인문논총 82권 2호(2025. 5. 31.), pp. 35~72
[DOI] https://doi.org/10.17326/jhsnu.82.2.202505.35

# 한국 근대문헌의 디지털 텍스트 편찬과 쟁점

## 『진단학보』를 중심으로*

장문석**
곽한나***
문예지****
심현지*****
안소연*****
이지훈****
최솔*****
허민석****
김지선******

**초록** 이 글에서는 한국의 인문학 학술지 『진단학보』의 제1호~제14호에 실린 논문을 대상으로 근대문헌 텍스트의 협업 기반 디지털 편찬 방법론을 탐색하였다. 보다 정밀한 데이터 처리를 위해서 『진단학보』의 문헌적 특징과 데이터 활용 가능성을 고려한 『진단학보』 XML 스키마를 설계하였다. 설계된 스키마를 기반으로 미디어위키를 활용하여 텍스트 입력 틀을 구축하고, 『진단학보』에 수록된 총 85편의 논문 가운데 74편에 대한 원문 및 현대문 데이터를 입력하였다. 텍스트 데이터 입력 과정에서 XML 설계의 보완 필요성, 현대문 텍스트 변환 규칙 정립, 편찬 도구의 효율성 등의 문제가 확인되었으며, 이는 향후 추가적인 논의가 필요한 과제로 남았다. 본 연구의 『진단학보』 텍스트 편찬 경험과 시행착오를 바탕으로, 향후 여러 연구자들의 협업을 통해 한국 근대문헌에 적합한 디지털 텍스트 편찬 방법론을 찾아갈 수 있기를 기대한다.

**주제어** 진단학보, 한국 근대문헌, XML 스키마 설계, 디지털 텍스트 편찬, 시맨틱 데이터 아카이브

---

* 이 논문은 서울대학교 인문학연구원이 지원한 집담회의 성과임.
** 경희대학교 국어국문학과 부교수 및 서울대학교 인문학연구원 객원연구원, 제1저자
*** 서울대학교 국사학과 박사과정, 공동저자
**** 서울대학교 국어국문학과 박사과정 수료, 공동저자
***** 서울대학교 국사학과 석사과정 수료, 공동저자
****** 고려대학교 문과대학 강사, 교신저자

www.kci.go.kr

# 1. 들어가며

이 글은『진단학보』아카이브 구축의 경험을 바탕으로 한국 근대문헌의 디지털 텍스트 편찬과 쟁점을 검토한다. 저자들은 "아날로그 문헌자료를 어떻게 온라인 환경에서 데이터의 형식을 반영하여 편찬하고 공유할 수 있을 것인가?", "인문학 연구자 및 시민이 효율적으로 활용할 수 있는 문헌자료의 디지털 아카이브를 어떻게 구축할 것인가?"와 같은 문제의식에 입각하여,『진단학보』제1호~제14호에 실린 논문을 대상으로 한 디지털 아카이브를 설계 및 구축하고 있다.

한국 근대문헌 아카이브의 경우, 2000년대 이후 국사편찬위원회를 위시하여 여러 국공립 도서관과 각 대학교 도서관 및 부설 연구소 등에서 구축 작업을 꾸준히 수행해 왔다. 그 결과물로서 구축되어 현재 서비스되고 있는 아카이브는 목차, 제목, 키워드 등의 메타데이터만을 제공하는 사례가 대부분이며, 원문 제공에 있어서도 이미지나 플레인 텍스트(plain text)로 제공하는 경우가 다수이다. 원문 텍스트 데이터가 제공될 경우, 그것을 활용해 계량적 분석 또는 딥러닝 기반의 디지털 인문학 연구를 시도할 수 있겠으나, 문헌 자체의 특성이 데이터에 충실히 반영되어 있지 않은 경우가 대부분이라, 텍스트를 정밀하게 분석하거나 효율적으로 데이터를 활용할 수 있는 여지가 크지 않다.

학술지를 대상으로 한 디지털 아카이브(또는 데이터베이스)의 경우, 국내에서는 그 본격적 사례를 찾아보기가 아직은 어렵다.[1] 해외의 경우 학술

---

1 한국학 분야에서 간행된 학술지를 대상으로 하여 그 안에 담긴 여러 정보를 학술데이터로 가공해서 공유-시각화하기 위한 차원의 실험적 연구로 류인태(2022),「인문학술 데이터 프로세싱에 관한 시론」,『한국학』45(2), 한국학중앙연구원 참조. 류인태의 연구는 한국학 분야의 학술지 논문을 대상으로 실제 데이터를 다루는 과정을 통해, 그 안에 담긴 학술 정보를 어떻게 XML 기반의 데이터로 가공 및 편찬할 수 있는지에 관한 논증을 전개하였다. 이 연구는 특정한 학술지를 본격적으로 아카이빙하는 데 목적을 두지는 않았다는 측면에서, 본 연구와는 성격이 다르다. 하지만 XML 기반의 텍스트 데이터 편찬을

지를 대상으로 한 데이터 구축 사례로 일본애니메이션학회(Japan Society for Animation Studies, JSAS)에서 구축한 Database for Animation Studies[2]를 거론할 수 있다. 이 데이터베이스는 애니메이션 연구 분야의 서적, 논문 및 기타 출판물의 서지 정보를 포괄적으로 수집하고 있으며, 국내외 애니메이션 관련 학술 문헌을 체계적으로 수록하여 저자별·작가별·작품별·키워드별 분류 체계와 추천문헌 리스트를 제공한다. 특정 학문 분야에 특화된 이 데이터베이스는 일본 문화청 위탁사업의 일환으로 지속적으로 업데이트되고 있어, 학회 주도의 체계적인 학술 데이터베이스 구축 모델을 보여준다.

이미 구축된 학술지 데이터를 활용한 연구 사례로는 Signs@40[3]을 거론할 수 있다. Signs@40은 1975년 창간된 여성학 및 젠더 연구 분야의 대표적 국제 학술지인 *Signs: Journal of Women in Culture and Society*의 40년간(1975-2014) 발행분을 대상으로 한 대규모 디지털 분석 프로젝트이다. 이 프로젝트는 약 20만 페이지에 달하는 방대한 텍스트를 JSTOR's Data for Research 플랫폼을 통해 분석하였으며, 특히 토픽 모델링과 동시인용 네트워크 분석을 통해 40년간의 학술 담론 변화 및 인용된 저자들 간의 학술적 연결망을 추적하고 시각화하였다. Signs@40은 이미 디지털화된 기존 아카이브(JSTOR)를 활용하여 인용 관계나 주제 중심의 거시적 동향 분석에 중점을 둔 사례로, 체계적으로 구축된 디지털 아카이브가 있을 시, 그것을 기초 자원으로 삼아 다양한 분석 연구를 수행할 수 있음을 보여주는 사례라 하겠다.

국내외 유관 연구 및 결과물 현황을 참고할 때, 가장 중요한 점은 구축 대상이 되는 학술지의 고유한 특징을 포착하고 그에 초점을 둔 데이터 모

---

시도한다는 점에서 이 연구의 기초적 문제의식과 그 결이 일정부분 맞닿아 있다고 하겠다.

2 〈Database for Animation Studies〉 https://database.jsas.net/mapping/

3 〈Signs@40: Feminist Scholarship through Four Decades〉 https://signsat40.signsjournal.org/

www.kci.go.kr

델링을 진행해야 한다는 사실이다. 특히 본 연구에서 대상으로 삼은 학술지 『진단학보』의 경우 근대기에 간행된 자료이기 때문에, 한국 근대문헌 특유의 원문-현대문 이중 구조를 효율적으로 처리할 수 있는 XML 스키마 설계가 기초적인 데이터 구축 작업에 선행되어야 한다. 저자들은 이러한 문제의식을 바탕으로, 종합 한국학 학술지 『진단학보』의 논문을 사례로 하여 한국 근대문헌의 특성에 최적화된 디지털 아카이브 구축 방법을 고민하였으며, 그 실행 및 결과에 관한 내용을 제시하고자 한다.

한국 근대문헌의 형식은 단행본, 잡지, 신문, 학술지 등으로 다양하다. 그중 저자들은 1930년대에 창간한 학술지 『진단학보』 제1호~제14호에 수록된 논문을 구축 대상으로 선택하였다. 『진단학보』 아카이브의 구축이 『진단학보』의 문화사적 위상을 특권화하는 것은 아니다. 하지만 『진단학보』의 문헌학적 특징과 학술사적 위상을 고려할 때, 『진단학보』 아카이브는 추후 여타의 한국 근대문헌 아카이브 구축 과정에서 참조할 수 있는 다양한 모색을 담을 수 있다고 판단하였다.

첫째, 학술지 『진단학보』에 실린 논문은 형식을 지키는 글쓰기에 기반한다. 신문 및 잡지의 기사 역시 산문이지만, 집필자의 개성에 따라 다양한 형식으로 쓰인다. 하지만 학술지 논문은 서론, 본론, 결론 등 의미적 구조를 갖춘 한편, 주석, 표, 그림 등 형식적 특징을 갖추어서 쓴다. 한국에서 근대적 아카데미즘이 성립하였던 1930년대 진단학회가 간행한 『진단학보』는 한국의 근대적 학술 규범 및 관습의 확립과 학술적 글쓰기의 정착 과정을 보여주는 학술지이다. 따라서 『진단학보』 아카이브를 구축할 경우, 한국 근대문헌의 다양한 형식적 특징을 반영한 데이터 설계가 가능하며, 추후 다른 아카이브 구축에서 활용할 수 있다.

둘째, 『진단학보』는 전체적인 형식을 느슨하게 공유하면서도 비균질적인 다양한 글쓰기에 기반한 논문이 실려 있다. 『진단학보』의 논문은 학술적 글쓰기라는 성격을 공유하면서도, 문학, 역사, 어학, 민속학 등 한국학 여러 분야의 논문이 실려 있다. 학술영역에 따라 논문의 글쓰기도 차이가 있는

www.kci.go.kr

데, 그러한 차이로 인해 다양한 특징을 아카이브 구축 과정에서 만나게 된다.

셋째, 『진단학보』의 데이터 규모는 아카이브 설계 및 구축, 특히 텍스트 데이터 편찬을 시도하기에 적당하다. 해방 이전 『진단학보』는 총 14호가 간행되었고, 80여 편의 논문이 수록되었다. 메타데이터뿐 아니라 텍스트 데이터를 제공하는 아카이브 구축을 시험하기에 적당한 규모이다.

넷째, 『진단학보』의 학술사적 위상을 고려할 때, 아카이브의 확장 및 다른 아카이브와의 연결을 시도할 수 있다. 80여 편의 논문은 데이터에 대한 계량적 분석을 시도하기에는 많지 않은 분량이다. 하지만 『진단학보』의 성격을 고려할 때, 시간 및 공간적 데이터의 확장이 가능하다. 『진단학보』의 한국학은 서양, 일본, 중국의 동양학 및 한국학 연구의 성과를 기반으로 성립하였다. 『진단학보』는 제국 일본의 인문학과 연동하는 한편, 청구학회의 일본어 학술지 『청구학총』과의 긴장 안에서 전개되었다. 그리고 1945년 이후 대한민국과 북한이 성립하면서, 『진단학보』에 관여한 인물 및 학술적 성취 역시 남과 북으로 분기하면서 확산한다. 『진단학보』 아카이브는 식민지 시기 한국의 학술 아카이브를 넘어서, 시간적으로 20세기 후반 대한민국 및 북한의 한국학 학술 아카이브로 확장할 수 있다. 또한 공간적으로는 한국학 및 동양학을 매개로, 유럽, 북미, 일본, 중국 등의 다양한 지역 및 국가에서 생산된 학술 문헌과 연결될 수 있다. 최근 전 지구적으로 다양한 문헌 아카이브가 구축되어 있다는 점을 고려한다면, 『진단학보』 아카이브 역시 세계의 다양한 아카이브의 데이터 및 정보와 연동할 수 있다.

