2024 디지털인문학대회 학술대회 자료집

- **일시 :** 2024년 12월 13일(금) 09:30-18:00
- **장소 :** 고려대학교 문과대학(서관) 202호
  (Zoom, YouTube 온라인 중계)
- **주제 :** 디지털 인문학의 넓이와 깊이
- **주관 :** 한국디지털인문학협의회(KADH), 고려대학교 인문사회 디지털융합인재양성사업단
  고려대학교 한자한문연구소, 서울대학교 AI연구원 인공지능 디지털인문학센터
  서울대학교 국사학과 BK21교육연구단, 성균관대학교 국어국문학과 BK21교육연구단,
  KAIST 디지털 인문사회과학센터

KADH KOREAN ASSOCIATION FOR DIGITAL HUMANITIES

http://www.kadh.org/

2024 디지털인문학대회를 개최하며...

아카데미아에서 '디지털 인문학'은 어느새 익숙한 용어가 되었습니다. 각 학문 분야에서 디지털 인문학을 표제로 한 학술 기획을 활발하게 진행하고 있으며, 그 결과로서 제출되는 논문도 매년 증가하고 있습니다. 이뿐만 아니라 대학에서 디지털 인문학 유관 전공을 개설하고, 개별 분과 학문을 망라하는 디지털 인문학 방법론에 관한 세미나와 워크숍이 인기를 얻는 등, 연구와 교육의 현장에서 디지털 인문학은 분명 새로운 바람을 일으키고 있습니다.

동시에 디지털 인문학에 대한 이해와 수용의 격차도 커지고 있습니다. 디지털 인문학에 대한 막연한 호기심부터 학문적 효용에 대한 의구심에 이르기까지, 디지털 인문학의 양적 확산에 따라 실로 다양한 인식의 틀이 공존한다는 사실이 이를 입증합니다. 물론 디지털 인문학의 내포와 외연을 한두 문장으로 정의하기란 난망한 일입니다. 그러나 이러한 인식의 다양성이 이해의 '격차'에 그친다면, 그 또한 문제적이라 하지 않을 수 없습니다. 이는 디지털 인문학이 포괄하는 다양한 양상을 포착하는 것을 넘어, 비판적 성찰과 논쟁을 통한 이론화가 필요한 시점임을 시사합니다.

이번 학술대회에서는 "디지털 인문학의 넓이와 깊이"라는 제목으로 한국의 학술장 내 디지털 인문학의 현 위치와 미래를 전망하고, 이를 통해 연구의 확산과 심화를 도모하고자 합니다. 각 분과 학문 내에서의 디지털 인문학 수용 양상과 이를 경유해 생겨나고 있는 새로운 질문들을 생생하게 담아내고, 대담을 통해 이러한 시도에 대한 수용적-비판적 입장의 목소리를 듣고자 합니다. 또한 디지털 인문학 연구를 지원하는 여러 기관의 상황과 보유 자원을 확인함으로써 연구자-기관의 실질적 네트워크를 구축할 수 있는 자리를 마련하였습니다.

본 학술대회가 디지털 인문학에 대한 개념적 성찰과 실천적 연구방법론이 공유되는 풍성한 담론의 장이 되기를 기대합니다.

한국디지털인문학협의회(KADH)
회장 박진호, 연구기획위원장 허수 올림

학술대회 일정

구 분	시 간	프 로 그 램
등록	09:30-09:50	등록 및 개회 / 온라인 접속
개회	**개회식**	
	09:50-10:00	환영사 / 정병호(고려대학교 DHUSS 사업단장) 개회사 / 박진호(한국디지털인문학협의회 회장)
세션1	**오전 세션: 디지털 인문학 최신 연구 동향** 사회: 이승은(고려대학교 국어국문학과)	
	10:00-10:25	**발표1:** 언어학과 디지털 인문학 / 정성훈(목포대학교 국어국문.문예창작학부)
	10:25-10:50	**발표2:** 양적 분석 방법을 통한 한국 문학 연구의 확장 / 심지섭(인하대학교 한국어문학과)
	10:50-11:15	**발표3:** 전근대 사회사 연구와 양적자료 분석: 전통과 도전 / 백광열(서울대학교 사회발전연구소)
	11:15-11:40	**발표4:** 데이터로 보는 동양고전 연구, 그 현황과 과제 / 서재현(성균관대학교 유학동양한국철학과)
	11:40-11:50	**휴식**
	종합 토론 좌장: 박진호(한국디지털인문학협의회 회장)	
	11:50-12:50	**〔지정토론자〕** **발표1토론:** 도재학(경기대학교 국어국문학과), **발표3토론:** 신은경(고려대학교 사회학과) **발표2토론:** 전성규(성균관대학교 국어국문학과), **발표4토론:** 이상엽(서울대학교 철학과)
점심	12:50-13:50	**점심식사**
학생 발표회	13:50-14:40	**학생 디지털 인문학 연구 포스터 발표 진행:** 6개팀
	14:40-14:50	**휴식**
세션2	**오후 세션: 기관에서의 디지털 인문학** 사회: 이승은(고려대학교 국어국문학과)	
	14:50-15:15	**발표1:** 주요 기관 제공 역사자료 데이터베이스: 특성, 지속가능성, 활용의 방향 **발표자:** 류준범(국사편찬위원회)
	15:15-15:40	**발표2:** 지역 소재 소문헌 자료 아카이브의 구축 및 활용 방안과 시사점 **발표자:** 김사현(한국유교문화진흥원)
	15:40-16:05	**발표3:** 국가유산 디지털 아카이브 구축 현황 및 향후 과제 **발표자:** 이종욱(한국전통문화대학교)
	16:05-16:30	**발표4:** AI시대 국가지식문화자원 데이터 허브, 국립중앙도서관 **발표자:** 김수정(국립중앙도서관)
	16:30-16:40	**휴식**
	16:40-17:40	**대담** 대담자: 유인태(전남대학교 중어중문학과) **※지정 토론자 없이 대담자를 중심으로 발표자 및 청중과 자유롭게 토론 진행.**
총회	**한국디지털인문학협의회(KADH) 2024년 총회**	
	17:40-18:00	폐회사 / 박진호(한국디지털인문학협의회 회장)

차 례

세션1: 디지털 인문학 최신 연구 동향

세션2: 기관에서의 디지털 인문학

세션1

디지털 인문학 최신 연구 동향

○ **(전산)언어학과 디지털인문학**
정성훈(목포대학교 국어국문.문예창작학부)

○ **양적 분석 방법을 통한 한국 문학 연구의 확장: 디지털 인문학의 동향 및 도전 과제**
심지섭(인하대학교 한국어문학과)

○ **전근대 사회사연구와 양적자료분석: 전통과 도전**
백광열(서울대학교 사회발전연구소)

○ **데이터로 보는 동양고전 연구: 그 현황과 과제**
서재현(성균관대학교 유학동양한국철학과)

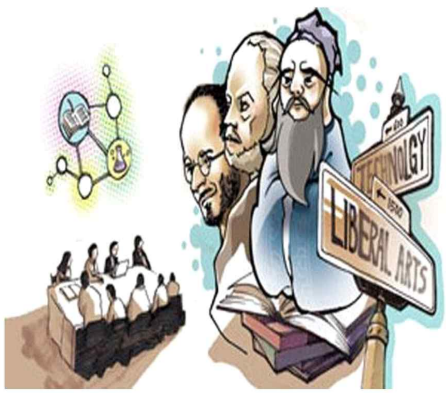
세션1 발표문 1

(전산)언어학과 디지털인문학

(Computational) Linguistics and Digital Humanities

정 성 훈(국립목포대학교)

1. 들어가기
   - 아날로그와 디지털
   - 언어학과 컴퓨터
2. 전산언어학
   - 전산언어학의 역사
   - 전산언어학의 최신 동향
3. 전산언어학과 디지털인문학
   - 언어학과 인문학
   - 디지털인문학과의 만남
4. 요약/정리



1. 들어가기
   - 아날로그와 디지털
   - 언어학과 컴퓨터

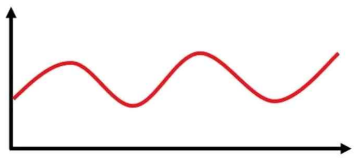
2. 전산언어학
   - 전산언어학의 역사
   - 전산언어학의 최신 동향

3. 전산언어학과 디지털인문학
   - 언어학과 인문학
   - 디지털인문학과의 만남

4. 요약/정리

1. 아날로그와 디지털

(1) 아날로그
- 영어의 'analog'
  : 라틴어 'analogus'에서 유래
  → 비례, 비율
- 아날로그(analog) 정보처리
  : 연속적인(continuous) 정보 처리
  → 다양하게 표현되며, 중간값이 있음
  ☞ 값의 범위가 따로 고정되어 있지 않음
  : 아날로그 신호
  → 진폭, 주파수, 위상으로 표현
  → 기본 신호, 사인파
  예) 인간의 목소리



1. 아날로그와 디지털

(1) 아날로그

- 아날로그 컴퓨터

: 지속적이고 연속적인 정보 처리

→ 기본연산, 덧셈과 미적분

☞ 속도는 빠르나, 정확도가 떨어짐

: 저장·기억장치와 프로그램 불필요

→ 즉각적인 정보처리

: 특수한 용도

→ 장치의 제어, 설계 등의 목적

☞ 아날로그 컴퓨터를 단독으로 사용하는 경우는 거의 없음

: 증폭회로 사용

: 상대적으로 낮은 가격

안티키티라 컴퓨터(기원전 1세기)

안티키티라 컴퓨터(2007, 복원)

1. 아날로그와 디지털

(2) 디지털

- 영어의 'digit'

: 라틴어 'digitus'에서 유래

→ 손가락/발가락, 10진법의 숫자

- 디지털(digital) 정보처리

: 이산적인(binary) 정보 처리

→ 주로 2진법으로 표현되며, 중간값이 없음

☞ 1비트(bit)는 0과 1로 구성된 디지털 정보 단위

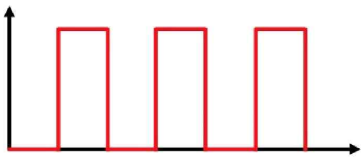
: 디지털 신호

→ 비트율(bps)로 표현

→ 기본신호, 구형파

예) CD 음악

디지털 정보



1. 아날로그와 디지털

(2) 디지털

- 디지털 컴퓨터

: 이산적이고 불연속적인 정보 처리

→ 기본연산, 산술과 논리

☞ 정밀도와 정확도 향상

: 저장·기억장치와 프로그램 필요

→ 정보처리의 용이성

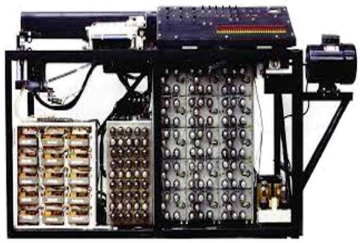
: 일반적인 용도

→ 여러 분야에서 일반적으로 사용

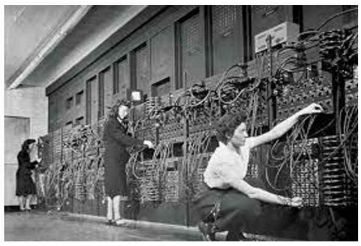
: 논리회로 사용

: 상대적으로 높은 가격

아나타소프-베리 컴퓨터(ABC, 1939년)



에니악 컴퓨터(ENIAC, 1946년)



2. 언어학과 컴퓨터

(1) 언어학

- 언어학(linguistics), 인간의 언어를 체계적이고 과학적으로 연구하는 학문

: 연구목적, 인간이 가지고 있는 (무의식적인) 언어 지식을 체계화·규칙화

: 연구대상, 과거와 현재에 존재한 인간의 모든 언어와 표기 체계 등

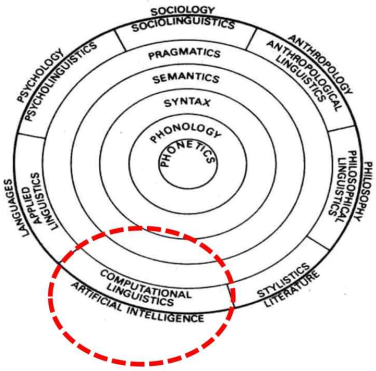
: 하위분야

→ 이론언어학

☞ 음성학, 음운론, 형태론, 통사론, 의미론, 화용론

→ 응용언어학

☞ 전산언어학, 코퍼스언어학, 사회언어학, 심리언어학 등



2. 언어학과 컴퓨터

(2) 전산언어학

- 전산언어학(computational linguistics)

: 컴퓨터로 인간의 언어(natural language)를 처리하기 위한 기초·응용 연구분야

→ 언어학과 컴퓨터공학의 학제적 연구 분야

☞ 언어학적 지식과 자연어처리(NLP) 지식 필요

- 전산언어학의 활용

: 음성 인식과 음성 합성

☞ 외국어교육, 자동차 네비게이션, 의료현장 등에서 활용

: 정보검색과 문서요약

☞ 인터넷, 정보검색시스템 등에서 필요한 텍스트 정보를 검색하고 자동으로 요약

: 기계번역·기계통역

☞ 전산언어학에서 가장 종합적인 응용시스템, 구글번역기·파파고 등에서 활용

2. 언어학과 컴퓨터

(3) 코퍼스언어학

- 코퍼스(corpus)

: 체계적으로 수집되고 구성된 텍스트의 집합

☞ 자연어처리를 전제로 한 기계가독형 데이터

- 코퍼스언어학(corpus linguistics)

: 코퍼스 구축을 위한 연구분야와 코퍼스에 기반한 언어학의 기초·응용 연구분야

→ 관찰된 언어에 대한 실증적인 연구

: 역사, 문학, 사회, 외국어 교육 등 인접 인문·사회과학 분야에서 활용

→ 코퍼스의 언어통계에 대한 중요성

: 전산언어학이나 자연어처리(NLP) 발전에 기여

☞ 코퍼스의 규모와 품질, 언어모델의 성능에 큰 영향

1. 들어가기
   - 아날로그와 디지털
   - 언어학과 컴퓨터

2. 전산언어학
   - 전산언어학의 역사
   - 전산언어학의 최신 동향

3. 전산언어학과 디지털인문학
   - 언어학과 인문학
   - 디지털인문학과의 만남

4. 요약/정리

1. 전산언어학의 역사

(1) 전산언어학

- 전산언어학(computational linguistics)
  : 자연어의 통계적 모형과 논리적 모델을 다루는 학문
  : 컴퓨터를 이용하여 화자의 언어 생성과 청자의 언어 해석을 적합한 유형으로 모형화함(modeling)으로써 자연어로 정보 전달을 재생하는 것

- 전산언어학의 연구 주제
  : 인간의 언어에 대해서 컴퓨터가 수행하는 언어적 처리를 위한 모델링
  → 인식(perception), 의사소통(communication), 지식(knowledge), 계획(planning), 추론(reasoning), 학습(learning) 등의 인공지능 주제 포함
  ☞ 현재, 인공지능과 전산언어학의 경계에 대한 명확한 구분은 없다

- 전산언어학의 목표
  : 인간의 언어에 대한 실제적이고 실용적인 알고리듬과 시스템 구현

1. 전산언어학의 역사

(2) 1940~50년대의 전산언어학

- 전산언어학의 시작
  : 기계번역에서 출발
  → Weaver(1949)의 번역에 관한 제안서 (Jones, 2001)

※ Weaver(1949) 이전에도 기계번역에 대해서 진행 중인 프로젝트가 있었다는 연구도 있음

- 번역에 대한 Weaver(1949)의 접근법
  : 통계적 의미 속성 관찰
  : 번역의 기계화
- McCulloch & Pitt(1943)
  → 인공신경망 개념 제안
- Alan Turing(1950), 튜링머신 제안
  → 인공지능의 시작

Weaver(1949)의 memorandum

The attached memorandum on translation from one language to another, and on the possibility of contributing to this process by the use of modern computing devices of very high speed, capacity, and logical flexibility, has been written with one hope only - that it might possibly serve in some small way as a stimulus to someone else, who would have the techniques, the knowledge, and the imagination to do something about it.

I have worried a good deal about the probable naivete of the ideas here presented; but the subject seems to me so important that I am willing to expose my ignorance, hoping that it will be slightly shielded by my intentions.

Warren Weaver
The Rockefeller Foundation
49 West 49th Street
New York 20, New York

1. 전산언어학의 역사

(2) 1940~50년대의 전산언어학

- 기계번역 프로젝트
  : MIT(1951), 기계번역 연구 프로젝트
  → 기계번역의 이론적인 문제 해결에 중점
  → 1955년, 촘스키(Chomsky)도 참여
  ☞ 언어학적 지식(통사적 분석)에 대한 중요성 강조
  : GAT(1952), 러시아어 → 영어 번역 (물리학 분야)
  → 초기, 언어학 이론과 지식 없이 단어와 단어를 1대1 대응
  → 1954년, 사전(6개의 문법규칙/250개의 단어)을 이용한 기계 번역
  ☞ 대량의 코퍼스에서는 적절히 작동하지 않았으나, 실제로 수행된 최초의 기계번역 시스템
- 다트머스 회의(1956), '인공지능(Artificial Intelligence)' 용어 처음 사용
- Rosenblatt(1957), 선형분류를 위한 퍼셉트론(perceptron) 최초 제안
  : 인간의 뉴런(neuron)을 모방한 인공신경망

1. 전산언어학의 역사

(2) 1940~50년대의 전산언어학

- 다트머스 회의(1956), '인공지능의 아버지'


존 맥카시


클로드 새넌


마빈 민스키


나다이엘 로체스터


아서 사무엘


레이 솔로모노프


올리버 셀프리지


앨런뉴얼


허버트 사이먼


트렌처드 모어

1. 전산언어학의 역사

(3) 1960년대의 전산언어학

- 전산언어학의 암흑기

  : ALPAC(1966) 보고서, '기계번역에 대한 미래 가능성은 매우 적다'

  → 기계번역 프로젝트에 대한 지원 중단

  ☞ 규칙 기반 시스템(rule-based systems), 언어의 복잡한 구조와 다양한 예외 처리의 한계

- 인공신경망 연구의 빙하기

  : Minsky & Papert(1969)

  → 퍼셉트론(perceptron)의 한계를 수학적으로 증명

  ☞ 인공지능 연구에서 인공신경망 연구는 관심의 대상에 멀어짐

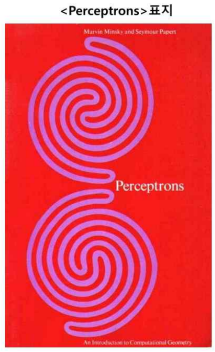
- 전산언어학의 관심 변화

  : 기계번역에서 질의응답 시스템(Q&A system)으로

  → 기업들의 수요

  ☞ 기본적인 고객 질문을 처리할 수 있는 시스템 탐색

<Perceptrons>표지



1. 전산언어학의 역사

(3) 1960년대의 전산언어학

- 전산언어학의 관심 변화

: ELIZA 프로그램(1964-66)

→ MIT의 Weisenbaum이 개발

☞ 최초의 대화형 컴퓨터 프로그램(chatterbot)

→ 컴퓨터 CRT를 이용한 진단 프로그램

☞ 간단한 패턴 매칭 기법 사용

※ 튜링 테스트를 통과한 최초의 프로그램으로 알려져 있음

Weisenbaum과 ELIZA(1964-66)

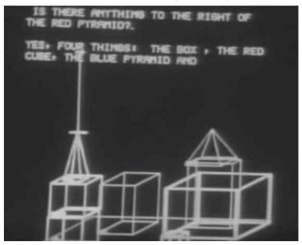
: SHRDLU 프로그램(1968-71)

→ T. Winograd가 개발한 질의응답 시스템

☞ DEC PDP-6 컴퓨터와 DEC graphics terminal 상에서 LISP 언어로 작성

※ '블록세계'에서 물체를 움직이게 하는 시뮬레이션

SHRDLU 프로그램(1968-71)



1. 전산언어학의 역사

(4) 1970년대의 전산언어학

- 전산언어학의 암흑기
- 1960년대의 흐름 지속

: 지식기반(knowledge-based) 전문가 시스템 개발

→ 도메인 내의 질문에 대해 반복적이고 유효한 응답 제공

: 기업들의 수요, 고객서비스 담당자를 대체할 자동 시스템 구축 (Wahlster, 1989)

- 통계학의 발달

: 기업의 R&D, 실용적인 통계 기법에 집중

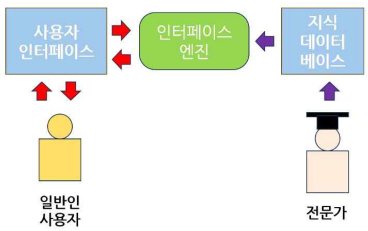
→ 현대적인 실험계획법 및 통계 분석

☞ 기업의 제조품질 및 생산효율 향상

→ 데이터마이닝의 시작

☞ 빅데이터 기술의 근간

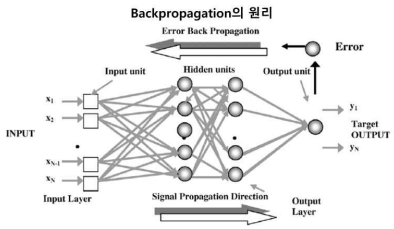
전문가 시스템의 개념도



1. 전산언어학의 역사

(4) 1970년대의 전산언어학

- 역전파(Backpropagation) 알고리즘 등장
  : Paul Werbos(1974) 최초 제안
  : 신경망의 오차를 다시 입력층으로 전파
  → 각 층의 가중치에 대한 기울기 계산
  ☞ 적절한 weight와 bias 학습



- 1971-73년, LUNAR 시스템
  : 아폴로 11호 달탐사선이 반환한 암석의 지질학적 분석에 관한 질의응답 시스템
  → 지질학에 특화된 도메인
  : 데이터베이스 검색 방식
  → 언어학 규칙기반 의미 해석
  → ANT 파서(parser) 사용

LUNAR 시스템의 ANT 파서(1973)

```
SENTENCE:
(WHICH COARSE GRAINED IGNEOUS ROCKS HAVE BEEN ANALYZED FOR COBALT)
PARSING:
RAW CONSES
4,647 SECONDS
PARSINGS:
S REQ
  NP DET WHICHQ
     ADJ COARSE
     ADJ GRAINED
     ADJ IGNEOUS
     N ROCK
     NU PL
     S QUES
       VP PRO SOMETHING
         AUX TNS PRESENT
              PERFECT
         VP V ANALYZE
            NP DET WHR
               N ROCK
               NU PL
            PP PREP FOR
               NP DET HIL
                  N COBALT
                  NU SG
```

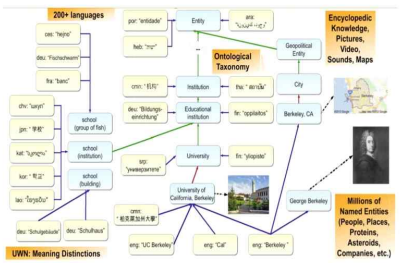

1. 전산언어학의 역사

(5) 1980년대의 전산언어학

- 기계번역으로의 '복귀'
  : 1940-50년대의 연구 집중 검토
  : 실용적이고 쉽게 접근할 수 있는 도구 개발
  → Alvey 프로젝트(1987), 형태소분석기·파서·문법과 어휘집 등의 도구 개발

- WordNet 프로젝트
  : 1985년, George Miller 개발
  → 사전(단어목록)
  → 시소러스(단어 간 관계)
  : 대규모 영어 어휘 데이터베이스
  → 전산언어학과 자연어처리에 유용한 도구
  ※ UWN(Universal WordNet) 등으로 확장

UWN의 지식 그래프



1. 전산언어학의 역사

(5) 1980년대의 전산언어학

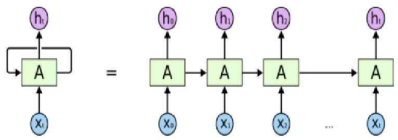
- Geoffrey Hinton의 등장
  : 볼츠만 머신 제안(1985)
  → 신경망 + 대규모 병렬연산
  : 다층 퍼셉트론과 역전파 알고리즘 증명(1986)

※ 2024년 노벨물리학상, 인공신경망을 통한 기계학습 연구

제프리 힌튼(1947~)

- RNN(Recurrent Neural Network) 개념 등장
  : David Rumelhart의 제안(1986)
  → 인식에 대한 형식적 언어 접근
  : 시퀀스 형태의 입력 처리
  → 요소 간 연결이 순환적 구조를 지님
  ☞ 네트워크 안에 루프(recurrent loop)로 정보를 지속시킴

RNN의 기본구조



※ LSTM(Long Short-Term Memory, 1997), GRU(Gated Recurrent Units, 2014) 등에 영향

1. 전산언어학의 역사

(6) 1990년대의 전산언어학

- 전산언어학의 변곡점
  : 전산언어학의 발전기
  → Data-driven 방법 + 통계·확률 기반 모형
  : 전산언어학의 정체기
  → 90년대 후반, 인공신경망 연구의 한계
  ☞ 신경망의 과적합 현상과 신경망 학습을 위한 파라미터 최적화에 대한 이론적 근거 無

- 대규모 데이터와 컴퓨팅 파워
  : 1990년대 인터넷의 보급과 발달
  → 대규모 텍스트 데이터 이용
  ☞ 대량의 데이터에 접근하는 것이 쉬워졌고, 이를 통해 언어 모델은 더 많은 패턴을 학습
  : 컴퓨터 성능의 급격한 향상
  → 복잡한 언어 모델을 학습하는 데 필요한 계산 능력 확보

1. 전산언어학의 역사

(6) 1990년대의 전산언어학

- HMM(Hidden Markov Model)

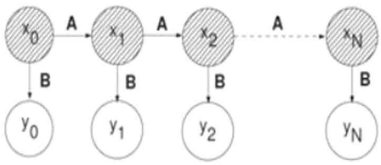
: 통계적 마르코프 모델

→ Andrey Markov(1906) 제안

→ 1970년대 음식인식 분야에서 활용

→ 1990년대 이후 생물정보학, 전산언어학 등에서 활발히 사용

HMM의 기본구조



☞ 다양한 추론 문제들(품사추론, 기계번역 등)과 깊은 관련

- N-gram 모델

: 빈도에 기반한 통계적 언어 모델

→ n개의 연속적인 단어 나열

☞ trade-off , n이 작으면 정확도는 현실의 확률분포에서 멀어지며 n이 크면 희소성의 문제가 심각

→ 통계·확률 기반 모형

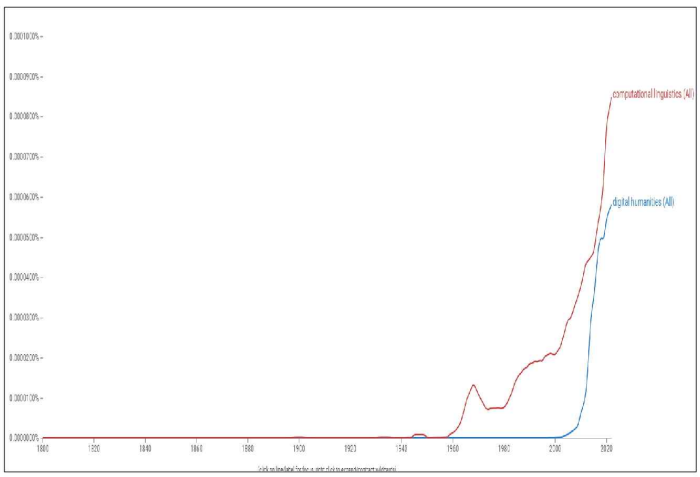
※ 데이터(코퍼스)의 축적 + 컴퓨터 하드웨어의 급속한 발전

1. 전산언어학의 역사

(6) 1990년대의 전산언어학

- 구글의 Books Ngram Viewer

Books Ngram Viewer: Computational Linguistics와 Digital humanities



2. 전산언어학의 최신 동향

(1) 2000년대의 전산언어학

- 인공신경망의 시대

: Geoffrey Hinton, 사전훈련(pre-training) 제안(2006)

→ 비지도학습을 통해 신경망 각 층을 사전 훈련

☞ 전체 네트워크에 대한 미세조정으로 신경망의 학습속도와 효율성 향상

→ 'Deep Learning' 개념 정립

: 다양한 알고리즘 발전

- LSTM 모델

: Hochreiter & Schmidhuber(1997) 아이디어 제안

: RNN의 기울기 소실 문제를 해결하기 위해 고안된 모델

☞ 메모리 셀을 통해 긴 시퀀스 데이터 처리

: Alex Graves 팀(2007), 필기체 인식에서 LSTM 모델로 우승

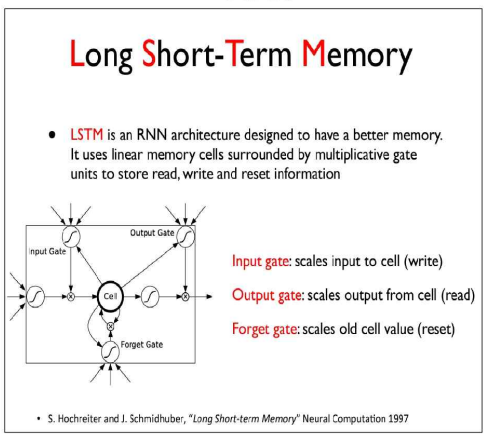
☞ 기계학습과 패턴인식에 대한 인공신경망 모델의 재평가

2. 전산언어학의 최신 동향

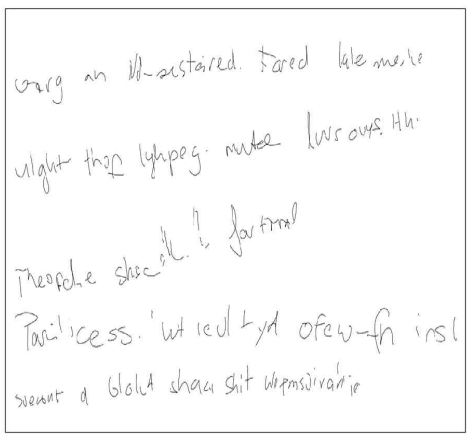
(1) 2000년대의 전산언어학

- 인공신경망의 시대

LSTM의 기본 원리



알렉스 그레이브 팀의 샘플



2. 전산언어학의 최신 동향

(2) 2010년대의 전산언어학

- 전산언어학의 전성기

: GPU 등의 컴퓨터 하드웨어 발전

→ 유사하고 반복적인 연산을 병렬로 처리

☞ 최근 GPGPU(General Purpose computing on GPU)도 등장

: 데이터의 폭발적 증가

→ 스마트폰의 보급과 사물인터넷의 발전

☞ 많은 데이터를 학습한 딥러닝 알고리즘의 정교화

- 단어임베딩(Word Embedding)과 Transformer 모델 등장

☞ 대규모 언어 데이터 처리와 대규모 언어 모델 등에 영향

2. 전산언어학의 최신 동향

(2) 2010년대의 전산언어학

- 단어 임베딩(Word Embedding)

: 각 단어를 고차원 공간에서 실수의 밀집 벡터로 표현

→ 단어의 의미 체계와 문맥 정보 표현

: 벡터(vector), 벡터 공간에서의 수치

→ 벡터의 상대적 위치, 단어 간의 의미론적 관계를 반영

- 빈도 기반 임베딩

: TF-IDF

- 예측 기반 임베딩

: Word2Vec

: GloVe

: fasText

2. 전산언어학의 최신 동향

(2) 2010년대의 전산언어학

- Word2Vec

: 구글의 Tomas Mikolov 외(2013) 제안

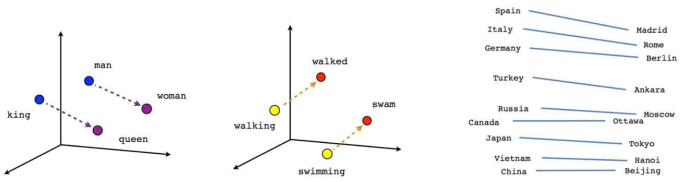
→ 단어의 의미, 텍스트에서 다른 단어들과 함께 나타나는 것으로 추론

: 간단한 인공신경망 모형

→ 분포 가설(distributional hypothesis)에 따른 분산표현

→ 학습과정의 병렬화로 학습 속도 개선

Mikolov et al., (2013) "Linguistic Regularities in Continuous Space Word Representations"



Male-Female　　Verb tense　　Country-Capital

2. 전산언어학의 최신 동향

(2) 2010년대의 전산언어학

- Word2Vec

: CBoW(Continuous Bag of Words)

→ 주변 단어들로 중심 단어를 예측

→ 윈도우(window), 주변단어의 크기

☞ 학습 데이터 셋을 위해 슬라이딩 윈도우 (Sliding Window) 사용

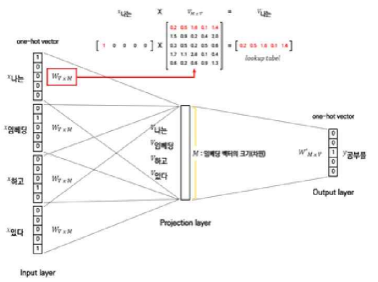
: Skip-Gram

→ 중심 단어로 주변 단어 예측

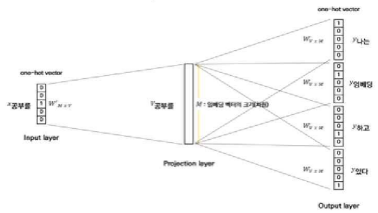
☞ CBoW와 반대

※ 일반적으로 Skip-Gram이 CBoW보다 성능이 좋음

CBoW 알고리즘



Skip Gram 알고리즘



2. 전산언어학의 최신 동향

(2) 2010년대의 전산언어학

- 구글의 Transformer 모델(2017)

: 순차적 데이터 내의 관계를 추적하여 맥락과 의미를 학습하는 양방향 모델

→ 8개의 NVIDIA GPU 사용, 10억개의 단어쌍 데이터 3.5일만에 훈련

☞ 2017년 컴퍼런스에서 기계번역 정확도 공개

: BERT(Bidirectional Encoder Representations from Transformers, 2018)

→ 레이블 없는 방대한 데이터, 사전학습(Pre Training) 수행

☞ 위키피디아(25억 단어) + BooksCorpus(8억 단어)

→ 레이블 있는 다른 작업, 추가적으로 전이학습(Transfer Learing) 진행

☞ 파라미터(parameter) 재조정으로 성능 향상

→ 특정분야(과학, 금융, 법률 등)에서는 성능 향상 필요

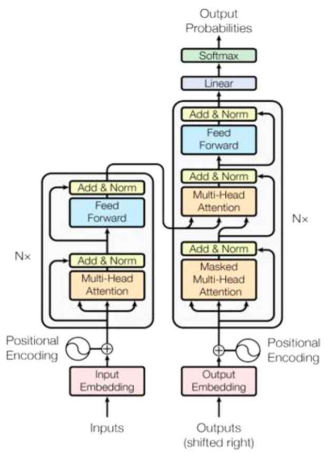
☞ 해당 분야의 데이터를 추가로 수집하고 학습시켜야 할 필요성

2. 전산언어학의 최신 동향

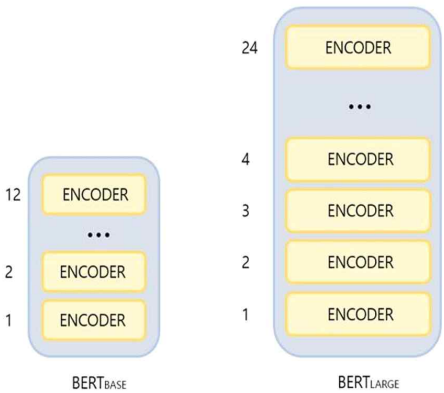
(2) 2010년대의 전산언어학

- Transformer 구조

Transformer의 구조



Bert Base와 Large 모델



2. 전산언어학의 최신 동향

(3) 2020년대의 전산언어학

- 전산언어학의 성숙기

: 생성형 언어모델의 시대

→ 대규모 언어모델(Large Language Model)

☞ 트랜스포머 기반의 신경망을 기반으로 방대한 데이터 학습

※ 전이학습(fine tuning)을 통해 추가 학습 가능

: OpenAI, 3세대 언어모델인 GPT-3(2020)

: OpenAI, 대화형 인공지능 ChatGPT 초기베타(2022.11.30)

: 마이크로소프트, 대화형 인공지능 Bing chat(2023.02.07)

: 구글, 대화형 인공지능 Bard(2023.03.21)

: OpenAI, 대화형 인공지능 ChatGPT 안정화 베타(2023.05.24)

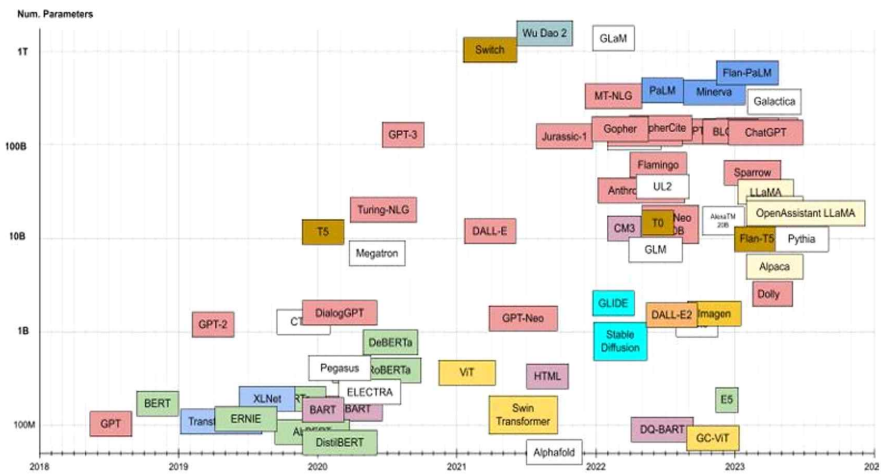
: 구글, Gemini(2024.02.08)

: 메타, Llama 3(2024. 4. 18)

2. 전산언어학의 최신 동향

(3) 2020년대의 전산언어학

- 대규모 언어 모델(LLM)



1. 들어가기
   - 아날로그와 디지털
   - 언어학과 컴퓨터

2. 전산언어학
   - 전산언어학의 역사
   - 전산언어학의 최신 동향

3. 전산언어학과 디지털인문학
   - 언어학과 인문학
   - 디지털인문학과의 만남

4. 요약/정리

1. 언어학과 인문학

(1) 인문학인가? 과학인가?

- 인문학(人文學)

: 인간과 인간 문화에 대한 학문

※ 언어, 문학, 역사, 철학 따위를 연구하는 학문(네이버사전)

: 인문학의 범주

→ 전통적인 범주, 문학·역사·철학

☞ 일반적으로 자연과학에 대비되는 개념

→ 인문학의 프리즘, 예술·사회과학·자연과학 등의 영역 포함

☞ 인간에 대한 통합적 시각 요구

- 인문학의 변화?

: 과학화, 근대 이후 엄밀한 탐구방법과 대상 요구

→ 인문학을 통한 성찰적 사고 필요

1. 언어학과 인문학

(1) 인문학인가? 과학인가?

- 언어학(Linguistics)

: '언어'에 대한 '(자연)과학'적 연구

: 인문학으로서의 언어학

→ 언어, 인간의 가장 중요한 특성

☞ 인간 정신활동의 기반이고 인간 문화의 가장 중요한 부분

→ 언어, 모든 인문학의 기초

☞ 언어를 통하여 '인문학'이라는 학문이 성립

→ 언어학도 '철학'에서 시작

☞ 언어학의 목적도 인간정신을 탐구하는 인문학의 목적과 일치

1. 언어학과 인문학

(1) 인문학인가? 과학인가?

- 언어학(Linguistics)

: '언어'에 대한 '(자연)과학'적 연구

: 과학으로서의 언어학

→ 19세기 역사비교언어학, 역사주의 표방

☞ 자연과학적 연구방법론 등장, 생물학·지질학·물리학 모델 도입

→ 20세기 미국의 구조주의, 객관적으로 관찰될 수 있는 언어만을 대상

☞ Bloomfield(1933), 정밀한 과학적 방법으로 언어 연구 시도

→ 20세기 후반 생성문법, 유한한 수의 규칙으로 무한한 언어 생성(generation)

☞ Chomsky(1957)의 생성문법, 수학적 개념 포함

→ 전산언어학·코퍼스언어학·신경언어학 등의 발달

※ 인문학으로서의 언어학, 과학으로서의 언어학 모두 중요시해야 함

2. 디지털인문학과의 만남

(1) 디지털 시대의 언어학

- 디지털 환경

: 1990년대 이후, 컴퓨터의 발달과 인터넷 환경으로의 변화

: 자연어처리의 중요성

→ 언어, 가장 정교하고 중요한 정보 시스템

☞ 인공지능과 자연어처리, 언어학의 중요성 증가

: 언어와 디지털 환경의 결합

→ SNS 등 디지털 플랫폼에서의 언어

☞ 인간과 사회에 대한 새로운 통찰(Insight) 제공

: 디지털 세상 속의 다양한 언어와 문화

→ 글로벌 시대의 의사 소통에 기여

☞ 지금도 디지털 세상에서 수많은 대화와 교류가 이루어지고 있음

2. 디지털인문학과의 만남

(2) 디지털인문학과 전산언어학

- 디지털인문학(Digital Humanities)

: 디지털 기술과 인문학의 만남

→ 디지털인문학, '인문학 연구를 목적으로 컴퓨터 기술을 활용한 연구분야'

☞ 전산언어학, '언어 연구를 목적으로 컴퓨터 기술을 활용하는 연구분야'

: 디지털인문학 연구의 접근법

→ 자료의 디지털화, 디지털 자료의 변환과 데이터베이스 구축

→ 자료의 분석, 전산언어학의 여러 모델과 기법 적용

→ 새로운 연구 패러다임 창출, 인문학의 새로운 통찰과 연구방법론 개발

: 디지털인문학에 필요한 플랫폼

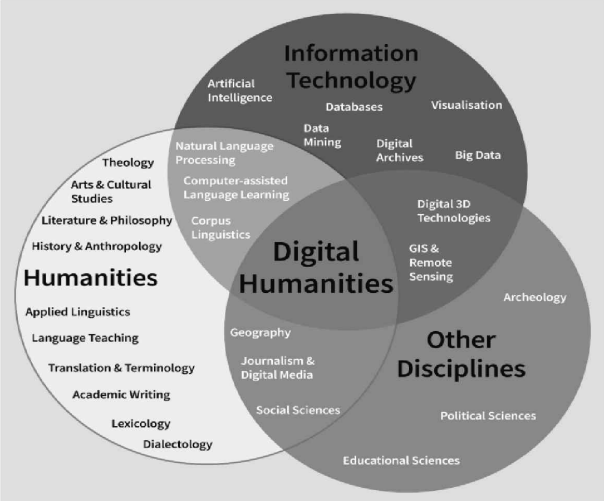
→ 데이터 수집 도구, 분석 도구, 시각화 도구 등

☞ 연구자의 작업 효율성 향상에 기여

2. 디지털인문학과의 만남

(2) 디지털인문학과 전산언어학

- 디지털인문학의 영역



2. 디지털인문학과의 만남

(2) 디지털인문학과 전산언어학

- 디지털인문학과 전산언어학의 만남

: 연구 범위의 확장

→ 대규모 데이터와 분석방법을 통해 지금까지 불가능해 보였던 연구 수행

☞ 분석 도구 또는 플랫폼 제작을 통해 거시적인 연구 수행도 가능

: 연구의 정밀성 향상

→ 컴퓨터 기반의 전산언어학적 분석을 통해 연구의 정확도 제고

☞ 정확도와 함께 분석 속도 향상

: 학제간·연구자간 협업 촉진

→ 학제간 연구로 인문학의 영역 확대

→ 연구자간 협업으로 집단 지성과 새로운 통찰 기대

☞ 디지털인문학의 접근방식, 학제간 연구를 촉진하고 인문학의 영역을 확대하는 데 기여

2. 디지털인문학과의 만남

(2) 디지털인문학과 전산언어학

- 디지털인문학과 전산언어학의 만남

: 사례 1, 한문역사자료 데이터 구축과 활용(최운호·정성훈, 2024)

→ 〈일성록〉, 정조 1776-1778년 자료

· 텍스트 분절(text token/word segmentation) 자동화

Unsupervised token extraction

N-gram + Branching Entropy (BE) + Accessor Variety (AV)

· 사전 및 표현(fixed-expression, Multi-Word Unit) 추출

· 한문 텍스트를 위한 가용한 언어 자원

탐색적 구축 또는 (텍스트의 가공)

재사용 가능하게 가공 (기존 사전 활용)

2. 디지털인문학과의 만남

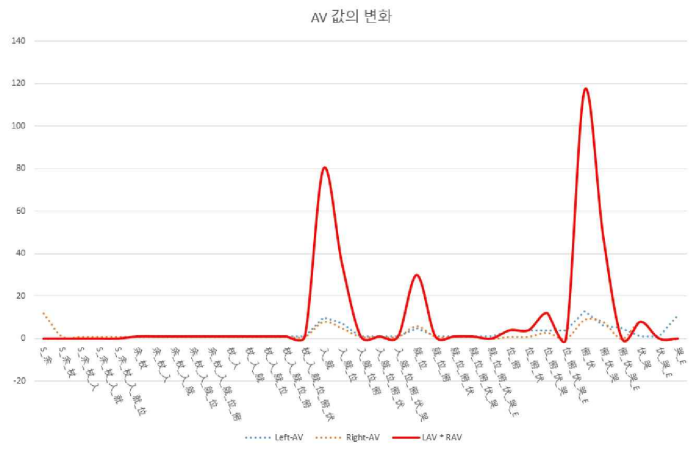
(2) 디지털인문학과 전산언어학

- 디지털인문학과 전산언어학의 만남

: 사례 1, 한문역사자료 데이터 구축과 활용(최운호·정성훈, 2024)

S_余_杖_入_賊_位_俯_伏_哭_E

Strings	Left-AV	Right-AV	LAV * RAV
S_余	0	12	0
S_余_杖	0	1	0
S_余_杖_入	0	1	0
S_余_杖_入_賊	0	1	0
S_余_杖_入_賊_位	0	1	0
余_杖	1	1	1
余_杖_入	1	1	1
余_杖_入_賊	1	1	1
余_杖_入_賊_位	1	1	1
余_杖_入_賊_位_俯	1	1	1
杖_入	1	1	1
杖_入_賊	1	1	1
杖_入_賊_位	1	1	1
杖_入_賊_位_俯	1	1	1
杖_入_賊_位_俯_伏	1	1	1
入_賊	10	8	80
入_賊_位	7	5	35
入_賊_位_俯	1	1	1
入_賊_位_俯_伏	1	1	1
入_賊_位_俯_伏_哭	1	1	1
賊_位	5	6	30
賊_位_俯	1	1	1
賊_位_俯_伏	1	1	1
賊_位_俯_伏_哭	1	1	1
賊_位_俯_伏_哭_E	1	0	0
位_俯	4	1	4
位_俯_伏	4	1	4
位_俯_伏_哭	4	3	12
位_俯_伏_哭_E	4	0	0
俯_伏	13	9	117
俯_伏_哭	6	8	48
俯_伏_哭_E	5	0	0
伏_哭	1	8	8
伏_哭_E	1	0	0
哭_E	11	0	0



2. 디지털인문학과의 만남

(2) 디지털인문학과 전산언어학

- 디지털인문학과 전산언어학의 만남

: 사례 1, 한문역사자료 데이터 구축과 활용(최운호·정성훈, 2024)

