공공기관 심층 웹기록물 아카이빙을 위한 메타데이터 설계

Metadata Design for Archiving Public Deep Web Records

차승준(Seung-Jun Cha)*, 최윤정(Yun-Jeong Choi)*, 이규철(Kyu-Chul Lee)**

초 록

웹 기술이 발전함에 따라, 공공기관에서는 웹을 이용하여 업무를 처리하고 또한 국가와 시민간의 연결통로로 사용하고 있다. 웹기록물은 공공기관에서 이용하는 웹 사이트상에서의 업무처리의 결과로, 정보로서 중요한 가치를 담고 있으나 보존의 방법과 도구가 부족하여 많은 양의 자원들이 소실되고 있는 실정이다. 본 논문은 웹기록물의 한 분류인 심층 웹기록물 아카이빙에 필요한 메타데이터 설계를 목적으로 하고 있다. 이를 위해 우선 국외 연구기관 및 연방정부에서 제공하는 심층 웹기록물에 대해 알아보고, 이를 바탕으로 국내 공공기관의 심층 웹기록물을 정의하였다. 정의된 심층 웹기록물을 바탕으로 아카이빙에 필요한 메타데이터 항목을 설계하고, 국내외 호환성을 위해 전자기록물 장기보존포맷과 더블린코어 메타데이터와의 관계를 설명하였다. 이는 국내 웹기록물 아카이빙의 기반기술로 활용될 수 있다.

ABSTRACT

According to the development of web sites' technologies, public institutions use web sites to carry out their business and also to utilize as pathway between government and the people also. Public web records means the result of business process over web sites in public institutions. Although there is much valuable information, it is vanished away easily because there is not yet proper methods and tools for preservation. The purpose of this paper is to design the metadata elements required when archiving deep web records, which is a kind of web records. For that, we first analyze oversea's related researches to define what public deep web records is. Then we define metadata elements about that and also explain the relationship on archival information package in Korea and dublin core metadata to support interoperability for them. The defined metadata can be used for the basis technologies in archiving domestic public web records.

키워드 : 심층웹, 웹기록물, 아카이빙, 공공기관, 메타데이터
Deep Web, Web Records, Archiving, Public Institution, Metadata

---

본 연구는 행정안전부 국가기록원의 지원을 받아 기록물 보존기술 연구개발(R&D) 사업의 일환으로 이루어졌으며, 이에 감사드린다. 또한 연구는 지식경제부 및 정보통신산업진흥원의 대학 IT연구센터 육성·지원사업(NIPA-2009-C1090-0902-0031)의 연구결과로 수행되었음.

* 충남대학교 공과대학 컴퓨터공학과
** 교신저자, 충남대학교 공과대학 컴퓨터공학과

2009년 09월 23일 접수, 2009년 10월 07일 심사완료 후 2009년 11월 06일 게재확정.

1. 서 론

개인이 손쉽게 접근할 수 있는 일반적인 웹 사이트들은 웹이 개발된 이후 기하급수적으로 증가하고 있다. 이러한 웹 사이트와는 다르게 기록으로서의 보존 가치가 있는 웹 사이트도 증가하고 있다. 이는 기록의 속성(내용, 구조, 맥락)을 통한 증거능력을 확보할 수 있는 특징을 가진 웹 사이트로, 특히 공공기관의 웹 사이트들이 이에 해당된다. 이는 전자정부의 발전, 특정 업무과정 및 국가와 시민간 상호 작용의 증거이다. 이는 단체나 개인이 업무 수행 과정을 웹을 기반으로 한 것으로, 생산 및 접수과정에 대해 법정 증거로 사용하거나 또는 그 자체 내용의 정보가치 때문에 유지 관리 및 보존해야 한다[1].

이러한 공공기관에서 이용하는 웹 사이트 상에서의 결과를 웹기록물 이라고 한다. 웹기록물은 전자적으로 되어 있다는 점에서 관계법령에서 정의하는 전자기록과 전자문서의 범주에 속할 수 있지만, 기록물이 웹사이트에 있다는 점에서 차이를 둘 수 있다.

웹기록물은 웹을 통해 접근되고 관리되기 때문에 웹과 같은 특성들을 가진다. 대표적인 특성으로는 지속적인 수정과 삭제가 발생하는 '휘발성', 하이퍼링크 기반의 불연속적인 연결로 이어진 '불연속성', 복제와 전송이 용이하여 여러 가지 형태로 증가하는 '증식성', 텍스트/이미지/오디오 등 동시에 존재할 수 있는 '다양성' 등이 있다[2, 3]. 특히 휘발성의 특성을 가진 웹기록물은 그 생성과 삭제가 빈번하게 이루어지기 때문에 보존의 가치가 있는 자원이지만 보존의 방법과 도구가 없어 많은 양의 정보들이 소실되고 있는 실정이다[8].

이러한 웹기록물은 웹에서와 마찬가지로 수집기로 접근 가능한 표면 웹기록물과 접근 불가한 심층 웹기록물로 구성된다[5]. 표면 웹기록물은 접근할 때마다 동일하게 표현되는 정적인 문서의 집합으로 구성된 웹사이트로 기관소개, 연혁 등의 자료가 이에 속한다. 심층 웹기록물은 사용자 요구가 변경될 때마다 저장된 데이터베이스 내용도 갱신되는 웹사이트로 키워드 검색, 공지사항, 게시물, 자료실 등이 이에 속한다.

