Data Mining and Text Reuse in the Collections of Japanese Buddhist Texts

Fiesole Retreat 26 — Session 2: Collection as Data

Gaetan Rappo — Professor

Faculty of Culture and Information Science

Dōshisha University, Kyoto

Tübingen, April 14, 2026

同志社大学
文化情報学部
〒610-0394 京田辺市多々羅都谷1-3
TEL：0774-65-7610　FAX：0774-65-7618
E-mail：jt-bnkjm@mail.doshisha.ac.jp
https://www.cis.doshisha.ac.jp

The Taishō Tripiṭaka

The Taishō shinshū daizōkyō (1924–34) is the standard modern edition of the East Asian Buddhist canon: 2,997 texts in 85 volumes, primarily in Classical Chinese.

What it contains

Indian texts translated into Chinese (2nd–11th c.)
Chinese commentaries and treatises
Korean commentaries, written in Chinese
Japanese commentaries (vols 56–84), written in Chinese
Historical catalogs and biographies

Why it matters

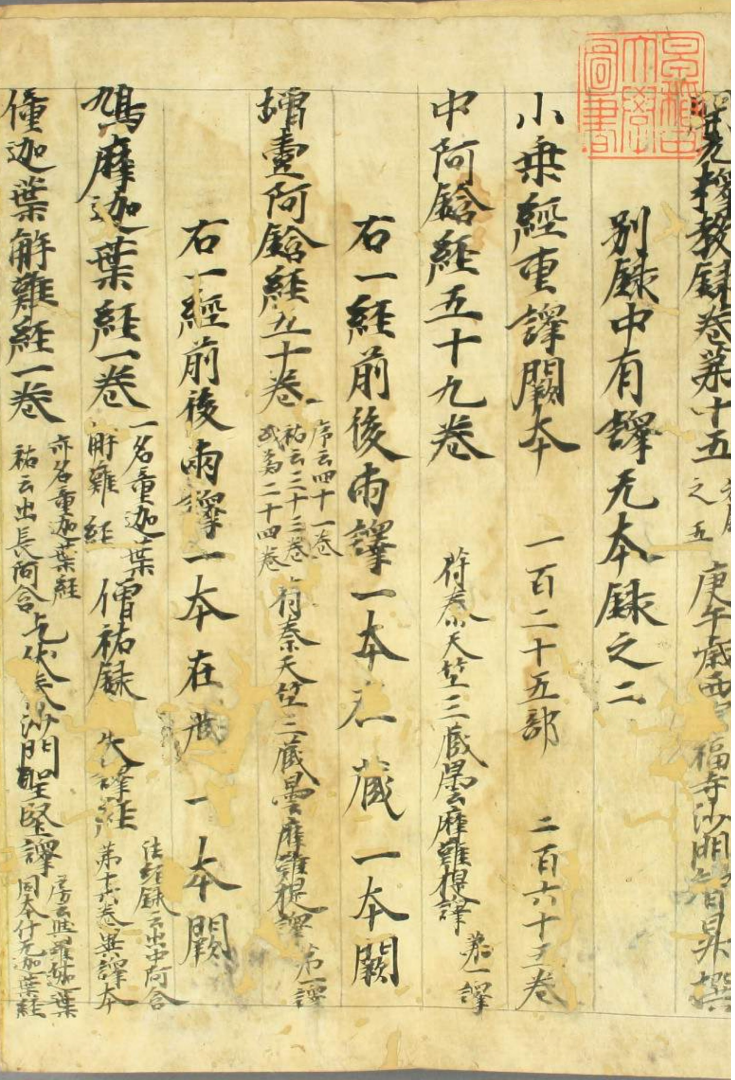
Standard citation system used worldwide
Fully digitized: SAT (U. of Tokyo) and CBETA (Taipei) provide free online access to the full text
Basis for most digital Buddhist text projects
Our starting point: the SAT database

2,997
texts

85
volumes

~73M
characters

2nd–18th c.
date range



From Buddhist catalogues to data

East Asian Buddhist catalogs record title, author, date, school, and classification for thousands of texts — structured metadata, created centuries before databases.

