Study of Analyzing Outcome of Building and Introducing System for Preserving Full-Text of e-Journal

Kwang-Young Kim*, Soon-Young Kim**, Hwan-Min Kim***

ARTICLE INFO

Article history:
Received 15 July 2012
Revised 5 October 2012
Accepted 20 November 2012

Keywords:
Full-text Preservation,
e-Journal Full-text,
Full-text Management System,
Archiving System,
Archiving Center

ABSTRACT

Today, most researchers conduct their studies through the full-text of e-journals. Therefore, an important base for domestic development of science and technology is to obtain the full-text of quality e-journals by overseas researchers and to provide it to Korea's researchers.
This study aims to build a system based on the National Archiving Center for the full-text of e-journals and to make a service system for providing them to the public by acquiring the full-text of quality overseas e-journals. To do this, an analysis was made of the outcome of introducing such a system for full-text of e-journals in comparison with the investment.
As a result, 112 more institutions, that is, from 47 institutions to 159 institutions, have introduced the system as of 2012, and the number of downloaded full-texts increased at least 2.17 times.

1. Introduction

Many researchers use e-publication data to conduct their studies by means of highly developed Internet and e-publication technology. Fernandez-Cano, Torralbo, and Vallejoa (2004) said that the volume of science and technology information including e-journals has steadily increased over the last 30 years. Although most of the previously available information was about academic bibliography data, users now want to read full-text. Jung (2008) said that the unit fee for e-journal subscription had not increased until e-media appeared, but then rose steadily, resulting in sharply increasing expenses required by libraries for e-journal subscription. It is difficult to quantitatively and objectively measure the fundamental value of this service, due to the various features, intangible assets, and public good generated by providing access to the original copies of e-journals.

Seol (2005) wrote that Korea is still in a blind spot of digital preservation. He also noted the

* Senior Researcher, Department of Overseas Information, Korea Institute of Science and Technology Information, Korea (kykim@kisti.re.kr)
** Senior Researcher, Department of Overseas Information, Korea Institute of Science and Technology Information, Korea (maya@kisti.re.kr)
*** Senior Researcher, Department of Overseas Information, Korea Institute of Science and Technology Information, Korea (mrkim@kisti.re.kr) (Corresponding Author)
International Journal of Knowledge Content Development & Technology, 2(2): 5-15, 2012.
[http://dx.doi.org/10.5865/IJKCT.2012.2.2.005]

discussion on encouraging awareness of improved preservation in the forum for digital heritage preservation held by the Korean National Commission for UNESCO and the Information Trust Center and supported by the National Digital Library (NDL), and there is no sign that the issue of preservation is dealt with for managing national knowledge resources although the NDL promotes the Digital Resource Preservation Project. Lee (2004) defined the purpose of digital document archiving as the safe preservation of digital documents, and all activities for enabling access to the documents and keeping the original copies thereof even after long periods of time elapse. Current public interest in e-document preservation stems from the sharply increased production and distribution of digital information, following the development of information technology.

To respond to this trend, the so called Paper to Digital movement, it is essential to build a digital archiving system for developing digital information resources which implements the dynamic and simple creation of added values and for systematically preserving and using the resources. It is also necessary to build a digital full-text preservation system for establishing a long-term preservation system for national e-information resources.

KISTI (Korea Institute of Science and Technology Information) tries to lower the subscription fee of globally recognized e-journals through a group subscription to e-journals of the KESLI (Korean Electronic Site License Initiative), and to enable users to join the consortium to use the full-text of e-journals although they do not subscribe to the printed journals thereof, for users' convenience.

Because the full-text of Korea's academic essays is collected and provided by the KISTI, users in Korea can access it easily. On the other hand, because most of the full-text of overseas e-journals is possessed by the relevant publishers, users must contact the publishers to get the full-text of overseas e-journals. If the publishers close their site or small-sized publishers stop business, it is almost impossible to get the full-text easily.