본 논문은 사라져가는 웹기록물들에 대한 보존을 위해 웹기록물 중 심층 웹기록물 아카이빙에 같이 저장되어야 하는 필수 사항인 메타데이터 요소 정의를 목적으로 하고 있다. 이는 장기보존에 있어 정보의 검색, 관리, 조작, 보존을 하기 위해 반드시 필요하다.

국내외 많은 기관에서 이미 아카이빙에 대해 연구하고 있지만 표면 웹기록물만 수집하거나, 정책적으로만 정의되어 있을 뿐 실제 수집이 이루어지지 않는 실정이다. 국내 심층 웹기록물의 아카이빙을 위해 우선 해외 아카이빙의 사례로 호주, 뉴질랜드, 스위스를 조사하였다. 분석된 내용을 바탕으로 국내의 대표적인 공공기관 중 협조가 가능한 기관의 웹 사이트 특성을 분석했다. 이를 바탕으로 우리나라 공공기관 웹사이트에서 보존해야 할 심층 웹기록물에 대해 정의하였고, 수집, 보존, 전달을 위해 같이 저장해야 하는 메타데이터 항목 요소에 대한 개발을 했다.

본 논문의 구성은 다음과 같다. 제 2장에서는 관련연구로 국외 연구기관 및 연방 정부에서의 아카이빙 정책에 대해 알아보고, 제 3장에서는 심층 웹기록물에 대해 정의하였다. 제 4장에서는 앞선 내용을 바탕으로 국내 공

공기관의 심층 웹기록물 메타데이터인 KoDeWeb을 정의하였다. 제 5장에서는 본 논문의 결론 및 기대성과에 대해 설명한다.

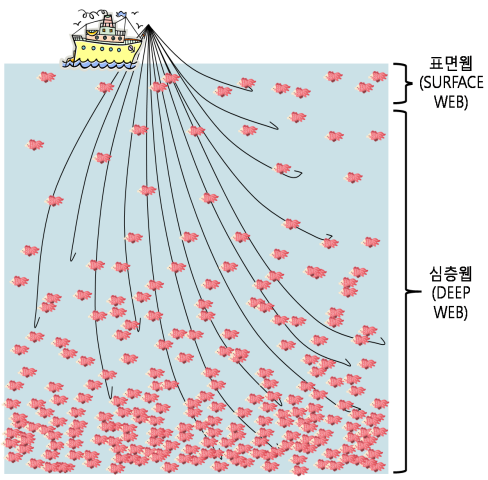
2. 관련연구

심층 웹기록물 아카이빙은 국외 연구기관 및 연방 정부에서 많은 관심을 가지고 연구를 진행하고 있다. 대표적인 연구기관으로는 브라이트플래닛(BightPlanet)와 심층웹 연구(Deep Web Research)가 있다. 또한 국외 연방 정부 중 호주, 뉴질랜드, 스위스에서는 웹기록물과 관련된 표준들을 제시하고 있다.

2.1 국외 연구기관의 사례

브라이트플래닛(BrightPlanet)[10]은 미국 사우스다코다주의 벤처기업으로 보통 검색 엔진의 힘이 닿지 않는 '웹의 심층부' 또는 '보이지 않는 웹'까지 검색하는 기술을 개발, 일반 검색 엔진의 한계를 넘어서는 일에 도전하고 있다. 웹의 구성요소는 <그림 1>[14]과 같이 일반적인 검색 엔진으로 접근 가능한 표면웹(Surface Web)은 현재 15억 개의 웹페이지 정도이지만, 중요한 내용을 담고 있지만 접근이 불가한 심층웹(Deep Web)은 무려 5500억 개에 달한다고 한다. 이러한 심층웹의 정보는 기업별 연봉 비교 통계, 미국특허청의 사진 데이터베이스, 미 증권거래위원회의 각종 기록, 학술논문, 인구통계자료, 의회도서관 기록, 의학연구, 예술사진, 음악파일 등 매우 다양하다.

따라서 브라이트플래닛은 이러한 심층웹으로



〈그림 1〉 표면웹과 심층웹의 비교

여겨지는 접근 불가능한 컨텐츠와 웹의 표면적 자원을 수집하는 선두적인 업체로, 관련된 백서를 2002년에 발간하였다. 수집 대상으로 표면웹의 문서 뿐만 아니라 범위를 넓혀서 심층웹의 문서를 포함시키며 뿐만 아니라 법률정보, 다우존스 지수와 같은 사유의 데이터 자원과 고객 자신의 내부 정보자원도 포함하고 있다. 또한 컨텐츠 발견, 수집, 관리, 분석을 위한 플랫폼으로 DQM(Deep Query Manager)을 제공하며 컴플리트플래닛(Complete Planet)를 통해 수집한 심층웹 데이터베이스에 접근을 제공한다.

심층웹 연구(Deep Web Research)[12]는 Marcus P. Zillman이 가상의 개인 도서관(Virtual Private Library)에 게시하는 정기 간행물로 2004년부터 매년 데이터를 추가하고 있다. 이는 심층웹 연구에 대한 역사를 더 잘 이해하기 위해 다양한 자료들로 구성된다. 최근 버전은 2008년 12월에 발간된 것으로 총 10개의 섹션, 500건 이상의 문서로 구성되어 있다. 각 문서는 주로 심층 웹 정보수집

도구에 대한 정보를 제공하며 또한 매년 발간하는 자료를 통해 관련된 기사나 논문에 대한 정보를 제공한다.

