

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

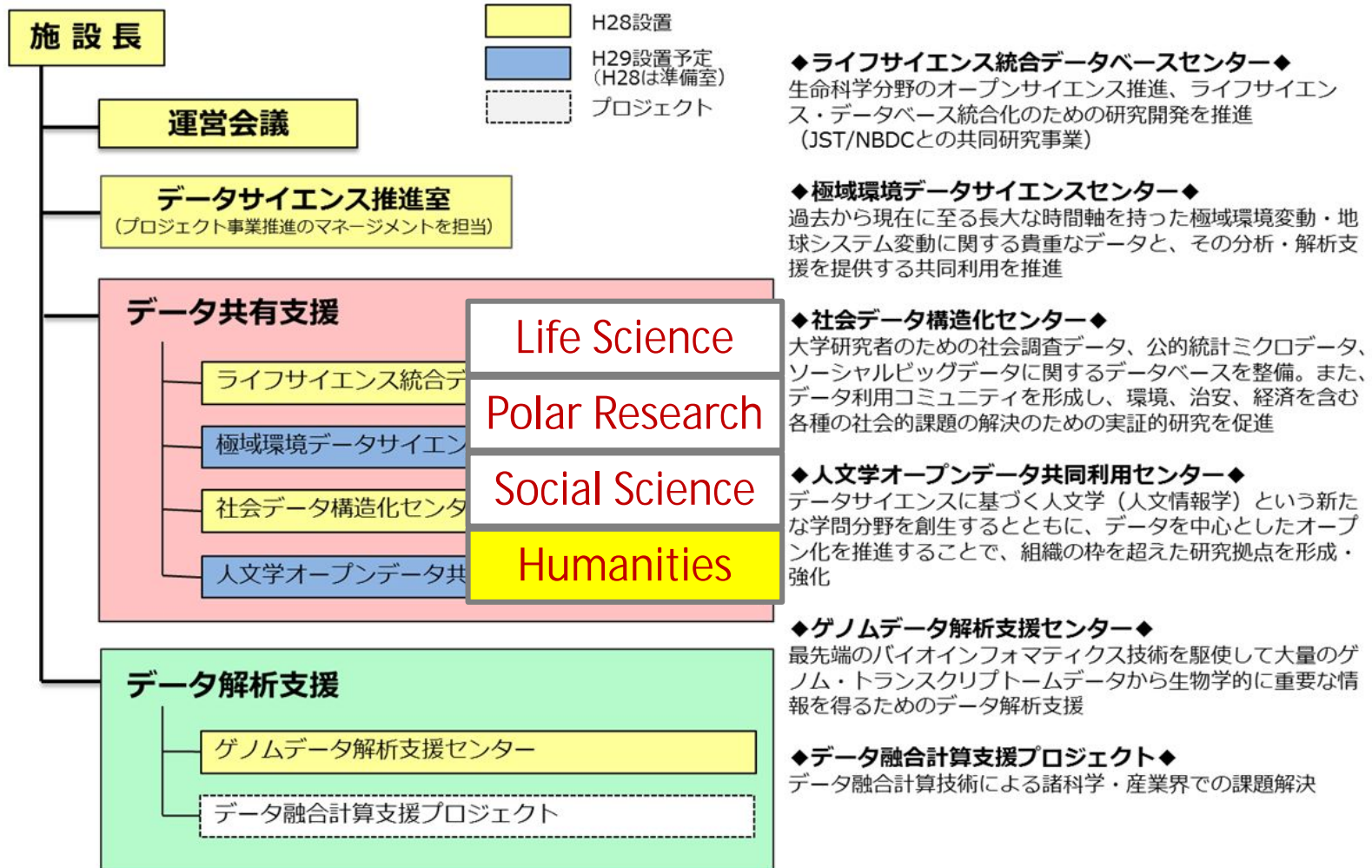


What is CODH?

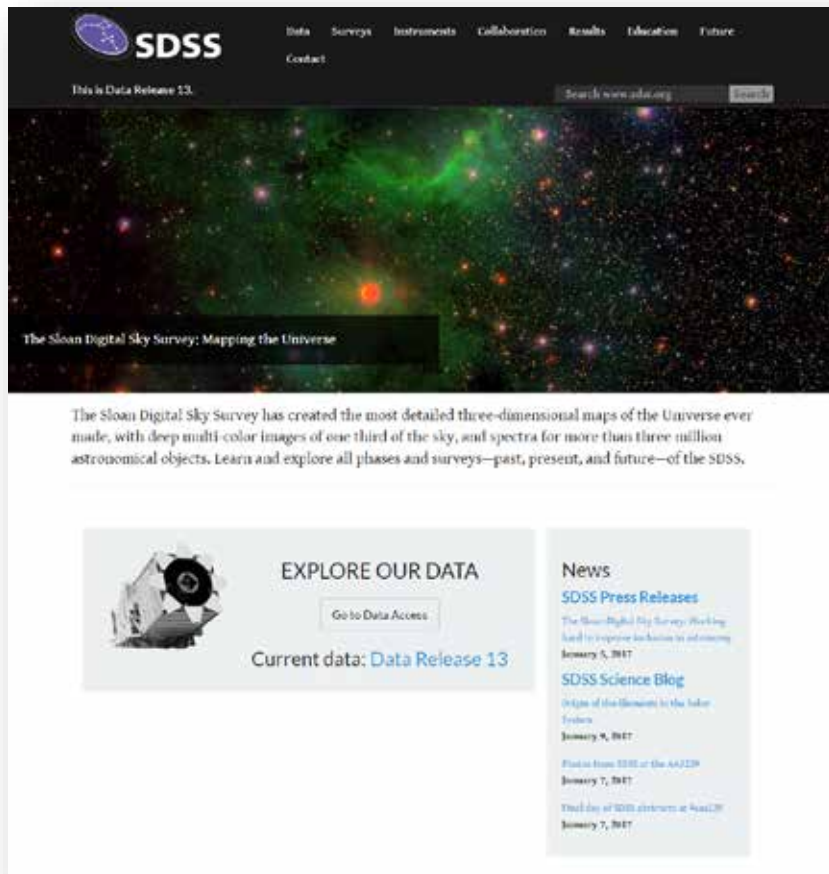
<http://codh.rois.ac.jp/>

- April 1, 2016: Established as a pre-center.
 - April 1, 2017: Officially launched (I hope).
 - ROIS > Joint 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



