Journal of Korean Library and Information Science Society
Vol. 55, No. 1

공공도서관 목록데이터의 중복검증에 관한 연구
- 부산 지역 G도서관 사례를 중심으로 -

A Study on Duplication Verification of Public Library Catalog Data: Focusing on the Case of G Library in Busan

송 민 건(Min-geon Song)*
이 수 상(Soo-Sang Lee)**

〈 목 차 〉

요 약 : 본 논문은 아이템 기반으로 작성된 공공도서관의 목록데이터에 대해 중복검증 알고리즘을 적용하여 서지레코드의 통합방안을 도출하고자 하였다. 이를 위하여 부산 지역에서 비교적 최근에 개관한 G도서관을 선정하였다. G도서관의 OPAC 데이터를 웹 크롤링을 통해 수집한 다음, 한국문학(KDC 800) 다권본 도서를 선별하고 KERIS의 중복검증 알고리즘을 적용하였다. 검증 결과를 바탕으로 2차에 걸친 데이터 교정 작업을 진행한 이후, 중복검증률은 95.53%에서 98.27%로 총 2.74% 상승하였다. 데이터 교정 후에도 유사/불일치 판정을 받은 24권은 개정판, 양장본 등 별도의 ISBN을 부여받고 출판된 다른 판본의 자료로 확인되었다. 이를 통해 목록데이터 교정 작업을 통해 중복검증률의 개선이 가능함을 확인하였으며, 공공도서관의 중복된 아이템 레코드들을 구현형 레코드로 전환하기 위한 도구로서 KERIS 중복검증 알고리즘의 활용 가능성을 확인하였다.

주제어 : 공공도서관, 목록데이터, 중복검증, 다권본, 통합도서관

ABSTRACT : The purpose of this study is to derive an integration plan for bibliographic records by applying a duplicate verification algorithm to the item-based catalog in public libraries. To this, G Library, which was opened recently in Busan, was selected. After collecting OPAC data from G Library through web crawling, multipart monographs of Korean Literature (KDC 800) were selected and KERIS duplicate verification algorithm was applied. After two rounds of data correction based on the verification results, the duplicate verification rate increased by a total of 2.74% from 95.53% to 98.27%. Even after data correction, 24 books that were judged to be similar or inconsistent were identified as data from other published editions after receiving separate ISBN such as revised versions or hard copies. Through this, it was confirmed that the duplicate verification rate could be improved through catalog data correction work, and the possibility of using the KERIS duplicate verification algorithm as a tool to convert duplicate item-based records from public libraries into manifestation-based records was confirmed.

KEYWORDS : Public Library, Catalog Data, Duplicate Verification, Multipart Monograph, Integrated Library

* 부산대학교 문헌정보학과 박사과정(mgs207@pusan.ac.kr / ISNI 0000 0005 1420 3658) (제1저자)
** 부산대학교 문헌정보학과 교수(sslee@pusan.ac.kr / ISNI 0000 0000 6434 9851) (교신저자)

• 논문접수: 2024년 2월 26일 • 최초심사: 2024년 3월 7일 • 게재확정: 2024년 3월 20일
• 한국도서관·정보학회지, 55(1), 1-26, 2024. http://dx.doi.org/10.16981/kliss.55.1.202403.1

Ⅰ. 서 론

1. 연구의 배경과 목적

국내 공공도서관의 목록데이터는 대체로 아이템 단위, 즉 개별 도서를 기반으로 하여 작성되어 있다. 즉, 각각의 개별 도서가 하나의 레코드를 차지하고 있는 것이다. 이러한 목록 작성 방식은 검색결과 레코드가 매우 과다하게 나타나게 되어 이용자가 도서를 검색하고 식별하는 데 있어 큰 장애요소가 된다. 예를 들어, G도서관의 홈페이지에서 '파친코'를 검색(검색일: 2024년 2월 4일)하면, 작가 이민진(이미정 역)의 2권짜리 소설 '파친코'의 1부 3건, 2부 3건, 총 6건의 레코드가 각각 별도로 출력된다. 통합기술방식의 목록데이터라면 1건, 개별기술방식이라면 2건의 레코드면 충분하지만, 아이템 기반의 목록데이터에서는 복본에 해당하는 도서들도 각각 별도의 목록데이터를 작성하기 때문에 검색 결과가 과다하게 나오는 현상이 일어난 것이다.

이러한 문제는 G도서관 하나만의 문제가 아니다. 노지현, 이은주(2023)의 연구에 따르면 부산 지역의 모든 공공도서관은 "1책당 1개의 레코드"를 원칙으로 서지데이터가 구축되고 있다. 이외에도 부산 지역을 제외한 16개 광역대표도서관을 조사한 결과 동일 자료(복본)를 책 단위로 분리 구축하는 방식이 절대적으로 많았으며, 서울도서관, 대전한밭도서관 등 일부의 도서관들만이 하나의 서지레코드에 복수의 소장정보를 연결하는 통합구축방식을 적용하고 있었다.

또한 이러한 아이템 기반의 목록데이터로는 지자체나 특정한 연합체 단위에서 통합목록 DB를 구축하는 데 어려움이 있으며, 향후 RDA, BIBFRAME 등의 개념이 적용된 목록데이터의 발전 방향에 대응하기 어렵다. 그렇기에 개별 도서관부터 아이템(개별자료) 기반의 목록데이터를 구현형 기반의 목록데이터로 통합하는 작업이 시급하다. 아이템 기반의 목록데이터에서는 하나의 도서에 대해 N+1개의 복본을 소장하고 있는 경우, 1개의 구현형 레코드가 아니라 N+1개의 개별 아이템 레코드가 작성된다. 처음 등록된 1건의 서지레코드를 기준 레코드라고 할 경우, N개의 복본 레코드가 작성된다는 것이다. 이러한 중복되는 복본 레코드들은 중복검증 알고리즘을 이용하여 기준 레코드와 중복 여부를 판단하고, 복본 레코드의 소장정보만 추출하여 기준 레코드에 추가하는 방식으로 하나의 서지레코드에 통합되어야 한다. 특히 다권본 도서의 경우, 목록데이터를 기술하는 과정에서 동일한 표제의 개별 도서들이 많은 다권본의 특성을 잘 반영하지 않기 때문에, 레코드를 통합하는 과정에서 더욱 세밀한 작업이 필요할 수 있다. 특히 다권본 문학 도서의 경우, e-book, 점자자료, 큰글자자료 등의 이형자료들이 많이 있으며, 권차만 다른 동일한 표제의 도서들이 많이 생성되기 때문에 이러한 과다한 레코드의 문제점이 더욱 크게 드러난다.

이에 본 연구에서는 부산 지역에서 비교적 최근에 개관한 공공도서관을 선정하여, 해당 공공도서관이 소장하고 있는 도서 중에서 다권본 도서가 풍부한 한국문학(KDC 800) 도서에 대해, 중복검증 알고리즘을 적용하여 서지레코드의 통합방안을 도출하고자 한다. 본 연구에서 설정한 연구 문제는 다음과 같다.

연구 문제 1. 공공도서관의 다권본 목록데이터에 중복검증 알고리즘을 적용하였을 때 어느 정도의 중복검증률을 나타내는가?

연구 문제 2. 공공도서관의 다권본 목록데이터의 중복검증률을 개선할 수 있는 방안은 무엇인가?

2. 연구방법

본 연구는 다음과 같은 과정을 통해 수행되었다. 첫째, 본 연구에 적용할 중복검증 알고리즘에 대해 파악하고 연구에 활용할 중복검증 알고리즘을 선정하였다. 둘째, 부산 지역에서 최근에 개관한 G도서관을 선정하고 OPAC 데이터를 통해 국내 다권본 문학 도서의 서지데이터를 웹 크롤링을 통해 수집하였다. 셋째, 수집한 데이터에 대해 선정한 중복검증 알고리즘을 적용하였다. 넷째, 알고리즘 적용 결과의 중복검증률을 분석하고 중복검증이 안 된 서지데이터의 원인을 파악하였다. 다섯째, 현행 공공도서관의 서지데이터에 관한 중복검증 알고리즘 적용의 개선방안을 제안하였다. 여섯째, 제안된 개선방안을 적용한 다음, 중복검증 알고리즘을 적용하고 기존 결과와 비교하였다.

Ⅱ. 이론적 배경 및 선행연구 분석

1. 중복검증 알고리즘

국내에서 실제 사용하고 있는 서지데이터 중복검증 알고리즘은 KERIS 알고리즘과 국립중앙도서관의 KOLIS-NET 중복검증 알고리즘 등이 있다. 그리고 현재 활용되고 있지는 않지만, 조순영(2003)이 KERIS 종합목록의 품질개선을 위해 새로운 유형의 중복데이터 색출 알고리즘을 제안한 바 있다.

조순영 알고리즘은 MARC 데이터 일치 여부를 비교하고 각 요소가 모두 해당 조건 이상이어야 중복으로 판정하는 KERIS 알고리즘과 다르게 비교 요소간의 유사성을 측정하고 각 요소의 중요도에 따라 가중치를 차등 부여하여 최종 중복값을 수치화하여 중복검증을 수행하는 방식의 알고리즘을 제시하였다. 해당 알고리즘은 KERIS 종합목록에 사용하기 위해 제시되었지만, 현재 활용되고 있지는 않다.

KERIS 담당자로부터 메일(2023년 11월 16일)로 확보한 KERIS 종합목록의 중복검증 알고리즘의 적용 절차는 다음과 같다. 첫째, 종합목록에 새로 추가하고자 하는 갱신요청 데이터에 대해, 중복 가능성이 있는 후보레코드들을 추출한다. 둘째, 그들을 대상으로 9가지의 MARC 데이터 요소를 비교하여, 각 비교요소별 점수를 산정한다. 셋째, 산정된 점수를 동일/유사 판정표와 비교하여 최종 판정(동일, 유사, 불일치)을 내린다. 9가지 비교요소는 서명, 저자명, 발행처, 발행년, 페이지, 판, 총서, 인식번호, 권차이며, 〈표 1〉과 같이 MARC 데이터 필드에서 추출한다.

KOLIS-NET 담당자로부터 메일(2024년 3월 11일)로 확보한 KOLIS-NET 알고리즘은 신규 작성된 레코드에 대해 고유 식별자에 해당하는 기관 제어번호와 ISBN을 검색엔진과 자체 DB에 검색하여 중복레코드를 파악하는 방식으로 진행된다. 1단계로 기관 제어번호를, 2단계로 ISBN을, 3단계로 표제, 발행자, 발행년을 검색하고, 각 단계에서 중복레코드를 파악한 경우, 다음 단계를 진행하지 않는다.

