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

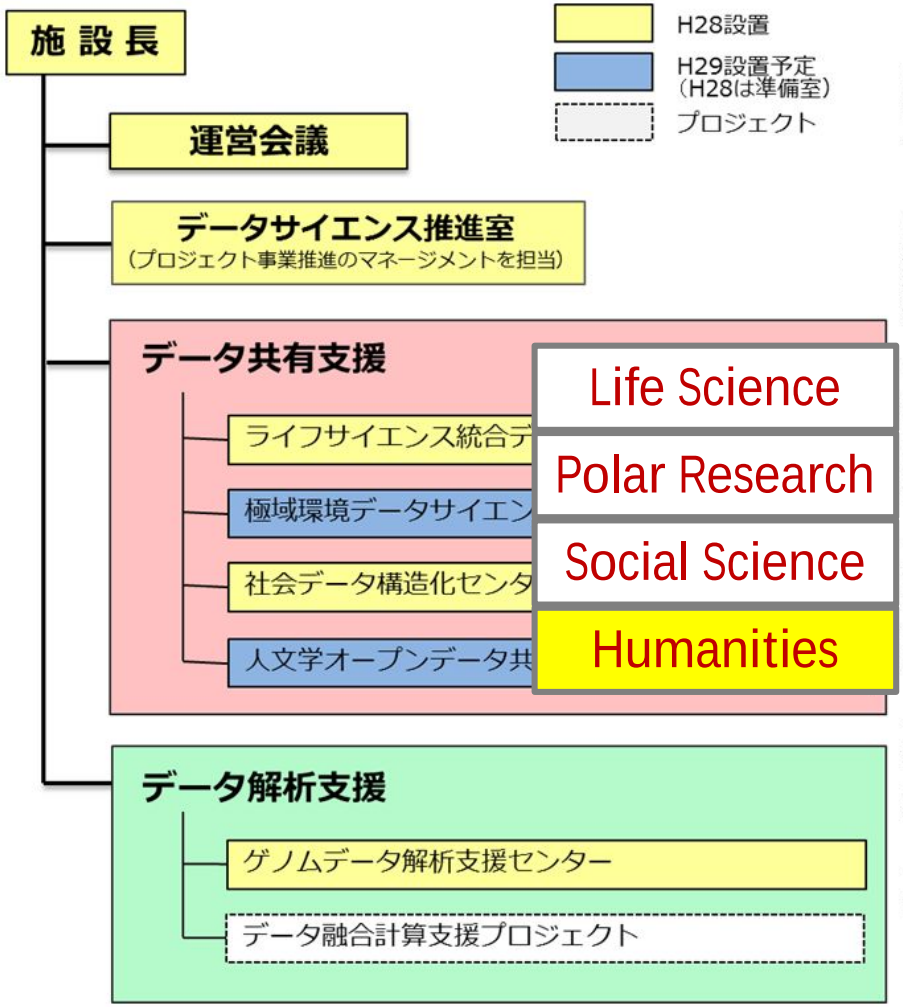
What is CODH?

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



◆ライフサイエンス統合データベースセンター◆
生命科学分野のオープンサイエンス推進、ライフサイエンス・データベース統合化のための研究開発を推進
（JST/NBDCとの共同研究事業）

◆極域環境データサイエンスセンター◆
過去から現在に至る長大な時間軸を持った極域環境変動・地球システム変動に関する貴重なデータと、その分析・解析支援を提供する共同利用を推進

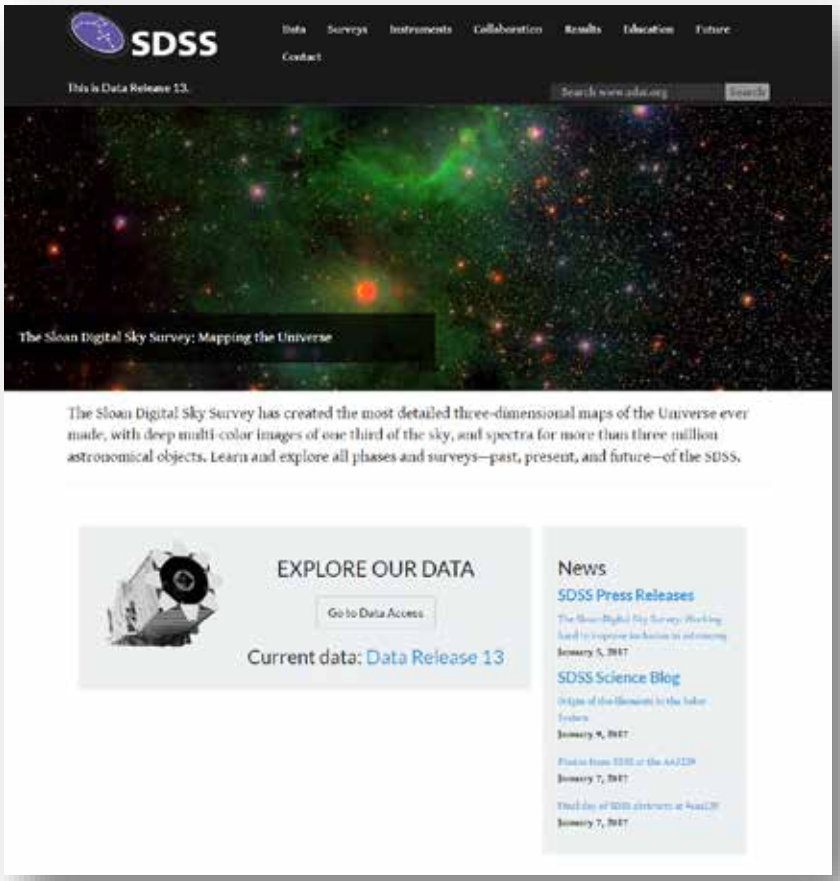
◆社会データ構造化センター◆
大学研究者のための社会調査データ、公的統計ミクロデータ、ソーシャルビッグデータに関するデータベースを整備。また、データ利用コミュニティを形成し、環境、治安、経済を含む各種の社会的課題の解決のための実証的研究を促進

◆人文学オープンデータ共同利用センター◆
データサイエンスに基づく人文学（人文情報学）という新たな学問分野を創生するとともに、データを中心としたオープン化を推進することで、組織の枠を超えた研究拠点を形成・強化

◆ゲノムデータ解析支援センター◆
最先端のバイオインフォマティクス技術を駆使して大量のゲノム・トランスクリプトームデータから生物学的に重要な情報を得るためのデータ解析支援

◆データ融合計算支援プロジェクト◆
データ融合計算技術による諸科学・産業界での課題解決

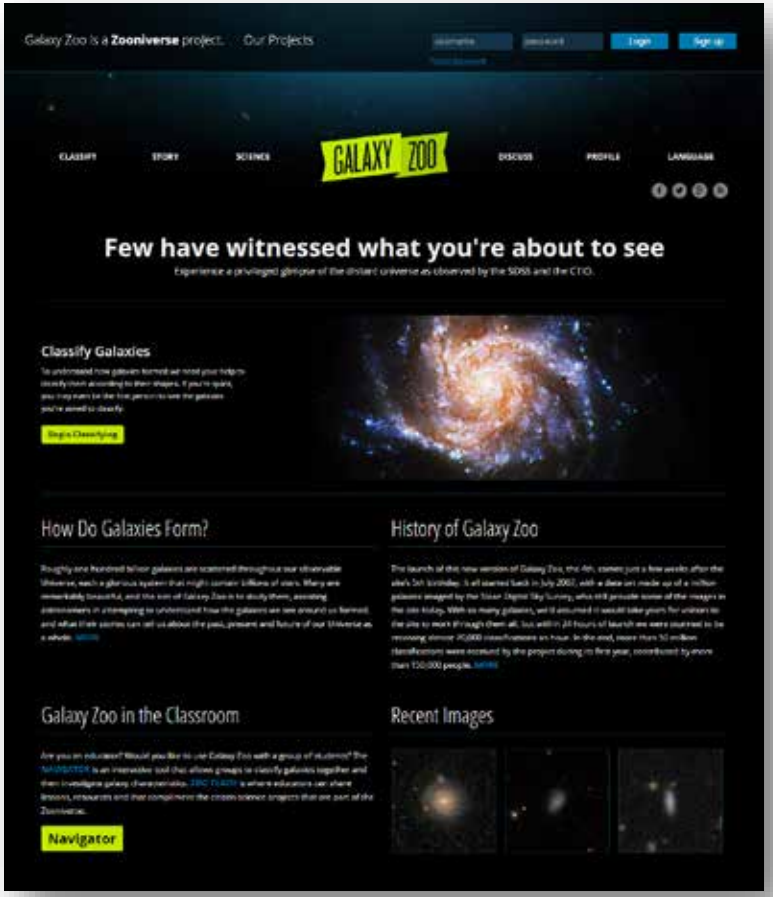
Sloan Digital Sky Survey



http://www.sdss.org/

- 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.