『진단학보』 아카이브는 한국 근대문헌 아카이브 구축의 범례가 될 수 있으며, 추후 확장 가능하고 여러 아카이브 구축의 참고사례가 될 수 있다. 저자들은 이 점에 유의하면서, 『진단학보』 제1호(1934.11.)~제14호(1941.5.)를 대상으로 한국 근대문헌의 디지털 아카이브 구축을 시도 중이다. 현재는 아카이브의 데이터 모델을 설계하고 그에 기반하여 기초 데이터를 입력한 상태이다. 이 글은 근대문헌의 디지털 텍스트 편찬의 과정을 XML 스키마

www.kci.go.kr

설계 및 편찬으로 나누어서 각각 쟁점을 살펴보고자 한다.

## 2. 『진단학보』 XML 스키마 설계

2000년 전후 한국 사회의 정보화가 진행되면서 다양한 문헌에 대한 정보화 역시 진행되었다. 근대문헌 자료를 디지털 환경에서 편찬 및 연구하기 위해서는 아날로그 근대문헌의 원문(plain text)을 기계가독형(machine-readable) 형태로 변환해야 한다. 여러 국공립도서관을 비롯한 국공립기관 및 대학 등은 근대문헌의 원문 텍스트를 디지털 환경에서 편찬할 때, XML(eXtensible Markup Language)을 활용하였다. XML을 활용할 경우, 데이터의 구조 및 의미를 명확하게 표현할 수 있으며, 개별적인 데이터 아카이브 사이에서 데이터의 교환, 공유, 연결이 더 용이해진다. XML은 문헌의 '기본 저장 단위의 구조화'가 가능한 동시에 '기본 저장 단위 내부 요소의 정보화'를 도모할 수 있다. 즉 문헌의 형식적 특징 및 논리적 체계를 반영한 데이터의 구조를 설계할 수 있으며, 편찬자가 태그를 직접 정의하면서 원문 문헌 내부에 혼재하는 다양한 성격의 정보 요소를 기계적으로 식별하고 활용할 수 있다.[4]

2000년대 이후 한국에서 구축된 다양한 아카이브들 또한 기본적으로는 XML을 활용하여 구축되었다. 한국 근대문헌보다는 한국 고전문헌의 아카이브 구축이 더욱 활발하게 이루어졌다. 한국 고전 및 전통에 대한 사회적 관심과 수요, 고전의 정리 및 번역에 대한 국가적 지원, 그리고 연구자의 역량 및 참여 등의 여건에 힘입어 현재 다수의 한국 고전문헌 아카이브가 구축되었다. 국사편찬위원회의 『조선왕조실록』, 『승정원일기』, 한국고전번역

4 김현(2006), 「고문헌 자료 XML 전자문서 편찬 기술에 관한 연구」, 『고문서연구』 29, 한국고문서학회, pp. 183-230.

www.kci.go.kr

원의 한국문집총간, 한국학중앙연구원의 디지털장서각, 성균관대학교의 한국경학자료시스템 등은 모두 XML을 활용하여 아카이브를 구축하였다. 또한 개별 문헌의 XML 데이터를 다운로드받을 수 있도록 하거나, 공공포털 등을 활용하여 XML 형식의 데이터를 공개하고 있다.[5] 시민 및 연구자는 공유된 XML 데이터를 재가공하여 새로운 작업을 시도할 수 있다.[6]

한국 근대문헌 아카이브 역시 XML 설계에 기반하여 원문을 제공하고 있다. 하지만 한국 고전문헌의 XML 설계가 개별 문헌의 구조적 특징을 반영하는 데 중점을 두고 있는 것과 달리, 한국 근대문헌의 경우 상대적으로 단순하고 범용적인 형식의 XML로 데이터를 입력하고 있다.[7] 또한 데이터의 공유도 한국 고전문헌만큼 활발하지 않다. 그동안 한국 근대문헌의 텍스트에 대한 분석은 주로 토픽모델링 등을 비롯한 계량적 연구의 방향으로 이루어졌다.[8] 저자들은 연구자들의 활용도가 높고 정교한 데이터 처리가 가능한 한국 근대문헌의 아카이브 구축을 위해서는 현재의 평면적인 XML

---

5 〈공공데이터포털〉 https://www.data.go.kr; 한국 국공립기관의 데이터 공유 현황에 관해서는 김바로(2022), 「〈공공데이터법〉과 인문데이터: 공공기관 보유 인문데이터 공개 신청 사례를 중심으로」, 『한국고전연구』 57, 한국고전연구학회 참조.

6 정성훈(2024), 「조선시대 한시의 경관 요소에 대한 계량적 분석: '소쇄원'과 '환벽당·식영정'을 중심으로」, 『한문학논집』 67, 근역한문학회; 김서윤(2024), 「조선시대 논어주석의 인용 분류체계」, 『민족문화』 67, 한국고전번역원; 지영원·최진경(2024), 「한국 한시 데이터 아카이브 구축을 위한 개념적 데이터 모델링 시론」, 『민족문학사연구』 85, 민족문학사연구소 참조.

7 국사편찬위원회가 구축한 한국근현대잡지자료의 데이터 스키마 분석은 김바로(2024), 「국사편찬위원회 한국근현대잡지자료 데이터(2024.03.27.)」, 『디지털인문학』 1(1), 한국디지털인문학협의회, pp. 145-151.

8 이재연(2014), 「작가, 매체, 네트워크: 1920년대 소설계의 거시적 조망을 위한 시론」, 『사이間SAI』 17, 국제한국문학문화학회; 허수(2018), 「언어연결망 분석으로 본 20세기 초 한국의 '문명'과 '문화': 주요 언론 기사에서의 논의 맥락을 중심으로」, 『개념과 소통』 22, 한림대 한림과학원; 이재연(2016), 「키워드와 네트워크: 토픽 모델링으로 본 『개벽』의 주제 지도 분석」, 『상허학보』 46, 상허학회; 김현주(2022), 「텍스트마이닝으로 본 『삼천리』의 문화」, 『문화』, 소화; 전성규(2023), 「한국 근대 잡지의 계량적 연구 방법에 대한 논의: 코퍼스 구축 및 데이터 분석의 사례를 중심으로」, 『민족문학사연구』 82, 민족문학사연구소.

www.kci.go.kr

스키마 설계를 넘어설 필요가 있다고 판단하였다.

데이터의 의미적(semantic) 연관을 고려한 한국 근대문헌 아카이브를 설계 및 구축하기 위해서는 먼저 근대문헌의 자료적 특성 및 연구 목적을 고려하여 XML 스키마를 설계하고[9] XML 데이터 전자문서를 편찬할 필요가 있다(① XML 데이터 설계 및 편찬). 동시에 근대문헌 텍스트의 형식 및 내용에 담긴 의미적(semantic) 체계를 반영하여 온톨로지(ontology)를 설계해야 한다. 문헌의 존재방식 및 문헌 구성 정보 간의 연결성을 반영한 온톨로지를 설계하는 과정에서 XML 스키마를 보완하고, 온톨로지 디자인을 완료한 후 XML 파싱을 통해 정형데이터를 구축하게 된다(② 온톨로지 디자인 및 데이터 파싱). 이후 구축한 정형데이터를 활용하여 그래프 데이터베이스(Graph DB)를 구현한다. 그래프 데이터베이스를 활용하여 데이터를 시각화하고, 양과 질 두 측면에서 데이터를 분석한다. 또한 분석 과정에서 XML 스키마 및 온톨로지 또한 함께 보완한다(③ 지식그래프 구현 및 데이터 분석).[10]

XML은 기본적으로 데이터의 계층구조를 표현하기에 적합한 마크업 언어이다. 국사편찬위원회 등에서 이미 구축된 한국근대문헌 아카이브가 XML 방식을 선택한 것도 이 때문이다. 하지만 기존 근대문헌 아카이브는 범용적인 XML 설계를 준용하느라 개별 문헌의 특징을 충분히 반영하지 못하였다. 따라서 저자들은 『진단학보』의 형태적·의미적 특징에 대한 충분한 검토를 바탕으로, 1) 『진단학보』의 문헌적 특징 및 2) 데이터의 활용 가능성을 고려하여 XML을 설계할 필요가 있다.

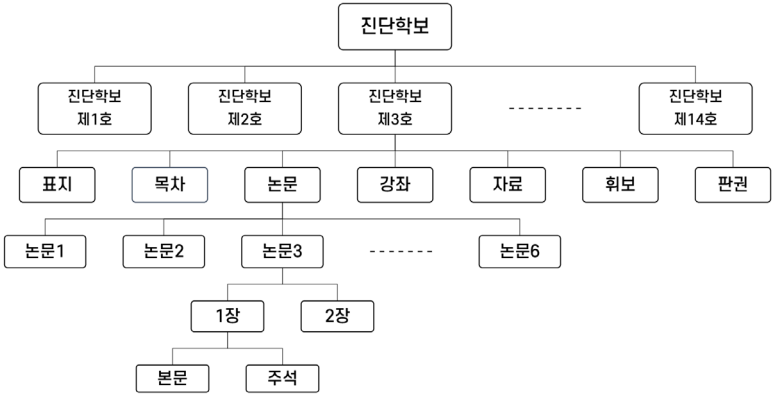
---

9 국제적으로는 TEI(Text Encoding Initiative) P5가 인문학 텍스트의 디지털 인코딩 표준으로 널리 사용되고 있으나, 본 연구에서는 『진단학보』 편찬과 분석에 특화된 XML 스키마를 설계하였다. 이는 협업 환경에서의 작업 효율성과 한국 근대문헌의 이중 텍스트 구조를 고려한 선택이며, 『진단학보』의 학술적 특성을 보다 세밀하게 분석할 수 있는 데이터 구조를 구현하기 위함이다.

10 ① XML 데이터 설계 및 편찬, ② 온톨로지 디자인 및 데이터 파싱, ③ 지식그래프 구현 및 데이터 분석 등으로 이어지는 시맨틱 데이터 설계 및 처리 과정에 관해서는 변은미·이동학·류인태(2024), 「尙書古訓과 尙書古注 연계 시맨틱 데이터 프로세싱 1: XML 데이터 설계 및 편찬」, 『한문학논집』 69, 근역한문학회, pp. 293-294 참조.

