

# 데이터세트 기록의 관리 방안

## A Study on the Management of Dataset as Records

현 문 수(Moonsoo Hyun)\*

### 목 차

1. 서론	3. 해외 데이터세트 기록의 관리 사례 분석
1.1 연구의 배경	3.1 NDAD
1.2 연구의 범위와 방법	3.2 AAD
2. 데이터세트 기록의 관리	3.3 NDAD와 AAD 조직 및 기술 비교
2.1 데이터세트 기록의 특성	4. 결론 및 제언
2.2 데이터세트 기록의 관리 필요성	

### <초 록>

정부기관 및 기업에서는 다양한 데이터세트가 생산된다. 업무활동의 일부로 생산되는 데이터세트는 현재 단순히 정보시스템 내에 저장되어 있고, 기록으로 획득되어 관리되고 있지 못하다. 원 데이터로서 업무활동에서 발생한 데이터세트 기록은, 시급히 기록관리 영역으로 편입되어 관리되지 않는다면, 오래지 않아 사장될 것이다. 데이터세트는 정부 정책 및 법제도 형성에 있어 기본 데이터로 활용 가능할 뿐 아니라 정책결정과정을 보여주는 증거로서 데이터세트 기록의 관리와 보존은 반드시 필요하다. 따라서 본 연구는 기록으로서의 데이터세트 관리 필요성을 지적하고 해외의 사례를 분석하여 데이터세트 기록관리의 기초 자료를 제공하고자 하였다.

주제어: 데이터세트, 데이터세트 아카이빙, 데이터세트 기록

### <ABSTRACT>

The purpose of the study is to propose the necessity of management and long-term preservation of dataset as records. Although government and corporate bodies produce various dataset in the regular course of the business, dataset have been stored and managed in the information system. Dataset as records should be captured into the record management system and managed in the overall system. They can provide a evidence of the decision-making process of the government and fundamental information of the process. If agents do not perform the right management, dataset records will disappear in the future.

Key words: dataset, dataset archiving, dataset record

\* 한국외국어대학교 정보·기록관리학과 박사과정(puremoon@hanmail.net)

## 1. 서론

### 1.1 연구의 배경

다양한 전자기록이 생산되고 있다. 업무활동에서 생산되고, 업무활동의 증거로 보존되어야 하는 이메일, 웹, 데이터세트 등, 텍스트 형태를 벗어난 기록 유형이 증가하고 있다. 이에 따라, 과거 기록관리 영역 외에 있었던 비텍스트 형태 기록에 대한 관심 역시 증가하고 있다. 우리나라의 경우 역시 이메일이나 웹을 기록관리 영역 안에서 보존하고 관리하려는 움직임이 상당하다. 반면, 데이터세트 기록의 경우 그 인식이 낮을 뿐 아니라 기록관리학 분야 밖의 문제로 여겨지고 있는 실정이다.

2005년 발표된 국가기록관리혁신 로드맵은 기록관리 혁신과제 중 행정정보시스템에서 생산된 데이터세트를 기록화하는 방안을 내세우고 있다. 청와대 기록관리 혁신 계획에서도 개별 업무시스템에서 생산된 데이터세트를 기록으로 획득하여 관리하기 위한 방안을 제기하고 있다. 이처럼 전 국가차원에서는 데이터세트를 기록의 한 유형으로 정의하고 기록관리체계 안에서 관리하려는 시도가 이루어지고 있다. 반면, 개별 정부기관이나 기업차원에서는 기록으로서의 데이터세트 보존과 관리에 대한 관심이 상대적으로 저조하다.

데이터베이스는 내용을 쉽게 접근하여 처리하고 갱신할 수 있도록 구성된 데이터의 집합이다. 컴퓨터의 사용으로 인해 인간은 대량의 데이터를 손쉽게 처리하고 구조화하여 저장할 수 있게 되었다. 구조화된 데이터 집합인 데이터세트는 데이터베이스에 저장되며 데이터베

이스가 속한 정보시스템 안에 존재한다. 이로 인해 데이터세트는 기록으로 관리하고 보존하기보다는 정보시스템의 관리 범위 안에서 백업 등의 다소 소극적으로 관리되어왔다.

그런데 최근 인구통계, 항공우주과학, 천문기상, 유전자공학 등 다양한 과학기술 및 사회과학분야에서 관찰 및 실험 데이터를 구축하여 공유하고 장기보존하기 위한 움직임이 진행되고 있다. ERPANET이나 영국 The National Archives(이하 TNA), 미국 National Archives & Records Administration(이하 NARA) 등 해외 기관들의 경우 데이터베이스 보존과 관련하여 논의를 시작했으며, 일정부분 성과를 발표하고 있다. 이들이 관심을 가지고 보존 전략을 연구하는 대상 역시 관계형 데이터베이스이다. 그러나 관계형 데이터베이스의 보존은 복잡한 문제를 지니고 있어, TNA나 NARA에서 구축하고 있는 데이터세트 아카이브와 같이 데이터베이스 안에 저장되었던 데이터세트를 대상으로 보존 및 이용을 보장하는 방향이 우선적으로 취해지고 있다.

본 연구에서는 기록관리의 대상으로 데이터세트를 인식하고 관리할 필요성을 지적하고, 진본 데이터세트 보존을 위한 기술(description)과 이의 서비스를 중심으로 살펴보고자 한다. 우선 기록으로서의 데이터세트는 무엇이며, 어떤 특성을 지니고 있는지 분석한다. 둘째, 데이터세트를 기록으로 획득하고 관리해야 할 필요성을 제시한다. 셋째, 데이터세트 기록을 진본 상태로 보존하기 위한 요건을 제시한다. 넷째, 영국과 미국의 국립기록관에서의 데이터세트 관리 및 서비스 사례를 비교분석함으로써 우리나라 데이터세트 기록의 관리 방향 모색에 도

움이 되고자 한다.

## 1.2 연구의 범위와 방법

첫째, 국내외 문헌과 보고서, 해외 데이터세트 보존 기관의 홈페이지 정보 등을 통해 기록으로서 획득·관리되어야 하는 대상으로서 데이터세트의 개념과 특성을 파악하였다.

둘째, 해외 데이터세트 보존 프로젝트, 우리나라의 국가기록관리혁신로드맵 및 대통령비서실 기록관리혁신을 위한 정보화 전략 계획 수립 보고서 등을 통해 데이터세트 기록의 관리 필요성을 제시하였다. 또한 진본 데이터세트 관리 및 보존을 위해 기존에 제안된 요구조건을 분석하였다.

셋째, 영국 TNA에서 구축·서비스하고 있는 National Digital Archive of Datasets(NDAD)와 미국 NARA에서 구축하여 서비스하고 있는 Access to Archival Database(AAD)를 비교 분석을 통해 국가 차원의 데이터세트 관리 및 서비스 사례를 분석하였다.

## 2. 데이터세트 기록의 관리

### 2.1 데이터세트 기록의 특성

#### 2.1.1 데이터세트 기록

데이터세트는, 컴퓨터 하드웨어와 소프트웨어의 잦은 갱신으로 인해 기존 데이터를 새로운 환경에서 읽을 수 없는 문제를 해결하기 위해, 데이터베이스에서 그 데이터만을 추출해 보관하는 원래의 데이터 묶음이다(Shepherd &

Smith 2000). 즉 데이터세트는 데이터베이스에 축적된 가공되지 않은 데이터의 집합이다. 컴퓨터와 특정 소프트웨어를 이용해 수집되고 처리된 설문조사결과, 관찰결과, 인구조사결과, 재고조사결과물 등 다량의 데이터들은 묶음으로 저장되며, 이러한 하나의 데이터세트는 동일한 업무활동의 결과물인 동시에 동일한 축적의 결과로 데이터세트 기록을 구성한다.

#### 2.1.2 데이터세트의 특징

장기보존과 관리를 위한 데이터세트의 추출은 원 데이터가 저장되어 있는 데이터베이스 유형에 따라 달라진다. 데이터베이스의 유형은 관계형(Relational) 데이터베이스, 계층형 데이터베이스, Native XML 데이터베이스, 객체형 데이터베이스, 네트워크형 데이터베이스 등 매우 다양하나, 관계형 데이터베이스가 다른 유형의 데이터베이스에 비해 널리 이용되고 있다. 따라서 데이터베이스 보존을 언급할 때, 이는 관계형 데이터베이스를 대상으로 하며, 특히 관계형 데이터베이스는 대규모 정부 시스템에서 이용되고 있기에 더욱 그러하다(Verdegem 2003).

