디지털 아카이브즈와 보존을 위한 실무 기술

Practical Technologies for Digital Archives and Preservation

수싱 첸(Su-Shing Chen)*

목 차

〈초 록〉

지난 몇 년 사이에 전자문화, 전자정부, 전자학습 및 전자 비즈니스의 디지털 아카이브즈들은 세계적으로 순조롭게 성장해 왔다. 우리가 이 아카이브즈들을 구축하고 관리하기 위해 상당한 시간과 노력을 투자해온 한편, 이 처리에 의해서 생산된 디지털 레코드들을, 미래의 기술로도 접근할 수 있게 하고, 사람들로 하여금 그것이 확실하고 신뢰할 수 있는 것인지 결정할 수 있게 하면서, 정보 기술의 여러 세대 전체에서 모두 이용할 수 있게 하는 능력은 갖지 못한다. 이것은 아직 솔루션이 개발되지 않은 심각한 문제이다. 이 논문은 디지털 아카이브즈와 보존이 성공하기 위한 실용적 기술에 대하여 논하고, 이 중요한 문제를 해결하기 위해 정보의 생명주기의 일반적 구조를 기술하여, 정량적 방법과 증진되는 방식으로 분석되고 평가될 수 있는, 디지털 레코드들을 보존하기 위한 타당한 방법을 발견할 수 있도록 한다.

주제어: 디지털 아카이브, 디지털 레코드, 보존

<ABSTRACT>

The digital archives of E-culture, E-government, E-learning, and E-business have grown by leaps and bounds worldwide during the last several years. While we have invested significant time and effort to create and maintain those archives, we do not have the ability to make digital records generated by the processes all available across generations of information technology, making it accessible with future technology and enabling people to determine whether it is authentic and reliable. This is a very serious problem for which no solutions have been devised yet. This paper discusses practical technologies for digital archives and preservation to succeed, and describes a general framework of the life cycle of information to address this important problem so that we may find reasonable ways to preserve digital records that can be analyzed and evaluated in quantitative measures and incremental manners.

Key words: digital archives, digital records, preservation

* 미국 플로리다대학교 컴퓨터공학과 교수(suchen@cise.ufl.edu)

1. 머리말

디지털 아카이브즈들이 등장하고, 디지털 커뮤니티는 전자문화, 전자정부, 전자학습, 전자비즈니스를 형성하긴 하였지만, 우리는 디지털 보존에 있어서 아직도 근본적으로 역설적인 상황에 직면하고 있다. 한편에서 우리는 디지털 정보가 생산된 그대로 완전히 관리되기를 원하지만, 다른 한편에서 우리는 역동적 이용 상황에서 접근할 수 있기를 원한다(Chen 2001). 오늘날 디지털 형태로 정보를 생산하고, 캡처하고, 처리하며, 커뮤니케이션 하는 데 있어서 정보기술이 급속히 진보한 것이 어째서 가까운 미래에 접근가능성을 위협하는가? 그것은 두 가지 이유 때문이다. 우선, 디지털 정보가 갑자기 확산되었고, 둘째로 하드웨어와 소프트웨어 제품들이 거의 매 18개월마다 개선되고 교체된다. 정보기술 부문에서 회사들은 그들이 제공하는 제품과 서비스의 대부분이 5년 전에는 존재하지 않았었다고 보고한다. 비용효과를 위해서 우리는 하드웨어와 소프트웨어 제품을 계속해서 바꾸어야 한다. IT 회사들이 그들의 제품과 서비스의 급속한 진보에 따라 번성하는 한편, 나머지 비즈니스, 산업 및 정부를 따라잡기 위해서 비싼 값을 치른다.

디지털 환경은 보존의 개념을 근본적으로 변화시켰다. 전통적으로 보존은 사물이 변화하지 않고 유지되는 것을 의미한다. 예를 들어, 프톨레마이오스 5세 당시의 로제타석을 상형문자, 고대 이집트어 및 그리스어로 오늘날에도 읽을 수 있다. 그러나 우리가 아무런 변화 없이 디지털 정보를 붙잡고 있는데 성공할 수 있다면, 정보는 접근하기 불가능하지는 않다 하더라도, 점점 더 어려워지게 될 것이다. 비록 물리적 매체가 디지털 정보를 완전하게 유지할 수 있다 하더라도, 정보가 디지털로 기록되는 포맷은 변화하고, 매체에서 정보를 검색하기 위한 하드웨어와 소프트웨어는 종종 시대에 뒤떨어지게 된다. 요컨대, 우리는 정보기술의 미래를 예측할 수 없고, 따라서 현재 상황에서 디지털 보존을 잘 계획할 수 없다.

Chen(2001)에서 우리 사회를 위한 기본적 정보 기념물로서 디지털 레코드를 보존하는 것이 얼마나 중요한지를 강조했고, 디지털 레코드를 보존하는 데 있어서 요구조건들과 전략들을 다루기 위한 정보의 생명주기를 소개했다. 장기적 요구조건과 전략을 확립함으로써 우리는 디지털 레코드와 아카이브즈의 발전을 안정시킬 수 있다. 이 논문에서 우리는 디지털 보존에서 아카이벌 안정성(archival stability)은 기관적 처리(organizational process) 및 기술 연속성(technology continuity)과 동반되어야 한다고 주장하고, 이 세 가지 기본적 요인들이 상호 연관되고 함께 고려되어야 하는 디지털 보존의 구조를 제안한다. 따라서 기록관리자들은 기관적 업무 흐름 처리(organizational workflow processes)와 기술 이행(technology implementations)을 동반하지 않고는 아카이벌 안정성을 달성할 수 없다. 마찬가지로 기관들과 기술회사들은 기록관리자들로 하여금 그들의 계획에 참여하도록 요구해야 한다. 이 세 가지 요인들에 대하여 다학문적 훈련을 받는 것이 디지털 부문에서 긴급히 요구되고 있다.

