ASKERIC 데이터베이스의 품질에 관한 연구

An Evaluative study on information quality of ASKERIC databases

이 명 희(Myeong-Hee Lee)*

목 차

초 록

국내에서 제작되는 데이터베이스의 품질 개선에 도움을 주고자 교육학 분야에서 널리 알려진 ERIC 데이터베이스에 대한 품질의 평가작업을 수행하였다. 선정된 데이터베이스는 웹상에서 검색이 가능한 ASKERIC이었으며, 데이터의 정확성, 일관성, 완전성, 현행성의 4가지 평가기준을 가지고 평가가 진행되었다. 정확성의 측정은 미국과 영국식 단어의 차이를 가진 글자를 가지고 수행되었는데 미국식 영어로 색인된 문헌의 검색결과가 영국식 영어로 된 문헌의 검색결과보다 훨씬 많은 것을 발견하였다. 틀리기 쉬운 10개의 단어를 가지고 철자에러를 체크해 보았을 때 상당한 양의 철자에러가 발견되었다. 일관성의 측정을 위해 색인어의 일관성을 조사하였는데 시소러스가 동의어 통제를 완전히 감당하지 못하는 것으로 드러났고 대소문자의 구별은 이루어지지 않는 것으로 나타났다. 완전성을 검증하기 위해 접근필드를 조사하였는데 ASKERIC은 상당히 다양한 접근 필드를 가지고 있었으나 매뉴얼에 나타난 접근필드와 검색창에 나타난 접근필드는 달랐으며, 검색창에 나타난 추가적인 필드를 가지고 검색하였을 때는 에러가 나타났다. 매뉴얼에서는 데이터의 갱신작업이 매월 이루어진다고 명시하였지만 현행성을 위해 실제로 검색해 보았을 때 데이터의 갱신주기는 매우 느린 것으로 나타났다. 그럼에도 불구하고 일반적으로 ERIC 데이터베이스의 품질은 대체로 양호한 것으로 나타났는데 이는 시스템의 자동 에러수정을 위한 부단한 노력과 이용자 피드백을 시스템의 품질 향상에 적극 반영하는 정책덕분인 것으로 보인다.

ABSTRACT

This study concerns information quality of the database which has been produced in the ASKERIC database. The measures used in this study were accuracy of the records, consistency, completeness and currency. Accuracy was measurred in terms of the keywords used in different ways in the US and Britain and the spelling errors in the records. Consistency was measured in terms of 'see also' and 'see reference' mechanism and character capitalization. Completeness was measured as follows : completeness of the search fields in the record and relevance of search fields. Currency was measured using the publication date. The experimental result showed that ERIC databases had some errors in terms of accuracy, consistency, completeness and currency. However, continuous striving for the automatic error checking functions and the policy of feedback from users have contributed to the improvement of the quality in ERIC databases.

* 상명대학교 문헌정보학과 전임강사
접수일자 1998년 12월 23일

1. 서 론

최근 인터넷을 비롯한 전자정보원의 중요성이 대두되면서 도서관을 비롯한 정보센터에서 과거의 인쇄매체를 어떻게 전자정보화하는가 하는 문제와 원문정보 자체를 처음부터 디지털화하는 작업에 많은 관심을 가지게 되고, 이에 따라 데이터베이스의 구축에 관한 인식과 중요성도 높아지고 있는 실정이다. 최근에 와서 우리나라에서도 지식산업의 중요성을 인지하고 공공기관이 중심이 되어 데이터베이스 구축에 박차를 가하고 있는 현상을 보이고 있다. 또한 컴퓨터와 통신기술의 발달로 데이터베이스는 온라인 전자정보원으로서 그 활용도가 높으며 신속정확한 정보검색의 기초적인 정보 인프라로서의 역할을 다하고 있다. 현재 국내에서 제작된 데이터베이스의 양은 1998년 현재 2,000여개가 넘는 것으로 나타나고 있으며 매년 30% 이상의 증가율을 보이고 있는 것으로 알려지고 있다. 또한 데이터베이스의 내용도 과거의 과학.기술 중심의 내용에서 도서, 언론, 인물, 생활, 교육, 일반법률, 인문.사회과학 전반, 규격, 산업, 재산권, 경제, 무역, 증권, 비지니스, 의학 등 학문의 전 분야에 걸쳐 다양하며, 한편 기업체 내부용으로 제작한 사내 데이터베이스까지 감안한다면 그 숫자는 엄청나다고 할수 있다. 이렇게 양적으로 증가하고 있는 국내제작 데이터베이스의 팽창은 질적인 수월성을 유지할 때 명실공히 데이터베이스 산업을 성장시킬 수 있으며, 이용자적 측면에서 고려하더라도 질적으로 우수한 데이터베이스에서 적합한 정보를 검색할 수 있는 것이다. 이러한 의미에서 데이터베이스 자체의 품질에 관한 평가는 양적팽창 못지않게 중요한 일이라 할 수 있다. 그럼에도 불구하고 지금까지의 데이터베이스에 관한 연구는 데이터베이스 그 자체에 관한 평가 연구보다는 데이터베이스 시스템 구축과 관련된 기술적인 문제, 데이터의 처리문제, 탐색서비스 및 이용자 문제에 편중되어 왔다고 해도 과언이 아니다. 또한 국내에서 많은 예산과 시간을 들여 제작된 데이터베이스의 품질을 평가한 최근의 몇몇 연구결과는 데이터베이스 구축을 위한 노력에 비해 생산된 데이터베이스의 질적인 면을 고려할 때 매우 실망스러운 결과를 낳고 있음을 보고하고 있다(김선형, 1997 : 이제환, 1997, 1998). 국내에서 제작된 데이터베이스의 제작 역사가 짧고 기술상의 노하우가 적은 현실을 감안해 볼 때 국내 제작 데이터베이스의 질을 향상시키기 위한 전제로서 우리보다 앞선 선진 외국에서 제작된 데이터베이스의 품질에 대해서 그들은 어떻게 평가하고 개선시켜왔는지를 이해하는 것은 매우 중요하다. 미국을 비롯한 선진국에서도 데이터베이스의 품질 그 자체에 관한 연구를 본격적으로 수행한 지는 약 10여년 밖에 되지 않지만 현재까지 꾸준히 연구가 진행되어 그 연구의 결과는 데이터베이스 제작자에게 피드백으로 돌아가고 데이터베이스의 품질 갱신작업에 크게 기여한 것으로 알려지고 있다.(Norton, 1981; O' Neill, 1988) 이 연구는 미국의 대표적인 서지데이터베이스 중의 하나인 ERIC 데이터베이스에 대한 품질평가가 어떻게 이루어져 왔는지를 실제 측정방법을 사용하여 평가해 보고자 한다. 이러한 측정방법의 시행은 우리에게 시사하는 바가 클 것이며, 추후 국내 제작 데이터베이스의 구축 및 평가에 기여하게 될 것으로 기대한다.

2. 이론적 배경

데이터베이스의 품질에 관한 연구는 기계가독형 정보의 발달이 급속히 이루어지기 시작한 1960년대부터 시작하였으며, 이때의 연구는 기계가독형 데이터의 철자오류 검출과 수정, 자체점검 데이터의 적용에 관한 것들이었다. 처음에는 주로 서지 데이터베이스에 대한 연구가 이루어졌으나, 점차 수치나 전문 데이터베이스 뿐 아니라 온라인 데이터베이스, CD-ROM 데이터베이스와 범용 데이터베이스 등이 제작, 보급됨에 따라 이에 대한 품질이 연구의 중요한 과제로 등장하게 되었다.(김지훈, 1996; Daniel, 1993; Dolan, 1992)

