디지털인문학의 미래, 스마트 빅데이터: 독일 트리어 디지털인문학 센터 CLS 연구 펠로우 후기

The Future of Digital Humanities, Smart Big Data: A Postscript from a Research Fellowship at Trier Center for Digital Humanities in Germany

김병준(독일 트리어 디지털인문학센터 연구 펠로우, KAIST 디지털인문사회과학센터 연구조교수)

目次

국문요약

이 글은 독일 트리어 디지털인문학 센터(TCDH)에서의 연구 펠로우십 경험을 바탕으로, 디지털인문학(DH) 연구의 새로운 패러다임인 "스마트 빅데이터" 개념을 소개하고 한국 DH 연구의 발전 방향을 제시한다. 최근 한국의 DH 연구에서 빅데이터 활용이 증가하는 추세이나, 데이터의 질적 수준과 인문학적 특성에 대한 고려가 부족하다는 우려가 제기되고 있다. 이에 대한 대안으로 스마트 빅데이터 개념을 제안하며, 이는 빅데이터의 규모와 다양성, 스마트 데이터의 질적 우수성을 결합한 것이다. 이 연구는 고급 기계가독형 데이터 포맷의 도입과 인공지능 기술, 특히 대규모 언어 모델(LLM)과 검색 증강 생성(RAG)을 활용한 스마트 빅데이터 구축 방안을 제시한다. 또한 해외 DH 연구 경험의 중요성을 강조하면서, 한국 인문학의 고유한 맥락을 고려한 창조적 수용과 발전의 필요성을 주장한다.

주제어 스마트 빅데이터, 전산문학연구, TCDH, LLM, RAG

Ⅰ. 들어가며

"해외 디지털인문학 선구자들은 어떤 생각을 할까?"

지난 10년간, 2015년 석사과정을 시작한 이래로 2024년 현재에 이르기까지 나는 디지털인문학을 연구하고 교육해왔다. 남들보다 일찍 디지털인문학 공부를 시작했다는 이유로 국내외 디지털인문학 연구사례를 소개하는 강연과 발표를 자주 하게 되었다. 그 자리에서 늘 제기되는 질문은 국내 디지털인문학(이하 DH)과 비교했을 때, 해외는 얼마나 더 앞선 연구를 하고 있느냐는 것이었다. 나는 언제나 해외의 유명 연구인 편지 공화국(Republic of Letters)이나 튜더 네트워크(Tudor Networks)와 함께 국내의 대표적인 성과물인 조선왕조실록, 삼일운동 데이터베이스 등을 균형 있게 소개했지만, 청중들에게는 DH라는 생소한 분야가 해외에서는 이미 자리 잡았을 것이라는 선입견이 있었던 것 같다. 해외 DH 연구와 교육을 다룬 논문(김바로, 2014)과 서적(이재연 외, 2019)이 지금도 많이 인용되는 걸 보면, 국내 연구자들 사이에서 DH는 아직 수입해와야 할 선진 학문으로 인식되는 게 아닐까? 나 역시 그런 프레임에서 자유롭지 못했고, 직접 해외 DH 연구를 경험해보고 싶었다. 그러던 중 좋은 기회가 찾아왔다. 바로 독일 트리어대학교 디지털인문학 센터(Trier Center for Digital Humanities, 이하 TCDH[1])에서 두 달간 연구 펠로우(Research Fellow)로 일할 수 있게 된 것이다. 이 글에서는 그 경험을 공유하고, 스마트 빅데이터(Smart Bigdata)라는 작은 제언을 하고자 한다.

가. TCDH와 CLS INFRA 펠로우십 소개

2023년 가을부터 해외에서 DH 연구 경력을 쌓고자 여러 기관을 조사하고 지원을 반복했다. 주로 1년 내외의 박사후 연구원이나 6개월 미만의 펠로우 및 방문연구자 프로그램에 지원했는데, 우연히 CLS INFRA(Computational Literary Studies Infrastructure, 전산문학연구 인프라)[2] 프로그램을 알게 되었다. 이를 통해 TCDH에 지원할 수 있었다. CLS INFRA는 유럽연구위원회(European Research Council, ERC)의 지원을 받아 시작된 프로그램으로, 전산문학 연구자에게 4주에서 12주까지 체재비를 지원하고 네트워킹 기회를 제공한다. CLS INFRA에는 유럽 10개국(독일, 프랑스, 스페인 등) 14개 대학 및 연구기관이 참여하고 있다.