	V1	V2	V3	V4	V5	V6
1	0	ISR_1776_03_n_10_01_s01	巳 時	巳_時	巳_時	巳_時
2	1	ISR_1776_03_n_10_01_s02	余 服 衰#服	余 服 衰#服	余 服 衰#服	余 服 衰_服
3	2	ISR_1776_03_n_10_01_s03	宗#親	宗#親	宗#親	宗_親
4	3	ISR_1776_03_n_10_01_s04	文#武#百#官	文#武#百#官	文#武#百#官	文_武_百_官
5	4	ISR_1776_03_n_10_01_s05	皆 衰#服	皆 衰#服	皆 衰#服	皆 衰_服
6	5	ISR_1776_03_n_10_01_s06	內#侍 設 靈 轝 於 靈#座 前	內#侍 設 靈_轝 於 靈#座_前	內#侍 設 靈_轝 於 靈#座 前	內_侍 設 靈_轝 於 靈_座 前
7	6	ISR_1776_03_n_10_01_s07	余 杖 入 就 位 俯#伏 哭	余 杖 入_就 位 俯#伏_哭	余 杖 入_就 位 俯#伏 哭	余 杖 入_就 位 俯_伏 哭
8	7	ISR_1776_03_n_10_01_s08	行#禮 如 儀 訖	行#禮 如_儀 訖	行#禮 如_儀 訖	行_禮 如_儀 訖
9	8	ISR_1776_03_n_10_01_s09	還 廬#次	還 廬#次	還 廬#次	還 廬_次
10	9	ISR_1776_03_n_10_02_s01	領#議#政 金#尙#喆 等 口 達	領#議#政_金#尙#喆 等 口 達	領#議#政 金#尙#喆 等 口 達	領_議_政 金_尙_喆 等 口 達
11	10	ISR_1776_03_n_10_02_s02	請 具 衰#服 出 次	請 具_衰#服 出_次	請 具 衰#服 出_次	請 具 衰_服 出_次
12	11	ISR_1776_03_n_10_02_s03	答 以	答 以	答 以	答 以
13	12	ISR_1776_03_n_10_02_s04	成#服 纔 過	成#服 纔_過	成#服 纔_過	成_服 纔_過
14	13	ISR_1776_03_n_10_02_s05	五 內 如 割	五 內 如 割	五 內 如 割	五 內 如 割
15	14	ISR_1776_03_n_10_02_s06	余 雖 上 不 敢 違 慈#敎	余 雖 上 不_敢 違 慈#敎	余 雖 上 不_敢 違 慈#敎	余 雖 上 不_敢 違 慈_敎
16	15	ISR_1776_03_n_10_02_s07	下 不 能 拂 群#情	下 不_能 拂 群#情	下 不_能 拂 群#情	下 不_能 拂 群_情
17	16	ISR_1776_03_n_10_02_s08	不 得 已 勉 從	不_得 已 勉_從	不_得 已 勉_從	不_得 已 勉_從
18	17	ISR_1776_03_n_10_02_s09	而 今 將 釋 衰 從#吉	而 今_將 釋 衰 從#吉	而 今_將 釋 衰 從#吉	而 今_將 釋 衰 從_吉
19	18	ISR_1776_03_n_10_02_s10	有 不 忍 焉	有_不 忍 焉	有_不 忍 焉	有_不 忍 焉
20	19	ISR_1776_03_n_10_02_s11	崩 隕 之 極	崩_隕 之 極	崩_隕 之 極	崩_隕 之 極
21	20	ISR_1776_03_n_10_02_s12	益 難 堪 抑	益 難_堪 抑	益 難_堪 抑	益 難_堪 抑

한국디지털인문학협의회(KADH)

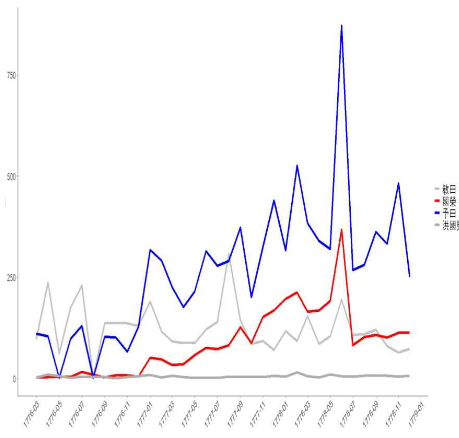
2024 디지털인문학대회

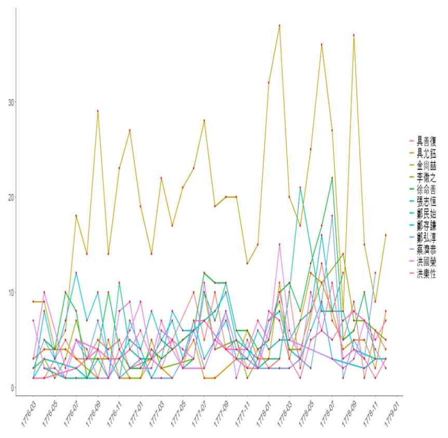
2. 디지털인문학과의 만남

(2) 디지털인문학과 전산언어학

- 디지털인문학과 전산언어학의 만남

: 사례 1, 한문역사자료 데이터 구축과 활용(최운호·정성훈, 2024)





한국디지털인문학협의회(KADH)

2. 디지털인문학과의 만남

(2) 디지털인문학과 전산언어학

- 디지털인문학과 전산언어학의 만남

: 사례 2, 만주어자료 데이터 구축과 활용(정성훈·최운호·도정업, 2023)

→ 〈만문노당〉 태조편 총 81권

· 만주어 전자 사전 구축

preprocessing 사전

어미 사전, 총 45개의 어미 형태 등록

동사/비동사 사전, 어휘 형태를 동사와 비동사로 구분

· 형태소 주석과 태킹(tagging)

문맥규칙 사전, 중의성을 가지는 단어의 좌우 어휘 문맥 의존 규칙 리스트 작성

총 68개의 규칙

· 주석코퍼스 구축

수직코퍼스 형태

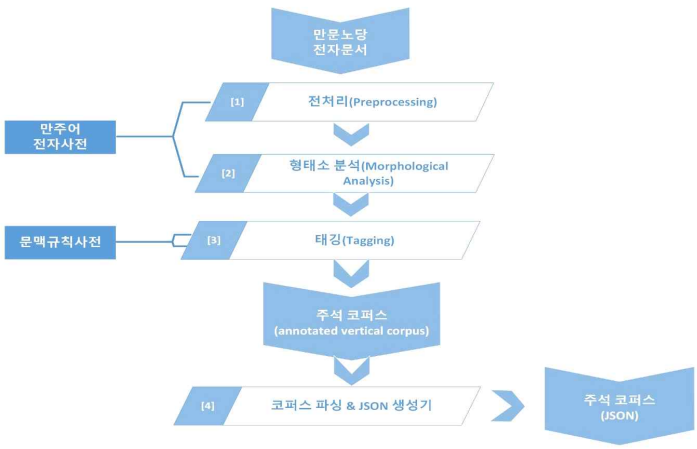
Json 파일로 배포, https://github.com/Kkamakpyel/manwenlaodang

2. 디지털인문학과의 만남

(2) 디지털인문학과 전산언어학

- 디지털인문학과 전산언어학의 만남

: 사례 2, 만주어자료 데이터 구축과 활용(정성훈·최운호·도정업, 2023)

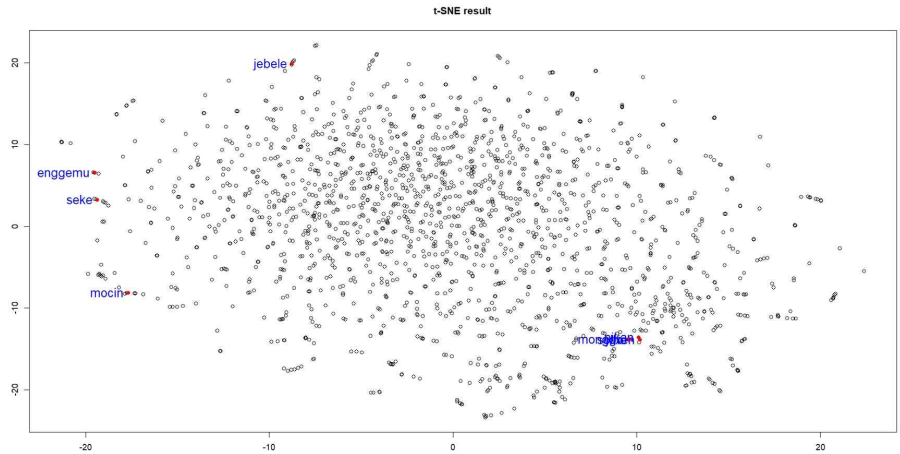


2. 디지털인문학과의 만남

(2) 디지털인문학과 전산언어학

- 디지털인문학과 전산언어학의 만남

: 사례 2, 만주어자료 데이터 구축과 활용(정성훈·최운호·도정업, 2023)



2. 디지털인문학과의 만남

(2) 디지털인문학과 전산언어학

- 디지털인문학과 전산언어학의 만남

: 사례 4, 토필모델링을 활용한 〈詩經〉 텍스트 분석 방법 연구(정성훈·양원석, 2021)

→ 조선시대 〈詩經〉에 대한 주해 자료 분석

· 조선시대 24명의 학자들의 註解 수집

성균관대학교 대동문화연구원의 〈한국경학자료시스템〉 활용

· 〈詩經〉 30편에 대한 註解 토픽모델 분석

· 토픽모델 분석 시각화

히트맵

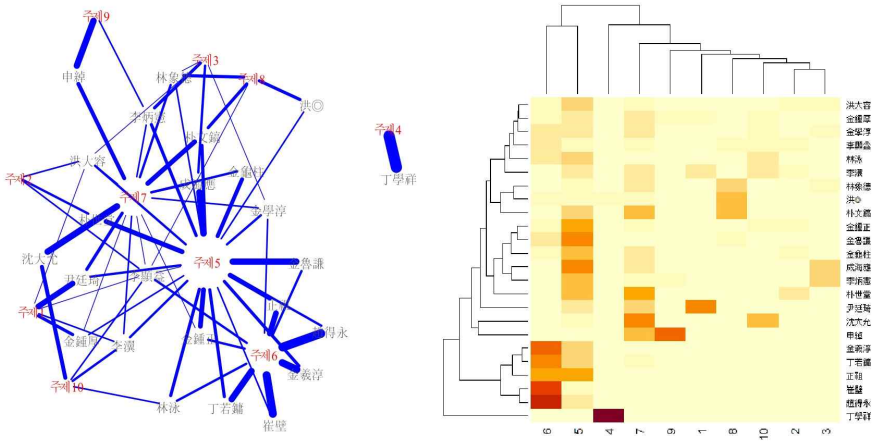
덴드로그램

2. 디지털인문학과의 만남

(2) 디지털인문학과 전산언어학

- 디지털인문학과 전산언어학의 만남

: 사례 4, 토필모델링을 활용한 〈詩經〉 텍스트 분석 방법 연구(정성훈·양원석, 2021)



2. 디지털인문학과의 만남

(2) 디지털인문학과 전산언어학

- 디지털인문학과 전산언어학의 만남

: 사례 5, 군집분석 기법을 이용한 텍스트의 계통 분석(최운호·김동건, 2009)

→ 판소리 〈수궁가〉 계통 분석

· 판소리 구술 기록

· 판소리 변이, '거리(distance)' 개념으로 환원

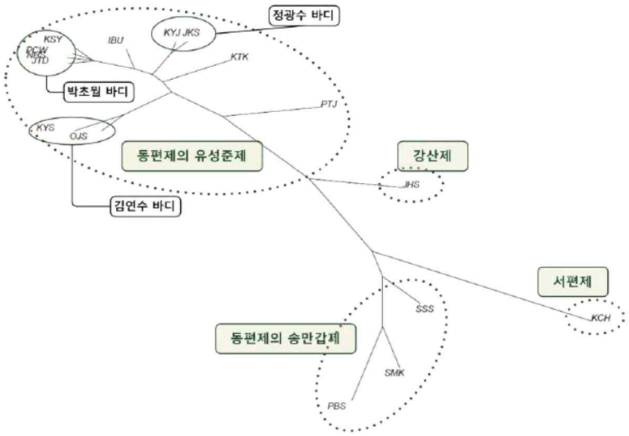
· 군집 분석을 통해 계보 시각화

2. 디지털인문학과의 만남

(2) 디지털인문학과 전산언어학

- 디지털인문학과 전산언어학의 만남

: 사례 5, 군집분석 기법을 이용한 텍스트의 계통 분석(최운호·김동건, 2009)



1. 들어가기
   - 아날로그와 디지털
   - 언어학과 컴퓨터

2. 전산언어학
   - 전산언어학의 역사
   - 전산언어학의 최신 동향

3. 전산언어학과 디지털인문학
   - 언어학과 인문학
   - 디지털인문학과의 교차점

4. 요약/정리


2024 디지털인문학대회
감사합니다!
한국디지털인문학협의회(KADH)

세션1 발표문 2

양적 분석 방법을 통한 한국 문학 연구의 확장 : 디지털 인문학의 동향 및 도전 과제

심지섭(인하대학교 한국어문학과)

〈목 차〉

1. 기술, 양(量), 문학 연구

디지털 인문학 연구는 정의상 다양하고 그 분야도 상이하다. 이 다양성은 기술과 인문학의 접목이라는 지점에 기인하고 있으며 디지털 인문학 연구는 기술 환경의 변화에 발맞춰 변화하며 수행되고 있다. 최근 다양한 분야에서 디지털 인문학 논문들이 제출되고 KADH에서도 2024년 디지털 인문학 전문 학술지 「디지털 인문학(Korean Journal of Digital Humanities)」을 창간하는 등 여러 연구 성과들이 축적되고 있다. 한국 문학 연구에서도 디지털 전환(Digital Turn) 이후, 프랑코 모레티(Franco moretti)의 『멀리서 읽기』[1]와 『그래프,지도,나무』[2]가 번역 출판되고 이 방법론에 대한 다양한 논의들이 이루어지고 있다. 양적인 방법론의 소개와 국내외 연구의 양적 증가가 이루어지고 있는 이 시점에 디지털 인문학과 한국 문학 연구를 다시 점검할 필요가 있다.

디지털 인문학 연구를 이해하는 방식은 기술에 대한 기존 인문학 연구의 인식과 태도를 반영한다. 디지털 인문학의 양적인 성과가 축적되고 있지만 최근까지도 디지털 인문학의 인식이 기술 변화에 반응하는 인문학계의 긍정 또는 부정의 형태로 논의되고 있는 지점도 적지 않아 보인다. 디지털 인문학을 향한 반발에는 가시적이고 명확한 성과에 대한 아쉬움이 있을 수 있

1) 프랑코 모레티, 『멀리서 읽기』, 김용규 역, 현암사, 2021.
2) 프랑코 모레티, 『그래프, 지도, 나무』, 이재연 역, 문학동네, 2020.

겠으나 한편으로는 디지털 기술 환경과의 접목을 꺼려하는 현상과 결부되기도 한다. 그럼에도 기술적 환경은 인간과 분리되는 것이 아니며, 기술과 인문학의 접점에서 순수한 인간이라는 관념은 대체될 필요가 있다. 도나 해러웨이(Donna Haraway)의 주장과 유사한 맥락에서, 어떤 의미로 인간은 이미 혼종적인 사이보그인 까닭이다.

디지털 인문학의 확산은 단순히 인문학에 기술을 적용하려는 시도에 그치지 않는다. 연구자와 연구 토대, 연구 대상 모두가 변화하면 새로운 사고의 도입이 요청된다. 류인태가 논문에서 언급하듯 "정보기술을 학술 활동의 도구로 보는 소극적 시선에서 벗어나 정보기술에 대한 이해를 학술 활동의 일환으로 인식하는 적극적 자세가 필요"한 시점이다.[3] 기술을 향한 열린 태도는 텍스트를 이해하는 방식의 새로운 변화를 의미하고, 이는 새로운 연구 토대 인식과 실천을 불러온다. 이 변화를 이해하고 기존 문학 연구와의 접점에서 새로운 논의 지점을 생성하는 연구는 한국문학 연구의 영토를 확장하고 변화를 생성하는 데 기여할 수 있음을 강조하고자 한다.

디지털 인문학과 인문학 연구의 열린 접근과 그 연결에서 생성해내는 힘들을 주목할 필요가 있다. 본 논문은 디지털 인문학 연구를 단순한 기술 적용이 아닌 연구 환경 토대 자체의 변화 속에서 인식하며, 정보 기술과 기계와 같은 새로운 요소들에 열린 태도로 얽히고자 한다.

특히 이 기술적인 발전은 양적인 문제와 문학의 관계를 이해하고 확장하는 데 밀접한 연관이 있다. 앤드류 파이퍼 (Andrew Piper)는 이러한 맥락을 "오늘날 문자와 숫자를 넘나드는 새로운 번역이 요구"되고 있으며 이는 곧 "텍스트를 양으로 번역하는 것"이라며 설득력 있게 주장한다.[4] 요컨대 시대적 변화에 따른 기술의 변화는 기술 환경의 변화를 의미할 뿐 아니라 문학과 수(數)의 관계를 새롭게 조명한다.

디지털 인문학의 발전에 따라, 최근 근현대 문학의 각종 말뭉치가 여러 차원에서 구축되고 있다. 연구 데이터의 구축은 디지털 인문학 연구의 선결조건임이 분명하다. 다만 또한 자료의 크기와는 다른 차원에서 텍스트를 양적으로 읽는 방법의 정립이 절실하게 필요한 시점이다.

이 논문에서는 이러한 주장을 기반으로 근현대 문학 텍스트의 차원에서 국내외 디지털 인문학의 연구 성과 중 양적인 차원에서 참고할만한 변화들을 개괄적으로나마 소개한다. 또한 그로부터 양적 분석과 한국문학 연구의 접목 방식을 점검하며 몇 가지 도전 과제를 제안하고자 한다.

---

3) 류인태, 「디지털 인문학은 인문학이다」, 『인문논총』 제77권 제3호, 2020, 377쪽.
4) Piper, Andrew. *Enumerations: data and literary study.* University of Chicago Press, 2018. pp.4-5.

2. 해외 양적 연구의 두 가지 흐름

(1) 멀리서 읽기와 거시 관점

해외에서는 양적 분석과 문학 연구를 접목하는 논문들이 상당히 축적되고 있다. 다만 최신 연구 동향을 소개하기 보다는 문학 연구의 양적 변화들에 주목하고자 한다. 양적 분석과 관련한 문학의 중요 분야로 '멀리서 읽기' 또는 '거시적인 읽기'의 관점을 먼저 간소하게 소개하고 언어학과의 접점에서 양적 방법이 문학 연구와 맺는 관계 양상들을 살펴본다.

양적 분석에는 먼저 프랑코 모레티로 대표할 수 있는 멀리서 읽기 연구 경향을 꼽을 수 있다. 멀리서 읽기는 한국 문학 연구에 이미 소개되었고 또한 디지털 인문학과 문학 연구에서 주요하게 논의된 까닭에 디지털 인문학이 겪고 있는 변화 지점에 참고할 만한 부분을 주목하는 방식으로 첨언하고자 한다. 프랑코 모레티의 멀리서 읽기는 세계 문학에 관한 통찰과 '문학의 도살장'과 관련한 연구로 논의되기 시작한다. 한국에서도 이 이론에 관한 번역과 해석적인 평가들이 연구된 바 있다.[5] 멀리서 읽기는 정전 중심의 연구 경향을 비판적으로 사고하며 소위 정전 작가와 정전 작품 위주의 문학 연구를 비판적으로 검토한다. 멀리서 읽기의 또다른 연구자로는 매튜 조커스(Jockers, Matthew)를 대표적으로 들 수 있는데, 그는 일종의 거시 경제처럼 새로운 문학 연구의 양적 차원을 여는 거시 분석 연구(Macroanalysis)를 수행하고자 한다.[6] 이러한 읽기의 유형에는 스탠포드 문학 연구소(Stanford literary lab)에서 수행한 다양한 연구나, 테드 언더우드(Ted underwood)가 수행한 연구들을 또한 포함할 수 있다. 이 연구는 디지털 인문학 연구에 중요한 방법론적 기반이 되고 있다.

다만 이 형식에 관한 이론들이 멀리서 읽기의 방법과 함께 점차 언어 데이터 분석과 컴퓨터 분석의 방향으로 변화하는 지점을 간과해서는 안 된다. 멀리서 읽기는 단일한 개념을 유지하기 보다는 미묘하게 변화한다. 특히 세계 문학 이론이나 진화론적인 프랑코 모레티의 관심사는 점차 말뭉치를 활용한 기술적인 지점을 탐색하는 방향으로 변화를 보인다. 캐서린 보드(Bode, Katherine)는 멀리서 읽기가 과거에는 읽히거나 읽지 않은 작품들을 드러내는 문학의 역사적인 시스템(literary historical systems)에 가까웠다면 이후로 점차 캐릭터화, 줄거리, 극적 형식을 조사하는 문학 연구 시스템(literary study system)으로 초점이 이동하는 미묘한 변화를 보인다고 말한다.[7] 멀리서 읽기의 방법의 중요성과 함께 문학 말뭉치 데이터를 활용한

5) 김용수, 「세계문학과 디지털 인문학 방법론: 한국 학계의 모레티 연구」, 『비평과 이론』 제24권 제3호, 한국비평이론학회, 2019, ; 김지선, 「멀리서 읽기와 디지털 인문학」, 『한국근대문학연구』 제24권 제1호, 2023,

6) Jockers, Matthew L. Macroanalysis: *Digital methods and literary history.* University of Illinois Press, 2013. 또한 이에 관한 주요 연구로 다음 두 책도 추후 소개될 필요가 있다. Piper, Andrew. *op.cit.*; Underwood, Ted. *Distant horizons: digital evidence and literary change.* University of Chicago Press, 2019.

7) Bode, Katherine. "The equivalence of "close" and "distant" reading; or, toward a new object for data-rich literary history." *Modern Language Quarterly* 78.1 2017, pp, 79-80.

다양한 방법들이 멀리서 읽기와 거시 분석에서 수행되고 있다는 점을 주목할 필요가 있다.

이러한 연구 방향은 가까이 읽기(close reading)와 대조되는 의미를 지향하던 것과는 다소 변화하는 지점이며, 설계나 해석의 차원에서 멀리서 읽기는 텍스트, 장르 등의 언어 말뭉치와 교차하면서 기존 논의들과의 간격이 미묘하게 재조정된다. 이는 문학 언어 데이터의 양적인 증가로 문학을 언어적인 구성 텍스트 차원에서 분석하고자 하는 여러 방법들이 멀리서 읽기와 교차하며 영향을 받고 있음을 요인으로 추정해볼 수 있다.

한편으로, 디지털 인문학에서 멀리서 읽기의 중요성이 강조되지만 또한 다른 양적 규모의 연구 방식도 함께 논의되고 있음을 언급할 필요가 있다. 예를 들어 캐서린 보드는 양적인 연구를 존중하면서도 거대한 양적인 규모에 의존하는 연구들의 문제를 언급하며 학술적이고 정확한 데이터 구축의 필요성과 다양한 양적인 크기에 적합한 연구 활용을 제안한다.[8] 호이트 롱(Hoyt Long)과 리차드 진 소(Richard Jean so)같은 학자들은 문학의 모더니즘, 영어 하이쿠들을 중심으로 장르의 패턴을 발견하고자 한다.[9] 이때 활용하는 방법으로 오히려 면밀히 읽기와 역사 비평, 기계 학습을 통한 계산 등을 동원한다는 점을 주목할 수 있다. 멀리서 읽기는 디지털 인문학의 중요한 방법론으로 중요한 이론적인 토대를 구축하지만 또한 디지털 인문학의 전부를 나타내는 용어는 아니며, 기술적 환경과 연구 방향성이 양적인 분석 방법의 차원에서 다양하게 나타나고 있음을 확인할 필요가 있다.

(2) 언어학과 문학 사이 연구

언어학과 문학은 상대적으로 다른 분과로 인식되면서 발전해온 경향이 있다. 조너선 컬러(Jonathan culler)는 이러한 현상에 대해 문학 연구를 '시학'과 '해석학'의 두 방향으로 분류한 바 있다. 그의 개념상 해석학은 "텍스트로부터 출발하여 이를 해석"하는 것으로 그 의미에 집중하는 반면 시학은 "언어학적 방법을 따르는 것"으로 의미가 가능하게 된 방법들을 설명하며 그 언어의 구성과 효과에 집중하는 것이다.[10] 조너선 컬러는 문학 연구의 흐름이 언어학적 분석보다 해석학적 연구에 집중되는 경향이 있음을 언급하기도 하는데, 그의 연구에서처럼 언어학과 문학의 연결고리는 분화된 듯이 보이기도 한다. 그러나 미학적인 언어를 갖춘 문학의 언어 구성은 언어학자들에게 중요한 것이고, 또한 문학자들에게 언어학적인 지식 역시 중요하다. 앞선 멀리서 읽기의 흐름이 언어 구성적인 부분과 결합하여 변화하는 방향과 유사하게,

8) Bode, Katherine. *A world of fiction: Digital collections and the future of literary history*. University of Michigan Press, 2019. 다만, 매튜 조커스의 연구에서도 거대한 양의 연구만을 강조하지는 않는다. 오히려 "거시적인 관점과 미시적인 관점의 혼합"이 필요함을 언급한다. 단, 이때의 미시적인 관점은 정밀한 해석적 읽기와는 다른, 미시정보의 차원을 의미한다. Jockers, Matthew L. Macroanalysis: Digital methods and literary history. University of Illinois Press, 2013. p.26.

9) Long, Hoyt, and Richard Jean So. "Literary pattern recognition: Modernism between close reading and machine learning." *Critical inquiry* 42.2 2016.

10) 조너선 컬러, 『문학 이론』, 조규형 역, 고유서가, 2016, 112-113쪽.

언어로 탐구하는 문학 연구는 시대적 흐름에 따라 충분히 조명되지 못했던 영역, 언어학과 문학 간 연구들을 주목하게 한다. 언어학과 문학 연구의 연결은 본격적으로 해결해야할 중요한 과제다.

문학의 언어를 분석하고 정량적으로 이해하는 연구는 문학보다는 언어학 분야에서 주목받고 있는 듯하다. 여기에는 스타일로메트리, 전산 분석, 코퍼스 언어학, 코퍼스 문체학, 문체학 등 다양한 영역이 매개되어 있으며, 디지털 인문학과 텍스트 연구 차원에서 가장 활발히 논의되고 있는 영역 중 하나라고 볼 수 있다.

여기서도 크게 두 가지로 나눠볼 수 있다. 먼저 언어 자질을 활용해 텍스트적인 특성과 그 효과를 연구하는 유형의 문체학이다. 이 연구의 뿌리는 로만 야콥슨(Roman Jakobson)의 러시아 형식주의와 프라하 학파 등으로 올라갈 수 있다. 텍스트와 문체를 다룬 문학 연구들은 1960년대 이후 번성하고 주목 받았는데, 이후 주목 받는 연구 성과로는 제프리 리치와 믹 쇼트가 쓴 『Style in Fiction』으로, 문학의 언어학적인 자질, 문체와 스타일로 문학 텍스트를 분석한 연구다. 다만 본 논문에서 보다 중시하는 바는 언어학과 문체학에서 적용된 양적인 변화들, 즉 양적인 분석과 측정을 통한 문학의 연구 방법이다. 제프리 리치와 믹 쇼트는 양적인 차원의 연구의 가능성과 한계들을 언급하면서도 새로이 장을 추가하면서 문체 연구에서 말뭉치 활용의 가능성을 추가로 언급한다.[11)]

이러한 문체학의 양적 전환은 말뭉치의 구축과 연관 깊다. 말뭉치는 문체학의 양적 전환과 새로운 가능성들을 생성한다. 코퍼스 언어학자 더글라스 바이버(Douglas Biber)는 연구에서 품사와 장르와의 관계 등을 여러 말뭉치들을 활용해 여러 장르를 언어학적으로 분석하는 작업을 수행한다.[12)] 이외에도 문학과 언어학 사이의 논의로 데이비드 후버(David Hoover), 조나단 컬페퍼 (Jonathan Culpeper), 피셔 스타르케는 (Fisher-Starke) 등 다양한 논의들을 찾을 수 있다. 이러한 연구들은 언어학적인 구성과 문학적인 효과의 관계를 말뭉치로 탐구하는 연구들이라 할 수 있다. 예컨대 데이비드 후버는 헨리 제임스의 스타일을 기간으로 분화한 후, 그 변화를 분석한다.[13)] 조나단 컬페퍼는 로미오와 줄리엣의 등장인물들의 대사를 분리한 후 워드스미스 툴즈를 활용해 키워드를 추출한다. 이를 통해 로미오에게서는 아름다움, 사랑, 부자와 같은 키워드를, 줄리엣에서는 만약과 같은 가정적인 어휘들을 통해 불안과 같은 정서를 읽는다.[14)]

---

11)Leech, Geoffrey, Mick short. *Style in fiction: A linguistic introduction to English fictional prose,* Pearson Longman,2007, p.286.

12) 더글라스 바이버, 수잔 콘라드, 랜디 레펜, 『바이버의 코퍼스 언어학』, 유석훈 김유영 역, 고려대 출판부, 2015.

13) Hoover, David L., Jonathan Culpeper, and Kieran O'Halloran. *Digital literary studies*. Routledge, 2016. pp.90-119.

14) Culpeper, Jonathan. "Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet." *International Journal of Corpus Linguistics* 14.1 2009,

이러한 양적인 연구들은 주관적이고 해석적인 방식의 문학 연구에 문제를 제기하며 상대적으로 객관적인 자료로 보완할 수 있는 방법을 보여준다. 이는 양적으로 언어학과 문학 사이에 교류가 이루어지는 대표적인 방법이다. 다소 차이가 있다면 말뭉치 언어학이 보다 언어학적이고 보편적인 문제를 분석하고자 한다면 말뭉치 문체학은 유사한 방법을 활용하지만 보다 문학적인 텍스트의 특수성을 포함해 질적인 해석과의 관계를 중시한다는 점을 들 수 있다.

이런 작업들은 문학 연구의 실증적인 기반을 확보하는 방법으로 활용되기도 하지만 또한 해석자나 비평가가 감지하기 어려운 미세한 언어적인 흔적과 패턴들을 발견하는 작업들을 포함하는 방식으로 연구 영역 확장을 돕는다. 이는 허인영이 제시했던 코퍼스를 증거로 활용하는 코퍼스 기반의 접근법과 코퍼스가 이끄는 접근법의 틀로 볼 수 있으며[15], 이 두 방법 모두 문학 연구의 언어적인 영역을 넓힐 수 있다.

또한 저자 판별과 전산 문체 분석은 저자 판별과 문체 분석 사이에 놓여있다. 저자 판별은 300년 전부터 다뤄지던 영역이며 서양에서는 특히 컴퓨터를 활용한 연구들은 20세기 중반 이후 본격화되었다.[16] 저자 판별은 모스텔러와 왈레스(Mosteller와 Wallace)의 연구를 중요한 분석으로 본다. 1990년대에는 저자 귀속의 문제는 문체적 특징들, 예를 들어 문장 길이, 어휘 빈도, TTR과 같은 문체적 방법으로 추론되었다. 보다 시일이 흐른 후에는 복잡한 방법을 활용하거나 기계 학습 등을 활용해 저자 판별을 수행하는 경향을 보인다.[17]

전산 문체 분석과 저자 판별은 모두 문체와 언어 구성에 관련 있는 영역이며, 언어학적인 기술들이 적용된다는 공통점이 있다. 다만 스타일로메트리, 저자 판별의 경우 특징적인 면을 부각해서 분류하는 작업에 가깝다면 문체 분석은 보다 포괄적으로, 텍스트의 언어 구성을 살펴보고 의미를 해석하는 방향에 가깝다는 점에서 차이가 있다. 따라서 저자 판별은 분류에 특화되어 있고 문학적인 성찰을 구하기는 어렵다는 한계가 있다. 저자 분석은 전산분석, 기계분석으로 보다 고도화된 저자 분석의 영역으로 접어들고, 말뭉치 문체와 같은 문체학적인 측면에서는 텍스트의 다양한 측정 방법에 관한 논의와 텍스트의 여러 문체적 조건을 확인하려는 논의 등으로 나아가는 경향을 보인다.

이는 작가의 영역뿐 아니라 장르나 성별, 특정 작가나 특정 작가의 시기에 관한 질문 등 다양한 양적 규모로 활용이 가능하다. 이외에도 피터 스톡웰(Peter stockwell)로 대표되는 인지시학은 과학과 문체 분석을 결합하고 있다는 점에서 흥미로운데, 이러한 영역에서는 언어의 구성 성분, 문체, 패턴, 스키마 등이 연구되며 보다 객관적이고 체계적인 분석이 강조된다.

15) 허인영, 「국어사 말뭉치의 활용 현황과 향후 과제」, 『국어사연구』 36, 국어사학회,2023,
16) 김일환, 「저자 판별을 위한 전산 문체론-초기 현대소설을 대상으로」, 『국어국문학』 170, 국어국문학회, 2015, 210-211쪽.
17) Stamatatos, Efstathios. "A survey of modern authorship attribution methods." *Journal of the American Society for information Science and Technology* 60.3, 2009, pp.538-539.

3. 국내 근현대 문학의 양적 텍스트 분석 연구

언어학과 문학, 기술과 문학의 접점에서 작품 스타일 또는 문체에 집중할 필요가 있다. 디지털 인문학 담론 연구들은 물론, 개념사 연구나 네트워크 분석 방법 또는 잡지 분석과 같은 연구들, 그리고 고전 문학 분석도 한국 문학의 경계를 넓히고 있음은 명확하다. 다만 본 논문에서는 거시적인 관점이나 기술적인 방법론을 활용해 언어와 관계 맺는 방식으로 근현대 문학 텍스트를 연구하는 논문들에 주목하고자 한다.

"디지털 인문학 연구보다 디지털 인문학 연구에 대한 연구가 더 많이 발표되고 있다"[18]는 다소 과장 섞은 표현을 한 정서현의 논문처럼, 디지털 인문학의 실질적인 논의를 담은 연구의 수는 상당히 적다. 말뭉치 구축 과정이나 개념의 조작화 방법 등 초입 단계에 있는 양적 분석 방법들을 수행하는 데 따른 어려움이 작용했을 것이라 본다. 이런 여건에서 양적 방법에 입각해 문학 텍스트 분석을 수행한 국내 논문들을 살펴보고자 한다.

먼저 문한별과 김일환의 「김남천 소설의 어휘 사용 양상에 대한 계량적 연구」는 소설을 구성하는 어휘를 양적으로 분석하는 방법을 적용한 시론적인 연구로 김남천을 연구하는 논문들에 "작품 세계를 살펴보는 시각이 주로 사적 맥락이나 이데올로기적 시각 등이 우선 전제"되어 있음을 지적하고 문체적인 요소를 실증적으로 고찰하고자 한다.[19] 방법으로는 명사의 출현 빈도와 예상 빈도를 추출하고 't-점수'로 계산하는 방법을 취한다. 큰 데이터를 활용하지는 않지만 '통계적 유의미성'을 갖는 어휘를 선별하고 공기어를 활용하는 분석적인 방법을 제안했다는 점을 주목할 수 있다. 김일환·문한별·이도길의 「계량적 전산 문체론 시고- 김남천, 이기영, 채만식의 작품을 중심으로」는 "글에 나타나는 개인 간의 구별되는 특성"을 '문체'로 정의하며 계량적 방법으로 김남천, 이기영, 채만식의 작품을 비교한다.[20] 이 논문에서 활용되는 방법은 작가의 어휘량, 어휘의 유형과 사용빈도다. 그러나 이 논문에서 한계로 언급하듯 "단어 차원의 사용 빈도와 유형을 중심으로 연구를 진행할 경우 방법론상의 한계"가 있고 "문장 차원에서 실현되는 문체적인 특성, 수사법"을 파악하기에 어렵다.[21] 다만 이러한 여러 방식의 기초 계량 연구들은 문학의 모든 현상을 설명할 수는 없지만 '문체'들을 구성하는 여러 요소들 중 일부이자 언어 구성 양상의 일단을 탐구한 연구로 의미가 있다. 전은진의 연구(2014,2016)는 신동문과 김수영, 윤동주의 시의 어휘 사용 빈도에 관한 논문으로 어휘 빈도를 분석하거나 네트워크 분석을 접목한 연구다.[22] 이 논문들은 방법론적 차원에서 앞선 논문들과 유사한 지점

18) 정서현, 「문학 연구의 고유성과 디지털 인문학의 가능성:연구 동향 분석 및 겸허한 제안」, 『근대영미소설』 30권 2호, 한국근대영미소설학회, 2023. 129쪽.
19) 문한별·김일환, 「김남천 소설의 어휘 사용 양상에 대한 계량적 연구」, 『한국현대소설학회』 48, 한국현대소설학회, 2011, 49쪽.
20) 김일환·문한별·이도길, 「계량적 전산 문체론 시고 - 김남천, 이기영, 채만식의 작품을 중심으로」,『한말연구』 33, 한말연구학회, 2013, 69쪽.
21) 위의 논문, 70-71쪽.
22) 전은진, 「신동문 시의 어휘 사용 양상과 공기어 네트워크 분석」, 『인문과학연구』, 강원대 인문과학연구소, 2014.; 「김수영 시에 나타난 어휘 연구」, 『청람어문교육』 61, 청람어문교육학회, 2017.

이 있다.

권은의 연구 「수량적 문체론과 기법의 문학사」도 주목할 만한 연구다.[23] 우선 문장 단위로 길이나 TTR 분석을 수행하며 이 방법을 '의식의 흐름' 기법과 관련해 측정한다는 점에서 기존의 어휘 단위의 분석과는 차별점이 있다. 또한 보다 특기할만한 지점은 이 연구에서 사용된 텍스트의 규모다. 한국 근대소설 1,535편, 문장으로는 120만 개에 해당하는 데이터를 활용한다는 점에서 놀라운 수준으로 분석 규모가 확장되었음을 알 수 있다.

또한 박진호의 논문 「두보 시, 한국 근현대 소설, 국어사 자료에 대한 텍스트 마이닝 시론」 역시 흥미로운 관점을 제시한다. 그는 "문학 작품의 문체를 주관적이고" "모호한 개념"으로 파악하는 관행에 문제를 제기한다.[24] 그의 지적처럼 기존 문체 연구는 다양한 주장에 따라 조금씩 다른 개념으로 활용되고, 측정하기도 어렵다는 문제가 있었다. 그는 이 논문에서 문체 개념의 조작화를 주장한다. 그는 크게 다섯 가지 방법으로 문체를 측정하고자 한다.[25] 이는 전통적인 연구에서 기본적으로 강조되어왔던 간결체와 만연체, 서사체와 묘사체 등 전통적으로 중요한 문체를 텍스트에서 측정하기 위한 방법이다. 보다 다양한 문체적인 특징을 측정하기 위해, 추후에 보다 다양하게 논의해야 할 필요성이 있지만 개념의 조작화는 문체의 조작화를 본격적으로 논의한다는 점에서 중요한 의미가 있다.

심지섭의 박사논문 「데이터 분석에 기반한 일제강점기 동화와 소년소설 연구」[26]는 일제강점기 동화와 소년소설의 말뭉치를 구축하고 연구를 수행했다. 모든 작품은 아니지만 주요 신문과 어린이 잡지의 텍스트들에서 추출한 800여 편 이상의 작품들을 자료로 구축하고 현대어로 변환했다. 논문은 장르와 작가 연구로 이루어져 있다. 이 연구는 먼저, 정전이나 특정 작가 위주의 연구 대신 다량의 작품을 확보하고 텍스트 연구를 수행했다는 점에서 거시적인 분석과 연구 맥락이 닿아있다. 이 연구로 수백 작품으로 구성된 장르 데이터에서 통계적인 분석을 수행해 장르나 시기 별로 언어 구성 차이를 통계적으로 알아보고자 했다. 이때 말뭉치 언어학적인 분석인 통계 어휘 분석과 언어프로그램 워드스미스 툴즈(Wordsmith tools)를 활용해 긍/부정 키워드로 상대 빈도를 분석하는 방법도 수행했다.[27] 또한 박진호의 문체 개념의 조작화를 통해 동화와 소년소설, 그리고 일부 소설의 문체 분석을 통해 세 장르 간 문체 차이를 비교하고, 문체를 기반으로 1920년부터 1945년까지의 장르의 문체 변화를 측정하기도 했다.

---

23) 권은, 「수량적 문체론과 기법의 문학사」, 『한국근대문학연구』 제24권 제1호, 한국근대문학회, 2023.
24) 박진호, 「두보 시, 한국 근현대 소설, 국어사 자료에 대한 텍스트 마이닝 시론」, 『UNIST-한국연구재단 공동연구 디지털 인문학 워크샵 자료집: 2023디지털 인문학 교육의 현재와 미래』, 한국연구재단 공동연구 네트워크형 디지털 인문학 교육 모델 개발팀 UNIST 디지털 인문학 센터, 2023, 53쪽.
25) 1. 연결어미 대 전성어미 비율 (paratactic 대 hypotactic 비율) 2. 동사 대 형용사 비율 (서사적 문체 대 묘사적 문체) 3. 보조용언 '있-' 비율 (서사적 문체 대 묘사적 문체) 4. 문장 길이(어절 수) (만연체 대 간결체) 5. TTR(type-token ratio)의 변형 (어휘 다양성), 위의 논문, 55쪽.
26) 심지섭, 「데이터 분석에 기반한 일제강점기 동화와 소년소설 연구」, 인하대 박사논문, 2024.
27) 키워드는 통계적인 핵심성을 의미하는 키니스(Keyness)와의 연관 속에서 측정된다. 빈도 상의 특이성을 측정할 수 있다.

후반부는 아동문학의 주요 작가 분석을 수행했다. 작가 분석에도 다양한 측정 방법이 있겠지만 이 연구에서는 구축한 장르 말뭉치의 일반 명사와 작가별 일반명사 구성의 차이를 상대 빈도의 차원에서 키워드 분석하고, 개념의 조작화로 문체 분석을 수행해 작가의 문체 특징을 구체적으로 비교-측정했다. 예를 들어 이주홍과 현덕의 작품은 비평가나 연구자들에게는 인상적으로 짧은 문장이 활용되었다고 추상적으로 평가되던 것에서 나아가 평균적으로 얼마나 짧고 긴지 구체적인 수치로 비교할 수 있었다.

이러한 작업으로 단어 빈도 차원의 논의에서 조금 더 다양한 방법을 수행하는 한편, 문체 개념의 조작화로 장르와 작가의 언어 구성과 그 양상들을 분석하고자 했다. 또한 양적인 차원에서 다량의 작품이란 측면에서는 거시 읽기지만 형태소적인 분석의 차원에서는 미시적인 읽기의 방식이기도 하다는 점을 언급할 필요가 있다. 인간은 암기할 수 없는 방대한 양과 형태소의 미세한 변화들을 체계화하는 이중적인 양적 방법을 활용하고, 문체 분석의 요소에 PCA를 활용해 양적인 비교 분석을 수행했다는 점, 그리고 가능한 한 문학적인 해석과의 긴장 속에서 수행한 연구라는 점 등을 특징으로 볼 수 있다.

또한 기계학습을 통한 저자판별로는 김일환·이도길, 최지명의 논문을 꼽을 수 있다.[28] 이 역시 몇 안 되는 한국 근현대 문학 관련 저자 판별 논문이다. 김일환과 이도길의 분석이 어휘의 구성과 문체적인 특징들을 기반으로 유사성을 판별하고자 했다면 최지명의 논문은 서포트벡터머신(Support Vector Machines)을 통해 저자 미상의 논설 텍스트의 저자를 판별하고자 한다. 방법은 다르지만 두 방법 모두 유효한 방법들이다. 다만 김일환·이도길의 작업이 현대문학 작가를 대상으로 하는 데 반해 최지명의 논문은 문학 텍스트를 대상으로 하지는 않고 있다는 점에서도 차이가 있다. 그럼에도 최지명의 논문이 문학 잡지 『개벽』을 대상으로 하는 데다, 한국어 텍스트의 기계 분석이라는 측면에서 참조점이 된다.[29] 최지명의 석사논문 역시 문학 텍스트를 직접 다루고 있지는 않지만 한국어 텍스트에 기반한 기계 분석, 저자 판별 방법을 사용하는 논문이라는 점, 그리고 기계학습 방법을 사용할 뿐 아니라 언어의 다양한 요소들을 활용하여 SVM, LDA 등 다양한 분석을 수행한다는 점에서 주목할 만한 논문이라 볼 수 있다. 아직 상당히 적은 연구가 수행 됐을 뿐이지만, 저자 귀속과 저자, 장르의 문체 분석 등 다양한 논문들이 해외의 사례처럼 풍부하게 논의되고 의미화될 필요가 있다.

4. 기술과 문학 연구의 얽힘, 그리고 확장

28) 김일환·이도길, 「저자 판별을 위한 전산 문체론: 초기 현대소설을 대상으로」, 『국어국문학』 제 170호, 국어국문학회,2015; 최지명,「기계학습을 이용한 역사 텍스트의 저자판별: 1920년대 개벽 잡지의 논설 텍스트」, 『언어와 정보』 22,한국언어정보학회, 2018.
29) 최지명, 「기계학습 알고리즘을 이용한 한국어 텍스트 저자 판별 : 블로그의 영화 리뷰를 대상으로」, 연세대 석사논문, 2015.

(1) 개념의 재사유와 문학 연구의 참여 가능성

연구 토대와 연구 대상, 연구 방법의 변화는 새로운 사고를 불러오며 기존의 방법들을 재검토하게 한다. 포스트휴머니즘의 '비인간'에 관한 사유가 인간을 재사유하게 하듯이, 기술 환경의 변화와 디지털 인문학의 사유 방식은 문학 텍스트에 대한 관점과 기존의 문학 연구의 관점을 재점검할 수 있는 가능성을 제공한다.

한국 문학 연구의 전체적인 상을 이 글에서 논의하기는 어렵지만 김병준과 천정환의 논의를 참고해볼 수 있다. 이들은 박사학위 논문에서의 경향을 데이터로 살피며 "적어도 '국어국문학과'의 틀 안에서 써진 박사논문 중에는 새로운 경향의 연구는 그 자체로는 소수며, 전통적인 의미의 연구 주제·방법이 대종을 차지"해 왔고, "그중에서도 대가 중심의 작가론 연구는 큰 비중"을 구성하고 있음을 밝힌다.[30] 물론 박사학위의 논문 특성도 적용되어 있고, 문화론, 비교문학, 페미니즘 기반 연구들이 증가하고 있지만 그럼에도 연구 풍토는 여전히 정전 작품과 대가들에 관심이 있음을 알 수 있으며 거시적이거나 텍스트에 기반한 양적 연구들은 다양하게 시도되지 못하고 있다고 볼 수 있다. 이런 측면에서 볼 때, 한국 문학 연구에서 멀리서 읽기, 거시적인 시각, 양적 연구 방법들은 유용한 참조점을 제공한다. 또한 언어학적이고 시학적인 연구, 계량적인 연구, 양적이고 실증적인 증거로 뒷받침되는 거시적인 관점 연구 등 양적 분석 방법들은 한국 문학의 다양한 가능성을 탐구할 수 있게 한다.

디지털 인문학 연구는 그 자체로도 새로운 관점을 상상하게 하지만 또한 그 관점으로 과거의 연구 방법들을 새로운 시각으로 조명하기도 한다. 예컨대 이재연은 멀리서 읽기에서 오래된 문학 사회학적인 연구의 전통들을 연결하는 테드 언더우드의 연구를 참고하며, 멀리서 읽기 연구 방법이 갖는 문학 연구에서의 사회과학적인 연구 방법들을 한국 문학의 연구 맥락과 연결한다.[31] 이러한 시도는 사회과학적인 논의를 한국 문학 연구와 결부하며 기존의 문학 연구의 의미와 맥락을 생성적으로 파악할 수 있게 한다는 점에서 중요하다. 오랜 기간 양적인 분석이 문학 연구에 부분적으로 활용되어왔다는 점을 고려할 때, 여러 방면의 선행 연구들을 디지털 인문학적인 연구의 맥락에서 확장하는 다양한 논의들도 필요하다.

또한 기술적 토대의 변화가 인문학을 다시 고찰하게 하고 새롭게 정립할 때 기존 문학 연구의 개념 역시 재고할 필요가 있다. 문학 연구가 디지털 인문학을 접하면서 새롭게 갱신해야하는 질문에 직면한 개념들이 있다. 이를테면 디지털 인문학은 저자의 개념을 상당히 다른 방식으로 파악한다. 문학적인 논의에서도 저자의 개념은 상당히 다양하기는 하지만, 롤랑 바르트의 '저자의 죽음'이나 강력한 사회구성주의에 기반한 개념 정의에서 저자는 디지털 인문학의 저

30) 김병준·천정환, 「박사학위 논문(2000~2019) 데이터 분석을 통해 본 한국 현대문학 연구의 변화와 전망」, 『상허학보』 60, 상허학회, 2020, 446쪽.
31) 이재연·정유경, 「국문학 내 문화사회학과 멀리서 읽기 - 새로운 검열연구를 위한 길마중」, 『대동문화 연구』 제111호, 성균관대학교 대동문화연구원, 2020, 295-337쪽.

자 개념과 충돌하는 듯이 보이는 지점이 있다. 저자판별, 저자 귀속 연구와 문체는 저자의 언어적인 특징이 있다는 전제로 이루어지기 때문이다. 이를테면 저자 귀속에는 남들과는 다른 어휘적인 사용, 예를 들어 어휘의 빈도나 배치의 차원에서부터 특정 어휘의 반복적이고 무의식적인 사용, 문장의 길이와 같은 문체적인 요소 등이 저자 마다 특정한 특징을 갖는다는 판단이 중요하게 여겨진다. 이런 의미에서 텍스트 역시 일종의 언어적인 지문처럼, 저자의 의식과 무의식, 장르, 성별, 시대에 따라 다양한 흔적을 남긴다는 가정이 전제되어 있다고 볼 수 있다.

