DOI https://doi.org/10.35506/jspace.2023..34.004

일본학 분야 데이터베이스 발전 방향에 대한 탐색적 연구

: 빅데이터 기반 문제 분석의 토대 제공을 위한 시론

허원영(제1저자) _ 단국대학교 정치외교학과 초빙교수
정재은(제2저자) _ 중앙대학교 소프트웨어학부 교수
최희식(교신저자) _ 국민대학교 일본학과 교수

목 차

국문초록

오늘날 기술의 발전으로 데이터 양이 폭발적으로 증가하면서, 대형언어모델을 통해 대규모 데이터를 학습하고 결과를 생성하는 능력을 갖춘 AI가 등장하기에 이르렀다. 이러한 발전은 동시에 ChatGPT의 '독도' 검색 오류와 같이 인공지능의 답변이 잘못된 정보를 제공하는 경우도 나타나면서 한일 양국은 물론 전세계에서 사회적으로 중요한 문제로 떠오르고 있다. 그러나 일본학 및 일본 연구에서 이러한 기술의 활용은 제한적인 편이며, 대규모 데이터를 활용한 일본 관련 연구는 문화, 언어, 언론,

* 본 연구 수행 과정을 함께 하면서 많은 아이디어와 의견을 제시해 주신 국민대학교 일본학과 이원덕 교수님, 박창건 교수님, 한림대학교 일본학연구소 김남은 박사님, 인하대학교 국제관계연구소 박경민 박사님에게 깊은 감사의 말씀을 드린다.

관광 등에 주로 한정되어 있다. 이러한 상황으로 인해 일본 및 일본학과 관련된 데이터베이스의 활용 및 발전 방향에 대한 연구는 부족한 편이다. 이상의 문제의식에서, 균형있는 데이터베이스를 구축하고 공유함으로써 불균형 데이터의 학습에 의한 인공지능의 부정확한 답변을 최소화하고 효과적으로 활용할 수 있는 방안을 모색해보고자 한다. 이를 통해 인문사회과학 분야의 연구 범위를 확장하고 새로운 방식의 연구를 가능하게 할 수 있을 것으로 기대된다.

주제어
일본학, 데이터베이스(DB), 빅데이터, 인공지능, 소셜미디어

Ⅰ. 서론

기술의 발전으로 인해 오늘날 인류가 생산하는 데이터의 양은 폭발적으로 증가하고 있다. 대략적인 추정에 따르면 인류가 종이에 기록을 남기기 시작한 이후로 2000년대 초까지 약 5000년 동안 생산된 데이터는 약 20엑사바이트(exabyte)라 한다.[1] 그런데 2000년대 초반부터 2021년까지 약 20여 년 동안에 생산된 전 세계 데이터의 총량은 약 50제타바이트(zettabyte)로 추정된다. 지난 5000년 생산 데이터의 2,500배를 넘어서는 양의 데이터가 생산된 것이다. 이러한 데이터 생산량의 폭증과 함께 이를 효과적으로 이용할 수 있는 인공지능(Artificial Intelligence, AI) 기술이 딥러닝·기계학습 등을 통해 진화하면서, 대형언어모델(Large Language Model, LLM)을 이용하여 기존에는 다루지 못했던 큰 규모의 데이터(hyper-scale data)를 학습하고 이용자의 특정 요구에 따라 결과를 능동적으로 생성해내는 단계까

1) 엑사바이트는 0이 18개 붙는 단위로 100경 바이트에 달한다. 제타바이트는 엑사바이트의 상위 단위이다. 「우리에게 '데이터'는 어떤 의미인가」, 『동아일보』, 2022.08.24., https://www.donga.com/ news/It/article/all/20220824/115122895/1 (검색일: 2023.10.05.)

지 발전하였다.[2] 가장 대표적인 생성형 인공지능 서비스라 할 수 있는 Open AI사의 챗GPT는 2022년 11월 말 출시된 지 두 달만에 월간 활성 사용자 수(MAU) 1억명을 돌파하면서, 산업계와 실생활은 물론 학술과 연구 분야에 있어서도 큰 변화를 가져오고 있다.

그런데 일본학 또는 일본 연구의 관점에서 보았을 때, 이러한 기술의 변화가 충분하게 수용 및 활용되고 있다는 인상을 받기는 어렵다. 이는 특정 대상에 대한 면밀한 연구의 축적을 통해 학문적 성과를 올려야 하는 인문사회과학 분야에서 두드러지는데, 기존에 일본과 관련하여 대규모 데이터를 활용한 연구는 주로 문화, 일본어 교육, 언론, 관광 분야의 수용 양상 및 트랜드 분석을 다루는 경우가 많다.[3] 한편 관련된 데이터 또는 데이터베이스 자체를 연구의 대상으로 하고, 이들이 현재 이용되는 양상을 분석하면서 일본학과 관련하여 이들의 향후 활용 방향에 주목하는 연구는 찾아보기 어렵다.

그러나 현재의 기술 수준과 장기적인 발전 추세를 고려했을 때, 일본 또는 일본학과 관련하여 현재까지 만들어진 데이터 및 데이터베이스가 어떻게 활용되고 있으며 그로부터 파생되는 문제점을 어떻게 보완해야 할 것인

2) 양지훈·윤상혁, 「ChatGPT를 넘어 생성형(Generative) AI 시대로: 미디어·콘텐츠 생성형 AI 서비스 사례와 경쟁력 확보 방안」, 『미디어 이슈&트렌드』 Vol.55, 한국방송통신전파진흥원, 2023, pp.3-4.

3) 김다흰·임찬수, 「소셜 빅데이터를 통한 국내에서의일본 애니메이션 영화 흥행 요인분석 : 「너의 이름은.」과 「귀멸의 칼날: 무한열차편」키워드를 중심으로」, 『일본문화연구』 80집, 2021, pp.37-65; 김혜연, 「캐릭터 애니메이션을 활용한 한일 어휘 교육 방안연구 — 빅데이터 분석결과를 바탕으로」, 『일본어교육연구』 59집, 2022, pp.119-132; 도해용·이해주, 「빅데이터를 이용한 한국·중국·일본의 레스토랑 브랜드개성에 관한 연구」, 『외식경영연구』 18(6), 2015, pp.227-248; 윤영일·조문식, 「중국관광객은 왜 일본에 열광하는가?: 중국 SNS 빅데이터 분석을 통한 한국관광에의 적용방안 탐색」, 『인문사회과학연구』 30(3), 2022, pp.267-293; 이지수·고영란, 「미디어 이슈를 통해 본 일본에서의 한류 및 K-컬쳐의 토픽 양상 및 변화」, 『일본연구』 38집, 2022, pp.207-242; 하성호, 「빅 데이터를 통해 본 일본 콘텐츠 수용 양상의 변화:〈가면라이더〉 시리즈에 대한 인식을 중심으로」, 『일본연구』 40집, 2023, pp.103-135 등.

가가 시급한 과제로 떠오르고 있다. 주목할 만한 예시로, 2023년 4월 시점에서 챗GPT에 '독도'와 관련한 검색을 했을 때 잘못된 정보가 출력되며, 이는 AI의 답변이 일본 외무성의 데이터에 기반했기 때문이라는 사실이 보도된 바 있다.[4] 이는 이제까지 일본과 관련하여 생산되고 축적되어 온 데이터 및 데이터베이스들이 그저 독립적으로 존재하면서 객관적인 검토를 거쳐 사용되는 것이 아니라, 특정 국가의 입장이 과도하게 반영된 데이터의 불균형성(imbalancing)이 인공지능의 학습 과정에 치우침을 초래하면서 이용자에게 잘못된 일본 관련 데이터를 제공할 수도 있음을 의미한다.

이러한 문제의식에서 본고는 일본학, 특히 일본과의 국가 간 관계에 있어서 대규모 데이터를 활용한 데이터베이스의 발전 방향을 모색해보고자 한다. 한 가지 언급해 두어야 할 것은, 본고의 목적이 일본학 관련 데이터베이스 활용의 정답을 찾는 것이 아니라 관련 전문가들이 이 문제를 다루는 데 있어 필요한 문제의식과 정보를 공유하는 데 있다는 점이다. 따라서 논문의 내용은 개념을 정리하고 아이디어를 제시하는 탐색적인 연구가 될 것이다. 즉 이제까지 데이터베이스와 관련하여 무엇이 이루어져 왔으며, 일본이라는 국가 및 일본학과 관련하여 데이터베이스 활용 방향을 마련하기 위한 실마리를 찾아보고자 한다.

II. 데이터, 데이터베이스, 아카이브

본고는 일본학 분야의 데이터베이스(이하 DB)에 빅데이터 분석을 접목하기 위한 방안을 모색하는 데 초점을 맞추고 있다. 따라서 먼저 DB와 관련된 개념들에 대해 간략하게나마 살펴보고자 한다. 이는 먼저 개념 정리를 통

4) 「챗GPT 장악한 일본의 독도 억지 주장」, 『YTN』, 2023.04.02., https://www.ytn.co.kr/_ln/ 0103_202304020512193383 (검색일: 2023.10.01.)

