Korean Web-Archiving Current Status and Prospects

Suh Kyungran

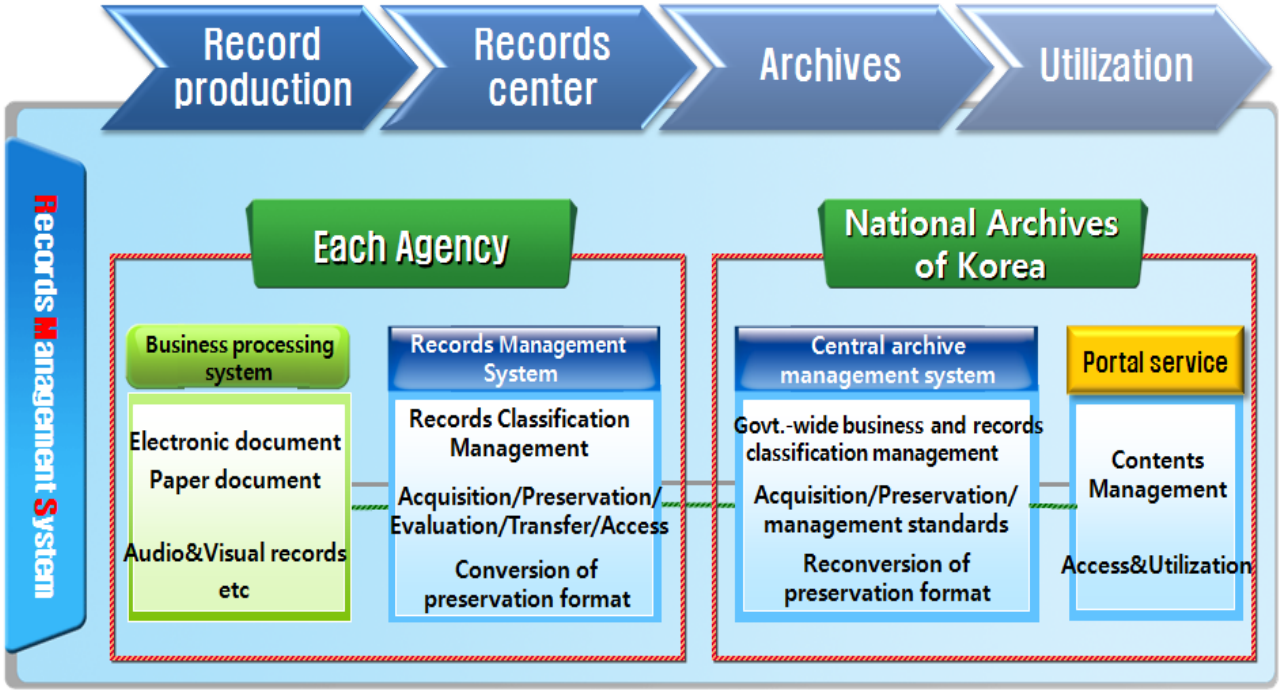
National Archives of Korea

1. Form of Electronic Record

In line with the Public Records Management Act of the Republic of Korea, electronic records are classified into three-fold: electronic documents, web archives (web sites) and dataset for administrative information.

In case of electronic documents, an electronic documents creation system was constructed in the early 2000s. Based on this system, records created from originating offices are transferred to a records center after a certain period of time. Then, permanent records having enduring value are transferred to a permanent archives, the National Archives of Korea (hereinafter referred to the NAK), for the long-term preservation after being preserved at a records center for a certain period of time. As of now, electronic documents created in the early 2000s are being transferred to the NAK.

<Korean Electronic Records Management System>



In case of dataset for administrative information, even though a tentative system is established under the pilot project, it is not easy to transfer them to the NAK after standardizing various forms of dataset. So a research into the transfer of dataset for administrative information is ongoing.

2. Web Archiving

Web Archiving is the process of collecting portions of the World Wide Web, preserving the collections in an archival format, and then serving the archives for access and use. Many records are

created in the digital format when compared to the past, and records which are made available in the homepage, blog, SNS etc. are increased exponentially. Despite of the quantitative increase of records on the web, the awareness of collection on them is not raised fully.

Records on the web are changeable, ephemeral, and volatile. That is, they are easily updated and do not survive for the next generation because of short-term life cycle. It is very unfortunate that key web sites having valuable information are changing and even disappearing, thereby no longer accessing information.

