 https://doi.org/10.23287/KJDH.2024.1.2.6

디지털과 인문학의 가교:
디지털 인문학에 관한 크리스토프 소흐 교수와의 인터뷰 *

Bridging the Digital and the Humanities: An Interview with Professor Christof Schöch on the Future of Digital Humanities

Schöch, Christof **
0000-0002-4557-2753

목차

1. 인터뷰 번역본

크리스토프 소흐(Christof Schöch) 교수님께.

인터뷰에 응해 주셔서 감사합니다. 특히 트리어 디지털인문학센터(Trier Center for Digital Humanities; TCDH)의 공동 소장이자 디지털인문학조직연합(Alliance of Digital Humanities Organizations; ADHO)의 전임 회장으로서 디지털 인문학 분야를 형성하는 데 중요한 역할을 해 주신 것과 시간과 통찰을 공유해 주신 것에 진심 어린 감사의 말씀을 드립니다.

이번 인터뷰는 디지털 인문학에 대한 한국 연구자 및 기관들의 관심이 고조되는 시기에 이루어졌습니다. 교수님의 답변은 영어와 한국어로 제공되어 한국 DH 연구 공동체에 교수님과 교수님

*인터뷰 번역문과 원문을 순차로 기재하였으며, 인터뷰어 정보는 다음과 같다. 김병준, 한국학중앙연구원 한국학대학원 인문정보학, 조교수, bjkim@byungjunkim.com.
The translated text and original text are presented sequentially, and the interviewer's information is as follows: Byungjun Kim, Assistant Professor, Cultural Informatics, The Graduate School of Korean Studies, the Academy of Korean Studies, bjkim@byungjunkim.com.

**1st Author: University of Trier, Trier Center for Digital Humanities, schoech@uni-trier.de

의 연구를 소개하는 데 도움이 될 것입니다. 교수님의 관점은 인문학 연구에 대한 디지털 접근 방식의 발전 중인 환경을 헤쳐나가는 이곳 학자들을 안내하고 영감을 주는 데 매우 유용할 것이라고 믿습니다.

특히, 교수님의 전산 문학 연구(computational literary studies)에 관한 전문 지식과 문학사 연구를 위한 "스마트 빅데이터(smart big data)" 및 링크드 오픈 데이터(Linked Open Data; LOD)와 같은 개념에 대한 혁신적인 연구가 흥미롭습니다. 이번 인터뷰를 통해 독자들이 이러한 접근 방식과 그 잠재력이 인문학 연구에 어떤 변화를 가져올 수 있는지 깊이 이해하는 계기가 마련되었으면 합니다.

그럼, 이제 더 이상의 지체 없이, 질문을 시작하겠습니다.

1.1 디지털 인문학: 관점과 영향

(1) *수년 간 디지털 인문학의 최전선에 서 계셨습니다. DH를 인문학 연구를 혁신적으로 변화시킬 잠재력을 지닌 '게임 체인저'로 보시는지, 아니면 전통적 방법을 보완하는 도구로 보시는지 궁금합니다. 일부 학자들은 DH가 결국 전통적 인문학을 대체할 것을 우려하기도 하는데, 이러한 우려에 대해서는 어떻게 생각하실까요? 더 나아가 DH가 인문학 연구의 미래에 어떤 역할을 할 것이라고 예상하시나요?*

먼저, 제 연구에 대해 이야기할 기회를 주셔서 대단히 감사합니다. 저에게 대단히 의미가 깊습니다.

디지털 인문학은 분명 인문학자들이 사용할 수 있는 도구를 확장하고 문화와 역사에 대한 다양한 접근 방식을 추가했지만, 단순히 디지털 도구를 추가한 것에 그치지는 않습니다. DH는 정량적 접근 방식과 정성적 접근 방식 간의 역동적인 상호 작용을 제공하여 전통적인 방법론을 더욱 풍부하게 만듭니다. 디지털 인문학은 인문학을 대체하기보다는 반복적(iterative) 과정을 연구에 도입하여 이를 확장한다고 볼 수 있습니다. 예컨대, 우리는 가까이 읽기(close reading)를 통해 가설을 개발하고 체계화할 수 있으며, 더 큰 데이터셋을 컴퓨터로 분석함으로써 이러한 가설을 추가 조사하거나 검증하는 방법을 개발할 수 있습니다. 그런 다음 다시 텍스트나 아티팩트(artifacts)로 돌아가서 보다 세부적이고 정교한 질문을 제기하거나 예외적 사례와 독특한 구성을 탐구할 수 있습니다. 이는 가까이 읽기와 데이터 분석의 강점을 바탕으로 끊임없는 탐구와 재해석을 가능하게 하는 순환적 과정입니다.

사실, DH는 인문학 분과와 함께 발전하고 있으며, 이를 대체하는 것이 아니라 강화하는 역할을 하고 있습니다. 시간이 지남에 따라 DH의 하위 분야는 점차 전문화되고 있습니다. 예를 들어, 전산 문학 연구(Computational Literary Studies; CLS)나 디지털 역사학(Digital History) 같은 분야는 독립된 하위 분야로 발전하고 있으며, 전산 언어학(Computational Linguistics)이나 고고학에서의 컴퓨터 응용(Computer Applications in Archaeology) 같은 분야는 이미 독자적인 영역으로 자리 잡았습니다. DH의 미래는 다양한 방향으로 전개될 수 있습니다. 일부 하위 분야는 새로운 방법론, 이론 및 비판적 접근 방식을 개발하기 위한 독립적인 학제 간 허브로 이어질 수 있으며, 기존 인문학과 컴퓨터과학의 접점을 형성할 것입니다. 다른 하위 분야는 인문학 부분과 결합되어 더 이상 "디지털"과 "기존의 정립된" 방법론 간의 차이를 느끼지 못할 수도 있습니다.

제가 보기에, 모든 접근 방식이 공존할 여지가 충분히 있습니다. 지향해야 할 것은 전통적 인문학과 디지털 인문학이 서로 보완하며 협업할 수 있는 환경을 조성하는 것입니다. 궁극적으로 DH는 인문학 연구에서 가능한 범위를 넓히고, 새로운 통찰의 문을 열면서도, 인문학적 탐구가 항상 문화와 사회를 이해하는 데 가져온 고유한 관점을 존중하는 데 그 목적이 있습니다.

또한, 우리는 같은 자원을 놓고 경쟁을 벌이기보다는 경쟁을 늘리는 게 정말 중요다고 생각합니다. 대신, 서로 협력해 '파이'를 키울 수 있어야 합니다. 제가 아는 한, 이를 잘 보여주는 사례로 독일의 국가연구데이터인프라(National Research Data Infrastructure; NFDI) 프로그램에서 인문학의 역할을 들 수 있습니다. 인문학계에서 디지털 지향적인 학자들이 힘을 모아 공학, 의학, 자연과학 분야와 마찬가지로 인문학에서도 최첨단 인프라가 필요하다고 성공적으로 논증하였고, 이로부터 인문학 전체가 큰 혜택을 받게 될 것입니다.