Therefore, this study aims to build an NAC (National Archiving Center) base system in order to obtain the full-text of overseas e-journals, and to build a system for providing the full-text of e-journals to users. To do this, an analysis is made of the outcome of introducing the systems in comparison with investment in the full-text of e-journals.

2. Related Studies

The NDSL is a portal site for providing a variety of digital resources including Korean and overseas essays, patents, and reports, managed by the KISTI (Korea Institute of Science and Technology Information). The current number of resources provided is more than 100 million essays, patents, trends, standards, and factual information. The provided full-text of e-journals includes 4,385 types and 5.7 million cases. The amount of meta-information of overseas e-journal essays is about 6,400.

In Korea, archiving digital content is still in a preliminary phase, and national institutions including the National Assembly Library, the KERIS and the KISTI, private institutions, including Booktopia and DBPIA, all store and manage digital contents. Choi and Lee (2005) wrote that the digital archive is not a true archive because there is no system for long-term preservation and access.

Kwak (2010) and Jung, Choi, and Choi,(2009) both observed that the National Library of Korea and the National Archives of Korea have a long-term preservation strategy, and the KISTI is currently establishing a strategy to enable permanent access to digital resources.

In the study conducted by Kwak (2010), exemplary cases of national libraries or memory institutes that are aware of the significance of digital information resources for collecting data on the basis of government or national support include the NDIIPP (National Digital Information Infrastructure and Preservation Program) of the U.S., and the DCC (Digital Curation Centre) and the JISC (Joint Information Systems Committee) of the UK. Other exemplary cases include the PANDORA (Preserving and Accessing Networked Documentary Resources of Australia) established by the National Library of Australia, and the Digital Archiving Project by the NII-REO (National Institute of Informatics - Repository of Electronic Journals and Online Publication) of Japan.

The National Library of the Netherlands (KB) archives digital publications from countries all across the world as well as its own digital repository data, web archives, and digitalized data in order to cope with aging media and the fast evolution of software and hardware platforms by means of the e-Deport. Most of the more than 12 million e-journals collected by the e-Deport from 2003 to December 2008 are publications by internationally recognized publishers including Elsevier, Springer, and the like.

Portico started with the Electronic-Archive Initiative launched by JSTOR with the financial support of the Andrew W. Mellon Foundation in 2002. The new e-archiving service began in 2005 as a nonprofit service for providing regular archives of e-journals, which is now financially supported by the LC (Library of Congress) as well as the Andrew W. Mellon Foundation. The type of archived e-journals include born-digital e-journals published in an electronic format provided by participating publishers, e-journals printed and electronically published, and e-journals digitalized from the original printed copies.

The LOCKSS (Lots of Copies Keep Stuff Safe) is an international program based on the Stanford University Libraries and provides digital preservation tools (open software) and supporting services so that libraries can easily collect and preserve their own digital contents at low cost. The LOCKSS member libraries have the preservation right for about 2700 e-journals published currently by approximately 400 academic document publishers, and including those in China, New Zealand, and Singapore in the Asia/Pacific region.

Journal@rchive is an archive site of science and technology information publications and distributions (J-STAGE) by the Japan Science and Technology Agency (JST), and is an e-archive project conducted since 2005. It has digitized and provided e-journals published by academic associations and societies in Japan from their first issue so that users can access the outcomes of research in Japan since the Meiji era; it also preserves many well-known essays which have influenced the world. E-journals for archives are selected in collaboration with associated institutions, for example, the Science Council of Japan, and include those in natural science, the humanities, and social science. The number of selected e-journals was 74 in 2005, 65 in 2006, 58 in 2007, 181 in 2008, and 266 in 2009. More selections from e-journals after 2010 will be made to expand the scope.

The Information Bridge provides people with about 210,000 full-text research reports and bibliography information without charge by the DOE (Department of Energy) itself or by supporting the research funders. The field of topics includes physics, chemistry, material science, biology, environmental science, energy technology, computer science and information science, and reproduction energy. Digitization of the research reports produced before and after 1991 is in progress.