Extracting this data turns printed reference works into searchable, structured metadata for enriching digital text collections.

The *Kaiyuan shijiao lu* 『開元釈教録』 (Zhisheng 智昇, 700-740) 730 CE): the foundational catalog of the Chinese Buddhist canon, listing 5,048 fascicles. Modern reference works inherited and systematized this bibliographic tradition.

5,048 fascicles of Buddhist texts cataloged

Preserved in Japanese temples libraries / University libraries

https://archive.wul.waseda.ac.jp/kosho/ha04/ha04_02725/ha04_02725.pdf

Beyond the Taishō: Japanese Collections

The Taishō covers the canonical corpus. But much of Japanese Buddhist scholarship — especially Shingon (esoteric) ritual texts and oral transmission records — was published in separate collections.

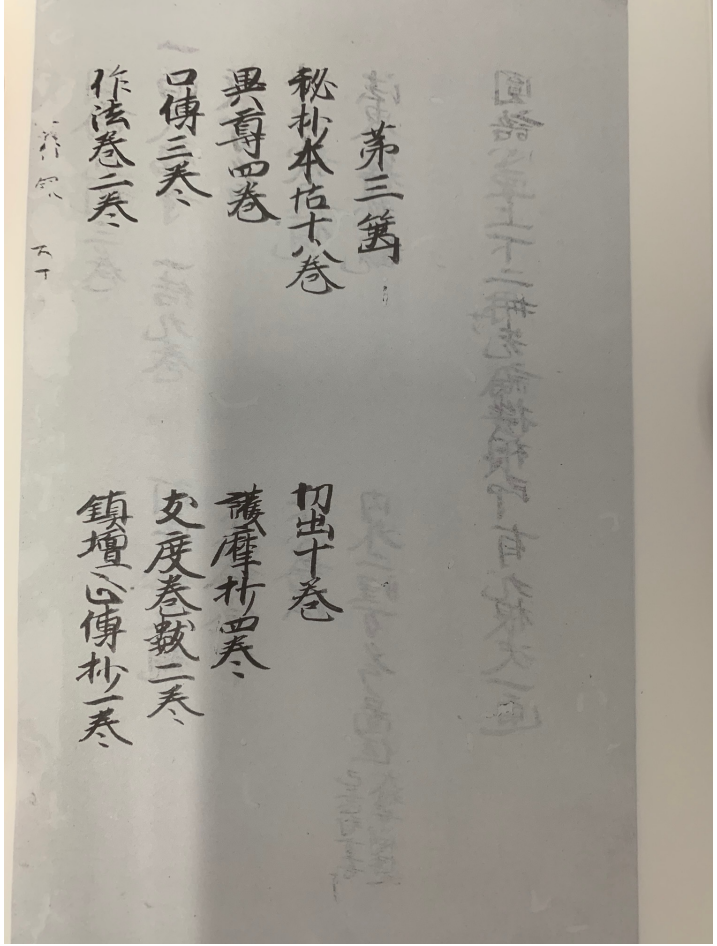
Personal manuscript editions	Texts established from temple manuscripts (Daigoji, Shinpukuji, Kōyasan)
Kōbō Daishi zenshū (KDZ)	Collected works of Kūkai (774–835). Meiji edition, public domain
Kōgyō Daishi zenshū (KGZ)	Collected works of Kakuban (1095–1143). Meiji edition, public domain
Dainihon bukkyō zensho (DBZ)	150-volume Japanese Buddhist canon (1912–22, public domain)
Nihon daizōkyō (NDZ)	100-volume Japanese canon supplement (1914–21, public domain)

several collections, 8th–19th centuries. Thousands of texts — but no unified metadata, no cross-collection search, no structured data.

From Catalogs to Corpus Japanese Reference Works as Data Sources

In addition to such collections, modern Japanese scholarship compiled comprehensive reference works.