## 2.1. 『진단학보』의 문헌적 특징

첫째, 『진단학보』의 문헌이 가진 특징을 XML 설계에 충분히 반영해야 한다. 『진단학보』의 문헌적 특징은 다시금 형식적 특징 및 내용적 특징으로 나눌 수 있다. 『진단학보』는 한국의 근대적 학술 규범 및 관습의 확립과 학술적 글쓰기의 정착 과정을 보여주는 학술지이다. 학술 규범이 확립되었다는 것은 학술지 및 논문의 체계가 균질적이라는 것을 의미한다. 『진단학보』의 형식적 특징을 살펴보면, 『진단학보』 각 호의 편집 체제는 시기에 따라 약간의 차이를 제외하면 전체적으로 균질적이라는 것을 알 수 있다. 학술지 각 호는 표지, 목차, 논문, 강좌, 자료, 휘보, 판권 등으로 구성된다. 논문, 강좌, 자료 등은 분류상 차이가 있지만, 각각이 의미적으로 분절된 장(章)으로 구성된다는 점은 동일하다. 그리고 각 장은 본문 문장, 주석, 표, 그림 등으로 구성된다. 『진단학보』 문헌의 계층구조는 [그림 1]과 같다.[11]




[그림 1] 『진단학보』의 계층 구조

11 물론 『진단학보』의 개별 논문은 형식적 특징이 균일하지는 않다. 장(章) 하위에 절(節)이 있는 경우도 있으며, 그림, 표 등의 유무 역시 개별 문헌에 따라 다르다. 또한 '논문'이 아니라 '자료'의 경우 글 앞에 '소개글'이 있기도 하다.

www.kci.go.kr

구체적으로 살펴보면 해방 이전 『진단학보』는 1934년 11월에 제1호를 창간하였고, 1941년 5월 제14호를 간행하면서 정간한다. 『진단학보』 제3호는 표지, 목차, 논문 6편, 강좌 1편, 자료 1편, 휘보, 판권 등으로 구성된다. 『진단학보』 제3호의 세 번째 논문인 손진태의 「지나민족의 웅계신앙과 그 전설」은 제1장 「웅계에 관한 신앙」과 제2장 「계신전설」로 구성된다. 아울러 각 장은 본문 및 주석으로 구성된다. XML 설계 시에 『진단학보』의 문헌학적 특징을 충분히 반영해야 한다. 『진단학보』가 학술지의 성격을 가지고, 개별 논문이 학술적 문헌이라는 점, 따라서 장이 나누어져 있고 주석으로 출처를 표시하거나 보충 설명을 제시한다는 점 등을 유의할 필요가 있다. 또한 책자 형태로 출판된 아날로그 문헌을 디지털화하는 것이기에, 쪽수 역시 반영할 필요가 있다.

다른 한편, 『진단학보』 문헌의 내용적 특징 또한 고려해야 한다. 『진단학보』에 실린 개별 문헌에서, 연구자가 연구의 대상이 된 '문헌'에 대한 검토를 바탕으로 다양한 '용어'를 사용하면서 자신의 '학설'을 제시한다. 동시에 연구자는 기존의 선행 연구 '문헌'에 대한 동의 및 비판 역시 수행하게 된다. 예컨대, 손진태는 『수서(隋書)』, 『태평어람(太平御覽)』 등('문헌')에 대한 검토를 바탕으로 고대 중국인의 '주술적 행위'('용어')를 기존의 통념과 달리 새롭게 해석하였다. 이 글을 통해 손진태는 "중국의 고대인이 웅계에게 주술력이 있다고 믿었다.", "닭이 울면 귀신이 숨는다고 믿었다.", "중국 고대의 웅계신앙은 음양오행설에 선행한다.", "웅계신앙은 동아시아 고대 여러 민족의 보편적 신앙이다." 등('학설')을 논증하였다. 논증의 과정에서는 '문헌' 외에 '인물', '시간', '공간', '기관', '단체' 등의 다양한 정보를 함께 활용한다. 따라서 『진단학보』 문헌의 내용적 특징을 고려할 때, '인물', '시간', '공간', '기관', '단체', '문헌' 등의 객관적인 정보와 '용어', '학설' 등은 주관적인 정보에 유의하여 XML을 설계할 필요가 있다.[12]

www.kci.go.kr

12 '인물', '시간', '공간', '기관', '단체', '문헌' 등은 문헌에서 정확하게 추출해야 할 객관정

## 2.2. 데이터의 활용 가능성

둘째, 데이터의 활용 가능성을 적극적으로 고려하여 XML을 설계해야 한다. 앞서 『진단학보』의 형식적 특징 및 내용적 특징을 살펴본 것 역시 정밀한 데이터 편찬 및 확장된 활용의 토대가 된다. 정교한 설계를 바탕으로 편찬된 XML 데이터는 RDB(Relational DataBase, 관계형 데이터베이스)에 적재하고, RDB로부터 JSON(JavaScript Object Notation) 및 RDF(Resource Description Framework) 등 웹 표준 포맷의 데이터로 변환한다. 표준적 형태로 변환한 『진단학보』 데이터를 널리 공유하여, 시민 및 연구자들이 자유롭게 데이터를 활용할 수 있다. 동시에 『진단학보』의 데이터 공유를 통해서 다양한 한국 근대문헌의 데이터와 연결하여 보다 거대한 규모의 데이터를 축적할 수 있다. 동시에 저자들이 설계한 XML의 유효성 및 타당성에 대한 토론 및 검증을 거쳐 보다 정밀하고 한국 근대문헌의 특징을 선명히 반영한 XML 설계로 보완이 가능할 것이다. XML 설계는 한 번으로 마무리되는 것이 아니라, 추후의 공유 및 사용자의 피드백에 기반한 보완 과정 등을 포함한 과정으로 이해할 필요가 있다.

또한, 문헌의 특징을 반영한 설계뿐 아니라, 연구자로서 아카이브 구축 시 어떠한 요소를 XML 설계에 반영해야 하는지 고민할 필요가 있다. 첫째, 개별 근대문헌 개별 페이지에 대한 정보의 표준화를 시도하였다. 현재 국사편찬위원회에서는 간략한 서지 정보 및 본문 플레인 텍스트 안에 쪽수만을 표시하고 있다. 저자들은 여기에서 더 나아가 메타데이터, 해제, 목차, 본문, 데이터, 주석의 형식으로 하나의 XML 문서를 설계하고자 한다. 메타데이터는 문헌의 접근성을 고려하여 상세하게 정리한다. 구체적으로는 '제목',

---

보이다. '용어' 및 '학설'은 개별 문헌 및 그것이 놓인 연구의 맥락에 따라서, 연구자들의 인문학적 통찰을 통해 추출할 필요가 있다. 따라서 객관정보의 범주 및 추출이 정확한지, 또한 주관정보의 범주 설정이 타당성을 가지는지 공개하고 지속적으로 검토할 필요가 있다.

www.kci.go.kr

‘원제목’, ‘학술지’, ‘수록권호’, ‘발행기관’, ‘저자’, ‘역자’, ‘집필일자’, ‘게재연월’, ‘시작쪽’, ‘종료쪽’, ‘전체쪽’, ‘연재여부’, ‘범주’, ‘분야’ 등을 설정하였다. 서지는 앞서 살펴본『진단학보』의 계층 구조를 참조하여 상세히 구성하였으며, 다양한 일자 및 쪽수정보는 정밀한 데이터 처리를 위한 기초 데이터가 된다. 또한 ‘범주’는 진단학보에서 구별한 양식인 논문, 강좌, 자료를 준용하였고, ‘분야’는 당대 및 현재에 통용되는 분류를 고려하여 역사학, 문학, 민속학, 고고학 등으로 분류하였다. 또한 ‘해제’를 편성하여 문헌의 핵심적인 용어 및 학설을 정리하고, 다른 문헌과의 연결 관계를 정리하였다. ‘목차’는 문헌 1편을 세부 장(chapter)과 절(section)별로 구조화할 수 있도록 설정하였다. ‘본문’은 문헌 원문의 본문 및 인용문을 단락별로 입력하였다. 또한 쪽수 및 주석 정보 역시 부기하였다. 근대문헌을 본문뿐 아니라, 메타데이터, 해제, 목차, 주석 등을 함께 입력하는 것은 문헌에 관한 다양한 상세 정보를 함께 편찬하여 추후 더 정교한 데이터 작업을 도모하는 한편, 상세 정보를 활용하여 사용자가 더 편리하게 문헌에 대한 정보를 얻을 수 있도록 하기 위해서이다.

둘째, 한국 근대문헌의 텍스트 구축 시 원문과 현대어 교열본을 함께 작성하였다. 한국 근대문헌은 20세기에 출판된 문헌이기는 하지만, 조선어학회가 한글맞춤법통일안을 제정한 것은 1933년이었으며, 그것이 공식적으로 채택된 것은 1948년이었다. 따라서 1948년 이전 한국 근대문헌은 표기법의 측면에서 균질적이지 않다. 당시에는 한글과 한자를 혼용하여 표기하는 경우가 많았으며, 오식 및 오류 역시 확인할 수 있다. 한국 근대문헌의 원문을 그대로 재현하여 입력하는 ‘원문’과 그것을 교열한 ‘현대문’을 모두 데이터로 입력하기로 하였다. 텍스트 입력과정에서 해당 분야에 전문적인 지식을 갖춘 연구자가 원문의 오식 및 오류를 바로잡고, 한글 표기(한자 병기)를 원칙으로 비평적 정본인 ‘현대문’을 입력하였다. ‘현대문’ 역시 ‘원문’과 함께 단락별로 입력하여, 필요시 두 판본의 특정 부분을 대조할 수 있도록 하였다. 이것은 추후 아카이브를 구축할 때, 원문과 현대문을 나란히 병

www.kci.go.kr

기하여 아카이브의 활용도를 높일 수 있다. 또한 상호 비교가능한 원문 데이터와 현대문 데이터를 함께 구축하여, 추후 원문 및 현대문에 대한 계량적 연구의 가능성을 열어두었다.

### 2.3. 한국 근대문헌 XML 스키마 설계

『진단학보』 논문의 문헌적 특징 및 데이터의 활용 가능성을 고려하여, 개별 문헌의 XML 스키마를 설계하였다. 『진단학보』 논문의 XML 문서는 ① 메타데이터 섹션, ② 목차 및 원문 섹션, ③ 현대문 섹션, ④ 주석 섹션으로 구성하였다. 각각의 섹션은 원문의 형식을 반영한 '형식 요소(Textual Element)', 편찬자가 개입하여 텍스트로부터 도출하여 정리한 '문맥요소(Contextual Element)' 등으로 나눌 수 있다. 아래에서는 「지나민족의 웅계신앙과 그 전설」을 통해 XML 구조를 살펴보겠다.

