한국기록관리학회지
22(1), 43-59, 2022.2
http://dx.doi.org/10.14404/JKSARM.2022.22.1.043
pISSN 1598-1487 eISSN 2671-7247

JKSARM
Journal of Korean Society of Archives and Records Management

이관 기록물 분류 자동화를 위한 목록 기반 이상치 판별 학습데이터 구축

Building the Outlier Candidate Discrimination Training Data based on Inventory for Automatic Classification of Transferred Records

정지혜(Ji-Hye Jeong)[1], 이젬마(Gemma Lee)[2], 왕호성(Hosung Wang)[3], 오효정(Hyo-Jung Oh)[4]

E-mail: j-jye@naver.com, gemma617@korea.kr, vvwang@korea.kr, ohj@jbnu.ac.kr

[1] 제1저자 전북대학교 일반대학원 기록관리학과 박사과정
[2] 국가기록원 디지털혁신과
[3] 국가기록원 디지털혁신과
[4] 교신저자 전북대학교 문헌정보학과 부교수, 문화융복합아카이빙연구소 공동연구원

OPEN ACCESS

논문접수 2022-01-20
최초심사 2022-01-25
게재확정 2022-02-07

ORCID
Ji-Hye Jeong
https://orcid.org/0000-0002-7771-5771
Gemma Lee
https://orcid.org/0000-0003-2605-2143
Hosung Wang
https://orcid.org/0000-0002-1955-7998
Hyo-Jung Oh
https://orcid.org/0000-0001-8067-2832

© 한국기록관리학회
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0/) which permits use, distribution and reproduction in any medium, provided that the article is properly cited, the use is non-commercial and no modifications or adaptations are made.

• 본 논문은 '2021년 국가기록관리 활용기술 연구개발(R&D) 사업'의 연구비를 지원받아 수행되었음.
• 본 논문은 2019년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (과제번호: NRF-2019S1A5B8099507).

초 록

전자적으로 생산된 공공기록물은 생산과 동시에 편철되고 보존기간이 부여되며 일정기간이 지나면 영구기록물관리기관으로 이관되어 보존된다. 이관 시 기록물관리 담당자가 기록물 분류정보를 확인하고 품질을 일정 수준으로 유지토록 해야 하지만, 이관된 기록물의 분류는 기록물 정리/기술 업무로 편성되어 있고, 대부분의 정리/기술 업무는 수작업에 의존하고 있어 당해 연도에 처리해야 할 기록물 수량을 맞추기 어려운 실정이다. 이에 본 연구는 이관 기록물 분류 업무의 효율화와 일관된 기준을 유지하기 위한 방안을 제안하고자 한다. 이를 위해 먼저 국가기록원에서 수행하고 있는 현행의 기록분류 업무 프로세스를 분석하고 개선 요구사항을 수렴하여 분류 업무의 수작업을 최소화하기 위한 방안으로 이관된 기록물의 편철 정보, 즉 목록에 기반한 분류 이상치 후보를 판별하는 과정을 도출·체계화하였다. 나아가 제안한 이상치 판별 프로세스를 실제 국가기록원으로 이관된 기록물을 대상으로 적용하고, 그 결과를 규격화하여 추후 기계학습에 활용 가능한 학습데이터 형식으로 구축하였다. 본 연구의 궁극적인 목적은 지능형 전자기록 관리 환경 구축을 위한 사전 단계로, 기록관리 업무 내 기계학습 기법이 적용 가능한 문제 유형을 선별하고 자동화하는 방안을 모색하고자 한다.

ABSTRACT

Electronic public records are classified simultaneously as production, a preservation period is granted, and after a certain period, they are transferred to an archive and preserved. This study intends to find a way to improve the efficiency in classifying transferred records and maintain consistent standards. To this end, the current record classification work process carried out by the National Archives of Korea was analyzed, and problems were identified. As a way to minimize the manual work of record classification by converging the required improvement, the process of identifying outlier candidates based on a list consisting of classified information of the transferred records was proposed and systemized. Furthermore, the proposed outlier discrimination process was applied to the actual records transferred to the National Archives of Korea. The results were standardized and constructed as a training data format that can be used for machine learning in the future.

Keywords: 이관 기록물, 기록분류, 자동화, 학습 데이터, 이상치 판별
transferred records, Rrecords classification, automation, training data, outlier discrimination

1. 서 론

1.1 연구 배경

기록은 조직의 기능과 업무에 따라 분류하는 것을 원칙으로 한다. ISO 15489에는 조직의 기능과 업무분석을 토대로 기록분류체계를 개발하는 것을 원칙을 명시하고 있으며, 우리나라의 중앙 및 지방 행정기관들도 공공기록물관리에 관한 법률(이하 공공기록물관리법)이 정하는 바에 따라 기능분류체계를 적용하고 있다. 우리나라에서는 1999년 공공기관의 기록물관리에 관한 법률이 제정되면서 업무를 반영한 기록분류의 원칙을 도입하였고, 2005년 동법이 공공기록물관리법으로 전면 개정되면서 업무 및 기능분류의 원칙이 더욱 강조되었다.

공적 업무 수행과정에서 생산된 공공기록물은 생산과 동시에 보존기간이 부여되고 업무 맥락을 반영하여 편철된다. 또한 일정기간 생산기관의 업무 목적을 위해 활용되다가 보존기간이 30년 이상으로 분류된 기록물은 공공기록물관리법 및 동법 시행령으로 정하는 기간 이내에 영구기록물관리기관으로 이관되어 보존된다. 영구기록물관리기관은 해당되는 법규에 따라 기록물을 이관받고, 특정 목적에 따라 기록물을 수집하며, 적법한 절차에 따라 소장 기록물을 평가하고 처분 한다. 이러한 정책은 향후 기록물을 우선순위에 따라 선별 수집하고 승인된 정책과 절차에 따라 평가·처분하여 기록물의 증거가치를 지속적으로 보유하게 함으로써, 해당 영구기록물관리기관에서 기록물관리의 설명책임성을 입증할 수 있도록 한다.

국내 대표 영구기록물관리기관인 국가기록원이 소장하고 있는 영구기록물에 대한 목록정보는 기록생산기관에서 이관 시에 제출하는 기록물 목록에 대한 정보가 그대로 중앙영구기록물관리시스템(CAMS: Central Archives Management System)에 적용되어 기록서비스에 활용되고 있다. 이는 처리과 생산 당시의 분절적·파편적으로 존재했던 기록이 국가기록원으로 이관되어도 그대로 존재하여 역사적·학술적 활용 가치가 있는 기록이 빛을 보지 못하고 사장되는 것을 의미한다(최철민, 2018). 국가기록원은 기록물의 이관 시 기록물관리 담당자가 기록물 분류정보를 확인하고 품질을 일정 수준으로 유지토록 해야 하지만, 이관된 기록물의 분류는 기록물 정리/기술 업무로 편성되어 있고, 대부분의 정리/기술 업무는 수작업에 의존하고 있어 당해 연도에 처리해야 할 기록물 수량을 맞추기 어려운 실정이다. 특히 2007년 본격적인 전자정부법 시행에 의해 공공기록의 전자적 생산 환경이 보편화되고, 그에 따라 2017년 이후 국가기록원으로 이관되는 전자기록물 역시 기하급수적으로 늘어나고 있는 상황에서 이관된 기록물의 분류체계를 검증하고 일관되게 관리하는 업무는 매우 중요하다.

이에 본 연구는 이관 기록물 분류 업무의 효율화와 일관된 기준을 유지하기 위한 방안을 제안하고자 한다. 이를 위해 먼저 국가기록원에서 수행하고 있는 현행의 기록분류 및 기술 업무 프로세스를 분석하고 문제점을 파악, 개선 요구사항을 수렴하였다. 또한 이관된 기록물의 분류를 검증하기 위해 담당자들이 참조해야할 자료의 유형을 규명하고, 현재 국가기록원에서 관리되고 있는 공통지식자원의 현황을 파악하였다. 분석 결과를 종합해 기록분류 업무의 수작업을 최소화하기 위한 방안으로 이관된 기록물의 편철 정보, 즉 목록에 기반한 분류 이상치 후보를 판단하는 과정을 도출·체계화하였다. 나아가 제안한 이상치 판별 프로세스를 실제 국가기록원으로 이관된 기록물을 대상으로 적용하고, 그 결과를 규격화하여 추후 기계학습에 활용 가능한 학습데이터 형식으로 구축하였다. 본 연구는 선행연구인 생산기관 직제분석 자동화(강윤아 외, 2021)의 후속 연구로, 연구의 궁극적인 목적은 지능형 전자기록 관리 환경의 토대를 마련하기 위해 기록관리 업무 내 기계학습 기법이 적용 가능한 문제 유형을 선별하고 자동화하는 데 있다.

1.2 선행연구

본 절에서는 기록물 분류 업무에 대한 연구와 기록관리 지능화와 관련된 선행연구를 살펴본다. 먼저 기록물

http://dx.doi.org/10.14404/JKSARM.2022.22.1.043

분류 업무와 관련한 연구로, 설문원(2013)은 정부기능분류체계가 기록분류에 어떻게 적응되고 있는지를 살펴보기 위해 기록관리전문직과 집단 면담을 진행하여 기록물 철 기반 분류의 실태와 문제점을 구조 및 운용 측면에서 분석하였다. 윤상우(2020)는 A기관을 중심으로 기록분류체계 및 기록물분류기준표 관리 현황과 업무기능을 분석하였으며, 이를 통해 기능 분류 모델 및 기록관리기준표의 도입과 역할과 구축 방안에 대해 연구하였다. 장현종, 노지현(2021)은 국립대학에서 사용하는 기록물 분류체계의 운영현황을 공통 단위과제 중심으로 분석하고 분류체계의 개선과 운영절차의 개선이 필요하다고 제안하였다.

