東洋學 第84輯(2021年 7月) 檀國大學校 東洋學硏究院

한국과 일본의 漢字 데이터셋(Dataset) 현황과 과제

김우정* · 박영미**

| 국문초록 |

본고는 한국과 일본의 한자 데이터셋 구축 현황을 살펴보고 그 개선방향을 제안한 것이다. 한자 데이터셋은 한자학과 한문학을 포함한 동아시아 인문고전학 연구 및 활용의 기초정보로 매우 중요하다. 하지만 데이터셋 구축방법의 일관성 결여, 관리체계의 부실, 데이터셋 간의 연계 및 호환성 부족 등으로 인해 활용가치가 크게 떨어짐을 확인하였다. 본고에서는 이를 해결하기 위한 방안으로 한자 속성정보 규정을 위한 전문위원회와 종합관리시스템을 구축하는 한편, 데이터셋 표준화 및 품질 제고를 위한 노력을 지속적으로 해야 할 것을 제안하였다.

[주제어] 한자, 한자 속성정보, 데이터, 데이터셋

| 목 차 |

Ⅰ. 들어가는 말

한자는 동아시아 지식정보의 기초다. 따라서 한자 정보를 디지털 형태로 정리하는 것, 곧 한자 데이터셋[1]을 구축하는 것은 동아시아 인문학 연구와 활용의 토대를 닦는 것이다. 한자 데이터셋은 좁게는 대표자형 선

* 제1저자, 단국대학교 한문교육과 교수 / rtoran@dankook.ac.kr
** 공동저자, 단국대학교 동양학연구원 연구전담조교수 / pdoma@hanmail.net

1) 데이터셋은 '관련 자료의 집합체'라는 점에서는 데이터베이스와 유사한 개념이다. 하지만 데이터베이스는 데이터의 저장은 물론 통합, 공용, 운영의 기능도 가지고 있는 반면 데이터셋은 특정 데이터들의 뭉치(integrated data)운용 '공유의 목적 아래 통합하여 관리되는 데이터들의 집합'을 말한다면, 데이터셋은 '유사하거나 유관한 자료를 컴퓨터에서 사용할 수 있도록 모은 뭉치'라는 점에서 구별된다. 따라서 하나의 파일도 데이터셋이 될 수 있으며, 데이터베이스에 제공되는 데이터의 일부가 될 수도 있다.

정, 사용빈도 조사, 共起 관계 파악, 시기별 지역별 한자 사용 양상 연구, 고문헌 자료 등 출판·인쇄 문화 연구 등의 기초자료로 활용할 수 있을 뿐만 아니라 최근 관심이 고조되고 있는 데이터 마이닝, 텍스트 마이닝, 코퍼스 구축, 지리정보시스템(GIS), 시각화 서술, 토픽 모델링, 맵핑(mapping), 비주얼라제이션(visualization), 의미망(semantic web), 네트워크 분석 등 디지털 기술과도 관계되어 있다. 또한 인간의 경험과 직관에 의존해온 자전·사전의 수준을 고도화·대량화하는 데도 긴요하며, AI 기술과 접목하여 한자 폰트 제작, 문자인식, 기계번역 등에도 없어서는 안 될 원천자료다.

하지만 가공되지 않은, 바꿔 말해 정형화되지 않은 한자 데이터는 정보로서의 가치가 없다. 정보로서의 가치를 지니도록 하기 위한 조작 또는 처리, 즉 데이터 가공이 필요한 이유이다. 데이터 가공이란 그림, 사진, 오디오 파일, 문서 파일 등의 정형화되지 않은 데이터의 특징을 추출, 구조화하여 디지털화한 정형 데이터로 바꾸는 것을 말한다. 그러나 정형화 과정을 거친 한자 데이터들의 상당수가 디지털 데이터로서의 활용 가치가 적다는 문제점을 보이고 있다. 이 글은 이와 같은 문제점을 살펴보고 그 개선책을 제안하고자 하는 목적을 지니는 바, 같은 한자문화권에 속한 국가이지만 우리와는 다른 방향에서 한자 데이터셋을 구축해온 일본의 사례도 함께 참고하고자 한다. 중화권 국가의 경우, 2000년에 시범서비스를 시작한 '異體字字典'(대만 교육부, https://dict.variants.moe.edu.tw/variants/rbt/home.do)을 비롯해 '中華經典古籍庫'에 포함된 '字符查詢'(중국 中華書局, 2020년 서비스 개시, https://mp.weixin.qq.com/s/SXwERVzvhw95DkmeT6v8Vw), '國學大師'(중국, http://www.guoxuemi.com), 하버드대학교 동아시아언어·문화학과 대학원생 Donald Sturgeon이 개발한 Chinese Text Project(중문명 中國哲學書電子化計劃, https://ctext.org/) 등 대량의 한자 데이터셋을 포함하고 있는 데이터베이스들이 있다. 하지만 이에 포함된 데이터셋의 구조, 프로세스, 속성정보 등이 공개되어 있지 않으며, 데이터의 수집·정리 과정이 불분명한 경우도 있다. 따라서 중화권 국가의 한자 데이터셋에 대해서는 관련 자료를 확보한 이후 별도의 논고를 통해 다루기로 한다.[2)]

Ⅱ. 한국의 한자·한문 자료 데이터 가공

한자는 시대와 지역, 書寫 습관 등에 따라 그 字形과 字體를 달리 하는 경우가 많다. 우리나라의 경우, 중국에서 건너온 서적을 復刻하거나 筆寫하는 과정에서 새로운 자형이나 자체가 발생하기도 했고, 자체적으로 새로운 자형을 만들어 사용하기도 했다. 또 復刻하거나 筆寫한 경우에도 당시 유행하는 자형이나 자체를 사용한 것, 復刻者나 筆寫者가 선호하는 자형을 사용한 것, 복각자나 필사자의 오류로 인해 誤刻하거나 誤寫한 것 등이 혼재되어 왔으며, 誤刻·誤寫字도 시간이 지남에 따라 正字나 異體字로 취급되기도 하였다.

2) 祝國忠·周亞民(2002), 李国英·周晓文(2009), 周亞民(2012; 2013; 2017), 王平(2014a; 2014b) 등에 한자 데이터셋에 관한 논의가 일부 포함되어 있으나, 주로 자료의 정제와 보완, 데이터베이스 구축의 필요성과 활용 가치 등에 초점을 맞추고 있을 뿐, 데이터셋의 구조나 특성 등에 대해서는 관심을 기울이지 않고 있다.

이처럼 다양한 이체자형의 존재로 인한 문자생활의 혼란과 불편은 동아시아 국가에게 공통된 문제였지만 한자를 버리고는 문자생활 자체가 불가능했던 중국에 있어 가장 심각한 것이었다. 따라서 字書나 韻書의 편찬사를 통해서도 알 수 있듯이 중국에서 먼저 한자에 관한 표준안 내지 규범안이 마련되면 한국과 일본이 이를 뒤따라 반영하는 과정이 반복적으로 재현되곤 하여왔다. 그러나 20세기 이후 상황이 급변하여 한국과 중국이 곤궁한 상황에 빠져있는 동안 일본에서는 『大漢和辭典』과 같은 기념비적 성과가 선을 보였다. 그러나 중국은 오랜 國共 內戰을 끝낸 뒤 簡化字 제정 등 문맹률을 낮추는 것이 당면 과제였던 때였으므로 한자를 체계적으로 수집・정리하는 데 관심을 둘 여력이 없었다. 한국의 경우, 광복 이후 한글전용론이 팽배하며 한자는 청산해야 할 과거의 유물로 취급되었다. 그 결과 한자 정보의 수집과 정리 역시 한동안 방치되었으며, 1990년대 후반 일련의 전산화 사업이 추진되면서 비로소 디지털화된 한자 데이터셋을 가지게 되었다. 즉, '한국역사정보통합시스템'에 포함되어 있는 '유니코드한자 검색시스템'과 '한국고전종합 데이터베이스'의 '이체자 정보', '한국학 디지털 아카이브'의 '한자자형전거', '한국학자료센터'의 '이두용례사전' 등이 그것이며, 네이버 디지털한자사전에 포함된 '한국한자어사전'(단국대 동양학연구원 제공) 등도 한자 데이터셋을 기반으로 한 것이다. 본고에서는 이 가운데 주요 한자 데이터셋에 대해 검토해보기로 한다.

