Implementation of Research Data Platform: in the Perspective of Data Transfer

Sungho Shin
Research Data Sharing Center
Korea Institute of Science and Technology Information
Daejeon, South Korea
maximus74@kisti.re.kr

Young Ho Shin
Research Data Sharing Center
Korea Institute of Science and Technology Information
Daejeon, South Korea
shinyh@kisti.re.kr

Jin Young Kim
Research Data Sharing Center
Korea Institute of Science and Technology Information
Daejeon, South Korea
jykim@kisti.re.kr

Min Ki Kim
Convergence Service Center
Korea Institute of Science and Technology Information
Daejeon, South Korea
mk.kim@kisti.re.kr

Min-Ho Lee
Research Data Sharing Center
Korea Institute of Science and Technology Information
Daejeon, South Korea
cokeman@kisti.re.kr

***Abstract*— Recent explosive increase in research data and increasing complexity of science and technology are continuing to expand convergence research and international collaboration among various fields. Developed countries are actively promoting open science policies to share research results and processes to create new knowledge and value through convergence researches. South Korea is also establishing and promoting similar strategies to encourage the sharing and utilization of national research data by the federal government. In 2018, a national research data platform was designed, and a prototype of the platform was constructed. It has been operated to verify the platform. It is important to link and collect more research data from various research institutions for the success of the national research data platform. We examine the communication and network technologies for the platform to link repositories outside and present some implications.

Keywords—Research data; Open science; Research data platform; Data linkage; GridFTP; OAI-PMH

I. INTRODUCTION

The fourth-generation research paradigm is data-intensive research that finds new theories or phenomena through data analysis. In particular, research data is growing rapidly as experimental equipment develops in advanced fields such as astronomy, aeronautics, space, and genetics. The number of repositories registered at Re3data.org, which provides information on research data repositories around the world, is more than 1,500 (as of 2016) and have steadily increased over the past four years. Due to the explosive increase in research data and the increasing complexity of science and technology, convergence researches and international collaborations among various fields are continuously expanding. The UK and Germany jointly published about 10,000 papers in 2011 [1]. Developed countries are actively planning and carrying their open science policies to open research results and processes to create new insight and value through convergence researches. The United States has set guidelines for the management and sharing of research data in 2013, and the EU has launched the OpenAIRE2020 Project since 2015. In addition, management policies of research data, research data sharing, utilization infrastructure support, researcher guidelines and education are being promoted.

In South Korea, the National Science and Technology Council, led by the Ministry of Science, Technology and Communication, announced in January 2018 a national research data sharing and utilization strategy to promote the sharing and utilization of national research data. In addition, the National Science and Technology Research Council is also promoting the research Big Data project as part of the plan to enhance the environment of government-funded research institutes from February 2018.

In that sense, Korea government has been planning a research data platform from 2018 at the national level. In this study, we analyze the factors to be considered in the construction of the research data platform and suggest ways for data linkage among repositories outside in particular. To do this, we introduce the concept of research data, research data platform, and open science in chapter 2. We also analyze advanced research data platforms and draw implications in chapter 3. Chapter 4 describes the design elements of how to apply the implications into Korea research data platform.

978-1-7281-1340-1/19/$31.00 ©2019 IEEE

II. BACKGROUND

A. Open science

Open Science is a concept that aims to develop scientific research by digitizing all kinds of scientific knowledge (dissertation, data, methodology, educational materials, etc.), opening and sharing them, and stimulating joint research [2]. It is a mechanism that makes it easier for the general public as well as the research community to access and utilize research results and research process data publicly supported by public funds in a digital format. International organizations and developed countries are actively promoting the open science program to create, share, and spread new insights and values. EU is also pursuing programs and policies aiming at promoting open science through the 2014 Horizon 2020 project. UNESCO provides the status of scientific information open access of 158 countries around the world through GOAP and provides scientific information (journals, conference papers, various types of data sets) from publicly funded research through ROAD services.

B. Research data

The main components of Open Science are Open Access, Open Data, and Open Collaboration. Open Data refers to research data, observation and experimental data, computer data, reference data, and metadata produced during the whole research process [2]. Open Data serves as a catalyst for collaborative research to support data-driven research and global problem solving. For example, environmental problems such as global warming, such as yellow sand and climate change, and marine pollution and ecosystem change, can be solved by analyzing data from various fields such as weather, ocean, astronomy, geographical information and ecology as well as one field of data. Similar to open data, research data is defined by the UK research committee that is produced, managed, shared and utilized through research activities, as well as data produced at the research stage as well as the final results [3]. The below are typical types of research data [4].