The NTRS (Technical Reports Server) provides NASA technical documents to students, teachers, researchers, and the general public, and the number of documents continues to increase as the science and technology information produced by NASA or other means of financial support increases. It is possible to search for the information included in 3 different collections (NACA Collection, NASA Collection, NIX Collection) through a single interface.

As described above, many projects by national libraries, universities, and research institutions in other countries are in progress and established. There is an increasing need for exchange and preservation of digital resources through agreements between countries or institutions for the safe preservation of digital resources.

3. Preservation system for full-text of e-journals

This study aims to develop a system for the permanent preservation of large digital full-text e-journals with fast, easy access. Therefore, a system based on one platform is developed to handle making an agreement with overseas publishers followed by acquisition, management, preservation, and services.

3.1. Full-text collection/management process

In this study, the process, from acquiring full-text of e-journals to provision thereof, is described below.

First, the KESLI (Korean Electronic License Initiative) prepares for the task of acquiring the full-text of e-journals through an agreement with overseas publishers.

Second, the KESLI acquires both bibliography metadata and the full-text from overseas publishers when an agreement with them is made. That is, KESLI acquires academic metadata and large full-text files in PDF or XML format provided by the publishers.

Third, the data analyzer analyzes both the acquired full-text and the bibliography metadata in order to map them. This process creates connection information between the acquired bibliography metadata and the full-text.

Fourth, the data analyzer uses the full-text mapped with the bibliography metadata to extract and store the element items required for the academic contents management system. It transmits the full-text to the full-text server to be stored and managed by the full-text preservation system.

Fifth, the academic contents management system conducts error and redundancy checks for the registered bibliography information for adding or updating it. The system manages full-text route information and local repository information for connection with the full-text service. The system stores and manages the full-text acquired for preservation as a dark archiving full-text, and converts it to a light archiving full-text if required to provide it.

A dark archiving full-text is acquired from a publisher only for preservation, but not for provision to users. A light archiving full-text is acquired from a publisher through an agreement for both preservation and provision.

Sixth, after finishing error and redundancy checks, the full-text is loaded onto the service DB for full-text provision. Individual users or institution users can search and read the index and load full-texts through the search service.

3.2. Full-text preservation system

As shown in Figure 1, the full-text preservation system is in charge of the entire process of systematic acquisition-management-verification-distribution. The system continues to monitor errors which may occur during the entire process from acquisition by means of history management. Functions of the full-text preservation system are described below in detail.

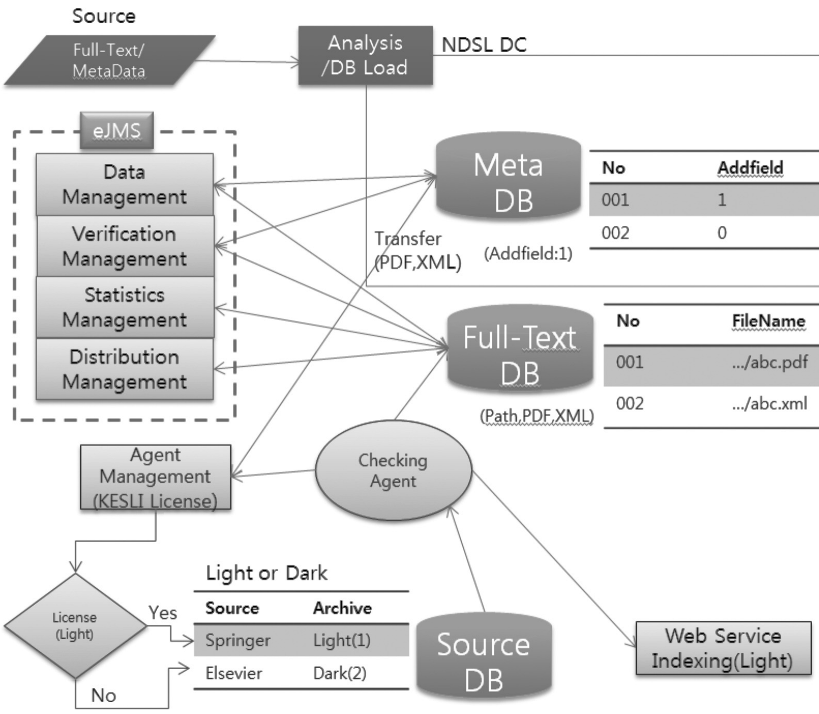

Fig 1. System Process

3.2.1. Collected data analyzer

The collected data analyzer analyzes the bibliography metadata of an essay acquired from a publisher. It automatically extracts bibliography metadata information required by the academic contents management system and the metadata information required for the search service to add it to the academic contents management system and to update it.

