Digital Humanities for Japanese Culture: Resources and Methods

Kiyonori Nagasaki
Asanobu Kitamoto
Yuta Hashimoto
Satoru Nakamura
Tatsuki Sekino
Taizo Yamada
Welcome to DHSI 2019!

Thanks for joining the DHSI community!

In this booklet, you will find essential course materials prefaced by some useful information about getting settled initially at UVic, finding your way around, getting logged in to our network (after you’ve registered the day before our courses begin), and so on.

Given our community’s focus on things computational, it will be a surprise to no one that we might expect additional information online for some of the classes - your instructors will let you know - or that the most current version of all DHSI-related information may be found on our website at dhsi.org.

Do check in there first if you need anything that’s not in this coursepak.

To access the DHSI wifi network, simply go into your wireless settings and connect to the “DHSI” network and enter the password “dhsi2019”.

And please don’t hesitate to be in touch with us at institut@uvic.ca or via Twitter at @AlyssaA_DHSI or @DHInstitute if we can be of any help....
Regional Map of Greater Victoria

Average Frequency

**Regional Route**
15–60 minute service with limited stops

**Frequent Route**
15 minute or better service
7am–7pm, Mon–Fri

**Local Route**
20–120 minute service
DHSI Wi-Fi

Network name: DHSI
Passkey: dhsi2019
The 2019 schedule is just taking shape nicely! A very few things to confirm, add, etc, still but this is the place to be to find out what is happening when / where ...

---

**Sunday, 2 June 2019 [DHSI Registration + Suggested Outings]**

If you're here a day or two before we begin, or staying a day or two afterwards, here are a few ideas of things you might consider doing ....

**Suggested Outing 1, Botanical Beach (self-organised; car needed)**

A self-guided visit to the wet, wild west coast tidal shelf (and historically-significant former research site) at Botanical Beach; we recommend departing early (around 8.00 am) to catch low tide for a better view of the wonderful undersea life! Consider bringing a packed lunch to nibble-on while looking at the crashing waves when there, and then have an afternoon drink enjoying the view from the deck of the Port Renfrew Hotel.

**Suggested Outing 2, Butchart Gardens (self-organised)**

A shorter journey to the resplendently beautiful Butchart Gardens and, if you like, followed by (ahem) a few minutes at the nearby Church and State Winery, in the Saanich Penninsula. About an hour there by public bus from UVic, or 30 minutes by car.

**Suggested Outing 3, Saltspring Island (self-organised; a full day, car/bus + ferry combo)**

Why not take a day to explore and celebrate the funky, laid back, Canadian gulf island lifestyle on Saltspring Island. Ferry departs regularly from the Schwartz Bay ferry terminal, which is about one hour by bus / 30 minutes by car from UVic. You may decide to stay on forever ....

**Suggested Outing 4, Paddling Victoria's Inner Harbour (self-organised)**

A shorter time, seeing Victoria's beautiful city centre from the waterways that initially inspired its foundation. A great choice if the day is sunny and warm. Canoes, kayaks, and paddle boards are readily rented from Ocean River Adventures and conveniently launched from right behind the store. Very chill.

And more!

Self-organised High Tea at the Empress Hotel, scooter rentals, visit to the Royal BC Museum, darts at Christie's Carriage House, a hangry breakfast at a local diner, whale watching, kayaking, brew pub sampling (at Spinnaker's, Swans, Moon Under Water, and beyond!), paddle-boarding, a tour of used bookstores, and more have also been suggested!

---

**9:00 to 4:00**

**Early Class Meeting:** 4. [Foundations] DH For Department Chairs and Deans *(David Strong Building C124, Classroom)*

Further details are available from instructors in mid May to those registered in the class. Registration materials will be available in the classroom.

**3:00 to 5:00**

**DHSI Registration** *(MacLaurin Building, Room A100)*

After registration, many will wander to Cadboro Bay and the pub at Smuggler's Cove OR the other direction to Shelbourne Plaza and Maude Hunter's Pub OR even into the city for a nice meal.

---

**Monday, 3 June 2019**

Your hosts for the week are Alyssa Arbuckle, Ray Siemens, and Jannaya Friggstad Jensen.

---

**7:45 to 8:15**

**Last-minute Registration** *(MacLaurin Building, Room A100)*
8:30 to 10:00

Welcome, Orientation, and Instructor Overview (MacLaurin A144)
- Welcome to the Territory
- Welcome to DHSI: Ray Siemens, Alyssa Arbuckle
- Welcome from UVic: Jonathan Bengtson (University Librarian), Alexandra D'Arcy (Associate Dean Research, Humanities)

10:15 to Noon

Classes in Session (click for details and locations)
- 1. [Foundations] Digitisation Fundamentals and their Application (Clearihue A103, Lab)
- 2. [Foundations] Introduction to Computation for Literary Criticism (Clearihue A102, Classroom)
- 4. [Foundations] DH For Department Chairs and Deans (David Strong Building C124, Classroom)
- 5. [Foundations] Developing a Digital Project (With Omeka) (Clearihue A031, Lab)
- 9. Out-of-the-Box Text Analysis for the Digital Humanities (Human and Social Development A160, Lab)
- 10. Sound and Digital Humanities (Cornett A120, Classroom)
- 11. Critical Pedagogy and Digital Praxis in the Humanities (Clearihue D132, Classroom)
- 12. Digital Humanities for Japanese Culture: Resources and Methods (McPherson Library A003, Classroom)
- 14. Retro Machines & Media (McPherson Library 128, Classroom)
- 15. Web APIs with Python (Human and Social Development A170, Lab)
- 16. Ethical Data Visualization: Taming Treacherous Data (Cornett A128, Classroom)
- 17. Linked Open Data and the Semantic Web (Cornett A132, Classroom)
- 19. The Frontend: Modern JavaScript & CSS Development (Human and Social Development A150, Lab)
- 20. Palpability and Wearable Computing (McPherson Library A025, Classroom)
- 21. Information Security for Digital Researchers (David Strong Building C114, Classroom)
- 22. Introduction to IIIF: Sharing, Consuming, and Annotating the World’s Images (Cornett A121, Classroom)
- 23. Modelling. Virtual. Realities. A Practical Introduction to Virtual (and Augmented) Reality (Human and Social Development A150, Lab)
- 24. Information Security for Digital Researchers (David Strong Building C114, Classroom)
- 25. Information Security for Digital Researchers (David Strong Building C114, Classroom)

12:15 to 1:15

Lunch break / Unconference Coordination Session (MacLaurin A144)
(Grab a sandwich and come on down!)
Discussion topics, scheduling, and room assignments from among all DHSI rooms will be handled at this meeting.

1:30 to 4:00

Classes in Session

4:10 to 5:00

Institute Lecture: Jacqueline Wernimont (Dartmouth C): "Sex and Numbers: Pleasure, Reproduction, and Digital Biopower"
Chair: Anne Cong-Huyen (U Michigan)
(MacLaurin A144)

Abstract: Drawing from Numbered Lives (MIT 2018), this talk will consider a long history of sex-number entanglement in Anglo-American Cultures. Drawing on historical and contemporary objects and practices, Wernimont will ask "in what ways do theories of biopower, critical gender and critical race studies, and media studies" suggest that we can understand this set of entanglements and their impacts. NB: While relevant, this talk will not include discussions of sexual trauma or violence. It will include frank discussion of sex acts and various ways of translating sexual behavior into numbers.

5:00 to 6:00

Opening Reception (University Club)

Tuesday, 4 June 2019
Wednesday, 5 June 2019

9:00 to Noon
Classes in Session

Lunch break / Unconference

"Mystery" Lunches

Presentation: An Introduction to Scholarly Publishing with Manifold (MacLaurin A144)
Lunch included for those who register here

This presentation introduces Manifold Scholarship, a Mellon-funded digital publishing platform developed by the CUNY Graduate Center, The University of Minnesota Press, and Cast Iron Coding. Manifold allows you to create beautiful, dynamic open access projects that can include text, images, video, embedded resources, and social annotation. We will provide an overview of Manifold and demonstrate how faculty, students and staff in the digital humanities can use Manifold to publish open access scholarly works, conduct and participate in peer review, and create custom edited versions of public domain course texts and OER.

1:30 to 4:00
Classes in Session

DHSI Conference and Colloquium Lightning Talk Session 2 (MacLaurin A144)
Chair: Kim O’Donnell (Simon Fraser U)

- Catherine Ryu (Michigan State U), "Tone Perfect: Developing a Multimodal Audio Database for Mandarin Chinese as an Open Source"
- Kenzie Burchell (U Toronto Scarborough), "Making Responsible Reporting Practices Visible: Comparing newswire coverage of humanitarian crises in Syria"
- Jessica Linzel (Brock U), "The Shopkeeper Aristocracy: Mapping Trade Networks in Colonial Niagara"
- Kirsten Painter (U Washington), "From Bogatyr to Bread: Digitization & Online Exhibition of Rare Russian Children's Books at the U Washington"
- John Barber (Washington State U), "A Mighty Span"

4:15 to 5:15

6:00 to 7:00
"Half Way There!" [An Informal, Self-Organized Birds of a Feather Get-Together] (Felicitas, Student Union Building)
Bring your DHSI nametag and enjoy your first tipple on us! [A great opportunity for an interest group meet-up ....]

Thursday, 6 June 2019

9:00 to Noon
Classes in Session

Lunch break / Unconference

"Mystery" Lunches

12:15 to 1:15

Instructor lunch meeting

1:30 to 4:00
Classes in Session

DHSI Conference and Colloquium Lightning Talk Session 3 (MacLaurin A144)
Chair: Kim O’Donnell (Simon Fraser U)

- Colleen Kolba (U South Florida), "What Comics can Teach our Students about Multimodal Literacy"
- Trish Baer (ETCL; U Victoria), "Preserving Digital Legacies: Archived Websites and Digital Discoverability"
- Suchismita Dutta (U Miami), "The Importance of Archival Transcription for Genre Building"
- Jeffrey Lawler (California State U, Long Beach), "Twining our way through the Past: Video Game Authoring as History Pedagogy"
Friday, 7 June 2019 [DHSI; ADHO Pedagogy SIG Conference Opening]

9:00 to Noon  
Classes in Session

12:15 to 1:15  
Lunch Reception / Course E-Exhibits (MacLaurin A100)

1:30 to 1:50  
Remarks, A Week in Review (MacLaurin A144)

2:00 to 3:00  
Joint Institute Lecture (DHSI and ADHO Pedagogy SIG Conference):
Matt Gold (CUNY Graduate Center and Association for Computers and the Humanities): “Thinking Through DH: Proposals for Digital Humanities Pedagogy”
Chair: Diane Jakacki (Bucknell U)  
(MacLaurin A144)

Abstract: How do we teach digital humanities, and how should DH be taught? What, indeed, should we teach when we teach DH? This talk will present a proposal for grounding digital humanities pedagogical practice in the research interests of our students and the epistemological foundations of our methods rather than through an approach grounded more central in data and methods.

3:30 to 5:00  
Joint Reception: DHSI and ADHO Pedagogy SIG Conference (University Club)  
E-Poetry Event (Chris Tanasescu)
Watch this space for details, including how to participate!  
DHSI Conference and Colloquium Poster/Demo Session  
• Pia Russel (U Victoria); Emily Stremel (U Victoria), “British Columbia’s Historical Textbooks Digital Library”  
• Cody Hennesey (U Minnesota); Rachael Samberg (U California, Berkeley); Stacy Reardon (U California, Berkeley), “Finding the Haystack: Literacies for Accessing and Using Text as Data”  
• Paula Johanson (ETCL; Independent Scholar), “Proving Seahorses and Juan de Fuca’s Travels in The Curve of Time”  
• Tara Baillargeon (Marquette U); Elizabeth Wawrzyniak (Marquette U), “FellowsHub: J. R. R. Tolkien Fanzine Portal”  
• Graham Jensen (U Victoria), “Canadian Modernist Magazines Project”  
• Caterina Agostini (Rutgers U), “Art at the Time of Syphilis: A First-Person Medical Narrative in Benvenuto Cellini’s Vita”  
• Lauren Elle DeGaine (ETCL; U Victoria), “Women at the Front: A Digital Exhibit of Victorian Frontispiece Illustrations”  
• Adam Griggs (Mercer U); Kathryn Wright (Mercer U); Christian Pham (Mercer U); Gail Morton (Mercer U); Stephanie Miranda (Mercer U), “Digitizing Middle Georgia’s History of Slavery”

Saturday, 8 June 2019 [Conference, Colloquium, and Workshop Sessions]

Conference / Workshop Registration (MacLaurin A100)

8:00 to 9:00  
The day's events are included with your DHSI registration. If you're not registered in DHSI, you're very welcome to join us by registering here as a Conference / Colloquium / Workshop participant. We'll have a nametag waiting for you!

Coffee, Tea, &c?  
Looking for some morning coffee or tea, or a small nibble? Options and hours of operation for weekend campus catering are available here. Mystic Market usually opens around 10.00.

9:00 to 4:00  
Conference and Colloquium Sessions  
ADHO Pedagogy SIG Conference Sessions  
Right2Left Workshop Sessions

9:00 to 4:00  
All Day DHSI Workshop Session (click for workshop details and free registration for DHSI participants)
• 55. Introduction to Machine Learning in the Digital Humanities [8-9 June; All day, each day] (David Strong Building C124, Classroom)

9:00 to 9:10  
Informal Greetings, Room Set-up (Lobby, outside Hickman 105)

Session 1

DHSI Colloquium and Conference (Hickman 105)  
Digital Humanities & Literature, Chair: Kim O'Donnell (Simon Fraser U)  
• Youngmin Kim (Dongguk U), “Transdiscursivity in the Convergence of Digital Humanities and World Literature”  
• Caroline Winter (U Victoria), “Digitizing Adam Smith’s Literary Library”  
• Kaitlyn Fralick (U Victoria); Kailey Fukushima (U Victoria); Sarah Karlson (U Victoria), “Victorian Poetry
9:10 to 10:30  
ADHO Pedagogy SIG Conference (Hickman 110)
Chair: Katherine Faull (Bucknell U)
- Aaron Tucker and Nada Savicevic (Ryerson U), "Write Here, Right Now: An Open Source eTextbook for the Flipped Classroom"
- Heather McAlpine (U Fraser Valley), "Digital Meters: Using Text Encoding to Teach Literature in the Undergraduate Classroom"
- Tiina H. Airaksinen (U Helsinki), "Digital Humanities in Cultural Studies: Creating a MOOC course for University Students and A-Level Students"

Right2Left Workshop (Hickman 116)
Keynote - Nathan P. Gibson (Ludwig Maximilianiens U, München): "Thinking in -JTR: Reorienting the Directional Assumptions of Global Digital Scholarship"

10:30 to 10:40  
Break

10:40 to Noon  
Session 2
DHIS Colloquium and Conference (Hickman 105)
Digital Humanities & Society, Chair: Eleanor Reed (Hastings C)
- Joel Zapata (Southern Methodist U), "Uncovering the Southern Plains' Mexican American Civil Rights Movement"
- Ayo Oseisan (U Ibadan), "Online Newspaper Construction of Agitation for the Sovereign State of Biafra in Nigeria"
- Joseph Jones (U British Columbia), "Testbed for an Approach to Distant Reading: Fictions That Represent Vietnam War Resisters in Canada"
- Brendan Mackie (U California, Berkeley), "Visualizing Long-Term Cultural Change: An Example From The Birth of Civil Society"

10:40 to Noon  
ADHO Pedagogy SIG Conference (Hickman 110)
Chair: Laura Estill (St Francis Xavier U)
- Jane Jackson (Chinese U of Hong Kong), "Interrogating digital spaces for intercultural meaning-making"
- Ryan Ikeda (UC Berkeley), "Disrupting Digital Literacy: Situating Electronic Literature Among Public Education Initiatives"
- Christopher Church, Katherine Hepworth (U Nevada, Reno), "We're STEAMed! A call for balancing technical instruction and disciplinary content in the digital humanities"
- Chelsea Milbourne (Cal Poly, San Luis Obispo), "Finding the Right Fit between Technology and Class Content: Reflections on Including Web Development in a Digital Storytelling Course"

Right2Left Workshop (Hickman 116)
- Edward "Eddie" Surman (Claremont Graduate U), "Qualitative Digital Text Analysis and #Right2Left Languages: A Demonstration of Atlas.ti using the Hebrew Bible"

Noon to 1:10  
Lunch (We recommend Mystic Market on weekends!)