- Mochizuki bukkyō daijiten 望月仏教大辞典 (11,000 entries, original ed. 1932–36, out of copyright)



Catalog of the Shinpukuji library, 14th century

Unifying and enriching disparate sources

Multiple sources, multiple formats

No unified format in the collections.

Missing metadata

Many texts list no author, no date, no classification. SAT provides titles and translators — but not dynasty, period, or school.

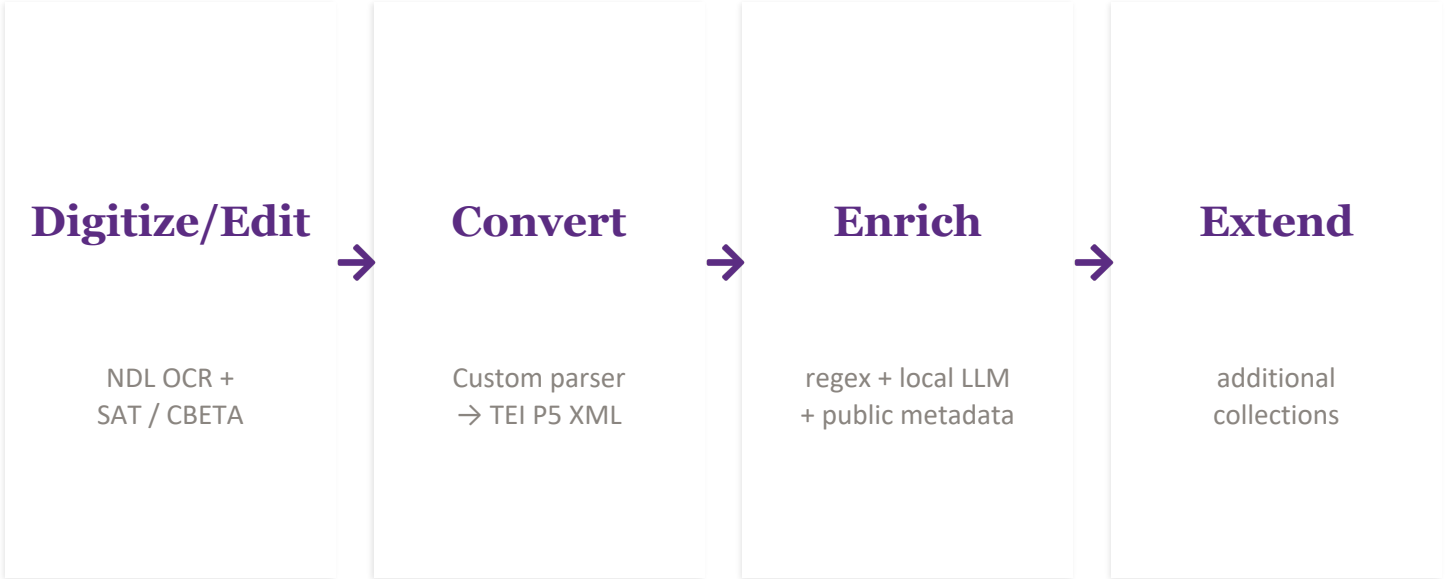
Reference collections locked in print

Reference works like the Mochizuki hold structured metadata — authors, dates, descriptions, classifications, cross-references — but none are searchable or linkable to digital text collections.

Collections cannot talk to each other

No unified digital infrastructure exists in Japan for cross-collection search or metadata enrichment. This forced the development of a local, personal pipeline.

The Pipeline: Collections → Data



Collections become data · Reference works enrich text collections · Cross-collection search becomes possible

Step 1: Digitization

SAT Daizōkyō Text Database

2,997 Taishō texts in custom plain-text format.
Metadata (title, translator, section) from SAT's publicly available database, combined with CBETA's open metadata.
Three encoding variants discovered during conversion.
One text (T1859) supplemented from CBETA.

Non-Taishō collections

Plain text files from printed editions. Using out of copyright editions. Public domain editions accessed through the National Diet Library Digital Collections (NDL) and Hathitrust.
OCR performed using the National Diet Library's NDL OCR engine (open-source, trained on pre-modern Japanese print).

