고문헌의 디지털화에 관한 연구

A Study on the Data Capturing of Oldbook

이지영, 최석두, 이화여자대학교 문헌정보학과

Jee-Young Lee, Suk-Doo Choi
Dept. of Lib. & Inf. Sci., Ewha Womans Univ.
sdchoi@mm.ewha.ac.kr

고문헌의 텍스트와 이미지를 디지털 방식으로 축적하여 제공하는 사례가 많아지고 있다. 전자화된 고문헌은 기존의 보존방안이었던 영인본이나 마이크로필름의 단점을 극복하고, 네트워크와 발달된 정보기술을 통해 매우 효율적인 정보이용수단을 제공할 수 있다. 본 연구에서는 영국 Electronic Beowulf와 한글과컴퓨터사의 삼국사기 CD 96의 사례를 분석하여 고문헌을 디지털 방식으로 축적하고자 할 때 고려해야 할 사항을 제시하고, 이렇게 축적된 정보자원이 어떻게 이용자의 정보요구를 충족시킬 수 있을지 활용방안을 논하고자 한다.

1 서 론

최근 고문헌 등 희귀자료에 디지털기술을 적용하는 국내외 사례가 증가하고 있으며 全文정보를 서비스하는 디지털 도서관의 목적과 부합하여 접근이 제한되었던 고문헌을 편리하고 쉽게 이용할 수 있는 길이 열리게 되었다. 디지털도서관은 全文의 디지털정보와 네트워크를 통한 정보서비스를 기본개념으로하는 새로운 도서관서비스의 한 형태이다. 국내의 몇 도서관에서는 고문헌의 디지털화를 디지털도서관 계획에 포함시키거나 이미 서비스를 제공하는 도서관도 나타나게 되었다.

본 연구에서는 고문헌의 보존과 이용을 활성화시킬 수 있는 방안으로서의 디지털기술의 특성을 파악하고 사례를 통하여 고문헌의 디지털화를 위해 고려해야 할 사항들을 제시한다. 또한 축적된 고문헌 정보가 어떻게 이용자의 정보요구를 충족시킬 수 있을지 활용방안을 논하고자 한다.

2 보존방안

엄밀하게 보존이란 원본 자체를 대상으로 하는 것이다. 그러나 각종 기술과 매체가 개발되면서 매체전환이나 재포매팅과 같은 방안까지 보존의 범주에 포함되는 추세이며, 특히 디지털기술의 발달로 원본의 내용과 이미지를 디지털 방식으로 축적하는 것에 대한 관심이 증가하고 있다.

자료의 보존은 크게 원본의 물리적 보존과 매체를 전환하여 원본을 대체하는 방안으로 나눌 수 있다. 본 절에서는 원본 자체를 보존하는 방안을 간단히 살피고, 원본을 대체하는 방안을 마이크로필름과 영인본의 경우, 디지털 보존의 경우로 나누어 각각의 장단점을 논한다.

2.1 원본보존방안

원본을 보존하는 방법은 문헌이 손상되는 원인에 따라 동서양의 차이가 있다.

동양에서는 주로 고문헌의 손상이 오랜 시간경과에 따른 용지의 부식과 해충의 피해로 인한 것이었고, 손상된 자료의 보존을 위해서 온도, 습도, 채광, 통풍 등 환경요인을 조절하거나 폭서, 훈증하는 방법을 사용하였다.

반면 서양에서는 산성용지의 부식에 의한 도서의 손상이 문제가 되었으며, 보존과학의 일부로 용지의 강화처리, 화학약품을 이용한 비산성화(deacidification) 처리를 실시하고 있다.

2.2 원본의 대체방안

완전한 보존방안이란 없으므로 어떤 방안을 적용하더라도 원본의 손상을 막을 수는 없다. 따라서 원본의 매체를 변환하여 대체 이용할 수 있는 수단을 강구하게 되었고 가장 보편적인 것이 영인본이나 마이크로필름으로 복제하는 방안이었다.