#### ① 메타데이터 섹션

메타데이터 섹션의 구조는 [표 1]에서 볼 수 있듯, 크게 형식 요소와 문맥 요소로 구분된다. 형식 요소인 기본 정보는 다시금 문헌 ID, 제목, 출처 정보, 저자 정보, 쪽수 정보로 구성된다. 제목, 출처 정보, 저자 정보, 쪽수 정보 등은 다시금 하위 정보로 구성된다. 문맥 요소인 맥락 정보는 분류, 분야, 해제로 구성된다. 「지나민족의 웅계신앙과 그 전설」의 '메타데이터 섹션'의 XML 구조 예시는 [표 2]과 같다.

www.kci.go.kr

[표 1]『진단학보』 논문 메타데이터 섹션의 XML 요소(element) 분류

| 섹션 | 분류1 | 분류2 | 분류3 | 분류4 | 설명 |
|---|---|---|---|---|---|
| 메타데이터 | 형식 요소 | 기본 정보 (Basic) | 문헌(Id) | | 문헌 ID |
| | | | 제목 (Title) | 대표명 제목(Rtitle) | 대표명 제목 |
| | | | | 원제목(Otilte) | 제목 원문 표기 |
| | | | 출처 정보 (Source) | 학술지(Journal) | 학술지 |
| | | | | 학술지권호(Volume) | 수록 권호 |
| | | | | 발행기관(Publisher) | 발행기관 |
| | | | | 게재연월(Ptime) | 게재연월 |
| | | | | 집필일자(Wtime) | 집필일자 |
| | | | 저자 정보 (Authorship) | 저자(Author) | 저자 이름 |
| | | | | 역자(Translator) | 역자 이름 |
| | | | 쪽수 정보 (PageInfo) | 시작 쪽(Spage) | 시작 쪽 |
| | | | | 종료 쪽(Epage) | 종료 쪽 |
| | | | | 전체 쪽(Tpage) | 전체 쪽 |
| | 문맥 요소 | 맥락 정보 (Context) | 분류(Category) | | 분류 |
| | | | 분야(Field) | | 분야 |
| | | | 해제(Commentary) | | 문헌 해제 |

www.kci.go.kr

[표 2] 『진단학보』 논문 메타데이터 섹션의 XML 구조

〈Metadata〉
〈Basic〉
〈Id〉JDA193504150〈/Id〉
〈Title〉
〈Rtitle〉지나민족의 웅계신앙과 그 전설〈/Rtitle〉
〈Otitle〉支那民族의 雄鷄信仰과 그傳說〈/Otitle〉
〈/Title〉
〈Source〉
〈Journal〉진단학보〈/Journal〉
〈Volume〉진단학보 3〈/Volume〉
〈Publisher〉진단학회〈/Publisher〉
〈Ptime〉1935년09월〈/Ptime〉
〈Wtime/〉
〈/Source〉
〈AuthorInfo〉
〈Author〉손진태〈/Author〉
〈Translator/〉
〈/AuthorInfo〉
〈PageInfo〉
〈Spage〉076〈/Spage〉
〈Epage〉092〈/Epage〉
〈Tpage〉017〈/Tpage〉
〈/PageInfo〉
〈/Basic〉
〈Context〉
〈Category〉논문〈/Category〉
〈Field〉민속학〈/Field〉
〈Commentary〉손진태의 「지나민족의 웅계신앙과 그 전설」은 『진단학보』 3권(1935.9.)에 실린 글이다. 이 글은 1장 "웅계에 관한 신앙"과 2장 "계신전설"로 구성된다. 〔…중략…〕 이청원은 「『진단학보』 제3권을 읽고」(『조선중앙일보』, 1935.11.9~14.)에서 이 글이 분석한 중국의 웅계신앙의 특징은 조선의 웅계신앙을 설명하는 데 유익하다고 주장하였다. 김태준은 「『진단학보』 제3권을 읽고」(『조선중앙일보』, 1935.11.15~19.)에서 이 글이 천계전설이 조선 고유의 전설이 아니라 이웃나라와 공유하는 것이라는 점을 보여주었다는 점에서 의미를 찾았고, 조선문화의 연구를 위해 세계문화 일반에 대한 연구가 필요하다고 주장하였다.〈/Commentary〉
〈/Context〉
〈/Metadata〉

www.kci.go.kr

### ② 목차 및 원문 섹션

[표 3]『진단학보』논문 목차 및 원문 섹션의 XML 요소(element) 분류

| 섹션 | 분류1 | 분류2 | 분류3 | 분류4 | 분류5 | 설명 |
|---|---|---|---|---|---|---|
| 목차 | 형식 요소 | 목차 정보 (index Info) | 목차 (index) | | | 문헌의 목차 (장 단위) |
| 원문 | 형식 요소 | 원문 텍스트 (Origin) | 원문 텍스트 장 (Ochap) | 원문 테스트 절 (Osect) | 원문 텍스트 단락 (Opara) | 전체 – 장 – 절 – 단락 단위의 원문 텍스트 |
| | | 이미지 (image) | | | | 원문 텍스트의 이미지 |
| | | 쪽수 (page) | | | | 원문 텍스트의 쪽수13 |
| | | 주석 (ref) | | | | 원문 텍스트에서 주석이 달린 단어나 구문을 표시 |

목차 섹션은 [표 3]에서 볼 수 있듯, 형식 요소인 목차 정보로 구성된다. 목차 정보는 문헌의 각 장/절으로 구성된다. 원문 섹션 역시 형식 요소인 원문 텍스트, 쪽수, 주석으로 구성된다. 원문 텍스트를 전체, 장, 절, 단락 등 4단계로 나누고 단락 단위로 입력한다. 또한 형식요소로서 원문의 쪽수와 주석이 달린 텍스트 부분(단어나 구문) 및 해당 주석의 순서 번호 역시 입력할 수 있도록 한다.「지나민족의 웅계신앙과 그 전설」의 '목차 및 원문 섹션'의 XML 구조 예시는 [표 4]와 같다.

13 하나의 단락이 두 개 이상의 쪽으로 나누어질 경우, number를 ○-1과 ○-2 등으로 구분한다.

www.kci.go.kr

[표 4] 『진단학보』 논문 목차 및 원문 섹션의 XML 구조

```
〈indexInfo id="JDC193504150"〉
 〈index order="1"id="JDC19350415001"〉웅계에 관한 신앙〈/index〉
 〈index order="2"id="JDC19350415002"〉계신전설〈/index〉
〈/indexInfo〉
〈Origin id="JDA19350415001"〉
 〈Ochap id="JDA1935041500101"〉
  〈Opara id="JDA193504150010101"〉〈page number="76-1"〉 漢 應劭의 「風俗通義」(漢魏叢書本) 卷八 雄雞條에〈/page〉〈/Opara〉
  〈Opara id="JDA193504150010102"〉〈page number="76-2"〉 俗說, 雞鳴將旦, 爲人起居, 門亦昏閉談開, 扞難守固, 禮貴報功, 故門戶用雞也, 青史子書說, 雞者東方之牲也, 歲終更始, 辨秩東作, 萬物觸戶而出, 故以雞祀祭也, 太史丞鄧平說, 臘者所以迎刑送德也, 大寒至, 常恐陰勝, 故以成日臘, 戍者溫氣也, 用其氣日, 殺雞以謝刑德, 雄著 門, 雌著戶, 以和陰陽, 調寒配水, 節風雨也,〈/page〉〈page number="77"〉 蓮按, 春秋左氏傳, 周大夫賓孟適郊, 見雄雞、自斷其尾, 歸以告景王曰, 憚其爲犧也, 山海經日, 祠鬼神, 皆以雄雞, 魯郊祀, 常以丹雞祀日, 以其朝聲赤羽, 去魯候之咎, 今人卒得鬼刺排, 悟殺雄雞以傳其心上, 病賊風者, 作雞散東門, 雞頭可以治蠱, 由此言之, 雞主以禦死壁惡也〈/page〉〈/Opara〉
                                〔…중략…〕
  〈Opara id="JDA193504150010105"〉〈page number="78-2"〉〈ref id="JDA193504150001" order="1"〉元旦, 縣官殺羊, 縣其頭於門, 又磔鷄以副之, 俗說以厭癘氣, 元以間河南代君, 伏君曰, 是月也, 主氣上升, 草木萌動, 齧百草, 鶴啄五穀, 故殺之, 以助生氣〈/ref〉〈/page〉〈/Opara〉
                                〔…중략…〕
  〈Opara id="JDA193504150010219"〉〈page number="91-6"〉 上述한 黃父傳說 重明鳥傳說 及 天雞(玉難) 傳說 等은 要컨대 그 思想根源 原始民俗思想에 發하였음을 짐작할 수 있으며, 門戶祭에 雄雞를 使用하는 咒述行爲는 磔雞를 쓰는 것이 原始型이오 木雞, 土雞, 畵雞, 鑄雞 等은 그 發達型일 것이며, 雞는 晨을 告하는, 異常性에 因하야 이것은 光明을 招來하고 鬼類를 隱伏케 하는 瑞鳥 또는 神聖한 鳥이라고 崇拜되었던 것을 알 수 있다. 나는 以上에서 原始宗教思想 異常性의 關係에 就하야 그 一瑞를 論한 것이다.〈/page〉〈/Opara〉
 〈/Ochap〉
〈/Origin〉
```

www.kci.go.kr

### ③ 현대문 섹션

[표 5] 『진단학보』 논문 현대문 섹션의 XML 요소(element) 분류

| 섹션 | 분류1 | 분류2 | 분류3 | 분류4 | 분류5 | 설명 |
|---|---|---|---|---|---|---|
| 현대문 | 형식요소 | 현대문 텍스트 (Trans) | 현대문 텍스트 장 (Tchap) | 현대문 텍스트 절 (Tsect) | 현대문 텍스트 단락 (Tpara) | 전체 – 장 – 절 – 단락 단위의 원문 텍스트 |
| | 문맥요소 | 객관정보 | 문헌 (Book) | | | 문헌에서 제시한 문헌 정보 |
| | | | 인물 (Person) | | | 문헌에서 제시한 인물 정보 |
| | | | 공간 (Location) | | | 문헌에서 제시한 공간 정보 |
| | | | 시간 (Time) | | | 문헌에서 제시한 시간 정보 |
| | | | 단체 (Group) | | | 문헌에서 제시한 단체 정보 |
| | | | 기관 (Institute) | | | 문헌에서 제시한 기관 정보 |
| | | 주관정보 | 학설 (Argument) | | | 문헌에서 도출한 학설 |
| | | | 용어(Term) | | | 문헌에서 도출한 용어 |