데이터베이스의 품질에 대한 주제는 전 세계적으로 많은 관심의 대상이 되어 왔는데 1990년대에 열린 National Online Meeting, Online/CD-ROM 90과 International Online Imformation Meeting 등의 회의에서는 데이터베이스의 품질을 평가하는 방법에 대하여 연구.발표하였다. SCOUG(The Southern California Online User's Group)는 1990년에 개최된 제4차 회의에서 서지 데이터베이스나 전문 데이터베이스, 명감 데이터베이스의 출력물이나 결과를 평가하는 방법을 개발하였다. 탐색자가 품질을 평가할 수 있는 요소로 11가지 기준을 제시하고 있는데, 일관성(consistency), 범위(coverage/scope), 시기적절성(timeliness), 에러율과 정확성(error rate/accuracy), 접근성/사용의 편이(accessibility/ease of use), 다른 데이터베이스와의 통합성(integration), 검색 결과(output), 도큐멘테이션(documentation), 이용자 지원 및 훈련(customer support and training), 비용대 가치(value to cost ratio) 등이 그것이다. SCOUG는 특히 질적인 평가점수를 계량화함으로써 유사한 데이터베이스들간의 비교가 가능하도록 지원하고 있으며, 데이터베이스 제작자나 시스템 벤더들로부터 자신이 제공하는 데이터베이스 평가방법에 대한 피드백을 받기도 했다.(Basch, 1990)

핀란드의 정보전문가들은 과거 데이터베이스 탐색경험을 바탕으로 하여 Finnish Society for Information Services 내에 전문가 그룹을 조직하였고, 데이터베이스의 질적평가를 위한 자료를 수집하여 연구를 수행하였다. 그들은 자신의 경험에 근거하여 데이터베이스를 평가하고 개선시키기 위한 12개 지침을 발표하였다. 그들에 의해 작성된 기준 리스트는 새로운 데이터베이스를 설계할 때 뿐 아니라 실제 운영되고 있는 데이터베이스의 평가 리스트로 사용되고 있다. 12개 지침은 1)평가그룹 설정(establish an evaluation group), 2)평가 프로젝트에 관한 자금 조성(organize the funding of the evaluation project), 3)기준이나 요구목록 개발(develop criteria or a wish list), 4)조정관 선정(find a coordinator), 5)데이터베이스 선정(choose the database), 6)패널그룹 조직(establish panel groups), 7)생산자와 호스트에게 통보(inform producers and hosts), 8)데이터베이스 테스트(test the databases), 9)결과 제시(present the result), 10)개선 및 수정물에 대한 검토(check for improvements and correction), 11)데이터베이스 모니터(monitor databases), 12)국제적 협

력(cooperate internationally) 등이다.

한편 Jacso는 CD-ROM 데이터베이스를 대상으로 하여 평가기준을 서술하고 있다. Jacso는 평가의 기준을 1)소프트웨어(software), 2)데이터웨어(dataware), 3)데이터베이스(database), 4)하드웨어(hardware)의 4개의 그룹으로 나누어 기술하였다. 특히 그 중에서 데이터웨어에 대해서는 질적인 면을 강조하고 있는데 정보의 주제적 범위와 내용적인 측면이 그들이며, 데이터베이스에 대해서는 정보에의 접근성과 데이터베이스가 문헌 및 이용자를 지원하는 정도, 전문용어 사용과 탐색비용 등을 강조하고 있다.(Jacso, 1993a, 1993b)

국내에서는 최근에 한국데이터베이스 진흥센터 등에서 품질 및 표준화에 관한 연구를 수행하였다. 이 연구센터에서 발간한 연구보고서에 의하면, 국내에서도 양적으로는 상당량의 다양한 형태의 데이터베이스가 개발 운영되고 있으며, 데이터베이스 품질기준 및 평가제도 등의 표준화 연구가 상당히 진척되었으나 이에 비해 질적인 발전은 아직까지 미흡한 상태인 것으로 지적되고 있다.(데이터베이스 진흥센터, 1996) 이 연구에서는 데이터베이스의 품질기준을 데이터베이스 그 자체의 품질기준과 데이터베이스 서비스의 품질기준으로 나누고 있는데, 전자에는 데이터의 정확성(accuracy), 완전성(completeness), 일관성(consistency), 현행성(currentness)을 거론하고 있으며, 후자에는 시스템의 검색성(searching), 사용의 용이성(ease of use), 사용자 지원성(customer support)을 언급하고 있다. 이들을 표로 나타내어 보면 다음과 같다.

이응봉(1996)은 상용의 온라인 데이터베이스 중 DIALOG 시스템을 통하여 제공되고 있는 데이터 파일을 중심으로 하여 데이터베이스의 품질에 영향을 미치는 요인을 크게 수록된 데이터 그 자체의 품질과 이용자의 검색기능의 두 가지로 구분하고 있다. 특히 그는 데이터베이스 품질 평가에 관련된 여러가지 요인을 아래와 같이 10가지로 구분하여 논하고 있다 : 1)수록 데이터의 철자 에러, 2)수록 데이터의 표기 변화, 3)수록정보의 정확성, 4)수록정보의 포괄성, 5)수록정보의 소급성, 6)수록정보의 갱신주기,

〈표 1〉 국내 데이터베이스의 품질기준 사례

구분	품질 기준	핵심판정
DB데이터 품질기준	정확성(accuracy)	데이터베이스 데이터가 실제값과 동일한가?
	완전성(completeness)	표현하고자 하는 실세계의 중요한 객체들과 품질 속성들이 담겨 있는가?
	일관성(consistency)	둘이상의 데이터가 불일치하지 않는가?
	현행성(currentness)	가장 최근의 데이터로 갱신되었는가?
DB서비스 품질기준	검색성(searching)	검색이 얼마나 신속하게 그리고 다양하게 서비스 이루어지는가?
	사용용이성(ease of use)	인터페이스를 통한 데이터베이스 접근과 산출정보 활용이 얼마나 쉽고 편리한가?
	사용자 지원성 (customer support)	documentation, help, training 등 사용자 지원 범위와 깊이가 적합한가?

7)검색키의 적절성, 8)도큐멘테이션, 9)탐색비용, 10)기타 부가기능 등으로 구분하여 이용자의 측면에서 분석한 데이터베이스의 품질에 관해 논하였다. 흔히 자주 사용되는 철자에러는 하이픈, 단락기호, 공백삽입, 입력실수 및 고유명사의 부정확한 기입 등으로서 DIALOG를 사용한 파일의 검색시에는 'EXPEND' 명령어를 사용하여 해당 용어의 철자를 확인하는 것이 필요하다고 그는 제언하였다. 수록 데이터의 표기변화는 약어의 표기 및 명칭의 변화로서 이와 같은 고유명의 표기에 대한 변화사항을 제어하기 위해서는 데이터베이스 제작기관에서 전거파일을 유지함으로써 문제점을 극복할 수 있다고 제안하였다. 수록정보 그 자체의 신빙성 내지 정확성에 관하여는 이용자 관점에서 보면 더욱 판단하기 어려운 것으로 데이터베이스 탐색자가 할 수 있는 유일한 방법으로는 가능한 한 지명도가 높은 데이터베이스를 탐색대상으로 선정하고 복수의 데이터베이스나 책자형태의 디렉토리를 참조하여 확인하기를 권하고 있다. 수록정보의 포괄성은 역시 검토하기 어려운 분야로서 데이터베이스 그 자체에 데이터의 수록기준을 명시하는 것이 좋은 것으로 나타났으며 수록정보의 갱신주기에 있어서 갱신주기가 짧으면 짧을수록 데이터베이스의 품질이 높다고 할 수 있기 때문에 정보의 갱신일을 명시하는 것이 중요한 사항이라고 주장하였다. 각 레코드 중에서 필드의 구별, 접근점의 제공, 자연어나 통제어를 사용한 색인어의 추출방법 등과 같은 검색키의 적절성은 데이터베이스의 품질을 높일 수 있는 중요한 요소가 된다. 탐색자의 입장에서 보면, 도큐멘테이션의 질적수준은 데이터베이스의 품질에 영향을 미치는 중요한 요소로서 수록주제의 범위 뿐 아니라 수록 정보원에 대한 상세한 언급을 하는 것이 바람직하다고 제시하고 있다. 또한, 상용 온라인 데이터베이스의 경우에 비용대 효과의 문제를 고려할 때 일반적으로 접속시간료, 출력료 등을 포함하는 데이터베이스 사용료가 낮은 데이터베이스가 가장 높은 평가를 받고 있지만 탐색자의 이용료에 대한 경험적 체득이 무엇보다 중요하다고 주장하였다.