기술 연속성을 유지하는 일관성 있는 하드웨어, 소프트웨어 및 어플리케이션의 발전은 IT 산업계의 책임이다. IT 산업은 기술 연속성이

효과적 비즈니스를 의미한다는 것을 인식해야 한다. 비록 분열시키는 기술이 IT 발전의 추진력이 되고 있긴 하지만, 기술 연속성은 고객을 장기적으로 유지함으로써 효과적 비즈니스가 되게 만든다. 산업계의 규범은 IT 산업에서 경쟁을 이기는 혼란한 새로운 제품과 서비스가 아니어야 한다. 최근 세계의 정부들은 이 결과에 영향을 미칠, IT 산업에 대한 재정지원 기회를 도입하고 있다(예를 들어, Mckemmish; Kim & Chen 2004; Shi 등 2004). 산업은 또한 잠재적으로 디지털 보존을 포함하는 미래의 비전을 발전시켜 왔다(예를 들어, Australian Standards and Framework).

2. 디지털 아카이브즈와 보존을 위한 구조

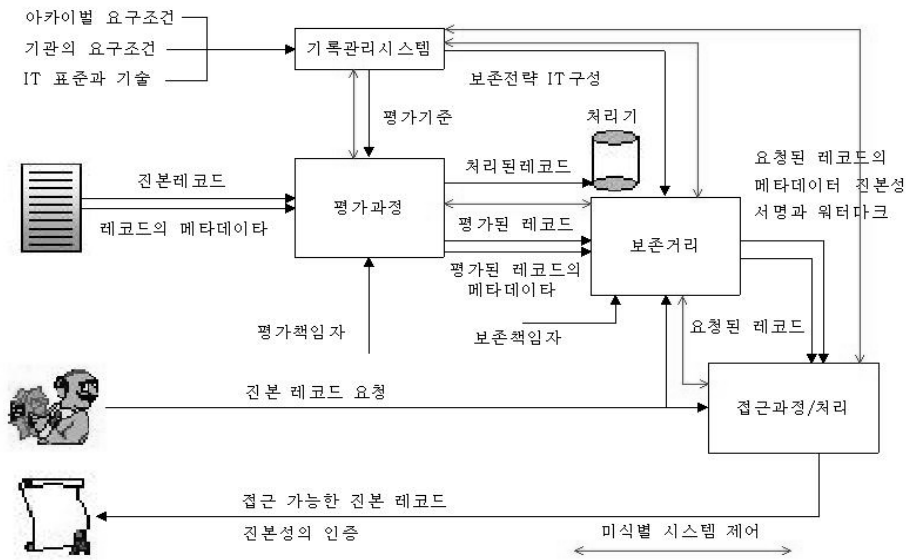
디지털 아카이브즈는 상당히 중요한 개념들을 갖는다. 그것들은 인터넷을 통하여 네트워크로 상호 연결되는 지식환경이다. 우선, 상호통신능력 지역(connectivity region)은 비교적 우수한 네트워크 상호통신능력을 갖는 일단의 사이트들이다. 둘째, 콜렉션 서비스(collection services)는 필요한 메타 정보를 제공하여, 일단의 디지털 아카이브즈 서버들이 콜렉션으로 상호 운용될 수 있게 한다. 셋째, 콜렉션 뷰(collection views)는 상호 통신 능력 지역에 일치하는 콜렉션의 구성을 나타낸다. 마지막으로, 다중언어 접근은 UNESCO 상황에서 다중언어로 탐색할 수 있게 한다.

디지털 보존에서 가장 중요한 문제들 가운데 하나는 무엇을 보존하며, 어떻게 보존할지를 아는 것이다. 새로운 구조는 보존에 있어서 기관적 처리를 맨 처음 강조하고, 디지털 보존의 전체적인 조망을 명확하게 한다. 보유 기관들 정부부서, 회사, 병원, 및 기관들의 책임은 궁극적 결과에 영향을 미칠 것이다. 기관적 처리 외에, 정보학계(예를 들어, 사서와 기록관리자)는 아카이벌 안정성을 제공해야 하고, 컴퓨터 업계는 동시에 아카이벌 안정성을 위한 기술 연속성을 발전시켜야 한다. 이 목적을 위해서 정보의 생명주기 내에서 디지털 보존 문제를 구성할 것이다. 정보의 생명주기는 역동적인 방식으로 입수, 보존, 콜렉션, 색인, 접근 및 활용에 미친다. 정보의 생명주기는 역동적 방식으로 입수, 보존, 콜렉션, 색인, 접근 및 활용하는 것을 포괄한다(Chen 1998). 만약 보존이 빠지면, 생명주기가 망가지고, 파괴될 것이다. 따라서 기관들은 정보의 생명주기에서 매끄럽게 디지털 아카이브즈의 보존을 설계하여, 기술이 필요한 연속성을 제공할 수 있도록 해야 한다.

전통적인 보존 요구조건들은 Chen(2001)에 나타난다(그림 1).