현대문 섹션은 [표 5]에서 볼 수 있듯, 형식요소와 문맥요소로 구분된다. 현대문의 형식요소는 현대문 텍스트를 전체, 장, 단락 등 3단계로 나누고 단락 단위로 입력한다. 현대문의 문맥요소는 객관정보와 주관정보로 구분된다. 편찬자들은 온톨로지 설계를 참조하여 텍스트에 객관정보로서 문헌, 인물, 공간, 시간, 단체, 기관 등의 정보를 마크업한다. 또한 주관정보로서 용어와 학설을 마크업한다. 마크업한 데이터는 추후 별도의 데이터 프로세싱이 가능하다. 「지나민족의 웅계신앙과 그 전설」의 '현대문 섹션'의 XML 구조 예시는 [표 6]과 같다.[14]

www.kci.go.kr

14 본 논문에서 제시한 XML 예시의 문맥요소 id 속성값은 임시로 부여한 값이다. 현재 『진

[표 6]『진단학보』논문 현대문 섹션의 XML 구조

〈Trans id="JDA19350415001"〉
〈Tchap id="JDA19350415001"〉
〈Tpara id="JDA193504150010101"〉한(漢) 〈Person id="P0001" name="응소"〉응소(應劭)〈/Person〉의 「〈Book id="B0001" name="풍속통의"〉풍속통의(風俗通義)〈/Book〉」(한위서총서본(漢魏叢書本)) 제8 웅계조에〈/Tpara〉
〈Tpara id="JDA193504150010102"〉俗說, 雞鳴將旦, 爲人起居, 門亦昏閉談開, 扞難守固, 禮貴報功, 故門戶用雞也, 青史子書說, 雞者東方之牲也, 歲終更始, 辨秩東作, 萬物觸戶而出, 故以雞祀祭也, 太史丞鄧平說, 臘者所以迎刑送德也, 大寒至, 常恐陰勝, 故以成日臘, 戌者溫氣也, 用其氣日, 殺雞以謝刑德, 雄著 門, 雌著戶, 以和陰陽, 調寒配水, 節風雨也, 蓮按, 春秋左氏傳, 周大夫賓孟適郊, 見雄雞、自斷其尾, 歸以告景王曰, 憚其爲犧也, 山海經日, 祠鬼神, 皆以雄雞, 魯郊祀, 常以丹雞祀日, 以其朝聲赤羽, 去魯候之咎, 今人卒得鬼刺排, 悟殺雄雞以傳其心上, 病賊風者, 作雞散東門, 雞頭可以治蠱, 由此言之, 雞主以禦死壁惡也〈/Tpara〉
〈Tpara id="JDA193504150010103"〉라 하야 한대(漢代) 민간에서 엽일(臘日)에 웅계를 죽이어 문호의 제(祭)에 저용(著用)한 습속에 취(就)하야 그것을 설명하는 당시의 속설과 청사자설(青史子說), 등평설(鄧平說) 등 소개한 후 응소 자신의 설을 끝으로 기록하였다. 그 소위 속설에 의하면 닭(계명장단(鷄鳴將旦) 이러한 것을 보면 웅계을 가러침이다)과 문호는 인간생활에 대하야 서로 유사한 성질의 공헌을 하는 것이므로 문호의 제사에 닭을 쓴다는 것이다. 〈Person id="P0002" name="청사자(青史子)"〉청사자(青史子)〈/Person〉의 설은 명백히 오행설(五行說)이니 원단(元旦)으로부터는 봄이 시작되고 봄은 농작물 산출 최시(最始) 준비기이므로 이것은 마치 만물의 촉출(觸出)하는 문호와 유사한 성질을 가지었다. 그런대 춘(春)은 오행설로 보면 동방에 속하는 계절이오 (예하면 동(東)을 동춘(東春), 춘풍을 동풍(東風), 춘작(春作)을 동작(東作)이라고도 한다) 닭도 동방의 생(牲)이므로 문호의 제에 닭을 쓴다는 것이다. 이 설을 쫓는다면 문제용계(門祭用雞)의 습속은 오행설이 생긴 이후의 것이라고 보지 아니할 수 없다. 그러나 우리는 그것을 믿을 수 없다. 등평설은 음양설적(陰陽說的) 해석이오 그 논지는 명확을 결(缺)하였으나 엽일에 닭을 쓰는것은 사형덕(謝刑德) 화음양(和陰陽) 조한배수(調寒配水) 절풍우(節風雨) 등의 의미를 가진 것이라는 것이다. 그러나 이것을 따른다면 문제용계는 음양설이 상당히 발달된 이후의 이론상 산물이리라고 볼 수밖에 없으므로 우리는 기종(斯種)의 민간신앙의 기원을 그러한 철학 상에 구할 수는 없다. 끝으로 보이는 응력(應力)의 설을 보면 그는 당시 민간에 있던 삼종(三種)의 〈Term id="T0001" name="주술적 행위"〉주술적(咒術的) 행위〈/Term〉를 예거하야 전(前) 삼설(三說)에 반대하고 문금용계는 요컨대 어사벽악(禦死辟惡)의 주술적 행위에 불과하다고 하였다.〈/Tpara〉
〔…중략…〕
〈/Tchap〉
〈/Trans〉

---

단학보』 전체 텍스트에 대한 정리 및 데이터 편찬 작업이 진행 중에 있으며, 향후 모든 문맥요소에 대한 종합적인 식별과 정리가 완료된 후에야 체계적이고 일관된 방식으로 최종 ID가 부여될 예정이다. 따라서 본문에 제시된 XML 예시는 전체 구조와 요소 간 관계를 보여주기 위한 참고용임을 밝혀둔다.

www.kci.go.kr

### ④ 주석 섹션

[표 7]『진단학보』 논문 주석 섹션의 XML 요소(element) 분류

| 섹션 | 분류1 | 분류2 | 분류3 | 분류4 | 분류5 | 설명 |
|---|---|---|---|---|---|---|
| 주석 | 형식 요소 | 주석 (Annot) | 각주 (Footnote) | | | 주석 (주석 데이터는 원문과 현대문 모두 입력) |

주석 섹션은 [표 7]에서 볼 수 있듯, 형식요소로서 주석 및 각주로 구성된다. 각주는 원문과 현대문 각각 입력한다. 원문 섹션의 주석번호 〈ref〉를 통해 텍스트에서 주석번호의 위치를 특정하며, 〈Footnote〉를 통해 주석(각주)의 내용을 입력한다.[15] 「지나민족의 웅계신앙과 그 전설」의 '주석' 부분의 XML 구조 예시는 [표 8]과 같다.

[표 8]『진단학보』 논문 주석 섹션의 XML 구조

〈Annot〉
〈Footnote type="original" id="JDA193504150001"〉唐 歐陽詢「藝文類聚」卷四 及 宋 李昉의「太平御覽」卷二九와 卷九一八 等에도 大同小異의 引文이 보이고 伏君은 藝文類聚에 任君으로 되어 있다.〈/Footnote〉
〈Footnote type="original" id="JDA193504150002"〉「說郛」所收「四時纂鏡」에도 同樣의 記事가 보인다.〈/Footnote〉
〈Footnote type="original" id="JDA193504150003"〉白은 衍字인 듯하다.〈/Footnote〉
〔…중략…〕
〈Footnote type="translation" id="JDA193504150001"〉당(唐) 구양수(歐陽詢)「예문유취(藝文類聚)」권4 급(及) 송(宋) 이방(李昉)의「태평어람(太平御覽)」권29와 권918 등에도 대동소이의 인문(引文)이 보이고 복군(伏君)은 예문유취(藝文類聚)에 임군(任君)으로 되어 있다.〈/Footnote〉
〈Footnote type="translation" id="JDA193504150002"〉「설부(說郛)」소수(所收)「사시윤경(四時纂鏡)」에도 동양(同樣)의 기사가 보인다.〈/Footnote〉
〈Footnote type="translation" id="JDA193504150003"〉백(白)은 연자(衍字)인 듯하다.〈/Footnote〉
〔…중략…〕
〈/Annot〉

15 『진단학보』의 문헌에는 본문의 주석 외에 해제의 주석, 번역문의 주석 등 다양한 주석이 있기에, 필요시 추가로 주석의 하위 종류를 추가한다.

www.kci.go.kr

# 3. 『진단학보』 디지털 텍스트 편찬 및 쟁점

저자들은 『진단학보』의 문헌적 특징을 기반으로 설계한 XML 스키마를 바탕으로 편찬 규칙을 정립하고, 이에 따라 텍스트 데이터를 체계적으로 편찬하고 있다.[16] 이 장에서는 현재까지의 편찬 진행 상황과 그 과정에서 도출된 텍스트 편찬의 주요 쟁점들을 차례로 살펴보도록 하겠다.

## 3.1. 텍스트 데이터 편찬

저자들은 『진단학보』 문헌의 특징을 반영한 XML 스키마를 설계하고, 데이터 입력을 시도하고 있다. 데이터 입력의 도구로는 미디어위키(MediaWiki)를 사용하였다. 위키는 공동의 작업자들이 실시간으로 연구 진행 상황을 확인하면서 데이터를 입력 및 보완할 수 있으며, 수정 이력을 자동으로 추적하여 이전 버전으로의 복원이 용이하다는 장점이 있다. 또한 간단한 마크업 언어(Markup Language) 또한 활용할 수 있다는 점에서 협업에 유리한 플랫폼이다. 또한 위키에는 'XML 내보내기' 기능이 있어서, 미디어위키로 편찬한 데이터를 기초 XML 전자문서로 변환할 수 있다.[17] 다만, 변환한 기초 XML 문서를 바로 데이터로 활용할 수 있는 것은 아니기에, 기초 XML 문서를 파이썬(python) 코드를 활용하여 앞서 설계한 표준 스키마에 근거한 Valid XML 데이터로 변환해야 한다. 저자들은 'XML 내보내기'와 Valid XML 변환을 염두에 두면서 『진단학보』 문헌 데이터 입력을 위한 위키 기본 틀을 만들었다.[18]

---

16 『진단학보』 원문 입력 작업 규칙에 대해서는 「JD작업규칙」을 참조할 수 있다. https://dh.aks.ac.kr/~nkh/wiki/index.php/JD작업규칙

17 김지선·장문석·류인태(2021), 「공유와 협업의 글쓰기 플랫폼, 위키」, 『한국학연구』 60, 인하대 한국학연구소 참조.

18 「지나민족의 웅계신앙과 그 전설」의 입력에는 다음 틀을 사용하였다. https://dh.aks.ac.kr/~nkh/wiki/index.php/JDArticle2 데이터 편찬 도구로서 미디어위키 및 XML 변환

이후 기본 틀을 활용하여 『진단학보』의 개별 문헌을 입력하였다. 현재 『진단학보』 제1호~제14호의 논문 가운데 '휘보'를 제외한 '논문', '자료', '강좌' 등 문헌 85편의 미디어위키 텍스트 데이터 입력을 목표로 하였으며, 현재 74편을 입력하였다.[19] 『진단학보』 문헌 85편의 분량은 아날로그 인쇄본 기준 2,527면이었다. 개별 문헌은 2~100면 분량이었으며, 평균 30면이었다. 작업 방식은 편찬자 6명이 400여 면(평균 분량 논문 13편)을 분담하여 입력하는 방식을 채택하였다. 데이터 편찬자의 전공은 역사학 3명 및 한국문학 3명이었다. 편찬자 1명은 ① 아날로그 문헌을 OCR 하여, 플레인 텍스트 초안을 확보한다. ② 확보한 플레인 텍스트를 아날로그 문헌 원문과 비교하여 정규화하여, 문헌 원문의 플레인 텍스트를 마련한다. ③ 원문의 플레인 텍스트를 규칙에 따라 교열하여 현대문 플레인 텍스트를 마련한다. ④ 원문 및 현대문 플레인 텍스트를 위키 입력틀에 입력한다. ⑤ 위키 입력틀에 메타데이터를 입력하고, 마크업할 정보요소를 태깅한다.[20] ⑥ 원문, 현대문을 검토하고 'XML 내보내기' 기능으로 표준 전자 XML 문서를 만든다. ⑦ 파이썬 코드를 활용하여 표준 스키마에 근거한 Valid XML 데이터로 변환한다. 입력자는 한 편의 문헌의 ①~⑤에 이르는 전 과정 데이터 편찬을 담당하였다. 미디어위키 데이터가 쌓이면 틈틈이 ⑥~⑦의 과정을 통해 위키데이터를 XML 데이터로 정리하고 있다. 「지나민족의 웅계신앙과 그 전

