분산체계로 구축된 통합 DB의 품질관리에 관한 연구 *

A Study on the Quality Management of a Union DB Built in a Distributed System

이 제 환(Jae-Whoan Lee)**

목 차

초 록

이 연구의 목적은 '분산체계로 구축된 통합 DB'의 품질관리를 위한 이론적 실천적 전략 및 방안을 도출해 내는데 있다. 이 목적을 위해, 먼저, DB 구축에 있어 분산체계의 의미를 정의하고, 분산체계로 구축된 대표적인 해외 DB인 OCLC Union Catalog의 품질관리 전략을 조사 분석하였다. 다음으로, 분산체계로 구축된 대표적인 국내 DB인 KORDIC의 SATURN DB의 품질을 검증하고, 품질관리 실태를 분석하였다. 끝으로, SATURN DB와 같이 분산체계의 형태로 구축되고 있는 통합 DB의 품질관리를 위한 실무 방안과 정책 조언을 제시하였다.

ABSTRACT

The purpose of this study is to discuss the theoretical and practical strategies for the quality management of a union database built in a distributed system. To the end, this study introduces the quality control methods employed for the OCLC union catalog, which has been built in a distributed system and also known as the most typical union DB. The main discussion is about the quality issue of KORDIC's SATURN DB, the most typical union DB in South Korea. The final recommendation includes the management strategies and practical guidelines for the efficient quality control of a union DB built in a distributed system.

* 이 논문은 1997년도 연구개발정보센터의 연구비에 의해 지원되었음.
** 부산대학교 문헌정보학과 조교수
접수일자 1998년 9월 1일

1. 서 론

1. 1 연구배경

'정보화'가 국가의 최대 목표로 등장하면서, 공공기관이나 민간업자 등 다양한 주체에 의해 많은 종류와 유형의 DB가 구축되고 있다. 특히, 과학기술분야는 '과학기술의 선진화를 통한 산업선진국으로의 진입'이라는 국가정책에 힘입어 공용 DB의 개발 및 구축에 있어 타분야를 선도하는 주도적 역할을 수행하고 있다. 가령, 과학기술분야의 대표적 정보유통기관인 KORDIC(연구개발정보센터)의 경우, 설립이후 6년 남짓한 짧은 기간에 25여종의 DB에 200만 건이 넘는 레코드를 구축하는 등, 국내에서 가장 대표적인 공용 데이터뱅크로 자리매김하고 있다 (연구개발정보센터, 1998).

국가차원에서 구축중인 DB의 양적 성장은 연구개발분야의 생산성 향상을 위해 매우 고무적인 일임에 틀림없다. 그러나, 우리의 시각을 DB의 품질면으로 돌릴 때, 우리는 '양적 성장 위주의 대한민국 정책의 예정된 폐해'가 이 분야에서도 예외 없이 나타나고 있음을 목도하게 된다. DB의 품질이 열악할 경우, 구축된 레코드의 양이 아무리 방대하더라도 고객들의 외면은 당연한 결과로 나타난다. 앞서 언급한 KORDIC의 경우도 예외는 아니어서, DB 구축에 들인 막대한 노력과 예산에도 불구하고 기대한 만큼 고객들에 의해 활발히 이용되지 못하고 있는 것으로 나타나고 있다 (이제환, 1997).

국가차원에서 추진한 사업의 결과물이 품질면에서 열악하여 투자효과가 기대치보다 낮다고 평가될 때, 비용효과분석을 국책 사업의 당위성과 지속성 평가를 위한 기준으로 삼고 있는 고위정책담당자들은 해당 사업에 대한 지원을 중단하거나 혹은 감소하는 결정을 종종 내려왔다. DB 구축분야 또한 이와 같은 정부의 관행으로부터 예외일 수 없다는 점을 고려할 때, 애써 구축한 DB가 폐기처분되거나 무용지물화되는 극단적인 상황이 발생하기 前에 DB의 품질을 총체적으로 검증하고 문제점을 파악하여 개선하려는 노력이 본격적으로 전개되어야 한다.

1. 2 연구목적

일년 전, 이 연구에 앞서 수행된 〈과학기술분야 서지 DB의 품질관리 및 평가방안〉에서 우리는 KORDIC이 구축하고 있는 DB중에서 그 투자규모나 구축량에 있어서 가장 핵심이라고 할 수 있는 SATURN DB의 품질문제를 논의한 바 있다. 당시 우리는 DB의 품질평가를 위한 기준을 마련하고, 그 기준에 의거하여 일정 수의 레코드를 표본으로 추출한 후 SATURN DB의 품질을 측정해 보았다. 측정 결과, SATURN DB의 품질은 기대이하로 열악한 것으로 나타났으며, 품질을 열악하게 만든 주요 원인으로, 구축전략의 문제, 전담조직의 문제, 전문인력의 문제 등이 지적되었다. 더불어, SATURN DB와 같이 분산체계하에서 구축된 통합 DB의 경우에는 DB 구축기관과 관리기관이 相異하여 품질관리가 수월하지 않다는 점을 언급한 바 있었다 (이제환, 1997).

1997년 현재, 국내에서는 분산체계의 형태로 통합 DB를 구축하고 있는 또는 구축계획을 세우고 있는 기관이 하나둘씩 늘어나고 있다. 이 기관들이 SATURN DB를 구축하면서 KORDIC이 저질렀던 품질관리와 관련된 실수를 반복하지 않도록 하려면, 계획의 입안자나 실무자들이 참조하여 활용할 수 있는 가이드라인이 반드시 필요하다. 이 연구의 궁극적인 목적은 이와 같은 가이드라인을 제공하는데 있으며, 이 연구에서는 특히 DB 선진국의 유사기관들의 사례를 선별적으로 수집 분석하는 것을 통하여 분산체계하에서 구축된 통합 DB의 품질검증을 위한 이론적 근거를 제시하고, 우리 실정에 적합한 품질 개선을 위한 보다 실질적인 방안을 도출해 내고자 한다.

1. 3 연구방법

이 연구는, 연구의 착수시점인 1997년 4월 현재, 분산체계로 구축된 유일한 국내 DB인 KORDIC의 SATURN DB를 사례로 하여 수행되었다. 연구를 위해 필요한 데이터의 수집은 다음 네 가지 방법에 의거하였다. 먼저, 분산체계하에서의 통합 DB의 구축전략 및 품질관리 방안을 다룬 국내외의 문헌을 수집하여 분석하였다. 다음으로, 분산체계로 구축 관리되고 있는 대표적인 해외 DB인 OCLC Union Catalog의 품질관리 실태를 파악하기 위하여 OCLC에 대한 방문조사를 실시하였다. 더불어, SATURN DB의 외형적 품질을 검증하기 위한 검색실험을 실시하였다. 검색실험에서는 SATURN DB의 구축에 참여한 14개의 기관별로 50개의 표본 레코드를 출력하여 완전성과 최신성 등을 검증하였다. 마지막으로, SATURN DB의 내용적 유용성을 파악하기 위하여 인용문헌분석법을 활용하였다. SATURN DB의 구축에 참여한 연구기관소속 연구원 120명을 선정하고 이들에 의해 생산된 대표적인 연구물을 연구원당 2편씩 모두 240편을 선정 수집한 후, 각 연구물에 포함된 참고문헌이 SATURN DB에서 검색되어지는 정도를 조사하였으며, 그 결과를 SCI에 대한 검색의 결과와 비교하여 분석하였다.

2. '분산체계로 구축된 통합 DB'의 定意

2. 1 DB의 구축단계

資料를 중심으로 볼 때, DB 특히 서지 DB를 구축하는 과정은 크게 세 단계의 업무로 구성된다: 그 첫째가 原문헌(original document)의 입수 업무이며, 그 둘째는 입수한 原문헌중에서 DB의 목적에 부합하는 원문헌을 선정하고, 선정된 원문헌을 분석하여 原문헌에 대한 대체물(surrogate)을 만드는 업무이며, 그 셋째는 만들어진 문헌대체물(document surrogate)을 미리 만들어 놓은 일정한 포맷을 가진 기계가독형 화일에 입력하는 업무이다. 물론 입력만으로 DB 구축이 끝나는 것은 아니다. 일단 구축된 DB는 일련의 품질 테스트를 거쳐야 비로소 독립된 DB로서 생명력을 지니게 되는데, 일반적으로, 입

력 레코드의 중복여부를 가려내고(duplication test), 입력 데이터의 표준성을 시험하며(consistency test), 입력 데이터의 완전성을 분석하고(plausibility test), 데이터 포맷의 준수여부를 조사한다(syntax test). 이와 같은 DB 품질유지를 위한 기본적인 테스트가 끝나야 비로소 이용자서비스에 들어가게 된다.

2. 2 각 단계에서 '분산체계'가 갖는 의미

그렇다면, DB의 구축단계와 구축된 DB를 이용한 정보서비스의 제공단계에서 '분산체계'란 무엇을 의미하는지에 대해 논의해 보자. 복잡한 컴퓨터 관련 모델이나 네트워크 관련 모델을 여기에서 인용하고 싶은 생각은 없다. 단지, 앞서 언급한 DB의 구축단계를 참조하면서, 각 단계에서 업무의 분산처리(distributed processing)는 어떻게 발생할 수 있는지 그 유형을 살펴보도록 한다.

① 분산체계의 첫 번째 유형은, 원문헌의 입수단계에서 나타난다. 어떤 한 기관에서 기관의 목적을 위해 DB를 구축할 경우, 그 기관에는 대부분 원문헌의 생산 혹은 출판여부를 파악하기 위해서 정기적이고 세밀한 문헌서베이를 전담하는 부서 혹은 인력이 있기 마련이다. 그러나, 여러 기관이 협조체제를 구성하여(networks or cooperatives) 공동의 목적을 달성하기 위한 도구로서 (대부분, 자료의 공유나 자료의 공동가공을 위한 목적에서) 하나의 통합 DB를 구축하고자 할 경우에, 상황은 다양하게 전개된다. 협조체제의 중앙에 원문헌의 파악과 입수를 전담하는 기관이나 부서가 있는 중앙입수형이 한 편에 있고, 멤버 혹은 가입기관들이 개별적으로 그들의 정보요구에 적합한 원문헌을 서베이하여 직접 입수하는 분산입수형이 다른 한 편에 있다. 일반적으로 前者가 後者보다는, 규모의 경제 측면에서 볼 때, 입수비용이 절약되고 서지통정이 효율적으로 이루어짐으로써 동일 문헌에 대한 중복투자가 감소되는 등 여러 장점이 있으나, 원문헌의 보관 및 관리에 대한 책임문제와 더불어 최종이용자에 의한 원문헌에 대한 접근과 이용이 시공간적으로 지연되는 등 단점도 만만치 않아, 어느 방법이 일반적으로 우월하다는 식의 논평은 여기서는 유보한다.

② 분산체계의 두 번째 유형은, 원문헌에 대한 가공 단계에서 발생한다. 여기에는, DB에 수록할 가치가 있다고 판단된 원문헌를 대상으로 하여, 중앙의 장소에 분석 가공전문가들이 모여 원문헌를 가공분석하는 중앙처리형과 주제별로 가입기관으로 원문헌를 분배한 후 (혹은 독자적으로 입수하게 한 후) 각 기관에서 원문헌를 가공분석하는 분산처리형이 있다. 가공단계에서의 중앙처리형과 분산처리형은 원문헌의 입수방법과 밀접한 연관을 갖는다. 가령, 중앙입수형의 경우는 대부분 중앙처리형태를 띠고, 분산입수형의 경우는 절대적으로 분산처리형태를 띠게된다. 간혹, 규모의 경제 논리를 따라 입수는 중앙집중형으로 하되 가공처리는 분산형으로 하는 경우도 있으나 흔한 사례는 아니다. 분산처리형에 비해 중앙처리형이 갖는 장점으로는, 무엇보다도 원문헌의 가공단계에서 개별 서지레코드에 대한 품질검증과 개선이 단계별로 체계적으로 이루어짐으로써

전체적인 DB의 품질관리가 수월하다는 점을 들 수 있다. 이 경우는 물론 중앙기관에 철저한 DB의 품질관리를 위한 절차와 규범 그리고 품질검증 및 개선을 위한 인적 자원과 기술적 지원이 마련되어 있음을 전제로 한다. 반면, 분산처리의 경우 철저히 통제 관리되지 않을 시, 가공되는 데이터의 양과 질 모두에 있어 종종 문제가 발생하여 DB 전체의 품질관리에 어려움을 겪게 되는 경우를 국내외의 여러 사례는 보여주고 있다.[1)]

③ 분산체계의 세 번째 유형은, 가공처리의 결과로 생산된 서지데이터를 중앙의 통합 DB에 입력하는 단계에서 발생한다. 중앙처리형의 경우 대개 가공처리단계에서 바로 온라인으로 입력이 이루어지나, 분산처리형의 경우는 온라인 입력과 오프라인 입력이 병행되는 경우가 일반적이다. 중앙시스템이 품질검증 및 수정을 위한 절차와 장치를 구비해 놓고 있는 경우에는, 분산처리형에서도 가공작업의 완료와 동시에 통합 DB로의 입력이 바로 이루어지기도 하나, 대부분의 경우는 가공처리된 자료를 일정량 모아 두었다가 오프라인 벳치(offline batch) 형태로 일괄 입력하는 경우가 많다. 그러나 이렇게 일정량을 모아 두었다가 일괄 입력할 경우, 입력 자료에 대한 품질검증이 제때에 실시되기가 어렵고 더불어 통합 DB의 최신성 유지가 어렵게 되어, 궁극적으로는, 통합 DB에 대한 이용자의 신뢰도 및 이용도를 저하시키는 요인이 되기도 한다.