장르 개념 역시 불가지론에서 벗어나 언어적인 데이터로 구성된 실험의 장으로 파악된다. 스탠포드 문학 연구소의 장르 실험, 언어학 계열에서 더글라스 바이버의 장르 실험 등은 장르에도 일정한 언어구성적인 특징을 발견하려는 연구의 일환이다. 이러한 연구 방식은 새로운 연구 방향을 제안하는 한편으로, 디지털 인문학에서 장르 개념의 재정의와 적용에 대한 기존 문학 연구의 여러 개입 지점을 의미하기도 한다. 개념의 정의에 따라 데이터 측정과 분석의 방법, 데이터 구축의 환경 등이 상이하게 변화할 수 있기 때문이다. 예컨대 장르를 규정하고 선택하는 방법에 따라 측정 모델이 변화할 수 있다. 또한 마치 지도학습, 비지도학습처럼 장르를 하향식(Top-Down)으로 장르 텍스트의 범위를 지정해두고 분석할 것인지, 상향식(Bottom-Up)으로 텍스트들에서 장르적인 특성들을 도출할 것인지 역시도 다양한 이론적인 탐구가 가능하며, 이 역시 실험 설계와 결과 도출에 상당한 영향을 준다.

스타일, 문체에 관해서도 재정의가 필요하다. 스타일, 문체, 문채 등 다양한 용어로 사용되는 용어들의 의미를 재구성할 필요가 있다. 스타일은 사상의 옷, 작가의 사상 또는 개인의 개성을 담은 글쓰기, 장르나 작가, 작품의 특징, 시점의 활용, 특정 시대와 문예 사조의 형식적 특징 등 폭넓은 스펙트럼으로 사용되고 있다. 말뭉치 문체학, 문체론 등 스타일과 문체 개념이 디지털 인문학에서 핵심적인 연구의 차원에 놓이고 있는 만큼 개념의 재정의가 선행될 필요가 있다. 텍스트의 분석과 해석 차원에서 스타일은 어휘, 구문, 의미론 수준의 언어적 특징을 의미하기도 한다. 여기에 기술적인 요소를 가미할 때 디지털 인문학에서는 어휘는 물론 어휘의 길이, 구두점, 형태소 유형 등 수많은 텍스트 데이터 요소들의 빈도와 배치, 패턴 등 역시 스타일의 대상이 된다. 이러한 스타일 정의는 보다 폭넓은 방식으로 확장하여 장르, 저자, 성별 등 다양한 요소들의 특징과 함께 교차하며 문학 연구에 열려 있다. 이러한 연구 속에서 스타일은 작가, 시대, 성별, 문예 사조 등 다양한 형태의 모든 영향들의 영향 속에 놓여 있는 개념으로 재인식되는 과정에 놓여있는 듯하다.

마지막으로 디지털 인문학 연구에서는 성별/젠더 연구도 중요하다. 이에 관한 여러 연구가 있으나 대표적인 연구로 테드 언더우드의 연구를 들 수 있다.[32] 그는 이 연구에서 1780년부터 최근에 이르기까지 소설에서의 성의 영향력을 측정하기 위해 남성과 여성 캐릭터 관련 어휘적

32) Underwood, Ted. *Distant horizons: digital evidence and literary change.* University of Chicago Press, 2019. pp.111-142.

인 특징을 비롯한 여러 성별 요인에 따른 시대적인 변화를 추적한다. 흥미로운 연구지만 그가 연구에서 언급하듯 이 연구는 이분법적인 성을 설계의 기반으로 하고 있다. 이를 인문학적인 층위에서 다양하게 비판할 수도 있겠지만, 케일린 랜드(Land, Kaylin)는 디지털 인문학 연구를 부정하기보다 인문학적 지식으로 알고리즘 설계와 해석적 작업에 개입하고자 한다. 그는 알고리즘과 해석에서 문학 연구와 비평적인 개입이 필요하며, 이를 통해 텍스트의 언어구성과 수행적이고 문화적인 젠더의 역할을 보다 폭넓게 탐구할 수 있다고 주장한다.[33] 그가 실천적으로 이분법적인 논의의 틀에 문제제기하는 방법을 선택함으로써, 인문학적 지식이 능동적으로 기술과 얽히는 현상을 볼 수 있다.

위와 같이 디지털 인문학의 연구 방법은 기존 인문학 연구에서의 관념을 재사유할 수 있게 하는 한편, 기존 문학 연구의 방법들이 개입할 수 있도록 열려있다는 점을 강조할 필요가 있다.

(2) 가설, 실험, 측정, 평가 : 체계적인 연구 영역의 확장

양적 분석과 문학적인 연구 사이에서는 다양한 긴장 관계가 형성된다. 이 문학 연구의 경계를 다소 넓히고 양적 연구 방법을 능동적으로 수용할 필요가 있다. 먼저 측정의 가치를 재고해야 한다. 앞선 연구 사례들에서 참고할 수 있듯 측정은 양적 연구에서 필수적인 요소다. 측정이 될 때, 보다 객관적인 공통의 수치를 확인하고 양적 분석을 수행할 수 있다. 실질적인 연구에 앞서 현 시점에서 다급히 필요한 것은 박진호의 앞선 연구처럼 무엇을, 어떤 방식으로 측정할 것인지에 대한 탐색적인 연구다.

정성적인 문학 연구자 입장에서 정량적인 문체 연구는 어휘 빈도나 문장의 측정이 텍스트를 단순화한다거나 정량적인 연구가 문체를 충분히 담아내지 못한다고 느끼는 바는 이해할 수 있는 지적이며 또한 참고할 수 있는 부분이 있다. 다만 여기서는 단일문체의 효과만으로 문학적인 효과를 크게 명확하게 드러내기 어렵다는 점을 먼저 언급해야 할 필요가 있다. 보다 다양한 문체적인 비교가 축적될 때 보다 선명한 문체가 드러날 가능성이 높다. 또한 그 방법을 발견함에 앞서 해외 사례에 비해 부족한 선행 연구의 문제 역시 감안해야한다. 이런 점에서 오히려 새로운 방법론의 영역을 탐험하며 문체 연구를 향해 나아갈 필요성이 제기된다. 문체의 복합적인 측정을 추구하는 '앙상블' 개념, 또한 단일 문체적인 효과를 PCA처럼 다양한 문체 요소의 집합으로 문체적인 특성을 발견하고자 하는 노력 등도 이에 해당된다.

다른 비판으로 유의미한 문체 연구의 경우 기존에 이미 '스타일'로 명확히 인식할 수 있을 만한 것을 재확인하는 데 그친다는 지적이 있다. 이는 디지털 인문학의 양적인 연구 방식이 일종의 순환적인 논법에 빠진다는 비판과 닿아 있다. 다만 스타일의 측정은 단순한 반복이 아니라 객관적이고 실증적인 증거의 영토로 연구 영역을 확장할 수 있는 가능성을 제시하기도 하

33) Land, Kaylin. "Predicting author gender using machine learning algorithms: Looking beyond the binary." *Digital Studies/Le champ numérique* 10.1, 2020.

며, 연구자나 비평가의 이론적인 근거가 되기도 한다는 점을 강조할 필요가 있다. 요컨대 가설을 재확인하더라도 그 효과는 다르다. 특히 인상주의적인 평가나 주관적인 평가 지점에서 이견이 발생할 경우, 측정은 논의에 보다 탄탄한 실증적 증거를 제안할 수 있다.

또한 구체적인 측정은 문체를 비교할 수 있게 하는 핵심적인 기반이 된다. 앞서 언급했지만 현덕과 이주홍의 동화는 둘 모두 비평가에게 간결체를 사용한다는 평을 받아왔다. 이 둘의 문장 길이의 평균을 구체적으로 측정하면 현덕이 평균 7.05어절, 이주홍이 8.52어절로 현덕이 이주홍 보다 약 1어절 이상 짧은 문장을 구사한다는 점을 확인할 수 있다.[34)]

또한 비평가의 인상주의적인 판단 또는 무수한 작품을 모두 읽을 수는 없는 한계에서 발생하는 문제, 그리고 인간이 인식하기에 어려운 미시적인 형태소 유형이나 어휘 패턴 등을 양적인 연구들은 탐색할 수 있게 안내한다. 이러한 연구들은 문학에 객관적인 측면을 보완하는 한편, 질적인 연구와 양적인 연구가 서로 영향을 주고 선순환하며 연구를 확장할 수 있는 모델로 드러나기도 한다.[35)]

이런 맥락에서 문학 연구의 다양한 실험, 측정, 가설 나아가 모델에 대한 평가까지 여러 방법으로 문학 연구의 틀을 확장할 필요성이 요청된다. 요컨대 수치로 실험하고 측정하는 계산적인 요소를 보다 능동적으로 받아들일 필요성이 있다. 또한 특정 영역에서는 양적 차원에서 과학적이고 체계적인 연구 방법 구축도 필요하다. 이제 양적인 연구의 초기 단계에서 기존 문학의 연구를 보완하거나 점검하며, 때로는 데이터 주도의 예측 불가능한 요소들을 흥미롭게 탐구하는 연구들이 고유의 가치를 갖고 다양하게 수행될 필요가 있다.

(3) 실질적인 방법의 개발과 공통 지표 탐색

문체, 장르, 작가, 작품, 성별 등의 언어 구성적인 특징을 분석하거나 문학사적인 질문을 양적으로 수행하는 문제와 같이 '문학'을 위한 실질적인 분석 방법에 대한 논의가 필요하다. 데이터로 문학을 읽고 분석하게 되면 수학과 통계처럼 과학적이고 기술적인 영역으로 확장이 필요할 때 실질적인 질문들을 마주하게 된다. 논문으로 설명할 때는 전부 설명하지 못하는 무수한 어려움이 코드 생성과 해석의 과정에서 생성되며 연구자는 이 연구를 수행하며 여러 판단의 순간에 직면해야 한다.

또한 이러한 분석 과정이 선행된 영문학에는 상당한 논의들이 구축되어 있지만 한국어의 형식과는 다른 지점이 많은 까닭에 단순 참고하기 어려운 상황이라는 점도 어려움을 가중하는 요소다. 이러한 실질적인 방법의 개발이 보다 이루어지기 위해서는 수학적이고 기술적인 방법의 개발로 문학 연구의 외연이 확장될 필요가 있다. 해외에서 TTR과 문장 길이의 실효성 논의가

34) 심지섭, 앞의 논문, 141쪽.
35) Piper, Andrew, *op.cit.*, p.10.

상당 수 이루어지고 있듯 한국어를 활용한 언어 구성과 문학 사이의 논의들이 수행되고 인정받을 수 있어야 한다. 또한 코드, 알고리즘, 개념의 조작화 방법 등이 다양하게 논의되고 다양한 문체 요소와 그 측정 방법들에 따른 다양한 논의들이 탄탄하게 수행될수록 연구의 성과가 향상될 것이다. 이러한 공유의 문제는 문체가 단일 요소로 선명하게 드러나지 않기도 하며, 비교 대상과 측정 방식에 따라 상이할 뿐 아니라 복수적으로 확인될 필요가 있는 개념이라는 점에서도, 또한 여러 측정 모델 역시 복수화 될 때 서로의 성능을 비교할 수 있다는 측면에서도 그러하다. 또한 P-Value, 로그라이클리후드 점수 등과 같은 통계와 수치를 문학에 적용하는 방법들을 고민할 필요도 있겠다. 도전해야 할 어려운 문제들이 많지만 이러한 논의들이 활성화될 때 비로소 그 성과들이 탄탄하게 형성될 수 있다.

복잡하고 어려운 문제들이 많지만, 결국 핵심적으로는 작품의 '문학적인' 문체를 측정하는 방법들을 다시 고민할 필요가 있다. 예컨대 해당 작가가 즐겨 사용하는 문체적 효과는 어떤 언어 구성 형식으로 구성되었는가? 또는 단문을 사용하는가, 복문을 사용하는가? 주로 활용하는 구문 구조는 무엇이며, 그 비율은 무엇인가와 같은 질문들을 떠올려볼 수 있겠다. 이런 문학적인 문제의식과 어떻게 측정할 것인가의 문제를 함께 이해할 필요가 있다.

나아가 일종의 공통 지표 개발 역시 필요하다. 이 분야에서 저명한 존 버러우(John Burrows)는 2002년 단어 빈도 기반 Delta 척도를 도입한다.[36] 이후 2006년에는 중간 빈도(Zeta)와 낮은 빈도(Lota)로 저자 여부를 판명하기 위한 연구를 수행하기도 한다.[37] 이는 추후의 연구에서 테스트 된다.[38] 또한 지속적으로 개선이 시도되기도 한다.[39] 이와 같은 방식으로 다양한 지표들을, 방법들을 공유하고 개선하며 언어와 기술, 그리고 문학적인 논의들을 이어나갈 필요가 있다. 여기에는 문학적인 아이디어, 양적인 텍스트에서의 문체 개념, 수학과 통계적인 지식 등의 조건이 함께 필요하다고 볼 수 있다.

5. 문학 연구와 기술의 조화를 향하여

근현대 한국 문학의 데이터가 구축되는 중요한 전환이 일어나고 있다. 상당한 양의 데이터와 국한문혼용체와 같은 문체적인 문제들이 그간 디지털 인문학이 한국 문학 연구에 접목되는 과정에서 실질적으로 맞닥뜨린 어려움이었기에 이러한 데이터 구축은 중요한 일이다. 다만 자료 구축은 탄탄한 기반을 마련하는 조건이 되지만 디지털 텍스트, 다양한 스캐일의 양적 연구가

36)Burrows, John. "'Delta': a measure of stylistic difference and a guide to likely authorship." *Literary and linguistic computing* 17.3 ,2002.
37)Burrows, John. "All the way through: testing for authorship in different frequency strata." *Literary and Linguistic Computing* 22.1 ,2006.
38)Hoover, David L. "Testing Burrows's delta." *Literary and linguistic computing* 19.4,2004.
39)Smith, Peter WH, and William Aldridge. "Improving authorship attribution: optimizing Burrows' Delta method." *Journal of Quantitative Linguistics* 18.1,2011.

자동으로 수행되지는 않는다. 그 안에는 다양한 방법과 확장을 위한 논의, 실질적인 방법과 논문 양식의 변화, 형식의 문제와 같은 문학 저변의 확장 등이 수반되어야 한다. 더 많은 양적 자료들을 축적하고 디지털화하여 연구 저변을 넓혀가면서 또한 캐서린 보드가 언급했듯 '전체'나 모든 데이터에 집착하기 보다는 연구 특성에 맞는 활용들을 염두하고 점차 그 방식들을 탐구할 필요가 있는 시점이다.

모든 한국 문학 연구자들에게 통계와 디지털 인문학적인 방법을, 양적 방법을 적용해야한다는 터무니없는 주장을 하려는 것은 아니다. 인문학 논의는 질적, 양적으로 다양하게 수행되어야 함은 물론이다. 다만 오히려 기술에 끌려가지 않고 능동적이되 비판적인 방식으로 기술과 얽히기 위해서, 한국 문학 연구의 성과 위에서 양적이고 질적인 연구들이 상호 보완하며 영역을 탄탄하게 확장해갈 필요가 있다. 즉, 그 어느 때보다 기술과 삶의 환경이 강력하게 연동되어 변화하는 시대에 그 흐름에 능동적으로 대응하기 위해서는 수와 문학, 양적인 변환에 따른 변화를 지각하고 그 원리나 문학적인 텍스트 분석을 수행하는 방법들을 고심해야 한다.

멀리서 읽기와 가까이 읽기, 언어학과 문학, 양적인 연구와 질적인 연구, 해석과 분석 사이에 문학 연구의 다양한 탐구 가능성이 있다. 중요한 것은 양적 연구의 방향으로 새로운 연구 영토들이 열려 있으며 그 영역들을 채워나갈 때 문학 연구가 보다 풍부하게 논의될 수 있다는 점이다. 이러한 의미에서 기술과 문학 연구의 관계를 고민해야할 필요가 있으며, 한국 문학 연구는 지금껏 이룩해온 다양한 통찰들을 양적 방법과 접목하면서 얽힐 필요가 있다.

참고문헌

저서

더글라스 바이버·수잔 콘라드·랜디 레펜, 『바이버의 코퍼스 언어학』, 유석훈·김유영 역, 고려대 출판부, 2015.

조너선 컬러, 『문학 이론』, 조규형 역, 고유서가, 2016,

프랑코 모레티, 『멀리서 읽기』, 김용규 역, 현암사, 2021.

프랑코 모레티, 『그래프, 지도, 나무』, 이재연 역, 문학동네, 2020.

Hoover, David L., Jonathan Culpeper, and Kieran O'Halloran. *Digital literary studies*. Routledge, 2016.

Jockers, Matthew L. Macroanalysis: *Digital methods and literary history.* University of Illinois Press, 2013.

Leech, Geoffrey, Mick Short, *Style in fiction: A linguistic introduction to English fictional prose.* Pearson Longman,2007.

Piper, Andrew. *Enumerations: data and literary study.* University of Chicago Press, 2018.
Underwood, Ted. *Distant horizons: digital evidence and literary change.* University of Chicago Press, 2019.

논문

권은, 「수량적 문체론과 기법의 문학사」, 『한국근대문학연구』 제24권 제1호, 한국근대문학회, 2023,73-106쪽.
김병준·천정환, 「박사학위 논문(2000~2019) 데이터 분석을 통해 본 한국 현대문학 연구의 변화와 전망」, 『상허학보』 60, 상허학회, 2020, 443-517쪽.
김용수, 「세계문학과 디지털 인문학 방법론: 한국 학계의 모레티 연구」, 『비평과 이론』 제24권 제3호, 한국비평이론학회, 2019, 59-78쪽.
김일환, 「저자 판별을 위한 전산 문체론-초기 현대소설을 대상으로」, 『국어국문학』 170, 국어국문학회, 2015, 207-239쪽.
김일환·문한별, 「김남천 소설의 어휘 사용 양상에 대한 계량적 연구」, 『한국현대소설학회」 48, 한국현대소설학회, 2011, 377-402쪽.
김일환·문한별·이도길, 「계량적 전산 문체론 시고 - 김남천, 이기영, 채만식의 작품을 중심으로」,『한말연구』 33, 한말연구학회, 2013, 69-105쪽.
김일환·이도길, 「저자 판별을 위한 전산 문체론: 초기 현대소설을 대상으로」, 『국어국문학』 제 170호, 국어국문학회,2015, 207-239쪽.
김지선, 「멀리서 읽기와 디지털 인문학」, 『한국근대문학연구』 제24권 제1호, 2023, 43-72쪽.
류인태, 「디지털 인문학은 인문학이다」, 『인문논총』 제77권 제3호, 2020. 365-407쪽.
박진호, 「두보 시, 한국 근현대 소설, 국어사 자료에 대한 텍스트 마이닝 시론」, 『UNIST-한국연구재단 공동연구 디지털 인문학 워크샵 자료집: 2023디지털 인문학 교육의 현재와 미래』, 한국연구재단 공동연구 네트워크형 디지털 인문학 교육 모델 개발팀 UNIST 디지털 인문학 센터, 2023.
심지섭, 「데이터 분석에 기반한 일제강점기 동화와 소년소설 연구」, 인하대 박사논문, 2024.
이재연·정유경, 「국문학 내 문화사회학과 멀리서 읽기 - 새로운 검열연구를 위한 길마중」, 『대동문화 연구』 제111호, 성균관대학교 대동문화연구원, 2020, 295-337쪽.
전은진, 「신동문 시의 어휘 사용 양상과 공기어 네트워크 분석」, 『인문과학연구』 , 강원대 인문과학 연구소, 2014. 175-200쪽.
_____, 「김수영 시에 나타난 어휘 연구」, 『청람어문교육』 61, 청람어문교육학회, 2017. 325-354쪽.
정서현, 「문학 연구의 고유성과 디지털 인문학의 가능성:연구 동향 분석 및 겸허한 제안」, 『근대영미소설』 30권 2호, 한국근대영미소설학회, 2023. 129-157쪽.
최지명, 「기계학습 알고리즘을 이용한 한국어 텍스트 저자 판별 : 블로그의 영화 리뷰를 대상으로」, 연세대 석사논문, 2015.
최지명,「기계학습을 이용한 역사 텍스트의 저자판별: 1920년대 개벽 잡지의 논설 텍스트」, 『언어와 정보』 22,한국언어정보학회, 2018, 91-122쪽.
허인영, 「국어사 말뭉치의 활용 현황과 향후 과제」, 『국어사연구』 36, 국어사학회, 2023,111-143쪽.

Burrows, John. "Delta': a measure of stylistic difference and a guide to likely authorship."

Literary and linguistic computing 17.3 , 2002, pp.267-287.

Burrows, John. "All the way through: testing for authorship in different frequency strata." *Literary and Linguistic Computing* 22.1,2006, pp.27-47.

Bode, Katherine. "The equivalence of “close” and “distant” reading; or, toward a new object for data-rich literary history." *Modern Language Quarterly* 78.1, 2017, pp, 77-106.

Culpeper, Jonathan. "Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare’s Romeo and Juliet." *International Journal of Corpus Linguistics* 14.1, 2009, pp.29-59.

Hoover, David L. "Testing Burrows's delta." *Literary and linguistic computing* 19.4,2004, pp.453-475.

Land, Kaylin. "Predicting author gender using machine learning algorithms: Looking beyond the binary." *Digital Studies/Le champ numérique* 10.1, 2020; https://www.digitalstudies.org/article/id/7358/

Long, Hoyt, and Richard Jean So. "Literary pattern recognition: Modernism between close reading and machine learning." *Critical inquiry* 42.2, 2016, pp.235-267.

Stamatatos, Efstathios. "A survey of modern authorship attribution methods." *Journal of the American Society for information Science and Technology* 60.3, 2009, pp.538-556.

Smith, Peter WH, and William Aldridge. "Improving authorship attribution: optimizing Burrows' Delta method." *Journal of Quantitative Linguistics* 18.1 ,2011, pp.63-88.

세션1 발표문 3

전근대 사회사연구와 양적자료분석: 전통과 도전

백광열(서울대학교 사회발전연구소)

〈목 차〉

1. 들어가며

이 글의 목표는 한국 전근대 사회사 연구 중에서 양적 자료를 다루는 연구의 동향을 살펴보고 향후 과제를 도출하는 것이다. 이러한 작업은 디지털 역사학 및 한국(사회)사의 콘텐츠화라는 관점에서 이미 많이 수행되고 있다.(김인호, 2023 ; 이상국, 2022 ; 노명환, 2021 ; 주성지, 2019) 이 글을 발표하는 필자는 역사사회학과 한국사회사학의 문제의식에서 장기적인 한국사회구조의 변동에 관심을 가져 왔다. 디지털 인문학과 디지털 역사학에 대해 체계적으로 논급할 수 있는 입장은 아니므로 관심을 제한하여, 한국에 있어 전근대 시기 사회의 구조와 장기적인 변동에 관한 역사상을 양적인 자료와 분석방법으로 다룬 연구들을 나름대로 선별해서 그 의미에 대해 생각해 보는 것으로 주어진 과제를 대신하기로 한다.

사회라는 것은 인간의 상호작용이 일어나는 공간을 의미하며 이것은 정치사만으로 제한하기 어려운 '생명력'의 공간이라고 할 수 있다. 이러한 것에는 미시적으로는 가족이나 또래집단, 촌락 등 생활세계도 있고, 거시적으로는 국민국가나 전세계('전지구화', '지구촌') 범위도 있고, 중범위적으로 지역사회나 도시, 기업 등 사회조직도 있다. 집단의 결속방법에 따라 공동체(가족, 민족 등)도 있고, 이익사회도 있다. 인식 방법에 따라서 객관적이고 물질적인 대상들만을 지칭하기도 하고,(경제사회 등) 현상학적인 세계(생활세계, 정신세계)가 관심이 되기도 한다. 이처럼 다양한 집단과 세계 속에서 계급투쟁이나 지배/피지배 관계가 일어나는 현상은 국가라

는 일원적인 대상을 중심으로 해서 그것을 영웅들의 거대서사를 중심으로 파악하는 정치사보다 한 단계 발전한 인식법이라고 할 수 있다. 나아가 사회사는 이후 문화사, 심성사, 미시사 등 다양한 인간 삶의 영역에 대한 역사학적 관심이 등장하는 계기가 되기도 하였다.

주지하듯이, 사회사라고 하면 20세기 서구 역사학에서 정통 정치사 방법의 한계를 비판하며 등장한 역사학의 조류이다.(조지 이거스, 1998) 대표적으로 프랑스의 아날학파를 들 수 있다. 세대별 차이는 있지만 일반적으로 아날학파의 역사론의 특징으로 흔히 거론되는 것은 전체사, 사회사(↔정치사), 구조사(↔사건사), 장기사(국면연속) 등이다. 아날의 사회사에 양적인 연구대상에 대한 관심이 컸다. 1세대 마르크 블로크의 봉건사회 연구부터 그러하지만, 특히 2세대 브로델의 경우 그의 세계사 개론서『물질문명과 자본주의』1권 1장의 제목이 '수의 무게'(즉, 인구)인 것에서 이를 알 수 있다. 또한, 아날 3세대에서는 심성사와 더불어 계량사 연구가 많이 등장하였다. 이처럼 사회사와 양적 분석은 서로 친화성이 있어 보인다. 또한 사회학에 있어서도 뒤르켐의 '사회학주의'는『자살론』에서 보듯이 사회라는 실체에 대해 그것을 보여줄 표상, 지표로서 양적인 현상을 많이 다루었다. 이처럼 사회사 연구가 양적인 대상, 분석과 친화성이 있다면 그것은 아마도 저자의 주관적 파악을 넘어서 전체를 객관적으로 먼저 파악한 이후에 그를 대상으로 분석을 해보고자 하는 관심때문이리라 생각된다. 대상을 양적으로 먼저 확정한 이후 다음 단계의 논의가 진행될 수 있다는 입장인 것이다. 이것은 정치사처럼 국가제도에 명시적으로 존재하는 연구대상이 아니라 인간의 행위와 상호작용 사에에 어렴풋이 존재하는 (하지만 '생명력'으로서 힘을 미치고 있는) 연구대상의 특성과도 관련이 있다고 보인다. 이것이 사건사나 제도사와 다른 '구조사', 그리고 단기 시간이 아니라 장기적인 연속 시간 규정과도 관계가 있으리라 생각된다.

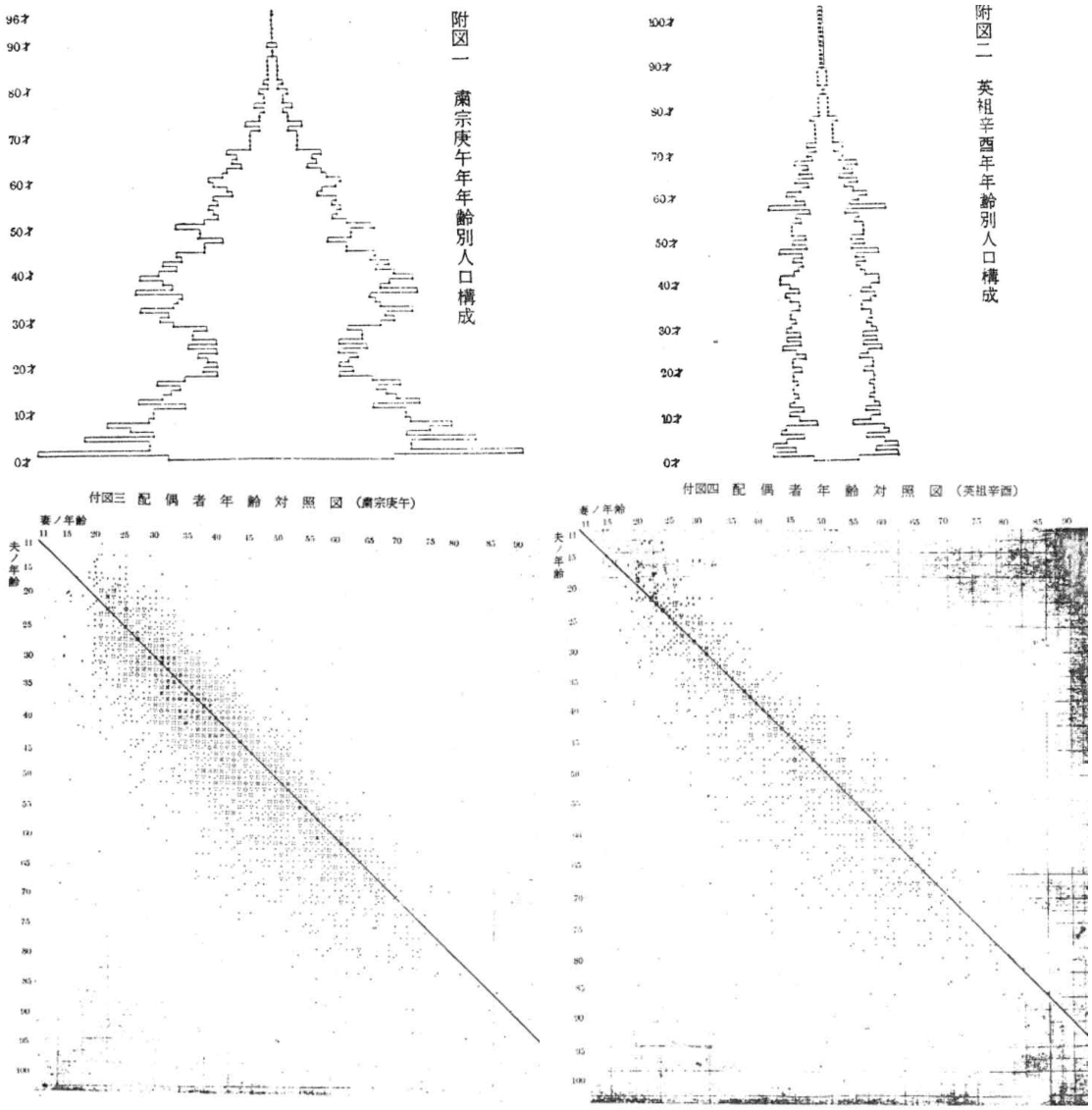
오늘날 사회사 연구에서 수량화와 관련해서 다양한 방법이 시도되고 있다. 이것은 목적과 상정하는 독자가 다르기 때문이다. 기술통계(빈도, 평균 등) 및 해석, 시각화, 대규모 비교, 교차분석, 표준화, 추론검정 등이 그것이다. 각 방법은 각기 특성에 따라 신뢰도, 타당도, 예측력 등을 달리한다. 또한, 전문가적 정보가치와 대중성 중 어느쪽을 중시하느냐도 다르다.

아래에서는 한국 사회사에 있어 양적 분석을 활용하는 연구들을 살펴보는데 있어서 우선 주된 자료인 호적대장, 족보, 방목 등을 중심으로 해서 살펴보되, 구체적인 연구방법을 중심으로 살펴보겠다. 사례의 조사는 국사편찬위원회의 한국사연구휘보 수록 연구 중 2020년 이후의 것에 대해 키워드 검색으로 실시했다. (관련도에 따라 그 이전도 포함)

2. 호적대장과 신분제, 인구 연구

전통적으로도 경제사나 재정사는 그 대상의 성격상 굳이 특화된 수량사가 아니더라도 양적인 자료가 활용되어 왔다. 예를 들어서 '자본주의 맹아론'으로 유명한 김용섭의 양안 연구는 토지

측량단위에 대한 실증 및 이를 통한 부의 양극화에 대한 논의로 한국사회경제사학의 전기를 마련했다고 할 수 있는데,(김용섭, 1960) 이 논의는 양안에 수록된 개인별 토지소유량을 계산해낸 것에 바탕한다. 1) 이후 양안 연구는 경제사학의 중요한 주제가 되었다. 주로 양안 자체의 성격을 논의하는 방향으로 발전해갔다고 생각된다.(이영훈, 1988 ; 미야지마 히로시, 2013)[40]



<그림> 四方博, 1937, 「李朝人口に関する一研究」, 京城帝國大學法学会論集第九冊 『朝鮮社會經濟史研究』, p22, pp31~32

김용섭은 이후 이 논문을 증보하여 양안-호적대장의 교차분석을 통해 신분별 토지소유 양극화

40) 이외에도 수량경제사 분야의 성과들이 많지만 여기서는 다루지 않기로 한다.

실태의 분석으로 나아갔다.(김용섭, 1963) 여기서 알 수 있듯이, 호적대장은 그 이전부터도 신분사 연구에서 활용되고 있었다. 그 시초는 경성제대 법문학부 교수 시가타 히로시의 대구호적(대구부 장적) 연구이다.(四方博, 1937 ; 四方博, 1938) 그는 이를 인구자료로 활용하여 몇 가지 집계와 분석을 시행했고, '신분계급별'로 분할표를 작성하기도 하는 등 양적으로 분석하여 조선후기 사회의 '문란상'을 주장하였다. 반면, 김용섭의 양안-호적대장 연구에서의 역사상은 '역동성'에 초점이 맞춰져 있다는 점은 흥미롭다. 하나의 자료가 맥락의 관심에 따라 상반되게 해석될 수 있는 사례라고 할 것이다.

호적대장은 정부(호조 및 한성부) 주도로 인민들의 호와 구를 매3년마다 파악해서 기록한 것이다. 이것의 역사는 유교경전(『주례』「지관사도」)에 나올만큼 오래되었고, 한국에서도 통일신라시대 이래 작성된 것으로 추정되며, 고려시대 개인별 신고 자료(단자)가 산발적으로 발견되고, 조선시대의 대장(장적)이 17세기 이후부터 일부지역(대구, 단성, 언양, 울산, 상주, 남원, 제주 등)의 것이 남아 있다. 호적대장은 군현별로 작성되는데, 1개 군현 대장은 면>리>통>호>구 단위로 기록되어 있다. 즉, 가장 기초 단위는 호와 구(개인)인데, 구의 기록 기준이 1명의 주호 아래에 호를 구성하는 구성원으로 기록되어 있기 때문에 호가 실질적인 최하단위라고 할 수 있다. 매 구의 주요 속성으로 '직역', '연령' 및 인간관계가 있다.

이후, 호적대장을 이용한 신분사 연구는 계속해서 이어졌다. 특히, 와그너의 양적 연구가 주목할만하다.(와그너, 2007[1974]) 1663년 한성부 북부 호적을 단순 집계 분석('빈도분석')하여 시가타 히로시 및 김용섭 연구에서 주장한 양란후 신분제 문란성·역동성 주장이 비판되고 안정성이 주장되고 있다. 또, 호적, 족보, 방목의 전산화에 관한 시론적 연구가 소개되고 있다.(와그너, 2007) 이것은 60년대 이래 보급되기 시작한 천공테이프 방식의 컴퓨터를 이용해서 자료를 입력하고 그것을 검색하는 방식, 그리고 빈도분석을 하는 방식 등이 주를 이룬다.

이후 연구는 호적대장의 빈도분석에 당대 사회 제도와 관행에 대한 맥락 분석을 더하는 방향으로 진행되었다.(이준구, 1993) 또, 특정 지역사회에 대해 호적-족보를 교차분석하여 서파(庶派)의 규모와 같은 훨씬 더 치밀한 지식에 도달한 연구도 등장하기 시작했다.(임학성, 2000) 그리고 호적대장의 전산화 프로젝트와 이를 기화로 한 대규모 본격 연구들이 등장하였다.(호적대장연구팀, 2003)[41] 빈도분석에 족보 및 지리 자료까지 포함하여 맥락을 분석하고 있다. 프로젝트의 결과물은 엑셀자료로 공개되었다. 자료는 두 지역의 것(단성, 대구)이 입력되었고, 각 지역의 면 단위로 파일이 작성되었다.(단성현 및 대구부 하위의 각 면별로 정리. 즉, 1개 파일은 '모년 모면'을 제목으로 달고 있음.) 내용은 연대별로(17세기부터 19세기까지) 정리된 각 파일 속에 인구와 개별 인물들의 속성을 표로 정리한 것이다.

이후, 호적대장팀이 공개한 DB를 이용한 연구도 등장하였다. 주로 검색을 통한 동일인 찾기가

41) 연구사 및 연구과제의 정리는 권기중(2017) 참조.

주된 방법이다.(이영훈·조영준, 2005 ; 권내현, 2014) 전자는 동일인의 가계를 여러 연대 속에서 찾아서 이어붙이는 방식을 써서 '家' 혹은 '戶'로 표현된 가족단위의 존속 기간을 조사하였다. 후자는 학술논문을 대중서로 펴낸 것인데, 한 인물을 여러 연대 속에서 찾아내어서 그 인물의 신분(실은 '직역' - 호적대장에 신분이라는 말은 안나옴) 상승을 추적한 것이다. 두 연구 모두 DB화를 통한 검색 방법이 활용되고 있다.

호적대장의 호의 성격이 문제가 되었다. 호적대장의 호가 자연호 그 자체를 반영한다는 주장(최재석, 1983 ; 이영훈, 2004)에 대해, 호적대장의 호는 자연호를 반영하는 것이 아니라 지역 행정단위에서 국가에 보고할 때 조정을 거친 것이라는 편제호설(정진영, 2002)이 대립하고 있으며, 편제호설의 주장자들은 논리적으로 맥락 분석을 중시하게 된다.(정진영, 2007)

이후, 호적대장 연구는 지역별 인구와 그 속성의 표를 활용하여 거기에 맥락을 덧붙이는 방식의 연구가 주종을 이루고 있으며, 지금도 이어지고 있다. 이른바 '광무호적'으로 불리는 1896년 신호적법 제정 이후 일제 민적부 이전(1896~1907) 시기의 호적에 대한 DB화와 분석도 진행되고 있다.

최근에는 호적대장을 인구학적으로 활용하는 연구도 등장하기 시작하였다.(손병규 외, 2016) 호적대장의 속성 중에 연령, 호구원 수, 자녀 수 등을 그 나름의 한계를 보정해 가면서 활용하는 것이다. 사실, 호적대장을 이용한 인구파악은 연원이 오래되었다. 즉, 현재까지 거의 유일(유이)한 조선시대 전 기간에 걸친 연도별 인구추산 연구는 호적자료를 이용한 것이다.(권태환·신용하, 1977 ; 미첼, 1989) 권태환·신용하의 연구는 조선 왕조 전 기간에 걸친 호구총수 기록이 존재하므로 이것과 현존 호적대장으로서 가장 신뢰할만한 호적인 한성부 호적을 비교하여 모형을 만들고 이를 이용해서 전 시기 인구를 추정한 것이다.(원리는 미첼 연구도 유사하다) 그러나 호구총수는 1810년대에 알 수 없는 이유로 큰 기록수 감소가 나타나기 때문에 자연히 이로 인해 19세기 인구가 하락하는 것으로 추정되게 되는데 이 점은 신뢰도 관련한 문제를 불러일으킨다. 그러나 현재까지 전 시기를 일관해서 인구를 추정한 연구는 이것이 유일(유이)하다. 최근의 역사인구학 연구는 여전히 호적대장을 이용하되, 총인구 추정은 잠시 보류한 상태에서 출산, 사망, 혼인 등 인구동태의 연구에 관심을 보이고 있다.

참고로, 호적, 족보 등은 유아 인구의 기록 부실 문제가 치명적인 한계이므로 이를 보완하기 위한 시도가 이어지고 있다. 주목할만한 연구로 지배층 인물들의 행장을 수집해서 거기서 평생 출산, 사망수와 출산, 사망율을 구하는 연구가 있다.(박희진, 김두얼) 주지하듯이, 행장에는 해당 인물의 출산 관련 정보가 빠짐 없이 기록되어 있기 때문이다. 다만, 행장 작성 관행이 최상층 인물에 한정되어 있었다는 문제가 있다.

한편, 호적대장의 신분 분석은 모두 직역명들을 분석한 것이라고 할 수 있다. 분명히 이 직역

들이 일종의 신분적인 서열관계를 나타내는 것만은 분명해 보이나 아직 전모가 밝혀졌다고 보기 어렵다. 이 때문에 직역명들을 체계적으로 분석하고자 하는 시도는 위에 언급한 연구(이준구, 1993)를 포함하여 지속적으로 이루어져 왔다. 예를 들어서 호적대장의 A라는 인물이 a라는 직역을 수행하다가 다른 식년에서 b라는 직역을 수행하고 있었다고 한다면, a라는 직역과 b라는 직역은 서로 유사한 위상에 있을 가능성이 높다고 할 수 있을 것이다. 이처럼 인물들을 매개로 한 직역의 서열구조를 파악하는 방법으로서 최근 머신러닝 기법을 이용한 시도가 등장하기도 하였다.(장효민, 2022) 이것은 머신러닝의 단어임베딩 및 계열화(sequence) 기법을 응용한 것인데, 향후에도 이러한 방식은 더 폭넓게 응용될 수 있는 여지가 있다고 생각된다. 이 연구에서는 호적대장의 동일인을 찾는 방법도 머신러닝을 이용함으로써 작업효율성을 크게 향상시키고 있다.

3. 족보와 지배층 연구

족보는 주로 지배층을 수록한 자료이기 때문에 분석 대상이 제한되어 있다. 사회의 기층을 분석하기 어렵다는 한계가 있다. 하지만 세대수와 인물간 관계가 분명하게 기록되어 있기 때문에 이것을 양적으로 표현하는 시도는 오래전부터 있어왔다. 그럼에도 국가권력에 의해 작성된 호적과 달리 민간의 기록이라는 점 등 여러 가지 한계가 지적되어온 자료이기도 하다.(와그너, 2007[1971] ; 송준호, 1987 ; 이기백, 1999; 노명호, 1999) 최근에는 비교적 신뢰성 있는 족보를 이용하여 역사인구학적 연구가 제출되기도 하였다.(손병규 외, 2016)

족보를 양적으로 분석한 연구의 출발점으로 주목할만한 것은 와그너의 연구라고 할 수 있다. 앞서도 언급하였듯이 와그너는 한국사 연구자료의 전산화를 제안한 인물이다. 그는『안동권씨성화보』,『문화유씨가정보』,『만성대동보』등의 전산화를 시도하였다. 물론, 이 방대한 자료들을 다 전산화하지는 못하였으나, 이후의 가능성을 열었다고 할 수 있다. 그의 전산화는 방목자료와 결합되어 결론적으로 여말~조선전기 지배층의 동질성을 주장하는데 근거로 활용되었다. 즉, 지배층내 권문세족-신진사대부 대립 및 조선전기 훈구-사림 대립이라는 역사상에 대해 동일 종합보 내에 상이한 세력이 먼 인친으로 등장하는 사례를 찾아냄으로써 반박하고 있는 것이다. 이 같은 사례는 기존의 사례 중심 연구를 조금더 객관적으로 파악함으로써 기존에 보지 못한 것을 보게 하는 효과를 가진다고 할 수 있다. 이 역시 양적 자료의 '검색' 기능을 활용한 분석이라고 할 수 있다. 이 방법은 이후에도 이어져서 인물 간의 인친 관계에 대한 조금더 일반적인 정량화로 발전하였다.(백광렬, 2017) 즉, 조선 전기 이래의 종합보를 모두 DB화한 다음, 이로부터 인물간의 인친 관계를 지수화하는 것이다. 지수화 근거로는 부-자 관계 및 부-부 관계를 1단계로 하여 두 인물 사이에 이것이 몇 번 이어지는지를 계산하는 방식이다. 이로부터 주요 인물들을 추출하여 이들 간의 연결망과 개인별 위상을 임베딩할 수 있게 되는데 이를 활용하여 지배층 인물 간의 시기별 거리 및 위계 관계를 분석할 수 있다. 이 방법은 머신러닝의 임베딩 기법과 네트워크 분석 기법 등이 활용된 것으로 이후에도 더 확장될

가능성을 가지고 있으나 신뢰성 있는 자료를 확보하고 이를 DB화하는 것의 방대함이 관건으로 남아 있다.

족보를 네트워크로 보는 것이 아니라, 족보의 인물들 각각을 추출해서 속성을 분석하는(일종의 ego-centric 방법도 시도되고 있다. 예를 들어서, 한상우(2014)는 대구서씨세보 등 19세기 벌열가에 기원하는 공신력 있는 족보를 활용하여 세대별 인원수를 '관직' 정보와 교차분석함으로써 지배층 재생산 구조(특히, 부측 영향과 모측영향의 비교)를 분석하였다. 빈도분석 및 분할표분석(명목변수의 위험비 분석)을 활용하였는데, 비교적 신뢰할만한 세대수 및 관직자수를 규정한 것이 비교분석의 밑바탕이 되고 있다고 생각된다.

조선 전기 이전의 인구자료는 거의 전무하다고 할 수 있다. 이때문에 족보 및 역사서, 문집 등을 최대한 검토해서 그 속에서 (주로 지배층인) 인물들을 그 속성(관직, 생몰, 인적 관계 등)과 함께 하나하나 추출하는 작업도 이뤄지고 있다.(이상국, 2002) 특히, 이러한 자료들을 속성간 관계를 분석함으로써 제한적이나마 당대 사회의 일면을 확인해가기도 한다. 여기에 속성들 사이의 상관성을 통계적으로 추정하는 작업이 수행되기도 한다.(예를 들어서 수집한 표본 n개로부터 관직과 출생순위 사이의 상관관계를 추정) 이것은 표본으로 추출된 인물들이 당대 사회의 인구구조로부터 비교적 고르게 추출된 것이라는 가정을 전제한 것이다. 자료의 부족을 극복하기 위해 이 같은 가정은 불가피하다고 할 수 있지만, 표본추출의 가정을 항상 주의하면서 작업을 수행해나갈 필요가 있을 것이라고 생각된다.

선원속보(전주이씨대동보) 등 대규모 족보가 전산화되고 있으며, 이에 따라 이들을 활용한 인구추정 연구도 진행되고 있다.(박경숙 외, 2023) 선원속보는 수록된 남성 인원만 130만에 이르는 대규모 데이터이기 때문에 이로부터 출생율, 사망률 등과 관련 속성 사이의 상관관계의 통계적 추정 등을 수행할 수 있다는 점이 주목된다. 다만, 성인 이전 인구 기록의 누락, 신뢰성 부족, 여성 기록의 부족 등 족보가 가진 일반적 한계를 극복하는 것이 관건이다. 하지만, 호적대장을 제외하고 전무후무한 수백만의 자료규모와 가족단위가 명시적으로 기록된 점 등이 큰 장점이므로 추후의 연구가 지속될 필요가 있다.

4. 방목 등 기타 자료

최근 역사자료의 전산화로 인해 정치사, 사회사에 양적 자료를 활용할 기반이 무르익고 있다. 대표적인 것이 문과방목의 전산화라고 할 수 있다. 이 역시 위에 언급한 와그너(그리고 송준호)가 선구적으로 수행하였다. 문과급제자 15472명 전원을 그들의 속성(합격 연대, 연령, 지역 및 인적 관계)과 더불어 입력한 자료는 큰 반향을 불러 일으켰고, 이후 부가적인 연구도 활성화시켰다. 예를 들어서 박현순(2012)의 경우, 과거 급제자의 시험 유형과 과목을 분석하여 '직부전시' 등 서울의 경화벌열 가문에 일방적으로 유리하게 된 시험 구조를 발견하기도 하였다.

이 역시 방목 DB의 빈도분석에 근거하고 있다.

와그너는 방목 분석을 족보 분석과 교차함으로써 지배층 벌열 가문의 장기지속성을 주장하였다. 이와 유사하게 조선후기 지배층 연구에서 방목 DB가 자주 활용되었다.(미야지마 히로시, 2013 ; 차장섭, 1997 ; 백광렬, 2017 ; 한영우, 2013) 한영우는 방목과 족보(종합보)를 전수 비교하여 방목 속에 비지배층이 다수 포함되어 있음을 실증하였다. 신분사 연구에 있어 함의가 크다고 생각된다. 그러나 이것은 양적분석이 주로 뒷받침하고 있기는 하지만 이러한 전수조사 방식을 개선할 효율적인 방법을 모색할 필요도 있지 않을까 생각된다.

최근에는 방목과『실록』의 데이터를 가공하여 그 속에서 서열화 지수를 개발하는 작업도 이뤄지고 있다. 예를 들어서 문과급제자를 배출한 성관별로 불평등 지수를 시기별로 산출하는 식이다.(백광렬, 2017) 이것을 좀더 확장할 가능성도 충분하다고 생각된다. 불평등의 단위들, 즉 지배층 내의 비교 대상이 되는 각 집단들을 설정하는데 있어 그 단위를 성관으로 비정하는 한계는 극복의 대상이다. 실상과 맞지 않기 때문이다. 위에서 차장섭의 벌열 연구는 8촌까지의 단위를 별도로 구성하는 방법을 활용한 바 있고, 유명 조상을 기준으로 해서 그 후손 숫자를 비교에 활용하는 연구도 있었다.(미야지마 히로시, 2013) 또한, 그러한 집단을 가정하지 않고 전체를 연결망으로 표현하여 그 속에서 군집을 찾고자 한 시도도 있었다.(백광렬, 2017) 그러나 여전히, 지배층의 결집 단위를 찾는 노력은 더 필요하다고 생각된다. 지배층 연구에서 지배층의 결집단위와 관계망, 재생산, 계보와 변동 등이 중요하기 때문이다.