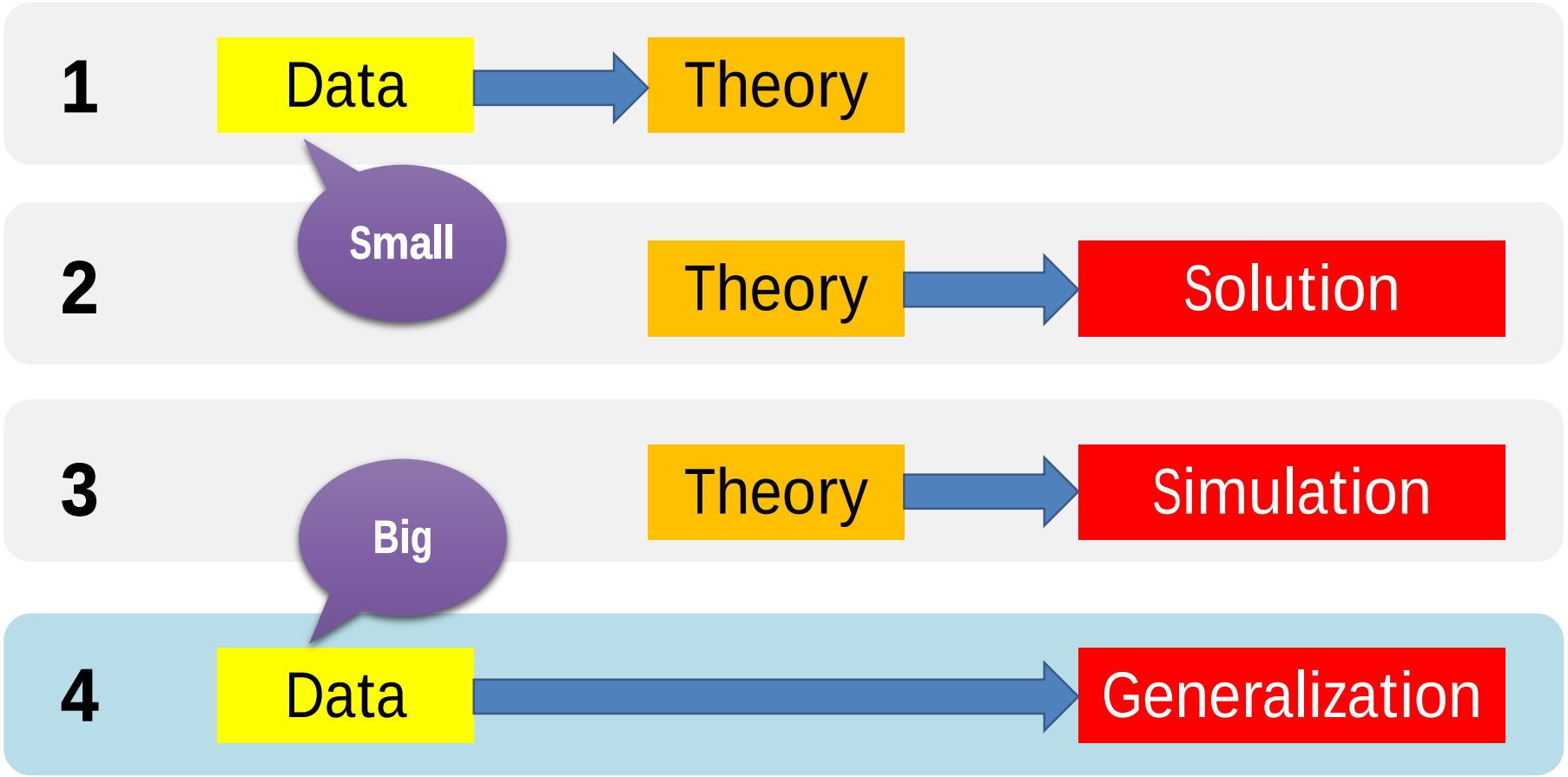
Citizen Science at Galaxy Zoo



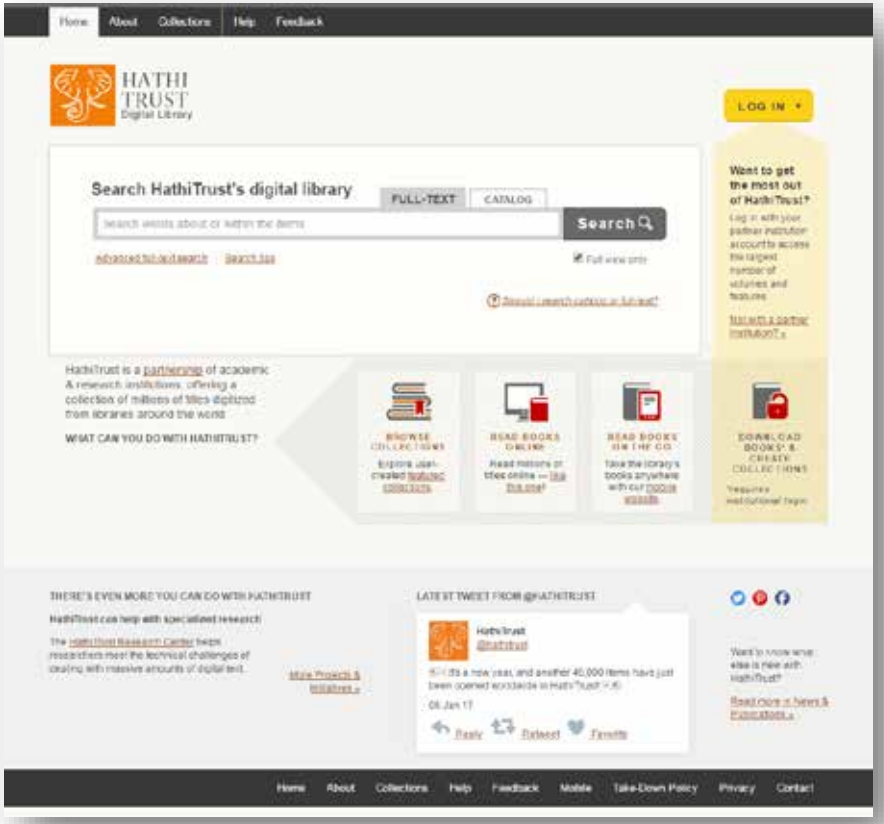
https://www.galaxyzoo.org/

- 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.

Data-Driven Science



HathiTrust Digital Library



https://www.hathitrust.org/

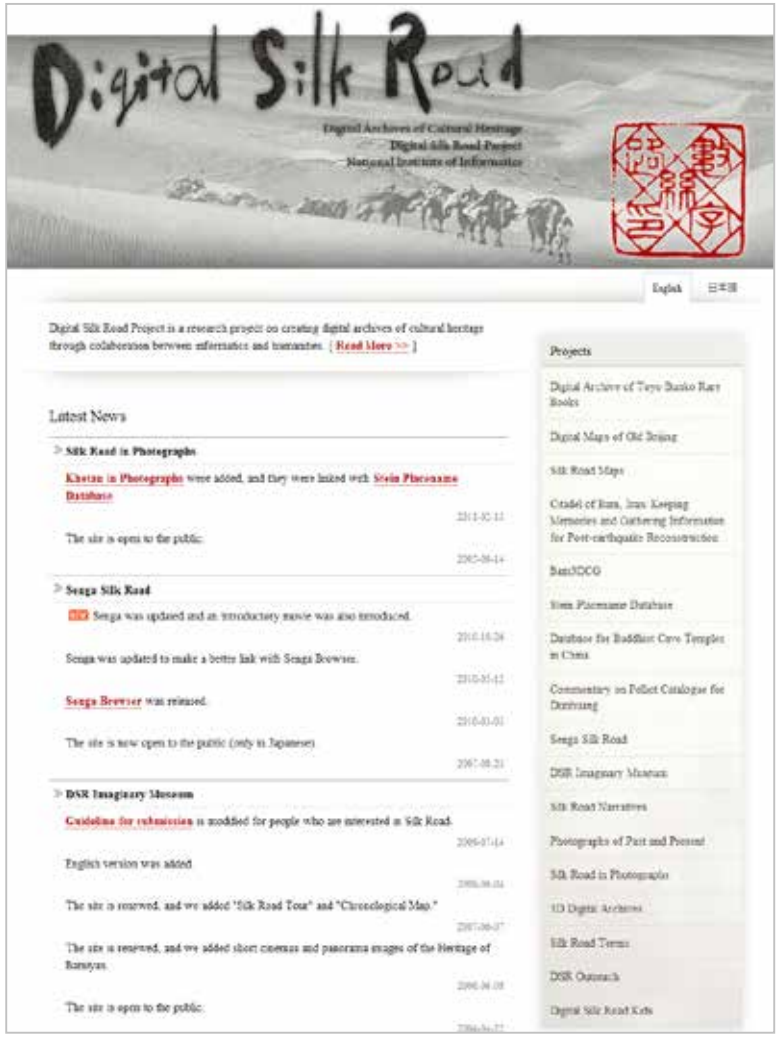
- Good example of data-driven humanities.
- 5,199,106,500 pages (as of Jan.22, 2017)
- A database where you can ask many interesting questions.
- Quantitative evidences can be obtained.