해 본고가 논하고자 하는 내용의 토대를 공유하기 위해 필요하기 때문이다. 그러나 이와 더불어 DB와 관련된 개념을 정리하는 과정에서 인문학 분야와 사회과학 분야(이하에서는 정성적 접근과 정량적 접근이라는 표현을 함께 사용함) 간에 DB를 다루는 관점을 두고 보이는 차이점을 보다 명확히 할 수 있다. 이 차이점을 확실히 이해함으로써, 기존의 학술 DB가 작성·운영되어 온 방식과 향후 지향해야 할 방향성을 명확히 할 수 있을 것이다.

먼저 데이터(data)란 "현실 세계의 사람, 사물, 또는 사건에 대한 추상물"을 말한다. 이 추상물들은 변수, 특징, 속성 등의 개별 요소를 가진다.[5] 여기에서 데이터가 원론적으로 전산화 또는 디지털화 된 것을 지칭하는 것은 아니라는 점을 상기할 필요가 있다. 오늘날 생산되는 데이터의 대부분을 차지하며 또 본고에서 주로 논하게 될 데이터는 물론 디지털 데이터이지만, 기본적으로 데이터는 반드시 디지털(이진수의 불연속적인 형태)인 것은 아니다. 아날로그 형태의 데이터는 전자적 또는 광학적인 방법이 아닌 물리적인 방법으로 기록되고 저장된 모든 데이터를 말한다.[6] 이렇게 나누면 데이터 자체는 우리가 세계에 대해 추상적으로 이해하는 내용들(원자료, raw material)이며, 이를 기록하는 방식에 따라 아날로그와 디지털 데이터로 구분됨을 알 수 있다.

이러한 "데이터들을 상호 관련성에 따라 모아 놓음으로써 데이터를 더 쉽게 관리하고 접근할 수 있도록 저장하고 구성하는 수단"을 데이터베이스(database)라 한다.[7] 아날로그 데이터 또한 상호 관련성에 따라 모아 놓으면 아날로그 데이터베이스가 될 수 있다. 어떤 식료품 상점에 있는 식료품의 종

5) 존 캘러허·브렌던 티어니 저, 권오성 역, 『데이터 과학』, 김영사, 2019, p.49.
6) Clement Luong and John Min, "Analogue and Digital Data," Data Handling, http://Clement&John.com (검색일: 2023.07.08)
7) Microsoft, "What are database: Definitions, types, and wxamples of databases," http://azure.microsoft.com (검색일: 2023.07.07)

류와 개수, 가격 등은 데이터이지만 이들을 모아 종이에 적은 식료품 목록 장부는 아날로그 데이터베이스로 이해할 수 있다. 한편 이러한 내용들이 컴퓨터 시스템에 저장된다면 디지털 데이터베이스라 할 수 있을 것이다.

앞서 데이터와 데이터베이스에 대한 아날로그와 디지털의 차이가 기록 수단에 있었다면, 아카이브(archive)의 경우에는 보다 복잡한 개념적 차이가 포함된다. 먼저 데이터과학 또는 소프트웨어 분야에서 아카이브는 "사용 중이 아닌 데이터를 장기 보존 하기 위해 모아두는 것"을 의미한다.[8] 여기에서 아카이브는 공간적인 개념이라기보다는 데이터 관리의 관점에서 데이터들을 '고정적인 상태'로 보존하는 행위를 지칭하며, 따라서 아카이빙(archiving)이라는 표현을 사용하는 경우가 훨씬 많다.[9]

한편 정성적 접근을 하는 학문 분야에서 아카이브는 그 단어의 어원에 가깝게 이해된다. 사전에서 아카이브는 "공적인 기록 또는 역사적 자료(문서 등)가 보관된 장소"로 풀이된다.[10] 이러한 시설로서의 아카이브(공문서관)는 프랑스 대혁명 시기에 구축되기 시작하였으며, 1789년 파리를 시작으로 1838년 영국, 1872년 캐나다, 1934년 미국 등 국립공문서관(National Archive) 설립이 이어졌다.[11] 아카이브에 대한 이러한 정성적 접근은 최근 활발하게 구축되고 있는 '디지털 아카이브'의 개념과 밀접하게 관련되어 있다. 정성적 접근에서 아카이브는 아날로그 데이터 중에서 어떤 것이 가치 있는지 선별하고 기록하는 작업이 선행하기 때문이다. 이는 자동적으로 아날로

8) Symantec, "Your Backup is Not an Archive, White Paper: Data Protection," symantc.com, 2010, pp.1-3.
9) druva, "Data Archiving," https://www.druva.com/glossary/what-is-data-archiving-definition -and-related-faqs (검색일: 2023.07.07)
10) Merriam-Webster, "archive," Merriam-Webster's Learner's Dictionary, https://www.merriam -webster.com/ (검색일: 2023.07.07)
11) 황동열, 「문화·예술아카이브의 효율적 운영방안」, 『기록IN』 제18호, 2012, p.22.

[표 1] 데이터, 데이터베이스, 아카이브의 정의

	아날로그(analog)	디지털(digital)
데이터 (data)	현실 세계의 사람, 사물, 또는 사건에 대한 추상물. 아직 가공되지 않은 원자료(raw material). 여러 속성(attribute)을 가짐.	
	물리적인 방법으로 기록되고 저장된 모든 데이터.	컴퓨터 시스템 및 소프트웨어에 저장된 숫자 코드.
데이터셋 (dataset)	–	여러 속성으로 묘사되는 관측치의 집합. 구조화된 데이터, 통계분석에 활용하기 위한 데이터의 집합체.
데이터베이스 (database)	상호 관련된 정보의 집합(collection). 데이터를 더 쉽게 관리하고 접근할 수 있도록 저장하고 구성하는 수단.	
	물리적 수단에 기록·저장되는 경우 e.g. 종이에 쓴 식료품 목록	컴퓨터 시스템에 저장되는 경우 e.g. 지역 식료품의 재고 파일
아카이브 (archive)	역사적 가치 또는 장기 보존의 가치를 가진 기록이나 문서들을 보관하는 장소, 시설, 기관.	사용 중이 아닌 데이터를 장기 보존하기 위해 한 곳에 모아둔 것.
	디지털 아카이브: 문화유산 및 아날로그 정보를 디지털화해 모아 놓은 시스템 또는 디지털상에 조성된 데이터 저장고.	

출전: 캘러허·티어니(2019), Luong and Min 등을 참고하여 저자 작성.

그 데이터의 디지털 데이터화 및 데이터베이스화 작업을 포함하게 된다. 따라서 디지털 아카이브는 "문화유산 및 아날로그 정보를 디지털화해 모아 놓은 시스템"을 의미하게 된다.[12] 이는 데이터과학 분야에서 사용하는 아카이브의 의미와는 크게 다른 것이다.

이상의 정리를 통해 데이터에 대한 정성적 접근과 정량적 접근의 차이가 보다 뚜렷하게 드러난다. 가장 큰 차이점은 정성적 접근에서 데이터, 데이터베이스, 아카이브를 '보존'과 '디지털화'의 관점으로 바라보는 반면, 정량적 접근에서는 이들을 '정제'와 '활용'에 초점을 맞추어 바라본 다는 점이다. 이를 대표적으로 보여주는 것이 '데이터셋(dataset)' 개념이다. 아날로그 데이터에서 잘 사용되지 않는 이 개념은 "여러 속성으로 묘사되는 관측치의 집

12) 한국정보통신기술협회·전자신문사, 「디지털 아카이브」, 『최신 ICT 시사상식 2021』, 2020, p.36.

합”이며, 데이터과학을 수행하기 위해 가장 먼저 확보되어야 하는 것으로 생각된다.[13]

이러한 정량적 관점을 따르면 데이터와 데이터베이스는 데이터셋을 만들기 위해 필요한 것이다. 데이터셋이 확보되어야만 정량적(또는 데이터과학적인)인 분석과 활용이 가능해지기 때문이다. 반면에 정성적 접근에서는 데이터셋을 만드는 것보다는 아날로그 데이터를 수집하고 디지털화하여 보존하는 것에 초점이 맞추어지는 경우가 대부분이다. 이러한 차이는 인문사회과학 분야가 DB를 활용하는 방식에 두 가지의 큰 조류를 만들게 된다.

Ⅲ. 인문사회과학 분야 DB 활용의 방향

디지털 데이터와 DB의 역사는 컴퓨터의 역사와 같다고 할 수 있다. 그러나 앞서 살펴보았듯이 정량적 접근을 하는 학문 분야와 정성적 접근을 하는 학문 분야의 접근은 그 방향은 크게 다르며, 이는 DB의 활용에서도 나타났다. 아래에서는 본고의 목적에 맞추어 그 배경과 차이를 분명히 하고, 발견되는 차이를 기준으로 삼아 오늘날 활용되는 주요 DB를 도식화함으로써 이해를 돕고자 한다.