여기서 보듯이 방목은 정치사 자료이기도 하지만, 지배층에 한정하여 그 속에도 존재한 불평등과 서열 및 개인간의 연결을 보여준다는 점에서 사회사의 일부라고 할 수도 있을 것이다.

최근 양적인 자료의 활용범위는『승정원일기』등 전산화의 범위가 넓어짐에 따라 계속 넓혀지고 있다. 일기, 편지 등 각종 고문서의 전산화와 관련 연구도 활성화되고, 지리 자료의 전산화와 GIS의 발전에 따라 이러한 자료를 종합적으로 재구성하여 하나의 전체상을 구성하고 그로부터 인물과 시대를 입체적으로 보여주는 시도도 이어지고 있다.(류인태, 2017) 이러한 연구에서는 인물의 속성 및 연결망의 다차원적인 관계를 보여주기 위해 인맥 네트워크 자료, 지리정보 자료 및 관련 속성들 자료들로 웹을 구축하고 이들을 연결한 시멘틱 웹을 구축한다. 이것은 인간의 삶을 입체적으로 복원하는 인문학 본연의 작업이라고도 할 수 있다. 여기서 가장 주가 되는 것은 역시 시각화라고 생각된다. 여기에 더하여 예컨대 네트워크 내의 중심도, 군집 등의 분석이 이뤄지고 있다. 이러한 시각화와 분석은 확장력이 매우 높다고 할 수 있다. 다만, 시각화 이외의 분석의 신뢰도를 높이기 위해서는 표본의 엄밀한 정의 등 통계적 고려가 더 필요하다고 생각된다.

5. 논의 및 결론

이상, 호적대장, 족보, 방목 등 몇 가지 자료를 예시로 하여 양적인 자료와 분석을 활용하는 사회사 연구의 동향을 살펴보았다. 역사자료로서 사회사 자료는 정치사와 달리 인간 삶의 풍부한 생명력을 표현할 지표가 필요하다. 그리고 그 지표는 주관적인 해석으로 추출하기보다는 객관적 기준에 의해 추출될 필요가 있다. 이것이 정량 자료의 의의라고 할 수 있을 것이다. 정량화에 드는 수고만큼 객관적 근거로서 활용된 자료가 기존의 사례 연구 등 주관적 연구들에서 놓친 점을 발견한다면 그것이 사회사에 대한 중요한 기여가 될 수 있으리라 생각한다. 이를 위해 자료 속에서 양화 데이터를 추출하는 다양한 개념 및 기법이 현재 실험되고 있는 중이라고 할 수 있다. 여기에는 기술통계(빈도, 평균 등) 및 해석, 시각화, 대규모 비교, 교차분석, 표준화, 추론검정 등 다양한 기법이 존재한다.

먼저, 자료가 신뢰성 있게 양화될수록 더 정밀한 분석의 여지가 생겨난다고 할 수 있다. 이 때문에 대표성과 신뢰성 있는 자료의 발굴이 선행되어야 하며 이것은 기존의 역사학 연구의 수준이 충분히 흡수되어야 가능한 일이라고 할 수 있다. 나아가 이로부터 객관적 기준으로 데이터를 추출하는 작업은 생각보다는 훨씬 더 품이 많이 드는 전처리 작업을 포함한다. 또한, 작업 과정에도 여전히 모호한 기준이 발생하게 되고 이를 처리하는데 있어서도 기존의 인문학적 지식이 필요하다.

이 같은 과정을 거쳐서 목적에 맞는 자료가 가공되었을때에야 비로소 신뢰할만한 분석이 수행될 수 있고, 이로부터 신뢰할만한 지식이 도출될 수 있다. 이 과정에 단순한 빈도분석이나 시각화 같은 기법도 활용되겠지만, 보다 더 복잡한 주장을 위해 조금 더 복잡한 모형이 채택될 수도 있다. 이것이 자칫 기법의 복잡화를 가져올 우려도 있다. 실제 문제의 발견과 해결에 어떤 도움을 줄 것인가가 항상 고려될 필요가 있다고 본다. 사회사의 문제의식은 역사적인 '사회' 속에서 규칙성과 의미를 찾아내는 것이기 때문이다. 양적 분석은 단순히 사례에 대한 주관적인 분석만으로는 얻을 수 없는 통찰을 얻기 위한 작업이기 때문에 이후 충분한 맥락에 대한 논의가 더해져야 함도 물론일 것이다.

참고문헌

국사편찬위원회 한국사연구휘보(https://db.history.go.kr/diachronic/level.do?itemId=hb)

권기중, 2017, 「조선후기 호적 연구의 현재와 향후 과제」, 『대동문화연구』 100
권태환·신용하, 1977, 「조선왕조시대 인구추정에 관한 일시론」, 서울대학교 동아문화연구소, 『동아문화』 14, 289-330쪽
권내현, 2014, 『노비에서 양반으로, 그 머나먼 여정 - 어느 노비 가계 2백년의 기록 -』, 역사

비평사
김용섭, 1960,[1995], 「量案의 硏究」, 『史學硏究』7·8 [『증보판 朝鮮後期農業史硏究I』, 지식산업사]
김용섭, 1963, 「조선후기에 있어서의 신분제의 동요와 농지점유-상주양안연구의 일단」, 『사학연구』 15, pp1-50
김인호, 2023, 「디지털 시대의 역사학, 무엇을 할 것인가」, 『역사와 현실』 130
노명호, 1999, 「한국사 연구와 족보」, 『한국사 시민강좌』 24
노명환, 2021, 「디지털 아카이브와 큐레이션에 기초한 디지털 역사학, 공공역사, 트랜스내셔널 역사: 다양성 속의 통일 원리에 기초한 세계의 평화·상생을 향하여」, 『역사문화연구』79
류인태, 2017, 「『덕천원생록』과 시맨틱 웹 DB 연구」, 『2017년도 한국선비문화연구원 학술대회 자료집』
미야지마 히로시, 2013, 『미야지마 히로시, 나의 한국사 공부 - 새로운 한국사의 이해를 찾아서 -』, 너머북스
미첼, 토니, 1989, 「朝鮮時代 人口變動과 經濟史: 人口統計學的 측면을 중심으로」, 『부산사학』 17, 75-107쪽
민경주 외, 2022, 「한국고전종합DB 개선방안 II -인물관계망을 이용한 家系圖를 중심으로-」, 한국고전번역원, 『민족문화』 62
박경숙 외, 2023, 「《선원속보(璿源續譜)》자료에 기초한 조선후기 출산율의 추이와 특성 연구」, 『한국인구학』 46(4)
박현순, 2012, 「조선후기 文科에 나타난 京鄕 간의 불균형 문제 검토」, 『한국문화』 58
백광열, 2017, 「조선후기 '양반지배네트워크'의 성격과 구조변동 : 상층양반의 친족연결망을 중심으로」, 서울대학교 박사학위논문(사회학)
四方博, 1937, 「李朝人口に関する一硏究」, 京城帝國大學法學會論集第9冊 『朝鮮社會法制史硏究 第3』(『사방박저작집』(中), 1976 재수록)
四方博, 1938, 「李朝人口に関する身分階級別的觀察」, 京城帝國大學法學會論集第10冊 『朝鮮經濟の硏究 第3』(『사방박저작집』(中), 1976 재수록)
손병규, 2013, 「시카타 히로시(四方 博)의 조선시대 '인구·가족' 연구에 대한 재검토」, 『한국사학보』 52
손병규 외, 2016, 『한국 역사인구학연구의 가능성』, 성균관대학교출판부
송준호, 1987, 「족보와 보학」, 『朝鮮社會史硏究; 朝鮮社會의 構造와 性格 및 그 變遷에 關한 硏究』, 일조각
와그너, 에드워드, 이훈상·손숙경 역, 2007, 『조선왕조 사회의 성취와 귀속』, 일조각
와그너, 에드워드, 이훈상·손숙경 역, 2007[1971], 「역사 자료로서의 한국 족보」, 와그너(2007), 일조각
와그너, 에드워드, 이훈상·손숙경 역, 2007[1974], 「17세기 조선의 사회계층 - 1663년의 서울 『북부장호적』에 대한 고찰」, 와그너(2007), 일조각
이기백, 1999, 「족보와 현대사회」, 『한국사 시민강좌』 24
이상국, 2022, 「한국사 연구와 디지털역사학 연구방법론 - 양적분석을 중심으로」, 『한국사연구』 197
이영훈, 1988, 『朝鮮後期社會經濟史』, 한길사
이영훈. 2004, 「朝鮮時代의 主戶-挾戶關係 再論」, 『고문서연구』 25: 1-32.

이영훈·조영준, 2005, 「18~19세기 農家의 家系繼承의 추이 - 경상도 丹城縣 法勿也面 戶籍에서」, 『경제사학』 39
이준구, 1993, 『朝鮮後期身分職役變動硏究』, 일조각
임학성, 2000, 「조선후기 戶籍에 등재된 兩班職役者의 身分 - 1786년도 丹城縣 縣內面의 사례 분석 -」, 『조선시대사학보』 13
장효민, 2022, 「불평등과 사회이동의 장기변주 : 18~19세기 호적대장으로부터」, 서울대학교 석사학위논문(사회학)
정진영, 2002, 「조선후기 호적 '戶'의 編制와 성격」, 『대동문화연구』 40: 179-216.
정진영, 2007, 「역사인구학 자료로서 호적대장 이용을 위한 기초 연구 - 『대구부호적대장』과 촌락문서의 비교 검토」, 『대동문화연구』 59
조지 이거스, 임상우·김기봉 역, 1998, 『20세기 사학사』, 푸른역사
주성지, 2019, 「디지털시대 한국사 연구의 확장과 과제」, 동국대 사학과 박사학위논문
차장섭, 1997, 『조선후기벌열연구』, 일조각
최재석, 1983, 『한국가족제도사연구』, 일지사
한영우, 2013, 『과거, 출세의 사다리 1,2,3,4』, 지식산업사
호적대장연구팀, 2003, 『단성 호적대장 연구』, 성균관대 대동문화연구원

세션1 발표문 4

데이터로 보는 동양고전 연구: 그 현황과 과제

서재현(성균관대학교 유학동양한국철학과)

〈목 차〉

1. 데이터 활용 연구와 데이터, 방법론에 대한 관점

최근 데이터와 디지털 방법론을 활용한 동양 고전 연구가 눈에 띄게 증가하고 있다. 이는 개발 환경의 접근성 향상, 디지털 인문학 연구가 제공하는 '협업 기회의 확대와 온라인 환경에서의 자료 공유 및 교육 기회 증대'가 주 요인으로 작용한 것으로 보인다. 동아시아학 분야에서 이루어진 디지털 연구의 발전은 다음 여섯 가지 영역에서 두드러진다. 1) 데이터베이스의 확장, 2) 연구를 위한 도구와 플랫폼 개발, 3) 디지털 동아시아학 연구 및 프로그램의 확산, 4) 언어적 다양성의 증대, 5) 비판적·이론적 접근 경향의 강화, 6) 연구 커뮤니티의 형성. 이러한 경향은 디지털 동아시아 연구의 지속적인 성장과 함께 그 주제와 방법론의 다양성을 보여준다.[42)]

본고에서는 디지털 동아시아학 연구의 여섯 가지 발전 영역 중 '연구'와 직결된 두 가지, 즉 1) 데이터베이스의 확장과 2) 연구를 위한 도구 개발을 중심으로 논의를 전개하고자 한다. 이를 위해 선행 연구들의 경향을 크게 두 가지 흐름으로 나눠 살펴볼 것이다. 첫 번째는 데이터베이스의 확장이다. 디지털 방법론을 도입하려는 기존 인문학 연구자들은 데이터의 구축과 활용에 대한 포부를 밝히며, 이를 통해 연구에 제공될 잠재적 이점에 큰 기대를 드러내고 있다.

이러한 기대는 모레티나 슬링거랜드의 연구 목표에서 잘 나타난다. 요약하자면, 기존 연구자

42) (Horváth, A., 2023)

들이 주로 다뤄온 '정전화된 텍스트(canon)'[43], 또는 선별적인 취사 선택(cherry-picking)으로 인해 발생하는 편향[44]에서 벗어나고자 하는 것이다. 그 대신, 대량의 데이터를 활용한 새로운 결론으로 나아가거나, 양적으로 확보한 객관적 근거들(데이터)을 바탕으로 보다 보편적이고 신뢰할 수 있는 결론을 도출하려는 것이 이들의 주요 목표라 할 수 있다. 이때 우리가 사용하는 데이터가 어떤 성격을 지니고 있는지에 대해 깊이 살펴볼 필요가 있다.

그렇다면 데이터란 무엇인가? 이 질문에 대해 오늘날 데이터는 너무나도 범용적으로 사용되기 때문에, 사람들은 쉽게 저마다의 답을 내놓을 것이다. 예컨대, 데이터는 '다양한 형태를 띠는 인간 지식의 집적물'이나, '전산 처리가 가능한 형태로 가공된 정보'라고 직관적으로 정의할 수 있다. 디지털 인문학 연구자들 또한 데이터에 대해 다양한 이해를 제시해왔다.

일부는 데이터가 태생적으로 편견을 내포하고 있다고 주장하며, 다른 이들은 데이터의 정의를 철저히 사전적[45]이거나 전산학적 맥락에서 접근하기도 한다. 또한 서양 철학사에서 형이상학적 보편자를 정보로 치환해 이해하는 관점[46]을 차용하면, 편견이 개입될 여지가 없는 순수한 정보나 데이터 영역이 존재한다고 볼 수도 있다. 이 관점에 근간하면, 정보 또는 데이터는 추론과 계산의 토대로서 그 자체로는 해석되지 않은, '유용성을 잠재한' 상태[47]로 존재한다. 이는 어떤 편견이나 해석도 개입되지 않은 순수한 형태로 이해된다.

그러나 데이터를 여기서 언급한 '정보'와 구분하여 본다면, 데이터가 위치한 영역은 철학적 근원에 가까운 '정보'의 영역과는 다르다. 정보는 해석자의 주관과 판단이 개입되기 어려운 순수한 형태로 간주될 수 있지만, 우리가 다루는 데이터는 우리의 심상(마음)에서 해석되거나 현실에 투영된 관찰과 기록의 산물이다. 이러한 데이터는 필연적으로 해석자의 주관과 맥락이 반영될 수밖에 없으며, 이는 데이터가 단순히 중립적인 상태로 존재하기 어려운 이유이기도 하다.[48]

---

43) 멀리서 읽기 방식을 통해 연구자는 가까이 읽기의 편향으로부터 비교적 자유로운 상태에서 기술통계적 지표들로부터 통찰을 얻게된다. 이를 통해 기존에 주목하지 못했던 지점들에도 관심을 쏟을 수 있게되는데, 이처럼 컴퓨터를 활용한 읽기 방식은 극소수의 정전(canon)에서 벗어나 관심의 주변부에 있었던 텍스트들에까지 접근할 수 있게끔 한다.

44) 대표적으로 로엘 스터크, 제인 기니, 마이클 나일랜의 주장을 비판대상으로 삼는다. 슬링거랜드는 이들이 마음-몸 일원론을 옹호하면서 체리피킹(cherry picking)한 몇 가지 근거들을 제시하며 사용한 표현들에 대해 문제를 제기한다. 이들은 '수적으로 우세하다', '항상 그러한 것은 아니다', '더 자주 가리킨다' 등의 표현을 사용하는데 이것이 실제로 정량적 근거를 통해 입증된 것이 아니라는 것이다. (Slingerland, 2013)

45) 데이터에 대한 정의 중 세 번째[3], '이론을 세우는 데 기초가 되는 사실. 또는 바탕이 되는 자료'의 의미(표준국어 대사전, 데이터 2024, 11월 15일 검색)

46) 서양 철학사상에서 '정보'는 근원적, 이데아적, 형이상자의 지위를 점하며, 세계, 마음, 언어는 이 정보라는 본질이 구현된 것이었고, 이를 처리하는 기술의 집약체인 오늘날의 컴퓨터는 서양철학의 한 결정체로 간주된다. (이승종, 2024)

47) Merriam-Webster's definition of "data" (accessed November 15).

48) 장석권은 관찰자의 '관점'이나 '마음'이 데이터 수집 과정에서 어떻게 데이터를 형성하고 왜곡할 수 있는지를 상세히 설명한다. 이 때, 왜곡은 현실의 실체가 본질적으로 불확정성을 띠고 있기 때문에, 우리가 실체 자체를 보는 것이 아니라 그것의 일부 단면만을 관찰하게 되는데에서 기인한다. 이는 '관찰 효과(observer effect)'로 나타나며, 데이터를 필터링하거나 특정 기준을 설정하는 순간부터, 또

이 때, 컴퓨터는 데이터를 처리하는 핵심적인 연산 장치로 기능한다. 그렇다면, 컴퓨터를 도구로 활용한 우리의 인문학 연구 작업은 어떻게 이해될 수 있을까? 이는 자연어 텍스트를 임베딩하고, 컴퓨터의 연산 능력을 활용해 구문론적(Syntax) 처리가 가능한 알고리즘을 통해 단어들 간의 연관 관계를 수치로 도출하는 과정으로 설명할 수 있다. 그러나 이 과정에서 우리는 두 가지 중요한 문제에 직면한다. 첫째는 데이터의 객관성 또는 편향성을 어떻게 이해하고 평가할 것인가에 대한 문제이며, 둘째는 방법론의 문제로, 데이터(혹은 데이터로 환원될 수 없는 것들)를 어떻게 다룰 것인가에 대한 문제이다.

데이터에 대한 관점: 데이터를 이해하는 관점에 따라, 수집된 데이터가 '수적으로 다량'이라는 사실 만으로 '객관성'은 자연스럽게 확보되는 것이 아님을 알 수 있다. 그보다, '데이터를 활용' 함으로써 기대할 수 있는 객관성은, 동일한 입력 값이 투입되었을 때, 동일한 출력 값이 담보되는 '기계적'인 객관성에 가깝다고 할 수 있다. 데이터가 형성되며 배태된 불확정성, 데이터가 품고 있는 여러갈래의 '인간의' 편견들, 그리고 그 데이터를 구성할 때 개입되는 구성자의 의도는 데이터를 주관적이고 편향적으로 만들 위험성을 내포하고 있다. 이는 구조적 편향성을 야기할 수 있는 중요한 문제로 주목해야 한다.[49]

텍스트와 방법론의 연결: 데이터 연구에서 인문학 연구자가 주요 분석 대상으로 삼는 것은 여전히 문자로 기록된 '텍스트'이다. 물론, 이미지나 기호로 분석 범위가 확대되고 있는 추세[50]이지만, 연구자들의 관심사는 여전히 고전 텍스트와 같은 전통적 연구 대상에 머물러 있는 경우가 많다. 이러한 상황에서, 문자 텍스트에 구문론적 분석 수단을 적용하여 의미론적 또는 해석학적 분석으로 나아가는 데는 여전히 큰 어려움이 따른다. 실제로, 이러한 시도가 가능하더라도 해석의 깊이와 정확성을 확보하는 데에는 많은 한계가 존재한다.

편의상 약칭하자면, '데이터의 편향 문제'와 '연구문제와 방법론 간의 적정 결합 문제'는 디지털 방법론을 활용한 인문학 연구 설계 단계에서 반드시 고려해야 할 문제라고 할 수 있다. 우선, '데이터의 편향 문제' 는 데이터 자체의 특성이 연구 성과를 직접적으로 저해하기보다는, 데이터에 대한 명료한 이해 없이 연구를 진행할 때 발생한다. 데이터가 제공할 객관성에 대한 막연한 기대와 그 기대에 부응하지 않는 결과값 사이에서 생기는 불일치는 연구에 혼란을 초

는 계층 구조, 시야, 감각기관의 설정값, 알고리즘의 하이퍼파라미터와 같은 특정 배율을 통해 데이터를 관찰하는 순간부터 데이터는 불확정성을 가지게 된다. (장석권, *1장-3, 2018*)

49) 이때 편견을 반드시 '부정적인 것'으로만 해석할 필요는 없다는 관점도 존재한다. 가령, (류정민, 2024) 연구에서는 (이희용, 2019)를 인용하며 편견이 근대 이전에는 가치중립적인 단어였다는 점에 주목하며, 이를 단순히 '검증되지 않은' 것으로 이해하는 것이 타당하다고 주장한다. 특히, 자료 수집 과정에서 객관성을 기계적으로 적용하기 어려운 인문학 데이터의 경우, 검증 이전의 "해롭지 않은" 편견은 오히려 연구 과정에서 유용하게 작용할 수 있다.

50) 장석권에 따르면, 메신저 공간에서, 언어 대화 뿐 아니라 감정 전달을 위한 여러 장치들(이모티콘, 동영상 클립)이 등장하여 표현의 다양성과 감정의 '깊이' 등을 데이터의 관찰 대상에 포함시킴으로써 데이터에 대한 정의 자체가 바뀌었다. (장석권, 2018)

래할 수 있다.

2. 연구 과정과 사례로 살펴본 데이터 기반 동양 고전 연구에서의 문제

다음으로, '연구문제와 방법론 간의 적정 결합 문제'는 특정 방법론, 예컨대 구문 분석 도구를 사용해 철학적 문제나 기존 연구 문제를 검증하려 할 때 직면한다. 여기서 중요한 것은, 채택한 방법론의 특성 및 한계와, 도출된 결론의 함의를 명확히 하는 것이다. 이를 통해 구문 분석에서 도출된 수치들이 연구 문제를 해결하는 데 어떤 방식으로 연관되고 기여할 수 있는지에 대한 구체적인 논의가 이루어져야 한다. 이러한 고민 없이는 연구 방법론의 적용이 단순한 기술적 시도로 그칠 위험이 있다.

이 두 가지 문제는 기존의 선행 연구에서 다양한 양상으로 나타나며, 많은 연구들에서 이를 극복하기 위한 시도가 이루어졌다. 이제부터는 디지털 인문학 연구 과정과 그 단계별 특성을 잘 반영한 대표 연구들을 중심으로, 데이터와 방법론이 얽힌 문제들을 살펴보고자 한다.

디지털 인문학 연구의 전 과정을 간략하게 정리하면 다음과 같은 순환적인 구조로 표현할 수 있다. 다음은 디지털 인문학 연구의 전 과정을 간략한 모식[51]으로 표현한 것이다.

> 대규모 데이터베이스(DB) 구축 작업: 연구를 위한 기본적인 데이터 인프라를 마련하는 단계 ➔ 소규모 데이터 추출 및 환경 구축: 대규모 데이터에서 연구자의 특정 연구 수요에 맞춘 소규모 데이터를 추출하고, 방법론 적용을 위한 환경, 코드, 설정값 등을 준비하는 단계 ➔ 방법론 정립 및 최적화: 데이터와 방법론을 최적화한 뒤, 도출된 결과를 연구 문제와 연관 지어 해석하는 단계 ➔ 디지털 인문학 교육을 통한 순환 환경 조성: 연구 과정과 결과를 교육 콘텐츠로 환원하여 새로운 연구자들에게 학습 환경을 제공하는 단계 ➔

이러한 단계는 디지털 방법론이 단순한 기술적 수단으로써의 의미를 넘어 인문학 연구에 기여하기 위해 반드시 거쳐야 하는 과정을 보여준다.

1) 연구를 위한 대규모 DB의 구축 작업: 국내에서 진행된 대규모 데이터베이스 구축 작업은 디지털 인문학 연구의 초석을 마련한 대표적인 사례들로 꼽힌다. 예를 들어, 국내 DB 작업의 기점으로 거론되는 *문과방목* 데이터베이스화 작업[52]과 *조선왕조실록* DB화 사업[53]

51) 여기서, 단계의 순서는 연구가 '발전된 정도'를 말하지 않는다. 다만, 연구의 수행 순서 상에 따라 배열하고 구분한 것이며, 선행하는 단계의 연구가 후행 단계의 연구에 기여할 수 있는 것과 같이 후행하는 단계의 연구 역시 선행 단계의 연구에 기여하는 것이 가능하다. 각 단계의 구분 및 단계별 예시로 제시된 사례들 각각이 적절히 선별된 것인가에 대해서는 충분한 사후 논의가 필요하다.

52) 이 프로젝트에서 에드워드 와그너와 송호준은 약 14,000여 명의 문과 합격자 및 친인척 데이터를 디지털화하여 분석하였다 (이재옥, 2021).

53) 온라인 플랫폼은 2005년 조선왕조실록 대국민 온라인 서비스 사업으로 구축된 DB이며, <조선왕조실록 CD-ROM> 사업 (https://sillok.history.go.kr/intro/peoplePopup.do?type=02)이 그 모태이다.

을 시작으로, *한국고전종합DB*, *한국학 DB*, *동양고전 DB* 등이 있다. 이러한 DB들은 구축 이후 지속적인 연구자들의 활용과 피드백[54]을 통해 현재도 발전을 거듭하고 있다. 해외 DB로는, *CBDB*(China Biographical Database)[55], *CTEXT*(Chinese Text Project)[56], 한적전자문헌자료고[57] 등[58]이 있으며, 이 DB들은 raw 데이터를 업로드 하거나, API를 제공하는 등 탁월한 접근성으로 인해 연구자들에게 애용되고 있다. 이 같이 진입장벽이 최소화됨으로써 현재는 누구나 마음만 먹으면 API로 호출해 개인 PC 리소스만을 가지고도 기본적인 디지털 방법론을 활용한 연구가 가능해진 상황이라고 할 수 있다.

본 발표에서는 이러한 DB의 구조나 기술적 특성보다는, DB를 활용한 연구에서 어떻게 앞서 언급한 데이터의 편향성과 연구문제-방법론 간 적정 결합 문제를 인식하고 대응했는지에 초점을 맞춘 논의를 전개한다.

2) 연구에 필요한 소규모 데이터 추출 및 환경 구축: 이 단계에서는 연구에 필요한 텍스트 판본을 특정[59]하고, 연구 문제와 직결된 타겟 키워드 딕셔너리를 구축하며, 설계자의 연구에 특화된 데이터 모델링을 통해 소규모 DB를 구축[60]하는 작업 및 양질의 데이터를 엄선하여 직접 연구용 데이터를 구성하는 작업[61] 등이 이루어진다. 이러한 작업에는 기존 대규모 DB에서 데이터를 추출하거나 편집하는 것뿐 아니라, 데이터를 새로운 형태로 가공[62]하거나 시각화[63]하는 과정도 포함된다. 특히, 연구 대상 데이터의 성격을 파악하고 이를 어떻게 처리할지 계획하는 작업이 중요한 부분을 차지한다. 예를 들어, 동양 고전의 대부분을 차지하는 한문 텍스트를 다룰 때, 토크나이징 과정에서 다양한 난제에 직면한다. 대표적으로, n음절 단어의 처리 문제[64]나 이형태 단어의 통합 문제[65]가 있다[66]. 이 같은 문제

이 DB는 1968~1993 까지 세종대왕기념사업회와 민족문화추진회가 번역한 실록을, 서울시스템이 1995년 CD-ROM으로 구축한 데이터를 바탕으로 하고 있다.

54) (민경주 외, 2022), (정만호, 2021), (문미진, 2019) (서정화, 2015), (정만호, 2018)에서는 한국고전종합DB의 설계, 구성, 활용과정 등에서 보이는 문제점 및 개선 방안을 제언한다.

55) China Biographical Database Project (CBDB) (https://projects.iq.harvard.edu/cbdb/home)

56) Chinese Text Project 中國哲學書電子化計劃 (https://ctext.org)

57) 漢籍電子文獻資料庫, (https://hanchi.ihp.sinica.edu.tw)

58) 그 외에도 많은 언어 모델들의 학습자료로 사용된 殆知阁古代文献txt大全集 등이 있다. (https://github.com/garychowcmu/daizhigev20/tree/master)

59) 판본의 판별에는 저자 등 해당 텍스트와 관련된 여러 메타 데이터에 관한 판별 과정이 포함된다. 기계 학습을 이용한 역사 텍스트의 저자 판별 연구는 다음을 참조 (최지명, 2017) 및 (박선영, 2022).

60) 연구용 소규모 db에 대한 구상은 '조선후기 한시 온톨로지'의 사례에서 잘 보인다.(류정민, 2024)

61) DRH(The Database of Religious History)는 에드워드 슬링거랜드가 주축이 되어 구축한 데이터베이스로, 종교사와 문화사를 중심으로 정량적 데이터와 정성적 주석, 참고자료를 통합하여 학술적 협력을 촉진하는 데이터베이스다. 학문적 엄격성을 유지하며 시각화 및 분석 도구를 통해 종교와 역사적 변수 간의 관계를 정량적으로 연구할 수 있도록 지원하는 플랫폼이다. 또한 학술적 의견 차이를 문서화하고 전문가 간 논의를 장려하며, 다양한 자료를 공유하고 접근할 수 있는 공간을 제공한다. (https://religiondatabase.org/landing)

62) 각 단어들에 딸린 메타데이터의 네트워크를 활용하여 ; 개념들간의 연결 관계를 시각화해주는 서비스로 고전용어 시소러스가 있다. 이 서비스는 한국문집총간에 나타난 고전용어를 기반으로 하여 당시 사람들이 여러 형태로 표현했던 용어들을, 표제어를 기준으로 유의어와 관련어, 상하위어 등의 관계를 규정해 줌으로써 효율적인 정보검색을 제공한다.

63) 한국 고전 종합DB 의 시각화 기능 개선을 도모하기 위한 제언 및 연구는 다음에서 보인다.(이병찬, 민경주, 2021)

64) 두 음절 이상의 결합어를 복합어로 처리할 것인지 혹은 '단음절' 단위로 일괄적으로 분리할 것인지

의식은 선진시기 동양 고전 텍스트를 대상으로한 연구들에서도 보이는데[67] 데이터를 효율적으로 처리하기 위해 토크나이저를 개발[68]하거나, 연구용으로 정제된 DB를 공유하기 위한 자체 레포지토리를 구축[69]하는 작업이 병행되기도 한다.

3) 방법론 정립: 이 단계는 직전 단계에서 수집한 데이터를 바탕으로 본격적으로 디지털 방법론을 적용하는 과정이다. 연구 문제를 해결하기 위해 필요한 방법론을 고안하고, 해당 방법론을 실행하기 위해 알고리즘의 여러 설정값[70]을 결정하는 것이 핵심이다. 초기 연구에서는 방법론을 안정적으로 적용하고 실험적으로 분석하며, 텍스트와 방법론 간의 기술적 접합을 이루는 토대를 마련하는 데 초점이 맞춰졌다.[71] 이어지는 후속 연구들에서는 '방법

에 대한 문제를 지칭한다. (정유경, 반재유, 2019)에서는 복수의 연구들에서 한국어에 존재하는 한자 형태소 처리에 대한 논의가 있었으나, 1음절 한자어의 다양한 의미, 분포 문제에 대한 해결책 제시가 부재한채로, 2 음절 한자어를 복합어로 볼지, 또는 분리된 단일어로 볼 지에 대한 논의가 첨예하게 대립하고 있다고 소개한다(281),이 연구에서는 국한문 혼용 텍스트의 한자어 처리에 있어 단음절 단위와 복합어 처리의 적합성을 논의한다. 연구는 근대 초기 국한문 신문 텍스트의 전처리 과정에서 전공자 6명이 수작업으로 색인어를 선정하고 검증하는 방식을 사용했다. 분석 결과, 전체 색인어 중 80% 이상이 2음절 단어로 나타났으며, bigram 분석으로 불용어 리스트를 생성했다. 한글 형태소 분석기 및 딕셔너리를 활용한 방식을 중국어 분석기(jieba)와 비교했을 때, 재현율은 약 8% 높았고, 정확률은 약 2배 우수한 성능을 보였다.

65) 이 문제는 같은 형태를 공유하면서, 다른 음과 뜻을 지닌 한자어들에서 발견된다. 대표적으로 악惡과 오惡, 악樂과 락樂 등을 꼽을 수 있다. 타개책으로 유니코드를 활용한 변별, 데이터에 직접 마킹 등의 방법이 있다. 특히 국가별 판본을 아울러 텍스트 분석을 할 경우 이 문제는 더욱 심화된다. 그렇다고 어느 한 유니코드 범위를 통해 한, 중, 일 한자 중 한 가지로 일원화 시키는 전략은 실현되기 어려운데, 특히 각 국가가 사용해온 한자 문화 혹은 정체성에 대한 이해와 합의 없이는 성립되기 어렵다.(The Digital Humanities and the History of Religion in Asia: An Introduction 13pg 에서 재인용, (Jing Tsu, 2022)

66) 이 마저도 일관된 여러 다른 경로로 수집된 텍스트들이 일관성 있는 유니코드로 표기되었다는 전제에서 가능한 것으로, 여의치 않거나 데이터에 해당 값들이 비일관적인 방식으로 기록되어 있을 경우 언어 임베딩을 통해 벡터 값을 활용하는 전략 등이 고려될 수 있다. 이에 대해 적용해 볼 수 있는 방안은 다음 연구를 참조해 볼 수 있다. (박진호, 2020)는 문장벡터(문장에서 공기한 단어들의 벡터)까지 고려하고 FSE를 활용하여 소수 문장에서 특징적인 단어에 가중치를 부여하는 방식, 혹은 공기어에 따른 미세조정에 유리한 BERT 방식 등을 검토하였으며, 기계학습 단계에서 중의성 해소를 요하는 타겟단어를 중심으로, (선조거리 기준이 아닌) 통사적으로 밀접한 관계 요소를 취사하여 활용하는 방식의 유효성을 실험한 바 있다. 다만 이 것이 벡터화된 한문 텍스트에서의 문자 판별에 어떻게 기여할 수 있는지에 대해서는 추가 논의가 필요하다.

67) 슬링거랜드, 김바로 등의 경우에는 모든 단어를 단음절로 고려하여 처리한 반면, 직접 연구에 필요한 키워드 딕셔너리를 구성해 사용한 것은 (서재현, 2020) (박선영, 2021), (하나, 2023)의 연구에서 보인다. 슬링거랜드의 경우, 동양 고전 텍스트에서 합성어 대비 단음절 단어의 비중이 압도적으로 많음을 이유로 들어 별다른 연어 처리 없이 한자어를 낱자 단위로 디지털 방법론에 투입하였다. 디지털 방법론을 활용한 고전 분석에 있어, 고대 중국어 동형동음이의어 및 합성어의 처리 과정에 대한 고찰은 (박선영, 2021)의 연구 58-68 쪽에 상술되어 있다. 기본적으로 단음어 처리하되, 일부 연어를 합성어 처리하는 방식으로 처리하였다. 고대 중국어 단어의 형태, 독음, 의미, 품사, 구조 및 사회적·문화적·내용적 맥락 등 여러 측면을 두루 고려하여 키워드를 직접 선정하였다.

68) 기계학습을 이용한 UD-Kanbun (Yasuoka, 2019), 대규모 고전 텍스트로 Finetuning 시킨 언어모델 XunziALLM, (GuwenBERT, 2020), GujiBERT, 2013), (甲言Jiayan, 2019)

69) 예시, 슬링거랜드 연구팀의 자체 레포지토리 (https://hecc.ubc.ca/quantitative-textual-analysis/data-repository/) 참조.

70) 가령, W2V 분석의 경우 window size, skipgram 또는 Cbow 방식 여부 등을 연구자가 결정해야한다. 이 과정에 대해서는 (서재현, 김병준 외, 2021)의 연구 데이터 특성에 맞춘 방법론 하이퍼파라미터 값의 세부 설정에 대해 상술된 부분(367-371 쪽)을 참조.

71) 김바로. (2019)에서는 CBETA 불경 데이터를 Word2Vec 방법론으로 분석하고 시각화하여 인공지능

론 최적화'를 통해 구체적인 방법론의 세부 설정의 최적화, 더 나아가 결과가 보이는 의외의 현상에 대한 서술, 실 맥락을 토대로한 교차 검증 등이 이뤄진다.

3-1) 방법론 최적화: 이 단계는 디지털 방법론을 개별 연구 문제에 효과적으로 적용하기 위해 적합한 방법론을 탐색하고, 실험적으로 설정값을 조정해 최적화하는 과정을 포함한다. 도출된 결과물은 연구 문제와 연관지어 해석되며, 이를 통해 데이터 분석 결과가 학문적 논의에 실질적으로 기여할 수 있도록 한다. 이에 대한 몇 가지 예시를 들자면, 주성분 분석(PCA)을 활용한 고전 편장에 대한 저자 판별 연구[72], 워드 투 벡터 분석(Word2vec)을 활용한 일제 강점기 신문 텍스트 검열연구[73] 및 조선 문인들의 사칠론에서 보이는 감정어휘의 양상 분석[74], 토픽 모델링 분석을 활용하여 논어, 맹자, 순자에 보이는 주제상 유사점을 분석한 연구[75] 및 동양에 심-신 이원론적 사유가 존재했음을 밝히는 연구[76], 네트워크 분석을 통하여 특정 개념과 얽힌 의미 영역별 맥락변화를 추적한 연구[77] 등을 꼽을 수 있다. 이 사례들은, 주로 인문학 분야를 중심으로 하여, 디지털 인문학 연구에서 주로 사용해온 대표적인 방법론들을 활용한 연구라고 할 수 있다. 그 중에서도, 특히 동양 고전을 대상 텍스트로 삼거나, 동양 사상과 관련한 연구들을 최대한 포함시키고자 하였다. 이는 단순히 새로운 방법론을 개발하는 단계를 넘어, 기존 연구 주제를 디지털 인문학적 관점에서 최적화된 방법으로 재해석하는 과정이라 할 수 있다. 디지털 인문학 연구에서 방법론은 데이터를 확대 관찰하고 오류를 발견하면 후처리 과정을 통해 개선한 뒤 다시 적용하는 반복적인 실험 과정을 통해 기존에 보지 못했던 새로운 패턴이나 통찰을 발견하는 데 중요한 역할을 한다. 이 과정은 연구 문제를 보다 심층적으로 탐구할 기회를 제공하며, 기존 데이터의 한계를 넘어 새로운 시각을 열어준다. 그러나 실험적 연구 과정에서 나타나는 한계 중 하나는 데이터 구축과 소규모 데이터 추출 과정에서 연구자의 주관이 개입될 수 있다는 점이다. 데이터의 성격에 대한 명확한 정의 없이 이루어지는 데이터 구축은 연구 결과의 신뢰성을 저해할 수 있다. 이러한 과정에서 데이터(및 데이터를 구성하는 방식, 데이터 전처리)에

을 활용한 불교학 연구의 가능성을 탐색한다. 분석 결과를 불교학 연구자가 쉽게 활용할 수 있도록 시각화 방안을 제시하는 한편, 불교학 디지털 온톨로지 구축을 제안한다.

72) (박선영, 2022)에서는, 주성분분석·BERT 등 군집화·분류 알고리즘을 활용하여 『한비자』의 편장에 대한 유사성을 판별한다. 특히 논쟁의 여지가 있어온 두 편장, 「초현진」과 「존한」편이 상호 간 높은 유사도를 보인다는점, 『전국책』과 높은 유사도를 보인다는 점을 근거로 들어 이 두 편의 위작 가능성 및 출처가 전국책에서 비롯되었을 가능성을 제기한다.

73) (이재연, 정유경, 2020) 에서는 Word2vec을 검열연구에 적용하여 OO나 XX와 같은 복자(후세지, 삭제된 글자 대신 표시한 OO, XX와 같은 기호)의 전후 문맥에서 반복되는 의미의 맥락을 살핀다.

74) (하나 외, 2024)에서는 성호 이익의 감정론을 중심으로 텍스트 범위를 설정하고, 퇴계와 율곡을 대조군으로 삼아 비교 분석을 수행하였다.

75) Nichols, R., Slingerland, E., Nielbo, K., Bergeton, U., Logan, C., & Kleinman, S. (2018). Modeling the contested relationship between Analects, Mencius, and Xunzi: Preliminary evidence from a machine-learning approach. The Journal of Asian Studies, 77(1), 19-57.

76) Slingerland, E., Nichols, R., Neilbo, K., & Logan, C. (2017). The distant reading of religious texts: A "big data" approach to mind-body concepts in early China. Journal of the american academy of religion, 85(4), 985-1016

77) (허수, 2016)

대한 이해가 충분히 선행되지 않는다면, 방법론이 도출한 결과는 단지 편향된 데이터를 반복 재생산하는 데 그칠 위험이 있다.

3-2) 또한, 연구문제와 방법론 간의 결합 과정에서 최적화의 문제가 발생할 수 있다. 예를 들어, 슬링거랜드 연구에서 심(心)과 신(身)의 관계를 검토하기 위해 토픽 모델링을 활용했으나, 상위 단어만을 근거로 심신 이원론을 주장한 것에는 한계가 있다. 하위 단어에 신체 관련 단어가 포함될 가능성을 배제하거나, 공출현 여부만으로 단어 간 관계를 입증하려는 접근은 논리적 비약으로 이어질 수 있기 때문이다. 심(心)이라는 단어가 다양한 맥락에서 나타나는 속성들을 통합적으로 검토하지 않고 개별 토픽만을 분석하는 접근은 결과의 타당성을 약화시킬 수 있다. 디지털 방법론을 적용할 때는 예상되는 결과를 충분히 고려하여 방법론과 데이터 구성을 최적화함으로써 데이터의 신뢰성을 높이고, 연구 문제를 보다 정밀하게 탐구할 필요가 있다. 동시에, 축적된 '편향된' 데이터로부터 도출된 결론은, 잘 정제된 한 문장에서 이루어진 논증보다 낮은 설득력을 가질 수 있다는 점을 항상 유념해야 한다.

4) 디지털 인문학 교육을 통한 순환 환경 조성: 10년 전만 해도 세계 디지털 인문학 학술대회에서 교육에 관한 논의는 찾아보기 어려웠으나[78], 디지털 방법론의 활용이 본격화된 오늘날, 학생들에게 '디지털 인문학'이라는 이름으로 무엇을 가르칠 것인지에 대한 담론이 활발히 이루어지고 있다. 국내외 대학에서는 정규 및 비정규 과정을 통해 디지털 인문학 개론이나 실습 기반의 방법론을 프로젝트 형식으로 교육[79]하고 있으며, 최근에는 전공 역량 강화를 위해 디지털 인문학 또는 AI 교육을 정규 커리큘럼에 포함시키는 사례가 증가하고 있다. 동양 고전을 활용한 교육 사례 역시 찾아볼 수 있다. 서울대학교 자유전공학부 강의에서 논어를 활용한 교육 사례[80] 및 성균관대학교 비정규 강좌 및 정규 강좌 사례에서 한국고전종합DB를 활용한 교육 사례[81]가 대표적이다. 이 단계에서 핵심적인 논의는 다시금 '멀리서 읽기(distant reading)'의 정의와 활용으로 귀결된다. 학생들은 디지털 방법론(멀리서 읽기)을 활용한 고전 분석 전공 지식과의 연계를 탐구하며, 디지털 방법론이 기존 학문적 접근법과 어떤 차별점을 제공할 수 있는지 고민하게 된다. 이 과정은 1단계(대규모 DB 구축), 2단계(소규모 데이터 추출 및 방법론 적용), 3단계(최적화 및 해석)에서 축적된 데이터를 기반으로 이루어진다. 교수자는 학생들에게 새로운 방법론과 기존 지식 체

78) (홍정욱, 김기덕, 2014)
79) 서울대, 연세대, UCLA, 라이덴대, 프린스턴신학대학원 외, 의 DH 교육 사례를 분석한 연구로는 다음(한미경, 2021)을 참조.
80) (류인태. 2021)의 수업 사례에서는, 논어 텍스트를 바탕으로 온톨로지를 구축하여, 그 안에 내재한 사유들을 구조화하고, 학생들의 다양한 문제의식을 해소하는 사례들이 소개되며, (하나, 2024)의 수업 사례에서는 학생들이 한국고전종합DB로부터 기존의 전통 음식,
81)(하나, 2024)에서는 <디지털 방법론으로 보는 한국문화와 철학> 수업과 <한국철학문화 디지털 융합인재 공모전>의 사례로 '한식의 해외 인기 요인 분석'과 '나만의 도설 만들기' 프로젝트를 소개하였다. 이는 디지털 방법론을 활용한 교육 사례로, 대량 데이터를 기반으로 유관 키워드를 추출하고 논증을 강화한 방식을 제시하였다.

계의 결합을 모색하며 이를 효과적으로 전달하고, 학생들은 이를 배우는 과정에서 고전에 대한 관심을 새롭게 발견한다. 동시에 학생들은 멀리서 읽기 기술을 익히며, 디지털 인문학이 강조하는 협업, 재현, 공유의 가치를 자연스럽게 체득해 나간다.[82] 이러한 교육 과정은 디지털 인문학 헌장을 통해 공유되는 가치와 결합하여, 기존 학문 분과와의 융합을 촉진한다. 그 결과, '디지털 문학', '디지털 한국학', '디지털 고전 연구' 등과 같은 학제 간 연구가 활발히 이루어지고 있다. 이를 통해 디지털 인문학은 학문의 확장성과 융합 가능성을 강화하며, 점차 여러 학문들을 연결시키는 가운데서 융합연구의 중심적인 위치를 차지해 나가고 있다. 이때, 교수자와 학생은 기존 선행연구에서 주목하지 않았던 주변적인 요소들, 혹은 더 큰 구조 속에서 드러나는 새로운 패턴에 관심을 기울이는 기회를 갖게 된다. 이를 통해 새로운 질문이 필요하다는 사실을 자연스럽게 깨닫게 되며, 비로소 '질문하기로서의 멀리서 읽기'가 가능해지는 것이다.[83]

3. 디지털 동양 고전 연구의 과제와 디지털 동양학의 미래

최근 간행된 디지털 인문학 학술서의 표지에는 다음과 같은 문구가 실려 있다:

"경제성이 곧 가치를 가르는 시대가 되면서 낡은 학문으로 간주되어 온 인문학에 디지털은 재앙이자 축복이다."

이 문구는 인문학이 처한 현실을 적나라하게 드러내며, 동시에 디지털 기술에 대한 기대와 부담을 여실히 반영하고 있다. 디지털 인문학이라는 이름 아래, 무언가 '새로운 것'이나 '편견에서 자유롭고 보편적인 것'을 기대하거나, 방법론이 담보하지 못하는 결과물에 막연한 기대를 갖기 쉬운 것이 현실이다. 이러한 양면성을 이 문구는 '재앙이자 축복'이라고 표현한 것이다.

따라서 우리는 이 전환의 기회 앞에서의 '기대'와 '부담'을 인식하고, 각자의 연구 과정에서 책임감과 엄격한 검증을 준비해야 한다. 방법론이 최적화되지 않았을 때 발생할 수 있는 오류, 데이터 구축 및 가공 과정에서 개입되는 편향성의 인지, 그리고 충분한 연구 과정을 거치지 못해 불충분한 결론에 타협하는 문제 등을 경계해야 한다.

82) 수업 과정에서 학생들이 발굴한 새로운 소규모 데이터와 발견들은 교수자와의 협의(튜터링)를 통해 검토되며, 이를 바탕으로 디지털 방법론의 설정값이 조정되거나 새로운 방법론이 도입된다. 도출된 결과물은 가까이 읽기를 통해 분석되며, 이러한 작업은 여러 차례 반복적으로 이루어진다. 특히, 이러한 과정은 학부생들이 생소한 고전에 처음 접근할 때, 진입장벽을 낮추고 고전에 대한 이해와 흥미를 높이는 데 효과적이다(하나, 2024)에 대한 논평문.

83) (김지선, 2023)은 모레티의 두 저서에서 파악될 수 있는 데이터 기반 문학 읽기의 본질이, '풀어낼 수 없는 문제'를 발견한다는 것에 있다고 보며, 적절한 질문을 요청하는 '멀리서 읽기'의 함의를 부각한다.

디지털 인문학 연구에서는 이러한 문제에 대응하기 위해 '협업'과 '(연구자원) 공유' 등을 적극적으로 활용하고자 노력한다.[84] 무엇보다도 연구자 개개인의 책임감과 성찰이 꾸준히 요구된다는 점은 아무리 강조해도 지나치지 않다.

디지털 방법론은 양적 연구를 통해 질적 연구를 보완하고 협업을 강화할 수 있는 가능성을 열어준다. 그러나 이는 양적 근거와 질적 근거 간의 경쟁을 의미하는 것이 아니다. 데이터를 활용한다고 해서 곧바로 객관성이 보장되는 것은 아니다. 오히려 데이터의 편향성을 인식하고 이를 보완하려는 노력이 중요하다.

특히 데이터 연구에서 흔히 발생하는 편향된 기대와 그로 인한 불일치는, 데이터를 분석하거나 인공지능 모델에 학습시킬 때 보정되지 않은 편견이 그대로 재출력되는 문제로 관찰된다.[85] 따라서 데이터를 연구에 활용할 때는 해석의 맥락화와 비판적 검토가 필수적이다. 데이터의 활용은 구문 맥락을 크게 벗어나지 않도록 하여 해석의 주관성을 일정 수준 통제할 수 있는 도구로 자리매김해야 한다.

