

Bunkankun: Digital Text as a Collection of Distributed Layers

(Late breaking)

Christian Wittern

Kyoto University Institute for Research in Humanities

Balisage 2026

Digital Text as Distributed Layers

- 1 Background
- 2 A new look at premodern Chinese text
- 3 Text as layers
- 4 Three months into the project...

Motivation

- I am responsible for two large digital collections of premodern Chinese texts:
 - Kanseki Repository
 - Thesaurus Linguae Sericae (TLS/HXWD)
- The respective data format is at the opposing ends of a scale measuring verbosity
 - ... they should not have been separate from the start,
 - ... for years, I have been trying to bring them together ...
- On top of that, interest in
 - Corpus wide research in intertextuality and parallel passages
 - Corpus aware deep reading, translation, linguistic description etc.

Kanseki Repository

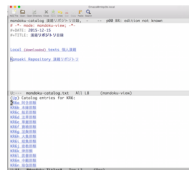
- Close to 10000 textual items, spanning more than 2000 years
- Minimal (invisible?) markup, based on Emacs org-mode
- Most texts do **not** have any punctuation
- Interpretative and documentary editions
 - realized as branches in git repositories
- Developed since 2011, first release 2015
- No central authority required, released on GitHub

Bird's eye view of the Kanseki Repository

Kanseki Repository



Stack room @kanripo / Github

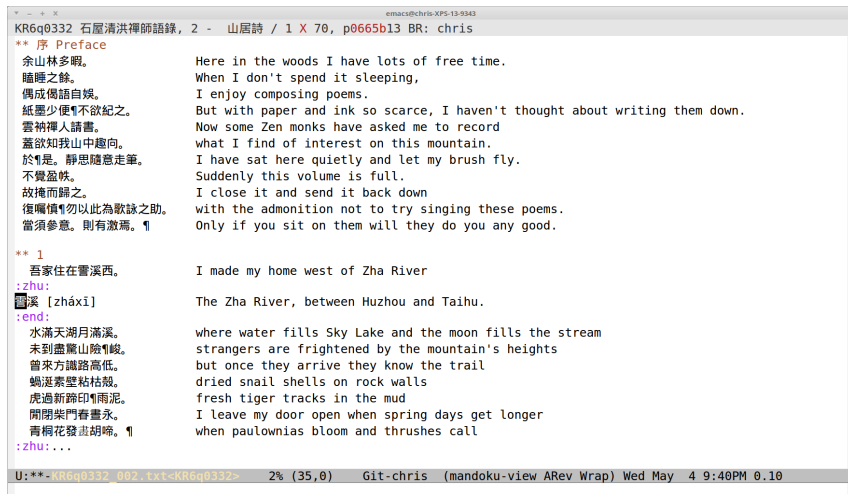


Study Mandoku / Emacs



Reading room www.kanripo.org
Website

Source text, translation and annotation



```
emacsr@chris-XP5-13-9343
KR6q0332 石屋清洪禪師語錄, 2 - 山居詩 / 1 X 70, p0665b13 BR: chris
** 序 Preface
余山林多暇。          Here in the woods I have lots of free time.
瞌睡之餘。            When I don't spend it sleeping,
偶成偈語自娛。        I enjoy composing poems.
紙墨少便不欲紀之。    But with paper and ink so scarce, I haven't thought about writing them down.
雲衲禪人請書。        Now some Zen monks have asked me to record
蓋欲知我山中趣向。    what I find of interest on this mountain.
於此是。靜思隨意走筆。 I have sat here quietly and let my brush fly.
不覺盈帙。            Suddenly this volume is full.
故掩而歸之。          I close it and send it back down
復囑慎勿以此為歌詠之助。 with the admonition not to try singing these poems.
當須參意。則有激焉。  Only if you sit on them will they do you any good.

** 1
吾家住在雲溪西。      I made my home west of Zha River
:zhu:
雲溪 [zháxī]          The Zha River, between Huzhou and Taihu.
:end:
水滿天湖月滿溪。      where water fills Sky Lake and the moon fills the stream
未到盡驚山險峻。      strangers are frightened by the mountain's heights
曾來方識路高低。      but once they arrive they know the trail
蝸涎素壁粘枯殼。      dried snail shells on rock walls
虎過新蹄印雨泥。      fresh tiger tracks in the mud
閑閉柴門春晝永。      I leave my door open when spring days get longer
青桐花發盡胡啼。  when paulownias bloom and thrushes call
:zhu:...
```