1:10 to 2:30  
Session 3
DHIS Colloquium and Conference (Hickman 105)
Digital Humanities & Community, Chair: Claire Carlin (U Victoria)
- Pia Russel (U Victoria); Emily Stremel (U Victoria), "Mentorship and disability: Supporting disabled employees in digital humanities"
- Amy Lueck (Santa Clara U), "Virtually Emplacing Indigenous Memory"
- Md. Shehabul Alam (National U Bangladesh), "Integrating Library Service with Union Information and Service Center: A Joint Initiative towards Digital Bangladesh"
- Veronica Gomez (Instituto de Humanidades y Ciencias Sociales (HuCSo) - UNL-CONICET), "Latin American E-literature and Location: The Nation Revisited in Electronic Literature Organization (ELO)"

1:10 to 2:30  
ADHO Pedagogy SIG Conference (Hickman 110)
Chair: Chris Tănăsescu (UC Louvain)
- Laura Estill (St Francis Xavier U), "One Assignment, Three Ways: Assessing DH Projects in a Literature Course"
- Felix Bayode Oke, Stella N. Kpolugbo (Anchor U Lagos), "The Multimodal Technique as a Pedagogical Tool in Pelu Awofeso’s White Lagos: A Definitive and Visual Guide to the Eyo Festival"
- Francesca Giannetti (Rutgers U, New Brunswick), "So near while apart: Correspondence Editions as Critical Library Pedagogy and Digital Humanities Methodology"

Right2Left Workshop (Hickman 116)
- Najla Jarkas (American U Beirut) and David Joseph Wrisley (NYU Abu Dhabi), "RTL Software Localization and Digital Humanities: the Case Study of Translating Voyant Tools into Arabic"
2:30 to 2:40 Break

Session 4

DHSI Colloquium and Conference (Hickman 105)
- Olivia Wikle (U Idaho), "Listening with Our Eyes: Using Topic Modeling, Text Analysis, and Sound Studies Methodologies to Explore Literary Soundscapes"
- Olin Bjork (U Houston-Downtown), "Dramatic Redundancy: Interactive Transcripts and Multimodal Performance Editions"
- Ashleigh Cassemare-Stanfield (U Chicago), "Sonifying Hamlet and Reading the Room"

ADHO Pedagogy SIG Conference (Hickman 110)
Chair: Aaron Tucker (Ryerson U)
Youngmin Kim (Dongguk U), "Teaching Digital Humanities and World Literature in Class"
Alice Fleerackers, Juan Pablo Alperin, Esteban Morales, Remi Kalir (Simon Fraser U, U Colorado Denver), "Online annotations in the classroom: How, why, and what do students learn from annotating course material?"
Andie Silva (York C and Graduate Center, CUNY), "Keeping it Local: Undergraduate DH as Feminist Practice"

Right2Left Workshop (Hickman 116)
- Joanna Byszuk (Institute of Polish Language, Polish Academy of Sciences, Warsaw/Computational Stylistics Group) and Alexey Khismatulin (Institute of Oriental Manuscripts, Russian Academy of Sciences, Saint Petersburg), "Attribution of Authorship for Medieval Persian Quasidas with Stylometry"
- Ilan Benattar (New York U), "#Right2Left Biblical Translations in Jewish Textual History: Case Studies in Judeo-Arabic and Judeo-Spanish"

Sunday, 9 June 2019 [Workshop Sessions]

DHSI Registration (MacLaurin Building, Room A100)

8:00 to 5:00
The day's events are included with your DHSI registration. If you're not registered in DHSI, you're very welcome to join us by registering here as a Conference / Colloquium / Workshop participant. We'll have a nametag waiting for you!

Coffee, Tea, &c?
Looking for some morning coffee or tea, or a small nibble? Options and hours of operation for weekend campus catering are available here. Mystic Market usually opens around 10.00.

9:00 to 4:00
All Day Workshop Sessions (click for workshop details and free registration for DHSI participants)
- 55. Introduction to Machine Learning in the Digital Humanities [8-9 June; All day, each day] (David Strong Building C124, Classroom)
- 56. Pedagogy of the Digitally Oppressed: Anti-Colonial DH Methods and Praxis [9 June; All Day] (Hickman 116, Classroom)
- 57. Natural Language Processing and Network Coding Apps for Text & Textual Corpus Analysis in the Humanities [9 June; All Day] (David Strong Building C114, Classroom)

AM Workshop Sessions (click for workshop details and free registration for DHSI participants)
- 59. 3D Visualization for the Humanities [9 June; AM] (Cornett A229, Classroom)
- 60. It's All Relational: AbTeC’s Indigenous Video Game Workshops as Storytelling Praxis [9 June; AM] (Cornett A121, Classroom)
- 61. Spatial DH: De-Colonizing Cultural Territories Online [9 June; AM] (Cleerihue D130, Classroom)
- 63. Creating a CV for Digital Humanities Makers [9 June; AM] (David Strong Building C198, Classroom)

Noon to 1:00
Lunch (We recommend Mystic Market on weekends!)

PM Workshop Sessions (click for workshop details and free registration for DHSI participants)
- 65. Indigenous Futurities in the Classroom and Beyond [9 June; PM] (Cornett A121, Classroom)
- 66. DHSI Knits: History of Textiles and Technology [9 June; PM] (Fine Arts 109, Classroom)
- 68. Linked Open Datafication for Humanities Scholars [9 June; PM] (McPherson Library A003, Classroom)
- 69. Stylo - WYSIWYM Text Editor for Humanities Scholars [9 June; PM] (McPherson Library A025, Classroom)

After the day, many will wander to Cadboro Bay and the pub at Smuggler's Cove OR the other direction to Shelbourne Plaza and Maude Hunter's Pub OR even into the city for a bite to eat.

Monday, 10 June 2019
Your hosts for the week are Ray Siemens and Jannaya Friggstad Jensen.

<table>
<thead>
<tr>
<th>Time</th>
<th>Event Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:45 to 8:15</td>
<td>DHSI Last-minute Registration (<a href="#">MacLaurin A100</a>)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>Welcome, Orientation, and Instructor Overview (<a href="#">MacLaurin A144</a>)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>Classes in Session (click for details and locations)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>28. [Foundations] Text Encoding Fundamentals and their Application (Digital Scholarship Commons, McPherson Library A308, Classroom)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>29. [Foundations] Understanding The Predigital Book: Technologies of Inscription (McPherson Library A003, Classroom)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>30. [Foundations] Databases for Digital Humanists (McPherson Library 210, Classroom)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>31. [Foundations] Music Encoding Fundamentals and their Applications (Clearihue A030, Lab)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>32. [Foundations] Digital Storytelling (Cornett A120, Classroom)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>33. [Foundations] Text Mapping as Modelling (Clearihue D131, Classroom)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>34. [Foundations] Stylometry with R: Computer-Assisted Analysis of Literary Texts (Clearihue A102, Lab)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>35. [Foundations] Databases for Digital Humanists (McPherson Library 210, Classroom)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>36. [Foundations] Digital Games as Tools for Scholarly Research, Communication and Pedagogy (Cornett A229, Classroom)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>37. [Foundations] Text Mapping as Modelling (Clearihue D131, Classroom)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>38. [Foundations] Queer Digital Humanities (David Strong Building C114, Classroom)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>39. [Foundations] Parsing and Writing XML with Python (Clearihue A108, Lab)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>40. [Foundations] Introduction to Electronic Literature in DH: Research and Practice (Cornett A128, Classroom)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>41. [Foundations] Surveillance and the Critical Digital Humanities (David Strong Building C108, Classroom)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>42. [Foundations] Text Analysis with Python and the Natural Language ToolKit (Clearihue A103, Lab)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>43. [Foundations] Creating LAMP Infrastructure for Digital Humanities Projects (Human and Social Development A170, Lab)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>44. [Foundations] Processing Humanities Multimedia (Human and Social Development A150, Lab)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>45. [Foundations] Digital Humanities Pedagogy: Integration in the Curriculum (Cornett A121, Classroom)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>46. [Foundations] Accessibility &amp; Digital Environments (Priestly Law Library 265, Classroom)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>47. [Foundations] Agile Project Management (Cornett A132, Classroom/Lab)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>48. [Foundations] XPath for Processing XML and Managing Projects (Clearihue A105, Lab)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>49. [Foundations] Endings: How to End (and Archive) your Digital Project (Priestly Law Library 192, Classroom)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>50. [Foundations] Text Processing - Techniques &amp; Traditions (McPherson Library A025, Classroom)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>51. [Foundations] Introduction to Humanities Data Analysis &amp; Visualization in R (HDA) (Human and Social Development A160, Lab)</td>
</tr>
<tr>
<td>8:30 to 10:00</td>
<td>52. [Foundations] Introduction to Network Analysis in the Digital Humanities (Clearihue D132, Classroom)</td>
</tr>
<tr>
<td>10:15 to Noon</td>
<td>Lunch break / Unconference Coordination Session (<a href="#">MacLaurin A144</a>)</td>
</tr>
<tr>
<td>10:15 to Noon</td>
<td>(Grab a sandwich and come on down!)</td>
</tr>
<tr>
<td>12:15 to 1:15</td>
<td><em>Mystery</em> Lunches</td>
</tr>
<tr>
<td>12:15 to 1:15</td>
<td>Lunch break / Unconference Coordination Session (<a href="#">MacLaurin A144</a>)</td>
</tr>
<tr>
<td>1:30 to 4:00</td>
<td>Classes in Session</td>
</tr>
<tr>
<td>4:10 to 5:00</td>
<td>Institute Lecture: Angel David Nieves (San Diego State U): &quot;3D Mapping and Forensic Traces of Testimony: Documenting Apartheid-Era Crimes Through the Digital Humanities&quot;</td>
</tr>
<tr>
<td>4:10 to 5:00</td>
<td>Chair: Constante Crompton (U Ottawa)</td>
</tr>
<tr>
<td>4:10 to 5:00</td>
<td>(<a href="#">MacLaurin A144</a>)</td>
</tr>
<tr>
<td>4:10 to 5:00</td>
<td>Abstract: In 1989 the killing of a queer, 14-year-old youth in Winnie Mandela's house named Stompie Seipel (an event that few in South Africa are willing to recall, let alone discuss, in any detail) -- is perhaps one of the most glaring examples where the queer and activist community was suppressed or erased from anti-apartheid/liberation histories. Digital humanities may actually help both reconstruct and recover a history that is still very early in the telling, despite what is commonly believed about the liberation struggle and the contributions of queer activists in the dismantling of apartheid. Perhaps it could explain why a youth such as Seipel was killed -- or at the very least, provide a more complex and messy narrative that permits one to know more how the history of queer anti-apartheid activists was suppressed. This talk outlines a methodology for &quot;messy thinking and writing&quot; in the digital humanities that -- through a queer and feminist intersectional framework -- permits a more complex layering of oral histories and 3D historical reconstructions.</td>
</tr>
<tr>
<td>5:00 to 6:00</td>
<td>Reception (<a href="#">University Club</a>)</td>
</tr>
<tr>
<td>9:00 to Noon</td>
<td>Classes in Session</td>
</tr>
<tr>
<td>12:15 to 1:15</td>
<td>Lunch break / Unconference</td>
</tr>
<tr>
<td>12:15 to 1:15</td>
<td>&quot;Mystery* Lunches</td>
</tr>
</tbody>
</table>

---

**Tuesday, 11 June 2019**
<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:30 to 4:00</td>
<td>Classes in Session</td>
</tr>
</tbody>
</table>
| 4:15 to 5:15    | DHSI Conference and Colloquium Lightning Talk Session 4 (MacLaurin A144) Chair: Lindsey Seatter (U Victoria)  
|                 | - Ashley Caranto Morford (U Toronto); Kush Patel (U Michigan); Arun Jacob (McMaster U), "Our DHI’s anti-colonial: Questions and challenges in dismantling colonial influences in digital humanities pedagogy"  
|                 | - Julia King (U Bergen), "Developing Network Visualizations of Syon Abbey's Books, 1415-1539"  
|                 | - Luis Meneses (ETCL; U Victoria), "Identifying Changes in the Political Environment in Ecuador"  
|                 | - Alicia Brown (Texas Christian U), "Digital Cartography of the Ancient World"  
|                 | - Laura Horak (Carleton U), "Building the Transgender Media Portal"  
|                 | - Andrew Boyles Peterson (Michigan State U), "Last Mile Tracking: Implications of Rental Scooter Surveillance"  |
| 6:00 to 8:00    | DHSI Newcomer's Gathering (Grad House Restaurant, Graduate Student Centre)  
|                 | - Come down, buy meal and a beverage, and make some new friends!  |
| 9:00 to Noon    | Classes in Session                                                   |
| 12:15 to 1:15   | Lunch break / Unconference                                           |
|                 | "Mystery" Lunches                                                    |
|                 | Presentation: An Introduction Jupyter Notebooks for Researchers (MacLaurin A144)  
|                 | - This presentation introduces Jupyter Notebooks for researchers, via a partnership between Compute Canada and the Pacific Institute for the Mathematical Sciences (PIMS) including a large number of Canadian institutions. Read more here. Presenting is James Colliander, PIMS Director and team.  |
| 1:30 to 4:00    | Classes in Session                                                   |
| 4:15 to 5:15    | DHSI Conference and Colloquium Lightning Talk Session 5 (MacLaurin A144) Chair: Lindsey Seatter (U Victoria)  
|                 | - Calin Murgu (New College of Florida), "Putting local metadata to strategic use: A Dashboard for visualizing 60 years of theses metadata"  
|                 | - Jason Lajoie (U Waterloo), "Queer Critical Making and the Logic of Control"  
|                 | - John Barber (Washington State U), "Zambezi River Bridge"  
|                 | - Kent Emerson (U Wisconsin-Madison), "Digital Mappa and the George Moses Horton Project"  |
| 6:00 to 7:00    | "Half Way There (yet again)!" [An Informal, Self-Organized Birds of a Feather Get-Together] (Felicitas, Student Union Building)  
|                 | - Bring your DHSI nametag and enjoy your first tipple on us! [A great opportunity for an interest group meet-up ....]  |

**Wednesday, 12 June 2019**

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 to Noon</td>
<td>Classes in Session</td>
</tr>
<tr>
<td>12:15 to 1:15</td>
<td>Lunch break / Unconference</td>
</tr>
<tr>
<td></td>
<td>&quot;Mystery&quot; Lunches</td>
</tr>
<tr>
<td></td>
<td>Instructor lunch meeting</td>
</tr>
<tr>
<td>1:30 to 4:00</td>
<td>Classes in Session</td>
</tr>
</tbody>
</table>
| 4:10 to 5:00    | Institute Lecture: Karina van Dalen-Oskam (Huygens Institute and U Amsterdam; Alliance of Digital Humanities Organizations): "The Riddle of Literary Quality: Some Answers" Chair: Aaron Mauro (Penn State, Behrend C) (MacLaurin A144)  
|                 | - Abstract: What is literature, and can you measure it? That is the key question of the project The Riddle of Literary Quality. "The Riddle" is a research project of the Huygens Institute for the History of the Netherlands (Amsterdam) in collaboration with the Fryeke Akademy (Leeuwarden) and the Institute for Logic, Language and Computation (University of Amsterdam). The Riddle combines computational analysis of writing style with the results of a large online survey of readers, completed by almost 14,000 participants. In my talk, I will go into  |

**Thursday, 13 June 2019**

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 to Noon</td>
<td>Classes in Session</td>
</tr>
<tr>
<td>12:15 to 1:15</td>
<td>Lunch break / Unconference</td>
</tr>
<tr>
<td></td>
<td>&quot;Mystery&quot; Lunches</td>
</tr>
<tr>
<td></td>
<td>Instructor lunch meeting</td>
</tr>
<tr>
<td>1:30 to 4:00</td>
<td>Classes in Session</td>
</tr>
</tbody>
</table>
some of the main results of the project.

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 to Noon</td>
<td>Classes in Session</td>
<td></td>
</tr>
<tr>
<td>12:15 to 1:15</td>
<td>Lunch Reception / Course E-Exhibits (MacLaurin A100)</td>
<td></td>
</tr>
<tr>
<td>1:30 to 2:00</td>
<td>Closing, DHSI in Review (MacLaurin A144)</td>
<td></td>
</tr>
</tbody>
</table>

Contact info:  
enstitut@uvic.ca  
P: 250-472-5401  
F: 250-472-5881
The Course Description of DHSI 2019
Offered by JADH and SIG-CH, IPSJ

Digital Humanities for Japanese Culture: Resources and Methods

As DH spreads globally, necessity for comprehensive information of DH for each local culture has been increasing. To address the issue, the Special Interest Group of Computers and the Humanities under the auspices of the Information Processing Society of Japan and the Japanese Association for Digital Humanities decided to collaboratively offer this course to the DHSI in order to provide core information of digital resources and methods on Japanese culture, which have been addressed since three decades ago, with some hands-on training. It will include lectures of characteristics of the culture in DH and methods for non-segmented texts, cursive glyphs, pictures, and so on, such as text analysis, crowd-sourcing tagging and transcription, image analysis, TEI and IIIF for the resources. Participants will not be required expertise on Japanese culture, but interest for a comprehensive knowledge of DH on a certain culture.

Requirements for participation
1. A laptop computer with USB-A and WIFI connection
2. Interests in Japanese studies

Summative day-by-day overview

* Each session has the main lecturer, but a part of the session may be given by other lecturers.
* While some related materials are included below, the lectures will be provided according to knowledge and experiences of the participants.

1. Basic information for Japanese digital resources and their situation in academic and social contexts
   - Morning:
     Self-introduction of the lecturers and the participants with their experiences
   - Afternoon 1:
     Lecturers: All lecturers
     Japanese digital resources have been accumulated by various cultural institutions not only in Japan but also in the world. The activities and the current achievements
will be introduced from the viewpoint to leverage for DH.

Related materials:
Kiyonori Nagasaki, Contexts of Digital Humanities in Japan (attached)

- Afternoon 2:
  Lecturer: Asanobu Kitamoto
  Introduction to practices, theories and ideas on digital humanities in Japan.
  Related materials:
  Asanobu Kitamoto, Digital Humanities for Japanese Culture: Resources and Methods (Slides attached)

2. Introduction to textual analysis for Japanese, principles, tools and methods
   - Morning&AAfternoon:
     Lecturers: Taizo Yamada and others
     Due to spread of Unicode, the barrier of character encoding has been almost resolved. But the language which doesn’t have word segmentation like Japanese include difficulties to be analyzed. The lecturer will explain theoretical aspects of textual analysis for Japanese texts. After that, some tools for word segmentation and integrated analysis tools will be introduced by hands-on.
     Related materials:
     Taizo Yamada, Detection of topics from newspaper and its analysis of temporal variations in regions (Attached)
     Taizo Yamada, et al., A Common Base of Knowledge for Japanese Historical Materials and its Application (Attached)

3. Introduction to IIIF and TEI for Japanese resources
   - Morning:
     Lecturers: Satoru Nakamura and others
     Introduction to IIIF, which has become a key issue in Japanese digital cultural materials.
     Related materials:
     Kiyonori Nagasaki and Satoru, Nakamura, IIIF Workshop (Slides attached)

   - Afternoon:
Lecturers: Kiyonori Nagasaki, Satoru Nakamura and others
While TEI is a key method to share textual resources, it has not yet been spread among researchers for Japanese studies. The lecturers will introduce usage of TEI to the participants who are not familiar with it with the latest achievement of the Special Interest Group of East Asian / Japanese Text in the TEI Consortium. Following the morning introduction, it will be continued to learn TEI. Participants will try to encode a critical edition of tales of Genji and a corpus of a Japanese modern novel. It will also be treated to connect TEI encoded texts with IIIF contents. After that, they will have an opportunity to discuss and to treat their favorite materials.