1. 정량적 접근: 통계적 처리를 위한 기초자료로서의 DB

데이터의 수집과 분석은 인류가 기록을 하기 시작한 이후로 행해져 온 것이다. 멀리로는 고대 메소포타미아의 거래 기록이나 이집트의 인구조사 통계로부터 오늘날의 전자거래 기록이나 소셜미디어 데이터에 이르기까지 다양하다. 그러나 본격적인 활용은 컴퓨터의 발전으로 디지털 데이터의 수

13) 캘러허·티어니(2019), pp.49-51.

집과 저장이 수월해진 1970년대부터 가능했다고 할 수 있다. 에드가 F. 코드(Edgar F. Codd)가 1970년 발표한 관계형 데이터 모델(relational data model)에 대한 논문이 데이터의 수집과 저장에 획기적인 변화를 가져왔다. 기존의 데이터 모델이 데이터가 물리적으로 어디에 저장되어 있는지 알아야 출력이 가능했다면, 코드의 모델은 사용자가 이를 모르더라도 미리 정의된 질의(query)만으로 이를 쉽게 가능하도록 했다.[14] 이는 현대의 관계형 DB와 SQL(structured query language)의 토대가 되었다.

이러한 변화는 경제학 등 다른 사회과학 분야는 물론 국가 간 관계를 다루는 국제정치학 분야에도 영향을 주었다. DB에 대한 체계적인 활용에 적극적이었던 것은 국가 간 전쟁에 대해 연구하는 일군의 학자들이었다. 미국 미시건 대학의 정치학자 데이비드 싱어(J. David Singer)는 전쟁에 대한 과학적 지식을 체계적으로 축적하기 위한 목적에서 1963년 '전쟁 상관관계 프로젝트(COW, Correlates of War)'를 발족시켰다. 이 프로젝트는 나폴레옹 시대 이후 국가의 국력, 동맹, 지리, 극성(polarity), 지위를 설명하는 요인들을 측정하기 위해 노력했고, 이러한 노력을 바탕으로 데이터셋을 확보했다.[15] 이 프로젝트는 초기부터 오늘날에 이르기까지 전쟁과 관련된 데이터 수집과 데이터셋 작성을 통해 연구 성과를 발표하고 있다.[16]

---

14) 캘러허·티어니(2019), p.18.

15) Corerelates of War Website, "History," https://correlatesofwar.org/history/ (검색일: 2023. 07.07)

16) COW의 대표적인 데이터·데이터셋 관련 논문은 다음과 같다. Melvin Small and J. David Singer, "Formal Alliances, 1815–1965: An Extension of the Basic Data," *Journal of Peace Research* 6, 1969, pp. 257–282; J. David Singer, "Reconstructing the Correlates of War Dataset on Material Capabilities of States, 1816–1985," *International Interactions* 14, 1987, pp. 115–32; Daniel M. Jones, Stuart A. Bremer, and J. David Singer, "Militarized Interstate Disputes, 1815–1992: Rationale, Coding Rules, and Empirical Patterns," *Conflict Management and Peace Science* 15, 1996, pp. 163–

그럼에도 국제정치 분야에서 디지털 데이터를 적극적으로 활용하기 위해서는 시간이 필요했다. 국가 간 관계를 구성하는 사건은 그 행위자가 복잡 다양하고 규모가 크고 자료도 제한적일 뿐만 아니라 경제학 및 사회학과 같은 인접 학문 분야에 비해 설문조사 등의 접근이 거의 불가능했기 때문이다. 이러한 어려움을 보여주는 대표적인 데이터셋은 러셀 렝(Russel J. Leng)이 주도하여 1987년 처음 발표한 '전쟁의 행위적 상관관계, 1816-1979 (BCOW, Behavioral Correlates of War)'이다. 이 자료는 1816년부터 1979년까지 164년 동안 발생한 45개의 국가 간 충돌(crises)에 대한 국가의 행동 데이터를 수집한 것이다.[17] 데이터셋 작성을 위한 코드북(codebook)과 코더 매뉴얼(coder's manual)에는 정성적인 아날로그 데이터(신문기사, 역사서, 연대기, 연표 등)을 수집하여 80개의 열로 나누어 데이터셋을 작성하기 위해 주의해야 할 점이 300페이지에 가깝게 상세히 서술되어 있다.[18]

국제관계에 대한 본격적인 대규모 디지털 DB인 ICOW(Issue Corrrelates of War)의 구축이 시작된 것은 1997년이었다.[19] 국가 간 분쟁 이슈가 거치는 갈등과 협상 과정을 이해하기 위한 목적에서 분쟁 데이터를

213; Glenn Palmer, Vito D'Orazio, Michael Kenwick, and Matthew Lane, "The MID4 Data Set: Procedures, Coding Rules, and Description," *Conflict Management and Peace Science*. 2015.

17) Russell J. Leng, Behavioral Correlates of War, 1816-1979. Inter-university Consortium for Political and Social Research [distributor], 2006-01-12. 여기에 데이터셋을 위한 프로그램의 일부가 VAX-11 Pascal로 작성되었음이 표시되어 있다.

18) Russell J. Leng, BEHAVIORAL CORRELATES OF WAR, 1816-1979 [Computer file]. 3rd release. Middlebury, VT: Middlebury College [producer], 1993. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 1995.

19) Paul R. Hensel and Sara McLaughlin, "Lessons from the Issue Correlates of War (ICOW) project," *Journal of Peace Research* 52(1), 2015, pp. 116-117. 이 논문에는 ICOW 프로젝트를 수행하는 데 있어서 국가 간 분쟁 데이터를 규명·정의하는 일의 어려움과 자료의 제약이 상세히 설명되어 있다.

수집한 ICOW는 2002년에 하천 및 해양 분쟁으로 그 대상을 넓혔고, 이 DB가 제공하는 데이터셋을 통해 국가 간 분쟁의 무력충돌 가능성, 평화로운 해결이 가능한 조건, 영토 분쟁의 무역에 대한 영향 등에 대한 다양한 연구들이 이루어졌다. ICOW는 1816년 이후의 시기에 주로 서유럽과 미주 대륙에서 발생한 분쟁을 대상으로 하고 있으나, 이를 통해 함의를 이끌어내는 국내의 연구도 다양하게 존재한다.[20] 1990년대와 그 이후 시기에 걸쳐 국제관계와 관련되어 셀 수 없을 만큼 많은 DB가 만들어졌고, 또 현재까지 수집, 제공되고 있다.[21]

이상의 흐름은 한국에서도 유사하게 전개되어, 정량적 접근을 사용하는 연구분야에서 많은 DB들이 작성되었고 국가적인 차원에서 이를 통합한 DB도 등장하였다. 1997년 11월 한국사회과학데이터센터(KSDC)가 설립되었고, 해당 센터는 1999년 1월부터 국제기구, 통계청 및 각종 연구소, 언론, 기업, 공공기관의 DB 수집과 표준화 작업을 진행하고 있다.[22] 이를 통해 공개되고 있는 한국 최대 규모의 사회과학 조사·통계자료 아카이브인 KSDC DB는 2,300건의 조사자료와 1,450건의 통계자료를 확보하고 있으며, DB의 역할로 '과학적인 양적연구 발전' '연구자의 원활한 자료탐색과 통계분석', '양적연

20) 대표적으로, 배진수·윤지훈 공저, 『세계의 영토분쟁 DB와 식민침탈 사례』, 동북아역사재단, 2008.

21) 대표적인 DB의 예로는 스웨덴 웁살라대학 평화분쟁연구학부 분쟁 데이터 프로그램(UCDP)과 노르웨이 오슬로 국제평화연구소(PRIO)가 공동으로 작성하고 있는 데이터셋(https://www.prio.org/Data/ Armed-Conflict/UCDP-PRIO/)과 무장 충돌 데이터셋(https://www.acleddata.com/), 스웨덴 스톡홀름 국제평화연구소(SIPRI)가 공개하는 군사지출 및 무기이전 데이터셋(https://www.sipri.org/ databases/milex), 비영리조직 CSP(Center for Systemic Peace)가 운영하고 있는 민주주의 척도 데이터셋(Polity)과 내전·테러 데이터셋(http://www.systemicpeace.org/inscrdata.html), 미 메릴랜드 대학이 시작한 글로벌 테러리즘 DB(https://www.start.umd.edu/gtd/) 등이 있다. (검색일: 2023.07.07)

22) KSDC Website, 「연혁」, https://ksdc.re.kr/bbs/content.php?co_id=history. (검색일: 2023. 07.07)

구 관련 강의지원' 등을 제시하고 있다.[23)]

이처럼 정량적 접근을 한 대표적인 사회과학 DB들의 특징은 데이터셋을 작성하기 위해 이용하는 원자료(raw material)의 내용을 보존하기보다는 이를 수치화 또는 코드화하는 것에 초점이 맞추어져 있다는 것이다. 이는 곧 이들 DB가 기존의 아날로그 데이터를 통계적으로 활용하기 위한 데이터셋 구축을 주요 과제로 삼고 있다는 것을 의미한다. 위에서 열거한 DB를 활용한 연구들 또한, 데이터셋에 회귀분석 등 통계적 처리를 가하여 함의를 찾는 것이 대부분을 차지하고 있다.

