도서관 자료의 디지털화에 대한 연구
A Study on Digitizing Library Materials

남영준, 전주대학교 문헌정보학과

Young-joon Nam, Dept. of Library & Information Science, Jeonju University

기존의 도서관은 이용자의 다양한 정보욕구와 정보량의 증가, 컴퓨터기술 및 정보통신기술의 획기적 발달에 따라 머지 않은 장래에 디지털 도서관으로 변신할 것이다. 그러나 소장자료의 디지털화는 많은 제약과 문제점이 있기 때문에 이에 대한 체계적인 검토와 방안이 마련되지 않을 때에는 디지털화자체가 이루어지지 않을 수 있다. 본 연구에서는 도서관 소장자료를 디지털화하는데 야기될 수 있는 문제점을 조사하고, 그 해결점을 제시한다.

1. 서론

전통적으로 도서관은 이용자의 정보욕구를 충족시키고 위해 많은 규칙과 정보검색도구들을 개발하였다. 한편, 컴퓨터를 이용하여 이러한 정보욕구를 좀 더 신속하고 정확하게 제공하기 위한 노력도 병행되었다. 인터넷과 같은 새로운 정보전달매체가 개발되면서 도서관은 이를 이용하여 이용자들에게 보다 효율적이고 편리하게 정보를 제공받을 수 있는 방법을 능동적으로 연구하기에 이르렀다. 즉, 이용자들이 원할 경우에 시간과 장소에 구애받지 않고 필요한 자료에 대한 원문을 열람할 수 있는 방안을 모색하기에 이르렀다.

현재, 기술적으로는 인터넷상에 수록된 (up-load) 일부 자료들의 원문을 통신선으로 이용자가 직접 데이터를 받아볼 수 (down-load) 있는 수준에 이르렀다.

통신선상에서 자료의 원문을 받아볼 수 있다는 것은 해당 자료들이 디지털화한 형태 (digitized materials)로써 각 사이트들의 호스트에 事前에 입력되어 있다는 것을 뜻한다.

도서관계에서 주창하고 있는 전자도서관도 소장 자료들을 디지털화하여 이용자에게 원문을 제공할 수 있는 全文데이터베이스(full-text database)를 구축하는 것을 의미한다. 도서관내 전문데이터베이스의 구축은 원문을 소급입력하는 것이며, 이 사업에는 엄청난 경비와 인력이 투입되어야 한다. 이 규모는 일반 도서관들로서는 감당하기 어려운 수준으로서 이에 대한 정확한 이해와 해결방법이 제시되지 않을 경우, 전자도서관은 탁상공론으로 머무를 수 있다.

즉, 자료의 디지털화는 엄청난 인력과 경비가 소요되며, 이에 대한 체계적인 통제나 협의가 없을 경우에는 국가 자원의 심각한 낭비를 초래할 것이며, 디지털화 사업자체가 불가능하게 될 것이다.

본 연구는 이 점의 중요성을 인식하고, 전자도서관 구축의 실제적인 구현에 있어 가장 기본적이면서 핵심적인 소장자료의 디지털화에 대한 방법과 문제점, 현실적인 해결방안을 제시하고자 한다.

1. 디지털화 방법

1.1 키인방법 (key-in method)

기초자료코퍼스를 구축하기 위한 가장 원시적이고 정확한 입력의 질을 확보할 수 있는 방안가운데 하나가 키인(key-in)방식이다. 이 방식은

방식명대로 기초자료로 선정된 데이터를 사람이 직접 키보드를 사용하여 입력하는 방식이다.

이 방법의 장점은 학술문헌에 나타난 특수기호 혹은 한자, 외국어원문을 입력자가 그대로 입력할 수 있으므로 디지털자료구성내용에 영향을 받지 않는다. 상대적으로 입력의 질이 가장 안정적이다. 또한, 1차 데이터 구축후 검수비용이 상대적으로 저렴하며, 검수자의 노동 양이 상대적으로 적은 수준이다.

1.2 후처리방법 (post-processing method)

소장자료의 디지털화에 노동 및 경비를 절감하기 위해 하드웨어를 사용하는데, 그 방안가운데 가장 일반적인 것이 스캐너를 활용하여 문자 인식을 하고, 최종적으로 검수자가 교정하는 방식이 있다. 이 방법은 완전한 기계처리는 아니며 많은 부분에서 컴퓨터가 문자인식프로그램을 이용하여 문헌인식을 하고 있다. 이 방안은 스캐너와 문자인식소프트웨어에 따라 효율이 크게 차이가 발생한다. 이 과정가운데 수작업으로 하는 부분은 ① 자료의 선정을 비롯하여 ② 스캐닝, ⑤ 결과편집, ⑥ 구축자료검수이며 이에 대한 수작업은 정확히 구분된 데이터는 없으며 통계적으로 ①~④ 과정의 경우 하루에 약 2권정도를 처리할 수 있다.

