논문

대량기록물 처리를 위한 영구기록물관리시스템의 디지털저장소 배치형상 연구

임 진 희* · 이 대 욱**

* 명지대 디지털아카이빙연구소 연구실장.
** 서강대 컴퓨터공학과 연구교수.

[국문초록]

2015년부터 국가기록원은 영구기록물관리기관으로서 다수의 기관으로부터 대량의 디지털 기록정보를 이관받게 된다. 이 논문에서는 영구기록물관리기관인 국가기록원이 2015년 이후를 대비하기 위해서 해결해야 할 여러 과제 중 영구기록물관리시스템에 필요한 디지털저장소의 배치형상에 대해 제안하고 있다. 논의의 순서는 영구기록관리기관에서 기록정보를 처리하는 절차를 검토하고, 각 단계에서 처리할 기록정보의 양을 추산한 후, 단계별 필요한 디지털저장소를 배치하는 것이다.

첫째, 영구기록물을 관리하는 단계를 입수단, 보존단, 제공단의 3단으로 구분하여 각 단계별로 디지털 기록정보의 흐름과 처리내역을 살펴봄으로써 최소한 3단계별로 영구기록물관리시스템이 배치와 형상 면에서 구별되어 구축되어야 한다고 제안한다. 둘째, 계산 결과 2015년에는 약 2.5테라바이트 정도의 디지털 기록정보가 입수될 것으로 추정되었다. '대량기록물'관리 과제의 특성을 점검하고, 데이터베이스로 관리할 디지털객체와 파일시스템으로 관리할 디지털객체의 종류를 정하고, 기록시스템에서 대량처리가 필요한 업무를 도출하여 대량처리 업무를 효율적으로 집행하기 위한 데이터베이스 설계 방향을 제안하고 있다. 셋째, 입수단, 보존단, 제공단별로 처리의 특성을 감안하여 디지털저장소의 개수를 달리하여 배치하도록 제안하고 있다. 이 논문은 2015년도를 대비하여 영구기록물관리시스템의 디지털저장소를 어떻게 준비해야 하는지 구체적이고 상세하게 논의할 수 있는 출발점을 제공한다.

주제어 : 대량기록물 이관, 영구기록물관리시스템, 디지털저장소, 배치형상, 전자기록

1. 머리말

2015년은 우리나라 공공기록물관리 분야에 매우 중요한 해이다. 국가기록원이 영구기록물관리기관으로서 기관들로부터 장기보존 대상 디지털 기록정보를 대량으로 이관받아 관리하게 되는 원년이기 때문이다. 참여정부 전자정부 31대 과제 중 하나인 '문서처리 전 과정의 전자화'의 일환으로 2004년부터 각급 기관들은 신전자문서시스템을 도입하여 공문서를 전자적 방식으로 생산하기 시작했으며, 기록관리법령에 따라 2004년도 생산기록이 2015년에 국가기록원으로 이관될 예정이다. 국가기록원은 각 기관으로부터 매년 기록물을 이관받아왔으나 디지털 기록정보를 대량으로 이관하는 일은 새롭게 시작되는 업무이므로 절차를 상세히 설계하고 디지털 환경의 특징을 고려하여 2015년을 꼼꼼히 준비해야 한다.

국가기록원은 2015년 이후를 대비하면서 전자기록관리 전략개발 태스크포스팀을 중심으로 전자기록 관리 정책과 프로세스, 시스템에 관한 전반적인 여건을 점검하고 있다. 또한, 보존복원연구과를 중심으로 장기적인 과제로 차세대 전자기록관리를 위한 프레임워크를 만들어가고 있다. 일차적으로는 대량의 디지털 기록정보를 효과적으로 입수하는 체계를 완비하는 것이 시급한 과제이나 입수 이후에 디지털 기록정보를 안정적으로 장기보존하면서 원활하게 내외부로 제공할 수 있는 체계를 준비하는 것 역시 중대한 과제가 될 것이다.

---

1) 국가기록원에서는 2015년부터 시작되는 대량의 전자기록물 이관 및 관리에 대응하기 위하여 "차세대 전자기록관리 기술기반 연구"를 다년간 사업으로 진행하고 있다. 이 논문은 2011년 명지대 디지털아카이빙연구소가 주관연구기관이 되어 (주)앱토스, (주)세미콘네트웍스와 공동수행한 "차세대 전자기록관리 인프라 응용기술 연구개발사업"(이하 "차세대프로젝트") 중 제1세부과제의 내용을 기반으로 하여 수정·보완한 것이다.

2015년 이후를 대비하기 위해서는 정책과 절차 개발, 각종 도구 개발, 데이터베이스 설계, 소프트웨어와 하드웨어 구축, 네트워크 개선, 전문 IT 인력 확충, 예산 확보 등 여러 측면에서의 준비가 필요하다. 특히, 디지털 기록정보의 이관은 각급 기록관리기관과 영구기록물관리 기관 간 협업으로 처리해야 하는 과제이기 때문에 양측 모두의 준비가 필요하다. 국가기록원 입장에서는 먼저 2015년부터 매년 이관되어올 디지털 기록정보의 양이 얼마나 될지 계산해보고, 품질은 어떠한지 검증해보고, 기록정보의 양과 품질에 맞춰 입수 세부절차를 정하고, 각 단계별 처리에 필요한 자동화도구를 개발하는 것이 주요 단기과제가 될 것이다. 각급 기록관리기관에서도 이관대상 디지털 기록정보를 자체 검사하여 정확한 양과 품질을 확인한 후 이관 규격과 절차에 맞게 디지털 기록정보를 준비해나가야 할 것이다.

이 논문에서는 영구기록물관리기관인 국가기록원이 2015년 이후를 대비하기 위해서 해결해야 할 여러 과제 중 영구기록물관리시스템에 필요한 디지털저장소(Digital Repository)[2]의 배치(Deployment)와 형상(Configuration)에 대해 제안하고자 한다. 영구기록물관리시스템에 입력되고 관리되며 출력되는 전체 디지털 기록정보는 임시로 혹은 장기간 디지털저장소에 저장된다. 이때, 기록정보의 하위요소인 메타데이터와 디지털컴포넌트 별로 어떤 정보구조로 어떤 저장소에 보관되어야 할지 살펴보는 일은 2015년을 대비하는 구체적인 작업이 될 것이다. 여기서는 기록정보 디지털객체를 데이터베이스로 구축하여 관리해야 하는 메타데이터 및 기타 데이터류와 파일시스템으로 관리해야 하는 디지털컴포넌트로 나누어 저장소를 몇 개로 구성하는 게 효과적, 효율적일지 제

2) 이 논문에서의 디지털저장소는 디지털객체를 입수, 관리, 보존, 제공하는 절차와 시스템, 조직과 인력을 갖춘 큰 범주로서의 정의가 아니라 정보시스템의 일부 구성요소로서 디지털 정보를 저장하는 정보구조와 저장매체, 그리고 이를 직접적으로 관리하는 파일시스템과 데이터베이스관리시스템을 의미하는 협의의 개념으로 정의하고자 한다.

안하고자 한다. 이 논문에서 제안하는 디지털저장소의 설계방향은 향후 영구기록물관리시스템 데이터베이스관리시스템과 저장매체의 형상과 배치를 결정하는데 활용될 수 있을 것이다.

논문의 내용은 영구기록물관리 단계의 기록정보 처리절차를 검토하고, 처리할 기록정보의 양을 추산한 후, 기록정보의 저장소를 배치하는 순서로 구성하였다. 2장에서는 영구기록물을 관리하는 단계를 입수단, 보존단, 제공단의 3단으로 구분하고, 각 단계별로 디지털 기록정보의 흐름과 처리내역을 살펴보면서, 3단별 특성에 기반하여 디지털저장소의 배치와 형상이 달리 설계되어야 한다는 점을 설명하고자 한다. 3장에서는 2015년 이후 국가기록원이 처리하게 될 디지털 기록정보의 양을 추정해봄으로써 '대량기록물'관리 과제를 양적 측면에서 점검하고, 데이터베이스로 관리할 디지털객체와 파일시스템으로 관리할 디지털객체의 종류를 정하고, 기록시스템의 대량처리 요건을 효율적으로 집행하기 위한 데이터베이스 설계 방향을 논의하고자 한다. 4장에서는 3단별로 디지털객체의 양과 처리내역을 고려하여 저장소의 개수와 종류를 정해 배치안을 제시하고자 한다.

2. 영구기록물관리 단계별 기록정보 처리절차와 특징

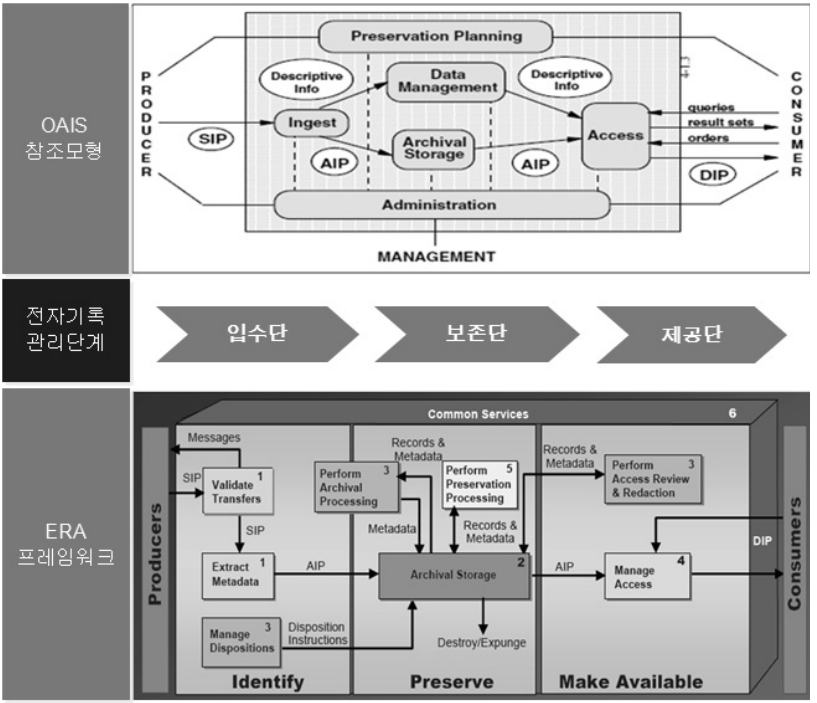
1) 영구기록물관리의 생애주기 단계 구성

영구기록물관리기관의 디지털 기록정보는 생애주기 동안 여러 단계를 거치면서 관리되며 전 단계에서 공통적으로 기록의 진본성, 신뢰성, 무결성, 이용가능성을 보증하는 방식으로 처리되어야 한다. 영구기록물관리시스템에서 기록정보를 처리하는 단위의 관점에서 살펴보면, 입수

단계에서는 주로 기관에서 이송된 디지털 기록정보 단위로 처리하게 되며, 보존단계에서는 주로 저장매체나 기록정보 유형별로 처리하게 되며, 제공단계에서는 주로 이용자가 검색, 열람 요청하는 기록 건이나 철, 건의 묶음 단위로 처리하게 되는 특성을 갖게 된다. 이처럼 단계별로 기본적인 처리단위와 관리기능이 다른 경우 기록정보의 품질을 보증하면서 처리의 효율을 높이기 위해서 단계별로 별도의 서브시스템을 구성하는 것이 필요하다. 이때, 각 서브시스템마다 디지털저장소의 종류와 개수 또한 다르게 설계될 것이다.

