# 웹기록물 보존을 위한 전자기록물 장기보존포맷 확장 설계

## Extension of the Long-term Archival Information Package for Electronic Records to Accommodate Web Records

박병주(Boung-Joo Park)*, 차승준(Seung-Jun Cha)**, 이규철(Kyu-Chul Lee)***

**초 록**

웹기록물은 공공기관의 업무활동이나 전자상거래에 대한 법적증거로 활용될 수 있기 때문에 보존할 가치가 있는 정보이지만 웹기록물의 특징 중 하나인 '휘발성'으로 인해 소실되고 있다. 따라서 이렇게 사라지는 웹기록물을 장기보존하기 위한 장기보존포맷이 정의되어야 한다. 웹기록물은 전자기록물의 일종이기 때문에 전자기록물 장기보존포맷에 보존할 수 있어야 한다. 하지만 현재 표준으로 제시된 포맷은 웹기록물의 특성을 고려하지 않고 정의되었기 때문에 웹기록물을 보존할 수 없다. 본 논문에서는 표면/심층 웹기록물 문서보존포맷으로 연구된 KoDeWeb/KoSurWeb과 전자기록물 장기보존포맷을 분석하고, 이를 바탕으로 웹기록물을 보존할 수 있는 확장된 전자기록물 장기보존포맷을 정의하였다. 정의된 포맷을 활용하면 웹기록물도 전자기록물들과 같이 보존되어 활용될 수 있고, 전자 상거래에 관련된 공공기관의 웹기록물을 보존함으로써 전자 상거래에 대한 법적 증거로서 활용될 수 있다.

**ABSTRACT**

Web records is valuable information to preserve, because it can be used as a legal evidence about business or e-commerce of a public institution, but it is easily disappeared because of its volatile characteristic. Therefore, archival information package should be defined for long-term preservation. Web records can be stored in the archival information package for electronic records, because web records is a kind of electronic records. However, the NEO(NARS Encapsulation Object), the archival information package for electronic records in Korea, can't able to store web records, because it was developed without consideration of the characteristic of web records. In this paper, we define extended NEO based on the analysis of KoSurWeb and KoDeWeb, that archival information package for document of surface and deep web as well as the NEO. Web records can be preserved and utilized along with electronic records by using the extended NEO. Also it can be used for record and legal evudence by archiving web records of public institution about e-commerce.

---

본 연구는 행정안전부 국가기록원의 지원을 받아 기록물 보존기술 연구개발(R&D) 사업의 일환으로 이루어졌으며, 이게 감사드린다.

* 충남대학교 공과대학 컴퓨터 공학과

** 충남대학교 공과대학 컴퓨터 공학과

*** 교신저자, 충남대학교 공과대학 컴퓨터 공학과

2010년 09월 27일 접수, 2010년 10월 17일 심사완료 후 2010년 11월 05일 게재확정.

**키워드** : 웹기록물, 장기보존, 전자기록물, 장기보존포맷, 전자상거래
Web Records, Long-Term Preservation, Electronic Records, Archival Information Package, E-Commerce

# 1. 서 론

1989년 3월, 유럽의 소프트웨어 공학자 팀 버너스 리(Tim Berners-Lee)의 제안으로 웹이 시작되었다. 웹이 발전함에 따라 웹 사이트에서 다양한 영역의 자료나 프로그램 등을 얻을 수 있게 되었고, 문화 공간으로 활용되거나 웹을 통해 전자상거래를 하는 등 점차 웹이 다루고 있는 데이터의 비중이 높아져가고 있다.

공공기관도 초기에는 웹을 단순한 기관의 홍보를 위한 목적으로 사용하였다. 그러나 이러한 흐름에 맞춰 행정업무의 일부를 웹 사이트에서 처리하고, 기업과 정부 기관(B2G : Business-to-Government)과의 전자상거래를 활성화하여 여러 기관의 각종 신청서 제공, 세무 양식 제공, 공과금 납부와 같은 업무를 웹에서 지원하고 기업이 정부 기관의 구매 요건을 파악한 뒤 제안서를 제출할 수 있도록 활용하고 있다.

웹기록물이란 이와 같이 공공기관의 웹사이트에 포함된 정보 뿐만 아니라 전자상거래 시 웹상에서 처리한 업무의 결과로서 보유되고 업무과정에서 생산 접수되는 정보이다. 이러한 정보들은 생산 및 접수과정에 대한 법적 증거일 뿐만 아니라 그 내용 자체가 정보가치가 있기 때문에 이를 유지 관리 및 보존해야 한다[1].

웹이 접근단위에 따라 표면 웹과 심층 웹으로 분류되는 것과 마찬가지로 웹기록물도 표면 웹기록물과 심층 웹기록물로 분류된다[2]. 표면 웹기록물은 접근할 때마다 동일하게 표현되는 정적인 문서들로 구성되어 있으며, 브라우저를 통해 사용자들이 접근할 수 있다. 심층 웹기록물은 일반적으로 사용자들의 접근이 불가능한 데이터베이스를 통해 구성되며, 사용자의 요구가 변경될 때마다 내용이 갱신된다. 이전 연구[3, 4, 5]에서는 이러한 웹기록물의 분류에 따라 표면 웹기록물 문서보존포맷인 KoSurWeb(Korea Surface Web)과 심층 웹기록물 문서보존포맷인 KoDeWeb(Korea Deep Web)을 정의하였다.