영인본은 원본의 페이지를 사진복제하여 현대적 장정으로 제본하여 이용하는 방식이다. 대량 복제가 가능하며 접근과 이용이 제한되는 고문헌을 도서관의 일반 장서와 함께 구성, 이용할 수 있다.

마이크로필름은 매우 안정적인 기술과 반영구적인 매체를 기반으로 한다. 도서관에서는 소장공간의 부족문제를 해결할 수 있고 인쇄본 형태의 자료에 비해 제작, 배포비용이

저렴하여 널리 이용되어 왔으며 고문헌을 비롯하여 국가문서, 학술논문, 신문기사 등 여러 정보들이 마이크로형태로 재작되어 왔다.

그러나 영인본과 마이크로이미지는 그레이스케일(gray-scale)과 디테일의 표현력이 떨어져 이미지의 정밀한 연구에는 부적절하다. 복제할 때마다 이미지 품질이 저하되며 마이크로필름의 경우 한번 복제할 때마다 이미지 품질이 10% 정도 떨어진다. 또한 이용이 불편하다는 단점이 있다.

2.3 디지털 보존방안

정보환경의 발전으로 이용자들은 자신의 연구실에서 모든 정보를 입수하고 이용할 수 있기를 기대한다. 그러나 고문헌의 원본 뿐 아니라 영인본이나 마이크로필름 형태도 이러한 이용자의 요구를 충족시키기에는 한계가 있다. 즉 정보이용에 시·공간적 제한을 가지고 있는 것이다. 디지털 기술은 다음과 같은 장점으로 인해 영인본이나 마이크로필름의 단점을 보완할 수 있는 효율적인 보존방안으로 인식되고 있다.

첫째, 디지털데이터는 네트워크를 통해 정보의 빠른 전송과 배포가 가능하며 복수이용자가 원격지에서 동시에 정보원에 접근하여 이용할 수 있다.

둘째, 디지털데이터는 마이크로필름, 종이, CD-ROM 등의 매체로 복제할 수 있으며 여러번 복제하더라도 원본이 손상되거나 내용이 변하지 않는다.

셋째, 디지털이미지는 고해상도의 컬러이미지를 지원하여 원본과 거의 동일한 이미지를 재현할 수 있고, 이미지처리 프로그램을 이용하여 손상된 원본의 이미지를 개선하거나 복구할 수 있으며, 연구목적에 따라 이미지를 조정할 수 있다. 이를 통하여 원본이나 마이크로필름 이미지를 육안으로 관찰함으로써 얻을 수 있었던 것보다 더 많은 성과를 기대할 수 있다.

넷째, 고문헌의 이미지를 디지털화하면 여러 데이터베이스에 분산저장되어 있는 고문헌 원본의 이미지와 관련 데이터를 네트워크를 통해 동시에 이용할 수 있다. 또한 권, 책, 낱장으로 분리되어 전해오던 고문헌의 완질을 전자적으로 완성할 수 있다.

그러나 장점과 함께 해결해야 할 여러 문제점이 있다.

첫째, 기술의 지속성 문제이다. 디지털기술은 빠른 기술발달에 노출되어 있고 하드웨어와 소프트웨어에 매우 의존적이다. 정보에의 접근을 결정하는 요소는 매체와 기술이며 디지털정보를 장기적으로 이용할 수 있기 위해서는 매체를 장기보존할 수 있어야 하고 기술의 지원이 뒷받침되어야 한다.

둘째, 원본과의 동질성 문제이다. 다양한 파일 포맷과 매체가 개발되면서 육안으로는 구별할 수 없는 여러 이종의 디지털 사본들이 존재하게 된다. 디지털정보의 유통과 이용이 활성화되기 위해서는 비트단위로 비교되는 디지털정보의 원본의 고유성 인증문제를 제도적, 기술적 측면에서 해결하여야 한다.

그 외에도 복잡한 저작권의 문제, 아직은 상대적으로 고가인 처리비용이나 기기비용의 문제 등도 고려되어야 한다.