Mochizuki 望月仏教大辞典

10 vols, ~11,000 entries. Original edition (1932–36), author d. 1948 — public domain. OCR'd to structured text.
Covers author, date, classification, descriptions, canonical references, and cross-references.
Structured fields extracted via regex + local LLM (Ollama / Qwen3) → JSON.

From Manuscript to Edition

Before any computational analysis, the texts absent from collections must be established from manuscripts. I produce my own critical editions from ritual and doctrinal manuscripts held in Japanese temples and libraries.

Written in Sino-Japanese (kanbun)

Not classical Chinese, but Japanese texts written using an adapted Chinese script. The grammar is Japanese, the characters are Chinese — a hybrid that standard NLP tools cannot parse.

Stored in temples across Japan

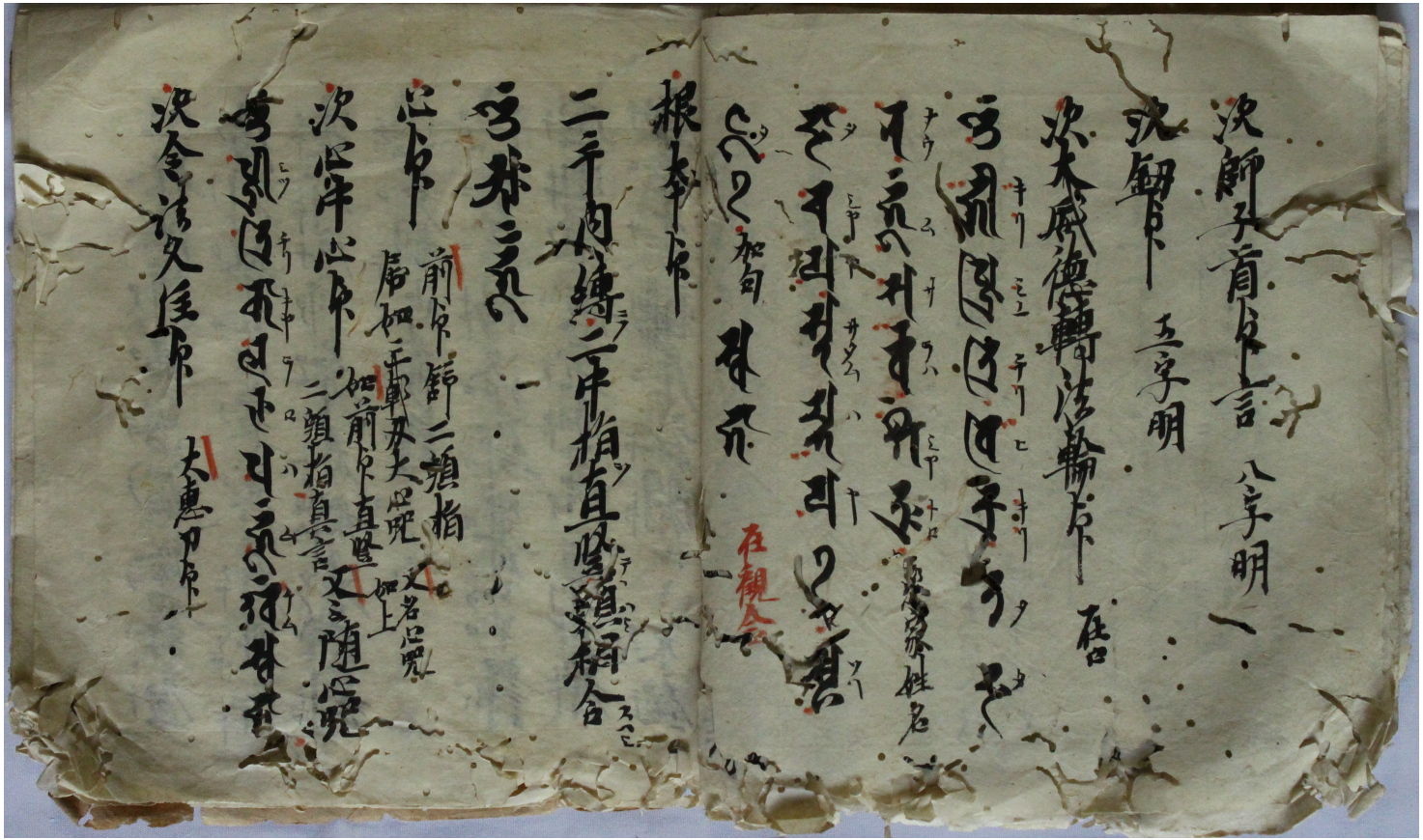
Often with restricted access. Not digitized, not cataloged in standard databases. Visiting the manuscript may require institutional introductions and travel.

