Development of Metadata Elements for Intensive Web Archiving

선택적 웹 아카이빙을 위한 메타데이터 요소 개발

Heejung Kim*, Hyewon Lee**

ABSTRACT

As digital preservation becomes increasingly important, interest in web archiving has correspondingly increased. The processes of web archiving depend on the types of acquisition methods employed, the organization and storage of data, their completeness, and their scope. This study develops metadata for intensive web archiving. Several web archiving projects are reviewed and analyzed. As a result, administrative metadata has been suggested in addition to the basic elements from the Dublin Core.

초 록

디지털 보존의 중요성에 대한 인식이 확산되면서 웹 아카이빙에 대한 관심도 높아지고 있다. 웹 아카이빙은 수집형태나 운영형태, 또는 아카이빙 대상 컬렉션 범위와 포괄정도에 따라서 설계지침이 달라지게 된다. 본 연구에서는 이 중 선택적 아카이빙을 중심으로 아카이빙을 수행하고자 할 때에 고려해야 할 메타에이터 요소들을 분석하였다. 선택적 아카이빙을 수행한 선행 프로젝트에서 제안한 메타데이터를 기반으로 필요한 요소들을 분석하였으며, 분석결과 더블린코어의 기본적인 요소들과 함께 관리적인 요소에 해당하는 메타데이터 내용들의 확충이 필요함을 확인할 수 있었다.

Keywords : web archiving, intensive archiving, extensive archiving, metadata, digital preservation
웹 아카이빙, 선택적 아카이빙, 포괄적 아카이빙, 메타데이터, 디지털 보존

* Lecturer, Dept. of Library & Information Science, Yonsei University (heejung@yonsei.ac.kr)
** Lecturer, Dept. of Library & Information Science, Seoul Women's University (hwlee@swu.ac.kr)

■ Received : 25 May 2007
■ Accepted : 23 June 2007

1. Introduction

The World Wide Web now serves as the most global monopolic communications channel. According to the dynamic characteristics of the web, 7 million web pages are created daily, 40 percent of which will exist for only 44 days on average, and cease to be connected within one year(Lyman 2002).

Because of this fragility, web archiving projects aimed at long-term preservation have been planned and adopted among develped countries since the mid 1990' s.

Proper management of Web-based records during their life-cycle is vital, and in effect, efforts must be expended to ensure Web sites are 'future-proof' (DCC 2006).

With the growing dependence on external digital asstets(ex. Web contents), libraries and archives are undertaking some measures to protect their continued use of these resources(Kenny et al 2002).

In addition to the importance and necessity of the way how to build up and use the web contents, importance of long term preservation is rapidly increasing nowadays.

In Korea, no representative web archiving project has yet been completed, and a few researches have been performed on web archiving. Only a few reviews and introduction on web archiving in the broad perspective have been undertaken so far.

Web archiving projects have been carried out in various realm in developed countries. Projects such as focused on special subjects, URLs, and organization.

On the other hand, web archiving can be classified into extensive archiving and selective archiving according to the web completeness. Extensive archiving means collects web pages horizontally, ignoring the site level. While Intensive archiving collects web pages vertically, contrary to extensive archiving(Masanés 2006).

Suh(2004) summarized merits and demerits of extensive and intensive archiving: In the extensive archiving case, management cost is lower than intensive archiving. However quality of archived contents are poor than intensive archiving, and dynamic or hidden web sites might be missed because of the technical limitation of web crawler.

On the contrary, management cost of intensive archiving would be much higer than extensive archiving because of the labor intensiveness. Assurance of the good quality of archived web is the merit of intensive archiving.

This study focused on intensive archiving because it assures the good quality of web contents and hidden webs. Metadata of intensive archiving has suggested also. Extensive web archiving, which use web crawlers or robots, usually use basic descriptive metadata of Dublin Core(DC) because of its simpleness. On the other hand, as the intesive web archiving, more detailed metadata elements are needed because of the selectiveness focused on the quality.