(2) *교수님의 연구는 종종 전산적 방법론을 전통적인 문학 분석과 연결합니다. 연구에서 이 두 접근 방식의 균형을 어떻게 맞추시나요? 전산적 분석이 오래 지속된 문학적 해석을 재평가하도록 이끈 사례를 하나 공유해 주실 수 있을까요? 반대로, 문학 연구에 대한 배경지식이 전산적 접근 방법의 한계를 파악하는 데 도움이 되었던 적이 있었을까요? 향후 문학 연구에서 이러한 방법론들이 어떻게 이상적으로 통합될 수 있을지 상상하시는 바가 있을까요?*

제 연구에서 저는 두 가지 유형의 질문에 이끌립니다. 하나는 인문학에서 직접적으로 도출되는 근본적인 질문이고, 다른 하나는 이러한 질문을 효과적으로 해결하기 위해 정교화한 전산적 방법론이 필요한 질문입니다. 예를 들어, "문학적 하위 장르를 어떻게 정의할 것인가?"나 "문학적 진화는 시간이 흐르면서 어떻게 전개되는가?"와 같은 전통적인 인문학의 질문이 제 연구를 많은 부분에서 추동합니다. 동시에, 이러한 질문은 자주 방법론적 문제로 이어지기도 합니다. 예를 들면, 전산 문체학(stylometry)에서 저자가 누구인지를 정확히 파악할 수 있게 도울 수 있는 방법은 무엇인지, 혹은 서로 다른 언어와 장르에서 이러한 질문에 답하기 적합한 코퍼스(corpus)는 어떻게 설계해야 하는지와 같은 문제를 들 수 있습니다.

이러한 질문은 흔히 서로 융합되곤 합니다. "문학적 하위 장르"를 예시로 들어 보겠습니다. 이를 전통적인 문학 연구와 잘 공명하면서도 전산적 분석에 적합한 방식으로 개념화하는 것은 복잡할 수 있습니다. 저는 문학적 하위 장르를 문학 연구자들에게 의미 있게 정의하면서, 동시에 대규모의 다양한 코퍼스를 디지털 방식으로 분석할 수 있는 운영 가능한 개념으로 만드는 것을 목표로 하고 있습니다. 제 경험상, 이 과정의 핵심은 복잡한 개념을 더 작고 관리 가능한 구성 요소로 "분해"하고, 이를 식별하고 분석할 수 있는 전산적 방법을 개발한 뒤, 상관성과 공기 패턴을 연구함으로써 분해한 요소들을 다시 합치는 것입니다. 이런 점에서 알고리즘적 접근의 필요성과 "컴퓨팅 사고"(computational thinking)의 개념적 접근이 제 문학적 개념에 대한 사고 방식을 근본적으로 바꾸어 놓았다고 할 수 있습니다.

이러한 작업은 곧잘 이론적 질문을 제기하기도 합니다. 가령, 전산적 방법은 종종 "핵심적인 요소"(keyness)를 순전히 통계적 방식으로 정의하지만, 저는 문학적 맥락에서 "핵심적인 요소"의 의미를 검토하는 것이 필수적이라고 생각합니다.(제 연구팀에서는 이 핵심적인 요소를 "독특성(distinctiveness)"이라고 부르는 것을 선호합니다.) 마찬가지로, 전산적 연구에서 "문체(style)"는 빈도 패턴으로 축소되는 경향이 있지만, 이는 문학 연구에서 길고도 복잡한 역사를 가진 개념입니다. 특히 대규모 언어 모델(Large Language Model; LLM)을 통해 단순한 단어 형태 이상의 텍

스트적 특징에 더욱 세밀하게 접근할 수 있게 된 지금, 전산적 연구를 수행할 때 이러한 맥락을 잊지 않는 것이 중요합니다. 그리고 전산적 연구에서 핵심적인 개념인 재현성(reproducibility)은 전산 문학 연구(Computational Literary Studies; CLS)에서 새로운 차원을 가집니다. 원래 비전산적이거나, 적어도 디지털 외적인 패러다임에서 수행되었던 과거의 연구를 재현하려고 하는 경우를 예시로 들 수 있는데, 이는 자연과학 분야에서는 좀처럼 상상하기 힘든 이야기입니다.

1.2 연구와 방법론

(3) *교수님의 연구에서 "스마트 빅데이터"(smart big data)라는 개념을 소개하셨습니다. 이 개념의 의미와 장점에 대해 자세히 설명해 주실 수 있을까요? 기존의 빅데이터를 활용한 접근 방식과는 어떻게 다른가요? 또 교수님의 연구 프로젝트에서 스마트 빅데이터가 문학 연구상의 새로운 통찰을 이끌어 낸 특정 사례가 있다면 소개를 부탁드립니다. 이 접근 방식을 구현하는 과정에서 어떤 어려움을 겪으셨는지도 궁금합니다.*

저에게 "Smarter bigger data"란 스마트 데이터의 정밀성과 방대한 빅데이터의 규모 사이에서 절충점, 그보다는 "제3의 길"을 찾고자 하는 것입니다. 기존의 "스마트 데이터"는 학술적인 디지털 편집 영역의 역사 비평 또는 유전적 편집(genetic edition)처럼 신중하게 선별된, 결함이 없는 데이터를 뜻합니다. 이렇게 고도로 연마된 데이터는 대단한 가치가 있지만, 모든 세부 사항에 대한 꼼꼼한 주의를 요하기에 제작 과정에서 상당한 시간이 소요될 수 있습니다. 반면에 빅데이터는 규모는 방대하지만 정리되지 않은 경우가 많습니다. 방대한 데이터(문학 데이터 포함)를 기반으로 학습된 대규모 언어 모델(LLM)과 같은 도구가 필요할 수도 있지만, 우리가 사용하는 데이터셋은 보통 그 정도의 규모일 수도 없고, 그럴 필요도 없는 경우가 많습니다. 인문학 데이터셋은 그 크기보다는 복잡성 때문에 어려움을 겪는 경우가 더 많기 때문입니다. 따라서 단순히 빅데이터 그 자체를 목표로 삼기보다는, 연구 질문에 따라 방법을 맞출 필요가 있습니다.

근본적으로 "Smarter bigger data"라는 개념은 스마트 데이터의 품질을 증진시키면서도 연구 목표를 놓치지 않는 것을 목표로 합니다. 알고리즘, 기계 학습(machine learning), 정보 추출을 결합해서 주석이 잘 달린, 중간 규모의 데이터셋을 만들면, 이러한 데이터셋을 통해 복잡하고 의미 있는 질문을 던질 수 있게 됩니다. 예를 들어, 우리는 이제 대규모 언어 모델(LLM)을 사용하여 텍스트의 초기 주석을 생성한 뒤, 이를 큐레이션하고 지식 그래프 안에 구조화할 수 있습니다. 이는 단순히 LLM을 활용해 질의(query)하는 것과는 다릅니다. 이는 데이터를 검사 및 정제하고 점진적으로 주석을 조정하는 명료하고 반복적인 과정입니다. 그 결과, 분석을 위한 깊이와 넓이를 모두 갖춘 풍부하고 구조화된 지식 기반이 만들어집니다.

이러한 접근 방법의 한 사례로 MiMoText 데이터베이스를 들 수 있습니다. 이는 Mining and Modeling Text 프로젝트에서 기계 학습과 신중하게 구조화된 지식 기반 설계를 활용해 개발한 데이터베이스입니다. MiMoText를 활용하면, 전통적인 정전(canon) 범주를 넘어서는 훨씬 광범위하고 포괄적인 데이터셋을 바탕으로 문학적 특징을 상호 연관시키는 작업을 수행할 수 있습니다. 이러한 접근 방법은 완전히 새로운 가능성을 열어 줍니다. 예를 들어, 소설을 매핑하거나 클러스터링하여, 개별 텍스트만으로는 볼 수 없었던 패턴과 연결성을 감지할 수 있습니다. 여기서 시각화가 중요한 역할을 하는데, 시각화는 데이터에 나타난 유사성이나 연결성을 볼 수 있게 해 줌으

로써 데이터베이스를 효과적으로 문학적 통찰을 위한 탐색 도구로 전환해 줍니다.