4-1. Introduction to historical GIS with HuTime and other tools

- Morning & Afternoon 1:
  Lecturers: Tatsuki Sekino and others
  Historical GIS is a powerful tool for digital humanities. The lecturer will introduce the basic idea of historical GIS with a useful tool for visualizing temporal information, HuTime (http://www.hutime.org/). After that, some tools for connecting maps, timelines and IIIF contents will be treated. Discussion will be followed.
  
  Related materials:
  Tatsuki Sekino, Basic Linked Data Resource for Temporal Information (attached)
  Tatsuki Sekino, Time Information System Web HuTime: Comparison with Existing Web Applications (attached)

4-2. Collaboration and Crowd-sourcing

- Afternoon 2:
  Lecturer: Yuta Hashimoto
  Crowd-sourcing has been a key workflow in recent DH, especially in Japanese materials because Japanese materials have been digitized but not transcribed so much. The lecturer will introduce a crowd-sourcing transcription project that has been successfully conducted in Japanese ancient earthquake documents. Moreover, some tools for crowd-sourcing or collaborative works will be introduced by hands-on.
  Related materials:
  Stuart Dunn and Mark Hedges, Crowd-Sourcing Scoping Study Engaging the Crowd with Humanities Research (attached)
Yuta Hashimoto, A Survey of Digital Approaches to the Large-scale Transcription of Pre-modern Japanese Documents (attached)

5. QA and discussion

After summarizing these day’s activities and supplementing them, it will have time for a QA between all lecturers and participants. It is expected that participants will ask not only about their own techniques and researches but also about starting and conducting their DH projects to learn from lectures’ rich experiences.
Digital Humanities for Japanese Culture: Resources and Methods

Asanobu KITAMOTO
Center for Open Data in the Humanities
National Institute of Informatics
http://codh.rois.ac.jp/
Twitter: @rois_codh

ROIS-DS Center for Open Data in the Humanities (CODH)

2016 Pre-center started.
2017 Officially launched.

**Member:** One director and four project researchers (NII and ISM).

1. **Data-driven Humanities:** Innovation in humanities research using computer science technologies and tools.

2. **Humanities Big Data:** Innovation in non-humanities research using humanities data.
Open science is the key to fill the gap between two cultures and solve world’s problems.

Open Data
http://codh.rois.ac.jp/dataset/

Pre-modern Japanese Text
Kuzushiji

Edo Cooking Recipe
Modern Magazine
Open Software and Service
http://codh.rois.ac.jp/software/

PMJT Curation and ICP

Bukan Complete Collection

NIJI-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...
Dataset of Pre-modern Japanese Text
http://codh.rois.ac.jp/pmj/  
As of Nov. 2018  
Books: 3,126  
Images: 609,631  
You can download all images in a ZIP file.

Question 1: Deep Access

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. Deep access to the content of books needs more human or machine readers.
Question 2: Books as Data

1. Books are not only for reading text, but also for analyzing data.
2. Non-linear analysis of books, such as fragmentation and recombination.
3. Close reading vs. distant reading: the latter focuses on the quantitative analysis of text.
4. Data structuring is the key technology for quantitative analysis of data.

Japanese Culture Projects

1. Kuzushiji Recognition Challenge  
   • http://codh.rois.ac.jp/char-shape/
2. Collection of Facial Expressions  
   • http://codh.rois.ac.jp/face/
3. Bukan Complete Collection  
   • http://codh.rois.ac.jp/bukan/
4. Edo Cooking Recipes  
   • http://codh.rois.ac.jp/edo-cooking/
5. Historical Big Data  
   • http://codh.rois.ac.jp/historical-big-data/
Kuzushiji Recognition Challenge

How can we train machines to read or automatically transcribe Japanese old books?

Kuzushiji Challenge!
http://codh.rois.ac.jp/kuzushiji-challenge/

Kuzushiji, The Tale of Genji

**Kuzushiji**: cursive Japanese characters.

**Usage**: around 8th century to the end of 19th century.

**Problem**: most Japanese people can’t read them.

Kuzushiji: cursive Japanese characters.

**Usage**: around 8th century to the end of 19th century.

**Problem**: most Japanese people can’t read them.
Technological Advancement

- Human Transcription
  
  http://www.honkoku.org/

- Machine Recognition
  
  http://codh.rois.ac.jp/char-shape/

Kuzushiji Dataset

http://codh.rois.ac.jp/char-shape/

- Created by NIJL; curated by CODH; released in 2016 as open data.
- 4,645 character types and 684,165 character images (Feb. 2019).
- One million character images expected in 2019.
Bounding Box of a Character

Variation of the same character U+304B or 'Ka'.

A bounding box of a character and a cropped character image is included in the dataset.

Coordinates CSV

Total 4,043 pages

Coordinates CSV has Unicode and (x, y, w, h) of the bounding box.
End-to-end Recognition Model
(by Tarin Clanuwat and others)
Kuzushiji-MNIST or KMNIST
http://codh.rois.ac.jp/kmnist/

- 10 classes; 28x28 grayscale; total 70,000 images.
- A drop-in replacement for the MNIST dataset, in the original MNIST format and the NumPy format.

Complexity of KMNIST
Future: Smartphone Kuzushiji OCR

You go to a museum, and find a document written in Kuzushiji. You want to know what is written, and use a smartphone to read it. Out of 346 characters in this page, the model transcribed 323 characters, or 93.35% of characters correctly.

Collection of Facial Expressions

How can we evaluate the similarity of facial expressions to identify creators?
Revolution in Image Access

1. International Image Interoperability Framework (IIIF): rapidly growing among leading museums, libraries and archives.

2. Scale: 1 billion (high quality) images are accessible via IIIF.

3. Community: Moravian Library (Moravská zemská)

IIIF Curation Viewer (for Timeline)

http://codh.rois.ac.jp/software/iiif-curation-viewer/

Developed on Leaflet-IIIF
Good Old Analogue World

1. Scissors
2. Paste


Frictionless Digital World

1. Draw a rectangle.
2. Add to favorites.
IIIF Curation Platform

Client: User Tools

1. ICViewer
2. JSONkeeper
3. ICManager

Server: IIIF resources

4. ICMelder

Client: Manager Tools

5. ICFinder
6. Canvas Indexer
7. ML Services
8. ICPlayer
9. IIIF Server

Collection of Facial Expressions

http://codh.rois.ac.jp/face/

Tag: 「牛若丸」 (https://en.wikipedia.org/wiki/Minamoto_no_Yoshitsune)

Curation: collection of relevant (sorted) images under a theme, which can be shared or exhibited.
Human-Generated Metadata

```
"metadata": {
  "label": "title", "value": "イメージ"
},
"label": "collection", "value": "コレクション"
"label": "description", "value": "説明"
"label": "creator", "value": "作成者"
"label": "rights", "value": "権利"
"label": "subject", "value": "科目"
"label": "spatial", "value": "空間"
"label": "temporal", "value": "時間"
"label": "description", "value": "説明"
"label": "language", "value": "言語"
```

IIIF Curation Finder

http://codh.rois.ac.jp/software/iiif-curation-finder/

Click the tag, and get the list of a part of canvases annotated by the tag.
Machine-Generated Metadata

• “agent”: “machine” indicates machine-generated tags.
• “score” is the output of a machine learning system.
• Canvas Indexer accepts an image tagging bot for machine-generated metadata.

Automated Image Tagging by Machine Learning

- ResNet 101 classifier learns 5000 tags from 9 million images (Open Images Dataset V2).
- We used the model already trained on general photographs, not on our dataset.
Serendipity by Machine Learning

1. Correct or not correct, that is not the question.
2. Serendipity: unexpected encounter with images.

Tag: Circle

Close vs. Distant Reading

- **Distant reading**: treat work as data, and study the collective characteristics of data, apart from the context of the work.
- **Close reading**: interpret work in the context of the background of the work and knowledge related to the work.
- **Distant reading** using IIIF Curation Finder and machine learning help scholars start their own **close reading** assisted by digital tools.
Bukan Complete Collection
How can we compile many versions of books and create the time-series database?

Time-Series Historical Sources
• Bukan: directory of state king families and bureaucrats of the central government in the Edo period (1603-1868).
• Time-series publications for 100 to 200 years with a peak frequency of a few times in a month.
• 381 versions of Bukan will be released as open data.
Kansei Bukan (1789)
http://codh.rois.ac.jp/pmjt/book/200018823/

Many iconographic elements are not only visually attractive, but also visual features for identification.

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

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

- **Text A**
  - Transcription
  - OCR
- **Text B**
  - Transcription
  - OCR
- **Digital image A**
- **Digital image B**

**Image-based difference**

**Non-textual difference**

---

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

Differential Reading

1. A new style of reading supported by machines, like distant reading.
2. Changes are enhanced by machines to help humans notice the change.
3. Machines could produce better results for simple, repetitive and tedious tasks.
4. Comprehensive survey of Bukan may reveal the correct order of multiple versions.

Edo Cooking Recipes
How can we transform old knowledge for the modern life of people
Easier Access to Old Recipes

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

PMJT Dataset
(from NIJL)

Edo Cooking Recipe Dataset
(Created by CODH)
Adapted Material on NIJL Dataset
(from NIJL)

Edo Cooking Recipes

- It consists of more than 100 cooking recipes of egg dishes.
- The book indicates culinary culture toward enjoying the variation of cooking.
Data Structuring Ladder

• We designed a workflow to reduce the barrier for citizens.
  1. Organize digital images in the Dataset of PMJT.
  2. Transcribe text from digital images.
  3. Translate old Japanese text to modern Japanese text.
  4. Structure modern Japanese text to the format of the cooking recipe service.
• The content of books is more human readable and accessible in the later stage of the workflow.

Translation for Reusability

• **Quantity**: time and amount is estimated to reduce the risk of failure and make the recipe actionable.
• **Ingredients**: Some of local ingredients are changed to similar ones that are easier to obtain now.
• **Tools**: Old cooking tools are changed to modern ones, such as refrigerator and food processor, so that we can take advantage of today’s civilization.
• **Photographs**: stepwise photographs and an attractive photograph for the final presentation convince people that the recipe is reproducible.
Stepwise Photographs

Edo Cooking Recipe Dataset  (Created by CODH)
Adapted Material on NIJL Dataset (from NIJL)

Photographs by Cooking Experts
Where to Deposit Data?

- Cookpad is the largest cooking recipe service in Japan.
- We decided to deposit recipe data to “Cookpad Edo Dishes,” run by Cookpad and The Japan Society of Home Economics.

http://cookpad.com/recipe/4153357

Typical Responses from Citizens

Me: "I will release Edo cooking recipes on our CODH website."
They: “That sounds nice” (without interests).

Me: “I will release Edo cooking recipes also on Cookpad.”
They: “That’s a good idea!” (with a brilliant smile).
Big Impact from the Release

Chain of Creations

1. “Tsuku-repo” is a service to submit derivative works by users.
2. Recipes show that they take the recipe into their daily lives.
3. Actionable data can stimulate users to make new actions.
Why Citizens were Excited?

1. Hearing the word “Cookpad” they can instantly imagine how to use the data.
2. Recipe with photographs convince people that the recipe is reproducible.
3. To see data in their familiar platform makes them feel that it is our data.
4. A gap between modern platform and old content attracts curiosity of people.

FAIRness for Citizens

• **Findable**: deposit data into a popular platform in the daily lives of citizens.
• **Accessible**: transform the content into modern format with additional information.
• **Interoperable**: interface and description does not require new training.
• **Re-usable**: convince people reproducibility and allow them link their derivative work.
Historical Big Data
How can we reconstruct the past world from old documents by quantitative analysis?

Historical Documents for Non-Humanities Discoveries

1. Historical documents contain the description of the world such as weather and season.
2. Historical documents have potential for non-humanities discoveries by structuring data.

Source: Kotenseki Sogo Database, Waseda University
Tweet vs. Old Diary

1. Social media is recording the current history. How about the past, e.g. using people’s diary?
2. How can we continuously extend the latest technology from the modern to historical data?

Historical Data from Human Recordings

Situation Record: human sensory observations of the world, either visual, auditory or tactile, recorded in the form of text.

1. Modern situation record: born-digital text on the internet or smart phones in the form of social media or communication tools.
2. Historical situation record: analog text on paper in the form of diaries and documents.
Historical Big Data

http://codh.rois.ac.jp/historical-big-data/

Historical Record

Modern Record

Historical Observation

Commonalities as record

Historical Evidence

Commonalities as historical data

What is Ansei Edo Typhoon?

A typhoon in the third year of Ansei (1856). Passed near the city of Edo that caused very high storm tide, and many casualties.

"Ansei fūburushū" published in 1856 (stored in Waseda University Library)
http://archive.wul.waseda.ac.jp/kosho/wo01/wo01_03959/wo01_03959_0002/wo01_03959_0002_p0011.jpg
#Ansei_Typhoon

**September 9, 1856 from noon to night**

- Officer@Nii-jima Island
  - 夜入風天気大雨

**September 23, 1856 4 PM**

- Harris@Shimoda 4 PM
  - The wind at four P. M. was S. S. E., and continued to haul to S. S. W., at which point the gale was heaviest.

- Harris@Shimoda 4 PM
  - Yesterday at four P. M. the wind began to blow fresh from E. S. E., with rain.
#Ansei_Typhoon
September 23, 1856 early night

Ishikawa@Hachioji
夜五ツ半時より八ツ時迄大風ニ而一世大あれも、尤上々度ハ大あれも、江戸人死多し

Sekiguchi@Yokohama
陰天、夕刻より雨降出し段々風強く、雷鳴両三声

September 24, 1856 early morning

Sanemon@Tamamura
昨夜大嵐丑寅夜七ツ頃北風ニ成る

Genbei@Soma
曉七ツ半頃より止風猛し、大柳武本根返り、水流通り、近年希成大時化、立木ハふれ又南之方向強し

Ootaka@Mito
明方辰巳風二相成漸く静二相成
Estimated Track of Ansei Typhoon

Integration of the change of wind direction helps to estimate the track of the typhoon. History is a lesson to learn about a significant event, e.g. maximum impact.

Data Structuring Workflow

1. **Unstructured data** (image)
2. **Unstructured data** (plain text)
3. **Semi-structured data** (markup text)
4. **Structured data** (list of records) | **Historical record**
5. **Structured data** (analysis-ready data)
Challenges for Historical Big Data

Personal or family diaries
- Weather, earthquake, ...
- Events, accidents, ...
- Human relationships, ...

Official records
- Economies, industry, ...
- Organization and promotion, ...

Travelogues and reports
- Places and regions, ...
- Human mobility, ...

Transcribed or OCRed → Filtered → Structured

Reconstruct the historical world and the society.

Summary
Japanese Culture Challenges

1. Japan has a huge amount of historical documents yet to be analyzed.
2. Due to delayed digitization and kuzushiji characters, electronic text is very few.
3. Computer vision and machine learning has potential for realizing deep access.
4. Human and machine can collaborate on data structuring for digital humanities.

Japanese Culture Projects

1. Kuzushiji Recognition Challenge
   • http://codh.rois.ac.jp/char-shape/
2. Collection of Facial Expressions
   • http://codh.rois.ac.jp/face/
3. Bukan Complete Collection
   • http://codh.rois.ac.jp/bukan/
4. Edo Cooking Recipes
   • http://codh.rois.ac.jp/edo-cooking/
5. Historical Big Data
   • http://codh.rois.ac.jp/historical-big-data/
Collaborators

• Center for Open Data in the Humanities (CODH)
  • Chikahiko Suzuki, Mika Ichino, Tarin Clanuwat

• National Institute of Japanese Literature (NIJL)
  • Kazuaki Yamamoto, Kumiko Fujizane, Center for Collaborative Research on Pre-modern Texts

• Montreal Institute for Learning Algorithms (Mila)
  • Alex Lamb

• University of Tokyo
  • Akira Takagishi

• FLX Style
  • Jun Homma

• Albert Ludwigs University of Freiburg
  • Tarek Saier

• AMANE LLC.
  • Hiroshi Hori, Misato Hori, Syouko Terao

• Teikyo University
  • Junpei Hirano

• Tokyo Metropolitan University
  • Kooiti Masuda
Chapter 4

Contexts of Digital Humanities in Japan

Kiyonori Nagasaki
International Institute for Digital Humanities, Japan

ABSTRACT

This chapter describes the brief history and recent trends in digital humanities in Japan, which had been led within the context of IT (information technology) and recently has strongly involved humanities researchers. According to the analysis of 991 technical reports by the Special Interest Group for Computers and Humanities (SIG-CH), the fields of linguistics and literary studies have been dominant while recently the history field has been increasing its number of the presentations, and many other fields in the humanities have been treated in a small percentage. Japanese texts have some difficulties in the digital environments. Although the recent developments in IT partially solve them, other attempts to improve the DH research environment have been activated. The policy of Japanese government to promote open science and open data will make DH in Japan more fruitful in the future.

INTRODUCTION

This chapter describes some contexts of digital humanities (DH) in Japan. With Japanese DH being so widely proliferating nowadays, it is difficult to introduce every aspect and every project. The Special Interest Group for Computers and Humanities (SIG-CH) has played a prominent role in Japan for about three decades and the Japanese Association for Digital Humanities (JADH) has since 2011 served as a counterpart in the realm of international DH activities. Additionally, a movement in the area of digital archiving has become popular among cultural institutions such
as libraries, museums, and industries, adapting to the Japanese environment. This movement bears relationship to activities concerning digital cultural heritage, but function much more broadly. As it worked in a complementary relation with the field of DH, its activities will be mentioned below.

A BRIEF HISTORY

As in other countries, efforts to establish digital scholarship in the humanities have been initiated starting a few decades ago by scholars in Japan who specialize in either the humanities or information technology. Activities to apply digital technologies in humanities have been carried out since establishing the Mathematical Linguistic Society of Japan in 1957. According to Sugita (1982), during the age of the mainframe computer, some research institutions such as National Institute of Language and Linguistics, National Museum of Ethnology, and the National Institute of Japanese Literature installed computers to digitize their resources and to leverage digitized data.

