

레이와 대장경(令和太藏經)과 차세대 인문학

현재 연구 매체의 디지털화(DX)와 인공지능(AI)이 결합됨으로써 새로운 연구 환경이 빠르게 형성되고 있다. 본 연구는 일본의 인문학이 이 새로운 환경에 대응하기 위한 요건을 명시하는 동시에, 이를 실현하기 위한 디지털 학술 공간 모델을 구축하여 인문학의 전통적 활동을 발전적으로 재편하기 위한 구체적 방향을 제시한다.

공식 웹사이트(일본어)



2025.8
Collaborative
Research Data
Infrastructure
for the Humanities

제1호

<https://reiwadzk.dhii.asia/>

<https://creativecommons.org/licenses/by-sa/4.0/deed.ja>

1 창간 인사

●연구대표자: 시모다 마사히로(下田正弘) [무사시노대학]



일본학술진흥회 과학연구비 지원사업 특별추진 연구 ‘디지털 연구기반으로서의 레이와 대장경 편찬—차세대 인문학을 위한 연구기반 구축 모델의 제시’를 시작하며, 이 활동 내용을 학계 여러분께 널리 알리는 동시에, 본 연구 과제가 분야를 넘어 ‘지(知)의 미래’를 창조하는 대화를 이끌어내는 계기가 되기를 바라며 본지를 창간합니다.

근대 이래 인류는 전문 분야의 세분화를 통해 지(知)의 심화를 추진해 왔습니다. 그 결과, 현재 방대한 지식이 각각의 의미 영역을 구성하여 병존하게 되었습니다. 과거 이러한 전문 지식을 종합하는 장(場)인 아카데미즘은 대체로 서로 독립적인 개별 전문가들로 구성되었으며, 비전문가가 전문 지식을 얻고자 할 때는 각 전문가에게 정해진 절차에 따라 문의하는 것 외에는 방법이 없었습니다. 하지만 지난 세기 후반부터 정보통신기술(ICT)이 급속히 진보하고, 무엇보다 인공지능(AI)이 폭발적으로 진화함에 따라 이러한 지식 환경은 극적으로 변화하고 있습니다. 가속화되는 데이터 공개와 고성능 컴퓨터를 통해 지(知)는 전문 분야의 벽을 넘어 연계되었고, 사람을 통하지 않고도 비전문가가 빠르게 이용할 수 있게 되었습니다. 이러한 환경에서 비롯된 새로운 가능성의 활용, 그리고 이로부터 생성되는 지(知)의 신뢰성 확보라는 과제에 어떻게 대응할 것인가, 이것이 지금 모든 과학 분야에 주어진 과제입니다.

특히 인문학에서는 이 과제가 다른 학문 분야에서는 볼 수 없는 섬세하고 복잡한 양상으로 제기되고 있습니다. 언어, 역사, 사회, 문화의 차이에 따라 축적된 지식 기반에서 생성되는 지(知)는 마치 그 축적 과정을 소급하듯이 의미를 재현합니다. 따라서 새로운 기술을 매개로 연구 지식 기반을 재구성할 때는 단순히 신기술을 도입하는 것이 아

니라 아카이브라는 작업 자체에 내재된 문제를 객관적이고 심도 있게 분석하여 문제를 해결하려는 자세가 중요합니다. 본 연구는 바로 이 과제에 대응하기 위해, 근대 동양학을 대표하는 불교학의 지식 기반을, 역사적 아카이브학의 경로를 명확히 밝히며 재편하고, 하나의 학문 분야 전체의 ‘지(知)의 기반’을 재편할 수 있는 모델을 제시하고자 하는 것입니다.