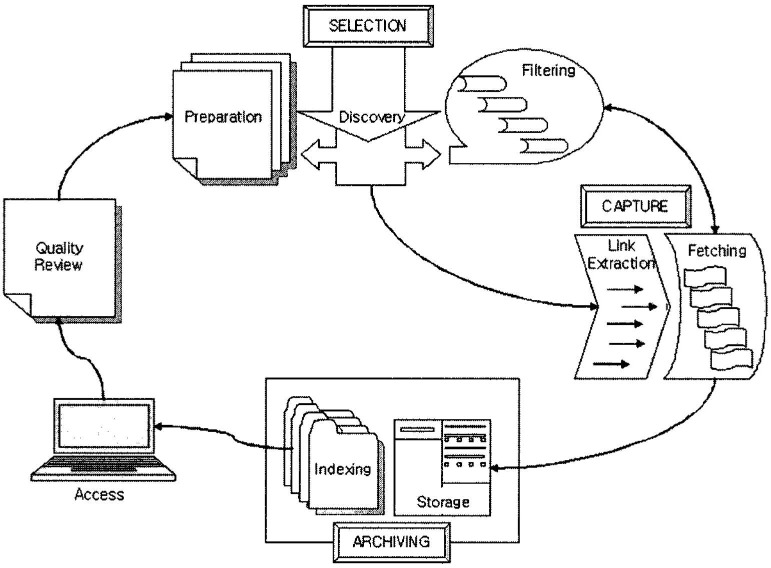
In this study, the processes and types of web archiving have been reviewed first, and then representative web archiving projects have been reviewed. After that, intensive web archiving projects have been selected on the basis of special topics or sites archiving by small size organizations. As a following process, metadata has been analyzed. The reason of selecting subject based intensive web archiving is that topic web archiving is becoming more and more popular, often driven by direct research needs(Messenes 2006). Researches focused on metadata of web archiving project by small size group has not been actively accomplished compared with large size group. Because of that reason, this study has focused on metadata of web archiving project by small size group. According to the common metadat elements among intensive web archiving projects, metadata elements for the intensive web archiving have been suggested.

2. Processes and Types of Web Archiving

2.1 Web Archiving Processes

Masanés (2006) described five steps of web archiving process: selection, capture, archiving, access, and quality review. This process is shown in ⟨Figure 1⟩.

In the selection process, basic policies and designs are built for web archiving. After selection, the capture process is undertaken based on the results of selection. Web sites are captured through the capture process, and after, that the archiving process is undertaken. Access is for future users who wish to access the archiving website; quality review is the follow-up step that occurs after use. And then the



⟨Figure 1⟩ Web Archiving Cycle

cycle repeats from selection.

The selection process is a key issue and the first step for web archiving, comprises three phases: preparation, discovery, and filtering. Further details on selection are described below(Masanés 2006):

① Preparation: The main objective of the preparation phase is to define the collection target, the capture policy and the tools for implementation. Input here is required from domain experts who can define the target information space, capture policies, and tools. There are four categories of tools which can be used for web archiving: hubs, search engines, crawlers, and external sources.

② Discovery: The main goal of the discovery phase is to determine the list of entry points that will be used for the capture as well as the frequency and scope of this capture. The usual frequencies are "once only", "weekly", "monthly", and "every x months". The scope of capture is important to define. The page and the site level can be used.

③ Filtering: The main goal of the filtering phase is to reduce the space opened by the discovery phase to the limits defined by the selection policy. Filtering can be done either manually or automatically. Manual filtering is necessary when criteria used for the selection cannot be directly interpreted by automatic tools or robots. Several evaluation axes can be used for filtering. They are quality, subject, genre, and publisher(Masanes 2006).

2.2 Web Archiving Types

Several archiving types have been suggested by researchers. Even though the words naming them are different, main frame of classification is almost similar.

There are Five web archiving types: extensive(unselective), intensive(selective), topic-centric(thematic), domain-centric and combined types. Details of five types are as follows(Massene 2006; Brown 2006; PADI; Suh 2004):

① Extensive Archiving: Also called as unselective archiving(Brown 2006), whole domain archiving(PADI), and comprehensive archiving(PADI). Extensive archiving collects web pages horizontally, ignoring the site level(Massenes 2006). Exetensive archiving collects surface web rather than deep web, and use automatic robots such as a web crawler or spider. Representative project of extensive web archivng type is Kulturarw3 (Cultural Heritage Cubed) and The US-based Internet Archive.

② Intensive Archiving: Also called as selective archiving(Brown 2006; PADI). This type of archiving collects web pages vertically, and collects deep web sites(hidden web sites) where access to the full content is not possible with crawlers. Threrefore manual archiving by the experts is inevitable(Massene 2006). Selection may be based on the significance or quality of resources, their theme or topic, or by targeting a related set of Web sites. Rather than attempting to archive all content(PADI).