(4) *교수님의 최근 연구는 데이터 기반 접근 방식으로, 특히 링크드 오픈 데이터(Linked Open Data; LOD)를 활용하여 문학사를 탐구하는 데 중점을 두고 있습니다. 이 방법론이 문학사에 대한 우리의 이해를 어떻게 변화시킬 수 있다고 보시나요? 우리가 어떤 새로운 질문을 제기할 수 있고, 어떤 전통적인 가정들에 도전하게 될 수 있을까요? 링크드 오픈 데이터(LOD)가 특히 문학사 연구에 적합한 이유가 무엇인지 설명해 주실 수 있을까요? 이러한 접근 방법이 제공하는 흥미로운 가능성과 잠재적인 한계로 무엇을 이야기해 볼 수 있을까요?*

링크드 오픈 데이터(LOD)를 활용해 문학사를 탐구하는 것은 새로운 시각으로 문학사의 영역을 조명하게 해 주고, 문학이 시간이 경과하면서 어떻게 발전했는지에 대한 새롭고 세밀한 질문을 제기할 수 있게 해 주는 혁신적인 접근 방법입니다. LOD는 데이터를 다루는 유연하고 명료한 구조를 제공하며, 이는 "Smarter bigger data"의 실행 사례로 볼 수 있으며, 이를 통해 정성적, 정량적인 다양한 출처의 정보를 통합하고, 그 정보를 여러 방식으로 분석할 수 있습니다.

LOD의 핵심적인 장점 중 하나는 그 다재다능함입니다. 우리는 정성적 통찰이나 정량적 지표에서 출발하여 데이터셋을 구축할 수 있으며, LOD의 구조를 활용해 이러한 데이터셋을 양쪽 관점에서 분석할 수 있습니다. 개별 항목을 탐색하고 검토하고 싶을 때는 그렇게 할 수 있고, 데이터셋 전체에 걸쳐 나타나는 추세나 패턴을 특정하기 위해 더 큰 규모의 쿼리를 실행할 수도 있습니다. 최근에는 Wikibase.cloud와 같은 플랫폼 덕분에 LOD 접근성이 훨씬 높아졌고, 연구자들은 더욱 풍부하고 연결된 데이터를 손쉽게 생성하고 관리할 수 있게 되었습니다.

LOD는 다국어 연구에도 흥미로운 가능성을 열어 줍니다. LOD 시스템의 각 개념에는 여러 언어로 라벨(label)을 달 수 있어 교차 언어적 탐구와 분석이 가능합니다. 또한 LOD는 연합적(federated) 방식으로 작동하기 때문에, 위키데이터(Wikidata)나 도서관 카탈로그(library catalog)와 같은 다른 LOD 자원에 연결함으로써 추가적인 정보로 연구를 향상시킬 수 있습니다. 반대로, 우리의 연구는 다른 연구자들도 재활용할 수 있도록 공개되어 협업적인 세계적 지식 네트워크를 구축하게 됩니다.

문학사 연구에서 LOD는 역사적 데이터에 대한 "원자화된(atomized)" 접근 방식이라는 특히 강력한 어떤 것을 제공합니다. 거대 서사를 사전에 정의하고 그에 맞는 증거를 수집하는 대신, LOD는 문학사를 사건, 인물, 출판물, 주제 요소 등 무수히 작은 구성 요소로 나누고 이러한 요소들이 시간에 따라 어떻게 상호 연결되는지 분석할 수 있게 해 줍니다. 이러한 접근 방법을 통해 우리는 전통적인 하향식(top-down) 서사에 이의를 제기하는 상관관계, 공기성(co-occurrence), 패턴을 관찰할 수 있습니다. 이는 고립된 "위대한 작품"만 살피는 것이 아니라 문학적 지형의, 종래에 숨겨져 있던 역학을 드러내는 보다 넓은 연결망을 살필 수 있게 합니다.

또한 위키베이스(Wikibase)와 같은 플랫폼은 역사적 주장에 맥락과 명료성을 더할 수 있는 프레임워크를 제공합니다. 이러한 데이터베이스의 각 항목에는 특정 진술에 대한 출처, 확실성 수준 및 시간 범위 등의 세부 정보가 포함될 수 있습니다. 이와 같은 명료성 층위는 지식 생산 과정을 명확히 문서화하는 데 대단히 중요한 "출처"(provenance)의 개념과 일치합니다. 우리는 단순히 "사실"을 제시하는 것이 아니라, 각 주장의 근거를 보여줌으로써 다른 사람들이 그 주장 뒤에 있는 증거와 맥락을 이해할 수 있도록 합니다.

물론, 도전 과제도 존재합니다. 문학사 연구를 위한 LOD 활용에서 가장 큰 도전 과제 중 하나

는 데이터 모델링과 문학 연구의 뉘앙스를 효과적으로 포착할 수 있는 공유 온톨로지의 사용 또는 생성입니다. 다양한 언어, 장르, 역사적 맥락을 수용할 만큼 유연하면서도 의미적으로 명확한 모델을 만드는 것은 학제 간 협업을 필요로 합니다. 또한 연합(federation), 즉 여러 데이터셋의 매끄러운 통합은 또 다른 자체적 난제를 제시합니다. 진정한 상호 운용성(interoperability)을 위해서는 연결된 데이터셋이 호환 가능한 표준과 온톨로지를 따라야 하며, 이는 기관 간, 그리고 기술적 인프라 간의 일치를 요합니다.

이러한 도전 과제는 저에게 텍스트 인코딩 이니셔티브(Text Encoding Initiative; TEI)를 떠올리게 합니다. TEI는 매우 광범위하고 유연하지만, 이로 인해 진정한 상호 운용성을 달성하는 데 어려움을 겪기도 합니다. TEI에서 Lex-0과 같은 하위 집합이 사전편찬 데이터 (lexicographical data)에 대해 표준화되고 제한된 옵션을 제공하듯이, LOD를 위한 온톨로지 또한 널리 공유되는 핵심 영역이 필요하다고 생각합니다. 이를 위해 더 확장된 더블린 코어(Dublin Core)와 같은 접근법이 유용할 수 있습니다. 이러한 접근은 자원을 더 폭넓게 접근 가능하고 상호 연결되게 만드는 동시에, 도메인별 특수성을 유지하게 도울 수도 있습니다. 연합을 실현하려면 넓은 연결성에 대한 열망과 특수성 및 학문적 엄격성에 대한 필요 사이에서 균형을 맞추어야 하며, 우리가 생성한 연결이 소음을 더하거나 오해를 초래하지 않는 대신, 연구를 향상시킬 수 있도록 보장해야 합니다.

(5) *재현성(reproducibility)의 개념은 최근 교수님의 연구, 특히 전산 문학 연구(CLS)에서 중요한 초점이 되고 있습니다. 재현성이 이 분야에서 왜 중요하다고 생각하시나요? 다른 분과와 비교했을 때, 인문학 연구에서 재현 가능성을 보장하려면 어떤 독특한 어려움이 수반될까요? DH에서 재현 가능한 연구 문화를 어떻게 조성할 수 있고, 재현성이 DH 분야의 발전과 신뢰성에 어떤 영향을 미칠 수 있을까요?*

재현성은 연구의 투명성과 개방성을 보장하는 핵심 요소입니다. 전산 문학 연구(CLS)에서는 단순히 다른 사람도 구축할 수 있는 데이터셋을 공유하는 것과 같은 작업이 이 분야를 발전시키는 데 필수적이라는 것을 이미 확인하고 있습니다. 연구 결과를 진정으로 이해하고 신뢰하려면, 데이터가 어떻게 구성되었는지, 그리고 특정한 방법이 데이터를 통해 어떻게 결과를 도출하였는지 그 과정을 명확히 알아야 합니다. 재현성이 없다면, 이러한 투명성이 무너지고 연구자들이 서로의 연구를 기반으로 삼거나 연구 결과를 검증하기가 어려워집니다.