디지털저장소의 배치와 형상을 달리 설계할 필요가 있는 영구기록관리의 단계를 생애주기 관점에서 살펴보기 위해 〈그림 1〉과 같이 ISO14721의 OAIS 참조모형과 NARA의 ERA프레임워크를 상호비교해 보았다. OAIS 참조모형은 디지털객체의 장기보존 시스템 기능을 6개로 구분하고 있는데, 보존계획(Preservation Planning)과 운영(Administration) 기능은 디지털객체의 전체 생애주기에 걸쳐진 기능이므로 논의에서 제외하고 살펴본다면, 입수정보패키지(SIP)를 대상으로 하는 입수(Ingest) 기능, 보존정보패키지(AIP)를 대상으로 하는 데이터관리(Data Management)와 영구저장(Archival Storage) 기능, 배부정보패키지(DIP)를 대상으로 하는 접근(Access) 기능 등 3개의 서로 다른 단계로 나눠볼 수 있다. ERA프레임워크에서는 공통서비스(Common Services) 기능을 논의에서 제외하고 살펴본다면, 관리 대상을 찾아내고(Identify) 이를 보존처리하여 저장하고(Preserve), 이를 가공하여 이용자가 이용할 수 있도록(Make Available) 해주는 3개의 단계로 구성되어 있다. 두 개 모형의 공통점을 참고하여 이 논문의 논의에 유용하도록 영구기록물관리기관의 기록정보 생애주기 단계를 도출해보면 다음과 같은 3단 구성이 가능해진다.

〈그림 1〉 전자기록 관리의 3단계



첫째, 입수단(Ingest)이다. ISO14721 OAIS 참조모형의 입수 기능과 NARA의 ERA프레임워크의 식별영역에 해당하는 단계이다. 둘째, 보존단(Preservation)이다. OAIS 참조모형의 데이터관리와 영구저장 기능, ERA 프레임워크의 보존영역에 해당하는 단계이다. 셋째, 제공단(Access)이다. OAIS 참조모형의 접근 기능과 ERA프레임워크의 이용 영역에 해당하는 단계이다.

위에서 정의한 3단별로 처리하는 기록정보의 단위와 관리기능이 다르고 양적 특성이 다르므로 디지털저장소의 구성방식도 달라져야 할 것이다. 한편, 효과적이고 효율적인 디지털저장소의 배치형상을 설계하

기 위해서는 각 단계별 처리절차를 상세히 살펴보면서 요건과 고려사항을 도출할 필요가 있다. 요건과 고려사항을 도출하기 위해 디지털 기록정보의 입수에서 제공에 이르는 주요 생애주기에 걸쳐 저장소에 임시 혹은 장기적으로 저장되어야 할 디지털객체와 이에 대한 처리과정을 살펴보고자 한다. 구체적인 논의를 위해 국가기록원의 영구기록 관리절차를 예시로 살펴보고자 한다.[3)]

2) 입수단(Ingest): SIP의 양과 품질

국가기록원의 입수단 프로세스를 디지털 기록정보 처리흐름을 중심으로 살펴보면 〈그림 2〉와 같이 7개의 과정으로 나누어 볼 수 있다. 이 그림에서는 디지털컴포넌트저장소(그림에서 'SIP저장소'와 'D · C저장소'로 표기)와 데이터저장소(그림에서 '입수단데이터저장소'로 표기)를 구분하여 표시하고 있다. 각 과정의 특징을 살펴보면 다음과 같다.

첫째, I1:입수계획 수립 및 이송과정이다. 국가기록원은 다수의 기관으로부터 매년 장기보존 대상 디지털 기록정보를 입수하게 된다. 생산기관과 협의를 거쳐 입수일정, 입수대상, 입수방법을 정하여 계획을 수립하는데, 기관별 준비정도와 정보통신환경을 고려하여 효율적인 입수방법을 모색해야 한다. 이때, 이송되어온 기록정보를 받을 SIP저장소의 개수와 배치 형상도 함께 결정하도록 한다.

둘째, I2:입수파일저장과정이다. 생산기관으로부터 온라인, 오프라인

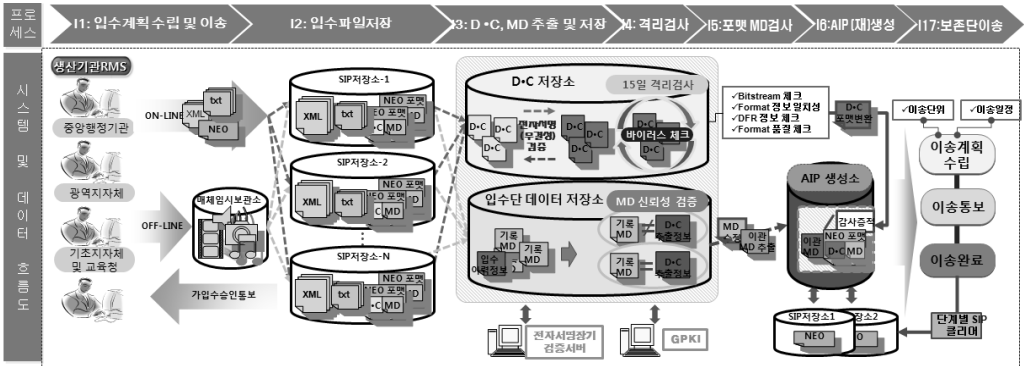
3) 이 논문의 2장 2절, 3절, 4절에서 제시하는 처리절차는 차세대프로젝트 과정에서 프로젝트팀과 국가기록원 전자기록 전략개발 태스크포스팀이 워크숍을 하면서 공유했던 절차도이다. 이 절차도는 디지털 기록정보의 처리와 저장의 관점에서 필요한 과정만 선별적으로 표시한 그림으로 영구기록물관리기관의 디지털 기록정보 관리업무 전체를 포괄하려는 의도로 작성된 것이 아님을 밝힌다. 이 절차도는 현행(AS-IS)과 미래(TO-BE)절차가 섞여서 표시되어 있으며, 향후 처리절차의 개선을 논의할 목적으로 작성되었다.

으로 이송되어온 기록정보 파일들을 입수단의 SIP저장소에 저장한다. 생산기관 RMS에 대용량 송수신 모듈을 구축한 경우 온라인 이송이 가능하다. 온·오프라인 이송방식에 따라 저장 처리에 걸리는 시간이 달라지며, 품질검사 과정에서 재전송을 요청하는 회수가 많아지고 재전송 양이 클수록 저장에 소요되는 시간이 길어진다. 생산기관에서 영구기록물관리기관으로 기록정보가 이송된 이력정보가 입수단데이터저장소에 누적되어 관리되어야 한다. 국가기록원으로 입수되는 기록정보는 NEO(NAK's Encapsulated Objects)포맷으로 패키징되어 있어야 하며, NEO들은 이관 XML파일에 담고 기록물 철, 건 메타데이터는 TXT파일에 담아 이송하게 되어 있다. 이 논문에서는 이관포맷이 초점은 아니므로 상세한 분석은 하지 않겠으나 다만, 현재의 이관지침 상 XML과 TXT 파일에 기록정보의 메타데이터가 중복하여 포함되는 것은 메타데이터 항목 간의 일치성 확인을 해야 하는 등 향후 대량의 전자기록 이관 시 작업 효율을 저해할 수 있으므로 재고할 필요가 있다고 본다. XML에는 기록관에서 NEO로 패키징된 기록정보를 담고, TXT에는 패키징 이후 기록관 RMS(Records Management Systems)에서 변경한 메타데이터나 감사증적 정보를 담아 보내도록 용도를 구분하는 것이 유용할 것이다.

셋째, I3:디지털컴포넌트(그림에서 'D·C'로 표현)와 메타데이터(그림에서 'MD'로 표현) 추출 및 저장과정이다. I2에서 입수한 NEO를 파싱(Parsing)하여 메타데이터와 디지털컴포넌트를 추출한 후 각각 '입수단 데이터저장소'와 'D·C저장소'에 저장한다. 이때, 디지털컴포넌트의 경우 XML의 문자열을 디코딩하여 비트스트림을 생성한 후 전자서명을 이용하여 무결성을 검증한다.

넷째, I4:격리검사과정이다. 디지털컴포넌트들이 바이러스에 감염되었는지 확인하고 잠복한 바이러스를 퇴치하는데 필요한 기간 동안 'D·C저장소'를 격리하여 조치한다. 이때, 디지털컴포넌트의 본문에 삽입되어 있는 하이퍼링크에 연결된 웹컨텐츠도 함께 저장하고 격리 조치할

〈그림 2〉 입수단 절차별 기록정보의 흐름



것인지 결정한다.

다섯째, I5:포맷 및 메타데이터 검사과정이다. 디지털컴포넌트별로 비트스트림을 검사하여 파일명 확장자에 제시된 포맷정보와 일치하는지를 확인하고 불일치하는 경우 확장자를 수정해준다. 국가기록원에서 운영하는 DFR(Digital Format Registry)에서 제공하는 포맷정보를 이용하여 디지털컴포넌트 포맷의 유효성을 검사한다. 디지털컴포넌트의 본문에서 내용 메타데이터를 추출하여 함께 입수된 메타데이터 값과 비교함으로써 신뢰성을 검증하며, 기록정보의 내용과 메타데이터가 다를 경우 필요시 메타데이터를 수정하도록 한다. 포맷검증도구, 본문키워드추출도구, 메타데이터검증도구 등 이 과정에 필요한 여러 도구의 개발이 요구되며 입수단 시스템에서는 여러 도구가 작동하는데 필요한 데이터 처리 공간을 확보해야 한다.

여섯째, I6:AIP(재)생성과정이다. 앞 절차들을 통해 디지털컴포넌트와 메타데이터에 대한 검증이 완료된 기록정보는 NEO포맷의 AIP로 재패키징 해준다. 이때, AIP에는 입수된 NEO와 수정된 디지털컴포넌트, 수정된 메타데이터, 입수관련 감사증적 정보가 포함된다. 이 과정에 AIP 생성기 도구가 필요하며 이 도구를 작동하는데 필요한 데이터 처리 공

간으로 'AIP생성소'를 확보해야 한다. 이 생성소에는 NEO포맷 규격정보가 제공되어야 하며, 비트스트림을 64Base로 인코딩하여 문자열을 생성해주는 도구가 탑재되어 있어야 한다.