2. 3 '분산체계로 구축된 통합 DB'의 意味

이 연구에서 '분산체계'는 앞서 언급한 DB의 구축과정에서 나타나는 세 단계의 작업형태가 모두 분산되어 이루어지는 경우에 한정한다. 즉, 하나의 DB를 구축하기 위하여 여러 기관들이 참여하되 원문헌의 입수과정/입수한 자료의 가공처리과정/가공처리된 자료의 입력과정이 모두 분산되어 이루어지고, 그 결과로 하나의 통합(union) DB가 구축될 때, 우리는 이 DB를 '분산체계로 구축된 통합 DB'로 정의한다. 물론 원문헌의 입수는 중앙기관에 의해 이루어지고, 자료의 가공처리 및 입력과정이 분산형태로 이루어지는 사례도 종종 있으나, 이 연구에서는 이에 대한 논의를 배제한다.

이렇게 볼 때, 국내에서는 KORDIC의 SATURN DB가 전형적인 '분산체계로 구축된 통합 DB'로 분류된다. SATURN DB의 관리 및 서비스기관으로 KORDIC이 존재하지만, DB의 구축과정에서 KORDIC의 역할은 극히 미미한 것이 현실이다. 현단계에서 KORDIC의 역할이라면, DB의 저장과 검색에 필요한 하드웨어와 소프트웨어 그리고 통신망의 제공과, 상위 부처인 과학기술부에서 이 작업을 위해 할당된 예산을 적절히 분배하여 DB 구축기관 (원자력연구소와 같은 정부출연연구소의 자료실)에 나누어주는 일에 국한된다. 물론, SATURN DB에 대한 관리와 SATURN DB를 이용한 서비스의 제공(가령, 원문제공서비스)이 KORDIC의 주요한 역할이기는 하지만, SATURN DB의 구

1) 후에 논의될 OCLC의 Union Catalog나 KORDIC의 SATURN DB는 전형적인 例가 된다.

축과정에서의 역할은 극히 한정되어 있음을 의미한다.

KORDIC의 SATURN DB처럼 주제영역별로 다양한 구축기관이 참여하면서 그 구축과정이 철저하게 분산형태를 띠고있는 事例를 찾는 것은 국내외를 막론하고 결코 쉬운 일이 아니다. 비교분석을 위한 목적에서 국내외의 여러 기관들의 정보시스템들, 특히, DB에 대한 면밀한 조사 결과, 미국 Online Computer Library Center(OCLC)의 Union Catalog가 여러 면에서 가장 근접해 있다는 결론에 도달했다(Amstrong, 1994; Mifflin, 1991; O'neil, 1988; Wang, 1995). 이에, SATURN DB에 대한 본격적인 논의에 앞서, 다음 파트에서는 OCLC의 Union Catalog의 구축과정과 품질관리를 위한 전략 등을 상세하게 소개하고자 한다. 이를 통해, 이 연구에서 논의하고자 하는 SATURN DB의 품질문제가 어디서부터 비롯되었는지를 추론하고, 어떻게 개선되어야 하는지를 이론적으로 밝혀보고자 한다.

3. 분산체계로 구축된 해외 DB의 사례

3. 1 OCLC Union Catalog의 현황

Online shared cataloging(온라인 공동편목)의 시대가 막을 올린 1960년대. 이 시기의 미국은 戰後의 흥청거림도 막을 내리고 국가재정의 긴축으로 인한 공공부문의 대폭적인 사업감축이 예고되던 시절이었다. 그 시절에, 몇몇 대학도서관들이 모여서 공동으로 목록을 작성하고 그 결과를 공유할 수 있는 기계적 메커니즘을 개발하고자 하는 노력을 시작하였다. 이들의 목적은 간단했다. 자료를 가공처리하는데 드는 비용을 절감하고 더불어 소장자료를 공동으로 활용함으로써, 재정의 긴축으로 야기된 도서관운영의 어려움을 극복하고자 하는 것이었다 (Martin, 1986). 자료의 공동처리와 정보자원의 공유 - 이것이 당시 그들의 목적이었고 30년후인 오늘날에는 전세계 정보관리기관들의 목적이 되고 있다.

1965년에 공유를 위한 최초의 노력이 New England 지방에서 6개 대학이 뭉치면서 일어났으며(New England Library Information Network), 이 노력은 Ohio로 이어져 54개의 Ohio소재 대학이 뭉쳐 OCLC(당시는 Ohio College Library Center)가 만들어졌다. OCLC의 등장은 본격적인 online shared cataloging network (온라인 공동편목망)의 시작을 의미했다. 1970년에 최초의 서비스가 시작되었을 때, OCLC는 카드목록을 공동으로 생산해 내기 위한 전형적인 batch processing system이었다. 그러나 일년후인 1971년, OCLC는 멤버도서관들이 OCLC system을 이용하여 OCLC의 중앙 데이터베이스(OCLC union catalog)에 자유로이 서지레코드를 입력하고 수정하고 검색하는 것을 허락하는 online shared cataloging system으로 변화하였다 (Davis, 1987 & 1989).

출발부터 OCLC는 철저히 분산체계의 형태를 띠고 있었다. 원문헌의 입수는 당연히 멤버도서관들에 의해서 분산적으로 이루어졌고, 원문헌의 가공작업도 멤버도서관 차원

에서 독립적이고 분산적으로 이루어졌다. 입수된 원문헌은 멤버도서관의 서가에 분산되어 보관되었다. 독립건물도 없이 오하이오주립대학의 한 공간에 OCLC system이 있었고 소수의 관리인이 그 시스템의 작동을 책임지었다. 당시 OCLC system은, 하드웨어와 OCLC Union catalog와 데이터의 처리를 위한 소프트웨어로 구성되어 있었다. OCLC Union catalog에 가공처리된 서지레코드를 입력하는 행위를 통해 실질적으로 DB를 구축하는 일은 멤버도서관들의 몫이었다 (Saylor, 1986).

OCLC Union Catalog가 최초로 설계되었을 때, 거기에는 후일 이 DB의 품질에 중요한 영향을 끼칠 몇 가지 특징적인 면이 있었다. 먼저, 오프라인이 아닌 온라인 데이터베이스라는 DB의 특성은 OCLC Union catalog의 설계자들로 하여금 기존의 카드목록에서는 반드시 필요로 했던 main entry의 개념을 불필요하게 느끼게 만들었고,[2] 이 잘못된 느낌은 전거화일(authority files)의 무용론으로 이어졌다. 그 결과로 OCLC Union catalog는 전거화일이 없이 설계되었다.[3] 두번째 특징은 주제탐색을 위한 기능이 배제된 채 설계되었다는 점이다. OCLC Union catalog는 주제에 의한 접근을 할 수 없도록 만들어졌으며, 이는 오랫동안 OCLC system의 최대 약점으로 남게 되었다. 세 번째 특징은 OCLC Union catalog는 하나의 문헌에 대해서 오직 하나의 서지레코드만을 고객의 단말기에 디스플레이 되도록 허용하는 구조를 지니고 있었다. 따라서, 멤버도서관들이 독자적으로 작성한 다양한 내용의 서지레코드를 온라인상에서 비교할 수가 없었다. 마지막으로, 서지레코드에는 원문헌의 소재정보를 수록하는 데이터필드가 존재하지 않았다. 소재정보가 누락된 상황에서 초기의 OCLC Union catalog는 ILL을 위한 서지도구의 역할에 근본적인 의문을 던지고 있었다 (O'neil, 1988).

그러나 초기의 약점에도 불구하고, 이후 통합 DB로서의 OCLC Union catalog의 성장은 눈부신 것이었다. 출발당시 OCLC Union catalog의 구축에 참여했던 54개의 멤버도서관은 1993년에는 4,867개의 멤버도서관으로 늘었으며, 편목작업을 수행하지는 않지만 OCLC에 가입하여 각종 서비스를 이용하고 있는 전세계의 도서관의 수는 1997년 말 현재 17,000여 개에 이르고 있다. OCLC Union catalog를 구성하는 서지레코드의 수 또한 기하급수적으로 늘어 1997년 말 현재 약 3,000만 레코드에 이르고 있으며, 매년 100만개의 서지레코드가 계속 추가되고 있는 실정이다. OCLC Union catalog를 이용한 ILL의 요구 또한 기하급수적으로 늘어 1997년 12월 현재 5000만건 이상의 상호대차가 이루어지고 있는 것으로 나타난다 (OCLC, 1997).

이처럼, 출발이후 불과 30년만에 OCLC는 명실공히 세계최대의 문헌정보망으로 우뚝 섰고, 그 핵심이 되는 OCLC Union catalog는 날로 그 규모가 커져가고 있다. 지난 30여년

2) Main entry의 필요성을 둘러싼 학자들의 논쟁은 오래 동안 이어졌다. 관련 문헌들은 1970년대 Library literature나 LISA 같은 색인/초록집에서 넘칠 정도로 발견된다.
3) 물론 지금의 OCLC database는 authority files를 갖추고 있다.

동안, 초기의 약점들이 대폭 치유되어 시스템의 기능도 개선되고, 경영체제도 주식회사로 탈바꿈하면서 경영의 효율화가 진척되고, 인력도 초기의 문헌정보학자중심에서 컴퓨터전문가와 경영전문가들의 계속적인 유입으로 적절한 인력상의 균형이 이루어졌으며, 이를 바탕으로한 다양한 정보서비스가 계속 개발되어 전세계적인 이용자커뮤니티의 기대에 부응하고 있다.

3. 2 Union Catalog의 품질관리

OCLC의 성장에 있어 OCLC Union catalog는 늘 중심에 있었다. 최대 문헌정보망으로서의 OCLC의 위상을 유지하기 위해서는 OCLC Union catalog의 양과 질은 최상의 상태를 유지할 필요가 있었다. 그러나, 1970년대 중반으로 접어들면서 Union catalog의 품질에 이상 징후들이 나타나기 시작하였다. 이 징후들은 분산체계로 구축되던 Union catalog에 서지레코드를 입력하는 멤버도서관의 수가 급격히 늘면서 (가령, 1967년 54개에서 1975년 559개로 다시 1980년에 2,429개로) 불거지기 시작하였다. 적절한 대책이 마련되지 못한 상태에서, 커져가는 OCLC의 덩치만큼이나 Union catalog의 양적 규모는 커져갔다. 결국 1970년대 후반, OCLC는 '경영혁신'이라는 방법 (즉, 주식회사로 조직형태를 개편)을 통해 기관경영전략은 수정하였으나, 급격히 증가되는 Union catalog의 서지레코드에 대한 품질검증 및 개선방안은 여전히 미흡한 상태로 남아있었다. 자연스럽게, 수준이하의 서지레코드들이 OCLC Union catalog에 늘어나기 시작하였고, 곧, Union catalog의 품질을 개탄하는 소리가 도서관계는 물론 OCLC의 창립자(F. Kilgour)의 입에서도 흘러나왔다 (Martin, 1986). 이후 1980년대 중반에 이르기까지, OCLC Union catalog의 품질문제는 미국 도서관계의 뜨거운 논쟁의 불씨로 남아있었다.