브라이트플래닛은 공공기관이 아닌 개인회사로 백서를 통해 심층웹에 대한 일반론을 제공하고 있지만 관련 도구 및 내용은 자세히 제공하지 않는다. 심층웹 연구에서 제공하고 있는 대부분의 문서들은 폼(form)을 이용한 방식으로 이는 데이터베이스의 일부만 가져오는 것이기 때문에 진본성을 제공해야하는 공공기관 웹 사이트에 적용하기에는 어려움이 있다.

2.2 국외 연방정부의 웹기록물 보존 방안

호주 국립기록보존소에서 작성된 '연방정부의 웹 기반 활동의 기록 관리에 관한 지침(Archiving Web Resources : Guidelines for Keeping Records fo Web-based Activity in the Commonwealth Government)'[15]은 연방기관의 웹기반 활동에 관한 기록 지침을 제시한다. 이 문서에서는 기관들이 웹기반 활동에 관한 진본이고, 신뢰성 있으며, 정확하고, 지속적인 증거를 생산하고 관리하는 것이 필수적이라고 정의하였다. 또한 각 기관이 가져야 할 웹기반 기록을 포함한 기록을 획득하고 관리해야 할 중요한 보존 책임을 명시하였다. 호주에서는 웹 사이트의 다양한 형태에 따라서 웹자원을 4가지로 구분하였다. 가장 기본적인 형태로 웹 사이트 서버의 폴더 내에 존재하며 하이퍼링크로 연결되어 있는 '정적 웹 사이트와 웹자원'이 있으며, 방문자의 의견과 요구사항 등 정보를 수집하기 위해 서식(Form)을 이용한 '서식에 기반한 상호작용이 있는 정적 웹사이트와 웹 자원'이 있다. 이 두 가지는 표면 웹기록물에 해당하는 자료이며, 아래 두 가지는 심층 웹기록물에 해당하는 자료이다. 사이트 사용자들에게 준비된 목록을 검색하거나, 데이터베이스의 컨텐츠에 대한 문의, 결과를 보여주는 '동적 데이터 접근에 기반한 웹 사이트와 웹 자원'이 있으며, 전체 페이지를 로딩시에 생성하여 사용자들의 선택사항, 접근권한, 현재의 요구사항, 기술성능, 제약사항에 따라 다른 결과물을 보여주는 '동적으로 생성된 웹사이트와 웹 자원'이 있다.

메타데이터의 적용은 웹기반 활동기록을 포함한 모든 기록을 성공적으로 관리하는데 중요한 것으로, 기본적으로 호주정부 위치식별 서비스(AGLS, Australian Government Locator Service) 메타데이터 표준을 준수하는 메타데이터를 사용할 것을 요구하고 있다. 또한 한편으로 온라인 자원을 포함하여 기록에 해당되는 정보자원을 기술하는데 사용하는 추가 메타데이터를 규정하고 있다. 웹 사이트 상의 기록과, 웹기반 활동 기록에 대한 추가적인 요소는 다음과 같다.

<웹사이트 상의 기록>

- 기록관리시스템으로의 기록의 생산 및 등록 일시
- 구조적 맥락
- 원 데이터의 포맷
- 웹 사이트의 위치를 포함하여, 기록을 사용한 용도
- 기록의 생산, 보유, 처분에 적용되는 규정
- 판결, 보존, 처분을 포함한 생산 이후의 기록 관리 이력

<웹 기반 활동 기록>

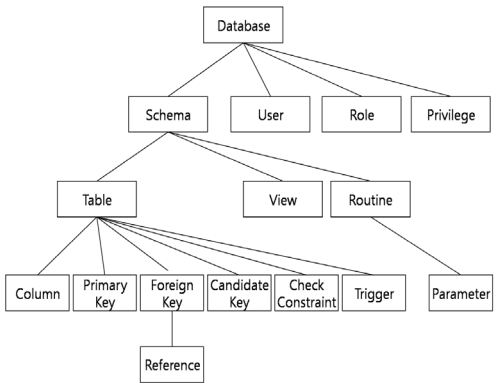
- 획득일시
- 특정 URI에 링크된 날짜 및 버전 정보를 포함한 인터넷 식별자(URI)에 대한 링크
- 웹 사이트 설계에 대한 기술적 세부사항
- 웹 자원을 생산하기 위해 사용된 소프트웨어의 세부사항
- 검색 엔진을 포함하여, 웹 자원을 보완하는 어플리케이션의 세부사항
- 웹 자원을 보여주는데 필요한 고객용 소프트웨어의 세부사항

웹 사이트 기록관리 전략 개발 지침(A Guide to Developing Recordkeeping Strategies for Websites)[9]은 뉴질랜드 국립기록보존소에서 웹 사이트에 대해 공공 기록이나 지방당국의 기록으로 관리하기 위한 전략을 개발하려는 목적으로 만들어진 문서이다. 이 문서는 웹 사이트를 기록으로 관리하기 위한 핵심원칙을 명확하게 제시하고, 조직들이 활용할 수 있는 기록관리 전략의 범위를 웹 사이트의 기술 분석에 근거하여 산정한다. 또한 웹 사이트를 정부기관이 웹상에 존재하는 기본적인 정보와 전달에서부터 중요한 정보에 대한 온라인 접근, 서식 다운로드, 이메일 교환, 업무처리 시스템 계열을 연속적으로 통합함으로써 온라인에서 발생하는 트랜잭션을 완료하는 것까지 다양하게 정의하고 있다.