유혜영(1997)은 산업기술원에서 제공하는 국내제작 온라인 데이터베이스인 BIST, DIGS, DIMD, INFO, KSMA의 품질을 정확성, 완전성, 일관성의 평가 기준에 따라 실험하였다. 정확성의 측면에서 오류가 가장 적게 발견된 것은 DIGS와 DIMD였음을 보고하였다. 일관성에 관해서는 상호참조의 유무, 대소문자와 띄어쓰기 인식여부를 실험하였는데, 일관성의 측면에서는 DIMD가 가장 적은 오류가 검색되었다. 완전성에 관해서는 데이터베이스 필드의 구성, 저자필드에서의 무명씨의 검색, 전체레코드 수와 각 필드별 레코드 총수의 비교실험을 하였는데, 완전성에서는 모든 데이터베이스가 같은 결과를 보이고 있음을 발견하였다. 전반적으로 오류가 가장 적게 발견된 것은 DIMD였고 그 다음이 DIGS의 순이었으며 나머지 3개 데이터베이스는 오류발견에 있어서 같은 결과를 보였다고 주장하였다. 전체적으로 정확성에서나 완전성에서 예상 이상의 오류가 많으며 일관성 역시 완벽하지 못하다는 것을 발견하였다.

김선형(1998)은 KORDIC에서 제작된 SATURN 데이터베이스와 UNION 데이터베이스를 중심으로 하여 국내 제작 데이터베이스의 내용적 품질평가를 수행하고 데이터베이스의

품질 향상을 위해 최우선적으로 제시할 수 있는 정책이나 대안을 제시하였다. 그는 데이터베이스 품질을 평가하기 위한 기준으로 정확성, 일관성, 완전성, 현행성의 4가지를 선정하였다. 정확성 측정의 첫 번째 방법으로서 같은 뜻의 외래어이지만 다양하게 표기됨으로써 검색결과가 달라지는 오류정도가 어느 정도인가를 측정하였다. 선정한 용어는 '디지털'과 '디지탈', '네트웍'과 '네트워크'로서 SATURN 데이터베이스와 UNION 데이터베이스에 대해 각각 비교한 결과 검색의 결과가 달라지는 것을 발견하였으며, 색인어를 브라우징한 결과 철자법의 오류 또한 심한 것을 발견하였다.

일관성의 측정에서 색인어의 일관성, 대소문자의 구별여부, 띄어쓰기의 여부에 대하여 알아본 결과, 색인어의 표기에 있어서 일관성이 없었으며 대소문자의 구별이 잘 되지 않았으나 띄어쓰기는 대체로 잘되고 있었음을 발견하였다. 완전성 측정의 방법으로는 레코드의 주제범위를 정의하고 있는가의 여부와 불완전성의 사전 알림 메세지 제공여부, 레코드내 필드값의 누락여부, 필드 구성의 완전성 여부 등이다. 검토결과, 두 데이터베이스 모두 레코드 주제범위에 대한 정의가 불투명하였으며 레코드의 필드값이 비어 있거나 필드구성에 있어서 많은 필드가 생략된 것을 발견하였다. 현행성의 측정에 있어서 수록정보의 갱신주기를 측정하였는데, UNION 데이터베이스의 경우는 월 1회, SATURN 데이터베이스의 경우는 분기별로 갱신이 이루어지는 것으로 나타났다. 그는 결론적으로 데이터베이스의 품질을 향상시키기 위해서 지속적인 품질관리 방안이 제시되어야 하며, 데이터베이스 생산자나 제작자들이 준거할 수 있는 품질평가 기준안이 공식적으로 마련되어야 한다고 제안하였다.

이제환(1997)도 국내과학기술계의 KORDIC이 구축한 KRISTAL 데이터베이스 중 UNION 데이터베이스와 SATURN 데이터베이스의 품질을 평가하였다. 먼저 데이터베이스의 품질을 분석하고 이를 이용한 정보서비스의 품질을 개선하기 위한 방안으로 KIST를 비롯한 정부출연연구소 연구원 50명을 대상으로 그룹 인터뷰의 형식으로 시스템에 대한 인지도, 시스템이 제공하는 정보의 내용, 서비스의 효용성에 대해 조사하였다. 인터뷰의 결과, 시스템에 대한 인지도는 매우 낮았으며, KRISTAL 시스템에 대한 접속비율 또한 극히 저조한 것으로 나타났다. 또한 시스템이 제공하는 정보의 내용을 포괄성 또는 적합성과 관련하여 데이터의 질, 데이터의 포괄성, 접근성, 탐색방법의 편이성, 데이터의 최신성, 비용 등을 조사하였는데 이용자의 기대에 부응하지 못하는 것으로 드러났다.

그는 또한 검색실험을 통해 레코드의 구조적 일관성, 데이터 필드의 적합성, 수록 데이터의 정확성, 레코드의 갱신주기, 레코드의 중복률을 조사하였다. 다양한 기관에서 분산입력을 하였으므로 레코드 구조의 일관성에 있어서 많은 문제점을 가지고 있었으며 데이터 필드의 적합성에 있어서도 필드가 빠진 부분이 발견되었고 수록 데이터의 표기방법 및 내용에 있어서도 누락된 부분이 많음을 발견하였다. 표기방법의 다양성, 언어처리의 비일관성이 드러났으며 레코드 갱신과 관련된 최신성 유지가 되지 않았고 입력에러 등의 수정이 이루어지지 않은 것도 발견되었다. KRISTAL 데이터베이스의 품질개선을

위한 제언으로 데이터베이스를 구축하기 전에 먼저 이용자 그룹의 정보요구에 대한 분석이 먼저 이루어져야 한다고 그는 제언하였다. 또한 분산형태로 이루어지는 단위 데이터베이스의 구축과정을 개선하여 통합 데이터베이스 구축이후에 KORDIC에 의한 데이터베이스 구축 레코드에 대한 평가작업 및 철저한 감독으로 입력 레코드의 품질을 검증하여 수정할 수 있는 정책적 고려가 필요하다고 주장하였다.

후속연구에서 이제환(1998)은 분산체계로 구축된 통합 데이터베이스인 SATURN 데이터베이스의 구축내용과 유사한 체계로 구축된 OCLC 데이터베이스의 구축상황을 살펴봄으로써 분산체계하에서 구축된 통합 데이터베이스의 품질검증을 위한 이론적 근거를 제시하고 품질개선을 위한 실질적인 방안을 도출하기 위해 연구를 수행하였다. 분산체계로 이루어진 OCLC의 UNION catalog의 내용을 분석하였다. 초창기 OCLC의 약점은 전거화일의 부재, 주제접근의 불가, 단일 서지레코드만의 허용, 서지레코드에 원문의 소재정보 수록 데이터필드 누락 등이라고 지적하였다. OCLC 회원도서관 사이에 품질의 저하문제가 크게 대두되었고 이를 해결하기 위해 인명전거와 주제전거 파일의 재작성, 회원 도서관에 의해 작성된 서지레코드의 품질 검사제도의 장치 마련, 또한 자동 에러 탐지 및 수정장치나 중복 레코드 탐지장치와 같은 기계적 시스템을 개발하여 품질 개선작업이 진행되었음을 발견하였다.

SATURN 데이터베이스의 품질을 유용성의 측정방법에 의해 점검하였는데, 자체제작 레코드의 비율, 레코드의 최신성 및 갱신주기, 레코드의 중복률, 레코드 구조의 일관성, 데이터 필드의 적합성, 수록 데이터의 완전성 등에서 아직도 미비한 점이 많다고 지적하였다. 인용문헌 분석법을 통해 본 유용성의 측정에서 SATURN 데이터베이스는 아직도 소장처를 알려주는 연구기관 사이의 원문복사 서비스를 위한 서지도구로서는 존재가치가 있는 것으로 판명되었다. 통합데이터베이스의 품질관리 방안으로서 현재의 분산체계를 중앙조정기관인 KORDIC에 의한 통합구축과정으로 대체하거나 목록 레코드 제작기관에 대한 감독 및 관리체계를 구축할 것을 제언하였다. 또한 조직구조의 개편에서 데이터베이스의 품질관리를 전담할 부서를 설치하여 운영하거나 한시적인 전담팀을 구성하여 운영하기를 제언하였다.

