Differential Reading by Image-based Change Detection and Prospect for Human-Machine Collaboration for Differential Transcription

Asanobu KITAMOTO (Center for Open Data in the Humanities (CODH), Japan)

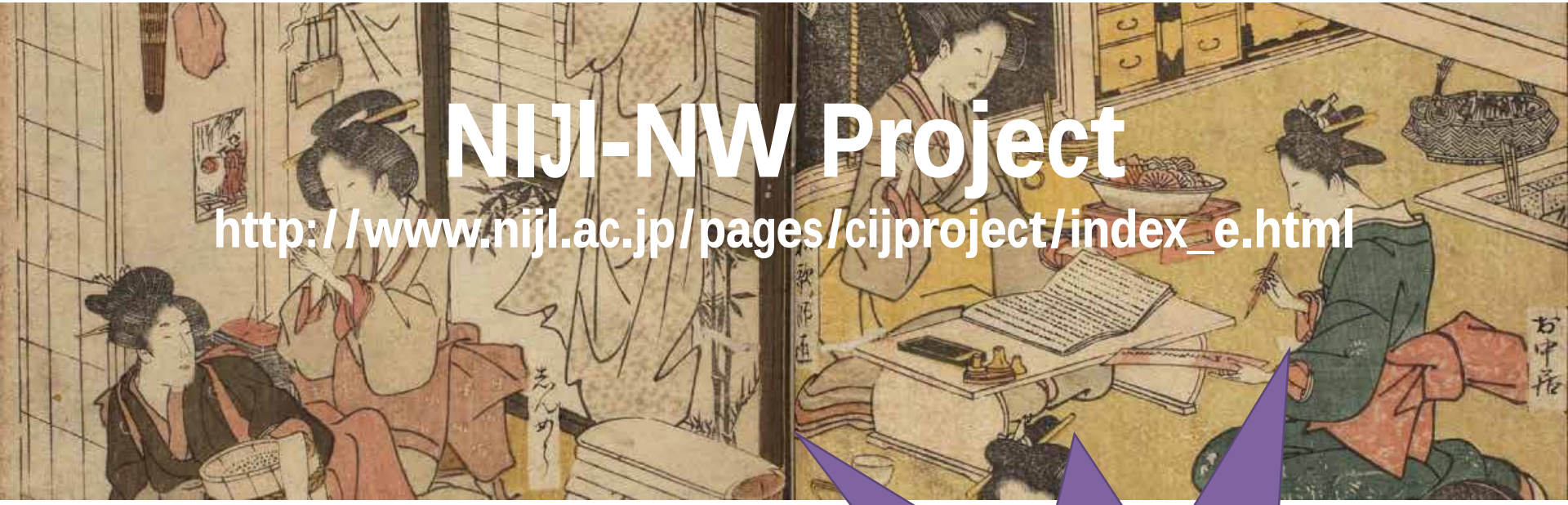
Hiroshi HORII, Misato HORII (AMANE LLC)

Chikahiko SUZUKI (CODH)

Kazuaki YAMAMOTO (National Institute of Japanese Literature)

Kumiko FUJIZANE (Notre Dame Seishin University)

http://codh.rois.ac.jp/ @rois_codh



NIJL-NW Project

http://www.nijl.ac.jp/pages/cijproject/index_e.html

300,000 Pre-modern Japanese Books (before 1868) are being digitized and released as open data.