3. 기관적 처리

상이한 기관들은 그들의 아카이브즈와 도서관들을 위한 상이한 요구조건들과 이행 방법을 가지며, 이들은 기관의 속성에 상당히 의존한다. 예를 들어, 병원은 그들의 환자 기록들을 관리할 것이고, 학교 시스템은 학생기록을, 그리고 회사는 그들의 재정기록을 관리할 것이다. 일반적으로, 기관들이 보존해야 할지, 어떻게 보존해야 할지, 그리고 무엇을 보존해야 할지



〈그림 1〉 전통적 보존 요구 조건

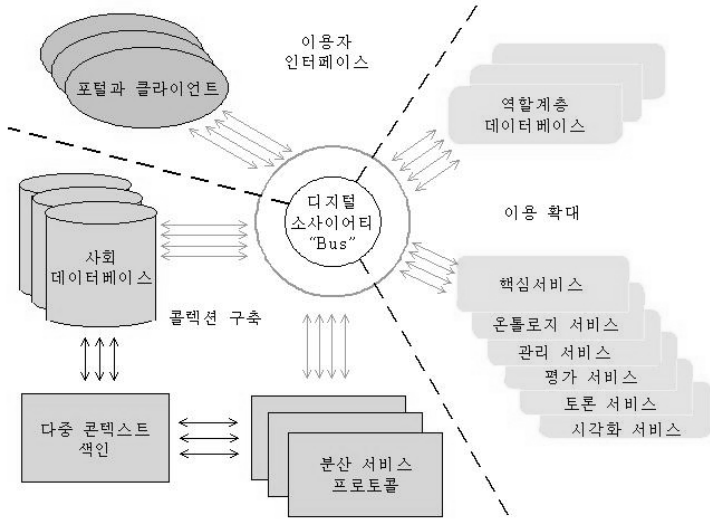
에 대한 표준은 없다. 디지털 사회라는 보다 집중된 상황에서, 기관들이 디지털 보존에 관련하여 처리하는 것이 그것들의 유지 가능성에 대해 분명하게 나타낼 것이다. 이 장은 이러한 유형의 처리에 관련된 여러 문제들을 검토한다. 이 처리를 아카이벌 변인으로 향후의 조직 관리 등식에서 고려해야 한다.

디지털 보존은 다양한 저장과 보존 기능들을 맡을 것이다. 전통적으로 보존된 레코드는 확대, 스캐닝, 재생과 영사장치들의 도움으로 직접 읽거나, 듣거나, 볼 수 있는 도서, 단행본, 보고서, 지도, 사진, 아날로그 음성 트랙과 필름의 형태로 되어 있었다. 물리적 및 아날로그 매체의 보존은 장기적 안정성과 접근성을 보장해야 한다. 디지털 레코드의 보존은 상당히 다른 방향을 취한다. 왜냐하면, 기술이 너무 빠르게 진보하여 하드웨어와 소프트웨어 제품들이 계속해서 개선되고 교체되고 있기 때문이다. 정보기술 부문의 회사들은 자기들이 제공하는 제품과 서비스들의 대부분이 오 년 전에는 존재하지 않았던 것들이라고 보고한다. 동시에 디지털 정보는 쉽게 손실되거나 훼손될 수 있기 때문에 디지털 형태로 된 정보의 폭발적인 증대는 기관들과 그들의 정보제공자들에게 심각한 당면 문제를 제기한다. 더 나아가 기술발전의 속도는 미래에 기관들을 위해서 정보를 표현하기 위한 기존의 데이터 구조나 포맷의 능력에 심각한 압력을 야기하고 있다. 디지털 정보를 보존하는 데 필요한 지원 정보는 원본 디지털 정보가 생산되는 당시에 이용할 수 있거나, 오직 그때만 이용할 수 있다. 보통 정보가 생산된 이후에, 작동하는 소프트웨어는 개선될 수도 있고, 그것의 버전이 바뀔 수도 있다. 기관들은 장기적으로 보존하기 시작해야 하고, 그렇지 않으면 그들의 정보가 영원히 손실될 것이다.

보존을 위해 기관들이 맡을 여러 기능들이 있다. 일부 국립 아카이브즈와 연구 그룹들을 제외하고, IT 종사자들은 일반적으로 이 기능들을 무시한다. 우리는 여기서 그것들 가운데 세 가지 기능을 강조한다. 기관들은 커뮤니티의 요구들을 지켜보고, 그들의 서비스 요구조건과 이용 가능한 제품기술에서 변화를 추적하기 위해 소비자들 및 제작자들과 상호 작용해야한다. 그 같은 요구조건들은 데이터 포맷, 매체 선택, 소프트웨어 패키지, 컴퓨팅 플랫폼 및 디지털 아카이브즈와 커뮤니케이션 하기 위한 메커니즘을 포함할 것이다. 이 첫째 기능은 피드백을 구할 수 있는 서베이, 정기적인 공식적 리뷰 처리, 커뮤니티 워크숍을 통해, 그리고 개인적 상호작용에 의하여 달성될 수 있을 것이다. 그것은 미래의 보존 전략과 표준을 발전시키기 위해 보고서, 요구조건 경보 및 새로 등장하는 표준을 제공한다. 그것은 보존 요구조건들을 디지털 아카이브즈 개발자와 관리자들에게 보낸다. 두 번째 기능은 아카이빙 컴퓨팅 환경에서 기술의 노화를 야기할 수 있고, 아카이브즈의 현재 보유자원 일부에 대한 접근을 막을 수도 있는 기술들을 식별하기 위해 새로 등장하는 디지털 기술, 정보표준 및 컴퓨팅 플랫폼(즉, 하드웨어와 소프트웨어)을 추적하는 책임이다. 이 기능은 새로 등장하는 기술을 더 잘 평가할 수 있기 위한 프로토타입으로 만드는 능력을 포함할 수도 있고, 디지털 아카이브즈 개발자로부터 모형을 입수할 수도 있다. 이 기능은 또한 디지털 아카이브즈로 하여금 일부 현재의 보유자원이나 새로운 입수자원의 이동을 요구할 수도 있는 커뮤니티 서비스 요구조건이나 기술 동향에서의 미래의 변화를 더 잘 예측할 수 있도록 전략과 표준을 개발하고 권장하는 데 책임이 있다. 세 번째 기능은 디지털 아카이브즈 관리자로부터 표준 및 이동 목표를 승인한다. 표준들은 포맷 표준, 메타데이터 표준 및 문서화 표준을 포함한다. 그것은 이 표준들을 보존 요구조건에 적용한다. 이 기능에 의하여 받아들여진 이동 목표는 기술 노화 때문에 생기는 접근의 손실을 회피하기 위해 콘텐츠를 변형하는 것을 포함하여, 보존 패키지의 변형을 포함한다. 이동 목표에 대한 반응은 새로운 보존 디자인, 프로토타입 소프트웨어, 시험 계획, 커뮤니티 리뷰 계획 및 이행 계획의 발전을 포함할 수도 있다.