3. 연구방법 및 평가 기준

본 연구는 미국 교육부 산하 교육자료정보센터인 ERIC에서 제작한 데이터베이스 중에서 information and technology 분야 데이터베이스를 대상으로 하여 그 데이터베이스에 수록된 정보내용을 평가의 기준으로 삼고자 한다. 특히 현재 웹상에서 검색할 수 있는 ASKERIC을 대상으로 하여 데이터의 정확성, 일관성, 완전성, 현행성의 평가기준을 가지고 조사하려 한다.

3.1 데이터베이스의 선정

이 연구를 위해서 선정된 데이터베이스는 ERIC Clearinghouse on Information & Technology이다. 이 데이터베이스는

1966년부터 1998년까지 ERIC의 Journal Abstract(CIJE), Document Abstract(RIE)와 ERIC Digest를 초록의 형태로 제공하고 있다.

3.1.1 ASKERIC의 탐색가능 필드의 정의

1)키워드 : 문헌의 어디에서나 나타나는 단어이다.

2)저자명 : 저자사항이 공저자로 되어 있는 경우에는 처음저자의 이름만 탐색이 가능하며, 단일 저자인 경우에도 성과 이름의 순서로만 검색이 가능하다. 단 성과 이름 사이에 콤마는 필요하지 않으며 '성 이름'의 순서로 된 경우만 검색이 가능하고 '이름 성'의 순서로 된 경우에는 검색결과가 0으로 나타난다.

3)제목 : 제목에서의 단어를 말한다.

4)ERIC-NO : ERIC 엔트리에 할당된 고유한 번호로서, ED는 ERIC 문헌번호이고 EJ는 ERIC 잡지번호이다.

5)출판년도 : ERIC에 의해 색인된 연도를 말하는 것으로서, 때에 따라서는 색인된 문헌이 출판된 연도와 다를 수 있다.

6)디스크립터 : ERIC 시소러스에 나타난 주제언어로서, 특정의 문헌이나 잡지의 주제를 정의하기 위해 ERIC 시스템에 의해 인정된 용어들이다. 이는 용어 관계사이의 표준화된 형태로서 시소러스에 의해 할당된 통제언어이다. 모든 ERIC 문헌이나 잡지에는 ERIC 디스크립터가 할당되어 있다.

7)아이덴티파이어 : 아이덴티파이어는 통제되지 않은 자연언어 주제로서 표준화된 정의를 가지고 있지 않으며 대체로 기술에 관련된 새로운 용어인 경우가 많다. 즉 아직까지 디스크립터에 나타나지는 않았지만 언젠가 디스크립터로 채택될 용어들이다. 예를 들면, 'search engines'는 현재 아이덴티파이어로서 사용되지만 곧 디스크립터로 채택될 것으로 예상되고 있다. 또는 고유명사들도 아이덴티파이어로서 사용되고 있는데, 예를 들면 'Department of Education' 이나 RIE 등이 여기에 해당된다.

8)초록 : 간단한 지시적 초록의 형태를 취하고 있는데 초록에 나타난 어떠한 언어도 키워드가 될 수 있다.

9)지리적 소스 : 국명, 주명, 도시명 등을 포함한다.

10)기관명 : 문헌이 출판된 기관 또는 저자가 소속된 기관을 의미한다.

11)출판형태 : 문헌의 형태나 기관을 지칭하는 광범위한 카테고리이다.

12)언어 :영어 이외의 다른 언어.

3.1.2 ERIC 데이터베이스의 내용

ERIC 데이터베이스는 미국 교육부의 지원을 받는 국가정보시스템으로서 국가 교육도서관의 한 분야이다. 이는 교육학 연구와 현장에 관련된 문헌과 잡지의 초록 950,000건을 수록한 세계 최대 규모의 교육학 관련 자료 데이터베이스로서 상용의 CD-ROM 형태와 Web 기반의 무료 데이터베이스와 인쇄본, 마이크로피시 형태로 제공되고 있다. ERIC 데이터베이스의 내용은 인쇄본인 Resources in Education(RIE)과 Current Index to Journals in Education(CIJE)에 수록된 내용과 동일한 것이다.

이 데이터베이스는 매달 내용이 갱신되고 있으며 정보의 내용은 정확하고 시기적절한 것으로 주장되고 있다.

ERIC 클리어링하우스는 ERIC 데이터베이스를 위해 교육 관련 자료를 수집, 초록, 색인하는 곳이며 각 전문 분야별로 이용자의 정보요구 및 질문에 응답하는 곳이다. 또한 현장 연구와 프로그램 등에 대한 전문적인 출판물을 제작하는 곳이다. ERIC 데이터베이스 중 Information and Technology 주제분야는 ERIC 클리어링하우스 중 시라큐스 대학교에서 담당하는데 문헌정보학과 교육공학에 관련된 1,000건의 문헌과 2,000종의 잡지의 내용을 매년 갱신하고 있다.

ASKERIC은 이용자가 ERIC의 서비스에 만족하는 경우에는 그들에게 알려주기를 원하는 피드백을 요구하고 있으며, 또한 이용자가 그들의 생산물이나 서비스에 만족하지 않은 경우에는 어떠한 항목이 만족스럽지 않은지 구체적인 코맨트와 제안을 해 줄 것을 바라고 있다. 그러면 전자우편을 통해 건의사항이 어떻게 처리되었는지를 상세히 알려주고 계속해서 이용자가 데이터베이스의 품질을 향상시키기 위해서 조정자의 역할을 해주기를 원하고 있다.

ASKERIC이 데이터베이스 내용의 품질 향상을 위해 주력하는 또 한가지의 작업은 '수월성 추구를 위한 노력(Striving for Excellence) : 국가 교육 목표'이다. 이것은 16개 클리어링하우스의 집단적인 노력을 나타내고 있는데, 각 클리어링하우스는 교육학의 각 전문분야를 포함하고 있다. 그들은 데이터베이스의 자료를 선정하고 출판물을 출판하며 교육자들과 일반 이용자들의 요구에 응답하고 있다. 또한 클리어링하우스의 스탭은 연구보고서, 회의록 보고서, 단행본, 잡지기사 등을 규칙적으로 종합하고 그 정보를 ERIC DIGEST의 짧은 내용으로 편집한다. 82개의 ERIC DIGEST는 '국가 교육목표'에 관련되는 이슈와 프로그램과 연구의 개관을 보여준다. 이들 다이제스트는 그 목표에 상응하는 8개의 섹션으로 나누어지며 각 섹션 내에서 알파벳순으로 조직되어 자유롭게 배포되고 있다.

ERIC 데이터베이스가 지원하는 8가지 국가 교육목표는 아래와 같다:

1) Ready to Learn (취학아동들에게 즉시 학습할 수 있게 지원)
2) School Completion (고등학생들에게 학교를 졸업할 수 있도록 지원)
3) Student Achievement and Citizenship (학업성취와 시민정신 고양 지원)
4) Teacher Education and Professional Development (교사교육과 전문적인 지식과 기술 습득 지원)
5) Mathematics and Science (수학 및 과학교육 학업성취도 향상 지원)
6) Adult Literacy and Lifelong Education (성인교육과 평생교육 지원)
7) Self-, Disciplined, and Alcohol- and Drug-Free Schools (폭력으로부터의 안전과 알콜 및 마약으로부터 안전한 학교 지원)
8) Parental Participation (부모참여 교육 지원)

이 데이터베이스의 검색엔진은 PL Web으로서 필드탐색과 cross-field 탐색을 제공해 주고 있으며 적합성에 대한 랭킹을 부여하고 있다.

탐색기능으로서는 인접연산자, 와일드 카드(*형태), 절단탐색, 불리안 연산자, 동의어, 개념접근, 필드제한, exact match의 기능을 가지고 있다. 또한 만약 디스크립터나 아이덴티파이어를 모를 때에는 탐색자에게 친숙한 자연언어 키워드를 가지고 검색할 수 있다. 그 검색결과를 브라우징해 보면서 거기서 자신이 선택한 키워드와 유사한 디스크립터 또는 아이덴티파이어를 선택할 수 있는 것이다.

3.2 데이터베이스 품질의 평가기준

이 논문에서 데이터베이스의 품질을 평가하기 위해 사용된 기준은 정확성, 일관성, 완전성, 현행성의 4가지이다.