기록관리 지능화와 관련된 연구로 장지숙, 이해영(2009)은 기 분류된 기록집합체뿐만 아니라 분류체계와 시소러스를 분류기준으로 함께 구축하여 상호 보완할 수 있도록 맥락정보를 이용한 자동분류시스템을 설계하였다. 오진관(2019)은 해외 기록시스템의 자동화, 지능화 사례를 살펴보고, 디지털 환경에 적합한 영구기록 관리 및 서비스 방안을 제안하였다. 김인택, 안대진, 이해영(2017)은 4차 산업혁명에 따른 가장 핵심이 되는 인공지능(이하 AI) 기술 중 세 가지 – 텍스트분석, 영상인식, 음성인식 – 를 선정, 기록관리 업무에 활용할 수 있는 방안에 대해 제안하였다. 또한 이들 기술을 활용하기 위한 선결해야 하는 과제들을 제시하였다. 최정렬(2018)은 표준화된 절차에 따른 학습데이터세트 구축 방안을 연구하였다. 학습데이터가 갖추어야 할 요구사항을 문제 유형과 데이터 유형별로 분석하였고, 이를 토대로 기계학습에 활용하기 위한 학습데이터세트 구축에 관한 참조모델을 제안하였다.

기록물 분류 업무 자동화 관련한 연구로, 김해찬솔 외(2017)는 문헌의 자동분류와 AI 학습방식이 발전해 온 과정을 살펴 본 후, 기계학습 중 특히 지도학습 방식을 통한 AI 기술을 기록관리 분야에 적용해야 하는 필요성을 도출하였다. 또한 서울시의 결재문서 원문을 대상으로 실제 학습데이터를 구축하여 자동분류 모델을 개발하고, 각 과정에서 고려해야 할 사항들을 제시하였다.

상기 연구들 대부분은 기계학습을 이용한 전자기록물의 자동분류, 원문인식, 공개재분류 등, 기록관리 개별 업무 단위로 AI 기술을 적용해 자동화를 꾀하고 있다. 대부분이 기록물 '원문' 전체를 대상으로 태깅하거나 정답 분류를 할당하는 방식으로, 본 연구에서와 같이 기록물의 일부 메타데이터로 구성된 '목록'기반 자동화와 관련된 연구는 전무하다. 특히 기록관리 업무에서 실무자들에게 공통으로 활용 가능한 기초자료들을 참조하여 오분류된 기록물 이상치 후보를 제안하고, 기록분류 실무자가 전문(full-text) 열람 여부를 사전에 판단케 함으로써 가능한 한 수작업을 최소화하는 업무 효율화를 꾀한다. 또한 기록분류 실무자의 업무 프로세스를 다수의 작업자가 답습하여 진행하는 과정을 통해 자동화할 수 있는 부분을 규명하고 기계학습에 활용 가능한 학습데이터를 구축하였다.

2. 기록분류 관련 업무 현황 분석

2.1 영구기록물관리기관 업무 운영 절차

영구기록물관리기관은 「공공기록물 관리에 관한 법률」에 명시된 고유의 업무 기능을 수행하여야 한다. <그림 1>은 영구기록물관리기관의 기록물 관리 절차(NAK 9:2021(v2.2))를 도식화 한 것으로, 운영 절차는 총 8단계로 구성되어 있다.

본 연구는 영구기록물관리기관으로 이관된 기록물의 분류 자동화를 위한 사전 연구로, 총 8단계의 업무 운영 절차 중 '기록물 기초데이터 관리', '기록물 수집', '기록물 정리・기술' 단계에 중점을 두었다. 먼저 '기록물 기초데이터 관리' 단계에서 생산현황 관리 및 활용과 소장기록물의 통계를 작성・관리한다. 기록물 생산현황은 향후 기록물의 수집 및 이관 방침 수립, 보존량 예측을 위한 기초자료로 활용하며, 기록물 생산량 변동 분석 등을 통해 생산단계에 대한 통계조치로도 활용한다(국가기록원, 2021b).

http://dx.doi.org/10.14404/JKSARM.2022.22.1.043

〈그림 1〉 영구기록물관리기관 업무 운영 절차

‘기록물 수집’ 단계에서는 기록물 수집정책을 마련하고 법적 이관 대상 기록물을 인수하며 이관지침에 따라 정리 및 편철되었는지를 점검한다. 영구기록물관리기관으로 이관된 기록물은 생산된 지 10년이 지난 기록들로, 그때의 생산기관의 특성과 직무 등 생산맥락을 이해해야할 뿐 아니라 기록물 관리기관으로 이관된 시점의 변천이력 등의 맥락도 병행해서 파악해야만 한다. 이때 참조할 수 있는 자료는 기록물 관리를 담당하는 실무자가 공동으로 참조해야하는 기초지식들로, 본 연구의 선행연구(강윤아 외, 2021)를 통해 다음과 같은 유형이 규명되었다. 생산기관의 실·국 단위의 흐름을 보여주는 ‘조사대상 하위기관 존립기간 목록’과 해당 기관의 직제 및 기능분석 업무의 최종 산출물인 ‘하위기관 변천내용’, ‘계열별 기능어 분류표’를 이 단계에서 활용할 수 있다. 본 연구에서는 이러한 자료를 ‘공통지식자원’으로 명명하고자 한다.

다음 단계인 ‘기록물 정리·기술’에서는 편철정보를 비롯해 해당 생산기관에서 작성한 메타데이터 값이 정확하게 입력되었는지를 확인하고 소장 기록물을 분류한 후 기술한다. 이 과정에서 수집된 기록물을 분석하여 기록물을 출처와 원본질서, 생산맥락에 따라 분류하는 업무를 수행한다. 상기한 바와 같이 이관된 기록물의 생산맥락은 최소 10년 전의 정보로, 이를 정확하고 완전하게 파악하기 위해서는 ‘조사대상 하위기관 존립기간 목록’, ‘그룹별 하위기관 목록’, ‘그룹별 직제령 비교표’, ‘하위기관 변천내용’, ‘계열별 기능어 분류표’ 등의 자료를 참조해야 한다(강윤아 외, 2021). 특히 활용되는 공통 자원 중 ‘그룹별 하위기관 목록’과 ‘그룹별 직제령 비교표’의 중요도가 비교적 높은데, 이 자료들은 이미 동일한 생산맥락을 갖는 하위기관을 그룹화하는 과정에서 산출된 결과물이기 때문에 기록물 분류·기술의 가이드라인 역할을 할 수 있다.

2.2 심층면담

상기한 기록물 분류와 관련된 업무 현황을 파악하기 위해 현재 해당 업무를 수행 중인 국가기록원 담당자와의 기초 인터뷰[1]를 진행하고 실제 CAMS 열람을 진행하였다. 기초 분석 결과, 이관 기록물에 대한 관리 과정 대부분이 수작업으로 진행되고 있고, 참조하고 있는 자료 역시 산발적으로 관리되고 있어 많은 애로사항이 있는 것으로 파악되었다. 특히 2016년 이후 전자적으로 생산된(born-digital) 기록물이 대량으로 입수되고 있으나, 기록분류 담당자가 매년 검토해야 할 기록물의 양이 너무 방대해 당해 연도 내 전수조사는 불가능한 실정이다. 이로 인해 일부 생산기관과 기록물의 중요도에 따라 우선순위를 부여하여 검수하고 있으며, 이 과정에서 기록의 맥락을 이해하기 위해 원문을 열람하거나 생산기관 전거데이터를 확인하는 데 많은 시간을 할애하는 것으로 파악되었다. 또한 새로운 생산기관을 담당할 때마다 기관의 직제 및 변천 정보 등을 파악하는 작업이 반복해서 수행되고 있었다(강윤아 외, 2021).

구체적인 업무 프로세스와 지능화 개선 요구사항을 수렴하기 위해 국가기록원의 기록분류 실무자 및 CAMS

1) 2021년 4월 29일 ~ 2021년 5월 21일 국가기록원 방문(대전, 성남) 총 3회 수행, 면담자: 5명

http://dx.doi.org/10.14404/JKSARM.2022.22.1.043

담당자와 심층면담(이하 FGI: Focused Group Interview)및 업무 참관)[2]을 진행하였다(국가기록원, 2021a). 기초인터뷰 및 심층면담이 다수로 진행된 이유는 실무자의 업무 과정을 작업자가 답습하는 과정과 실제 CAMS를 열람하는 과정에서 도출된 문제점이나 특이점에 대한 재확인 과정이 반복되었으며 세부업무별 담당자 역시 세분화되어 있어 개별 면담이 필요하였다. 면담 결과, 이관된 기록물의 분류를 검증하는 과정은 먼저, 이관시 작성된 목록을 통해 기록물에 부여된 메타데이터를 검토하여 맥락을 파악한다. 그러나 CAMS 열람을 통해 확인한 결과, 애초 기록생산기관에서 이관 시 제출한 기록물 메타데이터 값이 부실하여 목록만으로 내용을 파악하기 어려웠다. 특히 CAMS 내에서 관리되고 있는 생산기관의 BRM 정보와 실제 이관된 기록물의 편철정보가 매핑되지 않는 경우가 다수 발견되었으며, 이를 수작업으로 정리하고 있었다. 이로 인해 기록물 분류 검증을 '건'별로 진행하는 것에 많은 부담을 느끼고 있었다. 또한 해당 기록의 담당자가 기록분류 이후 보존기간 재평가 업무도 함께 수행하고 있어 '철' 단위 위주의 검토를 진행하고 있었다.