20세기 초에 편찬된 <다이쇼 신수대장경(大正新脩太藏經)>은 중국 송대 이후 약 1,000년에 걸친 불전 계승의 역사를 내포한 불교 성전을 근대적 문헌학 기법으로 집대성한 획기적인 사업으로, 오늘날까지 100년에 걸쳐 전 세계 동양학 연구의 확고한 지식 기반으로서 그 역할을 수행해 왔습니다. 본 연구는 인문학의 특성을 고려하여 ICT와 AI를 활용하고, 송판(宋版), 원판(元版), 고려판 등 모든 판본을 망라적으로 교감하여 <다이쇼 신수대장경>이 내포한 역사 전체를 시각화할 것입니다. 그리고 연구자 간에 실시간으로 공유할 수 있는 ‘공동창조적 디지털 학술 플랫폼 CRDIH(Collaborative Research Data Infrastructure for the Humanities)’를 구축하여 레이와판 <다이쇼 신수대장경>이라 할 수 있는 <레이와 대장경> 편찬의 기반을 마련하고자 합니다.

이러한 시도는 연구 프로세스의 시각화, 성과의 공유, 그리고 전문 분야 간의 경계를 넘는 협업의 촉진을 통해 새로운 학술 문화 창조의 구체적인 방향을 제시하며, 해석의 기초가 되는 텍스트 크리틱(원전 고증) 작업을 디지털 시대에 부합하는 형태로 실현하는 것을 목표로 합니다. 나아가 불교학을 넘어, 일본 인문학의 제반 영역에 기여하는 것을 중요한 목표로 삼고 있습니다. 이를 위해서는 연구자 여러분의 깊은 이해와 성원이 무엇보다 절실합니다. 본 연구가 성공적으로 완수될 수 있도록 여러분의 아낌없는 협조를 진심으로 부탁드립니다.

2 지금까지의 기반 정비와 향후 방향

● 나가사키 기요노리(永崎研宣) [일반재단법인 인문정보학연구소/게이오대학]

‘디지털 연구기반으로서의 레이와 대장경 편찬—차세대 인문학을 위한 연구기반 구축 모델의 제시(JP25H00001)’(이하, 본 연구)는 디지털 시대에 부합하는 새로운 대장경 편찬을 목표로 하는 연구 사업이다. 이 사업은 SAT 대장경 텍스트 데이터베이스 연구회(이하, SAT 연구회)가 지난 30년간 추진해 온 불전 연구를 위한 디지털 연구 환경의 기반 정비 활동을 토대로 기획되었다. 여기서는 먼저 그 기반 정비에 관한 내용을 소개하고자 한다.

2007년 완성, SAT 대장경 텍스트 데이터베이스

본 연구의 핵심은 SAT 연구회를 뒷받침해 온 많은 분들의 노력으로 구축된 1억 자 이상의 SAT 대장경 텍스트 데이터베이스(https://21dzk.l.u-tokyo.ac.jp/SAT/index_en.html 이하, SAT 텍스트)이다. 이는 <다이쇼 신수대장경> 중 텍스트 부분 85권을 전산화한 것으로, 2007년 처음 완성되었다. 그 후 2008년 4월에는 웹 서비스로서 전문 검색을 시작했으며, 찰스 윌러 교수의 Digital Dictionary of Buddhism(<http://www.buddhism-dict.net/ddb/>)과 일본인도학불교학회의 논문 데이터베이스 INBUDS(<https://www.inbuds.net/kor/>)와의 연계 검색 기능을 제공하면서 널리 활용되기 시작했다. 주요 이용자를 연구자로 상정했기 때문에, 연구 지원의 일환으로 다른 다양한 데이터베이스나 웹 서비스와의 연계를 추진해 나갔다. 구체적으로는 2009년에 공개된 CiNii의 웹 API를 활용한 논문의 PDF 링크 기능과, 컬럼비아대학의 Buddhist Canons Research Database(<http://databases.aibs.columbia.edu/>)와의 연계 검색 등 다양한 기능을 제공해 왔다.

텍스트의 정확한 전달—불전 한자의 유니코드(Unicode)등록

한편, 텍스트 데이터 자체에 대한 검토 작업도 병행했다. 먼저 문자 문제에 대해 살펴보자.