4. 방법론의 적절한 결합에 대한 성찰

디지털 방법론을 연구 문제에 결합하는 과정에서는, 채택한 방법론의 한계와 이를 통해 도출되는 결론의 함의를 명확히 이해해야 한다. 연구 설계 단계에서부터 방법론과 연구 문제 간의 연관성을 깊이 고민하고, 구문 분석을 통해 얻은 수치가 연구 문제를 어떻게 해결하며 어떤 기여를 할 수 있는지에 대한 선행적인 숙고가 필요하다. 이는 디지털 방법론을 도입한 연구 설계의 필수 과제이다.

오늘날 연구 접근성이 비약적으로 향상되면서 디지털 인문학은 새로운 전환점을 맞이하고 있다. 데이터의 속성과 디지털 방법론의 특성은 연구의 방향을 다각화하고, 다양한 주제와 접근법을 통해 학문적 가능성을 확장하고 있다. 특히 디지털 인문학은 데이터 철학, 미학, 예술학과의 융합을 통해 기존 학문의 경계를 넘어 새로운 형태의 학제 간 연구를 만들어가고 있다.

이 과정에서 디지털 방법론은 인간과 기계의 역할 경계를 재구성하는 중요한 계기를 제공한

84) 디지털 인문학은 협업과 공동 창작, 그리고 지식의 분산을 중시한다. 또한, 학제 간 팀워크를 통해 복잡한 학문적 과제를 해결하려는 대규모, 분산형 학문 모델을 지지한다 (Digital Humanities Manifesto, V2.0)

85) (김도일, 2023) 은 데이터의 특성 및 편향성에 대해 논의하며, 데이터의 양적 증가가 더 많은 사람들이 동의할 수 있는 데이터를 구축하고 데이터의 객관성을 높이는 데 기여할 수 있다고 설명한다. 그러나 전자화한 데이터가 비교적 짧은 시기에 축적된 현 시대의 데이터임을 감안할 때, 인공지능 학습에 사용된 데이터는 우리의 다양한 편향(bias)을 그대로 반영하고 있음을 인식해야 함을 강조한다. (허유선, 2024)에서는 이에 대한 구체적인 사례로 -Ai 면접자 모델이 비효율적인 근로자를 구분한 결과에 '여성'의특성을 반영한 사례- 등을 제시한다.

다. 예를 들어, 이전까지 인간의 고유 영역으로 여겨졌던 계산과 형식화가 어려운 정보 처리 과정에 디지털 기술이 개입하면서, 인간과 기계 간의 협업이 가능해졌다. 이는 단순히 기술적 도구로 데이터를 활용하는 수준을 넘어, 기계가 인간의 사고와 유사한 방식으로 데이터를 해석하고 분석할 수 있는 가능성을 열어가고 있다. 그 한 예로 예술 영역에서의 디지털 기술 확장을 들 수 있다.

같은 맥락에서, 실제로 많은 대학에서는 디지털 인문학의 한 분야로 디지털 아트를 다루고 있다. 디지털 아트는 디지털 환경에서의 창작과 해석 과정을 통해 새로운 미학적, 철학적 질문을 제기하며, 학문적 교류와 융합을 촉진하는 역할을 한다. 이러한 융합은 예술과 인문학의 경계를 허물고, 디지털 기술을 기반으로 한 새로운 학문적 상호작용을 만들어내고 있다. 무엇보다 중요한 것은, 텍스트화되기 이전의 '경험의 영역'을 디지털 아트를 통해 체험하고[86], 사용자와의 몰입형 상호작용을 통해 새로운 형태의 데이터로 수집함으로써 '계산되거나 기록될 수 없었던 영역'이 데이터로 수집되어 간다는 점이다[87].

결국, 디지털 인문학의 미래는 데이터와 방법론의 한계를 비판적으로 검토하면서도 그 확장 가능성을 탐구하여, 연구 문제와 방법론 간의 적절한 결합, 더 나아가 인간 경험에 보다 가까워진, 연구문제-데이터의 구축을 통해 학문적 가치를 실질적으로 높이는 데 있다. 이 학문적 가치의 향상이란 단순히 '디지털 인문학'이라는 새로운 형식을 통해 관심을 끌거나, 그로 인해 경제적 가치를 추구하는 데 매몰되는 것을 의미하지 않는다.

오히려 오랜 시간 축적된 기존의 학문적 문제들에 깊이 천착하고, 디지털 기술을 통해 새로운 구조, 관계, 질문을 발견하며 문제 해결의 실마리를 찾아가는 것이다. 앞서 데이터에 대한 관점의 정립과 방법론 최적화 과정을 강조한 이유도 여기에 있다. 이를 통해 우리는 맹목적으로 '새로운 것'이나 '가치 중립적이고 보편적인 것'을 기대하거나, 방법론이 담보하지 못하는 결과물의 함정에 빠지는 한계를 극복할 수 있다.

나아가 디지털 기술은 단순한 도구에 머무르지 않고, 인문학을 새로운 가능성의 장으로 변화시키는 계기가 될 수 있다. 이러한 변화가 가능하다면, 디지털 인문학은 과거에는 구성될 수 없었던 새로운 질문을 제기하며, 인문학과 함께 발전하고 동행해 나갈 것이다.

---

86) (DHQ, 2020)에서는 디지털 인문학(DH)이 공연 예술 분석과 인간 경험의 질적 접근을 결합하는 방식에 대해 논의한 연구. 오디오-비주얼 미디어의 다중 모달리티와 시간성을 중심으로 computational 분석-인간 경험 간의 조화를 이루기 위한 방법론을 제시하며, 디지털 기술이 예술적 감정과 역학의 학문적 이해를 심화시키는 도구로서 기능할 가능성을 살핀다.

87) (Teraes, M, 2016)에서는 몰입형 인터랙티브 기술이 디지털 인문학(DH)에서 연구 및 교육의 혁신을 가능케 하는 방안을 탐구한 연구. 가상 현실(VR)과 증강 현실(AR)과 같은 기술이 인문학 연구 및 학습 경험의 확장을 위한 도구로 활용될 가능성을 논의하며, 이를 위한 프레임워크를 제시한다. 특히 몰입형 환경에서의 상호작용성과 인간 경험의 역할에 주목하며, 학술 연구와 교육 프레젠테이션을 개선할 수 있는 가능성을 강조한다.

참고문헌

김도일, 김장현. (2023). "AI persona 재해석을 통한 AI tutor 개발의 초석 놓기 - 동아시아 고전 모델을 통해 보는 특화 모델의 한 방향성." 발표 도입부, AI융합연구지원사업(2단계) 연구결과 발표회, 성균융합원, 4월 26일.

김바로. (2019). 딥러닝으로 불경 읽기: Word2Vec으로 CBETA 불경 데이터 읽기. *원불교사상과종교문화, 80*, 249-279.

김지선. (2023). 멀리서 읽기와 디지털 인문학. *한국근대문학연구, 24*(1), 43-72.

김지선. (2023). 모레티의 '멀리서 읽기'에 대한 비판적 고찰. *문학과 데이터, 2*, 15-40.

류인태. (2021). 고전탐구세미나1: 데이터 기반의 고전 읽기 교육 - 논어를 대상으로 한 디지털 인문학 강의 사례를 중심으로. *서울대학교 자유전공학부, 인문논총, 78*(1), 43-73.

류정민. (2024). 조선후기 한시 연구를 위한 온톨로지 구축 시론: 추론 기능에 주목하여. *한국학논집*, *96*, 5-41.

문미진. (2019). '한국고전종합DB'에 보이는 몇 가지 문제점: 사에 나타난 오탈자 및 구독의 문제를 중심으로. *민족문화, 53*, 5-32.

민경주, 이병찬, 정만호, & 이향배. (2022). 한국고전종합DB를 이용한 한학 대중화를 위한 스마트 검색 시스템 개발: 각주 시스템을 중심으로. *민족문화, 60*, 413-446.

박선영. (2022). '멀리서 읽기'를 통한 [한비자][초현진],[존한] 진위 논쟁 재검토: 주성분분석을 중심으로. *인문과학, 84*, 39-70.

박진호. (2020). 문장 벡터를 이용한 동형어 구분. *한국 (조선) 어교육연구*, *16*, 7-48.

서재현, 김병준, 김민우, & 박소정. (2021). 멀리서 읽는 “우리”: Word2Vec, N-gram을 이용한 근대 소설 텍스트 분석. *대동문화연구, 115*.

서정화. (2015). 한국고전번역원의 주석DB 활용 및 개선방안 연구. *민족문화, 45*, 33-58.

이승종 (2024). 역사적 분석철학, 서울: 서강대학교 출판부

이재연, & 정유경. (2020). 국문학 내 문학사회학과 멀리서 읽기: 새로운 검열연구를 위한 길마중. *대동문화연구, 111*, 295-337.

이재옥. (2021). 온톨로지를 이용한 문과방목의 정본화 (定本化) 연구. *인문콘텐츠*, (60), 171-209.

이희용. (2019). 편견에 대한 해석학적 통찰. *현대유럽철학연구, 52*, 161-195.

장석권, (2018). 데이터를 철학하다, 서울:흐름출판

정만호. (2018). 고전 번역에 끼치는 '한국고전종합DB'의 영향과 문제 개선 방안. *Journal of Korean Culture, 42*, 59-84.

정만호. (2021). 한국고전종합DB 활용과 문제점. *민족문화, 60*, 447-472.

정유경. (2020). 디지털 인문학 분야의 국내외 연구 동향 분석. *정보관리학회지, 37*(2), 311-331.

정유경, 반재유. (2019). 국한문 혼용 텍스트 색인어 추출기법 연구 [시사총보] 를 중심으로. 정보관리학회지, 36(4), 7-19.

최지명. (2017). 기계학습을 이용한 역사 텍스트의 저자판별: 1920 년대 개벽 잡지의 논설 텍스트. *Language and Information*, *22*(1).

하나, 2024 <디지털로 철학하고 디지털로 교육하기>, 한국유교학회 추계학술대회: 디지털 전환(DX) 시대의 유교적 전망, 발표자료집.

하나, 서재현, 박소정. (2024). 성호 이익의 감정 스펙트럼에 관한 연구: Word2Vec을 활용한 성호와 퇴계ㆍ율곡의 감정 어휘 비교 분석을 중심으로. *대동문화연구, 125*, 319-351.

한미경. (2021). 디지털 인문학 교육의 사례 연구: 초기 내한 선교사 사료 수업과 연계하여. *인문과학, 123*, 7-39.

한미경. (2021). 디지털 인문학 교육의 현황과 과제: 국내외 사례를 중심으로. *인문학연구, 45*, 123-150.

허수. (2016). 네트워크분석을 통해 본 1980년대 민중: [동아일보]의 용례를 중심으로. *개념과 소통, 18*, 53-95.

허유선, (2024). "인공지능과 젠더 편향." 성균관대학교 유교문화연구소 비판유학.현대경학연구센터 학술대회: 포스트휴먼 시대, 여성주의적 유학, 성균관대학교, 8월 23일.

홍정욱, & 김기덕. (2014). 2014 세계 디지털인문학학술대회 및 한국의 디지털인문학. *인문콘텐츠, 34*, 53-75.

Digital Humanities Quarterly (DHQ). (2020). Matching Computational Analysis and Human Experience: Performative Arts and the Digital Humanities. *Digital Humanities Quarterly*, 14(4).

Franco Moretti, 김용규 역. (2021). *멀리서 읽기: 세계문학과 수량적 형식주의.* 서울: 현암사.

Jing Tsu, Kingdom of Characters: The Language Revolution that made China Modern (New York, NY: Riverhead, 2022), 263-264.

Koichi Yasuoka (安岡孝一): Universal Dependencies Treebank of the Four Books in Classical Chinese, DADH2019: 10th International Conference of Digital Archives and Digital Humanities (December 2019), pp.20-28.

Nichols, R., Slingerland, E., Nielbo, K., Bergeton, U., Logan, C., & Kleinman, S. (2018). Modeling the contested relationship between Analects, Mencius, and Xunzi: Preliminary evidence from a machine-learning approach. *The Journal of Asian Studies, 77*(1), 19-57.

Slingerland, E., Nichols, R., Nielbo, K., & Logan, C. (2017). The distant reading of religious texts: A 'big data' approach to mind-body concepts in early China. *Journal of the American Academy of Religion, 85*(4), 985-1016.

Teraes, M. (2016). Immersive Interactive Technologies in Digital Humanities: A Review and Basic Concepts. International Journal of Digital Humanities.

Wang, D., Liu, C., Zhao, Z., Shen, S., Liu, L., Li, B., ... & Wang, X. (2023). Gujibert and gujigpt: Construction of intelligent information processing foundation language models for ancient texts. *arXiv preprint arXiv:2307.05354*.

세션1 발표1 '(전산)언어학과 디지털인문학'에 관한 토론문

도재학(경기대학교 국어국문학과)

전산언어학은 언어 빅데이터인 코퍼스를 구축하고 활용하여 언어 연구를 수행하는 분야이고 디지털인문학은 '디지털 자원과 기술을 활용하는 인문학 연구'라고 할 수 있으므로, 전산언어학은 디지털인문학의 한 축이라고 할 수 있을 것입니다. 특히 여러 다른 디지털인문학적 연구에 적용 가능한 방법적 기반을 제공하는 역할을 한다는 점을 고려하면 핵심적이고 중요한 위상을 차지하고 있는 것이 아닌가 생각해 봅니다.

이번 정성훈 선생님의 발표는 전산언어학의 역사와 최신 연구 동향을 아우르고 있습니다. 구체적이고 상세한 설명을 통해 전산언어학, 그리고 디지털인문학에 대한 이해를 높일 수 있었습니다. 내용 정리가 잘 되어 있고 메시지가 명확하기 때문에 토론거리로 삼을 어떤 쟁점이 마땅치는 않았습니다. 이에, 평소 제가 궁금하기 여겼던 몇 가지 사항을 질문드리는 것으로 토론을 대신하고자 합니다.

일부 질문 내용에는 디지털인문학과 전산언어학에 대한 다소 부정적이고 비관적인 뉘앙스의 표현이 일부 들어가 있습니다만, 제가 그렇게 생각한다는 것은 아닙니다^^;; 방대한 데이터에 내재된 오류로 인한 부정확성, 텍스트의 세부적인 맥락을 제거한 채 파편적으로 얻어진 정보를 분석하는 일의 비구체성, 연구 결과가 결국은 우리가 이미 아는 바를 확인하는 수준에 그치는 경우가 많다는 비참신성을 지적받고 실의에 빠진 적이 있었는데요. 그래도 지금의 여건에서 할 수 있는 일을 해 보는 게 의미가 있지 않을까 하는 생각을 가지고 있는데, 발표자 선생님은 어떻게 보고 계시는지 궁금해서 질문드립니다.

1. 데이터 주도 연구(data-driven research)

1990년대의 전산언어학을 소개하는 부분에서 등장한 키워드입니다. 선생님께서는 데이터 기반 연구와 데이터 주도 연구가 어떤 면에서 구별된다고 보시는지 궁금합니다. 그리고 언어 연구에서 데이터 주도 연구라고 할 수 있을 만한 성과가 무엇이 있을지 말씀을 듣고 싶습니다.

저는 좀 극단적으로, '데이터 주도 연구는 불가능한 것이 아닌가' 하고 생각해 본 적이 있습니다. 어떤 연구이든 연구의 시작, 과정을 거쳐 결과에 이르기까지 곳곳에서 연구자의 판단과 해석이 동반됩니다. 그리고 그 판단과 해석에 있어서는 전문가의 도메인 지식이 깔려 있는 것

이고요. 데이터 주도 연구라는 표현에는 마치 연구 당사자의 주관이 배제되는 듯한 뉘앙스가 있는데, 과연 그러한 연구가 가능할 것인가에 대해서 회의감을 가졌던 것입니다.

언어 표현에 있어서 데이터 기반 연구와 데이터 주도 연구가 엄연히 다르고 직관적 느낌에 있어서도 뭔가가 정말 다를 것 같기는 한데, 실제로는 어떠한지에 대해 선생님의 경험에 입각한 말씀을 듣고 싶습니다.

2. 연구의 정밀성 향상

발표 자료에서는 디지털인문학과 전산언어학의 만남을 통해, '연구 범위의 확장', '연구의 정밀성 향상', '학제간 연구자간 협업 촉진' 세 가지가 제시되었습니다.

첫째와 셋째에 대해서는 별다른 이견이 없습니다만, 둘째 '연구의 정밀성 향상'은 정말로 그러한지에 대해 어떻게 입증할 수 있을지 궁금합니다. 보통 대규모 언어자원을 활용하는 연구에 비판적인(대체로 비관적인) 입장에서 제기되는 문제가 연구 자료 또는 내용에 대한 '정밀하지 못한 분석', '임의적인 해석' 등이었던 것 같습니다. 디지털인문학 분야의 연구자들은 여러 가지 통계적 기법을 적용한 데이터 처리를 통해 분석의 정밀성을 높이고자 노력하고 있지만, 그 통계적 기법을 통해 얻어진 데이터와 그에 대한 해석이 '인문학적(문학적, 어학적, 철학적 등)인 것이 맞느냐', '인문학적(문학적, 어학적, 철학적 등)으로 유의미하다고 입증될 수 있느냐'를 두고 이견이 있는 듯합니다.

흔히들 코퍼스가 대표성과 균형성을 가져야 한다고 하지만, 여전히 우리는 충분히 대표적이고 균형적인 '잘 정제된 코퍼스'를 갖고 있지 못합니다. 어느 정도 정제된 코퍼스라 하더라도 여러 가지 오류와 불필요한 정보들이 포함되어 있고요. 방대한 규모의 라벨링되지 않은 원시 데이터를 그대로 가져다 쓰는 방식이 일반화된 지금은 언어 자료의 편향성이 더욱 문제적일 수 있습니다.

요컨대, 데이터의 품질이 의심될 때 기법이 아무리 정교하더라도 정밀한 분석은 어렵다고 보아야 할 것 같은데, 이에 대한 선생님의 견해가 궁금합니다.

3. 인문학 對 디지털인문학

디지털인문학의 영역에 대해 표시한 벤다이어그램에서, 인문학과 디지털인문학의 관계에 대해, 인문학이 더 포괄적인 것으로 제시되었습니다. 선생님께서는 디지털인문학도 인문학에 속

하는 한 세부분야 혹은 영역으로 인식하고 계신 것으로 이해하였습니다. 사실 저도 이러한 입장입니다.

그런데 어떤 입장에서는 인문학과 디지털인문학을 대별해서 보기도 하는 것 같습니다. 특히 전통적 연구 내용과 방법을 견지(또는 고수)하는 입장에서는 디지털인문학이 연구의 태도와 방법이 전혀 다른 별개의 (생경한) 분야라고 보는 것 같고, 또 한편으로 고루한 전통적 인문학을 극복하는 일종의 대안적 인문학으로서의 디지털인문학의 위상과 독자성을 강조하는 입장도 있는 것 같습니다.

물론 이게 해결을 필요로 하는 중요한 문제는 아니라고 생각됩니다만, 선생님께서는 어떻게 보고 계시는지 궁금합니다.

세션1 발표2 '양적 분석 방법을 통한 한국 문학 연구의 확장 : 디지털 인문학의 동향 및 도전 과제'에 관한 토론문

전성규(성균관대학교 국어국문학과)

이 글은 디지털 인문학의 양적 분석 그중 특히 문제 분석과 관련하여 국내외 연구 동향을 살피고 여러 가지 제언을 해주고 있습니다. 탐색적인 연구방법론에 기반하여 다양한 방식으로 이루어지는 양적 연구를 축적해 가는 것이 질적 연구와 긴밀히 연동되어 있는 것이며 기술과 인문학의 "능동적 얽힘"이 디지털 인문학의 학문적 성숙도를 심화시킬 수 있다는 선생님의 기본적인 주지에 저 역시 동의하며 몇 가지 질문 혹은 의견, 선생님이 하신 여러 말들 중 초점화하고 싶은 부분을 말씀 드리고자 합니다.

선생님께서는 초입에서 디지털 인문학에 대한 인문학계의 반응들을 언급하고 계십니다. 문체분석에 대한 사례를 검토하고 계시지만 기본적으로 어떻게 소통할 것인가하는 문제가 전반에 놓여 있다고 생각합니다. "디지털 인문학의 순환론적 방법론"에 대한 의견, "혼종적 사이보그"라는 전제 등 글 곳곳에서 주변에서 들어온 질문들, 거기서 느꼈던 답답함, 문제의식, 이를 통해 만들어지는 답변들을 녹여놓고 계신다는 생각이 들었습니다. 글에서 많이 언급하신 어휘 중에 '다양성'이 있는데요(선생님의 문체적 특징일까요), 디지털 인문학의 학적, 기술적, 해석적 다양성을 위해서는 다양한 주체들이 참여하여 다양한 방법론을 가지고 대상을 탐색하는 다양한 길이 필요한데요, 이러한 자원의 확보, 열려있는 학문적 생태계를 확보하기 위해서는 보다 디지털 인문학이 넓은 범위에서 환대받아야 할 필요가 있다고 저 역시 생각합니다. 디지털, AI가 어느 곳에나 붙여지는 지금, 이것은 하나의 전제가 이미 되어 있습니다. 하지만 연구자들이 접속하기를 주저하거나 꺼려하는 이유는 기술적인 장벽도 물론 있지만, 방법론적인 문제가 보다 크다는 생각을 합니다. 여기서 박진호 선생님의 말씀을 빌려 선생님도 언급하셨습니다만, "무엇을 셀 것인가, 그것을 어떻게 셀 것인가의 문제"를 포함한 "개념의 조작화"의 문제를 초점화해서 이야기해볼까 합니다. "개념의 조작화"는 두 단계 즉 "개념화"와 "조작화"를 동시에 포함하는 말일 텐데요, 각각은 무엇을과 어떻게에 보다 해당한다고도 말할 수 있습니다. 디지털 인문학 연구의 개성, 개별성이 개념의 조작화 과정에서 비롯하는 것이며 다양성 또한 바로 이곳에서 창출한다고 생각합니다. 그리고 이 부분에 대한 고민이 디지털 인문학이 인문학과 깊숙이 만날 수 있는 가능성이 되며 디지털 인문학이 문화연구로 나아갈 수 있는 방향을 제시한다고 생각합니다. 무엇을, 어떻게 측정할 것인가는 언뜻 보면 쉬운 문제일 수 있지만 정말 그렇지 않습니다. 어떻게 측정할 것인가의 문제에는 기술에 대한 방법론이 포함될 테지만 간단한 정량화로도 가능한 부분이며 그것이 무엇을에 해당하는 것을 잘 보여준다면 간단한 계산만으로도 의미있는 방법이 될 수 있다고 생각합니다. 문학 텍스트는 쉽게 수치화할 수 없는 것들을 포함하고 있는데 예를 들면 비가시성, 비명시성 등, 구체적으로 제가 생각하

는 것 중에서 예를 들면 문학 안에서 여성의 공간은 비가시적인데 비가시성(무엇을, 개념화의 문제)을 어떻게 수치화할 수 있을까의 문제, 수치화할 수 없음 혹은 어려움의 성질을 정량적 방법으로 다룸으로써 기계가 읽을 수 있는 형식으로 그것을 변환하는 다양한 방법을 모색하는 일이 필요합니다. 이 과정에서 성격이나 수준을 다양한 기준을 활용하여 판별하고(조작화) 수치를 스펙트럼화하는 일 등이 요청됩니다. 분류기준과 경계를 어떻게 만들 것인가, 그리고 그것은 대상, 상황, 맥락에 따라 어떻게 달라져야 할 것인가에 대한 논의가 요청된다고 생각합니다. 디지털 인문학이 인문학과 보다 넓은 접점을 확보하기 위해서는 문화적 감수성에 대한 깊이 있는 연구가 가능하다는 것을 많이 보이는 일이 중요하다고 생각합니다. 이를 위해서는 메타데이터, 데이터셋이 보다 중요해질 텐데요, 개념의 조작화와 디지털인문학의 성숙도와 관련해 고민하고 계신 점 있으시면 조금 더 듣고 싶습니다.

양적 방법론을 통해 문체 분석을 다각도로 시도하시는 선생님께 문체를 어떻게 정의하시는지 묻고 싶습니다. 양적 방법론이라는 것이 개입했을 때 기존 문학 연구에서의 문체 분석과 어떤 지점이 같고 다른가요? ""문학적인" 문체"의 의미도 궁금하고요, 또 문체 분석을 통해 (한국) 문학의 무엇을 보고 싶으신지도 궁금합니다. 작가, 장르, 젠더 등의 '범주'들을 언급하신 것처럼 양적인 방법론은 점차 보다 상향식 방법으로 문체를 조명하는 것이 가능하게 할 것이고 그런 방향성을 자연스럽게 띠게 되지 않을까 생각합니다. 패턴을 찾아가는 것은 어떻게 보면 언어의 본질을 드러내는 것이라는 생각도 드는데요, 기본적으로 누군가의 언어는 반복적이기 때문입니다. 자주 쓰는 단어가 있고, 그것을 연결하는 방식이 있고, 그런 '구조' 속에 하고 싶은 말을 담아내는 '수행'이 있다고 생각합니다. 언어 구성적인 관점, 역사주의적 관점과도 관련이 된다고 할 수 있겠습니다. 문학이 개인의 수행이지만 개인의 수행으로만 설명되지 않는 지점 역시 여기서 발생하는 것이라고 생각합니다. 저자판별은 문체를 개별적인 것으로 인식하는 경향이 좀 강한 것 같고, 저는 언어가 규칙이지만 또한 개인의 언어와 규칙성 양쪽 사이의 어딘가 즈음에서 탄생한다고 보는 입장에서 특히 문학의 언어는 그 사이를 복잡하게 오고 가고 있다고 보는 입장에서 저자판별이 심층적 차원에서는 그렇게 필요한 부분은 아닌 것 같다고도 생각합니다. 문체라는 개념 자체가 개인적인 어휘집이라는 의미가 좀 강하다고 할 때 양적 방법론을 통해 그 개인성을 벗어나 문학적 패턴을 밝히고자 하는 시도가 이루어질 경우 그것을 명명하기에는 문체 분석이라는 표현이 한계가 있다고도 생각합니다.

어떻게 보면 제가 사이의 어딘가 즈음으로 애매하게 말한 그 부분 즉 수행성 속에 녹아있는 형성 과정에 대한 연구, 그러니까 이 작가는 이와 같은 문체적 특징을 가지고 있다라는 결정적 언급이 아니라 유동적인 과정으로서 확인하는 일 역시 양적 측정으로 필요하다는 생각입니다. 이것을 판단하기 위해서는 '변수'에 대한 논의가 보다 중요할 것으로 생각되는데요, 누구를 독자로 하느냐, 특정 시기 특정 작가나 작가 집단은 어떤 작가들의 글을 열심히 읽었는가, 혹은 어떤 사상의 영향을 받았는가, 등 문체를 형성한 차원의 변수들에 대한 고민 역시 다양한 차원에서 이뤄질 필요가 있다고 생각합니다. 이것이 개별성과 언어 구성적 시점을 교차하면서 변화 양상을 파악할 수 있는 방법이라고 생각합니다. 문체 분석에서 변수를 어떻게 설명

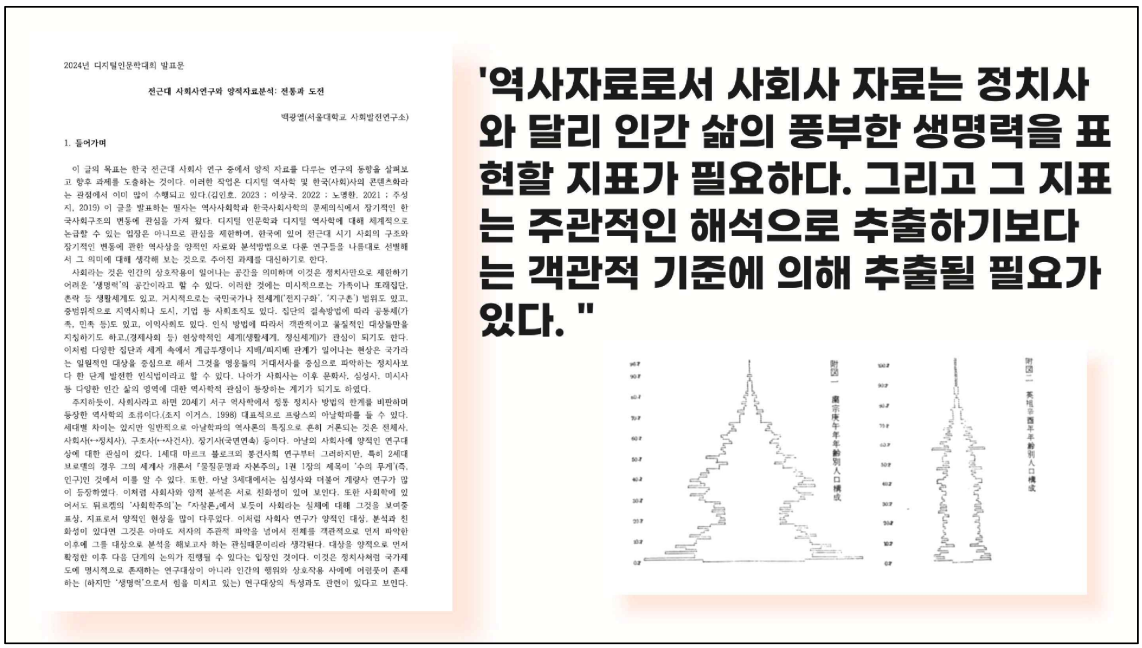
할 수 있고, 구분할 수 있을는지요. 변수를 확인할 수 있는 혹은 변수가 존재한다는 것을 암시받을 수 있는 양적 분석이 필요하고, 변수를 특정할 수 있는 질적 연구가 필요하고, 그것을 다양한 각도에서 규명할 수 있는 양적 연구방법이 필요할 것입니다. 구체적으로 어떤 방법들을 통해 각각의 단계가 가능할까요. 선생님께서 학위논문을 쓰시는 과정에서 해온 고민들 중 말씀해 주고 싶으신 부분이 있다면 부탁드립니다.

마지막으로 생성형 AI를 활용한 서사 자동 생성기 등 디지털글쓰기 환경이 문체에 미치는 영향에 대해서도 듣고 싶습니다. 저자의 죽음을 넘어서 디지털환경 자체가 문학을 생산하고 문학적 관행을 만들어내는 상황에서 문체 연구는 어떤 의미를 가지고 있는 것인지 의견 듣고 싶습니다.

감사합니다.

세션1 발표3 '전근대 사회사 연구와 양적자료 분석: 전통과 도전'에 관한 토론문

신은경(고려대학교 사회학과)



2024년 디지털인문학대회 발표문

전근대 사회사연구와 양적자료분석: 전통과 도전

백광열(서울대학교 사회발전연구소)

1. 들어가며

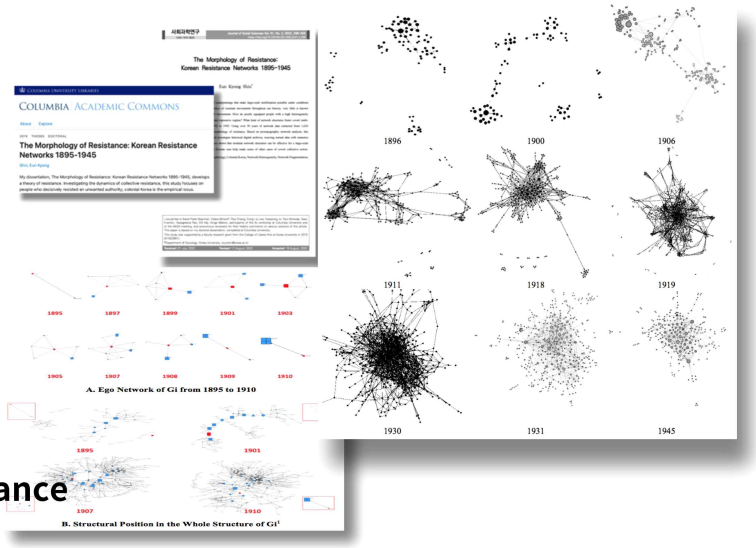
이 글의 목표는 한국 전근대 사회사 연구 중에서 양적 자료를 다루는 연구의 동향을 살펴보고 향후 과제를 도출하는 것이다. 이러한 작업은 디지털 역사학 및 한국(사회)사와 콘텐츠학라는 관점에서 이미 많이 수행되고 있다.(김인호, 2023 ; 이상국, 2022 ; 노명환, 2021 ; 주성지, 2019) 이 글을 발표하는 필자는 역사사회학과 한국사회사학의 문제의식에서 장기적인 한국사회구조의 변동에 관심을 가져 왔다. 디지털 인문학과 디지털 역사학에 대해 체계적으로 논급할 수 있는 입장은 아니므로 관심을 제한하여, 한국에 있어 전근대 시기 사회의 구조와 장기적인 변동에 관한 역사상을 양적인 자료와 분석방법으로 다룬 연구들을 나름대로 선별해서 그 의미에 대해 생각해 보는 것으로 주어진 과제를 대신하기로 한다.

사회라는 것은 인간의 상호작용이 일어나는 공간을 의미하며 이것은 정치사만으로 제한하기 어려운 '생명력'의 공간이라고 할 수 있다. 이러한 것에는 미시적으로는 가족이나 또래집단, 촌락 등 생활세계도 있고, 거시적으로는 국민국가나 전세계('전지구화', '지구촌') 범위도 있고, 중범위적으로 지역사회나 도시, 기업 등 사회조직도 있다. 집단의 결속방법에 따라 공동체(가족, 민족 등)도 있고, 이익사회도 있다. 인식 방법에 따라서 객관적이고 물질적인 대상들만을 지칭하기도 하고,(경제사회 등) 현상학적인 세계(생활세계, 정신세계)가 관심이 되기도 한다. 이처럼 다양한 집단과 세계 속에서 계급투쟁이나 지배/피지배 관계가 일어나는 현상은 국가라는 일원적인 대상을 중심으로 해서 그것을 행동들의 거대서사을 중심으로 파악하는 정치사보다 한 단계 발전한 인식법이라고 할 수 있다. 나아가 사회사는 이후 문학사, 심성사, 미시사 등 다양한 인간 삶의 영역에 대한 역사학적 관심이 등장하는 계기가 되기도 하였다.

주지하듯이, 사회사라고 하면 20세기 서구 역사학에서 정통 정치사 방법의 한계를 비판하며 등장한 역사학의 조류이다.(조지 이거스, 1998) 대표적으로 프랑스의 아날학파를 들 수 있다. 세대별 차이는 있지만 일반적으로 아날학파의 역사론의 특징으로 흔히 거론되는 것은 전체사, 사회사(↔정치사), 구조사(↔사건사), 장기사(국면변속) 등이다. 아날의 사회사에 양적인 연구대상에 대한 관심이 컸다. 1세대 마르크 블로크의 봉건사회 연구부터 그러하지만, 특히 2세대 브로델의 경우 그의 세계사 개론서 『물질문명과 자본주의』 1권 1장의 제목이 '수의 투계'(즉, 인구)인 것에서 이를 알 수 있다. 또한, 아날 3세대에서는 심성사와 더불어 계량사 연구가 많이 등장하였다. 이처럼 사회사와 양적 분석은 서로 친화성이 있어 보인다. 또한 사회학에 있어서도 뒤르켐의 '사회학주의'는 『자살론』에서 보듯이 사회라는 실체에 대해 그것을 보여줄 표상, 지표로서 양적인 현상을 많이 다루었다. 이처럼 사회사 연구가 양적인 대상, 분석과 친화성이 있다면 그것은 어떠도 서사의 주관적 파악을 넘어서 전체를 객관적으로 먼저 파악한 이후에 그를 대상으로 분석을 해보고자 하는 관심때문이리라 생각된다. 대상을 양적으로 먼저 확정한 이후 다음 단계의 논의가 진행될 수 있다는 입장인 것이다. 이것은 정치사처럼 국가체도에 명시적으로 존재하는 연구대상이 아니라 인간의 행위와 상호작용 사이에 어렴풋이 존재하는 (하지만 '생명력'으로서 힘을 미치고 있는) 연구대상의 특성과도 관련이 있다고 보인다.

'역사자료로서 사회사 자료는 정치사와 달리 인간 삶의 풍부한 생명력을 표현할 지표가 필요하다. 그리고 그 지표는 주관적인 해석으로 추출하기보다는 객관적 기준에 의해 추출될 필요가 있다. "

Digital Humanities 1.0



抵抗の形態学

Morphology of Resistance

ID 777: Gi, Wooman (기우만, 奇宇萬, 1846. 8. 17~1916. 10. 28) Hometown: Cheonnam

Gi, a grandson of a vice minister Gi Jungjin, was born on October 17th 1846. (……) In **January, 1896**, after receiving a manifesto from Yoo Inseok (ID 603), Gi made an appeal in writing to mobilize militia to protect Korean dynasty from Japan. **On 11th, February 1896**, Gi went to Naju with Go Gwangsun (ID 1145), Gi Samyeon (ID 1159), Gijuhyeon, Yang Sangtae, Gi Donggwan, Lee Seunghak (ID 1315), Gi Dongno. In Naju, Gi mobilized more people to the militia and performed ancestral rites at the local Kim's shrine. While the group marched to Gwangju province, Gi went into a mountain writing poems advocating traditionalism and patriotism. **In 1904**, Gi also moved to Gwangju. In January, Gi was appointed as official by the provincial governor Lee Dojae, but he didn't accept the offer because he didn't trust Lee. (……) **In 1911**, after Korea officially lost its sovereignty, **Gi retired to hermitage** in the mountain and didn't cooperate to any legal obligation. And several years later, he passed away in the underground tunnel of the mountain.

Texts

Biography of A: *A cooperated with B and C at T1* (Coding: entering A, B, C at T1)
Biography of B: *B had a meeting with A, D and E at T2 and got arrested and killed at T3* (Coding: entering E at T2 and exiting B at T3)
Biography of C: *C planned a terrorist attack at T3 with A and F* (Coding: entering Fat T3)
Biography of D: *D helped A and G at T1*(Coding: entering D, G at T1)

Matrices

Adjacency Matrix at Time 1

.	A	B	C	D	E	F	G
A	.	1	1	1	0	0	0
B	1	.	0	0	0	0	0
C	1	0	.	0	0	0	0
D	1	0	0	.	0	0	1
E	0	0	0	0	.	0	0
F	0	0	0	0	0	.	0
G	0	0	0	1	0	0	.

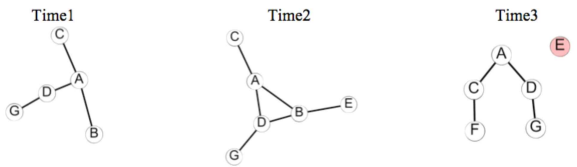
Adjacency Matrix at Time 2

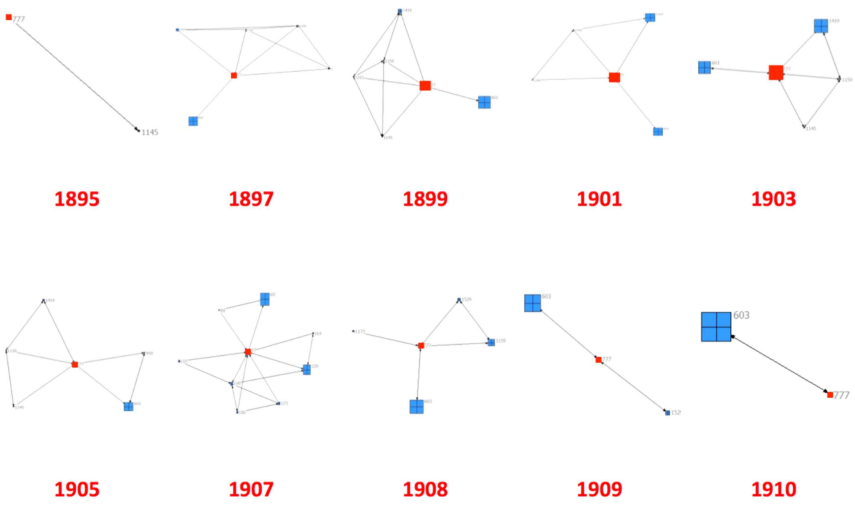
.	A	B	C	D	E	F	G
A	.	1	1	1	0	0	0
B	1	.	0	1	1	0	0
C	1	0	.	0	0	0	0
D	1	1	0	.	0	0	1
E	0	1	0	0	.	0	0
F	0	0	0	0	0	.	0
G	0	0	0	1	0	0	.

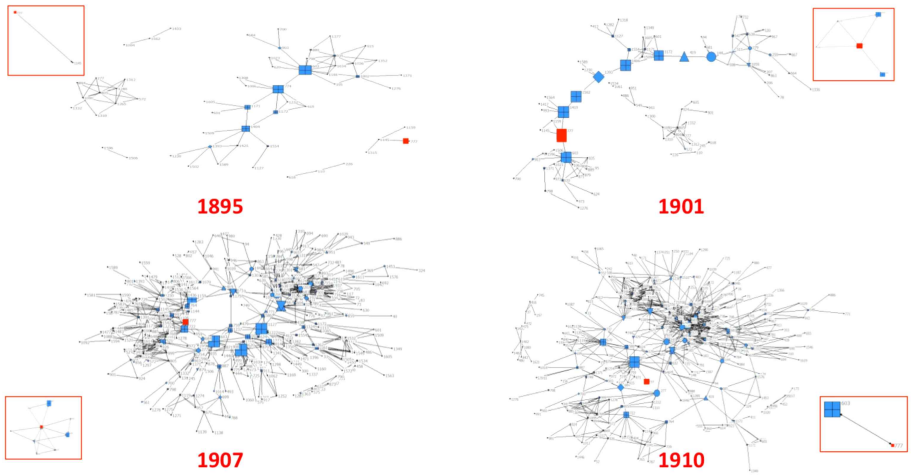
Adjacency Matrix at Time 3

.	A	B	C	D	E	F	G
A	.	.	1	1	0	0	0
B	.	.	.	.	.	.	.
C	1	.	.	0	0	1	0
D	1	.	0	.	0	0	1
E	0	.	0	0	.	0	0
F	0	.	1	0	0	.	0
G	0	.	0	1	0	0	.

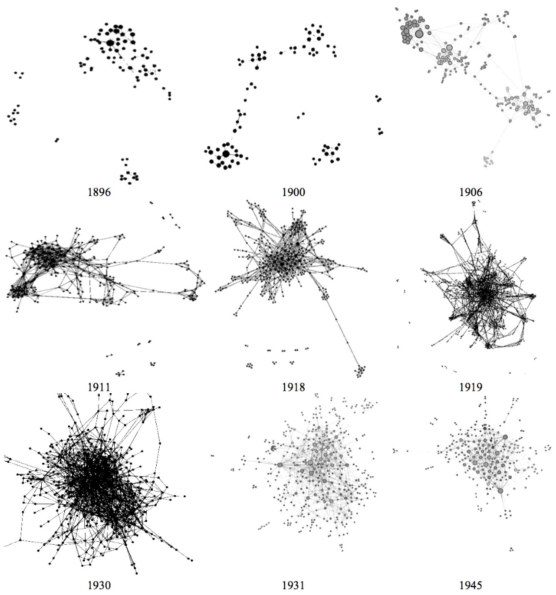
Graphs







Diamond: Domestic Resistance Movement; Thing: Enlightenment Movement; Up Triangle: March 1st Movement; Circle: Provisional Government of Korea in China; Box: Militia; Circle in Box: Army; Down Triangle: Activists in Russia; Square: Terrorists; Round Square: Activists in the US; Plus: Cultural activists; Unshaped: Student Movement / Japan




감사합니다
eunkyongshin.com

세션1 발표4 '데이터로 보는 동양고전 연구, 그 현황과 과제'에 관한 토론문

이상엽(서울대학교 철학과)

본고는 현재 동양고전을 대상으로 이루어지고 있는 디지털인문학 연구의 현황을 "1) 데이터베이스의 확장" 및 "2) 연구를 위한 도구와 플랫폼 개발"이라는 두 가지 측면에 초점을 맞추어, 비판적이며 전체적인 조망을 제공하는 글이다.

우선 데이터베이스의 확장을 논하면서는 연구자들이 데이터베이스에 축적하게 되는 "데이터"라는 것 자체가 "태생적으로 편견을 내포"하고 있으며 "필연적으로 해석자의 주관과 맥락이 반영될 수 밖에 없음"을 상기시키며 데이터에 대한 비판적이고 명료한 이해가 디지털 인문학 연구에 전제되어야 함을 주장한다. 이 문제는 특히 필자가 이 글에서도 지적하듯이 디지털 인문학에서 분석 대상이 되는 주 데이터가 "문자로 기록된 '텍스트'"라는 점과 연관하여 생각할 때 특히 절실하게 다가온다. 어디까지나 식자층이라는 소수의 특수한 경제적·사회적·제도적 집단의 관점에서 집적된 데이터이므로, 종교, 문화, 언어와 같이, 식자층이라는 집단의 테두리 내로 귀속되지 않는 전사회적 인문현상을 연구하는 데에 있어 현재 우리가 수집가능한 데이터가 대표성의 측면에 있어 한계를 지니고 있다는 것이다. 이러한 문제는 디지털 인문학 연구자들이 항상 주의해 온 바라고 생각하며, 앞으로도 이러한 비판적 의식과 성찰은 계속되어야 할 것이다. 또한 이러한 문자 데이터의 태생적 한계를 극복할 수 있는 방안에는 무엇이 있을지 함께 고민해봐야 하지 않을까 생각된다.

이 글의 "데이터베이스의 확장"에 대한 논의에서는 또한 현재 한국, 나아가 세계 곳곳에서 개발되고 있는 각종 동양고전 데이터베이스가 소개하고 있어 독자로서 큰 도움을 받았는데, 불교문헌의 전산화 및 불교문헌을 기반으로 한 데이터베이스에 대한 논의가 완전히 빠져있는 것은 조금 의아했다. 가령 대만의 중화전자불전협회(CBETA)는 1997년에 창설되어 2007년에 이미 성경의 130배 분량에 해당하는 『대정신수대장경』의 전산화를 TEI표준에 따라 완료한 바 있다. 이 데이터베이스는 현재 전세계의 한자문화권 불교 연구자들에 의해 매일과 같이 쓰이고 있으며, TEI표준을 따라 구축된 관계로 무한한 확장성을 가지고 2차적, 3차적 데이터베이스를 지속적으로 파생시키면서 전기 데이터베이스 구축, 네트워크 분석, n-그램 분석, 통계학을 활용한 코퍼스 분석, 인공지능을 활용한 다언어 문헌간 대응개소 발견 등 다양한 방식으로 활용되고 있다(Marcus Bingenheimer, Michael Radich, Sebastian Nehrdich 등 연구자의 성과 참조). 한문 불교문헌의 전산화와 한문 불교문헌의 디지털인문학 연구 현황을 이해하는 것으로써, 선진철학이나 성리학 문헌의 디지털인문학 연구도 보완될 수 있는 바가 있을 것이라고 생각된다. 이 점에 대해서 첨언하자면 중국고전의 경우 TEI 표준에 따라 전산화된 DB가 없고, 이로 인해 한 문헌의 다양한 판본을 한 번에 검색하는 방법도 존재하지 않다고 논평자는 알고 있는데, 논평자의 이해가 맞는지, 맞다면 향후 한국 학계에서 이러한 프로젝트를 주도할 계획은 없는지 묻고 싶다.

이 글의 두 번째 주된 주제인 연구 도구 및 방법론, 플랫폼에 대한 성찰에서는“문자 텍스트에 구문론적 분석 수단을 적용하여 의미론적 또는 해석학적 분석으로 나아가는 데”에나 “구문 분석 도구를 사용해 철학적 문제나 기존 연구 문제를 검증하려 할 때” 따르는 어려움에 필자는 주목한다. 이 역시 디지털인문학 방법론을 불교 지성사, 불교 제도사 연구에 활용해 온 논평자가 공감하는 문제이나, 논평자의 연구를 비롯하여 논평자가 알고 있는 많은 불교학의 연구는 이러한 문제를, DH 방법론을 통해 착상, 발전시킨 논지를 DH와는 무관한, 전통적인 수치적, 역사적 자료들의 제시와 논의를 통해 이중으로 뒷받침하는 방식을 통해 극복하려고 한다. 가령 엘리트 학승들이 신이(神異)나 기적과 같은 종교활동을 행하지 않았다는 패턴을 DH 방법론을 통해 발견하면, 우선 이러한 패턴을 DH 방법론에 익숙하지 않는 사람도 이해하기 쉬운 전통적인 수치로도 제시하고(학승 중 몇 퍼센트만이 신이를 행했지만 학승이 아닌 승려 중 몇 퍼센트가 신이를 행함 등), 또 이러한 패턴을 동시대인의 시점에서 기술하고 있는 기록을 찾아서 보충하는 식의 양갈래 전략(two-pronged approach)을 취하는 것이다. 적어도 논평자의 좁은 소견으로는 불교학의 분야를 벗어나서는 이와 같이 DH 방법론의 결론을 비-DH 방법론으로도 이중으로 입증하는 논문을 본 기억이 없는데, 이러한 접근법이 필자가 제시하는 연구 방법론 문제에 대한 하나의 가능한 해법이 될 수 있을지 생각을 여쭙고 싶고, 또 만약 그러한 좋은 연구가 있다면 이 자리를 빌려 소개받고 싶다.