CLS INFRA는 주로 유럽연합 소속 연구자들을 위한 프로그램이지만, 호라이즌 유럽(Horizon Europe) 관련 국가의 참여 또한 열려있다. 이러한 맥락에서, 대한민국이 2025년부터 아시아 최초

1) https://tcdh.uni-trier.de/
2) https://clsinfra.io/

〈표 1〉 CLS INFRA 참여 기관

기관	국가
Institute of Polish Language at the Polish Academy of Sciences	Poland
University of Potsdam	Germany
Austrian Academy of Sciences	Austria
National University of Distance Education	Spain
École Normale Supérieure de Lyon	France
Humboldt University of Berlin	Germany
Charles University	Czech Republic
Digital Research Infrastructure for the Arts and Humanities	France
Ghent Centre for Digital Humanities, Ghent University	Belgium
Belgrade Centre for Digital Humanities	Serbia
Huygens Institute for the History of the Netherlands (Royal Netherlands Academy of Arts and Sciences)	Netherlands
Trier Center for Digital Humanities, Trier University	Germany
Moore Institute, National University of Ireland Galway	Ireland
The Trinity Centre for Digital Humanities, Trinity College Dublin	Ireland

로 호라이즌 유럽의 준회원국 지위를 획득한 것은 큰 의미를 지닌다. 호라이즌 유럽은 2021년부터 2027년까지 약 130조 원 규모의 지원을 제공하는 대규모 연구 프로그램으로, 한국의 준회원국 가입으로 인해 국내 연구자들 역시 이 거대한 연구 생태계에 참여하여 연구비를 수주할 수 있는 기회를 얻게 되었다. 내가 한국인 최초로 CLS INFRA 연구 펠로우에 선정된 것도 한국의 호라이즌 유럽 준회원국 가입 덕분이 아닐까 싶다. CLS INFRA에서는 총 6차례 연구 펠로우를 모집하는데, 내가 지원한 5차 선발까지 총 48명의 펠로우가 선정되었다[3]. 5차 모집에서는 여러 기관 중 독일 트리어대학교의 TCDH를 선택했는데, 이는 공동 센터장인 크리스토프 쇼흐(Christof Schöch) 교수 때문이었다. 그는 세계 최대 디지털인문학 학술단체인 ADHO(Alliance of Digital Humanities Organizations)의 현 회장이자 전산문학 연구의 탁월한 성과를 보여준 연구자다. 나는 그가 유럽을 넘어 세계 디지털인문학의 현황과 미래를 잘 이해하고 있을 것이라 생각했기에, 직접 연락하여 연구 펠로우 지원을 타진했고 다행히 합격할 수 있었다[4].

3) 지금까지 뽑힌 연구 펠로우 목록은 다음을 참고할 것.
https://clsinfra.io/opportunities/tnafellowships/
4) 좀 더 자세한 연구 펠로우 지원 후기를 보려면 아래 유튜브 영상을 참고할 것.
https://youtu.be/EeAQBAJI2To?si=FDUfZ0xwE-a_6v4c

TCDH는 "인문학 전자 목록 및 출판 프로세스를 위한 역량 센터"라는 명칭으로 설립되었다. 디지털 사전 편찬과 디지털 에디션 분야에서 선구적이고 모범적인 솔루션을 제공하는 "디지털 아키텍트"로 명성을 얻은 곳이다. TCDH 연구진은 디지털인문학의 연구와 실무 분야에서 국내외적으로 활발한 기초 연구를 수행하며, 특히 디지털 에디션, 디지털 사전 편찬, 소프트웨어 시스템, 연구 인프라, 디지털 문학 및 문화 연구 등을 중점적으로 다룬다. 센터에는 전임교수 3명, 박사급 연구원 11명, 석박사과정 연구원 및 스태프 26명 등 총 40여 명이 근무하고 있다. 이는 국내 DH 연구 센터에 비해 상당히 큰 규모이며, DH 석박사 과정까지 운영하고 있다는 점에서 그 의의가 크다고 하겠다.