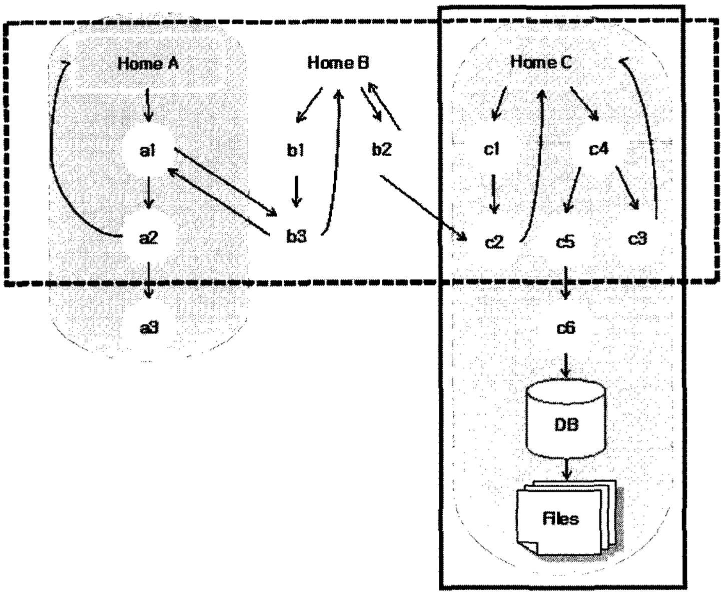
③ Topic-centric Archiving: Also called as thematic archiving(PADI). This type of archiving involves the collection and preservation of web content relating to a particular theme or event. Topic-centric archiving requires manual input by experts(Mannés 2006).

④ Domain-Centric Archiving: Domain-centric archiving collects web pages according to domain name. One can distinguish functional or generic types(".com", ".edu") and geographical types(".kr", ".ca"). Domain-centric archiving can collect web pages by a web crawler or automatic robots (Mannes 2006).

⑤ Combined approaches

A growing number of Web archiving programs reach the conclusion that there is no model entirely satisfactory. Therefore by employing a combination of extensive (comprehensive), intensive(selective) and topic-centric(thematic) types can be applied(PADI).

This study focused on intensive web archiving type among five web archiving types above. Extensive web archiving and intesive web archiving can be shown as following ⟨Figure 2⟩. Horizontal box shows extensive archiving, and vertical box shows intensive archiving.



⟨Figure 2⟩ Extensive and Intensive Web Archiving

2.3 Web Archiving Projects (Extensive/Intensive Types)

Many web archiving projects have been planned and developed since mid 1990's.

Among them, representative projects of extensive and intesive web archiving have been selected. They are selected on the basis of the frequencies of occurence in the research materials focused on web archiving.

These projects share not only a topic orientation but also the use of a network of informants(Lecher 2006).

Web archiving projects are described in 〈Table 1〉, according to lead organizations, methods, and domain/topic. Methods are split into extensive(E) or intensive(I) approaches. Domain/Topic refers to domain-centric archiving or topic-centric archiving.

3. Metadata Suggested in Web Archiving Projects

3.1 Web Archiving Projects

As mentioned in introduction, this study has focused on web archiving projects, which carried out by small size groups and focused on special sites or topics.

Among eight projects listed in 〈Table 1〉above, five projects focused on intesive web archiving have been selected and metadata of the projects have been analyzed.

PANDORA, which proceeded by large group, has been omitted.

Five selected projects and their basic contents are as follows;

3.1.1 Harvard University Library (LDI: Library Digital Initiative)

① overview: Harvard's LDI is a two-year pilot project addressing the acquisition of web sites for long-term

〈Table 1〉 Archiving Projects Described by the Method

Country	Initiative/ Lead Organization	Method	Domain/Topic
Sweden	Kulturarw3(KB)	E	Domain
USA	Internet Archive	E	Domain&Topic
Australia	PANDORA(NLA)	I	Domain
USA	Harvard Library	I	Topic
Australia	NLA	I	Domain
USA	SUNYIT/LC	I	Topic
USA	SIA	I	Domain
USA	Goddard Library	I	Topic

archiving where university project partners will provide content and curatorial perspective.

② technical focus: The key technical focus will be on the development of services for ingest, storage, preser vation and basic delivery of web sites. The pilot will focus on harvesting the web for content in a focused area to enable small-scale, curated collections.

③ services: During the course of this project, LDI staff and curatorial staff from the partner project units will address collection management issues specific to archiving web sites, such as establishing policies concerning intellectual property rights issues, exploring best practices for describing archived web sites, and determining searching and browsing requirements for delivery.

3.1.2 National Library of Australia (Web Archiving)

① overview: National Library of Australia has collected web sites it has hosted, together with selected metadata. The aim of assigning such metadata is to enhance access to resources on the site through the library's own site search facility and through services such as the Commonwealth Government Search Engine.