Following upon the earliest attempts, several communities were established at the end of the 1980’s based on the impetus of the proliferation of the IBM PC. One, the SIG-CH, was formed under the auspices of the Information Processing Society in Japan, the largest computer science society in Japan. (SIG-CH will be discussed below) The others were the Japan Society of Information and Knowledge and the Japan Art Documentation Society.

As the SIG-CH was founded under the auspices of an academic society of computer science, its orientation has tended to be toward the approach of computing rather than the humanities, although participants have always shown a strong interest in cultural heritage. This means that activities in the SIG-CH have been evaluated primarily from the viewpoint of computer science. On the other hand, there were fundamental difficulties facing humanities researchers who wanted to engage in the field of computer science and the related IT sector, since those areas had long-established methods of evaluation of academic achievements that differ from those in the humanities. Thus, if a humanities scholar made an IT-centric academic achievement, it was not duly evaluated in humanities academic circles. Despite these limitations, the number of humanities scholars who engage themselves with the computer science and IT communities has been gradually increasing to the extent that they have established the JADH.

Besides the activities of SIG-CH and other societies carrying out DH-like studies, several academic communities within traditional humanities including the field
of Buddhist studies have also addressed applying digital technologies in order to facilitate their research in the humanities since 1980’s. After that, several academic communities have been established mainly due to the availability of the Internet. It is especially noteworthy that digital scholarship societies in archaeology, English corpora, or Asian literature were formed in the 1990’s.

Large government grants have frequently provided impetus to the DH field, among which there have been several unusually large grants. The Ministry of Education awarded a 5-year grant totaling around one million USD to a project called “Computer and Humanities: Promotion of Humanities Research with Support from Computers,” which was formed by the core members of the SIG-CH and related influential researchers. It was a blockbuster of a grant that supported many symposia and the publication of a series of textbooks for the field. The Ministry of Education also started the Center Of Excellence (COE) program, which offers grants. The grants supported several centers addressing DH-like studies such as Kyoto University, Tokyo University of Foreign Studies, and Ritsumeikan University. Most notably, the Art Research Center at Ritsumeikan University not only digitized its Japanese cultural resources including Ukiyo-e, provided various implementations of the digitized data but also collaborated with centers and departments having Japanese collections. They also engaged themselves with international DH communities, led by Professor Ryo Akama. Institute for Research in Humanities in Kyoto University established the Center for Informatics in East Asian Studies, a leading center on DH to provide various useful DH tools and resources for the community and takes up the challenge of several significant developments including Morphological Analysis of Chinese classical texts. The Historiographical Institute at the University of Tokyo has been providing huge databases in order to compile Japanese history. Its activities for the databases started in 1980’s and were supported by the COE grant during its history.

Besides such activities, institutional and other types of supports have been conducted for various projects. Institutions which belong to the National Institute for Humanities have provided various databases for humanities and an integrated search system for the database so-called NIHU-INT. According to Muller and et al. (2018), the SAT Daizōkyō Text Database Committee (SAT project), led by Professor Masahiro Shimoda has been managed since 1994 by a community of Buddhist scholars through support not only of governmental grants but also Buddhist organizations and temples so that researchers who use Buddhist texts can research them easily through digital media.

After 2011–2012 when JADH was formed and approved as a constituent organization of ADHO (Alliance of Digital Humanities Organizations), the Japanese DH community steadily began to expand its cooperation with the international DH community. Recent trends will be explained below.
Tendency of Research in DH

An overview of past trends in DH in Japan can be shown through past presentations of SIG-CH. This data is useful as an overview, since its non-refereed workshops with technical reports had been held quarterly (recently tri-annually) and open to various fields, while it is still an IT-focused field as mentioned above. Figure 1 shows that the highest number of presentations is in the field of linguistics throughout the entire period. Recently it has grown in proportion to twenty percent within the 2009-2013 term. The reason for the growth may be that the environment for textual analysis of Japanese texts became easier as Unicode technologies have been popular.

Figure 1. A graph of the target fields of 991 presentations with technical reports until 2017. They are summarized every five years except the last term.

*For a more accurate representation see the electronic version.*
and large Japanese corpora have been released. Besides some fields that frequently presented, many fields were targeted in low number of times, such as area studies, religious studies, geography, archives, cognitive science, LIS, folklore, musicology, and ethnology.

The top 10 fields are summarized in Figure 2. History has recently increased its activities due to the efforts of several energetic young researchers in the field. Literary studies has recently decreased. Museum studies has decreased as compared to the initial stage; in the initial stage of this community, museum researchers and practitioners gathered in order to investigate efficient ways of presenting exhibitions. Dance studies also have increased due to two active research groups. Recently, as development for digital cultural resources has gradually been spread, presentations for all fields of the humanities have been increased.

Figure 3 is a graph of methods adopted in the presentations. The analysis of digitized resources has been significant since 1994, but recently decreased slightly. On the other hand, software development has been gradually increasing. The sum of the percentage of software development and digitization is greater than the analysis. Recently, the trends may have changed slightly due to technological developments, change in generation and change of environments in the humanities. For further research, it is recommended to see the online journal and conference abstract books of the JADH, find related papers in the technical reports of SIG-CH and the conference proceedings of the annual domestic conference.

Figure 2. Top 10 target fields of presentations

*For a more accurate representation see the electronic version.
As an interdisciplinary approach to education has been promoted in Japanese universities, applying digital technologies to humanities has been taught in several universities for over a decade at the University of Tokyo, Tokyo Institute of Technology, Ritsumeikan University, Doshisha University, Tsukuba University, among others. For example, the University of Tokyo established the Interfaculty Initiative in Information Studies with the Graduate School of Interdisciplinary Information Studies in 2000. In addition to its original characteristics, the graduate school started an interfaculty minor course for DH together with the graduate school of Humanities and Sociology, Information Science and Technology, Frontier Sciences since 2012. The course teaches basic DH technologies such as textual analysis, Text Encoding Initiative (TEI), International Image Interoperability Framework (IIIF), metadata and their backgrounds including license issues. As the course is open to every graduate student, students from various fields in humanities, such as linguistics, Western history, Eastern history, Buddhist studies, religious studies, and literary studies in various languages participate in it. Moreover, some students come from fields in the natural sciences and technology.

Doshisha University provides DH-like education in the Faculty and Graduate School of Culture and Information Science, which was established in 2005. The department includes the pioneer of textual analysis of Japanese literature, Professor Masakatsu Murakami. They provide consistent education to study cultural resources such as text data of classics via statistical analysis. Ritsumeikan University provides
Digital Humanities for Arts and Cultures as a course in the Graduate School of Letters. The curriculum includes various themes based on the results of several large grants and the tremendous digitized Japanese cultural resources in its Art Research Center. Besides the official curricula, seminars and workshops for DH have often been held to disseminate how to dive into DH, treating some DH technologies such as TEI, IIIF, Omeka, Zotero, Python and so on. Recently some of them have been held by graduate students.

**DIFFICULTIES OF JAPANESE TEXTS AS A NON-SPACE WRITING SYSTEM**

As most modern and pre-modern western writing systems explicitly represent the division of words in a sentence by spaces or breaks, it has been easy to use computers to analyze texts based on each word and its meanings. However, there are several modern and pre-modern writing systems that do not explicitly indicate word separation in texts; that is, all words in a sentence are contiguous. A major contemporary representative of this kind of writing is seen in the language system of East Asia. Moreover, a popular Japanese pre-modern writing system called kuzushi-ji (cursive style characters) had often been presented with undivided characters even in typesetting until the late nineteenth century (see Figures 4 and 5). Both figures are examples of contiguous writings. The example on the left was printed by woodcut typesetting. Then the types were made as contiguous characters. But this kind of typesetting suddenly became almost obsolete in Japan. After that, xylographic printings were prevalent for about 200 years. The example on the right was a typical xylographic printing including some difficult elements, which will be explained below.

The lack of word separation evokes not only ambiguity but also multiple interpretations, but creates an aspect of cultural richness in Japanese culture. However, as a result, Japanese texts have intrinsically presented difficulties: not only in the case of textual analysis but also in both manual and automatic transcription in the digital era. This presentation will discuss problems in these writing systems and the current situation of attempts to resolve them through the methods of digital humanities.

**DIFFICULTIES OF TRANSCRIPTION**

Recent Japanese texts do not have serious problem with Optical Character Recognition (OCR), due not only to the separation of each character but also accuracy and clarity of printing. However, it is difficult to OCR books printed even 100 years ago for
Figure 4. Pictorial tales of Genji, an example of contiguous writings.

Figure 5. Tales in Ise, another example of contiguous writings.
two reasons: first, most of them used relatively complicated characters which are
difficult to recognize, and second, they included parallel embedded small-font size
script (called ruby in HTML5), which express the pronunciation of a word, and are
physically too close to the explained word to OCR (see Figure 6), even though they
were printed by metal typesetting. Three hundred years ago, during the Edo period,
Japan isolated itself from other nations, up until 1868. During this period, most
Japanese publications were printed by xylography because it was easy to reprint
for printing companies at that time. In this printing style, characters were often
contiguous, and the writing style of characters was partially cursive (see Figure 7).

Recently, some researchers and practitioners are attempting to develop tools
for automatic recognition of kuzushi-ji based not only on the shape of individual
characters but also by continuous shapes of characters. They have not yet reached the
stage where they are able to transcribe all characters accurately, for both technical
and intrinsic reasons, but the technology can nonetheless assist in reading such texts
by showing candidates of characters. Figure 8 shows an attempt to automatically
recognize a cursive character by machine learning on heroku by cropping a character
from an image on the Web via IIIF image API. Figure 9 shows a search result of
an image search on a collaborative transcription software called “Smart-GS”. The
figure means that similar series of contiguous characters were found in a text. The

Figure 6. A Guidebook to practical cooking

79
software was implemented by Professor Hayashi of Kyoto University, adopting an image search algorithm developed by Professor Terasawa of Hakodate University. Moreover, Center for Open Data in the Humanities (CODH) has been attempting OCR of kuzushi-ji via deep learning by providing dataset developed with National Institute of Japanese Literature. CODH has held workshops and contests for developing kuzushi-ji OCR.

However, there are special difficulties presented when a needed character is not encoded in Unicode. It seems to be similar with the case of the Medieval Unicode Font Initiative, but the number of unencoded characters would be much more in the Japanese case included in East Asian culture. Especially, as Japanese culture has been involved with foreign cultures and developing them in its contexts, several writing systems are preserved in its cultural resources, including Kanji, Hiragana, Katakana, Hentaigana, and Siddham scripts.

There are already over 80,000 characters as CJK unified ideographs and its extension in Unicode, but this number will continue to increase. Hentaigana (including over 200 glyph shapes) was finally registered in Unicode 10.0. Siddham scripts were encoded in Unicode 8.0 with its variant characters (Figure 10) by the efforts of the Script Encoding Initiative, international experts, and the SAT project.
As the Siddham script was created in India, they only represent phonetic value in principle. However, usage of the scripts had been changed in Japan, probably due to the effect from Chinese scripts. As a result, some characters which have the same phonetic value but different glyph shape were assigned different meanings. This means that the system for Indic scripts must be changed for the usage in Japan. Over three years were necessary for the changes, but cultural diversity was implemented in the digital technology by the activity.

In order to make easy-to-use digital scholarly editions for Japanese texts, especially classics, this process will be continued.

Efforts in transcription, due to the commoditization of digitizing textual materials in hi-resolution digital image databases have also been developing in Japan. Especially, the National Diet Library in Japan has been addressing the publication of digitized collections including over 300,000 books--since over decades ago and recently re-iterated that most of them are to be released in the public domain. And some institutes such as Kyoto Prefectural Library and Archives and the University of Tokyo Library are publishing their digitized collections under open license. The Art Research Center in Ritsumeikan University and the National Institute of Japanese
Literature have released many digitized textual resources under academic licenses. The latter institute plans to distribute parts of their contents under open license in 2018 in its new comprehensive digitization project. Needless to say, these are useful to enhance the convenience of humanities research. Especially, in Japanese contexts, many humanities researchers mention that validation of research results has been made much more efficient by the increased use of the digitized images.

Crowd-sourced transcription has also recently emerged in Japan. The transcribe JP project has been conducted as a SIG of the JADH. It provides a Web service for transcription with Omeka and Scripto plugins (see Figure 11).

This started as a micro task crowdsourcing project on October 2015 in cooperation with the Crowd4U project\(^5\). Contributors can determine whether a character is accurately OCR-ed by comparing a candidate character with a piece of an image only by one click. (Figure 12) The first experiment was conducted from December 2015 to April 2016, adopting a book which was released from the digital collection.
of the National Diet Library in Japan as public domain. 10630 tasks were carried out on PCs and smartphones so far. As a result, the first section of the book was transcribed. But it seems that too much effort is required for the task. Therefore, the project is preparing for a more efficient way.

Transcribing Japanese pre-modern texts have recently been addressed successfully by a group studying historical earthquakes. Generally speaking, it is somewhat difficult for the public to read kuzushi-ji in such texts. However, according to Hashimoto (2017), by cooperating with a mobile application for learning kuzushi-ji, the project has successfully gathered people who can transcribe these scripts. The Web collaboration system Minna de Honkokù (Transcribing by Everyone) has implemented gamification such as a ranking system to fascinate the collaborators. This project has transcribed over five million characters to date. The progress is reported every night on a twitter account of Dr. Yuta Hashimoto who developed both the learning app and the Web collaboration system.
Figure 11. Web service for transcription provided by JP project

Figure 12. Web service for transcription provided by JP project
DIFFICULTIES IN WORD SEPARATION

In spite of the difficulties of transcription, there are many digitized texts in Japanese. Aozora-Bunko, a public domain Japanese text repository similar to the Gutenberg Project, provides over 10,000 texts on its website and GitHub. The National Institute of Japanese Language and Linguistics (NINJAL) publishes several encoded historical Japanese texts with Part of Speech (POS) tags on the Web and Web services of textual analysis on modern Japanese texts including 100 million words with POS tags each word in its original format as the Balanced Corpus of Contemporary Written Japanese (BCCWJ). The SAT project also provides digital texts of Buddhist scriptures and consists of 100 million characters mainly in Chinese and Japanese with some philological elements and useful tools on the Web.

The texts of NINJAL consist of separated words with POS tags, but most of the others do not use this method. Other methods for textual analysis are common in Japan: One is n-gram analysis, where a character is regarded as one “n”. The other is developing tools for automatic separation of words sometimes with POS tagger, such as Mecab, Chasen, and Kuromoji. These tools realize a high degree of precision, but sometimes produce errors. In this case, one has to manually correct the result of the tools if sharing exactly-processed texts are necessary. Moreover, even if a separation is not mistaken grammatically, it might support an interpretation in some cases. Such kinds of cases can also occur in word-separated corpora. For example, “北大通り” can mean both “Hokkaido University Street (北大通り)” and “North Large Street (北-大通り)”. This type of writing system includes such kinds of issues.

RENDERING OF TEXTS

In XML-formatted texts, such as those maintained in TEI, Journal Article Tag Suite (JATS), and so on, representation of breaks in source XML files seems to be regarded as a space as a separation between words in popular stylesheets. But in the case of non-separated texts, it causes problems such as inaccurate separation. The XSLT-processed Japanese text in the left of the Figure 13 must exclude spaces between characters in spite of line-breaks in the XML source (refers to the right side of Figure 13). Conversely, as a Japanese semi-governmental open access journal system, so-called J-Stage, adopting JATS ignores line breaks even in English, the words are contiguous in this case. This problem seems to be recognized in ePub with solution in CSS according to the target language. While it must already be discussed even in context of DH because non-spacing texts have been generated in various time and place, the differences of treatment of the line breaks in XML
source files should be carefully treated regarding not only representation but also analysis of texts. These issues will be addressed in the special interest group for East Asian/ Japanese Texts, which was established in the TEI consortium in 2016.

In the context of current DH, huge humanities resources still remain dormant. To awaken these resources and the field, these kinds of issues should be gradually revealed and need to be solved from both practical and abstract viewpoints. Through solving them earnestly under global communication, DH will come to better fruition.

**RECENT TRENDS**

As recent trends of DH in Japan except for the above mentioned, IIIF and Linked Open Data (LOD) have been focused on by many researchers and practitioners. IIIF seems to be emerging in the community of digital cultural heritage and Web technologies. Being based on the current Web technologies such as JSON-LD, it has been spreading rapidly in the world and gradually in Japan. The University of Tokyo, Keio University, Kyoto University adopted it relatively early for their digital collection of Japanese and East Asian classical books and some other materials;
the National Institute of Japanese Literature used it for its collection of Japanese classics. Moreover, this year, as the National Diet Library implemented IIIF for its digital collection including over 300,000 public domain books, the environment to treat Web-based hi-resolution images efficiently has been in place also for Japanese and East Asian resources. As IIIF allows the leveraging of hi-resolution Web images and other digital contents beyond borders of organizations and Web sites via some APIs, which are defined by their specifications, some DH researchers recently began to utilize them by developing various software and systems. CODH has released a specialized IIIF viewer²⁰ called “IIIF Curation Viewer,” which enables “curation” of Web images across the borders of Web sites. The Digital Humanities Initiative in the University of Tokyo participates in the development of a viewer for IIIF and provides a collaborative system using IIIF such as a research environment for Buddhist icons²¹ and a collaborative work bench²² for alignment of Buddhist texts released in various Web sites in the world.

Some institutions such as the Khirin (Knowledgebase of Historical Resources in Institutes) project²³ in the National Museum of Japanese History and the University Archives²⁴ in the University of Tokyo have been working to adopt not only IIIF but also LOD to utilize their digital collections. These kinds of works will show more important value by integration with spatiotemporal information such as calendrical period resources as discussed by Sekino (2017) and gazetteer²⁵ which have been released recently by efforts of the Humanities GIS Research Group²⁶.