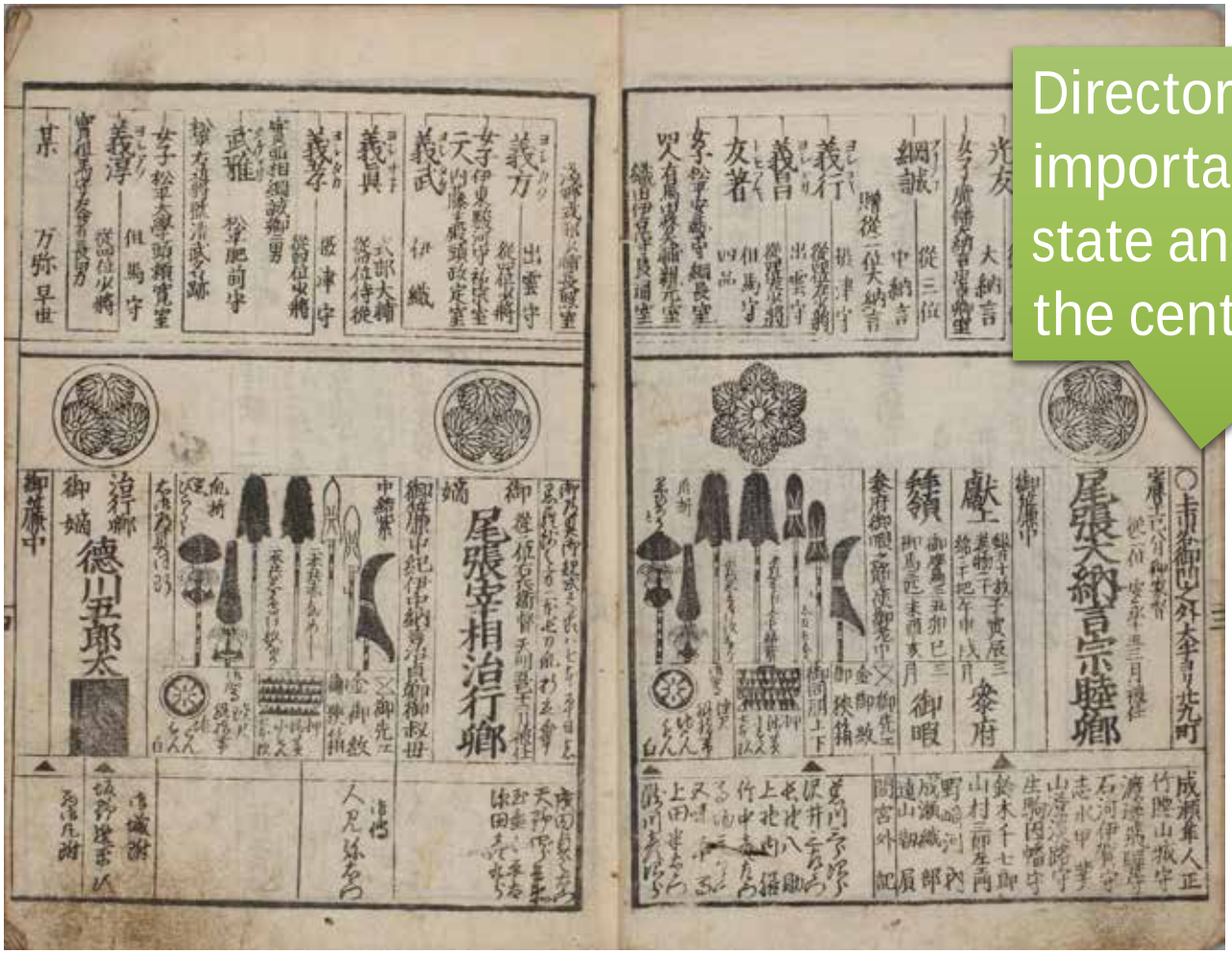
Japanese culture finally entered into big data era...

Problems of Old Japanese Books

1. Japan had very active publishing industry in the Edo period (1603-1868).
2. Characters and writing has changed, so native Japanese are not good readers now.
3. Too many books for too few readers. Books/readers ratio is the global worst?
4. Humans and machines should collaborate for deep access to the content of books.