② collection: The library will assign metadata to the following categories of information on its public web site:

- collections of descriptive or marketing information about the library, its collections, services and activities;
- policy and strategy documents;
- media releases and staff papers;
- exposure drafts (e.g. National Framework for Australian Subject Gateways);
- formal library publications (e.g. Gateways, Annual Report);
- bibliographies and subject guides;
- online exhibitions;
- entry points for significant initiatives (e.g., the Digital Services Project) or working group activities (e.g. Australian Library Collections Taskforce);
- service entry points (e.g. The Catalogue, Kinetica Web, PADI, Australian Libraries Gateway)

3.1.3 SUNY institute of Technology (Election Web Collection)

① overview: WebArchivist.org, a collaborative project at the University of Washington and the SUNY Institute of Technology, initiated the Election 2002 Web Archive project with the Library of Congress. The library contracted with the SUNY Institute of Technology to develop and implement tools and processes to support the identification, acquisition, collection and cataloging of web sites for inclusion in the Election 2002 Web Archive.

② scope: This project is an example of a thematic web archive. A thematic web collection is an archive of web objects identified and captured using a set of

URLs believed to be relevant to a specific theme or topic.

The scope of the project is as follows; to collect and process metadata associated with the sites in the collection; to identify site owners for notification purposes; to create an interface allowing users of the archive to identify sites of interest; and to provide a report on this project to facilitate future web archiving initiatives.

③ services: A portion of the Election 2002 Web Archive was made available to the public in March 2003. Nearly 1,200 sites produced by House, Senate and gubernatorial candidates were indexed, catalogued, and presented using an interface developed by WebArchivist.org. Research-based interface to the collection was created also by WebArchivist.org on its PoliticalWeb.Info site. Using this interface, visitors to the Election 2002 Web Archive can currently search campaign sites by five fields in addition to those provided by the MINERVA site.

3.1.4 Smithsonian Institution Archives(SIA)

① overview: SIA provides a set of recommended guidelines for the archival preservation of Web sites and HTML pages. The purpose of the guidelines is to help the SIA implement appropriate strategies and procedures that ensure the capture, management, and preservation of Smithsonian Institution (hereafter SI) web sites and HTML pages for as long into the future as may be required.

② collection: The resources of SIA focused in some detail on a smaller number of SI Web sites. Accordingly, the National Air and Space Museum (NASM), the National Museum of American History (NMAH), and the Freer Gallery of Art and Sackler Gallery were selected. These three museums are likely to contain most of the types of HTML pages and associated technology issues found in Web sites across the SI.

③ services: Senior SI officials are encouraging various offices, museums, and research programs to expand their use of the Internet to inform the public of various activities and programs, to offer more “virtual exhibits”, and to facilitate greater access to their wide ranging resources.

3.1.5 Goddard Library Web Archiving

① overview: In 2001, the NASA Goddard Space Flight Center (GSFC) Library began investigating ways to capture and provide access to internal project-related information of long-term scientific and technical interest. These activities were coincident with NASA GSFC’s enterprise-wide emphasis on knowledge management.

② collection: The GSFC is interested in a wide range of content types including project documentation such as progress reports, budgets, engineering drawings and design reviews; web sites, videos, images and traditional published materials such as journals articles, manuscripts and technical reports.

The GSFC is concerned about a small selective domain for which proper access restrictions and distribution

〈Table 2〉 Metadata of the Web Archiving Projects

DC	Harvard Library	NLA	SUNYIT/LC	SIA	Goddard Library
Creator	Author or Creator	Author or Creator		Author or Creator	Creator/Creator Employee
Publisher	Publisher	Publisher		Publisher	Creator Organizaton/Subject Organization/Publisher Organization/Publisher Code
Contributor	Other Contributor	Other Contributor			
Rights	Rights management	Rights management			Rights
Title	Title	Title(every page)	Title		Title
Subject	Subject or Keyword	Subject or Keyword	Subject or Keyword		Subject – Mission,Project/Competency/Instrument/Business Purpose/ Industries/Uncontrolled Subject (Keywords)
Description	Description	Description (every page)	Description		Description
Source	Source	Source			Source
Language	Language	Language	Language	Language	Language
Relation	Relation	Relation			
Coverage	Coverage	Coverage			Spatial Coverage
Date	Date(of creation/ modification)	Date(of creation /modification)		Date(mod. date separate)	Date
Type	Resource Type/ Genre	Resource Type/Genre	Resource Type/Genre		Content Type
Format	Format	Format		Format	Format
Identifier	Resource Identifier(URL)	Resource Identifier(URL)	Resource Identifier(URL)	Resource Identifier(URL)	Persistent Identifier/URL
	Function Descriptor	Function Descriptor			
	Availability	Availability			
	Audience	Audience			Audience
	Mandate	Mandate			
	Harvest File	Harvest File			
	Author or Creator Email				
	Expiration Date				
	Alternative Title		Alternative Title		
	Date Captured		Date Captured		
	Access Condition		Access Condition		
	Collection Title		Collection Title		
	Date Metadata Modified			Date Metadata Modified	
	Date Validated			Date Validated	

limitations are as important as the original content itself.