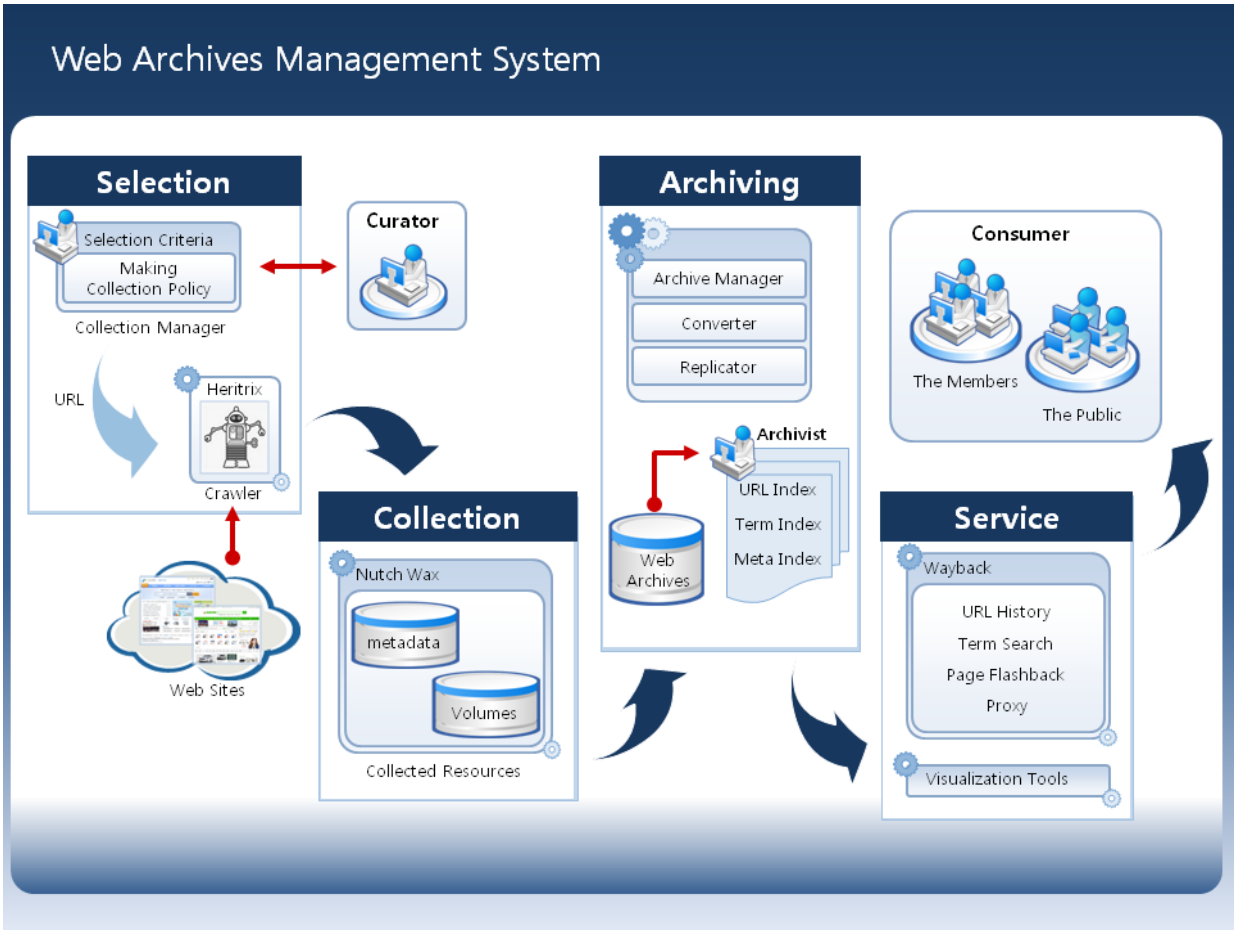
In particular, web sites of public institutions play a role of communicating between the government and the public. Korean government is changed every 5 years as a result of a presidential election, and government departments are rearranged or merged together with this change. If web sites of abolished departments are not collected and preserved in due course, information in these web sites is not available in the future.

3. Korean Web Archiving Current Status

In Korea, the National Library of Korea has collected key web sites such as part of central government agencies, universities, electronic journals etc under the name of a OASIS (Online Archiving & Searching Internet Sources) project since 2004. In case of the NAK, we carried out a research project in 2008 and constructed a web archives management system in 2011. A research project in 2008 analyzed various overseas cases such as Internet Archives of the United States, PANDORA(Preserving and Accessing Networked Documentary Resources of Australia) of Australia, National Assembly Library of Japan etc, and as a result we established underlying technologies such as a web records collector, a reproduction or display machine etc. and standard formats. The NAK adopts tools, software, and WARC formats served by IIPC (International Internet preservation Consortium) which develops standards and tools on web archiving and applies them into a NAK web archives management system.

The NAK collects web archives using a web crawler called Heritrix and stores them at storage in the format of WARC. Stored web archives are made available with a reproduction or display machine dubbed Wayback machine.

Since the establishment of this system, the NAK has collected and preserved web sites of around 50 public institutions.



< NAK Web Archives Management System>

4. Main Issues

Like other digital fields, the environment of web is constantly changing and disappearing on a regular basis. I would like to go through these controversial issues in terms of the web archives management.