KERIS 알고리즘에서는 신규 작성된 레코드에 대해 고유 식별자를 통해 중복 가능성이 있는 후보레코드 추출한 다음, 신규레코드와 후보레코드의 MARC 데이터에서 9가지 비교요소를 추출하여 직접 비교하여 각 비교요소별 점수를 산정하여 중복 여부를 판정한다. 즉, KOLIS-NET 알고리즘이 고유한 식별자를 최우선으로 고려하여 중복 여부를 판정한다면, KERIS 알고리즘은 다양한 비교요소를 종합적으로 고려하여 중복 여부를 판정한다는 차이가 있다.

본 연구에서는 G도서관의 중복검증을 위해 KERIS 알고리즘을 사용하였다. 그 이유로는 우선 1997년 구축된 KERIS 종합목록이 2001년 서비스를 시작한 KOLIS-NET에 비해 종합목록 구축과 중복검증을 오래 수행하며 보완을 거친 알고리즘이기 때문이다. 또한 기준이 되는 자료와 기관 제어번호가 존재하여 고유 식별자를 우선으로 비교할 수 있는 종합목록 데이터와는 다르게, 공공도서관의 서지데이터에는 같은 도서의 복본 자료에도 별개의 등록번호가 부여되고, OPAC에서 다권본의 세트 ISBN과 개별 ISBN을 모두 확인할 수 없다. 그로 인해, 기관 제어번호와 ISBN을 최우선으로 비교하는 KOLIS-NET 알고리즘을 G도서관에 직접 적용하여 중복검증을 수행하기에 어려움이 따르기 때문이다.

〈표 1〉 KERIS 알고리즘의 9가지 비교요소와 MARC 데이터 필드

비교요소	MARC 데이터 필드
서명	245a(관제관칭 제외), 245a(관제관칭 포함), 245ab, 245abp, 245ap, 245x, 245b, 245p, 246a, 740a, 940a
저자명	245d(또는 c)의 정보를 역할어까지 추출 100a, 110a, 110ab, 111a, 700a, 710a, 710ab, 711a, 900a, 910a, 910ab, 911a를 순서대로 추출 저자정보가 존재하지 않는 경우, 260b 정보를 저자정보로 추출
발행처	ISBN 번호를 출판사보다 먼저 비교 008TAG의 26~27번째 한국대학출판부호와 38~39 한국정부기관부호 추출 260b의 출판사정보 추출 502b의 학위수여기관정보 추출
발행년	008TAG의 07~10 ('||||'는 발행년 없는 것으로 판정), 260c에서 최초 발견되는 연속된 4자리 숫자 260c에서 최초 발견되는 연속된 숫자정보 (4자리 X) 008TAG의 07~10 정보가 숫자정보가 아니어도 해당 정보를 그대로 추출
페이지	300a
판	250a('[…]' 사이의 정보 제거 안 함)
총서	490a, 490v, 830a, 830v, 440a, 440v, 400a, 400v, 410a, 410v, 411a, 411v, 245a
인식번호	020a에서 10자리(13자리)의 ISBN 추출 020에서 추출 할 때, ISBN 계산식에 의해, 10자리 혹은 13자리 ISBN 판별, 10자리 ISBN은 13자리 추가 추출, 13자리 ISBN의 경우 10자리 ISBN 추가 추출, 이후 MARC 대 MARC비교에서 ISBN 비교에 모두 사용 022a, z, y에서 9자리의 ISSN 추출 010a, z에서 13자리의 LCCN 추출
권차	245n (245n이 존재하지 않는 경우, 090c 등의 정보를 참조하지 않는다.)

KERIS 알고리즘의 각 비교요소별 점수 산정 기준은 〈표 2〉와 같다.

〈표 2〉 KERIS 알고리즘의 각 비교요소별 점수 산정 기준

비교요소	설명	점수 산정
서명	• 정보비교는 동자이음어, 사전변환, 표준형 변환을 한 정보로 비교한다. • 서명에 한자가 있을 경우, 한자 독음이 2개 이상 나올 수 있을 경우, 각각을 따로 추출하여 비교	245ab, 245ap, 245abp 정보간 완전일치시 5점
		완전일치하고 두 정보 중 하나는 245ab, 245abp, 245ap 중 하나이고 다른 하나는 245x이면 4점
		완전일치하나 두 정보 중 하나라도 245a(b/p는 다르고 a만 일치), 245b, 245p, 740a, 940a, 246a이면 3점
		서명 부분일치시 (최소 서명 길이 6자 이상, 80% 일치시) 2점
		기타 0점
저자명	• 정보비교는 표준형 변환을 한 정보로 비교한다. • 245d/c의 경우, 역할어까지 추출 • 1xx, 7xx, 9xx의 경우, ( )내의 정보는 제거하고 추출 • 저자명 정보 추출시, ‘,’ ‘;’ 이후의 정보는 무시 • 저자명에 한자가 있을 경우, 한자 독음이 2개 이상 나올 수 있을 경우, 각각을 따로 추출하여 비교	245d/c가 일치하면 3점
		245d/c를 제외하고 첫번째 저자 추출정보가 일치하면 3점
		저자정보 중 하나라도 일치하는 게 있으면 1점
		해당사항이 없으면 0점
발행처	• ISBN 비교 점수가 4점 이상이면, 발행처 점수를 4점 • ISBN 비교 점수가 4점 미만이면, 이래의 순서로 출판사 점수 비교함. • 발행처 비교 점수와 ISBN 비교 점수 중 더 높은 것을 발행처 비교 점수로 평가	추출한 260b 정보가 동일하면 4점, ( )안의 데이터는 무조건 제외
		추출한 260b가 반복될 경우, 각각을 추출해서 점수 비교 -> 하나만 일치해도 4점
		추출한 502b 정보가 동일하면 4점
		260b 정보가 Head/Tail match일 경우 2점
		해당사항이 없으면 0점
발행년	• 4자리 숫자정보를 가지는 발행년에 대해서는 +/- 계산까지 처리 • 발행년이 ‘19uu’ 등의 문자정보를 포함하거나, 4자리 숫자정보가 아닌 경우, 완전일치만 처리	추출한 출판년 정보가 완전일치하면 4점
		4자리 숫자인 출판년 정보를 +/- 1 했을 때 출판년이 동일해지면 2점
		해당사항이 없으면 0점
페이지	• 300a의 숫자그룹을 모두 추출하여, 페이지 정보를 각각 비교	두 서지에 페이지 정보의 개수와 내용이 모두 일치하면 5점
		두 서지에 페이지 개수는 일치하지 않으나, 동일한 내용이 있으면 3점
		두 서지에 모두 페이지 정보가 존재하지 않거나 한쪽 서지만 페이지 정보가 존재할 경우 2점
		해당사항이 없으면 0점(페이지 정보가 완전히 다른 경우)
판	• 정보비교는 사전변환, 표준형 변환을 한 정보로 비교한다.	추출한 정보가 완전일치하면 3점
		두 서지에 판 정보가 모두 없으면 3점
		해당사항이 없으면 0점
총서	• 정보비교는 사전변환, 표준형 변환을 한 정보로 비교한다.	추출한 총서사항 정보의 $a와 $v가 모두 같으면 3점
		두 서지에 모두 총서사항이 없으면(245a는 제외) 3점
		추출한 총서사항 정보의 $a만 같거나, ‘총서태그a’와 ‘245a’가 같을 경우 2점 (245$a 끼리는 서로 비교하지 않음)
		해당사항이 없으면 0점
인식번호	• 정보비교는 표준형 변환을 한 정보로 비교한다. • ISBN의 경우, 중간에(‘-’문자가 있는 경우는 ‘-’문자를 제거한 후 비교 • 인식번호의 비교는 동일유형(ISBN, ISSN, LCCN)간에만 비교	▼a 의 추출갯수와 내용이 모두 동일하면 5점
		▼a 의 추출 갯수는 다르나, 일치하는 내용이 있는 경우 4점
		▼a, ▼z, ▼y에 상관없이 일치하는 내용이 하나라도 있으면 3점
		둘 다 없으면 2점
		해당사항이 없으면 0점
권차	• 정보비교는 사전변환, 표준형 변환을 한 정보로 비교한다. • 권차 정보에 로마자 숫자가 있는 경우, 일반 숫자로 변환한다. • 권차 정보의 맨 처음이 “제3권” 형식으로 되어 있으면, “제”문자를 제거하고 비교한다.	두 서지에 모두 권차가 있으면서 일치하면 3점
		두 서지에 모두 권차가 없으면 2점
		한 서지에만 권차가 있으면 1점
		두 서지에 모두 권차가 있으나 다르면 0점

각 비교요소별로 점수를 산정한 뒤(산정점수), 중복 여부를 최종 판정하기 위한 중복 판정 점수표를 활용하여 중복여부를 판정한다. 단행본, 연속간행물, 학위논문, 다권본, 고서, 비도서의 자료유형별로 다른 점수표를 적용한다. 이 중에서 다권본에 적용되는 판정 점수표는 〈표 3〉과 같다.

목록레코드에 부여된 산정점수를 역순으로 구성된 우선순위에 따라 기준 점수를 적용하여 비교한다. 각 비교요소별 점수를 모두 만족하면 해당 판정을 내리게 되고, 만족하지 못하면 다음 우선순위의 점수를 비교하는 과정을 반복한다. 유사 판정의 우선순위 1의 점수까지 만족하지 못할 경우, 불일치로 판정한다. 일부 비교요소의 경우 점수가 0점으로 산정되더라도 다른 데이터 요소가 모두 일치하면 동일 도서로 판정될 수 있다. 입력 과정에서의 사소한 오타 등으로 인해 일부 비교요소의 점수가 0점으로 잘못 산정되더라도 다른 데이터 요소가 모두 일치하면 동일 도서 또는 유사 도서로 판정될 수 있는 것이다. 예를 들어, 서명의 경우, 산정점수가 최소 3점 이상이어야 기준도서와 동일 즉, 복본으로 판정할 수 있으나, 발행년의 경우, 산정점수가 0점이어도 동일 도서(우선순위 3)로 판정할 수 있다.