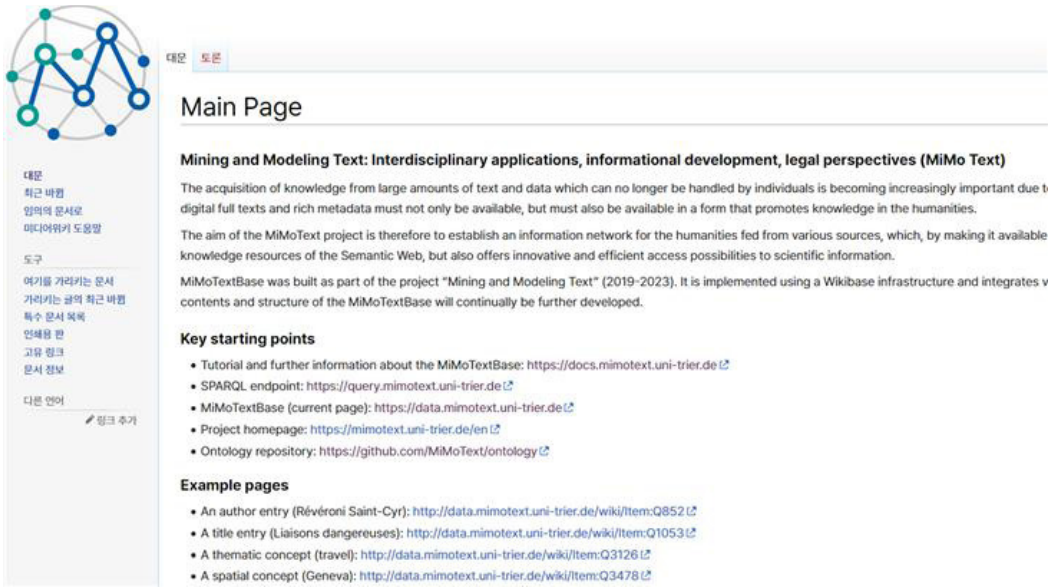
나. 대표 연구 프로젝트(MiMoText)

내가 참여한 프로젝트는 텍스트 마이닝과 모델링(Mining and Modeling Text, MiMoText)[5] 이었다. MiMoText는 1751년부터 1800년까지 프랑스 소설사 관련 디지털 자료를 대상으로, 방대한 텍스트와 데이터에서 지식을 추출하여 인문학 연구에 활용하는 것이 목표다. 이를 위해 다양한 정보원에서 추출된 정보를 Linked Open Data(LOD) 형식으로 변환하여 상호 연결하고, 시맨틱 웹상에서 자유롭게 이용 가능한 지식 네트워크를 구축함으로써 학술 정보에 대한 새로운 접근법을 제시한다. 국내에는 국립중앙도서관의 국가서지 LOD[6]나 한국사 LOD[7]가 대표적이나, MiMoText처럼 문학(사)에 특화된 LOD는 찾아보기 힘들다. 이 프로젝트는 문학 LOD 구축을 계획 중인 연구자들에게 훌륭한 참고 사례가 될 것이다. 특히 Wikidata 온톨로지를 활용해 문학(사)의 주요 메타데이터(저자, 제목, 개념 등)를 체계화했으며, SPARQL 쿼리[8]를 통해 데이터 탐색도 가능하다. 또한 프로젝트의 전체적인 구축 과정과 의의를 설명한 논문(Schöch et al., 2022)과 LOD의 기반이 된 18세기 프랑스 소설 200편의 TEI/XML 파일을 논문(Röttgermann, 2024) 및 깃허브 레포지토리[9]를 통해 공개했다. MiMoText의 가장 큰 장점은 문학(사) 자료(소설)의 수집 기준, 메타데이터 구성, 텍스트 주석(annotation) 작업 등이 모두 전공 연구자들의 토론과 전문 지식을 바탕으로 이루어졌다는 점이다. 데이터의 양적 규모뿐 아니라 질적 수준까지 고려한 스마트 빅데이터(Smart Big Data)의 모범 사례라 할 만하다.

MiMoText는 2023년 종료되었으나, 나는 이 프로젝트에 학술 데이터베이스인 OpenAlex[10]를 연결해 데이터를 확장하는 작업을 맡았다. MiMoText에는 문학 관련 학술 자료(논문, 단행본

---

5) https://mimotext.uni-trier.de/
6) https://lod.nl.go.kr/home/
7) http://lod.koreanhistory.or.kr/
8) https://query.mimotext.uni-trier.de/
9) https://github.com/MiMoText/roman18
10) https://openalex.org/



〈그림 1〉 MiMoText (출전:위키 페이지)
(https://data.mimotext.uni-trier.de/wiki/Main_Page)

등)에 대한 간단한 메타데이터(저자, 제목, 발행연도 등)가 포함되어 있는데, 이를 OpenAlex API의 검색 쿼리로 활용해 매칭되는 경우 OpenAlex 공유 URL을 추가하는 방식이었다[11]. 비록 두 달이라는 짧은 기간이었지만, 연구책임자인 쇼흐 교수를 비롯한 연구진과의 논의를 통해 종료된 프로젝트의 데이터베이스라도 계속해서 보완하고 확장하려는 의지를 느낄 수 있었다. 단순히 특정 연구과제의 결과물에 그치지 않고, 전산문학 연구라는 큰 흐름 속에서 데이터의 지속적인 발전을 도모하는 모습이 인상 깊었다.