Kansei Bukan（1789）

http://codh.rois.ac.jp/pmjt/book/200018823/

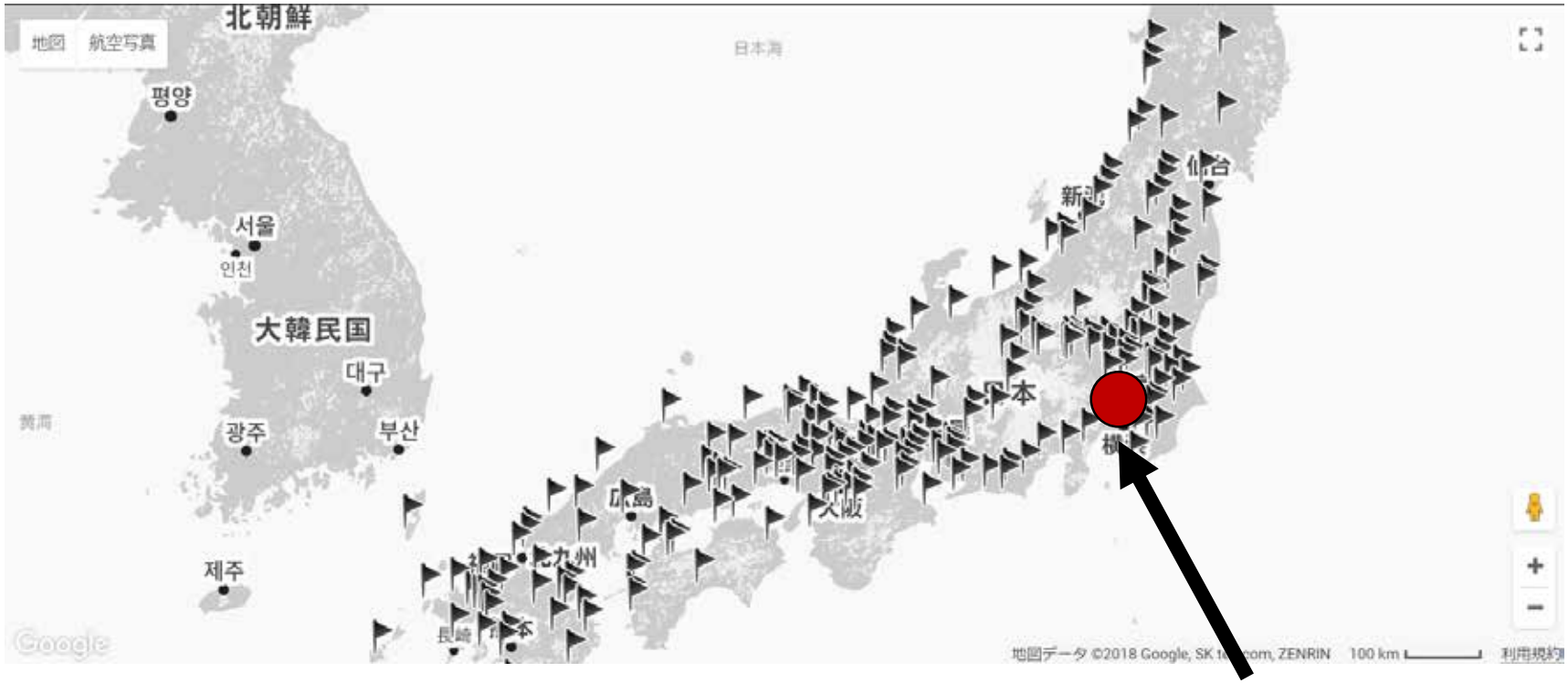


Directory of families and important people in each state and bureaucrats of the central government.

Dataset of pre-modern Japanese Text (archived in National Institute of Japanese Literature)

Administration in the Edo Period

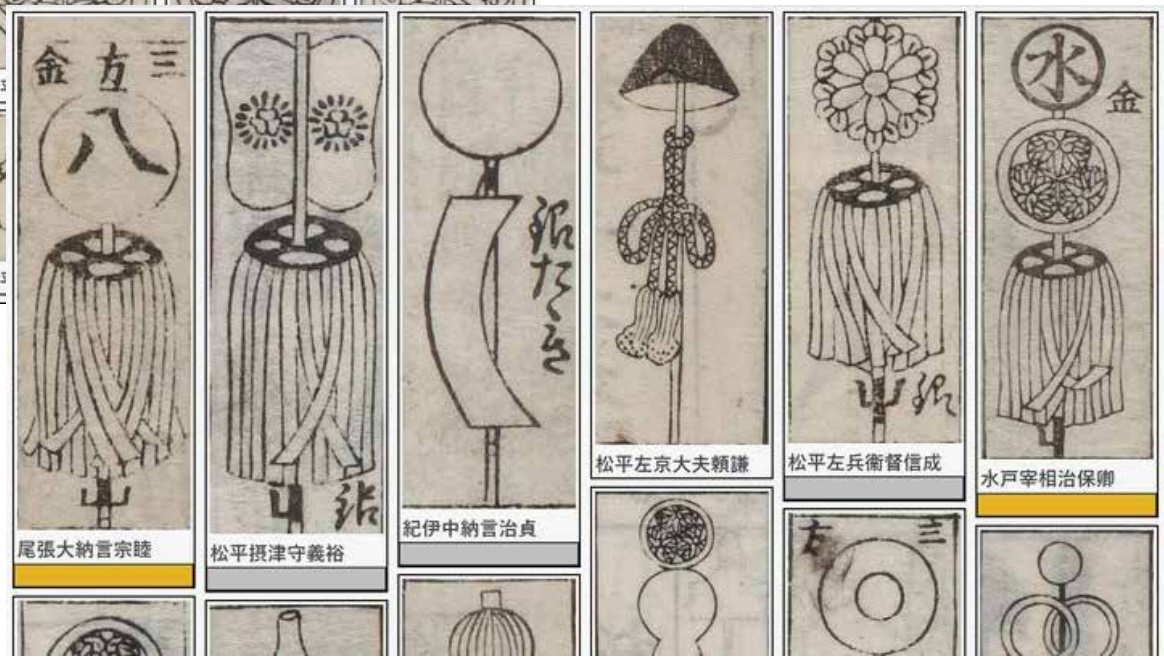
http://codh.rois.ac.jp/bukan/book/200018823/map/