세션2

기관에서의 디지털 인문학

○ **주요 기관 제공 역사자료 데이터베이스: 특성, 지속가능성, 활용의 방향**
류준범(국사편찬위원회)

○ **지역 소재 고문헌 자료 아카이브의 구축 및 활용 방안과 시사점**
김사현(한국유교문화진흥원)

○ **국가유산 디지털 아카이브 구축 현황 및 향후 과제**
이종욱(한국전통문화대학교 디지털헤리티지학과)

○ **AI 시대, 국가지식문화자원 데이터 허브, 국립중앙도서관**
김수정(국립중앙도서관)

세션2 발표문 1

주요 기관 제공 역사자료 데이터베이스 : 특성, 지속가능성, 활용의 방향

류준범(국사편찬위원회)

〈목 차〉

1. 들어가며

오늘 이 발표에서 이야기하고 싶은 기본 주제는 '디지털화된 역사 자료와 역사 연구의 확장'에 관한 것이다. 21세기 들어 역사 연구의 기본 자료인 문자 자료들이 종이 등의 매체를 떠나 디지털화된 형태로 제공되기 시작하였고 이제 그 수량이 매우 크게 늘어나 역사 자료의 이용이 기본적으로 디지털화된 형태의 것을 기반으로 하고 있다. 그리고 적어도 한국에서는, 국사편찬위원회를 포함하여 역사 자료 편찬에 관여하는 다양한 기관들이 디지털 역사 자료의 제공자 역할을 하고 있고 역사학계가 그 수요자로서 관계를 맺고 있다. 좀 더 시야를 넓혀, 이른바 '공공 역사' 분야까지 포함하여 역사 자료를 기반으로 다양한 역사 연구를 진행하고 그 결과물을 내는 집단과, 안정된 조직과 예산을 바탕으로 디지털화된 역사 자료를 제공하는 기관들 사이의 관계를 한번 검토해 보는 것이 역사학 연구 전반의 확장을 위해 필요한 일이 아닐까 생각된다. 그렇지만, 이미 이 발표를 들으시는 분들이 잘 알고 계시듯이 역사 연구 집단과 역사 관련 기관 사이가 단순히 수요자와 제공자의 관계인 것만은 아니다. 실제로 다양한 역사학 관련 종사자들이 다양한 프로젝트에서 생산자와 이용자의 자리를 번갈아 점하고 있고 각각의 자리에서 고마움과 아쉬움-때로는 그보다 더 강렬한 감정까지도-을 느끼고 있다.

이 발표에서 발표자의 의도는, 역사학 연구의 확장을 위해 디지털 역사 자료를 어떻게, 어떤 방향으로 활용해 볼 것인가를 이야기해 보는 것이다. 현재 한국사 관련 자료만으로도, 수천만 건, 수백억 자에 이르는 디지털 역사 자료가 웹을 통해 제공되고 있다. 이 같은 환경에서, 크게 보아 역사학 연구 분야의 종사자들이 현재의 역사 자료 기반 상황을 점검하고 이를 활용하여 역사학 연구를 어떻게 확장할 것인지 이야기를 나누는 약간의 계기가 되기를 바라는 마음이다. 이제 이어, ① '검색과 온라인 열람'이라는 기본적 측면에서 디지털 역사 자료의 현황을 간략히 살펴보고 ② 공개 데이터를 기반으로 역사 연구의 '공구(工具) 데이터'를 연구자가 제작하는 것이 역사학 연구 확장에 필요한 일이라는 의견을 제시하고 ③ 역사 관련 기관들이 기존처럼 1차 역사 자료의 디지털 기반 제공이라는 역할만에 집중하기는 어려우니 역사학계와 역사 관련 기관 간의 협조와 이해를 위한 좀 더 의식적인 노력이 필요할 것이라는 말씀을 드릴 것이다.

2. 웹을 통해 접근할 수 있는 디지털 역사 자료의 현황과 특성을 간단히 살펴보자

역사학 분야에서 본격적인 디지털화 작업은 2000년을 전후하여 한국전산원의 지식정보자원관리사업을 통해 이루어졌다. 이때 국사편찬위원회, 한국학중앙연구원, 서울대학교 규장각, 한국고전번역원 등을 중심으로 역사 자료 디지털화 사업이 시작되었고 이후 다양한 역사 관련 기관이 참가하게 되었다. 이들 기관의 역사 자료 데이터를 연계하여 통합 검색을 제공하는 서비스가 한국역사정보통합시스템이었다. 많은 비용을 들여 디지털화를 진행한 이유는 명백하다. 역사 자료에 대한 접근 편의를 높이기 위함이다. 그 핵심은 검색과 온라인 열람이다. 역사 자료를 필요로 하는 이들이, 굳이 도서관에 가거나 사료집을 사 모으지 않아도 사료를 빠르게 찾아서 보고, 읽게 하자는 것이다. 역사 자료를 필요로 하는 이들이 대상이므로 누구에게나 개방된 웹 기반으로 디지털화된 역사 자료가 제공되었다. 역사 유관 기관들의 이와 같은 작업과 병행하여 국립중앙도서관·국회도서관을 중심으로 전자도서관 사업이 진행되었고 이후 국가기록원의 아카이브 전산화 작업이 진행되면서 디지털 역사 자료의 웹 기반 서비스가 일반화되었다.

도서관, 기록관(문서관) 등은 소장 자료의 목록 디지털화를 바탕으로 디지털화된 원문을 제공하는 방식이 기본적이었던 데 비해 역사 유관 기관들은 역사 자료의 편찬에 기반하였다는 점이 차이라고 할 수 있다. 따라서 각 기관의 개별 데이터베이스를 살펴보면 그 구성과 기능에 차이가 꽤 컸다. 조선왕조실록 데이터베이스와 국립중앙도서관 고문헌 데이터베이스를 비교하면 간단히 그 차이를 알 수 있는데 도서관 고문헌 서비스가 정돈된 메타데이터와 품질 좋은 원문 이미지 제공을 중심으로 이루어진 반면 실록의 경우 문자 하나하나에 대한 검색 제공에 큰 의의를 두었다. 기관 자체의 운영 목적과 역사에 따라 데이터베이스 성격에도 약간의 차이가 있었다. 그렇지만 20여년의 기간 동안 이 프로젝트들이 진행되어 온 결과를 돌아보자면 데

이터베이스 결과물은 앞서 말한 검색과 온라인 열람이라는 주목적 하에 상당히 비슷해지는 경향도 크다. 원사료의 질서에 따른 데이터 구성을 바탕으로 적절한 데이터 단위를 나누고 그에 기반한 메타데이터, 원문 텍스트, 원문 이미지 제공이라는 구성으로 수렴하는 측면도 크게 작용하고 있다. 한편 각 기관들의 상황을 스케치하여 보면, 처음에는 역사 자료의 디지털화 사업이 각 기관의 중심 사업의 곁에서 보조 사업처럼 시작된 반면 현재는 디지털화라는 방식 자체가 각 중심 사업의 기반을 이루게 되었다. 국사편찬위원회를 예로 들자면 -국편의 주요 사업 중 하나가 사료의 편찬 사업인데, 사료의 편찬 업무와 사료의 디지털화 업무가 따로 진행되다가 사료 편찬 업무 자체가 디지털 기반으로 변화되었다. 현재는 사료 편찬이 디지털 형태로 이루어지고 그 결과물도 기본적으로 디지털 기반으로 웹을 통해 제공하는 방식으로 진행되고 있다. 이 상황은 지금껏 디지털 역사 자료 기반 제공자 역할을 해 온 주요 역사 유관 기관들이 앞으로도 지속해서 이 역할을 수행할 것인가라는 질문과도 연결되는데 이 주제는 따로 5번 항에서 다루어 보겠다.

역사 자료 활용을 위한 기본 기능인 '찾아서 본다', 즉 필요한 자료를 검색을 통해 찾고 찾은 자료를 온라인 상으로 본다라는 기능의 작동 범위는 '찾아볼 수 있는 자료의 양'에 달려 있다. 2023년 1월 무렵 한국역사정보통합시스템에서 검색 가능한 기사 수는 13,721천 건이었고 기관 수는 스무개였다. 한국역사정보통합시스템의 서비스가 본격적으로 개시된 2004년 이후 검색 가능 자료 수는 지속적으로 증가하였고 기관 수는 출입이 있기는 하였지만 2010년 이후 대체로 스무 기관 전후를 유지하였다. 현재 역사 자료 데이터 연계·통합 검색 기능은 한국역사정보통합시스템을 계승한 한국학자료통합플랫폼(한국학중앙연구원 운영)에서 담당하고 있다. 23년말에 27개 기관, 91개 데이터베이스를 연계하고 있었고 현재는 그 수가 늘었다. 한국학자료통합플랫폼 연계 기관·DB 검색 페이지(https://kdp.aks.ac.kr/service/organDbList)를 통해 확인할 수 있는데 현재 33개 기관으로 확인된다. 고대부터 현대까지 다양한 문자 사료들이 디지털화되어 웹을 통해 제공되고 있고 역사 연구의 전체 분야에서 이렇게 제공되는 자료를 활용하여 연구가 진행되고 있다. 여기서 역사 자료를 찾아서 본다는 것의 범위가 어느 정도인지 간략한 사례 하나를 들어 보자.

전라남도 신안 하의도는 조선시대 궁방전 절수에서부터 시작된 토지소유권의 오랜 분쟁 지역으로 유명하다. 하의도에서는 '내 땅 찾기 300년 역사의 섬'이라고 이야기한다. 인목대비의 딸로 인조반정 이후 중요한 인물이 된 정명공주와 그 남편 풍산 홍씨 홍주원의 제사 등을 위해 하의도 일대가 정명공주방으로 절수되었다. 당시 절수의 의미가 사여인지 수조를 위한 단순 절수인지 등이 문제가 되었고 또한 절수지의 범위 문제가 크게 대두되었다. 하의도는 조선 후기 때 다시 개간된 땅이 많은데 이때 정명공주방에서 절수지의 범위를 하의도 전체로 확장하려 하였다. 여기에 지방관청(나주)에서 이중 과세하는 문제까지 겹쳐 영조 때까지 갈등이 지속되었다. 정조 당시 궁방전 정리 때에 정명공주방 절수지가 유지된 것이 확인되고 철종 때의 것도 확인된다. 실록, 승정원일기, 일성록, 비변사등록 등 연대기와 등록 자료가 대부분 검색

가능하기에 정명공주방과 관련한 조선 후기 자료들은 거의 모두 확인할 수 있다. 해당 토지는 하의도 주민들의 소유권 주장에도 불구하고 1900년 무렵 내장원 토지로 귀속되었다가 제실재산정리 사업 때 홍주원의 후손 洪祐祿의 주장으로 왕실 재산에서 홍씨 개인 재산으로 넘어간다. 이때부터 하의도 주민들과 홍우록 이후 매매자들-홍우록이 해당 토지를 매매하여 이후 매매 지속-과의 다툼이 법정과 현지에서 지속된다. 이들에 관한 자료는 각군소장과 재산조사국 거래문을 비롯한 대한제국 시기 공문서 등록 자료와 황성신문, 대한매일신보 등 당시 발행 신문 기사에서 거의 확인된다. 결국 토지조사사업 때 소유권 판정이 당시 지주라고 주장하던 右近權左衛門에게 돌아 갔다는 것은 국가기록원 지적아카이브의 자료를 통해 확인 가능하다. 일제 시기에는 주민들의 소유권 요구와 함께 당시 현실을 인정한 소작료 인하 요구가 주민 운동으로 지속되며 이는 일부 일제 시기 경찰 자료와 동아일보 등 신문 자료를 통해 확인할 수 있다. 그리고 이 토지가 해방 후 신한공사 관리로 넘어가면서 다시 분쟁이 시작되었는데 이에 관한 연대기 기사는 자료대한민국사를 통해 확인 가능하다. 결국 농지개혁 당시 섬 주민에게 유상분배되는 것으로 하의도 농지의 소유권 문제는 마무리되는데 이는 국가기록원의 분배농지부 문서를 통해 확인할 수 있다. 조선 토지 및 조세 제도의 특징인 절수부터 개간과 소유권 문제, 근대 시기 토지에 대한 권리의 처리 문제와 권력 관계, 농민 운동과 농지개혁까지 300년에 걸친 하의도 농지 사건에 대해 웹을 통해 매우 다양한 자료를 다양한 출처로부터 얻을 수 있다.

역사 자료 디지털화가 본격적으로 시작되며 목표로 한 자료에 대한 접근성 확보는 이미 상당한 수준에 올라와 있다고 보인다. 역사 연구의 측면에서 보자면 접근 가능하고 활용 가능한 사료의 범위가 크게 확대된 것이며 관련 자료의 입수 비용(특히 시간)이 크게 줄어든 것이다. 자료의 범위는 크게 확대되고 자료 입수의 비용은 크게 줄어들어 보다 넓고 깊은 자료를 기반으로 역사 연구를 확대할 여건은 되지 않았나 생각된다.

3. 온전한 형태의 디지털 데이터가 제공되고 있다-확대가 필요하며 '관행' 또한 정돈될 필요가 있다

웹을 통한 검색과 열람 범위의 확대는 역사 연구에서 접근할 수 있는 자료의 양을 늘리고 그 시간을 단축하였다. 하지만 디지털화된 역사 자료를 연구 데이터로 활용하기 위해서는 기계가 읽을 수 있는 형태로 데이터 전체를 확보하여 활용할 필요가 있다. 데이터의 분석, 상호 연계, 재구성 등의 과정을 거쳐야 디지털 역사 자료가 역사 연구의 데이터로 온전히 활용될 수 있을 것이다. 그러려면 우선 데이터 자체를 확보할 필요가 있다. 이 같은 요구는 오래 전부터 제기되어 개별적인 형태로 기관에서 연구 집단으로의 자료 제공이 이루어진 사례 자체는 여럿이다. 그런데 기계가 읽을 수 있는 형태의 데이터 묶음 전체를 제공하는 일은, 소위 '공공데이터 개방법' 이후 공공데이터 포털이 활성화된 이후의 일이다. 현재 국사편찬위회는 공공데이터

포털을 통해 82건의 데이터 파일을 제공하고 있고 별도로 역사GIS 데이터를 역사지리정보DB 사이트를 통해 제공하고 있다. 디지털화된 역사 자료를 역사 연구에 온전히 활용하려면 이 수준의 데이터가 연구 집단에 제공되는 기회가 확대되어야 할 것은 물론이다.

그런데 기계가 읽을 수 있는 형태의 전체 데이터 제공이라는 일은 겉보기와는 달리 좀 더 세심한 접근이 필요한 일이다. 디지털 역사 자료의 제공과 활용이 더욱 확대되기를 바라는 마음에 이 작업과 관련하여 두 가지 점을 이야기하고자 한다.

하나는, 어쩔 수 없는 행정 상의 불합리와 그와 관련된 신뢰 부족의 문제이다. 법률과 규정에 따른 공공기관의 보유 데이터 개방이라는 절차는, 그 의도는 분명하며 문명적이지만 그 운영 실제에서는 어려움이 발생한다. '데이터 개방'이 이미 정규화된 테이블 형태의 데이터-제공 형식이 csv이든 json이든 기본적으로 현재 공공기관이 생산하는 데이터가 주대상이다 보니 현대 사회의 매우 규격적이고 정규화된 데이터가 주된 대상이 된다. 이에 맞추어 정비된 절차와 기준을, 생산 당시 전혀 정규적이지도 않고 현재의 관행과 수준과는 다른 역사 자료에 적용하다 보니 조금은 엉뚱한 일이 생기기도 한다. 여기에 우상향만 인정하는 각종 개방 성과의 적용 등이 이루어지면 데이터 개방 업무가 비효율적으로 운영되기 쉽다. 또 각 기관의 담당자에게는 이 업무는 부가적 업무일 수밖에 없다. 여기에 데이터 제공 요청을 일반 행정 기관의 민원처럼 처리하는 경우까지 겹치면 이 일은 역사 연구의 발전을 위한 협력 업무가 되기가 어렵다. 데이터 제공과 수취가 역사 연구를 위한 협력적 업무가 되도록 신뢰가 쌓일 필요가 있을 것이다.

두 번째는 제공되는 데이터에 관한 '제3자 권리'의 문제이다. 데이터 소스에 대한 저작권적 우려가 자료에 존재할 수 있다. 역사 자료의 대부분은 저작권이 보호하는 시기 이전의 자료이지만 근현대 시기 자료로 내려오면 이 부분은 민감해질 수 있다. 저작권리 이외에 -법적으로 명문화된 권리는 아니지만 소장자의 권리 문제도 제기된다. 특정 자료를 디지털화하면서, 예컨대 소장자 등에게 국사편찬위원회 편찬 업무와 한국사데이터베이스를 통한 연구용 제공에 한정하여 사용 허가를 받는 경우가 이에 해당된다. 최근에는 각 기관들 사이의 자료 교환도 일상적으로 이루어지는데 이 경우에도 '제3자 권리' 문제가 발생할 수 있다. 또 하나 더 역사학계를 비롯하여 기관들에서 검토할 문제는 저작인격권과 관련된 것이다. 주요 역사 자료 데이터베이스는 단순 입력으로 제작되는 것이 아니라 구성, 해설, 주석, 교감 등의 작업과 함께 제작된다. 이 작업에는 많은 관련 연구자가 개입되는데-기관의 연구직원만이 아니라- 대개 저작권 자체는 비용을 부담하는 기관이 공동 소유하지만 저작인격권까지 무시될 수는 없을 것이다. 해당 데이터를 그 수준으로 만드는 데 기여한 연구자들이 최소한 자신의 저작인격권을 지킬 수 있는 방안이 논의되기를 바란다. 삼일운동데이터베이스를 예로 들자면, 해당 데이터베이스는 삼일운동 관련 다양한 1차 사료를 개별 시위 사건 단위로 재구성한 데이터이다. 개별 시위 사건에 대한 해설 기사는 집필자의 이름을 밝힘으로써 저작인격권 보호를 고려하였다.

소요사건관계서류 같은 혼잡스러운 자료를 분석하여 개별 시위 사건 단위로 정보를 재구성하는 작업 또한 여러 연구자들이 수년간 노력한 결과이다. 딱히 꼭 집어 방안을 제시하기는 어렵지만 디지털 역사 자료 제작 프로젝트와 관련하여 저작인격권 보호 문제는 차분한 논의가 있기를 기대한다. 특히 한국연구재단과 한국학중앙연구원의 한국학 토대 사업 결과물에 관해서도 원천 데이터 공개 논의가 있는데 이와 관련하여서도 저작인격권 문제는 차분히 따져볼 필요가 있겠다.

디지털 역사 자료를 활용하여 역사학 연구를 확장하기 위해서는 기계가 읽을 수 있는 온전한 형태로 데이터가 제공되는 것이 필수적이다. 이를 위해 데이터 제작, 제공의 역할을 담당하고 있는 주요 역사 분야 공공기관의 적극성이 필요하다. 그리고 한편으로 데이터 권리 문제를 포함한 다양한 현안을 해결하기 위한 역사학계 전반의 협력적 노력도 필요하리라 생각된다.

4. 역사학 연구를 위한 데이터 확장을 위해 연구자들이 자신의 공구 데이터를 만드는 것이 필요하지 않을까

이 항목은 기계가 읽을 수 있는 온전한 형태의 디지털 역사 자료를 어떻게 활용할 것인가에 관한 것인데 사실 조심스러운 주제이다. 각각의 연구자, 연구 집단이 다양한 프로젝트를 통해 그 성과를 축적하고 있고 앞으로도 그럴 것이다. 조심스럽긴 하지만 디지털 데이터를 활용하여 연구용 공구 데이터를 만드는 작업이 널리 진행되어야 한다는 의견을 제시하여 본다.

한국고전번역원이 운영하는 한국고전종합DB에는 저자행력정보, 인물관계정보가 제공되고 있다. 문집의 행장을 모아서 만든 사전형 데이터베이스가 저자행력정보이고 고전종합DB를 만드는 과정에서 축적된 인물 정보를 바탕으로 인물 관계망까지 제공하는 것이 인물관계정보이다. 한국학중앙연구원이 운영하는 한국학자료통합플랫폼에서는 전통인물, 근현대인물이라는 주제로 주제 검색 서비스를 제공하는 데 인물 관련 여러 데이터베이스를 통합하여 제공하는 서비스이다. 한국국학진흥원에서는 필사본 자전이라는 이름으로 초서 자체의 데이터를 제공하고 있다. 각 기관들이 공구 자료 형태의 데이터 서비스를 제공하는 것은, 역사 자료 데이터 구축 과정에서 축적된 정보를 바탕으로 자료 이용자들의 공구형 데이터에 대한 수요에 대응하기 위한 것이다.

디지털 역사 자료를 모아서 엮고 정비함으로써 시도해 볼 수 있는 역사 연구 확장의 중심에 이 같은 공구 데이터 제작이 있다고 생각된다. 아마도 각 기관들이 위와 같은 공구형 데이터 제공을 늘려 나가겠지만 개별 역사 연구에 필요한 수요를 모두 맞추어 충족시킬 수는 없다. 오히려 각 연구 프로젝트에서 역사 연구자들이 자신에게 필요한 공구형 데이터를 직접 만드는 방식으로 역사 연구를 확장할 필요가 있다고 생각된다. 역사학 연구에서는 개별 연구 주제의

직접 대상 자료보다 그 주제를 이해하기 위한 주변 자료의 양과 규모가 중요하다. 인물, 지명, 용어의 용례 등 그 시대와 사건을 이해하기 위해 해당 주제와 직접 관련 없어 보이는 다양한 지식이 축적되어야 한다. 현재 축적되어 있는 디지털 역사 자료라면, 이제 이와 같은 시도가 개인 연구자 수준에서 가능하지 않을까. 각종 읍지의 토산을 물명 자료와 연결시키고 조선시대 법령 자료에 나오는 다양한 제도적 용어를 실제 용례와 연결하여 정리하는 작업이 가능한 수준까지 왔다고 보인다. 디지털화된 데이터의 양, 디지털 데이터를 다루는 도구의 기능 수준 등을 고려할 때 예전에는 생각이 있더라도 실행에 옮기기 쉽지 않은 일들을 지금은 시도해 볼 수 있다.

한편으로 데이터과학의 기법을 도입한 역사학 분야의 관련 연구들도 역사 연구와 관련해서는 공구로 쓸 수 있는 정보를 제시하는 것이라 생각된다. 다시 한번, 조심스럽지만 역사학 분야에서 디지털 데이터의 활용은 우선 대상 시대와 사건을 이해하기 위한 공구 데이터의 제작이 중심이 되지 않을까 제시하여 본다.

5. 역사 관련 기관들은 기반 시설 제공자로서의 역할을 지속하겠지만 그에 머물러 있기 힘들 것이다

국사편찬위원회 주관으로 역사정보화 관련 기관 협의회가 정례적으로 열리어 역사 자료 정보화의 현황을 점검하고 관련 의견을 교환하고 있는데 이에 참가하는 국사편찬위원회, 동북아역사재단, 서울대학교 규장각한국학연구원, 한국고전번역원, 한국국학진흥원, 한국학중앙연구원 등이 지금껏 디지털 역사 자료의 제작, 제공에 중심적 역할을 수행하여 왔다. 이들 기관이 한국학중앙연구원의 한국학자료통합플랫폼에도 주요 자료 제공기관으로 참가하고 있다. 도서관 정보화와 공공기록물 정리, 제공을 담당하는 국립중앙도서관과 국가기록원 그리고 소장 유물의 디지털 목록과 이미지를 제공하는 국립중앙박물관이라는 큰 기관을 제외한다면 앞으로도 디지털 역사 자료의 제공은 앞서 말한 역사 관련 기관들이 중심이 되어 진행될 것이다. 이 기관들은 기관의 임무 자체에 전통 역사 자료의 기반을 구축하는 일이 포함되어 있으므로 앞으로도 디지털 역사 자료를 제작하고 제공할 것이다. 여기에 독립기념관, 민주화운동기념사업회, 동학농민혁명기념재단처럼 특정한 역사적 기념과 관련하여 사료를 디지털로 정리하여 제공하는 기관들의 기여도 지속될 것이며 한국학호남진흥원, 한국유교문화진흥원처럼 지역 문화를 배경으로 전통 역사 자료를 정리, 제공하는 기관들의 역할 또한 확대될 것이다.

그런데 사료의 수집 정리 편찬 자체가 기관의 핵심 기능으로 명문화된 국사편찬위원회를 제외하면 대부분의 기관은 역사 자료의 제작 자체가 목적이라기 보다는 기관의 목적을 달성하기 위해 역사 자료의 디지털 제작을 주요 업무 중 하나로 삼았다고 볼 수 있다. 이를 염두에 두고 앞으로 역사 관련 기관들이 디지털 역사 자료 제공에서 어떤 특징을 보일지, 일종의 동향

탐색의 일환으로 간단히 세 가지 점을 제시해 본다. 디지털 역사 자료의 제작과 활용이라는 측면에서 역사 관련 기관들의 활동을 경향적으로 예상해 보는 것도 디지털 역사 자료를 활용하여 역사 연구를 확장하려는 이들에게 도움이 될 수도 있을 거라는 의도에서 이야기하는 것이며, 특히 이는 단지 개인의 의견일 뿐이다. 개인 의견일 뿐이라는 점, 다시 강조하여 둔다.

한국국학진흥원 웹사이트에 들어가 보면 고도서 한자인식, 고도서 이미지 검색이라는 메뉴가 있다. 고서의 이미지 파일을 올리면 문자를 인식하여 입력하여 주고 해당 문자가 포함된 고도서를 검색하여 주는 서비스이다. 한국정보화진흥원과 연계하여 진행된 인공지능 기반 사업의 결과물로 현재 시범 서비스 중이다. 국사편찬위원회에는 디지털 기반 편찬 사료를 제공하는 한국사데이터베이스 이외에 수집 사료를 정리하여 목록과 이미지 데이터를 제공하는 전자사료관이 있다. 인공지능 활용 이미지 내 문자 인식과 검색 제공 기능이 향후 도입될 가능성이 커 보인다. 이 경우 디지털 이미지 형태의 역사 자료(문자 자료)들의 검색 범위가 확대되게 된다. 역사 자료 편찬의 전통을 잇는 현재의 역사 자료 데이터베이스 제작 작업과는 조금 궤를 달리하지만 스캔, 촬영된 이미지 파일 형태의 자료들에 포함된 주요 문자들에 대한 검색 기능을 제공하는 작업은 머지않아 진행될 것으로 보인다. 판독율이 90% 이상 나온다고 해도 현재의 역사 자료 디지털화 작업-선정, 구성, 교정, 교열, 교감, 해설 등의 과정을 거쳐 품질을 확보한 역사 자료집을 제공하는 것을 목표로 하는 작업을 대체하거나 크게 영향을 주기는 힘들 것이다. 그렇지만 이미지 데이터 내의 주요 문자들이 검색 가능해지면, 검색과 온라인 열람이라는 면에서 활용 가능한 사료의 범위가 크게 증가될 것으로 생각된다. 책자와 문서, 특히 문서 형태의 역사 자료에 대해 서지-원문 이미지를 제공하는 기본 서비스에 검색 범위를 원문 이미지 내 문자까지 확대하는 작업은 여러 기관에 도입되지 않을까 예측해 본다.

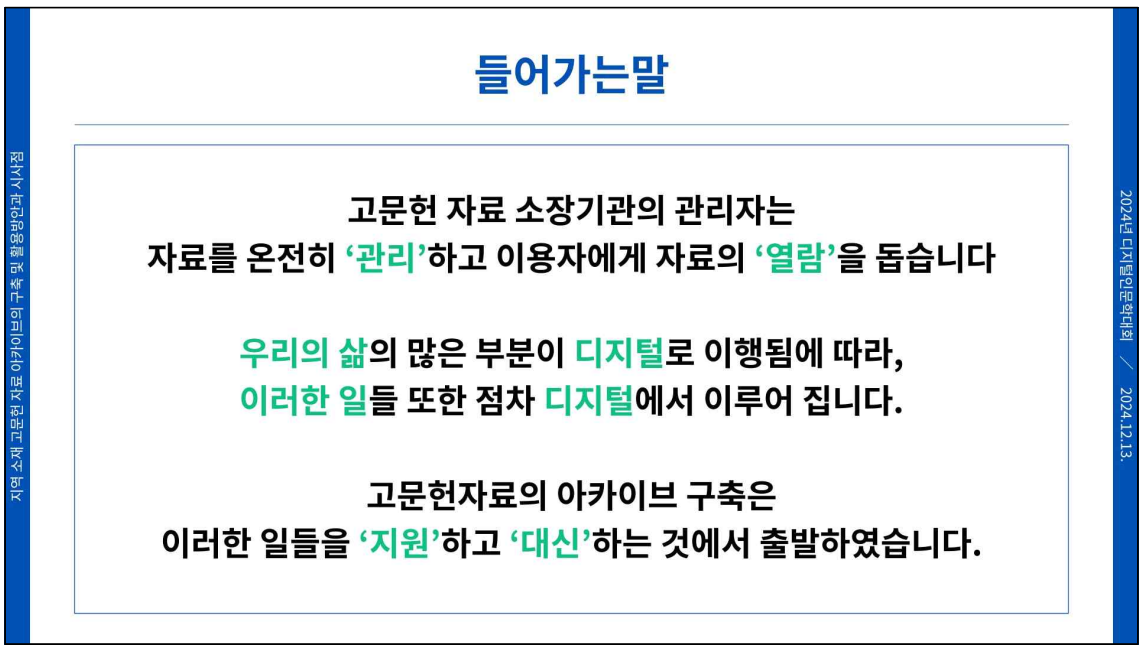
한편으로 디지털 역사 데이터 제작에서 연구 편찬적 성격이 강화될 것으로 보인다. 1차 사료의 디지털화를 바탕으로 보다 연구적 측면에서 가공된 역사 데이터 제작이 늘어날 것이다. 국사편찬위원회를 예로 들자면, 현재 연구 편찬 업무 자체가 디지털을 기반으로 하여 제작과 제공까지 모두 디지털 기반으로 이루어지고 있다. 한마디로 사료집 편찬 업무가 데이터베이스 제작 업무로 이루어지고 있다. 고대사 분야의 역주 사업, 금석문 통합 사업, 조선시대 법령 사업, 근대 사회단체 데이터베이스 제작 사업, 한국 헌정사 자료 제작 사업 등이 국사편찬위원회 한국사데이터베이스 사업으로 진행되고 있는데 대부분이 연구에 기반한 가공 데이터 제작 사업이다. 이와 같은 흐름은 한국학중앙연구원이나 규장각한국학연구원, 고전번역원 등에서도 관찰되는데 어찌 보면 한국학 토대 사업의 주제들과 유사한 면이 있다. 이와 관련하여 대학의 연구 역량과의 연계 또한 좀 더 긴밀해질 필요가 있을 것이다.

마지막으로, 각 기관들에는 연구를 위한 기반 제공에 더하여 일반 시민들이 효능감을 느낄 수 있도록 일을 하라는 압력이 존재한다. 압력이라 표현하였지만 부정적인 뜻은 아닌데 이는 각 기관이 유지되는 데 매우 당연한 일이기도 하기 때문이다. 학술 연구 기반 조성이라는 면도

중요하지만 각 기관이 운영됨으로써 시민들의 편의와 효용이 늘어났다는 사실도 증명할 수 있어야 하기 때문이다. 문화 향유적인 측면에서 보다 일반 시민이 직접적으로 대면할 수 있는 결과물에 대한 요구가 역사 관련 기관 내외에서 다양하게 증가하고 있다. 역사 기관 내부적으로 보자면 기초 자료 제공이라는 일과 일반 시민이 직접 효능감을 느끼게 하는 일의 배분 비율이 서서히 바뀌지 않을까 예측된다. 역사학계에서 보자면 디지털 역사 자료 제공이라는 기초적인 측면에서 서서히 그 양이 줄어들 수도 있는 일이다. 하지만 전체적으로 보자면 역사학 연구의 대중화, 확대라는 역사학계의 오랜 과제와도 연결될 수 있을 것이다. 조금 막연하긴 하지만, 디지털 역사 자료 제공자로서 역사 관련 기관의 역할과 그 변화에 대해서 앞으로 넓은 범위의 논의가 있기를 기대하며 마지막으로 이 문제를 제기하여 둔다.

세션2 발표문 2


2024 디지털인문학대회
지역 소재 고문헌 자료
아카이브의 구축 및 활용 방안과 시사점
김사현(한국유교문화진흥원)


들어가는말
고문헌 자료 소장기관의 관리자는
자료를 온전히 '관리'하고 이용자에게 자료의 '열람'을 돕습니다
우리의 삶의 많은 부분이 디지털로 이행됨에 따라,
이러한 일들 또한 점차 디지털에서 이루어 집니다.
고문헌자료의 아카이브 구축은
이러한 일들을 '지원'하고 '대신'하는 것에서 출발하였습니다.
지역 소재 고문헌 자료 아카이브의 구축 및 활용 방안과 시사점
2024년 디지털인문학대회 / 2024.12.13.

들어가는말

세상 모든 것이 그러하듯, 처음에는 그 방식이 단순했습니다.

그러나, 정보기술의 발전과 그에 따라 변화하는 환경에서
우리들은 새로운 경험을 하게 되고, 익숙해 지고
이전 보다 더 나은 무언가를 원하게 됩니다.

고문헌 자료 아카이브 또한,
나아가야 할 방향성과,
그 실제적 기능, 제공해야 할 정보는 점차 복잡하고 다양해 지고 있습니다.

들어가는말

물론, 많은 노력들이 있었습니다.
좀 더 정확하고 온전하게 고문헌 자료 정보를 기술하고,
자료를 찾고, 정보를 연결하고,
정보를 보기 위한 편리한 방식에 관한 것들입니다.

많은 논의점이 있겠으나, 오늘은
고문헌 자료 소장기관의 입장에서
고문헌 자료 아카이브의 고민과 미래 그리고 하나의 시도
에 대해 이야기 하고자 합니다.

목차

1 지역 소재 고문헌 자료 소장 기관

고문헌 자료 다량 소장 기관

- 공공기관/연구원 : 규장각, 장서각, 국립중앙도서관, 국사편찬위원회
- 대학 도서관 : 계명대, 경상대, 고려대, 대구카톨릭대, 연세대, 영남대, 전남대, 충남대 등

지역자료 중심의 소장기관 : 국학진흥협의체(권역별 국학진흥기관 연합)

- 문화체육관광부 국학진흥기반정책조성사업 참여 기관(2002~)
- 한국국학진흥원(경상) / 한국학호남진흥원(호남) / 한국유교문화진흥원(충청) / 율곡연구원(강원)
- '국학자료(고문헌 자료 포함)'의 조사, 수집, 정리, 아카이브구축 활용 등 공동 사업 추진
- 기증(소유권 이전), 기탁(관리권 이전) 제도 운영

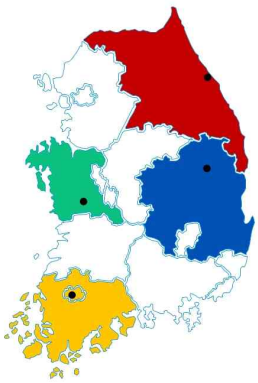
지역 소재 고문헌 자료 아카이브의 구축 및 활용방안과 시사점
2024년 디지털인문학대회 / 2024.12.13.

지역 소재 고문헌 자료 아카이브의 구축 및 활용 한국사사전

1 지역 소재 고문헌 자료 소장 기관

지역자료 중심의 소장기관 : 국학진흥협의체(권역별 국학진흥기관 연합)

구분	한국 국학진흥원	한국 유교문화진흥원	한국학 호남진흥원	율곡연구원
소재지	경북 안동	충남 논산	광주광역시	강원 강릉
설립연도	1995	2022	2017	1993
연고학문	퇴계학	기호유학	호남유학	율곡학
법인성격	재단법인	재단법인	재단법인	사단법인
출연기관	경상북도	충청남도	광주전남	-



2024년 디지털인문학대회 ／ 2024.12.13.

지역 소재 고문헌 자료 아카이브의 구축 및 활용 한국사사전

1 지역 소재 고문헌 자료 소장 기관

지역 고문헌 자료 조사 수집 정리 보존(예시 한국유교문화진흥원)

1 연구자 및 향토사학자 등 관련 분야 전문가의 자문 및 문헌 조사

4 문화재 전문 운송 차량을 이용하여 조사 수집 국학자료를 진흥원으로 이동

2 현장 방문 및 소장자 면담, 자료 보관상태 확인, 소장내력과 출처 확인

5 혼종 소독된 자료를 해포하고 정리하여 세부목록을 작성한 뒤, 수장고에 격납

3 기초목록을 작성한 후 보관의뢰서와 보관증을 소장자와 교환

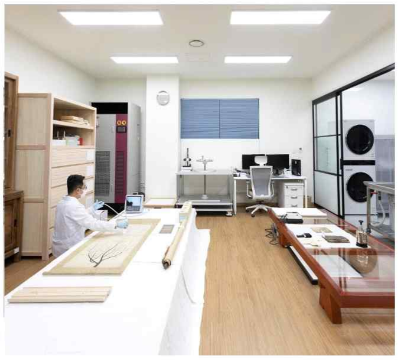
6 국학자료 수증·수탁 심의위원회 개최

수장고
국학자료가 최적의 환경에서
안전하게 보존될 수 있도록
특수 설계된 보안공간

2024년 디지털인문학대회 ／ 2024.12.13.

1 지역 소재 고문헌 자료 소장 기관

지역 고문헌 자료 조사 수집 정리 보존(예시 한국유교문화진흥원)



보존처리실
국학자료의 상태를 조사하여
손상상태에 따라 가장 적합한 재료와
기술로 보존처리하는 공간

유물정리실
국학자료를 학술적인 규칙에 따라 분류 정리 및 등록하는 공간

촬영실
국학자료 촬영 및 DB 구축 공간

1 지역 소재 고문헌 자료 소장 기관

지역 고문헌 자료 조사 수집 정리 보존(예시 한국유교문화진흥원)

이삼분무공신교서
(李森奮武功臣教書)

연 대 1728년
형 태 족자/1축/6.8*147
기탁처 함평이씨 함은군파 종중

포저선생유서
(浦渚先生遺書)

연 대 1692년
형 태 목판본/11권10책/30.4*20.8
기탁처 풍양조씨 오재 조지겸 후손

1 지역 소재 고문헌 자료 소장 기관

지역 고문헌 자료 조사 수집 정리 보존(예시 한국유교문화진흥원)

극재일기
(克齋日記)

연 대 1847년-1931년
형 태 필사본/19책/28.0*19.0
기탁처 원광어비문성균관 국제 어유적 종손

1978-1999年 전교일기장
(1978-1999年 典校日記帳)

연 대 20세기 초-2014년
형 태 필사본/5책/23.5*18.5
기탁처 연기향교

지역 소재 고문헌 자료 원문 활용을 위한 데이터베이스 구축 및 플랫폼 개발 사업

1 지역 소재 고문헌 자료 소장 기관

고문헌 자료 소장량

기준 : 한국국학진흥원(24.10.) / 한국학호남진흥원(24.3.) / 한국유교문화진흥원(24.8.) / 율곡연구원 (24.8.)

	합계	고서	고문서	목판현판	민속/기타
한국국학진흥원	651,581	206,265	365,835	69,026	10,455
한국학호남진흥원	73,927	16,248	52,945	143	4,591
한국유교문화진흥원	39,533	11,974	24,924	1,039	1,596
율곡연구원	24,830	11,089	13,181		560

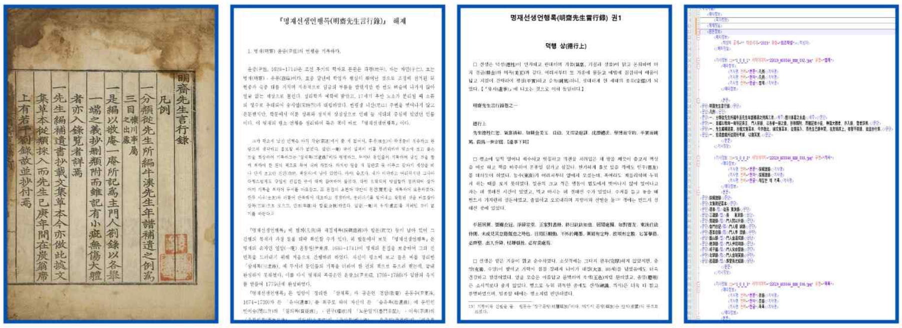
2024년 디지털인문학대회 ／ 2024.12.13.

지역 소재 고문헌 자료 아카이브의 구축 및 활용방안과 시사점

2024년 디지털인문학대회 / 2024.12.13.

2 고문헌 자료 아카이브 구축

고문헌 자료 아카이브 구축 ① 디지털화 – 이미지, 해제, 원문, 국역



이미지 | 해제 | 원문/국역 | 전자문서(XML)

지역 소재 고문헌 자료 아카이브의 구축 및 활용방안과 시사점

2024년 디지털인문학대회 / 2024.12.13.

2 고문헌 자료 아카이브 구축

고문헌 자료 아카이브 구축 ① 디지털화

- 이미지 구축
  - 과정 : 대상자료선정 → 촬영/스캔 → 보정 → 검수 → 파일네이밍
  - 고서(스캔), 그 밖에 문헌은 촬영. 원 소스는 최대한 무손실 압축(tiff)으로 획득.
  - 이후 보정 과정을 통해 최종적으로 보존할 파일 규격 확정(jpg, 해상도 조정)
  - 해당 자료와의 관리를 위해 파일 네이밍 규칙에 따라 파일 관리

2 고문헌 자료 아카이브 구축

2 고문헌 자료 아카이브 구축

- 원문/국역/해제
  - 과정 : 대상자료선정 → 원문입력 및 교정(반복)/국역/해제 → 검수
  - 원문의 경우 이체자 등 해결 방안 필요. 의미 단위의 표점 부여시 표점 기호 적용.
  - 국역 및 해제는 해당 자료에 대한 내용적 이해도가 있는 전문가에게 의뢰해 진행.
  - 전자문서는 주로 XML(eXtensible Markup Language)로 구축. 기관별 데이터 모델은 상이함.
- 디지털화는 대개, 하나의 자료를 동시에 이미지/원문/국역/해제 하지 않음.
- 그 이유는 각각의 과정에 따른 시간과 비용의 문제, 그리고 전문분야의 상이함 때문.

지역 대학 고문헌 자료 디지털아카이브 구축 및 활용 사업
2024년 디지털인문학대회 ／ 2024.12.13.

2 고문헌 자료 아카이브 구축

고문헌 자료 아카이브 구축 ② 정보시스템 구축/운영을 통한 자료 검색, 열람

한국국학진흥원
기록유산의 총아 고도서, 옛 일상 속 인간 관계- 고문서, 선인의 일상생활-일기 등

한국학호남진흥원
호남국학종합DB

한국유교문화진흥원
충청국학 디지털 아카이브

율곡연구원
강원한국학아카이브

지 현 제 고문헌 자료 아카이브 구축 및 활용을 위한 서사적 접근

2 고문헌 자료 아카이브 구축

주요 기능 : 자료 검색과 이미지, 원문, 국역 열람에 초점

- 자료 검색(텍스트 검색, 디렉토리 검색[유형, 소장처 등])
- 상세페이지(자료 목록/서지/해제), 이미지 뷰어, 목차, 원문 및 국역
- 이미지, 원문/국역 다운로드(일부)

특이사항

- 한국국학진흥원은 고문헌 자료 유형별 사이트 구축. 통합서비스(포털) 구축.
- 또한, 고문헌 자료와 관계 있는 콘텐츠성 사이트 구축 운영.
- 그외 나머지 기관은 각각 단일 사이트 구축 및 운영. 사이트 여러 유형의 자료 서비스.

2024년 디지털인문학대회 ／ 2024.12.13.

3 고문헌 자료 아카이브의 고민과 미래

고민-1 고문헌 자료 증가 폭에 비해 감소하는 디지털화 및 아카이브 구축 인력

- 지역 소재 고문헌 자료 증가폭에 비해 고문헌 자료 관련분야 고등교육자 감소
- 특히, 지역의 경우 이러한 문제가 더욱 심함. 고문헌 자료 1차 정보라 볼 수 있는 목록, 표점, 국역, 해제 등의 구축 한계
- 고문헌 자료 소장 기관에서 아카이브 구축 사업을 기획하고 추진할 수 있는 전문 인력이 부재함

고민-2 고문헌 자료 아카이브 이용층 확대 노력

- 역사분야 방송, 드라마, 웹툰 등 문화콘텐츠 인기로 이어지는 고문헌 자료 관심에 부응 필요
- 여전히 전문가만을 위한 아카이브. 고문헌 자료의 검색/열람을 위한 정확성과 전문성은 당연하나, 또 다른 고객인 일반인들에게도 열려 있는 아카이브 필요.
- 또한 한국 문화에 관심이 있는 외국인들을 위한 접근점 마련 필요

3 고문헌 자료 아카이브의 고민과 미래

고민-3 공공정보 공개, 디지털 인문학 등 고문헌 자료 정보 제공 방식 노력

- 공공데이터법(2013) 제정으로 공공분야 정보/데이터의 사회적 요구(디지털 기반 데이터 활용성)
- 정보기술을 적극 활용한 새로운 인문학 연구 방법인 디지털 인문학 연구에 대응
- 고문헌 자료 아카이브의 본연의 목적인 효율적인 데이터 구축, 신속한 공개, 공유 방안 마련 필요
- 고문헌 자료 분야 연구 성과의 아카이브 정보 및 데이터 환원 노력 필요

고민-4 지역 특성을 반영한 아카이브 구축 노력

- 소장 자료의 특성이 곧 아카이브의 특성이 되는데, 지역적 특성을 어떻게 함께 고려할 수 있는가?
- 단, 지역 주요 인물, 유산 등 지역성을 띠는 주제과 관계 있는 고문헌 자료를 우선적으로 디지털화 함

지역 소재 고문헌 자료 아카이브의 구축 및 활용방안과 시사점
2024년 디지털인문학대회 / 2024.12.13.

3 고문헌 자료 아카이브의 고민과 미래

미래-1 고문헌 자료 정보의 충실한 디지털 저장소

- 소장자료 디지털 자원 및 데이터의 안정적인 구축과 관리.
- 자원 구축은 향후 예상되는 고문헌 자료 관련 고등교육자 감소에 대비하여, 정보기술 기반 고문헌 자료의 디지털 자원 제작 도구의 개발과 활용 노력 필요.
- 현재 고문헌 자료 문자 판독(OCR), 문자 의미 해독(문장분절/표점), 국역 관련 연구 사례 축적
- 완전한 이행 보다는 단계적 발전을 염두에 두고 지속적 추진,
- 표준적 데이터 구축을 위한 논의 지속과 그에 따른 데이터 제공 방식과 형식의 갱신 노력

지역 소재 고문헌 자료 아카이브의 구축 및 활용방안과 시사점

3 고문헌 자료 아카이브의 고민과 미래

미래-2 고문헌 자료 정보와 지식정보 연계의 노력

- 고문헌 자료의 1차적 정보만으로 일반인의 접근과 활용을 기대하기에 어려움
- 유관분야 지식정보를 구축하고, 기능적으로 연결하여 보편적 관점과 시각에서 자료 접근성 향상
- 여러 방법이 있으나, 시맨틱 데이터 기반의 디지털 아카이브 구축이 하나의 좋은 모델
  ex) 백과사전적 아카이브(Encvyes, Encyclopedia + archives)

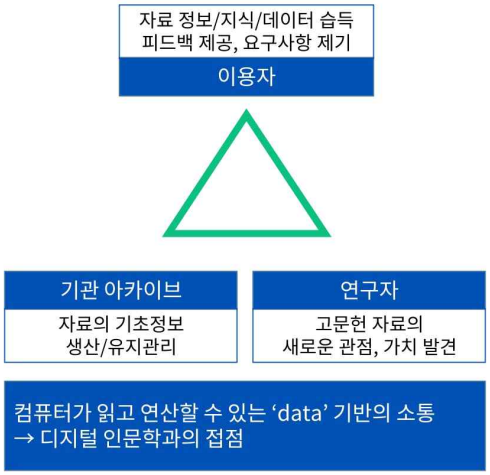
2024년 디지털인문학대회 / 2024.12.13.