이러한 웹기록물들은 웹을 통해 접근되고 관리되기 때문에 웹과 같은 특성을 가진다. 그 중 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[6, 7]. 특히 '휘발성' 특징을 가진 웹기록물은 생성과 삭제가 빈번하게 이루어지기 때문에 보존의 가치가 있는 자원이지만 많은 양의 업무 기록 및 전자상거래에 대한 법적 증거들이 소실되고 있다[8]. 따라서 본 논문의 목적은 이와 같

이 소실되는 웹기록물을 보존하기 위한 웹기록물 장기보존포맷을 설계하는 것이다.

국가기록원에서 전자기록물 보존을 위해 표준으로 제시한 전자기록물 장기보존포맷 기술규격(NARS : National ARchives Standard)[9]은 ISO 15489와 VERS(Victorian Electronic Records Strategy), CEDARS(CURL Exemplars in Digital ARchives), OCLC and PLG (Online Computer Library Center and Research Library Group), 그리고 우리나라의 메타데이터의 내용을 기반으로 구성된 전자기록물 장기보존포맷이다[10]. VERS에서 제공하는 형식에는 이미 웹기록물을 수용되어 있기 때문에 전자기록물 장기보존포맷에서도 웹기록물 문서보존포맷을 수용 가능해야 한다. 하지만 기존의 전자기록물 장기보존포맷은 웹기록물의 내용이 가지는 특수성과 웹기록물의 보존 및 복원에 관련된 정보들을 정의하지 않았다.

웹기록물은 웹상에 존재하는 정보들이기 때문에 그 내용 역시 웹과 관련되어 있다. 데이터베이스를 대상으로 한 심층 웹기록물 문서보존포맷을 전자기록물 장기보존포맷에 저장할 경우, 보존 대상이 데이터베이스에 관련된 문서보존포맷임을 명시할 수 있는 정보가 기존의 전자기록물 장기보존포맷에는 포함되어 있지 않다. 또한 웹기록물을 보존 및 복원하기 위해 필요한 정보들 역시 웹과 관련되어 있는데 기존의 전자기록물 장기보존포맷에는 이러한 정보들을 저장할 수 있는 메타데이터들이 정의되지 않았다.

이러한 문제점을 해결하기 위해서 본 논문에서는 전자기록물 장기보존포맷이 웹기록물에 관련된 내용을 명시할 수 있도록, 기존의 전자기록물 장기보존포맷의 메타데이터의 범위를 확장시키고, 웹기록물의 보존 및 복원에 관련된 정보들을 저장할 수 있도록 기술 메타데이터 항목을 추가시켰다.

본 논문의 구성은 다음과 같다. 제 2장에서는 관련 연구로 표면 웹기록물 문서보존포맷인 KoSurWeb과 심층 웹기록물 문서보존포맷인 KoDeWeb에 대해서 설명한다. 제 3장에서는 전자기록물 장기보존포맷에 대해 설명하고 제 4장에서는 전자기록물 장기보존포맷을 확장한 보존포맷에 대해서 설명한다. 제 5장에서는 공공기관의 웹 사이트를 대상으로 시행한 적용시험의 과정과 결과를 바탕으로 확장된 전자기록물 장기보존포맷에 대해 평가하며, 제 6장에서는 본 논문의 결론 및 기대성과에 대해 설명한다.

## 2. 관련 연구

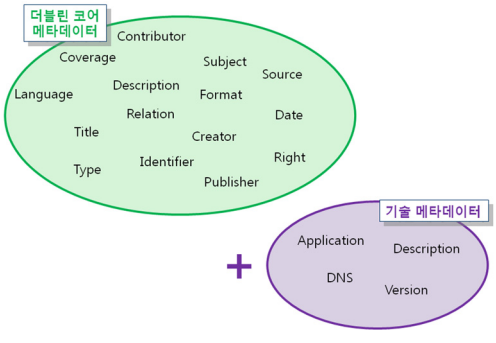
관련 연구에서는 전자기록물 장기보존포맷이 웹기록물을 보존할 수 있도록 메타데이터와 컨텐츠를 확장하기 위해서 웹기록물 문서보존포맷의 메타데이터와 컨텐츠를 살펴보았다. 웹기록물을 보존하기 위한 문서보존포맷으로는 표면 웹기록물을 보존하기 위한 문서보존포맷인 KoSurWeb과 심층 웹기록물을 보존하기 위한 문서보존포맷인 KoDeWeb이 있다.

### 2.2 KoSurWeb

KoSurWeb(Korea Surface Web)은 표면 웹기록물에 대한 문서보존포맷이다. KoSurWeb은 더블린코어(Dublin Core)[11]의 메타데이터를 기반으로 정의되었기 때문에 국내외적으로

호환성을 가지고 있으며 XML로 구성되어 있어 소프트웨어나 하드웨어에 독립적이고 개방적이다.

KoSurWeb은 표면 웹기록물에 대한 메타데이터들과 표면 웹기록물의 실제 내용에 해당하는 컨텐츠로 구성되어 있다.




〈그림 1〉 KoSurWeb의 메타데이터

KoSurWeb의 메타데이터는 <그림 1>과 같이 더블린코어를 기반으로 정의한 더블린 코어 메타데이터와 표면 웹기록물에 접근하고 보존/복원하는데 필요한 기술적 항목을 기반으로 정의한 기술 메타데이터로 구성되어 있다.

더블린 코어 메타데이터는 'Identifier', 'Format', 'Source', 'Title', 'Creator', 'Subject', 'Description', 'Relation', 'Type', 'Publisher', 'Date', 'Right', 'Contributor', 'Coverage', 'Language'의 15가지의 메타데이터로 구성되어 있으며 표면 웹기록물을 이해, 보존, 관리하기 위해 필요하다. 기술 메타데이터는 'Database', 'Application', 'DNS', 'Description'의 4가지 메타데이터 구성되어 있으며 표면 웹기록물에 접근하고 보존/복원하기 위해 필요하다.

