종합목록 데이터의 오류 유형에 관한 연구
- KERIS 종합목록의 학위논문 서지데이터를 중심으로 -

A Study on Error Data Types in the KERIS Union Catalog
- Focused on Dissertation Bibliographic Database -

조 순 영(Sun-Yeong Cho)*

목 차

초 록

KERIS의 대학도서관 종합목록은 330개 대학도서관이 가입하여 570만 레코드를 보유하고 있는 국내 최대의 목록이다. 그러나 각 회원기관이 이미 구축한 DB를 짧은 기간 내 통합하면서 중복 및 오류 레코드가 많이 포함되어 있기 때문에 검색의 효율성이 떨어진다는 지적이 있다. 따라서 본 연구에서는 전체 자료의 10%를 차지하는 학위논문 데이터 1000건을 대상으로 오류 데이터의 유형을 분석함으로서 데이터의 품질을 측정하고 기계적으로 오류데이터를 색출할 수 있는 방안을 제시하였다. 분석 결과 오류데이터는 전체 표본 DB의 30%를 차지하였고 발생하는 주요 원인은 1)입력오류 2)MARC의 사용 오류 3)목록규칙의 적용 오류가 가장 큰 원인으로 나타났다.

ABSTRACTS

The KERIS Union Catalog is the largest bibliographic database in Korea. It has over 5.7 million bibliographic records and at present 330 university libraries are participating in shared cataloging services. The database, however, includes various errors and a large number of duplicate data because separate databases from many university libraries were merged without proper quality control in a short period. This study intends to find error data types by analyzing dissertation bibliographic data. The results show that error data are mainly caused by input errors, and the mistakes in using MARC formats and cataloging rules.

키워드: 종합목록, 오류데이터, 중복데이터

* 한국교육학술정보원(KERIS) 학술연구정보화실장(chosy@keris.or.kr)
논문접수일자 2002년 8월 25일
게재확정일자 2002년 11월 10일

1. 서 론

정보자원이 폭발적으로 증가하면서 복잡한 MARC으로는 이제 더 이상 수용할 수 없다는 많은 연구자들의 주장에도 불구하고 도서관은 아직도 지난 30여 년간 구축한 MARC 데이터를 중심으로 종합목록을 구축하고 그 목록을 통한 자원 공유 체제는 더욱 활발해지고 있다. OCLC가 전세계 82개 국가의 4700백만 레코드를 종합목록으로 구축하여 중복적인 편목 작업을 최소화하고, 상호대차 서비스를 실시하는 것이 그 대표적인 예라 할 수 있다.

국내에서는 한국교육학술정보원(이후 KERIS)이 570만 레코드의 대학도서관 종합목록을 구축하여 2002년 현재 330개 도서관이 공동 활용하고 있다. 그러나 데이터의 표준화가 이루어지지 않은 상태에서 각 도서관이 개별적으로 구축한 DB를 단기간에 통합 구축했기 때문에 오류 및 중복데이터가 다량 발생하여 검색의 효율성을 떨어뜨리는 것이 큰 문제로 지적되고 있다. 종합목록의 역사가 상대적으로 오랜 OCLC에서는 수작업에 의한 오류 데이터의 검증 단계를 모두 생략하고 오직 프로그램만으로 데이터의 품질을 관리하고 있다. 그러나 KERIS 종합목록은 서지 데이터의 오류 유형이 다양하고 데이터의 입력 수준이 낮기 때문에 기계적인 방법으로 오류데이터를 여과할 수 있는 비율은 극히 저조하다. 따라서 현재는 기계적인 방법으로 오류 대상 레코드 그룹을 생성한 후 목록 경력자들이 육안으로 오류 여부를 식별하고 있다. 이에 본 연구에서는 오류데이터를 색출하여 그 유형을 분석하고 그 결과를 데이터 검증 프로그램에 반영함으로서 기계적인 데이터 품질 관리의 수준을 높이고자 한다.

종합목록을 구성하고 있는 서지데이터는 자료의 유형에 따라 입력 형식이 다르기 때문에 한국문헌자동화목록형식(이후 KORMARC)에서는 단행본, 연속간행물, 비도서자료, 고서 등으로 구분하여 기술하고, MARC21 Bibliographic Format의 경우 Books, Computer Files, Maps, Music, Serials, Visual Materials, Mixed Materials로 구분하여 기술한다. 그러나 실제 양적인 구성 분포를 보면 단행본 데이터가 전체의 약 91%[1]에 해당되기 때문에 단행본 데이터의 품질 수준을 측정한다면 전체 DB의 수준을 가늠하기에 충분할 것이다. 이에 본 연구에서는 단행본을 대상으로 오류 데이터 유형을 분석하고자 한다. 그러나 단행본 데이터 중에서도 종합서명이 없이 각 권 서명을 가지고 있는 다권본이라던가, 각 권 별로 소장하고 있는 총서, 번역서, 저작자의 역할 기술이 애매한 연구보고서 등은 입력자의 판단이나 도서관의 정리 지침에 따라 목록의 내용에 많은 차이가 나기 때문에 상대적인 비교가 어렵다. 따라서 본 연구에서는 자료의 출판 형식이 정형화 되어 있어 입력자의 수준이나, 도서관의 선택적 정리 지침이 크게 작용하지 않는, 즉 편목 작업이 가장 용이한 학위논문자료를 대상으로 오류 데이터를 색출하고 그 유형을 분석함으로서 모든 데이터의 기초 자료로 삼고자 한다.

