Center for Open Data in the Humanities (CODH): Activities and Future Plans

Asanobu KITAMOTO
National Institute of Informatics
Research Center for Open Data in the Humanities (CODH)
Research Organization and Information and Systems

http://codh.rois.ac.jp/
Twitter: @rois_codh
Introduction
What is CODH?
http://codh.rois.ac.jp/

• April 1, 2016: Established as a pre-center.
• April 1, 2017: Officially launched (I hope).
• ROIS > Join Support-Center for Data Science Research > CODH

1. Humanities research based on the technology of informatics and statistics.

2. Informatics and statistics research using humanities data.
Data Science Research Centers

Life Science
Polar Research
Social Science
Humanities
Sloan Digital Sky Survey

- The Sloan Digital Sky Survey has created the most detailed three-dimensional maps of the Universe ever made, with deep multi-color images of one third of the sky, and spectra for more than three million astronomical objects.

http://www.sdss.org/
Citizen Science at Galaxy Zoo

• It all started back in July 2007, with a data set made up of a million galaxies imaged by the Sloan Digital Sky Survey, who still provide some of the images in the site today.

• With so many galaxies, we'd assumed it would take years for visitors to the site to work through them all, but within 24 hours of launch we were stunned to be receiving almost 70,000 classifications an hour.

• In the end, more than 50 million classifications were received by the project during its first year, contributed by more than 150,000 people.

https://www.galaxyzoo.org/
Data-Driven Science

1. Data → Theory

2. Small Theory → Solution

3. Big Theory → Simulation

4. Data → Generalization
HathiTrust Digital Library

- Good example of data-driven humanities.
- 5,199,106,500 pages (as of Jan.22, 2017)
- A database where you can ask many interesting questions.
- Quantitative evidences can be obtained.

https://www.hathitrust.org/
Digital Silk Road and Digital Humanities

http://dsr.nii.ac.jp/
Digital Silk Road

http://dsr.nii.ac.jp/

• Started in 2001.

• Digital Humanities: Collaborative work among informatics + humanities scholars.

• Databases and digital resources are publicly accessible on the Web.
Digital Archive of Toyo Bunko Rare Books

http://dsr.nii.ac.jp/toyobunko/

• 245 books (72,591 pages) were digitized and released.
• Relevant books in the academic community of Silk Road were selected.
• Caption and table of contents were manually typed.
• Full text is obtained by OCR (without correcting errors).
Variety and Heterogeneity of Data

Text

Map

Photograph

Gazetteer
Stein Map (Silk Road)

- Stein’s map “Innermost Asia” was registered and displayed on Google Earth satellite images.
Question: “Missing” Ruins?

- Oi-tam, ruined fort
- Bögan-tura
- Buluyuk (Shipang, Sassik-bulak, Kazma)
- Murtuk-ruins
- Yoghan-tura
- Chikkan-köl
- Bedaulat’s town, Bēsh-kāwuk, Kosh-gumbaz
- Yutōgh

Missing Ruins
• Some ruins were reported by 20\textsuperscript{th} expeditions, but are missing in recent survey reports.
Matching Entities

Stein’s map and satellite images for the same area. Each source reports different ruins due to different conceptualization.
Linking Entities Across Sources

Textual criticism

Textual source A

Place name S

Textual criticism

Textual source B

Place name T

Geographic
And Visual
Relationship

Place name U

Non-textual Source C, D, E...

Place name V

Digital criticism

2017/01/23 1st CODH Seminar
Japanese Literature and Open Data

http://codh.rois.ac.jp/
CODH / NII / NIJL Collaboration

CODH
Promote data-driven research for humanities with infrastructure for open data and science.

NIJL-NW Project
Digitize 300,000 pre-modern Japanese books and make them open to promote international collaboration.

NII and ISM in ROIS are involved in the center.

NIJL in NIHU plays the central role.

Solve issues in Japanese literature through collaboration between informatics and humanities.

2017/01/23 1st CODH Seminar 19
Released on November 10, 2016

Open Data for Scholars

Pre-Modern Japanese Text Dataset (from NIJL)
Open Data for Machines

PMJT Dataset (from NIJL)

PMJT Character Shape Dataset (from NIJL and processed by CODH)
Released on November 24, 2016

Open Data for Citizens

PMJT Dataset
(from NIJL)

Edo Cooking Recipe Dataset
(Created by CODH)
Adapted Material on NIJL Dataset
(from NIJL)
1. Pre-Modern Japanese Text (PMJT) Dataset
http://codh.rois.ac.jp/pmjt/
Pre-Modern Japanese Text (PMJT) Dataset

- **November 2016** “Pre-Modern Japanese Text Dataset” (700 items) released from CODH.
- In addition to image files, bibliographic metadata and tags given by experts are also included.
- Transcribed text is added to a limited number of books.
- License is CC BY-SA 4.0.
Data Distribution

- IIIF (International Image Interoperability Framework) = protocol for images based on an international activity.
- Developed new IIIF viewer for multi-resolution browsing as open source.