③ services: The goal was to capture content of scientific and technical significance rather than information from human resources or advertisements from the employee store. Therefore, a mechanism that captured the whole domain was inappropriate. The GSFC system needed to be more selective.

The features of projects selected in this study have been reviewed. To summarize target of web archiving, Harvard Library has collected contents of university project partner. NLA and SIA collect documents and services provided by National Library of Australia and SIA. SUNYIT/LC collect contents related to election using Minerva. Goddard Library collects contents of NASA GSFC intranet.

3.2 Metadata Analysis of Web Archiving Projects

Metadata extracted from the five projects are suggested in <Table 2>. To compare and analyze each other, related elements are mapped based on dublin core metadata.

Five web archiving projects include basic descriptive metadata, which are based on DC. Harvard Library, NLA and SUNYIT/LC have more administrative elements than others. Especially, NLA has most adminstrative elements, such as Availability, Audience, Mandate, and Harvest File.

SUNYIT/LC added Alternative Title, Date Captured, Access Condition, and Collection Title. Access Condition provide information of resources. Because SUNYIT/LC has special subject such as election, it doesn't have descriptive metadata of author or publisher.

Policies, procedures, and systems which influence contents of SIA web site can be changed as time goes by and origital contents might be lost during that process. Because of those problems, web archiving projects have been carried out, and metadata of SIA is related to date.

Goddard Library has carried out web archiving project to capture and provide access to internal project-related information of long-term scientific and technical interest. Therefore, details of institutions and creator information are described. And subject has to be listed clearly for users to retrieve friendly.

4. Suggested Metadata for Intensive Web Archiving

4.1 Considerations for Defining Metadata

As digital preservation, accessibility to information is core interest, and issues of authenticity is important also. During the flow of information, reliability and continuous accessibility of information should be acquired.

Models such as OAIS divide information flows into process unit, and define descriptive and administrative elements of

〈Table 3〉 Considerations for Defining Metadata

CATEGORY 1	CATEGORY 2	Definition
Collection Scope	Define Collection	Content : What subject area will be covered?
		Audience : Who will access and use the archived material?
		Duration : How long will you provide storage and access to the archived material?
	Selecting what to collect	Formulate a collection schedule based on the nature of the content • For each site determine if it should be captured once, a limited number of times, or on an ongoing basis. • For sites that will be collected more than once.
Access and Intellectual Property Rights	Access Control	Accessible Contents • Will all material be publicly available or are there any restrictions? • Consider the web domains of sites selected for your project.
	Permission	Copyright Issues
Metadata	Provide persistence access	Identifier : At what level(s) is the Identifier assigned for discovery, e.g. at the collection level, web site level, individual web page level, etc.?
	Administrative metadata requirements	Administrative metadata : What metadata elements will be used, e.g. administration, technical metadata, process history/provenance metadata, rights management?
	Structural and relationship metadata requirements	Structural and relationship metadata • Will web site directory structure be explicitly documented? • How will different versions of a web site be documented, e.g. harvest date, modification date?
	Descriptive metadata requirements	Descriptive metadata : How will users discover the collection?
	Delivery requirements	Define interface requirements for the presentation of the collection
Acquisition	Collecting Methods	Considerations • Harvesting by an automated software program • Manual downloading • File transfer from another party • Vendor
	Tools	Software of web archiving tools and Integrated system
Implementation and Maintenance		Maintenance and Supplement of the Projects • Create and Maintain the list of objects • Create and Maintain a collection schedule • Undertake QA/QC on all acquisitions • Process and prepare the content

information.

In this study, metadata for the smaller realm than digital preservation, web archiving is suggested. Metadata suggested in this study is descriptive and administrative metadata which develop reliabilty of information through assurance of continuous access.

From this viewpoint, considerations for defining metadata for web archiving are presented in 〈Table 3〉. They are broadly categorized by collection scope, access and intellectual property rights, metadata and discovery, acquisition, and implementation and maintenance.

〈Table 3〉 is based on web archiving collection development issues from Harvard University. Elements suggested in 〈Table 3〉 reflect general metadata distinction.

Elements suggested by Gilliland-Swetland(2000), different types of metadata has been included in 〈Table 3〉 also. Gilliland-Swetland differenciate 6 types of metadata and details are as follows(Lazinger 2001).

- Administrative : metadata used in managing and administering information resource
- Descriptive : metadata used to describe or identify information resource
- Preservation : metadata related to the preservation management of information resources
- Technical : metadata related to how a system functions or metadata behaves
- Use : metadata related 데 the level and type of use of information resources

Elements which are administrative and necessary to maintain systems are suggested in category 1. Collection Scopes, Access and Intellectual Property Rights, Acquisition, Implementation and Maintenance are for those who practically manage systems.