전산 연구 분야에서는 데이터와 코드에 공개적으로 접근할 수 있게 하는 것이 재현성의 출발점이 됩니다. 그러나 문학 연구에서 “데이터”의 대부분은 늘 기계 가독형이 아닌 책으로 이루어져 있기에 독특한 어려움에 직면하게 됩니다. 책은 분명 데이터이지만, 디지털화되어 있지 않거나 디지털화되어 있더라도 저작권 문제로 접근이 제한되는 경우가 많습니다. 이는 접근성에 대한 중요한 의문을 제기합니다. 재현성을 높이려면, 더 많은 공개된 디지털 텍스트 판본을 구축할 필요가 있으며, 이를 위해 단기적으로는 공정 이용(fair use)과 같은 저작권 예외 조항을 활용하고, 장기적으로는 저작권법의 변화를 모색하는 등 저작권 문제에 대한 창의적인 해법이 필요합니다. 현재 저작권 보호 기간이 작가 사후 70년이라는 점은 제게는 시대에 뒤떨어진 제도로 느껴지며, 문학사라는 방대한 기록물에 대한 접근을 제한함으로써 연구를 저해하고 있습니다.

재현성을 향상시키기 위한 집단적 합의를 구축하고 이를 위한 전략을 개발하는 것 외에, 연구를 재현 가능하게 만드려면 연구자가 상당한 시간과 노력을 들여야 한다는 점을 또 다른 과제로

들 수 있습니다. 재현성에 대해 심사자가 관심을 갖게 만들거나 재현성을 출판 과정의 일부로 추가하는 것 역시 시간이 걸리는 작업입니다. 제가 에펠린 기우스(Evelyn Gius) 교수, 피어 트릴케(Peer Trilcke) 교수와 같이 공동으로 편집하고 있는 JCLS(Journal for Computational Literary Studies)에서는 데이터와 코드를 보관하는 것을 요구하며, 데이터와 코드에 대한 선택적인, 확장된 리뷰를 위한 가이드라인도 제공하고 있습니다. 모든 저자가 그러한 과정을 택하지는 않는데, 모든 저자가 그 과정을 거친다면 편집자와 심사자에게 막대한 부담이 될 것입니다. 동시에 재현성은 공동체 차원의 노력을 요구하고, 학술지가 그러한 공동체를 형성하는 결정적인 한 지점으로서 시작하기에 좋은 장소라고 생각합니다. 현재까지 이러한 시도에 대한 경험은 무척 긍정적이었습니다.

1.3 DH 공동체와 교육

(6) *TCDH의 공동소장이자 ADHO의 전임 회장으로서, 디지털 인문학 분야를 관찰하면서 동시에 형성하는 독특한 위치에 서 계셨습니다. 이러한 역할에서 가장 기억에 남는 경험이나 도전 과제는 무엇이었나요? 이러한 경험이 연구와 리더십 접근 방식에 어떤 영향을 미쳤나요? 오늘날 DH 커뮤니티가 직면한 가장 시급한 문제는 무엇이라고 보시나요?*

정말 광범위한 질문이네요. ADHO만 생각하더라도 그렇습니다. 솔직히 지금 되돌아보면, 우리가 무언가 함께 성취했던 순간들, 공동체에 긍정적인 영향을 미쳤다고 모두가 느낀 순간들이 특히 기억에 남습니다. 가령 ADHO의 행동 강령이나 학술대회 제출물에 대한 새로운 심사 기준 도입과 같은 성과들이 떠오릅니다. 물론 연례 학술대회를 경험하며, 다른 많은 사람들과 함께 학술대회 준비에 기여했다는 것도 생각납니다. 그런데 도전 과제는 더 일상적인 부분에 가까운 경우가 많았습니다. 멕시코시티와 버클리에서 시작해 몬트리올, 미국 동부 해안을 거쳐 파리, 로마, 서울, 도쿄, 캔버라에 이르기까지 전 세계로 흩어진 이사회 구성원들과 서로 다른 시간대 속에서 생산적이고 즐거운 온라인 회의를 진행한 것이 그 중 하나입니다. 또 사실, 결정이 제 뜻대로 되지 않아 낙담했던 경우도 있었습니다.

저 스스로는 비교적 합의 지향적인 리더십 스타일을 가지고 있었다고 생각합니다. 물론 애초에 그럴 수 없었던 경우가 많지만, 제가 할 수 있다는 이유만으로 사안을 밀고 나가는 것을 좋아하지 않습니다. 구체적으로 말하자면, 가능한 한 의견을 폭넓게 수렴하고, 이를 바탕으로 결론을 도출한 다음, 모임에 그 결론을 제안하는 방식을 선호합니다. 이러한 결론을 어떻게 도출하느냐는 개인의 가치관과 특정 순간에 추구하는 목표나 야망에 달려 있습니다. 따라서 그에 관한 공통점을 찾는 것도 중요합니다. 이러한 접근 방식은 합리적으로 들릴 수 있지만, 지난 몇 년간 현실적으로 모든 의견을 수렴하기는 어렵다는 점, 그리고 결국 회의 준비를 맡은 개인이나 모임이 결과와 방향을 상당 수준 좌우한다는 점을 알게 되었습니다. 이제 저는 이러한 사실을 받아들이게 되었고, 어쩌면 가끔 이 과정을 즐기고도 있습니다.

DH 커뮤니티의 도전 과제 중 상당수는 사실 최근 DH의 놀라운 성공과 성장의 결과로 발생한 것이라고 봅니다. 성장하는 와중에 어떻게 하면 안정성, 개방성, 그리고 일관성을 유지할 수 있을까요? 거대한 일종의 원심력이 작용해서 점점 더 많은, 독립된 하위 분야가 생겨나고, 그 증거로 전문 학술대회와 출판 경로가 늘어나는 상황에서, 우리가 공동체로서 정체성을 어떻게 유지할 수

있을까요? 저는 이러한 발전을 환영하며 어느 정도는 이러한 움직임을 주도하기도 하지만, 이는 글렌 레인-워시(Glen Layne-Worthey) 교수의 비유처럼 DH가 점점 더 큰 텐트가 되고 있다는 의미이기도 합니다. 마찬가지로 디지털 인문학 공동체의 급증하는 세계적 영향력과 그 범위는 매우 긍정적인 발전이며, ADHO는 이러한 지역 공동체의 다양성을 모두 수용하기 위해 설계된 구조를 가지고 있습니다. 하지만 이는 생산적으로 협업하기 위해서는 서로의 차이를 그 어느 때보다 신중하게 고려해야 한다는 뜻이기도 합니다.

