

NEWSLETTER

The New Tripitaka and Next-Generation Humanities

The convergence of digital transformation (DX) and artificial intelligence (AI) is rapidly creating a new research environment. This project identifies the requirements that Japanese humanities scholarship must meet in this new context, offering a model for a digital scholarly space that fulfills those requirements and sets out concrete pathways for the developmental reorganization of traditional practices in humanities fields.

2025.8
Collaborative
Research Data
Infrastructure
for the Humanities

NO. 1

Official Website [Japanese]



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

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

1 Inaugural Remarks

● Principal Investigator: Prof. Masahiro Shimoda (Musashino University)



With the start of the JSPS Grant-in-Aid for Scientific Research (Specially Promoted Research) project “Compilation of the New Tripitaka as a Digital Research Infrastructure: A Model for Constructing a Research Infrastructure for Next-Generation Humanities,” we are launching this newsletter to report our activities to the wider academic community and hopefully catalyze cross-disciplinary conversations that help shape the future of knowledge.

Since the modern era, knowledge has been advanced largely through specialization and subdivision. As a result, vast bodies of knowledge now coexist, constituting individual domains of meaning. Traditionally, academia—the place where specialized knowledge is supposed to be synthesized—has been constituted by individual experts whose connections to one another have remained largely unexamined; non-specialists seeking expert knowledge typically had no choice but to consult each specialist through the prescribed methods. Over the latter half of the 20th century, however, rapid advances in information and communication technology (ICT) and, more recently, the explosive development of AI have dramatically altered this landscape. Accelerated data sharing and high-performance computing have enabled linkages across disciplinary boundaries, and knowledge has become directly accessible to non-specialists as well. Across all fields of scientific study, we are now challenged both to harness these new possibilities and to ensure the reliability of the knowledge being delivered.

This challenge is particularly delicate and complex in the humanities. Knowledge generated from archives organized along linguistic, historical, social, and cultural lines reproduces meaning in ways that retrace the very processes by which they were accumulated. Reconstructing research knowledge bases with new tech-

nologies, therefore, requires more than adopting the latest tools; it calls for a rigorous analysis of the issues inherent in archiving itself and for strategies that resolve them. Our project addresses precisely this task by reorganizing the knowledge base of Buddhist studies—a pillar of the study of East Asia—while clarifying the historical pathways of its archival formation, thereby offering a model for refashioning the knowledge base of an entire academic field.

Compiled in the early 20th century, the *Taishō Shinshū Daizōkyō* (Taishō Revised Tripitaka) was a momentous enterprise that, using modern philological methods, assembled Buddhist scriptures encompassing roughly a millennium of canonical transmission from the Song dynasty onward. For the past century, it has served as a bedrock resource for the study of East Asia worldwide. On this foundation, our project employs ICT and AI in ways appropriate to the humanities to collate, comprehensively and comparatively, the principal transmitted editions of the Buddhist canon—Song, Yuan, Goryeo, and others—so as to render visible the historical strata embedded in the *Taishō Shinshū Daizōkyō*. We are developing the Collaborative Research Data Infrastructure for the Humanities (CRDIH), which enables real-time sharing between researchers and thereby lays the groundwork for compiling the *Reiwa Daizōkyō*—a Reiwa-era version of the *Taishō Shinshū Daizōkyō* that opens new possibilities.

By making research processes visible, sharing results, and promoting collaboration across fields, this endeavor charts concrete routes toward a new scholarly culture. Our aim is to implement textual criticism in forms suitable for the digital age. We seek to contribute not only to Buddhist studies but also to the broader field of humanities in Japan. To that end, we ask for the understanding and support of the scholarly community as we move this project toward completion.

2 What We Have Built—and What Comes Next

● Prof. Kiyonori Nagasaki (International Institute for Digital Humanities / Keio University)

The project “Compilation of the New Tripitaka as a Digital Research Infrastructure: A Model for Constructing a Research Infrastructure for Next-Generation Humanities” (JP25H00001) aims to compile a new Tripitaka suited to the digital era. It builds upon 30 years of work by the SAT Daizōkyō Text Database Committee (hereafter the SAT Committee) to establish a digital research environment for Buddhist textual studies. I begin with that foundation.