Ⅱ. 디지털인문학에서의 데이터의 유형

연구 펠로우 기간 동안 나는 특히 쇼흐 교수가 이끄는 전산문학 연구 분야 데이터베이스의 특징이 무엇일지 고민했다. 그의 논문을 찾아 읽어보니, 쇼흐 교수가 주창한 스마트 데이터와

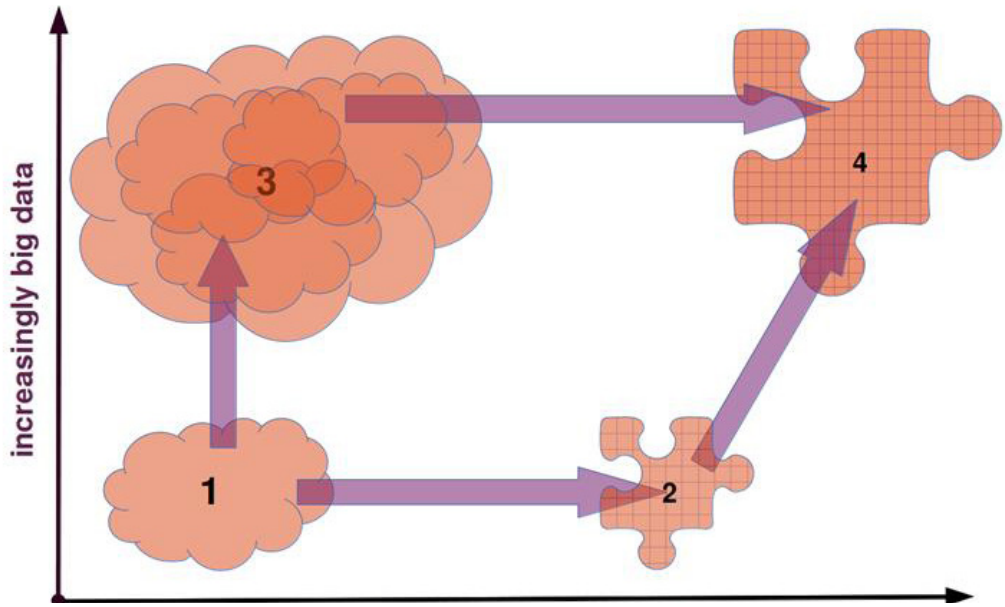
11) 해당 작업의 파이썬 코드와 과정 소개는 아래 깃허브 레포지토리와 링크를 참고할 것.
https://github.com/ByungjunKim/LODinG_OpenAlex
https://byungjunkim.com/talks/2024-05-02-talk-6

스마트 빅데이터로의 전환이야말로 이번 독일 연수를 관통하는 핵심 키워드였다. 2013년 발표된 그의 논문(Schöch, 2013)을 바탕으로 이 개념들을 소개하고자 한다.

가. 빅데이터와 스마트 데이터

DH에서 빅데이터란 대량의 비정형 데이터를 가리킨다. 디지털 기술의 발전으로 텍스트, 이미지, 영상 등 다양한 형태의 데이터가 기하급수적으로 증가하고 있는데, 이들은 대개 비정형적이라 기존의 데이터베이스 관리 시스템으로는 처리하기 어렵다(Mayer-Schönberger & Cukier, 2013). 반면 스마트 데이터는 체계적인 메타데이터와 주석(annotation)이 부가된 구조화된 데이터를 의미한다. 전통적인 인문학에서는 연구자가 직접 텍스트를 정독하고 분석하는 방식이 주를 이루었기에, 소량의 데이터를 심도 있게 다루는 것이 일반적이었다. DH에서도 이런 스마트 데이터의 중요성은 여전히 강조되고 있다(Schöch, 2013).

아래 그림에서 쇼흐 교수가 제시한 'The Story of smart and big data'의 과정을 볼 수 있다. 이 과정은 실제 작품에서 시작하여 디지털 표현으로 전환되고, 이후 스마트 데이터와 빅 데이터의 두 가지 경로로 발전한다. 최종적으로는 이 두 접근법을 결합하여 '스마트 빅 데이터' 또는 '더 큰 스마트 데이터'를 만들어내는 것이 목표이다. 이는 인문학 연구에서 데이터의 규모와 정밀도를 모두 확보하여 더욱 심도 있고 포괄적인 분석을 가능하게 한다.