Dense with specialized vocabulary

Ritual procedures, doctrinal arguments, and lineage records using Tendai and Shingon technical terminology. Reading them requires both philological training and domain expertise.

Multiple annotation layers

Reading marks (kaeriten), siddham characters (indic script), phonetic glosses (furigana, okurigana), red ink emphasis (shuten), corrections (misekechi), and marginalia (uragaki) — all carrying editorial information that published editions typically strip away.



Gyakuto taiji gomashidai (Homa ritual for the subjugation of rebels). 14th century
With permission from the Yoshino Nyoirinji temple

Step 2: TEI XML Conversion

TEI (Text Encoding Initiative) P5 is the standard XML schema for scholarly text encoding. Its structured headers make it possible to embed metadata from reference collections directly alongside the text — so that enrichment data travels with the text through every stage of the pipeline, including text reuse detection.

What the converter handles

- Page–column–line citation IDs
- Inline apparatus notes from Taishō editors
- Verse passages with line structure
- Siddham (bonji) script characters
- Gaiji (non-standard characters)
- Three SAT format variants

The TEI header captures

- Title (original script + romanization)
- Author / translator / editor
- Taishō volume + text number
- Section classification
- Source description and provenance
- Encoding documentation

Step 3: Metadata Enrichment

Many TEI headers remain incomplete: authors as ‘Unknown’, no dates, no classification. I combined SAT and CBETA public metadata, open-access databases (DDB, INBUDS), and data extracted via regex + local LLM from out-of-copyright reference works to fill these gaps. Matched to TEI files by title and by the LLM. All additions tagged resp="#enrichment" — original metadata preserved.

Reference Collections

Mochizuki (11K) + SAT + DDB
+ catalogs (7K) + own research
All public domain or open access

→

Local LLM

Ollama + Qwen3 (32B) locally
Also possible: Claude / Gemini API
Prompt engineering + post-processing

→

JSON Metadata

Title · Author
Dynasty · Date
Classification, Description

→

TEI Files

Title matching (Python)
Fill missing fields
resp="#enrichment"

Fields added to TEI headers:

- Authors and translators — replacing ‘Unknown’ placeholders
- Dynasty or period (e.g. Tang, Song, Heian, Kamakura, Edo)
- Composition or translation dates
- Descriptions and cross-references from reference works
- Text classification (sūtra, vinaya, treatise, commentary)

Enrichment: Preliminary Results

Mochizuki (10,898 entries, public domain) + catalogs from public domain editions (6,875) + monks metadata from own research (314) + SAT/CBETA public metadata. ~3500 TEI files enriched so far.

Field	Coverage	Notes
Title	100%	Extracted from entry headwords
Classification	87%	Sūtra / vinaya / treatise / commentary
Extant or lost	100%	Catalog survival status
Cross-references	76%	Catalog sources and related works
Dates	56%	Era-name format (CE conversion planned)
Authors	66%	Many texts genuinely anonymous
Description	28%	Only entries with substantive content
Dynasty	35%	Often inferrable from author dates

From Catalog to Enriched TEI: An Example

Catalog entry (before)

新編諸宗教藏總錄 (*Shinpen shoshū kyōzō sōroku*) / 義天 (Giten). One line in a catalog page — title and author, nothing else.