4.1 Comprehensive Collection vs. Selective Collection

Comprehensive collection is defined as collecting all web archives under a specific URL with no selective procedures. It features in fast process, and it is possible to preserve complete context of web sites. However the cost of storage and preservation for web archives is relatively expensive.

Selective collection is to collect web archives in line with a prepared guideline of an institution and makes a collector grasp in detail quantity and quality of archives to be collected. It costs a lot and is time consuming. Also a guideline to capture important web archives is subject and limited. Then records, which are excluded from collection range or policy, are in danger of disappearing in the future. So the next generation cannot use them.

As of now Australia, United Kingdom, and Japan adopt a selective collection policy, and United States, Sweden, and Finland adopt a comprehensive collection policy. The NAK chooses a comprehensive collection policy, but issues over which policy is good for preservation still remain controversial.

4.2 Direct Collection vs. Remote Harvesting

Direct collection is to request a target organization to copy and transfer web archives physically into an archives. For this, all information such as program source, DB data, environment configuration etc which are needed for restoration should be collected or captured. It might cost a lot because software licenses are required for restoration.

Remote harvesting is to use a web crawler or web robot software in order to capture web archives from a remote server. It costs relatively less, for an automated tool is used. However accuracy of web archives to be captured in terms of quality is relatively low.

The NAK uses basically a remote harvesting policy for capturing web archives and partly uses a direct collection policy in case of a remote collection's being impossible. However in case of presidential archives, a direct collection policy is primarily adopted. As of now, the NAK has collected the 14th to 18th presidents' related web sites and provided service to the public, but we face a problem of increase of service cost. So we are considering whether to use a web crawler software from now on. Like this, a direct collection and remote harvesting can be used at the same time considering conditions of an institution and network environment. So pros and cons are weighed up before selecting a adjustable policy.

4.3 Collection Cycle

As mentioned earlier, web archives are often changeable and easy to be deleted. So a collection cycle to capture web archives should be decided after considering a type of an institution, size of a web site, rate of change etc in combination.

In case of Korea, considering the government change repeated every 5 years, web sites of central government agencies are captured before and after reshuffle of a government, and web sites of affiliated institutions are captured once every 5 years. And collection time varies from one day to a few months according to size of a web site. There is also a case to shorten the collection time, once a week or once a month in case of international events on which the public has interest such as Asian game and an international exposition in order to check the change of a web site. Therefore, various attributes of a web site should be considered for the decision of the collection cycle.

4.4 Records Impossible to Collect

Main features of digital era are sharing and openness. Korean government as well as other countries' governments also does every efforts to improve transparency and trust of their own governments. Korean government proceeds a policy dubbed "Government 3.0" which is a new paradigm for government operation to promote active sharing of public information and removal of barriers existing among government ministries for better collaboration. However, despite of these efforts, Korean web openness index is not high as yet according to a result of a research to evaluate web openness index. One of the evaluation elements is "whether access to a web site by a web crawler is blocked or not". If an institution does not allow a web crawler to capture a web site in keeping with its own policy, it is impossible to perform web archiving. Therefore, the NAK recommends a target institution to unblock "denying access by a web crawler" and proceeds web archiving with no problem.

Aside from this case, there are web sites which are technically not captured or collected even with an updated version of a web crawler. For instance, if log-in is in need for collection, a web crawler cannot access to this web site because of impossibility of log-in session. Also these days, Ajax(Asynchronous JavaScript + XML) is often used in order to realize a dynamic web application,

but it is hard to capture this web site because a technology such as Ajax is not a standard of W3C. That is, a different web browser design and implementation method makes collection and display of a web site difficult. Also if a web site uses Javascript which creates URL dynamically, or Flash has another link, it is impossible to collect web archives, for a web crawler cannot extract URLs linked to Javascript or Flash. Other than this case, there are many cases that some problems occur at the stage of reproduction or display after completion of collection with no problem. Even though target institutions are recommended to use technologies corresponding to web standards, some cases explained above are caused by the technological limitation of a web crawler, Heritrix, so more technological development should be met in the future.

5. Prospects

Korea is well-known for E-government as she has ranked 1st in the UN E-Government evaluation for three consecutive terms. However, electronic records management methods to cope with various formats do not catch up with the quantitative and qualitative information expansion. In particular, to nurture professionals or experts and to further related research or study should be followed on in order to improve web archiving technologies which are neglected compared to other ones.

The NAK will supplement a web archives management policy through examining thoroughly web archives management related issues which are shown above. Moreover additional system extensions are needed in order to expand collection range and a collection cycle. That is because as of now web archiving is limited to some central government agencies. Also the NAK should cope with budget which is needed to introduce a web archives collector and secure storage in order to store and preserve web archives to be captured or collected.

The NAK should draw up measure not only to preserve web archives for the next generation but also to use them as materials for research and education. Also the NAK draws attention to collecting valuable private web archives having enduring value and should do every effort to develop a new service model.