지속 가능한 고문헌 자료 아카이브 구축 및 활용 방안 연구

3 고문헌 자료 아카이브의 고민과 미래

미래-3 고문헌 자료 아카이브 구축/활용 생태계 구상

- 기관 아카이브 , 연구자, 이용자의 상호 협력적 구조에 따른 생태계
- 1축 [고문헌 소장 기관 아카이브] : 고문헌 자료의 기초 정보를 생산하고 품질을 유지관리하고,
- 2축 [유관분야 연구자] : 고문헌 자료의 정보에 기초하여 자료의 새로운 관점과 가치를 발견하고, 그 정보가 고문헌 자료 아카이브로 환원되어 풍부한 정보를 가진 아카이브로 진보해 나가며
- 3축 [다양한 이용자 층] : 고문헌 자료 관련 정보, 지식 습득과 2차 가공을 위한 데이터 활용 및 피드백과 새로운 요구사항 제기



컴퓨터가 읽고 연산할 수 있는 'data' 기반의 소통
→ 디지털 인문학과의 접점

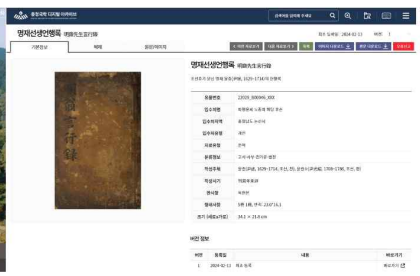
2024년 디지털인문학대회 / 2024.12.13.

지속 가능한 고문헌 자료 아카이브 구축 및 활용 방안 연구

4 지역 소재 고문헌 자료 아카이브의 하나의 시도

충청국학 디지털 아카이브(https://archives.ikcc.or.kr/)

- 충청국학 디지털 아카이브는 한국유교문화진흥원이 조사·수집한 충청권 국학자료를 디지털 환경에서 검색, 열람할 수 있는 사이트.
- 주요 기능은 국학자료 검색, 열람과 시맨틱 데이터 기반의 정보관계망 서비스.
- 2024년 2월 정식 오픈.
- 2024년 12월, 2천여건 고문헌 자료 공개.(수집자료 공개 지속)



2024년 디지털인문학대회 / 2024.12.13.

4 지역 소재 고문헌 자료 아카이브의 하나의 시도

자료와 지식의 연계를 위한 하나의 시도 : 시맨틱 데이터 구축과 관계망 서비스

- 고문헌 자료 아카이브의 활용성을 증진하고,
- 향후 예상가능한 다양한 지식정보 자원과의 연결을 염두에 두고 → 관계망 서비스를 구축
- 김현 교수와 ' 한국학중앙연구원 디지털인문학연구소(인문정보학과) ' 의 연구, 프로젝트에 영향 받음
- 현재 충청지역의 22개 테마(주제)의 관계망 공개, 서비스.
- 향후 시맨틱 데이터를 활용한 다양한 정보와 지식의 연결 방안 마련을 통해 종합적인 아카이브 구축.

시맨틱 데이터와 유관정보 구축

- 온톨로지 설계, 시맨틱 데이터 구축, 노드 설명 자원인 텍스트, 메타정보, 이미지, 아이콘
- 테마 단위(small data)의 시맨틱 데이터 구축. 각 테마는 서로 접점을 만들고자 노력함
- 일부 테마는 기관 소장 고문헌 자료를 중심으로 구축
- 새로운 테마와 시맨틱 데이터 구축시 온톨로지 설계에 반영하고 기존 데이터도 갱신

지역 소재 고문헌 자료 아카이브의 구축 및 활용방안과 시사점

4 지역 소재 고문헌 자료 아카이브의 하나의 시도

관계망 뷰어 주요 기능(충청국학 디지털 아카이브-지식관계망, https://archives.ikcc.or.kr/network/main)

- 다양한 관계망 접근점 마련 ① 테마 단위 접근, ② 노드 검색, ③ 소장자료 상세페이지
- 최초 관계망 생성, 이후 계속해서 노드 확장 가능. 노드 아이콘에서 확장 가능 여부 표시함
- 노드 클릭시, 간단한 설명, 속성정보(메타정보), 다른 노드와의 관계목록(전체) 정보 제공

2024년 디지털인문학대회 / 2024.12.13.

맺음말

- 근래에 지역에 고문헌 자료 소장 기관이 증가하는 추세
- 이 기관들은 고문헌 자료의 조사/수집/정리/보존/아카이브 구축 등 유사한 업무 수행
- 아카이브 구축은 디지털화, 정보시스템 구축과 운영으로 진행되며, 여전히 고문헌 자료의 검색과 열람 중심의 기능으로 '전문가' 중심의 서비스.
- 고문헌 자료 아카이브의 고민으로
  ① 고문헌 자료 관련 전문인력 부재로 디지털 자원 확보 어려움,
  ② 고문헌 자료 아카이브 이용층 확대 노력,
  ③ 공공정보 공개, 디지털 인문학 등 고문헌 자료 정보 제공 방식 노력,
  ④ 지역 특성을 반영한 아카이브 구축 노력

맺음말

- 미래 방향성으로
  ① 고문헌 자료 정보의 충실한 디지털 저장소,
  ② 고문헌 자료 정보와 지식정보 연계의 노력,
  ③ 고문헌 자료 아카이브 구축/활용 생태계 구상 제안
- 하나의 시도로서 충청국학 디지털 아카이브-시맨틱 데이터 기반 아카이브 구축 사례.
- 지역소재 고문헌 자료 아카이브는 지속적이고 안정적인 아카이브 구축과 '공공성'을 바탕으로 관련분야의 다양한 요구를 수용하여 개선, 발전되어야 함.

감사합니다

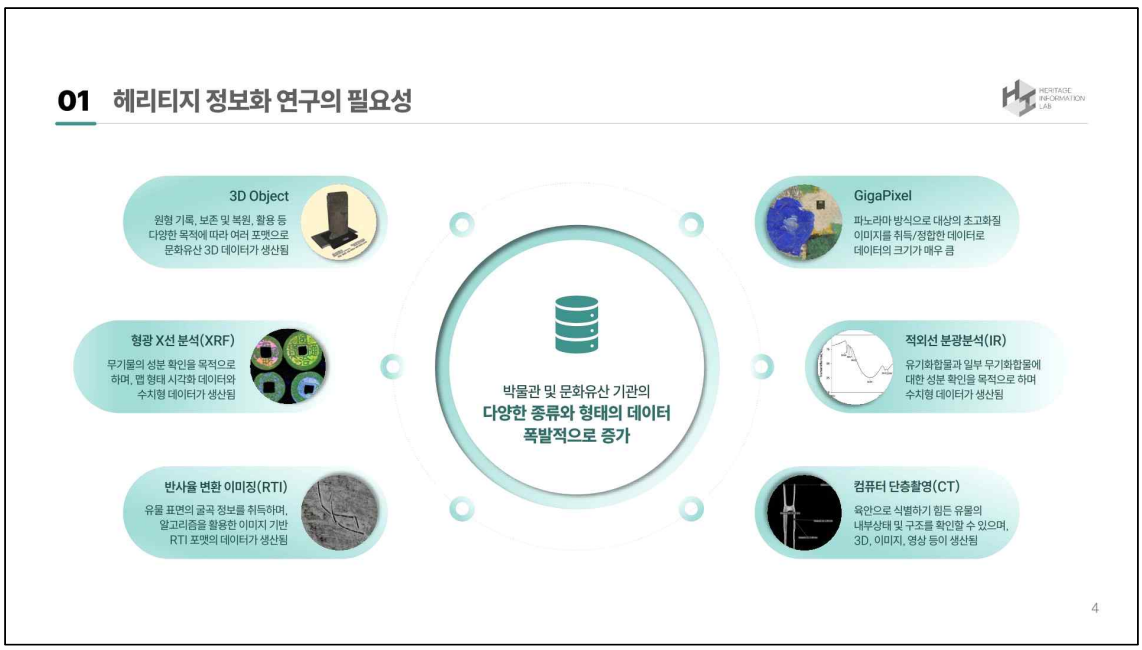
지역 소재 고문헌 자료 아카이브의 구축 및 활용방안과 시사점
2024년 디지털인문학대회 / 2024.12.13.

세션2 발표문 3


국가유산 디지털 아카이브 구축 현황 및 향후 과제
한국전통문화대학교 디지털헤리티지학과
이 종 욱 교수


RIJKS MUSEUM
이 종 욱
Jongwook Lee
한국전통문화대학교 디지털헤리티지학과장
디지털콘텐츠기획 전공 조교수
한국전통문화대학교 도서관장, 전산원장
학력
2019. 공학박사. KAIST 문화기술대학원
2012. 공학석사. KAIST 문화기술대학원
2010. 학사. 한국전통문화대학교 문화재관리학과
경력
국가유산청 문화유산위원회 공능문화재분과 전문위원
ICOMOS CIPA(디지털기록) 회원


Heritage Info Lab
문화유산 아카이브, 교육 및 전시 콘텐츠를 연구
박물관 및 문화유산 기관과 함께 디지털 큐레이션, 정보 시각화,
지능형 플랫폼, AI 스토리텔링을 개발
이집트에 디지털 유산 기록센터를 구축하고
사우디 아라비아에 디지털 유산 커리큘럼 설계 등 국제 용역을 수행
HERITAGE INFORMATION LAB


01 헤리티지 정보화 연구의 필요성
HERITAGE INFORMATION LAB
3D Object
원형 기록, 보존 및 복원, 활용 등 다양한 목적에 따라 여러 포맷으로 문화유산 3D 데이터가 생산됨
GigaPixel
파노라마 방식으로 대상의 초고화질 이미지를 취득/정합한 데이터로 데이터의 크기가 매우 큼
형광 X선 분석(XRF)
무기물의 성분 확인을 목적으로 하며, 맵 형태 시각화 데이터와 수치형 데이터가 생산됨
박물관 및 문화유산 기관의
다양한 종류와 형태의 데이터
폭발적으로 증가
적외선 분광분석(IR)
유기화합물과 일부 무기화합물에 대한 성분 확인을 목적으로 하며 수치형 데이터가 생산됨
반사율 변환 이미징(RTI)
유물 표면의 굴곡 정보를 취득하며, 알고리즘을 활용한 이미지 기반 RTI 포맷의 데이터가 생산됨
컴퓨터 단층촬영(CT)
육안으로 식별하기 힘든 유물의 내부상태 및 구조를 확인할 수 있으며, 3D, 이미지, 영상 등이 생산됨
4

01 헤리티지 정보화 연구의 필요성




POINT.

"디지털 헤리티지 데이터"

✓기하급수적으로 수량 증가

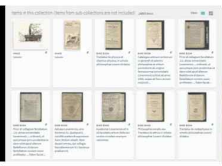
✓생성 목적과 유형 다양화

5

01 헤리티지 정보화 연구의 필요성

디지털 헤리티지.

인간의 지식과 표현의 고유한 자원으로서
문화, 교육, 과학, 행정적 자원 뿐만 아니라 기술적, 의학적, 법적 정보를 포괄하며,
디지털로 생성되거나 아날로그 자원으로부터 디지털 형식으로 변환한 것
- 유네스코(디지털 유산의 보존에 관한 헌장)


01 헤리티지 정보화 연구의 필요성

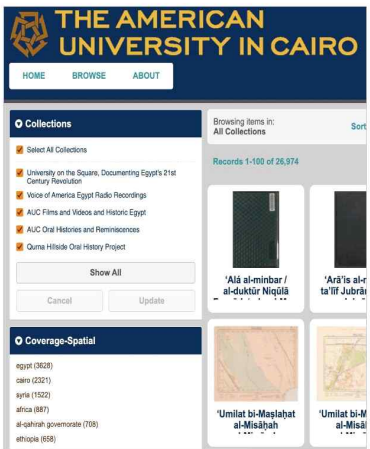
📌 *디지털 헤리티지의 유형에 따라 보존, 관리, 활용 방안이 달라지며,*
본 디지털 헤리티지의 경우 보존에 우선순위가 있음

7

01 헤리티지 정보화 연구의 필요성

📌
기존 아날로그 아카이브의 한계.

- 항구적인 수명
- 접근성의 향상
- 자료검색의 용이성



8

01 헤리티지 정보화 연구의 필요성

디지털 아카이브

디지털화된 자료를 정기적으로
저장, 유지, 접근시키는 시스템

디지털 아카이빙

- 지속적 가치를 가졌다고 판단되는 디지털 객체를 장기간 관리하는 활동
- 가치 있는 디지털 자원을 선별하여 그 내용 및 기능을 보존·관리하고
  장기간 접근할 수 있도록 하는 전반적인 활동

디지털 아카이빙 절차

데이터의 수집 → 아날로그 자료의 디지털화 → 자료고유번호 및 분류코드 부여 → 데이터베이스 구축 → 정보화 서비스

9

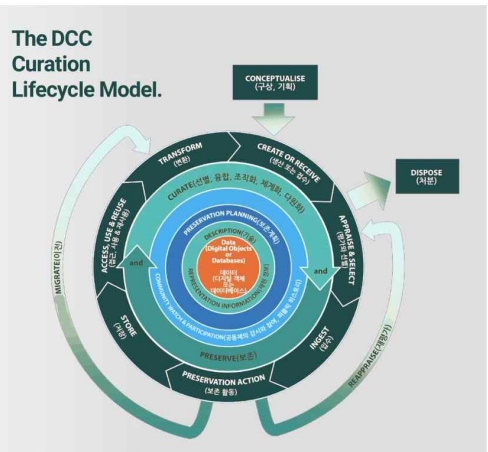
01 헤리티지 정보화 연구의 필요성

디지털 큐레이션

데이터를 적극적으로 관리·보존하고,
더 나아가 가치를 부가하여
현재와 미래에 이용·재이용 될 수 있도록만드는 작업

디지털 큐레이션 기술 개발에 대한
연구자 및 정보전문가들의 노력을 강조
(Beagrie, 2008)

→ 디지털 큐레이션 역량강화 프로그램의 필요



10

02 데이터모델

DublinCore

- 방대한 양의 웹 자원을 기술할 수 있는 단순한 구조의 메타데이터
- 간결성, 호환성, 확장성으로 자원의 유형과 시스템에 관계없이 확장 가능한 구조

Schema.org

- 다양한 인코딩을 통해 웹자원의 정보 마크업에 필요한 타입과 속성 제공
- 일반 사용자가 웹에서 관련 정보를 더 쉽게 찾을 수 있는 검색 엔진을 만들 수 있도록 지원

CIDOC CRM

- 문화유산 분야의 정보 통합을 위한 이론적·실용적 도구
- 문화유산 데이터의 질의 및 탐색에 대한 암시적·명시적 개념과 관계를 설명하기 위한 정의와 형식적 구조 제공

europeana carare project

- 유산 관련 기관의 온라인 컬렉션, Heritage Asset 및 디지털 자원에 대한 메타데이터 스키마를 제공
- 유적 및 유물에 대한 정보와 디지털 데이터를 함께 설명

bf BIBFRAME

- 도서관 커뮤니티 내외부에서 서지 정보를 유용하게 활용하기 위한 서지 정보 모델
- 자원의 저자, 책, 내용, 출판 형식, 책 사본에 대한 정보 등 다양한 연관 관계를 표현

1995	Year Founded	2011
DCMI, OCLC & NCSA	Founders	Bing, Google & Yahoo
DCMI	Maintained By	Open Community
Physical & Web Resources	Use For	Only Web Resources
Less SEO Friendly	SEO Friendly	Highly SEO Friendly
Abstract & Broad	Data Type	Highly Specific
Wide	Compatibility	Only With Web Resources
Here To Stay	Future Prospect	Rapid Growth

https://seopressor.com/blog/dublin-core-vs-schemaorg-metadata-comparison/

다양한 도메인의 웹 자원을 설명하기 위한 범용 스키마

2000	Year Founded	2010
ICOM	Founders	Europeana
Special Interest Group	Maintained By	CARARE
Physical & Web Resources	Use For	Physical & Web Resources
Information for Cultural Heritage	Domain	digital archaeological and architectural heritage

문화유산 도메인에서 널리 사용되는 온톨로지 어휘

CIDOC CRM이 아날로그 문화유산의 표현에 집중한다면,
CARARE는 고고학 및 건축 유산과 디지털 객체까지 설명함

디지털 헤리티지 분야에 응용 적용 가능한 서지정보모델

도서와 연관된 다양한 정보와 관련 객체를 서술하는 모델을 차용하여 문화유산과 디지털 헤리티지의 관계, 디지털 헤리티지 생성과 관련된 객체까지 설명할 수 있음

11

02 데이터모델 _ CIDOC CRM

ICOM CIDOC는 국제박물관협회(International Council of Museums, ICOM) 산하의 국제기록위원회(International Committee for Documentation, CIDOC)를 의미

이는 전 세계의 박물관과 문화유산 기관의 정보 기록화를 표준화하고 개선하기 위해 설립된 전문 조직

문화유산 정보 관리 표준 개발

- 박물관, 아카이브, 문화유산 관리 기관이 정보를 효율적으로 수집, 저장, 공유할 수 있도록 표준화된 프레임워크를 제공

지식 및 경험 공유

- 국제적 워크숍, 세미나, 학술대회를 통해 전문가들이 모여 지식과 경험을 공유할 수 있는 플랫폼을 제공

도구 및 리소스 제공

- 문화유산 기록 관리를 위한 지침, 소프트웨어, 데이터베이스 구축 방법 등을 지원

12

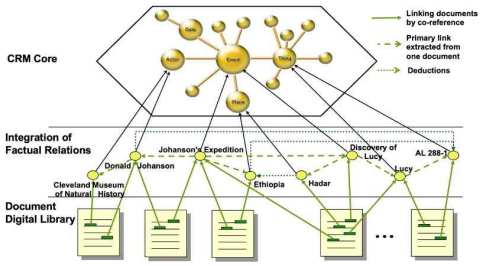
02 데이터모델 _ CIDOC CRM

❖ CIDOC CRM

- 문화유산 및 박물관 데이터를 통합적으로 표현하고 파운드 운용성을 지원하는 개념적 참조 모델
- ISO 21127:2006 국제표준으로 채택
- 데이터의 의미적 연결을 가능하게 하여, 서로 다른 시스템과 데이터를 공유하고 통합을 촉진

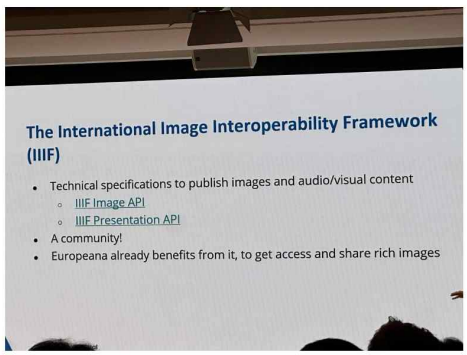
❖ 주요 특징

- 엔터티(Entity)와 관계(Relationships) : 문화유산 데이터를 표현하는 객체와 그것들 간의 관계를 정의
- 유연성과 확장성 : 다양한 데이터 형식과 주제에 적용 가능



13

02 데이터모델 _ Europeana



→유로피아나 재단은 최근 급증하는 고화질 이미지 데이터를 포함해서 3D를 포함 다양한 멀티미디어 데이터의 표준화에 관심
→기존의 고화질 이미지 데이터 표준인 IIIF를 발전시키고자 함
→한국의 다양한 멀티미디어 데이터 생성에 관심

02 데이터관리시스템 _ Research Space

ResearchSpace 프로젝트
(Oldman et al., 2018)

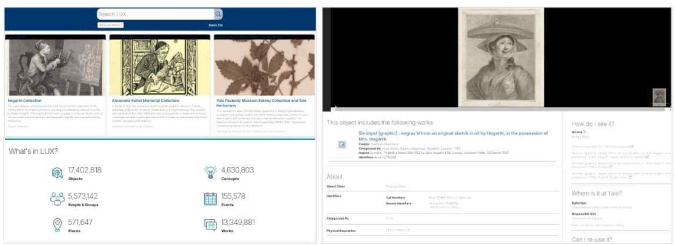
- CIDOC-CRM을 활용한 시맨틱 웹 기술 적용
- 연구자, 데이터, 연구 활동을 통합하는 오픈 소스 플랫폼
- 온톨로지를 통해 데이터 연계 및 주석 작업 수행
  → 데이터 공유 효율성 및 연구 심화 촉진

아마라 웨스트 연구 프로젝트
(Spencer et al., 2021)

- 상부 누비아 지역 생활 탐구 발굴 데이터 온톨로지 통합
- 다양한 발굴 및 과학 분석 데이터를 통합하여 새로운 통찰 제공

15

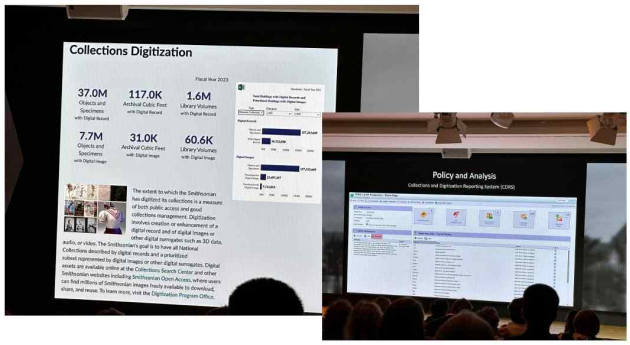
02 데이터관리시스템 _ Yale university



→ International Image Interoperability Framework(IIIF) 방식을 채택하여 이미지의 상호운용성을 극대화
→ 이를 통해 디지털 이미지의 원활한 수집과 주석이 가능하며, 연구자를 포함한 다양한 사용자가 고품질 이미지를 여러 방식으로 활용
→ CIDOC-CRM 기반의 Linked Art를 사용하여 수집 데이터를 Linked Open Data 형식으로 매핑함으로써, 사용자가 예술 작품 간의 관계와 맥락을 깊이 이해

16

02 데이터관리시스템 _ Smithsonian Institution

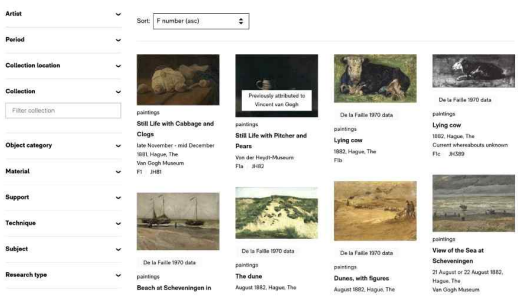


→스미소니언은 방대한 컬렉션 디지털 데이터를 포함
→ 특히 자연사 관련 디지털 데이터는 비상업적 용도로는 저작권 없이 모두 공개
→스미소니언 X3D는 고화질 데이터를 웹에서 서비스 하는 선도적인 역할

02 데이터관리시스템 _ 반고흐뮤지엄

자료현황
- Collection 25,800
- Library material 50,000
- Documentation 2.3tb(1,350,000 scanned pages)

자료관리 어플리케이션
- Axiell collections – collection, library, archive
- DAM(Commulus Comrads) – Digital asset(migration)
- Zylab – digitized documentation
- Zotero – research material

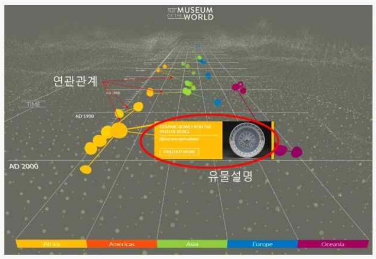


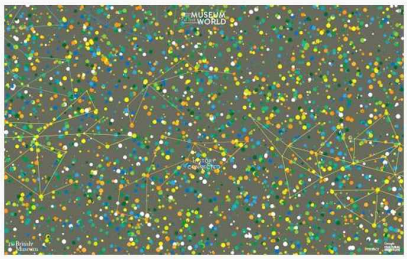
→ 유물관리, 디지털 자원 관리, 디지털 문서관리, 논문관리를 위해 각각 상용화된 프로그램을 활용
→ 각 개체들을 참조할 수 있는 형태로 개선 중
→ RKD(국립미술사연구센터)와 공동으로 작품과 자료연계 사업을 진행
→ 디지털 에셋 생성, 관리, 자료 디지털화를 위한 전문 업체와 연구소가 활발히 양성되고 있음

02 정보시각화 _ Museum of the world

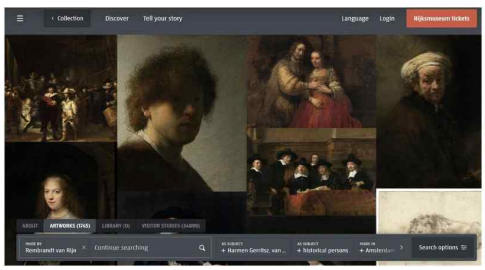
The Museum of the World

- ✓ 유물들을 지역별, 시간별로 구분하여 시각화함
- ✓ 현시점에 가까운 유물부터 시간을 거슬러 올라가는 형태로 각 지역의 유물을 제공
- ✓ 아프리카, 아메리카, 아시아, 유럽, 오세아니아와 같이 대륙별로 지역을 구분
- ✓ 삶과 죽음(living and dying), 힘과 정체성(power and identity), 종교와 믿음(religion and belief), 교류와 갈등(trade and conflict)의 키워드가 존재





02 정보시각화 _ 네덜란드국립박물관(Rijksmuseum)



→ CIDOC CRM을 기반으로 문화유산 및 미술 데이터의 특정 요구사항을 간소화하고 특성화한 애플리케이션 프로필인 linked art를 기반으로 작품을 연결
→ 모바일, 온라인 접근성이 매우 뛰어남
→ 물리적인 디지털 데이터 접근 및 체험 시스템은 간소화함

02 인공지능 _ 이미지분석기술

https://blogs.nvidia.co.kr/2019/02/07/deep-learning-receptors-historical-documents/

CNN를 사용하여 96% 정확도로
손글씨 문자를 인식하고,
언어 모델에 기반하여 각 단어에
가장 알맞은 문자를 결정하는 시스템

새로운 콘텐츠 발굴을 위한
고문서 인식·분석 기술 발전

문화유산
이미지 분석을 위한
인공지능 기술

- 자연어 처리 및 인식 기술의 발전에 따라 역사 기록에 대한 OCR 성능 대폭 향상
- 특정 단어 검색, 시간의 흐름에 따른 언어 진화 추이 분석, 인구 통계 및 사무 기록을 활용한 역사적 맥락 추적 등에 활용 가능
- 활자 외에도 삽화, 여백에 기록된 내용, 워터마크 등 문서에 담긴 추가 정보까지 분석할 수 있는 방법 모색
- 조선왕조실록 등 고문서 분석을 통한 디지털 인문학 연구 활성화 기대

Results from "Image Style Transfer Using Convolutional Neural Networks"

- 전통 예술의 고유한 스타일과 특징을 반영하여 예술 작품을 생성할 수 있는 스타일 변환 기법을 개발
- 딥러닝 기술을 활용하여 스타일 변환 알고리즘을 개선하고 효율적으로 학습시킬 수 있는 방법 모색
- 생성적 적대 신경망(GAN)이나 변이형 오토인코더(VAE)와 같은 딥러닝 아키텍처와 최근 이미지 생성모델로 각광받고 있는 확산형 모델(Diffusion model)을 사용하여 스타일 변환 모델 학습

새로운 콘텐츠 생산을 위한
Style transfer 기술

주된 형태는 Content Image를 유지하고,
화풍은 Style Image와 유사하게 바꾸어
새로운 이미지를 생성하는 기술

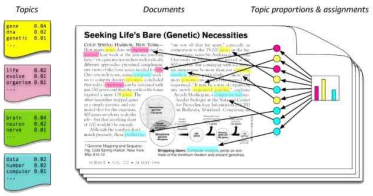
https://blogsht.wordpress.com/2017/04/27/style-transfer/

21

02 인공지능 _ 텍스트분석기술

토픽 모델링(Topic Modeling)

말뭉치에서 주제를 자동으로 발견해내는 기술

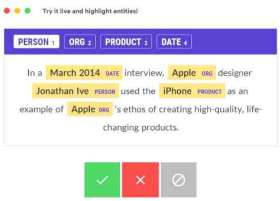


문화유산
문서 분석을 위한
인공지능 기술

- 주어진 문서-단어 행렬에서 문서별 주제 분포와 주제별 단어 분포를 발견
- 토픽 모델링을 문화유산 분야에 적용하여 통계적으로 문서의 주제 분포와 경향을 파악할 수 있음
- LDA(Latent Dirichlet Allocation) 모델이 대표적

개체명 인식 기술 (Name Entity Recognition)

미리 정의된 카테고리에 해당하는 개체명을 찾아서 분류하는 기술



https://prodi.gy/features/named-entity-recognition

- 문화유산 도메인 관련 단어 속성 및 관계 정의
- 문화유산 도메인 특화 말뭉치를 활용하여 NER Deep earning 시스템 구축
- 문화유산 관련 문서에 대한 자동 라벨링 및 NER

22

02 XR 콘텐츠

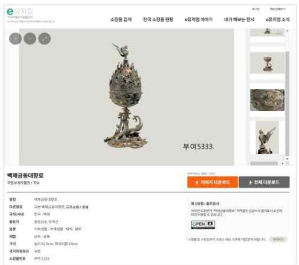
한양도성 타임머신

✓ 3D 모델로 제작되는 고건축물과 관련이 있는 역사 자료를 대상으로
한양 도성과 조선왕실 문화를 이해하는 데 필요한 지식 정보를
광범위하게 조사·추출하여 데이터화 하고,
데이터 요소 상호간의 관계를 명시적으로 표현함으로써
대상 자료의 내용을 분석·응용·확장할 수 있는
시맨틱 데이터베이스 기반의 데이터 아카이브를 구현

03 국가유산 디지털 아카이브 사업현황

이뮤지엄

- '문화유산표준관리시스템'은 대한민국의 유산 관련 기관이 보유한 유물 관련 정보를 클라우드 DB를 통해 통합 관리
- 724개 이상의 기관에서 수집된 약 680만 점의 소장품 정보를 관리
- 'e-뮤지엄'에는 263만 점의 유물 정보와 294만 점의 이미지가 포함되어 있으며, 이는 대중에게 공개

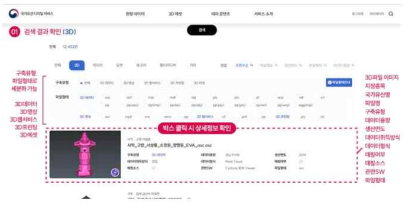


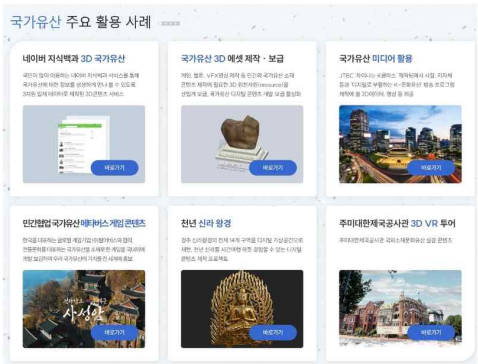
24

03 국가유산 디지털 아카이브 사업현황

국가유산포털

- 문화재 설명자료, 사진, 동영상, 도면, 조사·연구자료 등 문화유산 정보와 콘텐츠를 제공하는 서비스
- 그 중 국가유산디지털 서비스는 지정종목, 지역별, 시대별, 생산연도별 등 조건별 검색 기능을 통해 데이터에 쉽게 접근할 수 있고, '3D에셋', '테마 콘텐츠'와 같은 서비스를 통해 국가유산청에서 제공하는 3D 스캔영상, 웹 VR 콘텐츠 등의 정보를 쉽게 확인





25

03 국가유산 디지털 아카이브 사업현황

국가유산포털

- 저작권 표시
  - 공공누리 1유형:출처표시
  - 1. 온·오프라인 상에 공유 및 이용 : 온·오프라인을 통하여 공유 및 이용 가능
  - 2. 저작물 변경 : 2차적 저작물로 변경하여 이용 가능
  - 3. 이 저작물은 영리 목적으로 이용할 수 있습니다.

26

03 국가유산 디지털 아카이브 사업현황

국가유산 지식이음

- 국립문화유산연구원과 소속 지방연구소의 연구성과물을 제공
- 국가유산포털 검색시스템과 연계
- 저작권 정책과 open API 활용 매뉴얼을 제공

27

03 국가유산 디지털 아카이브 사업현황

무형유산디지털아카이브

- 국립무형유산원의 공연, 전시, 교육 관련 자료, 학술조사 연구 자료, 국가무형문화재 기록화 자료, 다양한 영상콘텐츠 등 생산 자료와 인간문화재, 연구자 등에게 수집한 자료를 공개하여 국민들이 무형유산에 대한 정보 및 자료를 한 곳에서 보고 이용하도록 서비스
- 2022년 11월 기준 무형유산 디지털 아카이브에 등록되어 있는 아카이브 자료들은 총 7만 5천여 건
- 아카이빙이 완료된 무형문화재 정보 검색, 공연영상, 도서, 사진, 구술채록 자서전, 학술조사연구 자료 및 소장품의 세부정보 검색이 가능

영상 No.2361

종묘제례악

제작자	국립문화재연구소 무형문화재연구실
생산년도	2008-01-01
지역	서울 종묘
공공누리	제 4유형
CCL	저작자표시-비영리-변경금지
키워드	종묘제례, 음악, 공연예술, 보존회, 기록화, 무형유산

28

03 국가유산 디지털 아카이브 사업현황

한양도성 타임머신

- 3D 모델로 제작되는 고건축물 (경복궁, 광화문, 6조거리, 사직단, 종친부)과 관련이 있는 역사 자료를 대상으로 한양 도성과 조선왕실 문화를 이해하는 데 필요한 지식 정보를 광범위하게 조사·추출하여 데이터화
- 데이터 요소 상호간의 관계를 표현함으로써 대상 자료의 내용을 분석·응용·확장할 수 있는 시맨틱 데이터베이스 기반의 데이터 아카이브를 구현
- 인문학 융합 연구 데이터 18만 건과 삼차원 복원·재현 데이터 2천300여 건, 개인과 개별 기관이 소장한 문화재 사진·도면·영상·문서 60만 건을 확보
- 데이터 기반으로 게임, 메타버스 콘텐츠를 제작, 공개하였지만 데이터를 공유하고 있지 않음

29

03 국가유산 디지털 아카이브 사업현황

국가유산 원형기록 통합 DB 구축 사업

- 2000년「공공기록물 관리에 관한 법률」이 제정되면서 이후 각종 문화재 서식이 완비되며 국가지정문화재를 중심으로 2D 이미지, 도면 뿐만 아니라 3D 스캔데이터와 같은 문화유산 3D 데이터를 생성
- 2019년에는 목조 건축문화재 CAD 도면 30건, 건축도면 아카이브 DB 100건, 건조물문화재 3D공간정보 50건, 3D프린팅 5점, HBIM 2건을 구축
- 2020년에는 문화유산 3D 스캔데이터 DB 27건 (자연유산 4건, 명승 18건, 국보·보물 건조물 5건), 목조·석조 건조물문화재 원형기록 통합DB 622건이 기록
- 2012년~2020년간 스캔데이터 DB 1,181건(원본파일, PTS), 3D 프린팅 데이터(PLY, STL) 170건, 기타 데이터(DWG 등) 32건 구축

30

03 국가유산 디지털 아카이브 사업현황

국립디지털문화유산센터 구축

- 세종시 국립박물관단지 내에 설립 계획 중인 국내 최초의 디지털 문화유산을 콘텐츠로 하는 복합문화시설
- 첨단 디지털 기술을 통해 문화유산을 기록화 하여 보존하고, 많은 사람들이 이를 접할 수 있도록 전시와 체험 프로그램을 제공
- 디지털 헤리티지 교육, 연구 기능

31

03 국가유산 디지털 아카이브 사업현황

이집트 국제개발협력(ODA)사업

- 이집트의 중요 6개 박물관 및 연구소(이집트박물관, 콥트박물관, 고고연구센터 등)가 소장한 유물들을 디지털 데이터베이스화
- 박물관 및 유물정보서비스 모바일 앱을 개발하여 전시실에 스마트뮤지엄을 구축

32

04 디지털 헤리티지 데이터 플랫폼 구축 기술

Project
Work
Instance
Item

국립박물관 소장 문화유산 3D 애셋 제작

- 공고일시 : 2023/07/27
- 수요기관 : 문화유산체육관광부 국립중앙박물관
- 입찰공고번호 : 20230735832-00

개별 파일 묶음은 인스턴스

개별 파일은 아이템

POINT.

데이터 계층 간 연관 구조를 통해
사용 목적과 활용 용도에 맞는 데이터에
효율적으로 접근할 수 있게 함

- 금동 반가사유상과 연결된 3D, Image, RTI 등 다양한 Instance와 연계
- 금동 반가사유상을 참조하는 프로젝트를 통해 데이터 구축 사업의 목적, 기간, 데이터 생산 업체, 동일 사업을 통해 산출된 다른 유물 정보 연계

33

04 디지털 헤리티지 데이터 플랫폼 구축 기술

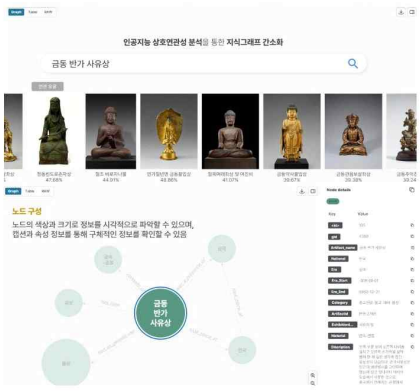
데이터 탐색 목적

반가사유상 중심
가상박물관 콘텐츠 기획

지식그래프 활용 방안

STEP 1
개별 유물의 속성 정보 확인

- 개체의 속성은 연결된 속성 개체로 확인
- 개체를 클릭하여 관련된 모든 상세 정보를 확인
- 연결된 속성 개체의 크기를 통해
  많은 유물들과 연결된 속성을 확인할 수 있으며,
  이를 통해 탐색 및 기획 방향 설정



활용 소프트웨어 : neo4j Gephi

34

04 디지털 헤리티지 데이터 플랫폼 구축 기술

데이터 탐색 목적

반가사유상 중심
가상박물관 콘텐츠 기획

지식그래프 활용 방안

STEP 2

속성을 매개로 관련 유물 탐색

- 반가사유상의 주요 속성 중 하나인 '불상'을 연관성이 있는 유물 탐색
- 유물 유형인 '불상' 외에도 시대인 '삼국', 재질인 '금동'을 매개로 관련이 있는 유물 확인

활용 소프트웨어 : neo4j Gephi 35

04 디지털 헤리티지 데이터 플랫폼 구축 기술

데이터 탐색 목적

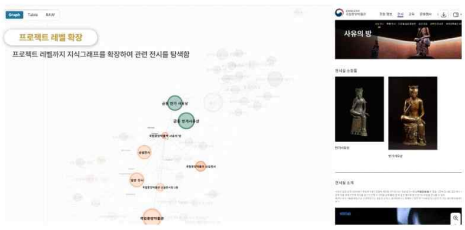
반가사유상 중심
가상박물관 콘텐츠 기획

지식그래프 활용 방안

STEP 3

관련 유물이 포함된 전시로 탐색 범위 확장

- 탐색 범위를 프로젝트로 확장하여 관련있는 유물들이 포함되었던 프로젝트 확인
- 같은 프로젝트에서 함께 전시되었던 유물 목록을 통해 가상 박물관에 함께 전시할 수 있는 유물 탐색



활용 소프트웨어 : neo4j Gephi 36

04 디지털 헤리티지 데이터 플랫폼 구축 기술

데이터 탐색 목적

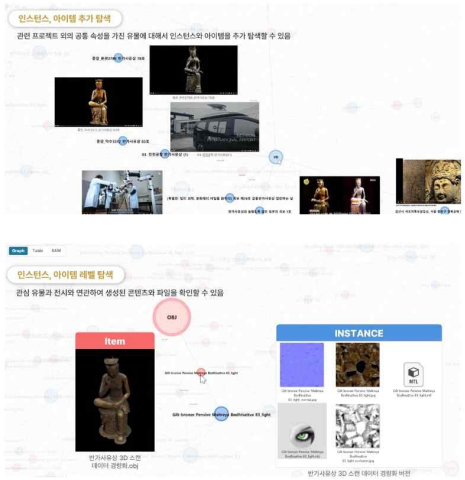
반가사유상 중심
가상박물관 콘텐츠 기획

지식그래프 활용 방안

STEP 4

유물, 전시와 관련하여 보유중인 데이터 탐색

- 탐색중인 유물을 모델링한 파일,
  관련 전시에서 함께 제작되었던 영상,
  작성된 설명 등
  가상 박물관에 전시할 수 있는
  미디어 파일을 함께 확인할 수 있음



37

04 디지털 헤리티지 데이터 플랫폼 구축 기술

AI 상호연관성 분석

전시-유물 연관성 분석
- 유물의 전시 이력과 유물 및 전시 설명 간의 유사성을 분석하여 유물과 전시의 연관성을 분석하는 모듈

유물-유물 연관성 분석
- 유물 속성, 전시 이력, 유물 설명 간의 유사성을 분석하여 유물 간의 연관성을 분석하는 모듈

전시 간 연관성 기반 유사 전시 추천
- 전시 특징 및 전시 설명 간의 유사성을 분석하여 유사한 전시를 추천하는 모듈

개체명 인식 기반 유물 키워드 도출 및 분류
- 개체명 인식을 통해 유물 설명에서 주요 키워드를 도출하고, 이를 기반으로 유물을 분류하는 모듈

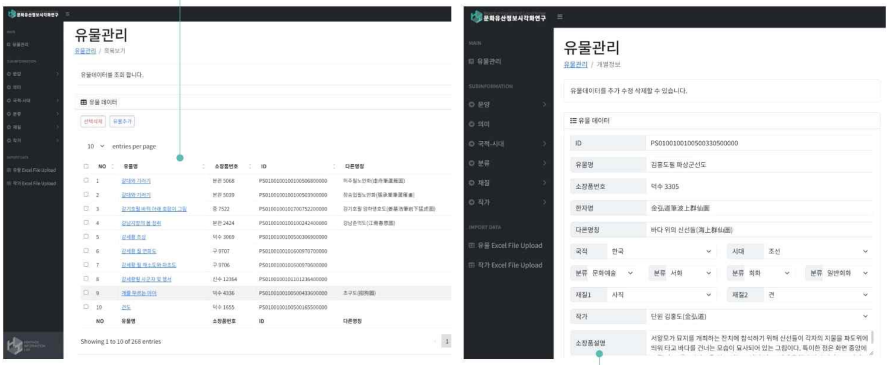
사용자 제시 키워드 기반 유물 군집 분류
- 사용자가 제시한 키워드를 기반으로 유물을 군집화하고, 각 군집의 적합한 세부 주제를 추천하는 모듈

유물 이미지 활용 연관성 분석
- 유물 이미지의 시각적 특징을 분석하여 유사한 유물을 추천하는 모듈

38

04 디지털 헤리티지 큐레이션 플랫폼 구축 기술

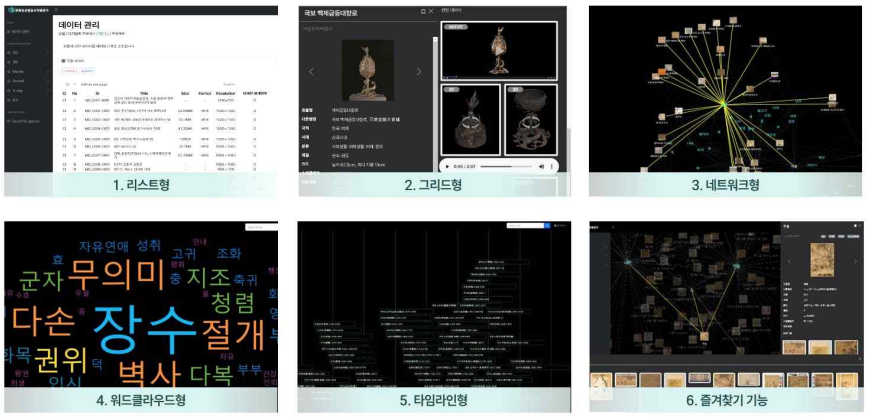
유물의 기본 메타데이터(유물명, 작가, 시대, 설명 등)를 비롯하여
각 유물에 묘사된 **문양 정보**를 선별, 관리할 수 있으며,
해당 유물과 관련된 **이미지**를 업로드하여 관리할 수 있음



데이터 관리, 편집 프로그램 (유물, 문양, 작가정보)

39

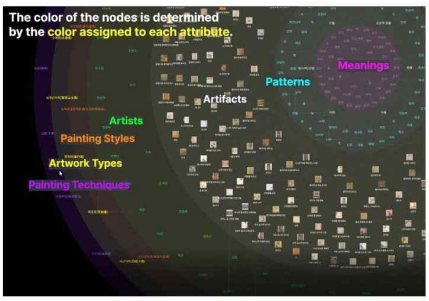
04 디지털 헤리티지 큐레이션 플랫폼 구축 기술

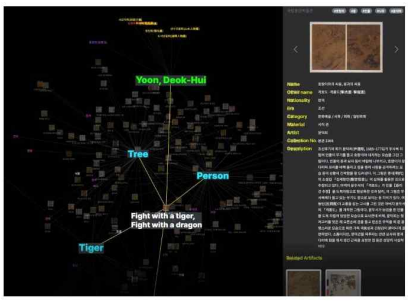


정보시각화 유형

40

04 디지털 헤리티지 큐레이션 플랫폼 구축 기술

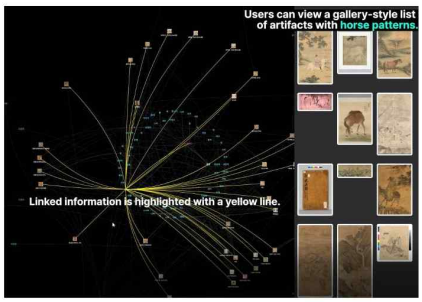


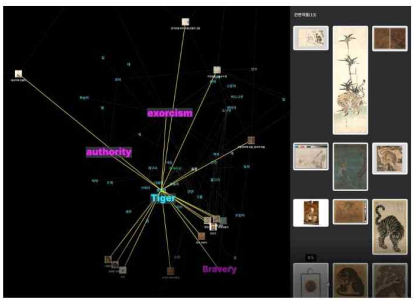


작품 정보
시각화
개요

41

04 디지털 헤리티지 큐레이션 플랫폼 구축 기술

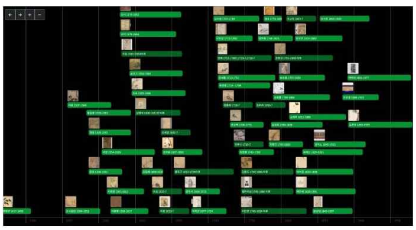


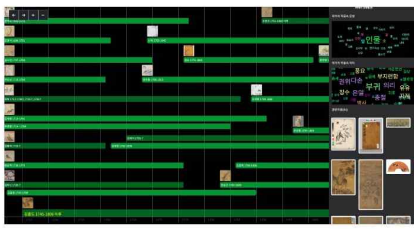


문양과 의미 기반
작품간
관계 시각화

42

04 디지털 헤리티지 큐레이션 플랫폼 구축 기술

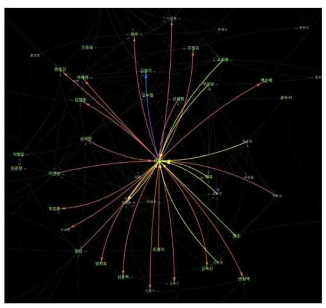


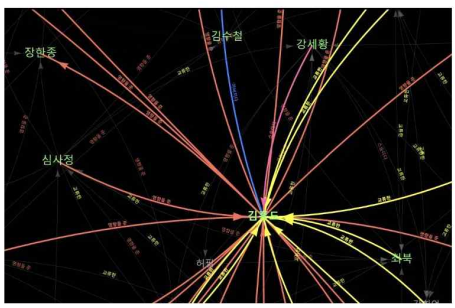


작가의 생몰연도
기반 타임라인
시각화

43

04 디지털 헤리티지 큐레이션 플랫폼 구축 기술

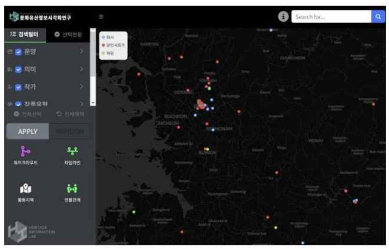


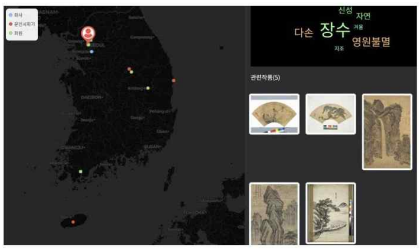


작가 간의 관계정보
시각화
(스승-제자, 가족, 영향)

44

04 디지털 헤리티지 큐레이션 플랫폼 구축 기술





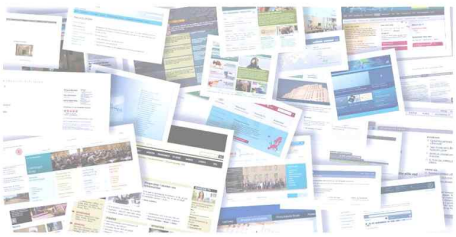
작가의 연관 장소 시각화
(출생지, 활동지)

45

05 향후 과제

❖ 디지털 유산 보존 개념 설정

- 자료들이 디지털형태로 존재하고 보관된다는 생각에 보존 활동이 의미가 없다는 보존 개념의 부재
- 디지털 형식이 대량생산되고 동일 내용에 버전이 다양한 콘텐츠가 생산됨에 따라 무엇이 중요한 것인지에 대한 판단이 어려워서 관리가 미흡
- 디지털 기술의 변화 주기가 단축됨에 따라 기존 소프트웨어에서 구축된 자료들이 호환되지 않아 손실될 가능성이 높음
- 태생 디지털 유산에 대한 보존·관리·활용 제도는 아직 전 세계적으로도 그 사례를 찾아보기 어려우며, 민간 또는 공공의 일시적 프로젝트 단위로 추진된 경우가 잦음

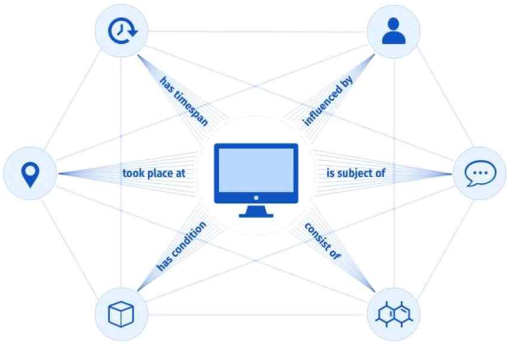


46

05 향후 과제

❖ 디지털 헤리티지 데이터에 대한 접근성 향상

- 디지털 문화유산 데이터 활용을 위한 다양한 포맷 지원 부족
- 건축물, 유적 등 문화유산의 개체 단위 활용 불가능
- 활용 사례가 3D 프린팅 등 특정분야에 편중
- 웹서비스에 활용하기에 데이터 용량이 큼
- 디지털 문화유산 데이터의 홍보 부족
- 검색 서비스 및 시각화 콘텐츠의 고도화

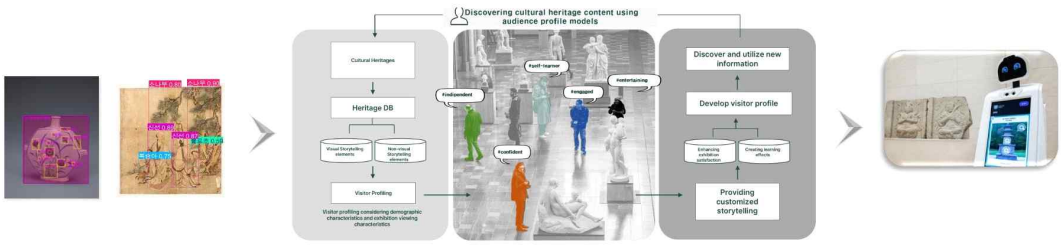


47

05 향후 과제

❖ 디지털 데이터의 접근성을 높이기 위한 인공지능 기술의 적용

- 이미지, 텍스트, 음성 등의 다양한 형태의 데이터를 통해 문화유산의 숨겨진 의미를 추출하는 AI 기반의 의미분석 필요
- 관람객 프로파일 모델 검증을 통해 기존 관람객 타입 고도화
- 관람객 유형별 선호를 반영한 문화유산 콘텐츠 추천 시스템 구축 등에 활용



48

세션2 발표문 4

AI 시대,
국가지식문화자원 데이터 허브, 국립중앙도서관

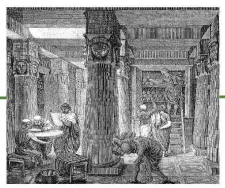
김수정(국립중앙도서관)
2024. 12. 13.