Based on considerations in 〈Table 3〉, optimal web archiving metadata elements are selected.

After selection of metadata, elements should be descibed specifically. Web crawler which collect web pages can extract elements automatically or the field can be filled manually.

Metadata creation includes three main components: the medata scheme, automatic metadata extraction, and human review and enhancement of the metadata(Senserini et al. 2004).

Most metadata elements of web archiving projects are developed based on DC element set. Such metadata can be used as access tools and user interface components of the web contents related to the projects. The process of defining such metadata elements is that of constructing the metadata scheme.

To extract metadata automatically from a web site, a web crawler can be used. A web crawler collects basic HTML tags automatically. Those tags contain such elements as title, description, keywords, related date, and content length.

The type of web page can be extracted also according to the file name or multimedia type(MIME type).

The occasion of collecting web pages of

certain institutions or domains can be extracted automatically by analyzing the directory of files, related offices òr web master.

The last step in the creation of metadata records is writing metadata, which is extracted automatically according to the metadata scheme.

The elements in the metadata scheme can be used to build the database to match the automatically extracted tags. When automatic extraction has not occurred, it can be made by inference. An example of a metadata based of inference is the Goddard Library web archiving project.

The three steps above constitute the process by which metadata is made automatically. Besides the basic DC elements, more specific metadata can be assigned. When administrative metadata is included, metadata can be manually assigned.

4.2 Intensive Web Archiving Metadata

Intensive web archiving metadata is presented in 〈Table 4〉. Besides the basic elements from the Dublic Core, other common elements from the projects have been adopted.

Adopted metadata can be categorized into DC-based elements and administrative elements.

According to the projects, DL project from Harvard university is a small scale plan which collects and preserves websites on research projects processing within university. The 2002 Election Web Collection from SUNY collects web sites on election. The SIA project aims to preserve contents through the SIA site. The Goddard project focused on research projects carried out by NASA in terms of knowledge management.

As shown from the projects above, it seems that intensive web archiving is increased among universities and research institutions. They focus on special subjects or are collected by certain institutions. The scale of the projects ranges from small to medium-size. Taking these things into consideration, metadata for intensive web archiving is suggested. For this suggestion, web archiving collection development issues and common elements from intensive web archiving projects are adopted.

Among the five web-archiving projects reviewed in section three, project metadata which is related to the special subjects and institutions are analyzed and suggested.

These elements can be used for the metadata of intensive web archiving.

From this analysis, metadata elements can be broadly categorized into administrative elements and descriptive elements. This study has adopted same criteria. Essentially, descriptive elements satisfied DC elements, and administrative elements were added through the analysis of those projects that were carried out in advance.

Intensive web archiving metadata suggested in this study is shown in 〈Table 4〉.

Intensive web archiving metadata is suggested in 〈Table 4〉. Besides the basic elements from the Dublic Core, other common elements from the projects have

〈Table 4〉 Metadata of Intensive Web Archiving

ELEMENT	DESCRIPTION
Creator	An entity primarily responsible for making the resource(DC' s definition).
Publisher	An entity responsible for making the resource available(DC' s definition).
Contributor	An entity responsible for making contributions to the resource(DC' s definition).
Rights	Information about rights held in and over the resource(DC' s definition).
Title	A name given to the resource(DC' s definition).
Subject	The topic of the resource(DC' s definition).
Description	An account of the resource(DC' s definition).
Source	The resource from which the described resource is derived(DC' s definition).
Language	A language of the resource(DC' s definition).
Relation	A related resource(DC' s definition).
Coverage	The special or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant(DC ' s definition).
Date	A point or period of time associated with an event in the lifecycle of the resource(DC' s definition). The "Date.Created" metatag should include the date on which the page was coded. The "Date.Modified" metatag should include the date on which changes were made to the page
Type	The nature or genre of the resource(DC' s definition).
Format	The file format, physical medium, or dimensions of the resource(DC' s definition).
Identifier	An unambiguous reference to the resource within a given context(DC' s definition).
Function Descriptor	Subjects of special area are suggested specifically(Thesaurus can be used)
Availability	How the resource can be obtained or contact information.
Audience	The group expected to use the resource.
Mandate	Use with the name of the specific Act, Regulation or Case or the scheme URI when appropriate.
Harvest File	Notice that file has collected from special institution and saved. Nodtice of description of metadata provided by the institution.
Alternative Title	Suggest common information of person or location which can be used for the alternaive title.
Date Captured	The date associated with site in archive.
Access Condition	The statement about the scope of the resources usages and from where it can be supplied
Collection Title	The name of special domain or project, or the name of information collection organized specially.
Data Medatada Modified	The date on which changes were made to the metadata.
Data Validate	The date on which the page was validated as being properly coded using the W3C Validation Service, HTML-Kit, or other similar services.
Collecting Method	Treat as default value: resource collection method(for example, automatic or manual, or trasferred)
Collecting Tool	Treat as default value: softwares necessary in the process of collecting resources

been adopted.