(7) *한국에서는 디지털 인문학에 대한 관심이 급증하고 있는데, 이는 인문학 전공 졸업자가 코딩 기술로 취업 전망을 높일 수 있다는 인식에 의해 부분적으로 영향을 받은 것입니다. 이러한 추세에 대해 어떻게 생각하시나요? 기술적 역량이 중요하긴 하지만, DH 연구자가 갖추어야 할 다른 중요한 역량은 무엇이라고 생각하시나요? 디지털화의 급류 속에서 인문학의 핵심 가치와 비판적 사고가 사라지지 않도록 하려면 우리는 어떻게 해야 할까요?*

분명히 프로그래밍, 데이터베이스 설계 경험, 데이터 분석 및 시각화 기술 등 디지털 인문학에서 배우는 기술은 취업 시장에서 자산이 될 수 있습니다. 코딩이나 기술적 측면이 진로 확대를 고려하는 학생들의 관심을 끄는 경우가 많다는 것도 사실입니다. 그러나 DH 연구자는 컴퓨터과학을 전공하는 학생도 아니고, 단순히 기술을 제공하는 사람도 아닙니다. DH 연구자가 제공하는 가치는 코딩을 넘어, 인문학과 디지털 기술의 세계를 연결하는 독특한 능력에 있으며, 이는 학계와 실무 양쪽에서 점점 더 중요해지고 있습니다.

기술적 노하우 외에도, DH 연구자는 복잡성을 다루고 빠르게 배우며 새로운 도구에 적응하는 데 능숙합니다. DH 연구자는 예를 들어, 데이터를 적합한 형식으로 변환하여 다양한 도구를 연결하는 데 능한데, 이는 상당히 강력한 역량입니다. 더욱 중요한 것은, DH 연구자가 컴퓨터과학의 기술적 언어와 인문학의 해석적 언어를 잇는 가교 역할을 수행할 수 있다는 점입니다. DH 연구자는 구현하기 쉽거나 어려운 것이 무엇인지 이해할 수 있을 뿐만 아니라, 기술 회사, 박물관, 기록보관소 등에서 사용자에게 중요한 것이 무엇인지도 인식할 수 있습니다. DH 연구자는 보통 커뮤니케이터나 중재자로 활동하며, 디지털 솔루션이 실제 사람들의 요구와 경험에 부합하도록 하는 역할을 하며, 이는 제품 관리자(PM), 교육 자원 개발자, 출판 관리자, 큐레이터 및 아키비스트 모두에게 매우 중요한 역할을 합니다.

그리고 기술적 역량이 중요하긴 하지만, 많은 게 반드시 더 좋은 것은 아닙니다. 도구와 알고리즘이 어떻게 구축되는지에 대한 이해와 비판적 사고가 더 핵심적인 것입니다. 제 수업에서는 선형 회귀든, K-평균 클러스터링이든, 나이브 베이즈 분류기든, 아니면 단순 신경망이든, 학생들이 단순하고 기초적인 형태의 알고리즘을 직접 만들어 보는 것이 중요하다고 강조합니다. 이러한 알고리즘을 한 단계씩 직접 구현해 봄으로써, 학생들은 알고리즘이 실제로 어떻게 작동하는지 더 깊이 이해하게 되며, 이를 바탕으로 보다 정교한 라이브러리를 활용해 보다 효율적으로 알고리즘을 구현할 수 있게 됩니다. 이렇게 기초를 이해하면, 기술이 오작동할 때 이를 진단하거나 "블랙박스" 모델에 내재된 편향을 식별하기 위한 준비를 갖춘 셈입니다. 마법처럼 보이던 신경망이 사실은 단순히 정교하게 배열된 행렬 연산이라는 것을 알게 되면, 오히려 현실감 있게 느껴지기도 합니다.

DH 연구자는 새로운 기술을 수용하고 이해하는 동시에, 인문학의 반성적이고 비판적인 가치를 잃지 않도록 이 관점을 발전시킬 수 있는 독특한 위치에 서 있습니다.

1.4 미래의 방향성

(8) *DH2026 학술대회가 대전에서 개최될 예정입니다. 이 행사에 대해 어떤 기대를 갖고 계신가요? 특히 아시아에서 전개되는 DH의 미래 방향성에 어떻게 반영되고 어떤 영향을 미칠 것으로 보시나요? 앞으로 교수님의 연구 계획과 우선 순위는 어떻게 되나요? 특히 재밌다고 생각하는 최신 DH 트렌드나 기술이 있을까요?*

우선 제 연구 우선순위부터 말씀드리겠습니다. 이미 여러 차례 언급한 부분이기도 하고, 이를 먼저 정리하는 게 좋을 것 같기 때문입니다. 우선, 저는 현재 대규모 언어 모델(LLM)과 지식 그래프/링크드 오픈 데이터(LOD)를 연결하는 연구를 계속 이어 나가는 데 관심이 있습니다. 동시에 DH의 다양한 영역에서 LOD의 활용도를 높이는 데 기여하고자 합니다. LOD를 심도 있게 다루다 보면, 여러 자원 간 연 계(federation) 문제와 어휘 및 데이터 모델의 상호 운용성 문제에 자연스럽게 직면하게 됩니다. 세 번째 우선 순위는 다국어 지원 능력(multilingualism)입니다. 저는 전산문체론 기반의 저자 식별이나 키워드 분석과 같은 영역에서 다언어 간에 효과적으로 작동하는 방법들을 계속 개발하고 평가할 계획입니다. 다국어 지원 능력은 DH가 전 세계적으로 성장함에 따라 매우 중요한 요소로 대두되고 있으며, 유럽의 DH에서 핵심적인 도전 과제이자 강점 중 하나이기도 합니다. 이는 제가 다양한 지역의 협업자들과 함께 이 주제를 탐구하기를 기대하는 이유이기도 합니다.

이러한 포부는 물론 더 넓은 추세와도 맞닿아 있습니다. DH는 데이터 공유의 강화, 의미론적 정교함, 그리고 언어 간 지원으로 나아가고 있다고 생각합니다. 이는 부분적으로 기계 학습(machine learning)의 발전 덕분일 수 있지만, 동시에 영어 중심의 자원 및 도구의 우세를 극복하고 그 너머로 나아가려는 의식적인 노력 덕분이기도 합니다. DH2026은 특히 아시아에서 이러한 발전의 중요한 이정표가 될 수 있을 것이며, 그 과정에 함께할 수 있기를 학수고대하고 있습니다.

2026년에 대전에서 개최될 디지털 인문학 학술대회(DH2026)와 거기서 파생될 다양한 기회에 대한 기대가 무척 큽니다. 이번 학술대회가 한국에서 개최되는 것은 매우 의미 있는 이정표입니다. 이는 아시아의 역동적이고 독창적인 디지털 인문학 공동체에 주목하게 할 뿐만 아니라, 북미와 유럽이라는 전통적 중심을 넘어 DH의 발전을 강화하는 계기가 될 것입니다. 최근 시드니와 멕시코 시티에서 학회가 열렸고, 2022년에는 도쿄가 온라인 DH 학회의 개최지였던 만큼, 이는 DH의 전 세계적 확장을 지속하는 중요한 흐름이라고 할 수 있습니다. 아시아, 특히 동남아시아와 동아시아는 DH가 역동적으로 성장하는 지역입니다. 지난 몇 년간 ADHO의 구성 조직이 확대되었듯이, 이제 학술대회 역시 이러한 변화를 반영할 때가 되었습니다.

2024년 봄에 한국을 방문했던 경험을 토대로, 저는 대전이 높은 수준의 전문성과 환대의 분위기를 모두 기대할 수 있는 도시라고 말할 수 있습니다. 한국은 기술 중심의 사회이면서도 동시에 문화적, 역사적 유산을 소중히 여기는 독특한 환경을 가지고 있으므로, 이런 점에서 DH2026은 무척 기억에 남는 이벤트가 될 것이라고 확신합니다. 개인적으로 저는 이번 학술대회에서 친구들과 재회하고 새로운 동료를 만날 생각에, 물론 예상치 못한 통찰과 협업 기회에 열린 마음으로, 들떠 있습니다. 이 학술대회는 지역 내 연구자들이 대륙 간 이동이라는 추가적인 어려움 없이 서로 연결될 수 있는 기회를 제공하여, 아시아 지역의 네트워크를 강화할 수도 있습니다. 일본, 대

만, 한국의 DH 공동체뿐만 아니라 인도네시아, 홍콩, 그리고 중국 본토와 같은 지역에서 새롭게 떠오르고 있는 공동체들 간의 더 많은 네트워킹을 기대합니다. 어쨌든 2026년 대전에서 한국 디지털 인문학 공동체를 만날 수 있기를 무척이나 고대하고 있습니다!