이 때 야기되는 문제로는 현재 국내에서 개발된 문자인식소프트웨어[1]의 경우에 문자 인식율이 약 60%미만에 머무르고 있어 구축자료 검수과정에 2차적으로 많은 시간과 경비가 투입되어야 하는 점이다. 스캐너 활용의 가장 큰 제한점으로는 한자와 같이 한글과 알파벳을 제외한 외래어와 특수문자의 인식자체에 어려움이 있기 때문에 학술문헌에서 자주 출현하는 수식이나 각주를 본문대로 표현할 수 없다는 점을 들 수 있다.

2. 디지털화 비용

디지털화에 소요되는 비용은 순수하게 디지털화에 따른 입력비용과 저작권료가 있으며, 그밖에 데이터베이스 구축에 소요되는 기계구입비와 운영유지비가 있으나 본 연구에서는 순수하게 디지털화에 소요되는 경비만을 조사하였다.

2.1 디지털화 비용

2.1.1 그래픽 처리

그래픽 처리는 문헌 자체를 사진형태의 파일로 저장하는 방법이다. 이는 전적으로 스캐너의 능력에 좌우되며, 또한, 플랫베드형과 양면인식형에 따라 많은 차이가 있다. 그러나 소장자료 보관의 측면이 있기 때문에 플랫베드형 스캐너로 결과편집과 구축자료검수에 소요되는 시간은 약 12시간정도가 소요됨에 따라 약 1.5일정도의 시간이 소요된다. 일반적으로 소요되는 경비는 수작업처리의 2/5이하정도의 비용과 노력이 투입된다.

2.1.2 ASCII 처리

ASCII 처리경비는 자료 입력비와 입력자료 검수비로 구분된다. 첫째, 입력비로서 학술문헌을 대상으로 한 권을 입력하는데 약 20시간이 소요된다. 이 작업량은 약 이틀의 노동력이 필요하며, 금액으로 환산하면 약 10만원이 소요된다.

이와는 별도로 입력자료 검수비로서는 한 권에 대해 약 8시간[2]이 소요되며 비용으로는 약 1만원이 소요된다. 그러므로 약 300페이지정도의 학술문헌 한 권을 전산입력하는데 약 11만원이 소요된다.

2.2 저작권료

도서관 소장자료의 디지털화에 가장 큰 문제점으로 저작권을 들 수 있다. 이에 대한 구체적인 법적 근거나 판례는 아직 국내외적으로 구체화된 것이 없는 실정이기 때문에 이에 소요되는 경비는 정확하게 추정될 수 없다. 단, 국내 온라인으로 전자서적을 열람 및 판매하는 회사의 경우, 원저작물의 인세의 형식으로 저작료를 지불하고 있으며, 저작료는 판매예상횟수와 인기도에 따라 책자형 자료 정가의 25%-40%를 지불하고 있다. 일괄적인 인세지급형식은 아니며 열람 및 다운로드 횟수에 따라 월별 지급을 하고 있다.

3. 디지털화의 문제점

디지털화과정에서 야기될 수 있는 문제점은 앞절에서 제시한 바와 같이 문헌의 내용을 그래픽형태로 저장할 것인지 혹은 ASCII형태로 저장할

1) 1996년도에 개발된 제품을 대상으로 실험하였음.

2) 현재 전주대학교에서 구축하고 있는 기초자료코퍼스의 평균데이터.

것인지에 따라 차이가 있다.

3.1 그래픽 파일일 경우의 문제점

도서관에서 소장하고 있는 문헌자료를 전문데이터베이스로 변환할 경우에 지금까지는 주로 이미지 형태의 데이터를 많이 사용하였다. 이때 도서관 소장자료가 흑백문서인가와 컬러문서일 경우에 따라 저장의 형태가 달라질 수 있다. 일반적으로 흑백문서일 경우, 문서의 처리는 산업계 표준인 TIFF(CCITT Group IV)형식을 사용한다. 대부분의 흑백문서는 이미지 처리의 경우, TIFF 방식을 선호하고 있기 때문에 기존 문서 처리와의 연관성 및 통일성의 측면에서 비용효과를 고려하면 TIFF가 적합한 표준이 될 것이다. 또한, 기존에 상품화가 되어 있는 이미지 처리시스템(이미지 브라우저 혹은 OCR처리기 등)의 대부분이 기본적으로 TIFF방식을 지원하고 있다.

