Trustworthy digital texts for many users

Christian Wittern

2015-12-02
Overview

- Introduction to the project "Kanseki Repository"
- Aim: new infrastructure for research in all fields of East-Asian Studies
- The methodology is based on some assumptions concerning
  - (1) philological theory
  - (2) scholarly communication habits
  - (3) technological possibilities
- Realization rests on three pillars:
  - A repository of texts: http://github.com/kanripo
  - A website for accessing the texts: http://www.kanripo.org
  - A tool for using the texts for research: http://www.mandoku.org
Philological Assumptions

- We need think about a text as represented through the different documents that carried the text throughout the ages.
- We collect all evidence of the text in these documents.
- We also collect information about this text, in catalogs, other works, about the persons that did have some relation to the text and anything else that could clarify some aspects of the text.
- We also want to collect annotations, translations and other secondary work related to the text.
- We want to study the differences we see between the different documents that carry the text.
- Finally, we want to be able to make new critical editions of the text. – at all times the distinction between record and interpretation should be clear.
Scholarly Communication

- As scholars, we get paid in recognition and impact
- We want a system that maximizes control every participant has over her contribution to this communication.
- We think not only of publication, but all kinds of communication, including comments on networks of all kinds, informal posts, (e)mail, annotation, commentary and translation.
- We also want to maximize control scholars have over the source and reference materials we use:
  - ideally, everything should be freely accessible and ready for reuse
Technical realization

- There is not one solution to these problems.
- We try to develop a modular, extensible set of conventions, protocols and applications.
- We also started to collect, curate or create data according to these protocols.
- Development is best done incrementally by a community of practitioners, builders and users.
- We want different applications to be able to use the same data, depending on the needs.
Conventions for the data

- We want both digital facsimile and transcribed documents
  - these are the two faces of digital text.
- We choose (rather arbitrarily) one of these documents as the base edition.
- For all transcribed documents, we mark the location of page (and possibly) line breaks.
- For documents other than the base edition, we mark also the location of page- and line breaks of the base edition.
Current state: Research project at Kyoto University

- April 2013 to March 2016
- Sketching of methods, data format, infrastructure
- Port existing tools to enable the research community to use it
- Desired output:
  - Data and programs
  - A user community
  - A handbook with explanations of how to use the system
  - Researchers’ testimony: reports on using the system for specific research projects
Most texts have a long history and come in many versions and editions

Over the centuries many commentaries have been written and emendations or annotations proposed

Scholars want to trace and evaluate these connections and discover new ones

No single publication/institution can solve this problem -> collaboration is needed across time and space
For casual browsing: Kanripo Website

- http://www.kanripo.org
- Early stage of development
- Searching and browsing of text and catalog
- Integration with the user’s account on github:
  - A user can annotate his own version and see it immediately on the website
  - A user can collect and refine data while working with the site
  - Some metadata, such as relevant dates for texts can be edited by the user and the user-specific version of the data will be used
For close reading: Mandoku

- A specialized environment and interface to the repository
- Quick lookup of phrases in the text database
- Integrates with Emacs as a powerful general purpose research environment
- More flexible than the website, but the two can be used together
- Additional sites can be integrated, texts can be easily incorporated
- Annotations, research notes, links to other texts, translation can all be assembled
- Scholarly communications of all types, including articles and books can be authored directly from this research environment
Technical digression

- Mandoku is based on Emacs and Org-Mode:
  - Why Emacs?
    - Is there a better tool for this purpose?
  - What is Org-Mode? (http://orgmode.org/)
  - An extension of Emacs for planning and organizing
  - Also used for research assistance and writing of academic papers, presentations etc.
  - Used also for note taking, blog articles etc.

- Viable perspective? Other possibilities?
Technical details: Repository

- The texts are kept in the repository under a version control system.
- This allows:
  - to keep multiple corresponding versions of a text.
  - to keep a record of what has been edited (by whom).
  - updating and collating of multiple copies in different locations.
- This also provides a stable identifier and location for the text.
- We hope that scholars working on specific texts will use this as a hub for communication.
Technical details: Editions, Versions, Branches

- This terminology can be confusing:
- An edition of a text available as a document is considered a version of this text.
- In the repository, the technical term for a version is branch.
- As a convention, we use the following conventions for branch names:
  - uppercase only for editions (sometimes specified with a "-" character), e.g. CK-KZ
  - uppercase with "@" character for editions reported through other editions, e.g. TK@T
  - lowercase only or mixed for branches derived from uppercase branches, e.g. T-chris
  - branch names starting with "_" (underscore) are used internally for metadata
Workflow for downloading and editing a text

- Find the repository on github.com/kanripo (or follow the link from www.kanripo.org)
- clone or fork the repository
- (fork allows better reporting of differences back to the owners)
- the text repository is now under control of the user
- if forked, all editions can be pushed to the users github account
- for corrections, the user can issue a pull-request to the kanripo owners
GitHub (http://github.com)

- A popular code sharing site, used by software developers.
- Provides some functionality for creating research groups, user interaction etc.
- Editing on-site is possible, but slightly awkward
- Not specially suited for working with texts -> additional/other access point is needed
Wrap up

- The project hopes to present itself to the public in March 2016
- Attempts will be made to create a welcoming community of practice around Kanripo and Mandoku
- We also hope to integrate with other projects:
  - Learn how to improve our data and processes
    - Add and exchange data, methods and technology
    - and finally...

Improve our understanding of the East Asian cultural sphere