SAT 텍스트는 2007년 완성 당시 주로 Shift-JIS를 기반으로 구축되었기에, 그 상태로는 표시할 수 없는 문자가 많았고 결과적으로 많은 외자(外字: 표준 문자 코드표에 없는 문자)를 포함하게 되었다. 당시 외자는 외자 번호와 문자 이미지를 조합하여 표시하는 방식을 취했다. 활판 인쇄본인 <다이쇼 신수대장경>은 전거가 된 목판본의 다양한 문자를 통합하여 편찬된 것이며, 문자의 종류는 고려판(高麗版)이나 송판(宋版), 원판(元版), 명판(明版)에 비해 상당히 적었지만, 그럼에도 불구하고 2007년 당시에는 약 10,000자의 외자가 포함되어 있었다. 그러나 이를 그대로 방치하면 복사·붙여넣기, 인쇄, 텍스트 분석 등 다양한 상황에서 적절한 처리를 할 수 없게 된다. 그 결과 텍스트의 전달이 원활하지 않아 오류가 혼입될 우려가 있으며, 결국 텍스트 데이터의 학술적 신뢰성을 충분히 담보할 수 없게 될 가능성이 커

지는 것이다.

현재의 컴퓨터 환경에서 이러한 문자를 적절히 처리하기 위해서는 국제표준화기구(ISO)가 전 세계의 문자를 고유하게 식별할 수 있도록 제정한 국제 표준 규격인 ISO/IEC 10646에 해당 문자를 등록하여 이를 유니코드 문자로 사용할 수 있게 해야 한다. 이에 SAT 연구회는 점차 늘어나는 유니코드 수록 한자(공식 명칭: CJK 통합 한자)와 SAT 텍스트의 외자를 매핑하는 작업을 진행하는 한편, 2012년부터는 국제표준화기구 산하에서 유니코드 한자 코드표를 제정하는 국제회의에 참가하였고, 해당 위원회의 연례회원으로서 정식 가입하게 되었다. 그 성과로 3,000자가 넘는 불전의 한자가 유니코드에 등록되었고, SAT 텍스트의 대부분을 일반적인 텍스트 데이터로서 다룰 수 있게 되었다. SAT 연구회가 유니코드에 한자를 등록하는 데에는 초기에 약 6년이 소요되었으나, 이는 단순히 SAT 텍스트의 편의성을 높인 것에 그치지 않고, 학술적 필요에 기반한 한자 등록 노하우를 축적하고 관련 절차를 확립했다는 의미를 갖는다(그림 1 참조). SAT 연구회는 문자 처리에 관하여 이처럼 실질적인 행동을 통해 대응 가능한 조직 역량을 갖추고 있으며, 바로 이것이 본 연구의 기초를 굳건히 지탱하고 있다.

이미지와 텍스트를 자유자재로 활용—IIIF를 통한 이미지 공개

SAT 텍스트의 데이터베이스로서의 신뢰성을 제고하기 위한 또 하나의 방안으로, 텍스트 데이터에만 의존하지 않고 판면에 자형(字形)이 실제로 어떻게 기재되어 있는지를 확인할 수 있는 환경 또한 중요하다. 이를 위해 앞서 언급한 유니코드 문자 등록과 병행하여 관련 작업을 추진해 왔다.

The Unicode Standard, Version 16.0, Copyright © 1991–2024 Unicode, Inc. All rights reserved. 3043

[그림 1] 유니코드에 등록된 문자와 그 제안 서류의 일례

즉, <다이쇼 신수대장경>의 전 페이지의 스캔 데이터(SAT 텍스트)와 이를 붙인 이미지와 연계하여 열람할 수 있도록 하는 시스템의 개발이다. SAT 텍스트는 페이지 번호와 행 번호를 기준으로 작성되어 있기 때문에, 해당 문자가 페이지 내 어느 위치에 있는지를 기계적으로 추정하기가 비교적 용이하다. 이에 따라 일부 예외를 제외하고, 2012년에 공개된 SAT 웹 서비스 버전에서는 텍스트를 클릭하면 해당 위치가 표시되는 기능이 제공되었다. 이 시점에서는 필자가 자체 제작한 이미지 표시 시스템을 사용했으나, 이 주제가 국제적으로도 중요한 사안이었기에 이 방식을 국제적인 공통 규격으로 만들어 효율성을 높이려는 움직임이 나타났다. 이것이 바로 2011년 구미의 주요 연구 도서관을 중심으로 시작된 IIIF(International Image Interoperability Framework, 트리플 아이에프), 즉 웹상에 공개된 이미지를 국제적으로 상호 운용하는 것을 목표로 하는 프레임워크이다.