일곱째, I7:보존단이송과정이다. AIP들을 보존단의 저장소로 이송하기 위해 보존담당자와 이송단위, 이송일정, 이송방법을 협의하여 계획을 수립한다. 보존단에 이송이 성공적으로 완료되면 입수단에 남겨진 SIP와 임시데이터들을 삭제하여 다음 입수에 대비한다.

종합해보면 입수단에 필요한 저장소의 종류와 개수는 입수되는 디지털 기록정보의 양과 품질에 기본적으로 좌우된다. 일정 시간 내에 입수처리를 완료하기 위해서는 I2과정에서 전체 입수 기록정보의 양을 시간단위, 입수저장소 단위로 배분하여 처리할 필요가 있다. I5과정에서는 기록정보의 품질이 나쁜 경우 보정을 위한 여러 도구가 동작해야 하므로 처리대상 기록정보를 임시로 저장해둘 저장소와 도구의 작동을 위한 공간이 추가로 요구된다.

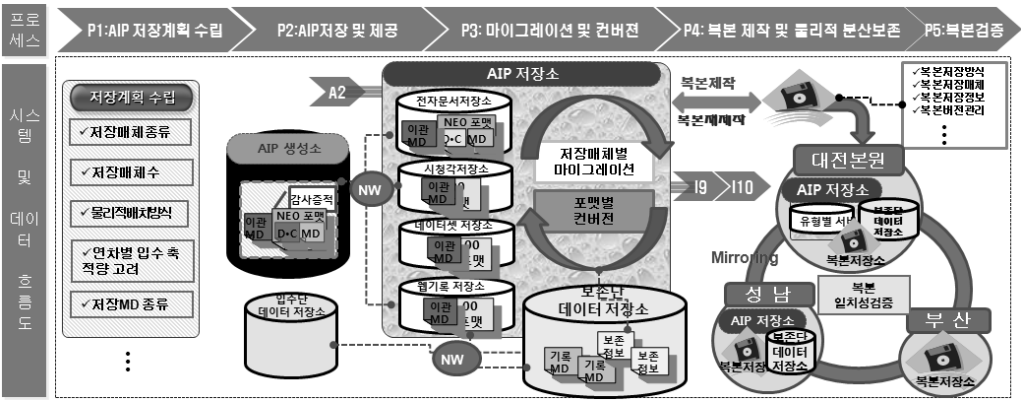
3) 보존단(Preservation): AIP유형과 복본 제작

국가기록원의 보존단 프로세스를 디지털 기록정보의 흐름을 중심으로 살펴보면 〈그림 3〉과 같이 5개의 과정으로 나누어 볼 수 있다. 이 그림에서는 'AIP저장소'와 '보존단데이터저장소'를 구분하여 표시하고 있다. 각 과정의 특징을 살펴보면 다음과 같다.

첫째, P1:AIP저장계획수립과정이다. 입수단을 거쳐 보존단으로 이송되어올 기록정보를 저장하기 위해 필요한 저장매체의 종류, 저장매체의 개수와 용량, 저장소 배치방식 등을 고려하여 AIP 저장계획을 수립한다.

둘째, P2:AIP저장 및 제공과정이다. AIP저장 계획에 맞춰 실제 AIP객체를 이송하여 저장매체에 저장하고, 제공단으로부터 요청이 오면 원하는 AIP를 찾아 제공단으로 전송해준다. 입수단에서 이송해온 AIP들은

〈그림 3〉 보존단 절차별 기록정보의 흐름



'AIP저장소'에 저장하고, AIP에 대한 정보들은 '보존단데이터저장소'에 저장한다. 보존단 'AIP저장소'에 NEO파일만 저장할 것인지 이후 국가기록원 내부의 접근용이성을 위해 디지털컴포넌트를 추출하여 함께 저장할 것인지 결정해야 한다. 〈그림 3〉에서는 NEO파일만 저장하는 것을 가정하고 있다. '보존단데이터저장소'에 저장되는 데이터에는 AIP의 기술정보(DI, Descriptive Information)뿐만 아니라 AIP 패키징정보, 디지털컴포넌트에 대한 메타데이터, 입수과정의 감사증적 정보, 보존과정의 감사증적 정보, 시스템 로그정보 등 다양한 종류의 데이터가 함께 저장된다. 보존단의 AIP저장소는 매년 누적되는 기록정보를 관리해야 하므로 대용량의 기록정보를 보존처리하기에 용이하게, 그리고 제공단 요청에 따라 쉽게 기록정보를 제공해줄 수 있도록 저장하는 것이 핵심전략이다. 저장소를 하나로 유지하기보다는 여러 개로 나누어 관리하되 저장소를 구분할 때는 디지털컴포넌트 포맷별로 나눌 것인지, 분류체계나 출처에 따라 나눌 것인지, 보존기간에 따라 나눌 것인지를 결정해야 한다. 하나는 포맷변환이나 처분과 같은 처리에 유리하며, 다른 하나는 매년 누적분에 따라 저장소를 증설하기에 유리하다. AIP에 대한 접근

요구의 빈도를 고려하여 저장소를 구성하는 저장매체를 온라인, 니어라인, 오프라인 종류로 나누어 배치해야 한다. 보존단의 핵심성과지표인 안정적인 보존을 위해 재난복구 전략수립이 필요하며 기록정보의 복본 제작 개수와 복본 저장 위치를 결정해야 한다.

셋째, P3:마이그레이션 및 컨버젼과정이다. 저장매체는 시간이 흐르면서 노후화되므로 기록정보는 기술의 변화에 따라 새로운 저장매체로 마이그레이션해야 한다. 마이그레이션 결과로 AIP의 물리적 저장장소가 변경되므로 관리메타데이터 값이 수정되며 감사증적정보가 생성된다. 현재 PDF/A로 정한 문서보존포맷이나 NEO로 정한 장기보존포맷이 새로운 포맷으로 변화된다면 디지털컴포넌트와 AIP들을 새로운 포맷으로 컨버젼해야 한다. 디지털컴포넌트만 새로 컨버젼해도 AIP를 재생성해야 되며, 컨버젼 후에는 메타데이터를 변경해 주어야 한다. 〈그림 3〉에서는 마이그레이션과 컨버젼은 보존단의 고유 작업이지만 AIP재생성은 입수단의 기능을 호출하여 수행하는 것으로 정의하였다. 보존단의 처리 단위가 주로 디지털컴포넌트의 포맷이나 저장매체 유형이 된다는 점을 고려하여 보존단 저장소를 설계할 필요가 있다.

넷째, P4:복본제작 및 물리적 분산보존과정이다. 기록정보의 안전한 보존을 위한 보존계획으로 재난 시 복구에 사용할 기록정보의 복본을 제작할 대상 범위, 복본의 개수, 복본의 저장 위치 등을 결정하고, 이에 따라 복본을 제작하여 분산 저장한다. 복본을 제작하기 위한 도구와 작업공간이 필요하다. 복본의 개수만큼 보존단 저장소가 추가로 요구되며 분산 저장하는 위치에 따라 보존단 저장소가 분산배치되어야 한다.

다섯째, P5:복본검증과정이다. 복본이 제작되고 나면 복본 간에 내용이 일치하는 지를 주기적으로 확인하고, 만약 불일치한 부분이 발생했을 경우 원인을 규명하여 해결하고 복본을 재제작하여 일치성을 유지하도록 한다. 복본의 일치성 확인을 위한 도구와 작업공간이 필요하다.

종합해보면 보존단에 필요한 저장소의 종류와 개수는 디지털 기록정

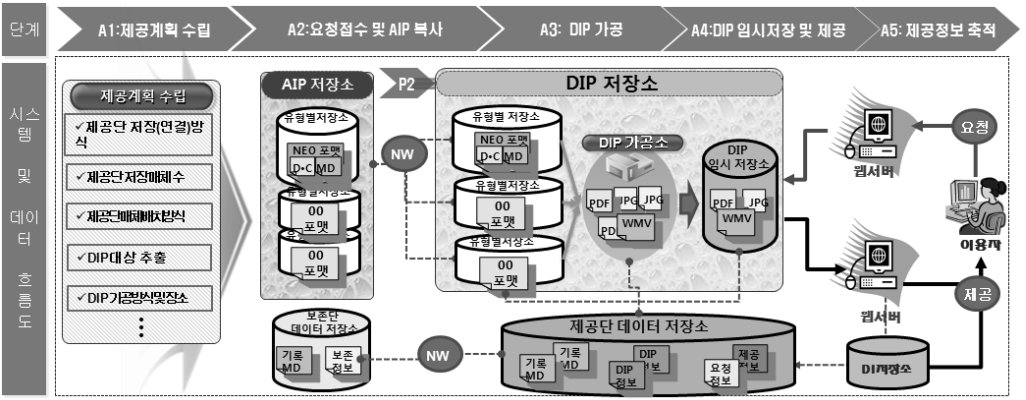
보의 유형과 재난복구를 위해 제작하는 복본 기록정보의 개수에 기본적으로 좌우된다. P3과정을 효과적, 효율적으로 수행하기 위해서는 P2과정에서 디지털컴포넌트의 포맷을 고려하여 'AIP저장소'를 복수 개로 구성할 필요가 있다. 전자문서류, 시청각류, 데이터세트류, 웹기록류 등으로 분리하는 것이 유리하다. P4과정에서 만든 복본의 수만큼 AIP저장소가 증가하게 되는데 원본과 동일한 서버사양으로 미러링하는 방식의 복본제작과 DVD와 같은 오프라인 저장매체에 복사하는 방식의 복본제작 등 다양한 저장소 배치 형상이 가능해진다. 복제되는 기록정보의 양에 따른 저장비용과 관리비용, 정보통신기술의 발전 수준을 반영하여 결정해야 할 것이다.

4) 제공단(Access): DIP제공 방식과 검색도구

국가기록원의 제공단 프로세스를 디지털 기록정보의 흐름을 중심으로 살펴보면 〈그림 4〉와 같이 5개의 과정으로 나누어 볼 수 있다. 제공단은 이용자에게 다양한 기록정보 디지털컨텐츠를 서비스하는 여러 서브시스템을 구성할 수 있겠으나 여기서는 기록정보의 원문과 메타데이터를 제공하는 기본 서비스만을 가정하여 절차를 살펴보고자 한다. 이 그림에서는 'DIP저장소'와 'DIP임시저장소', '제공단데이터저장소', 'DI저장소'를 구분하여 표시하고 있다. 각 과정의 특징을 살펴보면 다음과 같다.

첫째, A1:제공계획 수립과정이다. 공표대상과 공개로 분류된 기록정보를 이용자에게 제공하는 방식을 정하고, 공개가 제한되는 기록정보에 대한 이용자의 접근통제 방식을 정한다. 또한, 이용자가 요청한 기록정보를 DIP로 가공·제작하는 방식을 정하고, 이용자들이 기록정보를 탐색, 검색하는데 필요한 도구 개발 계획을 수립한다. AIP패키지에 대한 기술정보(Descriptive Information)를 어디에 제공할 것인지를 결정한다.