1.5 추가적인 읽을거리 제안

위에서 언급된 주제들에 대해 좀 더 깊이 살펴보고 싶은 독자들을 위해, 크리스토프 소흐와 그 동료의 명의로 발표된 몇 가지 논문을 참고용으로 여기 소개합니다.

1. On the theoretical investgations, one may consult "Revisiting Style, a Key Concept in Literary Studies" (2015), "From Keyness to Distinctiveness" (2021) or, on reproducibility, "Repetitive Research" (2023).
2. The idea of smart data was developed in "Big? Smart? Clean? Messy? Data in the Humanities" (2013) and combined with the Linked Open Data paradigm in "Smart Modeling for Literary History" (2022). Readers may wish to consult the MiMoTextBase as well.
3. Regarding methods, a summary on subgenre analysis is provided in "Computational Genre Analysis (2022), distance measures for stylometry are evaluated in "Understanding and Explaining Delta Measures for Authorship Attribution" (2017) and an investigation into multilingual authorship attribution is described in "Multilingual Stylometry" (2024), accompanied by an interactive showcase.
4. Generally speaking, recent, international work in Computational Literary Studies is published regularly in the *Journal of Computational Literary Studies*.

2. Interview in English

Dear Professor Christof Schöch,

Thank you for agreeing to this interview. We greatly appreciate your time and insights, especially given your prominent role in shaping the field of Digital Humanities as co-director of the Trier Center for Digital Humanities and former chair of the Alliance of Digital Humanities Organizations.

This interview comes at a time of growing interest in Digital Humanities among Korean researchers and institutions. Your responses will be made available in both English and Korean, serving to introduce you and your work to the Korean DH research community. We believe your perspectives will be invaluable in guiding and inspiring scholars here as they navigate the evolving landscape of digital approaches to humanities research.

Your expertise in computational literary studies and your innovative work on concepts like "smart big data" and Linked Open Data for literary history are of particular interest. We hope this interview will provide our readers with a deeper understanding of these approaches and their potential to transform humanities scholarship.

Without further ado, let's begin with our questions:

2.1 Digital Humanities: Perspectives and Impact

(1) *You've been at the forefront of Digital Humanities for many years. Do you see DH as a potential game-changer that will revolutionize humanities research, or more as a complementary tool to traditional methods? Some scholars worry that DH might eventually replace traditional humanities. How do you respond to these concerns? What role do you envision for DH in the future of humanities scholarship?*

First of all, many thanks for the opportunity to speak about my work here. It means a great deal to me.

Digital Humanities has definitely expanded the toolset available to humanities scholars, and added to the variety of approaches to culture and history, but it’s more than just an addition of digital tools. DH offers a dynamic interplay between quantitative and qualitative approaches that actually enriches traditional methods. Rather than replacing the humanities, it brings an iterative process to research – where we can develop and formulate hypotheses through close reading, develop ways of further investigating or even testing these hypotheses by looking at larger datasets with computational methods, and then return to the texts or artifacts for deeper, refined questions and for investigating the edge cases and unique configuratiosn. It’s a cycle that allows for constant exploration and reinterpretation, drawing on the strengths of both close reading and data analysis.

In fact, DH is evolving alongside humanities disciplines, not as a replacement but as an enhancement. Over time, we’re seeing DH subfields becoming more specialized, like Computational Literary Studies or Digital History, while disciplines such as Computational Linguistics or Computer Applications in Archaeology have long become fields of their own. The future of DH could take multiple directions: some subfields might continue as a standalone, interdisciplinary hubs for developing new methods, theories, and critical approaches, interfacing with both their humanities counterpart and Computer Science. Other subfields may become so integrated with their counterpart in the Humanities that we’ll barely notice the distinction between "digital" and "established" methods.

In my view, there’s plenty of room for all approaches. The aim should be to foster a collaborative environment where traditional and digital humanities complement each other, rather than competing. In the end, DH is about broadening the range of what’s possible in humanities research, opening doors to new insights while still valuing the unique perspectives that humanistic inquiry has always brought to understanding culture and society.

Similarly, I think it is really important that we not just compete for pieces of the same cake, creating increased competition. Instead, we should work together to increase the size of the cake. A successful example of this, as

far as I can see, is the role of the Humanities in the German National Research Data Infrastructure programme (NFDI). The digitally-oriented actors across the Humanities have successfully joined forces to argue for the need to take into account the Humanities' requirements for top-notch infrastructure, alongside those of fields like engineering, medicine, or the hard sciences. And the Humanities as a whole will benefit greatly from this down the road.

(2) *Your work often bridges computational methods with traditional literary analysis. How do you balance these two approaches in your research? Can you share an instance where computational analysis led you to reassess a long-held literary interpretation? Conversely, has your background in literary studies ever helped you identify limitations in computational approaches? How do you envision the ideal integration of these methodologies in future literary scholarship?*

In my research, I'm drawn to two types of questions: foundational questions that come directly from the humanities and those that require refining computational methods to address these questions effectively. For example, traditional humanities questions like "What defines a literary subgenre?" or "How does literary evolution unfold over time?" drive much of my work. At the same time, these often lead to rather more methodological questions, such as what methods help us accurately attribute authorship in stylometry or how we need to design corpora suited to answering these questions across different languages and genres.

Often, these questions blend together. Take "literary subgenre," for instance. Conceptualizing this in a way that resonates with established literary studies yet is suitable for computational analysis can be complex. I aim to define and operationalize literary subgenres so they're meaningful for literary scholars while also being analyzable across large, diverse corpora using digital methods. In my experience, the key here is to "deconstruct" complex concepts into their smaller, more manageable constituent parts; develop computational methods to identify and analysze these; and then put them together again through a study of the patterns of their correlations and co-occurrences. So in this case, the requirements of algorithmic approaches, but also the conceptual approach of "computational thinking", has fundamentally changed the way I think about literary concepts.

This work often also brings up theoretical questions. For example, computational methods often define "keyness" in a purely statistical way, but I find it essential to examine what "keyness" (or "distinctiveness", the term we have come to prefer in my team) means in a literary context. Similarly, while "style" is often reduced to frequency patterns in computational studies, it's a concept with a long, nuanced history in literary studies that we should not forget when working computationally, especially now that LLMs are giving us more nuanced access to textual features other than word forms. And reproducibility, a core idea in computational research, takes on new dimensions in Computational Literary Studies, for example when we try to perform reproductions of much earlier research that was originally performed in the non-computational, or at least non-digital, paradigm; a scenario that is rarely envisioned in the hard sciences.

2.2 Research and Methodologies

(3) *In your work, you've introduced the concept of "smart big data". Could you elaborate on what this means and its advantages? How does it differ from traditional big data approaches? Can you share a specific example from your research projects where smart big data has led to new insights in literary studies? What challenges did you face in implementing this approach?*

For me, "smarter bigger data" is a middle ground – or rather a "third way" – between the precision of smart data and the sheer scale of big data. Traditional "smart data" is carefully curated and flawless, like a historical-

critical or genetic edition in the domain of scholarly digital editing. While invaluable, creating such highly polished data can slow us down significantly, as every detail requires meticulous attention. On the other hand, big data is often vast but messy, and while we may need tools such as LLMs that are trained on huge amounts of data (including literary data), our own datasets rarely can or need to be at that same scale. Humanities datasets are often challenging due to their complexity, rather than their size. So, rather than pursuing big data for its own sake, we should be driven by our research questions and tailor our methods to those.

