A first look at Kanripo, a distributed repository for premodern Chinese texts

Christian Wittern

2013-12-10
Overview

▶ Aims of the project
▶ Preparations
▶ Current state
▶ Future plans
Towards a new digital infrastructure

Kanripo aims at

- Allowing the collecting of existing digital transcriptions and facsimile
- Describing them in a way suitable for scholarly work
- Realizing the unit of text, metadata and application interface
- It does not require a central authority
  - everybody can edit, but
  - authority is based on merit
Practice of digital text usage in the field of Chinese Studies

Scholars (especially aspiring young graduate students):

- Collect and exchange plain text versions of relevant texts that float around on the web
- no record about source, license, responsibility for the creation of text and method for digitization

Commercial publishers of large textual collections

- Many errors, no feedback circle
- Latest products do include facsimile and give information about editions
Annotation in traditional texts

Figure: Heavily annotated text of a Gozan 五山 edition of the Wudeng Huiyuan 五燈會元
More requirements

- reliability of the text

- multiple textual witnesses

- facsimile and transcription

- modern interpunctuated and textcritical edition

- sharing of intermediate results of the scholarly process
About Kanripo

The name stands for Kanseki Repository 漢籍リポジトリ

It tries to solve the problem of

- providing scholars access to reliable textual sources
- in a way that allows annotating and sharing
- without the need for a central authority

by defining a process for

and a reference implementation for this
More about Kanripo

- Kanripo works best with a text format that allows linking across different versions of a text. Such a format has been developed.
- Kanripo grew out of an earlier system for the study of Chinese called Mandoku.
  - This has now been reimplemented in a client/server model where
    - Kanripo serves as the backend/server and
    - Mandoku is one (of many) possible clients.
- Down- and uploading of texts is done using a distributed version control system called git.
- The overall design and web interface is modelled on Github, a site for open sharing of source code, which has recently been adopted for other types of data as well.
Kanripo and Metadata

- A key to understanding the texts lies in the information (data) about the texts
- Traditionally, these informations have been collected in catalogs (and commentaries)
- Kanripo provides means to maintain catalogs and combine them with the texts themselves to provide new insights
- There is a long tradition of 目録学 that can provide fruitful insights
The base edition of the Jingde chuandeng lu 景德傳燈錄 (CDL)
Sibu congkan 四部叢刊 edition

下。說法一會。度人四萬。神足二。一臥尼。二毘樓。侍者喜覺。子上聖。

*** 拘那含牟尼佛
拘那含牟尼佛(賢劫第二尊)偈曰。佛不見身知是佛若實有知別無佛。智者能知罪性空坦然不怖於生死。其長阿含經云。人壽三萬歲時此佛出世。種婆羅門姓迦葉。父大德。母善勝。居清淨城。坐烏鳴。婆羅門樹。說法一會。度人三萬。神足二。一舒槃那。二礦多樓。
侍者安和。子道師。

*** 迦葉佛
迦葉佛(賢劫第三尊)偈曰。切眾生性清淨從本無生可滅。即此身心是無生法。長阿含經云。人壽二萬歲時此佛出世。種婆羅門姓迦葉。父大德。母財主。居婆羅奈城。坐尼拘律樹。說法一會。度人二萬。神足二。一提舍。二婆羅婆。侍者善友。子集軍。

*** 釋迦牟尼佛
釋迦牟尼佛(賢劫第四尊)姓剎利。父浮飯天。母大清。浮妙。位登補處。生兜率天。名曰勝善天人。亦名護明大士。度諸天眾說補處行。亦於十方界中現身說法。普曜經云。佛初生剎利王家。放大智光明照於四方世界。地涌金蓮華自然捧雙足。東西及南北各行於七步。分手指天地作師子吼聲。上下及四維無能尊我者。即周昭王二十四年甲寅歲四月八日。至周四十二年二月八日。年十九欲
Dongchan si 東禪寺 edition

昆舍浮佛。　(莊嚴劫第一千尊。)　偈曰、
假借四大以爲身。　(心本無生因緣有。前境若
無心亦無。罪福如幻起亦滅。長阿含經云、
人壽六萬歲時。此佛出世。種種利、姓拘利若、
父善燈。母善戒。居無喻城、坐婆羅樹下。說法
二會。度人一十三萬。　(神足二、一扶軀、二鬱多
摩。侍者寂滅。)　偈曰、拘留孫佛
拘留孫佛。(見在賢劫第一尊。)
偈曰、(身無實是佛身。了心如幻是佛幻、
了得身心本性空。斯人與佛何殊別。長阿
含經傳。人　(壽四萬歲時。此佛出世。種婆羅門、
姓迦葉。父禮　(得。母善枝。居安和城、坐尸利沙
樹下。說法一　(會。度人四萬。神足二、一薩尼、二
毘樓。侍者善　(覺。子上勝。拘那含牟尼佛
拘那含牟尼佛。　(賢劫第二尊。)　偈曰、佛不見身知是佛。若實有知　(別
無佛。智者能知罪性空。坦然不怖於生死、
長阿含經傳。人壽三萬歲時。此佛出世。種婆
<md:T_T51n2076_0205b>婆羅門、姓迦葉。父大德、母善
勝。居清淨城、住烏　(暫婆羅門樹下。說法一會。度人三萬。神足二、
<pb:DCS_T51n2076_0002b>一舒榮那、二鬱多樓。侍者安和。子導師、
迦葉佛。(賢劫第三尊。)　偈曰、一切衆生性清淨。從
CBETA XML file with inline markup