또한, 기록의 맥락을 파악하기 위해 참조해야 할 공통지식자원 역시 일관된 지침이 부재한 채 관리되지 않고 있었다. 대표적으로 CAMS 내의 전거데이터는 2015년 이후 갱신되지 않아 현행화되지 않았으며, 이로 인해 생산기관의 변천이력을 파악하기 위해서는 담당자가 법령정보사이트(https://www.law.go.kr/)나 정부조직관리정보시스템(https://www.org.go.kr/) 등을 별도로 접속, 개별적으로 관련 지식을 구축하여 활용하고 있었다(강윤아 외, 2021). 또한 일련의 업무를 수행하는 과정이 각 담당자의 고유한 작업 경험에 의존, 나름의 개별적인 노하우로 수립되어 있었다. FGI를 통해 상기한 문제점을 포함하여 기록물 정리·기술 과업과 관련한 업무 프로세스 정립과 재편철, 이상치 판별 등 3가지 업무 지원 분야에 대한 요구 사항이 도출되었으며 자세한 내용은 <표 1>과 같다.

〈표 1〉 실무 담당자 심층 면담 결과

업무 기능	분류/기술
면담 대상	실무 담당자
면담 결과	1) 업무 프로세스 정립 - 분류/기술을 위한 참고자료 입수 - 필수 메타데이터 정의 및 자동 추출 방안 모색 - 작업 매뉴얼의 기계화
	2) 재편철 - '단위과제와 단위과제카드의 문제점'을 인식 후, 보정 작업 필요 - 보존기간 및 오분류의 재편철 요구 - 건에 대한 태깅 작업 진행
	3) 이상치 판별 - 분류체계 오류가 있는 기록물 판별 및 지능화 방안 필요(원문 열람 최소화) - 오분류인 경우 유사한 기능 추천

심층 면담을 통한 요구사항으로 먼저 '업무 프로세스 정립' 분야에서는 분류/기술을 위해 참고하는 자료를 발굴하고 이를 자동 입수할 수 있는 방안과 필수 메타데이터의 정의 및 자동 추출 방안이 필요하다고 하였다. 또한 실제 실무자의 업무 프로세스를 파악하여 하위계열을 신설하는 경우와, 전자기록 생산 단계 시 부여된 구조를 인정하고 이를 CAMS에 바로 적용할 수 있는 방안을 요구하였다. 마지막으로 반복되는 작업 중 기계화 가능한 부분을 파악하고 전체 프로세스를 체계화하여 매뉴얼 형태로 도출해주기를 요청하였다.

2) 2021년 6월 21일 ~ 2021년 7월 28일 국가기록원(대전, 성남) 방문 및 온라인 회의 총 7회 수행, 면담자: 7명

http://dx.doi.org/10.14404/JKSARM.2022.22.1.043

‘재편철’ 분야에서는 ‘단위과제와 단위과제카드의 문제점’을 인식하고, 이를 보정하는 것이 목표이며, 기록철 단위로 물리적 관리단위를 정한 것에 대한 재검토가 필요하다고 하였다. 보존기간 재평가와 관련하여 보존기간이 잘못 부여된 데이터를 처리하고, 오분류로 판단되는 경우 재편철을 요구하였으며, 각 건에 대한 보정결과를 태깅하는 작업이 필요하다고 하였다.

마지막으로 ‘이상치 판별’ 분야에서는 앞선 재편철 단계에서 오분류로 판단되는 사례 중 분류체계 자체 오류에 대해서는 생산부서, 철・건 제목 등을 참조하여 유사한 기능을 최대한 추천해주길 요구하였다. 무엇보다도 오분류 이상치 후보를 탐지함으로써 원문 열람을 최소화할 수 있는 방안을 필요로 하였다.

3. 이관 기록물 분류 이상치 판별 자동화

실무 담당자와의 심층 면담을 통해 도출한 개선 요구사항에 대한 대응 방안은 다음과 같다. 먼저 현재까지 기록분류 담당자별로 각자 진행하던 업무 과정을 수렴하고, 기록의 맥락을 이해하기 위해 참조해야 하는 기초자료들을 파악한다. 그 과정을 통해 기존에 반복적으로 수행되거나 수작업에 의존한 작업 중 자동화가 가능한 부분을 규명한다. 특히 기록물 분류 이상치를 판별하기 위해 건별 검수를 진행하기 전, 철 단위 검토를 먼저 수행하는 프로세스를 정립하고 철 목록 검토 사항을 승계하여 건별로 이상치 판별 작업을 진행한다. 또한 일련의 과정을 수행한 결과물을 취합하고, 검수하는 과정을 통해 오류 유형을 정리함으로써 향후 기계학습에 활용 가능한 고품질의 학습데이터 구축 방안을 제안하다. 나아가 제안한 프로세스를 다수의 작업자가 수행 시 업무 일관성이 유지될 수 있도록 매뉴얼을 제작한다.

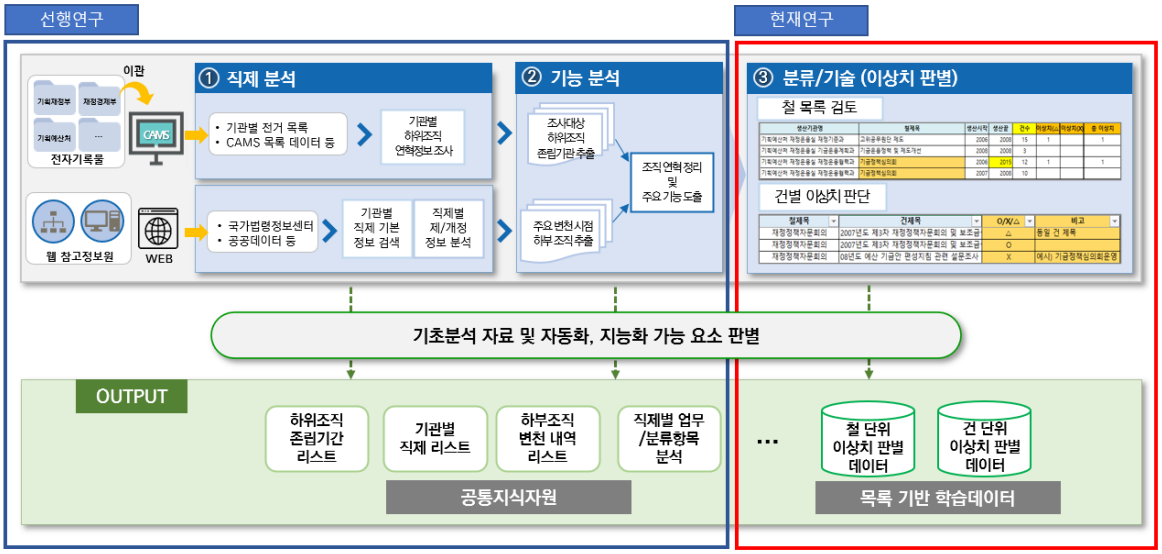
3.1 분류 이상치 판별 프로세스

<그림 2>는 본 연구에서 제안하는 기록물 분류 검증 프로세스를 도식화한 것이다. 이관된 대량의 전자기록물의 분류 이상치 탐지를 위해서는 먼저 해당 기록을 생산한 기관에 대한 직제분석이 선행되어야 한다. 이후 해당 기관에서 수행하는 기능분류체계를 참조하고, 주요 변천 내역을 파악해야 한다. 이러한 과정에서 공통으로 활용되는 기초 자료로는 하위조직 존립기간 리스트, 직제리스트, 기록관리기준표 등이 있다(강윤아 외, 2021). 이러한 공통지식자원을 기반으로 이관된 기록의 철 및 건의 주요 메타데이터를 자동으로 목록화하고 검증함으로써 오분류 가능성이 있는 기록물 건을 선별한다. 이후 최종 오분류 여부는 이상치로 선별된 기록물에 대한 원문 및 세부정보를 CAMS에서 참조하여 담당자가 판단한다.

<그림 2>의 ① 직제 분석 과정과 ② 기능 분석 과정은 본 연구의 선행연구(강윤아 외, 2021)를 통해 정립된 과업으로, 영구기록물관리기관에서 공통으로 활용 가능한 기초 자원을 구축하고 생산기관 직제분석 프로세스의 자동화 방안을 제시하였다. 영구기록물관리기관에서 수행하는 기록물 관리 절차 내에서 직제분석 정보가 필요한 단계별로 활용 방안을 제시하였으며, 선행연구에서 제안된 공통지식자원은 영구기록물관리기관의 일관되고 체계적인 업무 수행 지원을 목적으로 하고 있다.