KoSurWeb의 컨텐츠는 표면 웹기록물의 실제 내용에 해당하는 정보들을 WARC(Web ARChive file format)[12] 파일 포맷으로 저장한다. WARC 파일 포맷은 IIPC에서 제정한 웹 자원을 구조화하고 관리하는 표준화된 형식으로 표면 웹기록물에 대한 내용들을 설정한 용량의 크기에 따라 나누어 저장한다.

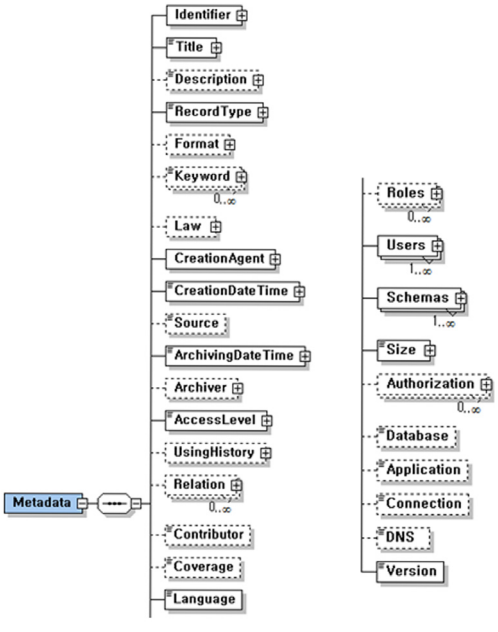
## 2.3 KoDeWeb

KoDeWeb(Korea Deep Web)은 심층 웹기록물에 대한 문서보존포맷이다. KoDeWeb의 메타데이터는 데이터베이스에서 추출한 심층 웹기록물의 메타데이터와 컨텐츠를 저장하기 위해서 SIARD(Software Independent Archival of Relation Database)[13]의 메타데이터를 참조하고, 국내의 심층 웹기록물을 분석하여 정의되었다. KoDeWeb은 KoSurWeb과 마찬가지로 XML로 구성되기 때문에, 소프트웨어나 하드웨어에 독립적이며, 하나의 포맷으로 구성되어 있기 때문에 문서보존포맷의 전송 및 보존과정에서 손실이 적다는 장점이 있다.

KoDeWeb은 심층 웹기록물에 대한 메타데이터 부분과 데이터베이스의 실제 내용에 해당하는 컨텐츠 부분으로 구성되어 있다. <그림 2>는 KoDeWeb의 메타데이터에 대한 스키마를 도식화 한 것이다.

'Identifier', 'Title', 'Description', 'Record Type', 'Format', 'Keyword', 'Law', 'Creation Agent', 'CreationDateTime', 'Source', 'ArchivingDateTime', 'Archiver', 'Access Level', 'UsingHistory', 'Relation' 메타데이터는 KoDeWeb을 검색, 이해, 보존, 관리하는데 필요한 메타데이터이다. 그리고 'Contributor', 'Coverage', 'Language', 'Roles', 'Users',

'Schemas', 'Size', 'Authorization' 메타데이터는 심층 웹기록물을 검색, 이해, 보존, 관리하는데 필요한 메타데이터들이다. 'Database', 'Application', 'Connection', 'DNS', 'Version' 메타데이터는 데이터베이스에 대한 기술(Technical)항목을 기반으로 정의한 메타데이터들로, 심층 웹기록물이 구성된 환경에 대한 이해와 보존/복원을 위해 사용되는 메타데이터이다.




〈그림 2〉 KoDeWeb의 메타데이터 스키마

KoDeWeb의 컨텐츠는 데이터베이스의 실제 내용을 저장하는 요소이다. 데이터베이스의 테이블 단위로 구성된 XML 파일에 데이터베이스의 테이블에 저장된 실제 내용을 행(Row)별로 저장하고, XML 파일들을 ZIP64 파일 포맷으로 압축하여 하나의 파일로 저장된다.

## 3. 전자기록물 장기보존포맷 기술 규격

전자기록물 장기보존포맷 기술 규격[7]은 전자기록물의 진본성과 무결성을 보장하는 포맷이다. 전자기록물 장기보존포맷은 메타데이터와 기록물의 구조상의 대상 범위에 따라서 2가지 포맷으로 분류되는데, 하나는 보존할 기록물을 전자기록물로 생성한 기록물건 장기보존포맷과 관련 있는 기록물건 장기보존포맷의 집합인 기록물철 장기보존포맷으로 나누어진다.

기록물건 장기보존포맷은 특정사안을 구성하는 문서에 대한 것으로, 웹기록물에서는 한번 수집된 표면/심층 웹기록물이 이에 해당한다. 기록물철 장기보존포맷은 특정 사안에 관련된 기록물건 장기보존포맷들을 모아 정리한 것으로 웹기록물인 경우 정책에 의해 정해진 특정 기간 안에 수집된 모든 표면/심층 웹기록물이 이에 해당한다. 본 논문에서는 한번 수집된 표면/심층 웹기록물을 저장할 수 있는 기록물건 장기보존포맷을 대상으로 메타데이터와 컨텐츠를 확장하였다.

전자기록물은 매체의 특성상 변형, 훼손, 유실이 되기 쉽다 하지만 오랜 기간이 경과해도 기록물이 생산된 당시에 가지고 있던 내용을 그대로 재현하여 접근할 수 있도록 보존되어야 업무활동에 대한 증거 및 업무에 대한 법적증거로서 효력을 갖는다. 이를 위해서 기록물건 장기보존포맷은 <그림 3>과 같이 기록물건 메타데이터, 문서보존포맷, 전자기록물 원문, 전자서명으로 구성되어 있다.