1)정확성 : 정확성이란 실세계에서 객체들이 가지고 있는 속성값과 데이터베이스내 데이터 값이 서로 일치하는 정도를 말하는 것으로서, 정확성의 척도는 데이터의 표현이 원정보의 형태에 얼마나 충실한가의 여부로 판단할 수 있다. 정확성은 편집 기법, 필드 구조의 설계, 색인작업 등과 관련이 깊은 것으로 알려져 있다. 흔히 발견되는 오류로서는 하이픈, 단락기호, 공백삽입 및 고유명사의 부정확한 기재 등을 들 수 있다. 데이터베이스의 부정확성을 체계적으로 평가해 볼 수 있는 방법은 몇 개의 색인어를 브라우징하여 검토해 보는 것이다. 검토해 보면, 색인어가 여러 가지로 잘못 기재되었음을 알 수 있다. 어떤 특정 필드의 경우에는 필드값의 범위가 논리적 관계를 벗어나 절대로 기재될 수 없는 문자가 쓰여질 수도 있다. 예를 들면, DDC 분류번호 필드에 999 이상의 번호라든지 출판년 필드에서 현재 연도보다 많은 수치의 출판년은 데이터의 오류일 가능성이 크다. 탐색자 입장에서 정확성을 높일 수 있는 방법은 가능한 한 지명도가 높은 데이터베이스를 탐색대상으로 선정하고 아울러 복수의 데이터베이스나 책자형태의 디렉토리를 참조하여 확인해야 한다. 이 연구에서 정확성의 측정방법으로는 철자법의 오류와, 같은 용어를 미국식 영어와 영국식 영어에서 다르게 사용하는 단어를 중심으로 다루려고 한다.

2)일관성 : 일관성이란 데이터베이스 내 둘 이상의 데이터가 서로 상충되지 않고 일관된 상태를 이루는 정도로 정의된다. 언어의 표현에 있어서 추상적인 개념 뿐 아니라 구체적인 개념에 있어서도 동일한 개념이 다양한 여러 용어나 용어구로 표현될 수 밖에 없는 경우가 있다. 다양한 어휘에 대한 철자, 맞춤법, 발음 등의 통제된 어휘들과 이들의 전거파일은 레코드에 대해 일관되고 신뢰할만한 정보 접근점을 제시해 준다. 색인어의 비일관적인 표기방식은 불완전한 검색의 요인이 되기 때문에 이점을 염두에 두고서 확인을 하면서 검색을 수행해야 한다. 데이터의 비일관성은 약어의 표기 및 명칭의 변화에서도 쉽게 찾아볼 수 있다. 사람의 이름도 결혼 등에 의해 성이 변경되는 경우가 발생한다. 이와 같은 경우에도 사전에 인명 디렉토리를 참조하여 해당인물의 경력을 조회할 필요가 있다. 동일한 사실의 표기에 대한 변화사항

을 체크하기 위하여 데이터베이스의 제작 기관에서는 포괄적인 전거파일을 작성하여 유지함으로써 해결이 가능할 것이다. 두번째는 상호참조와 시소러스 등으로 용어의 일관성있는 사용이 유지될 수 있다. 상호참조와 시소러스는 고유명을 변경하는 것, 동의어 아래 쓰여진 저자명이나 철자나 하이픈 등이 십입되어 있는 용어들에 대하여 통제된 용어사용을 이해시키는 기능을 한다. 이 연구에서 사용되는 일관성의 측정방법은 동일한 색인어를 다르게 표현하는 일관성 검증과 대소문자의 구별여부를 알아보는 것이다.

3) 완전성 : 완전성은 데이터베이스가 해당분야를 얼마나 폭넓고 깊이있게 망라하고 있는지에 관한 것과 레코드 각각이 개별적으로 충분한 정보를 갖고 있는지에 관한 측정이다. 완전성의 평가는 다음과 같은 질문을 통해 측정될 수 있다. 데이터베이스가 해당 분야의 중요한 객체들을 모두 포함하고 있는가, 객체에 관한 중요한 속성들을 모두 담고 있는가, 누락된 레코드나 필드값은 없는가 등이다. 이러한 문제는 해당분야의 전문가의 심층적인 분석과 판단이 요구되는 것으로 레코드의 불완전함은 그 내용이 직접 눈에 보이지 않기 때문에 정보검색에서 심각한 문제를 일으키게 된다.

구체적으로 완전성은 크게 범위, 구조, 접근성 등의 3가지 요소로 측정될 수 있다. 여기서 범위는 물리적 크기, 지리적 범위, 언어적 범위 등의 여러 측면을 포함한다. 범위에 있어서 주제가 제한되어 있거나 지극히 협의적이라고 해서 반드시 빈약한 것은 아니며 내용에 있어서도 레코드의 수량적인 측면이 꼭 기준이 되어서는 안될 것이며 이들의 범위를 데이터베이스 생산자가 분명히 정의하는 것이 필요하다. 구조에 있어서는 대부분 필드에 관한 것으로 표준 필드가 출판물의 내용전체를 완전하게 포함하는 다양한 것이어야 한다. 접근성은 데이터베이스 내에서의 레코드 내에서 부가적인 접근점을 제공하는 것이다. 시소러스를 활용하여 통제어휘를 사용함으로써 접근성을 향상시킬 수 있으며 초록 역시 서지 데이터베이스에 검색의 접근점을 증가시키고 전문 데이터베이스에서는 하나의 문단으로 전체의 주제를 모아 준다. 이 연구에서 사용되는 완전성의 측정방법은 주제범위의 완전성 또는 불완전성에 대한 언급 및 필드값의 누락여부와 접근필드의 다양성 및 적절성 여부를 알아보고자 한다.

4) 현행성 : 현행성은 데이터베이스의 데이터가 얼마나 최신의 데이터로 갱신되고 있는지를 의미한다. 시간이 지남에 따라 정확했던 데이터가 더 이상 정확하지 않는 것을 낙후된 데이터라 한다. 정보가 발생되어 데이터베이스에 수록되어 검색되어 나오기까지 시간의 간격, 즉 데이터베이스의 갱신주기는 짧으면 짧을수록 데이터베이스의 품질은 높다고 할 수 있다. 일반적으로 속보성을 중시하는 신문 등의 전문 데이터베이스는 갱신주기가 짧고, 소급탐색을 목

〈표 2〉 평가기준에 따른 측정방법

평가 기준	측정 방법
정확성	입력된 데이터의 철자 오류 동일한 용어에 대한 미국식 영어와 영국식 영어의 차이점
일관성	색인어의 일관성 검증 대소문자의 구별 여부
완전성	주제범위 소개에 대한 언급 여부 필드값의 누락 여부 접근 필드의 다양성 및 적절성 여부
현행성	수록 정보의 출판년도

적으로 하는 주제별, 분야별 서지 데이터베이스는 색인이나 초록과 같은 부가정보의 부여를 위해 다소간의 시간 지연은 불가피한 것으로 보인다. 최근에는 컴퓨터의 지원이나 분산 입력방식이 가능하기 때문에 데이터베이스의 갱신주기가 짧아지고 있는 추세이다. 이 연구에서의 현행성은 수록 정보의 출판년도를 측정하는 것이다.

4. 결과

4.1 정확성의 측정 결과

정확성 측정의 첫 번째 테스트는 미국식 영어와 영국식 영어의 차이에서 오는 철자가 다른 몇가지 글자를 비교하였다. 미국과 영국에서 각각 쓰이는 color와 colour, encyclopedia와 encyclopaedia, labor와 labour에 대하여 비교하였다. color에서 검색된 문헌은 9,211건이었으며 colour에서 검출된 문헌은 210건이었다. 또한 encyclopedia에서 검출된 문헌은 872건이었고 encyclopaedia에서 검출된 문헌은 37건으로 나타났다. 마찬가지로 labor에서 검출된 문헌은 9,211건이었고 labour에서 검출된 문헌은 1,204건으로서 이들은 각각 다른 문헌을 검출한 것으로 드러났다. 이들 두 용어를 검색한 결과를 비교해 보면, 모든 결과에서 미국식 영어를 가지고 검색한 결과가 영국식 영어를 가지고 검색한 결과보다 훨씬 많은 양이라는 것을 알 수 있다. 이것은 동일한 의미의 단어가 상이한 표기방법으로 인해 검색된 레코드의 총수가 다르다는 것을 의미한다. 만약 ERIC 시소러스가 제 기능을 다했다면, 다시 말해서 ERIC 시소러스가 통제된 디스크립터에 의해 동일한 의미를 가진 두 단어를 통제할 수 있었다면 어떤 단어를 입력하더라도 관련문헌이 모두 검색되도록 해야 하는 것이다. 그러나 불행히도 ERIC 시소러스는 그러한 통제어휘사전의 역할을 제대로 감당하지 못하였다.