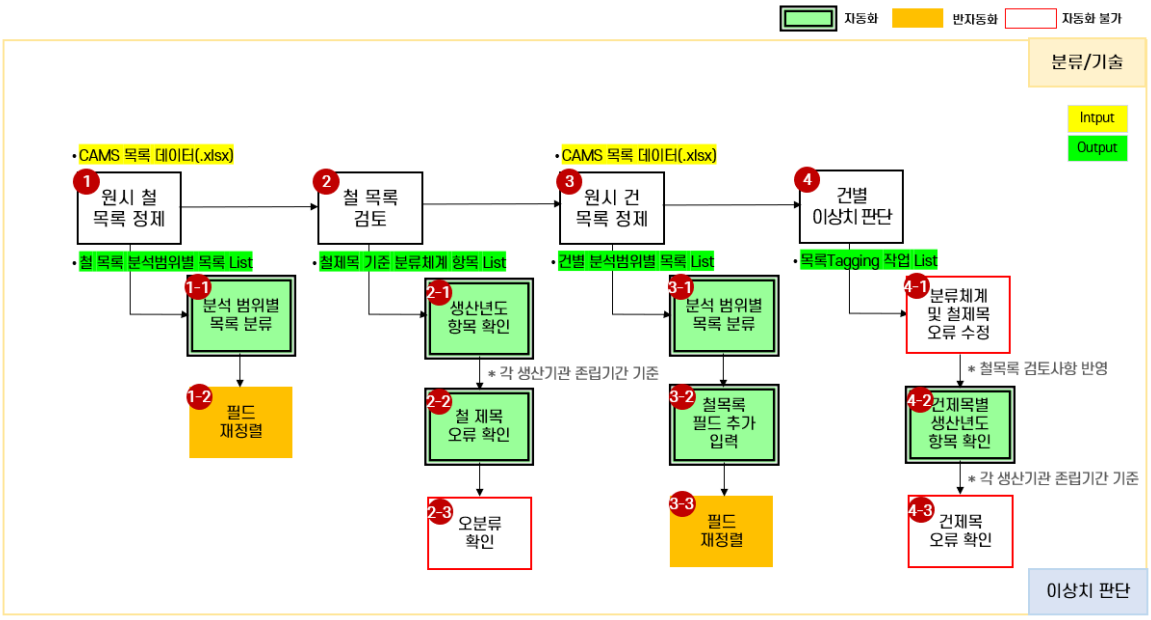
본 연구는 <그림 2>의 ③ 분류/기술 단계(붉은 박스)에 주안점을 둔 것으로, 기록물 철 목록 검토 및 건별 이상치 판단으로 구성된다. 이 단계에서는 강윤아 외(2021)의 연구에서 제안하고 있는 동일한 생산맥락을 갖는 하위기관을 그룹화하는 과정에서 산출된 결과물인 기초 분석 데이터를 참조해야 한다. 이 산출물에서는 생산기관의 변화 흐름과 업무 및 기능 정보를 담고 있으며 앞선 과정에서 수집하고 분석한 내용을 함축적으로 담고 있어 활용도가 높다. 이러한 공통지식자원에 기반하여 대상 철 목록을 검토하고, 단위과제명 및 기관의 존립기간 등의 거시적 검수를 진행하며 이후 세부 기록물 건에 대한 분류 검수를 수행한다.

http://dx.doi.org/10.14404/JKSARM.2022.22.1.043



〈그림 2〉 이관 기록물 분류 검증 프로세스 개요

<그림 3>은 상기한 ③ 기록물 분류 이상치 판별 업무의 세부 과정을 도식화 한 것으로, 이상치 판별에 필요한 데이터의 특성과 업무 수행 난이도에 따라 자동화・반자동화・자동화 불가로 구별된다. 세부 과정은 다음과 같다.



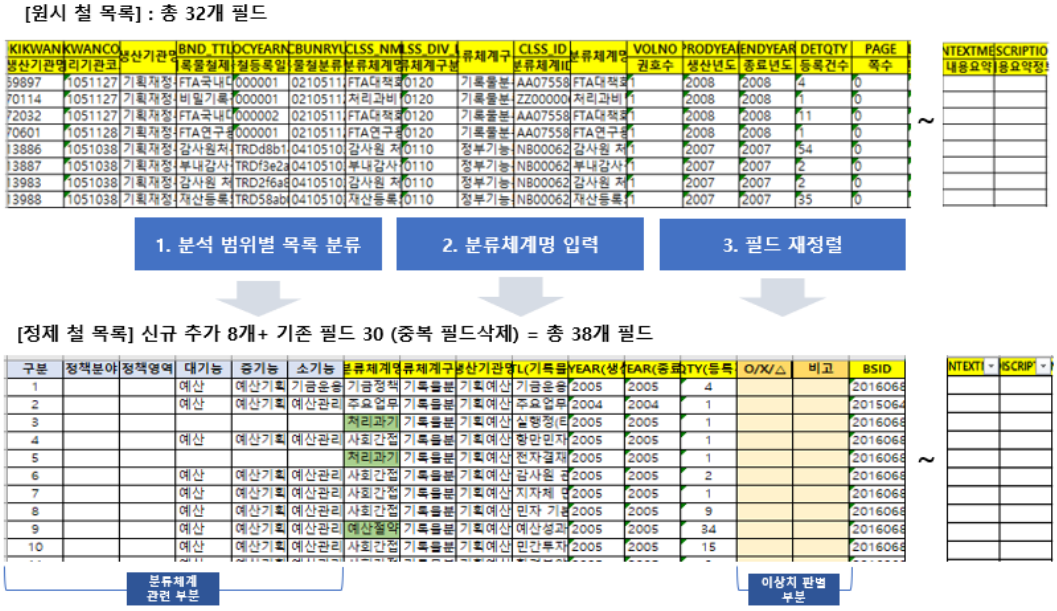
〈그림 3〉 이상치 판별 세부 프로세스

3.1.1 철 목록 검토

먼저 철 목록 검토에 앞서 국가기록원에서 제공한 CAMS 원시 철 목록에 대한 정제 작업이 필요하다(<그림 3>의 1 단계 참조). 원시 철 목록 정제를 위해 각 분석하려는 해당 생산기관별로 철 목록을 구분한 후 각 기관의 기록관리기준표, 기록물분류기준표를 참고하여 자동으로 추출하고, 엑셀(excel)에서 처리 가능한 형태로 출력

http://dx.doi.org/10.14404/JKSARM.2022.22.1.043

한다(<그림 3>의 1-1). 이후 분류체계명, 분류체계구분, 생산기관명, 기록물철제목, 생산년도, 종료년도, 이상치 구분, BSID[3]의 순서로 정렬한다(<그림 3>의 1-2). 해당 작업들은 간단한 코딩과 엑셀 기능으로 반・자동화가 가능하다. <그림 4>는 상기한 정제 과정을 거쳐 작성된 철 목록 예시이다. 기존에 제공받은 필드 32개 중 중복된 필드를 제외한 30개의 필드(노랑색 표시)와 신규 필드 8개를 추가하여 총 38개의 필드로 작업을 진행하였다. 신규로 추가한 필드는 총 8개로 분류체계와 관련한 필드 6개(하늘색), 이상치 판단을 위한 2개의 필드(주황색)이다.



〈그림 4〉 철 목록 정제 작업 예시

다음 단계는 정제된 목록을 대상으로 검토 작업이다(<그림 3>의 2). 가장 먼저 분석 대상 하위기관의 존립기관 목록과 전거데이터를 기준으로 생산년도 항목을 확인 한 후(<그림 3>의 2-1) 철 제목에 오타나 띄어쓰기의 오류가 있는지 확인(<그림 3>의 2-2)한다. 상기한 작업들은 간단한 문자열 비교 및 파일 처리로 자동화가 가능하다. 철 목록 검토를 마친 후 각 기관별 철 목록에 분류체계명을 구분하여 이상치 유형을 판별해야 한다(<그림 3>의 2-3). 이때 철 목록 이상치 판별 유형의 자세한 기준은 <표 2>와 같으며, 판단 기준 난이도에 따라 자동화 가능 여부를 구분하였다.

〈표 2〉 철 목록 이상치 판별 기준

구분		이상치 후보 유형	기준	자동화 가능 여부
X	1	생산년도 오류/종료년도 오류	조사대상 하위기관 존립기관 목록	○
	2	철명 오류	오타 및 띄어쓰기	○
△	1	전거데이터 부족	조사대상 하위기관 존립기관 목록	X

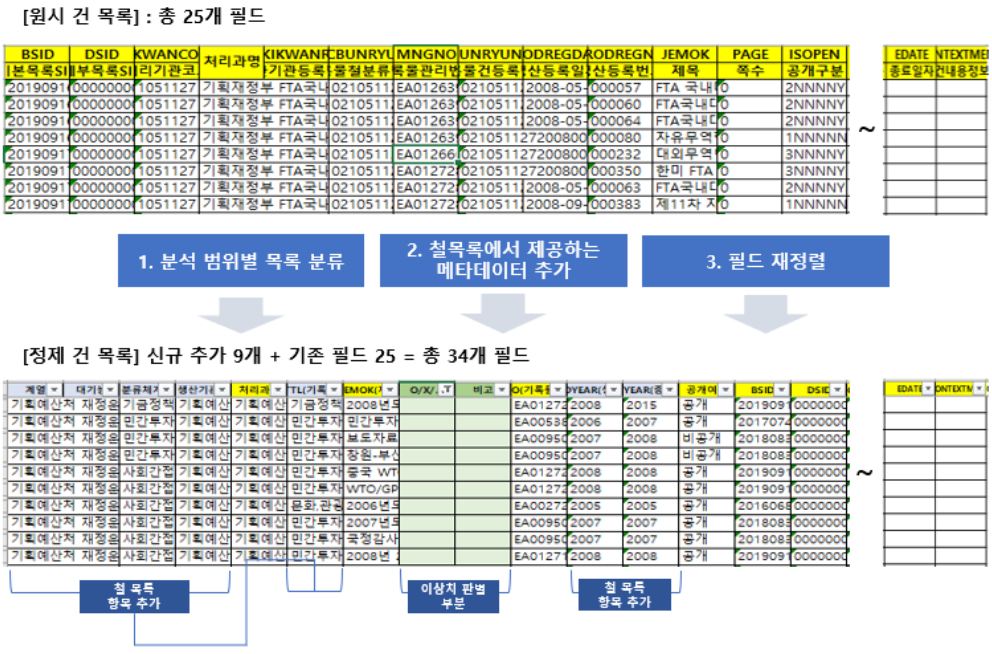
3) CAMS에서 기록물이 등록되고 관리되는 번호로 관리번호는 철단위로 부여되는 총 9자리 숫자로 알파벳 2자리+숫자 7자리로 구성된다. BSID는 철에, DSID는 건에 각각 부여되는 12자리 숫자이다.