기록물건 메타데이터는 전자기록물을 유지하고 이해하는데 필요한 정보들을 포함하는 메

타데이터이다. 기록물건 메타데이터는 AgentCon, MandateCon, IdentifierCon, TitleCon, DescriptionCon, StorageCon, ClassificationCon, IndexCon, CreationCon, PreservationCon, TranscationCon, RightManagementCon, ManagementHistoryCon, UseHistoryCon, RelationCon 등 총 15개의 상위요소를 가진다.

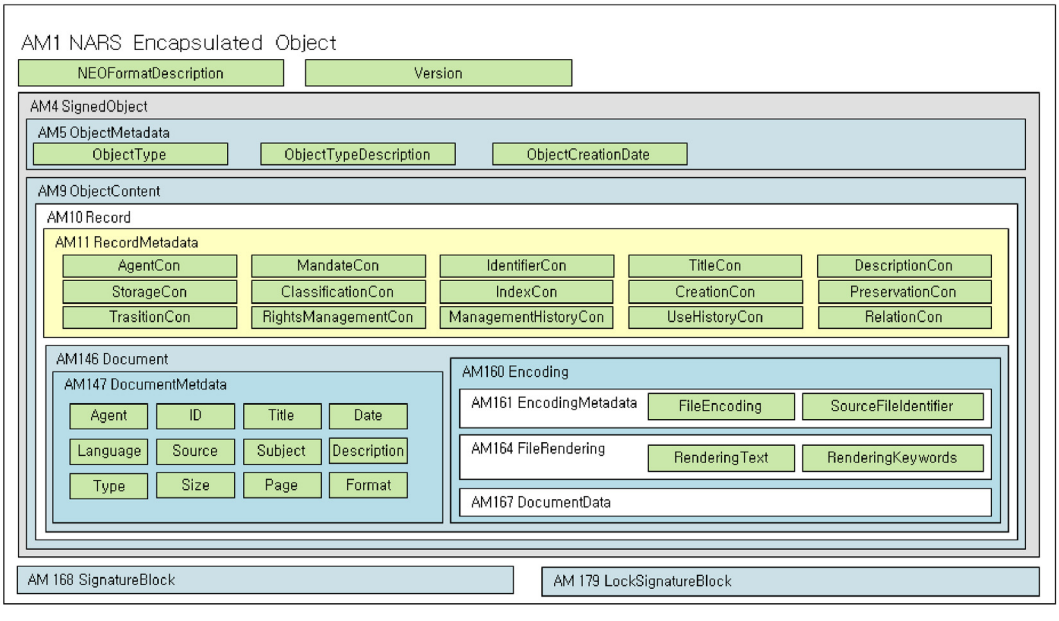
문서보존포맷은 문서가 생산된 당시의 애플리케이션이 없어도 해당문서의 내용과 외형을 그대로 재현하여 내용보기를 가능하게 하는 포맷이다. 문서보존포맷은 전자기록물 장기보존포맷의 'AM146 Document'와 그 하위 요소들에 저장된다.

'Document' 메타데이터는 하위 요소로 'Document Metadata' 메타데이터와 'Encoding' 메타데이터를 가진다. 'Document Metadata' 메타데이터는 'Agent', 'ID', 'Title', 'Date', 'Language', 'Source', 'Subject', 'Description', 'Type', 'Page', 'Format' 등 주로 전자기록물 원문을 이해, 관리 그리고 보존하는데 필요한 메타데이터들로 구성되어 있다. 'Encoding' 메타데이터는 실제 내용의 표현물과 인코딩된 기록물이 포함하는 내용을 어떻게 기술했는가에 대한 메타데이터들이 포함되어 있다. 여기에는 한 문서의 다양한 인코딩 방식을 포함할 수 있으며 인코딩 문서의 정보를 메타데이터와 함께 보존한다.

전자기록물 원문은 생산자가 처음으로 생산한 기록물로 진본성 보장을 위해 인코딩 메타데이터에서 정의한 인코딩 방법으로 전자기록물 장기보존포맷의 'DocumentData'에 저장된다.

전자서명은 전자기록물의 진본성 및 무결성 보장을 위해 포함되는데, 크게 인증정보와 잠김인증정보로 구성되어 있다.

전자기록물 장기보존포맷은 XML 형식의




〈그림 3〉 전자기록물 장기보존포맷의 구조

단일한 객체로 패키지화된다. 단일한 객체로 패키징은 진본성과 무결성을 유지해야 하는 전자기록물이 시스템과 기관의 전송 간에 일어날 수 있는 유실 및 훼손에 대한 위험성을 줄이고, 관리를 편리하게 한다.

## 4. 확장된 전자기록물 장기 보존포맷

전자기록물 장기보존포맷 기술 규격(NARS : National ARchives Standard)에는 VERS의 메타데이터의 내용을 기반으로 구성된 메타데이터들이 포함되고 있고, VERS에서 제공하는 형식에는 이미 웹기록물을 수용되어 있기 때문에 전자기록물 장기보존포맷에서도 웹기록물 문서 보존포맷을 수용 가능해야 한다. 하지만 기존의 전자기록물 장기보존포맷은 웹기록물의 내용이 가지는 특수성과 웹기록물의 보존 및 복원에 관련된 정보들을 정의하지 않았다. 따라서 본 논문에서는 전자기록물 장기보존포맷이 웹기록물 문서 보존포맷을 수용할 수 있도록 전자기록물 장기보존포맷의 메타데이터와 컨텐츠를 확장하여 설계하였다. <표 1>은 전자기록물 장기보존포맷의 일종인 기록물건 장기보존포맷의 메타데이터와 웹기록물 문서 보존포맷의 메타데이터를 비교 분석하여 전자기록물 장기보존포맷의 메타데이터를 확장하여 정의한 것이다.