보존 비용이 계속해서 증대하고 있는 것에 대응해서, 디지털 아카이브즈는 장기적으로 기본적 정보를 보존하기 위해 확고한 정책과 전략을 필요로 한다. 정책과 전략을 발전시키기 위해서, 디지털 아카이브즈는 일반적으로 인정되는 정보 구조나 생명주기를 필요로 한다. 정보의 생명주기는 보존과 접근을 위한 것만이 아니라, 디지털 아카이브즈의 완전한 비즈니스 모델을 위한 것이다. 다음에서는 조직 구조의 중요한 측면을 논하도록 한다. 정보의 구체적인 생명주기는 다음 장에서 논의될 것이다.

업무 흐름 처리는 비즈니스 처리를 컴퓨터화하여 표현한 것이다. 그것은 어떤 순서로 수행되어야 하는 비즈니스 처리의 다양한 활동들, 활동들 사이의 데이터의 흐름, 그리고 공통 목표를 수행하기 위해 활동을 하는 복수의 협력 기관들을 명시하고 있다. 업무 흐름 관리 시스템은 업무 흐름을 정의하고, 실례를 들어 보여주고, 실행하기 위한 소프트웨어 시스템이고, 현재 비즈니스 처리(예를 들어, 재정 시장, 은행,



〈그림 2〉 디지털 사회의 업무 흐름 체계

소매상, 교통 등)를 지원하는 주요한 기술이다.

기존의 업무 흐름 솔루션들은 프로젝트 정보와 어플리케이션들 사이의 정보 교환의 표현 능력에 초점을 맞춘다. 그러나 계속 변화하는 기관들의 속성은 정보(예를 들어, 레코드)를 역동적 엔티티로서 다룰 수 있는 능력이 있는 솔루션을 요구한다. 정보는 비즈니스가 사전에 계획된 활동들로부터 정상적으로 진보하는 것 때문에, 혹은 불확실성으로 인하여 발생하는 사건들 때문에 변화한다. 두 가지 경우에 다, 그것들은 정보의 생명주기의 결과이다. 우리는 융통성 있고 의미 있는 업무 흐름 시스템의 설계가 상호 운용될 수 있고, 역동적인 정보 처리 모델을 통하여 시도될 수 있고, 시도되어야 한다고 믿는다.

이 구조는 데이터와 서비스 제공자들이 업무 흐름 시스템에서 인터넷을 통하여 상호 작용할, 융통성 있는 업무 흐름 처리 구조에서 참여자들과 정보 객체들이 유동할 수 있게 허용한다. 그 같은 시스템 구조를 위해서, Open Archives Initiative가 제안한, 서비스 제공자들에 의한 데이터 제공자들의 메타데이터를 공개적으로 수확하기 위한 프로토콜이 있다(Lagoze 외 2001; Chen, Choo & Chow; InterPARES Project). 데이터 제공자들은 충분한 양의 메타데이터를 노출시킴으로써, 자기들의 콘텐츠를 다른 데이터 제공자와 서비스 제공자들이 검색하고 이용하도록 공개적으로 광고한다. 이 공개적 구조는 보안 메커니즘들에 의하여 확실히 보호되어야 한다. 이 논문에서 이 문제는 논의되지 않을 것이다.

데이터베이스 시스템은 더 나아가서(기관들의 기능성에 따라) 특정한 서비스 프로토콜을 통하여 업무 흐름 시스템에 접속된다. 업무 흐름 엔진은 그 기반이 되는 정보 모델을 갖고 업무 흐름 처리를 수행하고 관리한다. 어플리케이션들(예를 들어, 디지털 소사이어티 서비스)은 분야마다 특정한 미들웨어 층에 존재하게

되고, 그것도 이 논문에서는 논의되지 않는다.

업무 흐름 구조의 융통성은 방향이 정해진 도표로서 기본적 업무 흐름 구축 블록을 구축하고, 기본적인 업무 흐름들을 도표의 재작성으로서, 여러 개의 처리들이 보다 복잡한 처리들로 구성될 수 있는 저장소(리파지토리)로 색인 함으로써 달성된다.

업무 흐름이 실행되는 동안, 하나의 업무 흐름은 온라인의 제한 때문에 수정될 수 있고, 나중에 앞으로 사용할 새로운 엔트리로서 저장소에 받아들여질 수 있다. 업무 흐름을 아카이벌 구성요소로서 취급한다는 생각은 장기적으로는 보존 원칙을 대단히 개선한다. 보존을 위한 업무 흐름 시스템의 디자인은 명백히 많은 연구를 필요로 한다. 우리는 업무 흐름을 융통성 있고 역동적으로 구성하기위해서 도표의 재작성, 페트리넷(Petri-net), 및 확장 트랜잭션(Extended Transaction)과 같은 공식적 모델들의 적용 가능성을 연구하는 데 노력을 기울여 왔다. 보다 상세한 결과는 다른 곳에서 보고될 것이다. 그 같이 상호 운용 가능하고 역동적인 정보 모델은 분산된 이용자를 갖는 네트워크로 연결된 어플리케이션 서비스, 레코드(혹은 객체들), 업무 흐름 처리, 및 그것들의 저장소로 구성된다. 이런 유형의 어플리케이션 서비스는 역동적 레코드(혹은 객체) 모델에 기반하여 이행될 수 있다.