SAT 연구회 역시 이 프레임워크의 도입을 시도했으나, 실제로 일반 컴퓨터 환경에서 원활하게 사용할 수 있게 된 것은 2016년 무렵이었다. 당시에는 돈황 사본을 소장한 프랑스 국립도서관이나 풍부한 동아시아 고문헌을 보유한 하버드대학 도서관 등이 이미 IIIF를 지원하는 방식으로 이미지를 공개하고 있었고, SAT 역시 2016년 공개 예정이었던 <다이쇼 신수대장경> 도상편(圖像編)의 이미지 데이터베이스인 'SAT 다이쇼 대장경 도상 DB'에 이미지 어노테이션(각 불존 등을 설명하는 그림 색인)을 표시하기 위해 이를 도입하였다. 이는 일본 내 최초 공개 사례였으며, 이후 국립국회도서관, 국문학연구자료관, 각지의 대학 도서관 등 불전을 포함한 고문헌을 디지털로 공개하는 기관들 사이에서 IIIF 기반의 이미지 공개가 확산되었다. 이에 발맞추어 일본 내 IIIF 보급을 위한 세미나와 컨설팅이 각지에서 실시되었는데, 이 과정에서 SAT 연구회의 축적된 경험이 널리 활용되었다.

IIIF를 활용하면 웹상에 공개된 디지털 자료의 이미지 중, 페이지 내

특정 영역을 외부에서 직접 지정하여 참조하거나 추출하는 것이 가능하다. 이 덕분에 특정 자료에 등장하는 특이한 자형(字形)의 한자 용례를 자료 전체의 맥락 속에서 제시할 수 있게 된다. 이는 자료 간의 차이를 확인해야 하는 텍스트 판독 과정에서 유용한 기능일 뿐만 아니라, 앞서 언급한 유니코드 한자 등록 시 제안 한자의 근거 자료를 제시할 때도 매우 편리하다. 따라서 SAT 연구회의 유니코드 한자 등록 절차 또한 IIIF 도입과 함께 한층 더 효율화되었다(그림 1 참조).

레이와 대장경 편찬 시스템이란?

이러한 IIIF의 기능은 레이와 대장경 편찬에서도 매우 중요한 역할을 하고 있다. 현재 진행 중인 OCR 텍스트 수정 작업의 화면(그림 2 참조)에서는 송판(宋版)·원판(元版)·명판(明版)·고려판(高麗版)·궁내청 송판 일체경의 OCR을 수정하기 위해 텍스트와 이미지를 행 단위로 대조하고 있는데, 여기서 텍스트의 각 행을 표시할 때도 이 IIIF 기능을 활용하고 있다. 이 경우 각각의 판본을 공개하고 있는 조조지(増上寺, 송판·원판·고려판)와 도쿄대학 도서관(명판), 게이오대학 시도(斯道)문고(궁내청 송판)가 모두 IIIF 호환 방식으로 이미지를 공개하고 있기 때문에, 레이와 대장경 편찬 시스템을 구축하는 데 있어 단일 뷰어 프로그램을 개발하는 것만으로 모든 기관의 이미지에 대해 정확히 대응하는 행을 표시할 수 있다. 또한 추후 새로운 대장경이 추가되는 경우에도, IIIF를 지원한다면 동일한 시스템으로 처리가 가능하다.

본 연구에서는 이와 같이 관련 국제 표준 규격의 개선과 표준 기술의 보급 등 기반적인 부분을 공들여 정비해 온 점이 중요한 전제가 되고 있다. 여기서 언급한 것 외에도 특히 TEI화 작업은 매우 중요한 과제이나, 이에 관해서는 다음 기회에 논하고자 한다.