OCLC가 Union catalog의 품질관리를 위해서 채택한 기본 방법은 'master record' 제도였다. 분산체계하의 통합 DB인 관계로 동일한 문헌에 대한 여러 개의 서지레코드가 만들어질 수 있는 상황에서, OCLC는 초기에 입력된 비교적 완전한 형태의 서지레코드를 선정하여 해당 문헌에 대한 master record로 삼았으며, 이 master record만이 Union catalog의 이용자가 온라인 상에서 접할 수 있는 해당 문헌에 대한 유일한 서지레코드가 되었다. 따라서 master record의 품질은 곧 Union catalog의 품질을 의미하였다. 그러나 master record가 될 수 있는 서지레코드의 기준이 엄격히 정해져 있는 것이 아닌 상태에서, master record도 데이터의 정확성이나 완전성에서 문제가 되는 경우가 종종 있었다. 단 하나의 예외는 Library of Congress (美의회도서관)에서 작성한 서지레코드를 master record로 삼는 경우였는데, 이 정책의 白眉는 이미 master record로 선정된 서지레코드라고 할지라도 동일 문헌에 대한 LC의 서지레코드가 입력되면 그 자리를 LC 레코드에 양보하도록 하는데 있었다. 이 정책으로 Union catalog의 품질은 어느 정도 유지될 수 있었지만, 문제는 Union catalog에서 LC가 제작한 서지레코드가 차지하는 비중이 다수

가 아니라는데 있었다. 다양한 멤버도서관들은 다양한 수준의 서지레코드를 입력하였고 그 다양한 서지레코드들이 master record의 위치를 점하면서 OCLC Union catalog의 품질은 점차 관리하기가 어려울 정도로 일관성을 잃어갔다.

OCLC Union catalog의 품질저하를 유발한 보다 근원적인 원인은 앞서 언급한대로 '전거화일(authority files)'의 부재에 있었다. 멤버도서관에 의해서 입력하는 서지레코드에 포함되는 여러 데이터요소들, 특히 서명이나 저자 혹은 주제명 등은 형태의 일관성이 무엇보다도 중요한데, 전거화일의 부재는 서지레코드의 작성자들이 표목이 되는 데이터요소의 표준형태를 선택하는 것을 도와줄 도구가 없다는 것을 의미했다. 그 결과는 동일한 저자나 주제라 할지라도 서지레코드의 작성자에 따라 다양하게 표현되는 등, 일치성(confirmity)의 결여로 나타났다. OCLC가 이 문제의 심각성을 깨닫고 뒤늦게 name authority file과 series authority file을 만든 것이 (그것도 subject authority file은 손도 못대고) 1989년의 일이었으니, 표목의 일치성이라는 측면만을 놓고 보아도 Union catalog의 품질이 얼마나 열악했으리라 하는 것은 능히 짐작되고도 남는다.

이 근본 원인외에도, 멤버도서관의 서지레코드 작성자들에 대한 체계적인 교육과 훈련의 결여, 이들이 서지레코드를 작성하고 작성된 서지레코드를 Union catalog에 입력할 때 참고할 수 있는 각종 절차 및 법칙을 수록한 자료(documentation)의 미비, 그리고 입력된 서지레코드의 품질을 재검사하는 제도적 장치의 미비 등은, OCLC의 Union catalog의 품질을 열악하게 만든 주범이었다 (Barnett, 1993). 이렇듯 품질관리를 위한 환경과 준비가 열악한 상황에서, 앞서도 언급하였듯이, Union catalog의 규모는 급속도로 커져갔다.[4] 품질이 보장되지 못한 상황에서의 양적 팽창과 그로 인한 도서관계의 거친 비난에 OCLC 경영층의 고민은 커져갔다.

그러나, 1980년대 중반에 들어서야 비로소 Union catalog의 품질개선을 위한 전략과 구체적인 방법들이 마련되기 시작하였다. Union catalog의 품질을 검증하고 개선책을 마련하기 위한 여러 프로젝트가 수행되었고, 그 결과는 즉각 품질관리를 위한 정책에 반영되었다. OCLC가 Union catalog의 품질개선을 위해서 취한 전략은 크게 두 가지로 대별되었다. 그 하나는 제도적 장치의 마련이었고, 다른 하나는 기계적 장치의 개발이었다. 먼저, 〈품질강화 프로그램〉과 같은 제도를 마련하여 레코드의 완전성(completeness)을 도모하고자 하는 노력이 전개되었고, 다음, 자동에러탐지 및 수정 장치나 중복레코드 탐지장치와 같은 기계적 시스템을 개발하여 입력 데이터의 정확성(accuracy)을 높이고 더불어 중복 레코드의 비율을 줄이려는 노력이 전개되었다 (Amstrong, 1994; Barnett, 1993).

사실, 간단한 철자나 코드상의 에러를 자동적으로 탐지해 내거나 중복레코드를 탐지

4) 1974년 100만 레코드를 돌파하더니 1983년에는 1,000만 레코드를 넘어섰고, 향후 1년마다 거의 100만 레코드이상이 Union catalog에 첨가되는 상황이 지속되고 있다.

해내는 기계적 장치의 개발은 1970년대 후반 이미 시작되고 있었다. 단지 그 성능이 미흡하고, 레코드의 품질문제가 단순한 철자상의 실수(typographical errors)에 국한된 것만이 아니었기 때문에 그 실효성이 적었을 뿐이었다. 다시 말해, 데이터필드에 데이터요소가 미기입되거나 잘못 기입되는 등의 문제는 기계적 해결보다는 사람에 의한 수정을 필요로 하는 문제일 경우가 많았고, 기계적 해결이 가능하다고 하더라도 몇몇 소프트웨어를 개발한다고 해서 해결되는 문제는 아니었다. 가령, Online authority files이 없는 상태에서 주제표목(subject headings)의 옳고 그름을 기계적으로 판단하여 수정을 시도한다는 것은 어차피 불가능한 일이었던 것이다.[5)]

이 문제의 근본적인 해결을 위해 1987년에 OCLC는 LC의 Subject Headings List를 OCLC system에 올려 online 상태로 멤버도서관들이 이용하게끔 하였다. 이 조치는 멤버도서관들에 의한 서지레코드의 작성단계에서 아예 품질저하를 유발할 요인을 제거하자는 조치의 일환이자, 기존에 Union catalog에 수록된 서지레코드의 시대에 뒤떨어진 주제표목(subject headings)을 바로 잡자는 노력의 일환이었다. 더불어 OCLC는 Union catalog에 수록된 저자표목(name headings)들을 LC의 Name Authority file에 두 번 이상씩 대조하는 작업과정을 통하여 1987년 한해에만 약 530만 레코드를 수정하였다 (Davis, 1989).

기계적인 장치의 개발을 통한 품질관리에 더하여, 제도적 장치의 마련도 활발히 진행되었다. 대표적인 것이 1984년부터 시행된 the Project Enhance이었다. 이 프로젝트의 핵심은 OCLC의 멤버도서관중에서 서지레코드 제작능력이 높은 수준으로 판단되는 20개의 도서관을 선정하여, 이 도서관들로 하여금 다른 도서관이 입력한 품질이 낮은 서지레코드를 (master record라 할지라도) 수정하도록 권한을 부여한데 있었다. 이 'super libraries'의 숫자는 계속 늘어 1990년 현재 약 90개의 도서관으로 늘어나, 이들이 일년 평균 100,000개에 가까운 '문제가 있는' 서지레코드를 수정하고 있다 (OCLC, 1991).

Super libraries에 의한 품질검증 및 개선은 제도적 장치의 1차 단계로 볼 수 있다. 품질에 대한 또 한번의 검증과 개선은 OCLC의 내부조직인 Online Data Quality Control Section(ODQCS)에 속한 직원에 의해서 이루어졌다. 원래, Union catalog의 품질유지를 위한 전통적인 방법 중에 하나는 이용자로부터 Change Request Forms를 우송 받아 그 내용에 기초하여 ODQCS의 직원이 문제가 있는 것으로 판단된 master record를 수정 또는 삭제하는 방식이었다.[6)]

이 제도는 여러 장점이 있었으나 이용자

5) 참고로, OCLC에 의해서 행해진 한 이용자 서베이는, 1980년대 중반 당시, OCLC의 Online Union Catalog (OLUC)에서 발생하는 품질문제의 유형을 그 빈도에 따라 순위를 매겨놓고 있는데, 그에 의하면: Duplicate records (1위), subject headings errors (2위), name headings errors (3위), typographical errors (4위), classification errors (5위), incorrect applications of standard cataloging rules (6위), 그리고, MARC tagging (7위)로 나타났다 (Davis, 1989).

6) 이 방식을 사용하여 ODQCS staff는 1년 평균 약 60,000건의 레코드를 수정하였는데 이 양은 그들이 1년에 수정하는 총 125,000건의 절반에 약간 못 미치는 수치이다.

로부터의 협조가 해가 갈수록 줄어든다는데 OCLC의 고민이 있었다. 문제는 Change Request Forms를 이용하는 보고방식이 너무 시간소비적이고 또한 준비해야할 서류도 복잡하여 이용자들이 꺼려하는데 있었다. 이 문제는 OCLC가 online error reporting system을 도입함으로써 해결되었다. 이후 멤버도서관의 이용자들에 의한 에러보고의 숫자는 지속적으로 늘어나, 이 처리를 위해 ODQCS의 조직이 지속적으로 확대되는 상황을 맞고 있다 (OCLC, 1994).

이외에도, OCLC는 우수 멤버도서관과 우수 서지레코드 제작자 (사서)에 대한 포상제도와 같은 Union catalog의 품질을 개선하기 위한 각종 인센티브 제도를 도입하였는데, 그 결과, 미도서관계로부터 'dirty data'의 집합소라는 오명을 들었던 과거의 불명예로부터 어느 정도 벗어나고 있다. 그러나, OCLC의 교훈은 여기서 끝나지 않는다. 3000만 건에 달하는 Union catalog의 서지레코드에 대한 품질개선작업은 오늘도 여전히 진행되고 있으며, 특히, 소급자료에 대한 품질개선을 위해 계속해서 막대한 자금이 유입되고 있음에 우리는 주목해야 한다. OCLC의 사례는, DB는 처음 구축단계부터 첫 단추가 올바르게 끼워져야지 그렇지 못할 경우 DB의 가치가 반감되어 결국 이용자로부터 외면을 받거나, 품질개선을 위해 들여야 하는 비용이 종종 구축단계의 비용을 상회할 수도 있다는 값진 교훈을 우리에게 던져주고 있다.

자! 이제 이야기의 초점을 분산체계로 구축중인 대표적인 국내 DB인 KORDIC의 SATURN DB에 맞춰보자.

4. 분산체계로 구축된 국내 DB의 사례

4. 1 KORDIC SATURN DB의 현황

SATURN DB는 국내외를 막론하고 그 유례를 찾기 힘든 논문 단위를 대상으로 삼아 분산체계로 구축된 과학기술분야의 통합 DB이다. KORDIC이 1991년부터 1997년까지 50억에 가까운 비용을 쏟아 부었고, 목표대로라면, 1998년초에는 SATURN DB에 포함된 서지레코드의 수가 100만을 돌파하여 명실공히 단일 DB로는 이 분야에 국내 최대 DB로서의 지위를 얻게된다 (연구개발정보센터, 1997 & 1998). 총투자비용이 50억 가까이 들었으니 서지레코드당 5,000원도 안되는 저렴한 단가로 DB 구축에 성공한 셈이다. 국내외 유사기관의 경우와 비교해도 이 단가는 매우 저렴하여, KORDIC을 비롯한 여러 정부출연연구소의 DB 구축담당자들의 노고에 감사해야할 상황이다.[7]

그러나 이렇듯 저렴한 투자비용에도 불구하고, 아주 기본적인 비용효과분석(cost-benefit analysis)을 실시해 본 결과, "SATURN DB가 과연 성공작일까"하는 의문이 발생하고 있다.[8]

그렇다면, SATURN DB의 투자효과는

7) 문제는 DB의 품질이, 특히 original indexing의 결과로 생산된 서지레코드의 비율이 얼마나 되느냐 하는데 있다. 이 문제는 이 연구에 앞서 1996년도에 수행된 '과학기술분야 서지 DB의 품질관리 및 평가방안'에서 그 부정적 측면이 상세히 논의된 바 있으며, 다음 섹션에서 다시 한번 상세히 논의될 것이다.