The SAT Daizōkyō Text Database (Completed 2007)

At the core is the SAT Daizōkyō Text Database (https://21dzk.l.u-tokyo.ac.jp/SAT/index_en.html; hereafter SAT Text), a corpus of over 100 million characters created by numerous contributors supporting the SAT Committee. The SAT Text comprises transcriptions of the 85 textual volumes of the *Taishō Shinshū Daizōkyō* and reached initial completion in 2007. In April 2008, it launched as a web service with full-text search and cross-resource querying with Prof. Charles Muller’s Digital Dictionary of Buddhism (<http://www.buddhism-dict.net/ddb/>) and the INBUDS article database (<https://www.inbuds.net/eng/>) of the Japanese Association of Indian and Buddhist Studies, and it soon saw widespread use. Because researchers were the primary target users, the SAT Text developed numerous integrations, including CiNii-based links to article PDFs (using the 2009 CiNii Web API) and a linked search with Columbia University’s Buddhist Canons Research Database (<http://databases.aibs.columbia.edu/>).

Ensuring Faithful Textual Transmission: Registering Buddhist Characters in Unicode

Work on the text data itself proceeded in parallel. The following section addresses the issue of character encoding.

When the SAT Text was completed in 2007, it was primarily encoded in Shift JIS. Thus, many characters could not be displayed natively, leading to a large number of user-defined characters (gaiji) being included. They were displayed by combining external character numbers with character images. Although the letterpress edition of the *Taishō Shinshū Daizōkyō* unified many character variants found in its woodblock exemplars—and thus employs far fewer glyph forms than the Goryeo, Song, Yuan, and Ming editions—the SAT Text still included roughly 10,000 external characters at that time. Left unresolved, this impedes copying, printing, and computational analysis, and risks corrupting transmission and undermining the scholarly reliability of text data.

To properly handle such characters in current computing environments, they must be registered in the ISO/IEC 10646 universal character set—an international standard established by the International Organization for Standardization (ISO) to uniquely handle all the world’s characters—so they can be used as Unicode characters. Accordingly, the SAT Committee aligned its external

characters with the expanding Unicode repertoire (CJK Unified Ideographs) and, beginning in 2012, participated in the international meetings under the ISO that maintain the CJK code charts, officially joining the relevant ISO committee as a liaison member. As a result, more than 3,000 characters attested in Buddhist texts were added to Unicode, and the great majority of the SAT Text can now be handled as ordinary text. Although the SAT Committee initially required about six years to register characters to Unicode, beyond usability gains, this work established robust procedures—and institutional know-how—for registering academically necessary characters (see figure 1). The SAT Committee thus became an organization capable of addressing the handling of characters, supporting the foundation of this research project.

Working Seamlessly Across Text and Image: Publishing Images via IIIF

To enhance the reliability of the SAT Text as a database, at times, another important approach would be to verify how a given glyph was actually printed on the page, rather than rely solely on text data. This was handled in parallel with the Unicode character registration work mentioned above.

In other words, the SAT Committee therefore scanned every page of the *Taishō Shinshū Daizōkyō* and built mechanisms to link the transcriptions to page images. Because the SAT Text is keyed to page and line numbers, it is straightforward to estimate where a given character appears in an image. Since 2012, the web interface has allowed users to click a character to reveal the corresponding location, with some exceptions. While a viewer, which I built, ini-