Author + 6 alt titles (after)

Ŭich'ŏn 義天 (1055–1101), Koryŏ prince-monk. Birth/death from public databases. Alt titles include: 義天目錄 (*Giten mokuroku*), 新編高麗藏目錄 (*Shinpen koraizō mokuroku*), 諸宗教藏總錄 (*Shoshū kyōzō sōroku*).

Historical context (after)

Composed 1090. Dynasty: Koryŏ 高麗. Chinese teachers: Cibian 慈辨, Yuanzhao 元照, Jingyuan 淨源, Zongben 宗本. Copyist: 明空 (1176, Ninnaji 仁和寺).

Description + cross-references (after)

Description (from Mochizuki + public metadata): A catalog of 1,086 commentaries collected from China and Korea over 20 years by Ŭich'ŏn, listing titles and authors across three volumes (sūtra, vinaya, and śāstra commentaries). Related text: *Kaiyuan shijiao lu* 開元釋教錄 (730 CE). Cross-refs to Taishō vol. 55 and Zoku gunsho ruijū vol. 28.

Application: Text Reuse in Monkan's Works

A concrete use of this enriched corpus: tracing how a medieval author borrows, adapts, and reframes material from other texts.

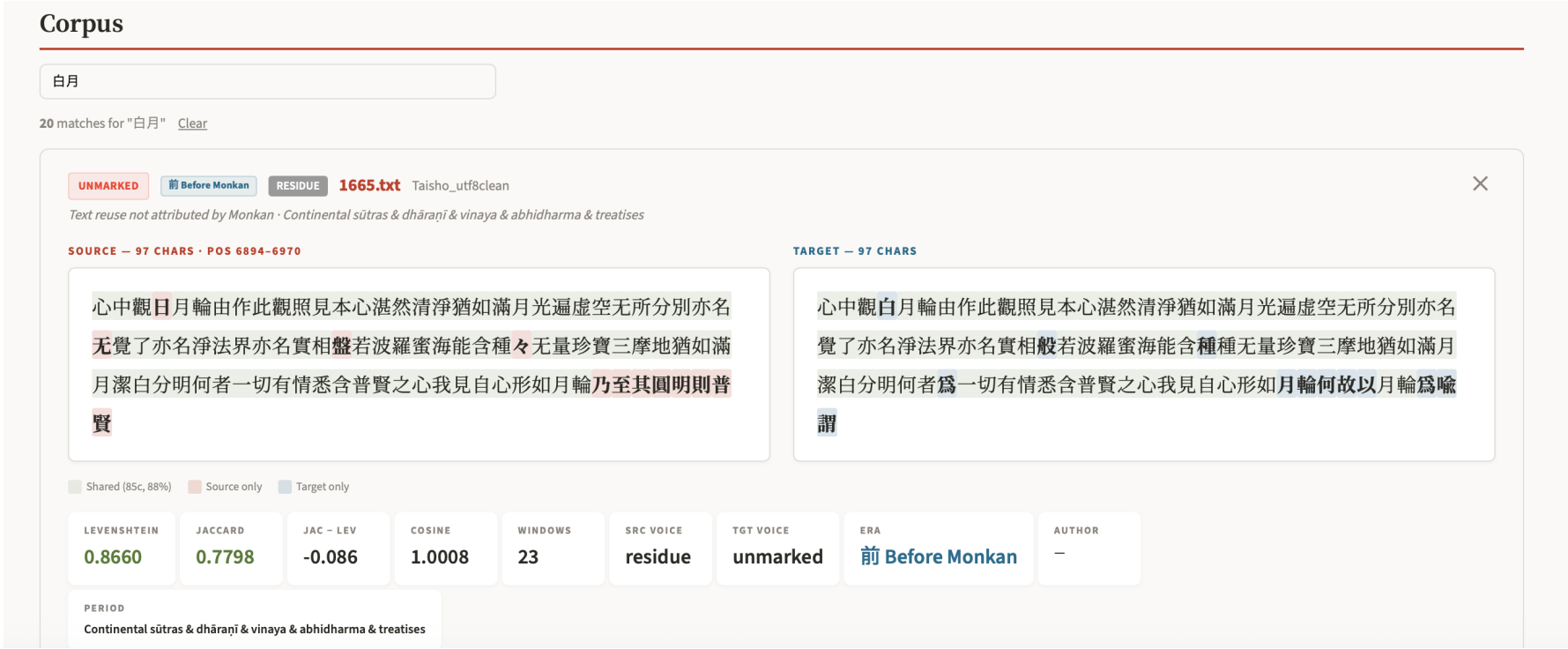
Monkan 文観 (1278–1357), a Shingon monk at Daigoji. Like most medieval Buddhist authors, his texts cite earlier sources — canonical sūtras, founder writings, commentaries. I selected 14 of his texts (~168K characters), including several of my own editions, to test the pipeline.