〈그림 2〉 The Story of smart and big data (출전: Schöch, 2013, Figure 4)

예를 들어, 셰익스피어의 작품을 연구하는 경우를 생각해보자. 위 그림의 번호를 예시에 부여했다.

1. 실제 작품과 디지털 표현: 셰익스피어의 원본 필사본이나 초판본을 디지털화된 텍스트
2. 스마트 데이터: TEI 가이드라인에 따라 마크업된 셰익스피어 텍스트. 여기에는 등장인물, 대사, 무대 지시문 등이 구조화되어 있으며, 언어적 특징이나 문학적 장치에 대한 주석이 포함됨.
3. 빅 데이터: 구글 북스나 인터넷 아카이브에서 수집한 셰익스피어 시대의 모든 영어 텍스트. 정제되지 않았지만 대량의 데이터를 포함.
4. 스마트 빅 데이터: 셰익스피어 시대의 모든 영어 텍스트에 대해 자동화된 방법과 크라우드소싱을 통해 기본적인 구조화와 주석을 추가한 대규모 코퍼스. 이를 통해 셰익스피어의 작품을 당대의 문학적, 언어적 맥락 속에서 포괄적으로 분석할 수 있음.

이러한 접근은 셰익스피어 작품의 언어적 특성, 주제의 고유성, 다른 작가들과의 관계 등을 더욱 깊이 있고 광범위하게 이해할 수 있게 해준다. 즉 스마트 빅데이터는 빅데이터의 규모와 다양성, 스마트 데이터의 질적 수준을 동시에 확보한 데이터를 가리킨다. 즉, 대규모 데이터를 수집하되 체계적인 메타데이터와 주석을 통해 품질을 높이는 것이 핵심이다. Schöch (2013)는 스마트 빅데이터야말로 DH가 지향해야 할 새로운 방향이라고 역설한다. 이를 통해 인문학의 질문을 보다 정교하게 설정하고, 통찰력 있는 해석을 도출할 수 있기 때문이다.

나. 빅데이터와 한국의 디지털인문학 연구

최근 한국의 DH 연구는 주로 빅데이터의 관점에서 접근되어 왔다. 비록 현재는 인공지능이라는 새로운 '시대정신'으로 초점이 이동하고 있지만, 대량의 데이터를 수집하고 처리하는 것이 DH 연구에서 매우 중요시되어 온 것은 사실이다. 그러나 이러한 접근은 '스마트 빅데이터' 개념과는 거리가 있었다. 많은 DH 연구가 데이터의 양적 확보에 치중한 나머지, 데이터의 질적 수준이나 심도 있는 분석과 해석은 상대적으로 소홀히 했다는 비판을 받았다.

이는 DH의 본질이 디지털 기술을 활용한 인문학 연구임에도 불구하고, 기술 중심의 사고에 경도된 결과로 볼 수 있다. 즉, DH를 단순히 빅데이터를 다루는 것으로 인식하는 근본적 오해에서 비롯된 것이다. 그러나 인문학 데이터는 자연과학이나 사회과학의 데이터와는 그 성격이 다르다. 인문학 데이터는 맥락 의존적이고 해석적인 속성을 지니고 있어, 단순히 대규모 데이터를 확보한다고 해서 의미 있는 연구 결과로 이어지기 어렵다(Owens, 2011).

이러한 맥락에서, 성균관대학교 대동문화연구원의 호적데이터 전산화 프로젝트(호적DB)[12]는

12) https://skb.skku.edu/ddmh/db/intro.do

‘스마트 빅데이터’의 좋은 예시라고 할 수 있다. 이 프로젝트는 한문으로 기록된 원문 텍스트를 디지털화하고, 72개의 메타 정보를 설계하여 엑셀 데이터로 작성하였다. 이는 방대한 텍스트에서 중요 정보를 관련 전공자들의 전문 지식을 바탕으로 추출하여 의미 있는 대규모 데이터셋을 만들어낸 것이다.

그러나 이 프로젝트에도 개선의 여지가 있다. 현재 데이터가 엑셀 형식으로 저장되어 있어, XML, JSON, LOD와 같은 표준화된 기계가독형 데이터베이스 형식으로 변환이 필요하다. 이러한 형식으로 변환되면 웹이나 다양한 프로그래밍 언어에서 더욱 효율적으로 데이터를 다룰 수 있게 되어, 데이터의 활용도와 접근성이 크게 향상될 것이다. 이와 관련한 자세한 제언은 다음장에서 다뤄보겠다.