2. 정성적 접근: 디지털화와 보존에 중점을 둔 DB

한편 아날로그 데이터의 보존 및 디지털화에 중점을 둔 DB 구축 또한 생각보다 긴 역사를 가지고 있다. 그 시초라 할 수 있는 것은 1971년 시작된 구텐베르크 프로젝트(Project Gutenberg)이다. 마이클 하트(Michael Hart)에 의해 시작된 이 프로젝트는 공공의 영역에 속하는 저작권이 만료된 책들을 모두가 읽을 수 있는 무료 전자판으로 만들자는 것이었다. 처음에는 하트 자신이 저작권이 없는 오래된 책들 수백 권을 직접 입력하면서 진행되었다. 때로 다른 이들의 도움을 받았지만, 프로젝트 발족 후 20여년 동안은 이러한 작업이 지속되었다. 그러나 인터넷이 보편화되기 시작한 1990년대 중반부터 상황은 바뀌었다. 1997년까지 1,000권에 머물렀던 구텐베르크 프로젝트는 2002년 4월 5,000권, 2003년 10월 10,000권, 2008년 4월 25,000권으로 기하급수적으로 늘어났으며, 2023년 현재 70,000권이 넘는 전자책이 전자문서의 형태로 기록되어 있다.[24)]

---

23) KSDC Website, 「KSDC DB 소개」, https://ksdc.re.kr/bbs/content.php?co_id=ksdc_db (검색일: 2023.07.07)

24) Marie Lebert, "The Project Gutenberg EBook of Project Gutenberg

이처럼 정성적 접근을 하는 학문 분야에서 디지털 DB에 대한 관심은 개인용 컴퓨터의 발전, 그리고 인터넷의 보편화와 밀접하게 관련되어 있다. 학술지를 디지털 문서의 형태로 수집하고 공개하기 시작한 JSTOR(1994년), Proejct Muse(1995년) 등의 전자저널 초기 형태 역시 90년대 중반에 등장했다.[25] 일본에서도 이러한 흐름은 유사하게 나타난 것으로 보인다. 구텐베르크 프로젝트의 일본 버전에 해당하는 아오조라 문고(青空文庫)가 시작된 것은 1997년이다.[26] 또한 일본 정치 및 국제관계에 관한 최대 규모의 DB로 잘 알려진 '데이터베이스『세계와 일본』(データベース『世界と日本』)' 또한 1994년부터 자료 수집을 시작하였다.[27] 다나카 아키히코(田中明彦)가 주도하여 만들어진 이 DB는 19세기부터 현재에 이르는 9,000건 이상의 자료를 수집하여 공개하고 있는데, 특이한 점은 정치학 분야의 DB임에도 불구하고 앞서 살펴본 정량적 접근을 하는 학문 분야와 달리 통계 처리를 위한 데이터셋을 제공하거나 이를 목표로 하고 있지 않으며 자료의 수집과 디지털화, 공유에 중점을 두고 있다는 점이다.

한국의 경우에는 주로 고전학 자료에 대한 DB 구축 사업이 출발점이 되었다. 대표적으로 16만 페이지가 넘는 '조선왕조실록'의 전산화는 1995년 이루어졌는데,[28] 이를 바탕으로 이후에 다양한 DB들이 구축되고 제공되었다. 2001년부터 한국고전번역원에서 서비스를 시작한 한국고전종합 DB에는 '조

(1971-2008)," Project Gutenberg, 2008. https://www.gutenberg.org/ebooks/27045 (검색일: 2023.07.07)

25) 조지형, 「인문학의 '위기'와 디지털 인문학」, 조지형 편『디지털 시대의 인문학, 무엇을 할 것인가』, 사회평론, 2001, p. 161.

26) 青空文庫ウェブサイト, 「青空文庫 FAQ」 https://www.aozora.gr.jp/guide/aozora_bunko_faq. html. (검색일: 2023.07.07)

27) 政策研究大学院大学 ウェブサイト, 「データベース『世界と日本』基金」 https://www.grips.ac.jp/ jp/about/gripsfund/ (검색일: 2023.07.07)

28) 이재연, 「한국 문학에서 본 디지털 인문학 연구」, 이재연 외, 『세계 디지털 인문학의 현황과 전망』, 커뮤니케이션북스, 2019, p.16.

선왕조실록', '승정원일기'를 비롯한 9억여 자의 텍스트와 78만 면의 이미지, 500만 건의 메타데이터가 집적되어 있다.[29] 국사편찬위원회가 운영하는 '한국사 데이터베이스'도 101종, 848만여 건, 15억 1천만여 자의 자료(2019년 3월 기준)를 구축하여 제공하고 있다.[30] 다른 한편으로는 여러 아날로그 데이터를 디지털 공간 안에 배열하고 시각화하는 전자문화지도 방식의 DB도 구축되었는데, 예를 들어 고려대학교 민족문화연구원의 '조선시대 전자문화지도'는 2002년부터 진행되었다.[31] 이렇게 데이터의 시각화에 중점을 둔 DB의 방향은 시멘틱 웹 및 백과사전식 아카이브 구축으로 이어졌다. 시멘틱 웹(Semantic Web)이란 "현재의 웹에 명확한 의미의 정보를 부가해 사람과 컴퓨터가 협업할 수 있도록 하는 것"으로, 웹 주소와 같은 명명법을 통해 두 대상 사이의 관계성을 정형화된 방식으로 서술한다. 이를 통해 두 대상의 시간, 공간, 인적·물적 관계를 다양하게 표현할 수 있다는 것이다.[32]

한국연구재단에서 토대연구 및 인문사회연구소 지원사업을 통해 구축되어 온 DB들 역시 기본적인 방향성은 크게 다르지 않다. 두 사업에서 구축되었거나 구축 중인 53개의 DB를 대상으로 분석한 결과, 수집·해제를 목적으로 하는 DB가 56.6%(30개)로 가장 많았다. 그 다음이 디지털 아카이브 및 전자문화지도(24.5%, 13개)였으며, 데이터셋 제공이나 네트워크 분석, 통합적 DB 구축 등 보다 정량적인 접근을 목적으로 하는 DB는 18.8%(10개)에 지나지 않았다.[33]

---

29) 한국고전종합DB 웹사이트, 「소개」, https://db.itkc.or.kr (검색일: 2023.07.07)
30) 한국사데이터베이스 웹사이트, 「한국사데이터베이스 소개」, https://db.history.go.kr. (검색일: 2023.07.07)
31) 김홍규 외, 『조선시대 전자문화지도와 문화연구』, 고려대학교 민족문화연구원, 2006.
32) 김현·임영상·김바로, 『디지털 인문학 입문』, HueBooks, 2016, pp.147-148; 정주영, 『소극장 연극 시맨틱 아카이브 구축에 관한 연구』, 보고사, 2018.
33) 한국연구재단, 「기초학문자료센터」, https://krm.or.kr의 내용을 바탕으로 저자가 정리하여 계산함.

3. 인문사회과학 분야의 DB 활용

이상의 분석을 통해 보면 정성적 접근을 하는 분야에서는 DB에 대한 활용이 제한적인 것처럼 보인다. 그러나 세계적으로 DB를 활용한 인문학 연구는 '디지털 인문학(DH, Digital Humanities)'이라는 이름으로 활발하게 전개되고 있다. 이를 상징적으로 보여주는 것이 Digital Humanities Awards (이하 DH Awards)이다. 이 상은 매년 초에 지난 한 해 동안 전 세계에서 발표된 디지털 인문학 연구의 결과물을 대상으로 하며, 2012년부터 2022년까지 총 11차례의 시상식이 열렸다. 디지털 인문학 분야의 전문 연구자들로 구성된 '국제지명위원회(International Nomination Committee)'가 연구자들이 등록한 연구 결과물을 대상으로 심사하며, 기준의 충족 여부에 따라 분야별 후보자를 결정한며, 최종 우승과 준우승은 웹사이트를 방문하는 사용자들의 투표로 결정된다.[34)]

본고에서 참고할 만한 것은 이 상의 시상 분야이다. '즐거움', '도구', '시각화', '웹 간행', '대중성', '교육', '데이터셋'이 그것인데, 이 기준들을 압축적으로 정리하면 만들어진 DB가 '얼마나 잘 사람들에게 공유되는가(즐거움, 도구, 대중성, 교육)', 그리고 '얼마나 잘 표현되었는가(시각화, 웹 간행, 대중성)'로 나누어 볼 수 있을 것이다.[35)] 이를 '공유'와 '표현'으로 정의하면, 이제까지 본고에서 논의된 인문사회과학 분야의 DB들을 2개의 축을 가진 좌표평면 위에 배치해 볼 수 있다.

---

34) 국내 DB로는 '조선시대표류기록시각화'(http://soundh.net/drift)가 2019년에, '지암일기 디지털 인문학 연구'(http://jiamdiary.info)가 2020년에 각각 데이터 시각화 분야의 수상 후보로 지명되었다. "Digital Humanities Awards," https://dhawards.org. (검색일: 2023.07.07)

35) 김지선·류인태, 「지식의 공유와 표현 그리고 디지털 인문학: 해외 디지털 인문학 연구 사례 검토」, 『인문논총』 79(2), 2022, pp.391-427.