한편, 데이터베이스는 개방형(open)과 폐쇄형(closed), 그리고 정적(static)과 동적(dynamic)인 것으로 나뉘며(Shepherd and Smith 2000), 이 네 가지 특성의 조합에 따라 데이터베이스 시스템의 보존 전략도 필수적으로 달라진다(Ashley 2004). 예를 들어 폐쇄형이면서 정적인 시스템이라면 데이터 입력이 완료되는 시점을 기다려 데이터베이스 시스템을 보존하면 문제는 비교적 간단하다. 개방형이면서 정적인 시스템 보존도 일정 시점마다 스냅샷을 해주는 방법이 있으므로, 그다지 복잡하지는 않다.

〈표 1〉 데이터세트 유형에 따른 보존 전략(Ashley 2004; Shepherd 2003)

	일회성 프로젝트 산출 데이터세트	연속적이며 정적(static)인 DB	종료시점이 부정기적인 동적(dynamic) DB	수정되지 않으며 열려 있는 DB
특징	<ul style="list-style-type: none"> <li>• 일정기간동안 수집 후 닫힘</li> </ul>	<ul style="list-style-type: none"> <li>• 특정 수집 기간동안 수집된 데이터세트</li> <li>• 종결된 데이터세트에는 데이터 추가불가</li> </ul>	<ul style="list-style-type: none"> <li>• 오래된 데이터 갱신</li> <li>• 정기적·지속적인 데이터 수정</li> </ul>	<ul style="list-style-type: none"> <li>• 미리 정한 데이터 수집기간 없음</li> <li>• 지속적 데이터 추가</li> </ul>
보존	단일 묶음 처리	일정 주기별 묶음 처리	스냅샷/로그파일	일정 주기별 전체 DB 스냅샷
예	1291~1292년간 세금 부과 데이터세트	3년 주기로 3번 설문조사한 런던 대 환경학부 학생 태도 데이터세트	개인 예금계좌용 은행 시스템의 계정 잔액	개인 예금계좌의 입출력 내역

개방형이든 폐쇄형이든 정적 데이터베이스 시스템의 경우, 보존 문제는 비교적 간단하다. 문제는 동적 데이터베이스 시스템이다. ERPANET 워크샵(2003)에서 언급된 데이터세트 장기보존에서 나타난 문제점 중 하나 역시 개방형 데이터베이스 혹은 동적 데이터베이스의 보존 문제였다. 계속 이용되고 있는 데이터베이스, 특히 정기적·비정기적인 갱신과 삭제가 이루어지는 데이터베이스의 경우, 이를 그대로 아카이브에 전송한다면, 전송된 데이터세트와 원본 데이터세트가 달라지는 경우가 발생하게 된다. 현재까지 가장 많이 이용되는 방법은 정적 데이터베이스 시스템 보존에도 이용되는 스냅샷 방법인데, 동적 데이터베이스 시스템의 경우, 시스템의 변화 주기에 따라 스냅샷의 실행 주기도 바뀌게 된다. 정기적인 변화주기가 없는 시스템에서 정기적인 스냅샷의 실행은 보존 데이터의 왜곡을 낳을 수 있으며, 또한 스냅샷 주기가 짧아질수록 비용이 증가하고, 데이터 중복의 정도도 증가하게 된다.

2.1.3 데이터세트 기록의 최소 관리 요소

현재 각국에서는, 데이터베이스에 담긴 기록을 진본 상태로 보존하기 위한 다양한 시도와

실험들을 진행하고 있다. 이를 성공적으로 장기보존 하기 위해서는 기본적으로 보존된 데이터베이스의 무결성, 진본성, 접근성, 명료성, 원본성을 지속적으로 유지하며 데이터를 보호해야 한다(ERPANET 2003). DPT(2003)에 따르면 데이터베이스의 진본성을 유지하며 보존하기 위해서는, 배경정보, 내용, 구조, 외형(Appearance), 행위(Behavior), 그리고 조직에서 이루어지고 있는 업무과정까지 저장되어야 한다고 말하고 있다.

데이터베이스의 어떤 요소를 관리하고 보존하느냐의 문제는 시스템을 운용하고 있는 기관에 따라 달라진다(Verdegem 2003). 따라서 현재는 데이터베이스의 특징을 규명하기 위해서 최소한 보존되어야 하는 요소를 파악하고 이를 관리하는 방안이 가장 현실적이다. DPT(2003)에서 밝히고 있는 데이터베이스 진본성 요건을 위한 요소 중 배경정보와 내용, 구조, 외형, 행위에 대한 정보를 대상으로 최소 관리 요건을 살펴보면, 첫째, 배경정보의 경우 조직의 명칭과 업무, 업무과정, 기관 및 조직의 배경정보, 데이터베이스의 기술적(technical) 배경정보, 분류코드, 다른 기록과의 관계, 보존행위 및 그 결과 등을 관리하도록 하고 있다. 둘째, 내용에

있어서는 해당 데이터베이스 내의 테이블에 저장된 실제 내용을 관리하도록 한다. 셋째, 구조는 테이블간의 관계나 필드 속성 등의 물리적 구조는 물론, 데이터 사이의 관계를 나타내 주는 ERD(Entity Relationship Diagram)나 XML 스키나 형태의 논리적 데이터 구조를 함께 관리하도록 한다. 넷째, 외형에는 디스플레이되는 방식을, 그리고 다섯째, 사용자 응용프로그램에서 일어나는 행위까지 관리하도록 제시하고 있다(DPT 2003).

어떤 관리 요소를 선택하는가의 문제의 답은 데이터베이스 시스템이 존재하는 기관이나 조직에 따라 다르며, 아직까지 데이터베이스 구조정보와 데이터베이스 운영프로그램과 응용프로그램 등 전체 데이터베이스 시스템의 다양한 특징들을 모두 보존할 뚜렷한 방안이 제시되고 있진 않다. 결국 최소한 데이터베이스의 내용은 보존되어야 한다는 것이 현재까지의 결론이다(Verdegem 2003). 데이터베이스 보존에 있어서 영국의 NDAD나 미국의 AAD 등이 데이터세트를 중심으로 국가 차원의 서비스를 제공하고 있고, 사회과학분야나 과학기술분야의 데이터세트 장기보존 논의가 우선적으로 이루어지는 것도 유사한 이유 때문일 것이다.

## 2.2 데이터세트 기록의 관리 필요성

정부 및 기관의 업무활동은 문서형태의 기록 외에 다양한 유형의 기록을 생산한다. 컴퓨터와 특정 소프트웨어를 이용해 수집되고 처리되는 데이터세트도 이에 포함된다. 따라서 정부와 기관의 업무활동을 보여줄 수 있는 데이터세트는 문서나 이메일 등 여타 기록과 동일하게 장기

보존되고 미래의 이용이 보장되어야 한다.

2005년 10월 정부혁신지방분권위원회 기록관리혁신 T/F 협의회(이하 혁신위)는 ‘국가기록관리혁신 로드맵’을 발표했다. 로드맵에 나타난 첫 번째 아젠다인 “공공업무 수행의 철저한 기록화”에는, “행정정보시스템을 통해 생산된 데이터세트(Dataset)를 기록으로 포착, 관리”한다는 혁신 방안이 제시되어 있다(정부혁신지방분권위원회 2005, 18). 2005년 종료된 대통령 비서실 ‘기록관리 혁신을 위한 정보화 전략 계획 수립’ 보고서 역시 미래 모형 설계의 한 장을 “개별업무시스템에서 생산된 데이터세트의 기록화 방안”으로 구성하여 데이터세트 기록 관리를 제안하고 있다(대통령비서실 2005). 혁신위와 대통령 비서실을 중심으로 이루어지고 있는 기록관리 혁신에 데이터세트 기록의 보존과 관리에 대한 인식 변화가 일부이나마 반영된 것이다.