---

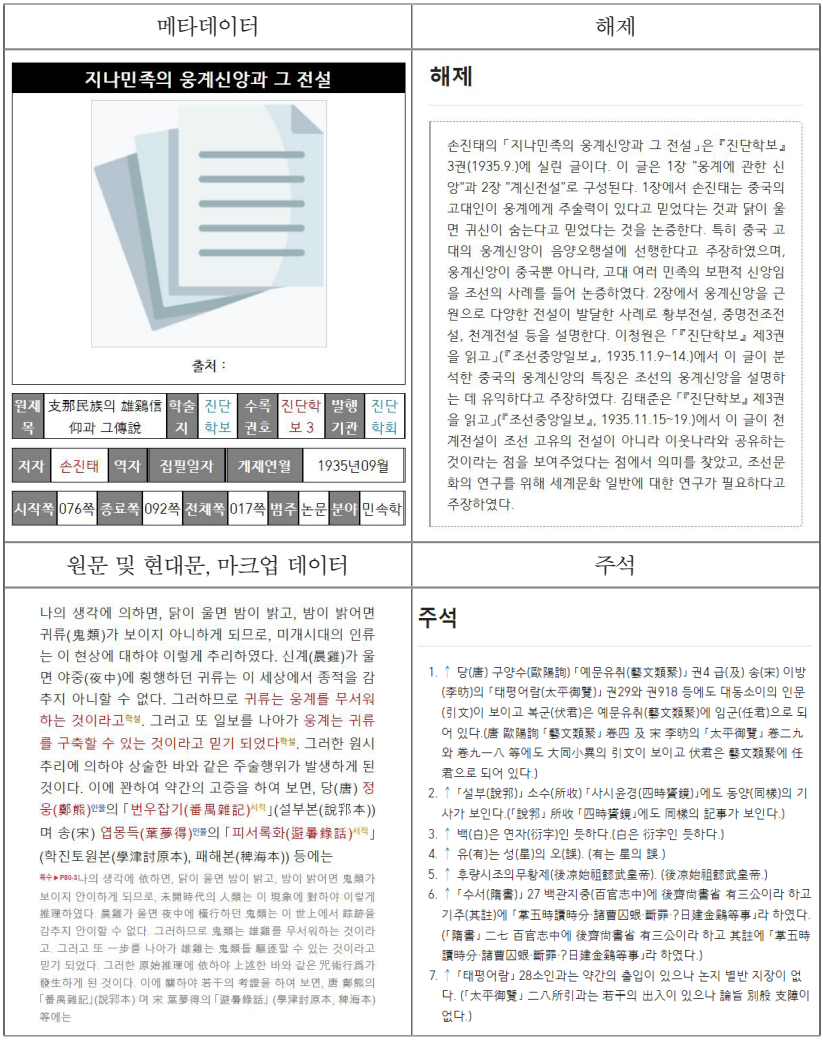
에 관해서는 추후 연구가 진행되면 다시금 정리할 예정이다.

19 현재 입력한 『진단학보』 논문의 위키데이터는 다음 페이지에서 확인할 수 있다. https://dh.aks.ac.kr/~nkh/wiki/index.php/분류:JDArticle

20 태깅할 정보요소는 XML 스키마 설계를 통해 확인한 '문맥요소' 중 객관정보로서 '문헌'(Book), '인물'(Person), '공간'(Location), '시간'(Time), '단체'(Group), '기관'(Institute)과 주관정보로서 '학설'(Argument), '용어'(Term) 등이다. 텍스트 데이터 편찬 과정에서 『진단학보』 온톨로지를 설계하고 이에 따라 개체 데이터를 마크업하였다. 『진단학보』 온톨로지 설계에 관해서는 홍종욱·장문석·류준필(2025), 「한국 근대문헌 디지털 아카이브의 설계: 『진단학보』를 중심으로」, 『인문연구』 110, 영남대 인문과학연구소 참조.

www.kci.go.kr

설」의 데이터를 미디어위키로 편찬한 예시는 [그림 2]와 같다.[21]




[그림 2] 『진단학보』 논문 Wiki 페이지 데이터

21 「지나민족의 웅계신앙과 그 전설」 미디어위키 편찬데이터. https://dh.aks.ac.kr/~nkh/wiki/index.php/지나민족의_웅계신앙과_그_전설


[그림 3] 『진단학보』 논문 XML 전자문서 데이터

『진단학보』 개별 문헌의 위키 데이터를 미디어위키의 'XML 문서 내보내기' 기능으로 기초 XML 문서로 내보낸다. 그리고 파이썬 코드를 활용하여 기초 XML 문서를 표준 스키마 기반의 Valid XML 데이터셋으로 변환한다. 「지나민족의 웅계신앙과 그 전설」의 XML 데이터는 앞의 [그림 3]과 같다.[22]

### 3.2. 데이터 편찬의 쟁점

데이터를 편찬하는 과정에서, 설계 과정에서 충분히 고려하지 못하였던 문헌의 특징이나 어려움을 만난다. 저자들 역시 데이터 편찬 과정에서 다양한 어려움을 만났으며, 그것을 해결하기 위해 다시금 토론하고 또한 XML 설계를 보완 및 수정하기도 하였다.

#### ① XML 설계의 보완

XML 설계 과정에서 『진단학보』 문헌의 특징을 검토하였으나, 실제 입력 과정에서는 초기 XML 설계로 입력하기 어려운 다양한 형식적 특징들이 발견되었다. 이러한 특징적 요소들은 디지털 텍스트 편찬에서 중요한 쟁점으로 부각되었다. 아래는 실제 편찬 작업에서 직면한 주요 쟁점들과 그 대응 방식이다.

첫째, 텍스트 계층 구조의 불규칙성을 처리하는 문제였다. 『진단학보』에 수록된 문헌들은 기본적으로 학술적 글쓰기라는 공통된 성격을 지니지만, 세부적인 텍스트 계층 구조에서는 불규칙성을 보였다. 대다수 문헌은 '장'(章) 단위로 구성되어 있으나, 일부 문헌에서는 '장' 구분 없이 단락 구

22 「지나민족의 웅계신앙과 그 전설」 XML 데이터. https://dh.aks.ac.kr/~nkh/xml/jd/JDA193504150.xml
위의 XML 데이터 예시의 문맥요소 id 속성값은 아직 최종 확정되지 않아 공란으로 두었다. 추후 『진단학보』 전체 텍스트의 데이터 편찬 작업이 완료된 후 일관된 체계로 ID가 부여될 예정이다.

www.kci.go.kr

분만으로 내용을 전개하기도 하였다. 특히 주목할 만한 사례는 도유호의 「중국도시문화의 기원」 1~3(제12호~제14호)이다. 이 논문은 1회분과 2회분에서는 '장'과 '절'(節)이 함께 사용되었으나, 3회분에서는 '절'만 등장하는 구조적 비일관성을 보였다. 또한 1회분에는 다른 글에서는 찾아볼 수 없는 '소개글'이 추가되어 있어, 3회에 걸친 연재분 모두가 서로 다른 구조적 특징을 가지는 특이성을 보여주었다. 이러한 다양한 계층 구조를 XML 스키마에 효과적으로 반영하기 위해 연구팀은 여러 차례 논의를 진행하였다. 그 결과, '장'(chapter)을 표현하기 위한 〈Ochap〉(원문 장) 및 〈Tchap〉(번역문 장) 태그와 '절'(section)을 표현하기 위해 〈Osect〉(원문 절) 및 〈Tsect〉(번역문 절) 태그를 새롭게 추가하였으며, section 태그는 chapter 태그의 하위 요소로 설정하여 계층적 관계를 명확히 하였다.

둘째, 다양한 형태와 성격의 주석을 일관성 있게 처리하는 방안이었다. 『진단학보』에는 각주, 협주, 역자주 등 여러 유형의 주석이 존재하여 이를 체계적으로 구분할 필요가 있었다. 최초 XML 설계 시에는 논문 본문의 주석에 대하여 주석의 위치를 〈ref〉 태그로, 주석의 내용을 〈Footnote〉 태그로 처리하고 고유 ID를 부여하여 상호 연계되도록 구성하였다. 그러나 실제 작업 과정에서 주석의 유형이 매우 다양하다는 것이 발견되었다. 특히 본문에 인용된 한문 사료 내에도 협주(夾註)가 달려 있는 경우가 다수 발견되었다. 이러한 한문 원사료의 협주는 본문 주석과 성격이 다르므로, 위키텍스트 입력 시 〈sup〉 태그를 사용하여 위첨자로 시각적 구분을 하고, XML 변환 시에는 〈originalNote〉 태그를 새롭게 설계하여 원문의 주석과 편찬자의 주석을 구분할 수 있게 하였다. 또한 이병도의 「난선 제주도 난파기(하멜 표류기)」 1~3회(제1호~제3호)에서는 또 다른 형태의 주석이 발견되었다. 이 글은 헨드릭 하멜의 표류기를 한국어로 번역한 것으로, 번역자 이병도는 본문 내에 괄호를 사용하여 역자주석을 직접 삽입하였다. 이에 대해 별도의 XML 태그로 처리할지를 논의하였으나, 『진단학보』 전체에서 이러한 역자주석은 예외적 사례로 판단하여, 별도 태그 없이 번역본문에 괄호 표시와

www.kci.go.kr

함께 역자주석을 그대로 입력하는 방식을 채택하였다.

셋째, 표와 도상 등 다양한 시각 자료를 효과적으로 디지털화하는 과제였다. 『진단학보』에포함된 시각 자료들은 원본의 특성을 유지하면서도 디지털 환경에 적합하게 변환하는 방법을 모색해야 했다. 도상 자료의 경우, 미디어위키의 '파일올리기' 기능을 활용하여 작업자들이 이미지를 업로드한 후 "[[파일:파일명.파일확장자명|섬네일|가운데|캡션입력]]" 형식으로 입력하였다. 이후 XML 문서 변환 시에는 〈image〉 태그를 사용하고, 도상에 해설문이 있는 경우 caption 속성에 해당 해설문을 체계적으로 정리하였다. 표의 경우에는 〈table〉 태그를 사용하여 처리하였으며, 표의 구조와 내용을 가능한 한 원본과 동일하게 유지될 수 있도록 하였다.[23]

넷째, 연구 활용성을 높이기 위한 편찬자의 해제 정보 설계 문제였다. XML 스키마에 논문에 대한 편찬자의 해제 정보를 중요한 요소로 포함하여 학술적 활용도를 높이고자 하였다. 해제는 〈Commentary〉 태그로 구조화하여 각 논문의 핵심 내용과 학술적 의의 등을 체계적으로 기술할 수 있도록 하였다. 해제 작성 과정에서는 해제 자체에 대한 주석 필요성이 제기되었는데, 이는 편찬자가 붙인 것으로 원문의 주석과는 성격이 다르기 때문에 명확한 구분이 필요했다. 이를 위해 기존의 〈ref〉 태그를 사용하되, 주석 내용은 〈Footnote type="commentary" id="고유ID"〉 형식으로 구분하여 해제 내 인용된 정보의 출처를 명확히 표시할 수 있게 하였다.

XML 설계의 지속적인 보완 과정은 『진단학보』 디지털화 작업에서 핵심적인 부분이었다. 초기 설계만으로는 예상하지 못했던 다양한 문헌적 특징들이 실제 데이터 편찬 과정에서 발견되었고, 연구팀은 각 사례에 대한 꾸준한 논의와 협의를 통해 문제를 해결해 나갔다. 이러한 섬세한 접근과

---

23 한국 근대문헌에 수록된 표는 기본적으로는 가로행 및 세로열 구조이지만, 각 표가 정리하는 정보의 특성에 따라 다양한 차이가 있다. 표의 다양한 형식을 위키의 표 입력 방식으로 충분히 표현하기 어려운 점이 있었다. 특히 김두헌(1935)의 「조선의 조혼 및 그 기원에 대한 일고찰」(『진단학보』 3, 진단학회)은 표 입력 과정에서 어려움이 상당했다.