〈표 3〉 KERIS 알고리즘의 다권본 중복 판정 점수표

판정	우선순위(역순)	판정점수	서명	저자	발행처	발행년	페이지	판	총서	인식번호	권차
동일	5	9	5	3	4	2	5	0	0	0	2
	4	9	5	3	4	4	5	3	3	5	0
	3	9	5	1	2	0	5	3	0	0	2
	2	9	4	3	4	0	5	3	0	0	2
	1	9	3	1	2	0	0	3	0	5	2
유사	6	6	2	3	2	4	5	0	2	0	2
	5	5	3	0	4	4	3	3	0	2	2
	4	4	0	1	2	0	5	0	2	0	0
	3	3	2	0	0	0	2	0	3	5	0
	2	2	5	0	2	0	0	0	0	2	0
	1	1	2	1	0	0	0	0	0	2	2

2. 선행연구

국내에서는 KERIS가 종합목록을 구축하기 위해 중복검증 알고리즘을 개발하여 현행 적용하고 있으나, 그에 관한 연구는 조순영(2003)의 연구를 제외하고는 거의 없는 실정이다. 조순영(2003)은 KERIS 종합목록의 품질 개선을 위해 새로운 유형의 중복 데이터 색출 알고리즘을 제안하였다. 당시 적용하고 있던 MARC 데이터 일치여부 비교 방식이 아닌 서지유형별 다른 비교방식을 적용하고 비교 요소간의 유사성을 측정하고 각 요소의 중요도에 따라 가중치를 차등 부여하는 방식을

채택하였다. 새로 개발한 알고리즘의 효용성을 입증하기 위하여 당시 종합목록에 업로드된 데이터 210,000건을 추출하여 실험용 마스터 파일을 구축하고 7,649건을 두개의 알고리즘으로 처리한 결과 새로운 알고리즘에서 중복레코드의 색출 비율이 36.2%p 더 높게 나타났다.

김선애, 이수상(2006)은 국가자료종합목록시스템(KOLIS-NET) DB의 품질을 검증하는 과정에서 중복레코드로 인한 심각성 정도와 유형에 대해 파악하였다. KOLIS-NET의 중복레코드가 발생한 유형으로는 출판년도의 차이, 저자사항 표기방식의 차이, 서명 기술방식의 차이, 형태사항 필드의 누락, ISBN 필드의 누락, 그리고 기타 순으로 나타났다.

해외에서는 다양한 기관들의 종합목록을 구축하는 과정에서 중복레코드에 관한 문제가 일찍부터 제기되어 활발한 연구가 진행되고 있다. Goyal(1987)은 영국 국가서지(British National Bibliography)와 OCLC의 서지레코드를 바탕으로 서지 데이터베이스의 중복 감지에 사용하기 위한 코드와 적용 가능성을 제시하였다. 이를 위하여 USBC(Universal Standard Book Code) 코드를 활용한 자동 중복 검증 알고리즘을 제시하였다. Toney(1992)는 국제 문헌 보존 네트워크인 BCIN(Bibliographic Database of the Conservation Information Network)에서 데이터 오류를 수정하고 중복 기록을 제거하는 프로젝트를 수행하였다. 비교 항목으로는 출판년도, 레코드 번호, 권차, 발행유형, 제목, 페이지의 항목을 추출하였다. 약 14만 개의 서지레코드에서 발생한 수많은 오류와 중복에 대해 컴퓨터 프로그램을 활용하여 오류를 인식하고 자동 수정하고, 자동으로 수정할 수 없는 항목에 대해서는 추후 수정을 위해 표시하였다.

O'Neill, Rogers, & Oksins(1993)은 OCLC의 온라인 종합 목록의 중복 레코드를 분석하여, 그 각 데이터 요소별로 그 특성을 파악하고 각 데이터 요소별로 중복검증 과정에서 일어날 수 있는 오류들에 대해 파악하였다. Cousins(1998)은 470만 개의 서지레코드를 보유하고 있는 영국의 종합목록 COPAC를 바탕으로 중복 감지 및 기록 통합 절차를 수행하는 방법과 다른 데이터임에도 불구하고 중복레코드로 잘못 판정한 경우와 중복레코드임에도 불구하고 판정하지 못하는 경우 등 직면한 문제 영역에 대해 설명하고 COPAC에서 사용되는 중복 기록 처리 메커니즘의 모델을 제시하였다. Sitas & Kapidakis(2008)는 ALEPH-ULM, OCLC, COPAC 등 다양한 기관의 중복 레코드 검증을 위한 알고리즘에 대해 조사하였다. 대부분의 알고리즘에서 중복 여부를 판정하는 방법은 필드 비교 및 가중치 부여 방식이었다. 또한, 알고리즘은 대체로 단일 단계 비교 방식과 다단계 비교 방식으로 구분할 수 있으며, 단일 단계 비교 방식의 경우, 빠르고 저렴하게 중복 제거를 진행할 수 있지만, 비교적 정확도가 떨어졌다.

Beall(2010)은 단행본과 논문의 차이점에 대해 설명하며, 레코드의 중복 문제는 기존에는 다양한 판본을 지닌 단행본의 문제였으나, 논문 메타데이터의 증가와 교차 라이센스 문제로 논문의 메타데이터에서도 중복 레코드 문제가 발생하고 있음을 지적하였다. 이에, 논문의 서지레코드의 중복도를 일관되게 측정하고 정의하기 위한 새로운 수식을 제안하였다.

이후에는 다양한 기술을 접목시켜 자동으로 중복레코드를 탐지하는 것에 관한 연구가 진행되었다. Taniguchi(2013)는 기존에 주로 사용된 서지레코드 기반의 중복검증 방법 대신 표제지와 그 뒷면을 OCR(Optical Character Recognition, 광학 문자 인식)로 변환하여 이를 기존의 서지레코드와 비교하여 중복 여부를 판정하는 방법을 제안하고 검토하였다. Liu & Zeng(2016)은 대규모 디지털 도서관들 간의 통합검색에서 중복레코드를 기계 학습을 통해 자동으로 탐지하는 방법에 대해 연구하였다. 학습 세트를 추출하고, 각 속성을 매핑하고, 속성의 가중치를 학습하여 최종적으로 서로 다른 디지털 도서관에서 일치하는 서지레코드를 식별하는 방법에 대해 연구하였다.

Ⅲ. 중복검증 알고리즘의 적용

1. 분석 대상 데이터의 선정

중복검증 분석의 대상 기관으로는 부산 지역에 위치한 G도서관을 선정하였다. G도서관은 2021년 개관한 공공도서관이며, 7만 권이 넘는 도서에 관한 아이템 기반의 서지데이터를 보유하고 있다. 중복검증 알고리즘은 기본적으로 MARC 데이터를 대상으로 중복검증을 수행하지만, MARC 데이터는 도서관 업무 시스템에서 반출해야 하는 내부데이터이기에, 확보에 어려움이 있다. 그렇기에 본 연구에서는 도서관 홈페이지에서 쉽게 확인할 수 있고 MARC 데이터를 기반으로 표시되는 OPAC 데이터를 사용하기로 하였다. 데이터 형식에는 차이가 있지만 그 내용은 MARC 데이터를 기반으로 하기 때문에, 중복검증 알고리즘을 적용하는 데는 무리가 없다.

G도서관의 OPAC 데이터는 2023년 12월 20일부터 22일까지 3일에 걸쳐 부산지역 도서관 자료검색 페이지(http://one-library.busan.go.kr/busanbooks/)에서 웹 크롤링으로 수집하였다. 구체적으로는 구글 colab 환경에서 python의 BeautifulSoup4를 통해 G도서관의 등록번호를 기준으로 웹 크롤링을 진행하였다. 전체 도서 74,301권의 데이터를 수집하고, 수집한 데이터를 openpyxl 라이브러리를 통해 엑셀 형식의 파일로 변환하였다.

G도서관 OPAC 데이터에 나타난 청구기호는 '별치기호(해당시) KDC분류기호-입수순기호-권차기호(해당시)=복본기호(해당시)'의 형태로 구성되어 있다. 이러한 청구기호 체계를 활용하여, 수집된 데이터 중에서 최종 분석 대상 데이터를 선별하였다. 첫째, 외국에서 출판된 국내 도서의 번역서나 점역서, 큰글자도서와 같은 동일 저작의 이형 판본들의 경우, 중복검증 과정에서 혼란을 일으킬 수 있으므로 별치기호를 확인하는 과정을 거쳐 총 7,756권의 데이터를 삭제하였다.[1)]

1) 별치기호가 몽골(26권), 베트남(45권), 빅북(179권), 아동몽골(55권), 아동베트남(110권), 아동소리책(594권), 아동영어(3401권), 아동인니(61권), 아동일본(256권), 아동중국(396권), 아동캄보디아(42권), 아동태국(48권),

둘째, 나머지 66,545권의 도서 중에서 청구기호의 KDC 분류기호를 기준으로 문학(800) 도서 30,421권을 추출하였다. 셋째, KDC 분류기호와 입수순 기호가 같고, 권차기호가 서로 다른 도서가 존재하는 16,630건의 도서들을 국내 문학 다권본 도서들로 선정하였다. 마지막으로 KDC 분류기호와 입수순 기호, 권차기호까지 모두 같고 복본기호가 다른 도서가 있는 2,485권의 도서를 2권 이상 복본이 있는 국내 문학 다권본 도서로서, 최종 분석 대상으로 하였다.

분석 대상 도서들을 대상으로 복본기호를 바탕으로 기준도서와 복본도서를 구분하였다. 이를 위하여 동일한 청구기호에서 복본기호가 있는 도서와 없는 도서 중에서는 복본기호가 없는 도서를 기준도서로, 복본기호가 있는 도서들은 복본도서로 분리하였다. 그리고 복본기호가 모두 존재하는 경우, 복본기호가 빠른 도서를 기준도서로 삼고, 나머지는 복본도서로 분리하였다. 이러한 과정을 통해 총 1,097권의 기준도서와 1,388권의 복본도서로 구분하였다. 평균적으로 1권의 기준도서에 대해 1.27권의 복본도서가 있는 것이다. 중복검증 알고리즘을 통해 복본도서들이 기준도서에 대해 중복이 확인되면, 복본도서의 소장정보를 추출하여 기준도서의 소장정보에 통합하는 것만으로 목록레코드를 통합할 수 있다. 이를 위해 중복검증 알고리즘을 적용하여 기준도서의 OPAC 데이터와 해당하는 복본도서의 OPAC 데이터를 비교하고, 중복 여부를 확인하는 작업이 필요하다. 다음 〈표 4〉는 G도서관의 전체 도서와 본 연구에 사용된 도서의 수를 정리한 것이다.