---

1) 단행본 91% 시청각자료 4.3% 연속간행물 1.8% 컴퓨터화일 1.5% 지도 0.04% 연속간행물 0.6% 녹음자료 0.7 %

2. 연구방법

KERIS종합목록은 공유형 데이터베이스를 원칙으로 하기 때문에 하나의 자료에 대해서는 하나의 레코드만 존재한다. 서명 · 권차 · 저자 · 출판사 · 출판년 · 면수 · 판사항 · 총서명 · 총서번호 · ISBN, ISSN, LCCN 등의 제어번호 필드를 비교한 후 기 구축된 자료와 동일한 것으로 판명되면 기존의 레코드에 소장정보를 추기하고 질적으로 우수한 서지 레코드만 유지한다. 그러나 본 연구에서는 일반 단행본에 비해 서지 요소가 적은 학위논문을 대상으로 서명, 저자, 출판년, 출판사, 논문주기 중심으로 비교하고자 한다.

종합목록에서 학위논문 데이터는 60만건으로 전체의 10%를 차지한다. 그 중 논문 1,000건을 추출하여 무작위 저자명 순으로 정렬하고 동일 저자 · 동일 서명이 두 개 이상인 것은 오류데이터로 인한 중복 레코드로 간주하여 선별하였다. 그 결과 251개 레코드가 색출되었고 실제 101건의 오류 유형으로 구분되었다. 이들 자료를 대상으로 데이터의 입력 수준을 비교하고 전체 MARC을 대상으로 각 필드별 차이 점을 비교함으로서 대표적인 오류 유형을 분석하고 그 결과를 데이터 검증 프로그램에 반영할 수 있는 방안을 모색하고자 한다. 서지 데이터만으로 비교하기에 충분치 않은 경우는 원문 DB를 통하여 표제지를 확인하는 방법을 병행한다.

3. 데이터 입력 수준

오류 데이터의 유형을 분석하기에 앞서 학위논문 서지 형식의 일반적 특성을 살펴보고 실제 데이터의 입력 내용에서는 어떤 차이점이 있는지 비교함으로서 오류 발생의 원인을 좀 더 구체적으로 분석하고자 한다.

3. 1 학위논문의 서지적 특성

학위논문은 다른 유형의 자료에 비해 서지 기술 수준이 간단하고 예외 사항이 거의 없다. 타 자료와 비교하여 주요 특징을 살펴보면

- 출판형식이 획일화 되어 있어 서지 정보의 위치가 동일하다.
- 단일 저자이고 저자명 기입 형식이 일정하다.
  99% 이상이 내국인이기 때문에 기술상의 어려움이 없고 대부분 첫 번째 저술이기 때문에 전거 통제를 위한 별도의 확인 작업이 필요 없다.
- 서명이나 저자 사항에 한자 기입이 많다.
- 서명이 전문용어로 구성되어 있어 일반 용어 보다는 오타의 가능성이 높다.
- KS CODE에서 지원되지 않는 분자식, 학명, 기호, 부호, 특수문자 등이 많다.
- 영문 대등서명이 많다.
- 서명 내 설명을 위한 괄호사용이 많다.
- 출판년도와 학위년도가 구분되지 않고 혼용되어 있다.
- 물리적 형태사항의 구성이 거의 동일하다.
  페이지 기입의 유형이 동일하다.
  (서문 페이지가 적고 본문 페이지의 기입이 명확하다.)
  책의 크기가 동일하다

3. 2 데이터 입력 수준

학위논문의 특성상 입력 지침이 표준화 되어 있다면 MARC의 고정장 필드나 가변장 필드의 입력 수준은 거의 유사해야 한다. 그러나 〈표1〉과 같이 유사그룹의 251개 레코드를 비교한 결과 시스템에서 자동으로 생성되는 코드 값을 제외하고 100% 일치하는 값을 가지는 데이터 필드는 하나도 없는 것으로 나타났다. 코드 값 자체가 빈칸으로 값을 가지는 경우가 있기 때문에 일부 데이터는 미 입력 데이터일 수도 있지만 평균 91.3%의 데이터가 입력 수준이 완전하고, 기술형식은 90% 이상이 KORMARC 기술 규칙을 사용하는 것으로 되어 있다. 그럼에도 불구하고 입력 수준이 모두 다르다는 것은 목록 데이터의 표준화가 전혀 이루어지지 않고 있다는 것을 입증하는 것이다.

MARC 형식에서의 리더 및 고정장 필드는 제한 검색이나 통계에서 아주 유용하게 사용되는 부호이다. 그러나 위와 같이 MARC 데이터의 표준이 없는 상태에서 정확한 통계는 거의 기대할 수 없다. 출판연도의 제한 검색 등 중요한 키 값이 되는 발행년이나, 학위논문의 대학별 제한 검색에 활용할 수 있는 대학부호 등의 값도 모든 데이터가 다 가지고 있지 못하기 때문에 제한 검색시 누락되는 데이터가 발생하게 된다. 기타 발행년2나 대상자수준, 연관레코드, 소설 등과 같이 불필요한 코드 값을 가지고 있는 것은 사실상 모두 단순 오류입력에 의한 것으로 확인되었다. 따라서 단순 입력에 의한 오류 데이터의 최소화를 위해서 자료 유형별 최소 입력 수준을 지정하는 방안도 함께 검토하는 것이 바람직하다.

리더 및 고정장 필드는 입력자들이 그 필요성에 대해서 절실하게 느끼지 않거나 코드값 입력을 번거롭게 느껴서 또는 시스템에서 자동 생성되는 것으로 오인하여 데이터를 누락하는 사례가 많다. 그러나 가변장 데이터 필드는 전적으로 사서들의 입력에 의한 부분으로 편목의 수준과 입력 수준을 측정 할 수 있는 기준이다.