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.

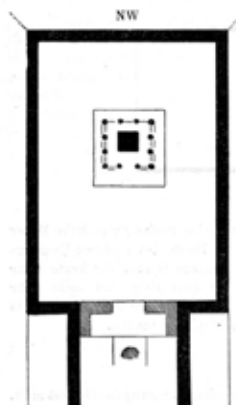
Digital Archive of Toyo Bunko Rare Books

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).

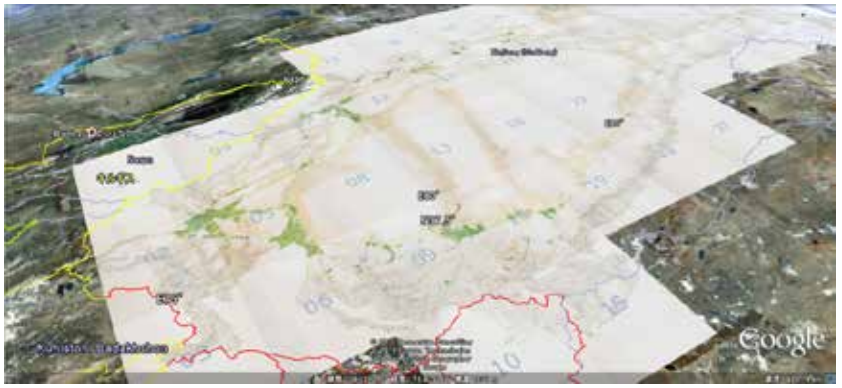
Variety and Heterogeneity of Data

Text



die obere sich wie eine in eine niedrigere 3,10 m tiefe Plattform eingepaßte Bank darstellt (auf der Skizze schraffiert) und die Mitte offen läßt. Vor dieser großen Unterstufe liegt der Rest eines mächtigen Sockels, in welchem ein tiefes Loch sich zeigt: hier hat also wohl eine große Statue oder eine Fahne gestanden. 12 m nach innen zu vom S.-Rand der Plattform des Hauptbaues, 5,50 m von den Seitenmauern und 7 m vor der Rückmauer, erhebt sich eine niedrige, 8 m ins Geviert betragende Stufe, auf deren Mitte ein jetzt zerstörter, 2 m großer, viereckiger Sockel steht; um diesen Sockel geht ein Gang herum, vorne und an den Seiten je 1,50 m breit, hinten aber nur 90 cm breit. Dieser Umgang ist nach außen von einer Mauer umgeben, welche durch zwölf kleine Säulen in kleine Abteile geteilt ist, von denen der mittlere der Frontseite den Eingang bildet. Auf der Rückseite ist dies aus zwei Eck- und zwei Mittelsäulchen bestehende System sehr zerstört. Vor den sechs Interkolumnien der Seiten und den zwei Interkolumnien neben dem Eingang sind je noch Sockel für Statuen erhalten: auch mancherlei dekoratives Bei-

Map



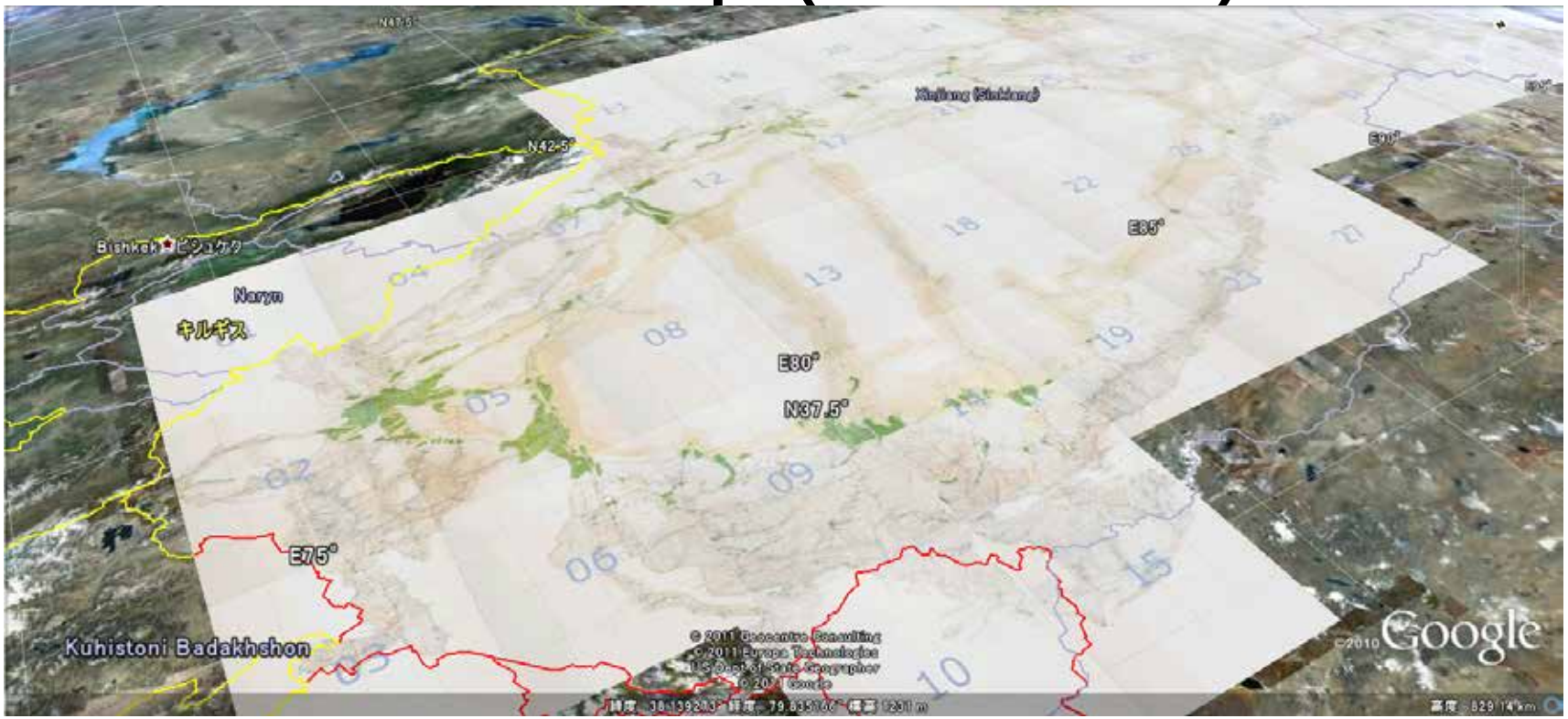
Photograph

1904(Le Coq, 1913, Tafel. 70, I)