U:**-KR6q0332_002.txt<KR6q0332> 2% (35,0) Git-chris (mandoku-view ARev Wrap) Wed May 4 9:40PM 0.10

Figure: The beginning of *juan 2*, with translation

TLS/HXWD

- a (smaller) text database, but with translations, phonetic glosses, and detailed linguistic, lexicographic and taxonomic annotations
- Started in the mid 90s as Filemaker DB, since 2016 transformed to TEI/XML
- Collaborative research platform used for original research and in classrooms
- Striving, but failing, to support the whole range of text-related research activities

Text representation in the TLS

- The texts are encoded according to the recommendations of the Text Encoding Initiative
- Within the TLS, TEI is used to describe the **textual hierarchy** of books, chapters, sections (including the respective headings) and paragraphs
- Within a paragraph, **<seg>** is used to represent a **phrase**, which is displayed as **line** in the Web interface
 - What is called phrase here is roughly speaking what appears between punctuation marks
 - This design is a legacy carried forward from the Filemaker DB through the conversion to XML
- All the other features mentioned are represented **externally** to the text and linked via the **<seg>** element, additionally using character offsets if necessary

First chapter of Laozi Daodejing in TLS

```
<div>
  <head xml:id="KR5c0057_tls_001-1a.3-h-h" n="1">
    <seg xml:id="KR5c0057_tls_001-1a.3-h">第一章</seg>
  </head>
  <p/>
  <p>
    <seg xml:id="KR5c0057_tls_001-1a.3">道可道<c n="," /></seg>
    <seg xml:id="KR5c0057_tls_001-1a.4">非恒道<c n="。" /></seg>
    <seg xml:id="KR5c0057_tls_001-1a.5">名可名<c n="," /></seg>
    <seg xml:id="KR5c0057_tls_001-1a.6">非恒名<c n="。" /></seg>
    <seg xml:id="KR5c0057_tls_001-1a.7" resp="#CH"
      modified="2023-08-10T00:55:40.863+09:00">無名，天地之始，</seg>
    <seg xml:id="KR5c0057_tls_001-1a.8" resp="#CH"
      modified="2023-08-10T00:55:47.379+09:00">有名，萬物之母，</seg>
    <seg xml:id="KR5c0057_tls_001-1a.9">故恒無欲<c n="," />以觀其妙<c n="。" /></seg>
    <seg xml:id="KR5c0057_tls_001-1a.10">恒有欲<c n="," />以觀其徼<c n="。" /></seg>
    <seg xml:id="KR5c0057_tls_001-1a.11">此兩者同出而異名<c n="。" /></seg>
    <seg xml:id="KR5c0057_tls_001-1a.12">同謂之玄<c n="。" /></seg>
    <seg xml:id="KR5c0057_tls_001-1a.13">玄之又玄<c n="," /></seg>
    <seg xml:id="KR5c0057_tls_001-1a.14">眾妙之門<c n="。" /></seg>
  </p>
</div>
```