지속적인 스키마 보완 작업이 있었기에 『진단학보』의 복잡하고 다양한 텍스트적 특성을 XML 스키마에 유연하게 반영할 수 있었으며, 원문의 특성을 충실히 보존하면서도 디지털 환경에서 효과적으로 활용 가능한 정밀하고 풍부한 데이터 구조를 구축해 나가고 있다.

### ② 현대문 텍스트의 정본화

저자들은 『진단학보』의 원문 텍스트와 현대문 텍스트를 함께 제공한다. 원문 텍스트 및 현대문 텍스트의 입력 규칙은 [표 9]와 같이 정하였다. 하지만 실제 현대문 변환 과정에서는 위의 규칙으로 판단이 쉽지 않은 다양한 사례를 만나게 되었다. 예컨대, 한자 '及'을 만났을 때, 소리를 살려 '급'으로 표기할지, 뜻을 살려 '및'으로 표기할지 선택한다. 저자들은 판단이 필요한 사례를 만나면, 토론을 하여 [표 10]과 같이 결정사항을 공유문서에 정리하면서 텍스트 변환이 일관성을 갖추도록 노력하였다.

1945년 이전에 출판된 근대문헌을 현대문으로 변환하는 규칙에 대해서는 아직 학계의 연구자들이 충분한 논의에 기반한 합의에 이르지 못하였다.

[표 9] 『진단학보』 원문 및 현대문 입력 규칙(초안)

| | 원문 | 현대문 |
|---|---|---|
| 맞춤법 | 원문 표기 존중 | 한글맞춤법을 따름 |
| 옛한글 | 원문 표기 존중 | 한글맞춤법을 따름 |
| 한자 및 한글 | 원문 표기 존중 | 한글 표기를 원칙으로 함. |
| 주요 개념 및 낯선 용어 | 원문 표기 존중 | 한글(한자)로 표기 |
| 띄어쓰기 | 한글맞춤법을 따름 | 한글맞춤법을 따름 |
| 오자 및 오식 | 바로 잡음 | 바로 잡음 |
| 숫자 | 원문의 한자는 한자로, 원문의 아라비아 숫자는 아라비아 숫자로 입력함 (한자 숫자 〇: 유니코드 3007) | 아라비아 숫자로 입력하는 것을 원칙으로 함 |
| 밑줄, 쉼표, 낫표 | 원문 표기 존중 | 한글맞춤법을 따름 |

www.kci.go.kr

[표 10] 『진단학보』 원문 및 현대문 입력 규칙 보완표(일부)

| 분류 | 변환의 쟁점 | 원문 | 현대문 | 비고 |
|---|---|---|---|---|
| 용언 및 수식언 | 맞춤법 | 대하야 | 대하여 | 현행 표기 |
| | | 일즉이 | 일찍이 | 현행 표기 |
| | 한자 및 한문 표기 | 作이 | 지음이, 짓는 것이 | 의미 해석 |
| | | 就하여는 | 대하여는 | 의미 해석 |
| | | 及 | 및 | 의미 해석 |
| | | 再言을不待할것이며, | 다시 말할 필요가 없으며, | 의미 해석 |
| | | 又 | 또, 또는, 또한 | 의미 해석 |
| | | 一云 | ~라고도 함 | 의미 해석 |
| | | 仍히 | 따라서, 인하여, 이어서 | 의미 해석 |
| | | 右注, 右文 | 위 주, 위 글 | 원문은 세로 쓰기이나, 현대문은 가로쓰기임. |
| | | 차절 次節 | 다음 절 | 의미 해석 |
| | | 전거前擧 | 앞서 든, 앞에서 들었던 | 의미 해석 |
| 체언 | 맞춤법 | 처-쵤 | 처칠 | 현행 표기 |
| | | 이·베·베루힌 | 이 베 베루힌 | 알지 못하는 인물은 중점 제거 |
| | 한자 및 한문 표기 | 松·栢의 實을 이름인 듯 | 소나무·잣나무의 열매를 이름인 듯 | 의미 해석 |
| | | 牛(牝牛)馬를 많이 畜養하여, | 소(암소), 말을 많이 축양(畜養)하여, | 의미 해석 |
| | | 吾人 | 우리 | 의미 해석 |
| | | 韃靼(滿洲淸人) | 달단(韃靼, 만주청인) | 한글로 변환, 필요 시 한자 병기 |
| | | 頁 | 쪽 | 의미 해석 |

www.kci.go.kr

| 분류 | 변환의 쟁점 | 원문 | 현대문 | 비고 |
|---|---|---|---|---|
| 기호 | 서지 낫표 | 「삼국유사」 | 「삼국유사」 | 그대로 |
| | 외래어 구분 낫표 | 「암스텔담」 | 암스테르담 | 삭제 |
| | 강조 낫표 | 「진단」 | ‘진단’ | 작은 따옴표 |
| | 줄표 | — | — | 그대로. 유니코드 2015 |
| | 말줄임표 | …<br>……<br>…… | … | 점 셋 말줄임표로 통일. 유니코드 22EF |

그동안 한국 근대문헌을 출판할 때에는 편자나 출판사에서 개별적인 규칙을 적용하여 현대문으로 문헌을 출판하였다. 앞으로 한동안은 근대문헌을 원문에서 현대문으로 변환할 때, 많은 어려움을 경험하게 될 것이다. 하지만 디지털 아카이브 구축을 계기로 학계에서 다양한 연구자들이 함께 논의하면서 근대문헌의 원문 표기에 관한 공동의 약속을 만들 수 있다. 디지털 아카이브 구축 시 표기 문제는 데이터 처리와 관련하여 중요한 문제인 동시에, 또한 디지털 환경에서 편찬한 텍스트의 경우 아날로그 문헌과 비교하여 용례의 확인 및 검토가 용이하기 때문이다.

연세대학교 근대한국학연구소는 1920~1930년대 한국 근대 신문 데이터베이스를 구축하면서, 표기가 통일되지 않은 명사 2,000여 개의 정규화 테이블을 구축하였다.[24] 연구의 기획 및 진행에 따라 연구자 개인, 연구소 등 다양한 연구 주체가 근대문헌 텍스트의 원문 변환을 수행하는 과정에서 발견한 다양한 문헌의 용례를 함께 논의하면서, 근대문헌 텍스트의 현대문 표기 원칙을 연구자들이 만들어갈 필요가 있다.

『진단학보』 원문의 현대문 변환 과정의 또 다른 쟁점은 원문의 오류 및

24 강범일(2025), 「근대 한국학 텍스트의 개체명 주석 연구: 1920~1930년대 신문 기사를 중심으로」, 『한국학』 48(1), 한국학중앙연구원, pp. 96–97.

www.kci.go.kr

오식이다. 『진단학보』를 최초에 간행 시, 원문에 오류가 있는 경우가 있다. 『진단학보』 논문의 원문 및 인용문의 분명한 오식은 어렵지 않게 바로잡을 수 있지만, 『진단학보』에는 무척 많은 언어 및 성격의 원문이 인용되어 있고, 인용 및 서술 과정에서 발생한 의미의 오류 역시 적지 않다. 『진단학보』 원문이 인용한 인용문의 경우, 정확한 텍스트 비평을 수행하기 위해서는 인용문의 원전과 대조하는 작업이 필요하지만, 널리 알려져 있거나 디지털화된 원전을 제외하면 인용문의 원전을 찾는 것은 거의 불가능하다. 예컨대, 이병도의 「진단변」(『진단학보』 1, 진단학회, 1934)의 서술에 등장하는 "혜원음의"(慧苑音義)는 한 단어로 보이지만, 실제로는 "혜원"(慧苑)과 "음의"(音義)라는 두 개의 불교용어를 잘못 이어 쓴 것이다. 표기의 오류를 넘어서 의미의 오류를 바로잡기 위해서는 해당 학문 영역의 전문적인 지식이 필요하다. 『진단학보』에 역사학, 문학, 고고학, 민속학, 종교학 등 다양한 학술 영역의 논문이 실리는 만큼, 다양한 학제의 연구자들이 텍스트 전환 과정에 참여하면 좋겠지만, 현실적으로 그것은 쉬운 일이 아니다. 이러한 상황에서 이미 해당 영역에서 비평판 전집을 출판한 연구자의 경우, 작업의 속도 및 신뢰성이 무척 향상되었다. 디지털 인문학 작업 역시 전통적인 인문학 영역의 성과를 토대로 진행된다는 것을 확인할 수 있다.

따라서 장기적으로 온라인 아카이브를 구축한 후, 데이터를 공개하여 관심과 전문적인 지식을 가진 시민 및 연구자들이 그 오류를 바로잡을 수 있는 경로를 마련할 필요가 있다. 또한 인문학 지식에 근거한 근대문헌 텍스트의 정본화 경험을 축적하면서, 인공지능과 협업하여 보다 효율적인 근대문헌 텍스트 편찬 가능성을 탐색할 필요가 있다.

#### ③ 기술적 문제

『진단학보』 제1호~제14호의 원문은 진단학회가 한국학술정보와 계약하여 KISS 사이트를 통해 제공하는 논문별 PDF 파일이 가장 널리 활용된다. KISS에서 제공하는 『진단학보』 원문은 이미지 파일 형태로만 서비스되

www.kci.go.kr

고 있으며, 이미지 처리 기술이 고도화되지 않았던 시기에 제작된 PDF 파일이어서 이미지 상태가 좋지 못하다. 또한 『진단학보』 논문 본문만 수록되어 있어 의 표지, 휘보 등의 일부 요소가 누락된 상태이다. 최근 국립중앙도서관에서는 『진단학보』 제2호~제12호, 제14호를 소해상도로 디지털화하여 제공하기 시작하였다. 저자들은 KISS 및 국립중앙도서관에서 제공하지 않는 『진단학보』 원문의 경우 직접 도서관을 방문하여 촬영함으로써 연구 자료를 확보하였다.

저자들은 『진단학보』 원문 이미지에 OCR 기술을 적용하여 플레인 텍스트 초안을 제작하였다. 이 과정에서 ABBYY 파인리더, 구글독스, Image to Text, Convertio, 누리IDT 등 다양한 OCR 프로그램을 활용하였다. 그러나 『진단학보』의 원문은 한글, 한자, 알파벳 등이 혼재되어 있고 세로쓰기 방식을 채택하고 있어, 현재의 OCR 프로그램으로는 정확한 문자 인식에 한계가 있었다.[25] 이에 따라 OCR 결과물과 『진단학보』 원문 이미지를 비교하는 텍스트 비평 작업이 필수적이었다. 이러한 교정 작업은 예상보다 많은 시간과 노력을 필요로 하였다. 영어문헌이나 일본어 문헌에 비해 한국 근대문헌의 OCR 인식률은 현저히 낮은 실정이다.