해외의 경우 수년 전부터 데이터베이스 및 데이터세트 관리 및 장기 보존에 관한 논의가 활발히 이루어지고 있다. 2003년 4월 스위스 연방 아카이브에서 데이터베이스 장기 보존을 위한 ERPANET 워크숍이 진행되었다. 프랑스의 CONSTANCE(CONSerVation et Traitements des Archives Nouvelles Constituées per l'Electronique: Preservation and Treatment of New Archives Derived from Computer Processing), 벨기에의 DAVID 프로젝트(Digital Archiving in Flemish Institutions and Administrations), 네덜란드의 DPT(Digital Preservation Testbed), 영국의 NDAD(National Digital Archive of Datasets), 스위스의 SIARD (Software Invariant Archiving of Relational

Databases)를 포함, 미국과 유럽지역 16개국의 데이터베이스 보존 프로젝트가 참가하여, 데이터베이스의 보존 정책과 전략, 보존과정, 데이터베이스 평가, 다큐멘테이션, 보존포맷, 접근, 보존기술, 효과적인 보존을 위한 협력 등을 논의했다(ERPANET 2003).

국내에서는 이소연(2002, 99)의 연구에서 전자기록의 기술을 언급하면서 특수매체 형태의 기록으로 데이터세트를 언급하고 있으며, 2004년 연구에서는 디지털유산의 장기 보존 정책 수립을 제안하면서 데이터세트의 장기보존 필요성을 말하고 있다. 더 나아가 설문원 외(2005b, 78-85)는 국가 차원의 디지털 아카이빙 체제 구축에 있어, 아카이빙 대상 매체 중 하나로 데이터베이스 보존의 필요성을 밝히고 있다.

그러나 실제 기록 관리 주체인 정부 기관들이 데이터세트 기록 관리에 기울이는 관심과 노력은 극히 제한적이다. 앞서 언급한 '국가기록관리혁신 로드맵'에 포함된 혁신 방안으로 데이터세트를 기록으로 인식하고 이를 등록해야 한다는 사항, 그리고 대통령 비서실을 대상으로 한 기록관리혁신 ISP 보고서에서 "개별 업무시스템 데이터세트 획득 방안"이 찾아볼 수 있는 전부이다. 이런 위에서부터의 혁신이 정부 부문 전반으로 확산될 것으로 여겨지나, 아직까지는 장기보존이나 구체적인 관리 방법까지 고려하고 있지 않을 뿐 아니라, 이런 움직임을 제외하면, 데이터세트 기록에 대한 관심은 미미한 수준이다.

인사기록이나 회계기록, 매일 상당량의 데이터가 축적되는 기상관측 데이터나 항공·위성 관측 데이터, 실험데이터 등 과거 아날로그 형

태 기록으로 관리되고 보존되던 데이터들이 현재는 데이터베이스에 저장되어 각종 정보시스템 안에서 관리되고 있다. 이에 대해 조직 내부의 행정정보시스템 등 각종 정보시스템 저장영역에 보존되어 관리되면서 백업 수준의 보존 조치만이 이루어지며, 기록으로 획득되지 못하고 있다(설문원 외 2005a, 27). 정부 및 각 기관의 업무활동에서 생산되는 문서나 이메일 등 여타 기록과 동일하게 업무활동의 결과로 생산된 데이터세트 역시 기록으로 인식해야 한다. 따라서 정부 및 기관 업무활동을 뒷받침하는 증거인 기록으로서 데이터세트를 관리하고 장기 보존함으로써, 미래의 이용을 보장해야 한다. 그러나 전반적인 인식 제고조차 이루어지지 않은 상태이므로 구체적인 관리 및 보존 방안을 수립하기 위해서는 많은 시간과 노력이 필요할 것이다.

문서형태나 이미지, 이메일 기록 등에 비해 데이터세트의 관리에 대한 관심이 저조한 것은 데이터세트가 저장되어 있는 데이터베이스의 특징에서 찾아볼 수 있다. 보존해야 하는 대상이 비교적 명확한 다른 기록 유형과 달리, 데이터베이스의 경우는 보존 대상을 결정하기가 어렵다(Ashley 2004; DPT 2003; Verdegem 2003). 이는 데이터베이스 시스템이 실제 콘텐츠로 구성된 데이터베이스 자체, 데이터베이스 운영프로그램인 DBMS, 그리고 GUI(Graphic User Interface) 및 검색을 위한 응용프로그램으로 구성되기 때문이다. 따라서 데이터베이스 시스템 보존 대상의 경우의 수는 전체 데이터베이스 시스템,<sup>1)</sup> 데이터베이스, 테이블에 저장된 데이터 개별 열(row), 여러 테이블에 분포된 데이터, 데이터베이스 내의 데이터로서 응용프로

그램에 의해 스크린에 나타난 정보 등의 다섯 가지로 나뉜다(DPT 2003; Verdegem 2003).

또한 Verdegem(2003)은 다른 기록 유형과 달리 데이터베이스 보존이 문제가 되는지를 5 가지로 설명하고 있다. 우선, 데이터베이스는 고유하므로, 보존된 데이터를 읽고 이용할 응용프로그램 역시 데이터베이스마다 커스터마이징되어야 한다는 문제가 발생한다. 둘째, 응용프로그램 환경 역시 데이터베이스마다 다르며, 이러한 시스템에 대한 지식 역시 공유되기 힘들다. 셋째, 데이터베이스는 구조적인 데이터의 집합이기 때문에 구조정보까지 보존하면서 데이터베이스를 보존 형태로 변환하는 작업에는 고도의 기술이 필요하다. 넷째, 갱신이나 추가 혹은 삭제 등이 이루어지는 데이터베이스의 경우, 그 내용이 자주 바뀌며, 특히 동적 데이터베이스의 경우는 더욱 그러하다. 마지막으로, 데이터베이스와 기록간의 관계가 불명확하며, 문맥 의존적이라는 점이다.

### 3. 해외 데이터세트 기록의 관리 사례 분석

#### 3.1 NDAD(National Digital Archive of Dataset)

##### 3.1.1 개요

NDAD는 영국 The National Archives(TNA)에서 구축·서비스하는 데이터세트 아카이브다. NDAD는 디지털 형태의 영국 중앙 정부부

처와 다양한 기관이 생산하는 데이터세트와 문서 중 TNA가 장기적으로 보존할 가치가 있는 데이터세트와 문서를 보존하고 대중의 온라인 접근과 이용을 보장한다. 영국 중앙정부는 1960년대부터 컴퓨터를 이용하기 시작했으며, 따라서 NDAD는 40여 년간의 데이터세트를 포괄하고 있다(The National Digital Archive of Dataset 2005).

이는 영국 정부의 전자기록 보존 및 아카이빙 프로그램의 일부로 운영된다. TNA의 기록관리 부서는 중앙 정부부처의 기록관리담당자 및 기타 공공기관과 함께 NDAD에 기록을 선별하고 데이터세트를 NDAD로의 이관작업을 수행한다. TNA가 법적으로 소유한 데이터세트를 대상으로 현재 ULCC(University of London Computer Centre)에서 NDAD를 관리하고 운영 중이다(The National Digital Archive of Dataset 2005).

##### 3.1.2 개발배경

현재 95% 이상의 영국 정부 기록이 디지털 형태로 생산되고 있다(The National Digital Archive of Dataset 2005). 워드프로세서로 작성된 문서나 이메일 등의 텍스트 문서와 함께, 데이터세트는 영국 정부 전자기록의 중요한 부분을 차지하고 있다. 데이터세트는 다양한 범위의 이슈를 넘나드는 영국 정부의 정책과 법제도에 대한 지식을 전달하고, 정책 및 법제도 형성에 도움을 주는 중요한 정보로 여겨진다. 각 정부기관에서 생산된 데이터세트는 고위 공직자들의 정책 형성이나 법제도 제정에 지대한

1) 전체 데이터베이스 시스템은 데이터베이스, RDBMS, 애플리케이션으로 구성된다(DPT 2003).

영향을 미치는 원 자료인 동시에, 이러한 정부의 정책 결정 과정을 보여주는 중요한 증거라는 인식이 증대하게 되었다. 또한 대체할 수 없는 데이터와 통계정보라는 유일한 자료로서의 데이터세트를 보존하고 접근을 보장해야 할 필요를 느끼게 된 것이다(The National Digital Archive of Dataset 2005).