컬러 문서는 표현해야 할 정보가 많기 때문에 TIFF방식으로 처리하는 것은 양질의 이미지 데이터를 처리하는데 부적절하다. 따라서 컬러문서의 처리는 산업계 표준으로 널리 사용되고 있는 GIF방식을 사용하고 있다. 만약, 문헌 자체가 사진으로 구성된 자료(圖鑑類 등)일 경우나 혹은 사진과 글이 함께 있는 자료일 경우는 JPEG형식을 사용한다. 왜냐하면 JPEG방식의 이미지 데이터는 파일 크기가 상대적으로 다른 저장방식에 비해 작지만, 파일의 재현수준은 이용자의 육안으로 원문과의 차이가 거의 없기 때문이다.

웹상에서 제공하고 있는 컬러 문서의 제공형태도 대부분 GIF나 JPEG방식을 준용하고 있으며, 이 두방식은 서로 변환이 매우 자유롭게 이루어지고 있다.

디지털화된 전문데이터베이스의 문제는 파일 크기에 따른 기억용량의 문제와 구축된 데이터베이스의 색인 및 검색의 문제로 크게 구분할 수 있다.

기억용량 및 처리속도에 관한 문제는 하드웨어적인 문제로서 현재와 같은 기술개발추세로 미래에도 계속 발전한다고 가정하면, 실제로 도서관에서 기억용량과 처리속도에 관한 것은 전문데이터베이스 구축에 있어 심각한 문제로 대두되지 않을 것이다. 오히려 색인과 검색에 대한 것이 중요한 문제로 제기될 것이다.

지금까지 전문데이터베이스의 검색은 불린식과 인접연산자 방식이 주로 사용되고 있다. 이 방식은 모두 한 문헌과 해당 문헌이 갖게 되는 색인파일과의 연결을 필요로 하며, 색인작업은 수작업 방식과 표제나 초록과 같은 곳에서 출현하는 (의미있는) 명사나 명사구를 색인어로 자동 연결하는 방식을 사용하고 있다. 이런 방식이라면 이미지 데이터 방식의 전문데이터베이스는 기존 도서관계에서 널리 사용되고 있는 서지데이터베이스를 사용하여, 적합하다고 판단된 문헌의 내용을 온라인으로 열람하는 것과 차이가 없다. 이러한 검색방식은 실제적인 전문검색이 아니라 현재 사용되고 있는 서지데이터베이스 검색방식의 보완일 뿐이다.

3.2 ASCII 파일의 문제점

도서관 소장 자료를 ASCII 파일의 형태로 데이터베이스를 구축할 경우의 문제점은 비용적인 면과 시간적인 면으로 구분할 수 있다. 앞에서 살펴본 바와 같이 이미지로 문헌의 내용을 저장하였을 경우는 각 문헌이 갖고 있는 폰트나 자간, 행간과 같이 문헌의 내용뿐만 아니라 자료로써 지니고 있는 외형적인 면까지 모두 이용자에게 전달될 수 있다. 또한, 원저자가 사용한 수식과 한자 혹은 특수문자를 전혀 어려움없이 원문대로 표현할 수 있다.

이에 비해 ASCII 형태의 텍스트 파일로 문헌의 내용을 저장할 경우, 원문의 형태와 전문데이터베이스의 수록될 전자문헌의 형태와는 달라질 수밖에 없다. 특히, KS로 지정되어 있지 않은 문자나 부호의 경우는 데이터베이스내에서 표현이 불가능하게 된다. 예를 들면, '똠'이나 '뙙'과 같은 글자는 앞으로 채택될 KS조합형 코드에서는 표현이 가능하지만, 가운데 점(·) 혹은 윗점(¨)등은 앞으로도 표현이 불가능할 것이다. 또한, 자주 사용되지 않은 수식기호나 부호, 화학식, 한자, 특수한 언어 등의 기호도 현실적으로 표현이 불가능하다.

한편, ASCII 형태로 디지털화하는 소요비용은 한 기관이 혹은 부서가 해결한다는 것은 지금의 기술력이나 혹은 예상할 수 있는 기술력으로 유추하여도 현실적으로 우리나라에서 생산되고 있는 모든 문헌의 디지털화는 불가능한 사업이 될

것이다.