CONCLUSION

As we have seen, DH has been developing in various ways in Japan like in other countries. Along with the evolution of information technology, DH has also been developed. Especially, the evolution of Web technology has recently been stimulating it. Due to the recent crisis of the humanities in Japan as discussed by Yoshimi (2016), DH will become more important not only for the evolution of humanities themselves but also for the public engagement. Moreover, DH has recently been promoted by the movement of open science, which has been supported by the government including the Cabinet Office in Japan so that wider range of people will engage in research of the humanities. As the result, the shared reusable data and environments will realize more rich and varied humanities in the future.
REFERENCES


ENDNOTES

2. https://kaken.nii.ac.jp/grant/KAKENHI-PROJECT-07207105/
5. http://wwwap.hi.u-tokyo.ac.jp/ships/db-e.html
12. The system was developed by Jun Homma.
17. http://pj.ninjal.ac.jp/corpus_center/chj/
http://21dkz.l.u-tokyo.ac.jp/SAT/index_en.html
http://codh.rois.ac.jp/software/iiif-curation-viewer/
https://dzkimgs.l.u-tokyo.ac.jp/SATi/images.php?alang=en
http://bauddha.dhii.jp/SAT/iiifmani/show.php
https://khirin-ld.rekihaku.ac.jp/
https://park.itc.u-tokyo.ac.jp/UTArchives/uta/s/da/page/home
https://www.nihu.jp/ja/publication/source_map
http://www.h-gis.org/
IIIF Workshop

Kiyonori Nagasaki
Satoru Nakamura

Today’s Agenda

- A brief introduction to IIIF
- Use cases of IIIF
  - Usage of Layer function
  - Collection of IIIF contents across institutions
  - Image Cropping & Annotations
  - Integration with transcriptions

- Tutorial:
  - Try some applications!
What is IIIF?

International Image Interoperability Framework™
Enabling Richer Access to the World’s Images

Learn how to get started

Draft IIIF v3.0 Specifications Released

Community Focused
Defined APIs
Plug ‘n’ Play
Difficulties for cultural institutions:

- To prepare budget to order custom-built application for Web images
  - While Web interface has rapidly been evolving.
- To extend usability of Web images across Web sites of various institutions
  - While a collection often has been separated in several institutions
  - While distributed Web contents should be evaluated by gathering in various contexts
- To embed scholarly achievements to Web images sustainably due to the short lifetime of Web technologies
  - While Web images should be valued by research results

Before IIIF:
Difficulties for users incl. researchers:

- To learn usage of each custom-built “convenient” (and unimproved) application for Web images in the world
  - While Web interface has rapidly been evolving.
- To look over interested Web images across Web sites of various institutions
  - While a collection often has been separated in several institutions
  - While distributed Web contents could be leveraged by gathering in various contexts
- To embed their achievements to Web images sustainably due to the short lifetime of Web technologies
  - While they want to value Web by their research results

Before IIIF:
• **IIIF** International Image Interoperability Framework

• Started in 2011 by:
  • BL, U of Oxford, BnF, National Library of Norway, Los Alamos National Lib, Stanford U Lib, and Cornell U Lib with a grant from the A. Mellon foundation

• Efficient interoperability of digital facsimiles delivered from cultural institutions in the world on Web
  • Corresponding with trend of Web technologies
  • Providing stable, usable, and open specifications
  • Accessing a part of a Web content from outside of the Web site, such as a part of an image, and a character.
  • Developing open source software to utilize the framework
  • Extending to multimedia such as sound, video, and 3D

**IIIF has offered a platform**

• Which enables to distinguish roles of contents/data providers from roles of users including agents who leverage and re-distribute the contents

• As the result:
  • Cultural institutions can provide their contents easily according to the international common way
    • Relieved people of considering a proper way of providing contents
  • Users can leverage all images and the other types of contents on Web with the unified method.
IIIF properly mediates providers and users

- Each viewer gives a unified interface for all IIIF content.
- Not only providers but also viewers can easily embed additional data/information to the contents.

Providing several APIs

- **API**: Application Programming Interface
- Everyone can develop (or make someone develop) applications leveraging Web contents via these APIs
- **IIIF APIs**:
  - **Image API**
    - To manipulate a Web image via URI (URL) when getting the image
  - **Presentation API**
    - To bind Web contents including images and annotations through the Image API and Web annotation defined by W3C
    - Also multimedia contents
  - **Authentication API**
  - **Search API**
Image API

Presentation API 2.1

Basic concept

A IIIF Manifest file represents a binding or an item
From another angle…

Presentation API 2.1

Relationship with other contents via Web annotation

The Presentation API must be written in JSON-LD format.
You can use any favorite Viewer:

- **Universal Viewer**
  - By BL, Wellcome Library and so on
- **Mirador**
  - By Stanford, Harvard, and other Web developers widely
- **IIIF Curation Viewer**
  - By CODH (Japan)

- And other various viewers and solutions...

Universal Viewer: http://universalviewer.io/
Mirador: demonstration

- Integration of distributed images in multiple institutions/Web sites
  - To represent the reconstructed page of medieval manuscripts keeping independency of each digital repository
  - Finally we’ve achieved it with a free software under an international open standard.
- Multi-layer representation
  - Vermeer’s picture in two tracks
  - The function reveals both pictures via multi-layers with X-ray and infrared photographs
IIIF Curation Viewer


IIIF drop icon or “IIIF drag&drop”

- The three viewers accept a function of IIIF drop icon.
- Please drag a 🖹 icon and drop into a window of a viewer.
License notification

- It’s required to be included in the “IIIF manifest”.
- Easy to check

An example in 2016

- SAT Taishozo Image DB:
  - https://dzkims.l.u-tokyo.ac.jp/SATi/images.php
  - To implement these functions, our project did only prepare JSON format data according to the IIIF specification, not need to develop a new software.

- Collaborative annotation system for the DB
Spread of IIIF

- Established the consortium in 2015
  - 54 institutions joined now.
  - While the framework is available for free!
- Many institutions provide their contents via IIIF in the world
  - Gradually increasing in Japan:
    - National Diet Library
    - National Institute for Japanese Literature (国文研)
    - University Libraries: Kyoto, Keio, Tokyo, Kyushu, Shimane
The (payment) members of the IIIF consortium

10,000 USD per year

No one need to pay for using it!


Institutions adoption IIIF in Japan

- University and National Libraries
  - National Diet Library, Kyoto-U, Keio-U, U of Tokyo, Shimane-U, U of Chiba, Kinki U
- Research Institute
- Cultural institutions in local governments
  - Okayama Prefectural Archives, Nerima City Public Library, Nakano City Library, Ueda City Multimedia Information Center
- Research Projects
  - SAGA University, SAT Daizokyo Text Database Committee...
There are several ways to adopt IIIF to your images as a provider

- Ask to a digital section in your organization
  - Strongly recommended!
- Ask to an IT specialist among your colleagues
- Set up a system by yourself using any open source software
  - There are many tools to support it.
- Put your images on IIIF hosting service
  - If you have a few budget.
- Put your images on FromThePage if you want to transcribe them with your colleagues or crowd-sourcing.

There are a lot of possibilities as a user of IIIF-compliant contents in the world

- Developing virtual collections gathering from any Web site
- Comparing any content on your browser easily
- Putting any content on a map and timeline
- Putting transcription on any content
- Citing a favorite part from any content
- And more…
- Through any convenient tool which has been developed in the world
- Some of them will be introduced to consider the possibilities
Use cases of IIIF:
Usage of Layer function

Layers with Automatic Image Colorization

http://da.dl.itc.u-tokyo.ac.jp/mirador/?manifest=https://nakamura196.github.io/colorization/data/hiraga/result.json&sidePanelOptions=on
Rebooting Memories

- Project by Hidenori Watanave (UTokyo)
- “... This work includes the creation of digital archives, and the colorization of black-and-white photographs. ... increase the information value of by converting “stocked” records in society into “flowing” images, thus transmitting past memories to the future based on emerging communication. ...”

Neural Network-based Automatic Image Colorization

“Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification” [Iizuka and Simo-Serra et al. SIGGRAPH 2016].
Heatmap which visualizes IIIF Image API call logs

Visualizing which parts of IIIF images are looked by users

- Conducted by
  - Chifumi Nishioka (Kyoto University)
  - Kiyonori Nagasaki (UTokyo)
- Abstract:
  - "... We developed a script that analyzes IIIF Image API call logs. The script generates heatmaps that visualize which parts of IIIF images are looked intensively. The generated heatmaps are shown over images using Mirador's layer functionality, ..."
Use cases of IIIF: Collection of IIIF contents across institutions

IIIF Discovery in JAPAN

http://iiif2.dl.itc.u-tokyo.ac.jp

News
• 2018-06-13: IIIF Discovery in Japan (Prototype ver.) released
IIIF Discovery in JAPAN

- Prototype system for accessing IIIF contents published by Japanese institutions
- Developed by:
  - Satoru Nakamura (UTokyo)
  - Kiyonori Nagasaki (UTokyo, Intl. Institute for Digital Humanities)
- Points:
  - Collect IIIF manifests published by Japanese institutions, without any theme
  - Display IIIF contents with several viewers, based on objectives and preferences

  - [https://docs.google.com/presentation/d/1EnTe1hpCbcRTTaeT49Pp_rojINPHh3MiXnzFcAkPVKA/edit?usp=sharing](https://docs.google.com/presentation/d/1EnTe1hpCbcRTTaeT49Pp_rojINPHh3MiXnzFcAkPVKA/edit?usp=sharing)

Digital Fujikawa

Digital Fujikawa

Kyoto University Library Network and Keio University Media Center concluded an agreement on the Joint Project on Digital Unification of Fujikawa Collection (Digital Fujikawa) and launched an integrated website that virtually brings together the holdings of both universities in one place by using the International Image Interoperability Framework (IIIF) protocol on September 28, 2018.

- Points:
  - Collect IIIF manifests with specific theme


Buddhist resources via IIIF in the world

- (3691) Bibliothèque nationale de France
- (1813) University of Tokyo
- (622) National Diet Library, JAPAN – it’s on the way to include
- (248) Kyoto University Rare Materials Digital Archive
- (147) National Institute for Japanese Literature
- (77) harvard.edu – it’s on the way to include
- (71) The Internet Archive – it’s on the way to include
- (67) Bayerische Staatsbibliothek
- (22) Kyushu University Library Collections
- (15) World Digital Library
- (13) Shimane University Library Digital Archive Collection
- (7) ubc.ca – U of British Columbia
- (7) Cambridge University Library
- (1) Images Copyright Biblioteca Apostolica Vaticana
- (1) e-codices - Virtual Manuscript Library of Switzerland
- (1) Vietnamese Nôm Preservation Foundation

We will find more!

The figures mean the numbers of items, not images
IIIF Manifest for Buddhist Studies

- Collaboratively aggregating IIIF Manifest files to share and add detailed information
- Providing the Manifest URIs with the added data
- We, Buddhist scholars can see the images of witnesses by a simple operation.

- It includes Dunhuang manuscripts and other Buddhist resources which might be connected each other.
Collaborative function to add information (ex. Line numbers of start and end in Taisho)

The added information can be displayed like:
A Recent Result of IIIF-BS

- The digital collection of rare books in Kyoto University Library includes Buddhist texts taking in fragments.
- Some of them didn’t identified due to lack of subject librarian.
- A collaborative work on IIIF-BS identified the materials and embedded the line number of Taisho tripitaka
- Kyoto-U Lib adopted the result of the collaboration in their Web site.
- It’s a collaborative achievement between a cultural institution and a research project.

IIIF JP-R: IIIF Discovery for Japanese Cultural Resources

http://iiif3.dl.itc.u-tokyo.ac.jp
IIIF JP-R: IIIF Discovery for Japanese Cultural Resources

- Prototype system for collecting IIIF japanese resources
- Developed by:
  - Satoru Nakamura (UTokyo)
  - Kiyonori Nagasaki (UTokyo, Intl. Institute for Digital Humanities)
- Points:
  - We can collaboratively collect IIIF manifests through the registration form.
  - You can download, leverage and re-distribute the list of the aggregated IIIF manifests.

Japanese IIIF Manifests for Fun

http://bauddha.dhii.jp/iiifws/show.php
Use cases of IIIF: Image Cropping & Annotations

Database of Pre-Modern Japanese Works / 新日本古典籍総合データベース

- Open license for its own materials (except deposited books).
- Aim to spread open license
- Adopting IIIF entirely.

- [http://kotenseki.nijl.ac.jp/?ln=en](http://kotenseki.nijl.ac.jp/?ln=en)
- [http://kotenseki.nijl.ac.jp/tagsearch/?q=動物](http://kotenseki.nijl.ac.jp/tagsearch/?q=動物)

- Due to its open license, the images have been leveraged in various ways in other Web site.
Database of Pre-Modern Japanese Works

- Tag search function with IIIF
  - http://kotenseki.nii.ac.jp/tagsearch/?q=%E7%8A%A2 (犬)

(Not yet published officially)
Digital Exhibit: Kunshujo

https://kunshujo.dl.itc.u-tokyo.ac.jp/
(Not yet published officially)
Digital Exhibit: Kunshujo

Tanaka Yoshio (田中芳男) (September 27, 1838 – June 22, 1916) was a Japanese civil servant and naturalist.

Kunshujo (捃拾帖) is scrap books of materials he collected from adolescence to later years.

The digital collection “Tanaka Yoshio, Natural History Collection” was launched by University of Tokyo Library System on August 6, 2018.

This digital exhibition site has been developed with the cooperation with librarians.

https://docs.google.com/document/d/1o5iAk3Ycf8SXnrHjJy5ubcsiPYSoP0lizDFxYSytPdM/edit#heading=h.gg7dkek3cl1
Collection of Facial Expressions

- Project that cuts out and collects the part of the face appearing in art works, and utilizes it for research on art history; works published by National Institute of Japanese Literature, Kyoto University Rare Materials Digital Archive, and Keio University Media Center Digital Collections.

- Developed by: Center for Open Data in the Humanities, Research Organization of Information and Systems
  - CODH is working on research and development for enhancing access to humanities data using the state-of-the-art technology in informatics and statistics, and at the same time constructing data platforms based on the idea of open science to promote trans-disciplinary participation of people with diverse backgrounds, thereby opening up new possibilities of research framework in digital humanities from data-driven perspectives.
Use cases of IIIF: Connection with transcriptions

Digital Exhibit: Tale of Lady Karaito (Karaito Soshi)

[Image of the digital exhibit]

http://digital.culturalresources.jp/omeka-yang/exhibits/show/karaito/karaito_intro
Digital Exhibit: Tale of Lady Karaito (Karaito Soshi)

Integration of TEI and IIIF = TEIIIF

Displaying annotation by mouse over

Showing a corresponding part of a IIIF image by clicking an icon
Linking from a line of a TEI encoded transcription to a corresponding zone of a IIIF image

Interests from TEI Consortium

- TEI Consortium: http://www.tei-c.org/
  - Has addressed how to makrup/encode digital resources for humanities since 1987.
  - Mainly focused on how to efficiently share texts
- TEI Consortium recently focuses Asian texts more:
  - Established two special interest groups:
    - East Asian / Japanese texts
    - Indic texts
- IDP community may collaborate with the SIGs.
Transcription @ UTokyo Library:
Line segmentation and manual annotations

Transcription @ UTokyo Library:
IIIF Search API

https://iiif.dl.itc.u-tokyo.ac.jp/repo/s/law/document/d507a810-cf7-4168-bc10-70a32a55920f
Welcome

The Ten Thousand Rooms Project (10万間居室) is a collaborative workspace for pre-modern textual studies being developed at Yale University with the support of the Andrew W. Mellon Foundation. Building on the Folger Project developed by Stanford University, the platform allows users to upload images of manuscripts, prints, inscriptions, and other sources and then organize projects around their transcription, translation, and/or annotation, both as a workspace for crowdsourcing textual research and as a publishing venue for scholarly contributions that are less well suited to conventional book formats, the Ten Thousand Rooms Project aims to establish an international online community committed to making the East Asian textual heritage more accessible to a wider audience.

All users are free to view projects on the site, and registered users are well come to create their own projects and/or contribute to others' projects (provided they do not violate the Terms of Use). If you initiate a project, you retain total control over your own project, including the ability to choose and vet project contributors. All original content uploaded to the Ten Thousand Rooms platform is licensed under the Creative Commons Attribution.

Minna De Honkoku: Learning-Driven Crowdsourced Transcription Of Pre-Modern Japanese Earthquake Records

This system is going to adopt IIIF near future.

https://honkoku.org/
An Alpha version of Scripto for Omeka S

Scripto is a community transcription tool.

The new Scripto will be available as a module for your Omeka S installation, with its own landing page and the ability to support multiple Scripto projects per installation.

Exercise 1. Gathering 3 IIIF manifest URIs

- Register the three into IIIF JP-R
- http://iiif3.dl.itc.u-tokyo.ac.jp
- Try to see the IIIF contents in the prepared IIIF viewers (UV, Mirador, and CV)

Exercise 2. Try to “curate” 衆鱗手鑑

- Curate two parts from pages of 衆鱗手鑑 on the IIIF JP-R
- Export the curation list (and see the curated images)

Exercise 3. Curate your favorite parts from the manifests URIs which you registered
Detection of topics from newspaper and its analysis of temporal variations in regions

Taizo YAMADA
Historiographical Institute
The University of Tokyo
Tokyo, Japan
t_yamada@hi.u-tokyo.ac.jp

Abstract—In the paper, we introduce a method of topic detection using topic model for Japanese newspaper and propose to visualize the time change of the detected topics. In the study, we detected topics from newspapers published by Mainichi Newspapers from 2010 to 2015. There are about six hundred articles (number of characters: about 300 million) in the text data. We performed to extracts nouns as characteristic words of the text. We characterized the text with a latent topic which is hidden in the text and can be detected by LDA (Latent Dirichlet Allocation) which is one of a topic model. There are very diverse topics including politics, sports, lotteries, Southeast Asian affairs, Japanese economics, academics, and so on. From them, we noticed earthquake topics and focused on them. In order to grasp the characteristics of the topic, we visualized the change of the frequency of the occurrence and the top words on a monthly basis. In order to calculate similarity between topics, we used cosine similarity in which the frequency of word occurrence per topic was used. Analyzing topics by region helps you to grasp the situation fluctuation in the region. If we investigate the posting position of the article and the topic variation, we can find out the importance of the topic or the article at that time.