〈표 4〉 최종 분석 대상 도서 선정 과정

(단위: 권)

구분	도서 수	비고
G도서관 전체 소장 도서	74,301	
국내 출판 단행본	66,545	외국 출판 도서 및 이형자료 7,756권 제외
문학	30,421	KDC 주류 800번대
다권본	16,630	KDC 분류기호와 입수순 기호가 같고, 권차기호가 서로 다른 도서가 존재하는 자료
2권 이상의 복본	2,485	기준도서: 1,097권 복본도서: 1,388권

2. 중복검증 알고리즘의 선정과 변경

KERIS 중복검증 알고리즘은 현재 국내에서 가장 오랫동안, 그리고 가장 방대하게 구축해 온 대학도서관 종합목록에 실제로 적용하고 있는 것이기에 본 연구의 중복검증 알고리즘으로 선택하였다. KERIS의 중복검증 작업은 MARC 데이터를 대상으로 종합목록에 새롭게 추가하고자 하는 갱신요청 데이터 중에서 기존의 수많은 데이터 중에서 중복된 레코드를 찾는 방식으로

아동필리핀(85권), 아트(457권), 영어(616권), 인니(28권), 일본(29권), 점자(217권), 점자라벨(123권), 중국(128권), 캄보디아(27권), 큰글자(754권), 태국(37권), 팝업(17권), 필리핀(25권)인 도서의 데이터를 삭제하였다.

이루어진다. 하지만 G도서관의 서지데이터는 MARC 데이터가 아니라 OPAC 데이터를 사용한다. 아이템 기반으로 구성된 G도서관의 OPAC 데이터를 청구기호를 통해 기준도서와 복본도서로 구분하고, 기준도서에 관한 복본도서의 중복 정도를 검증하기 전에, 기존 KERIS의 중복검증 알고리즘을 변경하는 과정이 필요하다. 따라서 기존의 KERIS 알고리즘의 기본적인 절차는 그대로 유지하되, OPAC 데이터의 특성에 맞도록 세부요소 데이터의 추출과 점수 산정 방식을 변경하였다.

가. 세부요소 데이터의 추출

G도서관 OPAC의 서지정보 데이터는 서명, 저자사항, 발행사항, 형태사항, 표준부호, 이용대상의 6개 요소로 구성되어 있었으며 그 외에 이용자 대출을 위한 소장정보(청구기호, 등록번호, 자료실, 반납예정일, 대출상태) 데이터가 있다. 이 중에서 분석에 필요한 데이터는 서지정보 5가지(서명, 저자사항, 발행사항, 형태사항, 표준부호)와 소장정보 2가지(청구기호, 등록번호)의 7가지를 수집하였다. 수집한 데이터는 구두점으로 구분하여 중복검증 알고리즘 적용을 위한 비교요소 추출에 활용하였다.

KERIS 변경 알고리즘을 새로 생성하기 위해, 우선 G도서관 OPAC 데이터 내용과 형식을 분석하여, 각 비교 요소에 해당하는 데이터를 매핑시키는 과정을 진행하였다. 우선 G도서관의 OPAC에서 확인할 수 있는 데이터의 형태는 〈그림 1〉과 같다.

셜록 홈즈 전집=Sherlock Holmes. 4, 공포의 계곡	
저자사항	아서 코난 도일 지음 ; 리하르트 거트슈미트 삽화 ; 백영미 옮김
발행사항	황금가지, 2010, 인쇄자료(책자형), ₩ 9000
형태사항	286 p. : 22 cm
표준부호	ISBN : 9788982734045
분류기호	한국십진분류법 : 843
이용대상	일반

푸른 사자 와니니 : 이현 장편동화. 02, 검은 땅의 주인	
저자사항	이현 지음 ; 오윤화 그림
발행사항	창비, 2019, 인쇄자료(책자형), ₩ 10800
형태사항	224 p. : 23 cm
표준부호	ISBN : 9788936443054
분류기호	한국십진분류법 : 808.3
이용대상	일반

〈그림 1〉 G도서관 OPAC 데이터 내용과 형식 예시

G도서관 OPAC 데이터에서는 구두점을 사용하여 각 비교요소별 세부요소를 구분하고 있는 것을 확인할 수 있어, 구두점을 기준으로 알고리즘에 적용하기 위한 세부요소를 추출하였다. 우선 서명 정보는 다음과 같은 구조로 구성되어 있다.

본표제($a)=대등표제($x) : 표제관련정보($b). 권차($n), 권표제($p)

위의 서명 구조에 나타나는 4가지 구두점을 기준으로 자동으로 구분하되, 반점(,)의 경우, 본표제 안에서도 흔하게 나타나는 기호이기 때문에 권차사항을 나타내는 온점(.) 구두점 이후에 나타나는 경우에만 권표제를 나타내는 구두점으로 판정하였다. 〈그림 1〉에 나타난 2건의 OPAC 데이터 사례의 서명을 알고리즘 적용을 위한 세부요소로 구분하면 〈표 5〉와 같다.

〈표 5〉 〈그림 1〉 데이터의 서명 세부 요소 구분

서명 구분	MARC 기호	사례 1	사례 2
본표제	245 $a	설록 홈즈 전집	푸른 사자 와니니
대등표제	245 $x	Sherlock Holmes	-
표제관련정보	245 $b	-	이현 장편동화
권차	245 $n	4	02
권표제	245 $p	공포의 계곡	검은 땅의 주인

저자사항, 발행사항, 형태사항, 표준부호 요소들에 대해서도 변경 작업을 수행하였다. 저자사항 요소는 구두점에 따른 세부요소 구분 없이 그 자체로 추출하여 전체 비교한다. 단, 각괄호([ ])와 '[공]'은 제외하고 비교하였다. 단 전체 비교에서 불일치하는 경우, 세부 비교를 위해 구두점 ' , '와 ' ; '를 기준으로 구분한다. 발행사항 요소는 '발행처, 발행년, 자료유형, 가격'의 구조로 구성되어 있다. 이에, 첫 구두점 ' , ' 이전에 나타나는 내용을 발행처 요소로, 첫 구두점 ' , ' 이후 나타나는 4자리 숫자정보를 발행년 요소로 추출하였다. 형태사항 요소는 '페이지 : 크기'의 구조로 구성되어 있다. 구두점 ' : ' 이전에 나타나는 숫자정보를 페이지 요소로 추출하였다. 표준부호 요소는 'ISBN : 숫자'의 구조로 구성되어 있다. 구두점 ' : ' 뒤에 나타난 10자리 또는 13자리 숫자정보를 인식번호 요소로 추출하였다.

KERIS 알고리즘의 9가지 비교요소 중에서 판사항과 총서사항의 경우, G도서관의 OPAC 데이터에서는 확인할 수 없어, 비교요소에서 제외하였다.

나. 점수 산정 방식 변경

KERIS 알고리즘의 점수 산정 방식에서 사용한 MARC 필드들은 G도서관의 OPAC 데이터에서 추출한 세부요소 데이터로 매핑시켜 변경하였다. 예를 들어, 서명 비교요소의 경우, KERIS 점수 산정표에서는 245ab, 245ap, 245abp 정보간 완전일치시 5점을 부여한다. 이를 G도서관 OPAC 데이터의 본표제, 표제관련정보, 권표제가 모두 일치하는 경우 5점으로 변경한 것이다. 최종적으로 위에서 설명한 각 비교요소별 추출 방식과 점수 산정 방식을 변경한 것을 모두 정리한 것은 〈표 6〉과 같다.

〈표 6〉 변경 적용한 비교요소별 추출 방식 및 점수 산정 방식

비교요소	추출 대상	점수 산정
서명	• 표제 전체를 추출한 뒤 구두점을 기준으로 구분 • 구두점가 나타나기 전 내용을 본표제로 추출 • 구두점 '=' 이후의 내용을 대등표제로 추출 • 구두점 ':' 이후의 내용을 표제관련정보로 추출 • 구두점 '.' 이후에 등장하는 구두점 ',' 이후의 내용을 권표제로 추출	본표제, 표제관련정보, 권표제가 모두 일치하면 5점
		내용은 모두 일치하되, 본표제, 표제관련정보, 권표제의 내용 중 하나가 다른 레코드에서는 대등표제의 형태로 나타날 때는 4점
		본표제나 표제관련정보, 권표제 중 하나만 완전일치하는 경우 3점
		본표제 부분일치시 (최소 서명 길이 6자 이상, 80% 일치시) 2점
		기타 0점
저자명	• 저자사항 항목 전체를 추출하여 점수 산정 • 저자사항 비교 과정에서 가장 먼저 각괄호([ ])와 '[공]'은 제외하고 비교	저자사항 데이터가 완전일치하면 3점
		저자사항 데이터가 완전일치하지 않은 경우, 구두점 ',', ';' 이후의 내용과 역할어를 모두 제거한 내용을 첫 번째 저자로 간주하여 완전일치한 경우 3점
		각 구두점 ',', ';'로 구분한 내용들 중 하나라도 일치하는 경우 1점
		기타 0점
발행처	• 발행사항 정보에서 첫 구두점(,) 이전에 나타나는 내용	인식번호 비교 점수가 4점 이상이면 4점
		인식번호 비교 점수가 4점 미만이면, 이래의 순서로 출판사 점수 비교하여 발행처 비교 점수와 인식번호 비교 점수 중 더 높은 것을 발행처 비교 점수로 평가
		추출한 정보가 동일하면 4점, ( )안의 데이터는 무조건 제외
		Head/Tail match일 경우 2점
		해당사항이 없으면 0점
발행년	• 발행사항 정보에서 첫 구두점(,)와 두번째 구두점(,) 사이에 나타나는 4자리 숫자정보	추출한 출판년 정보가 완전일치하면 4점
		4자리 숫자인 출판년 정보를 +/- 1 했을 때 출판년이 동일해지면 2점
		해당사항이 없으면 0점
페이지	• 형태사항 정보에서 처음 나타나는 숫자정보	페이지 정보의 내용이 존재하고 서로 일치하면 5점
		두 서지에 모두 페이지 정보가 존재하지 않거나 한쪽 서지만 페이지 정보가 존재할 경우 2점
		해당사항이 없으면 0점(페이지 정보가 완전히 다른 경우)
판	G도서관의 OPAC에서 확인할 수 없어 사용하지 않음	
총서	G도서관의 OPAC에서 확인할 수 없어 사용하지 않음	
인식번호	• 표준부호 정보에서 'ISBN : ' 이후 나타나는 13자리 숫자정보	일치하면 5점, 불일치하면 0점
권차	• 표제의 구두점 '.' 이후의 내용을 권차로 추출	두 서지에 모두 권차가 있으면서 일치하면 3점
		두 서지에 모두 권차가 없으면 2점
		한 서지에만 권차가 있으면 1점
		두 서지에 모두 권차가 있으나 다르면 0점

3. 중복검증 알고리즘의 적용

웹 크롤링을 통해 수집한 G도서관의 OPAC 데이터는 엑셀 형식이기에, 각 비교요소별로 위에서 서술한 세부요소 데이터 추출 과정에 따라 구두점을 기준으로 데이터 칼럼을 분리하였다. 이를 위해 엑셀에서 지정한 위치의 문자열을 반환하는 LEFT, MID 함수를 사용하였다. 분리하는 기준 위치를 잡기 위해 특정 문자의 위치를 출력하는 FIND 함수를 사용하여 구두점의 위치를 지정하였다.