최근 영남대학교 인문과학연구소는 1895-1945년에 이르는 시기의 한국 근대문헌 연구를 위한 OCR의 성능을 분석하였다.[26] 한국 근대문헌 가운데 원문이 충분히 공개된 경우가 많지 않고, 원문이 공개된 경우에도 주로 이미지로 제공되고 있는 상황에서 근대문헌의 원문 텍스트 데이터 구축의

25 작업 과정에서 누리IDT(https://ocr.nuriidt.co.kr/)가 한자와 한글 인식에 탁월하다는 것을 확인할 수 있었으나, 유료 구독이 필요하다는 점에서 접근이 자유롭지는 못하였다.

26 윤경애·이철우·김영철·이현주·김유정·김인환(2025), 「한국 근대 문헌 연구를 위한 OCR 성능 분석」, 『인문연구』 110, 영남대 인문과학연구소 참조. 이 연구는 한국 근대문헌의 OCR 성능을 정밀도(Precision), 재현율(Recall), F1 점수(F1-score) 등으로 분석하였다. 분석의 결과, 최근 표기법을 반영하거나 그것에 가까운 문헌의 OCR 성능이 가장 높았고, 띄어쓰기가 없거나 옛한글이 많이 포함된 문헌의 인식률이 낮았다. 또한 한글 및 한자 혼용 문헌은 카카오 OCR과 구글 OCR이, 순한글문헌은 네이버 OCR이 높은 인식률을 보였다.

www.kci.go.kr

방법에 대해서는 더 많은 연구자들의 고민이 필요하다.

또한 현재 연구자들은 XML 변환을 염두에 둔 미디어위키 문헌 양식에 원문 및 현대문 텍스트 데이터를 입력하고 있다. 미디어위키를 데이터 입력 도구로 선택한 이유는 사용법이 직관적이고, 연구자들이 서로의 작업 상황을 실시간으로 공유하며 오류를 즉시 수정할 수 있는 협업 환경을 제공하기 때문이다. 그러나 미디어위키의 고유한 특성과 기능적 제한으로 인해 입력 과정에서 여러 불편함이 발생하기도 하였다. 특히 편찬자들은 동일한 작업을 반복하는 과정에서 효율성 문제에 직면했다. 줄 번호의 자동 계산 기능이나 작업에 특화된 단축키 등이 제공된다면 데이터 입력 효율을 크게 향상시킬 수 있을 것이다. 또한 근대문헌 원문에는 한자가 많이 포함되어 있어, 한자를 한글로 변환할 수 있는 방법 역시 모색할 필요가 있다. 현재까지 약 70여 편의 문헌에 대한 텍스트 데이터 입력과 정보 요소 마크업은 연구자들이 직접 수행하였다. 데이터 편찬 및 마크업 경험을 축적해 나가면서, 향후에는 인공지능과의 협업 가능성도 적극적으로 탐색하고자 한다. 이를 통해 작업 효율성을 높이고 보다 정교한 디지털 아카이브 구축을 모색하고자 한다.

## 4. 나가며

이 글은 『진단학보』를 사례로 한국 근대문헌의 디지털텍스트 편찬 방법 및 쟁점을 검토하였다. 현재 온라인에서 서비스되고 있는 근대문헌 아카이브가 메타데이터나 원문 이미지를 제공하는 것에 머물러 있다는 것을 성찰하면서, “아날로그 문헌자료를 어떻게 온라인 환경에서 데이터의 형식을 반영하여 편찬하고 공유할 수 있을 것인가?”, “인문학 연구자 및 시민이 효율적으로 활용할 수 있는 문헌자료의 디지털아카이브를 어떻게 구축할 것인가?”라는 질문을 살펴보았다.

www.kci.go.kr

이 글은 한국의 인문학 학술지『진단학보』의 제1호~제14호에 실린 논문을 대상으로 근대문헌 텍스트를 입력하는 방법을 탐색하였다. 보다 정밀한 데이터 처리를 위해서『진단학보』의 문헌적 특징 및 데이터의 활용 가능성을 고려한『진단학보』XML 스키마를 설계하고, 이를 바탕으로 미디어위키를 활용하여 텍스트 입력 체계를 구축하였다. 이를 통해『진단학보』논문 85편 가운데 74편의 텍스트 원문 데이터 및 현대문 데이터를 입력하였다. 이 과정에서 XML 설계의 지속적인 보완, 원문의 현대문 변환 규칙 정립, 그리고 작업 도구의 효율성과 한계 등 다양한 쟁점들을 확인할 수 있었다.

향후 저자들은 XML 설계를 더욱 정교화하고, 텍스트 데이터 편찬을 완성하며, 온톨로지에 따른 정보요소 마크업을 추가로 수행할 계획이다. 또한『진단학보』데이터를 그래프 데이터베이스로 이관하여 학술적 의미가 있는 질문을 질의어(Query)로 발견하고, 적절한 데이터 시각화 양식을 개발하여 아카이브의 완성도를 높여갈 예정이다. 다만, 현재까지는『진단학보』라는 연구대상의 텍스트 데이터를 웹상에 편찬하고 공유하는 데에 힘을 쓰느라 저자들 역시 자신의 문제의식에 따라 구축한 데이터를 충분히 활용하지 못하였다. 애써 구축한『진단학보』데이터가 보다 널리 자유롭게 연구에 쓰일 수 있는 방법을 모색할 예정이다.

현재 저자들이 진행하고 있는『진단학보』디지털 텍스트 편찬 방법이 유일한 것은 아니다. 저자들 역시 보다 효과적이며 효율적인 텍스트 편찬의 방법을 탐색 중이다. 이 글의 목적은 현재까지의 경험과 시행착오를 학계에 공유하여, 연구자 및 시민들과 함께 더 나은 방법론에 대한 개선책을 모색하는 데 있다. 이러한 취지에서 저자들은『진단학보』편찬 규칙 및 텍스트 데이터를 웹상에 공개하고 있으며, 추후 XML 데이터를 포함한 완성된 데이터 세트 전체를 공개할 예정이다.『진단학보』디지털 텍스트 편찬 과정에서 축적된 경험과 시행착오가 한국의 디지털 인문학 연구자들 사이에서 공유되어 한국 근대문헌 아카이브 구축의 더 발전된 방법론을 함께 모색하는 토대가 되기를 기대한다.

www.kci.go.kr

# 참고문헌

## 자료

진단학회(1934~1941), 『진단학보』 제1~14권, 진단학회.
http://dh.aks.ac.kr/~nkh/jd.html

## 논저

강범일(2025), 「근대 한국학 텍스트의 개체명 주석 연구: 1920~1930년대 신문 기사를 중심으로」, 『한국학』 48(1), 한국학중앙연구원.
김바로(2022), 「〈공공데이터법〉과 인문데이터: 공공기관 보유 인문데이터 공개 신청 사례를 중심으로」, 『한국고전연구』 57, 한국고전연구학회.
김바로(2024), 「국사편찬위원회 한국근현대잡지자료 데이터(2024.03.27.)」, 『디지털인문학』 1(1), 한국디지털인문학협의회.
김서윤(2024), 「조선시대 논어주석의 인용 분류체계」, 『민족문화』 67, 한국고전번역원.
김지선·장문석·류인태(2021), 「공유와 협업의 글쓰기 플랫폼, 위키」, 『한국학연구』 60, 인하대 한국학연구소.
김현(2006), 「고문헌 자료 XML 전자문서 편찬 기술에 관한 연구」, 『고문서연구』 29, 한국고문서학회.
김현주(2022), 「텍스트마이닝으로 본 『삼천리』의 문화」, 『문화』, 소화.
류인태(2022), 「인문학술 데이터 프로세싱에 관한 시론」, 『한국학』 45(2), 한국학중앙연구원.
변은미·이동학·류인태(2024), 「尙書古訓과 尙書古注 연계 시맨틱 데이터 프로세싱 1: XML 데이터 설계 및 편찬」, 『한문학논집』 69, 근역한문학회.
윤경애·이철우·김영철·이현주·김유정·김인환(2025), 「한국 근대 문헌 연구를 위한 OCR 성능 분석」, 『인문연구』 110, 영남대 인문과학연구소.
이재연(2014), 「작가, 매체, 네트워크: 1920년대 소설계의 거시적 조망을 위한 시론」, 『사이間SAI』 17, 국제한국문학문화학회.
이재연(2016), 「키워드와 네트워크: 토픽 모델링으로 본 『개벽』의 주제 지도 분석」, 『상허학보』 46, 상허학회.
전성규(2023), 「한국 근대 잡지의 계량적 연구 방법에 대한 논의: 코퍼스 구축 및 데이터 분석의 사례를 중심으로」, 『민족문학사연구』 82, 민족문학사연구소.
정성훈(2024), 「조선시대 한시의 경관 요소에 대한 계량적 분석: '소쇄원'과 '환벽당·식영정'을 중심으로」, 『한문학논집』 67, 근역한문학회.
지영원·최진경(2024), 「한국 한시 데이터 아카이브 구축을 위한 개념적 데이터 모델링 시론」, 『민족문학사연구』 85, 민족문학사연구소.
허수(2018), 「언어연결망 분석으로 본 20세기 초 한국의 '문명'과 '문화': 주요 언론 기사

www.kci.go.kr

에서의 논의 맥락을 중심으로」, 『개념과 소통』 22, 한림대 한림과학원.
홍종욱·장문석·류준필(2025), 「한국 근대문헌 디지털 아카이브의 설계: 『진단학보』를 중심으로」, 『인문연구』 110, 영남대 인문과학연구소.

원고 접수일: 2025년 5월 16일, 심사완료일: 2025년 5월 27일, 게재 확정일: 2025년 5월 27일

www.kci.go.kr

ABSTRACT

# Compilation of Digital Texts in Modern Korean Documents and Its Issues

## Focusing on *Chin-Tan Hakpo*

**Jang, Moon-seok***
**Kwag, Hannah****
**Moon, Ye-ji*****
**Shim, HyunJi******
**Ahn, Soyeon******
**Lee, Jihun*****
**Choi, Sol******
**Heo, Minseok*****
**Kim, Jisun*******

This article explores the methodology of collaborative digital text compilation for modern Korean documents by focusing on articles published in the first through 14th issues of the Korean humanities journal *Chin-Tan Hakpo*. For more precise data processing, we designed XML schema of *Chin-Tan Hakpo* considering the philological characteristics and the possibility of utilizing the data. Based on this schema, we constructed a text input framework using MediaWiki and input both original and modernized text data for 74 out of 85 articles published in *Chin-Tan Hakpo*. The text data input process revealed several challenges requiring further attention, including the need for XML schema refinement, establishment of standardized rules for modern text conversion, and

* Associate Professor, Department of Korean Language and Literature, Kyung Hee University; Visiting Researcher, Institute of Humanities, Seoul National University (First Author)

** Ph.D. Candidate, Department of Korean History, Seoul National University (Co-author)

*** Ph.D. Candidate, Department of Korean Language and Literature, Seoul National University (Co-author)

**** M.A. Candidate, Department of Korean History, Seoul National University (Co-author)

***** Lecturer, College of Liberal Arts, Korea University (Corresponding Author)

www.kci.go.kr

improvement of input tool efficiency. We hope that the experiences and lessons learned from this *Chin-Tan Hakpo* text compilation project will contribute to finding digital text compilation methodologies suitable for modern Korean documents through future collaborative efforts among researchers.

**Keywords** *Chin-Tan Hakpo*, Modern Korean Documents, Designing XML Schema, Publishing Digital Text, Semantic Data Archive

www.kci.go.kr