1. '한국고전종합 데이터베이스'의 '이체자 정보'(http://db.itkc.or.kr/dch/)

민족문화추진회(현 한국고전번역원)에서는 1997년에 '한국문집총간'의 전산화사업 계획이 수립된 뒤 2020년까지 1,259종의 한국 문집 데이터베이스를 구축하였다. 2001년부터 서비스가 시작된 이 데이터베이스에는 9억여 자의 텍스트와 78만 면의 이미지, 500만 건의 메타데이터 등 방대한 규모의 한문고전문헌 자료가 담겨 있다. 이전까지 국학 원전자료는 대부분 데이터베이스로 구축된 바가 없으며, 구축되었더라도 이미지 프로세싱(image-processing) 기법으로 처리되어 있어 影印本과 다를 바 없었다. 이런 점에서 볼 때, '한국고전종합 데이터베이스'는 원전자료를 텍스트 파일로 구축함으로써 자료의 체계적 관리와 검색이 가능케 했다는 점에서 획기적인 것이었다. 다만 비정형 자료를 전자책화하거나 표점・교감 등을 더한 반정형 데이터 형태로 가공함으로써 원천 데이터만을 이용하고자 할 경우에는 재처리 과정을 거쳐야 하는 불편함도 있었다.

'이체자 정보'는 '한국고전종합 데이터베이스'를 구축하는 과정에 수집된 유니코드 확장한자 A영역(Extension A)을 비롯한 다양한 이체자 정보를 정리하고, 검색까지 가능하게 한 데이터셋이다. '지식정보자원관리사업'의 일환으로 2006년에 구축되었지만 실제로는 한국문집총간 전산화 과정에서 입력과 교정의 효율성을 높이기 위한 한자입력 방법을 모색하는 과정에서 비롯된 것이다. 처음에는 5백여 종의 고빈도 이체자만 정리하였다가 최종적으로는 4,850종 14,050자에 이르렀으며, 출전정보가 확인된 이체・이형자 3,000건은 출전 이미지를 추출하여 판종과 간행시기 등과 함께 제공하였다. 또한 '한국역사정보통합시스템'의 '유니코드한자 검색시스템'이 구축된 뒤, 이와 연계하여 '이체자 보기'를 클릭할 경우 '한국고전종합 데이터베이스'의 '이체자 정보'에 접근할 수 있도록 하였지만, 사용자가 '이체자 정보' 데이터셋 자체를 보거나 이용할 수는 없게 되어 있다.

한편 '이체자 정보'와 '유니코드한자 검색시스템'은 비슷한 시기에 시작되었으면서도 메타데이터(속성정보)

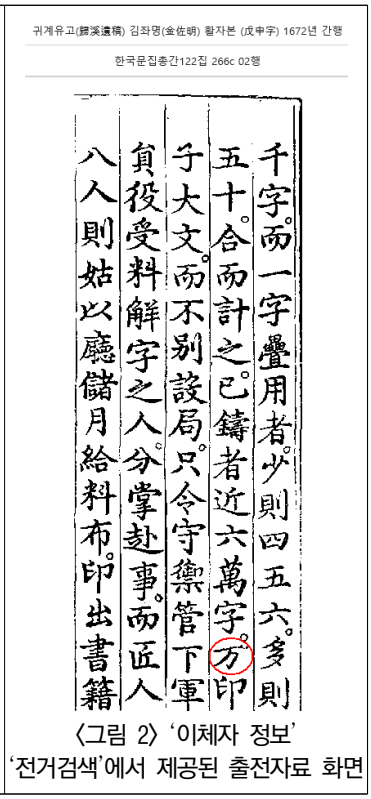
를 각자 구성한 결과, 상이한 필드(field)가 적지 않다. 메타데이터는 대량의 정보 속에서 찾고자 하는 정보를 효율적으로 찾아내기 위해 일정한 규칙에 따라 콘텐츠에 부여하는 정보로서, 콘텐츠의 위치와 내용, 작성자에 관한 정보, 권리 조건, 이용 조건, 이용 내력 등을 HTML 태그 등의 방법으로 기록한다. 따라서 데이터셋의 활용가치와 직결된다 하겠는데, '이체자 정보'와 '유니코드한자 검색시스템'은 유사한 정보를 다루고 있음에도 메타데이터의 필드 통합 작업은 진행하지 않아 효용성이 떨어지게 되었다. 즉, '이체자 정보'는 〈그림 1〉과 같이 部首, 유니코드, 四角號碼, 破字까지 17가지 속성정보만 제공되는 반면 '유니코드한자 검색시스템'은--구체적인 내용은 후술하겠지만--기관명, 사업연도, 위치 정보, 『康熙字典』 위치정보, 검정내역 등의 출전정보가 아울러 제공되고 있으며, Open API 호출규칙도 '유니코드한자 검색시스템'과 다르다. 데이터셋의 중장기적 운용 계획이 부재했기에, 비슷한 시기에 개발되었고 양쪽 모두에 관여한 전문가도 있었음에도 사실상 별개의 데이터로 구축되었다.


Home > 검색결과(코드/문자) > 상세보기

Unicode BMP

자형	万		
부수	一	유니코드	U+4E07
획수	2	총획수	3
KS코드	D8B2	JIS코드	969C
GB코드	4D72	BIG5코드	C945
한국뜻	1. 만;2.매우많은;3대단히.매우.전혀.절대로;4.만무;5.사천성에 있는 현 이름;6.성		
한글음가	만	한글로마자	MAN,MWUK
영어뜻	ten thousand; innumerable		
한어병음	MO4,WAN4		
일본어 훈독	YOROZU	일본어 음독	BAN,MAN
사각호마	1022,1270	파자	
비고			

〈그림 1〉 '한국고전종합 데이터베이스' 내
'이체자 정보' 검색결과 화면



〈그림 2〉 '이체자 정보'
'전거검색'에서 제공된 출전자료 화면

2. '한국역사정보통합시스템'의 '유니코드한자 검색시스템'(http://www.koreanhistory.or.kr/newchar/)

'유니코드한자 검색시스템'의 경우, 상당히 방대한 양의 문헌 및 시각 자료를 디지털화하여 제공하고 있는데, 23개 항목 정보를 갖춰 제출된 신출한자에 임시로 통합코드를 부여하고, 비트맵 이미지 형태로 신출한자 검색 결과를 제공하고 있다. 신출한자가 최종 등록되면 참여 기관은 최종 확정된 결과를 각 기관의 각종 데이터베이스에 반영하는 구조로 되어있으며, 기관에 따라 데이터베이스 문서 안에 비트맵 이미지를 삽입하거나 파자하여 표기하기도 한다. 그러나 데이터 구축과정에서 발생한 오류가 수정되지 않은 채 남아있기도 하고 수습되지 않은 자료도 적지 않다.

〈표 1〉 '유니코드한자 검색시스템'에 제공된 영역별 한자 목록

구분		영역 이름	코드 범위	문자 수	비고
표준	BMP 영역	韓中日統合漢字 (CJK Unified Ideographs)	0x4E00~0x9FA5	20,902자	UNICODE 2.0
		韓中日統合漢字 擴張A (CJK Unified Ideographs Extension A)	0x3400~0x4DB5	6,582자	UNICODE 3.0
	Ext.B 영역	韓中日統合漢字 擴張B (CJK Unified Ideographs Extension B)	0x20000~0x2A6D6	42,711자	UNICODE 3.1
비표준		新出漢字	KC00001~KC07355	7,360자	2004~2008
		新出符號	KS00001~KS00300	300자	2004~2008
		古漢字	OH00001~OH00190	190자	2005~2010

〈표 1〉은 '유니코드한자 검색시스템'에는 2015년 11월 현재 제공되는 한자의 영역과 수량이다. 이에 따르면 신출한자는 KC00001부터 KC07355까지 총 7,360자가 수록되어 있다고 밝혀져 있다. 하지만 실제로는 이보다 좀 더 많은 한자의 검색이 가능하며,[3] 코드에 해당하는 한자의 검색 결과가 없는 경우(예를 들어 KS00300)도 있어, 정보관리가 부정확하게 관리되고 있음을 알 수 있다.