If the all of the metadata is acquired from an overseas publisher, the task described above is concurrently conducted because the existing established academic contents management system includes deleted or new data. By doing so, the academic contents management system can keep the latest version of bibliography metadata information.

Table 1. Example of Springer Metadata

Name	Springer Metadata	DC
Publisher	PublisherInfo/PublisherName	DC.Publisher
Location	PublisherInfo/PublisherLocation	DC.Publisher.place
Homepage	PublisherInfo/PublisherURL	DC.Publisher.homepage
...	...	...
Electronic ISSN	Journal/JournalInfo/JournalElectronicISSN	DC.Identifier.electronicscheme=ISSN
Journal Title	Journal/JournalInfo/JournalTitle	DC.Relation.isPartOf.title
...	...	...
Paper URL	article-meta/self_uri@content-type	DC.Identifier.article

The collected data analyzer extracts the route information, the information source, the date of acquisition, and the title of the full-text for the full-text preservation management system. The collected data analyzer has to analyze metadata in various formats, acquired from each publisher. The analyzer transmits the original file to the full-text server for providing the acquired full-text. It also implements automatic classification and management according to the full-text access route by using the full-text and metadata mapping information. Therefore, the collected data analyzer plays an important role in the full-text preservation management system.

3.2.2 Data management

Data management is for full-text data. As shown in Table 2, various information is managed including control numbers, information sources, and routes for full-text data management. The control numbers are information for connection with the existing academic contents management system. The information sources concern the relevant publisher, and the route information concerns the physical full-text location in the full-text server.

Table 2. Full-Text Data Management

Name	Example	Description
Control_Number	5121311	system control number
Source	Springer	publisher
Path	Springer/a.pdf Springer/a.xml	full-text path name
AddField	10001S100a	full-text presence or not (first bit)
Archive	D	D(Dark) or L(Light): archive type
Scheme	XML	XML or PDF: file type
Date	20120302	date
...	...	..

The data management task sets the value in the Addfield of Meta DB as 1 in order to indicate whether the acquired full-text is stored in the local server. It is essential to classify whether the acquired full-text is for dark or light archiving when providing the full-text service. The full-text agreement for light archiving states that it can be provided to users, but the full-text acquired for dark archiving is for preservation only and therefore cannot be included in the local hosting service. However, if required, the full-text acquired for dark archiving preservation can be processed to be converted automatically to light archiving and provision to users.

If the license agreement changes, the change occurs only to the corresponding information source through the information source DB without changing the millions of full-text DB fields. For example, if Elsevier is converted to Light through an agreement, it is converted from Dark to Light in the information source DB and then re-indexing is conducted for the service in the NDSL service.

3.2.3. Managing verifier

Verifier management is required to minimize errors that may occur during the full-text service. The task of verifier management is to check full-text file size, and accurate physical routes, etc. in order to automatically check full-text errors.

As shown in Figure 2, the automatic full-text verifier uses the route information of the PDF file automatically to check the full-text file size and opens the file to check errors. If errors are found, they are stored in the error DB. If an operator wants manual examination, sampling for examination can be conducted.