### 3.1.3 데이터세트 조직 및 기술

NDAD의 아키비스트는 이용자가 데이터세트를 좀 더 잘 이해하고 쉽게 탐색하도록 참조번호를 할당하고, 생산조직의 연혁을 포함한 데이터세트의 배경과 콘텐츠 목록을 작성하는 등 검색도구를 생산한다(The National Digital Archive of Dataset 2005). NDAD의 기술단위는 시리즈(Series)이며, 기술은 ISAD(G)를 기본으로 작성된다. 각 시리즈별로 시리즈 목록, 시리즈에 대한 다큐멘테이션 목록, 해당 데이터세트 시리즈를 생산한 기관 및 부서의 행정연혁, 시리즈에 포함된 데이터세트 목록, 혹은 광범위한 시리즈의 경우 하위시리즈 목록까지 포함하고 있다. 검색도구는 계층별로 데이터세트의 생산 이유와 방법, 시기 등에 대한 정보를 제공한다. 시리즈에 속한 데이터세트의 양이 너무 광범위한 경우 연도별 구분 등을 통한 서브시리즈를 구성한다(The National Digital Archive of Dataset 2005).

NDAD의 기술요소는 ISAD(G)에서 나타난 것을 기본으로 데이터세트의 특징을 감안해 추가된 것으로 보인다. 기술요소는 10개 영역의 36개 요소에 생산부서 링크까지 포함하면 37개 요소로 구성된다.

<표 2> NDAD 시리즈 계층 기술 요소

기술영역	기술요소
배경	(생산부서 링크)
식별	제목 생산일자 내용일자 데이터세트 규모 다큐멘테이션 생산일자 다큐멘테이션 규모 최종 입력일 최종 이용일 ISAD(G) 기술 계층
행정배경	목적 및 용도 생산 책임 소재 관리 이력
특징과 내용	범위와 내용 처분일정 정보 추가 입수 계획 이전 참조
접근·이용조건	법적 지위 접근 조건 저작권 요건 데이터 보호법 요건 언어
관련 자료	관련 기술 단위 연합 자료 생산부서 제작 출판물 연구자 제작 출판물
원 시스템 특징	하드웨어 운영시스템 응용프로그램 이용자 인터페이스
구조	논리구조와 스키마 동적 혹은 폐쇄 여부 획득 및 검증 방식 데이터 신뢰성에 대한 제한
검증	이관 이후 수행된 검증
데이터세트목록	데이터세트 목록 연결
주기	기술에 이용된 정보원

<표 3> NDAD 데이터세트 계층 기술 요소

기술영역	기술요소
배경	(시리즈 링크)
식별	제목 데이터세트 생산일자 데이터세트 내용일자 최종 입력일 최종 접근일

	데이터세트 규모 ISAD(G) 기술 계층
행정배경	목적 및 용도 책임 소재
입수 정보원	입수 정보원
특징과 내용	범위와 내용 디지털 처리와 변환
이용 및 접근조건	접근 조건
관련 자료	관련 기술 단위 연합 자료 생산부서에서 제작한 출판물 연구자가 제작한 출판물
구조	논리 구조와 스키마 획득 및 검증 방법 데이터 신뢰성에 대한 제한
검증	내용 검증 변환 검증
연결 데이터세트	관련 데이터세트
주기	(기술에 이용된 자료)

〈표 4〉 NDAD 테이블 기술 요소

기술영역	기술요소
배경	(생산부서, 시리즈, 데이터세트 링크)
요약정보	제목 NDAD 참조 범위와 내용 이용 조건
식별	테이블명 레코드 수 파일 크기(Megabytes) 최대 레코드 크기(bytes) ISAD(G) 기술 계층
기술	테이블 정보
필드정보	(테이블 내 필드정보 기술)

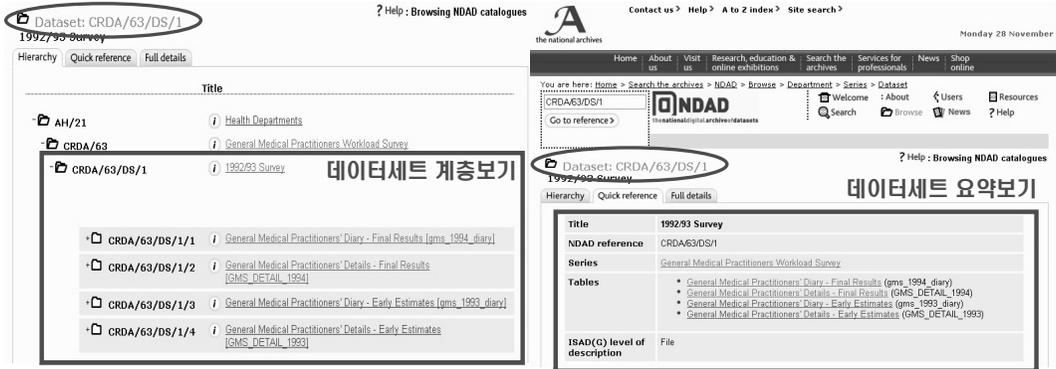
### 3.1.4 접근 및 이용

전 세계 이용자들은 NDAD 사이트를 통해 별도의 인증절차 없이 NDAD의 데이터세트를 자유로이 이용할 수 있다(The National Digital Archive of Dataset 2005). 시리즈에의 접근은 부서별 데이터세트 리스트 혹은 전체 데이터세트 리스트로부터 접근하는 ‘훑어보기(Browse)’

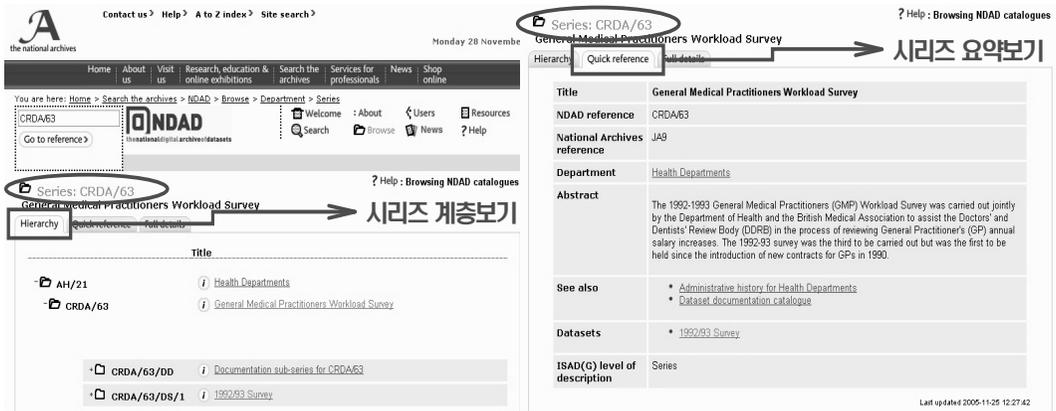
를 이용하거나, free-text 검색을 이용하여 시작된다. free-text 검색과 함께, 시소러스에 기초한 주제 색인을 이용하거나, 개인이나 기관 혹은 지명 색인을 이용하여 free-text 검색이 이루어질 수 있다. 시리즈 계층에서 데이터세트 계층과 테이블 계층으로 순차적으로 접근하여 실제 원 자료를 검색할 수 있다. 더불어 데이터세트의 기술 정보는 화면에 보이는 기술요소와 실제 기술정보, 혹은 서브시리즈를 구성하는 다큐멘테이션을 이용하여 얻을 수 있다.

각 시리즈와 데이터세트 및 테이블 기술은 ‘계층적 보기(Hierarchy)’, ‘요약보기(Quick reference)’, ‘전체보기(Full details)’ 탭 화면으로 나누어 표시되고 있어, 이용자가 다각도로 시리즈 기술 정보에 접근할 수 있다. 계층보기에서는 해당 시리즈 아래 포함된 데이터세트, 테이블, 데이터세트 다큐멘테이션 등의 하위 수준 기술단위 계층을 표현해주며, 요약보기에서는 간략한 시리즈 기술은, 전체보기에서는 완전한 수준의 시리즈 기술을 보여준다(The National Digital Archive of Dataset 2005).