정보통신부에서 2000년부터 추진해 온 '지식정보자원관리사업'은 현대적인 의미에서 한국 고전적 정보화의 본격적 출발이라 할 수 있다. 정보화는 필수적으로 지원하는 문자코드의 양과 종류에 따라 그 성패가 갈린다. 古典籍을 기준으로 볼 때, 모든 문자를 입출력할 수 있는지, 곧 기술적인 문제로 인한 변형 없이 원본 그대로를 입출력할 수 있는가 하는 점이 관건이다.

'유니코드한자 검색시스템' 구축 당시 대부분의 웹브라우저를 통해 지원될 수 있는 한자코드는 유니코드 Ext.A 영역까지였다. 따라서 원형 그대로 입력하기 위해 유니코드 Ext.B 영역의 한자를 사용한다 하더라도

3) 한국문집총간에서 수집한 한자(KC07356~KC08476) 외에 조선왕조실록(KC09900~KC09936), 『승정원일기』(KC10000~KC11315), 『명실록』(KC11400~KC13332), 한국사 목간자료(KC13333~KC13355) 등에서 수집한 한자들도 검색이 가능한데, '도움말'에는 이에 관한 내용이 반영되어 있지 않다.

사용자는 웹브라우저나 기타의 응용 어플리케이션 등을 통한 지원 없이는 유니코드 Ext.A 영역까지의 한자만을 사용할 수 있어, 同字로 치환하거나 이미지, 파자 등의 방법으로 처리해야 했다. 그런데 2015년 윈도우10이 출현한 뒤로 이러한 기술적 제약이 사라져, Ext.E 영역까지 거의 모든 유니코드 영역의 한자 입출력이 가능하게 되었다.[4] 따라서 대표자(同字)[5]로 치환하거나 이미지로 처리했던 한자도 유니코드를 이용하여 처리할 수 있게 되었으나, '유니코드한자 검색시스템'은 지금까지도 초기에 구축된 데이터를 수정 · 보완하지 않고 그대로 유지하고 있다.

한편 유니코드를 이용해 처리했다 할지라도 그 속성정보까지 함께 구축해야 활용 가치가 높아지는데, 이에 대해서도 관심을 기울이지 않고 있는 점도 문제다. 유니코드 확장한자 영역이 확대되면 될수록 해결해야 할 한자의 속성정보도 계속 늘어나게 될 것이므로, 이에 관한 속성정보를 즉각적으로 갖출 수 있도록 운영체제를 시급히 재정비해야 할 것이다.

한자의 속성정보에는 출전정보 외에도 形 · 音 · 義와 같은 기본정보 외에도 총획, 부수, 나머지 획수, 部件(構件), 부건의 조합 방식, 五筆字型, 四角號碼 등의 자형정보가 포함된다. '유니코드한자 검색시스템'의 경우, 다음과 같이 23개의 속성정보 필드를 갖추고 있다.

〈표 2〉 신출한자 데이터베이스 필드 내용

연번	필드명	필드 내용	예	비고
1	임시코드	1차 입력 단계에서 부여하는 각 기관별 신출한자 임시코드	A0089	필수
2	기관명	사업 기관 명칭	民推	필수
3	사업년도	사업 수행 연도	2004	필수
4	DB명	신출한자 출전 DB 명칭	韓國文集叢刊	필수
5	出典名	신출한자 출전 명칭	三淵集	필수
6	出典 묶음명	신출한자의 출전 묶음	kc_mm_a439	선택
7	위치정보	신출한자 출전의 위치 정보(면수, 행수)	mm_a167_231a	필수
8	파일명	신출한자가 입력된 파일명	kc_mm_a439_bv028	필수
9	전후 문맥	신출한자 전후 10여자의 원문 텍스트	白雲僧及海眼前導。緣東崖而上。越一嶄▼(山/絶)。至金	필수
10	원문 이미지	원문 이미지 파일명		필수
11	字形	원전 이미지 자형(60×60 내외)		필수

4) 윈도우10의 버전 정보는 https://technet.microsoft.com/en-us/windows/release-info.aspx를 참조.

5) 우리나라는 한자 자형을 공식적으로 표준화한 바가 없기 때문에 구성 성분이 다르거나(異構), 서사 방식이 다른(異寫) 한자 가운데 무엇을 표준자 또는 대표자로 할 것인지도 판단할 수 없다. 혼란을 피하기 위해서라도 자형 표준화에 관한 논의가 시작되어야 한다고 생각한다.

연번	필드명	필드 내용	예	비고
12	破字	자형의 조합 형태 표시	山/絶	필수
13	부수번호	부수 한자의 10진수 일련번호	046	필수
14	部首	부수 한자	山	필수
15	劃數	부수를 제외한 잔여획수	12	필수
16	總劃	부수를 포함한 전체획수	15	필수
17	字音	추정되는 자음	절	선택
18	字義	추정되는 자의	가파르다	선택
19	備考	인명, 지명 등의 부가 정보		선택
20	통합코드	신출한자 일련번호	KC01033	필수
21	檢定內譯	신출한자 판정 내역		필수
22	『康熙字典』 위치정보	『康熙字典』의 해당 면수	320.131	선택
23	五筆劃	한자의 첫획 모양	2	선택

그런데 국내에서는 한자학 방면의 연구가 워낙 취약한 탓에 자형, 자음, 자의 등의 기본정보조차 제대로 파악하지 못한 사례가 현저히 많다. 字音을 예로 들자면, 80,388자의 유니코드 등록한자 중 음가정보가 없거나 재검토해야 할 한자가 35,000여 자나 되는 것으로 조사된 바 있다.[6] 실제 우리가 가장 많이 사용하는 소프트웨어인 '흔글'의 경우 입출력이 가능한 한자음은 27,469자인데, 이 가운데 Ext.B 이하 한자의 자음은 전무하다(CJK영역 20,891개, Ext.A영역 6,578개). 이미 2004년에 Ext.B 영역 중 17,211자의 자음이 결정되었음에도 '흔글' 소프트웨어에는 아직까지 반영되지 않고 있는 것이다.

〈표 3〉 Unihan 소재 각국 음가 정보

구분	CJK영역	Ext.A	Ext.B	종합
kCantonese(광동어음)	13,882	5,608	453	19,943
kMandarin(북경어음)	20,204	5,050	146	25,400
kJapaneseKun(일본어훈독)	11,250	2	17	11,269
kJapaneseOn(일본어음독)	13,131	2	20	13,153
kKorean(한국어음)	8,762	0	0	8,762
kTang(唐音)	3,781	18	9	3,808
kVietnamese(베트남음)	3,936	125	4,236	8,297

6) 배은한 외(2016).

‘훈글’의 경우도 이러니 다른 응용 소프트웨어야 불문가지다. 이는 유니코드에서 지원하는 한자에 대한 개별 정보를 수집한 Unihan 데이터베이스 중 음가 관련 칼럼에 포함된 각국 음가정보의 데이터 수를 비교해보기만 하여도 알 수 있다.

‘훈글’의 경우 27,469자의 자음 정보가 있는 반면 Unihan 데이터베이스에 포함된 한국어음에 해당하는 칼럼에는 모두 8,762개의 자음만이 기록되어 있어 唐音과 베트남음에 이어 세 번째로 적다. 베트남의 경우, 우리와 큰 차이가 나지 않는데다가 Ext.B 영역도 포함되어 있어 더욱 대조적이다.

한자 속성정보의 문제는 현재 단국대학교 동양학연구원에서 편찬 중인 ‘한국고유한자자전’(가칭)과 한문교육연구소에서 수행중인 ‘한국 역대 한자자형정보 데이터베이스 구축사업’의 공통적인 애로사항이기도 하다. 학계의 중지를 모아 Ext.C 이하 한자 및 유니코드 미등록 한자들에 대한 속성정보를 조속히 확정하여야 할 것이다.