〈그림 4〉 제공단 절차별 기록정보의 흐름



둘째, A2:요청 접수 및 AIP복사과정이다. 이용자가 검색도구를 통해 특정 기록정보의 추가 데이터나 원문의 열람을 요청하면 이를 접수하고, 해당 기록정보가 제공단 DIP저장소에 존재하지 않을 경우 보존단에 요청하여 보존단 'AIP저장소'의 AIP를 제공단 'DIP저장소'로 복사해온다.

셋째, A3:DIP가공과정이다. 이용자에게 기록정보를 배부하기 위해 정해진 포맷대로 DIP를 제작한다. DIP제작도구가 필요하며, 이 도구를 이용하여 'DIP가공소' 공간에서 이용자가 선택한 디지털컴포넌트와 메타데이터를 중심으로 DIP패키지를 생성한다. 이용자가 요청한 일부 기록정보에 대한 가공뿐 아니라 공표대상 기록정보를 대량으로 한꺼번에 가공하는 작업도 가능하도록 DIP가공소를 설계해야 한다.

넷째, A4:DIP 임시저장 및 제공과정이다. DIP패키지는 'DIP임시저장소'에 저장되며 이용자가 원하는 채널을 통해 제공된다. 제공된 DIP는 임시저장소에 일정기간동안 보유하다가 임시저장소 공간이 부족해질 때 공간확보가 필요해지면 순서에 따라 삭제하게 된다. 공표대상 DIP패키지들은 공표를 위한 서비스시스템으로 이송되어 제공되며 이송직후 DIP임시저장소에서 바로 삭제한다. 기술정보(DI)는 이용자에게 제공할

기록정보에 관한 메타데이터를 보존단 데이터저장소에서 제공단 데이터저장소로 복사해 온다. 이 메타데이터 중에서 이용자에게 제공되는 검색도구에 사용되는 기술정보들은 별도의 'DI저장소'에 중복하여 저장된다. 검색도구는 여러 개를 개발할 수 있으므로 도구별 DI정보들이 'DI저장소'에 누적된다.

다섯째, A5:제공정보 축적과정이다. 기록정보에 대한 이용자의 요청과 AIP의 복사과정, DIP가공이력, DIP제공이력 등이 '제공단데이터저장소'에 저장된다. 향후 이용자가 DIP패키지의 진본여부를 확인받고자 한다면 진본인증을 위한 도구가 개발되어야 하며 이를 위해 '제공단데이터저장소'에 진본인증에 필요한 상세정보가 더 축적되어야 한다.

종합해보면 제공단에 필요한 저장소의 종류와 개수는 DIP패키지의 제공방식이 대량인지 소수의 건 단위인지, 그리고 디지털컴포넌트를 모두 포함하는 방식인지 서비스용 일부만 포함하는 방식인지, 메타데이터를 포함하는 범위 등에 따라 좌우된다. 검색도구가 여러 개 제공되면 그에 따라 기술정보(DI)가 추가되어야 하며, 제공되는 과정에 대한 감사증적을 남기는 범위에 따라 데이터저장소의 용량이 결정된다.

3. 대량기록물 관리를 위한 데이터베이스 설계

2장에서는 영구기록물관리 단계별 디지털 기록정보의 처리 흐름을 중심으로 저장소의 요건을 살펴보았다. 3장에서는 2015년에 국가기록원에 이관될 디지털 기록정보의 양을 추산해 봄으로써 2015년 이후의 입수단, 보존단, 제공단에 필요한 디지털저장소의 크기와 성능에 대한 고려사항을 살펴보고자 한다. 또한, '대량기록물' 처리의 구체적인 요건과 데이터베이스 설계 방향에 대해 검토해보고자 한다.

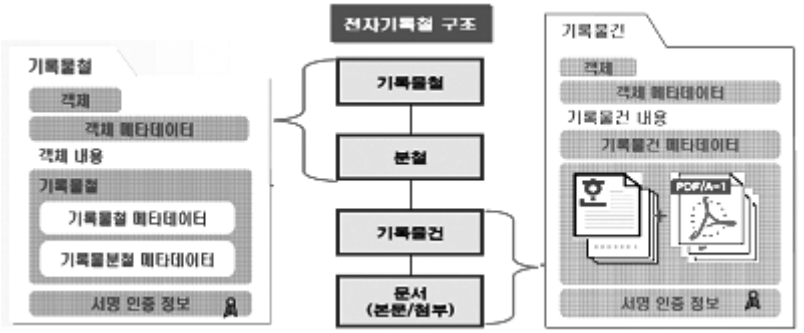
1) 2015년 국가기록원 관리대상 기록정보의 양

영구기록물관리기관의 디지털 기록정보 관리는 입수단에서부터 시작된다. 입수단 시스템을 구축하기 위해서는 입수된 기록정보의 종류와 양, 그리고 품질에 관한 정보가 필요하다. 국가기록원의 경우, 대량 이관의 원년인 2015년도의 입수량뿐만 아니라 그 이후 매년 입수될 기록정보의 양을 예측하는 것이 준비의 출발점이 될 것이다. 이 논문에서는 2015년에 국가기록원에 입수될 디지털 기록정보의 양을 데이터저장소와 디지털컴포넌트저장소 두 가지 종류의 용량으로 구분하여 추산해 보고자 한다. 먼저, 양의 추산을 위해 기록정보의 구조와 구성을 다음과 같이 전제하고자 하다.

첫째, 디지털 기록정보는 철, 건 구조를 가지고 입수된다. 2015년에 이관되어올 전자기록의 건(Item)들은 기록물철(Folder)에 편철된 상태로 입수된다. 전자기록 한 건은 본문을 구성하는 1개의 전자문서와 0개 이상의 첨부파일, 메타데이터 등으로 구성된다. 비전자기록의 경우도 디지털화를 수행한 경우 디지털컴포넌트가 포함된 디지털 기록정보로 입수된다. 전자기록의 철, 건 구조를 개념도로 표현하면 〈그림 5〉와 같다.[4)]

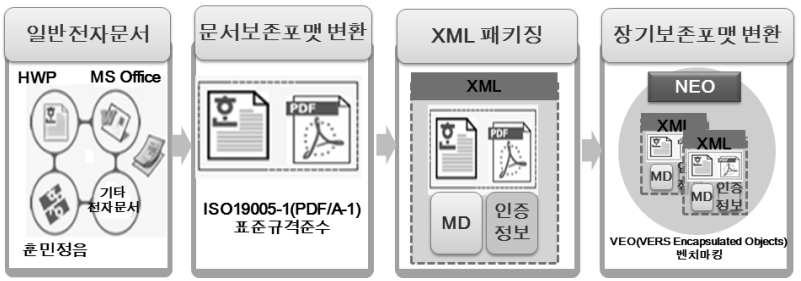
4) '전자기록 영구보존기술 적용을 위한 테스트베드구축사업' 최종보고회 발표자료, 국가기록원, 2006.

〈그림 5〉 전자기록의 철, 건 구조와 구성



둘째, 생산기관의 RMS를 통해 이관되는 디지털 기록정보는 NEO포맷으로 패키징된 것이다. 생산기관은 장기보존 대상 전자기록에 대해 본문과 첨부파일을 각각 문서보존포맷인 PDF/A로 변환하고, 디지털컴포넌트와 메타데이터, 인증정보들을 모아 64Base 인코딩하여 XML파일을 생성한다. 기록정보의 포맷변환 및 인코딩절차는 〈그림 6〉과 같다.

〈그림 6〉 전자기록의 보존포맷변환 절차



디지털 기록정보의 구성과 구조가 위와 같다고 전제를 하고 다음으로는 2015년부터 해결해야 할 과제의 핵심인 '대량기록물'의 실체에 대해 좀 더 접근해 보고자 한다. 우선 다음과 같은 분석을 출발점으로 삼

을 수 있을 것이다.

첫째, 기록시스템에서 관리하는 대상으로서의 '기록물'에는 디지털 기록정보와 물리적 기록정보가 모두 포함된다. 기록정보는 디지털컴포넌트와 각종 데이터로 구성되는데 OAIS참조모형의 정보모형 용어로 설명하자면 디지털 기록정보는 내용정보(Content Information)와 보존 및 기술 메타데이터(Preservation Description Information)로 구성되며, 물리적 기록정보는 내용정보 없이 보존 및 기술 메타데이터로만 구성된다.

둘째, '대량'의 의미는 처리할 기록정보의 건수가 많다, 메타데이터 양이 많다, 처리할 디지털컴포넌트의 수가 많다, 디지털컴포넌트의 총 바이트 수가 크다 등 여러 방식으로 해석할 수 있다. 디지털컴포넌트의 개수나 바이트 수는 메타데이터에 비해 '대량'인 것이 확실하다. 이러한 양적 측면을 고려할 때 영구기록물관리시스템의 디지털저장소를 데이터저장소와 디지털컴포넌트저장소로 구분하고, 내용정보인 디지털컴포넌트는 파일시스템으로 운영되는 디지털컴포넌트저장소에 저장하고 보존 및 기술 메타데이터와 감사증적(Audit Trail)데이터와 시스템로그(System Log), 임시데이터(Transient Data)등의 데이터는 데이터베이스로 운영되는 데이터저장소에 저장하는 것이 효율적일 것이다. 디지털컴포넌트를 데이터베이스에서 LOB(Large Object)유형의 칼럼으로 저장하여 관리할 수도 있으나 AIP의 처리과정에서 이득을 얻기 어렵다.

셋째, 디지털컴포넌트저장소에서 '대량'의 이슈는 분명하다. 전자문서시스템과 업무관리시스템에서 생산되는 전자기록 건 별 디지털컴포넌트의 수가 증가하고 있고, 디지털컴포넌트별 평균 사이즈도 증가추세에 있다. 종이문서를 스캐닝한 이미지파일의 경우도 현재 300dpi를 해상도 기준으로 하고 있으나 향후 600dpi 이상을 요구하는 추세로 나아가고 있다. 또한, 2장에서 살펴본 바와 같이 영구기록물관리 과정에서 하나의 디지털 기록정보를 처리하기 위해 NEO파일 내의 디지털컴포넌트를 추출하는 등 기록정보가 일시적으로라도 중복 저장되는 시점이 있으므

로 디지털컴포넌트저장소는 대용량의 스토리지가 필요할 것이다. 다만, 한 번 저장하면 장기적으로 유지되는 기록정보와 처리를 위해 임시로 저장할 필요가 있는 기록정보는 저장소를 분리하여 보관해야 한다.

넷째, 데이터저장소에서 '대량'의 이슈는 2015년 기준으로 보았을 때 중요 이슈는 아닐 것이다. 최근 컴퓨팅 분야에서 빅데이터(Big Data) 문제가 화제이지만 이때의 빅데이터는 테라급을 넘어 페타급의 데이터양을 의미하는 것이 일반적이다. 데이터의 양적 측면에서 본다면 국가기록원의 경우는 아직 빅데이터 수준까지는 미치지 않는다고 예측할 수 있다.