IIIF Curation Viewer
Core contributor: Jun HOMMA
Data Identifiers

• Scholarly information becomes the network of knowledge connected by global identifiers.

• DOI (Digital Object Identifier) : the basic identifier for research publications and data.

• NIJL : Planned to assign DOI for each book using the ID derived from their databases.

• CODH : Planned to assign DOI to derivative works from NIJL datasets and other datasets.
2. PMJT Character Shape Dataset

http://codh.rois.ac.jp/char-shape/
PMJT Character Shape Dataset
Training Data for People

- Check the variation of character shape by their eyes.
- Can be incorporated into educational apps.
- Virtuous cycle: more people can read the characters, more people can use the dataset.
Training Data for Machines

- Training data for machine learning research.
- A sample program using deep learning library Keras.
- Coordinate information may be useful for analysis beyond characters.
- **Mother characters** is left for future work.

<table>
<thead>
<tr>
<th>Types</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>し</td>
<td>3,929</td>
</tr>
<tr>
<td>に</td>
<td>3,147</td>
</tr>
<tr>
<td>の</td>
<td>2,908</td>
</tr>
<tr>
<td>て</td>
<td>2,398</td>
</tr>
<tr>
<td>り</td>
<td>2,193</td>
</tr>
<tr>
<td>を</td>
<td>2,021</td>
</tr>
<tr>
<td>か</td>
<td>1,910</td>
</tr>
<tr>
<td>く</td>
<td>1,739</td>
</tr>
<tr>
<td>き</td>
<td>1,715</td>
</tr>
<tr>
<td>も</td>
<td>1,463</td>
</tr>
<tr>
<td>1,521 types</td>
<td>86,176 characters</td>
</tr>
</tbody>
</table>
Deep Access and Scriptome Analysis

• **Deep access**: access to images should be enhanced with access to content.

• **OCR**: good for printed text, but pre-modern Japanese text has only limited success.

• **Many approaches**: deep access is not only about OCR but image analysis for search.

• **Scriptome analysis**: the whole written text analysis is comparable to genome analysis.
# History of Genome Analysis

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1953</td>
<td>DNA double helix model was proposed.</td>
</tr>
<tr>
<td>1980s decade</td>
<td>100 years for the whole genome sequencing?</td>
</tr>
<tr>
<td>About 1987</td>
<td>Japanese scientist proposed the automated analysis for speed-up the sequencing.</td>
</tr>
<tr>
<td>About 2003</td>
<td>Human genome sequencing was completed after spending 13 years and 3 billion dollars.</td>
</tr>
<tr>
<td>2016</td>
<td>Human genome can be sequenced around 1000 to 10000 dollars and the price is still going down.</td>
</tr>
</tbody>
</table>
3. Edo Cooking Recipe Dataset

http://codh.rois.ac.jp/edo-cooking/
Edo Cooking Recipe Dataset

1. Digitize cooking recipe books.
2. Transcribe old Japanese characters.
3. Translate them into modern Japanese.
4. Adapt translation into a recipe.
5. Release the recipe at Cookpad.
6. Share experiences at “Tsukurepo.”

Collaborated with AMANE LLC.
## 2. Transcription

<table>
<thead>
<tr>
<th></th>
<th>Transcription</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>は大角の赤干藻一本を水につけほどばかし</td>
</tr>
<tr>
<td>2</td>
<td>鍋にいれ水二合入して煎し布にて一へんはよく</td>
</tr>
<tr>
<td></td>
<td>くこし又鍋へ入てあつくして</td>
</tr>
<tr>
<td>3</td>
<td>たまご十貫をわり込よくよくとき是も布にてこし</td>
</tr>
<tr>
<td>4</td>
<td>切右の中へ黒砂糖を五十匁酒すこし入る是も布にてこし</td>
</tr>
<tr>
<td>5</td>
<td>此二色をかんてんの鍋の中へ入る</td>
</tr>
<tr>
<td>6</td>
<td>是もすこしちつ小杓子にてそろそるとかきまた</td>
</tr>
<tr>
<td></td>
<td>川しかきまわし入れるなり</td>
</tr>
<tr>
<td>7</td>
<td>皆入てより又葛粉をすこし水にてとき入れ</td>
</tr>
<tr>
<td>8</td>
<td>切鍋をぬき早く折敷にてもうちあげ平めに延し</td>
</tr>
<tr>
<td></td>
<td>入れ物とも水に入れ冷し遣ふ</td>
</tr>
</tbody>
</table>