아래 도표의 가로축 왼쪽은 데이터의 통계적 활용에 중점을 두는 '데이터셋'이며, 오른쪽은 데이터의 수집과 정리, 디지털화에 초점을 맞춘 '보존'이다. 한편 세로축은 DH Awards에서 중점을 두고 있는 '공유와 표현'으로 설정했다. 이 좌표평면 DB들을 배치해 보면, 정량적 접근과 정성적 접근을 하는 DB의 괴리는 여전히 크지만, 두 종류의 DB 모두 공유와 표현의 정도를 기준으로 동일하게 평가할 수 있음을 알 수 있다. 즉 정량적 DB와 정성적 DB 사이에 우열은 없으며, 사용자들에게 잘 표현하고 공유하는 것이 중요한 판단 기준이 되어야 한다는 것이다. 이는 일본학 및 일본 관련 주제를 다루는 DB를 구상하고 구축하는 데 있어서도 마찬가지로 고민하고 적용해야 할 문제이다.

〈그림 1〉 인문사회과학 분야의 DB 활용 도식(예시)

출전: 저자 작성.

4. 빅데이터와 데이터 과학

앞서 살펴보았듯이, 정량적·정성적 접근을 사용하는 인문사회과학 분야의 연구는 DB를 서로 다른 방향으로 활용해 온 경향이 있다. 정량적 접근을

하는 많은 연구들이 통계나 서베이 등 다양한 데이터를 활용할 수 있었던 반면에, 정성적 접근을 하는 분야에서는 아날로그 데이터의 디지털화에 초점이 맞추어졌다.[36] 이로 인해 정량적 접근을 하는 분야에서는 아날로그 데이터를 코드화·데이터셋화하는 것에, 정성적 접근을 하는 분야에서는 텍스트화·디지털화하고 데이터 간의 상관관계를 시각화하는 것에 초점을 맞추었다.

그러나 1990년대 후반 이후부터 상황은 크게 달라지기 시작했다. 미국에서 데이터과학이라는 말이 널리 쓰이기 시작한 것도 이 시기부터다. 1997년 제프 우(C.F. Jeff Wu)는 대중 강연에서 당시 통계학의 분명한 몇 가지 경향을 강조했는데, 대표적으로 1) 거대 데이터베이스의 크고 복잡한 데이터셋 활용, 2) 컴퓨터 알고리즘과 모델이 점점 더 많이 쓰이는 현상이었다. 우는 이를 바탕으로 통계학을 데이터과학으로 바꿔 불러야 한다고 주장했다.[37] 이러한 경향은 2000년대 중반에 들어 온라인 서비스와 소셜미디어가 발달하면서 급격하게 가속되었다. Facebook (2004년 2월), YouTube (2005년 4월), Twitter (2006년 7월) 등은 모두 비슷한 시기에 서비스를 시작했으며, 이로 인해 이전에는 상상하기 어려웠던 양의 데이터가 생성되기 시작했다.[38]

빅데이터라 부르는 이 막대한 양의 데이터는 특성상 규모나 양을 정확하게 한정하여 정의하기 어려운 개념이다. 대부분의 조직에서 데이터의 상당량은 DB에서 오며, 그 규모가 커질수록 다양한 출처와 포맷에서 온 데이터가 담기기 때문이다.[39] 따라서 일반적으로 양(volume), 속도(velocity), 종류(variety)를 빅데이터의 3대 요소(3V)로 규정한다. 즉 빅데이터는 1) 많은 양

36) 이재연(2019), p.49, 40.
37) 캘러허·티어니(2019), p.29.
38) Oracle 대한민국, 「빅 데이터란 무엇인가?」, Oracle Cloud Infrastructure, https://www.oracle. com/kr/big-data/what-is-big-data/ (검색일: 2023.07.07)
39) 캘러허·티어니(2019), p. 33.

의 데이터가 있어야 하고, 2) 그 데이터가 빨리 수신되고 처리되며, 3) 사용 가능한 데이터의 유형(텍스트, 사진, 오디오, 비디오 등)이 다양하다는 것으로 정의할 수 있다.[40]

이러한 빅데이터의 특징은 데이터 처리장치와 인터넷의 속도가 빨라지고 저장장치의 용량이 지속적으로 늘어나며 활용 분야가 확대됨에 따라 우리의 삶은 물론 학문 분야에서도 받아들이지 않을 수 없는 것이 되고 있다. 인문사회과학 분야에서 빅데이터가 주목받기 시작한 좋은 예는 2004년부터 시작하여 3000만 권 이상의 책을 디지털화한 '구글 북스 라이브러리 프로젝트(Google Books Library Project)'일 것이다. 구글은 보존과 공유를 위해 막대한 양의 책들을 스캔하여 디지털화했고, 2010년 에레즈 에이든과 장바티스 미셸이라는 두 과학자가 해당 데이터를 바탕으로 개발한 '구글 엔그램 뷰어 웹사이트'가 공개되면서 이 '빅데이터'들이 '다루어질 수 있는 것'으로 분석되며 각광을 받기 시작했다.[41] '엔그램 뷰어'는 800만 권 이상의 책에 들어있는 8000억 개 이상의 단어가 1520~2012년까지 사용되는 빈도의 추이를 그래프로 보여주는 프로그램이다.

정량적 접근을 하는 사회과학 빅데이터 DB로 현재 가장 대표적인 것은 GDELT(Global Database of Events, Language, and Tone)일 것이다. 이 DB는 전 세계 언론기사를 수집하고 각각의 기사에 포함된 인물, 국가, 단체, 주제 등의 정보를 추출하여 코드화한 후, 이를 일정한 형식으로 바꾸어 제공한다. GDELT에 활용되는 언론기사의 언어는 100개가 넘으며, 이들 중 일부는 자동 번역을 통해 DB 구축에 활용된다. GDELT의 원자료는 홈페이

40) 국립중앙과학관, 「빅데이터: 빅데이터의 속성」, https://terms.naver.com (검색일: 2023.07.07)

41) 에레즈 에이든·장바티스 미셸 저, 김재중 역, 『빅데이터 인문학: 진격의 서막: 800만 권의 책에서 배울 수 있는 것들』, 사계절, 2015.

지에 공개되어 있으며, 구글(Google)의 빅쿼리(BigQuery)를 통해서도 활용 가능하다.[42] GDELT는 여러 종류의 데이터를 다루는데, 그중에서도 국가 간 관계를 다루는 학문분야에서 자주 활용되는 데이터셋은 사건 DB(Event Database)이다. 이 DB는 1.0 버전과 2.0 버전으로 나뉘어 있는데, 1.0 버전은 1979년부터 시작하여 1일 단위로 업데이트 되며 2.0 버전은 2015년 2월 18일부터 시작하여 15분 단위로 업데이트 된다.[43]

GDELT에서 국가 간 관계를 살펴보는 데 있어 가장 효율적인 자료는 Goldstein Scale과 어조(avarage tone)이다. Goldstein scale은 DB의 사건 분류에 따라 -10에서 10 사이의 값을 부여하는데, 0은 중립을 의미하며 음수는 부정적 사건, 양수는 긍정적 사건으로 볼 수 있다. 한편 어조는 자동화된 텍스트 분석을 통해 해당 기사가 담고 있는 의미의 긍정·부정적 정도를 분석한 것으로, 대체로 -10에서 10 사이의 값을 가지는데 이 역시 0보다 작으면 부정적이고 0보다 크면 긍정적인 것으로 해석할 수 있다.[44]

한편 빅데이터에 대해 따로 코드화를 하기보다는 공유와 표현에 초점을 맞추어 정성적 접근을 하는 DB들도 생겨나고 있다. 대표적인 것은 문화체육관광부와 언론재단이 함께 개발하여 2016년 4월 출시된 '빅카인즈(BIGKinds)'이다. 이 DB는 1990년 이후 축적된 54개 매체 7천만 건 이상의 뉴스 콘텐츠를 분석할 수 있으며, 기사 내 키워드와 개체명, 정보원, 트렌드 분석, 네트워크 분석 등의 서비스를 제공한다. 2020년에는 구글의 인공지능 언어모델 '버트(BERT)'를 적용하여 동음이의어 구분, 이형태 인식률 등을 높

42) The GDELT Procject, https://gdeltproject.org. (검색일: 2023.07.07)
43) 박성준, 「동북아 지역의 국제 갈등 양상과 무역분쟁: GDELT를 중심으로」, 『국가미래전략 Insight』 13, 2021.03.04., pp.5-6.
44) 박성준, 「빅데이터(GDELT)를 통해 살펴본 국가 간 갈등의 변화」, 『국제전략 Foresight』 6, 2021. 11.25., p.8.