Gazetteer

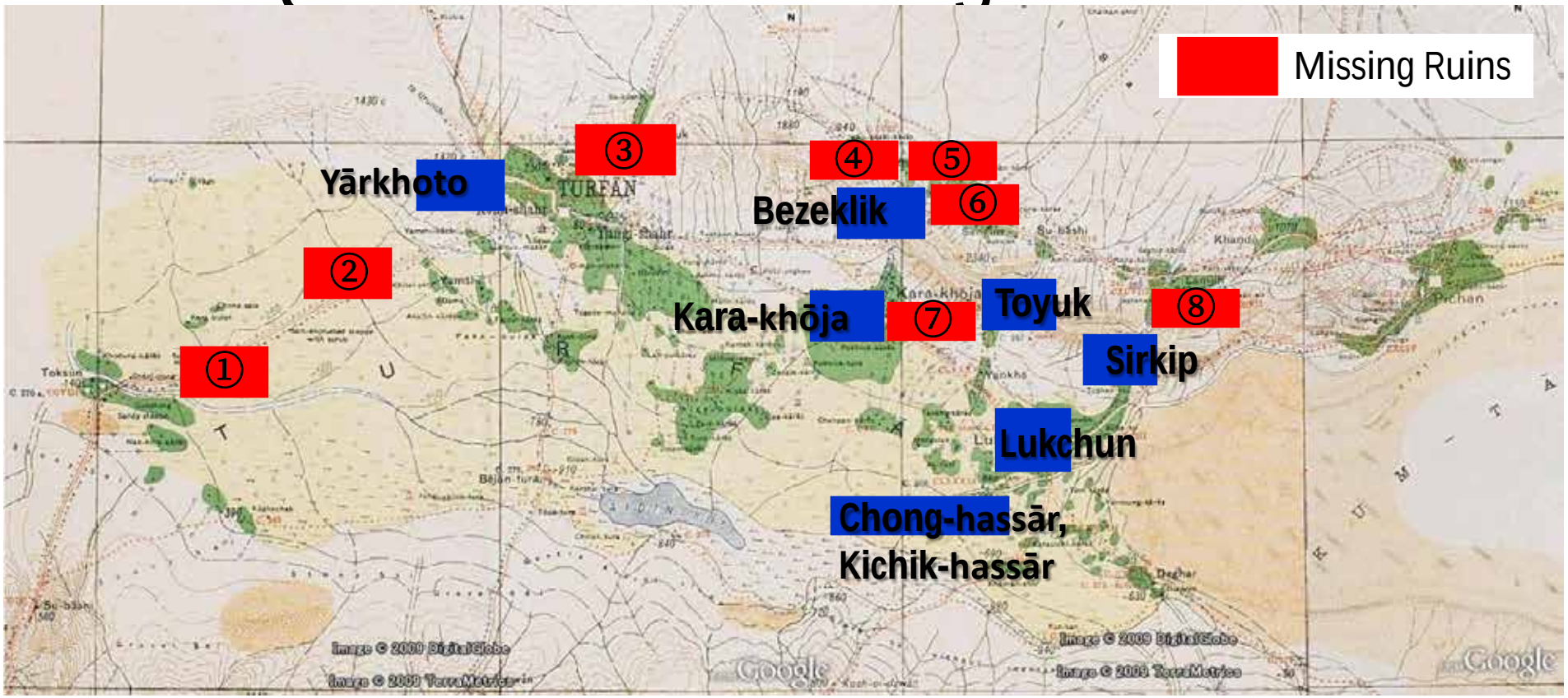
Abāb-langar, habit., 14. B. 3.
Ābād (of Ak-su), market-town, 12. A. 3.
Abād (of Kara-yulghun), vill., 12. B. 1.
Abād (of Karghalik), vill., 5. C. 4.
Ābād (of Kāshgar), vill., 5. A. 2.
Ābād (of Turfān), vill., 28. C. 3.
Ābād (of Yārkand), vill., 5. C. 2.
Ābād-jilga, valley, 12. B. 2.
Abdal, vill,. 30. B. 2.
Abdalkash-mazār, shrine, 14. C. 3.
Abdul-ghafūr-langar, loc., 10. C. 1.
Abdul-rahmān-jilga, valley, 9. A. 4.
Abshak-bēl, Pass, 2. B. 1.
Ach-tāgh, hill and vill., 7. C. 2.
Acha-dong (of Chizghān), hill, 19. C. 3.
Acha-dong (of Yārkand R.), loc., 7. D. 4.
Acha-kuduk, loc., 7. D. 4.
Acha-shīpang, loc., 22. D. 4.
Achak-aghzi, loc., 5. A. 4.
Achal (of Ak-su), vill., 12. A. 3.
Achal (on Charchak R.), loc., 21. C. 2.
Achchik-bulak (of Turfān), spring, 28. B. 4.
Achchik-bulak (of Yai-döbe), spring, 4. C. 4.
Achchik-daryā, river, 21. A. 2.
Achchik-dawān, pass, 9. B. 3.
Achchik-jilga (of Duwa), valley, 9. B. 3.
Achchik-jilga (of Kara-tāsh), valley, 2. D. 3.
Achchik-jilga (of Khotan), valley, 9. C. 3.
Achchik-jilga (of Sampula), valley, 14. A. 3.
Achchik-jilga (of Tawak-kēl), loc., 14. A. 1.
Achchik-kōl, lake, 15. D. 1.
Achchik-kuduk (of Kapa), well, 23. A. 1.
Achchik-kuduk (of Kuruk-tāgh), well, 28. C. 4.
Achchik-kuduk (of Marāl-bāshi), well, 5. D. 2.
Achchik-otan, loc., 7. C. 2.
Achchik-su, loc., 31. A. 4.
Achchik-tügemen, loc., 5. D. 2.
Achi-tāgh, hill, 32. B. 1.
Achik-aghzi, loc., 9. D. 3.
Achma (of Hanguya), vill., 14. A. 2.

Stein Map (Silk Road)



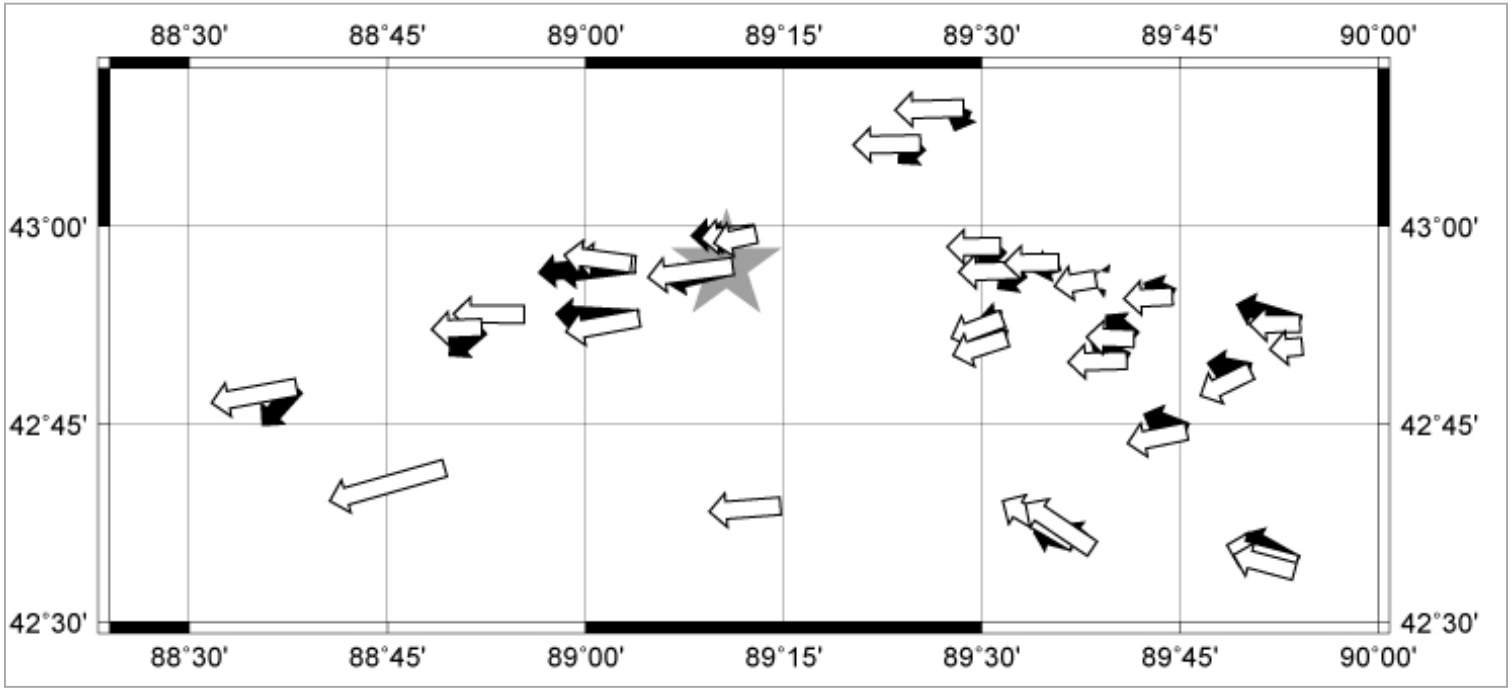
- Stein's map "Innermost Asia" was registered and displayed on Google Earth satellite images.

Question: “Missing” Ruins?