Sloan Digital Sky Survey

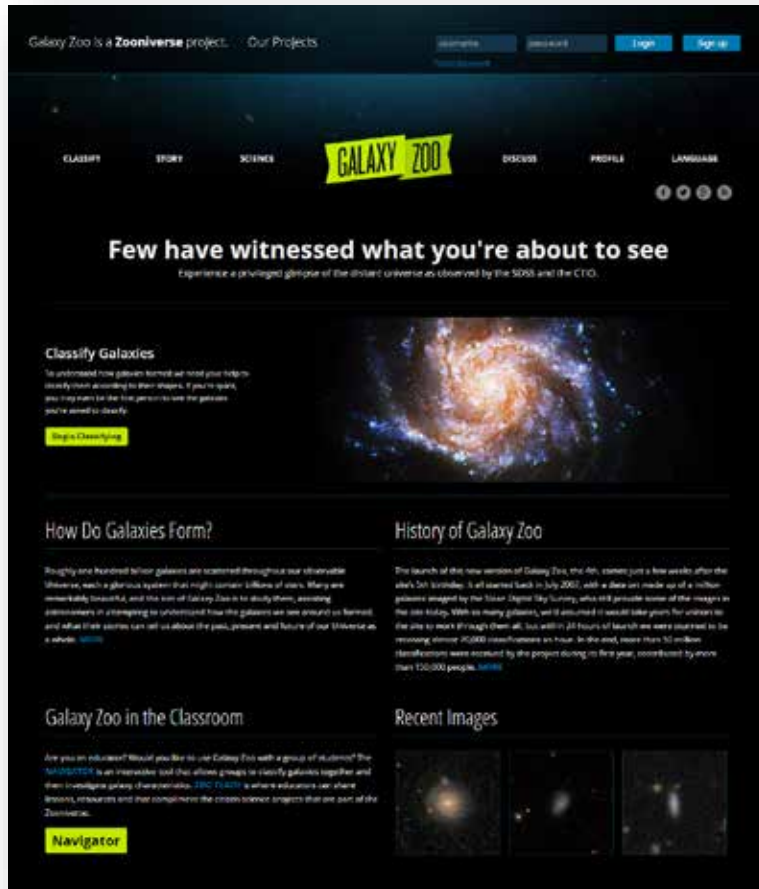


The screenshot shows the homepage of the Sloan Digital Sky Survey (SDSS). At the top, there is a navigation menu with links for Data, Surveys, Instruments, Collaboration, Results, Education, Future, and Contact. Below the menu is a search bar with the text "Search www.sdss.org" and a "Search" button. The main content area features a large, colorful image of a star field with a prominent green nebula. Below the image is a text box that reads "The Sloan Digital Sky Survey; Mapping the Universe". Underneath this is a paragraph of text: "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. Learn and explore all phases and surveys—past, present, and future—of the SDSS." Below the text is a section titled "EXPLORE OUR DATA" with a "Go to Data Access" button and the text "Current data: Data Release 13". To the right of this section is a "News" section with several links and dates: "SDSS Press Releases" (January 5, 2017), "SDSS Science Blog" (Origin of the Elements in the Solar System, January 9, 2017), "Photo from SDSS at the AAS129" (January 7, 2017), and "The Day of SDSS at AAS129" (January 7, 2017).

- 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

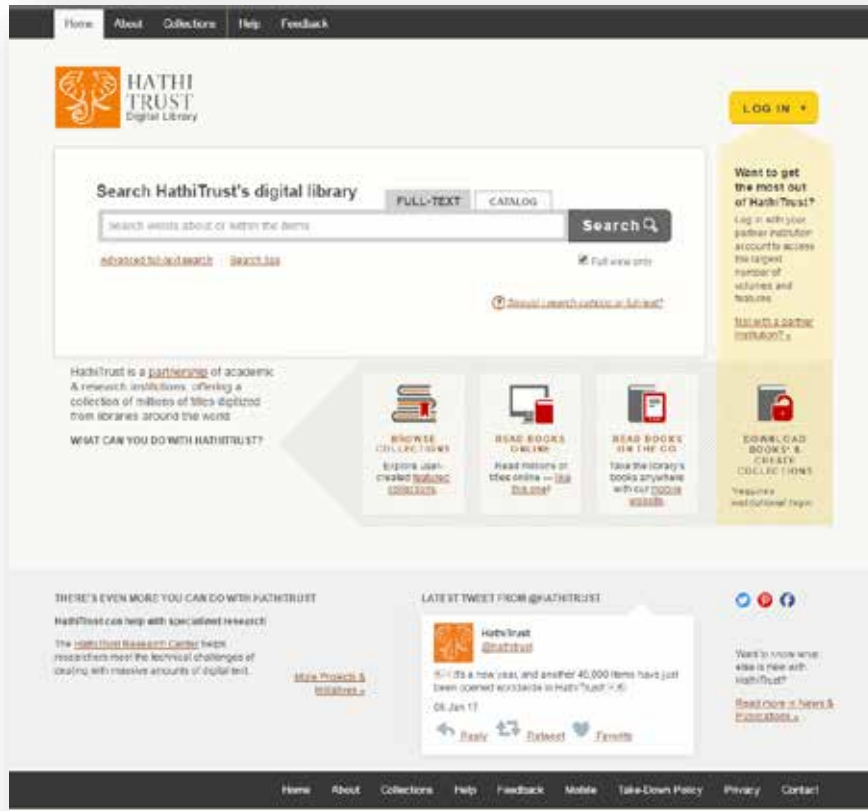
Small

2 Theory → Solution

3 Theory → Simulation

4 Data → Generalization

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.

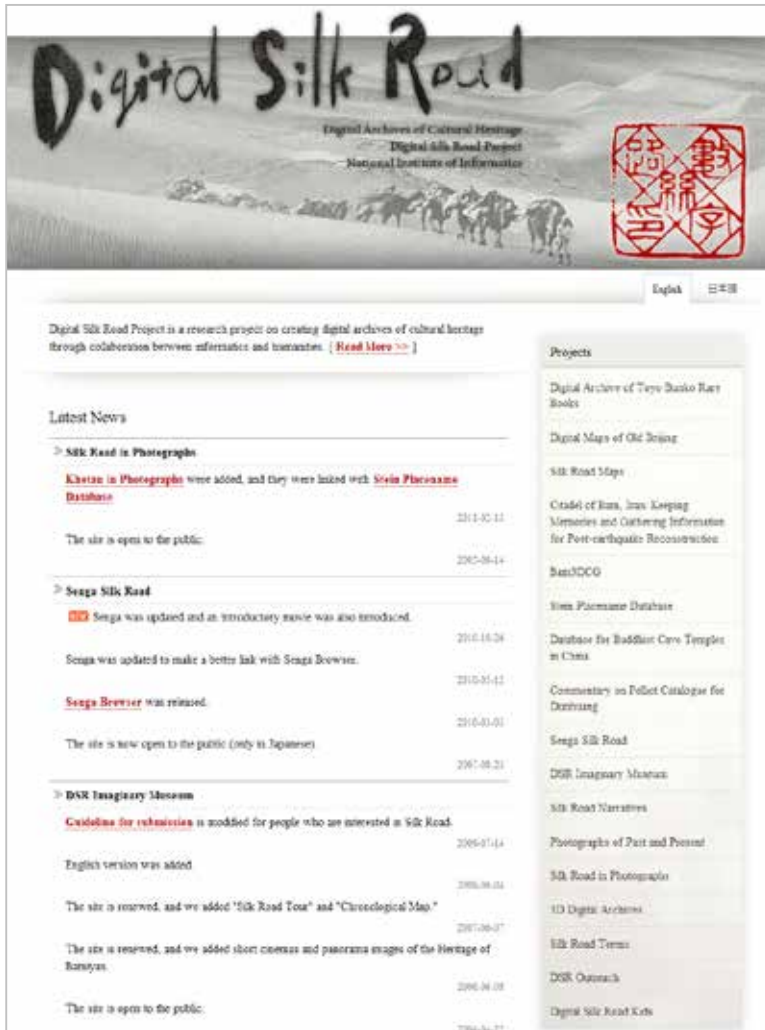
Digital Silk Road and Digital Humanities

<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.**





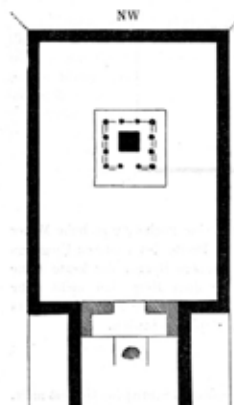
Digital Archive of Toyo Bunko Rare Books

<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).