Ⅲ. 스마트 빅데이터 구축을 위한 제언[13)]

결국 디지털 인문학(DH) 연구의 미래는 스마트 빅데이터에 달려 있다. Schöch et al. (2022)의 최근 연구에 따르면, DH의 현재 흐름은 크게 세 가지로 구분된다. 첫째, 엄선된 소량의 스마트(스몰) 데이터를 활용한 연구, 둘째, 토픽 모델링이나 워드 임베딩과 같은 빅데이터 방법론을 적용한 연구, 그리고 마지막으로 상대적으로 큰 규모이면서도 '스마트'한 데이터셋, 즉 스마트 빅데이터를 구축하는 연구이다. 이 중 두 번째 접근법은 전통적인 인문학 데이터의 특성과 맞지 않아 빅데이터를 강제하는 경향이 있다는 지적을 받고 있다.

쇼흐 교수가 지적했듯이, 차세대 DH 연구의 성패는 적정 규모와 고품질을 겸비한 데이터셋 구축에 달려 있다. 이는 곧 데이터셋의 질적 경쟁으로 이어질 것이며, 고품질의 데이터셋을 양적으로 구축하되 정성적인 관점에서도 세심하게 다듬어야 한다는 의미이다. 그렇다면 구체적으로 어떻게 스마트 빅데이터를 구축할 수 있을까?

먼저, 고급 기계가독형 데이터 포맷의 도입이 필수적이다. 김바로(2022)는 공공데이터법을 인용하며, 의미 있는 공공(인문)데이터는 기계 판독이 가능해야 하고, 모든 소프트웨어에서 자유롭게 활용할 수 있는 "오픈 포맷"이어야 함을 강조했다 (표 2). MiMoText 프로젝트의 사례에서 볼 수 있듯이, 스마트 빅데이터는 다른 데이터베이스와의 연결을 고려하여 구축해야 한다. 이를

13) 쇼흐 교수는 2024년 5월 22일부터 25일까지 한국을 방문해 경북대, 고려대, 동국대에서 디지털인문학(전산문학연구) 강연을 진행하였다. 그의 홈페이지에서 강연 발표자료를 확인할 수 있으며, 특히 고려대에서 진행한 "Bigger Smarter Data: Extracting, Modeling and Linking Data for Literary History" 발표는 그가 생각하는 스마트 빅데이터에 대한 전반적인 내용을 잘 보여주는 자료이다. https://christof-schoech.de/activities/?_sf_s=Korea

〈표2〉 기계 판독이 가능한 형태의 포맷 단계별 구분·비교 (출전: 김바로, 2022, 표1에서 재인용)

구분	1단계	2단계	3단계	4단계	5단계
기계판독이 가능한 형태	미충족포맷 (포털등록불가)	최소충족포맷	오픈포맷		
특징	특정 소프트웨어에서 읽을 수만 있는 데이터로 자유로운 수정, 변환 불가	특정 소프트웨어에서 읽고 수정, 변환 가능	모든 소프트웨어에서 읽고 수정, 변환 가능	URI를 기반으로 데이터 속성 특성 관계를 기술하고 있는 데이터 구조	웹상의 다른 데이터와 연결, 공유 가능
예시	PDF	HWP, XLS, JPG, PNG, WMV, MPEG, MP3, SWF	CSV, JSON, XML	RDF	LOD

위해서는 최소한 CSV, JSON, XML과 같은 3단계 이상의 포맷을 채택해야 하며, 더 나아가 RDF나 LOD와 같은 고급 포맷을 지향해야 한다. 예를 들어, 현재 엑셀 파일로만 구성된 호적DB를 3단계 이상의 포맷으로 변환한다면, 한국사 관련 LOD 데이터베이스와의 연계를 통해 더욱 확장된 연구가 가능해질 것이다.

그러나 기존의 스마트 데이터 구축 방식에는 한계가 있었다. 전통적으로 이는 전문 인력의 투입과 체계적인 주석 규칙 마련이 필수적이었으며, 방대한 데이터를 일일이 정제하고 구조화하는 데 엄청난 시간과 비용이 소요되었다. 이로 인해 스마트 데이터화는 소규모 데이터에 한정되었고, 스마트 빅데이터는 사실상 구현하기 어려운 개념으로 여겨졌다. 특히 국내의 경우, 인문학 연구자 수 감소와 대학원 진학률 하락으로 인해 대규모 인력 동원이 현실적으로 어려운 상황이다.