The central government (Bakufu) at Edo (Tokyo) ruled 264 states (Daimyo).

Curation of Graphical Elements

Matoi: Firefighter symbols



Kamon: Family emblems

Background and Problems

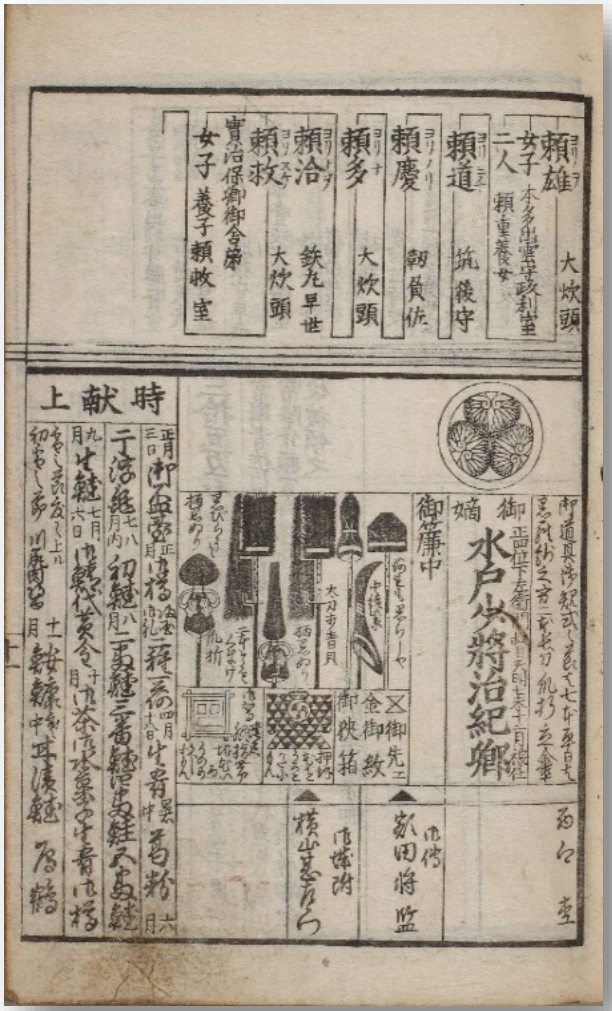
1. Historians did not use “Bukan” books as primary source because they are commercial, not official, publications.
2. Bukan was information source for celebrating people‘s promotion, identifying daimyo family, or used as a souvenir.
3. Best seller for 100-200 years, with update frequency of a few times in a month at peak. = **problem of many versions!**