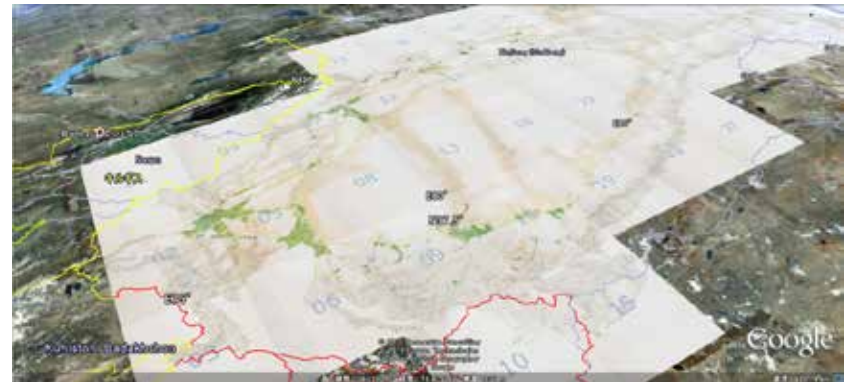
Variety and Heterogeneity of Data

Text



die obere sich wie eine in eine niedrigere 3,10 m tiefe Plattform eingebaute Bank darstellt (auf der Skizze schraffiert) und die Mitte offen liest. 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 Stäben in kleine Abteile geteilt ist, von denen der mittlere der Frontseite den Eingang bildet. Auf der Rückseite ist dies aus zwei Eck- und zwei Mittelstützen bestehende System sehr zerstört. Vor den sechs Interkolumnien der Seiten und den zwei Interkolumnien neben dem Eingang sind je noch Sockel für Statuen erhalten: nach mancherlei dekoratives Bei-

Map



Photograph

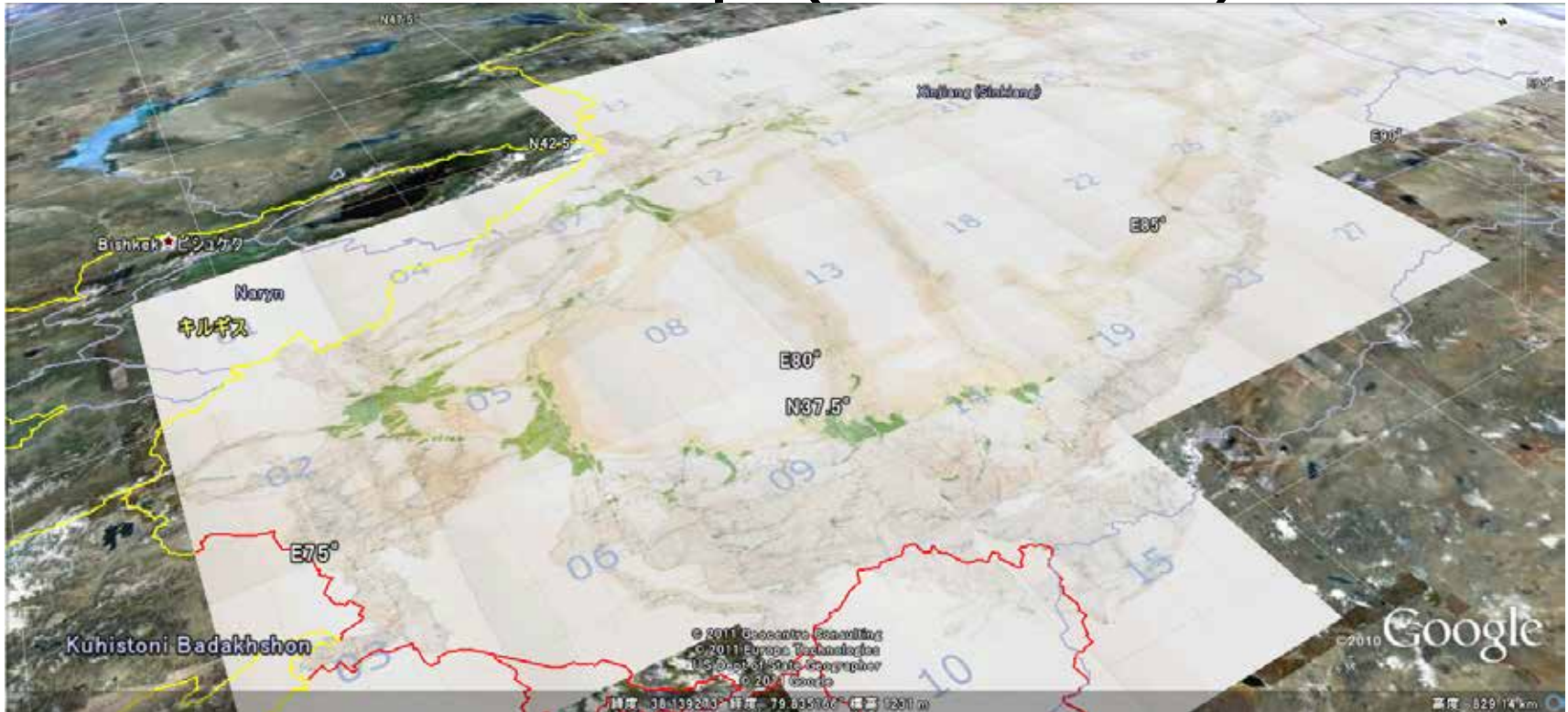


Y
1904(Le Coq, 1913, Tafel. 70, I)

Gazetteer

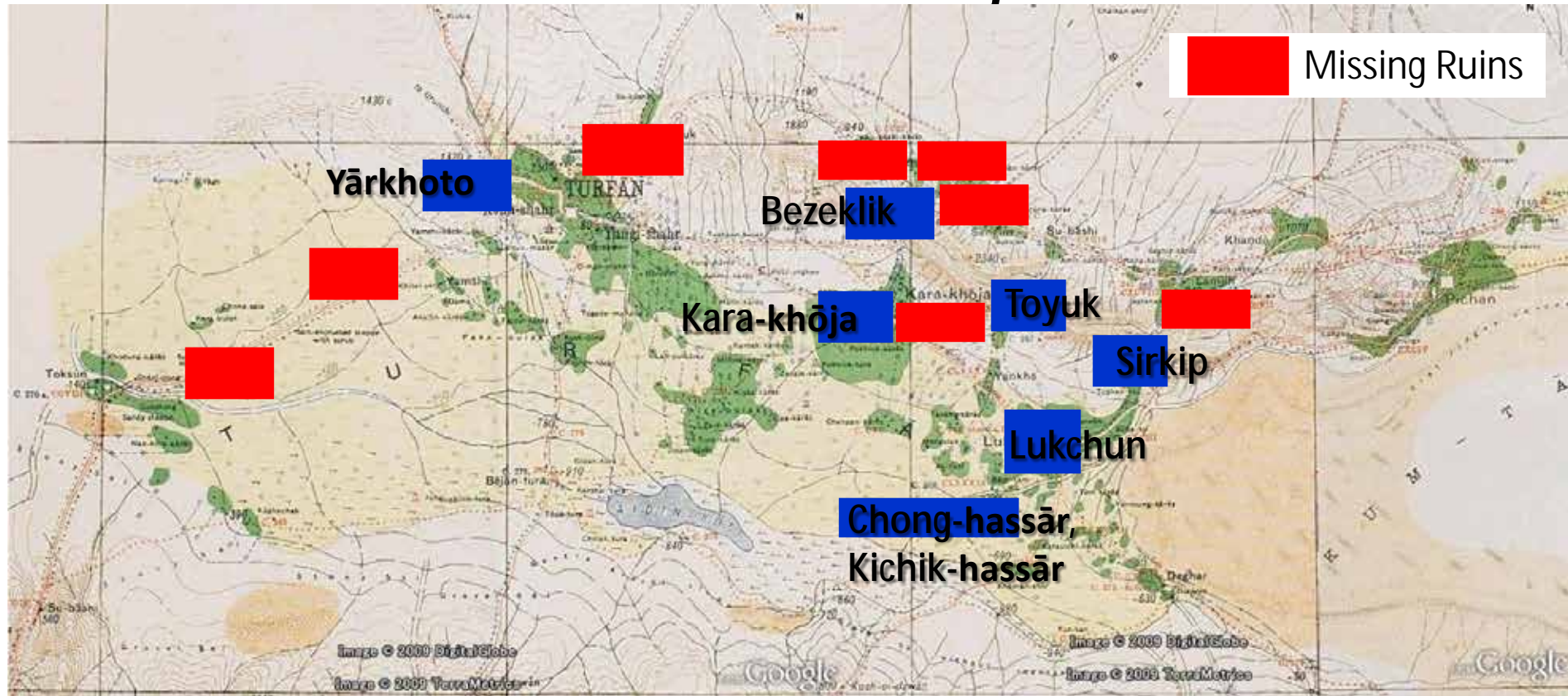
Abāb-langar, habit., 14. B. 3.	Aechhik-bulak (of Turfān), spring, 28. B. 4.
Ābād (of Ak-su), market-town, 12. A. 3.	Aechhik-bulak (of Yai-döbe), spring, 4. C. 4.
Ābād (of Kara-yulghun), vill., 12. B. 1.	Aechhik-daryā, river, 21. A. 2.
Ābād (of Karghalik), vill., 5. C. 4.	Aechhik-dawān, pass, 9. B. 3.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