위와 같은 분석을 염두에 두고 다음으로는 '대량기록물'의 구체적인 양을 2장에서 살펴본 3단별로 필요한 저장소 용량을 추산해 보고자 한다. 계산의 절차는 다음과 같다.

첫째, 차세대프로젝트에서는 2015년 국가기록원에 디지털 기록정보를 이관해야 하는 기관을 대상으로 보존기간 30년, 준영구, 영구로 책정된 디지털 기록정보의 보유현황을 살펴보았다. 아직 전수 통계자료가 만들어지기 전이라 국가기록원 전자기록 전략개발 태스크포스에서 2011년 10월 현재 조사 완료한 20개 기관의 보유현황을 기준으로 2015년 이관대상 총량을 추산해 보았다. 표1은 20개 중앙행정기관이 2004년부터 2010년까지 생산하여 보유하고 있는 디지털 기록정보의 수량으로 총량은 약 2.1GB이다.

〈표 1〉 20개 중앙행정기관 전자문서 보유현황(2004~2010년)

보존기간	기록물철 개수	기록물 건수	용량(단위GB)
30년	41,095	1,598,899	463.877
준영구	124,271	3,531,556	749.412
영구	78,253	2,427,906	953.745
총계	243,619	7,558,361	2,167.035

* 국가기록원 제공자료.

둘째, 2015년에 국가기록원으로 이관될 디지털 기록정보의 양은 총 2.5테라바이트로 추산하여 가정하였다. 생산현황 통계에서 빠진 기관의 수와 규모를 감안하여 표1 수치를 확장하여 양을 추산한 것이다. 2.1GB를 7년으로 나누고 20개 기관 수로 다시 나누어 한 기관의 평균 바이트 수를 구한 다음, 2015년 이관을 수행할 기관 수로 곱하였고, 여기에 전자문서를 NEO로 포맷변환하면서 증가하는 양과 이관파일을 작성하면서 추가되는 양을 포함하였다. 다만, 〈표 1〉의 통계는 전자문서만의 생산현황이라는 점을 감안하여 향후 시청각기록, 이메일기록, 웹기록, 데이터세트 등 다양한 유형의 기록을 함께 획득할 때는 입수량이 훨씬 커진다는 것을 염두에 두어야 할 것이다.

셋째, 입수단에서는 약 7.2테라바이트 정도의 디지털컴포넌트저장소와 약 52.6기가바이트 정도의 데이터저장소가 필요한 것으로 추산하였다. 총 2.5테라바이트의 이관 기록정보가 먼저 저장되고, 이관파일에서 NEO파일을 추출하여 저장해야 하며, 여기서 디지털컴포넌트와 메타데이터가 추출되어 다시 저장되어야 한다. NEO파일 크기의 대부분을 디지털컴포넌트가 차지하고 있기 때문에 디지털컴포넌트저장소의 용량은 7테라바이트 이상 요구된다. 여기에 디지털컴포넌트와 메타데이터의 품질 검사 결과 변경이 발생하면 NEO를 수정해야 하므로 저장공간이 더 필요해지며, 수정 가능성을 5%로 가정했을 때 2.5테라바이트의 5%인 125

기가바이트의 스토리지가 더 요구된다. 메타데이터 값들은 데이터베이스에 저장해야 하는데 현재 국가기록원 중앙기록물관리시스템 CAMS (Central Archives Management Systems)에서 기록의 건과 철을 관리하기 위해 사용 중인 테이블과 칼럼을 기준으로 〈표 2〉와 같이 산정했을 때 약 52.6기가바이트의 데이터베이스 용량이 필요한 것으로 계산된다.

〈표 2〉 기록정보의 메타데이터 저장에 필요한 용량산정

	기록물 철 테이블	기록물 건 테이블
CAMS의 테이블명	RG_DOCUMENT	RG_DETAIL
컬럼 수	139	105
레코드 최대 크기	17618바이트	17103바이트
입수 추정 레코드 수	10만건	320만건
테이블 추정 크기	1.64GB	50.97GB

넷째, 보존단에서는 2015년에 이관되는 2.5테라바이트 양만큼의 저장공간이 요구되며 이후 매년 10%씩 입수량이 증가한다고 가정했을 때 2.5TB * 1.1^n(n은 2015년 이후 연차)의 저장공간이 추가로 필요해진다. 국가기록원의 경우 재난대비 계획의 일환으로 대전 본원과 성남 분원 간 영구기록물관리시스템을 미러링하고 있으므로 입수량 두 배 크기의 저장소가 필요하게 된다. 재난대비 계획에 전체 기록정보를 대상으로 복본을 제작하기로 한다면 복본의 개수에 비례하여 추가로 저장소가 필요하게 된다. 미러링을 하고 복본 1개를 제작하는 것으로 전제했을 때 2015년에 보존단에 필요한 저장공간은 약 7.5테라바이트가 된다. 추후 보존단에 필요한 전체 저장소 공간을 추산할 때는 비전자기록을 디지털화한 컴포넌트도 저장해야 한다는 점, 보존기간 재평가를 통해 기록정보가 폐기되기도 한다는 점을 감안해야 하며, 보존단 저장소 관리자는 기록정보 폐기 후 저장소 공간을 재배열(re-arrangement)해야 한다

는 점을 고려해야 한다. 입수데이터 52.6기가바이트에 보존과정에서 추가되는 기술정보, 보존처리에 따라 입력되는 관리메타데이터, 감사증적정보와 로그데이터가 입수데이터의 약 50%가 된다고 추산해 보았을 때 필요한 데이터저장소 총량은 약 78.9기가바이트가 된다.

다섯째, 제공단은 사전공표나 자발적 공개 대상 기록정보의 범위, 그리고 기본 DIP구성 방식에 따라 불특정 이용자를 위해 기록정보를 제공하는데 필요한 저장공간의 양이 달라진다. 2015년 입수되는 기록정보 중 50%를 DIP로 제작하여 웹상에 제공한다고 전제하고 DIP제작 시 디지털컴포넌트 중 보존포맷으로 변환된 것만 포함하며 메타데이터의 일부만 포함하는 것으로 계산하면 약 1.2테라바이트 정도의 저장소가 필요하다는 결과가 나온다. DIP제작에 필요한 공간이나 DIP 임시저장 공간 등은 이용자의 기록정보에 대한 요청 시 지속적으로 재사용하는 공간이므로 1테라바이트 이하로 추산해도 충분할 것으로 판단된다. 제공단이 기록정보의 DIP뿐만 아니라 기록정보를 이용한 콘텐츠까지도 함께 서비스하는 역할을 하게 된다면 이에 소요되는 저장소는 콘텐츠의 특성을 고려하여 추가 계산되어야 할 것이다. 검색도구 하나 당 기술정보가 메타데이터의 30% 정도가 추출되어 중복저장되는 것으로 가정하면 검색도구를 4개 개발할 경우 약 63기가바이트의 DI저장소가 필요하다. 여기에 감사증적정보와 로그데이터는 입수 데이터의 약 20%로 추산해 보았을 때 필요한 데이터저장소 총량은 약 73.5기가바이트가 된다.

아래 〈표 3〉은 영구기록물관리시스템의 3단계별로 2015년에 필요한 데이터저장소와 디지털컴포넌트저장소의 수요량을 정리한 것이며, 〈표 2〉와 〈표 3〉은 보존단에 필요한 저장소를 2015년부터 2020년까지 6년간 누적분을 계산해본 것이다.[5]

5) 전제조건을 구체화함으로써 보존단과 제공단의 추정 양이 차세대프로젝트의 보고서와는 달라졌음.

〈표 3〉 2015년 단계별 필요 디지털저장소 예상치

단계	입수단	보존단	제공단
데이터저장소	52.6GB	78.9GB	73.5GB
디지털컴포넌트저장소	7.2TB	7.5TB	1.2TB

〈표 4〉 연도별 입수량 증가에 따른 보존단 데이터저장소 예상치

(단위: GB)	2015년	2016년	2017년	2018년	2019년	2020년
추가용량	78.9	86.79	95.47	105.02	115.52	127.07
누적용량	78.9	165.69	261.16	366.17	481.69	608.76

〈표 5〉 연도별 입수량 증가에 따른 보존단 디지털컴포넌트저장소 예상치

(단위: TB)	2015년	2016년	2017년	2018년	2019년	2020년
추가용량	7.5	8.25	9.08	9.98	10.98	12.08
누적용량	7.5	15.75	24.83	34.81	45.79	57.87

이 논문에서 추정한 수치는 여러 불확실한 가정에 기반해 있으므로 실제 2015년의 데이터양과 정확히 일치하지는 않을 것이다. 그러나 향후 국가기록원이 어느 정도 규모의 디지털 기록정보를 다루게 될지 가늠해볼 수 있다는 점에서 실용적 의미를 갖는다. 영구기록물관리시스템은 매년 대량의 기록정보가 누적된다는 특성을 갖는다. 설혹 한 해의 데이터 양은 크지 않다고 해도 100년 정도 장기간(Long-term period) 누적된 양은 큰 규모가 될 것이다. 지속적인 양적 증가를 감안한 데이터베이스 설계와 저장매체 구성 기법은 달라지게 된다. 장기적 관점에서 대량의 데이터를 효과적, 효율적으로 다룰 수 있는 데이터베이스 구조와 저장소의 배치 형상을 연구해야 하는 이유이다.

2) 기록시스템의 대량처리 요건에 따른 데이터베이스 설계 방향

앞 절에서 살펴본 바와 같이 2015년부터 국가기록원의 CAMS에서는 여러 처리과정에서 대량의 디지털 기록정보를 다룰 수 있어야 한다. 이 절에서는 CAMS와 같은 기록시스템에서 갖추어야 하는 대량처리(Bulk Operation) 기능요건을 살펴보고 데이터베이스의 설계 방향을 도출하고자 한다.

기록시스템에 대량의 기록정보를 대상으로 하여 다양한 일괄작업 기능을 구현하는 것은 기록관리의 효과성과 효율성을 크게 좌우한다.[6] MoReq2010에서는 기록시스템에서 갖춰야 하는 대량처리의 일반적인 기능 요건을 다음과 같이 제시하고 있다.

- 여러 엔티티(entity)를 선택하여 동일한 기능을 수행하도록 했을 때, 각 엔티티별로 같은 기능이 한 번씩 반복적으로 수행되어야 한다.
- 기능의 수행과정에서 실행이 실패한 엔티티가 일정한 양에 도달하게 되면 대량처리를 자동적으로 취소할 수 있도록 하는 방안이 제공되어야 한다.
- 대량처리의 진행과정이 기능 수행자에게 피드백될 수 있어야 한다.
- 대량처리를 수행시킨 상태에서도 시스템의 다른 기능을 사용할 수 있어야 한다.
- 대량처리를 수행하는 과정에서 언제든 임의로 대량처리를 취소할 수 있어야 한다.
- 대량처리가 취소되었을 때는 가능한 빨리 수행을 종료하되 시스템의 무결성(integrity)을 유지하여야 한다.
- 대량처리가 완료되거나 취소되었을 때 결과에 대한 요약정보를 수

6) 임진희, 「기록관리시스템 기능요건 표준의 실무적 해석」, 『기록학연구』 제18호, 2008, 171~175쪽.