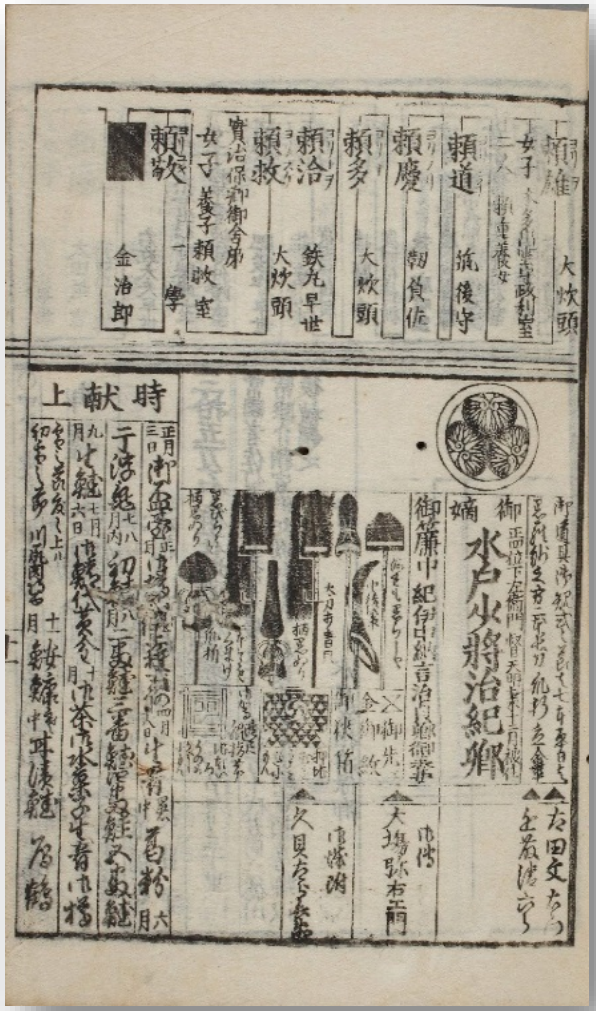
Definition of Versions

In the Edo period, **woodblock printing** was mainly used instead of movable type.

1. Publication: woodblock is completely recreated (=**major version**) .
2. Print: multiple prints are produced from the same woodblock (=**installation**)
3. **Correction**: woodblock is carved or patched by a small plate (=**minor version**).

Comparison of Different Versions





Left: Kansei Bukan **1789**.

Right: Kansei Bukan **1791**.

Dataset of pre-modern Japanese Text (archived in National Institute of Japanese Literature)