Index Terms—newspaper, topic model, earthquake, temporal information, Japan

I. INTRODUCTION

The newspaper notes not only the trends of news in a wide range of genres such as politics, economics, incidents, international affairs, culture, sports, but also reports on matters related to specific regions as well as the country or the world. It is now universal that various kinds of news and events can be reported on the web by SNS (twitter, facebook,...), blog and various websites, but still newspaper is highly reliable as the media to convey the news.

In the paper we introduce a method to automatically detect topics from the articles in the newspaper which reports local information while holding global information, and to analyze the time series changes of the topics. In the study we used LDA (Latent Dirichlet Allocation) which is one of topic models for topic detection. We also describe not only the change in the whole newspaper but also the change when narrowing down to the area.

II. NEWSPAPER DATA

In the paper, we used articles of The Mainichi Newspapers (written in Japanese) from 2010 to 2015. The newspaper data is a full-text article data collection (text data with tags) for the final edition of the morning evening paper of Tokyo and Osaka headquarters. The newspaper data have been managed in units of articles. The data consists of an ID number, an article header keyword (notation and reading), a main text keyword (notation and reading) a publication surface type code, presence or absence of photographs, publication date, a page, an index article number, an article headline, a paper distinction of morning or evening and an article text. The number of articles for the 6 years is as follows: 92,547 articles for 2010, 96,563 articles for 2011, 110,587 articles for 2012, 106,305 articles for 2013, 102,448 articles for 2014 and 98,474 articles for 2015.

III. TERM EXTRACTING AND TOPIC MODEL

A. Term extraction

In order to characterize newspaper data, we performed morphological analysis on the article text, and created Bag-of-Words for the article data based on the result. The newspaper data includes body keywords and article headline keywords as the items. However it is difficult to grasp the appearance frequency of their keywords, because in some case the article data has keywords which the article text dose not have. Therefore, we extract terms which can characteristic the article text from the text.

In the paper we used mecab2 which is one of well-known Japanese morphological analyzer, and IPADic3 which is one of a dictionary for Japanese morphological analyzer. In the process we used nouns as the term extraction target, but pronouns, suffixes, adverbs, adjective stems, conjunctional, non-autonomous were excluded. And we chunked consecutive nouns, suffixes immediately after the nouns to be extracted, and [a-zA-Z]+ sequences. In order to express extracted terms and their frequency of occurrence, we output the result as Bag-of-Words.

B. Topic Detection

In topic detection we used LDA (Latent Dirichlet Allocation) [1] which is one of the topic models. LDA treats a set of

2http://taku910.github.io/mecab/
3https://github.com/neologd/mecab-ipadic-neologd
terms which are subject to statistical co-occurrence as latent topics. In the topic model using LDA, it is assumed that there are multiple topics in one document. The LDA models the distribution of the topics.

Figure 1 shows the graphical model representation of the LDA used in the paper. Here, the blue circle indicates the observation variable, the white circle indicates the unknown variable, the rectangle indicates repeating, and the lower right numerical character indicates the number of the repeating times represented by the rectangle. \(w\) indicates the result of the term extraction mentioned above and is the only observed variable in here. \(z\) is the topics, \(\theta\) is the topic distributions for articles, and \(\phi\) is the term distributions. \(\alpha\) and \(\beta\) are \(\theta\) and \(\phi\) parameters and indicate hyperparameters of LDA. When the number of documents is \(D\) and the number of terms in document \(d\) is \(N_d\), \(\theta_d\) and \(\phi_k\) are generated by

\[
\theta_d \sim \text{Dir}(\alpha) \quad (d = 1, \ldots, D),
\]

\[
\phi_k \sim \text{Dir}(\beta) \quad (k = 1, \ldots, K).
\tag{1}
\]

Here, \(\text{Dir}(\cdot)\) represents the Dirichlet distribution. The topic \(z_{d,i}\) is generated by

\[
z_{d,i} \sim \text{Multi}(\theta_d) \quad (i = 1, \ldots, N_d).
\tag{2}
\]

Here, \(\text{Multi}(\cdot)\) represents a multinominal distribution. In addition, the term \(w_{d,i}\) is generated by

\[
w_{d,i} \sim \text{Multi}(\phi_{z_{d,i}}) \quad (i = 1, \ldots, N_d).
\tag{3}
\]

There are variational Bayesian method (VB) [1] and gibbs sampling (GS) [2] as mainly well-known inference methods of LDA model. Collapsed gibbs sampling (CGS) [2] which is improved version of GS. Because it is not necessary to calculate \(\theta_d\) and \(\phi_k\) if the method used, the calculation cost can be reduced greatly. Moreover, according to [3], the prediction performance of CGS (by perplexity) is better than VB. There is another inference method called collapsed variational Bayesian (CVB) method [4]. CVB is an improved version of VB which can performe marginalization of \(\theta_d\) and \(\phi_k\) like CGS in the inference. However, CVB algorithm is complex and requires a large amount of memory. In the paper we detected topics from newspaper data using CGS.

1. Initialize \(\alpha\) and \(\beta\)
2. Initialize \(z\)
3. Set \(S\) : the number of sampling
4. for \(s = 1, \ldots, S\)
5. for \(d = 1, \ldots, D\)
6. for \(i = 1, \ldots, N_d\)
7. Sample \(z_{d,i}\)
8. Update \(N_{d,z_{d,i}}\)
9. end for
10. end for
11. Update \(\alpha\) and \(\beta\)
12. end for

Fig. 1. Graphical model for LDA

Fig. 2. Procedure of collapsed Gibbs sampling

C. Sampling

The procedure of CGS used in this paper is shown in Figure 2. Here, \(N_{d,k}\) indicates the number of terms assigned topic \(k\) in Document \(d\). \(z_{d,i}\) can be sampled by

\[
z_{d,i} \sim \text{Multi}(p(z_{d,i} | W, \mathbf{Z}_{\setminus d,i})) \propto (N_{d,i} + \alpha) N_{k,w_{d,i}} + \beta N_k + \beta V
\tag{4}
\]

Here, \(W\) indicates all documents (whole article data). \(\mathbf{Z}_{\setminus d,i}\) indicates a set of topic except for \(z_{d,i}\). \(V\) indicates the number of kinds of terms.

\(\alpha\) and \(\beta\) are hyperparameters in LDA, and are parameters in Dirichlet distribution. In general, when each value of \(\alpha\) is not uniform (\(\alpha_k \neq \alpha_l, k \neq l\) and the value of \(\beta\) is uniform (\(\beta_1 = \beta_2 = \ldots = \beta V\)), the performance of the LDA improves [5]. We decided to use the hyperparameter setting. The hyperparameters can be inferred by maximizing marginal likelihood. When fixed point iteration can be used
for the iteration,
\[
\alpha^\text{new}_k = \alpha_k \frac{\sum_d D(N_{d,k} + \alpha_k) - D\Psi(\alpha_k)}{\sum_d \Psi(N_d + \sum_{k'} \alpha_{k'}) - \sum_{k'} \Psi(\alpha_{k'})}, \quad (6)
\]
\[
\beta^\text{new} = \beta \frac{\sum_k \sum_v \Psi(N_{k,v} + \beta) - KV \Psi(\beta)}{\sum_k \Psi(N_k + \beta V) - KV \Psi(\beta V)} \quad (7)
\]
Here \(\Psi(\cdot)\) indicates digamma function.

IV. EXPERIMENT

A. Experimental setup

We obtained terms from newspaper data by term extraction described in Section III-A. The number of terms was 286,288,248 and the kind of terms was 2,683,289. We assumed the number of topic is 200. In the topic detection we used CGS with LDA described in Section III-B and collapsed gibbs sampling described in Section III-C.

Determining the sampling frequency in CGS is a difficult task. Therefore, in order to evaluate the performance of LDA, we conducted preliminary experiments on perplexity. The results shows in Figure 4. From the result, we can obtain that the perplexity can stabilize with about 100 sampling, however, there was some variation until the 800 sampling. Therefore, we decided to use average from 900 to 1000 sampling for topic detection.

B. Results

Figure 3 shows monthly changes in frequency of appearance of terms assigned to a topic and major events and news in Japan. By the result we can expect that the occurrence of big events and impact news (such as Olympics, Football
Fig. 5. Top ten topics with variance values

Fig. 6. Topics similar to earthquake topic (V3)

World Cup, East Japan great earthquake, elections of House of Representatives and Councilors of Japan) and the frequency of topics related to that event are related. For example, such terms as 東日本大震災 (en: the Great East Japan Great Earthquake), 津波 (en: Tsunami), 被災地 (en: Afflicted Areas), 震災 (en: earthquake disasters), 被災者 (en: Victims) were assigned to Topic 3, and the frequency of the appearance was very high frequency in March and April 2011. Likewise, topic 120 indicates the national election voting result and topic 75 has terms related to the Olympic. The occurrences of the events coincides with the timing indicating the high appearance frequencies.

When sorting by frequency of occurrence of the assigned terms, the top five topics were as follows:

1) Topic 182: 人 (en: person), 自分 (en: myself), 仕事 (en: work), 声 (en: voice), 家族 (en: family), 母 (en: mother), ...

2) Topic 11: 日本 (en: Japan), 人 (en: person), 世界 (en: world), 言葉 (en: words), 時代 (en: era), 戦争 (en: war), 人々 (en: people), ...

3) Topic 166: 問題 (en: problem), 必要 (en: necessity), 調査 (en: investigation), 指摘 (en: indication), 国 (en: country), 説明 (en: description), 対応 (en: response), 検討 (en: consideration), ...

4) Topic 184: 首相 (en: prime minister), 民主党 (en: democratic party), 自民党 (en: Liberal Democratic Party), 党 (en: party), 選挙 (en: election), 国民 (Japanese citizen), 批判 (en: criticism), 国会 (en: Diet), 政府 (en: government), ...

5) Topic 176: ロシア (en: Russia), 米国 (en: USA), イラン (en: Iran), シリア (en: Syria), 大統領 (en: president), イスラエル (en: Israel), 可能性 (en: possibility), 死亡 (en: death), ...

Topics 182 and Topic 11 are considered to be related to columns in the Mainichi Newspaper which appears in daily newspapers. Although the frequencies were high, the variances were not high.

On the other hand, Figure 5 shows top 10 topics when sorted according to variance values. Top 10 topics in the result follows:

1) Topic 120: result of national election voting
2) Topic 3: earthquake
3) Topic 29: professional baseball in Japan
4) Topic 75: Olympic
5) Topic 184: political trend

Topic 3 was ranked 11th in terms of frequency of occurrences and 2nd in terms of variance. Topic 184 was ranked fourth in order of frequency of occurrences and fifth in terms of variance. In topics automatically detected by topic model, the topics with high occurrence frequency and variance value are highly likely to be events and news symbolizing Japan during this time.

Figure 6 shows the results of topic 3 which is related to earthquakes and some topics similar to the topic and similar topics. Here, the appearance frequencies of terms assigned to each topic (actually weighting by tf-idf) were used as the values in the feature vector of the topics, and the similarity between vector $u$ and vector $v$ was calculated using the following equation.

$$sim(u, v) = \cos(u, v) = \frac{\sum_i u_i v_i}{\sqrt{\sum_i u_i^2} \cdot \sqrt{\sum_i v_i^2}}$$ (8)

Topic 3 is the most similar to Topic 117 which relates to the announcement of the Japan Meteorological Agency and has terms such as 地震 (en: earthquake), 気象庁 (en: meteorological agency), 発生 (en: occurrence), 影響 (en: influence), 雨 (en: rain), 観測 (en: observation), 火山 (en: eruption), 発表 (en: announcement). The similar topic is Topic 64, which has 東電 (en: TEPCO; Tokyo Electric Power Co.,Inc.), 福島 (en: Fukushima), 放射性物質 (en: radioactive materials), 東京電力福島 (en: TEPCO Fukushima), 原発 (en: nuclear power plants) and 原発事故 (en: nuclear accident), and is related
to the Fukushima Daiichi nuclear disaster. The results are consistent with intuitive results. The time series data of these topics showed waveforms which may be correlated with topic 3. On the other hand, other topics (182, 173) cannot be said to correlate with time series change. Also these topics could be not related topics intuitively. We think that it is necessary to reconsider the similarity of topics based on the correlation of time series change. For example, topic 7 which was assigned to 核電 (en: nuclear power plants), 原発 (en: operation), 政府 (en: government), 東電 (en: TEPCO), 関電 (en: KEPCO; the Kansai Electric Power Co., Inc.), 電力 (en: Electric Power), 福島 (en: Fukushima), 電力会社 (en: Electric Power Companies), 必要 (en: necessary) is close to topic 64 and topic 3 for the waveform of the graph indicated by its appearance frequency. However, the similarity by the expression (8) is not higher than the topic 182 and others. If there is a way to describe the relationship between topic 7 and topic 3 well, we would be able to derive a more topic.

Figure 7 shows the time series changes of topics which cover only the articles data in which the specified keywords appear. In the result, the place names were designated as a keyword ((a)“Ofunato”, (b)“Sendai”, (c)“Kobe”, (d)“Kumamoto”). We expect that we can grasp the time series changes of the topics where the place names specified by the method appears. In (a) and (b) the occurrence frequencies of topic 3 was high immediately after the Great East Japan Earthquake occurred. Since topic 189 is assigned to 問い合わせ (en: inquiries), http, www, 毎日新聞社 (en: Mainichi Newspaper company), 開催 (en: holding), 無料 (en: free), 申し込み (en: request), ホームページ (en: homepage, website), はがき (en: postcard), 参加 (en: participation), 応募 (en: application), it seems to be related to a volunteer recruitment. The results of (c) and (d) showed a situation that is significantly different from (a) and (b). In the results of (c) topic 144 (related to “baseball”) and topic 50 (related to Japan national football team) were frequent mainly. Topic 3 (related to “earthquake”)

Fig. 7. Time-series changes of topics concerning place names
was also frequent, however unlike (a) and (b), it showed high frequency in January every year. The frequency was even more frequent in January 2015. Here, topic 3 is related to the Great Hanshin Awaji earthquake but not the Great East Japan Earthquake. Since the target period of newspaper article data is from 2010 to 2015, the Kumamoto earthquake in April 2016 is irrelevant.

Mainly topic 167 (related to “万能川柳” (Banno-senryu) which is a column of “川柳 (senryu)” in the newspaper), topic 76 (related to Japan tournament of high school baseball), topic 120 and topic 144 were frequent before the occurrence of the Kumamoto earthquake as topics related to Kumamoto. The reason for the results seems to be the use the newspaper articles of the Tokyo version and the Osaka version. If the local versions of the articles may be used, the results should be highly likely to change.

V. CONCLUSION

In the paper we introduced the method of topic detection using LDA for newspaper articles and described the visualizations to analyze the time series changes of topics from 2010 to 2015.

We think that the topic detections and their time series changes make easy to grasp the changes in topics and concerns. The method is also possible to grasp those changes only in specific articles with filtering by keywords. If the place names can be extracted from the articles of the newspapers and can be given latitude and longitude to, we can understand the changes of news or events easily. Moreover it lets us to compare between regionals and global. We believe it will be possible to treat newspaper data and the analysis results as a materials for sophisticated area studies if the above mention becomes possible.

LDA is one of the methods of unsupervised learning like k-means. As shown in the experiments, since LDA is possible to classify data without a classification indicator, it can be used to detect and estimate the essential structure which is latent in data. On the other hand, there are supervised learning such as neural networks and support vector machine (SVM) as machine learning methods. The supervised learning is a method of classifying input data in a state where some classification indicators exist. We think that new data can be classified by supervised learning with data trained by results of unsupervised learning such as LDA. We think that its manners can be applied to newspaper articles as well.

In our study we used only data of the Mainichi newspaper. In order to obtain even better results, we should consider to incorporate other resources such as other newspaper and SNS. Currently the sending by Individuals such as SNS affects various situations of society. Therefore, we plan to practice topic analysis including SNS not only newspapers.

ACKNOWLEDGMENT

A part of this work was supported by JSPS KAKENHI Grant Number 26730167, 26240049, 15H01722, 16H01897 and project of Japan-ASEAN Platform for Transdisciplinary Studies (by Center for Southeast Asian Studies, Kyoto University). In the paper we used data set of CD-Mainichi Newspapers.

REFERENCES


---

4 "川柳 (senryu)” is a Japanese form of short poetry similar to haiku in construction.
A Common Base of Knowledge for Japanese Historical Materials and its Application

Taizo YAMADA
Historiographical Institute
The University of Tokyo
Tokyo, Japan
t_yamada@hi.u-tokyo.ac.jp

Satoshi INOUE
Historiographical Institute
The University of Tokyo
Tokyo, Japan
inoue@hi.u-tokyo.ac.jp

Abstract—In the paper we describe the fundamental elements for handling and managing Japanese historical information such as catalogue, image, text, spatial, temporal, personal name, and introduce an application using the fundamental elements. Our Institution (The Historiographical Institute, the University of Tokyo) has been constructing and providing the information related to Japanese history and Japanese historical materials for more than 130 years. There is a diversity of data items for the materials, because data items added by some factors such as collections, eras and types are different. In order to make it easy to use the collected historical materials, we set fundamental items of them. We explain the features of each element. We will also describe the applications using that elements and their effects.

Index Terms—Japanese historical material, common base, historical knowledge, diversity

I. INTRODUCTION

The foundation of historical study consists of collecting, precise reading comprehension, and source criticism for historical materials. Historiographers can draw out a historical issue point and construct a historical model based on the historical information. In Japan, a huge amount of historical materials are held in private houses, temples, and shrines, not only public institutions such as archives, libraries, and museums. All of historical materials held in especially private houses, temples and shrines are not cataloged, although some catalogs may exist. For this reason, even new historical materials (even historical materials related to a very important person for Japanese history such as “豊臣秀吉 (TOYOTOMI, Hideyoshi)”) are still discovered. ¹ In Japanese history study, collection of historical materials is still one of the important research bases.