Stein Map (Silk Road)



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

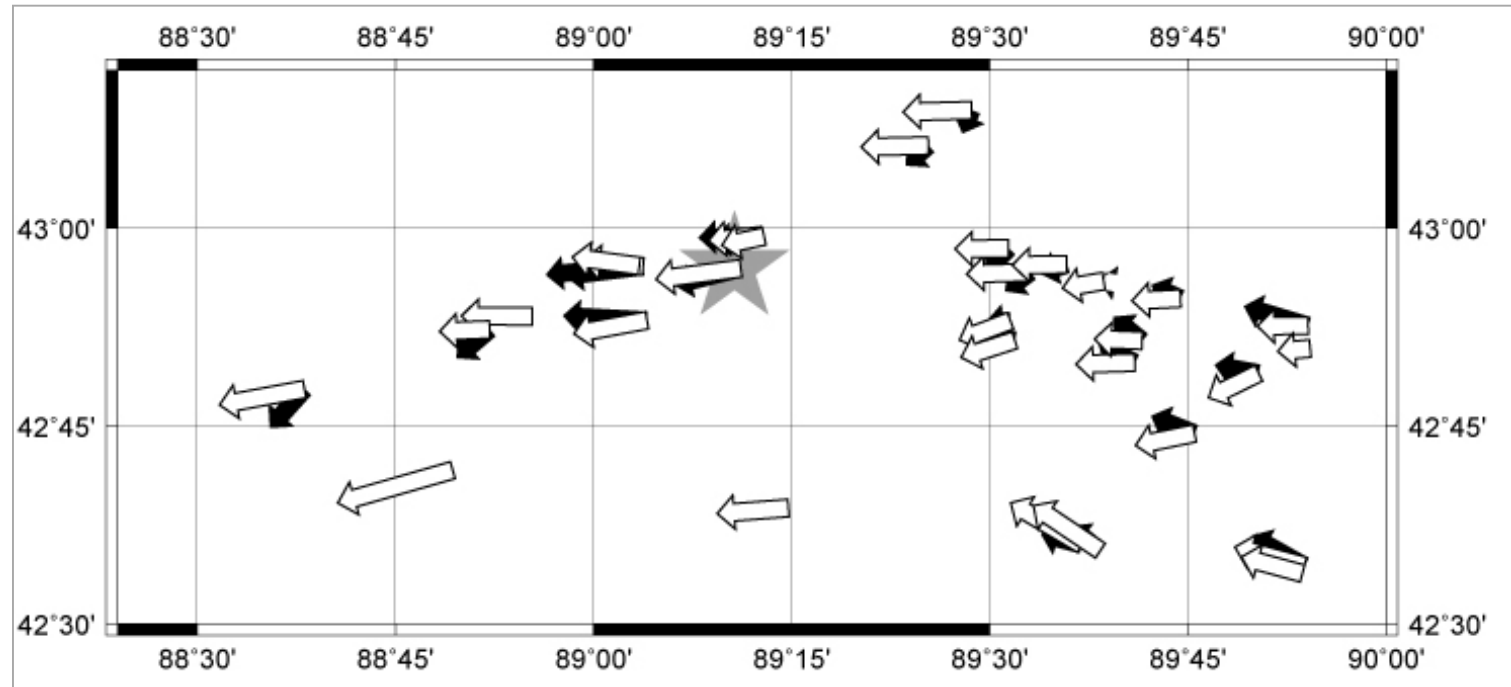
Question: "Missing" Ruins?



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

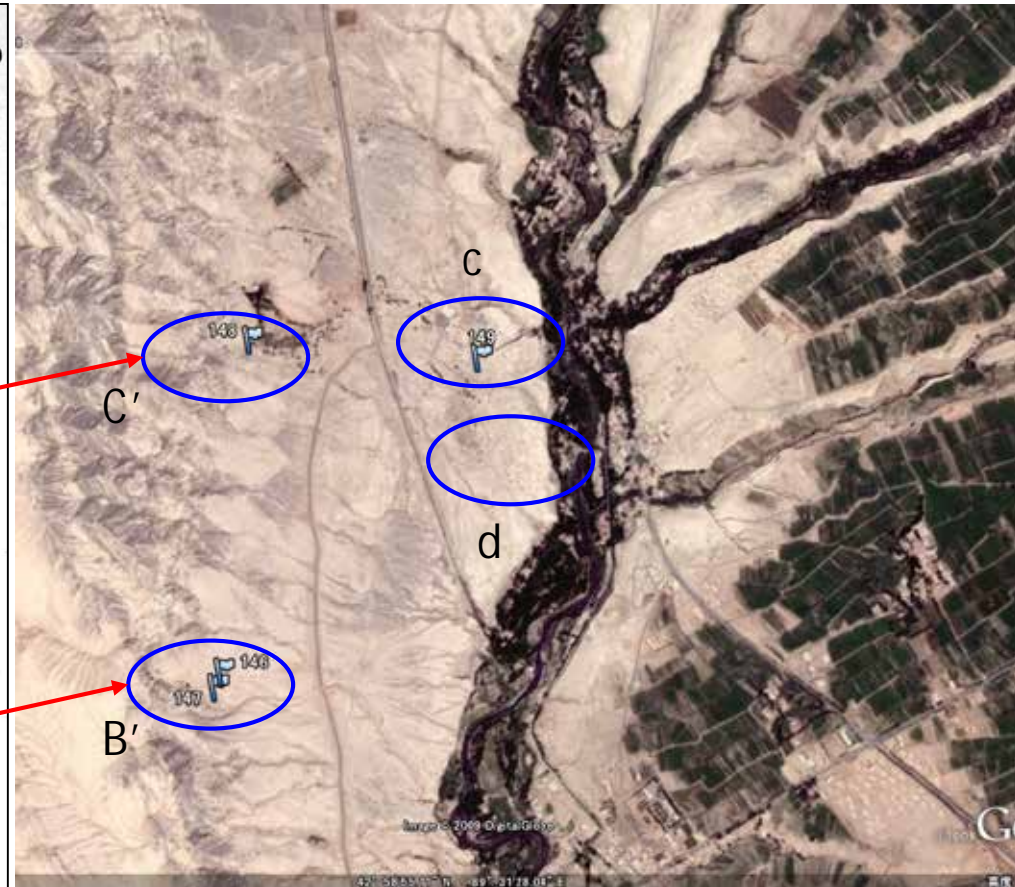
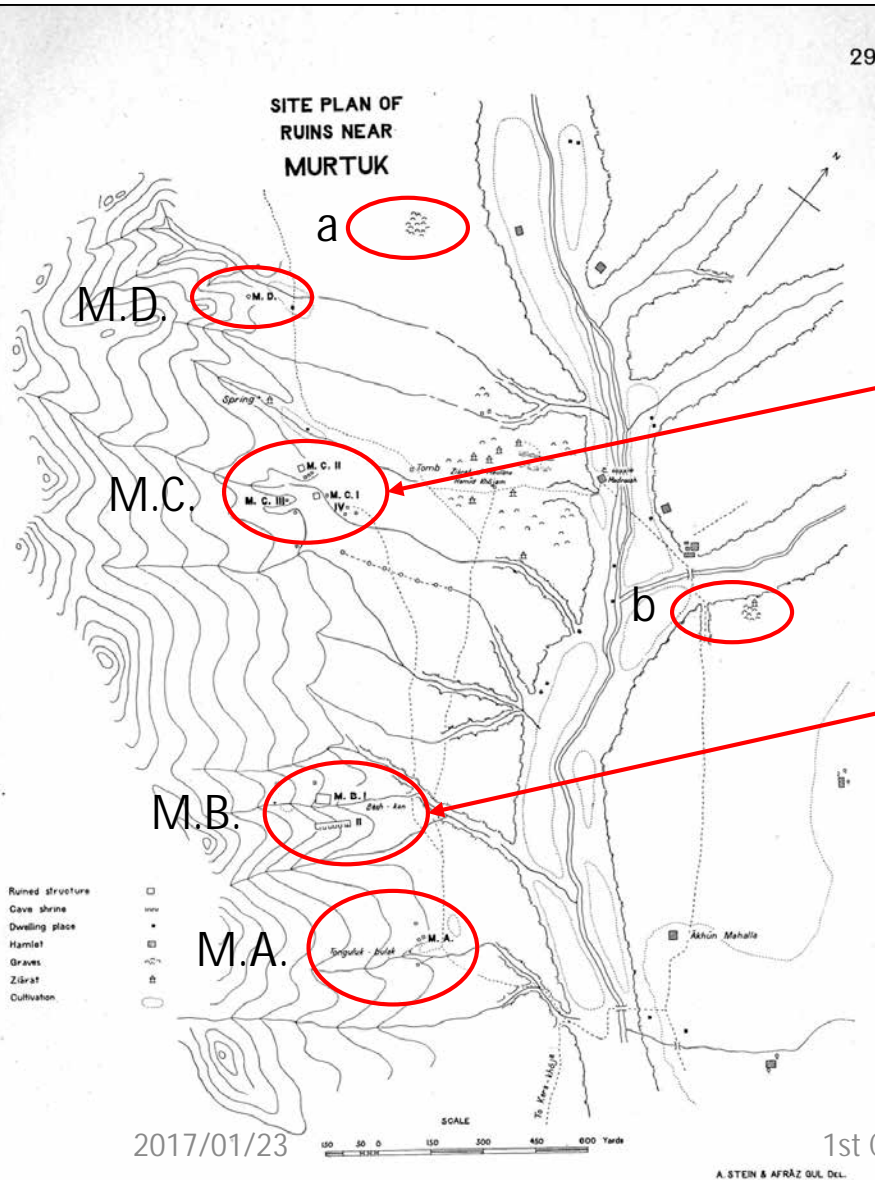
Error Distribution in Turfan