4. 레코드의 아카이브적 안정성

객체-중심 방법론은 정보 기술, 보존 및 따라서 기록관리학의 표준적 표현 스킴이 되었다(Shepard). 정보의 캡슐화 원칙은 디지털 레코드를 객체로 표현하는 방법을 제공한다. 그것은 정보 콘텐츠를, 필요할 때마다 적용되는 그것과 동반하는 절차로 둘러싼다. 정보의 생명주기에서 처리들은 "레코드" 객체들이라는 복잡한 객체들로 표현될 수도 있다(Chen 1998). 각 기관의 기반구조에서, 우리는 정보의 생명주기의 객체 중심 구조를 사용할 것이다. 생명주기는(앞 장에서 논의된 다양한 업무 흐름 처리들을 포함하여) 입수, 보존, 콜렉션, 색인, 접근 및 활용으로 구성되고, 보존은 생명주기의 중요한 요소이다.

4.1 객체로서의 디지털 레코드

정보의 캡슐화 정책은 생명주기 상의 다양한 멀티미디어 데이터 콘텐츠의 일반적인 표현 스킴을 제공한다. 멀티미디어 데이터 콘텐츠는 그것의 표현 정보 및 동반되는 소프트웨어 프로그램으로 둘러싸지고, 이들은 필요할 때마다 적용된다. 표현 정보는 데이터 콘텐츠의 비트 스트림을 특정한 포맷, 구조, 및 유형의 이해할 수 있는 정보로 매핑 한다. 따라서 데이터 콘텐츠, 표현 정보 및 소프트웨어 프로그램들은 모듈화되고, 재사용이 가능하게 된다. 디지털 객체들은 정보의 생명주기의 처리들 아래에서 변환될 수 있고, 변형될 수 있으며, 사용될 수 있다. 디지털 객체들이 캡슐화되기 때문에, 그것들은 활동성이 있고, 역동적이며, 확장될 수 있다. 그것들이 활동성이 있는 것은, 소프트웨어 에이전트들이 객체들에 삽입되어, 객체들에 의하여 활동들이 시작될 수도 있기 때문이다. 그것들은 객체들에 동반하는 소프트웨어 프로그

램과 관련된 어떤 처리 하에서도 역동적이다. 그것들은 멀티미디어 콘텐츠와 네트워크로 연결된 소스들이라는 의미에서 확장될 수 있다. 어떤 객체든 멀티미디어 콘텐츠의 다른 객체들에 의하여, 그리고 네트워크로 연결된 소스들로부터 확대될 수 있다. 디지털 레코드가 여러 개의 네트워크로 연결된 레코드들을 그 자체로 끌어오는 것은 매우 그럴듯한 일이다. 멀티미디어 디지털 콘텐츠는 하나 이상의 비트 시퀀스로 구성된다. 표현 정보의 목적은 비트 시퀀스를 보다 의미 있는 정보로 변환시키는 것이다. 표현 정보는 비트 시퀀스에 적용되어야 하고, 결과적으로 문자, 숫자, 화소, 배열, 도표, 등과 같이 보다 의미 있는 값을 가져오는, 포맷이나 데이터 구조 개념들을 기술함으로써 이것을 수행한다. 단순히 말하자면, 우리는 그 같이 활동성이 있고, 역동적이며, 확장될 수 있는 객체들을 한 마디의 형용사로, 즉 "역동적(dynamic)"이라고 표현한다. 따라서 이 논문에서 "역동적"이라고 하는 것은 활동성이 있고, 역동적이며, 확장될 수 있다는 특성들을 갖는다는 의미이다.

이 객체들을 위해서 우리가 어떻게 보존 전략을 발전시킬 수 있을 것인가? 데이터 유형, 그것들의 집합, 및 기초가 되는 데이터 유형으로부터 보다 높은 수준의 개념들로 매핑 하는 매핑 규칙들은 표현 정보의 구조 정보 구성요소로 여겨진다. 이 구조들은 보통 이름이나 연관된 비트 시퀀스 내에서의 상대적 위치에 의하여 식별된다. 구조 정보 구성요소에 의하여 제공되는 표현 정보는 일반적으로는 디지털 콘텐츠를 이해하는 데 불충분하다. 추가적으로 요구되는 정보는 의미 정보(semantic information)이다. 의미 정보는 상당히 복잡할 수도 있다. 그것은 구조 정보의 모든 요소들, 각 데이터 유형에서 수행될 수도 있는 처리들, 및 그것들의 상관관계들과 연관된 특수한 의미들을 포함할 수도 있다. 더욱이 표현 정보는 다른 표현 정보에 대한 더 많은 관련된 참조들을 포함할 수도 있다.

디지털 객체를 보존하기 위해서, 그것의 구조적 및 의미적 표현 정보 둘 다 또한 보존되어야 한다. 이것은 보통 표현 정보가 디지털 버전을 위한 ASCII 문자와 같이 널리 지원되는 표준을 사용하는 텍스트 기술로 표현될 때 달성된다. 만약 텍스트 기술이 모호하면, 우리는 데이터 구조를 기술하기 위해 잘 정의된 구문들을 포함하고 있는, 표준화된 공식적 기술언어들(예를 들어, XML 마크업 언어)을 사용해야 한다. 이 마크업 언어들은 표현 정보의 의미를 완전히 전달하기 위해 텍스트 기술을 보완할 것이다.