3.3 저작권법의 문제점

디지털화가 어떤 방식으로 결정되어도 도서관 자료의 디지털화에 따른 궁극적인 목적은 이용자에게 모든 정보를 신속 정확하게 제공하는 것이다. 이를 위해 도서관간의 협력은 필연적인 것이 되며, 이를 위해 도서관인들은 MARC와 같은 규칙과 원칙을 개발하고 있는 것이다. 소장자료의 디지털화는 기관간의 협력이 절대적이기 때문에 디지털화자료의 공유도 또한 필연적인 것이 될 것이다. 지금까지의 도서관 내에서의 대출과 열람은 물리적 형태의 자료를 대상으로 이루어졌으며, 이는 '저작권법의 제한' 규정에도 공정사용에 해당하고 있다. 이는 각 도서관들이 해당 자료를 구입(유상 혹은 무상과 관계없이)하여 비치하고 해당 정보의 전파 속도도 그렇게 빠르지 못한 수준이기 때문에 가능한 예외 규정이라고 판단한다. 이에 비해 온라인상에 해당 자료가 올라오면(up-load) 해당 자료에 대해 각 도서관들은 구입 동기가 현격히 줄어들며, 이는 원저작자의 문헌의 판매가 급감함을 의미하며, 최종적으로 원저작자의 재산상의 피해를 초래할 것이다. 이는 저작권법에서 제시하고 있는 원저작자의 보호라는 취지에 명백하게 위배되는 것이다. 즉, 자료의 디지털화는 저작권법상에서 제시하고 있는 도서관보호규정과는 괴리감이 있으며, 별도의 조치가 필요함을 의미한다.

국내에서 약 4개회사가 상업적으로 유명 작가의 문헌(주로 소설, 수필 등)을 통신망(천리안 등)에 올려 (up-load) 이용자에게 유료로 열람하게 하는 상업적 서비스를 실시하고 있다. 이 서비스는 상업적이기 때문에 당연히 인세형식의 저작료를 지불하고 있으며, 이때 인세는 인쇄형태의 저작물가격의 30%내외에서 결정되며, 열람과 다운로드 횟수에 따라 별도의 인세를 지급하고 있다. 도서관의 서비스는 공공형태의 서비스이기 때문에 상업적인 서비스와는 차이가 있겠지만, 인쇄물의 가격의 30%는 차치하더라도 열람 및 다운로드 횟수에 따라 별도의 비용을 부담하는 것은 디지털 도서관 존립의 문제로 지적될 수 있을 것이다.

4. 해결방안

자료의 디지털화는 대상 자료의 중요도를 고려하여, 난이도가 쉬운 자료부터 추진하는 것이 바람직하다. 중요도에 대한 기준은 이용자들의 관심도와 흥미가 높은 자료부터 전산화하는 방식이다. 이를 간략히 정리하면 다음과 같다.

가) 중복 디지털화를 방지한다.

나) 기관고유의 특화된 분야를 디지털화한다.

다) 지적재산권에 저촉되지 않는 자료를 우선 디지털화한다.

라) 자료의 유형에 따른 분담 디지털화한다.

마) 디지털 기대효과가 큰 자료부터 디지털화한다.

바) 공개된 어문저작물을 우선 대상으로 디지털화한다.

4.1 중복디지털화의 방지

앞 절에서도 조사된 바와 같이 일반 문헌을 디지털화하는데 소요되는 경비를 각 단위도서관이 소유한 모든 자료를 각 기관이 부담하는 것은 현실적으로 불가능하다. 이를 위해서는 타관에서 디지털화한 자료에 대해서는 절대 중복 디지털화하지 않는 방안이 필요하다. 이를 위해서는 다음과 같은 방법이 필요하다.

① 대한민국 도서관 소장 자료 종합목록을 작성하여 각 도서관이 소장한 자료들의 重複度를 조사한다.

② 국가전자도서관 구축협의회를 발족하여 디지털화할 자료를 조사하고, 디지털화에 따른 업무내용과 양을 기관별 특성에 따라 분배 및 관리를 하도록 한다.

4.2 표준안 작성

향후 디지털화는 각 도서관과의 유기적인 협조에 따라 구축되어야 하기 때문에 도서관과의 자료를 공유할 수 있는 국가 표준이 완성되어야 한다. 현재 각 도서관은 서지통제 도구에 대해서도 약간씩의 차이를 보이고 있으나, 대체적인 서지기술은 MARC를 준용하고 있다. 전문데이터베이스는 최종적으로 ASCII형태로 구축될 경우, HTML방식이 널리 활용되고 있으나, 구조적인 문서표현은 SGML이 국제표준으로 제정되고 있기 때문에 이에 대한 규정이 필요하다. 따라서

DTD에 대한 태그를 포함한 시소러스와 메타데이터 생산과 관리에 대한 전반적인 표준화와 지침을 앞에서 제시한 국가전자도서관 구축협의회에서 범국가적으로 제시 및 개발하여야 한다.