3. ‘한국학 디지털 아카이브’의 ‘한자자형전거’(http://yoksa.aks.ac.kr/jsp/hh/Directory.jsp?gb=1)

‘한자자형전거’ 역시 ‘이체자 정보’, ‘유니코드한자 검색시스템’과 마찬가지로 ‘지식정보자원관리사업’의 일환으로 개발된 한자검색 도구다. 조선시대에 간행된 『全韻玉篇』·『訓蒙字會』·『類合』·『新增類合』과 20세기 초에 간행된 『字典釋要』·『國漢文新玉篇』·『新字典』 등의 자전 정보를 모으고, 『字彙』·『康熙字典』·『千字文』류·『隸書集成』(손환일 편찬) 등에서 正字·俗字·辨似字[7] 등을 조사하여 정리하였다. 즉, 자전에 수록된 자형과 그 음·뜻을 비교해서 살펴볼 수 있도록 하는데 목적을 둔 것인데, 자형을 유니코드로만 보여줄 뿐 출전 이미지로는 보여주지 않고 있는 점이 한계다. 또한 검색 기능은 제공되고 있으나 데이터셋은 공개되지 않고 있으며, ‘해제’가 삭제되어 있어 사용자들이 어떤 자료를 어떤 방법으로 구축하였는지 확인할 방법이 없다.

2000년대 들어 ‘한국역사정보통합시스템’ 구축사업과 ‘한국학 디지털 아카이브’ 구축사업 등이 진행되었는데, 이를 통해 방대한 양의 한자 관련 데이터들을 집적할 수 있게 되었다. 또한 사립대학인 단국대학교에서 한국 고유한자와 한자어를 모은 『한국한자어사전』과 중·일의 대사전에 필적하는 규모의 대사전인 『한한대사전』을 출간하였으며, 아직 몇 가지 해결해야 할 사항이 남아 있기는 하지만 두 사전을 통합하여 디지털화한 자료도 존재한다. 따라서 한자 데이터셋 자체는 어느 정도 갖추어졌다고 할 수 있겠으나, 자료의 집적과 검색에 초점이 맞춰져 있고 오픈소스로 공개되지 않아 활용 가치가 떨어져 활용 가치가 떨어진다. 제반 연구의 기초자료로 삼겠다는 뚜렷한 목적의식과 충분한 사전 준비, 타당한 방향 설정, 사업 이후의 활용방안 등에 관해 충분히 숙고하지 못한 상태에서 사업에 착수하고, 종료 이후 관리도 이루어지지 않은 것이다. 대량의 자료를 오류 없이 디지털텍스트로 구축할 수 있다면 최선이겠으나 그것이 어렵다면 가치가 높은 전적류

7) ‘한자자형전거’에서는 ‘음과 뜻이 현저히 구별되나 형태는 매우 유사하여 혼용될 수 있는 한자’를 ‘辨似字’라 하였다.

(서지학적으로 의미 있는 목판이나 활자본, 희귀본 등)부터 선별하여 단계적으로 가공하는 것도 고려했어야 했다.

Ⅲ. 일본의 한자·한문 자료 데이터 가공

메이지 유신 이후, 일본에서는 근대화와 근대적 지식의 습득을 위해 한자를 폐지하자는 운동이 벌어졌으나 다른 한편으로는 한학의 붐이 불며, 다양한 한문 서적과 자전 등이 연이어 출판되었다. 이에, 근대적인 교육의 효율적인 전개를 위해 정부 당국은 교육 및 일반이 사용할 제한된 양의 '한자'와 표준 한자 자체를 고시하여 당대의 요구와 관습을 반영하고자 하였으며, 이러한 노력은 현재에도 계속되고 있다. 따라서 일본의 한자와 한자 자체의 개정 및 이에 수반하여 생성된 자료들은 이미 언어사적 의미를 지니고 있다고 할 수 있을 것이다.

전자화 이전의 자료는 일본의 근대적 한문 정책과 한문 문화의 산물이다. 그 정점은 1917년 시작하여 1943년 1권이 간행되고 2000년에 보충판까지 완성된『大漢和辭典』이라 할 것인데,『康熙字典』에서 시작되었던 이 사전은 지금도 일본 한자자체 데이터셋의 속성정보 처리에 활용되고 있다. 2000년 이후 일본의 한자 데이터셋 관련 현황은 다음과 같다.

1. '한자정보 데이터베이스'(漢字情報 データベース)와 '범용전자정보교환환경정비프로그램'(汎用電子情報交換環境整備プログラム)

일본은 2002년 전자정부를 목표로 'e・Japan重点計画―2002'를 발표했다. 이 계획에는 문자정보 및 코드 정비에 관한 내용이 들어 있었는데, 기존의 문자코드 규격(JIS X0208, JIS X0212)으로 처리하기 어려웠던 인명, 지명, 법인명 등을 해결하기 위한 것이었다. 이를 해결하고자 발족한 것이 '범용전자정보교환환경정비프로그램'으로, 대상은 총무성의 주민기본대장 네트워크 통일문자 21,000자, 법무성의 호적통일문자 55,000자, 법무성의 등기통일문자 67,000자로, 同字와 別字를 판정하고 각 글자의 정보, JIS X 0213(國內規格)와 ISOAEC 10646(國際規格)의 문자코드 정보,『大漢和辭典』문자번호 등의 속성정보를 부여한 '한자정보데이터베이스(漢字情報データベース)'를 축적하였다.[8)]

2. '한자자체규범사 데이터베이스'(漢字字體規範史 データベース, Hanzi Normative Glyphs : 약칭 HNG)

이시츠카 하루미치(石塚晴通)가 각 시대, 각 지역의 한자 자체의 표준과 그 표준의 시대별, 지역별 변천을

8) http://www.hng-data.org/ (검색일 : 2021년 6월 5일)

밝히기 위한 목적으로 20여년에 걸쳐 79종의 문헌에서 뽑은 50만 개의 용례를 정리한 『漢字字體資料』에 뿌리를 둔 데이터베이스다.[9)]

2000년 무렵 홋카이도대학 언어정보학강좌의 공동 작업을 통해 石塚의 『한자자체자료』를 데이터베이스화하여 온라인으로 공개하였는데, 종이카드 이미지와 텍스트로 처리하였으며, 텍스트화 작업에는 JIS漢字, UCS漢字, 大字典番號, 大漢和辭典番號를 사용했다.

이후 2004년 동경외국어대학에서 '한자자체규범 데이터베이스'로 명칭을 바꿔 구축되었는데, 이 데이터베이스는 '시대마다 지역마다 한자 자체의 표준가 존재하였으며', '시대마다 지역마다 존재하는 한자 자체의 표준은 변하였고', '한자 자체의 표준은 문헌의 성격, 속성과 관계가 있다'는 원칙에 따라 검색결과를 보여주는데 초점을 맞추었다.

'한자자체규범 데이터베이스'는 1단에 중국사본, 2단에 중국판본, 3단에는 일본사본・판본, 4단에는 한국사본・판본, 5단에는 중국 주변 사본・판본을 배치하였는데, 각각은 연대순으로 되어 있으며 6단은 字書의 용례를 보여주고 있다. '한자 자체의 표준은 문헌의 성격, 속성과 관계가 있음'을 보여주기 위해 異體率이 높은 비표준문헌과 이체율이 낮은 표준문헌과의 차이를 좌우에 배치하여 표현하였다.[10)]

'한자자체규범 데이터베이스'는 다시 '漢字字體規範史 데이터베이스'로 이름을 바꿔 지금에 이르고 있다. 이 데이터베이스는, 돈황문을 비롯한 唐代 이전의 중국 古寫本과 奈良, 平安시대 일본의 古寫本을 비교 연구하는데 유용하다. 初唐의 표준자체가 일본의 표준자체로 이입, 정착된 양상을 정치하게 기술할 수 있는 정보를 제공하고 있으며, 開城石經의 字體가 宋版을 통해 수용되었으며 이로 인해 초당의 표준자체와는 다른 자체가 새로운 규범자체로 정착되었음을 밝혀내는데 크게 기여하였다.