〈표 2〉는 251개 레코드를 대상으로 한번 이상 사용된 모든 가변장 데이터 필드를 정리한 것이다. 100%의 입력율을 보이고 있는 것은 만일 누락되면 업로드가 불가능한 표시기호 245와 학위논문 제한 검색시 조건으로 사용되는 표시기호 502 뿐이다. 실제 학위논문이면서

〈표 1〉 리더 및 고정장 데이터 필드의 기입 현황

표시기호	건수	%	표시기호	건수	%
입력수준	251(22)	100(8.7)%	대학부호	195	77.6%
기술형식	230	91.6%	회의간행물	238	94.8%
연관레코드	9	3.5%	기념논문집	239	95.2%
발행년1	247	98.4%	색인	232	92.4%
발행년2	9	3.5%	소설	16	6.3%
발행국명	244	97.2%	목록전거	189	75.2%
대상자수준	8	3.1%	언어부호	249	99.2%

도 502 필드의 미기입으로 누락된 자료도 발생할 수 있지만 여기서는 파악이 불가능하여 100% 기입된 자료만을 대상으로 한 것이다.

학위논문에서는 불필요한 010 미의회도서관 제어번호, 020 국제표준도서번호, 240 통일서명, 440 총서명 필드를 사용한 레코드가 있을 뿐 아니라 KORMARC에서는 정의되지 않은 246, 502 $e, 700 $h 등의 필드도 사용되고 있다.

반면 이용자나 목록자에게 필요한 주제명이나 분류기호 등의 기입은 누락된 것이 훨씬 많다. 우리나라 대학도서관의 대부분이 DDC와 KDC를 사용하고 있음에도 불구하고 실제 KDC를 기입하는 056 필드와 DDC를 기입하는 082 필드의 기입율은 30%에도 미치지 못한다. 대부분 자관 청구기호 필드의 $a를 분류기호 기입으로 대체하는 경우가 있으나, 도서관마다 분류표를 한국실정에 맞도록 임의 전개한 부분들이 있어 사실상 자관 기호를 표준으로 사용할 수는 없다. 특별히 미국 중심으로 되어 있는 DDC의 경우 종교, 문학, 예술, 역사, 지리, 법률 분야에 대해 우리 실정에 맞도록 변경하여 사용하는 기관이 많이 있다. 한국 분야 이외에도 학문의 발전 속도가 빨라서 지속적인 전개가 필요한 분야에 대해서는 추가 전개한 기관이 많이 있기 때문에 원래 분류표에서 어느 항목에 해당하는지에 대한 정보 없이 그대로 사용하는 경우 타 기관에서 목록을 다운로드 받을 때 혼란을 초래 할 수도 있다. 미국에서는 주제 접근시 검색어로 주제명 표목을 많이 사용하지만 국내에서는 사용하는 기관이 거의 없었고 653 비통제 주제명만 일부 도서관에서 입력하고 있었다.

저자 사항은 기본 표목의 채택 여부에 따라 표시기호 100과 700으로 나누어 사용하고 있다. 100필드를 사용하지 않는 기관이 전체의 17%에 이르렀다. 저자명 기입에서는 지시기호의 부적절한 사용으로 색인 생성의 오류가 발생하는 예도 있었다. 예를 들어 지시기호 없이 괄호 안에 한자 표기나 저자의 생몰년 등을 표기하여 잘못 색인이 만들어짐으로서 저자명의 완전일치 검색에서 누락되는 결과를 초래하였다. 또한 저자명 전거화일 구축을 대비하여 식별기호 $d 등을 사용한 예는 3%에 지나지 않았다. 전체적으로 저자, 서명, 발행사항, 형태사항, 논문 주기 이외 사항은 거의 기입하지 않았고 일부 기입된 데이터는 대부분 오류 데이터이다.

4. 오류레코드 분석

251개 레코드를 유사 데이터끼리 그룹핑한 결과 99개의 그룹이 형성되었다. 즉 1 그룹당 평균 2.5개의 레코드가 중복레코드 임에도 불구하고 데이터의 오류로 인해 신규 레코드로 처리가 된 것이다.

오류 데이터를 유형별로 분석한 결과 주요 원인은 1) 입력 오류 2) MARC의 사용 오류 3) 목록규칙의 적용 오류 등으로 나타났다.

4. 1 입력 오류 데이터

입력 오류는 데이터 자체가 길고 복잡한 서명에서 가장 많이 나타나고, 그 다음은 한자 사용이 많은 저자 필드이다. 특별히 서명에서

〈표 2〉 가변장 데이터 필드의 기입 현황

표시기호	지시/식별기호	건수	%	표시기호	지시/식별기호	건수	%
010	a	2	0.7	300	a	242	96.4
012	a	38	15.1		b	129	51.3
020	지시1	4	1.5		c	231	92.0
	a	8	3.1	440	지시1	1	0.3
	c	14	5.5		지시2	1	0.3
041	지시1	78	31.0		a	1	0.3
	a	86	34.2		v	1	0.3
	b	76	30.2	500	지시1	16	6.3
045	a	1	0.3		지시2	16	6.3
052	지시1	32	12.7		a	1	0.3
	지시2	32	12.7		m	3	1.1
	a	32	12.7		z	10	3.9
	b	32	12.7		j	3	1.1
056	a	72	28.6	502	지시1	239	95.2
	2	31	12.3		지시2	1	0.3
082	지시1	6	2.3		a	251	100.0
	지시2	6	2.3		b	242	96.4
	a	73	29.0		c	199	79.2
	2	55	21.9		d	233	92.8
085	a	19	7.5		e	4	1.5
	2	17	6.7	504	a	46	18.3
	e	1	0.3	505	a	2	0.7
100	지시1	209	83.2	520	b	13	5.1
	지시2	4	1.5	600	지시1	3	1.1
	a	209	83.2		지시2	3	1.1
	d	8	3.1		a	3	1.1
240	지시1	2	0.7		d	2	0.7
	지시2	2	0.7	650	지시2	6	2.3
	a	2	0.7		a	6	2.3
	l	2	0.7		x	1	0.3
245	지시1	251	100.0	653	a	86	34.2
	지시2	246	98.0	700	지시1	44	17.5
	a	251	100.0		지시2	1	0.3
	b	14	5.5		a	44	17.5
	c	35	13.9		e	2	0.7
	x	68	27.0	710	a	2	0.7
	d	228	90.8		b	2	0.7
	n	7	2.7	740	지시1	4	1.5
246	지시1	2	0.7		지시2	32	12.7
	지시2	3	1.1		a	37	14.7
	a	3	1.1	890	b	3	1.1
260	지시1	6	2.3	900	지시1	4	1.5
	a	249	99.2		a	4	1.5
	b	248	98.8		d	2	0.7
	c	249	99.2	940	지시1	4	1.5