4.3 법적 보완

도서관 자료의 디지털화는 현재 각 도서관이 소장하고 있는 자료의 소급입력에 해당한다. 한편, 현재 출판업계는 대부분 전산사식방식을 채택하고 있으며, 이는 어떤 형태라도 코드화된 데이터를 각 인쇄소 혹은 출판사가 보유하고 있다는 것을 뜻한다. 우리나라 납본에 관한 규정은 인쇄된 자료의 물리적 형태로 납본을 요구하고 있으나 바른 시일내에 물리적 형태뿐만 아니라 전산 코드화된 데이터도 함께 납본받을 수 있는 강제규정을 법으로 제정해야 한다. 이는 향후 디지털화에 따른 경비를 최소화할 수 있는 가장 확실한 방법이기 때문에 디지털도서관의 지속적인 자료보완이 가능하게 될 것이다.

현재, 저작권법에서는 도서관의 자료복제 및 이용에 대해 상당히 관대한 규정을 적용하고 있다. 이는 도서관의 공공성과 봉사성을 인정한 것으로서, 소장자료의 디지털화도 공공성과 사회봉사적 성격을 인정받을 수 있을 것이다. 따라서 보다 활발하게 디지털 도서관이 구축되기 위해서는 소장자료의 디지털화에 대한 법적인 근거와 보호조치, 소급자료의 디지털화에 따른 저작권법 적용의 유예에 대한 법적인 근거를 마련해야 할 것이다.

결 론

도서관의 봉사영역은 컴퓨터와 통신의 발달에 따라 계속해서 확대되고 있다. 또한, 급증하고 있는 정보의 양에 대한 수집방법과 처리방법에 대해 사서들의 연구영역도 계속 증가하고 있다. 이러한 상황을 모두 반영하여 이용자들의 정보욕구를 충족시켜 줄 수 있는 방안가운데 전자도서관의 구축은 당연한 것이 되었다. 전자도서관이 구축되고 발전되기 위해서는 소장 자료의 디지털화는 반드시 선행되어야 하는 작업이다. 소장 자료의 디지털화를 위해 소요되는 경비는 자료의 저장형태에 따라 많은 차이를 보이고 있다. 이미지 형태의 그래픽 파일로 문헌을 저장할 경우, 디지털화에 소요되는 경비는 약 5만원 내외가 소요된다. 이를 ASCII형태의 텍스트 파일로 저장할 경우는 최소·약 11만원이, 최대 20만원 정도가 소요된다. 이 액수는 각 도서관이 소장하고 있는 장서를 전부 디지털화한다고 가정하면, 천문학적인 금액이 되며, 디지털화에 대한 근본적인 검토가 필요한 액수이다. 이를 해결하기 위해서는 반드시 중복입력방지를 위한 국가차원의 별도 조직이 필요하다. 또한, 각 기관의 협의에 따라 디지털화에 대한 분담이 이루어지더라도 저작권적인 해결조항이 없을 경우에도 중대한 문제점에 직면하게 됨으로서 디지털화에 대한 법적 근거가 조속한 시일 내에 제정되어야 할 것이다. 법적인 해결책의 일환으로서 앞으로의 납본제도도 반드시 전자데이터의 납본도 함께 이루어져야 할 것이다.

한편, 도서관의 자료가운데에서 점차 그 종수와 양이 증가하고 있는 동화상자료나 혹은 오디오자료 대한 규정도 동시에 논의되어야 한다. 이러한 형태의 데이터는 저장과 서비스에 따른 표준형식이 정해질 경우, 일반 문헌자료의 디지털화에 비해 상대적으로 손쉽게 디지털화할 수 있을 것이다.

참고문헌

1. 남영준. 국어정보처리 기반구축사업 저작권 해결을 위한 연구. 문화체육부. 1997. 7
2. 남영준. 국어코퍼스 구축 방안. 우리말정보처리 규격 심포지움 보고서. KAIST. 1997.6.
3. 국회도서관. 국가전자도서관 구축 기본계획(시안), 1997.
4. 문화체육부. 국어 정보처리 기반구축을 위한 연구(3), 문화체육부. 1996.
5. 한국문헌정보학회. 국가디지털도서관 구축계획에 관한 연구. 최종보고서. 국립중앙도서관. 1996.