두 번째는 틀리기 쉬운 10개의 단어를 가지고 철자 에러를 체크해 보았다. 이 단어들은 1991년 Jeffrey Beall에 의해 제기된 Dirty Database Test(1991)에서 사용되었던 틀리기 쉬운 글자들이었다. Beall에 의해 사용된 10개의 단어는 Febuary(February), Guata-

〈표 3〉 encyclopedia를 키워드로 사용했을 때 검색된 문헌의 양

872 documents found (25 returned) for query : (encyclopedia)	
Score	Document Title
1000	ED392454. Purchasing an Encyclopedia: 12 Points To Consider. Fifth Edition.
1000	ED376815. Stevenson, Alice. Mastery of CD-ROM Encyclopedia Skills by Elementary Students.
1000	ED270110. Encyclopedia Roundup 1985.

〈표 4〉 encyclopaedia를 키워드로 사용했을 때 검색된 문헌의 양

37 documents found (25 returned) for query : (encyclopaedia)	
Score	Document Title
902	ED044036. Meetham, A. R., Ed.; Hudson, R. A., Ed.. Encyclopaedia of Linguistics, Information and Control. . 1969
902	ED039531. . L'inglese per gli italiani; un corso programmato autodidattico realizzato dalla Britannica. Laboratorio Linguistico Individuale Anglotutor [English for Italians; A Self-Instructional Programmed Course developed by Encyclopaedia Britannica]. . 1968
875	ED233713. . Federal Judge Rules Temporary Educational Use of Videotape Copies of Copyrighted Works Illegal. Phi Delta Kappan; v64 n10 p746 Jun 1983.
875	EJ070287. Kochen, Manfred. Progress in Documentation. WISE: A World Information Synthesis and Encyclopaedia Journal of Documentation; 28, 4, 322-343, Dec 72. 1972

mala(Guatemala), misssion(mission), goverment(government), Fransisco (Francisco), grammer(grammar), recieve(receive), Wensday(Wednesday), seperate(separate), conditons(conditions) 였다. 이 연구에서 이들 단어를 사용하여 검색하였을 때 나타난 철자에러는 다음과 같다 : goverment(55건), separate(19건), recieve(9건), grammer(11건), Wensday (1건), conditons(12건), misssions(0건), Fransisco(3건), Guatamala(0건), Febuary(1건)였다. Dirty Database Test 이후, 이들 단어가 데이터베이스 제작자에 의해 수시로 체크되어 수정된다는 것을 감안하면 상당한 양의 철자 에러가 있는 것으로 나타났다.

또한 다른 종류의 철자 에러는 출판년도에서 상당히 많이 나타났다. 예를 들면, 1993년도로 출판년도를 제한한 제한검색에서 검색된 문헌의 서지사항에는 1995년 또는 1997년 출판이라는 오류가 자주 발견되었다. 이러한 필드의 내용은 예상할 수 없는 명백히 틀린 철자오류의 한 형태인 것이다.

4.2 일관성의 측정 결과

일관성의 측정방법으로 먼저 대소문자의 구

〈표 5〉 goverment를 키워드로 사용했을 때 검색된 문헌의 양

55 documents found (25 returned) for query : (goverment)	
ERIC_NO:	ED398351
TITLE:	Dropout Rates in the United States: 1994.
AUTHOR:	McMillen, Marilyn M.; Kaufman, Phillip
LANGUAGE:	English
DESCRIPTORS:	*Academic Persistence; Cohort Analysis; Dropout Characteristics; *Dropout Rate; Dropouts; *Educational Attainment; Grade 8; High School Equivalency Programs; *High School Graduates; High School Students; High Schools; Junior High School Students; Junior High Schools; *School Holding Power; Urban Schools
IDENTIFIERS:	Current Population Survey; *National Education Longitudinal Study 1988; Time Series Analysis
ABSTRACT:	This report, which is the seventh in a series, presents data from 1994 on high school dropout and retention rates and examines high school graduation and completion rates. Included is an analysis of the 1994 high school completion status and subsequent life activities of members of the National Education Longitudinal Study of 1988 cohort of eighth graders. Time series data for the period from 1972 to 1994 are also included. The best and most current national data available were used to compile the report, including the Current Population Survey (CPS) of the Bureau of the Census. Data show that dropout rates have generally decreased over the last two decades, while completion rates have increased. In 1972, data from the CPS indicated that, of young adults under age 25, 6% dropped out of school that year, over 14% were dropouts, and about 83% of young adults aged 18 to 24 had completed high school with either a regular diploma or an equivalency certificate. In 1993, only about 5% dropped out, 11% were dropouts, and over 86% completed high school. Other findings of this report show that: close to one-half million students age 15-24 left school between October 1993 and October 1994; in October 1994 there were 3.7 million 16-24-year-olds who had not completed high school and were not enrolled in school; and in general, minority students were more likely than white students to have dropped out. Dropout rates werealso higher for low income students and students in the Southern and Western regions of the country. Three appendixes contain standard error and time series tables, technical noted, and supplemental tables. (Contains 6 figures, 38 tables, 47 tables in Appendix A, 3 in Appendix B, and 12 in Appendix C.) (SLD)
GEOGRAPHIC_SOURCE:	U.S.: District of Columbia
CLEARINGHOUSE_NO:	UD031282
INSTITUTION_NAME:	MPR Associates, Berkeley, CA.
PUBLICATION_TYPE:	010
PUBLICATION_DATE:	1996
AVAILABILITY:	U.S. Goverment Printing Office, Superintendent of Documents, Mail Stop: SSOP, Washington, DC 20402-9328.
EDRS_PRICE:	EDRS Price - MF01/PC06 Plus Postage.
COMMENTS:	146p.
PAGE:	146; 2
REPORT_NO:	NCES-96-863; ISBN-0-16-048717-X
LEVEL:	1

〈표 6〉 'database', 'data base', 'db'를 각각 키워드로 사용하여 검색한 문헌의 양

11038 documents found (25 returned) for query : (database)
3653 documents found (25 returned) for query : (data base)
26347 documents found (25 returned) for query : (db)

별여부를 살펴보았다. 이 연구에서 사용된 단어는 'DATABASE'와 'database', 또한 'UNION'과 'union'이었다. 검색된 결과, 'DATABASE'와 'database'는 모두 26,347건의 문헌을 검색해 내었고 'UNION'과 'union'은 모두 10,731건의 문헌을 검색해 냄으로써 대문자로 검색된 결과나 소문자로 검색된 결과는 모두 같은 양의 검색결과를 보여주었다. 따라서 ASKERIC은 대소문자를 구별하지 않고 동일한 것으로 인식하고 있는 것으로 나타났다. 만약 이러한 결과를 고유명사나 약어 등에 결부시킨다면 검색결과는 만족할 만한 것이 못된다. 따라서 시스템의 디폴트는 대소문자의 구별에 있어서, 대문자로 검색할 때에는 대문자로만 기입된 내용을 검색해 주고 소문자로 검색할 경우에는 대문자와 소문자로 기입된 내용을 모두 검색해 주는 것이 바람직할 것으로 보인다. (Tenopir, 1993)

색인어의 일관성 검증에 있어서 'database', 'data base', 'db' 등이 비교되었는데, 'database'는 11,038건을, 'data base'는 3,653건을, 그리고 'db'는 26,347건을 검색함으로써 검색결과는 각각 다른 것 또는 부분적으로 다른 것으로 나타났다. 또한 'Tenopir (author) and database'의 검색시에는 31건의 문헌이 검출되었으나 'Tenopir(author) and data base'의 검색시에는 6건의 문헌이 검출됨으로써 이들은 명백히 다른 것으로 이해되었다. 이 사실은 이 시스템이 이들 단어를 전혀 다른 용어로 인식하고 있었으므로 어떤 표기방식을 선택하느냐에 따라 검색된 문헌의 내용이 다르다는 것을 의미한다. 만약 시소러스가 제 기능을 다했다면 이들 세 단어를 모두 통제하여 각각의 단어로 검색했더라도 같은 결과를 검출해야만 하는데 시소러스가 이들을 통제하지 못하는 것으로 나타났다. 색인어를 통제하지 못하는 비일관성은 검색의 완전성을 유지할 수 없으며 탐색자 입장에서는 미리 색인어 표기방법을 검토하고 검색을 수행해야 한다는 어려움이 있다. 그러나 단수와 복수의 단어인 경우에는 모두 복수처리해 줌으로써 일관성을 유지할 수 있었다.