Searching his 14 texts (~168K characters) against ~1,100 source texts from the Taishō and other collections.

Text Reuse Detection

Tokenize	Embed	Search	Score
Trie-based compound merging (DDB + domain terms, 15K)	Yasuoka Classical Chinese BERT (1024d) + SWEMhier pooling	FAISS nearest-neighbor ~1,100 source texts (Taishō + additional collections)	Levenshtein + Jaccard similarity on merged passages

Window: 12 chars, stride 6. Default filtering at 0.85 Levenshtein and 0.3 Jaccard.



Matches with variations in word order or characters possible.

Temporal Analysis of Text Reuse

With author dates from enrichment, every text reuse match can be classified by temporal direction relative to the author Monkan 文観 (1278–1357).

Before Monkan

Borrowing from canonical texts, founder writings, and earlier commentaries.

Contemporary

Shared passages with monks active in the same period. Evidence for direct exchange, common training, or shared source materials.

After Monkan

Matches with later texts suggest later interpolation, shared tradition, or reverse attribution. Only detectable with dated metadata. Note: some matches reflect independent use of a shared canonical source rather than direct borrowing.

Why Metadata Matters for Text Reuse

Without enrichment metadata, a text reuse match is just a string overlap. With it, the same match becomes evidence for cross-sectarian borrowing, doctrinal fabrication, or lineage construction.

Without metadata

- Two texts share a passage
- Direction of borrowing unknown
- School affiliation unknown
- No way to filter by period or tradition

With metadata

- Author dates reveal temporal relationships and sometimes directionality
- School/lineage labels expose cross-sectarian absorption
- Composition dates enable chronological filtering
- Descriptions confirm whether a match is relevant

The enrichment pipeline transforms text reuse from pattern matching into historical argument — and opens the door to classification by school, lineage, and period.

School and lineage metadata

Each Buddhist school historically created its own text collections — Tendai, Zen, Ritsu, Jōdo each have their own zensho. Our corpus already includes several of these. This is still an early build, but the metadata infrastructure is in place to classify text reuse by school and tradition, not just by date.