의 입력 오류는 전체 오류 레코드의 95% 이상을 차지한다. 입력 오류로 인한 데이터 유형은 〈표 3〉과 같이 크게 동자이음어 중 잘못된 음으로 입력한 예, 한자의 오독으로 인한 예, KSCODE에서 지원 안되는 문자의 입력, 단순 입력오류로 구분할 수 있다.

한자 "見"이 "현"과 "견"으로 모두 발음되면서 한자로 표기되는 경우 또는 두음법칙 적용문자처럼 표준 한자 코드 내에서 복수의 음을 갖는 한자, "數字"처럼 사이 시옷이 첨부되는 경우, 표준 코드에 해당 한자음이 없는 경우, "五六月"처럼 다른 음을 빌려온 경우 등 한자와 한글을 1:1로 변환하면서 발생하는 문제는 한자 표기가 많은 학위 논문에서도 그대로 나타나고 있다. 245필드에서 동일한 한자 표기를 가지고 있지만 입력상의 오류로 색인에서는 다른 서명으로 만들어진 경우가 전체 오류데이터의 2%에 해당한다. 이 같은 문제는 시스템에서 동자이음어 사전으로 이중색인을 생성하는 방법이나 서명의 유사 값 측정을 통하여 최소화시킬 수 있지만 무엇보다도 정확한 입력이 가장 최선의 방법이다.

한자의 오독으로 인한 오류데이터 비율은 전체 오류데이터의 13%를 차지한다. 예를 들어 "癩病의 大食細胞리소짐에 對한 免疫組織化學的 硏究"라는 서명을 가진 레코드 3개중 두 개의 레코드에서 "癩病" 대신 "羅病" "拏病" 등으로 한자를 표기하였고 이 같은 사례는 주로 과학 분야에서 많이 발견되었다. 목록 시스템에서 입력자 들의 편의를 위한 한자사전 등을 보강하여 사용한다면 오류율은 크게 줄일 수 있을 것이다.

학위논문의 서명에는 KSCODE 5601에서 지원되지 않는 특수문자, 한자, 비로마자, 음독부호 등이 많이 포함되어 있다. 이들 문자는 KORMARC의 890 필드에 기록이 되거나 별도 관리가 되어야 함에도 불구하고 도서관마다 각기 다른 원칙을 적용하였거나 아무 처리 없이 입력에서 제외시킨 사례가 많이 나타났다. 일부 국립대학 도서관을 중심으로 가장 많이 사용되는 독일어의 우물라우트를 비롯하여 자주 출현하는 음독 부호에 대해서는 아래와 같이 서명 중에 포함 될 확률이 없는 1 byte 문자로 대체 입력하고 있다.

ü ⇒ u@(독어) β ⇒ ss ä ⇒ a@(중국어)
á ⇒ a〉 à ⇒ a〈 û ⇒ u^
Ñ ⇒ N* ê ⇒ e^ ç ⇒ c*

향후 UNICODE로의 변환에 대비하여 이들 문자에 대한 표준 입력지침을 제정하여 도서관내에서 공통적으로 적용해야만 데이터의 기계적 처리가 용이할것이다.

단순 입력의 오류로 인한 데이터는 주로 데

〈표 3〉 입력 오류 유형

1	동자이음어 중 잘못된 음으로 입력	2%
2	한자의 오독으로 인한 입력 오류	13%
3	KS CODE에서 지원 안되는 문자의 입력	2%
4	단순 입력 오류	19%

이터가 긴 서명에서 발견되는데 가장 많은 경우가 서명 중의 한 음절이나 단어를 빠뜨리거나 잘못 입력한 것이다. 이와 같이 오류레코드 중 단순한 입력에 의한 오류는 전체 오류데이터의 19%를 차지한다. 아래와 같이 251개 레코드에서 발견된 주요 사례를 보면 유사한 한자의 읽기 오류에 의한 것과 입력자의 부주의에 의한 것이 대부분이다.

결함 -> 결핍 숙성 -> 숙주 출혈 -> 실혈
공업 -> 기업 연소 -> 연쇄 토론 -> 토의
무풍 -> 무례 유지 -> 유대 투익 -> 손익
선방 -> 전방 을 -> 를 학교 -> 학급
수험 -> 수검 종합 -> 중합 혁신 -> 혁명...

4. 2 MARC의 사용 오류 데이터

〈표 4〉에서와 같이 오류데이터의 많은 부분이 MARC의 잘못된 이해에 원인이 있었다. MARC에서의 지시기호나 식별기호의 모든 값이 의미를 가지는 것임에도 불구하고 상당수의 데이터에서 기호를 사용하지 않은 채 전방의 식별기호에 이어서 사용하고 있다. 대표적인 사례로는 245 서명 필드에서 부서명에 $b를 사용하지 않고 구독점 ":"만 사용한 채 $a에 이어서 기술한 경우와 $c의 잡제와는 전혀 구별하지 않고 사용한 예가 전체의 18%였다. $x의 대등서명도 245 필드에 기입하지 않고 740 필드나 240 필드, 940 필드 등 다양한 필드에 독립적으로 기입하고 있다.