Essentially, the idea of "smarter bigger data" is to scale up the quality of smart data without losing sight of the research goals. It combines algorithms, machine learning, and information extraction to create well-annotated, mid-sized datasets that allow us to ask complex, meaningful questions. For example, we can now use large language models (LLMs) to generate initial annotations of text, which we then curate and organize within a knowledge graph. This isn't the same as simply querying an LLM; it's a transparent, iterative process where we can inspect and refine the data and adapt our annotations over time. The result is a rich, structured knowledge base that offers both depth and breadth for analysis.

One example of this approach is our MiMoText database, which we developed using machine learning and a carefully structured knowledge base design, in the Mining and Modeling Text project. MiMoText enables us to correlate literary features in a dataset that goes beyond the traditional canon, providing a much broader, more inclusive dataset. This approach opens up entirely new possibilities: we can, for instance, map or cluster novels, allowing us to detect patterns and connections that wouldn't be visible in individual texts alone. Visualization plays a key role here, as it helps us to see similarities or connections that emerge from the data, effectively turning the database into a kind of exploratory tool for literary insights.

(4) *Your recent work focuses on data-driven approaches to literary history, particularly using Linked Open Data. How do you see this methodology transforming our understanding of literary history? What new questions can we ask, and what traditional assumptions might we challenge? Could you explain why Linked Open Data is particularly suited for literary historical research? What are some of the exciting possibilities and potential pitfalls of this approach?*

Using Linked Open Data (LOD) to explore literary history is a transformative approach, one that allows us to see the field with fresh eyes and to pose new, nuanced questions about how literature evolves over time. LOD offers a flexible, transparent structure for handling data – it's like "smarter bigger data" in action – where we can integrate information from a wide range of sources, both qualitative and quantitative, and analyze it in multiple ways.

One of the key advantages of LOD is its versatility. We can build datasets that start from either qualitative insights or quantitative metrics, and the structure of LOD enables us to analyze these datasets from both angles. If we want to browse and examine individual items, we can do that. If we want to run larger queries to spot trends or patterns across the entire dataset, that's possible too. With platforms like Wikibase.cloud, LOD has recently become much more accessible, allowing researchers to create and manage rich, connected data more easily.

LOD also opens exciting possibilities for multilingual research. Each concept in an LOD system can have labels in multiple languages, enabling cross-linguistic exploration and analysis. And because LOD works in a federated way, we can link to other LOD resources (such as Wikidata or library catalogues), enhancing our research with additional information. Conversely, our own research becomes accessible for others to reuse, building a collaborative, global network of knowledge.

For literary history, LOD offers something particularly powerful: an "atomized" approach to historical data. Instead of collecting evidence for a predefined grand narrative, LOD lets us break literary history down into countless smaller components – events, people, publications, thematic elements – and analyze how these pieces interconnect over time. This approach enables us to observe correlations, co-occurrences, and patterns that

challenge traditional, top-down narratives. We're not just looking at "great works" in isolation but at a broader web of connections that can reveal previously hidden dynamics in the literary landscape.

Moreover, platforms like Wikibase provide a framework for adding context and transparency to historical assertions. Each entry in such a database can include details about the source, certainty level, and time range of any given statement. This level of transparency aligns with the concept of "provenance," which is crucial for documenting the process of knowledge production transparently. We're not simply presenting "facts" but showing the basis for each assertion, allowing others to see the evidence and context behind it.

Of course, there are challenges. One of the biggest challenges in using Linked Open Data (LOD) for literary history lies in data modeling and the use or creation of shared ontologies that can effectively capture the nuances of literary scholarship. Creating models that are flexible enough to accommodate different languages, genres, and historical contexts while maintaining semantic clarity requires collaboration across disciplines. Federation, or the seamless integration of multiple datasets, presents its own set of hurdles: for true interoperability, linked datasets must adhere to compatible standards and ontologies, a goal that requires alignment across institutions and technical infrastructure.

This challenge reminds me of the Text Encoding Initiative (TEI), which is very wide-ranging and flexible, to the point where achieving true interoperability becomes difficult. Just as with TEI, where subsets like Lex-0 for lexicographical data offers standardized, constrained options, I believe ontologies for LOD als need core areas shared widely. Something akin to Dublin Core, but more expansive, could help make our resources more widely accessible and interlinked, while maintaining specificity for different domains. Making federation a reality means balancing the desire for broad connectivity with the need for specificity and scholarly rigor, ensuring that the links we create enhance research rather than introducing noise or misinterpretations.

(5) *The concept of reproducibility has been a significant focus in your recent work, particularly in computational literary studies. Why do you think this is important for the field? What are the unique challenges of ensuring reproducibility in humanities research compared to other disciplines? How can we foster a culture of reproducible research in DH, and what impact might this have on the credibility and progress of the field?*

Reproducibility is fundamental to transparency and openness in research, and in computational literary studies, we can already see that it's essential for advancing the field, for example simply by sharing datasets for others to build on. To truly understand and trust research findings, we need to see how data was constructed and how specific methods lead from data to results. Without reproducibility, this transparency breaks down, and we can't effectively build on each other's work or validate findings.

In computational fields, reproducibility often begins with making data and code openly accessible. But in literary studies, where much of our "data" consists of books that aren't always machine-readable, we face unique challenges. Books are data, but they often aren't digital – or if they are, they aren't always accessible due to copyright restrictions. This raises critical questions about access. To foster reproducibility, we need to work toward making more textual editions both digital and open, which requires creative solutions for copyright issues in the short term, like fair use exceptions, and a shift in copyright laws in the longer term. The current duration of copyright – 70 years after an author's death – feels outdated to me and hinders research by limiting access to a vast archive of literary history.

One of the challenges, beyond building a collective agreement to improve reproducibility and to develop strategies on how to do so, is that making research reproducible takes time and effort for authors; adding reproducibility to the concerns of reviewers, and to the publication process, takes time and effort as well. At the Journal for Computational Literary Studies, which I co-edit with Evelyn Gius and Peer Trilcke, we require data and code to be deposited, and we have guidelines for an optional extended data and code review. Not all authors chose that path, and it would add a huge workload to editors and reviewers if they did. At the same time, I think

that reproducibility requires a community-level effort so a journal, as one cristallisation point of such a community, is a good place to start. Our experience with this is very positive, so far, in any case.

2.3 DH Community and Education

(6) *As the co-director of TCDH and former chair of ADHO, you've been in a unique position to observe and shape the field of Digital Humanities. What have been some of the most memorable experiences or challenges in these roles? How have these experiences influenced your research and leadership approach? What do you see as the most pressing issues facing the DH community today?*

That's a really far-reaching question, even when just considering ADHO. Thinking back on it now, with some honesty, brings back a memory of good moments when we achieved something together, something where we all felt that we were making a positive impact for the community. Things that come to mind are ADHO's Code of Conduct or the new reviewing criteria for the conference submission. Or of course experiencing the annual conference and knowing that you contributed your bit to making it happen, along with many other people. The challenges are more mundane, and include things like having online meetings together cheerfully and productively when the time zones of the board members stretch basically around the globe, from Mexico City and Berekely, via Montréal and the US East Coast, to Paris and Rome, all the way to Seoul, Tokyo and Canberra. Or, in fact, occasions when decisions did not go my way and I was disappointed about that.