주요내용

- AI 시대 국가도서관의 대응
- 국립중앙도서관 디지털화, AI 기반 프로젝트 소개
- 향후 추진 방향

2

AI 시대: 지식의 소유에서 접근, 인간과 AI의 데이터 활용과 생성



기원전 3세기경
알렉산드리아도서관
세상의 모든 책을 소유
장서 수*: 5만여권~70만권

*이레네 바예호, 이경민역,「갈대 속의 영원」, 반비, 2023, 75-76쪽.

1998년 설립
구글
세상의 모든 정보를 접근 가능
수십억개 웹페이지 인덱싱
Google Books 1억 3,000만 권 목표

*(https://en.wikipedia.org/wiki/Google_Books

2018년 이후 LLM 본격 성장
OpenAI의 GPT와 Google의 BERT 등장
방대한 데이터 학습, 유용한 정보생성
GPT-3 3000억개 토큰 학습(230만책)

*https://en.wikipedia.org/wiki/BERT_(language_model)/
https://en.wikipedia.org/wiki/Generative_pre-trained_transformer

3

ㅇ 디지털 자원의 폭발적 증가

- 지식의 생산과 확산의 속도는
  우리사회의 혁명적 변화를 가져옴

ㅇ 인간과 기계가 함께 학습하는 시대

- 고서, 고신문, 근현대자료들이 디지털화되며 새로운 르네상스를(k-고전의 재발견 등) 예고
- 데이터 처리 분석, 교육 및 학습의 혁신, 연구개발 가속화 등

4

사람과 기계 모두에게 낮은 도서관 데이터 장벽

* 장서 데이터화를 통해 기계와 인간 사이의 장벽을 낮추고, 정보 접근성을 높이려는 도서관계 노력

데이터로서 컬렉션에 관한 산타바바라 성명(2017)

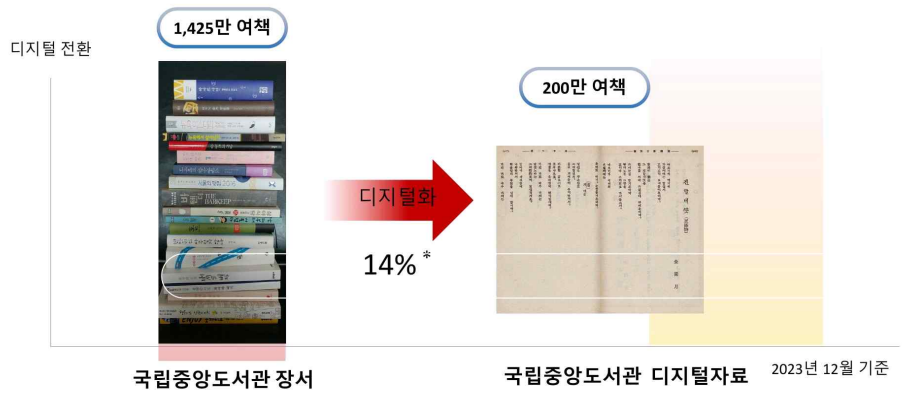
도서관 장서를 데이터로 인식
디지털컬렉션 데이터화 중요성 강조
데이터화된 컬렉션의 개방과 접근성을 높이기 위한 방향 제시
* https://zenodo.org/records/3066209

데이터로서 컬렉션에 관한 벤쿠버 성명(2023)

산타바바라 성명서를 업데이트한 문서로, **데이터 주권**과 **인공지능의 확산**을 고려한 책임감 있는 데이터 사용 촉구
https://zenodo.org/records/8342171

5

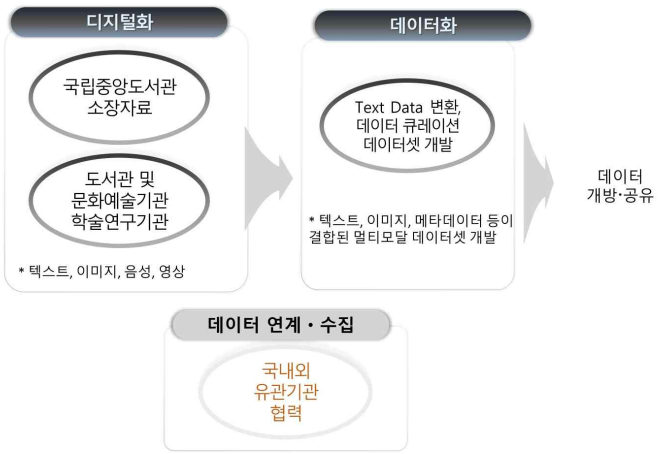
국립중앙도서관 소장자료 디지털화, 데이터화 추진



* 1,425만여책 중 복본, 문제집, 사전 등 디지털화 제외자료를 적용하면 약 47% 수준

6

국가지식문화자원 디지털화, 데이터화 추진



7

코리안메모리 프로젝트 참여기관 현황

- 122개관 26만 여책 디지털화 (2023년 12월 기준)

  - 도서관 43개관, 박물관 18개관, 미술관 14개관 등 국립중앙도서관 뿐만 아니라 전국 문화예술기관 소장자료도 디지털화, 컬렉션 구축, 데이터화 추진

구분	2018	2019	2020	2021	2022	2023	합계
참여기관(개)	14	19	48	43	46	42	211
구축 수량 책(점)	6,900	12,283	98,547	52,211	45,108	47,325)	261,093

참여기관 유형별							
도서관	박물관	미술관	문학관	교육기관	문화예술기관	단체・기관・개인	계
43	18	14	4	6	16	21	122*개관

8

디지털화 대상자료 : 실물도서, 영상, 사진, 도록, 원고, 리플릿, 팜플릿 등

다양한 자료를 디지털화하며 멀티모달데이터 구축

텍스트	도서, 기록문서, 논문, 신문기사, 원고, 악보, 리플릿, 팜플릿 등
이미지	회화, 삽화, 사진, 포스터, 지도, 그래픽 디자인 등
음 성	영화, 뉴스방송, TV프로그램, 구술, 음악(음원), 오디오 등
영 상	음악, 테이프, 디스크, 라디오 방송, 구술 등

9

해외사례: 미의회도서관

ㅇ 방대한 데이터 자원을 확보하고 API 제공 및 AI 기술 실험 등 다양한 시도를 통해
데이터의 허브이자 데이터 센터로서 역할 수행

- **(방대한 데이터 자원)** 다양한 디지털컬렉션과 광범위한 아카이빙을 통해 데이터 자원 확보
- **(데이터 접근성 강화)** API 및 데이터셋을 통해 외부 연구자들에게 데이터를 개방하고, AI 학습 데이터 활용 지원
- **(LC Lab)** 의회도서관의 디지털 전환 지원 및 AI와 머신러닝 기술을 실험하며 데이터 접근성, 활용성 강화

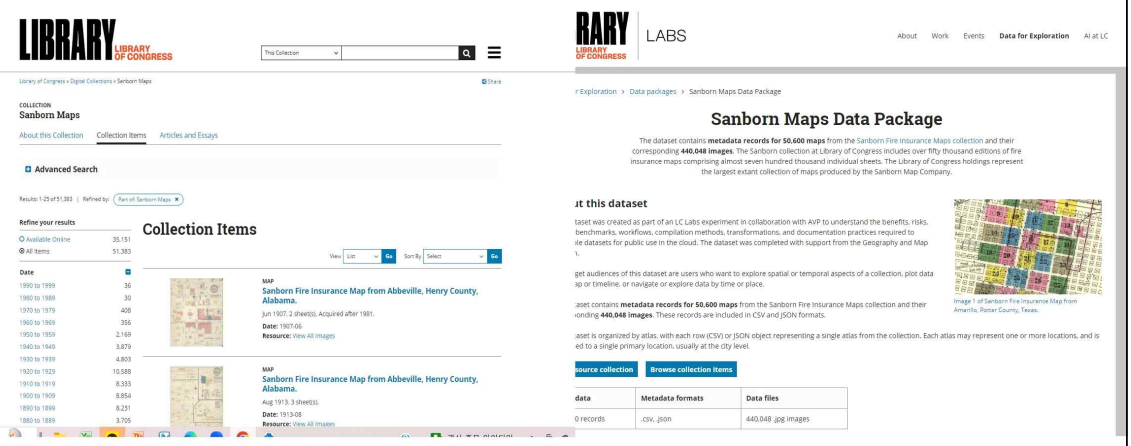
예시1) Chronicling America 신문아카이브 (1789~1963): 수백만페이지 신문 디지털화, OCR 텍스트데이터 구축, 본문검색, 분석과 활용을 위한 API 제공(Chronicling America API)

예시2) (Sanborn Maps Data Package) 컬렉션 **Sanborn Maps 컬렉션의 5만여개 지도와 44만여개의 이미지를 데이터 제공**

메타데이터	메타데이터 형식	데이터 파일
50,600개의 레코드	.csv, .json	440,048 .jpg 이미지

10

Sanborn Maps 컬렉션에 대한 데이터패키지 제공



11

해외사례: 영국 국립도서관 (British Library)

ㅇ 국가연구데이터 관리와 활용을 주도하고, 장서의 데이터화와 데이터 연구 혁신을 지원하며 데이터의 허브이자 데이터센터로서 역할 수행

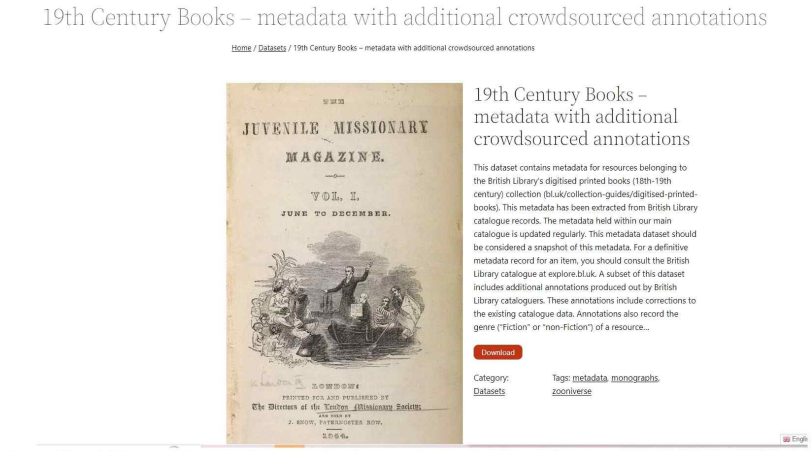
- **(국가 연구데이터)** 데이터 생애주기에 맞춘 전략을 수립하여 데이터를 관리하고, 연구자들의 데이터 활용 지원
- **(방대한 데이터 자원)** 디지털장서의 데이터화, 데이터 컬렉션화 를 추진하고 연구 및 AI 학습 데이터셋 제공
- **(BL Labs)** 디지털자료 분석을 위한 AI 기술개발,혁신적인 연구 지원, 이용자 참여 데이터 구축 등 데이터 활용 지원

예시1) 19th Century Books (19세기 책-- 크라우드소싱으로 추가된 주석이 포함된 메타데이터) : 19세기 출판된 책들의 서지정보와 크라우드소싱으로 추가된 주석 포함, 연구자들이 텍스트 분석과 같은 데이터 기반 연구에 활용 가능 https://labs.biblios.tech/item/19th-century-books-metadata-with-additional-crowdsourced-annotations/)

* 디지털 컬렉션 설명과 검색 기능 개선, 이용자들이 다양한 프로젝트에 참여하여 도서관 자료를 함께 구축할 수 있도록 LibCrowds (클라우드소싱 플랫폼) 운영

12

19th Century Books – metadata with additional crowdsourced annotations



https://labs.biblios.tech/item/19th-century-books-metadata-with-additional-crowdsourced-annotations/

13

국립중앙도서관 : 국가지식문화자원 데이터 허브

ㅇ 국가 디지털자원의 적극 수집 및 데이터 관리, 국가지식문화자원의 디지털화와 데이터화를 추진하고 데이터 접근성을 강화하기 위해 노력 하며 국가지식문화 자원에 대한 데이터 허브이자 데이터 센터로서 역할 수행

- **(방대한 데이터 자원)** 국립중앙도서관 소장자료뿐 아니라 문화예술기관, 학술기관, 정책연구기관 등의 광범위한 자료를 디지털화하고 데이터화 , 큐레이션을 통한 데이터 컬렉션화 추진

  * 국가지식문화자원 메타데이터, 콘텐츠데이터(텍스트, 이미지, 영상 등) 뿐만 아니라 빅데이터플랫폼 정보나루(공공도서관 대출정보 등), 학술자원 공유를 위한 OAK 운영, 학술지 저작권 정보 관리 등 방대한 데이터 수집 및 관리

- **(데이터 접근성 강화)** API 및 데이터셋을 통해 외부 연구자들에게 데이터를 개방하고, AI 학습 데이터 활용 지원
- **(NLK Labs)** 도서관 업무에 AI, RPA 등 신기술 적용을 실험하며 데이터 분석과 활용 지원

예시) 코리안메모리 프로젝트 데이터 서비스

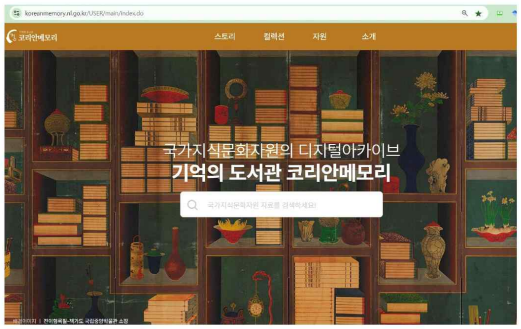
- 이현세컬렉션 : 이현세만화가의 원화, 구술채록, 자료 등 디지털자원에 대한 데이터셋(이미지+메타데이터 등) 제공
- 근현대한글잡지 컬렉션 : 한글잡지에 대한 메타데이터, 원문 텍스트파일, 이미지 파일 데이터셋 제공

  * https://koreanmemory.nl.go.kr/USER/contents/SV0301000.do?page=&schM=view&editorId=12&dataIdx=2079101

14

코리안메모리 프로젝트: 전국 문화예술기관 소장자료 디지털화, 데이터화 추진

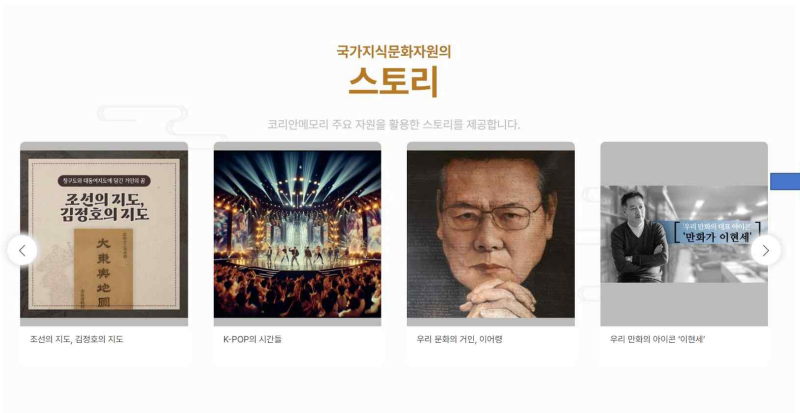
- **(코리안메모리 프로젝트)** 한국의 개인 및 국가, 단체의 기억을 담은 국가 지식문화자원에 대한 디지털 아카이브 프로젝트, 디지털 큐레이션 프로젝트



15

코리안메모리 플랫폼

국가지식문화자원을 큐레이션하여 인물, 사건, 장소 등 다양한 주제별 디지털컬렉션 제공



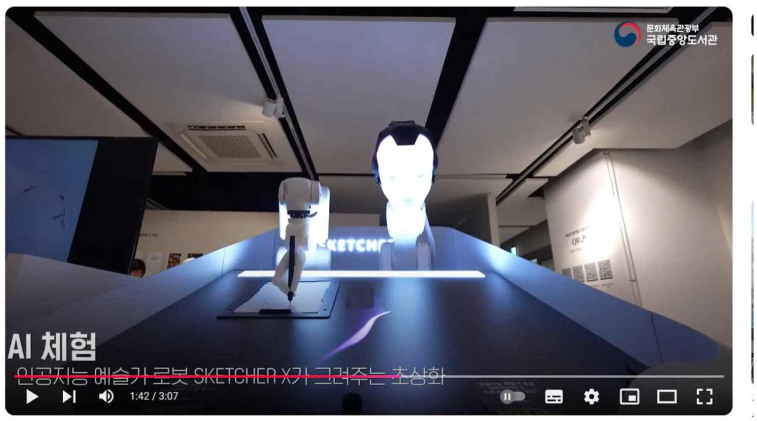
예) 이현세 컬렉션

16

이현세 컬렉션(디지털화, 데이터화) (관련 데이터셋 코리안메모리 플랫폼 등록 예정)

17

이현세 컬렉션 관련 전시



국립중앙도서관「이현세의 길 : K-웹툰 전설의 시작 특별전」(2024)

https://www.youtube.com/watch?v=aCemIfoGXio

18

(영화 컬렉션) 광고를 통해보는 한국영화의 발자취(2023년 구축)

ㅇ **제작: 국립중앙도서관, 관련 연구 및 집필: 이준엽**(한양대학교 연극영화학과 영화학 박사), (**감수) 함충범**(한국영상대학교 영화영상과 교수)

- 국립중앙도서관 (신문기사. 잡지, 단행본). 영화진흥위원회(시나리오, 포스터, 영상자료) 자료 디지털 큐레이션

- 광고와 영화, 시대의 의미정보와 관계정보를 구축하여 한국영화에 대한 체계적인 정보 제공

19

(영화 컬렉션) 2024년 구축 영화 컬렉션

ㅇ 영화와 문학사이

- 1920~1930년대 발행된 단행본, 잡지, 신문기사들을 시대별로 모아, 영화와 문학이 어떻게 서로 영향을 주고받았는지를 보여줌

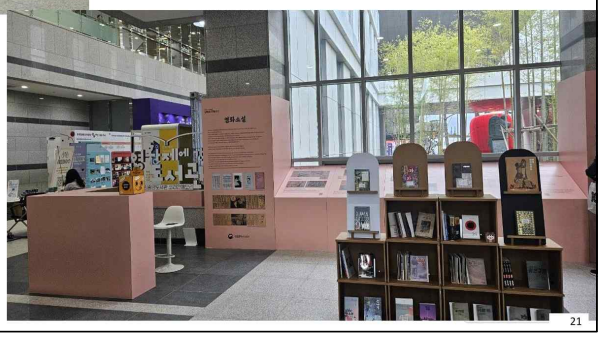
ㅇ 영화 증언 (임권택 감독)

- 1973년 영화진흥위원회 창립작품으로 제작한 극영화 <증언> 제작 기록

20

2024 the Busan International Film Festival

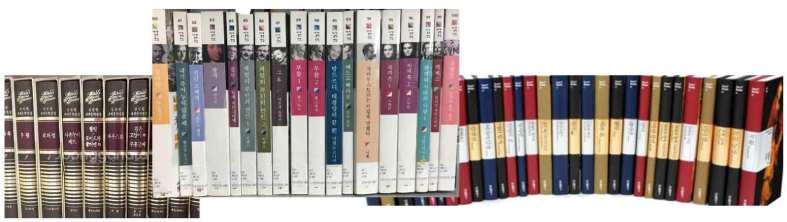
23/24년 부산국제영화제 기간 중 컬렉션 전시



21

(번역문학 컬렉션) 세계문학전집으로 보는 번역문학(1950~)

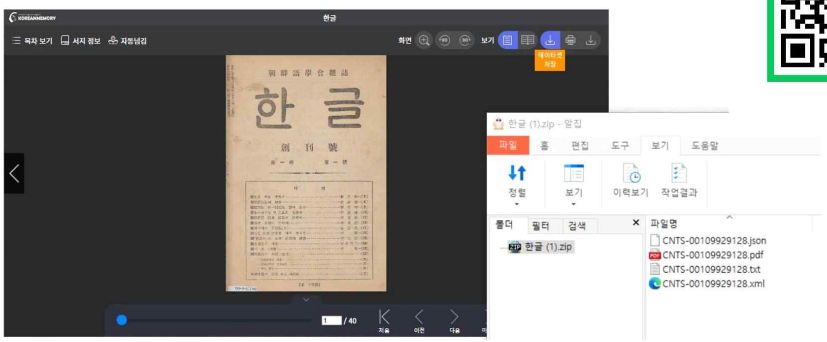
- **제작: 국립중앙도서관, 관련 연구 및 집필: KAIST 정서현 교수, 한상은 박사**
- **국립중앙도서관 (소장장서 및 목록). 기타 세계문학전집 출판정보 및 관련 자료 수집**
- 50년대부터 현재까지 세계문학전집류에 포함된 영미문학 번역서 정보를 체계적으로 정리, 아카이브 구축



22

한글(근현대 잡지): AI OCR 활용 텍스트 데이터 구축

○ 텍스트 데이터 및 관련 데이터셋 제공(텍스트 파일, 이미지, XML, JSON)



23

(근현대 잡지): AI OCR 활용 텍스트 데이터 구축

○ 2024년도 근대잡지 4종 55책 텍스트 데이터 구축 * 일부 권호 우선 구축
- (근대잡지 4종) 조광, 청춘, 문장, 카톨릭청년
- 표, 그림, 만화, 광고, 사진 등 이미지 데이터 구축

조광 잡지 1936년 1월호 : 광고

조광 잡지 1936년 1월호 : 사진

조광 1936년 1월호: 표지

24

도서관 데이터 현황(메타데이터, 콘텐츠데이터, 이용자/업무데이터)

구분	생성 형식
메타데이터	[국가서지] 2,654만 여건: 오프라인 695만여건 / 온라인 1,959만여건 * 국립중앙도서관 소장자료 [국가전거] 1,990천여건: 저자명 1,730천 여건(개인, 단체) / 주제명 25만여건 [LOD(링크드 오픈 데이터)] 데이터 2,768만 여건, 트리플 7억 3,164만 여건 *국가서지 LOD [식별체계] ISBN(도서) 485만 여건 / ISSN(연속) 2.8만 여건, ISNI 144만 여건(이름), UCI(콘텐츠) [소장사항/관리정보] 데이터베이스* 주로 고정된 필드에 저장된 정형 데이터로서 관계형 데이터베이스로 관리 [국가자료종합목록] 서지 1,930만 여건 / 소장데이터 8,531만 여건 * 공공도서관 등 전국 2,000여 개 도서관 자료 [한국고문헌종합목록] 서지데이터 50만 여건 * 국내외 고문헌 소장기관 자료 [정책정보종합목록] 182만 여건 / 콘텐츠 15만 여건* 정부부처 자료실 및 공공기관 자료
콘텐츠데이터	[원문] 200만여건/ 4억 3,396만 여면, 문화예술기관(박물관, 미술관 등) 자료 26만여건 포함, 발행 후 5년 지난 자료 디지털화 가능 [텍스트 데이터] 17,520건(책) / [그림, 표 이미지] 13.8만 여건 [목차] 1,481천 여건 (단행 50여만건, 연속 97만여건) / [기사색인, 초록] 1,069천 여건
이용자/업무 데이터	[정보나루] 장서데이터 1억 7,783만 여건, 이용자 데이터 3,603만 여건, 대출데이터 20억 9,127만 여건 * 공공도서관 데이터 * 이용자 데이터: 대출, 검색, 참고서비스, 소셜미디어, 프로그램 참여 등을 통해 생성된 데이터, 업무 데이터: 수서, 목록 작업 등 도서관 업무에서 생성된 데이터

* 국립중앙도서관 미래 비전(2024~2028) 데이터 분과 보고서를 참고하여 작성

25

국립중앙도서관 데이터 특징

- **(신뢰성 담보)** 데이터 표준화를 기반으로 전문사서들이 신뢰성 있는 데이터를 완전하고 포괄적 구축 제공
- **(다양성 확보)** 책, 학술지, 신문, 영상, 음원, 이미지 등 다양한 자료유형, 다양한 주제, 시대를 포괄하는 데이터
- **(접근성과 활용성 높음)** 국가지식문화자원 데이터를 통합 관리하여 인간과 기계가 접근 가능하게 개방

  * FAIR원칙 준수 노력: 탐색, 접근, 상호운용과 재사용 가능하게(findability, accessibility, interoperability, reusability, https://force11.org/info/the-fair-data-principles/)

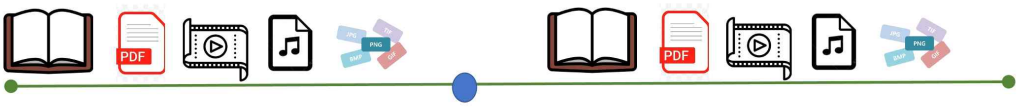
- **(방대한 데이터)** 국내외 국가지식정보자원에 대한 포괄적 수집과 디지털화로 방대한 데이터 보유
- **(국내외적 협력)** 국내 도서관, 연구기관, 문화예술기관, 해외 도서관 등과 교류 협력
- **(이용자 참여 가능)** 자원 기증, 이용자 참여 컬렉션/데이터 구축(AI OCR 공유서재, 도서관 프로그램 참여 등)

26

도서관 데이터 특성: AI 학습에 유용한 고부가가치의 원천 데이터

ㅇ 2020년 이전 출판된 도서관 소장자료는 AI 개입이 미미하여 인간이 만든 저작물로서 원본성과 신뢰성이 높아, 학술 연구나 AI 모델 학습에 유용한 고부가가치의 원천 데이터

* 2020년 GPT 3 등장으로 대형 언어 모델 본격적인 대중화 상업화 (https://en.Wikipedia.org/wiki/GPT-3)



대부분 인간이 만든 저작물

2020년
GPT 3 대형언어모델 등장

인간 + AI , AI 생성 가능성 존재

27

AI 활용 : AI OCR 문 프로젝트

ㅇ 딥러닝 기법을 활용한 OCR 기술을 통해 한글문서 인식개선 가능성 연구 수행(2020년)

ㅇ AI OCR시스템을 통해 텍스트 데이터 시범 구축(2021년)

ㅇ AI-OCR 텍스트데이터 구축 (1,716책, 2023년 기준)

- 금속활자본, 목판본, 목활자본, 필사본, 연활자본 등 다양한 판본을 대상으로 구축
- 텍스트 활용 연관 서비스에 필요한 형식으로 제작(XML 파일, Text 파일, JSON 파일)

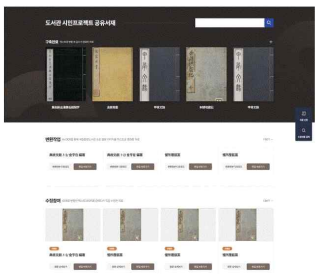
28

AI-OCR(문 프로젝트) 플랫폼 서비스 '공유서재'(2024년 12월 오픈 예정)

ㅇ 크라우드소싱 기반 텍스트데이터 구축 서비스

- AI OCR로 변환된 텍스트데이터를 이용자가 직접 수정 및 다운로드하여 활용

ㅇ (대상자료) 저작권이 해결된 자료



메인 화면

크라우드소싱기반 텍스트에디터

29

향후 추진 방향

디지털자원 확대, 데이터화 강화

- 텍스트 데이터, 이미지 등 비정형데이터 구축 확대, 이용자 참여 컬렉션/데이터 구축 강화

데이터 접근과 활용성 강화

- 다양한 API, 데이터셋 다운로드 제공 등

AI 기반 혁신 추진

- AI 활용 도서관 서비스 혁신 지속적 발굴

국가적 협력 및 글로벌 연계

- 국내외 기관 간 협력 강화, 국제적 데이터 표준 준수, 연구 협력 등 국제적 협력 촉진

30

세션2 '기관에서의 디지털 인문학'에 관한 종합 대담문
: 연구자와 기관은 어떻게 만날 수 있는가

유인태(전남대학교 중어중문학과)

디지털 환경에서 인문학 자료를 제공하기 위한 구체적 고민과 그것을 위한 여러 데이터 처리 과업을 가장 적극적으로 수행하고 있는 주체가 인문학 자료를 직접 다루고 있는 여러 기관임에도 불구하고, 그간 기관에 계신 여러 전문가 선생님들을 한 곳에 모시고 관련된 이야기를 여쭈어 들을 기회는 거의 없었던 것 같습니다. 기관에서 소장한 자료에 접근하는 연구 채널이 문학과 역사와 사상 등 여러 갈래로 나누어져 있기에 대학의 연구자를 중심으로 개최되는 학술대회의 경우 그 성격과 결이 저마다 달라 여러 기관에 계신 선생님을 한 자리에 모시기 어려운 까닭도 있을 것이고, 또 개별 기관이 처해 있는 현실적 상황이 저마다 다르기에 그러한 각 기관의 특수성을 고려하지 않은 상태에서 기관에서 진행 중인 과업에 관한 내용을 일반론적 차원에서 여쭈어 듣고자 하는 것이 어려운 탓도 있을 것입니다. 그 외에도 여러 이유를 찾아볼 수 있겠으나, 적어도 오늘만큼은 이 자리에 여러 기관의 전문가 선생님들께서 어렵게 와 주셨으니, 각설하고 길지 않은 시간 동안 지금 당장 그리고 앞으로 고민이 필요한 여러 이야기를 최대한 나누는 것이 좋겠다는 생각입니다. 연말 한창 기관의 일로 바쁘신 가운데 시간을 쪼개어 오늘 이 자리에 참석해주신 네 분 선생님께 감사하다는 말씀을 드립니다. 대담 형식으로 종합 논평을 중간에서 조율하는 역할을 제가 맡게 되었지만, 여러모로 저는 경험과 이해가 많이 부족한 연구자입니다. 논평의 중간자로서 제가 역할을 하는 가운데 혹여 잘못된 정보를 말한다든지 또 부연이 필요하겠다 싶은 지점이 있을 경우, 스스럼없이 지적해주시고 또 알려주시면 감사하겠습니다.

○국사편찬위원회에 계신 류준범 선생님께 드리는 질문

-기존에 역사 관련 기관들은 기초 사료를 디지털화해서 웹에서 그것을 볼 수 있게끔 제공하는 역할만 해왔는데, 이제는 그것만 하기 어려운 시대가 되었으니 그 다음 단계의 무언가를 고민하기 위해 학계와 기관 간의 협업이 필요하고, 또 그러한 기회를 마련하기 위한 의식적 노력을 해야 한다는 말씀을 하셨습니다. 기관에서 그동안 해온 사료 편찬 업무가 이제는 기본적으로 데이터베이스 구축 과업으로 이루어지고 있고, 그러한 측면에서 단순 사료 텍스트 데이터가 아니라 연구의 차원에서 가공된 역사 데이터 제작에 대한 수요가 향후 늘어나게 될 텐데, 대학의 연구 역량이 필연적으로 기관의 과업과 만나야 하는 채널이 더욱 확대되지 않을까 싶습니다. 관련해서 기관이 노력해야 할 지점과 대학의 연구자들이 노력해야 할 지점이 다를 듯 싶은데, 기관 차원에서 해야 할 노력이 무엇인지 그리고 대학에 소속된 연구자들이 노력해야 할 점이 무엇인지 말씀해주실 수 있는지요?

-전라남도 신안 하의도 사례를 통해 여러 기관에서 아카이빙한 사료를 통합적으로 살펴보는 과정이 향후 역사학 연구에 있어서 매우 중요한 프로세스가 되겠다는 생각이 들었습니다. 어

떻게 보면, 앞으로가 아니라 이미 역사학 연구가 그러한 지형 위에 있는 것이 아닐까 하는 생각도 듭니다. 개별 기관에서 구축한 유용한 아카이브 자원을 잘 찾아서 자신의 연구에 활용하는 것은, 연구자의 입장에서 중요한 역량이 될 것입니다. 한편으로 각각의 기관에서 구축한 사료 아카이브를 통합한 아카이브 서비스를 구현하는 작업 즉 여러 사료 아카이브의 아카이브로서 메타 아카이브 구현에 대한 논의도 향후 지속적으로 나올 듯한데, 이에 관해서는 어떤 생각을 가지고 계신지 여쭈어봅니다. 예컨대 발표문에서 말씀하신 한국학중앙연구원의 한국학 자료통합플랫폼도 그러한 성격에 해당하는 플랫폼이 아닐까 싶습니다.

-'저작인격권'에 대해 말씀해주신 부분이 매우 와 닿았습니다. 삼일운동데이터베이스를 예시로 말씀해주셨지만, 큰 규모의 디지털 인문학 연구는 필연적으로 다수의 연구인력이 참여하는 협업이 진행될 수밖에 없고, 그렇게 되면 결과물의 저작권을 어떻게 이해-수용해야 할 것인가의 문제가 발생할 수밖에 없습니다. 그러한 측면에서 저 또한 선생님께서 말씀하신 대로 저작인격권에 대한 논의가 앞으로 많이 이루어져야 한다고 생각합니다. 디지털 인문학적 관점에서 접근할 경우 이 문제는 데이터 편찬에 있어서 기여도의 문제와도 연결이 되는 것 같습니다. 결국 연구자 입장에서 그와 같은 과업에 참여하려면 저작권의 문제도 있지만 그 연장선상에서 그러한 과업에 참여한 이력이나 결과물이 연구 실적으로 인준될 가능성이 있는가의 현실적 문제와도 맞닿아 있는 것 같습니다. 가령 삼일운동데이터베이스 구축 과업은 거대한 협업 프로젝트로 진행된 것으로 알고 있습니다. 프로젝트를 진행하시면서 그와 같은 맥락의 고민이나 생각을 많이 하셨을 듯한데, 향후 역사학 데이터 편찬에 참여하는 과업 형식 및 결과물이 공식적인 연구 실적으로 인준될만한 가능성이 있는지, 만약 있다면 어떠한 단서 때문인지 없다면 또 어떠한 지점 때문에 그렇게 생각하시는지 여쭈어봅니다.

○한국유교문화진흥원에 계신 김사현 선생님께 드리는 질문

-발표문에서도 소개해주셨지만, 한국유교문화진흥원에서 현재 중심적으로 구축하고 있는 〈충청국학 디지털 아카이브〉는 모든 유관 정보를 시맨틱 데이터로 편찬해서 제공하는 방향으로 데이터를 가공하고 있는 것으로 짐작됩니다. 일반적인 아카이브 기관들의 경우 대체로 텍스트 데이터 중심의 문헌 데이터 가공 및 제공에 여전히 주력하고 있는 것으로 보입니다. 그러한 측면에서 한국유교문화진흥원의 행보는 상당히 돋보인다고 해야 할까요? 시맨틱 데이터를 기반으로 한 문헌 자료 아카이브를 구축하는 프로세스의 특징이 있다면 어떠한 것인지 궁금합니다. 예컨대 장점이라고 한다면 무엇인지, 한편으로 과업 프로세스에서 처할 수밖에 없는 현실적 어려움도 있을 듯합니다. 선생님의 경험에 비추어 그에 관한 내용을 잠깐 말씀해주시면 감사하겠습니다.

-선생님께서 계신 기관인 한국유교문화진흥원은 충청남도 논산에 있습니다. 발표 자료를 통해서도 강조해주셨지만, 지역 기관의 경우 디지털화 및 아카이브 구축을 위한 전문 인력을 구하는 데 있어서 여러 현실적 어려움이 있을 듯합니다. 아무래도 유관 분야에서 전문적인 교육을 받는 인력의 다수가 서울에 있을 가능성이 크고, 그들 가운데 지역에 내려와서 과업을 수행할 만한 인력을 찾는 것이 현실적으로 쉽지 않을 것이라 여겨지기 때문입니다. 한편으로 지역의 역사문화자료가 지닌 가치와 의미를 고려할 때 해당 자료를 대상으로 한 디지털 인문학적 성

격의 프로젝트에 직접 참여하는 경험은, 일선 대학이나 대학원의 교육 과정에서 쉽사리 경험하기 어려운 무척 희귀한 교육 체험이라는 생각이 듭니다. 사람은 서울에 있고, 일은 지역에 있는 상황이 문제가 되는 것인데, 관련해서 최근 지역 기관에서 시도되고 있는 디지털 인문학적 과업의 중요성이라고 해야 할까요 혹은 학술적 의미라고 해야 할까요? 교육의 차원이든 연구의 차원이든 여러 의미가 있을 듯한데, 발표를 통해서는 미처 강조하지 못한 그런 지점이 있지 않을까 싶습니다. 그에 관해서 조금 더 부연해주실 수 있을지요.

-지역 아카이브의 정체성은 아무래도 해당 지역의 역사-문화적 외연과 별개로 생각하기가 어렵습니다. 지역의 역사 인물이라든지 많이 알려진 문화유산 등이 아카이브 과업의 중심이 되거나 또는 그에 준하는 매개로서 작동할 수밖에 없을 듯합니다. 주로 특정 역사 인물을 선양하거나 특정 문화유산을 대상으로 한 콘텐츠를 기획하거나 제작하는 것 등이 그러한 대표적 양태가 아닐까 싶습니다. 한편으로 지역의 역사-문화 자원을 다룬다고 해서 꼭 해당 지역과 직접 관련된 무언가를 드러내고 강조하는 방향으로 과업을 기획-수행할 이유도 없다 생각합니다. 지역 자원이 또 다른 지역 자원과 데이터로 연결될 수 있다면 지역을 벗어나서 여러 의미의 연쇄 가운데서 해당 지역에 관한 이해를 새롭게 도모할 수 있는 단서로도 작용할 수 있을 것입니다. 선생님께서는 이러한 지점을 늘 고민하고 계시리라 생각합니다. 기관에 계시면서 여러 아카이브 사업을 기획하시고 또 실무를 맡아 일을 진행해 나가시는 가운데, 그러한 맥락의 고민이나 문제의식을 어떻게 풀어 나가고 계신지 여쭈어봅니다.

○국립중앙도서관에 계신 김수정 선생님께 드리는 질문

-국립중앙도서관은 그야말로 분야를 불문하고 온갖 성격의 자료를 모두 다루는 기관이기 때문에, 기관 차원에서 소장 자료를 디지털화하거나 데이터화하는 과업의 규모나 성격이 여타 기관과는 많이 다를 것이라는 생각이 듭니다. 예를 들어 그간 국립중앙도서관에서 구축한 가장 큰 규모의 디지털 자원으로는 국가서지 LOD(https://lod.nl.go.kr/home/)를 거론할 수 있지 않을까 싶은데, 1년 전인 2023년 12월 기준 트리플 데이터 건수만 해도 7억3천4백여 건입니다. 실로 어마어마한 규모라고 할 수 있겠습니다. 큰 규모의 자료를 디지털화하는 과업을 기획하시거나 혹은 그에 참여하시면서 아마도 여러 맥락의 어려움을 직간접적으로 경험하시지 않을까 싶은데, 관련해서 부연해주실만한 일화라든지 혹은 시사점 같은 것이 있다면 말씀을 부탁드립니다.

-국립중앙도서관은 큰 규모의 장서를 갖추고 있는 만큼, 자료 활용 서비스가 연구자들이 작업하고 있는 연구의 맥락과 직접적으로 연계한다면 매우 좋은 시너지가 날 것이라고 생각합니다. 몇 년 전부터는 전근대.근대기 문헌 자료 가운데, 연구자들이 연구에 활용하기를 원하는 자료들을 OCR 처리해주는 그런 사업 또한 진행하고 계신 것으로 알고 있습니다. 단순히 연구자들이 원하는 자료를 제공하는 수준에서 더 나아가, 도서관에서 소장하고 있는 중요한 문헌 자료를 대상으로 연구자들로 하여금 마크업 데이터를 편찬하게 한다든지 또는 거기서 더 나아가 데이터 분석이나 데이터 시각화를 진행하게끔 한다든지. 그와 같은 디지털 인문학적 성격의 새로운 협업 모델을 개발하는 것 또한 불가능한 일은 아닐 것이라 생각합니다. 관련해서 어떠한 생각과 고민을 하고 계신지 여쭈어 봅니다.

-멀티모달 데이터는 AI를 매개한 활용 가치가 앞으로 점점 더 커질 것이기 때문에, 아마도 국립중앙도서관은 자체적으로 소장하고 있는 다양한 자료를 멀티모달 데이터로 구축하는 작업을 향후 꾸준히 기획하고 또 진행하지 않을까 싶습니다. 무엇보다 자료의 규모라든지 성격이라든지 여러 측면에서 군소 규모의 기관에서는 좀처럼 시도하기 어려운 작업을, 할 수 있고 또 해야 하는 그러한 위치에 있는 곳이 국립중앙도서관이기도 합니다. 관련해서 작년(2023) 11월에 열린 '국가지식정보협의회 발족 총회 및 기념 국제 컨퍼런스'에서 해외 전문가 선생님들을 모시고 IIIF(트리플 아이에프)에 관한 논의를 중심적으로 진행했던 것으로 알고 있습니다. IIIF 또한 향후 멀티모달 데이터셋을 구축해 나가기 위한 장기적 도모에 해당하지 않을까 싶습니다. 멀티모달 데이터 구축과 관련해서 국립중앙도서관이 어떠한 검토와 노력을 하고 있는지 궁금합니다. 관련 내용을 조금 소개해주실 수 있는지요?

○한국전통문화대학교에 계신 이종욱 선생님께 드리는 질문

-선생님의 발표문은 디지털 헤리티지에 관해서 굉장히 다종다양한 정보를 전달하고 있습니다. 디지털 문화유산에 관한 관심을 평소에 늘 품어 왔지만, 해당 분야의 전공자가 아니기에 그 현황이라든지 맥락에 관한 이해가 항상 부족하던 터에 선생님 발표문 덕분에 많은 공부가 되었습니다. 관련해서 발표문을 통해 전달해주신 그러한 여러 지점이 모두 근래 박물관에서 중요한 이슈로 다루어지고 있는 것인지 궁금합니다. 아마도 선생님께서는 디지털 헤리티지와 관련해서 여러 박물관에 자문도 꾸준히 해오셨을 테고, 또 협업 경험도 있으실 테니, '디지털 박물관'이라고 해야 할까요? 여튼 그러한 방향에 대한 현실적 요구를 실제 박물관들이 어떻게 수용하고 또 대응하고 있는지에 대해서도 알고 계실 듯하여, 여쭈어봅니다.

-발표문 뒷부분에서 향후 과제를 말씀하시면서 디지털 유산 보존 개념의 설정이 필요함을 강조하셨습니다. 데이터 생산의 규모라든지 디지털 기술의 변화 주기라든지 여러 맥락의 현실적 문제가 얽혀 있는 것으로 보입니다. 관련해서 선생님 말씀에 따르면 본(born) 디지털 유산 보존.관리.활용 제도는 아직 전세계적으로도 그 사례를 찾기가 어렵다고 하셨는데, 식견이 얕은 제가 보아도 이 부분은 굉장히 중요한 문제라는 생각이 듭니다. 디지털 환경 특히 웹에서 곧바로 생산-유통되는 문화유산 유관 정보들이 많고 또 그만큼 제대로 수집-관리되지 않기에 휘발되고 없어지는 정보 또한 그만큼 많으리라 생각합니다. 그러한 측면에서 그에 관한 문제의식을 품고 계신 선생님의 경우, 현실에서 연구나 교육을 통해 그러한 문제점을 개선하기 위한 고민이나 시도가 될만한 노력을 하고 계실 듯한데, 관련 내용이 있다면 조금 여쭈어 들을 수 있을지요?

-향후 과제로 말씀하신 내용과 관련해서 한 가지 더 말씀을 꺼내어 보자면, 추후 디지털 데이터의 접근성을 높이기 위해 인공지능 기술을 적용하는 시도가 필요하다고 하셨습니다. 문화유산을 대상으로 만들어진 디지털 데이터의 접근성이란 곧 해당 문화유산을 향유하는 것과 직접적으로 관련이 있지 않을까 싶습니다. 문화유산 향유의 가장 큰 주체는 사실상 대중과 연구자이기에, 그와 같이 디지털 헤리티지 데이터에 인공지능 기술을 적용함에 있어서 문화유산에 관심을 둔 대중이라든지 실제 그러한 자원을 대상으로 연구를 하고 있는 연구자와의 호흡이

굉장히 중요하겠다는 생각이 듭니다. 박물관에서는 이미 그에 관한 여러 작업을 진행하고 있을 듯싶은데, 실제로 박물관에서 진행되고 있는 그러한 인공지능 기술 활용에 관한 고민이나 시도들이 박물관에서 제공하는 서비스의 효율성이나 기능성을 향상시키는 데만 초점을 두지 않고, 소위 관람자로서의 대중과 향유자로서의 연구자 입장에서 실질적으로 필요한 지점들 또한 고려되고 있는지 그 연장선에서 대중이나 연구자들의 요구 사항이라고 해야 할까요? 그러한 지점들이 반영되고 있는지 궁금합니다.

KADH KOREAN ASSOCIATION FOR DIGITAL HUMANITIES