- **Observation data**. Observations are captured by observing an action or activity. It is collected using methods such as human observation, open investigation, or the use of tools or sensors to monitor and record information. Observation data is captured in real time, so it is very difficult or impossible to recreate when lost.
- **Experimental data**. Experiment data is collected by researchers actively to create, measure, or make a difference when a variable changes. Experimental data are typically projected to a larger population, allowing researchers to determine causality. This type of data is often reproducible, but it is costly to do so.
- **Simulation data**. Simulation data is produced with a computer test model to follow the behavior of actual processes or systems over time. For example, predict weather conditions, economic models, chemical reactions, or seismic activity. This method is used to identify situations that may occur in a particular situation. The test models used are often more important than the data generated in the simulations.

III. COMPARIZON OF ADVANCED RESEARCH DATA PLATFORMS

Developed countries have long been promoting research data sharing through their research data platform. By analyzing the advanced research data platform from some developed countries, we can understand factors to consider when constructing a research data platform in Korea. Table 1 compares research data platforms of developed countries.

Table 1. Comparison of Research Data Platforms

	OpenAIRE	EUDAT	ANDS	RCOS
Ownership	EU	EU	Australia	Japan
Project Period	2009~	2012~	2009~	2017~
Core Services	Data registration, Data search, Data link, Data verification	Data search, Data archive	Data search, Data visualization, Data link, Data mgmt., Research-Vocabularies	Data search
Others	Dashboard	Virtualization	RIF-CS (Registry Interchange Format Collections and Services)	Cloud (service/SW/ storage)
Data Harvesting	Zenodo	CKAN	CKAN	RDM-COS
Data Sharing	API (OAI-PMH)	Large file transmission (GridFTP)	API (OAI-PMH)	-

The factors that should be considered when constructing a research data platform are as follows.

Firstly, as many research data as possible should be gathered into the platform. To do this, it is essential to link data with schools, research institutes, and communities of research fields that hold research data.

Secondly, it is necessary to provide the loaded data to the users through various types of search. Most of them provide data retrieval service through research data portals. It should be able to reflect various search conditions such as basic search, facet search, map search, etc.

Thirdly, an environment that can utilize (or reuse) collected research data is provided to users. With the exception of OpenAIRE, the rest of the platforms provide a virtual environment in which data can be utilized. The main purpose of researchers needing research data is to gain insights through analysis. This requires an infrastructure environment, such as computing resources, analytical tools, and storage.

IV. KOREA RESEARCH DATA PLATFORM

Based on the analysis results of the advanced research data platforms, Korea research data platform has been designed shown in Fig. 1. The main services are to provide one-stop meta search service for domestic and overseas research data, and an AI or other analysis environment for research data set from search results. In particular, the research data platform is designed to provide an infrastructure environment to support convergence analysis of research data set. Convergence analysis service is based on deep learning or artificial intelligence. The service provides a distributed deep learning framework that interfaces with a distributed storage framework to provide rapid analysis performance for multiple users. It also supports a user-friendly environment for easily configuring data, models, and analysis tools by using the workflow. A command line interface for advanced users is available.

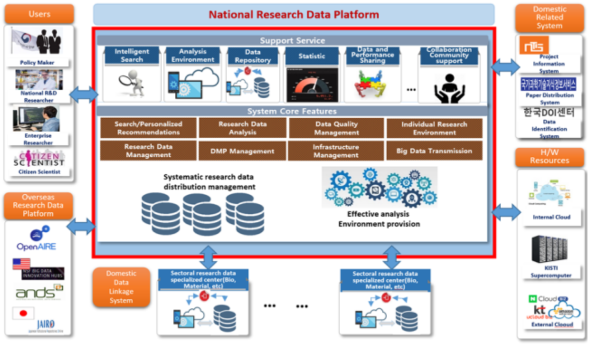
Users
Policy Maker
National R&D Researcher
Enterprise Researcher
CITIZEN SCIENTIST
Citizen Scientist
Overseas Research Data Platform
OpenAIRE
NSF BIG DATA INNOVATION HUBS
ands
JAIRO
National Research Data Platform
Support Service
Intelligent Search
Analysis Environment
Data Repository
Statistic
Data and Performance Sharing
Collaboration Community support
...
System Core Features
Search/Personalized Recommendations
Research Data Analysis
Data Quality Management
Individual Research Environment
Research Data Management
DMP Management
Infrastructure Management
Big Data Transmission
Systematic research data distribution management
Effective analysis Environment provision
Domestic Data Linkage System
Sectoral research data specialized center(Bio, Material, etc)
... ...
Sectoral research data specialized center(Bio, Material, etc)
Domestic Related System
NTIS
Project Information System
국가과학기술지식정보서비스
Paper Distribution System
한국DOI센터
Data Identification System
H/W Resources
Internal Cloud
KISTI Supercomputer
Cloud
kt
ucloud biz
amazon
External Cloud