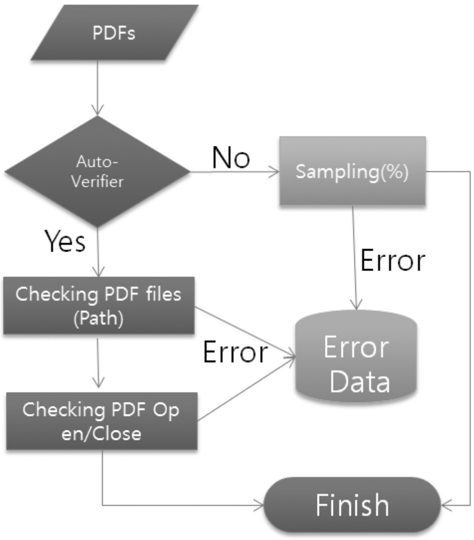

Fig. 2. Verifier Process

3.2.4. Statistics management

Statistics management is for yearly/monthly/daily statistics information for the number of entire full-texts. The error information which occurs during the entire process from acquisition to verification of the relevant full-text is also included.

3.2.5. Distribution management

Distribution management is implemented to download large full-text data to provide it to other research teams or other institutions. Therefore, with distribution management, it is possible to select an information source and years to automatically distribute the full-text files (PDF/XML) and to manage history information for the distribution.

3.3. Currently acquired full-texts

Figure 3 and 4 shows the currently held full-texts of e-journals, acquired in order to establish the full-text NAC base.

The number of acquired full-texts is 5.7 million for the entire 4,385 types of e-journals including 1,841 types from Springer, 61 types from IOP, 72 types from IOS, 2,286 types from Elsevier, 85 types from KARGER, and 40 types from AR (Annual Review). Only IOP and Springer are for light archiving, meaning the full-text is provided to KESLI member institutions for free.

The full-texts from the rest of the publishers are for dark archiving and managed for preservation to be converted to light archiving if required. That is, the system has a mechanism for automatically converting the full-texts for dark archiving to light archiving if necessary.

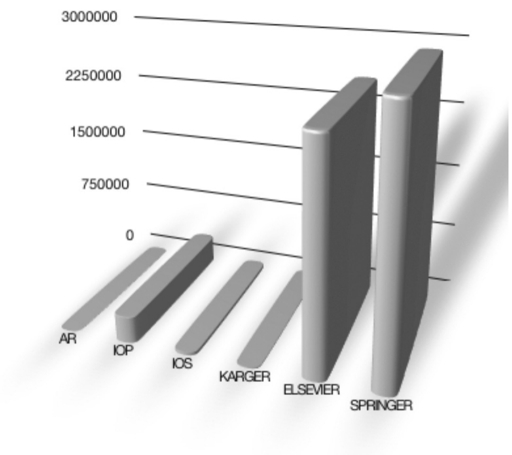

Fig. 3. The number of acquired full-texts

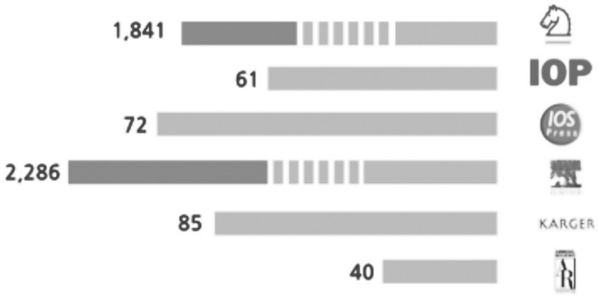

Fig. 4. The number of acquire journal

4. Results

This study aims to build a base system for e-journal full-text NAC (National Archiving Center) by acquiring overseas full-texts of quality e-journals, and establishing a system for providing free one-stop full-text service to KESLI member institution users and inexpensive paid full-text service (pay per view service) to individual users by acquiring the public service right for the full-texts of e-journals. To do this, an analysis was made of the results before and after introducing the full-texts of e-journals.

The results shown in Figure 3 reveal that only a small number of institutions could view the overseas full-texts of e-journals, but various KESLI member institutions can now view them. That is, while 47 institutions could view the full-text before the overseas full-text service, 159 institutions can now access them after the system's introduction.

While the number of downloaded full-texts was 197,634 before introducing the system, the number is 430,304 after introducing the system, increased by more than 2.17 times. The above analysis results cover only the Springer articles which are for light archiving. The service for IOP light archiving is currently in preparation.