KoDeWeb과 KoSurWeb은 각각 표면/심층 웹기록물의 내용과 외형을 보존/복원하기 위한 문서보존포맷으로, 앞에서 살펴본 기록물건 장기보존포맷 구성요소 중 문서보존포맷이 이에 해당된다.

따라서 기록물건 장기보존포맷이 웹기록물 문서보존포맷을 포함하기 위해서는 기록물건 장기보존포맷의 문서보존포맷이 웹기록물 문서보존포맷을 포함해야 한다. 이를 위해서 본 논문에서는 먼저 기록물건 장기보존포맷의 문서보존포맷과 웹기록물 문서보존포맷인 KoSurWeb과 KoDeWeb의 메타데이터를 먼저 살펴보았다.

문서 메타데이터는 기록물을 이해하고 보존, 관리하는데 필요한 메타데이터로 웹기록물 문서보존포맷에는 표면 웹기록물의 'Creator', 'Subject', 'Type' 메타데이터와 심층 웹기록물의 'CreationAgent', 'CreationDateTime', 'Keyword', 'RecordType', 'Size', 'Schemas', 'User', 'Roles', 'Authorization' 메타데이터, 그리고 표면 웹기록물과 심층 웹기록물에 공통적으로 포함되어 있는 'Identifier', 'Title', 'Language', 'Application', 'Description', 'Format', 'Version', 'Contributor', 'Coverage' 메타데이터가 여기에 속한다.

이 중 기록물 원본의 생산자에 관련된 정보인 'Creator'와 'CreationAgent'는 'Agent'에, 기록물의 고유식별자인 'Identifier'는 'ID'에, 기록물의 제목인 'Title'은 'Title'에 포함되고, 기록물 원본의 생산일시에 대한 정보인 'CreationDateTime'은 'Date'에, 웹기록물의 구성언어에 대한 정보인 'Language'는 'Language'에, 웹기록물을 생성한 도구에 대한 정보인 'Application'은 'Source'에 포함된다. 또한 기록물 검색에 사용되는 'Subject'와 'Keyword'는 'Subject'에, 기록물에 대한 설명 정보인 'Description'은 'Description'에, 기록물의 유형에 대한 설명인 'Type'와 'RecordType'은

〈표 1〉 확장된 전자기록물 장기보존포맷의 메타데이터를 정의

| KoSurWeb | KoDeWeb | 분류 | 전자기록물 장기보존포맷 | 확장된 전자기록물 장기보존포맷 |
|---|---|---|---|---|
| | | 기록물건 메타데이터 | AgentCon | AgentCon |
| | Law | | MandateCon | MandateCon |
| | | | IdentifierCon | IdentifierCon |
| | | | TitleCon | TitleCon |
| | | | DescriptionCon | DescriptionCon |
| | | | StorageCon | StorageCon |
| | | | ClassificationCon | ClassificationCon |
| | | | IndexCon | IndexCon |
| | Source | | CreationCon | CreationCon |
| Date | ArchivingDateTime | | PreservationCon | PreservationCon |
| Publisher | Archiver | | | |
| | | | TrascationCon | TrascationCon |
| Right | AccessLevel | | RightManagementCon | RightManagementCon |
| | | | ManagementHistoryCon | ManagementHistoryCon |
| | UsingHistory | | UseHistoryCon | UseHistoryCon |
| Relation | Relation | | RelationCon | RelationCon |
| Creator | CreationAgent | 문서 메타데이터 | Agent | Agent |
| Identifier | Identifier | | ID | ID |
| Title | Title | | Title | Title |
| | CreationDateTime | | Date | Date |
| Language | Language | | Language | Language |
| Application | Application | | Source | Source |
| Subject | Keyword | | Subject | Subject |
| Description | Description | | Description | Description |
| Type | RecordType | | Type | Type |
| | Size | | Size | Size |
| | | | Page | Page |
| Format | Format | | Format | Format |
| Version | Version | | | Version |
| Contributor | Contributor | | | Contributor |
| Coverage | Coverage | | | Coverage |
| | Schemas | | | Schemas |
| | User | | | Users |
| | Role | | | UserRoles |
| | Authorization | | | SystemRoles |
| | Connection | 기술 메타데이터 | | Connection |
| DNS | DNS | | | DNS |
| | Database | | | Database |

‘Type’에 포함되고, 웹기록물의 용량을 나타내는 정보인 ‘Size’는 ‘Size’에, 기록물의 저장매체에 대한 설명인 ‘Format’은 ‘Format’에 포함된다.