3 사례

고문헌을 디지털화하는 사례는 도서관이나 기타 기관에서 소장고서를 대상으로 수행하는 경우와 특정 고문헌을 선정하여 수행하는 경우로 나눌 수 있다. 특정 자료만을 대상으로 디지털화하는 경우에는 대상자료의 특성과 이용자의 요구를 구체적으로 반영할 수 있다는 장점이 있다. 아래에서는 특정 고문헌을 대상으로 디지털화한 영국국립도서관의 Electronic Beowulf 프로젝트와 한글과컴퓨터사의 삼국사기 CD 96을 조사하여 앞으로 고문헌의 디지털화를 수행하고자 하는 기관에서 고려해야 할 사항들을 파악하고자 한다.

3.1 Electronic Beowulf

영국국립도서관은 2000년까지 이미지와 네트워크 기술을 소장 정서에 적용하여 이용자의 정보접근을 향상시킨다는 목표를 수립하고 *Initiatives for Access* 프로그램을 시작하였다. 1993년에 시작된 Electronic Beowulf 프로젝트는 이 프로그램의 일환이며 현재는 익명 ftp(영국국립도서관:othello.bl.uk, 캔터키대학:beowulf.engl.uky.edu)나 캔터키대학 웹 사이트(http://www.uky.edu/~kiernan/welcome.html)에서 이미지를 이용할 수 있다.

Beowulf는 서양문화사의 중요한 내용을 담고 있는 중세 초기의 문학작품이다. 현재 단 한부만 전하는 11세기의 필사본은 도서관 화재와 재제본등의 수리과정에서 많은 글자들을 잃거나 알아볼 수 없게 되었다. 연구자들이 필사본을 연구하기 위해 개별적으로 불빛을 비추며 연구할 경우 원본의 손상이 가속화될 것이고 번번이 영국을 방문하기도 어려운 일이어서 영국국립도서관에서는 Beowulf 필사본 이미지의 디지털데이터 구축계획을 수립하게 된 것이다.

우선 필사본 원본의 이미지와 1817년과 1824년의 대조본 이미지를 스캐닝하고 번역문이나 관련 문헌의 텍스트를 입력하였다. 기본이 되는 보존용 이미지는 디지털카메라를 이용하여 24비트 컬러, TIFF포맷으로 스캐닝하였다. 스캐닝된 이미지는 이미지 처리 프로그램으로 손상을 복구하고 원본의 순서에 따라 링크하였다.

Electronic Beowulf 프로젝트에서는 육안으로 분별할 수 없는 많은 정보를 디지털카메라를 통해 포착할 수 있었다. 의도적으로 지우거나 종이 프레임에 가려진 글자 등은 자외선 램프조명으로 스캐닝한 이미지를 통해 복구되어 자연광에서 스캐닝한 이미지와 비교할 수 있도록 하였다. 연구자들은 이미지를 조작하여 색상, 원본의 재질 등을 자세히 조사할 수 있고 희미해진 이미지를 이미지처리 프로그램으로 복구하거나 개선하여 가독성을 높일 수 있다. 즉 필사본 원본을 연구하는 것보다 더 나은 결과를 기대할 수 있게 되었다(Kiernan, 1995).

3.2 삼국사기 CD 96

한국사사료연구소는 1992년부터 삼국사기 正德本 (1512년)을 저본으로 표점, 교감하여 입력하였고 이를 한글과컴

퓨터사에서 CD-ROM으로 제작하였다.

삼국사기 CD 96에서는 전체 50권의 삼국사기를 권수, 본기·연표·지·열전, 왕조, 왕의 재위년도로 분류하여 분류번호를 부여하고, 교감·표점본 삼국사기 한자 원본과 함께 한글번역문, 정덕본 원본이미지, 한자사전을 함께 입력하고 링크하였다.