디지털 객체들과 일반적으로 연관된 소프트웨어 프로그램들은 표현 변환(rendering) 소프트웨어와 접근 소프트웨어이다. 표현 변환 소프트웨어는 레코드를 인간 가독형으로 변환시키는 PDF 디스플레이 소프트웨어와 같이, 표현 정보를 인간 가독형으로 디스플레이 할 수 있다. 접근 소프트웨어는 디지털 객체의 정보 콘텐츠를 일부 혹은 전부 인간이나 시스템이 이해할 수 있는 형태로 제시한다. 그것은 또한 디스플레이, 조작, 처리 등과 같은 접근 서비스의 일부 유형들을(예를 들어, 시계열이나 다차원 배열을 지원하는 과학적 시각화 시스템과 같은) 다른 객체에 제공할 수도 있다. 다시 한번, 그것의 미래의 존재와 이동은 예측하기가 매우 어려운 기술 연속성에 의존한다. 표현 변환 소프트웨어와 접근 소프트웨어는 데스크탑에서 제공되기 때문에, 그것들의 보존은 각 객체 수준에서 필요

하지 않고, 환경 수준에서만 필요하다.

비용 효과적 방법으로서 일부 표현 정보를 통합하기 위해 인터넷-기반 접근 소프트웨어를 사용하는 것은 솔깃한 일이다. 많은 웹-기반 서비스는 실제로 완전한 표현 정보로서 웹 접근 소프트웨어를 사용하고, WWW 컨소시엄이 이러한 노력에 대해 탁월한 작업을 하고 있다. 접근 소프트웨어 소스 코드는 적어도 그 디지털 객체들의 부분적 표현 정보가 된다. 우선, 그 같은 정보는 다양한 다른 처리 및 디스플레이 알고리즘과 혼합될 수도 있고, 그 코드가 특정한 기초가 되는 운영 환경을 취하기 때문에 불완전할 수도 있다. 둘째, 만약 접근 소프트웨어의 실행 가능한 기능들이 사용되면, 소스 코드가 없이는 그 같은 아카이브즈는 표현 정보를 상실하는 상당한 부담을 갖게 된다. 장기간에 걸쳐 레코드를 이동하는 것보다 소프트웨어용 운영 환경을 유지하는 것이 더 어렵다. 만약 기관의 컴퓨팅 환경이 소프트웨어를 지원하면, 보존된 패키지에 접근하는 것은 어렵지 않다. 환경은 기초가 되는 하드웨어와 운영 시스템, 운영 시스템을 효과적으로 보완하는 다양한 유틸리티, 그리고 저장과 디스플레이 장치 및 그것들의 드라이버로 구성된다. 이것들 중 어느 것이라도 변화하면 소프트웨어가 더 이상 적절하게 기능하지 못하게 만들 것이다. 이것 때문에 소프트웨어의 보존이 복잡하고 까다롭다. 한 수준 더 나아가기 위해, 표현 정보는 디지털 콘텐츠를 표현하는 데 사용된 어떤 자연언어(예를 들어, 영어)의 사전과 문법을 포함할 필요가 있을 수도 있다. 자연언어 표현의 의미는 오랜 시간 동안 일반적 및 특정한 학문분야에서 상당히 발전할 수 있다.

4.2 보존과 패키지 정보

보존의 중요한 단계는 필수적인 보존 정보를 디지털 객체에 묶어서(bundling), 하드웨어, 소프트웨어, 및 매체 이동이 어떻게 발전하든, 그 콘텐츠를 그래도 접근하고 검색할 수 있게 하는 것이다. 보존 패키지는 콘텐츠 정보와 보존 정보라는 이 두 가지 유형의 정보를 개념적으로 담는 그릇이다(National Digital Information Infrastructure & Preservation Program). 콘텐츠 정보와 보존 정보는 패키지 정보에 의하여 캡슐화 되고 식별될 수 있다. 그 결과인 패키지는 보존 패키지의 기술 정보에 의하여 접근될 수 있다. 콘텐츠 정보는 보존의 원본 타깃 정보이다. 그것은 디지털 콘텐츠 및 그것과 연관된 표현 정보와 소프트웨어 프로그램으로 구성된다. 보존 정보는 콘텐츠 정보에 적용되고, 콘텐츠 정보를 보존하고, 그것이 명확하게 식별되도록 보장하며, 그것이 생성된 환경을 이해하기 위하여 필요하다.

보존 정보는 네 가지 유형의 보존용 정보로 구분된다: 출처(provenance), 상황(context), 참조(reference), 및 고정성(fixity). 간략히, 이들은 다음과 같은 네 개의 범주들에서 기술된다:

1. 출처(Provenance)는 콘텐츠 정보의 역사와 소스를 기술한다: 그 정보가 생산 된 이래 누가 관리해 왔는가 및 이력(처리 이력을 포함). 이것은 향후의 이용자들에게 콘텐츠의 가능한 신빙성에 대해 어느 정도의 확신을 준다. 출처는 특수한 유형의 상황(context) 정보로 간주될 수 있다.
2. 상황(Context)은 콘텐츠 정보가 정보 패키지 외부의 다른 정보와 어떻게 관련되

는지를 기술한다. 예를 들어, 그것은 콘텐츠 정보가 왜 생산되었는지를 기술하기도 하고, 그것이 이용할 수 있는 다른 콘텐츠 정보 객체와 어떻게 관련되는지에 대한 기술을 포함할 수도 있다.