'漢字字體規範史 데이터셋'은 '한자자체규범사 데이터베이스'에 공개된 한자 자형을 원본에서 오려낸 것과 메타데이터를 더한 것이다. 데이터베이스는 지속적으로 공개하는 것이 어렵기에 안정적인 데이터 공개를 위해 데이터셋 자체를 공개하는 것이 효과적이라고 판단하여 2018년 데이터셋으로 공개하였다.[11)]

3. '탁본문자 데이터베이스'(拓本文字データベー, http://kanji.zinbun.kyoto-u.ac.jp/db-machine/imgsrv/takuhon/)

'탁본문자 데이터베이스'는 2004년 교토대학 인문과학연구소에 소장된 漢代부터 中華民國 초기까지의 탁본자료를 대상으로 구축한 방대한 양의 文字畫像 데이터베이스다. 데이터베이스의 구성 요건은 '탁본화상의 데이터베이스', '釋文의 전문 데이터베이스', '漢~淸에 이르는 문자 데이터베이스'로, 이를 위해 ttext-kanbun이라는 문자적출 도구를 개발하였다. 기본적으로 검색을 목적으로 하며, 검색 결과를 시대순으로 한 글자씩 이미지로 보여준다.[12)]

9) 박영미・하지영(2019).
10) 高田智和(2013).
11) http://www.hng-data.org/
12) http://kanji.zinbun.kyoto-u.ac.jp/db-machine/imgsrv/takuhon

4. 木簡庫(http://mokkanko.nabunken.go.jp/ja/)

출토문자자료 전반의 연구 거점인 奈良文化財研究所에서 조사・정리한 목간 자료(일본 전체의 약 70%인 25만점)를 기초로 구축된 데이터베이스다. 목간의 釋文을 가로로 써야 했기에 자료인 목간과 문자의 유기적 관계를 파악하기 어려웠는데, 이를 해결하기 위해 목간 釋讀 지원 시스템인 'Mokkanshop'을 개발하였고, 이를 이어 목간 문자이미지 데이터베이스인 '목간자전'을 만들어 2007년 공개하였다.

이외에 '목간 인명 데이터베이스'를 작성하여 2011년 공개하였으며, 출토지점정보를 작성하여 목간자전과 링크를 하였고, '일본고대연구문헌목록 데이터베이스'와 연계하여 연계검색 시스템을 구축하였다. 외부 데이터베이스와의 연계를 중시하여 나라문화재연구소와 동경대학사료편찬소 간의 연계검색도 가능하도록 설계되었다.[13] '목간고'는 이상의 목간데이터베이스와 목간의 문자 화상데이터베이스・목간자전을 통합하여 검색창을 하나로 만든 것이다.

5. '『大字典』 데이터베이스'(大字典データベース)

『大字典』은 1917년 上田萬年, 岡田正之, 飯島忠夫, 榮田猛猪, 飯田傳一 등의 국어학자와 한학자가 협동하여 편찬한 漢和辭典으로 약 18000자의 新字도 수록하였다. 이 사전은 국어사전과 전근대일본의 한자자전의 틈을 메우는 것을 목표로 하여 일본식 해석(和訓), 國字, 일본에서 통용되는 字體를 많이 채록하였다. 현재 한자자체규범사 데이터베이스의 데이터 셋화와 검색 서비스의 재구축, 새로운 기술을 사용한 시도등이 행해지고 있는데 한자자체규범사데이터베이스와 그 전신인 이시즈카 한자자체 자료는『大字典』에 기반하여 문자를 정리하였다.

2010년 '한자자체규범사 데이터베이스'의 백업데이터에 포함된『大字典』의 데이터베이스를 엑셀 파일 daijiten_DB_20061008.xls을 tab-separated values(TSV) 형식으로 변환한 것을 기초로 하였으며, 字種コード,『大字典』번호(『大字典』의 문자 번호), 部首番號(『大字典』의 부수 번호), 文字, x0208(區点番號), 包攝・備考, UCS,『大漢和辭典』番號로 이루어졌다. 여기에서 '文字'는 'JIS 漢字'에 해당하고, '包攝・備考'는 '문자 포섭에 관한 정보' 즉, JIS 포섭기준 적용에 따른 連番 포섭, 호환 포섭 등을 기록한 것과 비고의 난을 둔 것이다. UCS는 JIS X 0221:2001으로 BMP 범위 내의 것만을 기록하였다.[14]

또한『大字典』데이터셋을 CHISE(CHaracter Information Service Environment) 문자 온톨로지와 통합하여 CHISE의 웹서비스 '이체자규범사 데이터베이스' 單字 검색에서『大字典』데이터셋 정보를 이용할 수 있게 하였으며,『大字典』의 전문 화상도 연계하여 제공하고 있어,[15] 향후 단국대학교에서 개발하고 있는 '통합디지털한한대사전'의 웹서비스가 실행될 경우 참고로 삼을만하다.

---

13) 渡辺晃宏 외(2012).
14) 守岡知彦(2019).
15) CHISE IDS 漢字検索(https://www.chise.org/ids-find) 참고.

6. '헤이안시대 한자자서 종합데이터베이스'(平安時代漢字字書総合データベース)[16]

이 데이터베이스는 한자자서종합 데이터베이스의 일환으로 기획되었으며, 다음과 같은 중국과 일본의 字書를 기초로 구축되었다.

① 中國의 字書 : 『玉篇』(梁・顧野王 撰, 543년, 現存2,087자, 原本玉篇), 『大廣益會会玉篇』(宋・陳彭年 等撰, 1013년, 약22,800자, 宋本玉篇), 『龍龕手鏡』(遼・行均 撰, 997년, 약26,000자, 高麗版)
② 日本의 字書 : 『篆隷萬象名義』(空海 撰, 9세기 초, 약16,000자), 『新撰字鏡』(天治本, 昌住 撰, 10세기 초, 약 21,000자), 『類聚名義抄』(圖書寮本, 11세기 초, 약3,600항목), 『類聚名義抄』(観智院本, 12세기 후반, 약32,000항목)

'헤이안시대 한자자서 종합데이터베이스'는 크게 두 가지 단계를 거쳐 구축되었는데, 첫 번째 단계에서는 일단 표제자(掲出字) 이미지 데이터베이스를 구축한 다음, 표제자 이미지를 오려내고 유니코드 부호를 붙이는 작업을 하였고, 소재 정보에 대응하는 이미지 파일명을 부여하였다. 두 번째 단계에서는 전문 텍스트베이스를 구축하였는데, 중국 한자 자서는 音注, 義注, 字體注로 구성하였다. 음주는 대개가 反切로 표시되었고, 의주는 경전 본문과 그 뜻을 주석한 것이다. 자체주는 '正', '俗', '亦' 등을 注記하였다. 일본의 헤이안시대 한자 자서는 單字는 기본적으로 大字로 추출하였지만, 2자 이상의 숙어도 보인다. 주문은 음주, 의주, 자체주에 和訓이 더해졌다. 『類聚名義抄』의 경우, 聲點, 가타가나 傍訓, 오코토점(ヲコト点)이 있어 다른 것보다 더욱 복잡하다.[17]

'헤이안시대 한자자서 종합데이터베이스'는 언어사 연구, 한자 자체사 및 한자 자체 편찬사 연구에 활용되고 있으며 화훈과 자음 연구의 자원으로도 활용될 수 있도록 설계되었다. 특히 築島裕의 『訓点語彙集成』(2007~2009)과 연계하는 것을 향후 과제로 삼고 있다.

7. 인문학 오픈데이터 공동이용센터(人文学オープンデータ共同利用センター, CODH)의 '일본 고전적 쿠즈시지 데이터셋'(日本古典籍くずし字データセット)」

일본에서는 한자를 이용한 표기 수단으로 '쿠즈시지(崩し字)'가 오랜 기간 사용되었다. 쿠즈시지는 한자 및 가나(假名)를 해서가 아닌 초서나 행서의 서체로 쓴 것이다. 가나의 경우, 지금과 달리 하나의 음에 다수의 헨타이 가나(變體假名)로 쓰였다. 變體假名는 1900년 소학교령 시행 규칙이 반포되어 히라가나와 가타가나에 현재와 같이 하나의 자체를 정하기 이전에 사용되었던 것으로, 하나의 문종에 다양한 자체가 사용되었다. 예를 들면 安에 자원을 둔 あ가 현행의 것이지만, 安 외에 𛀀(阿), 𛀁(悪) 등이 병용되었다. 그러므로 쿠즈시지 데이터베이스는 초서와 행서를 중심으로 한 다양한 한자 자체 데이터베이스를 포함하고 있다.