저자사항에 있어서 ASKERIC은 '성 이름'만으로 검색하거나 '성'만으로 검색한 경우에는 문헌을 검출할 수 있었으나 '이름 성'의 순서로 검색했을 때에는 검출된 건수가 0건으로 나타났다. 예를 들면, 'Harter'를 가지고 검색했을 때는 36건의 문헌이, 'Harter Stephen'을 가지고 검색했을 때는 6건의 문헌이 각각 검출되었으나 'Stephen Harter'를 가지고 검색했을 때는 전혀 검색이 되지않음으로써 성과 이름의 순서에 매우 민감한 반응을 보이는 것을 알 수 있었다. 보다 고급의 기능을 가진 데이터베이스라면 '이름 성'의 순서로 된 인명이나 '성, 이

름'의 순서로 된 인명을 동일한 것으로 인식해야 할 것이다.

4.3 완전성의 측정 결과

완전성을 검증하기 위해 먼저 접근필드의 다양성 여부를 점검하였는데, 접근할 수 있는 탐색필드가 매뉴얼에 나타난 필드의 종류와 실제 검색창에서의 필드종류가 서로 달랐다. 매뉴얼에서는 키워드, 저자, 서명, ERIC 번호, 디스크립터, 아이덴티파이어, 출판형태, 출판년 등이 검색필드였으나 검색창에서 집근할 수 있는 필드로는 위의 필드에 더하여 잡지의 인용, 초록, 지리적 소스, 기관명, 접근성, 언어 등의 추가필드가 있었다. 그런데 실제 이들 추가필드를 가지고 검색창에서 검색하였을 때에는 검색 건수가 0이 되었으며 매뉴얼에 나타난 8개의 필드만이 실제검색에서 효과적으로 사용되는 접근필드로 나타났다. 다시 말하면 매뉴얼에 나타난 검색필드인 디스크립터나 아이덴티파이어, 제목, 저자명 등을 검색필드로 제한했을 때는 상당한 양의 검색결과가 나타났으나 추가 검색필드인 지리적 소스, 기관명, 언어, 잡지 인용, 초록, 접근성을 가지고 검색하였을 때에는 검색결과가 없는 것으로 드러났다. ED 378344의 문헌을 예로 들면, 저자명과 키워드인 "cooperation and technological advancement and Forrester Keith (author)"를 입력하여 검색하면 ED 378344가 검색되었으나 추가 검색필드인 지리적 소스와 기본 필드인 출판년도와 저자를 조합하여 "Vermont and 1993 and Forrester Keith(author)"로 검색하였을 때에는 0건이 검출되었다. 또한 추가 검색필드인 언어를 포함한 출판형태, 언어, 지리적 소스를 조합한 "010 and English and Vermont"로 검색하였을 때에도 0건이 검출되었다. 그러나 매뉴얼의 탐색필드에 포함되지 않은 검색창에서 제한할 수 있는 검색필드인 잡지 이용, 초록, 지리적 소스, 기관명, 접근성, 언어의 탐색필드로 제한하였을 때는 검색결과가 왜 0이 되는지 그 이유는 밝혀지지 않았다. 그러므로 추가적인 검색필드는 검색결과에서 완전성을 오히려 제한하는 요소로 나타났다.

전체적으로 필드구성에서는 많은 내용이 다양하게 구성되어 있었다. 매뉴얼에 나타난 접근필드는 키워드, 저자명, 제목, 디스크립터, 아이덴티파이어, 출판년도, 출판형태, ERIC 번호 등으로 구성되어 이용자적 측면에서 접근할 수 있는 대부분의 필드를 광범위하게 망라하고 있었다. 그러나 1차검색에서 단일주제에 대한 결과를 찾은 후에 다시 2차검색에서 '제목'이나 '저자'필드로 접근해서 검색범위의 폭을 줄여가는 단계적인 접근은 불가능하였다. 결국 전체 문헌에서 검색한 레코드의 총수와 필드별 해당 레코드의 총수를 비교할 수가 없었다. 완전성의 측면에서 볼 때, 각각의 검색단계에서 원하는 필드로 접근한 후, 다시 원하는 필드로 결과를 좁혀가는 2차검색, 3차검색이 이루어져야 한다. 이러한 관점에서 생각할 때 ASKERIC 데이터베이스는 유연성이 부족하다고 생각되었다.

4.4 현행성의 측정 결과

현행성은 데이터베이스가 얼마나 최신의 데이터로 갱신되고 있는지를 의미하는 것으로 이

〈표 7〉 ASKERIC에서 검색된 ED 378344의 내용

ERIC_NO:	ED378344
TITLE:	Trade Unions and Social Research.
AUTHOR:	Forrester, Keith, Ed.; Thorne, Colin, Ed.
LANGUAGE:	English
DESCRIPTORS:	*Cooperation; Developed Nations; Foreign Countries; Labor Economics; *Labor Education; *Labor Relations; Research Design; *Research Methodology; Research Opportunities; Research Projects; *Researchers; *Social Science Research; Technological Advancement; Unions
IDENTIFIERS:	Canada; Great Britain; Sweden
ABSTRACT:	These 15 papers originate from a conference that brought together researchers and trade unionists from a number of countries to examine the potential for a more democratic, active, and participative form of collaboration or "research as engagement." Trade Unions in the 1990s" (Waddington, Whitston); "Notes towards the Development of a Liberatory Research Project" (Gottfried); "The Unions and Research on Working Life in Sweden: What Can We Do Together?" (Danvind, Mortvik); "The Canadian Experience: Establishing and Strengthening Links between Trade Unions and Researchers" (Kumar); "Unions and Technological Change: Labor Research Strategies and Structures in the USA and Canada" (Haddad); "The Research Circle: A Way of Cooperating" (Holmstrand); "Developing Workers' Research Skills: Collaborative Research and Trade Union Education" (Somerton, Vulliamy); "A Programme for Collaboration between Trade Unions and University in Workplace Development" (Jarnegren); "Research as Engagement: Political Questions in Collective Research with the Rank and File" (Grossman); "Working with Researchers: Stress at Work" (McDonald); "Developing Research Initiatives in a Local Economy" (Sivorn, Watson); "Action Research with Homeworkers in West Yorkshire" (Tate); and "New Technology and Trade Union Activities: Experiences from Two Projects" (Winterton). A list of contributors with addresses precedes the papers. An index concludes the book. (YLB)
GEOGRAPHIC_SOURCE:	U.S.; Vermont
CLEARINGHOUSE_NO:	CE067921
PUBLICATION_TYPE:	010
PUBLICATION_DATE:	1993
AVAILABILITY:	Avebury, Ashgate Publishing Company, Old Post Road, Brookfield, VT 05036 ($59.95).
EDRS_PRICE:	Document Not Available from EDRS.
COMMENTS:	215p.
PAGE:	215
REPORT_NO:	ISBN-1-85628-354-2
LEVEL:	3

것의 측정방법 중의 하나는 접근점을 출판년도에 제한시켜 레코드를 검색해 보는 것이다. 현행성을 측정하기 위해서 1990년부터 1998년까지 수록된 문헌의 양을 측정해 보았다. 측정한 결과, 1990년에 38,997건, 1991년에 35,122건, 1992년에 35,566건, 1993년에 35,758건, 1994년에 31,943건, 1995년에 34,766건, 1996년에 31,468건, 1997년에 24,452건, 1998년에 1,908건이 검색되었다. 이 연구가 수행된 현재 시기를 1998년 11월 30일로 감안한다면 1998년에 1,908건밖에 수록되지 않았다는 것은 갱신주기가 매우 느리다는 것을 의미한다. ASKERIC의 매뉴얼에서는 수록문헌의 갱신시기를 매월 1회로 잡고 있는데, 예년의 경우와 비교해 보았을 때 1998년 11월 30일 현재에는 적어도 20,000-30,000건의 문헌이 검색되어야 하는 것으로 예상할 수 있다. 따라서 1,908건밖에 수록되지 않았다는 것은 수록문헌의 내용이 매월 갱신된다는 매뉴얼의 내용과는 상당히 다르다는 것을 알 수 있으며 실제의 갱신주기는 매우 느리다는 것을 의미한다.