데이터 입력시 문제가 된 점은 한자형 선택과 표점, 교감에 대한 것이다. 정덕본 자체의 誤字와 異體字의 경우는 그대로 입력했을 때 검색의 효율성이 떨어지므로 수정하여 입력하였고 부수가 혼용되는 한자는 현대 한자로 바꾸어 입력하였다. 그러나 古體字의 경우 원본대로 입력하는 것을 원칙으로 하였다(한글과컴퓨터, 1996).

입력에 사용된 한자코드는 아래아한글 자체의 한자코드이며 1만5천여자로 구성되어 있다. 여기에도 포함되지 않은 한자의 경우 사용자글꼴로 처리하였다.

삼국사기의 디지털화를 통해 이전에는 접하기 어려웠던 목판본 원본의 이미지와 함께 번역본, 한문 원전, 한자사전, 주석 데이터를 데스크탑에서 이용할 수 있게 되었다.

4 데이터 입력시 고려사항

고문헌의 디지털화시에 고려해야 할 점은 다른 자료들과 큰 차이는 없다. 그러나 한자, 한글고어, 제책 형태 등 현대서와 다른 여러 특징을 포착할 수 있도록 고려하여야 한다. 고문헌은 국가적 귀중본이거나 희귀본인 경우가 많아 취급에 유의하여야 하고 여러번 스캐닝하는 것은 오히려 원본의 손상을 가속화하게 된다.

4.1 아스키 형태 입력

4.1.1 입력방안

텍스트의 입력방안은 자판 입력, OCR, 또는 이미 입력된 텍스트파일을 입수하는 것이나 고문헌의 원문은 주로 직접 입력하는 방안을 선택하고 있다. OCR 결과는 원본의 상태나 문자의 폰트, 크기 등에 영향을 받는다. 고문헌은 보존상태가 양호하더라도 활자의 형태와 크기가 다양해 현재의 OCR기술로 고문헌의 문자를 인식하기에는 무리가 있다.

우리나라에서 텍스트를 입력한 사례는 전술한 삼국사기 CD-ROM의 예가 있다. 국역 조선왕조실록을 간행했던 서울시스템에서도 현재 조선왕조실록의 원문을 입력하고 있고(이남희, 1997), 고려대장경 연구소에서는 팔만대장경 전산화사업 '21세기 팔만대장경'을 추진하고 있으며 한자 원문을 입력중이다. 1993년에 시작되어 현재 한자 원문의 90% 이상을 입력하였다(http://www.iWorld.net/~hederein/).

4.1.2 표준코드의 문제

한자 원문을 직접 입력한다 하더라도 우리나라 표준코드의 제한이라는 문제가 남아 있다. 한자표준은 KS C 5601 (1987)의 완성형 한자 4,888자이며 KS C 5657 (1991)에서 모자라는 인명용, 지명용 한자 2,856자를 추가하였다. 그러나 이 한자집합들은 일상에서 사용하는 한자조차도 수용하지 못하고 있는 실정이며 고문헌 한자의 입력 뿐 아니라 교육, 연구, 출판에 사용하기에는 턱없이 부족한 현실이다.

현재 고문헌 원문을 입력하는 기관들은 자체적으로 제작한 한자코드를 사용하고 있다. 그러나 현재와 같이 기관마다 다른 코드를 제작하여 사용할 경우 독립적으로 시스템을 운영한다면 별 문제가 없으나, 이후 데이터의 교환이나 공유에 장애가 발생하게 될 것이다.

문제를 해결하는 가장 궁극적인 방안은 표준을 개정하는 것이다. KS C 5601(1992)과 KS C 5657(1991)의 한자집합을 확장하는 문제를 검토할 수 있겠으나 완성형 한글의 표현에도 부족한 한정된 코드영역을 한자에만 할애할 수는 없는 실정이고 코드의 개정을 빠른 시일내에 기대할 수도 없다. 현재로서 표현가능한 문자의 집합이 크고 아직 널리 보급되지 않은 KS C 5700(1995)를 수정, 보완하여 사용하는 방안을 고려할 수 있다.