다행히도 최근 인공지능 분야의 발전이 이러한 한계를 극복할 수 있는 가능성을 제시하고 있다. 특히 거대언어모델(LLM)과 retrieval-augmented generation(RAG) 기술의 발전은 스마트 빅데이터 구축에 새로운 지평을 열고 있다. LLM은 방대한 텍스트 데이터를 학습하여 뛰어난 문맥 이해와 텍스트 생성 능력을 보여주며, RAG는 LLM에 외부 지식베이스를 연동하여 더욱 정확하고 풍부한 정보 생성을 가능케 한다(Lewis et al., 2020)[14].

14) RAG가 LLM의 할루시네이션 현상을 막고, 인문학같은 전문 지식에 특화된 LLM을 만드는데 현재 최적의 방법으로 주목받고 있다. 아쉽게도 디지털 인문학 분야에서 RAG의 직접적인 적용 사례는 아직 많지 않지만, 유사한 개념이나 관련 기술을 활용한 연구들이 진행되고 있다. Manjavacas와 Karsdorp (2019)는 대규모 언어 모델을 사용하여 고대 문서를 분석하는 방법을 제시했으며, Blanke 등 (2020)은 AI 기술을 활용하여 홀로코스트 생존자의 증언을 분석했다. Meroño-Peñuela 외 (2022)는 문화유산 데이터를 링크드 데이터로 구축하고 이를 언어 모델과 결합하여 질의응답 시스템을 개발했는데, 이는 RAG의 기본 개념과 유사한 접근 방식이라고 볼 수 있다. 최근에는 Simeone과 Okereke (2023)가 GPT와 같은 대규모 언어 모델을 디지털 인문학에 적용하는 방법을 광범위하게 조사했다. 이러한 연구들은 RAG와 직접적으로 연관되지는 않지만, 대규모 언어 모델과 정보 검색 기술을 디지털 인문학에 적용하는 방법을 보여주고 있다.

국내 DH 연구자들도 이러한 기술의 중요성을 인식하고 있다. 김현(2023)은 온톨로지 기반의 시맨틱 데이터 편찬이 인공지능, 특히 LLM과 함께 발전해야 함을 강조하였다. 그는 협업을 통해 구축한 지식 그래프를 LLM에 학습시켜 활용하는 것이 인문지식의 디지털 큐레이션 교육의 미래 방향이라고 제시하였다. LLM과 RAG를 활용한다면, 대규모 데이터를 반자동화된 방식으로 정제하고 주석화할 수 있을 것이다. 예를 들어, 인문학 지식베이스를 LLM에 주입하여 해당 분야에 특화된 모델을 만들 수 있다. 이는 그동안 불가능하다고 여겨졌던 스마트 빅데이터의 구현을 현실화하며, DH 연구의 판도를 바꿀 혁신적인 접근법이 될 수 있다.

물론 이러한 접근법은 아직 초기 단계에 있으며, 해결해야 할 과제들이 많이 남아있다. 그러나 스마트 빅데이터가 DH의 미래를 이끌 핵심 개념이 될 것이라는 점은 분명하다. 방대한 데이터를 대상으로 정교한 질문을 제기하고 통찰력 있는 해석을 이끌어내는 능력, 이것이야말로 스마트 빅데이터가 인문학 연구에 새로운 지평을 열어줄 열쇠가 될 것이다.

Ⅳ. 나오며

이 글은 독일 TCDH에서의 연구 펠로우 경험을 토대로, 스마트 빅데이터라는 새로운 데이터 패러다임을 소개하고 이를 통해 한국 DH 연구의 현주소와 나아갈 방향을 짚어보고자 했다. 최근 한국의 DH 연구에서 빅데이터 활용이 증가하는 추세이지만, 일부 연구에서는 데이터의 질적 수준과 인문학 고유의 특성에 대한 고려가 충분하지 않을 수 있다는 우려가 제기되고 있다. 이러한 상황에서 스마트 빅데이터라는 새로운 비전은 양적 접근과 질적 접근의 균형을 도모할 수 있는 대안이 될 수 있다. LLM, RAG 등 인공지능 기술을 활용한 스마트 빅데이터의 구현은 DH의 발전을 위한 중요한 과제가 될 것이다.

특히 고품질 데이터셋 구축은 인공지능 자동화가 보편화된 시대를 맞아 인문학자의 전문성이 더욱 빛을 발할 수 있는 분야가 될 전망이다. 인공지능 기술이 고도화될수록 데이터 품질이 연구 결과에 미치는 영향력은 더욱 커질 것이기 때문이다. 따라서 인문학자들은 데이터의 수집, 정제, 주석 작업 등에 있어 그들의 전문 지식을 적극적으로 활용해야 할 것이다.