Text-based and Image-Based Change Detection

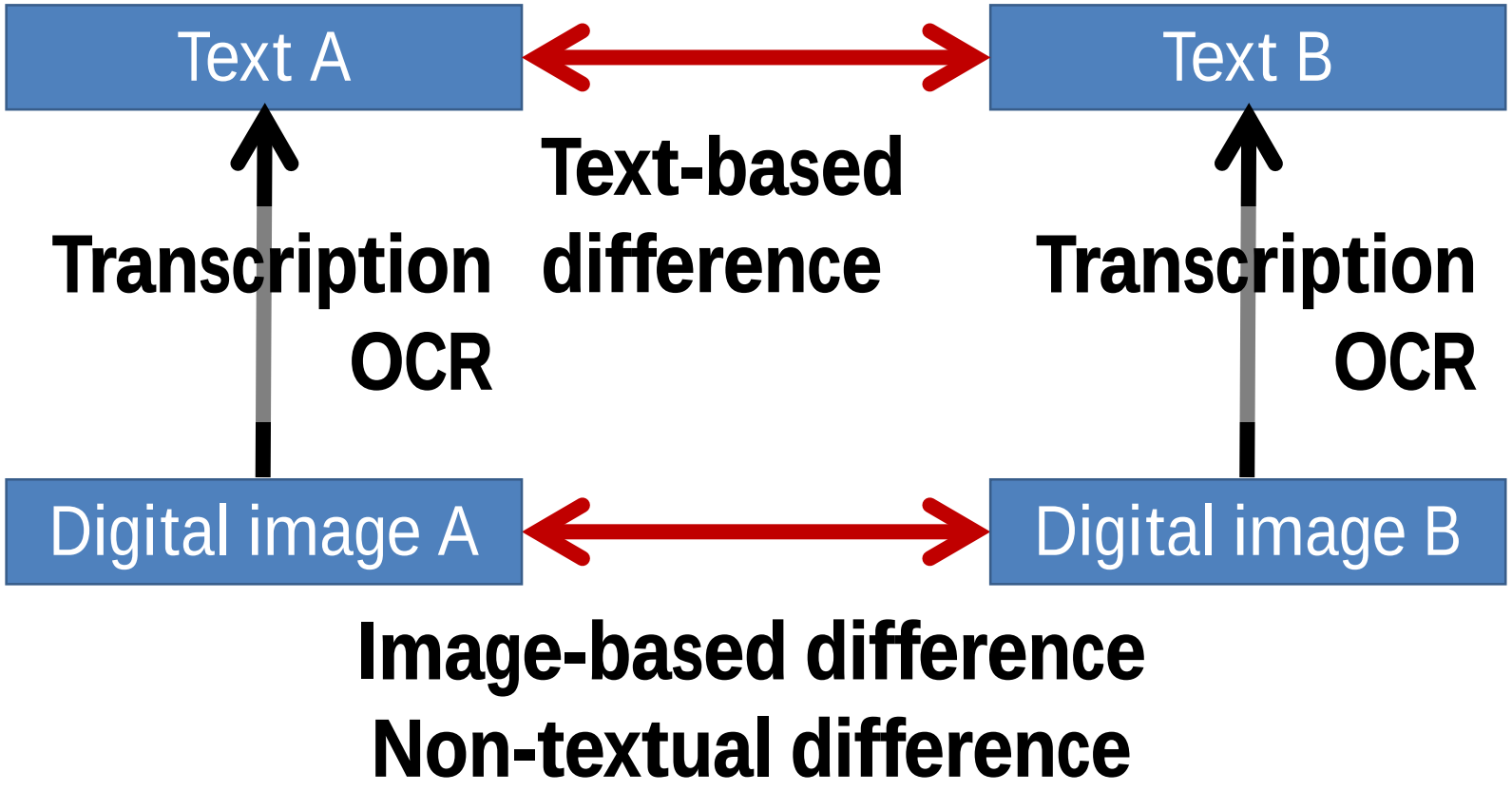
