determined by "versioning" entry. Table 6 shows versioning support across the repositories. At the time of this study (February 2018), among 2,029 registered data repositories 47.6% (965 repositories) support data versioning. Accordingly, 52.4% (1,064) do not provide versioning services. Among the examined repositories, 83.4% (805 repositories) support versioning, while 16.6% (160 repositories) do not.

Korea manages independently three repositories: International Space Environment Service, AMODS, and Korean Labor & Income Panel Study, all of which support data versioning. China manages independently 25 repositories, including China Forestry Scientific Data Center, of which 28.0% (8 repositories) provide version management services, and 4% (1 repository) do not. For the remaining 68.0% (17 repositories) versioning information is not available. Japan manages independently 31 repositories, including Japan Space Systems, of which 38.7% (12 repositories) support versioning, and 6.5% (2 repositories) do not. For the remaining 54.8% (17 repositories) versioning information is not available.

Table 6. Data repository versioning support (as of February 2018)

Item	NO NULL / NULL / YES / NO / TOTAL	TOTAL NULL / YES+NO / YES / NO (rate, incl. NULL)	YES / NO (rate, w/o NULL)
Global	965 / 1064 / 805 / 160 / 2029	52.4 / 47.6 / 39.7 / 7.9	83.4 / 16.6
KOR	0 / 3 / 0 / 0 / 3	100.0 / 0.0 / 0.0 / 0.0	N/A
CHN	8 / 17 / 7 / 1 / 25	68.0 / 32.0 / 28.0 / 4.0	87.5 / 12.5
JPN	14 / 17 / 12 / 2 / 31	54.8 / 45.2 / 38.7 / 6.5	85.7 / 14.3

4.5 Quality management

Whether or not a repository supports quality management is determined by "qualityManagement" entry. Table 7 shows quality management support across the repositories. Across all repositories, only 20 (1.0%) do not provide information about quality management. 730 repositories, however, indicate that this property is "Unknown." At the time of this study (February 2018), among 2,029 registered data repositories 99.0% (2,009 repositories) provide quality management info. Quality management info is not available only for 1.9% (20) repositories. Among repositories supporting quality management (including "Unknown" status), 61.8% provide quality management services. If we do not consider "Unknown," 97.0% (1,241 repositories) provide product quality management.

For three Korean repositories marked for quality management, one provides quality management services, while for the other two information is not available. For 24 Chinese repositories marked for quality management, 9 (36.0%) provide quality management services, for 15 repositories (60.0%) information is not available. For 31 Japanese repositories marked for quality management, 11 (35.5%) provide quality management services, 3 (9.7%) do not, for 17 repositories (54.8%) information is not available.

Table 7. Data repository quality management support (as of February 2018)

Item	NO NULL / NULL / YES / NO / UNKNOWN / TOTAL	NULL / YES+NO+UNKNOWN / YES / NO / UNKNOWN rate (incl. NULL)	YES / NO / UNKNOWN rate (w/o NULL)	YES / NO rate (NULL, w/o UNKNOWN)
Global	2009 / 20 / 1241 / 38 / 730 / 2029	1.0 / 99.0 / 61.2 / 1.9 / 36.0	61.8 / 1.9 / 36.3	97.0 / 3.0
KOR	3 / 0 / 1 / 0 / 2 / 3	0.0 / 100.0 / 33.3 / 0.0 / 66.7	33.3 / 0.0 / 66.7	100.0 / 0.0
CHN	24 / 1 / 9 / 0 / 15 / 25	4.0 / 96.0 / 36.0 / 0.0 / 60.0	37.5 / 0.0 / 62.5	100.0 / 0.0
JPN	31 / 0 / 11 / 3 / 17 / 31	0.0 / 100.0 / 35.5 / 9.7 / 54.8	35.5 / 9.7 / 54.8	78.6 / 21.4

4.6 Subject tagging

Only 17 repositories among 2,029 do not provide information about the subject. The number of subjects for a repository varies from 1 (8 repositories) to 29 (1 repository). The mode is 4 subjects per repository (472 repositories); the average is 6.

re3data uses German DFG "Research Area and Scientific Discipline" code for subject. As for scientific discipline, the most popular subject is "Life Sciences" (1,054 repositories), followed by

"Natural Sciences" (1,007), "Humanities and Social Sciences" (602), and "Engineering Sciences" (289). In the "Research Area" category, "Life Sciences" sub-category "Biology" (716) is the most frequent. Then go "Natural Sciences," "Geosciences" (including Geography, 629), "Medicine" (493), "Social and Behavioural Sciences" (316). Table 8 summarizes the number of occurrences of "Subject Area" code in all and for Korea, China, and Japan. In global repositories, the most frequent "Subject Area" is "Basic Biological and Medical Research" (424 repositories). Then go "Atmospheric Science and Oceanography" (347), and "Medicine" (317).

For Korea, the most frequent is "Geosciences" with most popular subject being "Atmospheric Science and Oceanography" (3 repositories). For Japan, the most frequent are "Physics" and "Geosciences" with most popular subjects being "Astrophysics and Astronomy" (17) and "Atmospheric Science and Oceanography" (15). For China, unlike Korea and Japan, the most frequent is "Life Sciences" with most popular subjects being "Basic Biological and Medical Research" (12) and "Zoology" (9).