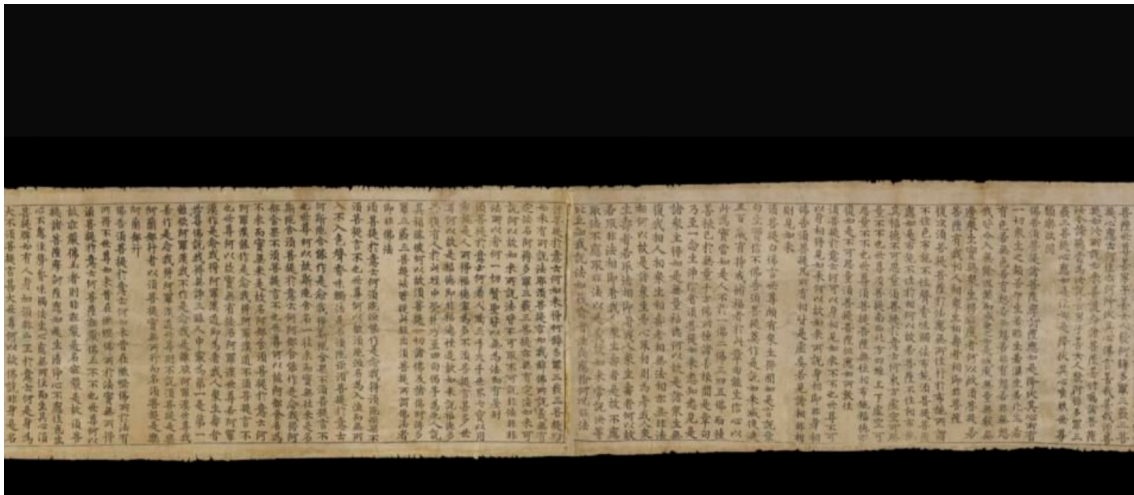
Digital Text as Distributed Layers

- 1 Background
- 2 A new look at premodern Chinese text
- 3 Text as layers
- 4 Three months into the project...

What is the most basic form of a text?

- A (long!) string of characters
 - (in practice) arranged in lines, due to the limitations of the medium
- In the Chinese case, since the invention of paper:
 - Sheets of papers glued together to form longer scrolls:
 - Scrolls become meaningful units of larger texts
- A text then, is a collection (bundle) of scrolls (*juan* in Chinese)

Diamond Sutra of ca. 868



Scrolls with Chinese text in a box



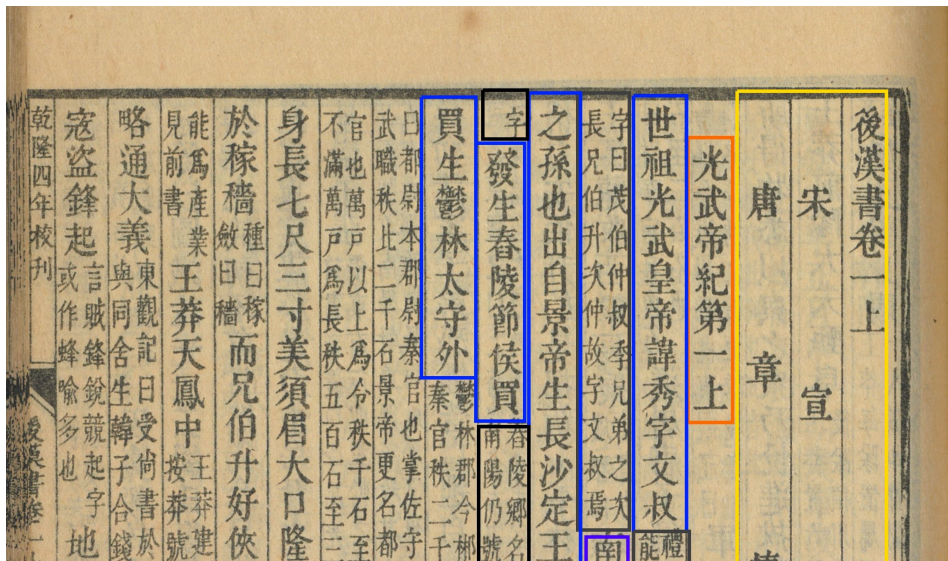
There is more: Text as multiple layers

- A careful look shows: There is more!
- The text – string of characters – of a *juan* can be seen as base layer
 - some textual features are hinted at from layout (indent, space)
 - in practice we might want to distinguish paratextual features:
 - front / body / back
- Other features of a text will be discovered and can link to this base layer
- This is an iterative process of interpretation and notation

Some complications

- Historically, most of the time, most (important) texts have been transmitted **with commentary**
- And later periods added yet new commentaries
- Finally, commentaries have been collected and assembled: many voices in the text
- Others added sound glosses for rare characters or foreign expressions
- All of this will be dealt with using markers **external** to the text string

Beginning of the record of Emperor Guangwu of the Later Han Dynasty



Digital Text as Distributed Layers

- 1 Background
- 2 A new look at premodern Chinese text
- 3 Text as layers
- 4 Three months into the project...