http://dx.doi.org/10.14404/JKSARM.2022.22.1.043

이상치 판별 결과는 2가지로, 세부 철 목록을 참조하지 않아도 오분류로 판단 가능한 경우에는 X, 세부 목록 및 원문 확인이 필요한 경우는 △로 표기하였다. 이상치 X 사례로는 생산년도 오류/종료년도 오류가 있으며, 공통 활용 기초분석 자료 중 조사대상 하위기관 존립기관 목록을 기준으로 자동으로 판별 가능하다. 철명 오류 역시 기록물분류기준표나 기록관리기준표를 참조하면 쉽게 판단할 수 있다. 그러나 애초 전거데이터가 부족하거나 CAMS에 이관된 정보에 결락이 발생해 작업자가 오분류 여부를 판단하지 못하는 경우가 다수 발생하였는데, 이러한 유형은 기계학습 기법 적용이 불가하다.

3.1.2 건별 이상치 판단

건별 이상치 판단을 위해 먼저 철 목록 검토와 마찬가지로 국가기록원에서 제공한 원시 건 목록의 정제 작업이 필요하다(<그림 3>의 3단계 참조). 분석하려는 해당 생산기관별로 건 목록을 구분한 후(<그림 3>의 3-1), '철 목록'에서 제공하는 생산기관명, 분류체계명, 철제목, 생산년도/종료년도를 '건 목록'에 승계, 추가 기술한다(<그림 3>의 3-2). 이 후 분류체계명, 생산기관명, 처리과명, 기록물철제목, 건제목, 이상치 구분, 기록물관리번호, 생산년도, 공개여부, BSID, DSID의 순서로 정렬한다(<그림 3>의 3-3). 해당 작업들 역시 철 목록 정제와 같이 간단한 코딩과 엑셀 기능으로 반·자동화가 가능하다. <그림 5>는 정제 파일 부분이 상기한 과정을 거쳐 작성된 건 목록 예시이다. 기존에 제공받은 필드 25개(노랑색)와 신규로 9개의 필드를 추가하여 총 34개의 필드로 작업을 진행하였다. 신규로 추가한 필드는 총 9개로 철 목록에서 가져온 7개 필드(하늘색)와 이상치 판단을 위한 2개의 필드(주황색)이다.



〈그림 5〉 건 목록 정제 작업 예시

건별 이상치 판별 작업(<그림 3>의 4)은 앞선 철 목록 검토사항을 반영하여 분류체계 및 철 제목 오류 사항을 분류하고(<그림 3>의 4-1), 건 제목별 생산년도 각 기관의 생산기관 존립기간을 기준으로 항목을 확인한다(<그림 3>의 4-2). 이 후 건별 이상치 기준(<표 3> 참고)에 따라 O/X/△를 기준에 맞게 선택하여 표기한다(<그림 3>의 4-3).

http://dx.doi.org/10.14404/JKSARM.2022.22.1.043

〈표 3〉 건 별 이상치 판별 기준

구분		이상치 후보 유형	기준	자동화 가능 여부
X	1	분류체계 오류	계열별 기능어 분류표	X
	2	생산년도 오류/종료년도 오류*	조사대상 하위기관 존립기관 목록	○
	3	철_건 생산기관명 불일치	'생산년도 오류/종료년도 오류'인 경우 철 목록의 생산기관명과 건 목록의 처리과명 비교	X
	4	철명 오류*	오타 및 띄어쓰기	○
△	1	메타데이터만으로 확인 불가	메타데이터 부족으로 인해 건 제목만으로 해당 철과 분류체계에 해당되는 내용인지 확인 할 수 없는 경우	X
	2	메타데이터만으로 확인 불가/공통	건 제목이 단순 기재되어 있는 경우	X
	3	전거데이터 부족*	조사대상 하위기관 존립기관 목록	X

*: 철 목록 이상치 승계

건 별 이상치 후보 유형으로는 철 목록 검토 사항을 반영한 후 건 제목 오류를 확인하여 이상 없는 경우 O, 오분류로 확인한 경우 X, 확인이 필요한 경우 △로 구분하였다. 이상치 X는 분류체계 오류, 생산년도 오류/종료년도 오류, 철_건 생산기관명 불일치, 철명 오류 등의 사례가 있으며, 이상치 △는 메타데이터만으로 확인 불가, 메타데이터 확인불가/공통, 전거데이터 부족 사례가 있다. 이상치 후보 중 CAMS 원문을 열람해서 확인해야 하거나 애초 메타데이터 등 관련 데이터가 결락되어 판단이 불가한 경우에는 기계학습 기법 적용 역시 불가하므로 자동화할 수 없다. 다만 철 목록 검토 결과와 마찬가지로 생산년도 오류/종료년도 오류, 철명 오류 이상치는 자동화가 가능한 유형이다.

3.2 분류 이상치 판별 학습데이터 구축

본 연구는 궁극적인 목적은 이관된 기록물의 분류 이상치를 자동으로 판별하기 위한 기계학습 기술 개발에 있다. 이를 위해서는 원천이 되는 학습데이터 구축이 필수적이다. 본 연구에서는 3.1장에서 정의된 기록물 분류 이상치 판별 프로세스 과정을 실제 국가기록원에 이관된 전자기록물을 대상으로 적용, 그 결과를 학습데이터 형태로 구축하였다.

3.2.1 학습데이터 구축 대상 기록물 선별

본 연구 대상 기록물은 전자문서만을 대상으로 하였으며, 2016년부터 2020년도까지 이관된 기록물 중 경제 관련 기관에서 생산한 기록물 22,021건을 학습데이터 구축 기록물로 선별하였다. 경제 관련 기관을 특정한 이유는 먼저, 실제 국가기록원에서 해당 업무를 담당하고 있는 실무자의 요청이 있었으며, 분류 이상치 판별을 위해 필요한 공통지식자원인 생산기관 변천정보 및 직제 분석 자료 등의 완성도가 상대적으로 높게 관리되고 있었기 때문이다. 더불어 국가기록원으로 이관된 기록물 수량이 비교적 많고 국민들의 관심도가 높은 기관이라는 점도 고려되었다.

<표 4>는 학습데이터 후보 집합을 기관 이력 및 규모에 따라 4그룹으로 나눈 것으로, 전체 22,021건 중 각 그룹별 건수는 [1] 재정운용 예산관리 관련 그룹이 1,607건, [2] 공공혁신 관련 그룹 1,189건, [3] 세제실 관련 그룹 15,304건, [4] 국고국 관련 그룹 3,921건으로 집계되었다. 이 중 [3] 그룹의 경우 1만5천 건 이상의 기록이 편중되어 있어, 실무자와의 협의를 통해 철 별 등록 건수가 100건 이상인 경우에는 각 15건, 100건 미만과 10건 이하 철은 각각 5건, 1건씩 산정하였다. 최종 학습데이터 구축 대상 기록물은 2,407철의 10,409건으로 선별되었다.

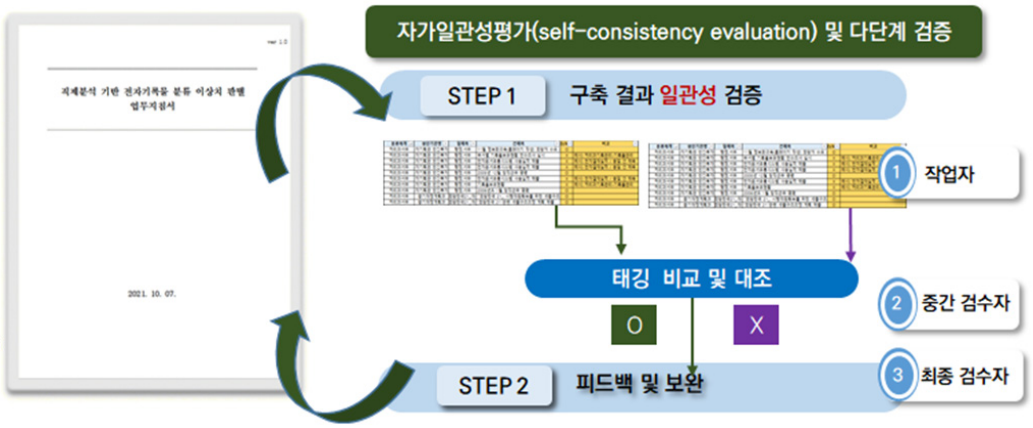
http://dx.doi.org/10.14404/JKSARM.2022.22.1.043