As far as I can tell myself, I have always had a rather consensus-oriented leadership style. I don't like to push for things just because I can (and often, of course, I wouldn't be able to do so anyways). Specifically, this means that I like to collect arguments broadly, draw conclusions from them and then propose a decision to the group that I believe follows from the arguments. How you draw those conclusions depends on your values and on the objectives or ambitions you pursue at any given moment. So finding common ground on those is key as well. What I have learned in the past few years is that, as much as this is rational-sounding theory, the reality is that it is hard to collect arguments from everyone, and that in the end, the person or group preparing a meeting shapes the outcomes, and sets the directions, to a considerable extent. I've come to accept this, maybe even enjoy this a bit, occasionally.

In terms of the challenges for the DH community, I think many of them are in fact a result of the extraordinary success and growth of DH recently. How can we maintain stability, openness, and coherence while growing? How can we keep our identity as a community, when there are so many centrifugal forces, not least the cristallisation of more and more distinct subfields, evidenced also by the multiplication of specialized conferences and publication venues? I welcome these developments, of course, and to some extent also drive them, but it does mean that DH is becoming a bigger and bigger tent, in Glen Layne-Worthey's metaphor. Similarly, the increasingly global reach and coverage of the Digital Humanities community is a very positive development, and with ADHO we have a structure that is designed to accommodate these regional communities in all their diversity. But it does mean that, more than ever, we need to mindful of differences in order to work together productively.

(7) *In Korea, there's been a surge of interest in Digital Humanities, partly driven by the perception that it can improve job prospects for humanities graduates through coding skills. What are your thoughts on this trend? While technical skills are important, what other competencies do you think are crucial for DH scholars? How can we ensure that the core values and critical thinking of humanities are not lost in the rush to digitize?*

Absolutely, Digital Humanities skills – like programming skills, experience with database design, or data analysis and visualization skills – can be a real asset on the job market. It's true, also, that the coding and technical

aspects often draw interest among the students, especially among those looking to increase their career opportunities. But DH scholars aren't computer science students, nor should they be seen as just a source of technical skills. What they bring to the table goes far beyond coding; it's a unique ability to bridge the worlds of humanities and digital technology in ways that are increasingly valuable in both academic and professional settings.

Beyond technical know-how, DH scholars are adept at managing complexity, learning quickly, and adapting to new tools. They're skilled, for example, at connecting various tools by transforming data to the right formats, a pretty powerful skill. More importantly, they can translate between the technical language of computer science and the interpretive language of the humanities. This means they not only understand what's easy or hard to implement, but they also recognize what matters most to users, whether in a tech company, museum, or archive. They're often the ones who serve as communicators and negotiators, ensuring that digital solutions align with the needs and experiences of real people – a role that's crucial for product managers, educational resource developers, publication managers, curators and archivists alike.

And while technical skills are valuable, more isn't always better. It's essential to focus on understanding processes and thinking critically about how tools and algorithms are built. In my teaching, for example, I emphasize the importance of creating simple, foundational versions of algorithms ourselves – whether it's linear regression, k-means clustering, a Naive Bayes classifier, or even a simple neural network. By building these step by step, students gain a deeper insight into how things work, which is invaluable once they start using more sophisticated libraries to implement these algorithms efficiently. When you understand the basics, you're better equipped to diagnose why something might go wrong or spot biases that could be embedded within these "black box" models. It's also quite grounding, in a way, to see that even a magically-seeming neural network is actually just some matrix multiplications arranged in a very clever way.

DH scholars are uniquely positioned to bring this perspective forward, ensuring that while we embrace and understand the new, we don't lose the reflective, critical values of the humanities.

2.4 Future Directions

(8) *The DH2026 conference will be held in Daejeon. What are your expectations for this event? How do you think it might reflect or influence the future direction of DH, particularly in Asia? Looking ahead, what are your own research plans and priorities? Are there any emerging trends or technologies in DH that you're particularly excited about?*

I'll start with my own research priorities, to get that out of the way and because I have already touched on a lot of it. First of all, I'm currently interested in continuing to bridge large language models (LLMs) and knowledge graphs / Linked Open Data. At the same time, I'd like to help increase uptake of LOD in a range of areas of the Digital Humanities, and as soon as you take LOD seriously, the issue of federation between multiple resources, and interoperability of vocabularies and data models appears. My third priority is multilingualism. I'll continue to develop and evaluate methods that work across languages, including in domains like stylometric authorship attribution or keyword analysis. Multilingual support is critical for DH as it grows globally, it's one of the key challenges but also strengths of DH in Europe, and it's something I hope to explore with collaborators across regions.

These ambitions of course intersect with the broader trends, and I do think DH is moving towards enhanced data sharing, semantic precision, and cross-language support, thanks in part to advances in machine learning, but in part also thanks to conscious efforts to work against and beyond the dominance of English-oriented resources and tools. DH2026 could be a milestone for these developments, especially in Asia, and I can't wait to be part of

it.

I'm thrilled, of course, about the Digital Humanities Conference 2026 in Daejeon and the opportunities it offers. Having the conference in South Korea is a highly significant milestone – not only does it bring the spotlight to a dynamic and unique Digital Humanities community in Asia, but it also reinforces DH's evolution beyond its traditional centers in North America and Europe. With recent conferences held in Sydney and Mexico City, and Tokyo being the host of the online DH conference in 2022, this marks a continuation of expanding DH's global reach. Asia, and South-East and East Asia especially, is an area of vibrant growth in DH. It's high time that, after the expansion in terms of ADHO's constituent organizations over the last years, the conference also reflects this shift.

From my trip to South Korea in the spring of 2024, I can say with confidence that we can expect a high level of professionalism and a welcoming atmosphere, when coming to Daejeon. The unique combination of a society very oriented towards technology, but at the same time cherishing and celebrating its cultural and historical richness, will make this event unforgettable. Personally, I'm excited to reconnect with friends, meet new colleagues, and, of course, keep an open mind for the unexpected insights and collaborations that every DH conference brings. There's a real possibility here for bolstering networks in Asia, offering a chance for those in the region to connect without the added challenges of intercontinental travel. I'd love to see more networking between DH communities in Japan, Taiwan, and South Korea as well as those emerging in places like Indonesia or Hong Kong as well as mainland China. In any case, I look very much forward to meeting the South Korean Digital Humanities community in Daejeon in 2026!

2.5 Suggestions for further readings

For readers interested in diving a little deeper into some of the topics raised above, a few pointers are provided here to articles published by Christof Schöch and colleagues.

1. On the theoretical investgations, one may consult "Revisiting Style, a Key Concept in Literary Studies" (2015), "From Keyness to Distinctiveness" (2021) or, on reproducibility, "Repetitive Research" (2023).
2. The idea of smart data was developed in "Big? Smart? Clean? Messy? Data in the Humanities" (2013) and combined with the Linked Open Data paradigm in "Smart Modeling for Literary History" (2022). Readers may wish to consult the MiMoTextBase as well.
3. Regarding methods, a summary on subgenre analysis is provided in "Computational Genre Analysis (2022), distance measures for stylometry are evaluated in "Understanding and Explaining Delta Measures for Authorship Attribution" (2017) and an investigation into multilingual authorship attribution is described in "Multilingual Stylometry" (2024), accompanied by an interactive showcase.
4. Generally speaking, recent, international work in Computational Literary Studies is published regularly in the *Journal of Computational Literary Studies*.