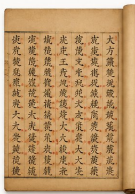
Text as Layers

Separating concerns to preserve, understand, and reuse.

Bunkankun

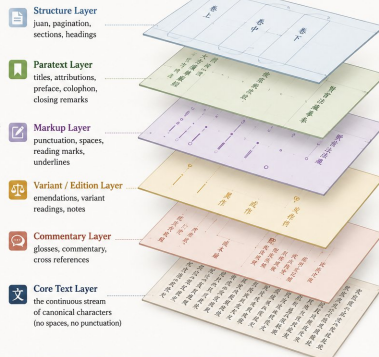
Source: a manuscript

(e.g., a juan from a larger work)



- A** Characters are the core. Everything else is a layer.
- Layers can be added, removed, or replaced independently.
- Different editions become different layer sets over the same core.

The layered model



Rendered: a reading view

(one of many possible views)

大方廣佛華嚴經卷上
賢首法藏奉 詔譯
如是我聞。
一時，佛在王舍城耆闍崛山中，
與大比丘眾千二百五十人入菩薩萬二千人，
皆是已得大菩提者，爾時，世尊食時，著衣持鉢，
入王舍城，乞食已還，
到本處，飯食訖，收衣鉢，洗足已，
敷座而坐。

Clients assemble layers according to user needs: different editions, layouts, languages, and devices.

Same core text + Different layers = Many presentations

Search across layers

Compare editions

Re-use & compose

Preserve faithfully

Figure: Bringing it together

Archival format for text bundles

- All files are serialized in YAML format
- The top level describes the 'bkk surface' (master) edition
- Entry point: **Manifest** file, which enumerates the available files and gives some additional metadata
- **Juan** ('scroll') files: 'text' in 'front', 'body' and 'back' buckets, table of contents, administrative metadata
- 'assets' folder: 'Marker' files, per juan standoff markup referring text points or chunks in the buckets by offset
- 'editions': same structure as the surface edition for the editions mentioned in the manifest
- optionally: mapping for characters not encoded in Unicode, represented using codepoints from the PUA area

Manifest file (excerpts)

canonical_identifier: bkk:krp/KR6c0128/v1

canonical_set:

 identifier: bkk:charset/cjk-v1

 hash: sha256:37995e65f5e6cdc8cfbc980fea50ec86663e3e7de950809f7e37f0037541

metadata:

 title: 般若波羅蜜多心經

 identifiers:

 alt_id:

 - T08n0251

 source:

 repository: cbeta

 path: /home/chris/src/xml-p5/T/T08/T08n0251.xml

 edition:

 short: bkk

editions:

Juan file

```
canonical_identifier: bkk:krp/KR6c0128/t/v1/juan/1
seq: 1
front:
  text: ''
  hash: sha256:0000000000000000000000000000000000000000000000000000000000000000
body:
  text: 唐三藏法師玄奘譯觀自在菩薩行深般若波羅蜜多時照見五蘊皆空度一切苦厄
  hash: sha256:1ce340ae8259ab4494ecdc1198197518dca7488c1ea3571b4fef3e54a2a8
markers:
- {type: 'cbeta:juan-start', offset: 0, content: 般若波羅蜜多心經, id: KR6c0128-t-v1-juan-1-0}
- {type: 'cbeta:juan-end', offset: 268, content: 般若波羅蜜多心經, id: KR6c0128-t-v1-juan-1-1}
metadata:
  title: 般若波羅蜜多心經
  edition:
    short: t
```

Marker file

canonical_identifier: bkk:krp/KR6c0128/bkk/v1/markers/1

seq: 1

markers:

front: []

body:

- {type: line-break, offset: 0, content: '', id: KR6c0128_T_001-0848c04,
- {type: line-break, offset: 0, content: '', id: KR6c0128_T_001-0848c05,
- {type: xml-element, offset: 0, content: '', id: KR6c0128_T_001-bkkxmle
- {type: variant, offset: 0, length: 1, content: 唐, 宋: ''}
- {type: variant, offset: 6, length: 1, content: 樊, 宋: 樊奉詔, 元: 樊奉
- {type: xml-element, offset: 8, content: '', id: KR6c0128_T_001-bkkxmle
- {type: line-break, offset: 8, content: '', id: KR6c0128_T_001-0848c06,
- {type: xml-element, offset: 8, content: '', id: pT08p0848c0601, name: p
- {type: punctuation, offset: 22, content: , , id: KR6c0128_T_001-bkkpn1

Character encoding

- Unicode contains many near-duplicates of East Asian characters
- In BKK, a **canonicalization** process for import and input constrains the characters to a subset of explicitly allowed characters
- Characters outside of this set will be replaced with their canonical equivalents, but markers will retain the original form to be restored for rendering or export on request
- Non encoded characters will be assigned a PUA codepoint and added to the canonical character set

Multiple reading sequences (voices) in the text

- The text as written or printed usually has some visual hints to mark voices in the text, such as spaces or smaller characters
- For BKK, all of these hints have been relegated to the markers
- multi-voiced text thus can not be read without taking the markers into account

0 4 13 17
text = '色不異空調色相虛幻不離空性空不異色調空性遍周不離色相'

markers:

- {type: voice, offset: 0, length: 4, name: root, id: r1}
- {type: voice, offset: 4, length: 9, name: commentary, responds-to: r1}
- {type: voice, offset: 13, length: 4, name: root, id: r2}
- {type: voice, offset: 17, length: 9, name: commentary, responds-to: r2}

(Excurs) Text identifiers

- A classified catalog is shared by all three projects
- Text identifiers are derived from the position in the catalog
 - eg. The Heart (of the Prajñāpāramitā) Sūtra is KR6c0128
 - 'KR6': Buddhist texts, 'KR6c': Prajñāpāramitā (Perfection of Wisdom) literature, '128' sequential number in this section
- References to texts are constructed by adding juan ('seq'), bucket and offset
 - This refers to the first 6 characters of this text in the surface edition
 - 'KR6c0128/1/body@0+6/v1' or shortened to 6c128/1/@0+6 with default values omitted

Recipes

- Recipes bring together layers as requested, define the scope and shape of the output
- Layers need not to be limited to the text or the corpus, a user can make additional data available
- Recipes have different parts and can be split for convenience, separating the collecting (input) from the forming (output)
- The input recipe will use references like the one above to select input, possibly from different sources
- Markers can be also used, for example for selecting only certain voices in a text

Recipe example: Input

```
kind: bkk.recipe-input/v1
for_template: punc-report
target:
  textid: KR5a0009
  juans: {0, 1, 2}
  bucket: body
layout:
  width: 40
sets:
- sigil: gem
  label: Gemma-4-26B-A4B-it-MTP-GGUF
  source: llm-punctuation
  model: Gemma-4-26B-A4B-it-MTP-GGUF
- sigil: lun
  label: gpt-5.6-luna
```

Recipe example: Render

```
template: |
  # Punctuation comparison — {{ input.target.textid }}
  Bucket: `{{ datasets.report.target.bucket }}` · width: {{ datasets.repo

Sets:
{% for s in datasets.report.sets %}
- `{{ s.sigil }}` — {{ s.label }}
{% endfor %}
{% for g in datasets.report.groups %}

## juan {{ "%03d"|format(g.juan_seq) }} @{{ "%06d"|format(g.offset) }}
...

{% for line in g.lines %}
{{ line.sigil }}~^I{{ line.text }}
{% endfor %}
```

Digital Text as Distributed Layers

- 1 Background
- 2 A new look at premodern Chinese text
- 3 Text as layers
- 4 Three months into the project...

Text corpus

- All of the target texts from three separate collections have been converted to bkk format and posted to `github/bkkbooks`, 12500+ repositories and counting
- The conversion proceeds in a number of separate steps, allowing for validation and correction at every stage
- The computability arising from the new format allowed for the remedy of some long known, but so far intractable problems in the Kanseki Repository
- All texts and tools are publicly available on GitHub, but this is work in progress...