뉴질랜드에서의 기록관리 메타데이터는 기록의 생산, 관리, 이용을 지속적으로 가능하게 하는 데이터로 정의하고 있다. 기록관리 메타데이터는 기록을 생산하고, 관리하고, 유지하고, 활용하는 사람과 프로세스와 시스템뿐만 아니라 기록을 식별하고, 진본임을 입증하며, 배경정보를 제공하는데 사용될 수 있다.

스위스 연방 정부에서는 빠르게 변하는 IT 환경에서 웹 자원들에 대한 손실없는 보존을 위해 웹 사이트를 보존의 필요성을 강조하였다. 웹 사이트에서 제공하는 정보에 대한 보증을 제공하여 정부 활동에 대한 문서화를 제공하기 위해 아카이빙을 법으로 규정하였다. 특히 웹 사이트에서 심층 웹에 해당하는 관계형 데이터베이스를 통해 제공되는 자료를 SIARD 포맷(Software Independent Archival of Relation Database)[16]을 통해 아키이빙 하도록 권고 하였다. SIARD는 스위스 정부의 SFA(Swiss Federal Archives)의 디지털 아카이빙을 위한 ARELDA 프로젝트의 일부분으로, 관계형 데이터베이스의 정보를 오랜 기간동안 보존할 수 있도록 정의한 표준적인 파일 포맷이다.

SIARD 포맷은 헤더(Header) 폴더와 컨텐츠(Content) 폴더로 구분하여, 헤더에는 데이터베이스의 컨텍스트와 메타데이터를 저장하며 컨텐츠에는 데이터베이스의 구조에 따라 실제 데이터가 저장된다. <그림 2>는 메타데



〈그림 2〉 SIARD 메타데이터 구조

이터의 계층 구조를 도식화 한 것이다. 데이터베이스는 하나 이상의 스키마타(Schemata), 사용자(user), 역할(role)에 대한 정의로 구성된다. SQL : 1999에서 사용자와 역할은 권한(Privilege)으로 사용될 수 있다. 각 스키마는 테이블, 뷰, 루틴의 값으로 정의된다.

SIARD에서는 아카이빙된 자료의 장기보존을 위해 표준화를 많이 고려하여, 데이터와 메타데이터는 ISO 표준인 SQL : 1999을 이용하였으며 각 데이터는 W3C 표준인 XML 1.0을 이용하여 저장하였다.

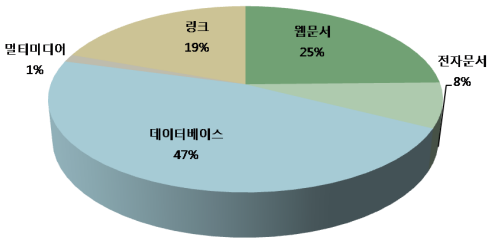
3. 심층 웹기록물의 정의

심층 웹기록물을 정의하기 위해 우선 공공기관 웹 사이트의 특성을 분석하였다. 또한 심층 웹기록물은 아카이빙 관점에서 표면 웹기록물 아카이빙에서 수집되지 못한 요소들에 대한 아카이빙이기 때문에 표면 웹기록물 아카이빙에 대해 살펴보았다. 이와 더불어 일반적인 정의론을 바탕으로 심층 웹기록물에 대해 정의하였다.

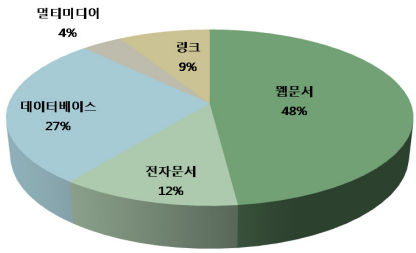
3.1 공공 홈페이지 특성 분석

공공기관의 홈페이지는 대표적인 기관 중 협조가 가능한 기관의 것으로 테스트베드를 선정하였다. 이에 해당하는 국가기록원과 행정안전부 웹 사이트를 대상으로 각각이 포함하고 있는 자원의 특성을 분석하였다. 각각의 결과는 <그림 3>, <그림 4>와 같다.

'웹문서'는 수집기를 이용하여 수집할 수 있는 텍스트와 이미지이며, '전자문서'는 워드, 엑셀 등과 같은 문서 파일을 제공하는 페이지이다. '멀티미디어'는 사운드, 이미지를 제공주는 페이지를 의미하며, '데이터베이스'는 실제 데이터베이스를 통해 동적으로 생성되는 페이지이다.



〈그림 3〉 국가기록원 홈페이지 분석



〈그림 4〉 행정안전부 홈페이지 분석

웹문서는 물론이고 전자문서와 멀티미디어는 표면 웹기록물 아카이빙을 통해서 수집해서 보존할 수 있다. 하지만 데이터베이스는 데이터의 구조적 특성으로 표면 웹 기록물 아카이빙을 통해서는 수집할 수 없다.