① Oi-tam, ruined fort
② Bögan-tura
③ Buluyuk (Shipang, Sassik-bulak, Kazma)
④ Murtuk-ruins
⑤ Yoghan-tura
⑥ Chikkan-köl
⑦ Bedaulat’s town, Bēsh-kāwuk, Kosh-gumbaz
⑧ Yutōgh

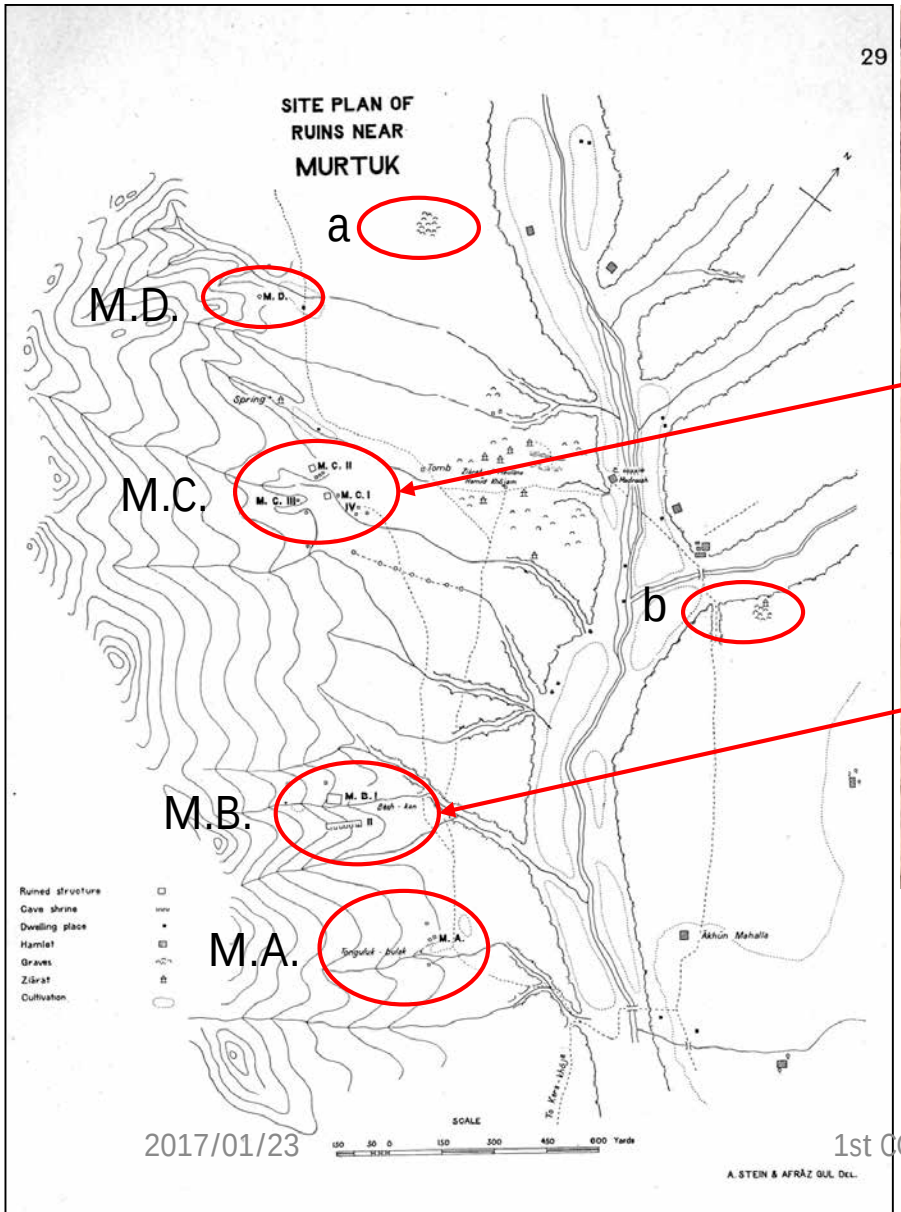
Error Distribution in Turfan

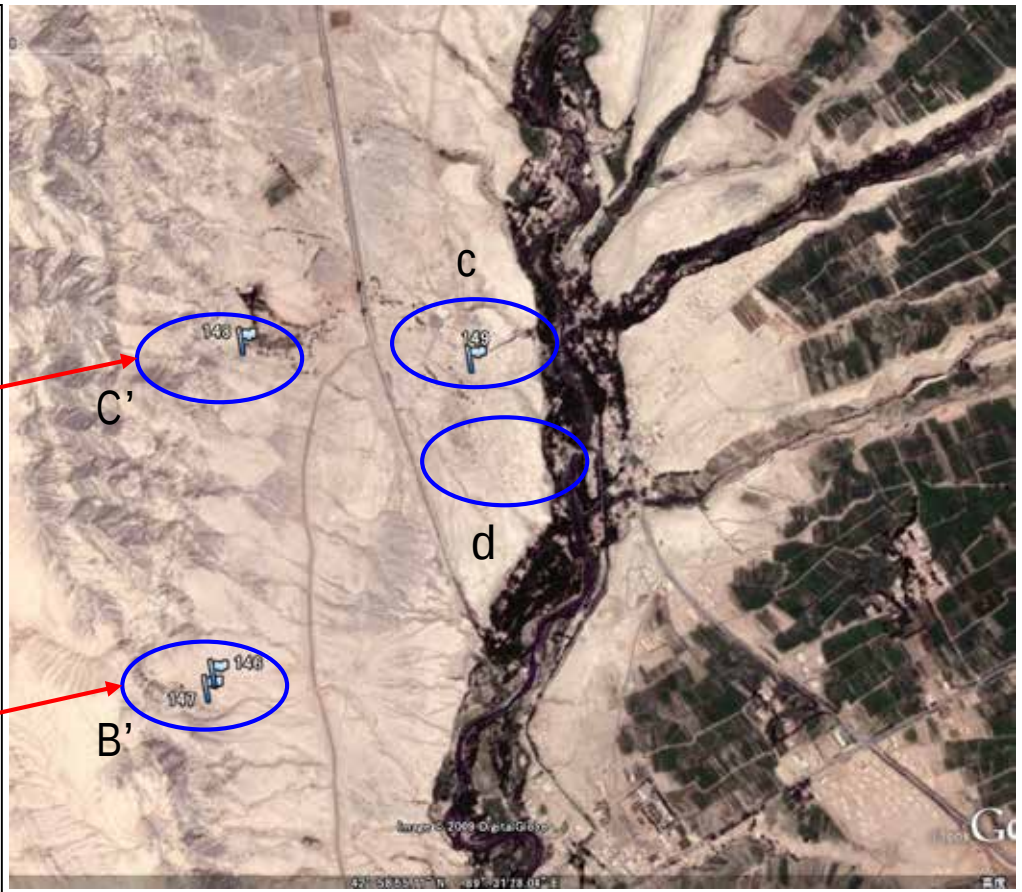


Error Distribution in Turfan Basin / White: Innermost Asia / Black: Serindia

- Some ruins were reported by 20th expeditions, but are missing in recent survey reports.