Historiographical Institute, the University of Tokyo (HI for short) is the only academic institution of university specializing in Japanese historical study, and has been providing data related to Japanese history for more than 130 years. HI has following three research missions which are shown in Fig. 1: investigating and collecting historical materials, compiling and publishing primary source materials dealing

—with Japanese history. These results have been publishing not only in books but also through the database system called SHIPS DB (DataBase of Shiryohensan-jo (史料編纂所) Historical Information Processing System).² SHIPS DB consists of 30 databases on the location and administration of historical materials, events, texts, personal names, images, facts. In 2017 the number of accesses to the databases was 6,884,004 and the number of accesses to the image was 8,019,930, SHIPS DB is now becoming one of the essential research infrastructures in Japanese history research.

In this paper we describe the fundamental elements of Japanese historical information which has been constructed and provided by HI. In addition, we describe an application using the fundamental elements.

II. FUNDAMENTAL ELEMENTS OF HISTORICAL MATERIALS

We describe the method of our investigating and collecting historical materials. Japanese historical materials have been dispersed throughout Japan or around the world, so it is necessary to collect historical materials. Because it is a basic principle of HI that source materials should be preserved where they are found, we spend several weeks each year investigating and examining historical materials wherever they are kept. We have over 130 years’ experience conducting these examinations, and our journeys have taken them all over Japan as well as to many different parts of the globe. The materials are reproduced by several different means, such as microfilm, calligraphic reproduction, and photography. Over the decades, we has accumulated an unparalleled collection of duplicate documents as well as tens of thousands of originals. We carefully catalog and index all source materials and list their findings in a variety of reference works.

Main types of historical materials that we have collected and accumulated are old documents, old diaries, classic books, picture maps and so on. There is a diversity of data item for the materials, because data items added by some factors such as collections, eras and types are different. In order to make it easy to use the collected historical materials, we set fundamental items of them as follows: catalogue, image, text,
spatial, temporal, personal name. Fig. 2 shown an example of the values of the items in old document "Dai Nihon Komonjo". This section describes the handling of these elements in SHIPS DB.

A. Catalogue

The catalogue is most important element to find the materials and the management is one of eternal issues. SHIPS DB has a hierarchical catalogue database system. The hierarchical catalogue has been described according to “Model of the levels of arrangement of a Fonds” in ISAD (G) 2nd [1]. Historical materials are given identifiers according to each hierarchical level (from Fond level to Item level) of the materials. The identifier is systematically and permanently managed, because it can be managed by the librarian of HI. Since the identifiers can be referred from other databases, it is possible to obtain the catalogue data. The database for the catalogue can manage not only materials held by HI but also reproductions made by us in which others hold the original.

B. Image

HI mainly has 200,355 items which indicates primary resources, 117,875 items which indicates reproductions including the calligraphic reproductions created by HI, and 66,990 items which indicates microfilms. From 2000 until 2008, we had digitized the materials of which the number is about 7200 volumes. Due to the digitalization, we took the materials with microfilm and we made an image file by scanning the film. The specification of the image is as follows:
- format: TIFF,
- condition: monochrome and binarization,
- resolution: 400 ppi.
From 2008, we have been shooting with digital camera. The specification is as follows:
- format: jpeg,
- size: about 31.3 mega pixel,
- resolution: 400 ppi.

For historical materials from other institutions, we had microfilm photography from 1970 to 2009. About 12,000 reels and 8,000 sheet files were produced by the shooting. The digitization of all them was carried out between 2008 and 2015 (8 years). The digitized image was given an identifier according to the catalogue and stored in our storage according to the hierarchy of the catalogue.

C. Text

As shown in Fig. 2, the text is created by reprinting the material. In addition, as shown in Fig. 1, in SHIPS DB and the publications of HI, person names, place names, event names and annotations on text formats such as corrections and deletions are added to the text and mark up.
The structure of the text data in SHIPS DB is an XML-like independent description rule or plaintext. Although the structure is human-readable, it is not a machine-readable method. It is unsuitable for performing rich search and text analysis. Therefore, it is necessary to introduce text structuring according to text guidelines like TEI (Text Encoding Initiative) P5 [2].

D. Temporal data

The temporal descriptions in the metadata and/or the text in the historical material, are very diverse. For example, on Japanese calendar "文禄三年八月一日" which indicates A.D. 1594-09-15 is same as "文禄三年八月一日". In some cases, the date of the historical material may be ambiguous or unknown. For example, because the historical materials are often not described in the year especially from the late 15th to the end of the 16th century, it is often described as "文禄三年八月一日" indicating estimated date. In order to resolve the ambiguity and the diversity of the description related to the data, we have created a tool which can convert from Japanese calendar to A.D. code. The A.D. code can be represented by eleven digits which five digits indicate year code, next three digits indicate month code, later two digits indicate day code and last digit indicates option code. For example, the code of "文禄三年八月一日" is "1594008010". The starting year is set for a regnal year. For example the starting year of "文禄" is 1592, therefore, the year code of "文禄三年" is 1594. The last digit of the year code is provided for the option and the default value is 0. "一日" means first day of the month and the code is 01. The last digit is optional and the default value is 0. If the date is the estimated such as "文禄三年八月一日", the value is 1. In one year we have constructed the rules concerning the date and decided the coding rules. We modularized the conversion tool and embed it in to SHIPS DB. Thus, the users can obtain historical data regardless of the ambiguity and the variety of the date description.

E. Spatial data

To find out where the historical events occurred is an important factor in understanding history. For example, a picture map has a place name when the map was created, a method of using the land, an owner, and a relative positional relations between the place names described. As shown in Fig. 2, a place name an old document in the map are one of the factors indispensable in history understanding. Moreover, in historical study, the location of historical materials is one of important elements. The reason is not only reading and accessing of the historical materials but also that the transmission and history of the historical material are indispensable data in order to grasp the characters of the historical materials. Since we sometimes handle place names over 1000 years ago, it is difficult to easily associate the place name with the current map. Therefore, we store the location (point with latitude and longitude) estimated together with the place name written in the material in the database.
F. Personal name

A historical material often has information related to a person. For example, an old document including a letter has sender(s) and receiver(s), and sometimes there is a personal name related to the document as shown in Fig 2 (in the example “平信兼” has appeared). In order to understand the old document, the information of these persons is indispensable.

We have been extracting from collected historical materials and databases holding personal names (which are managed by HI, stored into SHIPS DB). We are storing these personal names into “personal name repository” which can manage personal name with the material or data in the database, because the origin or the source of the names can be clarified. The repository has authoring tool which personal name data can be created and stored in. A user can freely create an item and can enter the value according to the item with the repository. The data can be changed to RDF (Resource Description Framework) format and represented. Currently, the repository has 419,680 name data.

A personal name has appeared in a variety of the historical materials such as an old diary, an old document, a classical book, a family pedigree, a document related to appointment, and so on. There is a diversity in the name representations in the materials such as a real name or an original name, a nickname, an epithet, a role name which indicates a person, a Kao (which is a stylized signature or a mark), and other related things. In order to handle the diversity, we drew up “basic common items” based on SHIPS DB. The items are as follows:

- identifier: in the repository
- name: personal name, label (controlled)
- temporal: date which the name appears in the source
The data structure can also hold original items.

[3] and [4] introduced a method of personal name against Chinese historical materials (like a “地方志 (difangzhi)”), based on China Biographical Database (CBCB) as a biographical dictionary. In Japanese History, there are no exhaustive an encyclopedia or a dictionary for a name of a historical person.

G. Other Items

We also have been servicing other data: characters written in historical materials and historical events in pre-modern Japan. The character database provides characters on the image of the historical material and their character codes. Because it is related to data on the catalogue, the database also has catalogue data such as the identifier, name, temporal, and so on. Also, if the author of the material can be known, the database has its author name and attributes (such as gender,
The database, it is possible to search and sort using these attributes.

Series of “大日本史科 (Dai Nihon Shiryo)” which is the series of the chronological source books of Japanese history is one of the most important publications of HI. We compile chronologically organized material collections which incorporate materials from many different sources in the order in which events occurred. The series is an ongoing and comprehensive attempt to cover all major events mentioned in the historical materials between the years 887 and 1867. We have been publishing the series since 1901 and servicing the database related to the series.

III. APPLICATION USING TEMPORAL DATA

In the section we explain the applications using the fundamental elements, especially temporal data.

A. Event Application

We have been managing pre-modern Japanese historical events in the compilation and publication of the series of “大日本史科”. Using the results, we have been constructing and servicing the historical event database called “大日本史科 (The Dai Nihon Shiryo) Unified Database”.

As mentioned in the above section, the data has historical events and temporal data related to the events. Fig. 3 shows the search results of the database, and the search results includes the events on and later than “延徳二年十二月一日” which is represented by “14900120010” as A.D. code. In the search result, the historical events can be arranged in chronological order, because A.D.Code allows temporal data of the events to be serialized.

B. Personal Name Application

We have been constructing database which can store a list of a personal name appeared in a historical material, a well-known the name of the person and the material catalogue as mentioned in Sec.II-F. In the search results of the database, the personal names stored in the personal name repository can be integrated with the based on the well-known name. Because the personal name repository has an RDF store \(^3\) and a SPARQL \(^5\) endpoint, the integration can be performed with the SPARQL in which the rule can be defined.

The database supports a keyword search for the data in the personal name repository, sorting in chronological order for search results. For example, if user searches with “足利 (Ashikaga)”, about 9,000 authority records related to 66 persons would be found. Fig. 4(a) shows the appearances of the personal names every years. By A.D. code and the conversion module, the search results could be arranged in chronological order. If choosing “足利義満 (ASHIKAGA, Yoshimitsu, 1358 - 1408)” which the amount of the results was fourth, it is possible to display his related items (such as a alias name, signature, role name) appearing in the historical materials in chronological order as shown in Fig. 4(b).


足利義満 (ASHIKAGA, Yoshimitsu, 1358 - 1408) was the third shogun of Muromachi shogunate. He became the top of samurai and court noble in both name and reality, and made a success of the unification of Northern and Southern Dynasties in Japan. There are 20 kinds of descriptions related to him and 392 authority records. He was dead in 1408. From the result, it turns out that descriptions related to his name changed after his death.

IV. Conclusion

In the paper we describe the data related to pre-modern Japanese historical materials our institution HI has following three research missions which are shown in Fig. 1: investigating and collecting historical materials, compiling and publishing primary source materials dealing with Japanese history. In the paper we explained the data related to pre-modern Japanese historical materials collected by our institution HI and the fundamental elements of the materials (such as catalogue, image, text, spatial, temporal, personal name) to cope with the diversity of the data for the management. Moreover, we introduced some databases handling with our data which are the applications utilizing the fundamental elements, especially temporal data. Although it is necessary to generalize the method such a [6], for the purpose, we need a mathematical model to solve the issues of the temporal data.

ACKNOWLEDGMENT

A part of this work was supported by JSPS KAKENHI Grant Number 26240049, 26220402, 15H01722 and 18H03576.

REFERENCES


Basic Linked Data Resource for Temporal Information

Tatsuki Sekino
Research Institute for Humanity and Nature
Kyoto, Japan
sekino@chikyu.ac.jp

Abstract—RDF Resources of calendrical periods such as date, month, and year were defined, and provided as Linked Open Data. These resources clarify the temporal ranges of various temporal expressions and enable their comparison on a common temporal axis, even if they are based on different calendars. These functions cannot be realized using only ISO 8601 (widely used as standard date expression). It is expected that these calendrical period resources will be applied to new and existing Linked Data, and bring new possibility into the Linked Data world.

Keywords—RDF; semantic Web; ISO 8601; calendar; date expression; linked open data

I. INTRODUCTION

Linked Data has become popular as a method to provide various kinds of data [1]. There are also some attempts to distribute Linked Data of historical resources [2][3][4]. To link these data, spatiotemporal information is important as a contact point. GeoNames [5], a provider of information about place names and locations, is widely used and normalizes location names in the Linked Data world. However, available basic data resources for temporal information are not enough, while there are useful resources for spatial information.

ISO 8601 [6] is a standard format to describe date and time (e.g. “2017-11-08”). It is also used widely for Linked Data, because use of this format for date expression is stipulated by the Resource Description Framework (RDF) [7][8]. Various other formats based on calendars are also used. This situation is the same for expressions of other calendrical periods such as month, year, and era. This issue is more serious in calendrical period expressions in regions outside of Europe, especially in historical documents, owing to the use of different characters for date expression (e.g. Arabic and Chinese numerals); year specification using era is equally complicated. There is no available service which enables the comparison on a common temporal axis of expressions based on different calendars.

Another problem is that the ISO 8601 format for date expression used in RDF is a literal node. A literal node only expresses value. It is difficult to associate dates expressed according to ISO 8601 with related information, because RDF prohibits the use of a literal node as the subject in a RDF triple. For example, 8th November 2017 is a Wednesday, but describing this fact in RDF requires complex triples.

ISO 8601 requires the expression of dates according to the proleptic Gregorian calendar (which extends the Gregorian calendar backwards to dates preceding its introduction). There is a nine day gap between 16th century dates and the proleptic Gregorian calendar dates for the same period; therefore the expression of the date “1st January 1500” as described in a historical document, must be expressed as “1500-01-10” if adhering to ISO 8601. This regulation makes date expression of ISO 8601 equivocal. When a date expressed as “1500-01-10” is given, a user cannot know whether this is strict ISO 8601 expression indicating 1st January 1500 or ISO 8601-like expression indicating 10th January 1500. Owing to this, the application of ISO 8601 may cause confusion in Linked Data of historical resources.

If a date is expressed as an RDF resource rather than a literal, the date resource can be associated with various kinds of information about the date using RDF. This information includes position on the common temporal axis, the previous and next date resources, hierarchically higher (e.g. month and year) resources, and the calendar on which the date is based. As days of the week are also associated with date resources, the above example of day of week can be expressed simply as one RDF triple. This mechanism can apply not only to date but also to other calendrical periods such as month and year.

The purpose of the present study is to define URIs of various kinds of calendrical periods as RDF resources, and to construct Linked Data providing related information about those calendrical periods. Given a known relative position on a common temporal axis among calendrical period expressions including ISO 8601, we aim to enable users to obtain related information about the given calendrical period.

II. DEFINITION OF RESOURCE

A. Common temporal axis

Calendrical period resources are associated with a common temporal axis to compare temporal ranges between them. The common temporal axis must meet the following requirements: 1) it covers a wide temporal range from the past to the future, 2) it is continuous in all ranges, and 3) its expression is simple and easy to process on computers.
Here, Julian Date is a calendar which can act as the temporal axis satisfying the above requirements. The Julian Date is defined as the total day count from January 1, 4713 BCE [9]. It is simply expressed as numerical values, and can be applied from the past to the future continuously. Therefore, the present study uses the Julian Date as the common temporal axis for calendrical periods.

B. URI of Calendrical Period Resources

URI of calendrical period resources is defined as follows.

http://datetime.hutime.org/calendar/

calendar ID / resource type / JD | period expression

The first segment in the path is “calendar”, and indicates that this URI is of the calendrical period resource. The calendar ID of the second segment in the path is an identifier to specify calendar (TABLE 1). For example, “2.1” is the proleptic Gregorian calendar used in ISO 8601, and “101.1” is the Julian / Gregorian calendar (reformed in 1582), generally used in the world. The number after the dot in a calendar identifier shows variation in the same calendar. Two kinds of Julian / Gregorian calendar which are different in the day of calendar reform (in 1582 and in 1752) are distinguished by the number, their identifiers being “101.1” and “101.2” respectively. The latest information about available calendar identifiers is listed on the web (http://datetime.hutime.org/calendar/).

<table>
<thead>
<tr>
<th>Calendar ID</th>
<th>Calendar Name</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Julian Date</td>
<td></td>
</tr>
<tr>
<td>2.1</td>
<td>proleptic Gregorian Calendar</td>
<td></td>
</tr>
<tr>
<td>101.1</td>
<td>Julian/Gregorian Calendar</td>
<td></td>
</tr>
<tr>
<td>2010-2019</td>
<td>proleptic Gregorian Calendar</td>
<td></td>
</tr>
<tr>
<td>101.2</td>
<td>Julian/Gregorian Calendar</td>
<td></td>
</tr>
<tr>
<td>2010-2019</td>
<td>Julian/Gregorian Calendar</td>
<td></td>
</tr>
<tr>
<td>102.1</td>
<td>Julian Calendar</td>
<td></td>
</tr>
<tr>
<td>103.1</td>
<td>Hijiri Calendar</td>
<td></td>
</tr>
<tr>
<td>104.1</td>
<td>Hebrew Calendar</td>
<td></td>
</tr>
<tr>
<td>1001.1</td>
<td>Japanese Calendar (Southern Court)</td>
<td></td>
</tr>
<tr>
<td>1001.2</td>
<td>Japanese Calendar (Northern Court)</td>
<td></td>
</tr>
<tr>
<td>1002.1</td>
<td>Japanese Lunisolar Calendar (after Meiji Era)</td>
<td></td>
</tr>
<tr>
<td>1003.1</td>
<td>Thai Buddhist Calendar</td>
<td></td>
</tr>
<tr>
<td>1004.1</td>
<td>Minguo Calendar</td>
<td></td>
</tr>
</tbody>
</table>

* Test version for experimental usage

The type of calendrical period is specified by resource type in the third segment. TABLE 2 lists available types of calendrical period resources. Resources that can use the “era”, “century”, and “millennium” types are limited.

Differences in period among the same resource types and calendars are distinguished by the last segment in two ways. When the Julian Date is placed in the last segment, it indicates the beginning point of the period. Another way is to place into the last segment a string expressing the beginning point of the period indicated by the resource. The following URIs are examples of date resources indicating a date of 9th November 2017 according to Gregorian / Julian calendar.

(1) http://datetime.hutime.org/calendar/101.1/date/2017-11-08
(2) http://datetime.hutime.org/calendar/101.1/date/2017_November_8
(3) http://datetime.hutime.org/calendar/101.1/date/2017_Nov_8
(4) http://datetime.hutime.org/calendar/101.1/date/2017-11-08

Example (1) shows the Julian Date method. Examples (2) to (4) show the period expression method. As there are various expressions indicating the same period, it is required that the system providing calendrical period resources interpret and normalize these expressions. This interpretation feature is essential in the East Asia, where more varied expressions are used to indicate calendrical period. The following examples are date resources from the Japanese calendar.