3.2 표면 웹기록물 아카이빙의 제약사항

[4]에서는 표면 웹기록물 아카이빙을 위한 도구인 웹 크롤러를 개발하였다. 일반적인 웹기록물을 아카이빙 할 수 있는 도구를 분석하여, 이 중 Heritrix를 국내 웹기록물 아카이빙에 맞게 확장하였다.

이러한 도구를 이용하여 표면 웹기록물을 아카이빙 할 때의 기술적인 한계점으로는 자바스크립트 또는 플래쉬와의 관계, 수집기(크롤러)의 제약사항, 로그인 후 가져와야 하는 데이터들, 데이터베이스로부터 생성된 동적인 페이지 정보, 운영중인 사이트에 존재는 하지만 사용되고 있지 않아 접근이 되지 않는 페이지들의 존재 등을 들 수 있다. 이러한 대부분의 한계점은 표면 웹기록물 아카이빙을 위한 수집기의 향상을 통해 해결이 가능하다. 하지만 데이터베이스의 동적으로 구성되는 데이터들은 표면 웹기록물 아카이빙을 통해서는 수집이 불가능하다.

3.3 심층 웹기록물 정의

심층 웹기록물을 정의하기 위해 다른 기관의 정의를 살펴보았다. 위키피디아 및 다른 아카이빙 기관에서는 표면 웹의 부분이 아닌 웹의 컨텐츠로 일반적 검색 엔진으로 접근되지 않는 것이라고 하였다. 특히 영국 도서관에서는 보이지 않는 웹(Invisible Web)이라고도 불리우며 온라인 데이터베이스와 같이 큰 저장소에 있는 컨텐츠로 웹 크롤러들이 접근할 수 없는 것으로 기존 검색 엔진의 제약으로 사용되지 않고 있다고 하였다. 하지만 SIARD에서는 여러 데이터베이스 중 관계형 데이터베이스라고 하였는데, 이는 관계형 데이터베이스가 데이터와 코드에 독립적이어서 모든 데이터들이 테이블에 존재하여 실제로 아카이빙은 테이블에 존재하는 자료를 추출하면된다. 또한 현존하는 시스템의 90% 이상이 관계형 데이터베이스를 이용한다.

뿐만 아니라 보존 대상인 웹기록물은 표면 웹기록물과 심층 웹기록물로 구분하여, 심층 웹기록물을 표면 웹기록물이 아카이빙 하지 못하는 나머지 부분으로 정의할 수 있다. 이는 데이터베이스를 통해서 동적으로 생성되는 데이터이다.

따라서 보존대상인 심층 웹기록물이란 표면 웹기록물 아카이빙에서 수집 불가한 관계형 데이터베이스에 저장되어 동적으로 페이지를 생성하는 데이터라고 정의할 수 있다.

4. 메타데이터 설계

공공기관의 웹기록물들은 선별정책에 의해 보존가치의 여부가 결정되어지고 보존가치가 있는 웹기록물을 대상으로 아카이빙하여 만들어진 심층 웹기록물은 시간의 흐름에 따라 그 규모와 양이 증가하게 된다.

메타데이터는 구조화된 정보로 시간이 흘러도 정보에 대한 검색, 관리, 조작, 이해, 보존이 가능할 수 있게 해주는 데이터이다[13]. 따라서 심층 웹기록물에 대한 분류, 보존자료에 대한 무결성 검토, 관련 기술자료 검토, 검색을 위해 기록물을 설명하는 메타데이터를 포함해야 한다.

국내 및 국외 웹 아카이빙 사례들을 살펴보면 더블린 코어를 중심으로 메타데이터를 기술한 것을 확인할 수 있다[6]. 또한 웹기록물은 전자적으로 되어 있기 때문에 국내 법령에서 정의하고 있는 전자문서 범위에 속하게 된다. 따라서 국내외 아카이빙들과의 호환성을 위해 더블린 코어로 정의되며, 국내 전자기록물과의 호환성을 위해 장기보존포맷으로 정의되어야 한다.

4.1 더블린 코어 메타데이터

더블린 코어(Dublin core)는 OCLC(Online Computer Library Center)와 NCSA(National Center for Supercomputer Application)의 지원 아래, 분산된 정보자원의 관리를 목적으로 도서관 관련 연구자와 교수, 프로그램 개발자의 의견을 수렴하여 네트워크 환경 기반을 염두에 두고 마련한 새로운 메타데이터 표현 기준이다[11].

더블린 코어에서 정의하는 요소(Elements)란 데이터가 나타내는 자원의 일반적인 사항을 담고 있는 것으로, 데이터를 검색하기 용이하게 하기 위한 요소들을 포함한다. 다음의 총 15개의 엘리먼트가 정의 되어 있다.