세부요소 데이터를 분리한 뒤, 각 세부요소별 점수 산정 방식에 따라 IF 함수를 적용하여 기준도서와 복본도서의 세부요소를 비교하고 점수를 산정하였다. IF 함수는 주어진 조건이 참인지 거짓인지를 판단하여 그에 따른 결과를 반환하는 함수로, 기준도서와 복본도서의 세부요소 데이터가 각 점수 산정 기준을 만족하는지에 따라 점수를 산정하는 데 활용되었다. 위 과정을 거쳐 산정된 점수를 바탕으로 〈표 3〉의 중복 판정 점수표를 참고하여, 동일 도서에 해당하는 점수 기준 5가지, 유사 도서에 해당하는 점수 기준 6가지의 IF 조건문을 작성하였다. 이렇게 작성된 조건문을 산정점수에 적용하여 각 도서의 중복 여부를 최종 판정하였다.

이렇게 G도서관의 OPAC 데이터에 중복검증 알고리즘을 적용한 결과는 〈표 7〉과 같다. 총 1,388권의 복본 중에서 1,326권(95.53%)을 기준도서와 동일한 복본으로 판정하였으며, 25권(1.8%)은 유사한 도서로, 그리고 37권(2.67%)은 복본이 아닌 불일치 도서로 판정하였다.

〈표 7〉 G도서관 데이터에 관한 KERIS 변경 알고리즘 적용 결과

(단위: 권)

구분	전체	기준도서	복본도서	판정 결과		
				동일	유사	불일치
건수	2,485	1,097	1,388	1,326 (95.53%)	25 (1.8%)	37 (2.67%)

중복검증 알고리즘에서 사용한 7가지 세부요소에 관한 산정점수의 구체적인 내용은 다음과 같다.

가. 서명 산정점수

점수 산정 방식에 따르면 서명 산정점수는 4가지(5점, 4점, 3점, 2점, 0점)로 산정할 수 있지만, 분석 대상 데이터에서는 3가지(0점, 3점, 5점) 점수만 나타났다. 〈표 3〉의 판정 점수표에 의해 서명 산정점수가 0점인 17권은 모두 동일 도서로 판정되지 않았으며, 유사 도서(13권) 또는 불일치 도서(4권)로 각각 판정되었다. 판정 점수표에 따라 동일 도서로 판정되기 위해서는 서명 점수가 최소 3점이어야 한다. 서명 산정점수 3점을 부여받은 도서는 48권이었으며, 75%인 36권이 동일 도서로 판정되었다, 서명 산정점수 5점을 부여받은 도서는 1,323권이었으며, 97.51%인 1,290권이 동일 도서로 판정되었다. 서명 산정점수를 기준으로 판정 결과를 정리한 것은 〈표 8〉과 같다.

〈표 8〉 G도서관 국내 문학 다권본 도서의 서명 산정점수별 판정 결과

(단위: 권)

점수	동일	유사	불일치	계
0점	0	13	4	17
3점	36	7	5	48
5점	1,290	5	28	1,323
계	1,326	25	37	1,388

나. 저자명 산정점수

저자명 산정점수가 0점일 경우, 판정 점수표에 의해 유사 도서(우선순위 5)로까지 판정할 수 있다. 동일 도서로 판정하기 위해서는 저자명 점수가 최소 1점이어야 한다.

저자명 산정점수가 0점인 경우의 사례의 경우, 필명을 사용하는 저자에 대해 본명과 필명을 각각 작성한 경우였다. 해당 사례로 '내 어린고양이와 늙은 개. 1'의 저자는 '정솔'이고 필명이 '초'이다. G도서관에서 이 도서의 저자사항이 '초(정솔)' 또는 '정솔'로 혼용되고 있으며, 해당 도서 3건은 모두 유사 도서로 판정되었다. 이 경우, 저자를 동일 인물로 판정하여 저자명 산정점수 3점을 부여할 경우, 동일 도서(우선순위 3)로 판정된다. 저자 산정점수를 기준으로 판정 결과를 정리한 것은 〈표 9〉와 같다.

〈표 9〉 G도서관 국내 문학 다권본 도서의 저자명 산정점수별 판정 결과

(단위: 권)

점수	동일	유사	불일치	계
0점	0	3	0	3
1점	2	1	1	4
3점	1,324	21	36	1,381
계	1,326	25	37	1,388

다. 발행처 산정점수

발행처 산정점수가 0점일 경우, 최대 유사 도서(우선순위 3)로 판정할 수 있다. 동일 도서로 판정하기 위해서는 발행처 점수가 최소 2점이어야 한다.

발행처 산정점수의 경우, 인식번호 산정점수가 4점 이상일 경우에는 일괄적으로 4점을 부여하고, 괄호 안의 데이터는 비교 대상에서 제외하는 등, 판정 점수를 부여하는 기준이 후한 편이다. 발행처 산정점수가 0점이라는 것은 곧 인식번호 점수도 0점이라는 뜻이다. 이에, 발행처 산정점수가 0점으로 산정된 경우는 2건에 불과하였으며, 모두 불일치 도서로 판정되었다. 발행처 산정점수를 기준으로 판정 결과를 정리한 것은 〈표 10〉과 같다.

〈표 10〉 G도서관 국내 문학 다권본 도서의 발행처 산정점수별 판정 결과

(단위: 권)

점수	동일	유사	불일치	계
0점	0	0	2	2
4점	1,326	25	35	1,386
계	1,326	25	37	1,388

라. 발행년 산정점수

발행년 산정점수가 0점일 경우, 최대 동일 도서(우선순위 3)로 판정될 수 있다. 발행년 산정점수가 0점인 도서는 52권으로, 이 중 57.69%인 30권이 동일 도서로 판정되었다. 발행년 산정점수를

기준으로 판정 결과를 정리한 것은 〈표 11〉과 같다.

〈표 11〉 G도서관 국내 문학 다권본 도서의 발행년 산정점수별 판정 결과

(단위: 권)

점수	동일	유사	불일치	계
0점	30	8	14	52
2점	27	1	3	31
4점	1,269	16	20	1,305
계	1,326	25	37	1,388

마. 페이지 산정점수

페이지 산정점수가 0점일 경우, 최대 동일 도서로 판정될 수 있다. 단, 인식번호 산정점수가 5점이어야만 동일 도서(우선순위 1)로 판정할 수 있다.

최종적으로 불일치로 판정된 도서 중에서 대부분은 페이지 점수가 0점이었다. 최종 불일치 판정 건수 37건 중, 35건이 페이지 점수가 0점인 경우였다. 페이지 산정점수를 기준으로 판정 결과를 정리한 것은 〈표 12〉와 같다.

〈표 12〉 G도서관 국내 문학 다권본 도서의 페이지 산정점수별 판정 결과

(단위: 권)

점수	동일	유사	불일치	계
0점	131	2	35	168
2점	1	0	0	1
5점	1,194	23	2	1,219
계	1,326	25	37	1,388

바. 인식번호 산정점수

인식번호 산정점수가 0점일 경우, 최대 동일 도서(우선순위 5)로 판정할 수 있다. 최종 불일치 판정 건수 37건 중, 35건이 인식번호 점수가 0점인 경우였다. 인식번호 산정점수를 기준으로 판정 결과를 정리한 것은 〈표 13〉과 같다.

〈표 13〉 G도서관 국내 문학 다권본 도서의 인식번호 산정점수별 판정 결과

(단위: 권)

점수	동일	유사	불일치	계
0점	101	7	35	143
5점	1,225	18	2	1,245
계	1,326	25	37	1,388

ISBN 데이터가 개별 ISBN과 세트 ISBN으로 나누어 입력된 경우와 실제 다른 판본의 ISBN이 입력된 경우를 나누어 파악하였다. 143건의 ISBN 불일치 사례 중에서 106건은 각각의 레코드에

개별 ISBN과 세트 ISBN이 입력되어 ISBN이 불일치하였다. 35건은 실제로 개정판, 양장본 등 별개의 ISBN을 부여받은 다른 판본의 도서였다. 나머지 2건의 경우, ISBN의 체크기호에 해당하는 마지막 번호만이 다른 경우와 다른 권차의 ISBN이 잘못 입력된 오기입 사례였다. 각 인식번호 불일치 유형별 판정 결과는 〈표 14〉와 같다.

〈표 14〉 G도서관 국내 문학 다권본 도서의 인식번호 불일치 유형별 판정 결과

(단위: 권)

불일치 유형	동일	유사	불일치	계
개별/세트	87	3	16	106
다른 판본	12	4	19	35
오기입(추정)	2	0	0	2
계	101	7	35	143

사. 권차 산정점수

권차 산정점수가 0점일 경우, 최대 동일 도서(우선순위 4)로 판정할 수 있다. 단, 동일 도서(우선순위 4)로 판정하기 위해서는, 다른 모든 점수가 최고 점수를 만족해야 한다. 권차 산정점수가 0점인 도서는 9권이었으며, 그 중에서 4권은 동일 도서로, 4권은 유사 도서로 판정되었다. 권차 산정점수를 기준으로 판정 결과를 정리한 것은 〈표 15〉와 같다.

〈표 15〉 G도서관 국내 문학 다권본 도서의 권차 산정점수별 판정 결과

(단위: 권)

점수	동일	유사	불일치	계
0점	4	4	1	9
1점	1	2	1	4
2점	859	9	23	891
3점	462	10	12	484
계	1,326	25	37	1,388

Ⅳ. 데이터 교정과 알고리즘의 재적용

1. 불일치 판정 도서의 비교요소별 문제점 파악

G도서관의 다권본 문학 도서에 KERIS 변경 알고리즘을 적용한 결과, 불일치 판정이 내려진 전체 37권의 도서와 7개 비교요소에 관한 세부 산정점수는 〈표 16〉과 같다. 이 중에서 페이지 점수에서 0점으로 판정된 도서는 35권이고, 인식번호 점수에서 0점으로 판정된 도서 역시 35권이었다. 페이지 점수와 인식번호 점수가 모두 0점인 도서는 33권이었다. 발행년 점수가 0점으로 산정된 경우는 14권이다. 불일치 판정을 받은 도서들의 비교요소별 세부 산정점수를 확인하고, 우선적으로 0점으로 산정된 데이터

요소들에 대해 0점이 나타난 원인을 파악하면, 중복검증률을 개선할 수 있을 것으로 판단할 수 있다.