어느 정도일까? 논의의 초점을 먼저 이용율 쪽으로 돌려보자. KORDIC이 잠정적으로 설정한 정보서비스대상 고객 중에 어느 정도의 인원이 SATURN DB를 이용하고 있으며, 그 실질적인 효용성은 얼마나 될까? KORDIC이 발간한 〈1996년 종합평가자료〉에 의거하면, KORDIC이 정보서비스의 대상으로 설정한 '고객'의 범위는 대략 12만명 정도로 되어있다. 이들은 정부출연연구소소속 연구자 1만명 정도, 대학소속 연구자 5만명 정도, 그리고 민간연구소소속 연구자 6만명 등으로 구성된다. 그렇다면 이들 중에 과연 몇 퍼센트가 SATURN DB를 이용하고 있는지를 알아보자. 먼저, 위의 〈종합평가자료〉에 나타난대로 사용횟수에 따른 통계를 분석해 보니, 1996년도 한해 동안 SATURN DB에 대한 접속횟수는 5,092회로 나타난다. 이 숫자는 KORDIC의 기관홍보가 본격화되기 시작한 1997년으로 들어서면서 대폭 늘어나, 6월말 현재까지 총 5,240회의 접속횟수를 기록중에 있다 (연구개발정보센터, 1997 & 1998).

그러나, SATURN DB의 이용율을 좀더 실질적으로 산출하려면 단순한 접속횟수보다는 '원문서비스 실적'을 참조하는 것이 나으리라는 판단에서 관련 통계를 조사하였다. 원문서비스가 SATURN DB에 대한 검색의 결과로 파생되는 2차적인 정보서비스임을 고려하면, 이 통계는 보다 실제적인 이용율을 나타낸다고 볼 수 있기 때문이다. 원문서비스는 1996년 12월 본격적인 서비스를 시작한 이래 1997년 6월 말 현재 월 평균 4,000건에 가까운 서비스 실적을 보이고 있다 (연구개발정보센터, 1998). 고무적인 일이다. 실질적인 고객이 증가하고 그들에 의한 이용횟수가 증가하고 있음을 보여주는 자료이다.[9] 그러나, 다시 생각해 보면, 총 투자액수 50억원과 100만 건의 서지레코드를 저울의 반대편에 올려놓을 경우 월 평균 4,000건의 원문서비스 제공이 과연 효용성 있는 투자라고 할 수있을지는 의문이다.

앞서 언급하였던 OCLC Union Catalog와 SATURN DB를 비용효과분석을 통해 비교하는데는 많은 무리가 있다. 하지만, 전체 서지레코드의 수에 대한 상호대차 (ILL) 신청회수의 비율을 산출하여, SATURN DB에서의 전체 서지레코드 수에 대한 원문복사신청률과 비교해 보는 것은 나름대로 의미를 갖

8) DB의 투자효과를 산출한다는 것은 쉬운 일이 아니다. 특히, DB 구축이 진행되고 있는 초기 단계에는 투자효과의 산출이 불가능하고, 일정 기간이 지난 후에도 그 효과가 점증적 누적적으로 나타나는 특성이 있기 때문에 정확한 산출이 어렵다. 물론 상용 DB의 경우는 투자분에 대한 이익회수분을 일정 기간이 지난 후에 통계적으로 분석하면 투자효과의 산출이 물리적으로 가능하지만, 문제는 SATURN DB와 같은 공용 DB의 경우이다. 공용 DB의 경우 투자효과를 산출하는 방법은 극히 제한될 수밖에 없다. Benefit, 즉 효과의 측정을 위한 지표의 개발이 결코 쉽지 않기 때문이다. 선행연구를 분석해 보면, 공용 DB의 투자효과 측정을 위해 가장 널리 쓰이는 방법은 분석대상 DB가 서비스대상으로 설정했던 고객에 의해 실질적으로 얼마나 이용되고 있는지를 산출하는 것이었다. 즉, 전체 서비스대상 고객중에 얼마나 많은 고객들이 얼마나 자주 해당 DB를 활용하느냐가 투자효과를 산출하는 지표로 흔히 쓰였다. 물론, 이용률이나 이용빈도 역시 추상적인 통계여서 진정한 '효용성'이라는 측면에서의 투자효과를 산출해 내는데는 문제가 있지만, 보다 실질적이고 구체적인 지표의 산출이 현실적으로 용이하지 않은 상황에서 '효용성 분석'을 위한 주요 지표로 활용되고 있다.

9) 참고로, 당시에는 서비스의 제공이 무료였으나, 1997년 후반기부터 유료서비스로 전환되었다. 이후, 이용률도 절반정도로 감소한 상태에 있다.

는다. OCLC Union Catalog의 1993년 통계에 의하면, 서지레코드의 수 2,700만건에 ILL 신청회수는 4,200만건에 이르고 있음을 보여준다 (OCLC, 1994). 어림잡아도 서지레코드 한 건에 대한 ILL 신청회수가 연평균 1.5회를 상회하고 있는 것으로 나타난다. SATURN DB의 경우는 어떠한가? 앞서의 〈1996년 종합평가자료〉에 의하면, 1996. 12. 24부터 1997. 6. 25까지 6개월 동안에 원문복사 신청회수는 23,642건이며, 1997년 6월 현재 SATURN DB의 서지레코드 수는 70만건에서 100만건 사이에 있음을 보여준다 (연구개발정보센터, 1998). 이 통계에 근거하여, SATURN DB의 서지레코드 한 건당 원문복사 신청회수 비율을 연평균치로 산출해 보면, 0.15-0.2회 정도에 머무르고 있음을 알 수 있다. 이는, OCLC Union Catalog에 비교할 때 1/8-1/10에 불과한 수치이다.

물론 SATURN DB의 경우 DB의 역사와 원문서비스의 역사가 일천하고 그 이용율이 지속적으로 증가하고 있음을 고려할 때, 이와 같은 단선적인 비교의 결과를 놓고 '효용성'을 운운하는 것에는 많은 무리가 있다. 그러나, 이 비교의 결과는 SATURN DB에 대한 투자와 투자방식 그리고 분산체계라는 구축방법이 과연 옳은 것이었을까 하는 근본적인 의문을 던져준다. 그리고 이와 같은 의문은, SATURN DB의 품질과 효용성을 보다 구체적인 지표를 활용하여 분석해 본 다음의 여러 실험의 결과에서 보다 구체화된다.

4. 2 SATURN DB의 품질검증 결과

SATURN DB의 품질에 대한 논의는 앞서 언급하였던 〈과학기술분야 서지 DB의 품질관리 및 평가방안〉에 관한 연구에서 이미 이루어진 바 있다. 당시 연구방법으로는 SATURN DB를 검색하여 레코드를 출력한 후 에러율(error rate)을 측정하는 방법과 이용자그룹과의 인터뷰를 통해 만족도를 측정하는 방식이 병행하여 사용되었다 (이제환, 1997). 이번 연구에서는 앞서의 연구에서와는 다소 다른 방법이 사용되었다. 물론, 검색실험을 통해 레코드를 표본으로 수집한 후 지난 1년 동안 외형적 품질이 어느 정도 개선되었는지를 점검하기도 하였지만, 이번 품질검증의 초점은 이용자의 관점에서 본 '유용성'의 측정에 주어졌다. 이를 위해 인용문헌분석법이 활용되었고 이용자에 의한 직접적인 검색실험과 실험에 이은 인터뷰가 행해졌다.

4. 2. 1 검색실험을 통해 본 외형적 품질

일년의 시차를 두고 두 번째로 행해진 SATURN DB에 수록된 서지레코드에 대한 품질평가의 결과는 한마디로 실망스러운 것이었다. 지난번 연구에서 지적된 문제점들이 여전히 수정되지 않은 채 반복해서 나타나고 있었다. 당시 설정한 품질평가를 위한 7개 항목 중에서,[10] 국내문헌에 대한 서지레코드의 비율을 늘리려는 노력의 흔적은 보였지만, 여타 품질상의 근본적인 문제는 여전히 남아 있었다. 다음은 7개 항목을 중심으로한

10) 품질검증은 다음의 일곱 항목을 중심으로 행하였다: ① 국내문헌에 대한 레코드와 해외문헌에 대한 레코드의 비율; ② 자체제작 레코드의 비율; ③ 레코드의 최신성 및 갱신주기; ④ 레코드의 중복율; ⑤ 레코드구조의 일관성; ⑥ 데이터필드의 적합성; ⑦ 수록 데이터의 완전성.

평가의 내용을 요약한 것이다.

① 국내문헌에 대한 레코드와 해외문헌에 대한 레코드의 비율: 기관별로 차이는 있었지만, SATURN DB를 구성하는 전체 서지레코드 중에서 국내문헌을 대상으로한 서지레코드의 비율이 증가하고 있음을 느낄 수 있었다. 일부 기관에서는 여전히 해외문헌을 대상으로한 서지레코드가 100%를 점유하는 기관도 있었지만, 여러 기관이 국내 생산 문헌을 대상으로한 서지레코드의 제작에 노력을 기울이고 있는 것으로 나타났다. 후자의 경우, SATURN DB의 토착 DB化를 위한 노력을 가시적으로 보여주고 있다는 점에서 SATURN DB의 미래와 관련하여 그 역할이 주목된다.

② 자체제작 레코드의 비율: 해당 기관에서 원문헌을 대상으로 자체적인 가공분석과정을 거쳐 생산된 것으로 보이는 서지레코드의 비율은 여전히 낮게 나타났다. 해외문헌을 대상으로한 서지레코드의 경우, 해당 기관에서 자체적으로 제작했다고 판단하기 어려운 서지레코드들이 상당수 검출되었다. 검색된 700개의 서지레코드 중에, 약 30% 정도만이 해당 기관에서 자체적인 가공분석과정을 거쳐서 제작된 것으로 보이며, 나머지는 외부의 (특히 해외의) 서지 DB들을 참조하여 제작된 것으로 판단된다. 한편, 대부분의 기관에서 외부의 서지 DB를 참조하는 단계를 넘어 그 내용을 다운로드받아 SATURN DB에 입력한 것으로 추정되는 서지레코드가 상당수 발견되었다.

③ 레코드의 최신성 및 갱신주기:

SATURN DB는 레코드의 최신성과 갱신주기에 있어서도 여전히 많은 문제를 안고 있었다. 서지레코드의 대상인 原文獻의 출판년도와 서지레코드의 입력년도를 비교분석한 결과, 구축기관별 차이가 심각하게 나타났다. 많은 기관들이 출판된지 4-5년이 훨씬 지난 문헌에 대한 서지레코드를 이제서야 입력하고 있었다. 특히, DB 구축에 참여하고 있는 14개 기관이 기관마다 서지레코드 제작 대상 문헌의 출판년도와 제작된 서지레코드의 입력시기에 대한 기준이 서로 달라, SATURN DB에 수록된 레코드의 주제별 최신성은 그야말로 천차만별이었다. 특히, 품질평가를 위해 본 연구팀이 검색해낸 700개의 레코드 중에 1997년 들어 출판된 문헌은 단 1건도 찾아볼 수 없어, SATURN DB의 최신성과 최신성 유지를 위한 레코드 갱신작업이 어느 정도 수준에 있는지를 짐작하게 해주었다.

④ 레코드의 중복율: 레코드의 중복율은 어느 정도 개선이 이루어진 것으로 나타났으나, 몇몇 기관의 경우 여전히 노력이 필요해 보였다. 기관별 소장저널의 중복여부를 조사한 결과, 그 동안 KORDIC의 자체적인 노력으로 14개 기관들 사이에 서로 중복되던 학술저널들의 기관별 정리가 어느 정도 이루어진 것으로 나타났다. 그러나, 각 기관에 의해서 이미 제작되어 SATURN DB에 입력된 레코드들 사이의 중복현상은 아직도 개선의 여지가 남아 있었다.