이고 분석 품질을 향상시켰다.[45]

2021년 9월 서비스를 시작한 '국회회의록 빅데이터' 역시 유사한 접근을 하고 있다.[46] 이 DB는 제16대~제21대 국회를 대상으로 약 2만 건의 국회회의록 원문을 의원별 발언 단위로 분리하여 1,200만여 건의 데이터를 제공한다. 기존 국회회의록 DB가 회의록명과 검색과 PDF 원문 파일 확인 및 출력 기능에 머물렀다면, '국회회의록 빅데이터'는 빅데이터 기술을 활용하여 발언자나 키워드 등으로 검색할 수 있을 뿐만 아니라 의원별 발언 내용을 분석해 시각화된 데이터를 제공한다. 발언자·키워드 검색은 미국, 일본, 영국 등 주요국 DB에서도 제공되는 것이지만, 발언 내용을 분석하여 시각화된 데이터를 제공하는 것은 이 서비스만의 특징이다.[47]

이상의 흐름을 살펴보면, 이제까지 정성적 또는 정량적 방식으로 구축되어 온 DB의 양상이 크게 바뀌면서 결국 빅데이터라는 하나의 흐름 안에서 다양한 방식으로 수량화되고 시각화되어 활용되고 있음을 알 수 있다. 인문사회과학 분야 또한 빅데이터라는 기술의 급격한 변화와 함께 기존의 방식과는 다른 데이터 및 DB 활용이라는 과제에 직면하고 있으며, 이에 대해 각자의 방법으로 대응하고 있다는 것이다.

45) 한국언론진흥재단, 「빅카인즈 홈페이지」, https://bigkinds.or.kr (검색일: 2023.07.07); 한국언론진흥재단, 「한국언론진흥재단, 빅카인즈 서비스 개편」, 『신문과 방송 뉴스레터』, 2020.05.29.

46) 대한민국국회, 「국회회의록 빅데이터」, https://dataset.nanet.go.kr (검색일: 2023.07.07)

47) 「국회도서관, '국회회의록 빅데이터' 서비스 시작」, 『국민일보』, 2021.09.01. https://www. kukinews.com/newsView/kuk202109010032 (검색일: 2023.07.07)

Ⅳ. 일본학 분야 DB의 발전 방향

앞에서 살펴본 빅데이터의 등장과 여러 분야에서의 활용은 이를 일본 및 일본학과 관련하여 어떻게 결합시킬 것인가라는 과제를 안긴다. 일본과 관련된 문제를 다룰 때는 과연 어떠한 방향으로 빅데이터를 활용해야 할 것인가? 이를 보다 구체적으로 생각해보기 위해, 이하에서는 일본과의 관계에 있어 가장 첨예하게 대립하는 주제인 동시에 기존에 가장 많은 양의 DB가 구축되어 있는 역사문제에 초점을 맞추어 탐색을 시도해 보고자 한다.

구체적인 활용 방법에 대해 생각해보기에 앞서, 역사문제 관련 DB를 구성하는 방식에 대해 생각해 볼 필요가 있다. 첨단 과학 기술의 끝에 서 있는 빅데이터와 한일 간의 역사문제는 얼핏 보면 쉽사리 접점을 찾기 어려운 것처럼 보인다. 그러나 자주 회자되는 랑케(Leopold von Ranke)와 카(E. H. Carr)의 대비로 익히 알려져 있듯이,[48] 역사는 과거의 사실 그 자체뿐만 아니라 그것을 받아들이고 대하는 현재와의 상호작용도 중요하다. 역사문제를 다루는 DB에 대한 접근 역시 1) 역사문제 그 자체에 대한 것과 2) 역사문제에 대한 사람들의 인식에 대한 것으로 나누어 살펴볼 수 있다.

먼저 역사문제 그 자체에 대한 DB를 생각해보자. 이들 DB는 기본적으로 역사문제 관련 자료의 디지털화, 공유 및 표현에 집중한다.[49] 따라서 그 목적

48) 조지형, 『랑케 & 카: 역사의 진실을 찾아서』, 김영사, 2006.

49) 근대 이후 역사문제를 다루는 대표적인 국내 DB로는, 고려대학교 글로벌일본연구원 '일제강점기 자료 아카이브(http://archive.kujc.kr/?c=rare)', 국민대학교 일본학연구소·동북아역사재단 '한일회담 일본외교문서(http://contents.nahf.or.kr/item/level.do?itemId=kjj)' '한일회담 한국외교문서 DB (http://contents.nahf.or.kr/item/level.do?itemId=kj)' '한일회담 관련 미국무부문서 DB(http://contents.nahf.or.kr/item/level.do?itemId=kju)' '한일회담 미해결 과제 관련 자료 DB(http:// ffr.krm.or.kr/base/td044/index.html)', 국사편찬위원회 '한국사 데이터베이스'의 '일제강점기·대한민국 시기 자료 DB(https://db.history.go.kr/)'를 들 수 있다.

또한 수집, 보존, 탈초(읽기 쉬운 필체로 바꾸는 작업), 해제(자료의 정보와 내용에 대한 대략적 설명), 정본화, 아카이브화 등이며, 이러한 작업은 현재에도 매우 많은 인문사회과학 연구자들이 활발히 작업 중이다. 그러나 이러한 접근법을 특정 역사문제에 적용하여 해당 자료들을 빅데이터 수준으로 모으는 것은 극히 지난한 작업일 것이다.

한편 역사문제와 현재의 상호작용, 즉 역사문제에 대한 사람들의 인식을 다루는 DB는 쉽게 곧바로 이미지가 떠오르지 않는다. 특정한 또는 여러 역사문제에 대한 여론조사 및 서베이를 축정해 놓은 DB 정도가 이에 부합할 것이지만, 많은 비용을 요하는 조사 특성상 DB를 구축할 정도로 많은 자료를 확보하는 것은 어려울 것으로 생각된다. 빅데이터라 할 정도로 많은 양을 구축하는 것은 불가능에 가까울 것이다.

그렇다면 과연 빅데이터는 역사문제 관련 DB와 어떻게 연결될 수 있을 것인가? 역사문제 그 자체에 대한 DB를 빅데이터와 연결할 수 있는 실마리는 빅데이터의 세 가지 기준, 즉 3V에 있다. 즉 오늘날에는 '여러 종류의 다른 DB를, 많은 양으로 구성하여, 빠른 속도로' 처리할 수 있다. 단일한 기관 또는 집단이 빅데이터로 불릴 만한 역사문제 자료를 수집하고 DB로 구성하는 것은 불가능에 가깝지만, 기존에 구축된 여러 DB들을 연결하고 결합하여 분석하고 그 안에서 통찰을 이끌어내는 것은 가능하다는 것이다. 예를 들어 기존에 구축된 '한일회담 관련 미국무부문서' '한일회담 관련 일본문서' '한일회담외교문서' DB들을 하나의 DB로 구성함으로써, 한일회담을 둘러싼 한미일의 상호작용을 전혀 다른 관점에서 바라볼 가능성이 존재한다. 더 나아가 과거사 문제에 대한 일본측 또는 일본어 DB와 한국측 DB를 결합하고 분석함으로써, 한일 양국의 주장 사이의 차이점과 공통점을 이전과는 다르게 바라볼 수도 있다.

한편 역사문제와 현재의 상호작용과 관련해서는 매우 활용성이 뛰어난 빅데이터가 매일, 매시간 생성되고 있다. 다양한 소셜미디어 플랫폼에서 만들어지는 막대한 양의 데이터들이 그것이다. 이 내용들은 각 플랫폼의 사용자들이 생산하는 비정형 데이터들이며, 이를 통해 사람들의 인식을 읽어내는 것이 가능하다. 이를 지속적으로 수집한다면 한일 간에 역사문제를 두고 어떤 사건이 일어났을 때 사람들의 반응을 보고 카테고리화 할 수 있다. 여기에 적용이 가능한 연구는 현재도 이루어지고 있다.[50] 반대로 소셜미디어 내에서 일어나는 역사문제에 대한 반응을 보고 한일관계에서 정책적으로 활용 가능한 실마리를 얻을 수 있을지도 모른다.

〈그림 2〉 한일 간 역사문제의 DB 활용 유형

역사문제	역사문제 인식
(up to date)	
수집, 보존, 해제, 정본화, 디지털화, 시각화	여론조사, 서베이 등
⇩ ⇩ ⇩ (빅데이터·인공지능 활용) ⇩ ⇩ ⇩	
(to-be)	
DB들의 결합을 통해 활용 가능한 빅데이터 DB 구축	소셜 미디어 데이터 등 현재진행형 빅데이터를 활용

출전: 저자 작성.

이처럼 인문사회과학에서 빅데이터를 활용한다는 것은 해당 분야의 연구자가 직접 데이터과학을 수행해야 함을 의미하지는 않는다. 다만 데이터과학과 어떻게 협업하여 무엇을 발견하고 제공할 것인가를 고민해야 한다. 레빗

50) Giang T. C. Tran, Luong Vuong Nguyen, Jason J. Jung, Jeonghun Han, "Understanding Political Polarization Based on User Activity: A Case Study in Korean Political YouTube Channels," *Sage Open*, Vol. 12, No. 2, 2022.