인식번호 점수의 경우, 대부분 데이터 표기상의 오류가 원인이었다. 다권본은 그 특성상 전체 시리즈의 ISBN과 개별 권차의 ISBN을 함께 가지고 있다. 하지만 G도서관 OPAC 데이터에는 인식번호를 하나만 확인할 수 있었다. 또한 그 인식번호에 개별 ISBN과 시리즈 ISBN이 혼용되어 있어, 실제 도서의 데이터와 다르게 불일치 판정이 일어나는 경우가 많았다. 〈표 16〉에서 볼 수 있듯이, 기준도서와 비교하여 ISBN이 불일치한 사례는 총 143건이지만 이 중에서 판본이 다르거나 ISBN이 오기입되는 등 실제로 ISBN이 불일치한 사례는 37건으로 파악되었다. 106건은 한 레코드에서는 개별 도서의 ISBN이, 다른 레코드에서는 세트 전체의 ISBN이 기입되어 있어 불일치 판정이 된 것이다. 그리고 인식번호 점수가 0점인 불일치 판정 사례 35권 중에서 16권이 위와 같은 개별 ISBN과 시리즈 ISBN의 표기 오류 때문이었다. 이것으로 볼 때, G도서관 데이터에서 도서의 ISBN을 교정하는 것만으로도 중복검증률을 크게 높일 수 있을 것이다.

〈표 16〉 불일치 판정 사례의 세부 산정점수

번호	표제	서명	저자	발행처	발행년	페이지	인식번호	권차
1	노인과 바다	5	3	4	0	0	0	2
2	개인적인 체험	5	3	4	0	0	0	2
3	젊은작가상 수상작품집. 2020(제11회)	5	3	4	4	0	0	3
4	곰탕. 1	3	3	4	0	0	0	3
5	칵테일, 러브, 좀비(리커버)	0	3	4	0	0	5	2
6	(설민석의) 삼국지. 1	5	3	4	4	0	0	3
7	(설민석의) 삼국지. 2	5	3	4	4	0	0	3
8	마음에 쏙 드는 엄마를 원하세요?	5	3	4	4	0	0	2
9	수염 전쟁	5	3	4	0	0	0	2
10	마틸다	5	3	4	0	0	0	2
11	박씨 성을 가진 노비	0	3	4	0	0	0	2
12	빨리 놀자 삼총사	5	3	4	0	0	0	2
13	악당 우주 돼지가 수상해	5	3	4	0	0	0	2
14	Go! 생물 탐험	0	3	4	4	0	5	2
15	귀신 감독 탁풍운	5	3	4	4	0	0	1
16	고양이 학교. 1-1, 수정 동굴의 비밀	5	3	4	0	0	0	0
17	(코믹)메이플스토리. 94	5	3	4	4	0	0	3
18	(코믹)메이플스토리. 95	5	3	4	4	0	0	3
19	(코믹)메이플스토리. 96	5	3	4	4	0	0	3
20	간니닌니 마법의 도서관. 2, 이상한 나라의 앨리스	3	3	4	4	0	0	3
21	간니닌니 마법의 도서관 : 명작 속으로 떠나는 판타지 동화 여행 . 3, 알라딘과 요술 램프	3	3	4	4	0	0	3
22	흔한 남매. 5	5	3	0	4	5	0	3
23	흔한 남매. 5	5	3	0	4	5	0	3
24	130층 나무 집 : 13층씩 커지는 상상! 유머! 모험! 책장이 술술 넘어가는 재미!	3	3	4	2	0	0	2
25	꽁꽁꽁 좀비	5	3	4	4	0	0	2
26	리디아의 정원	5	1	4	0	0	0	2

번호	표제	서명	저자	발행처	발행년	페이지	인식번호	권차
27	추 선생님의 특별한 미술 수업	5	3	4	0	0	0	2
28	테푸할아버지의 신기한 요술 테이프	0	3	4	0	0	0	2
29	나무야 넌 혼자가 아니야	5	3	4	4	0	0	2
30	오, 나의 푸드 트럭	5	3	4	4	0	0	2
31	곰의 부탁 : 진형민 소설	5	3	4	4	0	0	2
32	사랑에 빠질 때 나누는 말들 : 탁경은 장편소설	5	3	4	4	0	0	2
33	호수의 일	3	3	4	4	0	0	2
34	알로하, 나의 엄마들	5	3	4	4	0	0	2
35	야채호빵의 봄방학. 4	5	3	4	0	0	0	3
36	까칠한 재석이가 폭발했다	5	3	4	2	0	0	2
37	까칠한 재석이가 결심했다	5	3	4	2	0	0	2

한편, 인식번호 산정점수가 5점인데도 불일치로 판정된 사례는 『칵테일, 러브, 좀비(리커버)』와 『Go! 생물 탐험』의 2권이며, 페이지 산정점수와 서명 산정점수가 모두 0점이기 때문에 〈표 3〉의 판정 점수표에 따라 불일치 도서로 판정된 것이다. 기준도서의 서명은 『칵테일, 러브, 좀비』이지만, 복본도서의 서명이 『칵테일, 러브, 좀비(리커버)』이기에, '(리커버)'의 유무가 차이로 나타났다. 그리고 기준도서와 복본도서의 발행년과 페이지가 모두 다르기에, 다른 도서로 유추할 수 있다. 『Go! 생물 탐험』의 경우, 기준도서의 서명은 『GO! 생물 탐험』이기에, 'GO'와 'Go'처럼 대소문자의 차이로 서명의 불일치가 발생하여 서명 점수가 0점으로 산정된 것이다. 특히 이 사례는 서명과 페이지를 제외한 모든 비교요소의 점수가 높게 산정되어 있어, 해당 오류를 교정하면 판정 결과가 바뀌고 중복검증률이 높아질 수 있을 것이다.

서명 산정점수가 0점인 2가지 사례는 『박씨 성을 가진 노비』와 『테푸할아버지의 신기한 요술 테이프』이다. 『박씨 성을 가진 노비』의 경우, 기준도서의 서명이 『(박팽년의 후예)박씨 성을 가진 노비』이기에, 관제의 유무에 차이가 있었다. 『테푸할아버지의 신기한 요술 테이프』의 경우, 기준도서의 서명은 『테푸할아버지의 요술 테이프』로, 재출판되면서 표제에 '신기한'이 추가된 사례로 확인되었기에, 다른 도서로 보아야 한다.

저자 점수가 0점인 사례는 없고, 1점으로 산정된 사례는 『리디아의 정원』이다. 기준도서의 레코드에는 저자사항이 '사라 스튜어트 글 ; 데이비드 스몰 그림 ; 이복희 옮김'이며, 복본도서의 레코드에는 '데이비드 스몰 그림 ; 사라 스튜어트 글 ; 이복희 옮김'이다. 글 작가와 그림 작가의 순서가 다르게 작성되어 있다. 하지만, 이 사례는 발행년, 페이지, 인식번호 산정점수가 모두 0점이기 때문에, 저자사항의 교정만으로 판정 결과를 바꾸고 중복검증률을 높일 수 있을 것으로 기대하기 어렵다.

한편, 『흔한 남매. 5』의 경우, 발행처 점수가 0점으로 산정된 유일한 사례이다. 기준도서의 레코드에는 발행처가 '아이세움'이며, 복본도서의 레코드에는 발행처가 각각 '미래엔', '미래엔아이세움'이기에, 발행처의 불일치가 발생한 것이다. 그러나 '미래엔 아이세움'은 출판사 '미래엔'의 아동 출판 전문 하위 브랜드로, 실제로는 같은 출판사에서 발행한 자료임에도 불구하고 표기의 차이로 인해 발행처 점수 산정

에 문제가 발생한 것이다. 다만, 발행처 점수 산정 기준에 따르면, 인식번호 산정점수가 4점 이상일 경우 발행처 산정점수도 4점으로 산정하기 때문에 인식번호의 교정만으로 충분히 교정이 이루어질 수 있다.

권차 점수가 0점인 사례는 『고양이 학교. 1-1, 수정 동굴의 비밀』이다. 기준도서의 레코드에는 권차가 '1부 1권'으로, 복본도서의 레코드에는 '1-1'로 표기되어 있었다. 권차의 숫자는 일치하지만, 표기 양식의 차이 때문에 0점으로 산정된 것이다. 하지만 이 사례 역시 발행년, 페이지, 인식번호 산정점수가 모두 0점이기 때문에, 권차의 교정만으로 판정 결과를 바꾸고 중복검증률을 높이기 어렵다. 권차 점수가 1점인 사례로 『귀신 감독 탁풍운』이다. 이는 기준도서의 레코드에는 '[1]'이라는 권차가 입력되어 있으나, 복본도서의 레코드에는 권차가 입력되어 있지 않아 한 서지에만 권차가 있으면 1점이라는 점수 산정 기준에 따라 권차 점수가 1점으로 산정되었다. 이 경우, 실제 도서에서는 1권이라는 권차를 확인할 수 없었으나, 2권의 존재로 인해 해당 도서에 각괄호([ ])와 함께 1권이라는 권차를 기재한 것으로 보인다.

그리고 발행년과 페이지 요소의 경우, 점수가 낮게 산정된 사례는 많지만 해당 요소들의 경우, 실제로 다르게 표기될 여지가 많다. 예를 들어, 새로운 개정판이나 증보판이 발행될 경우, 새로운 발행년도가 부여된다. 또한 같은 책이라도 판본의 종류나 제본 방식(양장본, 페이퍼백 등)에 따라 다를 수 있다. 발행년과 페이지 요소에서 기준도서와 복본도서의 차이를 오류로 판단하고 교정을 진행하기 위해서는 도서관이 실제로 보관하고 있는 책을 일일이 확인해야 한다. 〈표 3〉의 판정 점수표에 따르면, 발행년과 페이지 요소의 산정점수가 모두 0점이어도 동일 도서(우선순위 1)로 판정이 가능하기 때문에, 다른 요소들의 교정 작업을 통해 충분히 중복검증률을 높일 수 있을 것으로 기대할 수 있다. 이에, 두 요소에 대해서는 별도의 교정 작업을 진행하지 않도록 하였다.