16) 池田証壽(2014)를 바탕으로 정리한 내용이다.
17) https://hdic.jp/top/와 https://github.com/shikeda/HDIC에 이 프로젝트에 관한 자료가 공개되어 있다.

CODH는 정보학과 통계학 기술을 활용하여 인문학 연구를 행하는 것을 목적으로 설립된 연구기관으로, 연구자와 시민들이 인문학 데이터를 편리하게 이용할 수 있도록 오픈소스화를 추구하였다. 그 일환으로 2016년에 '일본 고전적 자형 데이터셋'이 공개되었는데, 飜刻의 부산물인 문자의 위치정보를 인간과 기계를 위한 데이터로 이용할 수 있도록 한 것이다. 기계용 학습데이터는 2017년 현재 86,176자를 갖추고 있으며, 출현 빈도가 낮은 것도 포함되어 있다. 좌표정보를 사용하였기에 개별 문자보다 좀 더 큰 단위에서 인식하는 것도 가능해졌는데, 다만 変體仮名의 자모를 구별하지는 않았다. 데이터의 종류는 다음과 같다.

① 원본 보정 이미지 데이터(原本補正画像データ) : 일본 고전적 데이터셋으로 공개될 화상에 대해 飜刻作業을 용이하게 하기 위한 전처리 과정이다. 나아가 두 면으로 되어있는 것은 페이지를 분할하거나 회전시키거나 하여 위치를 바로잡게 하는 처리까지 더해진 화상이다.
② 문자 좌표 데이터(文字座標データ) : 원본보정화상 데이터에 문자를 에두른 장방형의 좌표와 문자의 유니코드 코드 포인트, Block ID, 문자 ID를 기록한 것이다. 장방형의 좌표에는 문자영역의 XYWH를 정리하여 두었다. Block ID는 텍스트 영역마다 부여하는 ID이며, 문자 ID는 Block ID 내의 문자 출현순으로 부여한 ID이다.
③ 자형 이미지 데이터(字形画像データ) : 원본 보정화상 데이터에 문자 좌표 데이터를 적용하고 오려낸 화상으로, 문자종마다 자형을 쉽게 열람할 수 있도록 하였다. 데이터셋에 포함된 文種은 빈도순 문종 리스트 혹은 코드순 문종 리스트로 열람할 수 있다. 쿠즈시지 각각의 문자 형태의 차이는 물론 쿠즈시지의 원자와 자모의 상이함에 의한 이체자의 변종 등 실제 자형을 화상으로 확인하면서 쿠즈시지 학습에 이용할 수 있도록 되어있다.
④ 작업보고문서 : 작업에서 읽을 수 없었던 문자에 대한 정보, 혹은 그 밖의 주의사항을 기록한 것이다.[18]

8. 인문학 오픈데이터 공동이용센터(人文学オープンデータ共同利用センター, CODH)의 '기계학습용 KMNIST 데이터셋'

이 데이터셋 역시 CODH가 제작한 것으로, 기계학습 연구에 활용되는 'MNIST 데이터셋'과 호환할 수 있는 쿠즈시지 데이터셋이며 '일본 고전적 쿠즈시지 데이터셋'에서 파생된 데이터셋이다. 'MNIST 데이터셋'에 대응한 기계학습 소프트웨어가 있을 경우, 설정을 변경하는 것만으로도 'KMNIST'를 시험해볼 수 있으며, 목적에 따라 Kuzushiji-MNIST, Kuzushiji-49, Kuzushiji-Kanji의 3종류가 있다.

9. '전자 쿠즈시지 자전 데이터베이스'(電子くずし字字典データベース)

東京大學史料編纂所는 1984년부터 '歷史情報處理시스템'(Siryohensansho Historical Information Processing

18) 北本朝展(2017)를 바탕으로 정리한 내용이다.

System, 약칭 SHIPS)을 구축하기 시작했다. 역사사료편찬을 목적으로 설계된 SHIPS는 사료목록계DB, 사료풀텍스트계DB, 색인계DB, 편년계DB 등 다양한 DB를 작성했으며, 공통 폰트를 통해 종합적으로 유지・운용할 수 있도록 설계하였다.

2000년 개발에 착수하여 2001년 데이터를 추가하고 2006년 대외에 공개되었다. 대상은 동경대 사료편찬소의 일본사 사료, 나라시대부터 에도시대(17세기 전반)의 자료부터 착수하였다. 데이터베이스 설계 전에 '難讀 자형과 특수한 자형은 물론 가능한 모든 것을 망라'하며, '단문자만이 아니라 어휘도 채록'하고, '자형 이미지의 출전을 명시할 수 있는 것을 선정'하며, '연구자가 수시로 등록할 수 있도록' 하며, '비슷한 자형을 참조할 수 있는 기능을 갖추는' 것을 원칙으로 삼았는데, 데이터의 신뢰성과 활용가치를 제고하는 동시에 지속적인 수정・보완까지 사전에 고려하였음을 알 수 있다.

동경대 사료편찬소는 역사 사료 이미지를 디지털 아카이브로 배치하고 각 데이터베이스로부터 모듈을 통해 참고가 되도록 하였는데, '전자 쿠즈시지 데이터베이스'는 기존의 시스템을 통해 데이터를 축적하는 방식이다. 소장 사료목록DB 혹은 유니온카타로그DB에 엑세스하여 수집 대상 사료를 검색하여 이미지로 표시하면 전용 등록시스템이 작동하는데, 등록화면에 등록하고 싶은 자형을 지정하고, 자형화상에 대응하는 문자와 어구 정보를 순차적으로 입력하도록 되어 있다. 단문자 입력의 경우, 문자코드를 통해 음독과 훈독, 부수 등이 자동적으로 부여되며, 사료에 관한 메타정보, 예를 들면 史料名, 史料群名, 所藏處, 出典, 和曆 등도 소장사료목록DB 혹은 유니온카탈로그 DB에서 자동적으로 등록되도록 설계되었다. 그 외에 대상사료 특유의 정보와 공개범위를 정하는 세큐리티 코드 등은 필요에 따라 입력할 수 있게 하였다. 이처럼 입력환경이 변화함에 따라 문자코드도 SHIFT-JIS에서 UTF8로 변경하였다.[19]

이러한 작업을 바탕으로 2000년 '전자 쿠즈시지자전 데이터베이스' 개발에 착수하여 2006년에 공개하였다. 이 데이터는 일본사 사료 가운데 나라시대부터 에도시대 초기까지 102종류의 사료군을 대상으로 하였으며, 자형이미지데이터의 집적이 주된 목표다. 단문자만이 아니라 어휘도 채록 대상으로 하며 자형이미지의 출전을 명시하고, 유사자형을 참조할 수 있는 기능을 갖추도록 하였다.[20]

10. '木簡・쿠즈시지 解讀시스템'(MOJIZO : Image matching search for mokkan or cursive characters, https://mojizo.nabunken.go.jp/)

동경대 사료편찬소는 奈良文化財研究所와 연계하여 나라문화재연구소의 '목간이미지 데이터베이스・목간자전(木簡画像データベース・木簡字典)'과 동경대 사료편찬소의 '전자 쿠즈시지 자전데이터베이스'를 연계 검색할 수 있게 하였다.[21] 두 곳의 데이터를 합하면 약 17만 건의 이미지 데이터로 문자 자형을 표시할 수

19) 井上聡(2015).
20) 山田太造(2013).
21) http://clioapi.hi.u-tokyo.ac.jp/ships/ZClient/W34/z_srch.php

있으며, 고대 목간부터 에도시대의 고문서에 이르기까지 약 1000여 년에 걸친 문자 변천을 확인할 수 있다.