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

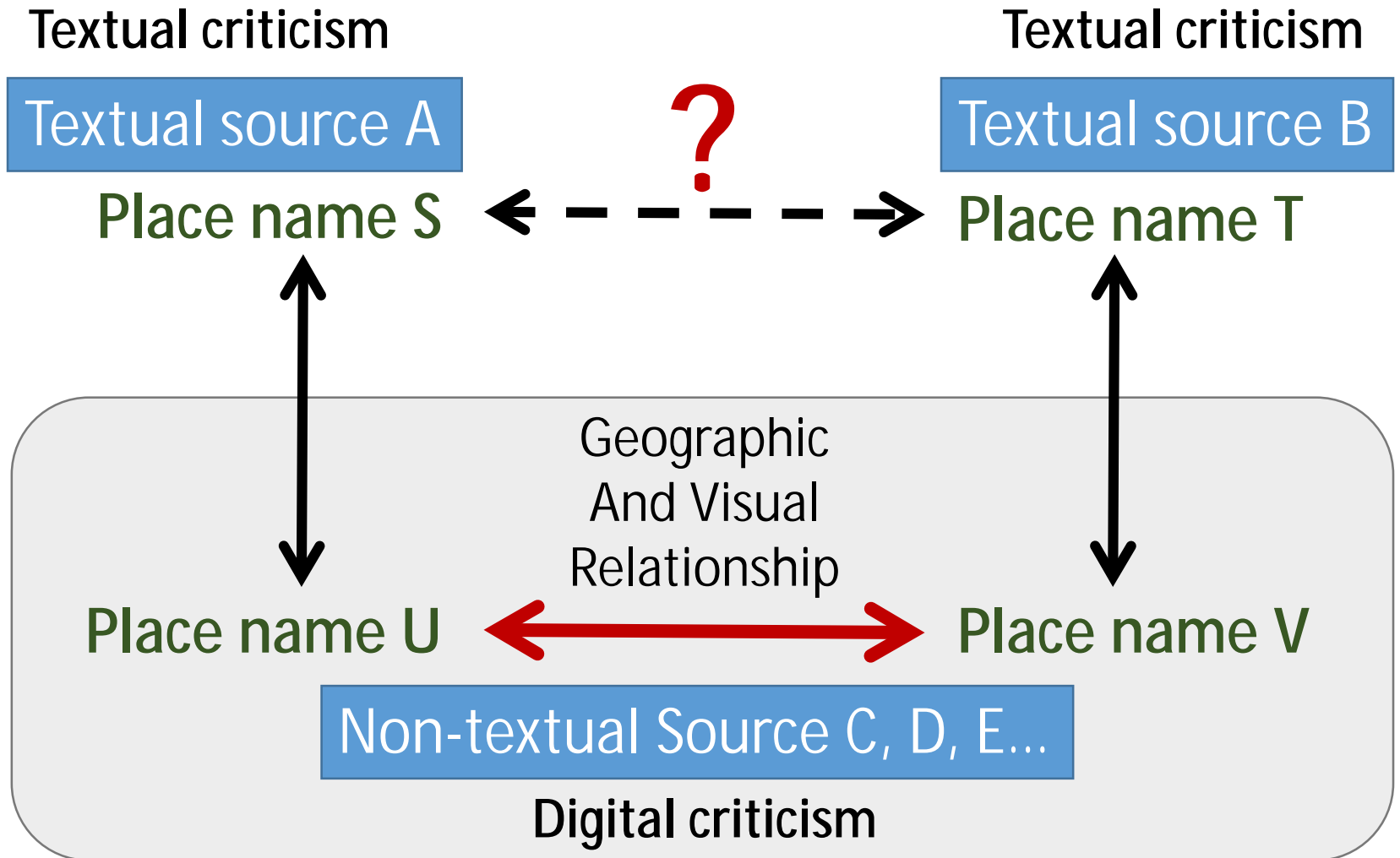
- Some ruins were reported by 20th expeditions, but are missing in recent survey reports.