한편, 해외 DH 연구기관이 국내보다 언제나 앞서 있다고 단언하기는 어렵다. 그들 역시 DH의 방향성과 방법론을 두고 다양한 고민을 거듭하고 있기 때문이다. 다만 주목할 만한 점은 해외 연구자들 사이에서 빅데이터 활용에 대한 신중한 접근이 강조되고 있다는 것이다. 이는 국내 학계에서도 참고할 만한 시사점을 제공한다.

이런 맥락에서 해외 DH 연구기관에서의 직접적인 연구 경험은 매우 귀중한 자산이 된다. 최신

연구 동향을 파악하고 국제적 연구 네트워크를 구축함으로써, 국내 DH의 미래 방향을 모색하는 데 큰 도움을 얻을 수 있을 것이다. 물론 외국의 사례를 무조건 좇을 수는 없다. 한국 인문학의 고유한 맥락과 특성을 면밀히 고려하는 가운데, 해외의 노하우를 창조적으로 수용하고 발전시켜 나가는 지혜가 요구된다. 이를 통해 한국 DH만의 독자적인 길을 개척해 나갈 수 있으리라 믿는다.

| 참고문헌 |

김바로, 「해외 디지털인문학 동향」, 『인문콘텐츠』, 33 (2014).

김바로, 「〈공공데이터법〉과 인문데이터 - 공공기관 보유 인문데이터 공개 신청 사례를 중심으로」, 『韓國古典硏究』, 57 (2022).

김현, 「디지털 큐레이션: 인공지능 시대의 인문학 연구 방법」 『2023 인문콘텐츠학회 동계학술대회』 (2023).

이재연 외, 『세계 디지털 인문학의 현황과 전망』 (커뮤니케이션북스, 2023).

Tobias Blanke, Michael Bryant, & Mark Hedges. "Understanding memories of the Holocaust—A digital humanities approach to testimonies," Digital Scholarship in the Humanities, 35-1 (April. 2020).

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riede,l & Douwe Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," arXiv preprint arXiv:2005.11401 (May. 2020).

Manjavacas, E., & Karsdorp, F, "Computational humanities and ancient documents: A case study," Digital Scholarship in the Humanities, 34-Supplement_1 (December. 2019).

Viktor Mayer-Schönberger, & Kenneth Cukier, *Big data: A revolution that will transform how we live, work*, and think, (Houghton Mifflin, 2013).

Meroño-Peñuela, A., Victor de Boer., & Hoekstra, R, "Linked Data for Digital Humanities: The CultureSampo Project," Journal on Computing and Cultural Heritage, 15-1(Feburary. 2022).

Trevor Owens, "Defining data for humanists: Text, artifact, information or evidence?," Journal of Digital Humanities, 1-1 (2011), 1-3.

Julia Röttgermann, "The Collection of Eighteenth-Century French Novels 1751–1800," Journal of Open Humanities Data, 10-1, (Feburary. 2024).

CHRISTOF SCHÖCH, "Big? smart? clean? messy? Data in the humanities," Journal of Digital Humanities, 2-3 (2013).

Christof Schöch, Maria Hinzmann, Julia Röttgermann, Katharina Dietz, & Anne Klee, "Smart Modelling for Literary History," International Journal of Humanities and

Arts Computing, 16-1 (March. 2022).

Simeone, O., & Okereke, C, "Large language models for digital humanities: A survey," arXiv preprint arXiv:2308.15237(2023).

<Abstract>

The Future of Digital Humanities, Smart Big Data: A Postscript from a Research Fellowship at Trier Center for Digital Humanities in Germany

Kim, ByungJun
(Trier Center for Digital Humanities, Research Fellow)

This paper introduces the concept of "smart big data" as a new paradigm in Digital Humanities (DH) research, based on the author's fellowship experience at the Trier Center for Digital Humanities (TCDH) in Germany. It addresses the current trend of increasing big data utilization in Korean DH research, while highlighting concerns about the lack of consideration for data quality and humanistic characteristics. As an alternative, the paper proposes the concept of smart big data, which combines the scale and diversity of big data with the qualitative excellence of smart data. The study suggests methods for constructing smart big data, including the adoption of advanced machine-readable data formats and the use of artificial intelligence technologies, particularly Large Language Models (LLM) and Retrieval-Augmented Generation (RAG). Furthermore, it emphasizes the importance of international DH research experience while arguing for the need for creative adaptation and development that considers the unique context of Korean humanities.

Keywords Smart Big Data, Computational Literary Studies, TCDH, LLM, RAG