〈표 4〉 그룹별 구축 대상 철/건 수

구축 범위(그룹)	후보		학습데이터 구축 대상	
	철 수	건 수	철 수	건 수
[1] 기획예산처 재정운용실, 예산실, 예산관리국, 기금정책국, 성과관리본부 및 기획재정부 예산실	170	1,607	170	1,607
[2] 기획예산처 공공혁신국, 공공혁신본부 및 기획재정부 공공정책국	145	1,189	145	1,189
[3] 재정경제부 세제실 및 기획재정부 세제실	1,804	15,304	1,804	3,332
[4] 재정경제부 국고국 및 기획재정부 국고국	288	3,921	288	3,921
총계	**2,407**	**22,021**	**2,407**	**10,049**

3.2.2 학습데이터 검증

고품질의 학습데이터를 구축하기 위해서는 다수의 작업자들의 결과를 취합, 검증하는 과정을 통해 오류 유형을 정리하고 이를 재학습시킴으로써 일관성을 유지하는 과정이 필요하다. 본 연구에서도 <그림 6>과 같이 3.1 절에서 정립한 이관 기록물 분류 이상치 프로세스를 매뉴얼화하여 다수의 작업자에게 배포하고, 실제 적용시 명확한 기준을 수립하도록 교육하였다. 또한 개별 작업자의 결과물을 취합, 중간검수, 최종 검수자로 나누어 다단계 비교함으로써 결과의 일관성을 검증하였다. 최종적으로 이상치 판단 결과를 실제 국가기록원 실무자에게 제시하여 이상치 분류 정확도를 평가받고 제안한 프로세스의 유용성과 학습데이터의 일관성을 검증받았다.



〈그림 6〉 학습데이터 검증

본 연구에서는 중간 검수자 2명에 각각 2명의 작업자를 할당, 최종 검수자까지 총 7명의 작업자가 학습데이터를 구축하였으며, 초기에는 작업자 모두 같은 기록물을 판별하여 결과를 검증하는 단계를 거쳐 오류를 보정하는 재학습 과정을 반복 수행하고 업무지침서 및 학습데이터 구축 가이드라인을 보정해 나갔다.

3.2.3 학습데이터 구축 결과

<표 5>는 상기 과정(<그림 3> 및 <그림 6>)을 통해 연구 대상 4개 그룹의 이관 기록물의 이상치를 판별한 결과이다. 먼저 그룹별 철 목록의 이상치 수는 전체 2,407철 중 1,971철(81.8%)이 이상치 후보로 판별 되었다. 오분류로 확인된 경우(X)는 155철, 세부 확인이 필요한 경우(△)는 1,816철로 집계되었다. 건별 이상치 건수는 전체 10,049건 중 7,818건(77.7%)이 분류 이상치 후보로 판별되었다. 특히, [3]그룹의 경우 다른 그룹에 비해 이상치율이 약 95% 이상으로 높게 나타났다. 이상치 유형을 세부적으로 분석해본 결과, 건별 이상치(△)사례 2,625건

http://dx.doi.org/10.14404/JKSARM.2022.22.1.043

중 2,035건이 '전거데이터 부족'인 사례로 파악되었다. 이는 [3]그룹인 재정경제부 세제실 및 기획재정부 세제실의 대부분의 전거데이터가 제대로 구축되어 있지 않아서 생긴 문제로, 이 같은 결과는 기록관리를 위한 공통지식자원이 정확하고 일관되게 관리되어야 할 필요성이 있음을 반증하는 예이다.

〈표 5〉 분석범위 별 이상치 철/건수

구축범위(그룹)	철 수				건 수			
	대상	이상치 (△)	이상치 (X)	이상치율 (이상치/대상)	대상	이상치 (△)	이상치 (X)	이상치율 (이상치/대상)
[1]	170	46	15	35.8%	1,607	454	289	46.2%
[2]	145	54	6	41.3%	1,189	698	220	77.2%
[3]	1,804	1,619	97	95.1%	3,332	2,625	585	96.3%
[4]	288	97	37	46.5%	3,921	1,291	1,656	42.2%
총계	2,407	1,816 (75.4%)	155 (6.4%)	1,971 (81.8%)	10,049	5,068 (50.4%)	2,750 (27.3%)	7,818 (77.7%)

전체 1만여 건 중 오분류로 확인된 기록물(X)은 총 2,750건(27.3%), 세부 확인이 필요한 경우(△)가 5,068건(50.4%)으로 집계되었다. 이는 이관 기록물에 부여된 분류에 대해 담당자의 검수를 필요로 하는 기록물의 수량이 78%에 달함을 의미하는 수치로, 기록분류 업무의 난이도가 그만큼 높음과 해당 업무의 효율화, 지능화 필요성을 반증한다. 또 다른 한편으로 기존에 전수조사에 대비해 오분류가 비교적 명확한 경우(X, 27%)와 분류 이상치가 발견되지 않은 정상치(100-78=22%)는 담당자의 검수 우선순위를 후순위로 하거나 제외함으로써 수작업을 최소화할 수 있음을 함의한다.

오류 유형 분포는 <표 6>과 같이 '생산년도 오류/종료년도 오류' 2,038건 (31%), '철_건 생산기관명 불일치' 2,020건 (31%), '분류체계 오류' 938건(14%), '전거데이터 부족' 883건(13%), '메타데이터만으로 확인 불가' 736건 (11%), '철명 오류' 1건(0%)의 순서로 나타났다. 건 목록 이상치(X) 결과에는 철 목록 검토 사항을 승계 받아 '생산년도 오류/종료년도 오류', '전거데이터 부족', '철명 오류' 사항이 포함되어 있으며, 이상치(△)의 사례 기준인 '전거데이터 부족'이 결과에 반영된 이유는 해당하는 기록물이 오분류로 판별되고 더불어 생산기관의 전거데이터가 미구축 된 사례가 포함되어 있기 때문이다.

〈표 6〉 건 목록 이상치(X) 결과

	철 목록 판별 승계			건 별 판별		
구분	1) 생산년도 오류 / 종료년도 오류	2) 철명 오류	3) 전거데이터 부족	4) 분류체계 오류	5) 철_건 생산기관명 불일치	6) 메타데이터만으로 확인 불가
총계	2,038 (31%)	1 (0%)	883 (13%)	938 (14%)	2,020 (31%)	736 (11%)
	중복계상 6,616건(총 이상치X: 2,750건)					

* 자동화 가능

특히 철 목록에서 승계받은 이상치 결과 중 '생산년도 오류/종료년도 오류'의 이상치가 다소 높아 업무 담당자와 인터뷰를 진행하여 원인을 파악한 결과, 전자문서시스템으로 생산된 기록물의 경우 종결 시점이 생산년도/종료년도 철 정보로 입력되기 때문에 온나라 문서시스템으로 생산된 기록물의 생산년도/종료년도와 차이가 날 수 있으며, 이러한 생산시스템 정보와 차이가 있음을 알고 있으나 이관시 이를 그대로 승계해 발생한 문제로 파악되었다.

http://dx.doi.org/10.14404/JKSARM.2022.22.1.043

구체적인 이상치(X) 예시는 <그림 7>과 같다. 먼저 '1) 생산년도 오류/종료년도 오류'는 조사대상 하위기관 존립기관 목록을 참조해 판별되었다. 이 기록물의 생산년도, 종료년도는 2004-2004(년도)로 입력되어 있었지만 기록물 생산기관 변천정보시스템과 국가기록원 전거목록을 확인 하였을 때, 기획예산처 재정운용실의 존립기간은 2005-2008(년도)로 나타나 오분류로 판별되었다. '2) 철명 오류'는 오타 및 띄어쓰기 기준으로 판별 가능하다. '3) 전거데이터 부족'은 기록물생산기관변천정보시스템, 국가기록원 전거목록을 확인한 후 조사대상 하위기관 해당 존립기관 목록에 존재하지 않는 경우에 해당한다. '4) 분류체계 오류' 사례의 경우, 이 기록물은 기존에 '처리과업무계획(보고)평가'로 편철되어 있었다. 그러나 앞선 <표 3>에서 제시한 기준에 의해 '계열별 기능어 분류표'를 기반으로 기록물의 생산기관, 처리과, 철 제목, 건 제목 등을 확인한 결과, 오분류로 판별되었으며 '민간투자사업 기본계획 운용'이라는 정답 분류체계를 추천하였다. '5) 철_건 생산기관명 불일치' 사례는 '생산년도 오류/종료년도 오류'로 판별이 난 기록물을 대상으로 철 목록의 생산기관명과 건 목록의 처리과명을 비교하여 이상치를 판별되었다. 마지막으로 '검토 부탁드립니다'라는 제목의 기록물의 경우, 제목만으로는 해당 철과 분류체계에 적합한 내용인지 확인 할 수 없어 '6) 메타데이터만으로 확인 불가'의 사례로 판별하였다.