[그림 2] 텍스트와 이미지를 행 단위로 비교

3 키포프 심포지엄 보고

●박현진 [일반재단법인 인문정보학연구소 불전텍스트연구부문 연구원]

지난 2025년 6월 14일(토)부터 15일(일)까지 이틀에 걸쳐, 본 연구 과제의 키포프 심포지엄이 무사시노대학 아리아케 캠퍼스에서 온·오프라인 하이브리드 형식으로 개최되었다. 본 심포지엄은 연구 협력자들과 본 연구 과제의 구상을 공유하고 연구의 본격적인 시작을 알리는 뜻깊은 자리가 되었으며, 참석자들의 깊은 관심과 호응 속에 성황을 이루었다.

첫날은 연구대표자 시모다 마사히로 교수의 개회사 및 과제 소개로 시작되었다. <다이쇼 신수대장경> 편찬 100주년을 맞이하는 시점에서, 본 연구의 목적은 단순한 디지털화를 넘어 AI와 디지털 기술을 전제로 하는 시대에 '지(知)의 기반'을 어떻게 구축할 것인가를 묻는 차세대 인문학 연구 모델을 제시하는 데 있음이 강조되었다. 이어 이 모델을 구현하는 핵심 시도로서 SAT(대장경 텍스트 데이터베이스 연구회: <https://21dzk.l.u-tokyo.ac.jp/SAT/>)의 역사와 사회적 지원의 경위가 소개되었으며, <레이와 대장경>이 전통적인 대장경 편찬의 역사를 디지털 기반 위에서 계승하고 재편하는 과업임이 강조되었다.

둘째 날에는 공동연구자인 나가사키 기요노리 교수가 본 연구의 핵심인 AI-OCR을 활용한 불전의 고정밀 텍스트화, TEI 가이드라인에 준거한 구조화,

IIIF를 활용한 이미지와 텍스트의 연계 등 기술적 기반을 발표하였다. 특히, 조조지(増上寺) 소장 3대장(3대 대장경) 등 귀중한 목판본의 AI-OCR 텍스트와 기존 SAT 본문의 차이를 하이라이트로 표시하는 교정 지원 사용자 인터페이스(UI)가 소개되었으며, 단계적 구조화를 위한 구체적 워크플로가 제시되었다.

이번 심포지엄의 가장 큰 특징은 참가자 전원이 참여한 활발한 토론이었다. 작업 분담의 구체화, 신진 연구자 육성, 국제 표준에 준거한 성과 공개 모델 등 다각적인 과제에 대해 건설적인 의견 교환이 이루어졌다. 이를 통해 본 연구가 단순한 데이터 구축을 넘어, 지속 가능한 지식 기반 형성, 신진 연구자 육성, 그리고 인문학의 미래를 향한 도전이라는 공통 인식이 확립되었다. 이로써 프로젝트의 방향성을 재확인하고, 본격적인 출범을 향한 중요한 첫걸음을 내디뎠다.



4 공개 학술 행사 리포트—DH의 최전선

●나가사키 기요노리

본 연구 과제에서는 <레이와 대장경> 편찬과 관련된 다양한 요소에 대해 국내외에서 폭넓게 논의하고, 불교학뿐만 아니라 인문학 전반에 기여하고자 불교학 및 디지털 인문학에 관한 공개 학술 행사를 개최해 왔다. 또한 논의의 장을 널리 공유하기 위해, 원칙적으로 온라인 실시간 송출을 병행하고 있다.

2025년 4월 11일(금) [후원]

디지털 인문학 국제 워크숍 '디지털 텍스트 연구의 새로운 전개'

호르닝엔대학의 페데리코 피안줄라(Federico Pianzola) 교수를 기조 강연자로 초청하여, 연구대표자 시모다 마사히로 교수, 공동연구자 나가사키 기요노리 교수와 더불어 국문학연구자료관에서 텍스트 분석을 담당하는 다케우치 아야노(竹内綾乃) 연구원의 강연이 진행되었다. 피안줄라 교수는 EU의 유럽연구위원회(ERC)가 지원하는 디지털 문학 연구 프로젝트 'GOLEM(Graphs and Ontologies for Literary Evolution Models) <https://golemlab.eu/>'을 이끌고 있으며, 이날 강연에서는 독서 경험을 통해 문학 작품을 분석하는 GOLEM 프로젝트에서의 그래프 모델과 온톨로지 기반 구조화에 대해 발표하였다. 이러한 시도는 불전 연구에서의 응용 가능성을 실감하게 하여, 현장에서는 활발한 토론이 이루어졌다.