⑤ 레코드구조의 일관성: 1차 평가시에 지적한 문제점의 대부분이 그대로 남아 있었다. 레코드의 제작기관에 따른 레코드구조의 상이함은 물론이고, 동일한 기관에서 제작한

레코드의 구조가 제작년도에 따라 상이한 경우도 여전하였다. 자료의 유형(잡지기사, 회의록기사, 보고서 등)에 따라 레코드구조가 상이한 점은 어쩔 수 없더라도, 일부 자료(특히, 보고서의 경우)는 제작기관에 따라 레코드구조가 천차만별이어서 시급한 시정을 요하고 있었다. 가장 빈번하게 나타나는 레코드구조와 관련한 문제점으로는: 특정 데이터필드가 존재하다 없어지거나 없다가 새로이 나타나는 경우, 같은 레코드내에서 특정 데이터필드가 중복해서 출현하는 경우, 그리고, 동일한 데이터필드를 가지고 있으면서도 그 순서가 뒤바뀌어 나타나는 경우 등이 있었다.

⑥ 데이터필드의 적합성: 먼저, 불필요한 것으로 보이는 데이터필드에 대한 정리가 여전히 이루어지지 않고 있었다. 가령, 데이터필드가 중복되어 나타나거나, 다른 데이터필드에 수록된 내용이 서로 중복되거나, 의미를 전혀 알 수 없는 데이터가 수록되어 있는 등, 그 존재이유가 의심스러울 정도로 불필요해 보이는 데이터필드가 그대로 방치되어 있었다. 한편, 검색과 자료의 입수 혹은 적합성 판단 등을 위해서 매우 필요한 데이터필드임에도 불구하고 누락되어 있는 경우도 여전히 많았다. 가령, 소장정보나 원문수록처 필드와 같이 SATURN DB의 존재가치와 직결되는 데이터필드들이 누락되어 있는 경우도 흔히 발견되었다. 또한, 수록된 데이터의 내용과 데이터필드명이 맞지 않는 경우도 여전히 발견되었다.

⑦ 수록 데이터의 완전성: 레코드에 수록된 데이터의 불완전성은 SATURN DB의 가장 치명적인 약점이었다. 1차 평가시에도 누차 지적하였건만 안타깝게도 개선의 흔적이 뚜렷해 보이지 않았다. 사례는 다양하였다. 가령, 레코드의 기본구조에는 포함되어 있으나 데이터필드 자체가 없거나 데이터필드는 존재하는데 그 내용이 비어 있어서 해당 데이터필드로는 검색이 불가능한 레코드가 상당수 발견되었다. 이와 같은 데이터필드 혹은 데이터내용의 ommission 문제는 기계적인 데이터수정방법으로 치유될 수 있는 부분이 아니어서 그 문제의 심각성이 더욱 커 보였다. 또한, 데이터필드에 수록된 내용이 불완전한 事例도 다양하게 나타났는데, 가장 빈번하게 나타나는 에러의 종류로는: 오자/탈자/띄어쓰기와 같은 typographical 에러, 데이터의 표기방법이 통일되어 있지 않아서 나타나는 에러 (특히, 외국어 표기와 저자명 표기시), 주제분석이 지나치게 일반적이어서 주제명을 이용한 검색이 검색효율에 오히려 부정적 영향을 주는 경우, 데이터의 技術이 불완전한 경우 등이 있었다.

4. 2. 2 인용문헌분석을 통해 본 유용성

먼저, SATURN DB에 포함되어 있는 14개 주제분야의 서지레코드 제작을 담당하는 12개 출연연구소[11]를 대상으로, 각 연구소에서 연구생산성이 상위그룹에 속하는 책임급 연구원을 10명씩 선정하여 모두 120명을 의

11) KORDIC과 기계연구원 창원분원을 제외한 12개 연구소: KAIST, KIST, KIMM-대전, KRIBB, KRISS, KIER, KRICT, KERI, KARI, KAERI, KIGAM, KRISO.

도적 표본으로 삼았다.[12] 다음, 이들의 연구저작물 중에서 최근에 생산된 것들을 2편씩을 선정하여 모두 240편의 연구저작물을 수집하였다.[13] 이렇게 수집된 연구저작물을 이용하여 연구원들이 연구저작물의 생산을 위해 참조한 총 6,185건의 '인용문헌'을 분석하였다. '인용문헌'에 대한 분석의 목적은, 연구자들이 연구저작물을 생산하는데 있어서 SATURN DB의 유용성이 과연 얼마나 되는지를 추정해 보는데 있었다.

(1) 수집된 '인용문헌'의 특성

수집된 연구저작물 240편에 수록된 인용문헌의 총수는 6,185개로 집계되었고, 수집된 연구저작물당 평균 인용문헌의 수는 25.8개로 조사되었다. 연구원들이 연구저작물의 생산을 위해 주로 이용하는 자료의 형태는, 학술저널에 실린 논문의 이용율이 가장 높았고(55.8%), 단행본, 보고서, 회의록논문, 학위논문 등이 뒤를 이었다. 특허자료나 법률자료의 이용율도 결코 무시할 수 없는 비율을 보여, 연구활동을 위한 이들 자료의 체계적인 공급이 필요함을 보여주었다. 한편, 인용문헌의 생산지별 (언어별 혹은 국가별) 분석의 결과는, 영어권 국가에서 영어로 출판된 문헌이 82.6%로 압도적이었고, 그 뒤를 한국, 일본, 독일 등이 잇고 있었다.

다음으로, 인용문헌의 출판년도를 분석하여 최신자료에 대한 선호도를 알아보았다. 과학기술분야의 연구자들을 대상으로한 정보이용행태에 관한 연구들은, 과학기술자들은 주로 5년 이내에 생산된 최신 자료에 대한 의존도가 다른 학문분야에 비해 매우 높다는 결론을 내놓았었다. 그러나, 이번 인용문헌에 대한 분석결과는 상이하게 나타났다. 이번 연구에서 수집하여 분석한 전체 인용문헌 중에서, 연구자가 이용했던 시점을 기준으로하여 5년 이내에 생산된 자료는 모두 합쳐도 35.6%에 불과하였고, 생산된지 5년 이상된 자료가 전체의 64.4%를 점하는 것으로 나타났다. 이중에는 10년이전에 생산된 자료도 36.4%나 차지하여, 단일 항목으로는 가장 높은 수치를 기록하였다.

한 가지 흥미로운 결과는, 인용된 국내문헌의 출판년도는 해외 문헌에 비해 현격한 차이를 보였다. 국내 문헌의 경우 활용 시점에서 5년 이내에 출판된 문헌의 인용율이 80%이상으로 집계되었다. 이 통계는, 적어도 국내문헌에 있어서는 '최신성'이 참고자료의 선정에 있어 중요한 요인이 되고 있음을 의미한다. 그러나, 해외문헌의 경우는 5년 이내에 출판된 자료의 인용율이 32-3%대에 머물러, 해외문헌에 있어서는 자료의 '최신성'이 자료선정의 최우선 조건이 아닌 것으로 나타

12) 정부출연연구소의 성격상, 연구프로젝트의 실질적인 책임자는 주로 책임급 또는 선임급연구원들이 맡고 있었으며, 인력 DB를 탐색한 결과 책임급연구원들에 의해서 생산된 연구저작물의 수가 선임급이나 일반 연구원에 의한 연구저작물에 비해 월등히 많은 것으로 나타났다. 이에 연구생산활동이 활발한 연구자그룹이 타그룹에 비해 정보요구도 높고 정보자료의 이용률도 높을 것이라는 가정하에, 책임연구원그룹을 이 실험을 위한 sample group으로 선정하였다.

13) 이때, 연구저작물은 1996년 말을 기준으로 3년 이내의 자료로 제한하였으나, 다만, 자료를 구할 수 없는 경우는 5년 이내까지로 한정하였다. 연구저작물의 유형은 연구자들의 신분특성상 연구보고서를 중심으로 수집하였다.

났다.

(2) 인용문헌분석에서 나타난 유용성

수집한 인용문헌을 이용하여 SATURN DB의 유용성을 측정하였다. 연구자들이 연구저작물의 생산을 위해 인용한 문헌 중에 과연 몇 퍼센트나 SATURN DB에서 검색되어 질 것인지를 파악하기 위한 목적의 실험이었다. 이 실험에서는 인용문헌에 나타난 서명을 이용하여 검색하였으며, 단, 서명이 기입되어 있지 않은 경우에는 저자명에 의한 검색을 병행하였다. 먼저, 모두 6,185건의 인용문헌중에서 SATURN DB에서 검색에 성공한 문헌은 단 139건에 불과하였다. Overall hit rate = 2.3%. 이 수치를 역으로 해석하면, SATURN DB의 일차적 서비스대상인 출연연구소소속 연구자들이 연구프로젝트의 수행을 위해 자료탐색행위에 들어가서 SATURN DB를 이용하였을 때, SATURN DB로부터 서지정보를 얻을 수 있는 문헌은 그들이 필요로 하는 문헌의 2.3%에 불과함을 의미한다. 이는 다시, 그들이 50개의 문헌을 필요로 한다면 단 1개의 문헌에 대한 서지정보만을 SATURN DB로부터 탐색해 낼 수 있음을 의미하는데, 이 결과는 그들에게 있어 정보시스템으로서의 SATURN DB의 의미는 극히 미미함을 보여준다.

(3) 소장처 분석에서 나타난 유용성

다음은, 검색실험에서 SATURN DB에 포함되어있는 것으로 확인된 139개의 문헌을 소장처에 따라 분석하였다. 이 분석의 목적은 통합 DB로서의, 즉, 연구소간에 상호대차서비스를 위한 정보유통도구로서의 SATURN DB의 유용성을 파악하는데 목적이 있었다. 분석결과, 전체적으로 보아 139개의 문헌 중 약 56%에 해당하는 78개의 문헌이 연구원들이 소속된 연구기관에 소장되어 있었으며, 나머지 44%에 이르는 61개의 문헌은 다른 연구기관들에 소장되어 있는 것으로 나타났다. 이 결과는, 분석대상 문헌의 수가 너무 적어서 일반적 결론을 유도해내는데는 문제가 있으나, 적어도 관련 연구기관 사이의 원문복사서비스를 위한 서지도구로서의 SATURN DB의 유용성과 관련하여서는 긍정적인 메시지를 주고 있다.[14] 적어도, 이 결과는 연구활동에 필요한 정보자료의 소장처를 알기 위한 서지 도구로서의 SATURN DB의 존재가치를 보여주고 있기 때문이다.

(4) SATURN DB와 SCI의 유용성 비교

동일 주제분야의 유사 DB와의 비교분석을 통해 SATURN DB의 유용성을 다시 한번 평가해 보자는 의도에서, 과학기술분야의 종합 색인인 Science Citation Index를 선택하였다. 실험조건은, 수집한 인용문헌 중에서 저

14) 연구원들이 연구활동에 필요한 정보자료를 소속된 연구기관의 장서로부터 충족하는 例는 전세계 어느 기관을 막론하고 존재하지도 가능하지도 않다. 통합 DB는 정보공유를 위해 가장 기본이 되는 도구이나, 그 성패는 통합 DB의 구성에 참가한 멤버들이 어느 정도 체계적으로 자기 몫의 정보자원을 개발하고 수집하여 그에 대한 서지정보를 통합 DB에 구축하여 주느냐, 나아가, 어느 정도 적극적으로 자관 보유의 정보자원에 대한 타기관의 이용에 협조하느냐의 여부에 달려있다. 과학기술정보의 체계적 유통을 위한 서지 도구로서의 SATURN DB의 가치와 유용성도 결국은 이점과 맞물려 있음을 생각할 때, 이 분석결과는 SATURN DB의 미래를 위해 그나마 긍정적인 의미를 던져주고 있는 것이다.