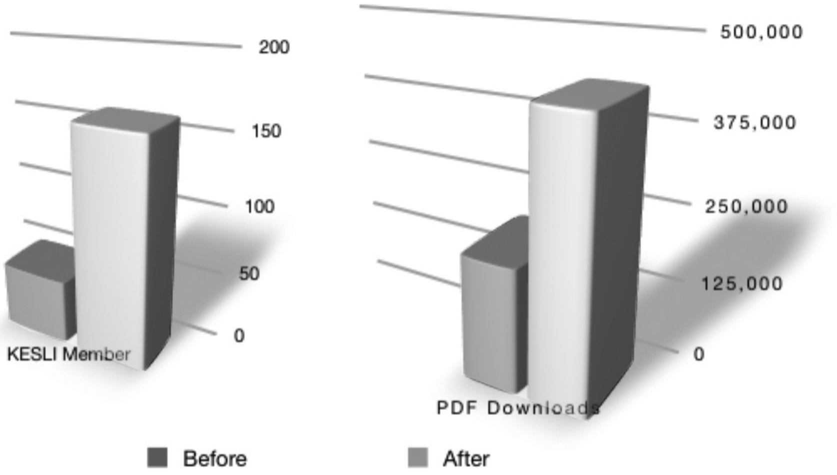

Fig. 5. Analysis Result

5. Discussion and Conclusion

While the Internet and e-publication technology continues to develop today, using overseas academic information of high quality is very beneficial. There is a high demand for the original full-texts of quality e-journals from Korean researchers. Digital archiving is required for safe preservation and provision of e-journals between different countries.

This study aimed to build a base system for e-journal full-text NAC (National Archiving Center) by acquiring the full-texts of quality overseas e-journals, and establishing a system for providing the public full-text service for e-journals. Doing this contributed to establishing a long-term preservation system for national e-information resources to provide free one-stop full-text service to KESLI member institution users and inexpensive paid full-text service to individual users. The analysis of the outcome of introduction in comparison with investment in the full-text service for e-journals reveals that 112 more institutions, that is, from 47 institutions to 159 institutions, have introduced the system as of 2012, and the number of downloaded full texts increased by at least 2.17 times.

As a future study, it will be necessary to continue to introduce full-text service of various quality e-journals for each publisher or journal and to study various service models. It is also necessary to study the metadata models for permanent preservation of full-texts.

References

Choi, H-N., & Lee, E-B. (2005). A Study on the strategies for building a digital archive of electronic journals. *Journal of the Korean Society for Library and Information Science*, *39*(2), 161-183.

Fernandez-Cano, A., Torralbo, M., & Vallejoa, M. (2004). Reconsidering price's model of scientific growth : An Overview. *Scientometrics*, *61*(3), 301-321.

Journal@rchive. Retrieved from http://www.journalarchive.jst.go.jp

Jung, Y-I., Choi, H-N., & Choi, S-H. (2009). Study on strategies for improving application of archived data: focused on international and domestic journal articles. *Journal of the Korean Society for information Management*, *27*(1), 185-206.

Jung, Y-M. (2008). Development of an economic valuation methodology and model for the DDS of foreign journals. *Journal of the Korean Society for information Management*, *25*(4), 243-267.

Kim, K. H., Morningstar, M. E., & Erickson, A. G. (2011). Strategies for successfully completing online professional development. *International Journal of Knowledge Content Development & Technology*, *1*(2), 43-51.

Kwak, S-J. (2010). *Best Practices for Archiving Digital Content by Life Cycle*. Daejoen : KISTI.

Seol, M-W. (2005). *A study on developing national digital archiving strategies*. Daejeon: KISTI.

The Information Bridge. Retrieved from http://www.osti.gov/bridge

The LOCKSS. Retrieved from http://www.lockss.org/lockss/Home

The National Library of the Netherlands (KB). Retrieved from http://www.kb.nl/hrd/dd/index-en.html

The NTRS (NASA Technical Reports Server). Retrieved from http://ntrs.nasa.gov

The Portico. Retrieved from http://www.portico.org/digital-preservation