2025년 5월 17일(토) [공동개최]

제138회 '정보처리학회 인문과학과 컴퓨터 연구회' 발표회

정보처리학회 산하 분과회로서 1989년부터 이어져 온 '인문과학과 컴퓨터 연구회(<https://www.jinmoncom.jp/>)'의 정례 발표회가 게이오대학 미타 캠퍼스에서 개최되었다. 본 연구회는 매년 5월 학부생·대학원생을 대상으로 한 '학생 세션'을 운영하고 있는데, 이번 행사에서는 일반 발표뿐만 아니라 각지의 학생들이 모여 포스터 발표를 진행하는 등 총 18건의 발표가 이루어져 성황을 이루었다. 본 연구 과제에 관해서는 공동연구자 나가사키 교수가 'CH 연구회에서의 디지털 대장경 연구의 성과와 전망'이라는 특별 강연을 통해, 이 연구회에서 진행된 디지털 대장경 관련 연구로부터 본 연구 과제로 이어지는 흐름을 총괄하고 향후 전망을 제시했다.

2025년 6월 3일(화) [공동개최]

DH 국제 심포지엄 '디지털 이미지와 텍스트의 새로운 전개: 자동 문자 인식의 최신 동향과 그 활용'



OCR, 그중에서도 필기체 OCR의 비약적인 발전에 힘입어 디지털 이미지와 텍스트 데이터의 관계가 새로운 국면을 맞이하고 있는 가운데, 관련 연구자와 개발자들이 모여 현황을 공유하고 향후 전개를 논의하는 귀중한 자리가 마련되었다.

무료 오픈소스 필기체 OCR 'eScriptorium(<https://gitlab.com/scripta/escriptorium>)' 개발 프로젝트를 주도하는 PSL대학의 피터 스토크스(Peter Stokes) 교수를 비롯한 4명의 연구자와, 본 연구 과제에서도 도입한 국립국회도서관의 고문헌 OCR을 개발하고 있는 아오이케 도루(青池亨) 씨 등이 연사로 나서, 텍스트와 이미지의 새로운 관계에 대해 다양한 과제와 사례를 발표하였다. 본 연구 과제와 관련해서는 공동연구자 나가사키 교수가 일본의 쿠즈시지(흙림제) 지원 OCR 소프트웨어의 최신 현황과, 이러한 흐름 속에서 <레이와 대장경> 편찬 프로젝트가 갖는 의의에 대해 발표하였다. 동시통역과 함께 온라인으로 개최된 이번 심포지엄은 100명 이상의 참가 신청을 기록하며 대성황을 이루었다.

2025년 6월 13일(금) [공동개최]

DH 국제 워크숍 '불전 AI 자동 번역의 최첨단: 전용 LLM의 개발과 이를 둘러싼 인문학 연구자의 커리어 형성'

불교학 분야 AI 활용의 최전선을 이끄는 UC 버클리의 세바스찬 네르디히(Sebastian Nehrlich) 박사를 초청하여, Dharma mitra 프로젝트(<https://dharmamitra.org/ja/>)의 AI 활용의 최신 현황에 대한 워크숍을 개최하였다. 네르디히 박사는 급속히 발전하는 생성형 AI를 활용함으로써 Dharma mitra 프로젝트가 자동 번역뿐만 아니라 불전 이미지 OCR 또한 높은 정확도로 수행할 수 있게 되었음을 발표하였다. 아울러 네르디히 박사는 기존의 불교학 연구에서 디지털 불교 연구로 연구 분야를 전환한 경위에 대해서도 소개하였다. AI의 활용이 바야흐로 다양한 국면에서 가능해지고 있음을 실감할 수 있었던 뜻깊은 자리였다.



DH의 최신 동향 및 관련 행사 정보는 <인문학 정보 월보>를 참고하시기 바랍니다.