Figure 1. Overall Structure of Korea Research Data Platform

Above all, it is most important to link research data possessed by domestic and foreign institutions. Domestic and foreign institutions have accumulated research data in their repository. There is no standardized repository but repositories fitted to each institution. It is not easy for the platform to link data of research institutions because of repository environment and different communication protocols. Since research data linkage is now in its infancy, it is difficult to expect physical support from research institutions. Therefore, the platform actively needs to implement the linkage for each institution. As a result, uncertainty increases as it reflects various environments. To solve this problem, the types of organizations to be linked according to communication protocols are divided into several groups.

Fig. 2 defines the linkage type for the institutions. There are two major communication protocols among the institutions. One is Restful API which is commonly used for the purpose of reflecting data characteristics. The other is Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), which has been widely used for gathering metadata. OAI-PMH is a protocol developed to collect metadata descriptions of records in an archive so that services can be constructed using metadata [5]. Except for the linkage to research institutions, the platform can collect research data from users directly. Users can upload their data through web interface in the portal.

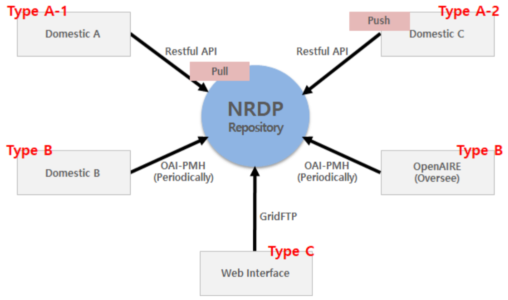

Figure 2. Types of Linkage for Korea Research Data Platform

Another factor that needs to be considered is the application of a high-speed transmission for big research data set. Unlike OpenAIRE platform, which has only metadata, EUDAT has data set files, so GridFTP is applied for fast transmission of data set files.

GridFTP is an extension of the File Transfer Protocol (FTP) for Grid Computing [6]. The purpose of GridFTP is to provide a more reliable and high-performance file transfer, for example, to enable the transfer of very large files. GridFTP uses multiple concurrent TCP streams to achieve much more bandwidth use than traditional data stream technology [7]. Files can be downloaded piece by piece from multiple sources at the same time. Conventional FTP does not support transferring only certain parts of a file. GridFTP allows the transfer of subsets of files. These features are useful for applications that require only a small section of a very large data file for processing.

Korea research data platform is also designed to contain research data set files as well as meta data, needs to support the transfer and upload/download of research data through the GridFTP or a advanced protocol. The platform has been implemented as a demonstration system in 2018 and is under test operation. Most of the services and functions we have designed are implemented, but they are at the prototype level to verify the functions and services. Data linkage has been done with four domestic platforms and one overseas platform. Total number of meta data collected is over 660,000 (Some of them have their data set). OAI-PMH and Restful API are all implemented for linking with the outside, and GridFTP is also applied for large data upload and download for web users.

V. Conclusion

Sharing and reuse of research data is necessary for the presence of open science. Developed countries have been implementing open science through research data platforms for many decades. Korea research data platform is now in its early stages of construction. First of all, much more research data must be accumulated. To this end, it is necessary to improve the data linkage and transmission technology with the institutions that have research data repositories. More various interconnection methods need to be devised and faster transmission technology needs to be developed and applied in the future.

Acknowledgment

This research is supported by Korea Institute of Science and Technology Information (KISTI).

References

[1] J. Adams, “Collaborations: the rise of research networks,” Nature 490, pp. 335–336, 2012. (doi: 10.1038/490335a)

[2] OECD, “Making open science a reality,” 2015.

[3] UK multi-stakeholder group, “Concordat on open research data,” 2016.

[4] Research guides, https://libguides.macalester.edu

[5] https://en.wikipedia.org/wiki/Open_Archives_Initiative_Protocol_for_Metadata_Harvesting

[6] W. Allcock, J. Bresnahan, R. Kettimuthu, and M. Link, "The Globus Striped GridFTP Framework and Server," ACM/IEEE SC 2005 Conference (SC'05). pp. 54.

[7] Luis Manuel Sarro,Laurent Eyer, William O'Mullane, and Joris De Ridder, “Astrostatistics and Data Mining,” Springer, 2012th Edition.