100 1 ▼a권순긍 (權純肯)
245 10 ▼a1910년대 活字本 古小說 硏究: ▼c그 改作.新作의 歷史的 性格= ▼x(A)study on Hwalzabon classical novel of 1910s/ ▼d權純肯
100 0 ▼a權純肯
245 10 ▼a1910년대 活字本 古小說 硏究 : 그 改作·新作의 歷史的 性格 ▼d權純肯

또한 저자명의 경우 $a의 한글표기 다음에 아무 식별기호 없이 원괄호 안에 기입한 사례가 8%나 발생하였다. 이들 데이터는 시스템에서 한자의 자동 한글 색인 생성에 의하여 동일인명이 두 번 반복되어 만들어지는 결과를 초래하게 하였다. 저자명의 경우 전거 파일에서 한자의 읽기를 입력하여 처리를 해야 함에도 불구하고 전거 파일의 미비 또는 MARC에 대한 기본적 이해 부족 등으로 인한 오류 레코드가 생성되었다.

영미권에서는 학위논문이 별도의 자료로서 다루어지지 않기 때문에 MARC21 Format for Bibliograhic Data의 tag 502에서는 석박사를 구분하는 지시기호의 사용이 없고 학위수여기관이나 학과 및 전공, 학위년도 등을 기술하는 식별기호도 사용하지 않는다. 또한 원

〈표 4〉 MARC의 사용 오류 유형

1	데이터필드의 생략 (예: 100 $c, 245 $b,c,)	18%
2	데이터필드 사용의 오류 (예: 502 $b, c)	3%
3	MARC21과 KORMARC의 혼용	25%

래 학위논문이 아닌 변형된 편집 또는 출판 형식의 논문은 tag자체도 500 일반 주기를 사용하도록 정의되어 있다. 그러나 KORMARC에서는 표시기호 502 학위논문 필드의 유무로 학위논문을 제한 검색하는 주요 필드로 사용하고 있다.

502 필드에서 $b는 학위를 수여하는 대학명이 기술되어야 하고 그 대학명은 검색키로도 활용될 수 있도록 전거 통제를 받은 표기로 사용하는 것이 바람직하나 실제 60%의 데이터는 xx대학교 대학원까지 기술하고 40%는 xx대학원으로 기술하고 있다. 또한 $c의 학과 및 전공은 대학마다 학과명과 구분 체제가 다르고 학과명이 변경된 사례도 많아서 석사 논문과 박사 논문의 학과명이 다르게 기술되는 경우도 많다. 따라서 전혀 통제되지 않은 형태이기 때문에 유사한 학과와 전공들이 혼용되고 있다. 더욱이 502 필드 자체가 전거통제를 받도록 정의되어 있지 않기 때문에 논문에 표기된 그대로 기술하는 경우에도 전공을 생략하는 사례가 25% 이상 발견되었다. 따라서 기계적으로 전공명까지 일치 여부를 확인하려면 완전일치 방식으로는 불가능하고 전방일치 방식으로만 가능하다. 또한 502 필드가 표시상수로 정의되고 있지 않음에도 불구하고 시스템에서 자체적으로 표출어를 생성하기 때문에 MARC 데이터 자체에 "학위논문"이라는 표출어를 생략한 레코드도 5%를 차지하였다. 이와 같은 경우 데이터 교환시 호환성의 문제가 있기 때문에 표시상수로서의 임의 지정은 바람직하지 않다. 아래의 예는 동일한 자료이지만 학위논문 주기사항이 다르게 기입된 대표적인 경우이다.

502 1 ▼a학위논문(박사) ▼b慶北大學校 大學院: ▼c電子工學科 回路 系統工學 專攻, ▼d1992

502 1 ▼a경북대학교 대학원 : ▼c회로계통공학전공, ▼d1991.

502 1 ▼a학위논문(박사) - ▼b大邱大學校 大學院 : ▼c地域社會開發學科 國際開發專攻, ▼d1988

502 1 ▼a학위논문(박사) ▼b대구대학교 대학원, ▼d1988

502 ▼a학위논문(박사) ▼b대구대학교 대학원: ▼c지역사회개발학과, d1988

502 1 ▼a학위논문(박사) ▼b대구대학교: ▼c지역사회개발학과, d1988

4. 3 목록규칙의 적용 오류 데이터

영미목록규칙에 비해 상대적으로 목록자의 판단이 요구되는 부분이 많은 한국목록규칙의 수준을 고려한다면 서지 데이터의 표준화가 이루어지지 않는 것이 어쩌면 당연한 결과일 것이다. 따라서 대부분의 도서관에서는 특별한 목록 규칙을 적용하지 않고 단순히 서지정보를 기술하는 수준에 머물거나 영미목록규칙을 국내서에 적용하는 사례가 적지 않다. 따라서 오류 데이터의 상당 부분은 〈표 5〉에서와 같이 목록 규칙의 이해부족으로 인한 것이었다.

연감이나 연보에서 내용에 대한 연도와 실제 출판년도가 차이가 나듯이 학위논문에서도 학위 수여년도와 발행년도는 다르다. 따라서 필드 260 $c에는 출판년도를 기입해야 하고 필드 502 $c에는 학위를 수여 받은 연도를

〈표 5〉 목록규칙의 적용 오류 유형

1	출판년도 (학위수여연도와 혼동)	42%
2	한자의 한글 변환 기입	38%
3	서명 중의 괄호안에( ) 원어명 등이 표기된 경우 생략	3%
4	분자식, 원소기호 등에서 첨자의 입력 오류	1%
5	면수 입력 (빈칸 페이지 등 임의 적용)	78%

기입해야 한다. 대부분 학위 수여 이전에 출판을 하기 때문에 1년이 차이가 나는 경우가 많다. 실제 동일한 자료 임에도 불구하고 발행년도가 다르게 기입된 예가 전체 오류데이터의 42%에 해당한다.