그러나 표준이 확장, 개정되기 전까지는 자체적으로 코드를 제정하여 사용할 수 밖에 없다. 도서관이나 컴퓨터시스템 제작사들간의 협의를 통해 공동으로 사용할 수 있는 코드를 제작하여 우선 사용한다면 이후 국가 표준코드 개정시에 참고자료로 활용할 수 있고 코드체계의 혼란을 어느 정도는 막을 수 있을 것이다.

4.1.3 텍스트 데이터의 활용방안

고문헌의 텍스트를 입력하여 얻는 잇점은 다음과 같다.

첫째, 원문접근가능성과 全文의 검색기능이라는 全文데이터베이스의 장점을 살릴 수 있으며, 이미지보다 빠른 처리와 전송이 가능하다.

둘째, 원문의 텍스트를 입력하고 이를 번역본과 링크하면 키워드나 주제어로 해당하는 원문의 위치에 쉽게 접근할 수 있게 된다.

셋째, 고문헌의 논리적 구조를 분석하고 이에 SGML을 적용하여 색인, 검색할 수 있다. 고문헌마다 차이는 있지만 어느 정도 공통의 구조를 가지고 있으며 기존의 태그집합을 응용하거나 고문헌용 태그집합을 제정하여 SGML로 태깅하면 기종과 시스템에 독립적으로 데이터의 호환이 가능해 진다. 또한 문헌의 구조에 따라 검색할 수 있다.

넷째, 입력된 고문헌의 텍스트는 컴퓨터를 이용한 통계처리나 분석이 용이해 고대의 언어습관, 한자와 한글고어의 이용통계, 문법적 특성 등을 파악할 수 있다.

다섯째, 한자나 한글고어의 용례를 모아 코퍼스 구축이나 사전 제작시 활용할 수 있다.

4.2 이미지 형태 입력

4.2.1 해상도

이미지의 품질은 해상도로 측정한다. 해상도는 이미지의 표현에 사용된 전체 픽셀의 수를 의미하는 공간해상도(spatial resoultion)와 한 픽셀의 표현에 사용되는 그레이스케일의 단계 수나 색상의 범위를 의미하는 강도해상도(intensity resolution)가 있다.

이미지 파일의 크기는 공간해상도와 강도해상도가 높아질수록 커지며 파일의 용량은 저장매체 비용, 처리와 전송에

드는 시간, 비용, 그리고 효율성에 큰 영향을 미치므로 입력 시 해상도의 결정은 이미지 데이터의 관리에 중요한 부분이다.

이미지 파일은 보존용 이미지와 서비스용 이미지로 구분하여 관리한다. 보존용 이미지는 600dpi 이상의 고해상도의 이미지를 확보한다. 보존용 이미지는 이용자가 디스플레이하여 사용할 수 있지만 연구 목적에 따라 이미지 처리용으로 사용할 수도 있다. 욱안으로는 1피트 거리에서 600dpi 이상의 이미지는 구별하지 못하지만 이미지처리 필터에서 고해상도를 요구할 경우가 있으므로 추후 이미지의 활용범위를 고려하여야 한다. 서비스용 이미지는 보존용 이미지의 해상도를 낮추고 압축하여 확보하거나 고문헌 이미지의 색상이나 디테일이 중요하지 않을 경우 처음부터 저해상도의 이미지로 확보할 수 있다.

디지털이미지의 저작권을 보호하기 위해서는 출력 장소를 제한한다든지 워터마킹, 디지털 서명 등의 기술을 이용할 수 있다.

4.2.2 스캐너

스캐너에는 평판스캐너, 드럼스캐너, 디지털카메라, 슬라이드 스캐너 등이 있으며 일반적으로 고문헌을 스캐닝할 때에는 평판스캐너나 디지털카메라를 사용한다.