**PMJT Dataset (from NIJL)**

**Edo Cooking Recipe Dataset (Created by CODH)**

2017/01/23 1st CODH Seminar
### 3. Translation

<table>
<thead>
<tr>
<th>1</th>
<th>大きな赤寒天を1本水に付けてふやかす。</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>鍋に寒天と水2合（360cc）を入れて煮溶かす。</td>
</tr>
<tr>
<td>3</td>
<td>抹を一度布で素早く漉し、再び鍋に入れて熱する。</td>
</tr>
<tr>
<td>4</td>
<td>生卵10個をよく溶き、布で漉す。</td>
</tr>
<tr>
<td>5</td>
<td>抹の中に黒砂糖50両（200g）と酒少しを入れ、布で漉す。</td>
</tr>
<tr>
<td>6</td>
<td>抹を寒天の鍋に入れる。小さな杓子で少しずつそろそろと混ぜながら入れる。</td>
</tr>
<tr>
<td>7</td>
<td>抹を全て鍋の中に入れたら、葛粉を水で溶き、鍋に入れる。</td>
</tr>
<tr>
<td>8</td>
<td>鍋を火から上げ、素早く中身を容器（折敷）に広げ、平たく延ばし、容器ともに水で冷やす。</td>
</tr>
</tbody>
</table>

PMJT Dataset (from NIJL)

Edo Cooking Recipe Dataset (Created by CODH)
4. Adaptation

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>寒天を水につけて、ふやかします。</td>
</tr>
<tr>
<td>2</td>
<td>生卵をよく溶きます。</td>
</tr>
<tr>
<td>3</td>
<td>溶いた生卵を布でこします。</td>
</tr>
<tr>
<td>4</td>
<td>黒砂糖と酒を入れ、溶かします。</td>
</tr>
<tr>
<td>5</td>
<td>4を3に入れ、再びこします。</td>
</tr>
<tr>
<td>6</td>
<td>鍋に寒天と水（180cc）を入れて煮とかします。</td>
</tr>
<tr>
<td>7</td>
<td>6を布などでこし、再び鍋に入れて熱します。</td>
</tr>
<tr>
<td>8</td>
<td>7の熱した寒天の中に、5の卵液を少しずつ入れます。</td>
</tr>
<tr>
<td>9</td>
<td>全て入れ終えたら、水でいたした片栗粉を鍋に入れてさっと混ぜ合わせます。</td>
</tr>
<tr>
<td>10</td>
<td>鍋を火からあげ、中身を容器に入れます。</td>
</tr>
<tr>
<td>11</td>
<td>冷蔵庫で、2時間程度冷やします。</td>
</tr>
</tbody>
</table>

PMJT Dataset (from NIJL) | Edo Cooking Recipe Dataset (Created by CODH)
Photographs for the Recipe

Edo Cooking Recipe Dataset (Created by CODH)
Edo Recipe Cooking Dataset

1. Transcription: 107

Released on the website as open data (CC BY-SA).

http://codh.rois.ac.jp/edo-cooking/
Dataset Release at ‘Cookpad’

Joint work with Cookpad and The Japan Society of Home Economics, Division of Food Culture.

Deposit and release the data from a web service (app) where people are already well familiar with.

http://cookpad.com/recipe/4153357
Unexpectedly Large Impact

https://twitter.com/caille2006/status/802575840819089409

https://twitter.com/jouhouken/status/801693251052781568
TV show to reproduce the dish

Lessons Learned

• Open data for citizens should be well prepared for immediate use, and should be released on the platform they love. The response is surprisingly different.
• Where to deposit data is an important issue, just as where to submit a paper is important.
• Put old data into a new platform gives an impact and can be generalized to other cases.
Open Science and CODH
Machine Citizen Scholar

Deepen Increase

Participatory and citizen science

Competition and cooperation between human and machines

Trans-disciplinary data platform

Data-driven science

Expand
1. Scholar

• Answer research questions by deeper interpretations of sources enabled by tools.
• We particularly focus on non-textual sources such as maps, photographs and images.
• Collaboration: good questions and best technologies are key to success.
• Our role: we ultimately work with communities, not individual scholars.
2. Machine

• Answer research questions by (quantitative) evidences supported by increasing data.

• We particularly focus on deep access technologies such as character recognition.

• Artificial intelligence: deep learning and other algorithms increased the potential.

• Our role: we ultimately develop new technology inspired by humanities data.
3. Citizen

• Answer research questions with the power of expanded supporters of research.

• We particularly focus on data collection in the field using mobile apps and other tools.

• Education: citizen science involves the training of people for better activities.

• Our role: we ultimately develop new platform so that citizen can share new data.
More Data Professionals

• **Data librarian**: organize data (offer the fundamental value).

• **Data curator**: arrange and order data (offer value-added services).

• **Data analyst (data scientist)**: analyze data (algorithmically).

• **Data engineer**: design and build data-related systems.
Summary

• **Mission**: data-driven approaches to humanities to explore new possibility.

• **Achievement**: Digital Silk Road, NIJL-NW, and other smaller projects.

• **Direction**: scholar (deepen), machine (increase), and citizen (expand) dimensions.

• **Wanted**: we are looking for good partners, both in terms of technology (informatics) and problems (humanities).
Related Websites

• Center for Open Data in the Humanities
  • http://codh.rois.ac.jp/

• Digital Silk Road
  • http://dsr.nii.ac.jp/

• Joint Support-Center for Data Science Research
  • http://ds.rois.ac.jp/

• Open Science
  • http://agora.ex.nii.ac.jp/~kitamoto/research/open-science/