행자에게 피드백해주어야 한다.

o 대량처리 결과에 대한 요약정보에는 처리에 성공한 엔티티, 처리에 실패한 엔티티, 처리를 시도하지 않은 엔티티가 구별되어야 한다.

o 대량처리 과정에서 실행에 실패한 경우 수행자가 시스템을 통해 오류에 대한 자세한 정보를 얻을 수 있어야 한다.

이와 같은 대량처리 기능은 기록시스템에서 관리하는 주요 엔티티별에 대한 다양한 처리과정에서 필요하다. MoReq2010에서 제시하고 있는 엔티티별 대량처리 기능 요건 몇 가지를 살펴보면 다음과 같다.

o 기록시스템에 경고(Alert)를 동시에 여러 개 동작시키고, 재호출하며, 주석을 남길 수 있어야 한다.

o 여러 개의 기능(Function)을 선택하여 이벤트 이력(Event History)을 남기거나 남기지 않도록 설정할 수 있어야 한다.

o 여러 사용자(User)나 그룹(Group)을 선택하여 활성화시키거나 비활성화시킬 수 있고, 역할(Role)을 부여하거나 회수할 수 있어야 한다.

o 분류체계에서 여러 클래스(Class)를 선택하여 활성화/비활성화시키거나, 메타데이터 값을 일괄 변경하고 템플릿을 적용하거나, 처분일정을 지정/해제할 수 있어야 한다.

o 분류정보를 담고 있는 데이터파일을 읽어들여 분류체계에 새로운 클래스를 여러 개 동시에 생성할 수 있어야 하며, 이때 분류정보를 검증하고 새로운 ID를 부여할 수 있어야 한다.

o 여러 집합체(Aggregation)을 선택하여 클래스에서 분류하거나 클래스로부터 제거할 수 있어야 하며, 닫기(Close)와 다시열기(Reopen)할 수 있어야 하고, 위치를 변동할 수 있어야 한다. 또한 집합체 내의 기록물 복사본을 생성할 수 있어야 한다.

o 기록(Record)과 컴포넌트(Component)를 여러 개 선택하여 동일한 템플릿을 적용할 수 있어야 하고, 메타데이터 요소에 동일한 값을 줄

수 있어야 하며, 새로운 클래스를 적용하거나 제거할 수 있어야 하고, 처분일정을 적용할 수 있어야 한다.

o 메타데이터와 템플릿 여러 개를 선택하여 기본값을 동일하게 부여할 수 있어야 하며, 기록물이 폐기될 때 삭제되어야 대상으로 설정할 수 있어야 하며, 요소를 활성화/비활성화할 수 있어야 한다. 또한, 메타데이터에 관해 기술되어 있는 데이터파일을 읽어들여 대량으로 메타데이터 요소에 대한 정의를 추가하고, 템플릿을 추가할 수 있어야 한다. 이때, 자동으로 새로운 ID가 부여되도록 해야 한다.

o 처분일정(Disposition Schedule)을 여러 개 선택하여 활성화/비활성화할 수 있어야 한다. 데이터파일로부터 처분일정을 대량으로 읽어들여 추가할 수 있어야 하며, 이때 자동으로 새로운 ID가 부여되도록 해야 한다.

o 검색을 통해 선택한 여러 개의 기록 건들에 대해 동일한 처분일정을 적용하고, 리뷰하며, 리뷰결과를 입력할 수 있어야 한다. 또한, 여러 기록 건들을 하나의 묶음으로 이관하거나 폐기하도록 승인할 수 있어야 하며, 처분에 관한 주석을 추가할 수 있어야 한다.

이상의 요건을 기록시스템에 기능으로 구현하기 위해서는 기록관리 어플리케이션, 데이터베이스관리시스템, 데이터저장소와 디지털컴포넌트저장소 등 기록시스템의 여러 구성요소가 함께 조정되어야 한다. 요건 중 상당수는 데이터저장소에 있는 메타데이터를 변경하는 작업으로 이는 데이터베이스의 설계와 데이터베이스관리시스템의 기능 설계 시 반영되어야 한다.

데이터베이스의 설계란 정보시스템 구축에 필요한 데이터를 구조화하여 담을 수 있는 데이터베이스 구성용 설계도를 만드는 것으로 ERD (Entity Relationship Diagram)를 만드는 일로부터 시작한다. 데이터베이스의 구축과정은 〈그림 7〉[7]과 같이 업무에 대한 분석, 데이터에 대한

7) DB 설계 백서, http://blog.naver.com/ilsooni3/50108431072.

분석을 수행한 후, 논리적 데이터모형, 물리적 데이터모형을 개발한 후 데이터베이스관리시스템에 데이터베이스를 구현하고 업무 및 정보 요건에 맞게 구축되었는지 시험을 거쳐 확정하게 된다.

데이터베이스가 성공적으로 구축되기 위해서는 다음 세 가지 핵심목표를 달성해야 한다.

첫째, 정보모형의 정확성이 확보되어야 한다. 구축과정의 앞 단계인 업무와 데이터에 대한 분석이 정확히 이루어져야 하며, 분석 결과 정보시스템에서 관리할 정보의 종류와 단위 및 관계가 ERD에 제대로 반영되어야 한다. 관리해야 할 대상 정보를 빠뜨리거나 정보간의 관계를 부적절하게 정의하게 되면 정보의 품질을 보장할 수 없게 된다.

둘째, 정보처리의 성능이 확보되어야 한다. 데이터베이스 설계과정이 논리모형과 물리모형으로 단계적 접근을 취하는 이유는 최종적인 정보시스템에서 정보처리의 과정이 원하는 속도와 산출량을 보장해야 하기 때문이다. 처리성능에 대한 목표치를 설정하고 이를 보장할 수 있는 데이터모형과 데이터베이스관리시스템을 선택해야 한다. 데이터모형의 경우 성능을 위해 적절한 수준에서 반정규화를 실행하기도 하며, 최근 대용량 데이터 처리를 위해 제공되는 데이터베이스관리 시스템의 특수한 기능을 채택하기도 한다.

셋째, 데이터 사전을 유지해야 한다. 기록관리에서 기록에 대한 메타데이터가 중요하듯이, 데이터베이스에서도 데이터에 대한 기술정보(Descriptive information)인 데이터 사전(Data Dictionary)이 잘 정의되고 현행화되어야 한다. 기록의 메타데이터 항목들은 최종적으로 데이터베이스의 필드로 정의되어 저장되므로 필드에 관한 정보가 자체기술(Self-Described)되는 것은 기록의 메타데이터 관리를 위한 데이터베이스의 필수 요건이라 할 수 있다.

〈그림 7〉 데이터베이스의 구축과정별 산출물

단계	분석 단계		모델링 단계		개발 및 테스트 단계	
상세 절차	비즈니스 분석	데이터 분석	논리 모델	물리 모델	DB 구현	시험
산출물	사용자 요구사항분석서	시스템 평가서	현행 시스템 모델 정의서, 논리모델 설계서 (ERD)	물리모델 설계서	스키마 생성용 스크립트, 운영관리지침서	시험계획서, 시험결과서
주요 내용	• 대상조직 업무파악 • 비즈니스 프로세스 분석 • 운영환경 분석 • 사용자요구 사항분석	• 엔터티 주제 영역별 분류 • 주식별자 및 속성 식별 • 데이터 접근 권한 분석 • 데이터정규화	• 기존 시스템 데이터 모델 분석 • DB기본설계 • 데이터관계를 ERD로 표현 • 정규화 및 비정규화	• 데이터 물리 구조 및 분산 설계 • 사용 제품별 특성에 맞도록 설계 수정 • 인덱스 정책 • 무결성,접근 경로별 설계 검토	• 개발용DB구축 • 물리설계시 요구 사항 충족 여부 검토 • DB오류 발생 여부 검토 • 개발자기술지원 • DB성능관리툴을 이용한 튜닝 수행	• 요구사항 만족정도 시험 • 설계상 오류 사항 발생 여부 시험 • 검증을 통한 논리/물리 모델 수정 반복 • 퍼포먼스 시험

앞에서 살펴본 대량처리에 대한 요구사항은 물리모형을 설계하고 데이터베이스를 운영하는 단계에서 수용할 수 있다. MoReq2010의 요건대로 대량처리가 가능하기 위해서는 데이터베이스관리시스템에서 제공하는 트랜잭션 통제 기능과 워크플로우 엔진(Workflow Engine)에서 제공하는 작업(Job) 관리 기능이 전제되어야 한다. 예를 들어, 복수의 기록에 대한 관리메타데이터를 수정하는 대량처리 작업의 경우 데이터베이스관리시스템의 트랜잭션 통제 기능을 통해 작업 전체를 하나의 트랜잭션으로 인식함으로써 요건에서 요구하는 대로 작업에 오류가 발생했을 때 처리된 내역을 모두 취소하여 기록정보의 무결성을 유지할 수 있다. 또 다른 예로, 복수의 기록에 대해 일련의 처리과정을 반복하여 수행해야 하는 대량처리의 경우 워크플로우 엔진에 각각의 처리작업을 정의하여 순차적으로 작업대기열에 넣어둠으로써 작업에 대한 수행 과정과 결과를 용이하게 추적할 수 있다.

대량의 기록정보를 관리하는 기록시스템은 조회가 자주 일어나는 데이터군을 분석하여 데이터베이스에 인덱스를 생성하거나 클러스터링,

반정규화 등의 기법을 사용하여 조회속도를 높여줘야 한다. 성능을 고려한 이러한 조치는 조회 속도는 높여주는 반면에 위에서 살펴본 대량처리와 같이 데이터의 입력, 변경, 삭제 등의 조작 속도는 느려지게 하는 트레이드오프(Trade-Off)가 있다. 따라서 특별한 대량처리가 필요한 시점에 해당 작업의 성능을 제고하기 위해 일시적으로 데이터베이스 내의 인덱스를 삭제하거나 비활성상태로 전환하는 등의 조치를 취할 수 있다.