스캐너를 선택할 때는 고문헌의 제책형태나 크기, 보존상태 등을 고려하여야 한다. 평판스캐너는 스캐닝할 수 있는 이미지의 크기가 제한되고 제본이 약한 고문헌의 경우 제본된 곳을 눌러 스캐닝할 경우 손상될 수 있다. 디지털카메라도 크기에 따른 제한이 있는 것은 마찬가지이나 카메라의 위치를 조정하여 어느 정도의 크기의 조절은 가능하다. 그러나 일일이 초점을 맞추어야 하는 불편이 있다.

또한 스캐너에서 발생하는 열과 광선은 고문헌의 조직을 손상시킬 우려가 있다. 스캐닝하는 장소의 온도와 습도를 조절하여 원본의 손상을 최소화 시킬 수 있는 방안을 강구하여야 한다.

4.2.3 고문헌 이미지의 활용방안

이미지에 포함된 문자의 가독성이 보장된다면 이미지 만으로도 고문헌데이터를 구축할 수 있다. 고문헌의 디지털이미지는 다음과 같이 활용된다.

첫째, 원본의 이미지는 문자코드로 입력이 불가능한 誤字, 古體字 등의 한자를 보완할 수 있다.

둘째, 원본의 체제, 레이아웃 등 외형적 요소를 유지하여 원본을 대체하는 연구대상이 된다

셋째, 텍스트와 링크하면 원하는 정보가 있는 원본의 이미지에 쉽게 접근할 수 있다. 이미지에 포함된 문자는 색인, 검색이 어렵지만 이미지의 문자를 분할하여 텍스트의 해당 문자와 링크한다면 이미지 자체를 색인하는 것과 동일한 효과를 얻을 수 있다.

넷째, 연구자의 의도대로 이미지를 조작할 수 있다. 원본이 손상되어 문자를 식별하기 어려울 경우 이미지의 대조, 밝기 등을 조정하여 가독성을 높이고 손상을 제거하여 고문헌 원본의 초기상태를 재현할 수도 있다. 또한 Electronic Beowulf의 사례와 같이 육안으로는 식별하기 어려운 정보를 파악할 수 있다.

다섯째, 활자의 분석이 가능하다. 활자는 고문헌 연구의 중요한 부분이지만 여러 고문헌을 두루 살펴며 활자를 연구하는 것은 매우 어려운 일이다. 연구자들은 여러 판본에 사용된 활자들을 한 화면에서 불러 관찰, 분석할 수 있고, 디지털이미지처리 필터를 적용하여 활자별 특성을 분석할 수도 있다.

5 결론

전자화된 고문헌의 텍스트와 이미지는 종래 원본이나 마이크로필름 등이 제공하지 못했던 다양한 활용방안을 제공한다.

네트워크를 통해 시·공간적 제약없이 고문헌과 희귀자료 원본의 이미지 뿐만 아니라 관련된 참고정보까지 함께 이용할 수 있게 되어 연구와 교육정보원으로서의 가치를 높일 수 있다. 연구자들은 고문헌의 디지털데이터를 이용하여 더 심도있은 연구를 수행할 수 있으며 일반인들도 쉽게 접하기 어려웠던 우리나라 고문헌을 원본의 모습 그대로 볼 수 있게 된다.

그러나 이러한 장점들이 있지만 입력시 한자코드의 문제, 이미지 품질의 문제, 디지털기술의 여러 문제 등 해결해야 할 과제들이 있으며 점차 문제를 해결해 나가면서 고문헌의 디지털 정보자원을 구축해야 할 것이다.

참고문헌

이남희(1997). 전산화를 통해서 본 조선왕조실록-서지학적 측면을 중심으로. *1997년도 제1회 서지학회 학술발표회 발표문*

한글과컴퓨터(1996), 삼국사기 CD 96. 서울; 同社.

Kiernan, K. S.(1995). The electronic Beowulf. *Computers in Libraries*. Feb, 1995. 14-15.

KS C 5601(1987)-정보교환용부호(한글 및 한자).

KS C 5657(1991)-정보교환용부호 확장세트.

KS C 5700(1995)-국제문자부호계(UCS) 제 1부: 구조 및 다국어평면.