구분	분류체계명	생산기관명	처리과명	BND_TTL(기록물철제목)	JEMOK(제목)	)/X/	비고	록	생산	종료
1	처리과서무	기획예산처 재정운용실 민간투자	기획예산처 예산관리국 민간투	인프라펀드 설립'운영	연기금의 SOC 투자 활성화 방안 연구용역	X	생산년도 오류 / 종	EA	2004	2004
2	처리과서무	기획예산처 재정운용실 중기재정	기획예산처 재정기획실 기획총	국가재정운용계획 수립	"공공갈등과 참여적 의사결정" 포럼 개최	X	철명 오류(띄어쓰기	EA	2005	2005
3	농어업용면세유 한도량	기획재정부 세제실 재산소비세정	재정경제부 세제실 재산소비세	농어업용면세유한도량배정	"귀금속, 보석산업 발전방안" 발전방안 마련	X	전거데이터 부족 /	EA	2007	2007
4	처리과업무계획(보고)평	기획예산처 재정운용실 민간투자	기획예산처 재정운용실 민간투	BTL교육 대외,대외협력	서산시 하수관거정비 임대형 민자사업(BTL	X	분류체계 오류 / 예	EA	2007	2007
5	처리과서무	기획예산처 재정운용실 재정기준	기획예산처 예산실 예산기준과	세출예산집행지침작성	2004년도 세출예산집행 추가지침 통보	X	철_건 생산기관명	EA	2004	2004
6	WTO/DDA 관세협상	기획재정부 세제실 관세정책관 다	재정경제부 세제실 관세국 다자	세계무역기구(WTO) DDA 관	검토 부탁드립니다.	X	메타데이터만으로	EA	2007	2007

〈그림 7〉 이상치(X) 유형별 사례

목록만으로는 최종 판단이 불가하고 실제 기록물 원문을 열람하거나 CAMS의 상세 정보 조회를 통해 확인이 필요한 기록물 이상치(△)는 총 5,068건으로 자세한 분포는 <표 7>과 같다. 사례의 분포는 전거데이터 부족 4,394건(74%), 메타데이터만으로 확인 불가 1,443건(24%), 메타데이터만으로 확인 불가/공통 123건(2%), 철명 오류 1건(0%) 순으로 나타났다. 또한 철 목록 검토 사항을 승계 받아 '전거데이터 부족', '철명 오류' 사항이 포함되어 있다.

〈표 7〉 건 목록 이상치(△) 결과

	철 목록 판별 승계		건 별 판별	
구분	1) 전거데이터 부족	2) 철명 오류	3) 메타데이터만으로 확인 불가	4) 메타데이터만으로 확인 불가 / 공통
총계	4,394 (74%)	1 (0%)	1,443 (24%)	123 (2%)
	중복계상 5,961건(총 이상치△: 5,068)			

* 자동화 가능

구체적인 이상치(△) 예시는 <그림 8>과 같다. '1) 전거데이터 부족' 사례는 기록물생산기관변천정보시스템, 국가기록원 전거목록을 확인한 후 조사대상 하위기관 해당 존립기관 목록에 존재 하지않아 목록만으로 최종 판단이 불가능한 경우였고, '2) 철명 오류'는 오타 및 띄어쓰기 기준으로 판별되었다. '3) 메타데이터만으로 확인 불가'는 '공무원 제안 검토'라는 건 제목으로 해당 철과 분류체계에 해당되는 내용인지 확인할 수 없는 사례로, 추가 확인이 필요한 기록물로 판별하였다. 마지막으로 '4) 메타데이터만으로 확인 불가/공통'에 해당하는 예시는 '관보게재 의뢰'라는 건 제목으로 다양한 철 목록에 동일한 건제목이 다수 존재하여 해당 철과 분류체계에 해당되는 내용인지 확인 할 수 없어 이상치 △로 판별하였다. 이렇듯 이상치(△) 기록들은 실제 오분류 여부 판단을 위해 담당자가 전문을 열람하거나 CAMS 내 상세 데이터를 추가로 검토하는 작업을 요한다.

http://dx.doi.org/10.14404/JKSARM.2022.22.1.043

구분	분류체계명	생산기관명	처리과명	BND_TTL(기록물철제목)	JEMOK(제목)	)/X/	비고	록	생산	종료
1	세계무역기구(WTO)관세	기획재정부 세제실 관세정책관 다	기획재정부 세제실 관세정책관	WTO DDA 비농산물 협상	DDA 비농산물시장접근(NAMA) 협상(secto	△	전거데이터 부족	EA	2008	2008
2	예산관계법령제개정및해	기획예산처 예산실 예산제도과	기획예산처 예산실 예산제도과	예산관계법령제개정및해석경	과징금 환급에 대한 질의	△	철명 오류 / 전거데	EA	2005	2005
3	부가가치세법령 입안	기획재정부 세제실 재산소비세정	기획재정부 세제실 재산소비세	부가가치세법령 입안	공무원 제안 검토	△	메타데이터만으로	EA	2008	2008
4	특별긴급관세	기획재정부 세제실 관세정책관 산	기획재정부 세제실 관세정책관	특별긴급관세	관보게재 의뢰	△	메타데이터만으로	EA	2008	2008

〈그림 8〉 이상치(△) 유형별 사례

3.3 분류 이상치 판별 자동화 방안

본 절에서는 <그림 2>와 <그림 3>에서와 같이 분류 이상치 판별 프로세스를 정립하는 과정을 통해 자동화가 가능한 세부 업무를 도출, 세부적인 구현 방안에 대해 기술한다. 판단 기준은 반복적으로 수행되고 기계적으로 학습 가능 혹은 처리 가능한 업무는 '자동화'로, 관련 자료 수집 및 정리 등 일정 부분만 자동화가 가능하고 이외는 기록물 담당자의 작업을 필요로 하는 업무는 '반자동화'로 구분하였다. 또한 기록물 담당자의 고차원적인 사고과정을 필요로 하여 현재의 AI 기술로는 적용이 불가능하거나, 기계학습에 들어가는 비용이 너무 커 비효율적인 업무는 '자동화 불가'로 구분하였다.

먼저 자동화가 가능한 영역은(<그림 3> '자동화' 녹색 겹선 박스 표시)는 원시 목록을 정제하는 과정 중 철·건 목록의 분석 범위별 목록 분류, 건 목록 정제 과정 중 철 목록 필드 추가 입력 과정이 있다. 이는 비교적 단순하면서도 반복적으로 수행되는 작업으로, 해당 작업들은 문자열 처리를 수행하는 간단한 코딩과 엑셀에서 제공하는 정렬 기능을 활용하는 방식으로 자동화가 가능하다. 또한 철 목록 검토와 건별 이상치 판별 프로세스에서 생산년도 항목을 확인하는 과정 역시 공통지식자원 중 조사대상 하위기관 존립기관 목록을 기준으로 자동 판별이 가능하다. 마지막으로 철 목록 검토 프로세스에서 철 제목 오류를 확인하는 과정은 기록물분류기준표나 기록관리기준표를 참조하여 웹 크롤링 기법을 포함해 기계학습 기반 언어분석 기술 등을 적용하여 자동화가 가능하다.

반자동화가 가능한 영역(<그림 3> '반자동화' 주황 음영 박스 표시)은 각 철과 건 목록의 필드를 재정렬하는 과정이다. 분석 범위별 목록 자체는 자동으로 추출, 엑셀(excel)에서 처리 가능한 형태로 출력되나, 이후 기록물 담당자가 이상치 판별을 위한 필드를 추가하고 재정렬하는 검수 단계가 필요하다. 이 과정은 자동화와 기록물 담당자의 역할이 공존하는 과정이기 때문에 반자동화에 해당한다.

한국고용정보원(2016)은 인공지능과 로봇기술에 따른 자동화의 직무 대체는 단순반복 과업 중심으로 대체될 것이고 중요한 의사결정과 감성, 지적능력에 기초한 직무는 여전히 인간이 맡게 될 것이라고 제언했다. 즉, 사람의 지각력과 사고방식, 지적능력이 필요한 업무는 자동화되기 힘들다는 것이다. 이상치 판별 과정 중 대체되기 어려운 업무는 철 목록 검토 과정에서 오분류를 확인하는 영역과, 건 별 이상치를 판단하는 과정에서 분류체계 및 철 제목과 건 제목 오류를 확인하는 영역이다. 상기한 과정은 기록물 담당자의 고차원적인 판단이 필요로 하는 영역으로, 공통지식자원 외에 담당자의 배경 지식과 생산기관별 세부 이력이나 규정 등 추가 정보를 활용하는 등 사전에 정의되지 않은 다양한 정보들을 필요로 하게 된다. 이러한 유형의 오분류 판단은 규격화되지 않은 추가 과정을 필요로 하는 것으로, 기계학습을 적용하기 불가능한 유형이다. 그 외 <표 2>와 <표 3>에 나타난 바와 같이 오류 이상치 유형의 판단 기준이 비교적 일반적이고 단순해 기계에 학습시킬 수 있는 부분들은 <그림 4>와 <그림 5>와 같이 각각 철 목록과 건 목록에 태깅된 이상치 유형을 학습데이터로 활용, 기계학습 모델을 구축하면 된다.

4. 결 론

4차 산업혁명 시대가 도래하면서 점차 일상생활과 업무의 영역에 영향을 미치고 있다. 특히 전자기록 환경에서

http://dx.doi.org/10.14404/JKSARM.2022.22.1.043

지능형 문서 처리는 기록관리 업무에 광범위한 변화를 초래할 수 있다. 학습데이터를 활용하여 반복되는 업무의 처리를 자동화할 수 있으며, 기록물의 기술 업무에도 문서 요약 엔진 활용 및 색인어 추출 자동화 등을 적용할 수 있다. 또한 전거정보의 활용 및 기록물 분류 업무에도 자동화 기술의 적용 가능성이 높다. 이러한 기록관리 업무 변화에 맞추어 기록물 담당자는 수행해야 하는 업무 내용을 확인하고 이에 걸맞은 역량을 준비할 필요가 있다. 그러나 매년 이관되는 기록물의 방대한 양에 비해 관리 과정은 대부분 수작업으로 진행하고 있으며, 참조하고 있는 자료 역시 산발적으로 관리되고 있어 어려움이 매우 크다.