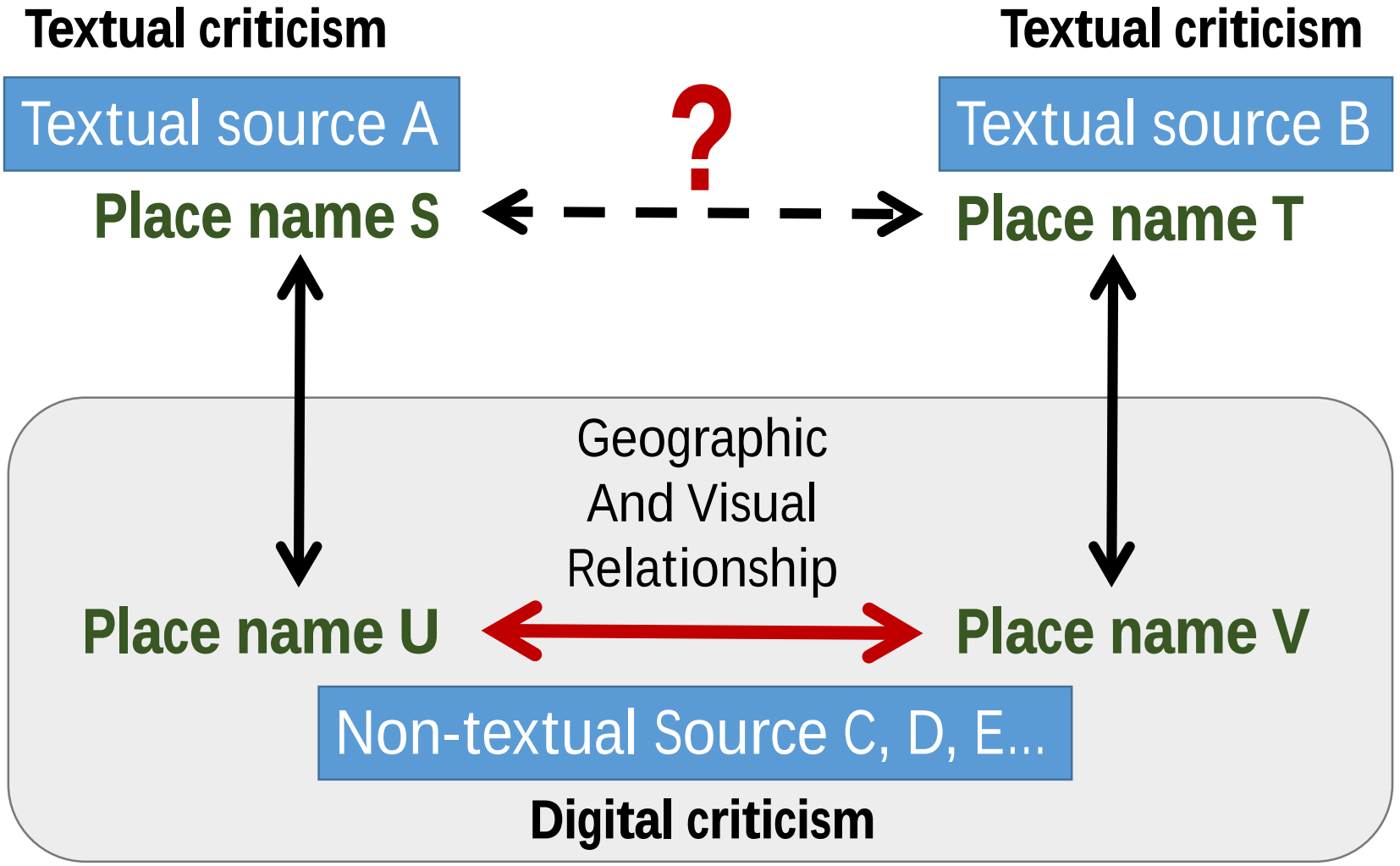
Matching Entities





Stein's map and satellite images for the same area. Each source reports different ruins due to different conceptualization.

Linking Entities Across Sources



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.

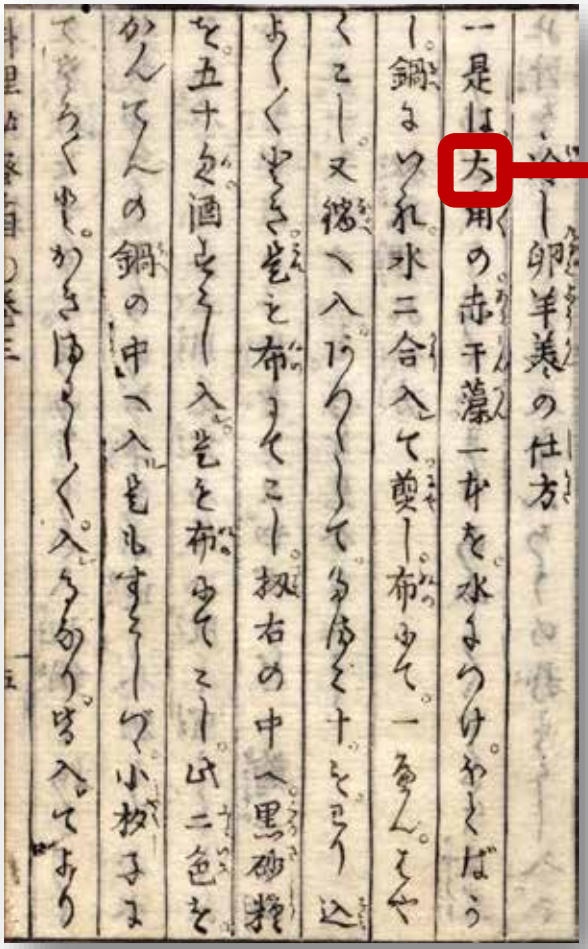
Released on November 10, 2016

Open Data for Scholars

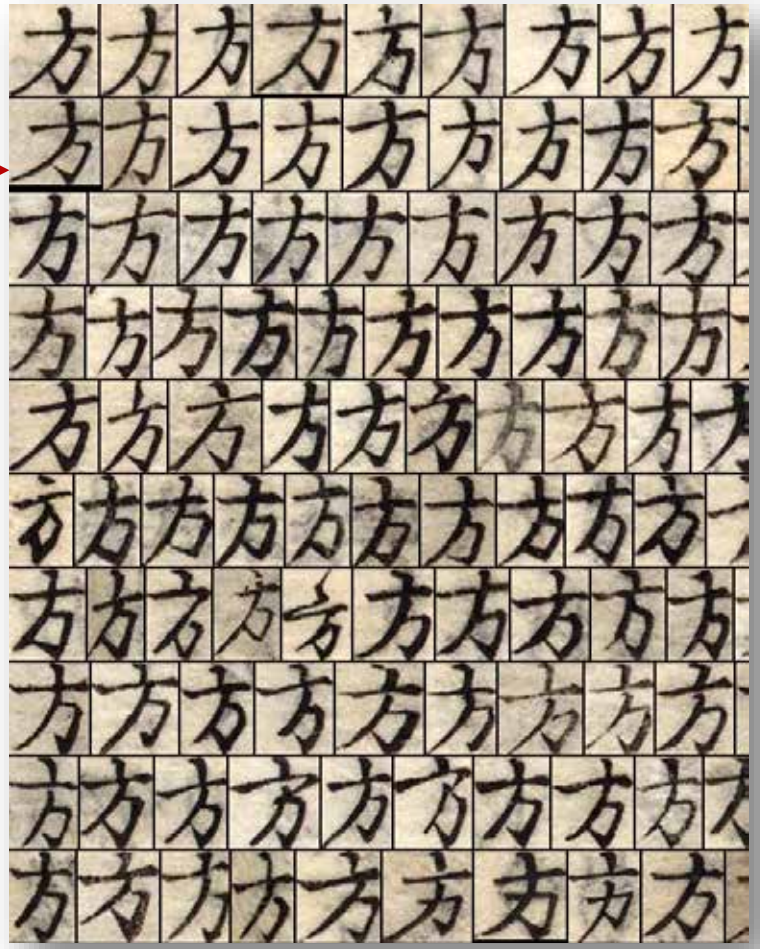
Pre-Modern Japanese Text Dataset (from NIJL)

Released on November 17, 2016

Open Data for Machines



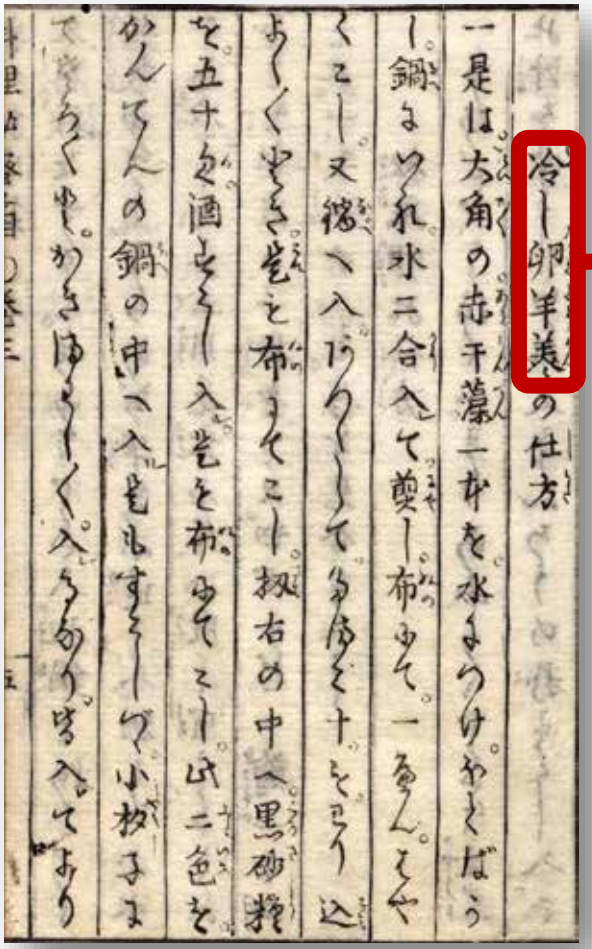
PMJT Dataset (from NIJL)



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 Curation Viewer
Core contributor: Jun HOMMA