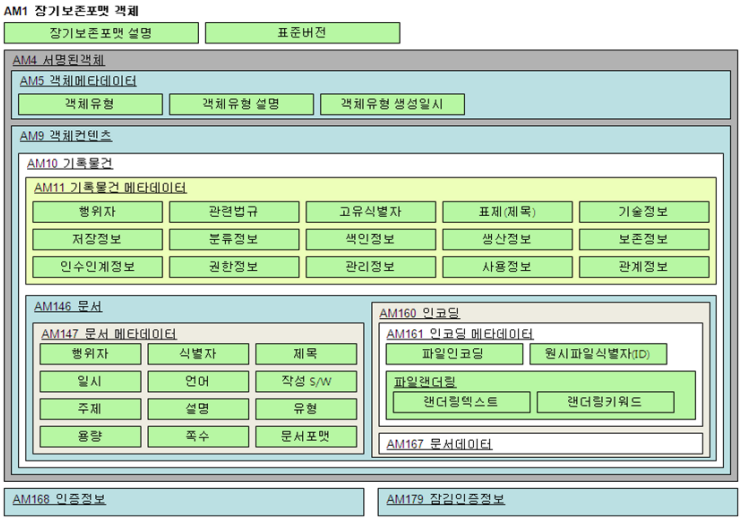
- 표제(Title) : 널리 알려진 자원에 부여한 명칭
- 생성자(Creator) : 자원의 내용을 작성하는데 주된 책임을 진 개체
- 주제(Subject) : 자원의 내용이 지닌 주제
- 요약정보(Description) : 정보의 내용에 대한 설명정보
- 발행처(Publisher) : 해당 자원을 이용할 수 있도록 책임을 진 개체
- 기여자(Contributor) : 자원의 내용이 기여한 개체
- 날짜(Data) : 자원의 존재기간동안 어떠한 사건과 관계된 날짜
- 자료유형(Type) : 해당 자원의 내용에 관한 성격이나 장르
- 파일형식(Format) : 자원의 물리적・디지털 구현형
- 식별자(Identifier) : 특정한 상황에서 자원에 대한 분명한 참조
- 정보원(Source) : 현재의 자원이 유래한 자원에 대한 참조
- 언어(Language) : 자원의 지적 내용의 언어
- 관련자원(Relation) : 관련된 자원에 대한 참조
- 내용범위(Coverage) : 자원의 내용에 대한 외연이나 범위
- 이용조건(Rights) : 자원에 관한 권리에 관한 정보

4.2 전자기록물 장기보존포맷 기술 규격

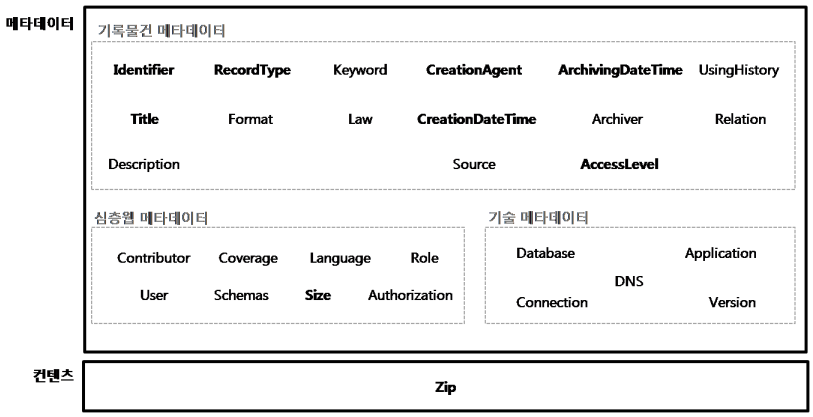
전자기록물 장기보존포맷 기술 규격[7]이란 전자기록물의 진본성과 무결성을 보장하고 장기간 안전하게 보존하기 위해 전자기록물 원문, 문서보존포맷, 메타데이터, 전자서명을 하나의 패키지로 구성한 포맷이다.

전자기록물은 매체의 특성상 변형, 훼손, 유실되기 쉬우므로 안전하게 보존할 수 있어야 한다. 그리고 오랜 기간이 경과해도 기록물이 생산된 당시에 가지고 있던 내용을 그대로 재현하여 접근할 수 있도록 보존되어야 하며, 업무활동에 대한 증거이며, 업무에 대한 책임소재를 분명하게 밝혀주는 기록물의 법적증거를 확보할 수 있도록 하여야 한다. 이러한 기록물의 법적증거는 진본성과 무결성이 유지되어야만 확보할 수 있다.

장기보존 포맷은 <그림 5>와 같이 원문, 문서보존포맷, 장기보존포맷 메타데이터, 전자서명으로 구성되며 이 구성요소를 XML을 이용하여 단일한 객체로 패키지 한다. 단일한 객체로 패키징하는 이유는 시스템과 기관을 옮겨다니면서 장기간 보존되는 전자기록물의



〈그림 5〉 전자기록물 장기보존포맷 기술규격



〈그림 6〉 KoDeWeb 구성

관리를 편하게 해주며, 기록물의 유실 및 훼손에 대한 위험성을 줄여주기 때문이다.

원문은 생산자가 처음 생산한 기록물로 진본성을 보장하기 위해 포함하고, 문서보존포맷은 시간과 기술변화에 상관없이 이용자가 접근할 수 있도록 하는 문서보존포맷을 포함한다. 장기보존포맷 메타데이터는 향후에 기록물에 대한 이해와 활용이 가능하므로 원문을 그대로 유지하고, 원문을 이해하는데 필요한 메타데이터를 함께 포함한다. 전자서명은 전자기록물의 진본성 및 무결성 보장을 위해 전자서명을 포함한다.