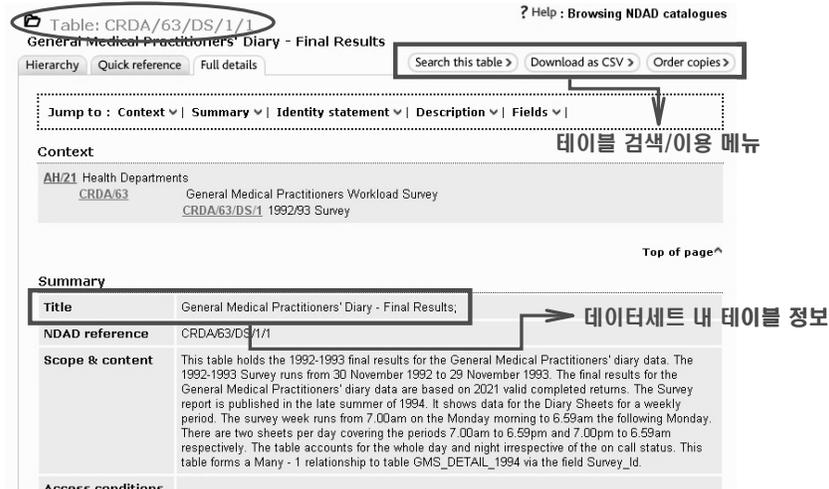
테이블 내 검색단계는 4단계로 이루어진다. 테이블 계층 전체보기 화면에서 “Search this table” 버튼을 눌러 원 데이터 검색이 시작된다. 첫 번째 탭에서 데이터 화면보기 옵션을 선택한다. 두 번째 탭에서는 데이터세트 내 검색 대상 테이블과 조합하여 검색 가능한 관련된 테이블을 선택한다. 세 번째 탭에서 결과에 포함될 필드를 선택한다. 마지막 탭에서, 검색식을 세우고 검색버튼을 눌러 검색을 완료한다. 물론, 이 과정 전체를 밟아야 하는 것은 아니며, 첫 번째에서 네 번째 탭까지 모든 과정에서 검색을 실행할 수 있다.



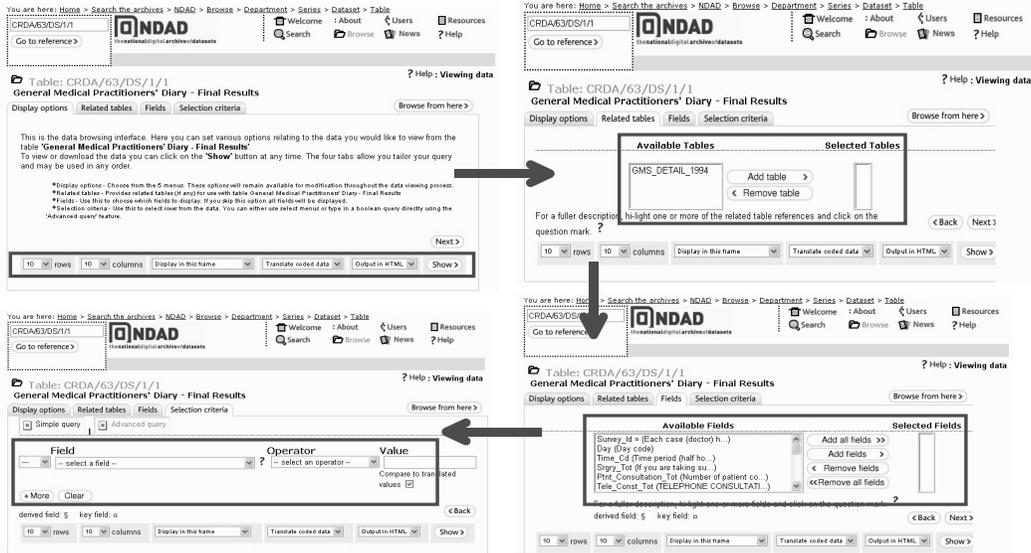
<그림 1> 데이터세트(CRDA/63/DS/1) 정보 보기



<그림 2> NDAD 시리즈(CRDA/63) 단위 정보 보기



<그림 3> 데이터세트 내 테이블(CRDA/63/DS/1/1) 정보 보기



〈그림 4〉 테이블 검색 순서

## 3.2 AAD(The Access to Archival Database)

### 3.2.1 개요

AAD는 미국 NARA(National Archives & Records Administration)에서 구축하여 운영하고 있는 데이터세트 아카이브로 2003년 2월 12일부터 서비스되고 있다(SAA 2003). NARA는 20여개 연방정부기관의 데이터베이스 내에 저장되어 있는 데이터세트 중, 특정 개인이나 지역, 기관, 시간범위별로 역사적 가치를 지닌 350개 이상의 데이터파일을 선별하여 제공하고 있다(Access to Archival Database 2005). AAD 시스템은 ERA (Electronic Records Archives) 프로그램의 후원 하에 개발된 첫 번째 공개 접근 응용프로그램이다(SAA 2003). 연방정부 업무활동에서 생산된 데이터세트를 보존하고 서비스하기 위해 구축되었으며, 주로 군사 분야 데이터세트를 중심으로 구축되어 있다(Darlington 2004).

### 3.2.2 개발배경

AAD는 특정 사실 데이터에 접근하고자 하는 이용자 요구를 충족시키기 위해 개발되었다(Adams 2003). 개인 PC의 보급 확산과 유비쿼터스 환경은 전자기록을 온라인으로 접근하고자 하는 대중의 기대치를 상승시켰다. 전통적인 기록 서비스 형태는 덩어리 기록이 아닌 특정 기록이나 사실을 찾으려 하는 이용자의 요구를 충족해 줄 수 없었다. 만약 기록관리담당 직원이 이런 개별 기록 서비스를 해야 한다면, 엄청난 노동이 수반될 것임은 자명하다. 결국 NARA는 필수적인 증거에 대한 준비된 접근을 제공하고 있는 것이다(Adams 2003).

### 3.2.3 데이터세트 조직 및 기술

AAD는 시리즈, 파일단위, 테이블 계층 등 3계층으로 기술된다. 상위 단위는 하나 이상의 하위 단위를 포함한다. 즉 하나의 시리즈는 하나 이상의 파일단위를, 또한 하나의 파일단위는

하나 이상의 테이블을 포함할 수 있다(Access to Archival Database 2005). 그런데, AAD의 대부분 시리즈의 경우, 파일단위는 하나의 테이블을 포함하고 있었다.

NDAD와 동일하게 기술단위는 시리즈이며, 해당 시리즈의 스캔된 인쇄형 검색도구가 존재할 경우, 해당 검색도구로 연결해주는 링크 페이지를 제공한다. 별도의 영역 구분 없이 시리즈 수준에서는 시리즈의 제목 및 다른 제목, 생산자, 기록물 유형, 기술 계층, 기록물 위치, 생산일자, 포함일자, 데이터세트 시리즈가 속한 레코드 그룹으로서의 상위계층, 기능 및 이용, 범위와 내용주기, 접근조건, 일반주기, 검색도구 유형, 검색도구 정보원, 이형 제어번호, 규모, 색인어 등을 기술하고 있다. NDAD의 데이터세트 계층에 해당하는 파일단위 계층에서는 제목과 다른 제목, 기록물 유형, 기술계층, 기록물

위치, 포함일자, 상위계층, 접근제한, 일반주기, 이형 제어번호 등을 기술하며, 테이블 수준에서는 테이블명과 상위계층명만을 기술하고 있다.

### 3.2.4 접근 및 이용

AAD의 데이터세트 역시 AAD 웹사이트를 통해 자유롭게 이용할 수 있다. AAD 데이터세트 검색은 크게 3단계로 이루어진다. 우선 시리즈 리스트, 개인별검색, 색인, 주제, 지역, 기관, 시간범위, 생산자 등의 검색 방법 중 하나를 이용하여 시리즈와 파일단위를 선택한다. 검색하고자 하는 정보를 담고 있을 시리즈와 파일단위를 결정하는 단계다. AAD 검색은 한 번에 하나의 데이터파일, 즉 하나의 테이블을 대상으로 이루어진다. 원하는 시리즈 제목을 클릭해서 시리즈 기술정보와 하위 파일단위 기술정보를 확인한다.