통제를 받는 필드 이외에는 자료의 표제지에 나타난 정보를 그대로 기술해야 함에도 불구하고 한자 서명을 한글로 변환하여 기입한 예가 전체의 38%에 이르렀다. 한자의 난이도가 높은 서명이 대부분이었다. 또한 서명의 특징으로 영문 대등서명을 사용하는 예가 전체의 90%이나 실제 기입을 한 경우는 27%에 불과하다. 대등서명을 통일서명이나 읽기 서명 등으로 잘못 기입한 예도 3%를 차지한다. 한편 필드 100부터 6XX까지의 데이터 필드에는 기입하지 않고 740 서명 부출 필드에만 기입한 사례도 7%에 해당한다. 그 외에도 괄호 안에 표기된 설명구를 생략하고 기술한 예와 띄어쓰기를 전혀 고려하지 않고 입력한 사례가 빈번하였다.

발행사항의 경우 시 단위를 원칙으로 하고 확인되지 않은 경우 괄호 안에 표기해야 하나 아래의 예와 같이 목록자의 상식 수준에서 임의로 기입하는 예가 많다.

260 ▼a하양 : ▼b曉星女子大學校, ▼c1993.

260 ▼a경산 : ▼b효성여자대학교, ▼c1993

260 ▼a大邱 : ▼b曉星女子大學校 大學院, ▼c1993.

자료의 형태 사항은 중복 여부를 가리는 결정적인 요소 임에도 불구하고 목록자들에게는 그 중요성을 거의 인정받지 못하는 필드이다. 실제 자료에 면수 표시가 되어 있지 않아 목록자가 임의로 세어 쓸 경우 괄호 안에 기입해야 하나 괄호를 생략한 경우, 면수를 세지 않고 간단히 1 v.으로 표시하는 경우, 장, 면, p등 단위에 대한 표준 없이 다양하게 사용한 경우, 또한 단면 인쇄본은 면수 표시가 된 것을 그대로 기입한 후 double leaves등의 설명을 해주어야 하나 임의로 두 배를 계산하여 기입한 사례 등 실제 동일 자료에 대한 레코드의 78%가 면수 표기를 다르게 하고 있다.

300 ▼a107p. ; ▼c26 cm.

300 ▼aix, 107장 : ▼b삽도 ; ▼c26 cm

300 ▼axvi,167 L.: ▼bill.plate ; ▼c26 cm.

300 ▼axv, 162 leaves: ▼bill.; ▼c30 cm. (실제 동일한 크기의 책)

300 ▼a1 책 : ▼b삽도 ; ▼c26 cm

300 ▼a36 p; ▼c26 cm

300 ▼a37 p: ▼b도표 ; ▼c26 cm

논문주기에서의 표출어는 지시기호에 의해 학위논문(박사), 학위논문(석사)로 시스템에서 자동생성하게 되어 있음에도 불구하고 임의로 "박사학위논문" 등과 같이 다른 표출어를 입력한 경우도 4%에 달한다.

4. 4 오류 데이터의 실례

〈그림 1〉은 동일 자료에 대해 각기 다른 도서관에서 작성한 서지레코드로서 복합적인 오류 유형을 모두 포함한 대표적 사례이다. 표제지는 간단한 정보를 제공하고 있으나 입력오류, MARC의 잘못된 적용 등으로 아래와 같이 8개의 레코드가 존재하는 경우이다.

[레코드 1]에서는 서명을 한글 표기로 채택하고 영문을 대등서명으로 처리하면서 245 $c의 저자명은 표제지가 아닌 다른 부분에서 발췌하여 영문 표기를 한 사례이다. 또한 학위논문의 특성상 주요 항목으로 사용되는 학위주기 부분에 학위를 구별하는 지시기호 및 전공분야 등의 정보를 기입하지 않고 도리어 발행기관에 학과 및 전공명을 기입하고 있다. 그 외에도 주제명 등의 입력 오타로 인하여 잘못된 색인이 생성될 우려가 있다.

[레코드 1]

035 ▼aKRIC01008174

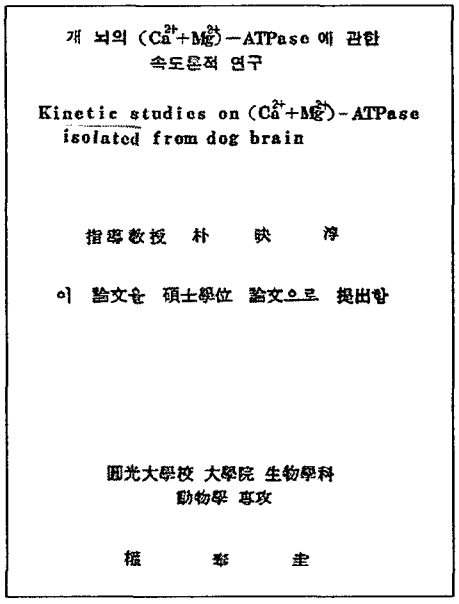
082 ▼a591.18

100 1 ▼a권봉규

245 10 ▼a개뇌의 (Ca^{2+} + Mg^{2+})-ATPase에 관한 속도론적 연구 = ▼bKinetic studies on (Ca^{2+} + Mg^{2+})-ATPase isolated from dog brain / ▼cby Bong-Kyu Kwon.