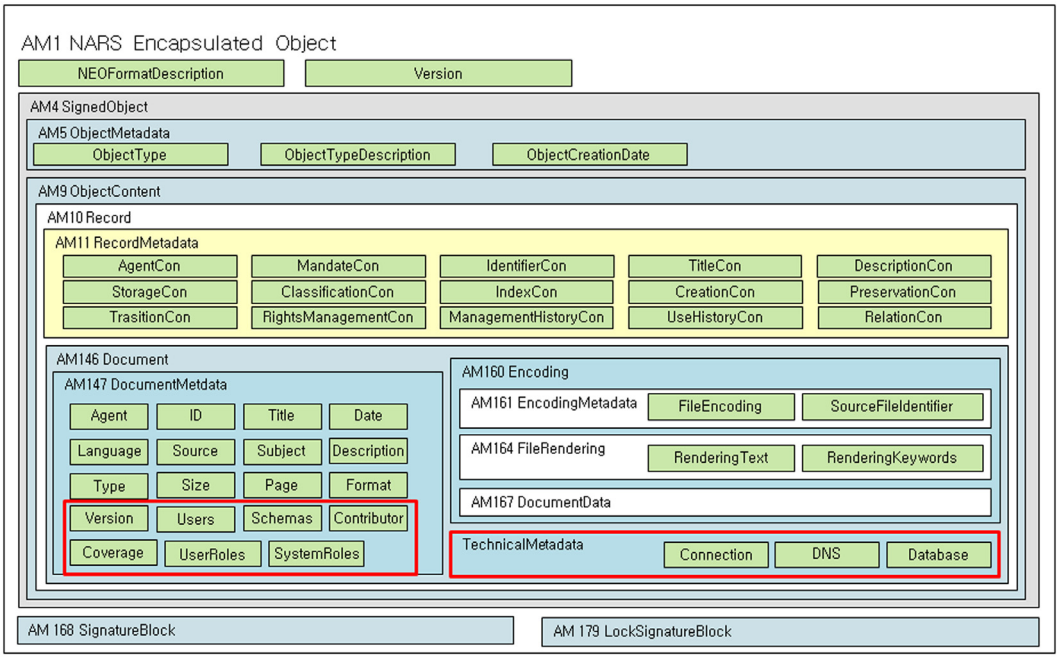
하지만 ‘Version’, ‘Contributor’, ‘Coverage’, ‘Schemas’, ‘User’, ‘Roles’, ‘Authorization’ 메타데이터는 웹기록물의 컨텐츠를 설명하기 위한 메타데이터들로 전자기록물 장기보존포맷에서 고려되지 않은 메타데이터이다. 따라서 해당 메타데이터를 기록물건 장기보존포맷에서 포함할 수 있도록 문서의 버전정보인 ‘Version’을 포함하기 위한 ‘Version’, 기여자 정보인 ‘Contributor’를 포함하기 위한 ‘Contributor’, 웹기록물의 수용 범위에 대한 정보인 ‘Coverage’를 포함하기 위한 ‘Coverage’, 데이터베이스의 스키마에 대한 정보인 ‘Schemas’를 포함하기 위한 ‘Schemas’, 데이터베이스 사용자에 대한 정보인 ‘User’를 포함시키기 위한 ‘Users’, 데이터베이스 사용자 권한에 대한 정보인 ‘Role’을 포함하기 위한 ‘UserRoles’ 메타데이터, 데이터베이스의 시스템 권한에 대한 정보인 ‘Authorization’를 포함하기 위한 ‘SystemRoles’ 메타데이터를 추가하였다.

기록물건 메타데이터는 기록물건 장기보존포맷을 이해, 보존, 관리하는데 필요한 메타데이터이다. 웹기록물 메타데이터 중 이러한 성격을 지니는 메타데이터로는 표면 웹기록물의 ‘Date’, ‘Publisher’, ‘Right’ 메타데이터와 심층 웹기록물의 ‘Law’, ‘Source’, ‘Archiving DateTime’, ‘Archiver’, ‘AccessLevel’, ‘Using History’ 메타데이터 그리고 표면 웹기록물과 심층 웹기록물에 공통적인 메타데이터인 ‘Relation’ 메타데이터가 있다.

기록물의 관련된 법규정보인 ‘Law’는 ‘MandateCon’에, 웹기록물의 출처에 대한 정보인 ‘Source’는 ‘CreationCon’에, 기록물의 보존을 수행한 일시에 대한 정보인 ‘Date’와 ‘Publisher’와 기록물의 보존을 수행한 행위자에 대한 정보인 ‘ArchivingDateTime’ 그리고 ‘Archiver’는 ‘PreservationCon’에 기록물 사용권한에 대한 정보인 ‘Right’와 ‘Access Level’은 ‘RightManagementCon’에 포함되고, 기록물 사용에 대한 정보인 ‘UsingHistory’는 ‘Use HistoryCon’에, 기록물과 다른 기록물간의 관련 정보인 ‘Relation’은 ‘RelationCon’에 포함된다.

웹기록물의 메타데이터 중 ‘DNS’ 그리고 심층 웹기록물의 ‘Connection’, ‘Database’는 웹기록물의 보존/복원에 필요한 기술적인 정보이다. 하지만 전자기록물 장기보존포맷에는 이러한 역할을 하는 메타데이터가 고려되지 않았다. 따라서 전자기록물 장기보존포맷이 이러한 메타데이터들을 포함할 수 있도록 상위 메타데이터로 ‘TechnicalMetadata’를 추가하고 기술 메타데이터의 하위 항목으로 데이터베이스 연결 정보인 ‘Connection’을 포함하기 위한 ‘Connection’, 아카이빙을 수행한 컴퓨터의 도메인 정보인 ‘DNS’를 포함하기 위한 ‘DNS’, 데이터베이스에 대한 제품 및 버전 정보인 ‘Database’를 포함하기 위한 ‘Database’ 메타데이터를 추가하였다.