〈표 5〉 AAD 계층별 기술 요소

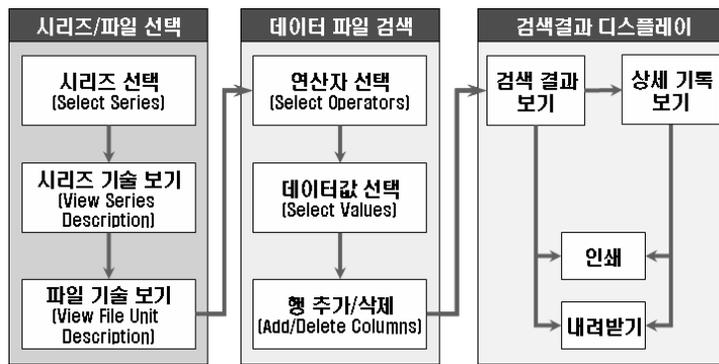
	시리즈	파일단위	테이블
	제목	제목	테이블명
	생산자	-	-
	기록물 유형	기록물 유형	-
	기술 계층	기술 계층	-
	다른 제목	다른 제목	-
	기록물 위치	기록물 위치	-
	생산일자	-	-
	포함일자	포함일자	-
	상위계층	상위계층	상위계층
기술요소	기능 및 이용	-	-
	범위와 내용주기	-	-
	접근조건	접근제한	-
	일반주기	일반주기	-
	검색도구 유형	-	-
	검색도구 정보원	-	-
	이형 제어번호	이형 제어번호	-
	규모	-	-
	색인어	-	-

두 번째는 실제 데이터 파일을 검색하는 단계다. 검색하고자 하는 하나의 데이터 파일을 선택하고, 필드 값을 제한하여 검색식을 완성한 후 검색을 수행한다. 탭 페이지를 이용해 검색 순서를 제시하는 NDAD와 달리, AAD는 데이터 파일 단위 화면에서 검색식을 입력하는데, 검색 대상 필드 추가나 삭제는 데이터 파일 화면에 링크된 팝업 화면에서 지원한다. 마지막으로 검색결과를 확인하고, 화면보기를 하거나 인쇄 혹은 다운로드하여 데이터를 이용할 수 있다.

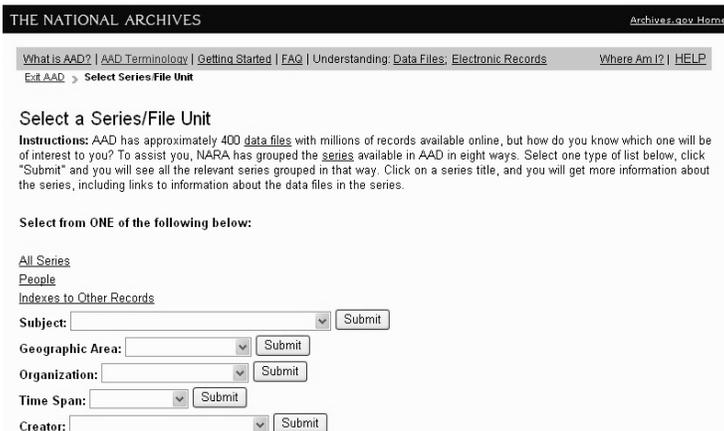
### 3.3 NDAD와 AAD 조직 및 기술 비교

#### 3.3.1 계층 및 기술

NDAD와 AAD는 모두 3단계로 계층을 나누어 기술하고 있었다. NDAD는 '시리즈-데이터세트-테이블'의 3계층 기술을, AAD는 '시리즈-파일개체(데이터세트)-데이터파일(테이블)'의 3계층을 이룬다. 그런데, AAD의 전체 시리즈를 살펴본 결과, 하나의 시리즈를 제외하면 파일단위 계층에는 단지 하나의 데이터



<그림 5> AAD 검색 흐름도(Access to Archival Database 2005)



<그림 6> AAD 검색 메인 화면



〈표 6〉 식별 기술요소

기술계층	NDAD	AAD	
시리즈	제목	제목 다른 제목	
	데이터세트 생산일자	포함일자	
	데이터세트 내용일자	범위일자	
	데이터세트 규모	규모	
	다큐멘테이션 생산일자	검색도구 유형 (검색도구로의 링크)	
	다큐멘테이션 규모		
	최종 입력일	-	
	최종 접근일	-	
	ISAD(G) 기술계층	기술계층	
	-	기록물 유형	
	-	이형 제어번호	
	데이터세트	제목	제목
		데이터세트 생산일자	-
데이터세트 내용일자		범위일자	
최종 입력일		-	
최종 접근일		-	
데이터세트 규모		-	
ISAD(G) 기술계층		기술계층	
-		기록물 유형 이형 제어번호	
테이블	테이블명	데이터 테이블명	
	레코드 수	-	
	파일 사이즈	-	
	최대 레코드 사이즈	-	
	ISAD(G) 기술계층	-	

〈표 7〉 행정배경 기술요소

기술계층	NDAD	AAD
시리즈	생산부서	생산자
	책임 소재	
	목적 및 용도	기능 및 이용
	관리이력	-
	-	상위계층
데이터세트	상위시리즈(링크)	상위계층
	목적 및 용도	-
	책임소재	-
	입수 정보원	-
테이블	생산부서 및 상위계층(시리즈/데이터세트링크)	상위계층

면, NDAD의 경우 데이터세트가 속한 시리즈  
로의 링크와 함께, 목적과 용도, 책임소재, 입수

정보원을 기술하고 있었다.

데이터세트 내용에 대해, NDAD와 AAD 모

두 데이터세트 시리즈에 대한 범위와 내용주기를 기입하고 있었다. NDAD는 ISAD(G)의 요소를 차용하여 처분일정정보와 추가입수계획, 이전참조 등을 시리즈 기술에 적용하고 있었으나, AAD에서는 이 같은 요소는 보이지 않았다.

데이터세트에의 접근 및 이용조건 기술에 있어, AAD의 경우 보존하고 서비스하는 모든 자료는 영어로 작성되었으므로(Access to Archival Database 2005) 언어 기술 요소가 빠져 있었다. 그러나 추후 자료 입수를 고려한다면, NDAD처럼 언어 기술요소를 마련하는 것이 바람직할 것으로 여겨진다. 이외에, NDAD의 경우 시리즈 수준에서 해당 기록의 법적 지위

와 접근 조건, 저작권 요건, 데이터보호법 요건, 언어 등을, 데이터세트와 테이블 수준에서 각각 접근 조건 기술 요소가 마련되어 있고, AAD의 경우 시리즈 수준에서 접근 제한과 위치 요소가, 데이터세트 수준에서는 접근 제한과 위치 요소가 마련되어 있다.

NDAD는 시리즈와 데이터세트 수준에서 관련 기술단위와 연합자료, 데이터세트를 이용하여 생산부서에서나 연구자가 제작한 출판물에 대한 기술을 포함하고 있는 반면, AAD는 관련 자료 기술요소가 마련되어 있지 않았다. 다만, 시리즈 단계에서, 해당 시리즈가 속한 레코드 그룹만 언급되어 있었다. 개별 테이블 수준에서

〈표 8〉 특징 및 내용 기술요소

기술계층	NDAD	AAD
시리즈	범위와 내용	범위와 내용 주기
	처분일정 정보	-
	추가 입수 계획	-
	이전 참조	-
	-	색인어
데이터세트	범위와 내용	-
	디지털 처리와 변환	-
테이블	범위와 내용	-
	테이블 기술정보	-

〈표 9〉 접근 및 이용조건 기술요소

기술계층	NDAD	AAD
시리즈	법적 지위	-
	접근 조건	접근 제한
	저작권 요건	-
	데이터보호법 요건	-
	언어	-
	-	위치
데이터세트	접근조건	접근 제한
	-	위치
테이블	접근조건	-

는 NDAD와 AAD 모두 관련자료 기술을 하지 않고 있었다. 개별 아이템 계층이라 할 수 있는 테이블 단계에서의 관련 자료는 시리즈나 데이터세트와 중복되므로, 생략한 것으로 보인다.

NDAD는 시리즈 계층에서, 해당 데이터세트가 원래 저장되어 있었던 환경, 즉 하드웨어, 운영시스템, 응용프로그램, 이용자 인터페이스 등을 기술하고 있었고, AAD의 경우 하드웨어 및 소프트웨어 환경에 대한 기술을 하지 않고 있었다.