Tool chain

- A CLI tool `bkk` has been created for interaction and maintenance of BKK files has been created.
- For all of the basic tasks, sub commands such as `import`, `export`, `index`, `validate`, `serve`, `repair`, `voice`, `recipe`, `chars`, `repo`, `llm` have been created
- This has been used to convert and curate the corpus
- Some of this is exposed through the `serve` command as a REST style API
- A UI for interaction with `serve` has been created, accessible at bunkankun.org

LLM based punctuation

- A pipeline for LLM based punctuation was established.
- Marker files with punctuation from different models have been produced for a few test files
- Rendering to output files, with the potential to off-line batch editing is ready.

First results

Punctuation comparison — KR5a0009

Bucket: ▶body4 · width: 40 chars/line

Sets:

- ▶gem4 — Gemma-4-26B-A4B-it-MTP-GGUF
- ▶lun4 — gpt-5.6-luna
- ▶cla4 — anthropic-claude-haiku-4.5
- ▶ds44 — deepseek-deepseek-v4-flash
- ▶fug4 — fugu

▼ **juan 001 @000000–000040**

gem	序品。爾時，海空智藏見元始天尊，遍遊他土，至極樂國，妙嶽之山，安坐正心，在騫林之下，與諸弟子千三	Plain Text
lun	序品爾時，海空智藏見元始天尊遍遊他土，至極樂國妙嶽之山，安坐正心，在騫林之下，與諸弟子千三	
cla	序品。爾時海空智藏見元始天尊遍遊他土、至極樂國妙嶽之山、安坐正心在騫林之下。與諸弟子千三	
ds4	序品爾時海空智藏見元始天尊遍遊他土至極樂國妙嶽之山，安坐正心在騫林之下，與諸弟子千三	
fug	序品爾時海空智藏見元始天尊遍遊他土，至極樂國妙嶽之山，安坐正心，在騫林之下，與諸弟子千三	

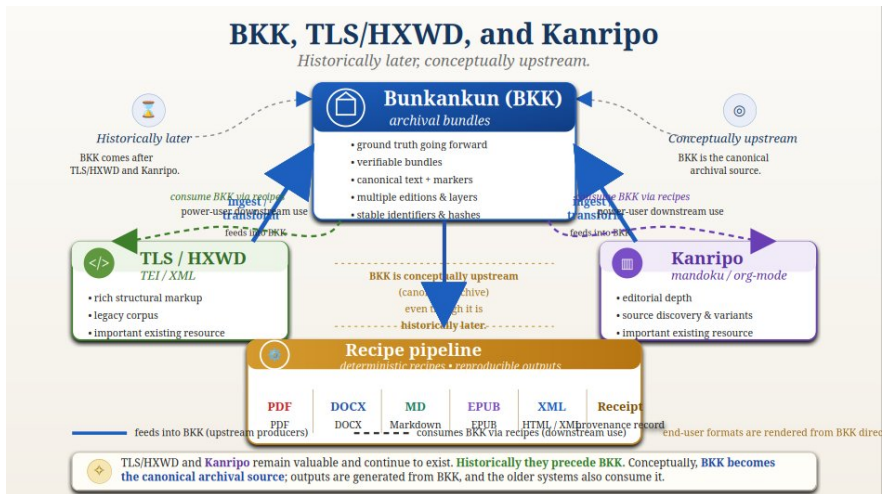


Figure: Possible connections between the three projects

Reflections

- The BKK format proved to be surprisingly efficient and apt for the tasks it was intended for
- There are drawbacks though:
 - it can not be immediately edited in a normal editor
 - it is not based on any standard, so every tool, including an editor has to be created from scratch
- It might serve as an upstream format for Kanripo and TLS/HXWD
- For the time being, it will be used to produce punctuated versions of Kanripo texts

What is a (digital) text, again?

- Part of a larger system, that includes data files, rendering and display
- Focusing solely on the data format, without taking the other parts into account misses the point
- The bkk processing system can support the whole range of interactions that are required for digital text
- Chinese texts as presented here are a very special case and dealing appropriately with them required a look beyond established methods