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

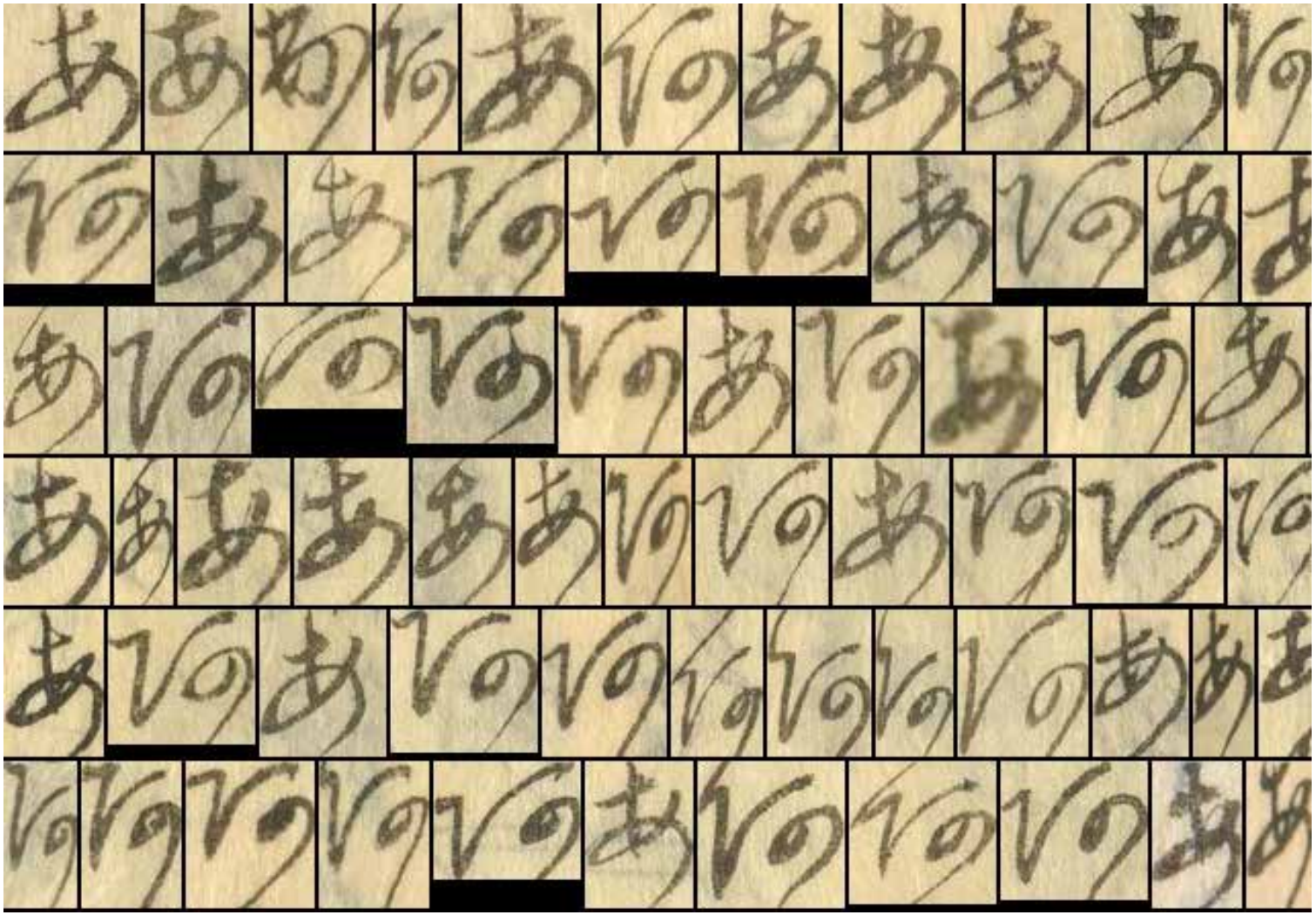
Data Identifiers

- 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.

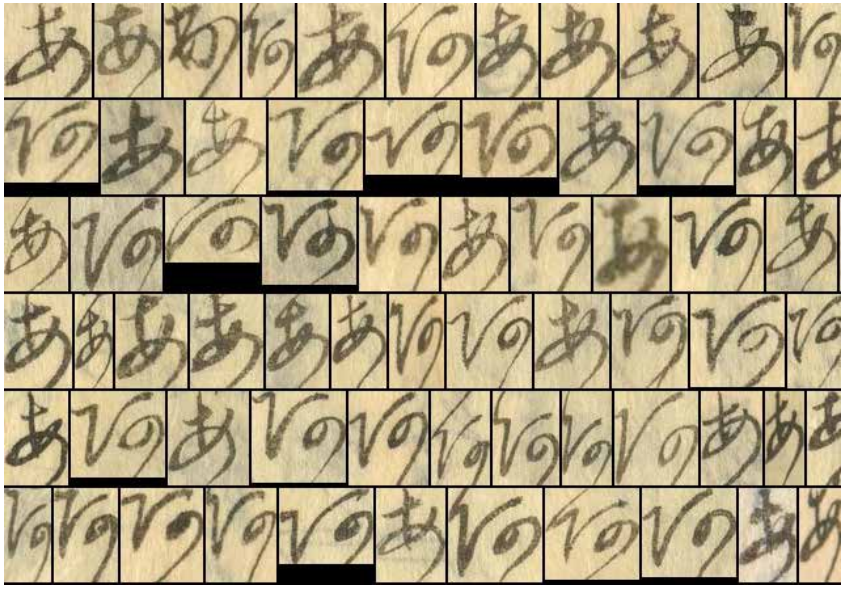
2. PMJT Character Shape Dataset

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

Types	Frequency
し	3,929
に	3,147
の	2,908
て	2,398
り	2,193
を	2,021
か	1,910
く	1,739
き	1,715
も	1,463
1,521 types	86,176 characters

- 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.

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

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

3. Edo Cooking Recipe Dataset

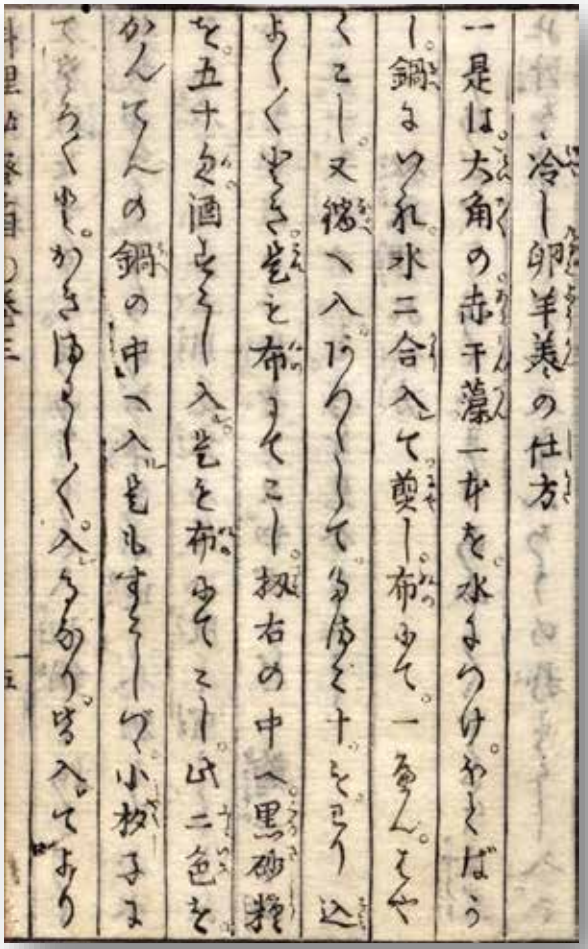
http://codh.rois.ac.jp/edo-cooking/

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

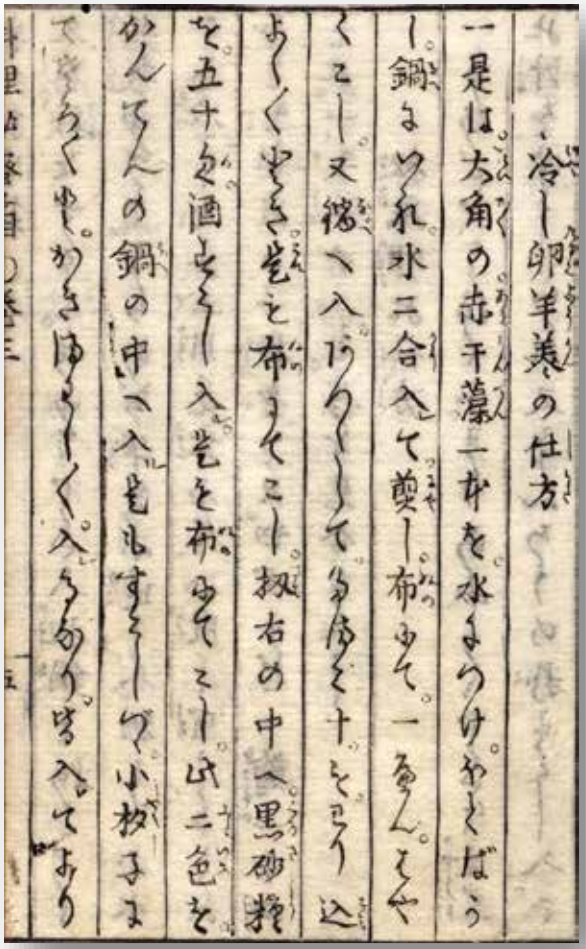


PMJT Dataset (from NIJL)