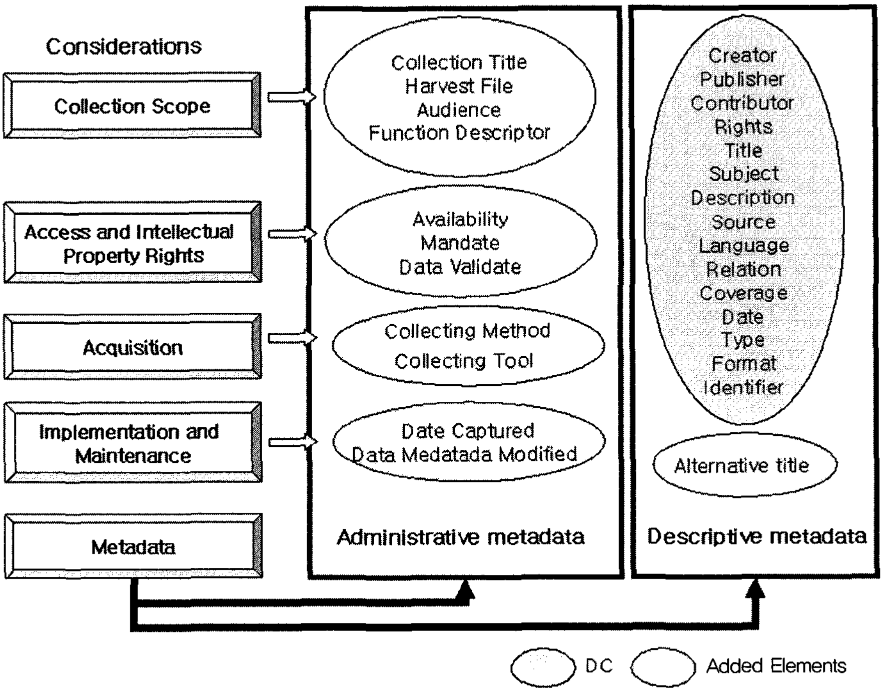
The scope of considerations are within general metadata and comprehensive digital preservation metadata, and it can be applied to web archiving area also.

Administrative metadata focused on convenient access for users are added and suggested in 〈Table 4〉.

elements in addition to the descriptive metadata are suggested as follows: Availability focused on usage, Function Descriptor describes special domain or subject, Mandate which can include related regulations, Harvest File which informs that collected, saved, and described by special institutions, Alternative Title which prepare for the several titles.

Addition and deletion of web pages occurred very often. Therefore the concept of time need to be suggested more specifically. The date when the web page has been captured and the date when the web page is effect has been suggested. moreover, date when metadata has modified has been described. Resources and provenance can be shown through the Collection Title. Creator and name of the collection can be suggested also. Collecting tool can be processed by default value, and through Collecting Method, collecting method, related software, and system configurations can be suggested.

Suggested metadata has been connected to considerations presented in chapter 4.1.



〈Figure 3〉 Suggested Metadata Connected to Considerations

Collection Scope relates to Collection Title, Harvest File,

Audience and Function Descriptor. Access and Intellectual Property Rights has related to Rights in DC, but Availability, Mandate, and Data Validate are added.

For the considerations of Acquisition, Collecting Method and Collecting Tool have been added. For Implementation and Maintenance, Date Captured and Data Medatada Modified are added.

Based on the features of web, administrative elements related to date are added. In Alternative title, descriptive element has been suggested.

As a conclusion, 〈Figure 3〉 has been presented as follows:

5. Conclusion

Although web archiving is still being undertaken by only a few representative organizations, many researchers and practitioners agree on its importance. According to research conducted by RLG, 60 % of members replied that web archiving is a very critical issue(RLG 2006).

Through the analysis of web archiving projects, it has been found that metadata of the intensive web archiving project is based on DC.

Considering the amount of resources which are for archiving, descriptions of detailed elements are difficult. Therefore, in this study only the most essential and common elements based on DC have been selected.

Recently, the importance of the effective use and interoperability of archived resources has increased. In light of this trend, administrative metadata has to be suggested. Related to administrative metadata, elements such as to whom archived resources are to be supplied, and the scope and name of the collection might be considered as the important elements.

For web pages which are more sensitive to change through the passage of time, variable concepts of time can be included in the metadata element.

The final thing to consider is the limit(scope) of rights of the usage. From this, archiving policies can be set and the main users identified.