'MOJIZO'는 검색 대상의 화상을 해석하고 나라문화재연구소가 소장한 목간의 자형・자체와 동경대학 사료편찬소가 수집한 고문서, 고기록, 전적류의 자형・자체에서 유사한 문자화상을 표시한다. 읽을 수 없는 문자의 경우『大漢和辭典』어휘 데이터를 이용해 비슷한 문자를 찾아볼 수 있도록 설계되어 있는 점도 특기할 만하다.[22)]

이상의 자료를 연계하여 검색할 수 있는 시스템이 '역사문자 데이터베이스 연계시스템(史的文字データベース連携検索システム)'이다. 이를 통해 '木簡庫', '電子くずし字字典データベース', '國文研字形検索', '漢字字体規範史データセット単字検索' 등 일본 국내 한자 데이터베이스는 물론 대만 중앙연구원 歷史語言研究所 数位文化中心의 '簡牘字典 : 史語所藏居延漢簡資料庫'까지 연계하여 검색할 수 있다.[23)]

이상에서 살펴본 바와 같이 일본의 한자 데이터셋은 기획 단계에서부터 자국내 데이터 간의 호환・연계만이 아니라 대만과 연계가 가능하도록 하는데 유의하여 설계하였으며, 데이터 리소스까지 꼼꼼하게 정리하였음을 알 수 있었다. 또한『大漢和辭典』,『大字典』등의 旣刊 辭典 데이터와 연계된 難讀字 추정검색기능을 제공하거나 데이터셋을 오픈하여 시민, 연구자, 기업 등에 이르는 다양한 분야의 사람들이 이용할 수 있도록 하고 있었으며, AI 기계학습에 활용할 수 있도록 재가공된 데이터셋도 일부 갖추고 있음을 확인하였다.

Ⅳ. 한자 데이터셋 구축의 과제

앞서 언급한 바와 같이 한자 데이터셋은 동아시아 인문고전학의 기초자료로서도 중요하지만 최근 화두로 떠오르고 있는 데이터 마이닝, 텍스트 마이닝, 코퍼스 구축, 지리정보시스템(GIS), 시각화 서술, 토픽 모델링, 맵핑(mapping), 비주얼라이제이션(Visualization), 의미망(Semantic Web), 네트워크 분석 등에도 필요한 자료다. 또한 인공지능 기술을 기반으로 한 한자 폰트 제작, 문자인식, 기계번역 등에도 없어서는 안 될 원천자료이므로, 향후에도 다양한 목적과 특성을 지닌 데이터셋들이 개발될 가능성이 높다. 따라서 본고에서는 상술한 내용을 바탕으로 향후 한자 데이터셋을 구축할 때 유의하여야 할 사항 몇 가지를 결론을 대신하여 제안하고자 한다.

첫째, 한자 속성정보 규정을 위한 전문위원회 설치 및 종합관리시스템 구축 : 한자는 장구한 시간에 걸쳐 동아시아 각국에서 사용되어왔기 때문에 실로 다양한 형태의 자형이 존재할뿐더러 자음과 자의 역시 확정하기 어려운 경우가 많다. 그래도 한자를 사용하지 않을 수 없는 중국이나 일본은 관련 연구를 활발하게 진행해온 편이나 우리나라는 이에 관한 연구가 매우 영성한 실정이었다. 이런 실정 속에서도 '유니코드한자 검색

22) https://mojiportal.nabunken.go.jp/ja/
23) https://mojiportal.nabunken.go.jp/ja/ (검색일 : 2021년 6월 5일)

시스템', '이체자 정보', '한자자형전거' 등 상당한 규모의 한자 데이터셋이 구축된 점은 퍽 다행스러운 일이나, 관련 연구가 충분히 뒷받침되지 못하고 장기적인 활용계획 없이 단기적 성과를 내는 데 급급한 결과 데이터의 신뢰성, 호환성 등에서 문제를 노출하였다. 그러나 빈약한 인적 연구기반을 단기간에 끌어올리는 것은 무리인 만큼 일단 소수의 전문가라도 규합하여 당면한 문제를 해결할 수 있는 협의체 내지 위원회를 설치하는 것이 시급하다. 연구자 개인이 각개약진하지 말고 중지를 모을 수 있도록 해야 하며, 이 위원회에서 확정된 속성정보를 다른 기관이나 개인 연구자들도 공통적으로 사용할 수 있도록 오픈소스 형태로 제공해야 할 것이다.

둘째, 데이터의 품질 제고 : 데이터셋은 볼륨(Volume)보다 질이 관건이다. 데이터가 아무리 방대하더라도 신뢰할 수 없다면 무용지물이다. 한자 자형 데이터는 비정형 상태의 데이터를 정형 또는 반정형 데이터로 가공하는 과정에서 원천 정보의 변형이나 왜곡이 발생하지 않도록 하는 것이 중요하다. 또한 사용자가 검색하거나 처리하는 데 장애가 있어서는 안 된다. 따라서 데이터가 추가되거나 변경되었을 경우 그 이력과 내용을 추적할 수 있도록 지속적으로 관리하는 것도 중요하다.

셋째, 한자 데이터셋 표준화 : 용어, 도메인, 코드, 필드 구성, 필드명, 속성정보 등에 대한 공통안을 마련하여 데이터셋 간의 연계 및 활용 가치를 높이도록 하여야 한다. 〈표 4〉와 〈표 5〉는 일반적으로 데이터셋이 갖추어야 할 속성정보와 데이터 리소스 속성정보인데,[24] 이러한 자료를 표준안 제정의 기초자료로 삼을 필요가 있다.

〈표 4〉 데이터셋의 속성정보

항목	설명
Title	데이터셋의 이름
Ontology	데이터셋에 사용된 스키마
Number of Triples	전체 데이터 유닛(Triple 단위)의 개수
Class of Instance set	각 클래스별 데이터 인스턴스
Current Version	데이터 공개버전
Description	데이터에 대한 설명
Organization	공개한 단체명 및 URL
Direct Download Link	데이터 다운로드를 위한 URL 주소 및 가능여부

24) 데이터 가공의 방향에 관해서는 허철(2020)에서도 거론된 바 있다.

〈표 5〉 데이터 리소스의 속성정보

항목	설명
Type of Class	데이터 리소스에 사용된 스키마이름 및 URL
Example	데이터 데이터 리소스의 예시(HTML, RDF, N3 형태)
Number of Instance	데이터 셋에 포함된 데이터 리소스의 개수
Data Properties	값을 가지는 속성명과 상세 설명
Link Properties Link	대상을 가지는 속성명과 상세설명
Statistic of Properties	각 속성별 사용빈도에 대한 통계정보

유니코드 홈페이지에서 제공하는 한자의 종합정보인 Unihan DB에는 이러한 속성정보들을 [표 6]과 같이 분류하여 제공하고 있다.

〈표 6〉 Unihan DB 필드 내용

Fields within file	정보데이터
Unihan_DictionaryIndices.txt 사전 정보	kCheungBauerIndex, kCowles, kDaeJaweon, kFennIndex, kGSR, kHanYu, kIRGDaeJaweon, kIRGDaiKanwaZiten, kIRGHanyuDaZidian, kIRGKangXi, kKangXi, kKarlgren, kLau, kMatthews, kMeyerWempe, kMorohashi, kNelson, kSBGY
Unihan_DictionaryLikeData.txt 검색 정보(예, 창힐코드, 사각호마 등)	kCangjie, kCheungBauer, kCihaiT, kFenn, kFourCornerCode, kFrequency, kGradeLevel, kHDZRadBreak, kHKGlyph, kPhonetic, kTotalStrokes
Unihan_IRGSources.txt (IRG 코드 정보)	kCompatibilityVariant, kIICore, kIRG_GSource, kIRG_HSource, kIRG_JSource, kIRG_KPSource, kIRG_KSource, kIRG_TSource, kIRG_USource, kIRG_VSource, kIRG_MSource, kRS유니코드
Unihan_NumericValues.txt (갖은자 정보)	AccountingNumeric, kOtherNumeric, kPrimaryNumeric
Unihan_OtherMappings.txt (기타 전산 코드 정보)	kBigFive, kCCCII, kCNS1986, kCNS1992, kEACC, kGB0, kGB1, kGB3, kGB5, kGB7, kGB8, kHKSCS, kIBMJapan, kJa, kJis0, kJis1, kJIS0213, kKPS0, kKPS1, kKSC0, kKSC1, kMainlandTelegraph, kPseudoGB1, kTaiwanTelegraph, kXerox
Unihan_RadicalStrokeCounts.txt (부수, 획수 정보)	kRSAdobe_Japan1_6, kRSJapanese, kRSKangXi, kRSKanWa, kRSKorean
Unihan_Readings.txt (독음, 자의 정보)	kCantonese, kDefinition, kHangul, kHanyuPinlu, kHanyuPinyin, kJapaneseKun, kJapaneseOn, kKorean, kMandarin, kTang, kVietnamese, kXHC1983
Unihan_Variants.txt (이체자 정보)	kSemanticVariant, kSimplifiedVariant, kSpecializedSemanticVariant, kTraditionalVariant, kZVariant