과 더브너는 그들의 저서에서 현대인의 삶을 이해하는 핵심은 '무엇을 어떻게 측정할지를 아는 데 있다'고 말한 바 있는데, 이 구절은 본고의 고민과 맥이 통한다.[51] 인문사회과학 영역은 데이터과학에 '무엇을 측정할 것인지'를 요구해야 하고, 데이터과학은 '어떻게 측정할 것인지'를 제공해야 한다는 것이다.

이는 데이터과학에서 생각하는 유용성의 정의를 보면 보다 분명해진다. 데이터과학자들은 데이터과학을 이용해 추출한 패턴이 문제를 해결하는 데 필요한 통찰을 줄 때만 유용하다고 말하고, 이를 '실행 가능한 통찰(actionable insight)'이라 부른다.[52] 여기에서 '통찰'이란 '데이터과학을 통해 얻어낸 패턴이 분명하게 드러나지 않은 문제와 관련된 정보를 주어야 한다'는 뜻이며, '실행 가능성'은 '그러한 통찰이 현재 보유한 역량으로 어떤 식으로든 활용할 수 있는 것이어야 한다'는 뜻이다. 이를 본고의 문제의식에 비추어보면, 데이터과학과 협업을 통한 결과물이 인문사회과학 분야에 긍정적인 것이 되기 위해서는 이제까지 인문사회과학의 방법론으로 분명하게 드러나지 않은 문제와 관련된 정보들을 얻을 수 있어야 한다는 것이다.

빅데이터를 한일 간의 역사문제에 활용할 때 데이터과학이 줄 수 있는 통찰에 어떤 것들이 있는지 '상상'하기 위해서는, 데이터과학이 어떤 것들이 주목하는지 알아보는 것이 도움이 될 것이다. 데이터과학자들은 막대한 양의 데이터 안에서 기존과 다른 종류의 패턴을 추출하는 것에 초점을 맞춘다.[53] 그 예들은 다음과 같다.

1) 군집화(clustering): 비슷한 행동을 보이는 고객 집단을 찾아내는 패턴
2) 연관분석(association rule mining): 고객들이 어떤 제품과 함께 구매

51) 스티븐 레빗·스티븐 더브너, 『괴짜경제학』, 웅진지식하우스, 2007, p.14.
52) 캘러허·티어니(2019), p.16.
53) 캘러허·티어니(2019), pp.14-15.

하는 제품의 패턴

3) 이상(anomaly) 또는 극단값 탐지(outlier detection): 보험금 청구 사기와 같이 이상하거나 예외적인 사건들의 패턴 추출

4) 예측(prediction): 사물들을 분류하는 패턴의 발견. 현재의 어떤 속성에서 누락된 값을 예측(e.g. 스팸메일 분류)

이렇게 인공지능을 이용한 데이터과학의 장점과 특징을 역사문제와 관련된 DB에 적용한다고 하면 어떤 분석들이 가능할 것인가? 아래 표는 이에 대해 사고해 본 예시들이다.

[표 2] DB 간 연결 및 소셜미디어를 통한 빅데이터 활용 예시

	역사문제 관련 DB (DB들의 결합 등)	역사인식 관련 DB (소셜미디어 활용 등)
군집화	• 한국과 일본의 DB들 사이에 특정 사안에 대해 유사한 논조가 보이는 경우 파악하기	• 역사문제에 대해 유사한 반응을 보이는 유저들을 국가별로 묶어보기 • '혐한 발언'을 많이 하는 계정의 공통점 및 패턴 알아보기
연관분석	• 여러 DB들을 통해 역사문제에 대한 한일 양국의 연관 패턴 파악 • 한일 학술 DB가 한일 과거사 문제, 혐한 문제 등을 다루는 패턴 파악	• 소셜미디어의 특정 시점 데이터가 한일관계의 사건과 가지는 관련성과 패턴 파악 • 군집을 이루는 한일 소셜미디어 계정들의 연관성에 대한 파악
이상 탐지	• 방대한 양의 한국사 DB에서 일본과 관련하여 특정 단어의 빈도가 많아지는 시기를 탐지	• 소셜미디어에서 역사문제 키워드가 들어간 데이터가 급격히 많아지는 경우를 파악
예측 (기타)		• 소셜미디어의 역사문제 발언 급증을 통해 현실의 역사분쟁 예측

출전: 저자 작성.

V. 결론

위에서 살펴본 바와 같이, 인문사회과학 분야의 DB는 다양한 방식으로 발전을 거듭해 왔으며 기술의 발전으로 인해 빅데이터라는 기회이자 과제와

만나게 되었다. 일본 및 일본학 관련 분야에서도 이러한 변화는 현재진행형이며, 인공지능과 소셜미디어의 발달 상황을 고려했을 때 부정적인 영향을 최소화하면서 새로운 발견이 가능한 활용 방향을 찾아야 할 필요성이 분명히 존재한다. 그리고 새로운 일본 및 일본학 관련 DB를 구축함으로써 학계는 물론 한국 사회 및 한일관계에 기여할 수 있을 것이다. 새로운 방식의 DB 구축과 공유의 필요성에 대해서는 다음 세 가지로 나누어 설명할 수 있다.

첫째, 일본 관련 연구 및 한일관계에 대한 기술 발전의 부정적인 영향을 줄일 수 있다. 서론에서 언급했듯이, 이제까지 구축된 데이터와 DB는 인공지능을 통해 무작위적으로 학습되면서 특정 국가 또는 주장에 편향된 결론을 도출하고 한일 양국 사회에 부정적인 영향을 줄 수 있다. 이는 단순히 인공지능 기술의 한계에서 비롯되는 문제가 아니며, 인공지능의 적절한 학습을 위해 필요한 균형있는 DB의 구축이 선행되지 않았기 때문에 발생하는 문제이다. 이러한 일본학 분야의 데이터 불균형(imbalancing) 문제를 해결하기 위해 한국 또는 한일 양국의 다양한 DB들을 한데 묶어내어 DB로 구축하고 여기에 인공지능 분석을 적용함으로써 보다 객관적이고 통합적인 활용을 가능하게 할 수 있다.

둘째, 소셜미디어의 확대에 따른 현실의 영향력을 정확하게 파악하고 대안을 모색할 수 있다. 이제까지 소셜미디어의 영향력에 대한 우려와 경계는 항상 존재했으나, 이것을 DB의 관점에서 파악하고 선제적으로 활용하고자 하는 노력을 그다지 찾아보기 어려웠다. 특히 한일 양국 대중의 온라인 활동이 현실의 한일관계에 큰 영향을 주어 왔음에도 불구하고, 이에 대한 연구가 제한적이었다는 점은 지적되어야 한다. 온라인 소셜미디어의 데이터들을 정제하고 DB로 구축하여 활용함으로써, 일본과 관련된 현안을 보다 빨리 파악하고 선제적으로 예측까지 시도하는 토대를 마련할 수 있을 것이다.

셋째, 학술적으로 연구의 범위를 확장하고 새로운 방식의 연구를 가능하게 할 수 있다. 예를 들어 일본을 연구 대상으로 하는 인문사회과학 분야에서는 특정 주제에 대한 하나의 DB를 대상으로 이루어지는 경우가 일반적이었다. 소셜미디어의 내용 역시 항상 사후적으로 검토되며, 현재진행형 사안 또는 예측과 관련된 연구는 경원시되기 마련이었다. 기존 DB들을 한데 묶어 더 큰 DB로 만들어내고, 소셜미디어와 인공지능 분석 기법을 결합함으로써 이러한 한계에서 벗어나 새로운 연구의 영역을 개척할 수 있을 것으로 생각된다.

그러나 본고에서 살펴본 내용은 어디까지나 이와 같은 문제의식을 공유하고 아이디어를 제시하기 위한 시론에 불과하며, 일본·일본학 및 한일관계와 역사문제를 바라보는 관점에 따라, 그리고 데이터과학이 제공할 수 있는 수단의 다양성에 따라 얼마든지 많은 활용법을 모색할 수 있을 것이다. 일본 및 한일관계, 그리고 역사문제를 다루는 학자들이 기존에 없었던, 또는 필요할 것이라고 생각되는 문제를 상상해내고 제기함으로써 데이터과학으로부터 필요한 수단을 제공받을 수 있을 것이다. 인문사회과학의 연구자가 빅데이터의 특성을 어느 정도 이해하고 상상력을 발휘했을 때, 활용법의 지평이 더 크게 넓혀질 수 있을 것이다. 日本空間

논문 투고일 : 2023년 11월 16일
논문 심사일 : 2023년 11월 18일
게재 확정일 : 2023년 11월 28일

참고문헌

〈한글문헌〉

김다흰·임찬수, 「소셜 빅데이터를 통한 국내에서의일본 애니메이션 영화 흥행 요인분석 :「너의 이름은.」과 「귀멸의 칼날: 무한열차편」키워드를 중심으로」, 『일본문화연구』 80집, 2021.

김지선·류인태, 「지식의 공유와 표현 그리고 디지털 인문학: 해외 디지털 인문학 연구 사례 검토」, 『인문논총』 79(2), 2022.