In this study, metadata elements focused on various intensive projects have been analyzed. For future studies, metadata on extensive projects should be analyzed and compared with those of intensive metadata. The qualitative evaluation of metadata automatically extracted using a web crawler should also be studied.

References

Archival Preservation of Smithsonian Web Resources; strategies, principles, and best practices 〈http://siarchives.si.edu/pdf/dollar_report.pdf〉.

Borghoff, U. M., P. Rodig, J. Scheffczyk, and L. Schmitz 2005. *Long term preservation of digital documents : principles and practices*. New York: Springer.

Brown, Adrian. 2006. *Archiving Websites. facet publishing.*

Day, M. 2003. Collecting and Preserving the World Wide Web; A Feasility Study Undertaken for the JISC and Welcome Trust. [cited. 2007. 3. 2]. 〈www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf〉.

DCC. 2006. *Future-Proofing Web Sites.* Wellcome Library Workshop. 〈http://www.dcc.ac.uk/events/fpw-2006〉.

Dollar Consulting. 2001. Archival Preservation of Smithsonian Web Resources; strategies, principles, and best practices. [cited. 2007. 3. 11].〈http://siarchives.si.edu/pdf/dollar_report.pdf〉.

Fitch, K. 2003. *Web Site Archiving.* AusWeb:The Ninth Australian World Wide Web Conference.

Gilliland-Swetland, Anne J. 2000. *Setting the Stage: Defining Metadata. in Introduction to Metadata: Pathways to Digital Information*. Murtha Baca. ed. Los Angeles: Getty Information Institute.

Gladney, H. M. 2007. Preserving digital information. Berlin: Springer.

Harvard University Library. [cited 2007. 2. 8].〈http://hul.harvard.edu/ois/projects/webarchive/〉.

Harvard University Library. 2006. Web Archiving Collection Development Issues. [cited. 2007. 3. 3]. 〈http://hul.harvard.edu/ois/projects/webarchive/final_coll_dev.pdf〉.

Kenny, A. R., N. Y. McGovern, P. Botticelli, R. Entlich, C. Largoze, and S. Payette. 2002. "Preservation Risk Management for Web Resources; Virtual Remote Control in Cornell's Project Prism." *D-Lib Magazine.* 8(1). [cited. 2007. 3. 2]. 〈http://www.dlib.org/dlib/january02/kenney/01kenney.html〉.

Lazinger, S. S. 2001. *Digital Preservation and metadata: history, theory, practice*. Englewood, Colo.: Libraries Unlimited.

Lyman, P. 2002. *Archiving the World Wide Web*. Council on Library and Information Resources and the Library of Congress.

Masanés, J. 2002. "Towards Continuous Web Archiving." *D-Lib Magazine.* 8(12).

Masanés, J. 2005. "Web Archiving Methods and Approaches: A Comparative Study." *Library Trends.* 54(1):72-90.

Masanes, J. 2006. *Web Archiving*. Springer.

National Library of Australia. [cited 2007. 2. 23]. 〈http://www.nla.gov.au/metadata.html〉.

OCLC/RLG Working Group on Preservation Metadata. 2001. Preservation Metadata for Digital Objects; A review of the state of the art. [cited. 2007. 1. 3]. 〈www.oclc.org/research/pmwg/presmeta_wp.pdf〉.

PADI 〈www.nla.gov.au/padi〉.

Pennock, M. and Kelly, B. 2006. "Archiving Web Site Resources: A Records Management View." *WWW*. May. 23-26.

Rauber, A. et al. 2002. "Uncovering Information Hidden in Web Archives." *D-Lib Magazine*. 8(12). [cited. 2007. 6. 20].

Schneider, S. M. 2004. SUNY Institute of Technology for Library of Congress' 2002 Election Web Collection. [cited 2007. 3. 9]. 〈http://www.webarchivist.org/Election-2002-Web-Archive-Final-Report.pdf〉.

Senserini, A., R. B. Allen, G. Hodge, N. Anderson, D. Smith, Jr.2004. "Archiving and Accessing Web Pages; The Goddard Library Web Capture Project." D-Lib Magazine. 10(11). [cited. 2007. 3. 12]. 〈www.dlib.org/dlib/november04/hodge/11hodge.html〉.

Smithsonian 〈http://siarchives.si.edu/〉.

Suh, H. R. 2004. "Web Archiving: What We Have Done and What We Should Do", The *Journal of the Korean BIBLIA Society for library and Information Science* 15(1)5-20.

The Goddard Library. [cited 2007. 3. 11]. 〈http://library.gsfc.nasa.gov/public/〉

The Goddard Library; Metadata Element Set. [cited 2007. 3. 11]. 〈http://library.gsfc.nasa.gov/mrg/Goddard_Core.htm〉