Table 8. Global data repository subjects across China, Japan, and Korea (February 2018)

Subject Area	Global	KOR	CHN	JPN
Basic Biological and Medical Research	424	1	12	14
Atmospheric Science and Oceanography	347	3	7	15
Medicine	317	1	5	10
Geophysics and Geodesy	222	1	8	12
Zoology	204	1	9	4
Social Sciences	201			
Microbiology, Virology and Immunology	182		6	
Plant Sciences	163		7	
Astrophysics and Astronomy	153	1	3	17
Economics	147			
Geography	132			6
Water Research	122		2	6
Agriculture, Forestry, Horticulture and Veterinary Medicine	94			
Linguistics	82			
Geochemistry, Mineralogy and Crystallography	73	1		4
Optics, Quantum Optics and Physics of Atoms, Molecules and Plasmas	69	1		
Geology and Palaeontology	63	1	3	
Neurosciences	60			
History	57			
Computer Science	57			
Physical and Theoretical Chemistry	29	1		
Molecular Chemistry	39			4

5. Conclusion

In this study we collected and analyzed data from e3data.org, which is a global registry of data repository services. We analyzed data profile for three leading Asian economies–Korea, China, and Japan–against the reference data for the countries, which participate in repository management. In particular, we examined how individual countries contribute to the repository, organizational type, versioning and product quality management, and subject tagging.

We come to the conclusion that all three Asian countries still fall short in terms of involvement. They participate in 95 international repositories; of which 46 repositories involve more than two countries among them. At the same time, there are 2,613 institutions from the USA, 766 from Germany, and 561 from England. Korea is represented by 7 institutions, China 64, and Japan 120. Among Chinese organizations, 3 are profit, 61 non-profit, and 37 organizations (which yields 1.8%) are involved in repository building. In Japan, there is 1 is commercial and 119 non-profit organizations, of which 57 (3.0%) are involved in repository building. All 7 organizations from Korea are non-profit, and 6 of them (0.3%) are involved in repository building. As for versioning management, 47.6% (965 repositories) provide corresponding information. Accordingly, 52.4% (1,064 repositories) do not provide versioning services. Among the examined repositories, 83.4% (805 repositories) support versioning, while 16.6% (160 repositories) do not. Among 2,029 registered data repositories 99.0% (2,009 repositories) provide quality management info. As regards versioning and product quality management, China, Japan, and Korea are up to par with other countries. Subject analysis reveals that Korea contributes more to geosciences, Japan to physics and geosciences, while China, unlike Korea and Japan, is more active in life sciences.

In this study we only mention some problems related to the quality of the collected data, and perform simple statistical analysis. Speaking of source data quality, we expect that re3data.org can work on this. There are many fields such as ‘PIDsystem’, ‘policy’ and software etc. which contains the information not analyzed in this article. In the future study, these fields are going to be analyzed. For one thing, it is important for the future studies not simply apply basic statistics, but being able to carry out full-fledged comparative analysis of R&D activities by country.

At the moment, we believe that analyzing repositories based on re3data.org source data is a challenging task. In any case, it is hoped that this study will help planning domestic infrastructure for research data repositories with proper consideration for specific research domains and national characteristics. We believe this study can serve as a starting point for future works that will perform comprehensive analysis for different countries.

References

Elger, K., Pampel, H., Vierkant, P., & Witt, M. (2016, February). New Features of the re3data Registry of Research Data Repositories. In *AGU Fall Meeting Abstracts*, IN41D-04.

Hayslett, M. (2015). Data World Does Not Lack Standards. *Journal of Librarianship and Scholarly Communication, 3*(2), eP1245.

Kim, S., & Choi, M. S. (2017). Registry Metadata Quality Assessment by the Example of re3data. org Schema. *International Journal of Knowledge Content Development & Technology, 7*(2), 41-51.

Klump, J., & Huber, R. (2017). 20 Years of Persistent Identifiers–Which Systems are Here to Stay?. *Data Science Journal, 16*.

ZHANG, S., HUANG, G., & GENG, Q. (2017). Research on UK Scientific Data Publishing Platforms Based on Re3data. In *Digital Library Forum* (Vol. 6, p. 005).

Pampel, H., Vierkant, P., Elger, K., Bertelmann, R., Witt, M., Schirmbacher, P., ⋯ & Ulrich, R. (2016, April). re3data. org-a global registry of research data repositories. In *EGU General Assembly Conference Abstracts* (Vol. 18, p. 16765).

Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., ... & Dierolf, U. (2013). Making research data repositories visible: the re3data. org registry. *PloS one, 8*(11), e78080.

Vierkant, P., Spier, S., Rücknagel, J., Pampel, H., Fritze, F., Gundlach, J., Fichtmüller, D., Kirchhoff, M. A., Goebelbecker, H., Klump, J., Kloska, G., Reuter, E., Semrau, A., Schnepf, E., Skarupianski, M., Bertelmann, R., Schirmbacher, P., Scholze, F., Kramer, C., Witt, M., Fuchs, C., & Ulrich, R. (2014). Schema for the Description of Research Data Repositories.

[About the authors]

Suntae Kim currently 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 professor also. 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. He may be contacted at HYPERLINK mailto:stkim@kisti.re.kr.