4. 처리단계별 디지털저장소 형상 제안

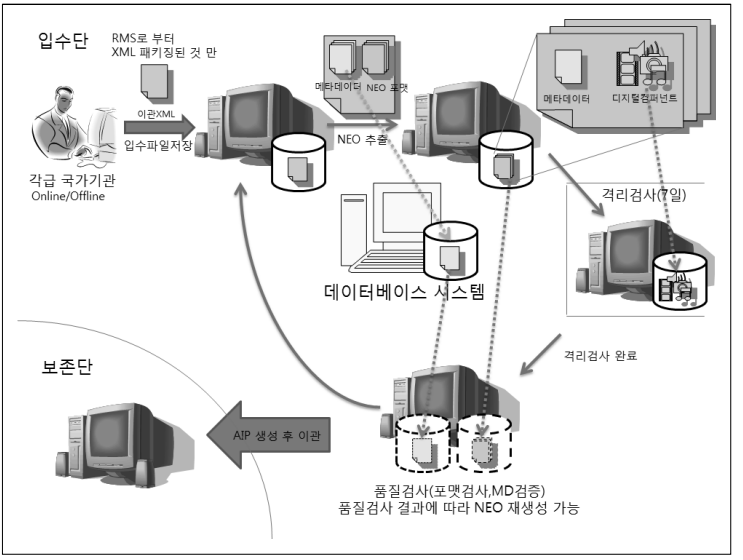
이 장에서는 영구기록물관리 3단계별로 디지털저장소가 어떻게 배치되어야 하는지 관해 국가기록원을 사례로 제안해보고자 한다. 2장에서 전제한 바대로 디지털컴포넌트저장소와 데이터저장소를 별도로 배치하되 3장에서 추정한 기록정보의 양을 처리하는데 필요한 개수를 제시하고자 한다. 여기서 저장소는 디지털 데이터가 보관되는 스토리지를 중심으로 하며 저장매체만으로도 가능하고 때론 데이터를 처리하기 위한 관리소프트웨어가 포함된 서버의 하위구성요소로 구성되는 것으로 전제한다.

1) 입수의 병렬 처리

입수처리에 필요한 저장소 개수를 결정하는데 고려해야 할 사항을 점검해보면 다음과 같다. 디지털 기록정보의 이관 방식은 온라인 이송과 오프라인 이송이 모두 가능해야 하며 영구기록물관리 기관 입장에서는 동시에 많은 양의 기록정보가 한꺼번에 이송되지 않도록 부하를 분산해 입수계획을 세워야 한다. 계획에는 몇 개의 서버를 두고 입수처

리할 것인지 포함해야 하며, 구체적으로는 디지털컴포넌트저장소와 데이터저장소의 개수가 정해져야 한다. 온라인 이송의 경우에는 생산기관과 영구기록물관리기관 간의 네트워크 경로별로 용량과 품질을 확인해야 한다. 입수는 해마다 이루어지므로 일 년 내에 일 년치의 기록정보를 모두 입수하는데 충분한 서버를 갖추는 것은 최소의 요건이 된다.

〈그림 8〉 입수단의 처리절차



입수단의 시스템 형상에 대해 제안을 하기 위해서는 〈그림 8〉에서 다음과 같은 몇 가지 전제가 필요하다.

첫째, 기록정보의 이관량은 2015년 기준으로 약 2.5테라바이트이며 이관 XML파일 파싱을 위해 동일한 양이 더 요구되는 것으로 가정한다. 매년약 10% 정도씩 이관 기록정보 양이 증가할 것으로 예상한다.

둘째, 네트워크 전송 속도는 3.3Mbps로 가정한다. 2011년 성남-대전 간 망의 속도를 테스트해본 결과 평균 속도가 26.7Mbps로 측정되었다.

32GB의 테스트 데이터를 23차례 전송한 평균 시간이 2시간 45분이었으며 이는 약 3.3MB/s에 해당한다. 향후 각 생산기관에서 국가기록원으로 온라인 이관을 하는 경우 네트워크 전송 속도는 망 부하나 품질에 따라 속도에 편차가 있을 것이나 평균적으로 3.3MB/s를 기준으로 계산하도록 한다.

셋째, 입수처리 서버의 하루 처리량은 100GB로 가정한다. 서버의 처리량은 CPU, 메모리와 같은 하드웨어의 성능에 좌우되며 이는 서버의 구매비에 관련된다. 고성능의 메인프레임급 서버 한 대를 둘 것인가 낮은 성능의 여러 서버를 둘 것인가를 고려했을 때 TCO(Total Cost Ownership)가 낮은 후자의 방향으로 설계하도록 한다.

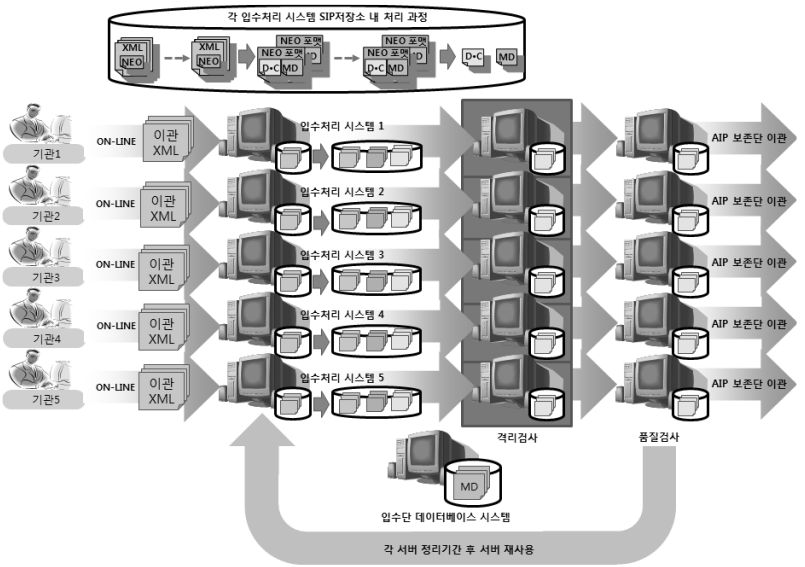
넷째, 2015년 온라인/오프라인 이관의 비율을 7대 3으로 가정한다. RMS표준모델을 구축한 기관의 경우 대용량 송수신 모듈을 장착하면 온라인 이관이 가능하다. 모듈 구축이 늦어지거나 네트워크의 부하로 인해 온라인 이관이 불가해질 경우를 고려하여 약 30% 정도는 오프라인으로 이관하는 것으로 계산한다.

다섯째, 입수단에서 한 해의 기록정보 이관을 8개월 안에 마치는 것으로 목표를 정한다. 이는 이관된 기록정보에 오류가 발생하여 생산기관에 재전송을 의뢰하는 오류수정 보완기간을 모두 합한 기간이다. 생산기관에 오류를 통보하고 관련 기록정보를 재전송받기까지 한 번의 피드백에 소요되는 시간은 20일로 가정한다.

여섯째, 디지털컴포넌트의 바이러스 격리검사를 위해 필요한 기간은 7일로 가정한다.[8] 영구기록물관리기관에서는 신형 바이러스의 동향을 주시하여 격리검사의 필요성과 격리 기간에 대한 정확한 정보를 취득해야 한다.

8) 〈그림 2〉에서는 격리기간을 15일로 표시하였다. 현재 7일 이내에 잠복한 바이러스를 검출할 수 있다는 바이러스 전문가의 견해에 따라 7일로 가정해보았다.

〈그림 9〉 5개 입수처리 서버를 배치한 형상



위의 전제하에 입수단에 필요한 서버의 개수를 산출해 보면 최소한 5개가 필요하다는 계산이 나온다. 이 5개의 서버는 디지털컴포넌트저장소를 중심에 둔 서버이다. 8개월 이내에 입수처리를 종료하기 위해서는 240일 내에 한 해의 입수처리가 끝나야 한다. 따라서 총 2.5테라바이트의 기록정보를 처리하기 위해서는 한 번에 100기가바이트의 기록정보를 처리할 수 있는 서버를 최소한 5대 갖추어 5회에 걸쳐 처리해야 한다. 이때 각 회마다 45일 이내에 처리를 마쳐야 하며, 각 회차 간에 데이터를 정리하는 기간은 3일을 둘 수 있다. 45일이라는 시간은 입수 기록정보에 문제가 발생하여 최소한 한 번 이상 재전송받는데 걸리는 시간까지 포함한 것이다. 하지만, 기록정보의 품질에 문제가 많아 여러 번 재전송을 받아야 한다면 8개월 이내에 입수처리의 종료가 어려워질 수도 있다.

입수단에는 한 개의 입수단데이터저장소가 필요하며 이 저장소에는 5개의 입수서버가 5번 회전하면서 처리하는 입수기록 전체에 대한 메타데이터와 처리이력정보가 누적하여 저장된다. 이 저장소는 입수과정 전체를 통제하기 위한 시스템의 역할을 해야 하므로 한 해의 입수량 전체에 관한 데이터를 모두 저장할 수 있는 데이터베이스관리시스템과 데이터베이스 설계가 필요하다.

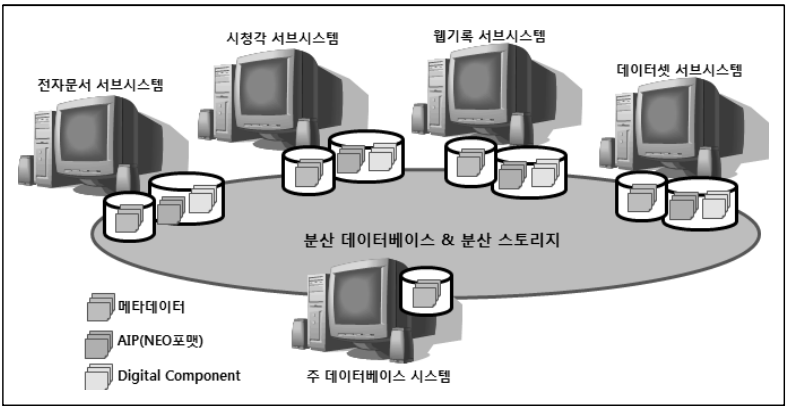
2) 기록유형별 저장과 복본 분산 저장

이 논문에서는 보존단 저장소를 AIP유형별로 구분하여 배치할 것을 일관되게 제안해 왔다. 즉, 〈그림 10〉에서 보는 바와 같이 전자문서보존서브시스템, 시청각보존서브시스템, 데이터세트보존서브시스템, 웹기록보존서브시스템 등 기록정보의 유형별로 저장소를 구성하는 것이다. 이 구성의 장점은 AIP유형별로 보존처리가 용이해진다는 것이다. 예를 들어, 시청각기록의 경우 보존단에서 동영상 멀티미티어 기록정보를 실행해 볼 수 있어야 하며, 의미 단위로 나누어 기술할 수 있어야 한다. 데이터세트기록의 경우 내용정보에 대해 질의가 가능하도록 저장해야 할 것이다. 웹기록의 경우 페이지 간 하이퍼링크가 작동하도록 연결하여 저장해야 할 것이다. 결국 AIP유형별 서브시스템에는 해당 유형의 기록정보를 다루기 위한 특수한 소프트웨어와 하드웨어를 설치하고 성능을 튜닝해주어야 한다. AIP유형별로 구분하여 서브시스템을 구성하는 것을 전제로 했을 때, 2015년에 전자문서보존서브시스템에는 2.5테라바이트의 저장소가 필요하게 된다.