널논문만을 검색대상으로 하되,[15] 출판년도는 1992-1996년으로 언어는 영어로 쓰여진 것으로 제한하였다.[16] 총 6,185개의 인용문헌 중 저널논문의 수는 3,452개였고, 이중 1992~1996년에 출판된 저널논문의 수는 396개로 조사되었다. 따라서, 이 396개의 저널논문을 대상으로 실험을 진행하였다. 실험의 결과, SATURN DB에서의 hit rate이 6.3%에 불과한데 비해 SCI에서의 hit rate는 58.3%에 이르는 것으로 나타났다. 이 수치는, 1997년 12월 현재, 정부출연연구소소속 연구원에게 있어 SATURN DB의 유용성은 SCI의 1/9도 안된다는 사실을 적나라하게 보여주고 있다. 이처럼, 7년의 기간에 걸쳐 약 50억을 투자하여 구축한 DB의 유용성이 이 정도에 불과하다면, SATURN DB에 대한 투자와 구축방법이 과연 합리적인 것이었는지에 대한 의문을 갖는 것은 어찌 보면 당연할지도 모른다.

4. 2. 3 이용자가 말하는 포괄성과 적합성

일년 전 연구에서 인터뷰에 응했던 대상자중에서 KRISTAL system을 지속적으로 사용하고 있는 것으로 밝혀진[17] 12명을 선정하여, 이들이 직접 참여하는 SATURN DB에 대한 검색실험을 실시하였다. 12개 주제분야에서 각 연구원들의 전공 및 현재 수행중인 프로젝트와 관련하여 가장 관심있는 2개의 질의를 스스로 개발하도록 하고, 이 질의에 적합한 문헌을 SATURN DB와 SCI(online version)를 이용하여 각각 탐색하게 하였다. 탐색에 필요한 키워드는 표제키워드로 한정하되 본 연구자와 참여 연구원사이에 논의를 거쳐 선정하였다. 수집된 데이터는 이용자의 관점에서 본 SATURN DB의 내용적 포괄성과 적합성에 대한 평가를 목적으로 분석되었다.

검색실험 직후 나타난 연구원들의 종합적인 견해는, "일년 전에 비해 크게 달라진 것이 없다"는 것이었다. KRISTAL system 전체로 보아서는 여러 면에서 개선이 있었으나, SATURN DB만을 독립적으로 놓고 볼 때는 무엇이 개선되었는지 파악하기 힘들다는 반응이었다. 이들이 SATURN DB를 이용하는 주목적은 프로젝트 혹은 관심분야에 대한 포괄적인 문헌조사에 있으나, SATURN DB는 이들의 이러한 목적을 충족시키기에는 여전히 내용적 포괄성에 있어 많은 한계를 지니고 있다는 지적이었다. 그럼에도 불구하고 이들이 SATURN DB를 계속해서 이용하고 있는 이유는, 소속연구소에서 구할 수 없는 원문헌를 원문서비스 신청을 통해 비교적 용이하게 입수할 수 있다는 점 때문인 것으로 조사되었다.

이들이 느끼는 SATURN DB의 내용상의

15) Science Citation Index는 저널논문을 주로 수록하는 서지 DB인 까닭에, 비교를 위한 조건을 일치시키자는 의도에서 검색실험용 인용문헌을 저널논문으로 제한하였다.

16) 이유는 실험에 사용된 SCI의 CD-ROM 판이 1992~1996년 동안 영어로 발표된 논문위주의 서지레코드를 수록하고 있었기 때문이었다.

17) 본격적인 검색실험에 앞서 1997년 9월 현재 접촉이 가능한 연구원들 (39명)을 대상으로 간단한 전화인터뷰가 실시되었다. 전화인터뷰 결과, 작년 인터뷰 실시이후 KRISTAL system을 지속적으로 (정기적으로는 아닐지라도) 사용해오고 있는 연구원들 (26명) 중에서, 검색실험에 협조할 의사가 분명한 연구원을 기관별로 1명씩 모두 12명 선정하였다.

문제점을 보다 구체적으로 파악하기 위해, 검색실험직후 간단한 인터뷰가 행해졌다. 먼저, SATURN DB의 내용적 포괄성에 대한 느낌을 물었다. "자신의 질의에 해당하는 문헌 정보를 어느 정도나 얻었느냐"는 질문에, 연구원들의 대답은 "충분하지 않다, 이 정도로는 어림없다, 다른 DB에 대한 탐색이 필요하다" 등으로 나타났다. 5단위 Likert scale로 충족도를 표하라고 하였더니 대부분 1 혹은 2에 표시하여 평균치가 1.71 (매우 만족한다를 5로, 전혀 만족하지 못한다를 1로 하여)에 불과하였다. 작년 실험때 보다는 0.11 포인트 상승한 결과이나 그 차이는 미미하였다.

다시, 자신의 정보요구에 대한 검색결과의 적합성을 물었다. 5단위 Likert scale로 나타난 충족도는 1.34 (모두 적합하다를 5로, 적합한 것이 전혀 없다를 1로 하여)에 불과하였다. 작년 실험때 보다 0.14 포인트 정도 상승하였으나, 이 또한 그 의미가 크지 않았다. 이번에는, 검색된 문헌 목록을 이용하여 정확율(precision)의 측면에서 적합성을 산출해보았다. 결과는 평균정확율 17.4%.[18] 평균 23건의 검색문헌중에 약 4건만이 자신들의 연구수행을 위해 '아마도' 필요한 문헌으로 선정되었다.

한편, SATURN DB에 대한 검색후에 실시된 SCI에 대한 검색의 결과는, SATURN DB에 대한 연구원들의 평가를 더욱 부정적으로 만들었다. 포괄성에 대한 만족도 3.9, 적합성에 대한 만족도 4.3, 그리고 60%를 상회하는 정확율(precision). SCI를 이용한 검색실험후에 연구자들에 의한 평가 점수이다. "SATURN DB의 구축 목적이 무엇입니까? 차라리 그 돈이 있으면 原文이 수록된 저널을 한 종이라도 더 구입해 주지..." 검색실험을 마친 후, 한 연구원이 본 연구자를 향해 던진 질문이자 건의였다.

이상의 결과는 SATURN DB에 포함된 서지레코드의 수가 1996년에 비해 20% 가량 (약 20만 레코드 정도) 증가했음에도 불구하고, 그 내용적 포괄성이나 적합성에 대한 연구원들의 반응은 여전히 부정적임을 보여주고 있다. 그러나, 이 검색실험의 결과로 얻게 된 다음의 사실은 SATURN DB의 미래와 관련하여 매우 중요한 의미를 지니고 있었다. 그 하나는, '아마도' 필요할 것으로 선정된 4건의 문헌의 소재지였다. 4건 중에서 평균 2건은 다른 연구기관에 소장된 문헌이었으며, 따라서 원문의 입수를 위해서는 원문서비스 신청을 요하는 대상이었다. 다른 하나는, '아마도' 필요할 것이라는 17.4%의 문헌의 내용이었다. 이 문헌 중에 약 1/3 정도는(정확히 34.2%) 국내에서 생산된 문헌이었으며, 이는 SATURN DB에 포함된 국내 문헌의 비율이 극히 소량임에도 불구하고 그 소량의 국내 문헌에 대한 연구원들의 '필요성'은 의외로 높음을 의미하고 있었다.

4. 3 SATURN DB의 품질관리 실태

SATURN DB의 품질이 형태나 내용면에서 지금처럼 열악해진 근본 요인은 무엇일까? SATURN DB는 국가차원에서 과학기술정보의 체계적이고 효율적인 유통을 위한

18) 총 549건의 검색문헌중에서, 95건만이 검색에 이용된 질의에 적합한 문헌으로 선정되었다.

'통합 서지 도구'의 성격이 짙은 DB이다. 구축 초기의 계획안을 살펴보면, KAIST를 비롯한 25개의 정부출연연구기관이 기초과학분야에서부터 농업관련분야에 이르기까지 전담 주제영역을 설정하고[19], 각 주제분야에서 연구개발활동을 위해 필요한 정보자료를 수집 분석 가공하여 주제분야별로 Sub-DB를 구축하여 운영하되, 각자 구축한 Sub-DB를 중앙정보센터(KORDIC)로 보내 통합 DB를 구성하고, 이 통합 DB를 대상으로 자료검색을 할 수 있는 정보체계를 구축하고자 했음을 알 수 있다. 이처럼, 초기 계획안에서 유추할 수 있는 SATURN DB의 구축목적은 정부출연연구소들간의 과학기술정보의 공유를 위한 통합형 서지 DB를 구축하고자 하는 것이었고, 그 목적을 이루기 위한 구축전략은 분산체계이었다.

앞서 OCLC union catalog에 대한 설명에서도 밝혔듯이, 분산체계로 구축하고자하는 통합형 서지 DB는, 그 주목적이 참여기관들 사이의 정보자원의 효율적인 공유에 있다. 따라서 DB의 구축전략은 각 참여기관과 중앙조정기관의 역할과 기능을 고려하여 세밀히 작성되어야 한다. 구축전략에는 구체적으로: 통합 DB에서 커버하고자 하는 주제영역의 범위, 입력대상자료의 유형, 입력대상자료의 시간적 범위, 대상자료에 대한 가공분석의 정도, 서지레코드의 구조, 그리고 데이터의 표기방법 등에 대한 원칙과 기준이 설정되어야 하며, 나아가, 참여기관들 사이에 데이터의 중복방지와 표준성 혹은 통일성 유지를 위한 방안이 세부적으로 마련되어야 한다. 그런 다음, 이 원칙과 방안에 근거하여, 참여기관별로 전담주제영역의 설정, 입력대상자료의 선정, 입력대상자료의 가공일정, 그리고 가공된 자료의 입력일정 등이 종합적으로 조정되어야 하는데, 이와 같은 구축전략에 따른 DB의 실질적인 구축과정은 통합 DB의 관리를 담당할 중앙조정기관이 주도적으로 관리 감독해 나가야 한다.

그런데 SATURN DB의 경우는 어떠하였는가? 구축전략의 마련과 실제적인 구축과정 그리고 구축된 DB에 대한 관리는 체계적이었으며, 특히 구축전략의 수립기관이자 집행기관이며 동시에 관리기관이어야 하는 KORDIC은 그 역할과 의무를 충실히 이행하였는가? 본 연구팀이 그 동안 몇 차례에 걸쳐 실시했던 SATURN DB의 품질검증을 위한 실험의 결과는, SATURN DB의 구축전략과 구축과정 모두가 체계적이고 일관적이지 못하였으며, 특히, 계획의 주도기관이자 작업의 집행기관이며 동시에 중앙조정기관인 KORDIC은 부여된 역할과 의무를 제대로 이행하지 못했음을 보여준다.

무엇보다도 KORDIC은 분산체계를 앞세우며, 스스로 자신의 역할을 지나치게 한정하는 우를 범하였다. 1991년부터 1996년까지 6년 동안 KORDIC은, DB 구축을 위해 편성된 예산의 분배자로서, 제작 입력된 레코드의 건수를 헤아리는 집계자로서, 그리고 사업결과를 모아 상부기관인 과기처에 보고하는 중계자로서의 역할에 지나치게 충실했던

19) 김창근. "SATURN 데이터베이스의 효과적인 이용을 위한 분산시스템의 구축," Proceedings of the 1996 KOSTI Workshop. pp.10.

것은 아닐까? 물론 정보검색을 위한 도구(KRISTAL 검색엔진)의 개발 등 여러 사업이 KORDIC의 기관차원에서 수행되었고, 특히 1997년 들어서는 원문서비스의 제공과 같은 정보서비스의 강화가 이루어지기도 하였지만, 적어도 SATURN DB의 구축 및 관리에 관한 한 KORDIC의 역할은 한없이 초라하고 빈약한 것이었다.