이용자에게 원 데이터를 공개하고 있음에도, 데이터세트와 테이블의 구조정보는 해당 데이터세트와 테이블을 이해하기 위한 중요한 기술 영역이다. 데이터세트 구조에 대한 기술은 온라인 상에서는 NDAD에서만 보여진다. 시리즈 계층에서는 논리구조와 스키마, 데이터세트의

동적/폐쇄 여부, 획득과 검증 방법, 데이터 신뢰성과 관련된 제약사항을 기술하고 있었고, 데이터세트 계층에서는 동적/폐쇄 여부 요소를 제외하고 시리즈에서와 동일한 요소를 기술하고 있었다. 테이블 계층에서는 테이블을 구성하고 있는 필드 구조를 기술하고 있었다. 단, AAD는 개별 테이블 수준에서, 데이터 레이아웃에 대한 정보를 추가정보 링크를 이용하여 보여주고 있었다. 웹에서 제공되지 않는 AAD 데이터세트의 구조에 대한 추가 정보는 인쇄본 다큐멘테이션을 참조해야 하는 불편이 있다.

장기보존을 위해서는 이관 이후 데이터세트를 지속적으로 관리하여 기록으로서의 진본성을 입증할 수 있어야 한다. 그런 의미에서 보존 기관으로 데이터세트가 이관된 이후 수행된 검

〈표 10〉 관련 자료 영역 기술요소

기술계층	NDAD	AAD
시리즈	관련 기술단위	-
	연합자료	-
	생산부서 제작 출판물	-
	연구자 제작 출판물	-
데이터세트	관련 기술단위	-
	연합자료	-
	생산부서 제작 출판물	-
	연구자 제작 출판물	-
테이블	-	-

〈표 11〉 원 시스템 환경 영역 기술요소

기술계층	NDAD	AAD
시리즈	하드웨어	-
	운영시스템	-
	응용프로그램	-
	이용자 인터페이스	-
데이터세트	-	-
테이블	-	-

〈표 12〉 구조 기술요소

기술계층	NDAD	AAD
시리즈	논리구조 및 스키마	-
	동적/폐쇄 여부	-
	획득 및 검증 방법	-
	데이터 신뢰성 관련 제약	-
데이터세트	논리구조 및 스키마	-
	획득 및 검증 방법	-
	데이터 신뢰성 관련 제약	-
테이블	필드 구조	데이터 레이아웃

〈표 13〉 검증 기술요소

기술계층	NDAD	AAD
시리즈	이관 이후 수행된 검증행위	-
데이터세트	내용 검증	-
	변환 검증	-
테이블	-	-

증행위나 내용 검증, 변환검증 등에 대한 기술이 NDAD에서와 같이 기술된다면, 데이터세트가 기록으로서 무결하고 신뢰성 있게 보존되고 있음을 보여줄 수 있다.

NDAD와 AAD의 시리즈에 속한 하위 데이터세트 혹은 데이터세트에 속한 하위 테이블

기술 화면으로 연결해주기 위한 요소가 마련되어 있다.

NDAD는 주기 요소에서 기술에 참고한 정보원을 기술하고 있고, AAD에서는 검색도구의 정보원을 기술하며, 별도의 일반주기를 기입하고 있다.

〈표 14〉 하위 단위 기술요소

기술계층	NDAD	AAD
시리즈	하위 데이터세트 링크	하위 파일단위 링크
데이터세트	하위 테이블 링크	하위 데이터파일 링크
테이블	-	-

〈표 15〉 주기정보 기술요소

기술계층	NDAD	AAD
시리즈	기술에 이용된 정보원	검색도구 정보원
	-	일반주기
데이터세트	기술에 이용된 정보원	-
	-	일반주기
테이블	-	-

### 3.3.2 데이터세트 접근 및 검색

문서형태의 기록과 달리 데이터세트 기록은 테이블에 저장된 데이터를 대상으로 한 검색과 이용, 데이터 다운로드를 필요로 한다. NDAD와 AAD 모두 데이터세트의 원 데이터의 검색과 이용을 지원하고 있었다. 그러나 그 방식에서는 다소 차이가 있었는데, 한 예로, NDAD는 하나의 데이터세트 내에서 관련된 여러 테이블을 대상으로 한 번의 검색식을 세울 수 있으나, AAD의 경우는 하나의 테이블만을 대상으로 검색을 제한하고 있었다. NDAD와 AAD의 접근방식과 검색방식의 차이는 <표 16>으로 정리하였다.

## 4. 결론 및 제언

현재 상황에서 데이터세트 기록의 관리는 다양한 과제를 지닌다. 우선 국가 차원의 기록 관리 혁신 움직임을 비롯해 기록관리 대상인 데이터세트에 대한 인식 확산이 첫 과제일 것이다. 이를 위해서는 데이터세트를 생산하는 업무과정을 파악하고, 어떤 행정정보시스템에서, 어떠한 유형으로 생산되는가에 대한 연구가 필요하다. 더불어 데이터세트에 대해 기록관리차원의 업무가 이루어져야 하는 주기 역시 고려되어야 한다.

<표 16> NDAD와 AAD 검색 비교

구분	NDAD	AAD
시리즈 검색방식	<ul style="list-style-type: none"> <li>• Free text 검색</li> <li>• 시소러스·인명색인 보조 이용</li> <li>• 전체 시리즈리스트 및 생산기관별 시리즈리스트로 접근 가능</li> </ul>	<ul style="list-style-type: none"> <li>• 주제, 지역, 기관, 시기, 생산자 등 미리 통제된 분야 및 값으로 검색 가능하며 조합검색 불가</li> <li>• 전체 시리즈리스트로 접근 가능</li> </ul>
기술 정보 화면	<ul style="list-style-type: none"> <li>• 계층보기, 요약보기, 전체보기 등 이용자의 필요에 따른 기술정보 제공</li> </ul>	<ul style="list-style-type: none"> <li>• 전체보기만 지원</li> </ul>
계층 보기	<ul style="list-style-type: none"> <li>• 계층보기를 통해 해당 시리즈 전체 구조를 디렉토리 형태로 보여주고, 접근 가능</li> </ul>	<ul style="list-style-type: none"> <li>• 디렉토리형 계층보기 없음</li> </ul>
다큐멘테이션	<ul style="list-style-type: none"> <li>• 다큐멘테이션 하위시리즈 구성</li> </ul>	<ul style="list-style-type: none"> <li>• 이미지화된 검색도구 유형 다큐멘테이션 지원</li> </ul>
데이터 검색방식	<ul style="list-style-type: none"> <li>• 테이블 계층에서 검색식 구성</li> </ul>	<ul style="list-style-type: none"> <li>• 데이터파일 계층에서 검색식 구성</li> <li>• 일부 데이터파일의 경우는 검색식 구성 단계 생략, 직접 원 데이터파일을 화면에 보여줌</li> </ul>
테이블 필드구조	<ul style="list-style-type: none"> <li>• 테이블 계층 전체보기에서 필드구조 제공</li> </ul>	<ul style="list-style-type: none"> <li>• 데이터파일(테이블) 추가정보 링크를 통한 팝업창으로 데이터 레이아웃 제공</li> </ul>
데이터 검색순서	<ul style="list-style-type: none"> <li>• 탭 화면을 이용하여 일련의 순서대로 검색식 구성 가능</li> <li>• 화면보기 설정 → 데이터세트 내의 관련 테이블 추가 → 출력할 필드 선택 → 필드값을 입력하여 검색 실행 → 검색결과 보여줌 → 화면보기 혹은 다운로드 가능</li> </ul>	<ul style="list-style-type: none"> <li>• 데이터파일 계층 화면에서 검색식 구성</li> <li>• 디폴트 검색 필드 설정됨</li> <li>• 팝업 옵션 링크 기능을 통해 검색에 이용할 필드 선택 가능</li> <li>• 팝업 'Sample records' 링크 기능을 통해 실제 데이터파일의 일부를 화면보기 가능</li> <li>• 검색 필드 우측 'Sample value', 'Sample code list' 링크를 통해 검색을 제한할 필드값 입력 도움</li> </ul>
결과 및 다운로드 파일 유형	<ul style="list-style-type: none"> <li>• HTML 화면보기 및 CSV 파일로 다운로드 지원</li> </ul>	<ul style="list-style-type: none"> <li>• HTML 화면보기 및 CSV 파일로 다운로드 지원</li> </ul>
검색결과 크기	<ul style="list-style-type: none"> <li>• 크기 제한 없음</li> </ul>	<ul style="list-style-type: none"> <li>• 150KB를 넘을 수 없음. 그 이상은 전체 데이터파일의 주문하거나 검색식을 재구성해야 함</li> </ul>

두 번째는 데이터세트 기록의 획득과 등록의 문제다. 행정정보시스템에 저장되어 있는 데이터세트를 어떤 방식으로 획득하고 등록해야 할 것인가는 이후 관리와 보존을 결정하는 중대한 과제다. 전체 데이터베이스 시스템, 개별 데이터베이스, 하나의 테이블에 저장된 개별 열, 여러 테이블에 분포된 데이터 집합, 화면에 나타난 정보 등 그 대상이 결정되어야 한다.