2. 불일치 판정 도서의 데이터 교정과 알고리즘 적용

KERIS 변경 알고리즘의 적용 결과 불일치 판정 도서에 나타난 데이터의 문제점을 위와 같이 파악하고, 다음과 같이 각 비교요소에 대하여 엑셀 데이터를 대상으로 1차 데이터 교정 작업을 수행하였다. 첫째, 우선 가장 큰 문제점이었던 개별 ISBN과 세트 ISBN의 차이로 인해 불일치가 발생하는 것에 대해 교정하였다. 레코드에 입력된 ISBN을 국립중앙도서관의 ISBN · ISSN · 납본 시스템 페이지(https://www.nl.go.kr/seoji/)에서 검색하고 이를 통해 파악한 개별 ISBN과 세트 ISBN을 모두 레코드에 입력하였다. 둘째, 대소문자의 구분으로 인해 서명의 불일치가 발생한 것에 대해 교정하였다. 『Go! 생물 탐험』의 사례가 이에 해당한다. 셋째, 저자사항에 입력된 순서의 차이로 인해 불일치가 발생한 것에 대해 교정하였다. 『리디아의 정원』의 사례가 이에 해당한다. 저자사항에 입력된 저자 정보의 순서를 실제 도서의 표제지와 표지에 표기된 순서에 따라 교정하였다. 넷째, 권차의 입력 형태의 차이로 인해 불일치가 발생한 것에 대해 교정하였다. 실제 도서의 표제지와

표지에서 확인할 수 있는 권차의 표기 형태에 따라 교정하였다. 『고양이 학교, 1-1, 수정 동굴의 비밀』과 『귀신 감독 탁풍운』의 사례가 이에 해당한다. 다섯째, 발행처의 출판사 하위 브랜드의 표기 여부에 따라 불일치가 발생한 것에 대해 교정하였다. 『흔한 남매. 5』의 사례가 이에 해당한다.

위와 같이 1차 데이터 오류 교정 작업을 거친 이후 KERIS 알고리즘을 다시 적용한 결과는 〈표 17〉과 같다.

〈표 17〉 1차 데이터 교정 후 KERIS 알고리즘 적용 결과 비교

교정 여부	동일	유사	불일치	계
교정 전	1,326 (95.53%)	25 (1.80%)	37 (2.67%)	1,388
1차 교정 후	1,343 (96.76%)	25 (1.73%)	20 (1.44%)	

1차 데이터 교정 작업 전후로 바뀐 판정 결과는 다음과 같다. 첫째, 인식번호 때문에 발생한 35권의 불일치 판정 사례 중에서 16권에 대해 데이터 교정이 이루어졌다. 16권 중 1권은 유사 도서로, 15권은 동일 도서로 판정이 수정되었다. 둘째, 『Go! 생물 탐험』의 서명 수정으로 동일 도서로 판정이 수정되었다. 셋째와 넷째, 저자와 권차 교정 결과 『리디아의 정원』과 『고양이 학교, 1-1, 수정 동굴의 비밀』은 각각 저자와 권차 산정점수가 수정되었지만, 그대로 불일치 도서로 판정되었다. 『귀신 감독 탁풍운』의 경우, 인식번호의 교정으로 유사 도서로 판정이 이미 수정되었으나, 권차의 교정으로 동일 도서로 판정이 다시 수정되었다. 다섯째, 『흔한 남매. 5』에 관한 발행처의 교정이 이루어졌으나, 해당 도서는 이미 인식번호의 교정으로 동일 판정으로 수정되어 발행처의 교정이 판정에 영향을 미치지는 못했다.

결국 불일치 판정 도서의 데이터 오류 교정 작업을 통해 17권의 불일치 도서가 동일 도서로 재판정되었으며, 중복검증률이 기존 95.53%에서 96.76%로 1.23%p 상승하였다.

3. 유사 판정 도서의 요소별 문제점 파악

불일치 판정 도서에 대해 1차 데이터 교정 작업을 수행한 다음, 유사 판정 도서에 대해 2차 데이터 교정 작업을 수행하고자 하였다. G도서관의 다권본 문학 도서에 KERIS 변경 알고리즘의 2차 적용결과, 유사 판정이 내려진 도서는 전체 25권이며, 7개 비교요소에 관한 세부 산정점수는 〈표 19〉와 같다.

유사 판정 도서 중에서 서명 점수가 0점인 경우는 『초대』 시리즈 2권, 『해리 포터 : 죽음의 성물』 시리즈 4권, 『임경업전』, 『왜냐면…』 시리즈 3권, 『카레가 보글보글』, 『책 민들레 엄대섭, 모두의 도서관을 꿈꾸다』, 『나는 책의사 선생님』으로 총 13권이었다. 해당 사례의 각 기준도서의 서명과 복본도서의 서명을 비교한 것은 〈표 18〉과 같다.

서명 점수가 0점인 13권의 유사 판정 사례는 모두 관제와 표제관련정보의 처리를 다르게 적용

한 경우이다. 『초대』의 경우 기준도서에서는 '첫번째', '두번째'를 권차로, 복본도서에서는 표제의 관제로 작성하고 권차를 별도로 작성하였다.

〈표 18〉 서명의 불일치로 인한 유사 판정 사례

번호	기준도서 서명	복본도서 서명
1	초대. 첫번째	(첫번째)초대. [1]
2	초대. 두번째	(두번째)초대. [2]
3	해리 포터와 죽음의 성물. 1	해리 포터 : 죽음의 성물. 1
4	해리 포터와 죽음의 성물. 2	해리 포터 : 죽음의 성물. 2
5	해리 포터와 죽음의 성물. 3	해리 포터 : 죽음의 성물. 3
6	해리 포터와 죽음의 성물. 4	해리 포터 : 죽음의 성물. 4
7	(아동문학가 고정욱 선생님이 다시 쓴 우리 고전)임경업전	임경업전
8	왜냐면	왜냐면… : 안녕달 그림책
9	왜냐면	왜냐면… : 안녕달 그림책
10	왜냐면	왜냐면… : 안녕달 그림책
11	우당탕탕 야옹이. 07, 카레가 보글보글	카레가 보글보글
12	(책 민들레 엄대섭)모두의 도서관을 꿈꾸다	책 민들레 엄대섭, 모두의 도서관을 꿈꾸다
13	(구름빵)나는 책의사 선생님	나는 책의사 선생님

〈표 19〉 유사 판정 사례의 세부 산정점수

번호	표제	서명	저자	발행처	발행년	페이지	인식번호	권차
1	프랑켄슈타인 : 메리 셸리 장편소설	3	3	4	0	5	0	2
2	곰탕. 2	3	3	4	0	5	0	3
3	제명 공주 : 일본의 천황이 된 백제 공주. 1	3	3	4	4	5	0	3
4	죽고 싶지만 떡볶이는 먹고 싶어	5	3	4	4	0	5	1
5	죽고 싶지만 떡볶이는 먹고 싶어	5	3	4	4	0	5	0
6	(첫번째)초대. [1]	0	3	4	4	5	5	0
7	(두번째)초대. [2]	0	3	4	4	5	5	0
8	해리 포터 : 죽음의 성물. 1	0	3	4	4	5	5	3
9	해리 포터 : 죽음의 성물. 2	0	3	4	4	5	5	3
10	해리 포터 : 죽음의 성물. 3	0	3	4	4	5	5	3
11	해리 포터 : 죽음의 성물. 4	0	3	4	4	5	5	3
12	악플 전쟁	3	3	4	0	5	0	2
13	임경업전	0	3	4	4	5	5	2
14	간니닌니 마법의 도서관. 5, 빨간 머리 앤	3	3	4	2	5	0	3
15	올림포스 여신스쿨. 2, 페르세포네의 거짓말	3	3	4	0	5	0	0
16	왜냐면… : 안녕달 그림책	0	3	4	4	5	5	2
17	왜냐면… : 안녕달 그림책	0	3	4	4	5	5	2
18	왜냐면… : 안녕달 그림책	0	3	4	4	5	5	2
19	카레가 보글보글	0	3	4	4	5	5	1
20	책 민들레 엄대섭, 모두의 도서관을 꿈꾸다	0	3	4	4	5	5	2
21	고구마구마 : 특별판	3	3	4	0	5	0	2
22	나는 책의사 선생님	0	3	4	4	5	5	2
23	내 어린고양이와 늙은 개. 1	5	0	4	0	5	5	3
24	내 어린고양이와 늙은 개. 1	5	0	4	0	5	5	3
25	내 어린고양이와 늙은개. 3	5	0	4	0	5	5	3

저자명 점수가 0점인 경우는 3건이었으며, 『내 어린고양이와 늙은 개』의 경우가 이에 해당한다. 3권 모두 기준도서의 저자사항은 '초(정솔) 그리고 씀'으로, 복본도서의 저자사항은 '정솔 지음'으로 작성되어 있었다. 저자명 서술 방식 자체에도 '씀'과 '지음'으로 차이가 있으나, '정솔'이라는 저자가 '초'라는 필명을 사용하고 있어 본명만이 작성된 레코드와 본명과 필명이 함께 작성된 레코드가 불일치 판정을 받았다. 해당 사례의 경우, 저자와 발행년 점수를 제외한 모든 비교요소의 점수가 만점으로 판정되어, 저자명의 오류를 교정하고 점수를 다시 산정하여 판정을 개선할 수 있을 것이다.

인식번호 점수가 0점이 나온 유사 판정 사례는 7권으로, 불일치 판정 사례와 마찬가지로 데이터 표기상의 오류가 원인이었다. 불일치 판정 사례와 마찬가지로, 도서의 개별 ISBN과 세트 ISBN을 모두 입력하는 교정 작업으로 중복검증률을 높일 수 있을 것으로 기대할 수 있다.

권차 점수가 0점으로 판정된 사례는 4권이 있었다. 이 중에서 『죽고 싶지만 떡볶이는 먹고 싶어』의 기준도서의 권차는 '[1]'이고, 복본도서의 권차는 없었다. 『초대』 시리즈의 기준도서는 권차가 없었으나, 복본도서의 권차는 각 '[1]', '[2]'로 입력되어 있었다. 전자의 경우는 권차가 별도로 존재하지 않으나, 2권이 출판되면서 1권이라는 정보를 추가로 유추하여 권차에 입력한 사례이다. 후자의 경우는 권차가 숫자 형태가 아닌 한국어 관제의 형태로 입력되어 있어, 숫자 형태의 권차를 별도로 입력한 사례이다. 하지만 두 사례 모두 처리 과정에서 일관성 있게 처리하지 못하여 점수 산정에서 오류가 발생하였다.

『올림포스 여신스쿨. 2:, 페르세포네의 거짓말』의 경우, 기준도서의 권차사항은 '2:'로, 복본도서의 권차사항은 '2'로 작성되어 있었다. G도서관의 OPAC을 통해 확인할 수 있는 기준도서의 서명이 『올림포스 여신스쿨. 2:, 페르세포네의 거짓말』이다. 권차사항에는 ' : '이 일반적으로 사용되지 않는 점에서 이는 표기의 오류라고 볼 수 있다.

발행년 점수가 0점인 유사 판정 사례는 8권, 페이지 점수가 0점인 유사 판정 사례는 2권이다. 발행년 점수와 페이지 점수는 불일치 판정 사례에서처럼 별도의 교정 작업을 진행하지 않도록 하였다. 발행처 점수가 0점인 유사 판정 사례는 없었다.