가령, SATURN DB의 성격과 방향 설정을 위한 작업, SATURN DB의 제작에 참여하는 기관의 선정 및 그 역할과 의무에 대한 조정, SATURN DB의 구축을 위한 세부 전략의 마련 (앞서 언급한, sub-DB를 위한 주제영역 설정부터 데이터의 표기방법에 대한 원칙과 기준의 설정에 이르기까지 전체 과정의), 참여기관이 그 역할과 의무를 제대로 이행하고 있는지에 대한 감독, 그리고 구축된 DB의 관리 및 유지를 위한 방안의 마련 등, 관계 기관들 사이에 묵시적 합의에 의해 부여된 SATURN DB의 중앙조정기관이자 관리기관으로서의 역할 중 어느 것 하나 KORDIC이 주도적이고 체계적으로 수행하였다는 흔적이 보이지 않는다.[20]

그 결과, 일부 관계자들 사이에서 SATURN DB의 구축 중단을 논의해야 할 정도로 이 사업은 '뜨거운 감자'로 등장하고 있다. 문제점을 구체적으로 살펴보자: 먼저, 당초 계획에 포함되었던 25개 대상기관 중 1997년 말 현재 13개 기관만이 DB 구축에 참여함으로 인해 처음에 의도했던 과학기술 분야의 통합 DB로서의 주제적 포괄성에서 한계가 나타나고 있으며; 다음, 신규 참여기관 및 주제영역의 선정에 있어 원칙과 기준의 미비로 인해 통합 DB로서의 성격이 불투명해졌으며; 또한, 당초 기관별로 할당되었던 주제영역이 중도에 변경되거나 입력대상 자료가 변경됨으로 인해 참여기관들 사이에 작업의 중복율이 높아졌으며; 특히, 통합 DB 구축을 위한 세부 원칙과 규정 그리고 도구를 준비하는데 소홀함으로 인해 DB의 품질유지에 실패하였으며; 나아가, 참여기관의 작업내용과 일정에 대한 감독 소홀로 인해 DB의 품질이 그 형태와 내용면에서의 더욱 조악해지는 결과를 낳고 있다.

이와 같은 SATURN DB가 갖는 품질 문제의 대부분은 KORDIC의 역할 소홀 (작업의 주도기관이자 집행기관이며 중앙조정기관이자 감독기관으로서의 직무 소홀)에서 비롯되었다. 물론, DB 제작에 참여한 기관들의 책임의식 결여로 인한 의무불이행이 SATURN DB의 품질을 조악하게 만든 주범이었음은 분명하지만, 근본적인 요인은 KORDIC이 자신에게 부여된 역할과 의무를 충실히 이행하지 못한데 있었다. 부여된 역할과 의무를 소홀히 한 대가로, KORDIC은 지금 SATURN DB라는 막대한 투자의 *産物*을 살려야 하는지 아니면 죽일 것인지에 대한 최후 결정을 내려야 하는 심각한 처지에 놓이게 되었다. SATURN DB를 폐기처분하자니 기왕에 투자한 비용이 아깝고 지속적으로 구축하자니 조악한 품질로 인해 그 효용성에 근본적인 의문이 제기되는 등,

20) 1996년 말에 이르러서야 비로소 SATURN DB에 대한 운영자로서의 관심이 증대하기 시작하였지만, 1997년 말 현재까지 여전히 SATURN DB의 미래에 대한 청사진이 분명해 보이지 않는다.

SATURN DB의 미래를 놓고 KORDIC은 중대한 결정을 내려야 하는 시점에 와있다.

5. 결론: 통합 DB의 품질관리 방안

KORDIC이 SATURN DB의 미래를 놓고 취할 수 있는 선택은 두 가지로 압축된다. 첫째는, SATURN이라는 통합 DB의 구축 자체가 잘못된 결정이었다면 혹은 환경의 변화로 인하여 더이상 국가단위의 통합 DB의 필요성이 없어졌다면, 기투자액이 50억이 되었든 혹은 100억이 되었든 KORDIC은 스스로의 우를 솔직히 인정하고 당장 작업을 중단하는 것이 최선일 것이다. 둘째는, 통합 DB의 필요성이 여전히 존재하고 장차 비용효과를 증대시킬 수 있는 방안을 마련할 수 있다면 폐기처분 보다는 지속적인 투자가 당연한 해답일 것이다. 이 연구에서 실시한 여러 실험의 결과는, 품질상의 문제점과 그로 인한 이용자들의 높은 불만에도 불구하고 SATURN DB의 필요성은 여전히 남아있는 것으로 나타난다.[21]

즉, 최소의 비용을 들여 DB의 품질을 획기적으로 개선할 수 있는 방안을 찾는다면, SATURN DB의 장래가 그렇게 비관적인 것만은 아니라는 이야기이다. 이에 여기서는, SATURN DB의 품질을 근본적으로 개선하기 위한 방안을, 앞서 소개하였던 OCLC Union Catalog의 事例를 참조하면서, 다음 두 가지 관점에서 제시하고자 한다: 먼저, 미시적 관점에서 SATURN DB의 품질상의 제문제 (특히 형태의 완전성과 내용의 유용성 문제)를 해결하기 위한 실무적 방안을, 다음, 거시적 관점에서 SATURN DB의 품질저하를 유발한 근본 요인을 제거하기 위한 정책적 조언을 제시하고자 한다.

5. 1 실무 방안

앞서 우리는 SATURN DB의 품질을 크게 내용과 형태의 두 관점에서 분석하고, 각 관점에서 문제점을 파악하였다. 여기서는 파악된 문제점들을 다시 인용하면서, 특히 내용의 유용성 증대와 형태의 완전성 제고를 통한 SATURN DB의 품질개선 방안에 대해서 논의하고자 한다.

(1) 내용의 유용성 증대를 위해서:

'인용문헌분석'의 결과에 의하면, 연구개발활동을 위한 SATURN DB의 유용성은 매우 낮은 것으로 나타났다. 그러면서도 한편으로는, 참여기관들 사이에 정보자원의 공유를 위한 서지 도구로서의 역할은 여전히 중요한 것으로 나타났다. 특히 원문헌의 생산지별 분석결과는, SATURN DB에 포함된 국내문헌이 해외문헌에 비해 연구자들에게 보다 유용하다는 해석을 가능하게 하였다. 검색실험에 참여한 연구원들도 SATURN DB의 전반적인 유용성에 대해서는 부정적인 견해를 보였지만, SATURN DB에 포함된 국내문헌의 '잠재적 유용성'에 대해서는 매

21) 특히, 인용문헌의 소장처 조사를 통한 SATURN DB의 유용성 분석 결과는, SATURN DB는 당초의 구축 목적 (즉, 자원공유를 위한 서지 도구)을 위해서라도 여전히 '존속이 필요하다'는 결론에 이르게 한다.

우 긍정적인 태도를 보였다. 이러한 여러 실험의 결과는 SATURN DB가 유용성의 증대를 위해서 향후 어떤 성격을 띠어야 할 지 그 방향을 제시한다.

여기서 본 연구팀의 제언은 SATURN DB는 내용적 유용성을 증대하기 위해서 국내문헌에 대한 비중을 강화할 필요가 있다는 것이다. 이를 위한 실천 방안은, 현재의 SATURN DB를 해외문헌 파트와 국내문헌 파트를 분리하여 구축하되, 통합하여 운영하자는 것이다. 구체적으로, 해외문헌 파트는 소장자료 중심으로 재구축하되 원문헌에 대한 가공처리를 독창적으로 수행하기보다는 기존의 해외 DB를 활용하여[22] 구축시간과 비용을 절약하고, 대신에, 소장자료를 이용한 원문제공능력을 강화하자는 것이다. 그리고 가공처리과정에서 절약된 비용을 원문헌(특히, 저널)을 확충하는데 투자하되, 여러 주제분야를 동시에 커버하는 학제적 저널과 주제분야별로 특성화된 저널의 종수를 동시에 증가시키자는 것이다.

한편, 국내문헌 파트는 SATURN DB를 궁극적으로 '토착 DB化' 한다는 목표로 추진하는 것이 바람직하다. 특히, DB의 포괄성과 적합성 그리고 최신성을 강화하는 것을 통해서 DB의 전반적인 품질을 고급화하는 것이 필요하다. 구체적으로 보면, 주제분야를 보다 세분화하고 다양화하여 가능한 과학기술의 모든 분야를 포함하도록 하되(정책분야도 포함하여), 단행본을 제외한 모든 유형의 자료를 포함시킴으로써 DB의 주제적 자료유형적 포괄성을 강화해야만 한다. 특히, 저널류는 種별로 back issues와 결호를 완벽하게 보충하여 시간적 측면에서의 포괄성과 완전성을 강화해야만 한다. 다음으로, 국내 원문헌의 가공분석 과정을 심화하여 서지레코드에 포함되는 내용을 고급화함으로써 유사기관들에서 제작된 DB들과의 품질상의 차별화를 시도해야만 한다. 이를 위해서는 초록 형태로의 완전한 전환이 바람직할 것으로 보인다. 더불어 DB의 최신성을 증대하는 것도 매우 중요하므로, 가공분석 일정을 최신자료 중심으로 편성하는 것이 바람직하며, back issues의 가공을 위해서는 별도의 작업팀에 의한 별도의 일정을 마련하는 것이 필요하다. 한편, 원문제공능력의 강화는 필수적이며, 이를 위해서는 국내문헌 파트에 포함된 모든 문헌을 궁극적으로는 全文 DB화한다는 목표로 전반적인 계획을 수립하는 것이 바람직해 보인다

(2) 형태의 완전성 제고를 위해서:

두 번에 걸친 검색실험의 결과에 의하면, 형태의 완전성 제고는 SATURN DB의 품질개선을 위해 가장 시급한 과제이다. 형태의 불완전함으로 인해 검색효율이 저하됨은 물론 검색된 자료에 대한 이용자의 신뢰도가 낮아지고, 궁극적으로는, SATURN DB와 KORDIC에 대한 이용자의 불신감이 커지기 때문이다.

형태의 완전성 제고를 위한 제안은 다음 다섯 항목으로 요약된다: 〈중복 레코드의 제거〉, 〈레코드구조의 일관성 확보〉, 〈데이터필드의 내용 보충 및 수정〉, 〈표기방법의 일관

22) 가령, ISI (Institute for Scientific Information)와 같은 해외 DB vendor들과의 직접적인 협약을 통해서.

성 확보〉, 그리고 〈오자와 탈자 등의 철자표기상의 오류 수정〉. 이상의 다섯 항목 중에서 〈중복 레코드의 제거〉와 〈철자표기상의 오류 수정〉은 기계적인 방법을 활용하여 부분적 개선을 기대할 수 있는 항목이나, 나머지 항목들은 기계적인 해결보다는 제도적인 해결을 절실히 필요로 한다. 특히, 〈데이터필드의 내용 보충 및 수정〉은 철저히 전문가에 의한 수작업에 의존할 수밖에 없는 항목이다. 이처럼, 다섯 항목의 작업들은 전문 인력과 많은 시간, 궁극적으로는, 추가 예산의 확보를 요구한다. 그러면서도 작업과정 및 결과에 대한 신뢰도는 여전히 담보하기 어려운 상태에 있다.

다섯 항목의 각 항목별로 구체적인 방안에 대해 논의해 보자. 먼저, SATURN DB에서 〈중복 레코드의 제거〉작업은 다른 항목에 비해 비교적 빠른 시일내에 해결될 수 있을 것으로 전망된다. 그 까닭은, KORDIC의 기관 차원에서 DB 구축에 참여하고 있는 기관들이 소장하고 있는 저널의 중복을 수정하는 작업이 이미 시행되어, 적어도, 앞으로 제작될 서지레코드에서 중복 레코드의 발생 가능성이 대폭 줄었기 때문이다. 문제는 현재 SATURN DB에 포함되어 있는 중복 레코드의 제거 작업인데, 이 방안은 유사한 경험을 했던 OCLC의 사례에서 배울 점이 있어 보인다. OCLC는 'master record' 제도의 실시, 중복 레코드의 자동적 발견을 위한 기계적 장치, 그리고 'the Project Enhance'와 같은 제도적 장치를 복합적으로 활용하면서, 이 문제를 해결해 왔다. 이처럼, 중복 레코드의 제거를 위해서는 기계적 장치를 활용하여 중복 레코드를 자동적으로 파악하되, 제도적 장치를 통하여 폐기할 레코드를 선정하는 복합적 방안이 요구된다. 따라서, 이 방안은 결국 서지레코드 제작기관의 차원에서가 아니라 중앙조정기관인 KORDIC 차원에서 마련되어야 한다. KORDIC은 기계적 장치의 조속한 개발에 이어, OCLC의 ODQCS(품질관리부서)와 같은 제도적 장치를 시급히 마련해야만 하는 시점에 와있다.