세 번째는 데이터세트 기록의 저장에 대한 과제다. 어떤 구조로 저장할 것이며, 어떤 저장 매체를 이용할 것인가 등의 연구가 이에 해당할 것이다. 이는 데이터세트의 획득 방법과 밀접하게 연결된다.

네 번째는 데이터세트 기록의 조직 및 기술 관련 과제다. 데이터세트 기록은 어떤 분류 체계를 지녀야 하며, 어떤 분류 체계로 편입되어야 하는가의 문제, 혹은 데이터세트 기록을 생산한 업무의 다른 기록과는 어떻게 연계해야 하는가, 데이터세트 기록의 처리 방법은 어떠한가 등의 연구가 이에 해당한다. 더불어 그 기술과 검색도구 작성, 메타데이터 요소 결정 등의 연구과제도 해결되어야 한다.

다섯 번째는 접근상의 문제로, 이용자 인터페이스 방식이나 주 이용자 그룹의 결정 및 그 이용목적에 대한 연구, 접근 권한 및 제한에 대한 연구 등이 해당되겠다.

마지막으로 장기보존과 관련된 과제이다. 데이터세트 기록의 진본성을 유지하면서 이를 장기보존하기 위한 종합적인 연구가 앞서 언급한 문제들과 병행하여 수행되어야 할 것이다.

정부기관 및 기업에서는 다량의 다양한 데이

터세트가 생산된다. 통계데이터, 실험데이터, 관측데이터, 설문조사데이터 등 과학기술 및 사회과학분야만 하더라도 엄청난 양의 데이터가 생산되고 있다. 업무활동의 일부로 생산되는 데이터세트는 현재 단순히 정보시스템 내에 저장되어 있고, 기록으로 획득되어 관리되고 있지 못하다. 그러나 가공되지 않은 원 데이터로서 업무활동에서 발생한 데이터세트 기록은 시급히 기록관리 영역으로 편입되어 관리되지 않는다면, 오래지 않아 사장될 것이다. 기상청의 예를 보아도, 과거 인쇄 형태로 생산된 관측 데이터는 영구보존기록으로 관리하는 반면, 데이터베이스에 저장되어 있는 데이터세트의 경우, 기록관리 영역으로 편입하여 관리하고 있지 못하며, 심지어 기록으로서의 데이터세트에 대한 인식도 미미하다.<sup>2)</sup>

데이터세트는 연구자나 정책기획담당자들에게 막대한 양의 통계정보와 지식자원을 제공해 준다. 또한 데이터세트는 정부 정책 형성에 있어 기본 데이터로 활용 가능하다. 데이터베이스에 저장된 데이터세트는 고위층에서의 정책 및 법제도 형성에 기초 자료로 지대한 영향을 미칠 수 있기 때문이다. 더불어, 주요 정책결정과정을 보여주는 증거로서도 데이터세트의 관리와 보존은 반드시 필요하다(설문원 외a 2005).

이런 의미에서, 기록으로서의 데이터세트를 정의하고, 그 특징을 살펴보았다. 더불어 영국과 미국의 데이터세트 아카이브인 NDAD와 AAD를 살펴보고 그 기술과 접근방식을 중심으로 비교해 보았다. 영국의 NDAD와 미국의 AAD는 데이터세트 관리 및 보존의 한 방법을

2) 2005. 10. 29. 기상청 기록관리담당자와의 대화에서 인용하였다.

제시해 주는 사례이다. 데이터베이스의 복잡성 때문에 데이터베이스 시스템 전체를 기록을 획득하여 보존하기 위한 최선의 방안은 아직도 각국에서 실험중이다. 다만, 데이터베이스 내에 저장된 일련의 데이터세트의 경우는 NDAD와 AAD에서처럼 현재 상황에서 데이터베이스를 기록으로서 보존하고 관리하는 하나의 방안이 될 것이다.

미래 데이터세트의 보존 및 관리, 더 나아가 데이터베이스의 보존과 관리는 다양한 학문간 협력이 요구되는 분야다. 관리 및 보존 대상이

될 데이터베이스는 전체 시스템을 분석하고 목적과 구조, 그리고 이용 맥락을 파악해야 하며, 데이터베이스 시스템 관련 기술에 대한 지식 수집은 물론 보존상의 비용까지도 고려하여 평가하여 결정되어야 한다(ERPANET 2003). 결국 기록 관리와 보존의 궁극적인 목적은 미래 이용을 보장함에 있으므로, 데이터세트 기록을 우선적으로 관리 보존하면서도 완전한 데이터베이스 시스템의 관리와 보존 및 미래 접근을 고려한 모든 가능성을 실험해야 할 것이다.

## 참 고 문 헌

- 대통령 비서실. 2005. 『기록관리혁신을 위한 정보화 전략 계획 수립』. 서울: HandySoft 컨소시엄. [미간행].
- 설문원 외. 2005. 『과학기술 정책자료의 효율적 활용을 위한 기록관리 강화방안』. 과천: 과학기술부.
- \_\_\_\_\_. 2005. 『국가 디지털 아카이빙 체제 구축에 관한 연구』. 서울: 한국과학기술정보연구원.
- 이소연 외. 2002. 『ISAD(G)를 적용한 한국기록물기술규칙 개발에 관한 연구』. 서울: 한국국가기록연구원.
- 이소연. 2004. 디지털유산의 장기적 보존: 국가정책 수립을 위한 제안. 『기록학연구』, 10: 27-64.
- 정부혁신지방분권위원회. 2005. 『기록이 없으면 정부도 없다: 국가기록관리혁신 로드맵 이해하기』. 서울: 정부혁신지방분권위원회.
- Adams, Margaret O. 2003. *Online Services for Electronic Records: U. S. National Archives and Records Administration*. IASSIST 2003.
- Ashley, Kevin. 2004. "The preservation of databases." *VINE: The Journal of Information and Knowledge Management Systems*, 34(2): 66-70.
- Darlington, Jeffrey. 2004. "A national archive of datasets." *Ariadne*, 39. <<http://ariadne.ac.uk/issue39/ndad>>.
- DPT. 2003. *From Digital Volatility to Digital Performance : Preserving Databases*. Hague: Digital Preservation Testbed. [cited 2005. 10. 2] <<http://www.digitaleduurzaamheid>.

- nl/bibliotheek/docs/volatility-permanence-databases-en.pdf>.
- ERPANET. 2003. "The long-term preservation of databases." *ERPANET Workshop, Bern*. 2003. 4. 9 - 11.
- SAA. 2003. "National archives ADD system now operational." *SAA Electronic Records Section Newsletter*, 2003 Spring. <<http://www.archivists.org/saagroups/ers/03spnara.asp>>.
- Shepherd, Elizabeth. 2003. *The Application of ISAD(G) to the Description of Archival Datasets*. ERPANET seminar.
- Shepherd, Elizabeth and Charlotte Smith. 2000. "The application of ISAD(G) to the description of archival datasets." *Journal of the Society of Archivists*, 21(1): 55-86.
- Verdegem, Remco. 2003. *Databases preservation issues*. [cited 2005. 10. 21] <[http://www.digitaleduurzaamheid.nl/bibliotheek/docs/longterm\\_preservation\\_of\\_databases.pdf](http://www.digitaleduurzaamheid.nl/bibliotheek/docs/longterm_preservation_of_databases.pdf)>.
- The National Digital Archive of Dataset. [cited 2005. 11. 3] <<http://www.ndad.nationalarchives.gov.uk/>> .
- Access to Archival Database. [cited 2005. 11. 4] <<http://www.archives.gov/aad/>>.