1	是は 大角の 赤干藻一本を 水につけ ほとばかし
2	鍋にいれ 水二合入レて 煎し 布にて 一へん はやくこし 又鍋へ入レ あつくして
3	たまご十ウを わり込よくよくとき 是も布にてこし
4	扨右の中へ 黒砂糖を 五十匁 酒すこし入ル 是も布にてこし
5	此二色を かんてんの鍋の中へ入ル
6	是もすこしづつ 小杓子にて そろそろと かきまわしかきまわし 入レるなり
7	皆入レてより 又葛粉をすこし 水にてとき入レ
8	扨鍋をぬき 早く折敷にても うちあげ 平めに延し 入レ物ともに 水に入レ 冷し遣ふ

Edo Cooking Recipe Dataset (Created by CODH)

3. Translation

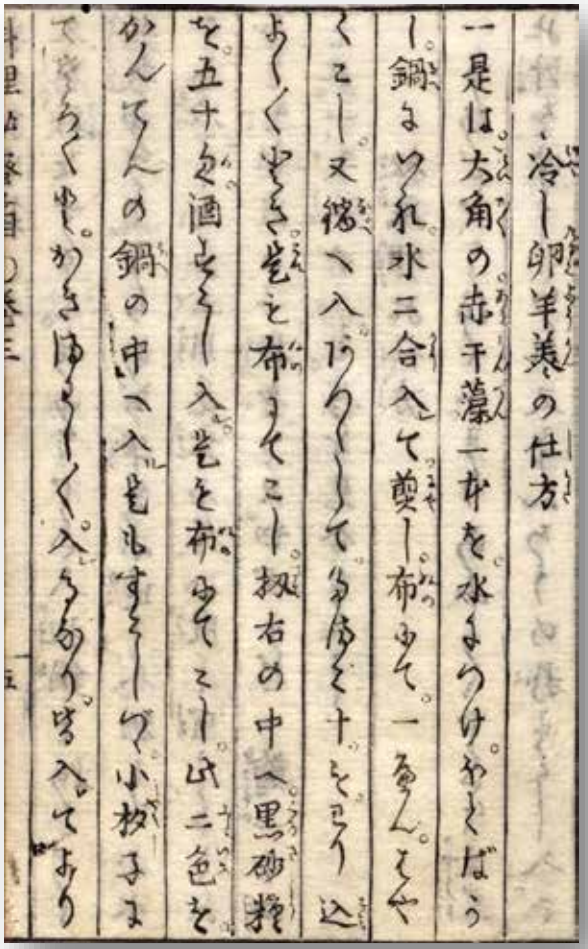


PMJT Dataset (from NIJL)

1	大きな赤寒天を1本水に付けてふやかす。
2	鍋に寒天と水2合（360cc）を入れて煮溶かす。
3	②を一度布で素早く漉し、再び鍋に入れて熱する。
4	生卵10個をよく溶き、布で漉す。
5	④の中に黒砂糖50匁（200g）と酒少しを入れ、布で漉す。
6	⑤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。
7	⑤を全て鍋の中に入れたら、葛粉を水で溶き、鍋に入れる。
8	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。

Edo Cooking Recipe Dataset (Created by CODH)

4. Adaptation



PMJT Dataset (from NIJL)

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

Edo Cooking Recipe Dataset (Created by CODH)

Photographs for the Recipe

Edo Cooking Recipe Dataset (Created by CODH)

Edo Recipe Cooking Dataset

http://codh.rois.ac.jp/edo-cooking/

1. Transcription: **107**
2. Translation: **20** out of 107.
3. Adaptation: **5** out of 20.

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

Dataset Release at ‘Cookpad’

http://cookpad.com/recipe/4153357

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.

Unexpectedly Large Impact

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

https://twitter.com/jouhouken/status/801693251052781568

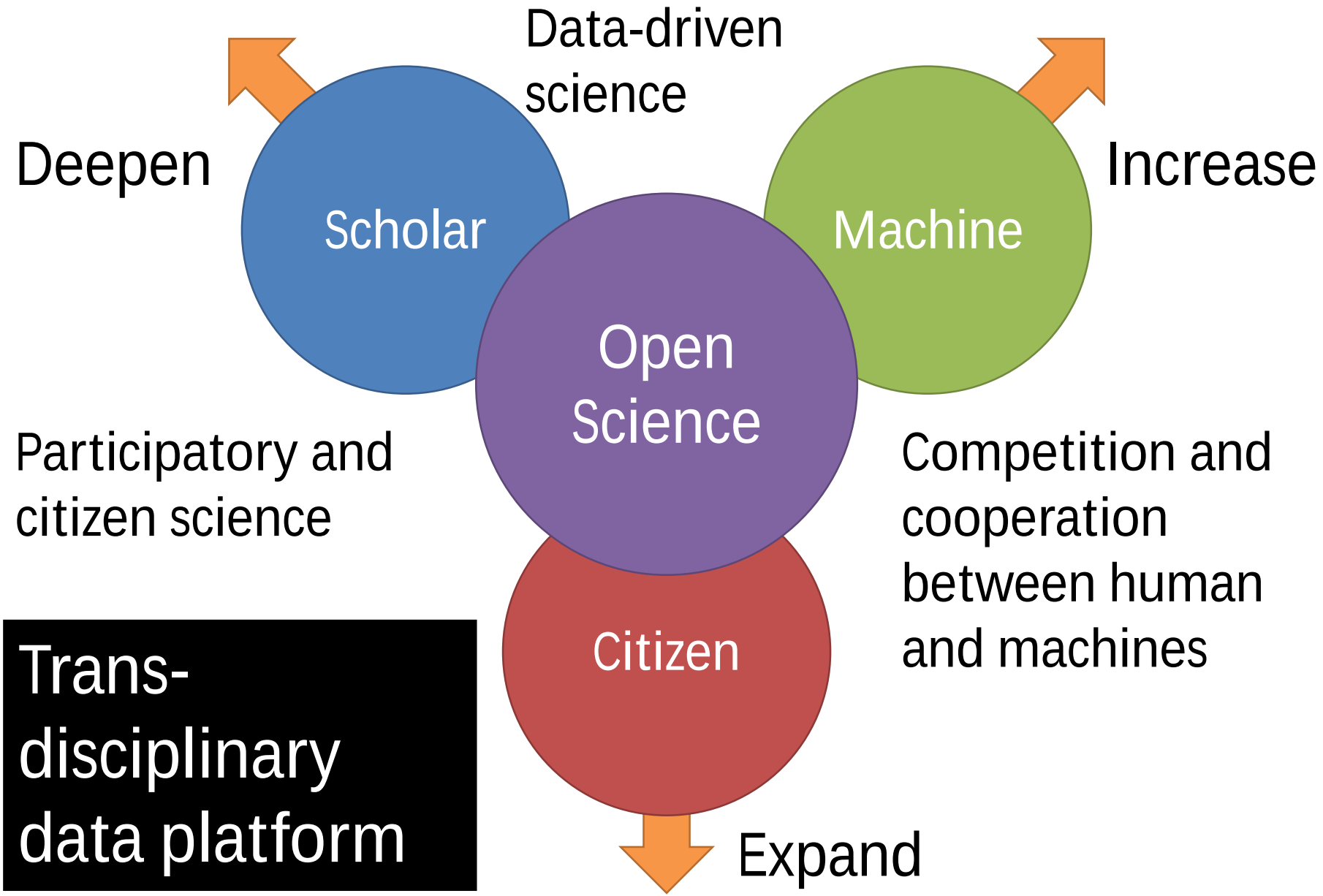
TV show to reproduce the dish

http://www.news24.jp/articles/2016/11/30/07347892.html

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


Data-driven science
Deepen
Increase
Scholar
Machine
Open Science
Participatory and citizen science
Competition and cooperation between human and machines
Citizen
Trans-disciplinary data platform
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/