246 11 ▼aKinetic studies on (Ca^{2+} + Mg^{2+})-ATPase isolated from dog brain

260 ▼a이리 : ▼b원광대학교 대학원 생물학과


개 뇌의 (Ca^{2+}+Mg^{2+})-ATPase 에 관한
속도론적 연구

Kinetic studies on (Ca^{2+}+Mg^{2+})-ATPase
isolated from dog brain

指導敎授 朴 映 淳

이 論文을 碩士學位 論文으로 提出함

圓光大學校 大學院 生物學科
動物學 專攻

權 奉 圭

〈그림 1〉 학위논문 표제지 사례 1

동물학전공, ▼c1983.

300 ▼a36 p.

502 ▼a학위논문 (석사) - ▼b원광대학교.

504 ▼aBibliography: p. 31-35.

653 ▼a뇌 ▼aca2+ ▼amg2+ ▼aatpase ▼a속도론 ▼aknetic studies ▼abrain

740 0 ▼aKinetic studies on (Ca^{2+} + Mg^{2+})-ATPase isolated from dog brain

[레코드 2]는 대등서명을 생략한 사례이고 100 필드에서 저자명의 한자형을 그대로 기입한 사례이다. 석사와 박사를 구별하는 지시기호를 기입하지 않아 학위별 제한 검색이나 통계산출에 문제가 있다.

[레코드 2]

035 ▼aKRIC03206610

041 0 ▼akor ▼beng

056 ▼a499.746

100 1 ▼a權奉圭

245 10 ▼a개 뇌의 (Ca^{2+}+Mg^{2+})-Atpase에 관한 속도론적 연구 / ▼d權奉圭

260 ▼a이리: ▼b圓光大學校, ▼c1983

300 ▼a36장; ▼c26cm

502 ▼a학위논문(석사) ▼b圓光大學校 大學院: ▼c生物學科 動物學 專攻, ▼d1983

653 ▼a개 ▼a뇌 ▼aCA2MG2ATPASE ▼a속도론 aAPTASE

[레코드 3]는 표제지에 있음에도 불구하고 대등서명을 생략한 예이다. 502 학위 주기에서도 학위 수여기관 및 전공명에 대한 구체적 기입을 하지 않아 전공별 논문 검색시 누락할 우려가 있다.

[레코드 3]

035 ▼aKRIC04176327

056 ▼a499.74604 ▼23

100 1 ▼a권봉규

245 10 ▼a개뇌의 (Ca^{2+} Mg^{2+})-ATPase에 관한 속도론적 연구 / d 權奉圭

260 ▼a이리: ▼b원광대학교, ▼c1983

300 ▼a36p.; ▼c26cm

502 0 ▼a학위논문(석사) ▼b원광대학교, ▼d1983

[레코드 4]는 서명에서 기호 및 저자명, 학위명, 발행지명 등 여러 필드에서의 오기입이 발생한 사례이다. 또한 석박사 논문을 식별하는 지시기호도 생략되어 있다.

[레코드 4]

035 ▼aKRIC04416747

100 10 ▼a권봉규

245 10 ▼a개 뇌의 (Ca^{2} +Mg^{2})-ATPase에 관한 속도론적 연구 = ▼xKinetic studies on (Ca^{2} +Mg^{2})-ATPase isolatde from dog brain / ▼cby 權奉圭

260 0 ▼a익산 : ▼b圓光大學校, ▼c1983.

300 ▼a36 p. : ▼bill. ; ▼c26 cm.

502 ▼aThesis(Ph. D.) ▼b圓光大學校 大學院: ▼c生物學科(動物學專攻), ▼d1983.

[레코드 5]는 서명에서 입력이 난이한 단어와 대등서명, 학위논문의 전공 분야를 모두 생략하고 주제명에서도 철자의 오기입이 발생하였다.

[레코드 5]

012 ▼aKDM199114946
035 ▼aKRIC04966803
041 0 ▼akor ▼beng
056 ▼a527.41 ▼23
100 1 ▼a권봉규
245 10 ▼a개 뇌의 APTase에 관한 속도론적 연구 / d權奉圭.
260 ▼a이리 : ▼b원광대학교, ▼c1983.
300 ▼a36장 ; ▼c26 cm.
502 0 ▼a학위논문(석사) ▼b원광대학교, ▼d1983
653 ▼a개뇌 ▼a속도론적 ▼aAptase

[레코드 6]은 쓰여진 번역서가 아님에도 불구하고 영문 번역서로 처리하면서 서명에서는 대등서명도 생략하였다. 또한 서명의 철자와 발행지가 잘못 기입되었다.

[레코드 6]

035 ▼aKRIC05154554
041 1 ▼aeng ▼bkor
056 ▼a495.13
100 1 ▼a권봉규
245 10 ▼a게 뇌의 (Ca^{2+}+ Mg^{2+})-ATPase에 관한 속도론적 연구 / ▼d權奉圭
260 ▼a전주: ▼b圓光大學校, ▼c1983
300 ▼a36p.: ▼b삽도,악보; ▼c26cm
502 0 ▼a학위논문(석사) ▼b圓光大學校 大學院: ▼c生物學科 動物學專攻, ▼d1983

[레코드 7]은 041필드에서 한글이 주 언어로 기술된 것을 표기하면서 국문 서명을 본서명으로 채택하지 않았을 뿐 아니라 국문서명을 500 10의 $z 정의되지 않은 표출어를 사용한 것은 잘못된 기입이다. 또한 학과 및 전공사항이 생략되었다.

[레코드 7]

035 ▼aKRIC05977399
041 0 ▼akor ▼beng
100 1 ▼a권봉규
245 10 ▼aKinetic studies on (Ca^{2+}+ Mg^{2+})-ATPase isolated from dog brain / ▼d權奉圭 著.
260 ▼a이리 : ▼b圓光大學校 大學院, ▼c1983.
300 ▼a36 p. : ▼b圖 ; ▼c26 cm.
500 10 ▼az국문서명: 개 뇌의 (Ca^{2+}+ Mg^{2+})-ATPase에 관한 속도론적 연구.
502 0 ▼a학위논문(석사) ▼b圓光大學校, ▼d1983.
504 ▼a참고문헌 : p.31-35.