다음으로 〈레코드구조의 일관성 확보〉작업의 경우, 상황은 다소 복잡해진다. 최소한 OCLC에게는 이 문제에 대한 고민은 없었다. Union catalog에 포함되는 자료의 형태도 일정하였고, 구축에 앞서 이미 OCLC MARC format이라는 통일된 레코드구조를 만들어 모든 참여도서관들이 준수하도록 유도한 결과였다. SATURN DB의 경우에 이 작업은 KORDIC의 주도아래 총체적으로 시행되어야 하는 수정 작업이다. KORDIC이 마련한 자료유형별 표준 포맷에 따라 서지레코드의 제작기관들이 수정 작업을 실시하되, 작업의 결과를 KORDIC이 재검하는 체계적인 장치의 마련이 필요하다. 이 장치를 구체적으로 어떻게 만들지는 KORDIC에 달려있다. 앞서 제안한 것처럼, 이 작업 또한 KORDIC에 전담팀을 만드는 제도적 장치의 마련을 통해 해결하는 것이 지금으로서는 가장 바람직해 보인다.

〈데이터필드의 내용 보충 및 수정〉 작업은 전적으로 사람에게 의지해야만 하는 작업이다. 특히, 데이터의 보충이 절실히 필요한 주제명 필드의 경우 주제전문가에 의한 색인어 선정이라는 고도의 지적 작업이 요구된다.

따라서 현재 SATURN DB에 포함되어 있는 자료중에서 일부만 개선하려해도 그 비용과 시간이 막대할 것으로 예상된다. 더욱이 이 작업을 수행하여 SATURN DB를 진정으로 고급화된 서지레코드의 집합체로 거듭나게 할 인력의 확보는 가장 큰 문제이다. 그렇다면, 이 작업을 위해 구체적으로 어떠한 방안이 있을 수 있을까? 본 연구팀은 앞서 '유용성 증대 방안'에서 제기하였던 SATURN DB를 해외문헌과 국내문헌으로 분리하여 구축하되 통합하여 관리하자는 안을 여기서 다시 제안하고자 한다. 즉, 해외문헌에 대한 서지레코드는 외부 DB를 참조하여 수정하거나 재구축하고, 양이 비교적 적은 국내문헌에 대한 서지레코드만을 대상으로 수정작업을 전개하자는 것이다. 국내문헌에 대한 레코드의 수정작업은 제작기관별로 작업을 할당하여 1차 수정을 하도록 한 후, KORDIC의 전담팀이 평가와 2차 수정을 시행하는 방안이 바람직해 보인다.

〈표기방법의 일관성 결여〉. 단일 항목으로는 SATURN DB의 품질을 조악하게 만든 가장 큰 요인이다. 문제의 원인은 표기방법의 원칙과 기준을 수록한 메뉴얼이 존재하지 않는다는 것과 저자명이나 서명 그리고 주제명의 표준형태를 수록한 전거화일이 없다는데 있다. 이 잘못은 전적으로 중앙조정기관인 KORDIC이 자신의 의무를 망각한데서 비롯된다. 따라서 이 문제는 중앙조정기관인 KORDIC이 이제라도 주도적으로 해결해 나가야 한다. KORDIC은 통일된 표기방법에 대한 매뉴얼을 시급히 (그러나 완전하게) 제작하여 배포하고, 관련자에 대한 교육을 실시하여야 한다. 더불어, 저자명이나 기관명 그리고 잡지명 등에 대한 전거화일과 과학기술분야의 종합적인 주제명 전거화일의 제작을 신중히 고려해야만 한다. 이는 국가차원의 정보유통기관으로서 KORDIC이 당연히 해야할 일이며, 특히, SATURN DB의 품질관리를 위해서는 반드시 필요한 사안이다.

마지막으로, 〈오자나 탈자 등의 철자표기상의 오류를 수정〉하는 작업이다. 데이터의 정확성 확보라는 측면에서 볼 때, 이 작업의 절실성은 앞서 언급한 어느 작업에 못지 않다. OCLC의 경우, 자동에러탐지와 수정을 위한 기계적 장치를 활용함으로써 Union catalog에 포함된 데이터의 정확성을 일정 수준 이상으로 끌어올릴 수 있었다. 이처럼, 이 부분이 품질개선을 위한 여러 작업 중에서 기계적 방법의 효과가 가장 크게 나타나는 분야임에는 틀림없지만, 이 또한 전적으로 기계에만 의존하는데는 무리가 있다. SATURN DB의 품질개선을 위해 KORDIC은 기계적 방법의 개발을 서두르고 있다. 1997년 초에 Oracle을 이용하여 철자상의 오류를 수정하기 위한 도구의 개발에 들어갔고, 하반기에는 실험적이기는 하지만 이 도구를 실무에 투입하기 시작하였다.[23] 그러나, 이로써 충분하다는 생각은 위험하다. 철자상 오류문제의 완벽한 해결을 위해서 OCLC도 제도적 장치의 활용을 병행하였음을 인식할 필요가 있다. 전담 부서를 만들어 활용하든 제작자 인센티브 제도를 활용하든 제도적 장치의 보완이 반드시 필요함을 관계

23) KORDIC Newsletter. 제6호 (1997.5). pp. 5.

자들은 인식하여야 한다.

5. 2 정책 제언

위에서 제시한 方案들을 효과적으로 추진하기 위해서는 무엇보다도, DB 구축체계와 관리체계의 개편 내지는 조정이 필요하다. 이 문제는 현재의 분산체계가 바람직한지 아니면 중앙집중처리형으로의 전환이 필요한지에 대한 논쟁을 야기할 소지가 있는 만큼, 직접적인 논의는 뒤로 미루고자 한다. 그러나 한 가지 분명한 것은 현재의 분산체계에는 많은 문제가 있다는 사실이다. 원문헌의 입수방법도 그렇고, 입수된 자료의 가공처리 과정도 그렇고, 완성된 서지레코드의 입력방법도 그렇고, 많은 문제가 내재되어 있다. 특히, 작업결과에 대한 책임을 분명히 묻지 않는 현재의 관행과 풍토에서 분산체계란 허울좋은 이론에 불과하다는 사실이 SATURN DB의 품질검증 과정에서 이미 노출되지 않았는가?

선행 연구에서 우리는 이미 현재의 분산체계를 중앙조정기관인 KORDIC에 의한 통합구축과정으로 대체하거나 혹은 레코드 제작기관들에 대한 철저한 감독 및 관리체계를 구축할 것을 제언한 바 있다 (이제환, 1997). 당시 우리는 후자의 현실성이 높다고 판단하여, DB구축기관(연구소의 자료실들)에 대한 DB관리기관(KORDIC)의 철저한 감독과 작업결과에 대한 정례적인 평가가 필요하다는 점과, DB구축기관이 제작한 레코드를 DB관리기관에서 검증하여 수정할 수 있는 제도적이고 기술적인 장치가 필요하다는 점을 역설한 바 있다. 다행히 KORDIC이 기술적 장치의 개발을 서두르고, 구축실적을 평가하여 예산을 차등 지원하는 등 부분적이나마 제도적 장치를 마련해 나가고 있으니, 그 결과는 곧 가시적으로 나타날 것이다.

그러나 앞서도 지적하였듯이, 현재와 같은 분산체계하에서는 근본적인 제도적 장치의 마련을 전제로 하지 않고서는 SATURN DB의 품질을 완벽하게 (내용적 유용성과 형태의 완전성을) 개선한다는 것은 불가능해 보인다. OCLC의 사례에서도 보듯이, 이중삼중의 품질검증을 위한 장치를 만들어 놓아도 불량품이 끼여드는 것이 현실이다. 특히, 국가차원의 주요 정보유통기관으로서의 KORDIC의 위상을 고려할 때, 조직구조의 개편을 통한 제도적 장치의 마련은 매우 시급해 보인다. 조직구조의 개편이 반드시 거창한 구조개혁을 의미하는 것은 아니다. DB의 품질관리를 전담할 부서를 설치하여 운영하거나, 그것이 어렵다면 한시적 품질관리 전담팀(task force)을 구성하여 운영하는 것도 하나의 방법이 될 수 있다. 이처럼, 이 문제의 해결은 기계적 장치에 지나치게 의존하고자 하기보다는 제도적 장치를 통한 품질관리를 'DB 관리'의 기본 원칙으로 삼겠다는 경영진의 철학과, 제도적 장치에 바탕한 엄정한 실무처리를 통해 'DB의 품격'만은 반드시 유지하겠다는 실무진의 의지에 달려있다. 이제, DB의 품질에 대한 논의를 마치면서, SATURN DB의 사례가 유사한 환경에서 유사한 작업을 계획중이거나 추진중인 여러 기관의 관계자들에게 자그마한 도움이 될 수있기를 기대해 본다.

참 고 문 헌

연구개발정보센터. 1997. 1996년도 자체평가 보고서. 대전: 연구개발정보센터.

_____. 1998. 1997년도 자체평가보고서. 대전: 연구개발정보센터.

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

Amstrong, C. 1994. "The Centre for Information Quality Management (CIQM): a single 'phone number for all your woes!." Library Technology News 12(Apr. 94): 3-5.

_____. 1996. "Database: fit for use or fit for us." Management (GB) 17(2): 40-42.

Arnold, S. 1992. "Information manufacturing: the road to database quality." Database 15(5): 32-34, 36-39.

Ayres, F. 1994. "QUALCAT: automation of quality control in cataloging." British Library. Research and Development Department BLRD Report 6068. p.112.

Ballard, T. et al. 1992. "Prediction of OPAC Spelling Errors through a Keyword Inventory." Information Technology and Libraries 11(June 92): 139-145.

Barnett, J. 1993. "OCLC cataloging peer committees: an overview." Cataloging & Classification Quarterly 16(4): 67-76.

Basch, R. 1990. "Measuring the Quality of Data: Report of the Fourth Annual SCOUG Retreat" Database Searcher 6(8).

Bland, R. et al. 1986. "Quality Control in a Shared Online Catalog Database: The Lambda Experience." Technical Services Quarterly 4(2): 43-58.

Calk, J. 1986. "Quality Control in Developing an Inter-institutional Database." Library Hi Tech 4(1): 85-90.

Chapman, A. 1994. "Up Standard - A Study of the Quality Records in a Shared Cataloging Database." Journal of Librarianship and Information Science 26(4): 201-210.

Davis, B. 1987. "Managing the Online Bibliographic Database for an Integrated Library system." Technical Service Quarterly 5(1): 49-56.

Davis C. 1989. "Result of a survey on record quality in the OCLC database." Technical Services Quarterly 7(2): 43-53.

Griffiths, J. et al. 1986. "The contribution of

online database services to the productivity of their users." In: 10th International Online Information Meeting London 2-4 December 1986, pp.66-76.

Martin, S. 1986. Library Networks, 1989-87. White Plains, NY:Knowledge Industry Publications.

Medawar, K. 1995. "Database Quality - A Literature-Review of the Past and a Plan for the Future." Program-Automated Library and Information Systems 29(3): 257-272.

Mifflin, I. et al. 1991. "Online Catalog Maintenance: The Role of Networks, Computers, and Local Institutions." Information Technology and Libraries 10(Dec. 91): 263-274.

Mintz, A. 1990. "Quality control and the Zen of database production." Online 14(6): 15-23.

OCLC. 1990-. OCLC Newsletter. Doublin, OH: OCLC.

O' Neill E. 1988. "Quality Control in Online Databases." Annual Review Information Science and Technology 23(1988): 125-156.

_____. 1989. " The Impact of Spelling Errors on Databases and Indexes." In : ed. by Carol Nixon and Lauree Padgett, National Online Meeting: Proceedings pp.313-320. Medford, NJ: learned Information.

Riner, H. 1995. "On the information highway, take your eyes off the road: evaluating database content." In: ed. by Martha E. Williams, Proceeding of the 16th National Online Meeting New York: Learned Information Inc., New Jersey: Learned Information Inc. pp.333-338.

Saylor, L. 1986. "Cooperative Cataloging Quality Control in the OCLC Pacific Network." Information Technology and Libraries 25(Sept. 1986): 235-239.

Stankowski, R. 1991. " Bibliographic Record Maintenance and Control in a Consortium Database." Cataloging & Classification Quarterly 12(2): 47-62.

Wang, R. 1995. "A Frame for Analysis of Data Quality Research." IEEE Transaction on Knowledge and Data Engineering 7(4): 623-640.

Zeng, L. 1993. "Quality control Chinese-language records a rule-based data validation system - Part 1: an evaluation of quality of Chinese-language records in the OCLC OLUC Database." Cataloging & Classification Quarterly 16(4): 25-66.