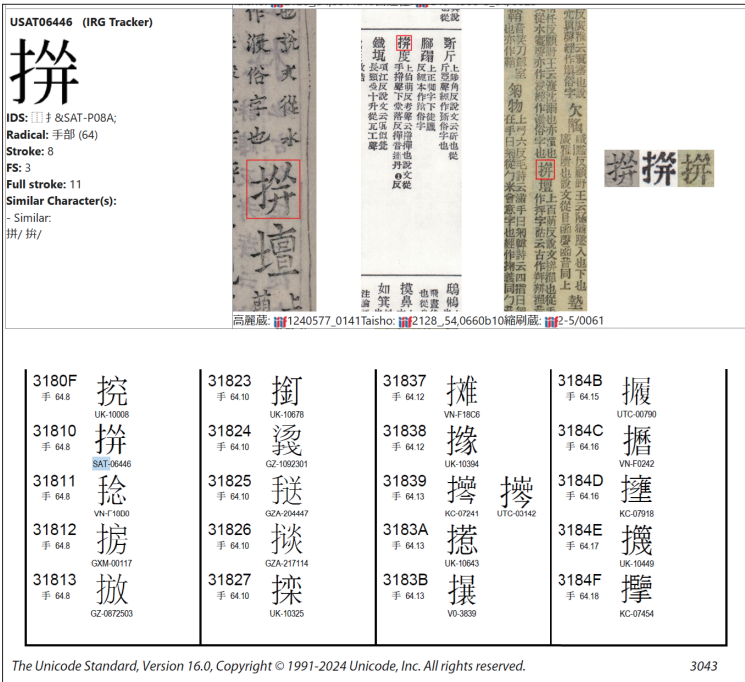


Figure 1. Examples of characters registered in Unicode with proposal documentation

tially handled this, a shared framework emerged due to the worldwide importance of this issue: the International Image Interoperability Framework (IIIF; pronounced triple-I-F), launched in 2011 by leading European and North American research libraries.

While the SAT Committee was also experimentally using this framework, it was around 2016 that IIIF matured for general use. Institutions such as the Bibliothèque nationale de France (with its Dunhuang manuscripts) and Harvard Library (with extensive East Asian holdings) began IIIF-compliant publication. SAT adopted IIIF that same year for the SAT Taishōzō Image Database, including image annotations (iconographic “picture-index” entries for deities, etc.). This was Japan’s first public IIIF deployment, which helped catalyze broader domestic adoption by organizations releasing classical texts, including Buddhist scriptures, digitally—such as the National Diet Library, the National Institute of Japanese Literature, and numerous university libraries. The SAT Committee’s accumulated know-how supported nationwide seminars and consultations that accelerated uptake.

IIIF enables external references to any location on any page image of web-published materials, making it possible to gather instances of unusually-shaped Chinese characters found in a specific manuscript across a corpus—vital for identifying differences between manuscripts when reading a text—and to supply docu-

mentary evidence in Unicode proposals. The SAT Committee’s Unicode workflow became markedly more efficient after adopting IIIF (see figure 1).

The *Reiwa Daizōkyō* Compilation System

IIIF is likewise central to compiling the *Reiwa Daizōkyō*. In the collaborative interface for correcting OCR output (see figure 2), editors compare transcriptions and images line-by-line across the Song, Yuan, Ming, and Goryeo canons, as well as the Imperial Household Agency Song edition. Because the holding institutions—Zōjō-ji Temple (Song, Yuan, Goryeo), the University of Tokyo General Library (Ming), and the Keio Institute of Oriental Classics (Imperial Household Agency Song)—all publish IIIF-compliant images, a single program can reliably handle each repository’s images and display the corresponding line. Additional Buddhist canons can be incorporated seamlessly so long as they are IIIF-enabled.

In summary, the project’s planning has depended on meticulous groundwork at the infrastructural layer, such as improving related international standards and promoting common technologies. Among further essentials is compliance with the TEI Guidelines, which we will discuss on another occasion.