[레코드 8]은 국문 서명이 본서명임에도 불구하고 영문서명을 채택하고 국문서명에 대해서는 대등서명이나 기타부출 서명으로 전혀 기입하지 않고 있어 서명으로는 찾을 수가 없다. 또한 영문 서명 내에서도 철자 오기입이 여러 군데 나타나고 있다.

[레코드 8]

035 ▼aKRIC06998545
100 1 ▼a권봉규.
245 10 ▼aKinetic studies on (Ca+Mg)-ATPase insolaned from brein / ▼d權奉圭 著.
260 ▼a이리 : ▼b원광대학교, ▼c1983.
300 ▼a36 p. : ▼b插圖 ; ▼c26 cm.

502 0 ▼a학위논문(碩士) ▼b元光大學校 大學院: ▼c생물학과 동물과 전공, ▼d1983.

5. 결 론

이상에서와 같이 학위논문은 서지 형식이 가장 단순한 자료 임에도 불구하고 오류의 유형은 다양하게 나타나고 있다. 그 원인은 크게 세가지로 구분될 수 있으나 대부분 목록자의 부주의에 의한 것이 많고 일정한 규칙이 없기 때문에 기계적으로 오류율을 줄이는 방법에는 한계가 있을 것이다. 그러나 오류유형을 통계적으로 분석하고 그 결과를 아래와 같이 필드별 비교 조건에 반영한다면 보다 많은 양의 오류 레코드를 기계적으로 색출할 수 있을 것이다.

① 데이터의 입력 수준이 미약해서 OCLC에서와 같이 고정장과 가변장에서의 관련 된 코드 값을 상호 비교하는 방식은 적절치 않고 상대적으로 입력자가 비중을 두고 있는 가변장 데이터 중심으로 비교한다.

② 서명 비교시 전체를 대상으로 완전일치 방식으로 비교하지 않고, 빈칸까지 모두 붙여서 1-5-10-20과 같이 건너뛰기 방식으로 일치 여부를 비교한다. 비교 범위는 부서명, 잡제 등을 모두 포함하되 생략한 경우에 대비해 앞에서 40 byte까지만 비교한다. 학위논문은 서명의 길이가 40 byte 이상인 경우가 80%를 차지하기 때문에 20문자까지 비교하는 것이 적절하다. OCLC에서는 3-2-2-1방식으로 각 단어의 두문자를 비교하지만 한글의 성격상 띄어쓰기가 일정치 않아 단어 단위의 비교는 적합치 않다.

③ 서명 비교시 괄호 안의 내용을 모두 생략하여 비교한다.

④ 모든 한자는 한글로 변환하여 비교하되 동자 이음의 경우 다른 음에 대한 색인을 동시에 생성하여 비교한다.

⑤ 출판년도의 비교는 학위 수여년도와의 혼용을 고려해 -1<year<1범위는 모두 동일한 것으로 인정한다.

⑥ 형태사항의 비교는 P.면,장 등 단위를 표기하는 문자는 비교 요소에서 제외하고 $a의 면수 표기 중 가장 큰 숫자 데이터만 비교한다.

⑦ 학위논문 주기에서는 $b의 대학명 중 앞에서 6자리만 비교하여 대학원을 생략하거나 학과명 기입의 오차로 누락되는 사례를 줄인다.

⑧ 모든 데이터 필드의 비교는 완전일치 방식보다는 유사 값을 부여하여 일정한 값 이상의 경우 오류 및 중복 데이터로 분류하고, 각 필드별 가중치를 다르게 부여한다.

참 고 문 헌

『2001년도 종합목록서비스 운영현황보고서』. 서울: 한국교육학술정보원, 2001

국립중앙도서관 편. 1993. 『한국문헌자동화목록법형식: 단행본용』. 서울: 국립중앙도서관, 1993

김지훈. 서지데이터베이스의 품질관리-k관의 MARC 레코드 분석을 중심으로 『도서관학논집』, 제21집(1994): 401-429

이지은. 1999. 『첨단학술정보센터 종합목록데이터베이스 품질관리에 관한 연구』. 석사학위논문. 숙명여자대학교 대학원. 문헌정보학과.

OCLC. 1990. Bibliographic input standards. 4th ed. Dublin, Ohio: OCLC.

Cousins, S. A. "Duplicate detection and record consolidation in large bibliographic databases: the COPAC database experience in Great Britain." *Journal of Information Science"* v. 24 no4 (1998) p. 231-40

Determining Duplicate Records: Kinetica Duplicates Guidelines for Monographs. 〈http://www.nla.gov.au/kinetica/det duplicates.html〉

OCLC. Bibliographic Formats and Standards: Chapter 4. When to Input a New Record. 〈http://www.oclc.org/oclc/bib/chap4.htm.〉

_____. Bibliographic Formats and Standards: Chapter 5. Quality Assurance / OCLC 〈http://www.oclc.org/oclc/bib/chap5.htm〉

_____. OCLC Cataloging Service User Guide. 3 ed. 〈http://www.oclc.org/oclc/cataloging/guide/frameset.htm.〉

O'Neill. E. A. "Duplicate detection". *Annual Review of OCLC Research.* (1988/89): 15-16, (1989/90): 13-14

_____, Sally A. Rogers, and W. Michael Oskins. "Characteristics of duplicate records in OCLC's online union catalog. LRTS, 37(1): 59-67

Ridley, M. J. "An Expert system for quality control and duplicate detection in bibliographic databases". *Program* 26(1) 1992: 1-18

Rittberger, M. Rittbeger, W. "Measuring quality in the production of databases". *Journal of Information Science,* 23(1): 25-37