3. 참조(Reference)는 그것으로 인해 콘텐츠 정보가 고유하게 식별될 수 있는, 하나 이상의 식별자나 식별자 시스템을 제공한다. 그 실례로는 책의 ISBN 번호나 한 인스턴스의 콘텐츠 정보를 다른 것과 구별하는 속성들의 집합을 포함한다. 더 예를 든다면, 분류 체계, 참조 체계 및 등록 체계 등을 포함한다.
4. 고정성(Fixity)은 콘텐츠 정보가 인증되지 않고 변형되는 것을 막아주는 포장 혹은 보호막을 제공한다. 그것은 특정한 콘텐츠가 인증되지 않은 방식으로 수정되지 않도록 보장하기 위해 사용되는 데이터 무결성 체크 혹은 인증/확인 키들을 제공한다.

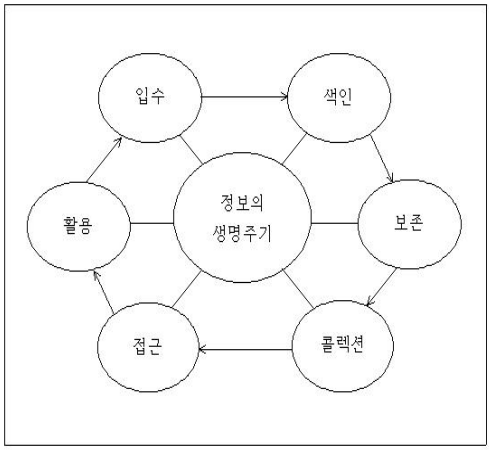
패키지 정보는 실제적으로 혹은 논리적으로 콘텐츠 정보와 보존 정보를 엮고, 식별하고, 관계를 맺는 정보이다. 패키지의 기술 정보는 어떤 패키지가 관심 대상인 콘텐츠 정보를 갖는지 발견하는 데 사용되는 정보이다. 그것은 목록 서비스에서 검색 가능한 완전한 세트의 속성들이나 메타데이터일 수도 있다. OAIS (CCSDS & ISO 2001) 에서, 무한정한 기간 동안의 전체 아카이벌 정보는 보존 기술 정보(preservation description information, PDI)라고 불린다. 패키지 정보는 그것이 콘텐츠 정보나 PDI에 기여하지 않기 때문에, 반드시 보존될 필요가 있지는 않다. 보존은 또한 디렉토리나 파일명 구조들의 명명 방식으로만 PDI나 콘텐츠 정보를 보유하는 것을 피해야 한다. 이 구조들은 패키지 정보로서 사용될 가능성이 가장 많다. 패키지 정보는 이동에 의하여 보존되지 않는다. 파일명이나 디렉토리 구조에서 저장된 어떤 정보든 패키지 정보가 변경될 때 손실될 수도 있다.

4.3 정보의 생명주기

정보의 생명주기는 또한 객체-중심 구조에서 표현된다. 생명주기는 적어도 다음과 같은 정보의 처리들로 구성된다: 입수, 보존, 콜렉션, 색인, 접근, 및 활용. 특히, 활용은 기관 내에서 많은 다른 업무 흐름 처리들을 더 포함할 수도 있다. 어떤 기관에서든, 정보는 외부의 거래 소스에서 입수되고, 그 자체의 업무 흐름 처리에서 생산되고, 혹은 정보원들(예를 들어, 도서관과 아카이브즈)에서 접근된다. 생명주기는 레코드가 기관에 의하여 입수된(유입되거나 창안된) 직후에 시작된다. 입수된 레코드는 장기적 이용을 위하여 보존될 것이다. 모든 보존된 레코드들은 적당한 콜렉션에 저장되고, 각각은 향후의 접근과 활용을 위해 적절하게 색인된다. 입수가 반복적 패턴으로 활용을 뒤따르기 때문에, 생명주기는(선형이기보다는) 나선형 주기를 취한다. 활용에 의하여 생산된 정보는 자주 생명주기로 유입된다.

생명주기에는 다음과 같은 보다 이차적 처리들이 있다:

- 변경과 변형(Conversion and transformation),
- 커뮤니케이션과 전송(Communication



〈그림 3〉 정보의 생명주기

and transmission),

- 중개와 통합(Brokerage and integration),
- 전달과 제시(Delivery and presentation).

이차적 처리들을 완전히 열거하는 것은 불가능하다. 이 이차적 처리들은 생명주기의 다양한 단계들에 삽입될 수도 있다. 이 논문에서 업무 흐름 처리들은 일반적인 형태로만 가정되고, 상세히 기술되지는 않는다. 업무 흐름 처리들은 이차적 처리들을 포함할 수도 있다. 예를 들기 위해, 몇 가지의 이차적 처리들에 대하여 논하도록 한다. 디지털 객체들의 포맷과 구조의 변환과 변형은 이들 간의 상호 교환을 허용한다. 그것들은 예를 들어 보존에 사용된다. 커뮤니케이션과 전송은 디지털 객체들을 컴퓨터 하드웨어와 저장 시스템으로부터 커뮤니케이션 네트워크를 통하여 보낸다. 그것들은 아마도 생명주기의 모든 처리에서 필요할 것이다. 중개와 통합은 네트워크 소스들로부터 온 쿼리 결과들을 이용자들을 위해 통합된 객체들로 중개한다. 그것들은 접근 처리에서 필수적이다. 전달과 제시는 이용자들이 접근하고 활용하는 데 있어서 유용한 방식으로 정보를 제시한다.