4.3 KoDeWeb

KoDeWeb(Korea Deep Web)은 우리나라의

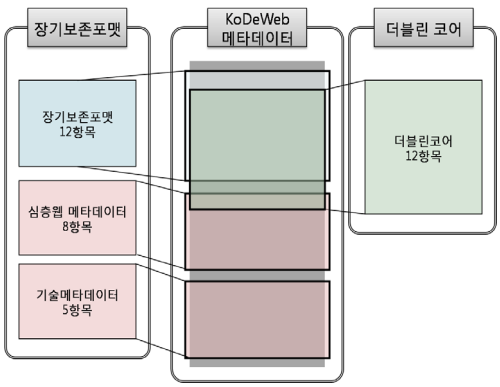
공공기관의 심층 웹기록물을 의미하는 것으로, <그림 6>과 같이 심층 웹기록물의 메타데이터와, 실제 데이터베이스의 내용을 담고 있는 컨텐츠 부분으로 나누어 구성하였다.

기본적인 메타데이터 항목은 전자기록물 장기보존포맷의 '기록물건'에 정의된 15가지 항목을 기반으로 정의하였으며, 심층웹 문서에 대한 메타데이터 항목으로 8가지 항목을 정의하였다. 또한 보존/복원을 위해 사용되는 데이터베이스에 대한 기술(Technical) 메타데이터 요소로 5가지를 정의하였다. 각 메타데이터의 항목은 <표 1>에서 확인할 수 있다.

<그림 7>는 전자기록물 장기보존 포맷, KoDeWeb 메타데이터, 더블린 코어 메타데이터와의 관계를 보여준다.

정의된 KoDeWeb 메타데이터 항목은 전자기록물 장기보존포맷의 필수항목을 중심으로 12개의 요소가 매핑되며, 문서 메타데이터에 심층웹 메타데이터 8개의 항목이 매핑되고 추가적인 기술 메타데이터 5가지 항목이 저장된다.

더블린 코어는 전자기록물 장기보존 포맷에 해당하는 12개의 항목과 심층웹 메타데이터에서 3개의 항목이 매핑되어 저장된다.



〈그림 7〉 KoDeWeb 메타데이터 구성

KoDeWeb의 메타데이터 항목 요소를 기술할 때 텍스트 기반인 소프트웨어나 하드웨어에 의존적이지 않고 개방적인 표준인 XML을 이용한다. 또한 메타데이터 요소들간의 관계를 설정하고 이를 의미 있게 하기 위해 메타데이터는 스키마에 의해 구조화될 필요가 있다. 따라서 XML Schema를 사용해서 정의했다.

5. 결 론

웹은 급속하게 변화하는 현대사회에서 정부와 시민들의 주요 의사소통 채널이 되고 있다. 정부의 업무처리도 점점 이러한 웹 환경으로 변해가고 있다.

하지만, 웹 사이트상에서의 업무처리의 결과인 웹기록물은 보존의 방안과 도구의 부재로 쉽게 사라지고 있는 실정이다. 따라서 본 논문은 웹기록물의 한 종류인 심층 웹기록물 아카이빙시 같이 저장되어야 하는 메타데이터 요소의 설계에 관한 것이다.

우선 국외 기관 및 연방 정부에서 정의하고 이용하는 웹기록물에 대해 살펴보았다. 또한 국내 공공기관 웹 사이트의 특성을 살펴보고, 표면 웹기록물 아카이빙에서 아카이빙 할 수 없는 요소들을 살펴보았다. 이들을 바탕으로 아카이빙의 대상이 되는 심층 웹기록물을 정의하였다.

메타데이터는 데이터의 장기보존에 있어 정보의 검색, 관리, 조작, 보존을 하기 위한 필수적인 요소로, 심층 웹기록물을 아카이빙 하는데 반드시 필요하다. 따라서 본 논문에서는 심층 웹기록물 아카이빙에 필요한 메타데이터를 설계하였다. 이는 국내 전자기록물과

〈표 1〉 KoDeWeb 메타데이터 항목

구분	요소	설명	장기보존포맷 요소	더블린 코어 요소
장기보존포맷 메타데이터	Identifier	아카이빙 식별자	고유식별자	Identifier
	Title	아카이빙 제목	공식표제	Title
	Description	수행한 내용에 대한 설명정보	기술정보내용	Description
	RecordType	유형	유형구분	Type
	Format	저장매체 및 데이터 포맷	포맷	Format
	Keyword	대표할 수 있는 키워드	키워드	Subject
	Law	관련된 법규에 대한 정보		
	CreationAgent	기록물 생산자 정보	생산자정보	Publisher
	CreationDateTime	기록물 생산 일시		
	Source	출처정보	출처정보	Source
	ArchivingDateTime	아카이빙 수행 일시	보존처리일시	Date
	Archiver	아카이빙 행위자	보존행위자	Creator
	AccessLevel	열람범위, 접근권한	보안분류	Right
	UsingHistory	사용한 기록의 정보		
	Relation	다른 아카이빙과의 관계	관계정보	Relation
심층웹 메타데이터	Contributor	자원의 기여자		Contributor
	Coverage	자원의 범위		Coverage
	Language	사용된 언어		
	Roles	데이터베이스 사용자 역할 정보		
	Users	데이터베이스 사용자 정보		
	Schemas	데이터베이스 내 스키마 정보		
	Size	아카이빙의 용량		
	Authorization	데이터베이스 내부 사용자 권한 정보		
기술메타데이터	Database	데이터베이스 제품의 버전		
	Application	아카이빙 생성 프로그램		
	DNS	아카이빙 수행 컴퓨터 DNS 정보		
	Connection	데이터베이스 연결 설정 정보		
	Version	아카이빙의 버전		

의 호환성을 위해 전자기록물 장기보존포맷으로 정의되며, 국외 기관들과의 호환을 위해 더블린 코어로도 정의하였다.