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

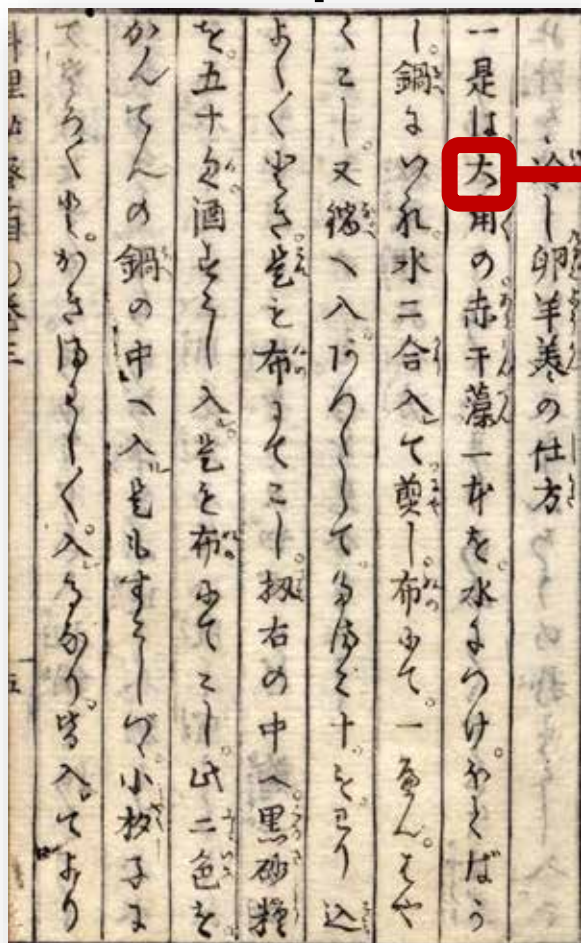
Open Data for Scholars



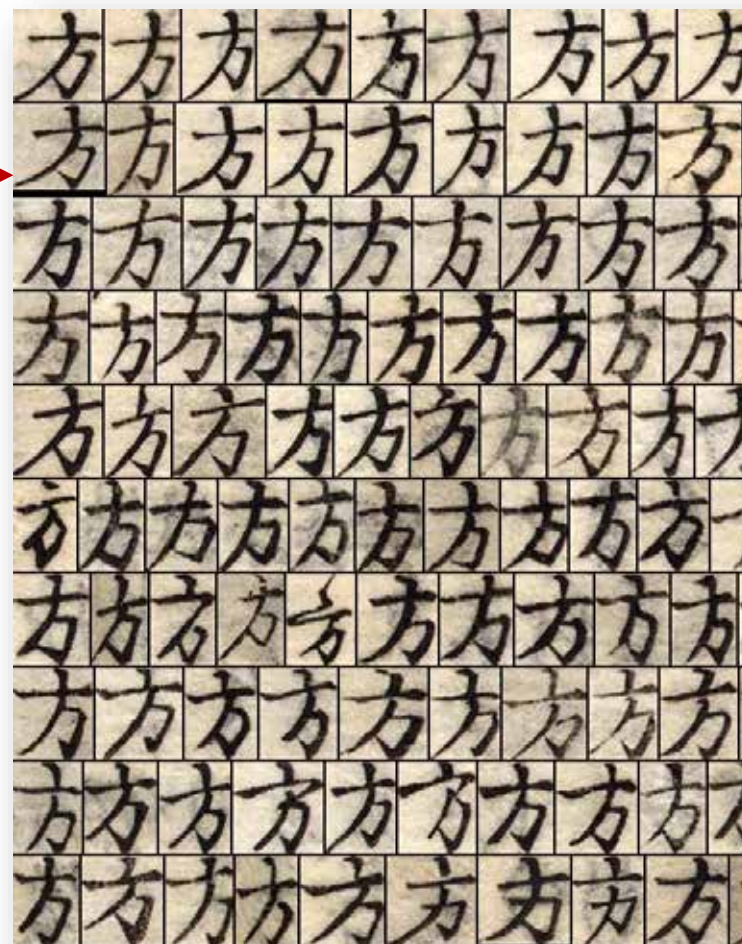
Pre-Modern Japanese Text Dataset (from NIJL)

Released on November 17, 2016

Open Data for Machines



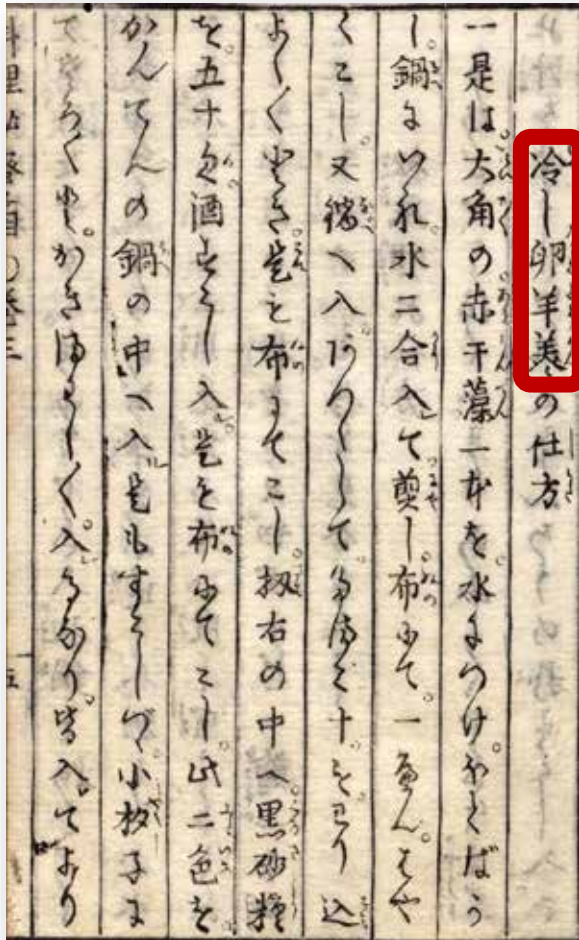
PMJT Dataset (from NIJL)