5. IT 산업에서 기술 연속성

IT 발전은 개인과 기관들이 그들의 디지털 레코드와 콜렉션을 보존하는 데 있어서 승리를 가져온 경이적인 것이다. 개인 시민들이 자신의 콜렉션을 만들 수 있을 뿐 아니라, 인터넷에서 공정한 이용을 조건으로 다운로드할 수 있는 많은 자료들이 있다. 그러나 디지털 레코드의 기관적 보존은 복수의 선택조건들이 있고, 무엇을 보존해야 할지, 혹은 디지털 사회에서 새로운 비즈니스 처리 방법이 무엇일지 알지 못한다는 사실 때문에 마치 "바벨탑"과 같다. 예를 들어, "구글"에서 디지털 레코드들이 보존될 것인지 알지 못한다. 정보 폭발은 유례 없는 디지털 레코드의 양을 만들고 있고, 기관들은 이들을 관리하고 보존하기에 잘 준비되어 있지 않다. 디지털 콜렉션들의 숫자가 증대하고 있기 때문에, 그것들에 대한 정보 접근의 상호 운용성 및 견고성을 관리하는 것 또한 복잡하다. 이것은 IT 산업이 파트너로서 해결하도록 기여할 수 있는 디지털 보존 문제의 난제이다.

6. 결 론

우리는 디지털 보존 문제에 잠재적 해결책을 제안하였다. 우리의 논의에서 명백한 것은 기관적 처리, 아카이벌 안정성 및 기술 연속성이

상호 작용하고, 이 방향으로 보다 혁신적인 연구가 수행되어야 한다는 것이다.

▷ 번역: 윤정옥(청주대학교 문헌정보학과 교수)

참 고 문 헌

CCSDS and ISO TC20/SC13, "Reference Model for an Open Archival Information System," July 2001. 〈http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf〉.

S. Chen, The paradox of digital preservation, IEEE Computer, March 2001.

S. Chen, Digital Libraries: The Life Cycle of Information, BE Publisher, 1998.

S. Chen, Digital preservation and the life cycle of information, Advances in Computers, M. Zelkowitz(ed.), volume 57, Academic Press/Elsevier Science, 2003.

S. Chen, Digital preservation and workflow process, in ICADL(International Conference on Asian Digital Libraries), Lecture Notes in Computer Science, Springer Verlag, Dec. 13-17 2004, Shanghai, China.

C. M. Dollar, Archival Theory and Information Technologies: The Impact of Information Technologies on Archival Principles and Methods, Macerata: University of Macerata Press, 1992.

L. Duranti, Diplomatics: New uses for an old science(Part V), Archivaria, 32 1991, pp.6-24.

J. Garrett and D. Waters, Preserving digital information, Report of the Task Force on Archiving of Digital Information, May 1996. 〈http://www.rlg.org/ArchTF/〉.

M. Hedstrom, Descriptive practices for electronic records: Deciding what is essential and imaging what is possible, Archivaria, 36, 1993, pp.53-63. InterPARES,〈http://www.InterPARES.org/〉.

J. Rothenberg, Ensuring the longevity of digital documents, Scientific American, 272, 1995, pp.42-47.

J. Rumbaugh, M. Blaha, W. Premerlani, F. Eddy, and W. Lorensen, Object-Oriented Modeling and Design, Prentice Hall, 1991.

T. Shepard, UPF User Requirements. 〈http://info.wgbh.org/upf/〉.

Victorian Electronic Records Strategy Final Report, Public Record Office Victoria 1999. 〈http://www.prov.vic.gov.au/vers/〉.

National Digital Information Infrastruc- ture and Preservation Program. 〈http://www.digitalpreservation.gov/〉.

OCLC. 〈http://www.oclc.org/services/preservation/default.htm〉.

IBM Vision of Autonomic Computing (IEEE Computer January 2003). 〈http://www.ibm.com/research/autonomic/〉.

NARA ERA. 〈http://www.archives.gov/electronic_records_archives/research/research.html〉.

Australian Standards and Framework: Records Management and Metatagging of Web Pages. 〈http://www.lester.boisestate.edu/metatags/Australian_Standards_and_Framework1.htm〉.

Sue McKemmish, Describing Records in Context in the Continuum: the Australian Recordkeeping Metadata Schema. 〈http://www.sims.monash.edu/research/rcrg/publications/archiv01.htm〉.

H. Kim and S. Chen, Ontology Search and Text Mining of MEDLINE Database, Conference on "Data Mining in Biomedicine," February 16-18, 2004, University of Florida, Gainesville, FL; Book published by Kluwer.

S. Shi, O. Rodriguez, S. Chen and Y. Shang, Open learning objects as an intelligent way of organizing educational material, International Journal on E-Learning, Vol. 3 No. 2, 2004, pp.51-63.

H. Kim, C. Choo, and S. Chen, An Integrated Digital Library Server with OAI and Self-Organizing Capabilities. In Proceedings of the 7th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2003), Trondheim, Norway, August(2003)

C. Lagoze and H. Van de Sompel, The Open Archives Initiative: Building a low-barrier interoperability framework. In Proceedings of the First ACM/IEEE Joint Conference on Digital Libraries, Roanoke, VA, 2001. Pages 54-62.

C. Lagoze, H. Van de Sompel, M. Nelson and S. Warner, TheOpen Archives Initiative Protocol for Metadata Harvesting. Open Archives Initiative, 2001. 〈http://www.openarchives.org/OAI/openarchivesprotocol.htm〉.

S. Chen, C. Choo, and Y. Chow, Internet Security: A Novel Role/Object-Based Access Control for Digital Libraries, Journal of Organizational Computing and Electronic Commerce, Lawrence Erlbaum Publisher, under revision.

InterPARES Project. 〈http://www.interpares.org〉.