이 데이터셋에는 각국에서 제출한 자형에 해당하는 사전 정보뿐만 아니라, 검색정보와 코드 정보, 부수와 획수 정보, 독음과 자의 정보, 이체자의 정보까지 모든 항목을 망라하고 있으므로, 향후 한국 한자 데이터셋의 표준안을 구성할 때에도 참고할 만하다.

현재 단국대학교 동양학연구원에서는 다양한 형태로 산재되어 있는 유니코드 미등록 한자 중 한국 고유한자(국자 이외에도 일부 국음, 국의자 포함)를 수집 · 정리하는 사업을 교비로 수행하고 있다. 그러나 기 구축된 데이터의 데이터 리소스를 확인하기 어렵고, 변화된 데이터의 이력을 추적하기도 어려워서 데이터 수집부터 구축까지 모든 과정을 반복해야 하는 경우가 다반사다. 또한 한문교육연구소에서는 Mask-RCNN 알고리즘을 활용하여 한자를 자동으로 인식하는 시스템을 개발하고 이를 이용해 한국 역대 문헌에 담긴 3억 자 이상의 한자를 자형 · 자체별로 분류 · 정리하는 데이터베이스 구축 사업을 진행 중이다. 우리보다 앞서 한자인식기술을 개발, 고도화 단계에 접어든 중국이나 일본에 비하면 매우 늦었지만 시행착오과정을 줄여 빠르게 따라잡고 있는 중이나, 분류 · 정리에 필수적인 대표자, 자음, 구건 표시 방식 등에 대한 국내의 연구가 거의 없어 많은 어려움을 겪고 있는 형편이다.

이러한 어려움을 해결하려면 무엇보다도 양질의 한자 데이터셋을 시급히 구축해야 한다. 양질의 한자 데이터셋은 한국학 및 동양학 분야의 연구에 활용하는 것은 물론 고전자료 데이터의 품질을 전반적으로 향상시켜 활용가치를 크게 끌어올릴 수 있다. 또한 문자인식 기술에 기반한 한자폰트 이미지 데이터셋 자동생성 및 이를 통한 다양한 한자 폰트 제작, 확장현실(XR) 기술을 적용한 한자 학습, 모바일 한자검색 애플리케이션 개발 등도 '질 좋은 데이터셋'만 있다면 얼마든지 구현할 수 있는 일이다.

〈참고문헌〉

배은한 · 김우정 · 조성덕 · 허철 · 주성일, 「한국 漢字音 표준화 방안 연구」, 한국고전번역원 기획연구과제 최종보고서, 2016.

박영미 · 하지영, 「일본의 한적 정보화와 코퍼스 구축 현황」, 『한문학논집』 53, 근역한문학회, 2019.

허철, 「한문자료 가공현황과 지향 탐색」, 『제1회 INDI학술대회 발표자료집』, 단국대 한문교육연구소, 2020.

李国英 · 周晓文, 「字料库建设的必要性和可行性」, 『北京师范大学学报(社会科学版)』, 2009年第5期, 2009.

王平, 「字典聯合檢索系統的信息組織與分類研究 : 以韓中日傳世漢字字典爲中心」, 『한자연구』 10, 2014.

王平, 「基于数据库的中日韩传世汉字字典的整理与研究」, 『中国文字研究』 19, 2014.

周亞民, 「中日漢字知識庫:漢字傳播與擴散觀點」, 『東吳中文學報』 24, 東吳大學中國文學系, 2012.

周亞民, 「結合中日漢字知識庫和臺灣教育部異體字字典在中日漢字比較的運用」, 『東吳中文學報』 25, 東吳大學中國文學系, 2013.

周亞民, 「結合數位典藏和字典在正體字保存及推廣的應用」, 『華語文教學研究』 14: 4, 2017.

祝國忠・周亞民,「淺談新世代資料庫系統的研究趨勢」,『景文技術學院學報』12(下), 2002.
山田太造,「電子くずし字字典データベースにおける現状と展望」,『研究報告人文科学とコンピュータ』, 2013.
井上聡,「東京大学史料編纂所「電子くずし字字典データベース」の概要と展望」,『情報の科学と技術』65, 2015.
北本朝展,「日本古典籍字形データセットの公開と活用への期待」, 第2回CODH-セミナー 발표자료, 2017.
守岡知彦,「大字典データベースのCHISEとの統合の試み」,『じんもんこん2019 論文集』, 2019.
池田証壽,「平安時代漢字字書総合データベースの構築」,『北海道大学文学研究科紀要』142, 北海道大学文大学文学研究科, 2014.
高田智和,「漢字字體と典籍の性格との関係－「漢字字體規範データベース」が主張するもの－」,『研究報告人文科学とコンピュータ』, 2013.
渡辺晃宏 외,「出土文字資料の画像データベースの構築」,『奈良文化財研究所年報』, 2012.

Website/Database:
한국역사정보종합시스템 Unicode한자 검색시스템'http://www.koreanhistory.or.kr/newchar/
한국학 디지털 아카이브 한자자형전거 http://yoksa.aks.ac.kr/jsp/hh/Directory.jsp?gb=1
한국학자료센터 이두용례사전 http://kostma.aks.ac.kr/dic/dicMain.aspx?mT=A
한국고전종합 데이터베이스 이체자 정보 http://db.itkc.or.kr/dch/
漢字字體規範史データセットhttp://www.hng-data.org
京都大学人文科学研究所所蔵 石刻拓本資料 http://kanji.zinbun.kyoto-u.ac.jp/db-machine/imgsrv/takuhon
CHISE IDS 漢字検索 https://www.chise.org/ids-find
平安時代漢字字書研究 https://hdic.jp/top/
HDIC Database Project https://github.com/shikeda/HDIC
Kuzushiji-MNIST https://github.com/rois-codh/kmnist
『電子くずし字字典データベース』『木簡庫』連携検索 http://clioapi.hi.u-tokyo.ac.jp/ships/ZClient/W34/z_srch.php
史的文字データベース連携検索システム https://mojiportal.nabunken.go.jp/ja/

* 이 논문은 2021년 5월 28일에 투고되어,
2021년 6월 11일에 심사위원을 확정하고,
2021년 6월 30일까지 심사하고,
2021년 7월 7일에 게재가 확정되었음.

Abstract

A Study on the Current Status and Improvement of Han characters(漢字) Dataset in Korea and Japan

Kim, Woojeong* · Park, Youngmi**

This paper examines the current status and problems of the Han character dataset in Korea and Japan, and suggests the direction of improvement. Han character dataset is very important as basic information of East Asian classical humanities research and utilization, including Han characters and Chinese literature. However, many problems were found due to the uniformity of the dataset construction method and poor management system, and the value of utilization was greatly reduced due to poor connection and compatibility between dataset. As a way to solve this problem, it was suggested that efforts should be made to standardize dataset and improve quality while establishing a specialized committee and a comprehensive management system for the provision of Chinese character attribute information.

[Keywords] Han character, Han character property information, data, dataset

* lead author, Professor, Dankook University
** co-author, Professor, Dankook University