표면 웹기록물은 컨텐츠를 WARC 파일 포맷으로 저장하며, 심층 웹기록물은 컨텐츠를 ZIP64 파일 포맷으로 압축된 파일로 저장한다. 하지만 전자기록물 장기보존포맷의 경우는 전송 및 보존에 따른 손실이 적게 하기 위해 컨텐츠를 단일화하여 ‘DocumentData’에 BIT Stream 형태로 저장한다. 따라서 WARC



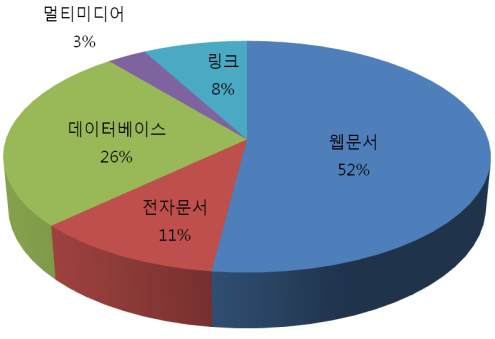

〈그림 4〉 확장된 전자기록물 장기보존포맷의 구조

파일포맷 형태로 저장된 KoSurWeb의 컨텐츠와 ZIP64 파일포맷 형태로 저장된 KoDeWeb의 컨텐츠를 전자기록물 장기보존포맷에서 포함할 수 있도록 Base64 데이터 포맷 형태로 인코딩하여 저장한다. <그림 4>는 확장된 전자기록물 장기보존포맷에 대한 구조로, 박스 안의 메타데이터가 웹기록물을 보존할 수 있도록 확장된 메타데이터들이다.

이렇게 확장한 전자기록물 장기보존포맷은 전자기록물 장기보존을 기반으로 하기 때문에, 국내 전자기록물의 표준을 준수하고, 기존의 전자기록물과도 호환이 가능하다. 또한 XML로 구성되어 있기 때문에 소프트웨어나 하드웨어에 독립적이며, 단일화된 패키지로 구성이 되어 있기 때문에 전송 및 보존에 따른 손실이 적다는 장점이 있다.

## 5. 확장한 전자기록물 장기보존포맷의 적용 시험

확장된 전자기록물 장기보존포맷이 웹기록물 문서보존포맷을 잘 수용할 수 있는지 확인하기 위해서 적용 시험 대상은 여러 가지 다양한 형태의 데이터들이 존재하고 있으며, 시스템적인 정보들 역시 다양하게 존재해야 한다. 여러 공공기관 중 국가기록원의 웹사이트의 경우, 다양한 형태의 컨텐츠를 포함하고 있고, 시스템적인 정보들 역시 풍부하여 이러한 조건을 충족한다. 또한 문서보존포맷으로 구성 시 다양한 메타데이터를 충족시킬 수 있기 때문에 적용 시험 대상으로 선정하였다. <그림 5>는 국가기록원 웹 사이트의 구성요소들을 도식화한 것이다.




〈그림 5〉 국가기록원 웹 사이트의 구성

‘웹문서’는 웹 사이트를 구성하는 일반적인 텍스트와 이미지로, HTML로 구성되어 있다. ‘전자문서’는 워드, 엑셀 등과 같은 문서 파일을 제공하는 페이지를 의미하며 ‘멀티미디어’는 배경음악 같은 사운드나 동영상을 제공하는 페이지를 의미하고, ‘데이터베이스’는 실제 데이터베이스를 통해 동적으로 생성되는 페이지를 의미한다. 이 중 표면 웹기록물은 접근 시마다 동일하게 표시되는 웹문서, 전자문서, 멀티미디어들로 이러한 정보는 표면 웹기록물 수집기를 통해 수집되어 표면 웹기록물 문서보존 포맷인 KoSurWeb으로 구성된다. 또한 심층 웹기록물의 정보들이 저장되어 있는 데이터베이스는 심층 웹기록물 수집기를 통해 수집되어 심층 웹기록물 문서보존 포맷인 KoDeWeb으로 구성된다.

본 논문에서는 이렇게 구성된 국가기록원의 KoSurWeb과 KoDeWeb을 확장 설계한 전자기록물 장기보존포맷에 적용하였다. <표 3>은 KoSurWeb의 메타데이터 일부를 발췌한 예시이고, <표 4>는 KoDeWeb의 메타데이터의 일부를 발췌한 예시이다. <표 5>는 <표 3>의 KoSurWeb의 메타데이터가 확장된 전자기록물 장기보존포맷에 저장된 모습이고, <표 6>은 <표 4>의 KoDeWeb의 메타데이터가 확장된 전자기록물 장기보존포맷에 저장된 모습이다.

WARC 파일 포맷으로 저장된 표면 웹기록물의 컨텐츠와 ZIP64 파일 포맷으로 저장된 심층 웹기록물의 컨텐츠는 Base64 데이터 포맷으로 인코딩되어 XML에 저장된다.

적용 시험 결과 웹기록물 문서보존 포맷의 메타데이터 및 컨텐츠가 확장된 전자기록물 장기보존포맷에 손실 없이 저장된 것을 확인하였다.