<p id="pT51p0205a1901">拘那含牟尼佛</p>

<p id="pT51p0205a2001">佛不見身知是佛，若實有知別無佛。</p>

<p id="pT51p0205a2201">長阿含經云，人壽三萬歲時此佛出世。種婆</p>

<p id="pT51p0205a2301">羅門姓迦葉。父大德。母善勝。居清淨城。坐烏</p>

<p id="pT51p0205a2401">暫婆羅門樹下。說法一會。度人三萬。神足二。</p>

<p id="pT51p0205a2601">迦葉佛</p>

<p id="pT51p0205a2701">一切眾生性清淨，從本無生無可滅。</p>

<p id="pT51p0205a2901">長阿含經云，人壽二萬歲時此佛出世。種婆</p>

<p id="pT51.2076.0205b">羅門姓迦葉。父梵德。母財主。居波羅奈城。</p>

<p id="pT51.2076.0205b">坐尼拘律樹下。說法一會。度人二萬。神足二。</p>

<p id="pT51.2076.0205b">一提舍。二婆羅婆。侍者善友。子集軍。</p>

<p id="pT51.2076.0205b">羅門姓迦葉。父梵德。母財主。居波羅奈城。坐尼拘律樹下。說法一會。度人二萬。神足二。</p>

<p id="pT51.2076.0205b">一提舍。二婆羅婆。侍者善友。子集軍。</p>
Digitizing of Chinese text

- Modern character encodings for Chinese are developed with the modern understanding of characters
- This makes it very difficult to use for premodern text
- It is thus next to impossible to study changes in orthography and character usage in digital text
- A new model for digitizing Chinese text is needed
Establish a new model for digital Chinese text

- Facsimile or Transcription -> Facsimile and Transcription
- Not character codes, but position of the characters in the text as primary reference
- Multiple representations can then be related positional
Requirements for a scholarly digital edition

▶ establishing of a reliable text
▶ all major witnesses available as facsimile
▶ user has full control over the text
▶ annotation, commentary, translation has to be possible
  ▶ private, shared or public
  ▶ the user retains control over his work
▶ Distributed version control systems to the rescue?
Distributed version control systems (DVCS)

- DVCS evolved out of traditional (centralized) version control systems
- DVCS do not require a central server or central control
- Users can freely create their own branches, but retain the connection to the master repository
- Quick and wide adaption in the open source community
- Seems ideal suited for the research community
  - But there might be barriers to the entry
Example of a workflow with Kanripo

**Figure:** Interaction of public and private repositories
Text format

- For many purposes, using TEI text directly is difficult
- The text format used here is an extension of the Emacs org-mode format:
  - almost plain text
- For every version (edition) of a text, a new branch is created
- for retrieval, comparison and computational access, these are collected and held in a NoSQL database
Text ID

- ZB1a0006 周易注疏

A text ID has three parts:
- The identifier of the repository, "ZB" in this case
- The section within the repository, 1a in this case, which is
  - 1: 經部
  - a: 易類
- The serial number of the text within this section, in this case "0006"

This will allow to uniquely identify every text within every repository, while retaining some kind of common classification.
Catalog of the ZB 漢籍 repository

- The arrangement of the catalog mostly follows the *Siku quanshu* 四庫全書
- Daoist and Buddhist texts gain separate main categories, following the example of the *Suishu jingji zhi* 隋書經籍志
- 漢籍リポジトリ目録
- Alignment with images from Archive.org still needs to be done
- DZ and DZJY not yet integrated
- ZB catalogs are available on github
Mandoku

- Moved to a client / server architecture
- Most users will need only some texts
- Multiple servers are possible
- Mandoku code is available on github
- preview version now available
8180 texts have been processed and moved to new repositories
Every text requires its own repository
Not possible on github, needs separate hosting
Kanripo on a gitlab server

- Gitlab is an open-source variant of Github
- Migration of the data to a Gitlab instance is under way
- A first test version is available
- http://gl.kanripo.org
An example of a collection in Kanripo

A preliminary version is live at
http://www.daozangjiyao.org/dzjy/texts/dzjy

It is planned to merge this edition into a larger institutional repository in the not so distant future
Future Work: Sustainable resources

▶ In order to be usable in an academic environment, resources need to have a credible pedigree and carry it with them wherever they go

▶ In accordance with the LOCKSS principle, more than one way of distribution would be required
  ▶ LOCKSS: Lots of Copies Keep Stuff Safe

▶ A distributed model, in the form of a network of stakeholders, such as institutions and individuals seems favorable

▶ Kanripo is still in an early stage of development. We hope to start releasing test versions within the next year