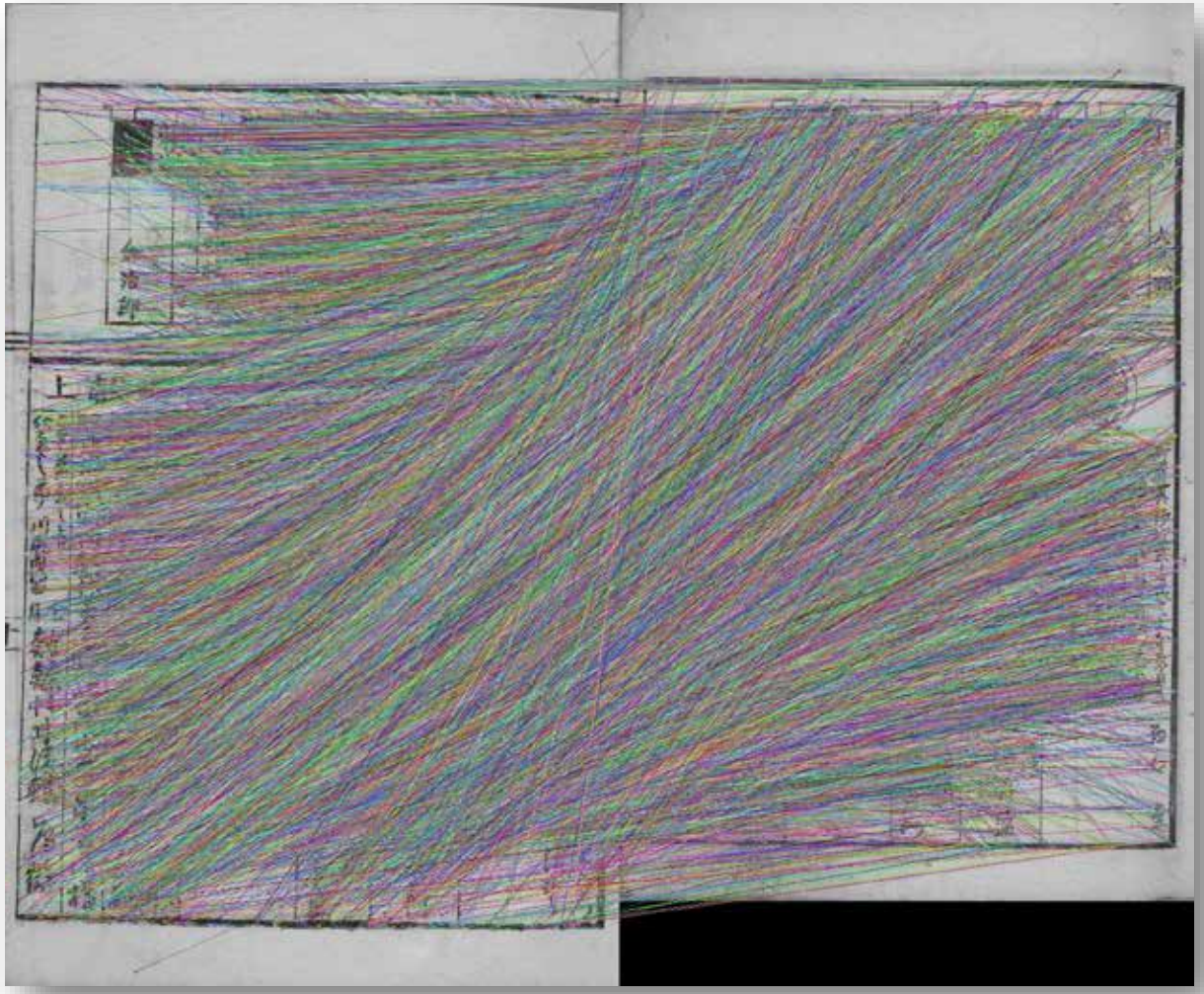
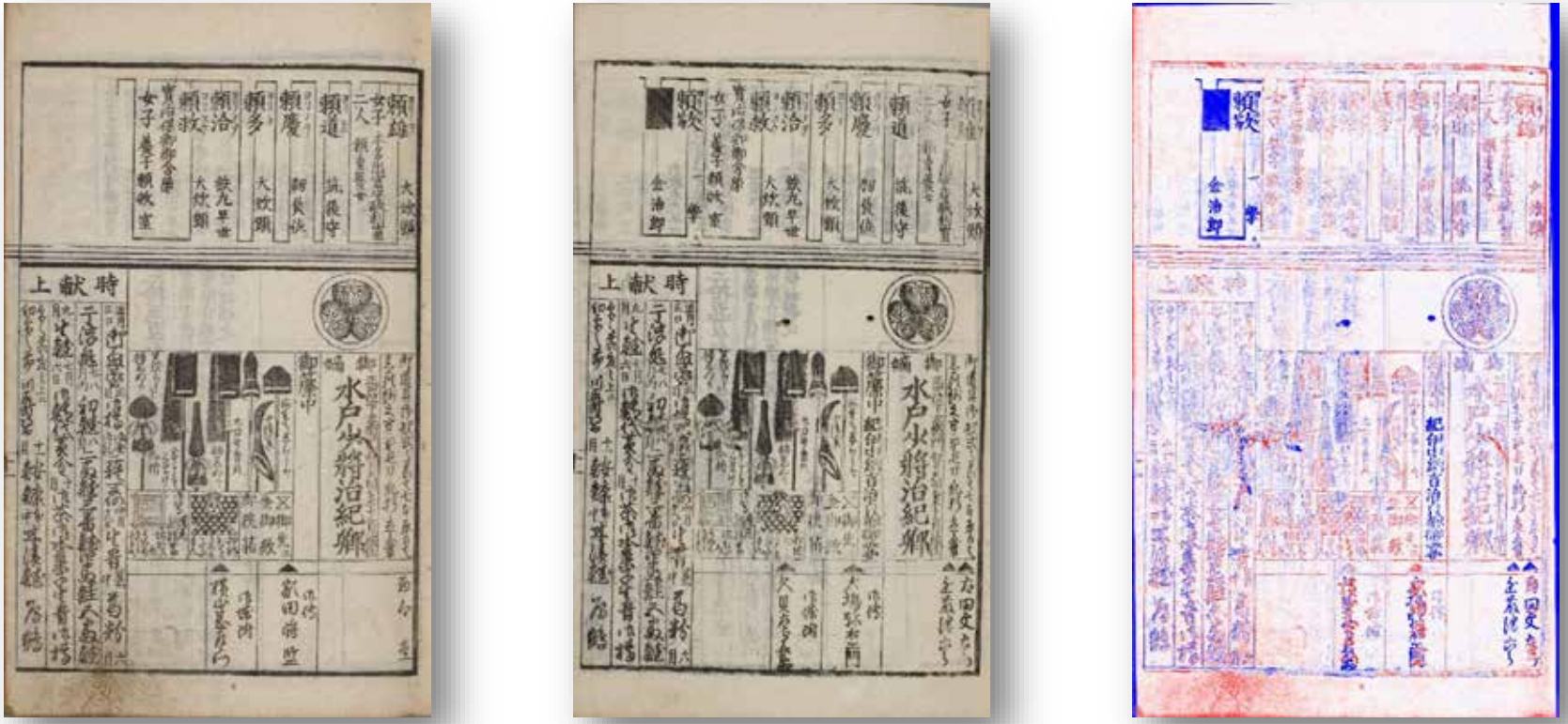