〈표 3〉 KoSurWeb의 메타데이터의 일부

```
<Title>국가기록원 웹사이트</Title>
<Type>웹기록물</Type>
<Creator>홍길동</Creator>
<Subject>국가기록원, 표면 웹기록물</Subject>
<Description>2009년 국가기록원의 웹사이트에 대한 표면 웹기록물</Description>
...
<Language>EUCKR</Language>
<DNS>10.XXX.XX.XX</DNS>
<Version>KoSurWeb v1.0</Version>
...
```

〈표 4〉 KoDeWeb의 메타데이터의 일부

```
<Title>국가기록원 웹사이트 데이터베이스</Title>
...
<RecordType>웹기록물</RecordType>
...
<CreationDateTime>2006-05-27T18:30:00</CreationDateTime>
...
<Language>KO16MSWIN949</Language>
<Size>3914424</Size>
...
<Database>Oracle9i Enterprise Edition Release 9.2.0.1.0-Production</Database>
<Application>KoDeWeb v1.0 </Application>
<Connection>jdbc:oracle:thin@localhost:1521:DBArchive </Connection>
<DNS>10.XXX.XX.XX</DNS>
<Version>KoDeWeb v1.0 </Version>
```

〈표 5〉 확장된 전자기록물 장기보존포맷에 저장된 KoSurWeb

```
<Agent>홍길동</Agent>
<Title>국가기록원 웹사이트</Title>
<Type>웹기록물</Type>
<Subject>국가기록원, 표면 웹기록물</Subject>
<Description>2009년 국가기록원의 웹사이트에 대한 표면 웹기록물</Description>
...
<Language>EUCKR</Language>
<DNS>10.XXX.XX.XX</DNS>
<Version>KoSurWeb v1.0</Version>
```

〈표 6〉 확장된 전자기록물 장기보존포맷에 저장된 KoDeWeb

```
<Title>국가기록원 웹사이트 데이터베이스</Title>
<Date>2006-05-27T08:00:00</Date>
<Language>KO16MSWIN949</Language>
<Source>KoDeWeb v1.0</Source>
<Type>웹기록물</Type>
<Size>3914424</Size>
<Pages>0</Pages>
<Version>KoDeWeb v1.0</Version>
<TechnicalMetadata>
    <Connection>jdbc:oracle:thin@localhost:1521:DBArchive</Connection>
    <DNS>10.XXX.XX.XX</DNS>
    <Database>Oracle9i Enterprise Edition Release 9.2.0.1.0-Production</Database>
</TechnicalMetadata>
```

## 6. 결 론

본 논문에서는 보존의 가치가 있지만 자체의 특성으로 인하여 소실되고 있는 웹기록물을 보존하기 위해서 KoSurWeb과 KoDeWeb의 메타데이터들을 분석하고 그리고 이를 바탕으로 기존의 전자기록물 보존포맷인 전자기록물 장기보존포맷을 확장하여 웹기록물의 보존에 적합한 확장된 전자기록물 장기보존포맷을 정의하였다.

또한 확장 설계한 전자기록물 장기보존포맷의 적용성을 확인하기 위해서 국가기록원의 웹기록물을 대상으로 확장된 전자기록물 장기보존포맷에 웹기록물 문서보존포맷을 보존하였다. 그 결과 확장 설계한 전자기록물 장기보존포맷이 웹기록물 문서보존포맷을 온전히 보존할 수 있다는 것을 확인하였다.

확장 설계한 전자기록물 장기보존포맷은 전자기록물 장기보존을 기반으로 하기 때문에, 국내 전자기록물의 표준을 준수하고, 기존의 전자기록물과도 호환이 가능하다. 또한 XML로 구성되어 있기 때문에 소프트웨어나 하드웨어에 독립적이며, 단일화된 패키지로 구성이 되어 있기 때문에 전송 및 보존에 따른 손실이 적다는 장점이 있다. 뿐만 아니라 전자 상거래에 대한 공공기관의 웹기록물을 보존함으로써 전자 상거래에 대한 기록 및 법적 증거로서 장기간 보존되고 활용될 수 있고, 나아가 웹 정보검색 기술의 활용 및 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] 차승준, 이규철, “Extension of the NARS Encapsulated Object to Accommodate Surface Web Records in Public Sector”, 한국정보과학회 데이터베이스 소사이어티, The Second International Conference on Emerging Database, 2010, pp. 184-188.

[3] 차승준, 이규철, “공공기관 심층 웹기록물 아카이빙을 위한 메타데이터 설계”, 한국전자거래학회지, 제14권, 제4호, 2009.

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

[5] 차승준, 이규철, “웹기록물 아카이빙 기반기술 연구 개발”, 지식정보산업연학학회 창립기념 학술대회, 2009, pp. 359-368.

[6] 유효림, “정부부처의 웹 아카이빙 방안 연구”, 명지대학교 석사학위논문, 2007.

[7] 이지은, “공공기관의 웹기록 관리방안 연구”, 한국 외국어대학교 석사학위논문, 2006.

[8] Adrian B., “Archiving Website : a practical guide for information management professionals,” facet publishing, 2006.

[9] 행정안전부 국가기록원, “전자기록물 장기보존포맷 기술규격(Standard of Archival Information Package)”, 2008.

[10] 이규철, 황윤영, 임혁수, “국가 전자기록물 영구보존을 위한 보존 메타데이터 설계”, 국가기록원, 기록보존 제18권, 2005.

[11] Dublin Core, http://www.dublincore.go.kr/documents/dcmi-terms/.

[12] J. Kunze, A. Arvidson, and G. Mohr, “The WARC File Format(Version 0.16),” IIPC Framework Working Group, 2007, p. 3.

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

## 저 자 소 개


박병주 (E-mail : dcrein@naver.com)
2009년 충남대학교 공과대학 컴퓨터공학과 (학사)
2009년～현재 충남대학교 공과대학 컴퓨터공학과 석사과정 재학중
관심분야 데이터베이스, 웹 아카이빙

차승준 (E-mail : junii@cnu.ac.kr)
2006년 충남대학교 공과대학 컴퓨터공학과 (학사)
2006년～현재 충남대학교 공과대학 컴퓨터공학과 석박사통합과정 재학중
관심분야 데이터베이스, 웹 서비스, GIS, 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 초빙연구원
1995년～1996년 미국 Syracuse University 초빙교수
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