What school metadata enables

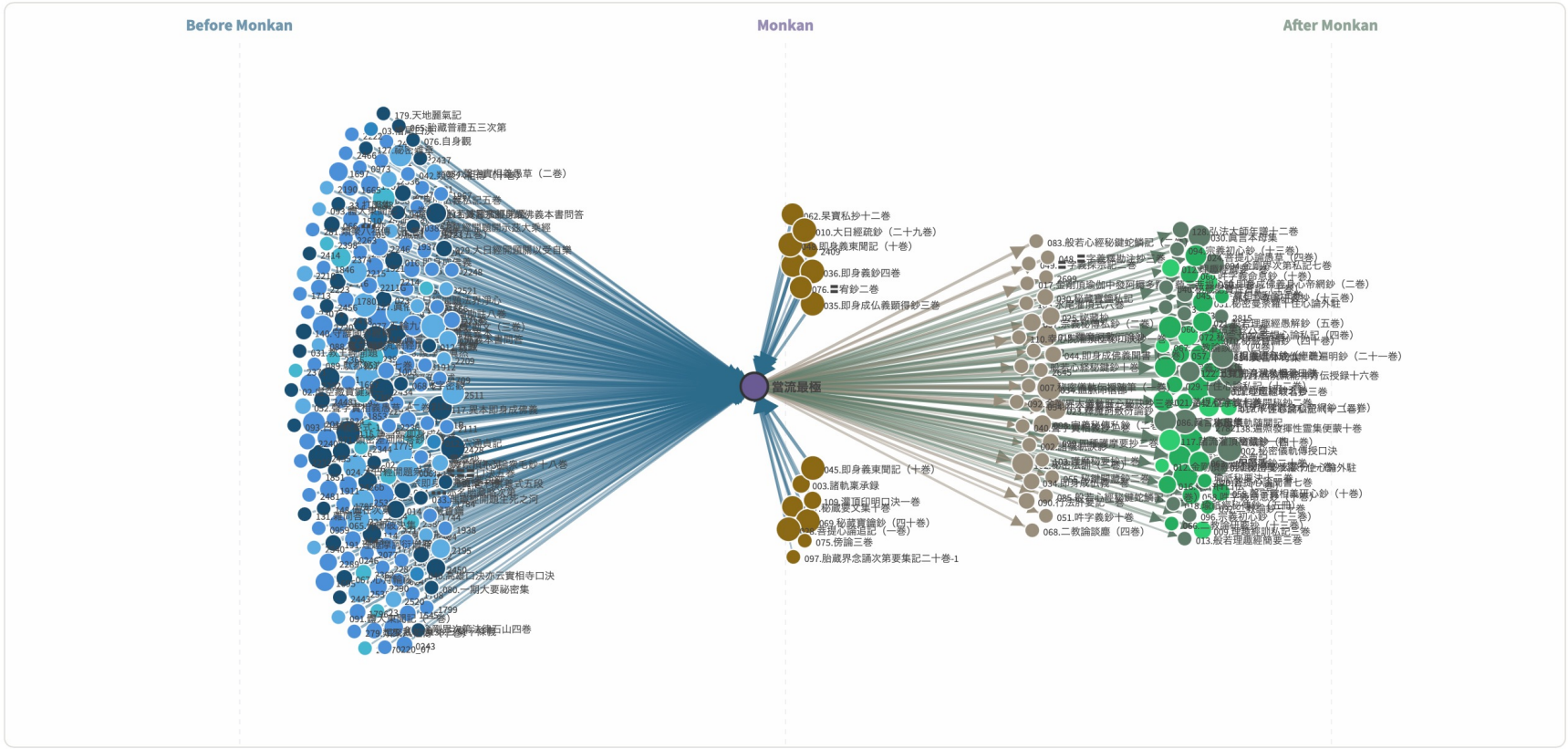
- Classify each text reuse match by the school of the source text
- Trace how ideas move across school boundaries over centuries
- Distinguish school-internal transmission from cross-sectarian borrowing
- Use school metadata to contextualize any text analysis, not just text reuse

Next steps

- Public domain collections from multiple schools (Tendai, Zen, Shingon, etc.)
- More precise school and lineage classification from enrichment metadata
- Text reuse detection across collection boundaries
- A long-term perspective: 9th–16th century, not just one author

Reuse Network

221 before→Monkan · 58 Monkan→after · 32 Monkan→uncertain · 312 texts · 311 edges



From Catalogs to Computation

None of this would be possible without centuries of cataloguing by Buddhist institutions, academic libraries, the systematic compilation work of modern Japanese scholars, and the digitization efforts of projects like SAT and CBETA.

Each layer of the pipeline rests on prior library work:

- Temple catalogs preserved the texts and recorded their authors and lineages
- Modern reference works (Mochizuki dictionary) systematized this knowledge into searchable metadata
- SAT and CBETA made the full texts freely available online
- TEI encoding carries this metadata into every stage of analysis
- Text reuse detection turns collections into evidence for historical argument

Without collections, there is no corpus. Without catalogs, there is no systematic metadata. Without metadata, text reuse is just pattern matching. The pipeline is built on library infrastructure.