각 서브시스템의 데이터저장소에는 기록정보 유형별 특수한 메타데이터를 저장하여 관리하며, 모든 유형의 기록정보에 대한 공통 메타데이터는 유형별 서브시스템과 별도의 주 데이터베이스시스템에 저장하도록 한다. 주 데이터베이스시스템에는 보존대상 기록정보 전체에 관

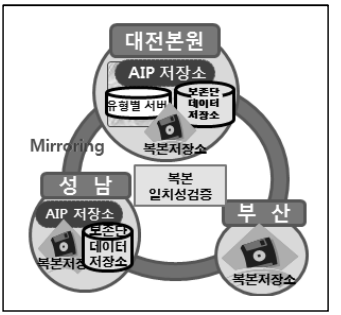
한 감사증적정보와 로그정보를 저장한다. 이때, 기록유형별 서브시스템의 데이터와 주 데이터베이스시스템의 데이터는 분산데이터베이스로 구축하여 통합하도록 한다. 결과적으로 입수단으로 들어오는 데이터 52.6기가바이트는 서브시스템과 주 데이터베이스시스템에 일부는 중복되면서 분산 저장된다.

〈그림 10〉 기록의 유형별 보존서버를 배치한 형상



보존단의 배치 형상을 결정하는 중요한 요소는 재난복구계획이다. 현재 국가기록원은 대전 본원에 영구기록물관리시스템을 두어 운영하고 있으며, 성남 분원에 미러링 사이트를 구축하여 운영 중이다. 여기에 기록정보의 복본을 별도의 저장매체에 3부를 더 제작하여 〈그림 11〉과 같이 대전 본원, 성남 분원, 부산 분원에 분산저장하려는 계획을 갖고 있다. 3장에서는 복본을 하나 제작하는 것으로 보존 기록정보의 양을 추산했으나 만약 세 개로 확정된다면 저장소 용량 산정결과가 달라져야 할 것이다.

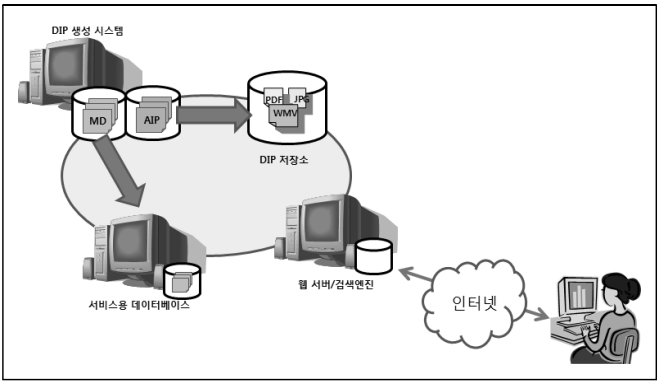
〈그림 11〉 복본 전자기록의 분산 저장 형상



3) 이용자 서비스를 위한 제공처리

정책적으로 공표하기로 결정한 기록정보 AIP는 입수완료 즉시 보존단에 저장됨과 동시에 제공단으로도 이송되어야 한다. 특정 기록정보에 대해 이용자가 열람을 요청한 경우에는 제공단에서 보존단 쪽으로 AIP전송을 요청한 후 이송받은 AIP와 메타데이터를 제공단에 임시저장할 수 있어야 한다. AIP가 이송될 때 관련 메타데이터 중 일부가 보존단 데이터저장소로부터 제공단 데이터저장소로 함께 이송되어야 한다.

〈그림 12〉 이용자에게 기록을 제공하는 형상



〈그림 12〉에서 보는 것처럼, 제공단에는 이용자에게 제공할 DIP를 제작하기 위한 DIP 생성시스템과 그 결과를 저장하기 위한 DIP저장소가 배치되어야 한다. 또한, DIP가 제작되면 색인을 생성하여 서비스용 데이터베이스에 저장하고 이를 이용자가 접근할 수 있는 웹서버나 검색엔진에 연계해주어야 한다. 서비스용 데이터베이스에는 모든 기록정보의 기술정보가 탑재되어 있어 이용자가 검색을 통해 원하는 기록정보를 찾아볼 수 있어야 한다.

5. 맺음말

이 논문에서는 영구기록물관리시스템에서 '대량기록물'을 효율적으로 처리하기 위해 입수단, 보존단, 제공단별로 데이터저장소와 디지털컴포넌트저장소를 어떻게 배치할 지에 대해 국가기록원의 2015년을 사례로 탐구해 보았다. 2015년의 상황에 대한 여러 전제와 산술식을 동원하여 입수 기록정보의 양을 추산하고 입수처리 서버를 5개 배치하여 8개월 안에 입수를 종료한다는 시나리오를 제시해 보았다. 이러한 시도는 비록 2015년도의 정확한 수치는 아닐지라도 영구기록관리기관시스템 입수단의 특성을 밝히고 설계요건을 구체화하기 위한 출발점으로 중요한 의미를 갖는다. 입수량의 정확한 사이징을 통해 시스템의 가용성을 최대화하는 방식으로 시스템을 구축하여 구매비용을 절감할 수 있으며, 연도별 누적량의 정확한 사이징을 통해 시스템의 확장 모형을 최적으로 설계하여 구축할 수 있다. 국가기록원을 포함한 영구기록물관리기관들은 기록정보의 입수주기, 생산기관의 수, 기록정보의 유형과 수량 등 디지털저장소를 설계할 때 고려해야 하는 주요 입력파라미터와 이들을 입력값으로 하는 정밀한 용량추산 산술식을 정의하고 적용함으로

써 기관에 필요한 디지털저장소의 형상배치를 과학적으로 구상해 나가야 할 것이다.

'대량기록물'의 양적 측면에 중점을 둔 이 논문에서는 디지털저장소 설계에 필요한 여러 전제를 가정하였다. 지나치게 낙관적일수도, 비관적일 수도 있는 전제와 가정일 수 있다. 이와 관련하여 향후 2015년을 대비하기 위해 점검해야 할 사항을 몇 가지 추가 제시하고자 한다.

첫째, 디지털 기록정보 품질의 문제이다. 2015년 이관이 관심의 대상이 되는 이유 중 하나는 '대량기록물'을 다루어야 한다는데 있다. 최근 여러 분야에서 빅데이터에 많은 관심이 쏠리고 있다. 각종 정보시스템을 통해 생산되는 대용량 데이터를 새로운 가치를 가진 자원으로 만들기 위해 빅데이터를 용이하게 조직화하고 관리하는 기술에 관심과 투자가 이루어지고 있다. 그러나 현재 빅데이터에 관한 최대 이슈는 대량의 데이터를 다루는데 소요되는 컴퓨팅기술보다 데이터의 저품질에 있다는 점에 주목할 필요가 있다. 이는 '대량기록물'을 다루는 공공부문의 기록관리기관에 시사하는 바가 크다. 2장에서 살펴본 기록정보의 처리절차, 3장에서 살펴본 데이터 양에 대한 추정, 그리고 4장에서 살펴본 저장소의 배치는 국가기록원에 입수되는 기록정보의 품질이 약 5%이내에서만 수정보완이 필요한 수준으로 가정한 것이었다. 만약, 생산기관에서 입수되는 기록정보의 품질이 이보다 훨씬 낮은 수준이라면 국가기록원의 2015년 이관은 대란으로 치달을 수도 있다. 국가기록원은 매년 정해진 기간 내에 정해진 양의 입수를 완료해야만 하는데 품질검증 과정에 소요될 시간과 노력의 규모를 예측하기 어렵기 때문이다. 2015년 이관을 대비하기 위해서는 이런 불확실성을 제거하는 것이 급선무이다.

둘째, 각급 기록관리기관의 디지털 능력에 관한 것이다. 이관대상 기록정보의 품질을 확인하는 과제는 국가기록원과 생산기관 공통의 과제이다. 국가기록원은 그간 디지털 기록정보의 생산, 관리, 보존 프로세스

를 안정화하고 각급 기관에서 사용할 RMS(Records Management Systems)를 개발하여 보급하는 등 여러 노력을 기울여 왔다. 그러나 국가기록원의 노력만으로는 2015년의 성공을 기대할 수 없다. 각급 기록관리기관에서 단지 기록관리시스템을 도입했다고 이관이 대비되는 것이 아니며, 품질이 검증된 디지털 기록정보를 국가기록원에 이관할 수 있느냐가 성공의 관건이기 때문이다. 기록정보 품질을 검증하기 위해서는 디지털 정보의 속성을 알아야 하고 이를 시스템을 통해 보완하고 통제할 수 있어야 한다. 기록관리기관이 스스로 디지털 기록정보에 대한 지식수준을 높이고 기관의 IT부서와 긴밀히 협업하여 디지털 기록정보 관리 전반에 걸쳐 통제력을 높여가야 할 것이다.

셋째, 차세대 디지털 기록정보 인프라의 방향이다. 대용량, 분산보존, 클라우드 컴퓨팅 기술이 날로 발전하고 있으며 기록관리를 위한 인프라도 이러한 기술환경에 조응하여 미래의 설계 방안을 모색해볼 때가 되었다. 예를 들어, 디지털 기록정보를 물리적 이관으로 처리하는 것은 정보의 이송과 복사 과정에서 오류가능성이 상존한다는 점과 디지털 정보의 위치투명성 관리라는 효율성을 포기한 처리방식으로 기록연속체론과도 거리가 멀다고 볼 수 있다. 대안으로 각급 기록관리기관에서 장기보존 대상 디지털 기록정보가 생산되면 별도의 '장기디지털저장소'에 따로 저장하도록 하고, 이관시점에 해당 저장매체에 대한 관리권만 영구기록물관리기관으로 넘김으로써 기록정보의 물리적인 위치 변동 없이 논리적으로 이관을 처리할 수 있다. 디지털 기록정보를 디지털 본래의 특성에 맞게 현재의 가장 안정화되고 미래지향적인 컴퓨팅 환경에 맞춰서 이관처리하는 방안을 모색해야 할 것이다.

ABSTRACT

A study on configuring deployment of digital repositories for the archives management systems

Yim, Jin-Hee · Lee, Dae-Wook

The National Archives of Korea(NAK) has a mission to ingest large-scaled digital records and information from a number of different government agencies annually from 2015. There are important issues related to the digital records and information transfer between NAK and agencies, and one of them is how to configure deployment of digital repositories for the archives management systems. The purpose of this paper is to offer the way to design it by examining the checkpoints through the whole life cycle of digital records and information in the archives management systems and calculating the amount of ingested digital records and information to the systems in 2015 and deploying the digital repositories configured according to the amount the records and information.

Firstly, this paper suggests that the archives management systems in NAK should be considered and examined into at least three different parts called Ingest tier, Preservation tier and Access tier in aspects to the characteristics of the flow and process of the digital records and information. Secondly, as a results of the calculation the amount of the digital records and information ingested to the archives management systems in 2015 is sum up to around 2.5 Tera bytes. This research draws several requirements

related to the large-scaled data and bulk operations which should be satisfied by the database or database management system implemented on to the archives management systems. Thirdly, this paper configures digital repositories deployment according to the characteristics of the three tiers respectively.

This research triggers discussion in depth and gives specific clues about how to design the digital repositories in the archives management systems for preparing the year of 2015.

Key words : Large-scaled records and information transfer, Archives management systems, Digital repository, system deployment, Electronic records