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



IIF Curation Viewer
Core contributor: Jun HOMMA

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

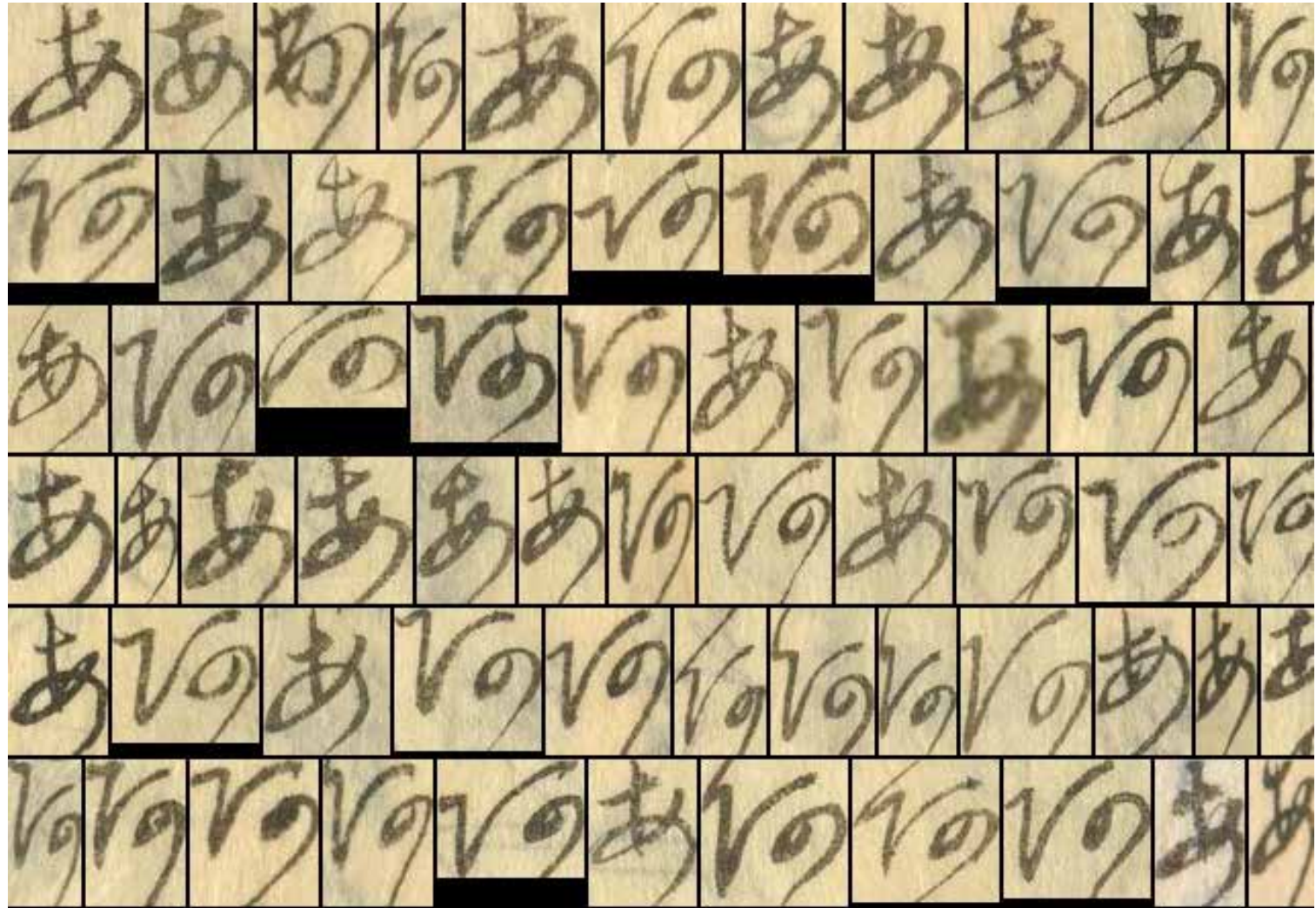
Data Identifiers

- 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

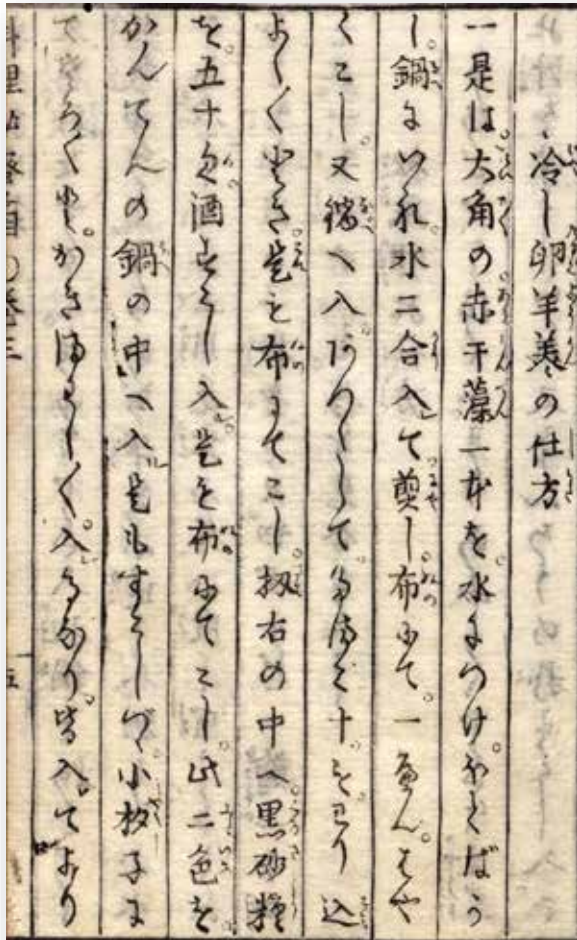
<http://codh.rois.ac.jp/edo-cooking/>

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

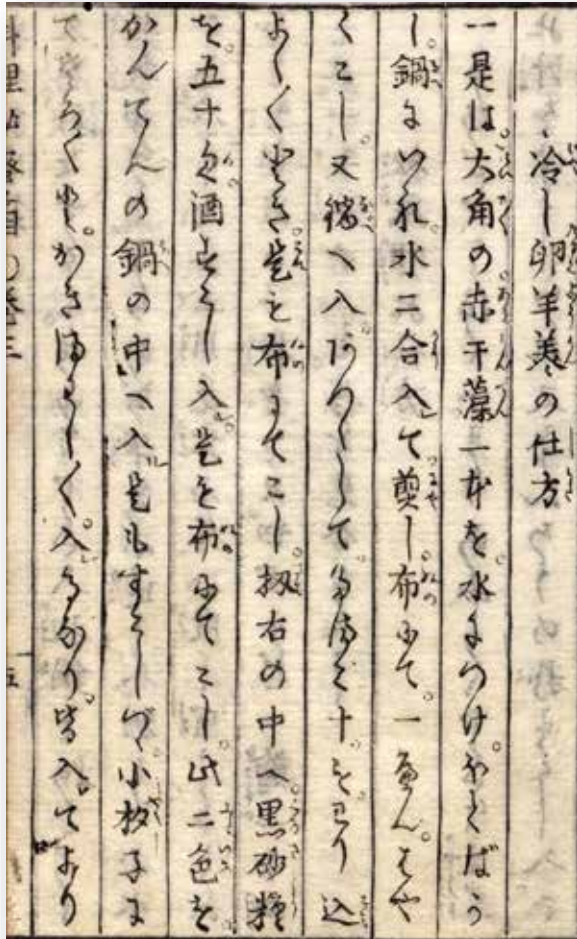


PMJT Dataset (from NIJL)

1	是は大角の赤干藻一本を水につけほとばかし
2	鍋にいれ水二合入し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	扱右の中へ 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	扱鍋をぬき 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

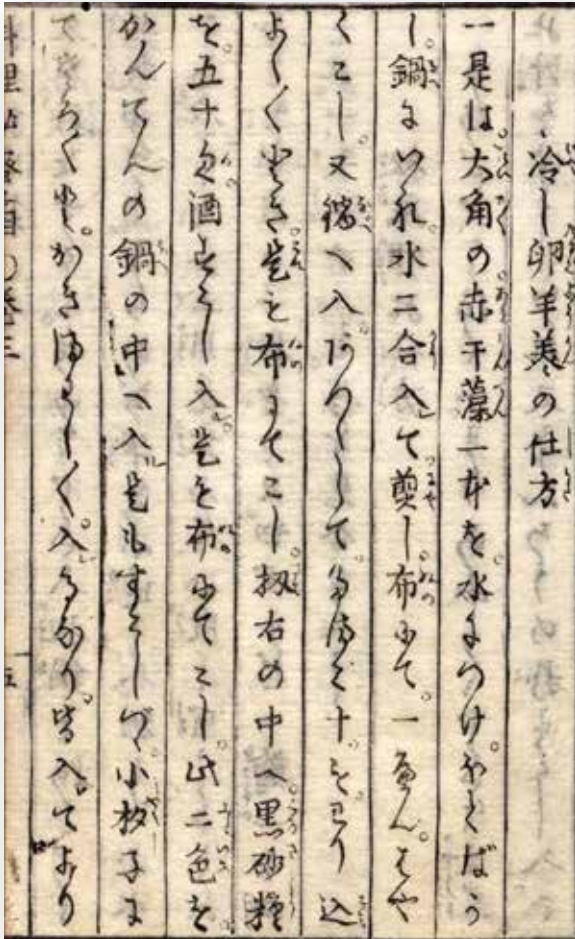


PMJT Dataset (from NIJL)

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	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	の中に黒砂糖50匁 (200g) と酒少しを入れ、布で漉す。
6	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	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	全て入れ終わったら、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

人文学オープンデータ共同利用センター事務局
Center for Open Data in the Humanities

江戸料理レシピデータセット

日本古典籍データセットに含まれる江戸の料理本を、現代の生活にも取り入れるために、現代レシピに変換して提供します。

最初の江戸料理レシピとして、100種類以上の鉄料理を集めた『万宝料理秘録 御百珍』を取り上げます。

「万宝料理秘録 御百珍」の江戸料理レシピ

くずし字を読める日本人が少ないという中で、日本古典籍データセットのようなデジタル画像を提供するだけでは、市民によるオープンデータ活用を進めることは難しいのが実情です。古典籍を日常生活にどのように活用していけばいいか、と考えているところで思い当たったのが江戸時代の料理本でした。これを現代語訳すれば現代でも料理を作って楽しめるのではないかと考えました。

雑煮などの季節の料理や地方色豊かな料理などは、日本人の生活に深く根ざしたものです。そして日本の料理としての和食は、単なる料理法を超えて自然の尊重という日本人の精神に基づく文化を表すとも言われています。平成25年には「和食：日本人の伝統的な食文化」がユネスコ無形文化遺産に登録され、和食文化に対する国際的な認知度も高まってきました。そんな和食という自身の文化をより深く理解するには、過去の料理について学び、気に向ければ作ってみることもできるようなレシピデータが必要だと考えました。そこで以下のような「レシピ化」のプロセスに取り組みました。

データ概要

原本画像データ	日本古典籍データセットで公開する画像です。くずし字を読み、かつ江戸時代の日本語や料理法を知っていれば料理が作れます。
個別テキストデータ	原本画像のくずし字をテキスト化したデータです。江戸時代の日本語や料理法を知っていれば料理が作れます。
現代語訳データ	個別テキストデータの内容を現代の日本語に翻訳したデータです。江戸時代の料理法を知っていれば料理が作れます。
現代レシピデータ	現代語訳データの内容を、現代の道具や食材でも作れるものに変換し、食材の分量や写真を加えてより具体化したデータです。手順に従えば料理が作れます。

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).