4. 유사 판정 도서의 데이터 교정과 알고리즘 적용

KERIS 변경 알고리즘의 적용 결과 유사 레코드의 사례에 나타난 데이터의 문제점을 위와 같이 파악하고, 다음과 같이 각 비교요소에 대하여 2차 데이터 교정 작업을 수행하였다. 첫째, 우선 불일치 판정 사례와 마찬가지로 국립중앙도서관의 ISBN・ISSN・납본 시스템의 데이터를 참고하여 개별 ISBN과 세트 ISBN을 모두 입력하였다. 둘째, 서명을 실제 도서의 표지와 표제지에 나타난 형태를 참고하여 교정하였다. 이 과정에서 서명의 권차 표기 역시 표지와 표제지에 나타난 형태에 따라 교정하였다. 셋째, 본명과 별도로 필명을 사용하는 저자에 대해 저자명을 일치시켰다.

위와 같이 2차 데이터 오류 교정 작업을 거친 이후 KERIS 알고리즘을 다시 적용한 결과는 〈표 20〉과 같다.

〈표 20〉 2차 데이터 교정 후 KERIS 알고리즘 적용 결과 비교

교정 여부	동일	유사	불일치	계
교정 전	1,326 (95.53%)	25 (1.80%)	37 (2.67%)	1,388
1차 교정 후	1,343 (96.76%)	25 (1.80%)	20 (1.44%)	
2차 교정 후	1,364 (98.27%)	4 (1.80%)	20 (1.44%)	

데이터 오류 교정 작업 전후로 바뀐 판정 결과는 다음과 같다. 첫째, 인식번호 데이터의 교정이 유사 판정 사례 3권에 대해 이루어졌다. 이 중에서 2권이 인식번호의 교정만으로 동일 판정으로 수정되었다. 둘째, 서명 데이터의 교정이 13권에 대해 이루어졌으며, 해당 사례들 모두 동일 판정으로 수정되었다. 또한 권차 데이터의 교정을 통해 3권이 동일 판정으로 수정되었다. 이 중 1권은 앞서 인식번호 교정만으로 판정이 수정되지 않은 사례로, 인식번호와 권차 두 비교요소의 교정을 모두 거쳐 판정이 수정되었다. 셋째, 서명 데이터의 교정 작업을 3권 진행하였으며, 3권 모두 교정을 통해 동일 판정으로 수정되었다. 총 21권의 유사 판정 도서가 데이터 교정 작업을 통해 동일 도서로 판정이 수정되었다.

이를 종합하면 데이터 오류 교정 작업을 통해 21권의 유사 판정 도서가 동일 도서로 재판정되었으며, 중복검증률이 기존 96.76%에서 98.27%로 1.51%p 상승하였다. 1차 교정 작업과 2차 교정 작업을 거쳐 중복검증률이 총 2.74%p 상승하였다.

데이터 교정 후에도 불일치 판정을 받은 20권 중에서 19권은 실제로 개정판, 양장본 등 별도의 ISBN을 부여받고 출판된 다른 판본의 자료들이었으며, 1권은 ISBN은 동일하게 입력되어 있었으나 발행년도가 다르고 복본도서의 표제에 관제인 '(리커버)'가 입력되어 있어 서명 점수가 낮아 불일치 도서로 판정되었다. 데이터 교정 후에도 유사 판정을 받은 4권 역시 모두 별도의 ISBN을 부여받고 출판된 다른 판본의 자료였다.

Ⅴ. 논의 및 결론

본 연구는 KERIS 중복검증 알고리즘을 실제 공공도서관의 데이터에 적용하여, 중복 도서를 판정하였다. 판정 결과로 나타난 불일치 판정 사례와 유사 판정 사례의 도서들을 대상으로 각 비교요소별로 판정 결과에 나타난 문제점을 분석하였다. 그리고 분석한 문제점을 토대로 목록데이터를 교정하면, 중복검증률이 개선될 수 있는 것을 확인하였다. 이를 통해 공공도서관의 중복된

아이템 레코드들을 구현형 레코드로 전환하기 위한 도구로서 KERIS 중복검증 알고리즘의 활용 가능성을 확인하였다. 개별 도서관에서 KERIS 중복검증 알고리즘을 우선 적용하고, 불일치 판정과 유사 판정의 사례들을 분석하여 문제가 있는 목록데이터의 내용을 교정하고, KERIS 중복검증 알고리즘을 재적용하는 방식으로 중복검증률을 높일 수 있다. 최종적으로 중복 자료로 판정된 자료들에 대해 복본도서의 소장정보를 기준도서의 소장정보에 통합하는 것만으로 아이템 기반의 공공도서관의 목록레코드를 타이틀 기반으로 통합할 수 있다.

다만 이 연구는 실제 도서관의 MARC 데이터를 사용하지 못하고 검색페이지에 나타난 OPAC 데이터를 이용하였다는 한계가 있다. 물론 OPAC 데이터는 MARC 데이터를 기반으로 표기하는 것이기 때문에 OPAC 데이터를 활용한 중복검증과 MARC 데이터를 활용한 중복검증률은 큰 차이가 나지 않을 수 있다. 하지만, OPAC 데이터에서는 판사항과 총서사항 요소를 확인할 수 없고, MARC 데이터 필드에서 다양한 필드에서 나타나는 데이터 요소를 모두 파악할 수 없다는 한계가 있다. 이에, MARC 데이터를 직접 추출하고 이를 KERIS 중복검증 알고리즘에 적용하여 더욱 정밀하게 알고리즘을 적용하는 후속 연구가 필요하다.

또한 이 연구는 부산 지역의 G도서관이라는 개별 도서관에 한정하여 중복검증을 진행하고 유사/불일치 판정 사례를 파악하였기 때문에 다른 도서관에서도 발생할 수 있는 중복검증 과정에서 발생할 수 있는 모든 오류 사례를 파악하지 못했다는 한계가 있다. 그리고 이 연구는 중복검증 과정에서 적용할 수 있는 원칙을 제시하기보다는 KERIS 중복검증 알고리즘이 공공도서관의 다권본 문학 도서에 적용이 가능하며, 약간의 데이터 교정만으로도 중복검증률을 확실하게 높일 수 있다는 점을 확인했다는 것에 그 의의가 있다.

개별 도서관에서 중복검증을 통한 아이템 기반의 서지레코드를 구현형 기반의 서지레코드로 통합하는 작업을 수행한 다음에는 지자체나 연합체 등에서 구축하는 다수의 통합도서관 목록, KOLIS-NET 등과 같은 대규모 통합도서관 수준에서 중복검증을 통한 통합목록 DB의 구축이 진행되어야 한다. 따라서 본 연구에서 수행한 개별 도서관 대상의 중복검증 작업뿐 아니라 통합목록 DB의 구축을 위해 복수의 도서관들을 대상으로 하는 중복검증 작업에 관한 후속연구도 필요하다.

참 고 문 헌

국립중앙도서관 (연도미상). KOLIS-NET 종합목록 중복검증 알고리즘. [KOLIS-NET 종합목록 담당자로부터 메일(2024. 3. 11.)로 전달받음]

김선애, 이수상 (2006). KOLIS-NET 종합목록 DB의 품질평가. 한국문헌정보학회지, 40(1), 95-117. http://uci.or.kr/G704-000226.2006.40.1.005

노지현, 이은주 (2023). 공공도서관 서지데이터의 품질 제고 방안 - 부산시 공공도서관을 중심으로 -. 한국도서관·정보학회지, 54(3), 105-128. https://doi.org/10.16981/kliss.54.3.202309.105

조순영 (2003). 종합목록의 중복레코드 검증을 위한 알고리즘 연구. 한국문헌정보학회지, 37(4), 69-88. http://uci.or.kr/G704-000226.2003.37.4.001

한국교육학술정보원 (2006). 중복검사 및 품질평가(종합목록). [KERIS 종합목록 담당자로부터 메일(2023. 11. 16.)로 전달받음]

Beall, J. (2010). Measuring duplicate metadata records in library databases. Library Hi Tech News, 27(9/10), 10-12. https://doi.org/10.1108/07419051011110595

Cousins, S. A. (1998). Duplicate detection and record consolidation in large bibliographic databases: the COPAC database experience. Journal of Information Science, 24(4), 231-240. https://doi.org/10.1177/016555159802400402

O'Neill, E. T., Rogers, S. A., & Oskins, M. W. (1993). Characteristics of duplicate records in OCLC's Online Union Catalog. Library Resources & Technical Services, 37(3), 59-71.

Goyal, P. (1987). Duplicate record identification in bibliographic databases. Information Systems, 12(3), 239-242. https://doi.org/10.1016/0306-4379(87)90002-0

Liu, W. & Zeng, J. (2016). Duplicate literature detection for cross-library search. Cybernetics and Information Technologies, 16(2), 160-178. https://doi.org/10.1515/cait-2016-0028

Sitas, A. & Kapidakis, S. (2008). Duplicate detection algorithms of bibliographic descriptions. Library Hi Tech, 26(2), 287-301. https://doi.org/10.1108/07378830810880379

Taniguchi, S. (2013). Duplicate bibliographic record detection with an OCR-converted source of information. Journal of Information Science, 39(2), 153-168. https://doi.org/10.1177/0165551512459923

Toney, S. R. (1992). Cleanup and deduplication of an international bibliographic database. Information Technologies and Libraries, 11(1), 19-28.

• 국한문 참고문헌의 영문 표기
(English translation / Romanization of references originally written in Korean)

Cho, Sun-Yeong (2003). A study on duplicate detection algorithm in union catalog. Journal of the Korean Society for Library and Information Science, 37(4), 69-88. http://uci.or.kr/G704-000226.2003.37.4.001

Kim, Sun-Ae & Lee, Soo-Sang (2006). Quality evaluation of a shared cataloging DB:

the case of KOLIS-NET. Journal of the Korean Society for Library and Information Science, 40(1), 95-117. http://uci.or.kr/G704-000226.2006.40.1.005

Korea Education and Research Information Service (2006). Duplicate Verification and Quality Assessment(Union Catalog). [Received via email from KERIS officer (2023, November 16)].

National Library of Korea (n.d.). KOLIS-NET Duplication Verification Algorithm [Received via email from National Library of Korea officer (2024, March 11)].

Rho, Jee-Hyun & Lee, Eun-Ju (2023). Improving the quality of bibliographic data in public libraries: focusing on public libraries in Busan Metropolitan City. Journal of Korean Library and Information Science Society, 54(3), 105-128. https://doi.org/10.16981/kliss.54.3.202309.105