Image Matching



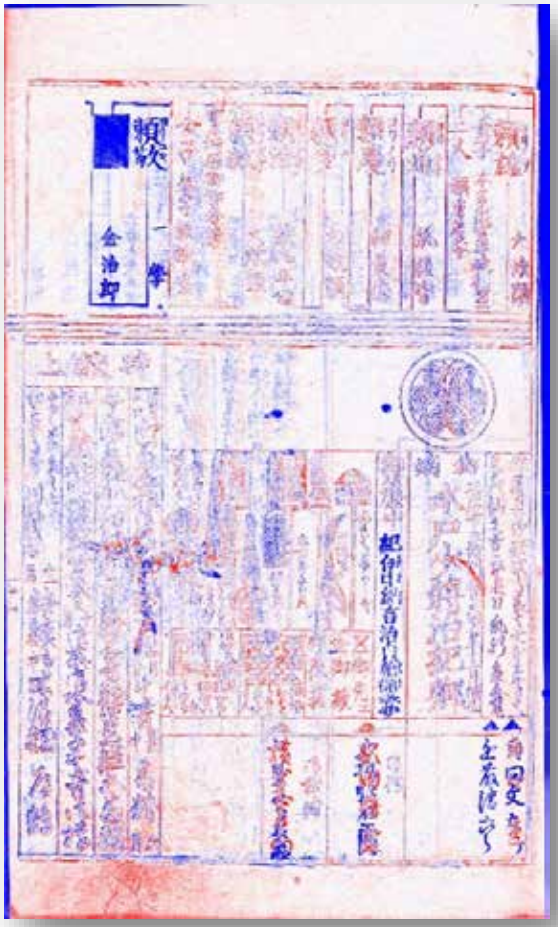
- OpenCV 2.4
- Feature detection: FAST
- Feature description：BRIEF
- Matching：Hamming distance
- Homography matrix：RANSAC
- Visualization：coloring scheme from blue to red.

Image-Based Change Detection



Left: Kansei Bukan (1789). Middle: Kansei Bukan (1791). Right: comparison, 1789 = red, 1791= blue.

Differential Transcription



- Base transcription: the whole page is transcribed.
- Differential transcription: only changes are transcribed.
- Recreation detection: woodblock recreation triggers base transcription of the page.
- Advantage: The amount of transcription is reduced to the percentage of change.

Differential Reading

1. A new style of reading supported by machines, such as distant reading.
2. Subtle changes are enhanced by machines to help humans notice the change.
3. A task that humans do not like, but machines are good at. Machines can produce better results than humans.
4. **What can we read from difference?**

New Research Questions

1. Complete ordering: given two books, which is the newer? correction history gives hints.
2. Publishing industry: how fast the error was fixed? how long the woodblock was used?
3. Human resource: how many and how often people were promoted or disappeared?
4. Historical big data: how economic situation affected changes in human resource?

Summary

1. “Bukan” is a unique historical source in volume, in frequency and in variety.
2. Woodblock printing allows us to pursue image-based change detection.
3. Differential transcription takes advantage of human-machine collaboration.
4. Differential reading on many versions allows us to pursue new research questions.

Related Resources

- **Bukan Complete Collection**
  - http://codh.rois.ac.jp/bukan/
- **Bukan books (381 versions) in the Database of Pre-modern Japanese Text**
  - http://codh.rois.ac.jp/pmjt/book/?**武鑑**
- **IIIF Curation Viewer**
  - http://codh.rois.ac.jp/software/iiif-curation-viewer/