<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.

cookpad

江戸時代のスイーツ 甘さスツキリ冷卵羊羹

江戸の料亭本から見つけた和風スイーツです。プリンの様ですが、牛乳不使用でさらっとした菓種の甘さがやみつきになります。

材料 (約4名分)

餅	5個
寒天(赤)	1本(4g)
黒砂糖	100g
水	180cc
片栗粉	適量
酒	適量

カロリー: 276kcal/人 @ 糖質: 0.3g/人

1. 寒天を水につけて、ふやかします。

2. 生卵をよく溶きまです。

3. 溶いた生卵を布でこします。

4. 3に入れ、再びこします。

5. 鍋に寒天と水(180cc)を入れて煮とがします。

6. 6を布などでこし、再び鍋に入れて熱します。

7. 7の熱した寒天の中に、5の糊液を少しずつ入れます。

<http://cookpad.com/recipe/4153357>

Unexpectedly Large Impact

人文学オープンデータ共同利用センターさんがリツイート

うずら @caille2006 · 11月26日
このプロジェクトがすごいのは、古文書の情報をさらに現代の生きた情報にするために、クックパッドにアカウントを開設してレシピを公開し「つくれぽ」も受け付けていること。江戸ご飯とつくれぽというこの未来感バネい。 cookpad.com/kitchen/146046...



クックパッド江戸ご飯 のキッチン

プロフィール

トップ	レシピ 32	つくれぽ 0	献立 0
-----	-----------	-----------	---------

レシピを検索

7478 retweets



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

2017/01/23

人文学オープンデータ共同利用センターさんがリツイート

NII 国立情報学研究所(NII) @jouhouken · 11月24日
[プレスリリース]
江戸の文化を現代に取り込む「江戸料理レシピデータセット」を整備～江戸時代の料理本を「レシピ化」し、クックパッドでも公開～
nii.ac.jp/news/2016/1124



1 1,074 971

1074 retweets

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

1st CODH Seminar

41

TV show to reproduce the dish

The image shows a screenshot of a news website, specifically the 'news24' section. The main headline is '「江戸ご飯レシピ」再現、作って食べてみた' (Reproduction of Edo Rice Recipe, Made and Eaten). The article is dated 2016年11月30日 21:05. The main image shows a person's hands using chopsticks to place a halved soft-boiled egg on top of a bowl of dark, sticky rice (likely 'kani gohan' or 'kani gohan'). The article is categorized under '社会' (Society) and 'エンタメ' (Entertainment). The website header includes navigation links for various news categories and a search bar. On the right side, there are advertisements for clothing, including 'GAUCHO PANTS' and 'soulberry 2016 Hello Autumn'.

0テレ NEWS24
2016年(平成28年) 12月5日(火) 今日の実況 仙台

UR賃貸住宅 社宅も UR
礼金 ¥0 手数料 ¥0 更新料 ¥0 保証人 不要

トップ 社会 エンタメ 国際 政治 経済 スポーツ 投稿 ブログ 特集 地方 天気 地震 防災

トップ > 社会 > 「江戸ご飯レシピ」再現、作って食べてみた

0テレ NEWS24 「江戸ご飯レシピ」再現、作って食べてみた
ツイートする シェアする 2016年11月30日 21:05

江戶ご飯レシピ Hot Word

全文

11月28日、「江戸ご飯レシピ」という言葉を含んだツイートが大きく伸びました。

「いま見るべき」「いま知るべき」厳選ニュースをメルマガでお届け

アクセスランキング 社会

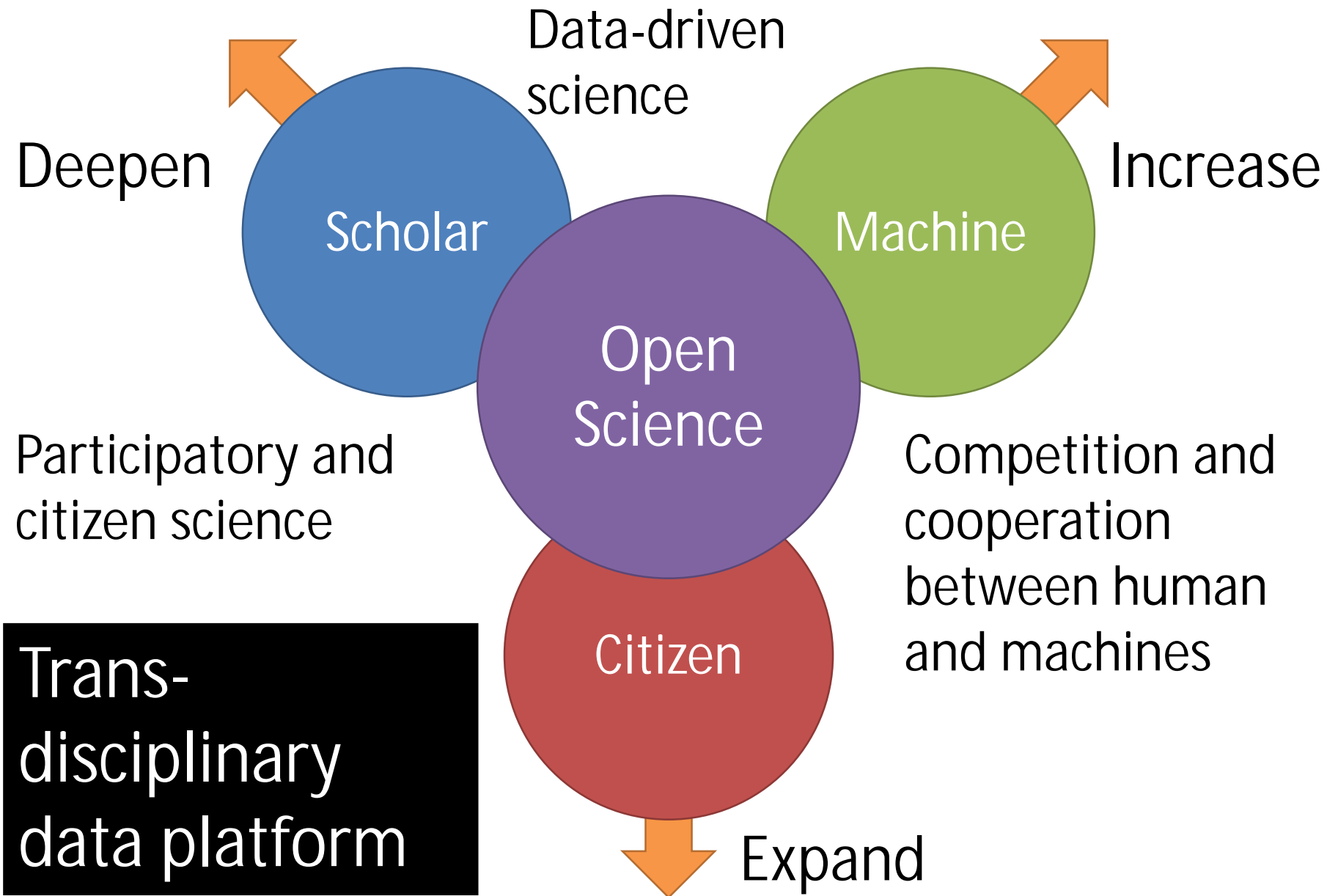
- 1 豊中市妊婦刺殺「経歴はなく事故」と否認
- 2 電車で女性の髪切り「アクションで売る」
- 3 千葉大医学部3人の英名公表 医師も逮捕
- 4 女性の神楽橋「ばらまくぞ」航空自衛官逮捕
- 5 東京の異端スタッフ祭り「おっぱい」

<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



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/>