5. 결론 및 제언

이 연구는 국내에서 제작되는 데이터베이스의 품질을 향상시키기 위한 기초연구 자료로서 미국에서 구축된 대표적인 서지 데이터베이스의 하나인 ERIC 데이터베이스에 대한 품질 평가가 어떻게 이루어져 왔는가 하는 것을 알아보기 위하여 수행되었다.

이 연구를 위해서 ERIC 데이터베이스를 웹상에서 검색할 수 있는 ASKERIC 데이터베이스가 선정되어 데이터의 정확성, 일관성, 완전성, 현행성의 4가지 기준에 의해 평가되었다. 정확성의 측정방법에서는 철자법의 오류와, 같은 용어가 미국과 영국에서 달리 사용되는 단어를 체크해 보았다. 일관성의 측정방법은 동일한 색인어를 다르게 표현하는 일관성 검증과 대소문자의 구별 여부를 알아 보았다. 완전성의 측정방법에서는 주제검색 접근 필드의 다양성 및 적절성 여부가 조사되었으며 현행성의 검증에서는 수록정보의 출판년도가 측정되었다.

연구결과, 정확성 측정을 위해 미국식 영어와 영국식 영어에서 차이를 가지고 있는 color와 colour, labor와 labour, encyclopedia와 encyclopaedia를 가지고 각각 검색해 보았을 때 검색결과에는 분명한 차이가 있었으며 미국식 영어를 가지고 검색한 결과가 영국식 영어를 가지고 검색한 결과보다 훨씬 많았다. 이는 시소러스가 통제언어를 모아주는 기능을 제대로 발휘하지 못한다는 것을 의미한다. 틀리기 쉬운 10개의 단어를 가지고 정확성을 측정해 보았을 때 상당한 양의 문헌이 검색되었다. 이는 또한 시스템의 자동철자 에러기능이 완벽하지 않은 것으로 판명되었다.

일관성의 측정을 위해 대소문자의 구별을 살펴보았을 때, 시스템은 대소문자의 구별을 하지 않는 것으로 나타났다. 그러나 고유명사, 약어 등의 검색을 위해 시스템은 대문자로 검색할 때에는 대문자로 색인된 문헌만을 검출해 주고 소문자로 검색할 때에는 대소문자 구별없이 검출해 내는 디폴트 기능이 바람직한 것으로 보인다. 색인어의 일관성 검증을 위해 'database', 'data base' 'db'를 가지고 검색했을 때, 각각 다른 검색결과가 나옴으로써

시스템은 이들을 전혀 다른 용어로 인식하고 있었다. 이는 역시 시소러스가 각각의 단어를 하나의 통제어휘로 변환하는 통제기능을 다하지 못하는 것으로 생각된다.

완전성의 검증을 위해 접근필드의 다양성 여부를 점검하였는데, 매뉴얼에 나타난 탐색 필드의내용과 검색창에 나타난 탐색필드의 내용은 약간 차이가 있는 것으로 나타났으며 검색창에 나타난 부가적인 접근필드는 오히려 검색결과를 저해하는 적절하지 않은 필드로 나타났다. 전체적으로 검색필드는 다양하게 구성되어 있었으나 이용자적 측면에서는 1차검색의 결과를 2차검색, 3차검색으로 제한할 수 있는 기능이 없어서 시스템의 유연성이 부족한 것으로 보였다. 현행성의 측정을 위해 출판년도별로 검색 내용을 비교해 보았을 때, 1998년도의 문헌은 소수밖에 검색되지 않아서 현행성이 떨어지는 것으로 판단되었다.

그럼에도 불구하고 ERIC 데이터베이스는 대체로 여러 가지 측면에서 양호한 것으로 드러났는데 이는 시스템의 성능을 향상시키기 위한 제작자의 부단한 노력의 결과인 것으로 생각된다. 먼저 ERIC 데이터베이스의 제작 스탭은 자동에러 체크 기능을 가진 시스템을 유지함으로써 수시로 데이터 자체의 에러를 교정하고 있었다. 또한 시스템은 확고한 목표를 가지고 운영되고 있었는데 미국 국가교육목표의 10가지 하위목표를 지원한다는 분명한 목적의식이 그들에게는 있었다. 그들은 이용자의 잠재적인 또는 현실적인 요구를 잘 이해하고 있었으며 데이터베이스의 구축시에 이러한 목표에 근접하는 데이터를 수록하기 때문에 이용자의 요구에 접근하는 정보자료를 제공할 수 있다는 것을 말한다. 또한 ERIC이 미국교육부 (Department of Education)라는 확실한 모기관을 가지고 있었기 때문에 예산이나 인력 등 여러 면에서 우수한 자원을 가지고 운영될 수 있었다. 이러한 외적 여건도 양질의 데이터베이스를 구축, 운영, 유지하는데 크게 기여하는 것으로 생각된다. 또한 ASKERIC은 이용자의 피드백을 적극 활용하여 데이터베이스의 품질개선에 크게 기여하는 것으로 나타났다. 그들은 웹상에서 이용자들에게 ERIC의 서비스에 만족하는지, 만일 만족하지 않으면 구체적으로 어떤 점이 개선되어야 하는지에 관한 제언을 적극 권고하고 있다. 이러한 개선점에 대한 이용자의 권고를 받는 즉시 이를 시행하여 이 일이 어떻게 업무에 반영되었는지를 그들에게 통보해 주며 추후에도 적극적인 모니터링의 역할을 해주기를 당부하고 있다. 이처럼 이용자의 피드백을 데이터베이스의 품질 개선에 적극 반영함으로써 이용자의 만족감이나 유용성에 크게 기여하는 것으로 보인다.

참 고 문 헌

김선형(1997). 과학기술정보 데이터베이스의 품질 평가에 관한 연구. 석사학위 논문. 서울 : 서울여자대학교 대학원.

김지훈(1996). 정보서비스의 품질평가에 관한 고찰. 도서관학논집. 제25집 : 441-474.

데이터베이스 진흥센타(1996). 데이터베이스 품질 기준 및 평가제도 표준화.

유혜영(1996). 국내 제작 데이터베이스의 평가에 관한 연구. 석사학위 논문. 서울 : 서울여자대학교 대학원.

이제환(1997). 과학기술분야 서지 데이터베이스의 품질관리 및 평가방안 : KORDIC의 KRISTAL 데이터베이스를 중심으로. 한국문헌정보학회지. 31(3) : 109-134.

이제환(1998). 분산체계로 구축된 통합 데이터베이스의 품질관리에 관한 연구. 한국문헌정보학회지. 32(3) : 179-206.

AL-Aside(1991). Ideas : The Dirty Database Test. American Libraries. Vol 22, No 3. 197.

Basch, R.(1990). Measuring the Quality for the Data : Report on the Fourth Annual SCOUG Retreat. Database Searcher. Vol 6, No 8 : 18-23.

Daniel, E(1993). Quality Control of Documents. Library Trends. 41(4) : 644-664.

Dolan, D(1992). Quality Control at System Level. Online. 16(2) : 30-35.

Jacso, P.(1993)a. Searching for Skeletons in the Database Cupboard. Part I: Errors of Omission. Database. Vol 16, No 1 : 38-49.

Jacso, P.(1993)b. Searching for Skeletons in the Database Cupboard. Part II : Errors of Commissison. Database. Vol 16, No 2 : 31-36.

Norton, N(1981). Dirty Data:A Call for Quality Control. Online. 5(1) : 40-41.

O' Neill, E and D. Vizine-Goetz(1988). Quality Control in Online databases. ARIST. 23 : 125-156.

Tenopir, C(1993). Quality of Abstracts. Online. 17(4) : 44-55.