이에 본 연구는 먼저 국가기록원을 대상으로 기록물 담당자와 FGI를 수행하여 세부 업무를 분석하였으며, 심층 면담 결과에 따라 요구사항을 도출하였다. 업무 프로세스 정립, 재편철, 이상치 판별 등의 요구사항에 따라 기록물 분류 이상치 판별 프로세스를 정립한 후 학습데이터를 구축하였다. 이를 통해 업무 효율화를 꾀하고, 업무 프로세스의 정립을 통한 체계적이며 일관된 업무 수행을 도모하였다. 연구 결과를 정리하면 다음과 같다.

먼저 분류 이상치 판별 프로세스를 정립하는 과정을 통해 자동화가 가능한 세부 업무를 도출, 자동화, 반자동화, 자동화 불가로 구분하였다. 이후 정립된 이상치 판별 업무 프로세스를 실제 국가기록원으로 이관된 기록물을 대상으로 적용한 결과, 구축 대상 전체 1만여 건 기록물 중 약 78%(7,818건)의 기록물이 분류 결과에 대한 재고가 필요하다고 판별되었다. 철 목록으로 기준으로 전체 2,407철 중 1,971철, 약 82% 철에 오분류 가능성이 있는 기록물이 포함된 것으로, 이는 애초 국가기록원으로 이관된 기록물의 편철 정보가 명확히 이해되는 경우가 적다는 것을 함의한다. 다시 말해 실제 기록물 원문 전체를 열람하지 않고 주요 메타데이터를 기반으로 작성된 목록만을 대상으로 판단할 수 있는 부분이 매우 제한적임을 의미한다. 이는 본 연구의 목적이 기존의 기록물 분류·기술 업무의 지능화 방안을 도출하는 데 있다는 점을 감안하면 매우 큰 괴리가 있는 결과로, 애초 해당 업무의 상당 부분이 기계학습에 부적합한 난이도의 업무임을 반증한다.

그럼에도 불구하고 다른 한편으로는 분류 이상치 판별 프로세스를 적용하여 도출된 학습데이터를 통해 입수된 이관 기록물의 메타데이터와 기록관리기준표 등과 같은 공통지식자원만을 참조하여 분류 오류 이상치로 판별된 유형은 기계적으로 학습 가능하며 이는 자동화할 수 있는 영역임을 다시 한 번 검증할 수 있었다. 또한 필수 메타데이터로 구성된 목록에 기반해 이상치를 판별, 원문 열람 후보를 추천함으로써 담당자의 수작업을 최소화하는데 일조할 수 있음을 보였다.

본 연구는 기록물 담당자의 개별적인 노하우와 수작업에 의존하여 진행하는 기록물 분류 업무의 지능화를 위한 사전 작업으로, 추후 새로운 담당자가 해당 업무를 명확히 이해하고 수행할 수 있는 지침서를 마련하고 실질적 업무 경감에 따른 효율성을 높이기 위해 시작되었다. 향후 연구 방향으로는 현재 경제 관련 주요 생산기관을 대상으로 연구를 수행한 점에서 확장해 다양한 계열의 공공기관 이관 기록물을 대상으로 적용함으로써 학습데이터 구축 프로세스의 실효성을 검증하고자 한다. 나아가 현재는 자동화가 어렵다고 판단한 이상치 후보의 세부 유형을 분석해 학습 가능한 요소를 규명, 이에 대한 학습데이터 구축을 확장하고자 한다. 본 연구의 결과가 이관 기록물의 분류 자동화를 위한 기초 자료로 활용되고 나아가 현업에서 직접적으로 적용할 수 있는 실효성 있는 연구 결과이길 고대한다.

참고문헌

강윤아, 박태연, 김현진, 오효정 (2021). 생산기관 직제분석 자동화 및 공통 활용 방안. 한국기록관리학회지, 21(4), 81-99. http://dx.doi.org/10.14404/JKSARM.2021.21.4.081

국가기록원 (2021a). 기록관리 AI 기술적용을 위한 공통 학습데이터 세트 구축 연구.

국가기록원 (2021b). 영구기록물관리기관 표준모델: 기능 및 업무절차(v2.2).

김인택, 안대진, 이해영 (2017). 인공지능을 활용한 지능형 기록관리 방안. 한국기록관리학회지, 17(4), 225-250. http://dx.doi.org/10.14404/JKSARM.2017.17.4.225

김해찬솔, 안대진, 임진희, 이해영 (2017). 기계학습을 이용한 기록 텍스트 자동분류 사례 연구. 정보관리학회지, 34(4), 321-344. http://dx.doi.org/10.3743/KOSIM.2017.34.4.321

설문원 (2013). 기록분류를 위한 정부기능분류체계의 적용 구조 및 운용 분석. 한국비블리아학회지, 24(4), 23-51. https://doi.org/10.14699/kbiblia.2013.24.4.023

오진관 (2019). 영구기록 관리와 서비스를 위한 자동화, 지능화 기술. 한국기록관리학회 춘계 학술대회 논문집, 69-74. http://dx.doi.org/10.14404/PKSARM.2019.S.069

윤상우 (2020). 업무기능분석을 통한 공공기관 기록분류체계 구축방안: A기관 기록관리기준표 수립사례를 중심으로. 석사학위논문, 강릉원주대학교 대학원 기록관리협동과정.

장지숙, 이해영 (2009). 맥락정보를 이용한 기록 자동분류시스템 설계. 한국기록관리학회지, 9(1), 151-173. https://doi.org/10.14404/JKSARM.2009.9.1.151

장현종, 노지현 (2021). 국립대학 기록물 분류체계의 운영현황과 개선방안에 관한 연구. 한국기록관리학회지, 21(2), 115-134. http://dx.doi.org/10.14404/JKSARM.2021.21.2.115

최정렬 (2018). 기계학습 활용을 위한 학습 데이터세트 구축 표준화 방안에 관한 연구. 한국디지털정책학회논문지, 16(10), 205-212. https://doi.org/10.14400/JDC.2018.16.10.205

최철민 (2018). 국가기록원 소장 기록의 논리적 재편철 연구. 석사학위논문, 한남대학교 대학원 기록관리학과.

한국고용정보원 (2016.3.24.). 인공지능(AI), 로봇과 사람의 협업시대.
출처: http://www.keis.or.kr/user/bbs/main/137/3963/bbsDataView/32721.do?page=38&column=&search=&searchSDate=&searchEDate=&bbsDataCategory=

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

Choi, Cheol-min (2018). A Study on the Logical Reorganization of Records Owned by the National Archives of Korea. Master's thesis, Department of Archival Studies, Graduate school of Hannam University, South Korea.

Choi, Jung Yul (2018). A study on the standardization strategy for building of learning data set for machine learning applications. Journal of Digital Convergence, 16(10), 205-212. https://doi.org/10.14400/JDC.2018.16.10.205

Jang, Hyun-Jong & Rho, Jee-Hyun (2021). Problems encountered by and improvement strategies of the records classification system for national universities. Journal of Korean Society of Archives and Records Management, 21(2), 115-134 http://dx.doi.org/10.14404/JKSARM.2021.21.2.115

Jang, Ji-Sook & Rieh, Hae-Young (2009). Design of automatic records classification system using contextual information. Journal of Korean Society of Archives and Records Management, 9(1), 151-173. https://doi.org/10.14404/JKSARM.2009.9.1.151

Kang, Yoona, Park, Tae-yeon, Kim, Hyunjin, & Oh, Hyo-Jung (2021). Automation and common utilization plans of job and organization analysis of producing institutions. Journal of Korean Society of Archives and Records Management, 21(4), 81-99. http://dx.doi.org/10.14404/JKSARM.2021.21.4.081

Kim, Hae Chan Sol, An, Dae-Jin, Yim, Jin Hee, & Rieh, Hae-Young (2017). A study on automatic classitification of record text using machine learning. Korean Society for Information Society, 34(4), 321-344. http://dx.doi.org/10.3743/KOSIM.2017.34.4.321

Kim, Intaek, An, Dae-Jin, & Rieh, Hae-Young (2017). Intelligent records and archives management that applies artificial intelligence. Journal of Korean Society of Archives and Records Management, 17(4), 225-250. http://dx.doi.org/10.14404/JKSARM.2017.17.4.225

Korea Employment Information Service (2016, March 24). Artificial intelligence(AI), the era of collaboration between robots and humans. Available: http://www.keis.or.kr/user/bbs/main/137/3963/bbsDataView/32721.do?page=38&column=&search=&searchSDate=&searchEDate=&bbsDataCategory=

National Archives of Korea (2021a). Study on Common Training Dataset Construction for appling AI technology for Records Managements.

National Archives of Korea (2021b). Standard Model for Archives: Function and Procedure Version 2.2.

Oh, Jin Kwan (2019). Automated and intelligent technology for archives management and services. Proceedings of Korean Society of Archives and Records Management, 69-74. http://dx.doi.org/10.14404/PKSARM.2019.S.069

Seol, Moon-Won (2013). An analysis of the application framework of the business reference model to records classification schemes in Korean central government agencies. Journal of the Korean Biblia Society for Library and Information Science, 24(4), 23-51. https://doi.org/10.14699/kbiblia.2013.24.4.023

Yun, Sang-Woo (2020). A Plan to Establish a Record Classification System for Public Institutions through Business Function Analysis: Focused on the Example of Establishment of A Organization Record Management Standard Table. Master's thesis, Department of Archival Studies, Gangnung-wonju National University Graduate school, South Korea.