김현·임영상·김바로, 『디지털 인문학 입문』, HueBooks, 2016.

김혜연, 「캐릭터 애니메이션을 활용한 한일 어휘교육 방안연구 — 빅데이터 분석결과를 바탕으로」, 『일본어교육연구』 59집, 2022.

김흥규 외, 『조선시대 전자문화지도와 문화연구』, 고려대학교 민족문화연구원, 2006.

도해용·이해주, 「빅데이터를 이용한 한국·중국·일본의 레스토랑 브랜드개성에 관한 연구」, 『외식경영연구』 18(6), 2015.

박성준, 「동북아 지역의 국제 갈등 양상과 무역분쟁: GDELT를 중심으로」, 『국가미래전략 Insight』 13, 2021.03.04.

박성준, 「빅데이터(GDELT)를 통해 살펴본 국가 간 갈등의 변화」, 『국제전략 Foresight』 6, 2021.11.25.

배진수·윤지훈 공저, 『세계의 영토분쟁 DB와 식민침탈 사례』, 동북아역사재단, 2008.

스티븐 레빗·스티븐 더브너, 『괴짜경제학』, 웅진지식하우스, 2007.

양지훈·윤상혁, 「ChatGPT를 넘어 생성형(Generative) AI 시대로: 미디어·

콘텐츠 생성형 AI 서비스 사례와 경쟁력 확보 방안」, 『미디어 이슈& 트렌드』 Vol.55, 한국방송통신전파진흥원, 2023.
에레즈 에이든·장바티스 미셸 저, 김재중 역, 『빅데이터 인문학: 진격의 서막: 800만 권의 책에서 배울 수 있는 것들』, 사계절, 2015.
윤영일·조문식, 「중국관광객은 왜 일본에 열광하는가?: 중국 SNS 빅데이터 분석을 통한 한국관광에의 적용방안 탐색」, 『인문사회과학연구』 30(3), 2022.
이재연, 「한국 문학에서 본 디지털 인문학 연구」, 이재연 외, 『세계 디지털 인문학의 현황과 전망』, 커뮤니케이션북스, 2019.
이지수·고영란, 「미디어 이슈를 통해 본 일본에서의 한류 및 K-컬쳐의 토픽 양상 및 변화」, 『일본연구』 38집, 2022.
정주영, 『소극장 연극 시맨틱 아카이브 구축에 관한 연구』, 보고사, 2018.
조지형, 「인문학의 '위기'와 디지털 인문학」, 조지형 편 『디지털 시대의 인문학, 무엇을 할 것인가』, 사회평론, 2001.
조지형, 『랑케 & 카: 역사의 진실을 찾아서』, 김영사, 2006.
존 캘러허·브렌던 티어니 저, 권오성 역, 『데이터 과학』, 김영사, 2019.
하성호, 「빅 데이터를 통해 본 일본 콘텐츠 수용 양상의 변화:〈가면라이더〉 시리즈에 대한 인식을 중심으로」, 『일본연구』 40집, 2023.
한국언론진흥재단, 「한국언론진흥재단, 빅카인즈 서비스 개편」, 『신문과 방송 뉴스레터』, 2020.05.29.
한국정보통신기술협회·전자신문사, 「디지털 아카이브」, 『최신 ICT 시사상식 2021』, 2020.
황동열, 「문화·예술아카이브의 효율적 운영방안」, 『기록IN』 제18호, 2012.

〈인터넷 자료〉

KSDC Website, 「연혁」, https://ksdc.re.kr/bbs/content.php?co_id=history. (검색일: 2023. 07.07)

KSDC Website, 「KSDC DB 소개」, https://ksdc.re.kr/bbs/content.php?co_id=ksdc_db (검색일: 2023.07.07)

Oracle 대한민국, 「빅 데이터란 무엇인가?」, Oracle Cloud Infrastructure, https://www.oracle. com/kr/big-data/what-is-big-data/ (검색일: 2023.07.07)

국립중앙과학관, 「빅데이터: 빅데이터의 속성」, https://terms.naver.com (검색일: 2023.07.07)

대한민국국회, 「국회회의록 빅데이터」, https://dataset.nanet.go.kr (검색일: 2023.07.07)

한국고전종합DB 웹사이트, 「소개」, https://db.itkc.or.kr (검색일: 2023.07.07)

한국사데이터베이스 웹사이트, 「한국사데이터베이스 소개」, https://db.history.go.kr (검색일: 2023.07.07)

한국언론진흥재단, 「빅카인즈 홈페이지」, https://bigkinds.or.kr (검색일: 2023.07.07.)

「국회도서관, '국회회의록 빅데이터' 서비스 시작」, 『국민일보』, 2021.09.01.

「우리에게 '데이터'는 어떤 의미인가」, 『동아일보』, 2022.08.24., https://www.donga.com/ news/It/article/all/20220824/115122895/1 (검색일: 2023.10.05.)

「챗GPT 장악한 일본의 독도 억지 주장」, 『YTN』, 2023.04.02., https://www.ytn.co.kr/_ln/ 0103_202304020512193383 (검색일: 2023.10.01.)

〈영어문헌〉

Daniel M. Jones, Stuart A. Bremer, and J. David Singer, "Militarized Interstate Disputes, 1815–1992: Rationale, Coding Rules, and Empirical Patterns," *Conflict Management and Peace Science* 15, 1996.

Giang T. C. Tran, Luong Vuong Nguyen, Jason J. Jung, Jeonghun Han, "Understanding Political Polarization Based on User Activity: A Case Study in Korean Political YouTube Channels," Sage Open, Vol. 12, No. 2, 2022.

Glenn Palmer, Vito D'Orazio, Michael Kenwick, and Matthew Lane, "The MID4 Data Set: Procedures, Coding Rules, and Description," *Conflict Management and Peace Science*. 2015.

J. David Singer, "Reconstructing the Correlates of War Dataset on Material Capabilities of States, 1816–1985," *International Interactions* 14, 1987.

Melvin Small and J. David Singer, "Formal Alliances, 1815–1965: An Extension of the Basic Data," *Journal of Peace Research* 6, 1969.

Paul R. Hensel and Sara McLaughlin, "Lessons from the Issue Correlates of War (ICOW) project," *Journal of Peace Research* 52(1), 2015.

Russell J. Leng, Behavioral Correlates of War, 1816–1979. Inter-university Consortium for Political and Social Research [distributor], 2006-01-12.

Russell J. Leng, BEHAVIORAL CORRELATES OF WAR, 1816-1979 [Computer file]. 3rd release. Middlebury, VT: Middlebury College [producer], 1993. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 1995.

Symantec, "Your Backup is Not an Archive, White Paper: Data Protection," symantc.com, 2010.

〈인터넷 자료〉

Clement Luong and John Min, "Analogue and Digital Data," Data Handling, http://Clement&John.com (검색일: 2023.07.08)

Corerelates of War Website, "History," https://correlatesofwar.org/history/ (검색일: 2023. 07.07)

"Digital Humanities Awards," https://dhawards.org (검색일: 2023.07.07)

druva, "Data Archiving," https://www.druva.com/glossary/what-is-data-archiving- definition -and-related-faqs (검색일: 2023.07.07)

Marie Lebert, "The Project Gutenberg EBook of Project Gutenberg (1971-2008)," Project Gutenberg, 2008. https://www.gutenberg.org/ebooks/27045 (검색일: 2023.07.07)

Microsoft, "What are database: Definitions, types, and wxamples of databases," http://azure.microsoft.com (검색일: 2023.07.07)

Merriam-Webster, "archive," Merriam-Webster's Learner's

Dictionary, https://www. merriam-webster.com/ (검색일: 2023.07.07)

The GDELT Procject, https://gdeltproject.org. (검색일: 2023.07.07.)

〈일본어 인터넷자료〉

青空文庫ウェブサイト,「青空文庫 FAQ」https://www.aozora.gr.jp/guide/aozora_bunko_ faq. html. (검색일: 2023.07.07)

政策研究大学院大学 ウェブサイト,「データベース『世界と日本』基金」https://www.grips. ac.jp/jp/about/gripsfund/ (검색일: 2023.07.07)

Abstract

An Exploratory Study on the Database Development in the Field of Japanese Studies : Seeking a Foundation for Big Data-Based Problem Analysis

Wonyoung Hur · Jason J. Jung · Heesik Choi

Recent technological advancements have led to a surge in data, and artificial intelligence (AI) technology, particularly enabled by large language models (LLMs), has demonstrated the ability to analyze extensive datasets and generate outcomes. However, despite the prevalence of these advancements, their application in Japanese studies is limited, primarily focusing on culture, language edcuation, media, and tourism research. This creates a research gap in the effective utilization and development of databases specifically related to Japan. Addressing these challenges, this paper aims to examine the current state of database utilization in Japanese studies and proposes strategies to enhance accuracy and effectiveness through innovative database construction and sharing. This initiative is expected not only to improve quality of databases but also to open new research avenues, expanding the scope of inquiry in humanities and social sciences.

Keyword

Japanese Studies, Database (DB), Big Data, Artificial Intelligence (AI), Social Media