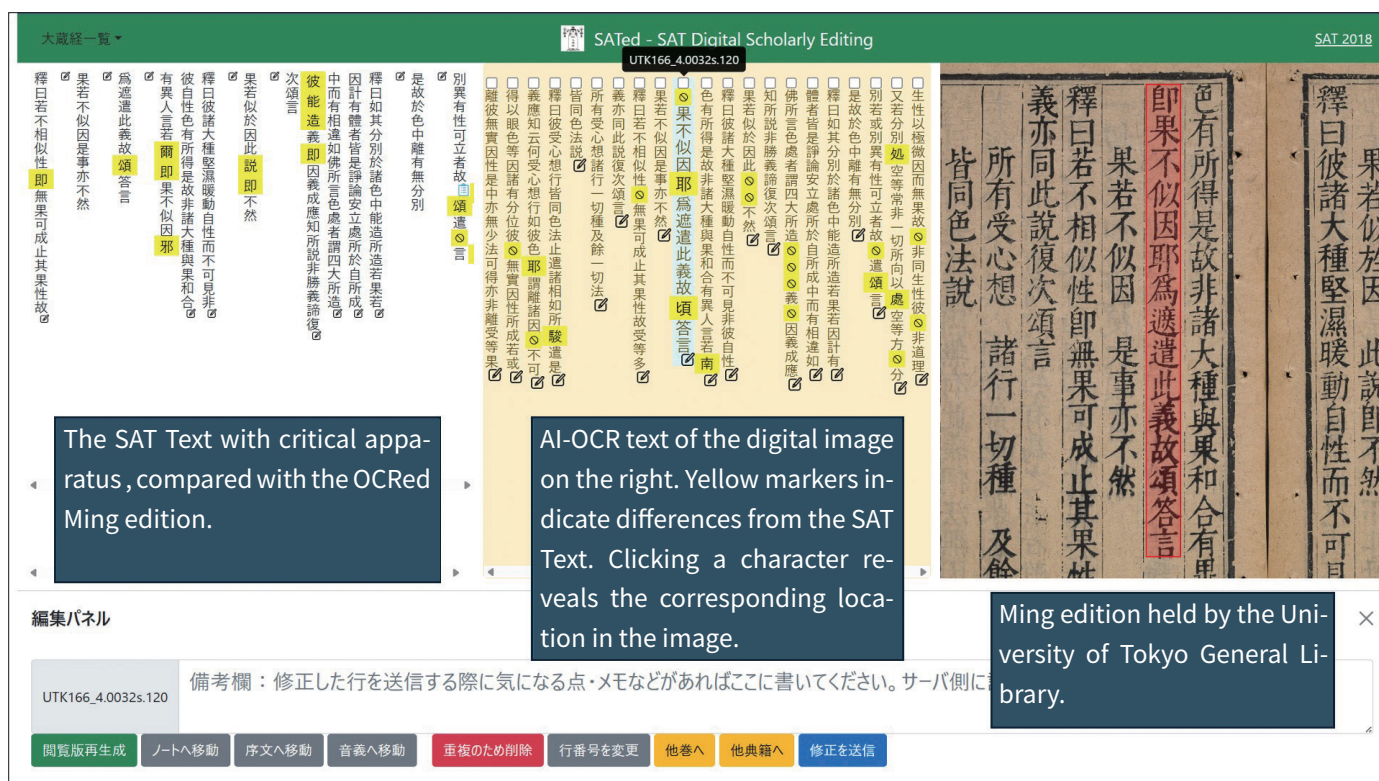


Figure 2. Line-by-line comparison of text and image

3 Report on the Kick-Off Symposium

● **Dr. Hyunjin Park** (Researcher, Department of Buddhist Textual Studies, International Institute for Digital Humanities)

The project's kick-off symposium was held in hybrid format at Musashino University's Ariake Campus on Saturday–Sunday, June 14–15, 2025. It provided a venue to share the project vision with collaborators and to announce the full-scale launch; the atmosphere was notably energetic.

Day 1 opened with remarks by principal investigator Prof. Masahiro Shimoda. As we approach the centennial of the *Taishō Shinshū Daizōkyō* compilation, he stressed that the aim goes beyond mere digitization to ask how a durable “foundation of knowledge” should be constructed in an era predicated on AI and digital technologies—presenting a next-generation model for humanities research. As a core instantiation of that model, he reviewed the history and societal support behind the SAT Committee (https://21dzk.l.u-tokyo.ac.jp/SAT/index_en.html), demonstrating how the *Reiwa Daizōkyō* builds upon and reorganizes the tradition of Tripitaka compilation on a digital foundation.

On day 2, co-investigator Prof. Kiyonori Nagasaki presented the core of the project: high-accuracy text digitization of Buddhist scriptures; TEI-conformant structural markup; and linkage of images and text using IIIF. He

introduced a text-comparison interface for proofreading that highlights differences between AI-OCR output from rare woodblock canons—such as Zōjō-ji Temple's *Sandaizō* (Three Editions of Buddhist Sacred Canons)—and the existing SAT Text, and he went over a staged workflow for progressively adding structure.

A major feature of the symposium was the active, room-wide discussion. Participants exchanged constructive views on task allocation, cultivating early-career researchers, and publication models aligned with international standards. Through these exchanges, a shared understanding emerged: this is not merely a data-building exercise, but an effort to form a sustainable knowledge base, cultivate early-career researchers, and orient the humanities toward the future. The meeting clarified the project's trajectory and was a crucial step toward its full-scale rollout.