본 논문에서 설계한 심층 웹기록물 메타데이터는 추후 관련 도구의 개발을 통해 국내 웹기록물 아카이빙의 기반기술로 활용할 수 있다.

또한 심층 웹기록물의 보존은 차기 연구인

웹 정보검색 기술의 활용과 행정정보 데이터베이스 보존기술로 활용될 수 있다. 또한 민간이 운영하는 웹사이트에 대한 영구보존에 활용될 수 있다.

참 고 문 헌

[1] 김유성, "공공기록물 관리에 관한 법률의 제정 의의와 개선방안", 한국기록관리학회, 제8권, 제1호, 2008, pp. 5-25.

[2] 유효림, "정부부처 웹 아카이빙 방안 연구", 명지대학교 석사학위 논문, 2007.

[3] 이지은, "공공기관의 웹기록 관리방안 연구", 한국 외국어대학교 석사학위 논문, 2006.

[4] 차승준, 정준선, 이규철, "공공기관 웹기록물 아카이빙을 위한 웹 크롤러 연구 개발", 한국정보과학회 데이터베이스 연구, 제26권, 제2호, 2009. 8, pp. 1-15.

[5] 차승준, 이규철, "웹기록물 아카이빙 기반기술 연구 개발", 지식정보산업연합학회 창립기념 학술대회, 창간호, 2008, pp. 369-377.

[6] 차승준, 천동석, 이규철, "웹기록물 아카이빙을 위한 워크플로우 및 메타데이터 연구", 제30회 한국정보처리학회 추계학술발표대회, 제15권, 제2호, 2008, pp. 1379-1382.

[7] 행정안전부 국가기록원, "전자기록물 장기보존포맷 기술규격(Standard of Archival Information Package)", 2008.

[8] Adrian B., "Archiving Website : a practical guide for informatio management professionals," facet publishing, 2006.

[9] Archives New Zealand, "A Guide to Developing RecordKeeping Strategies for Websites," 2008, http://www.archives.govt.nz/continuum/documents/publications/g20.pdf.

[10] BrightPlanet, http://www.brightplanet.com/.

[11] Dublin Core Metadata Initiative, "Dublin Core Metadata Element Set. version 1.1," 2008, http://dublincore.org/documents/dces/.

[12] DeepWebResearch, http://www.deepwebresearch.com/.

[13] Heejung, K. Hyewon, L. 2007. Development of Metadata Elements for Intensive Web Archiving., 정보관리학회지, 제24권, 제2호, pp. 143-160.

[14] Michael K. BERGMAN, "The Deep Web : Surfacing Hiidden Value," Bright Planet White Paper, 2001, http://www.brightplanet.com/images/uploads/12550176481-deepwebwhitepaper.pdf.

[15] National Archives of Australia, "Archiving Web Resources : Guidelines for Keeping Records of WEb-based Activity in the Commonwealth Government," 2001, http://www.naa.gov.au/Images/arch-web_guide_tcm2-903.pdf.

[16] Swiss Federal Archives, "SIARD Format Description," 2009, http://www.bar.admin.ch/themen/00532/00536/00818/index.html?lang=en.

저 자 소 개

차승준	(E-mail : junii@cnu.ac.kr)
2006년	충남대학교 공과대학 컴퓨터공학과 학사
2006년～현재	충남대학교 공과대학 컴퓨터공학과 석박사통합과정 재학중
관심분야	데이터베이스, 웹서비스, GIS, 웹 아카이빙

최윤정	(E-mail : yunjeong@cnu.ac.kr)
2004년	한남대학교 공과대학 컴퓨터 멀티미디어 전공 학사
2004년～2005년	솜씨 디자인 학원 강사
2005년～2008년	한전원자력연료 연구원
2008년～현재	충남대학교 공과대학 컴퓨터공학과 석사
관심분야	데이터베이스, 시맨틱 웹, 웹서비스, 웹 아카이빙

이규철	(E-mail : kclee@cnu.ac.kr)
1984년	서울대학교 공과대학 컴퓨터공학과 학사
1986년	서울대학교 공과대학 컴퓨터공학과 석사
1990년	서울대학교 공과대학 컴퓨터공학과 박사
1994년	미국 IBM Almaden Research Center 초빙 연구원
1995년～1996년	미국 Syracuse University 초빙 교수
2001년～현재	전자상거래 표준화 통합 포럼 전자거래 기반 기술위원회 위원장
2003년～현재	한국전자거래학회 편집이사
2003년～현재	웹 코리아 포럼 부위원장
2005년～현재	한국정보과학회 논문편집위원
2005년～현재	한국 기록관리학회 이사
2006년～현재	충남대학교 소프트웨어연구소 소장
2007년～현재	국가기록원 기록관리평가위원회 위원
현재	충남대학교 공과대학 컴퓨터공학과 교수
관심분야	데이터베이스, XML 웹 서비스, 시맨틱 웹 서비스, 유비쿼터스 컴퓨팅, 웹 아카이빙