4 From Public Academic Events: The Forefront of DH

● **Prof. Kiyonori Nagasaki**

To promote broad discussion—domestically and internationally—of components related to compiling the *Reiwa Daizōkyō*, and to benefit the humanities beyond Buddhist studies, we have organized open academic events on topics in the field and the digital humanities (DH), generally with simultaneous online streaming for public accessibility. Brief reports follow.

April 11, 2025 (Friday) [Supported by our project] International DH Workshop: “New Developments in Digital Text Research”



With keynote speaker Prof. Federico Ponzola (University of Groningen), talks were also given by principal investigator Prof. Masahiro Shimoda, co-investigator Prof. Kiyonori Nagasaki, and Dr. Ayano Takeuchi, who works on text analysis at the National Institute of Japanese Literature. Prof. Ponzola leads the digital literary research project Graphs and Ontologies for Literary Evolution Models (GOLEM; <https://golemlab.eu/>) funded by the EU's European Research Council (ERC), and his talk focused on structuring based on graph models and ontologies in the GOLEM project, which works on the analysis of literary works through reading experiences. The initiative's applicability to Buddhist textual research was evident, and a lively discussion followed.

May 17, 2025 (Saturday) [Jointly held] 138th Meeting of the IPSJ Computer and Humanities Special Interest Group

A regular meeting of the Computer and Humanities Special Interest Group (SIG) of Information Processing Society of Japan (IPSJ) that has continued since 1989 (<https://www.jinmoncom.jp/>). Held at Keio University's Mita Campus. Each May, the SIG runs a student session for undergraduates and graduate students; alongside general papers, students from around the country gathered for a poster session. Eighteen presentations were given in total, and the meeting was a success. For this project, Prof. Nagasaki delivered a special lecture, “Digital Tripitaka Research in the Computer and Humanities Special Interest Group: Past and Future,” summarizing the process from computer and humanities-related research to the present project in this SIG and outlining future prospects.

June 3, 2025 (Tuesday) [Jointly held] DH International Symposium: “New Frontiers in Digital Images and Texts: Recent Advances in Automatic Character Recognition and Their Applications”



With rapid advances in OCR—especially for handwriting—the relationship between image and text is reaching a new stage. This was a valuable opportunity for researchers and developers to gather and discuss the current situation and future developments.

Four researchers, including Prof. Peter Stokes (Université PSL; lead of the free software eScriptorium [<https://gitlab.com/scripta/escriptorium>] handwriting OCR project), together with Toru Aoike and others who develop the National Diet Library's NDLkotenOCR adopted in this project, presented current issues and case studies on the new relations between text and image. From this project, co-investigator Prof. Nagasaki reported on recent *kuzushiji* (cursive)-capable OCR in Japan and positioned the *Reiwa Daizōkyō* within that landscape. Hosted at Keio University's Mita Campus with simultaneous interpretation, the event drew over 100 registrations (including online) and was very well received.

June 13, 2025 (Friday) [Jointly held] International DH Workshop: “The Cutting Edge of AI Translation for the Buddhist Canon: Dedicated LLMs and Career Paths for Humanities Researchers”



Dr. Sebastian Nehrlich (UC Berkeley), a leader in the application of AI in Buddhist studies, introduced the latest developments from the Dharma mitra project (<https://dharmamitra.org/ja/>), demonstrating how rapidly improving generative AI is enabling not only higher-quality automatic translation but also high-accuracy OCR for canonical images. He also outlined his trajectory from conventional Buddhist studies to digital Buddhist research. Participants came away with a concrete sense that AI is finally becoming usable across a widening range of research contexts.

To keep up with the latest happenings in the digital humanities, see the *Digital Humanities Monthly*.

Newsletter “The New Tripitaka and Next-Generation Humanities” No. 1 **Publication date:** December 20, 2025 **Person responsible for editing and publishing:** Principal Investigator Prof. Masahiro Shimoda **Publisher:** Office of the “Compilation of the New Tripitaka as a Digital Research Infrastructure” Project **Website:** <https://reiwadzk.dhii.asia/> **Published bimonthly** **Research project:** “Compilation of the New Tripitaka as a Digital Research Infrastructure: A Model for Constructing a Research Infrastructure for Next-Generation Humanities” **Project number:** 25H00001