(1) http://datetime.hutime.org/calendar/1001.1/date/平成29年11月8日
(2) http://datetime.hutime.org/calendar/1001.1/date/平成二十九年十一月八日
(3) http://datetime.hutime.org/calendar/1001.1/date/平成丁酉年拾壱月己亥

Arabic numerals are used to express day, month, and year in example (1), while Chinese numerals are used in example (2). URI (3) is an example of date expression in which year (丁酉) and day (己亥) are expressed by sexagenary cycle (combinations of the ten celestial stems and the twelve earthly branches), which is often used in historical documents. The feature enabling interpretation of these expressions is useful for users who cannot understand the characters of certain languages, such as Chinese letters, since they can obtain the calendrical period resource and related information by pasting

TABLE 1. List of available calendars

<table>
<thead>
<tr>
<th>Calendar ID</th>
<th>Calendar Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Julian Date</td>
</tr>
<tr>
<td>2.1</td>
<td>proleptic Gregorian Calendar</td>
</tr>
<tr>
<td>101.1</td>
<td>Julian/Gregorian Calendar</td>
</tr>
<tr>
<td>101.2</td>
<td>Julian/Gregorian Calendar</td>
</tr>
<tr>
<td>102.1</td>
<td>Julian Calendar</td>
</tr>
<tr>
<td>103.1</td>
<td>Hijiri Calendar</td>
</tr>
<tr>
<td>104.1</td>
<td>Hebrew Calendar</td>
</tr>
<tr>
<td>1001.1</td>
<td>Japanese Calendar (Southern Court)</td>
</tr>
<tr>
<td>1001.2</td>
<td>Japanese Calendar (Northern Court)</td>
</tr>
<tr>
<td>1002.1</td>
<td>Japanese Lunisolar Calendar (after Meiji Era)</td>
</tr>
<tr>
<td>1003.1</td>
<td>Thai Buddhist Calendar</td>
</tr>
<tr>
<td>1004.1</td>
<td>Minguo Calendar</td>
</tr>
</tbody>
</table>

TABLE 2. Types of Calendrical Period Resources

<table>
<thead>
<tr>
<th>Type</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>era</td>
<td>Resource type of calendar era. This type is valid for resources of some kinds of calendars.</td>
</tr>
<tr>
<td>year month</td>
<td>Resource type of calendar year, month and date.</td>
</tr>
<tr>
<td>date</td>
<td>Resource type of calendar date.</td>
</tr>
<tr>
<td>10y</td>
<td>Specific period which has predetermined range such as 2010s. 10y type resource of 2010s indicates the period between 1st Jan 2010 and 31st Dec 2019.</td>
</tr>
<tr>
<td>100y</td>
<td>Specific century and millennium. These types are valid for resources of Gregorian and Julian calendar (Calendar ID: 2.1, 101.1, 101.2 and 102.1).</td>
</tr>
</tbody>
</table>
the string that seems to indicate the period onto the last segment in the URI.

These URIs specified by different expressions indicate the same period, and are equivalent RDF resources. Therefore, they can be linked to each other by relation of “owl:sameAs”. However, from the viewpoint of normalization, it is necessary to decide the standard expression. In this study, I selected URIs specified by the Julian Date as the standard, as this allowed every user to describe the expression easily, independent of available language and character set.

C. URI of Resources for the Common Temporal Axis

Resources indicating position on the common temporal axis are required, as well as calendrical period resources. These resources are used to combine calendrical period resources based on different calendars, or to describe the relationship between a calendrical period and ISO 8601 expression. The URI of resources for the common temporal axis is defined as follows.

http://datetime.hutime.org/date/ JD | ISO 8601 expression

The first segment in the path is “date”, and indicates that this URI is a resource for the common temporal axis. The last segment indicates the position of a date on the common temporal axis. The Julian Date is the standard in this segment, but for convenience, ISO 8601 expression is also possible to use in the segment.

III. INFORMATION ASSOCIATED WITH CALENDRICAL PERIOD RESOURCES

A. Information Associated with Calendrical Period Resources

Calendrical period resources are associated with various kinds of information using RDF triples. Fig 1 shows an example of information associated with a date resource, each node shown as a resource or literal value.

The label of this resource is “C.E. 2017 November 08”, indicating the standard period expression. The resource type is “hutime:CalendarDate” (Prefix “hutime:” shows http://resource.hutime.org/ontology/). This label and type are associated with the resource by properties of “rdf:type” and “rdfs:label”, respectively.

Two Julian Dates are associated with the resource by the properties “hutime:jdBegin” and “hutime:jdEnd”, which indicate the strict temporal range of the date resource on the common temporal axis.

The properties “hutime:previous” and “hutime:next” show the previous and next resources of the same calendar and type. These associations are useful when there is discontinuity between dates because of leap days or calendar reform. For example, in the Julian / Gregorian calendar (Calendar ID: 101.1), the next date following 4th October 1582 (JD = 2299159.5) is 15th October 1582 (JD = 2299160.5) because of calendar reform. As this association is the same in the other calendrical periods, it is useful to know discontinuous period expressions.

Whether the resource is a leap day or not is shown by the property of “hutime:isLeap”. This value is a Boolean literal. It is set to “True” in year resources including a leap day. The property of “hutime:dayOfYear” shows day count from the beginning of the year, and is “312” for the example.

Hierarchically higher resources (i.e. month and year resources) are associated with the date resource by the properties “hutime:ofMonth” and “hutime:ofYear”. The property “hutime:ofCalendar” indicates the calendar on which the resource is based.

Calendrical period resources are also associated with external resources, though this is not shown in Fig 1. For example, DBpedia resources [10] which indicate year or date having overlapped temporal range, are associated by the property “dcterms:relation”. As DBpedia is the largest hub of the semantic web world, association allows users to reach a large variety of resources, and obtain information by following links.

![Fig 1. Example of nodes associated with a date resource.](http://datetime.hutime.org/calendar/101.1/date/2017_November_8)
Information associated with calendrical period resources is more important for resources that are not date resources. Fig 2 shows an example of information associated with an era resource of the Japanese calendar (Calendar ID: 1001.1). In the Japanese calendar, a year is specified by the combination of an era and a year count from the beginning of the era, and therefore, information about the era is important. However, there is no regularity in names and periods of eras. Therefore, users require information about eras to specify a year. Fig 2 shows an example of the era resource, referring to the Showa (昭和) era.

The label and type of the resource are “昭和” and “hutime:CalendarEra”, respectively. The calendar resource on which the era resource is based has calendar ID “1001.1”. Properties to show these values are the same as for the date resource.

Strict temporal range on the common temporal axis is shown by the properties “hutime:jdBegin” and “hutime:jdEnd” in the same way as the date resource. Two date resources, indicating the 25th day of the twelfth month in the first year of Showa era (昭和1年12月25日) and the 7th day of the first month in the 64th year of the era (昭和64年1月7日), show the beginning and end date of the era in the same calendar. These are associated with the era resource by the properties “hutime:begin” and “hutime:end”, respectively.

The duration of the Showa era is also shown, and is 62 years 0 months and 14 days according to the same calendar, or 22660 days in day count. Day count is a literal, while durations according to a calendar are shown as resources, because it is necessary to associate these with a calendar.

The previous and next eras of the Showa era are Taisho (大正) and Heisei (平成), respectively, and can be determined from the resources associated with the Showa era resource by the properties “hutime:previous” and “hutime:next”.

As mentioned above, name, period and duration of Japanese eras are irregular. Therefore, this kind of information associated with an era resource is important to users.

![Fig 2. Example of nodes associated with an era resource.](http://datetime.hutime.org/calendar/1001.1/era/昭和)

**B. Role of Common Temporal Axis Resources**

A calendrical period resource can be converted into resources based on other calendars through a common temporal axis resource (Fig 3). Common temporal axis resources indicating the beginning and end days of the period can be obtained by following the links. Because these obtained resources are associated with date resources based on different calendars by the property “hutime:tEquals”, a user can convert a calendrical resource into resources for another calendar.

A relationship between a calendrical period resource and ISO 8601 expression can be determined in the same way. A common temporal axis resource and ISO 8601 date expression have a one-on-one relationship, and are associated with one another by the property “hutime:iso8601”. Therefore, when a user obtains common temporal axis resources associated with the period of a resource, they will consequently also know the expression of the period in ISO 8601.

Conversely, users can also associate ISO 8601 expressions with calendrical period resources. An ISO 8601 expression can be used for the URI of common temporal axis resources instead of a Julian Date. Therefore, if a user obtains a specific common temporal axis resource in this way, it is easy to determine calendrical period resources which have the same period. As mentioned above, it is usually difficult to determine the real date indicating a strict ISO 8601 expression “1500-01-10”. Using this mechanism it is easy to determine that the expression “1500-01-10” indicates the 1st of January 1500 of the Julian calendar which was used at the time.
IV. LINKED DATA OF CALENDRICAL PERIODS

The calendrical period resources and common temporal axis resources, and information related to these resources, are available as Linked Data (Fig 4). Additionally, a feature has been implemented to interpret temporal expression. Since these data are released under the Creative Commons CC-By license, they may be freely used (i.e. Linked Open Data).

When a user directly accesses the URI of a resource, the access is redirected to the appropriate content based on the accept header in the HTTP request (303 redirection mechanism) [11]. Access from a tool for RDF such as a Linked Data browser is redirected to RDF data in XML format; access from a web browser is redirected to a web page. Although the web page is written in HTML, it contains links to download RDF data in a variety of formats such as Turtle, JSON and CSV. Additionally, the web page is scripted with XHML+RDFa.

V. USAGE OF CALENDRICAL PERIODS

The following code is an example of the usage of calendrical period resources in RDFa + HTML.

```html
...
<p about="http://dbpedia.org/page/Christopher_Columbus">Christopher Columbus (1451 - 20th May 1506)
He discovered the Americas in the voyage of the 1490s.
</p>
...
</html>
```

RDFa + HTML is a method to embed RDF into HTML using attributes of elements [12]. When a user loads the above code into a web browser, the following text is displayed.

Christopher Columbus (1451 - 20th May 1506)
He discovered the Americas in the voyage of the 1490s.

Calendrical resources are embedded in “span” elements using the “resource” attribute, and annotate period expressions “1451”, “20th May 1506” and “1490s” in the text. These
resources are associated with the person and the event by the properties “schema:birthDate”, “schema:deathDate”, and “dc:date”, using the “property” attribute, respectively. These embedded RDF triples can be extracted by tools in external websites (e.g. RDFa 1.1 Distiller and Parser [13]) or by plugins installed into a web browser (e.g. OpenLink Data Explorer [14]). Fig 5 shows a graph of extracted RDF triples from the above code.

Not only the date expression “20th May 1506” but also the expressions “1451” and “1490s” are associated with calendrical period resources, and then the user can determine ISO 8601 expressions indicating the beginning and end dates of these periods by following links. It is difficult to describe these dates using only ISO 8601, but the calendrical resources make it easily possible.

However, these RDF triples are not strictly correct. Most vocabulary used for RDF is designed on the assumption that date and time are expressed in ISO 8601. Therefore, the properties “schema:birthDate”, “schema:deathDate” and “dcterms:date” are defined such that the literal value of an ISO 8601 expression is associated as an object in an RDF triple, but a resource such as a calendrical period resource cannot be the object [15][16][17][18]. Although the calendrical period resources are useful, this challenge remains to be solved in the future.

VI. CONCLUSIONS AND FUTURE WORKS

In this study, calendrical period resources that interpret and normalize temporal expressions, and are associated with various kinds of information, were proposed. These resources have been already implemented and provided as Linked Open Data. This mechanism enables users to use various kinds of calendrical period expressions in RDF appropriately. The temporal range indicated by the period expressions becomes clear and comparison on the common temporal axis becomes possible, irrespective of whether the expressions are based on different calendars or ISO 8601.

These resources are available to appropriately describe periods including dynasties, events, and ages (e.g. “middle ages”, “Renaissance”), and bridge the described temporal names to usual ISO 8601 expressions. Additionally, events that occurred on a specific date or in other calendrical periods can be listed using RDF because the calendrical period resources can be the subject in an RDF triple. These usages of date, month, year, and the other periods are impossible to express using only ISO 8601.

Methods to indicate temporal range have mainly been developed to describe events in the present, but historical events have not been sufficiently considered. Moreover, although dates based on local calendars are strongly tied to local events and cultures, ISO 8601 cannot express the relationships. The present study solves these issues, advancing semantic web technology and the Linked Data world.

Several issues are not addressed in this study. Values indicating the past (e.g. “5730 BP”) are often used in archeology and geology. A mechanism that associates these values with calendrical period resources is required. Resources indicating the time in a day are also a remaining issue. Expression of the time is more complex, as there are differences in hour systems between countries, especially before the 19th century. Future work also includes analysis, performed using Web API or SPARQL, of relationships between temporal expressions based on Allen’s time interval algebra [19].

It is expected that calendrical period resources will become essential for retrievals and analyses in the Linked Data world, as they play the same key role as spatial resources for GeoNames.

ACKNOWLEDGMENT

This study is mainly supported by JSPS KAKENHI “Semantic chronology - Construction of infrastructure for visualization and utilization of knowledge along temporal axis” (15H01723), and partially supported by JSPS KAKENHI JSPS KAKENHI 16H01897, 16H01830, 16H02918, 17H00773 and 23300097.

REFERENCES

Time Information System Web HuTime: Comparison with Existing Web Applications

Tatsuki Sekino

Abstract: Web HuTime, a web application developed by the HuTime Project, can be embedded into users’ web pages. The application can display text and numerical data on timelines and in charts, respectively. Further, it enables users to stack these timelines and charts on the same temporal axis simultaneously. This feature of Web HuTime gives it an advantage over other web applications, and makes it appropriate for historical science, which requires the displaying and comparing of various kinds of information in chronological order.

Keywords: HuTime, web application, timeline

1. Introduction

The HuTime Project (http://www.hutime.org/) is developing the time information system HuTime and other related tools to establish an analysis environment for spatiotemporal information in cooperation with GIS (Sekino 2015). Although HuTime is being developed for the humanities (the “Hu” in the name HuTime means “Humanity”), it is widely used for temporal analysis in various scientific fields such as area study, history, and environmental studies (Sekino 2009a, Sekino 2009b, Shibayama 2012, Hara and Sekino 2012). However, it has to be installed on the user’s PC.

Various kinds of information are distributed through the internet nowadays, and data, documents, and photo images associated with historical studies are no exception. In addition, information technology on the web, such as cloud computing and linked open data (LOD), is rapidly evolving (Bizer et al. 2009, Armbrust et al. 2010). If technologies developed in the HuTime Project are applied to the information on the web, various kinds of information for historical study could be easily displayed on a webpage in chronological order. Consequently, to achieve this goal, the project is engaged in the development of a time information system comprising tools and data that are available on the web. A typical product resulting from this is Web HuTime, which is based on the original HuTime (Desktop HuTime), and can be embedded in web pages. This paper reviews Web HuTime, compares it with other similar applications, and considers the advantages of Web HuTime in historical study.

![Fig. 1](http://www.hutime.org/) Web HuTime, developed in the HuTime Project, is a time information system that facilitates visualization of time series data. The most unique feature of the application is a function to stack different kinds of time series data. In this figure, a chart of temporal changes in world population and two timelines of major historical events in the world and Japan are stacked on the same temporal axis. Users can move or zoom the temporal range on which the charts and the timelines are displayed using an interface displaying a year scale (http://www.hutime.org/).

1 Research Institute for Humanity and Nature, Japan.
2. Web HuTime

Web HuTime is a time information system that can be embedded in web pages (Fig. 1). It displays timelines (chronological tables) and time series charts according to user data.

A Web HuTime timeline contains event records expressed by title texts and bars showing the time interval of the event. Users can define style (color, font, and size) of title texts and bars. When a user clicks a bar in the timeline, detailed information about the event is shown in a popup window (Fig. 2). The information in the popup window is in accordance with user data. It can contain text, location (latitude and longitude), images, web links, and other embedded objects such as Google Map images and YouTube movies.

These timelines, charts, and the time rulers can be stacked on the same temporal axis in Web HuTime. This function facilitates comparison of different types of information in chronological order.

Users can change the temporal range displayed in Web HuTime with an interface placed under the data display panels. When the user moves a slider in the interface or drag data panels, displayed temporal range is scrolled. When the user changes the slider width, the displayed temporal range is enlarged or shrunk.

Web HuTime only has a visualization function, although Desktop HuTime has analysis functions as well as visualization. However, the data format is common to both versions of HuTime, and therefore, users can easily publish data and analysis results composed by Desktop HuTime on the web using Web HuTime.

Fig. 3 Web HuTime has functions for charting that can be stacked on the same temporal axis. Two line charts about population dynamics in regions of Japan are shown (regions are separated by color).

Web HuTime has functions to create three kinds of charts (bar, line, and plot charts) as well (Fig. 3). Users can define the style of symbols, and can relate detail information shown in the popup window to plots in the chart. Web HuTime can display two types of time rulers additionally defined by users. Time ruler with ticks is appropriate for showing the date or other numbers related with the time axis. The other type of time ruler consists of bars showing time intervals, and is appropriate for showing historical periods (e.g., era, kingdom, and regime).

Fig. 4 Two methods of providing JSON data originating from Desktop HuTime to Web HuTime. (a) XML files converted into JSON data via a service using a HuTime Project server. (b) User downloads JSON files from a server of the data provider.
Desktop HuTime data consist of three types of files. The data body for a layer is described in a file. A gtm file contains the metadata of a layer, and is linked with the data body file. Pairs of data body and metadata files are integrated as a project in a gts file. Although these types of files are in XML format, users can create them from comma separated (CSV), tab separated (TSV), or MS-Excel files using an editor tool that can be downloaded from the HuTime website. However, these files for Desktop HuTime are inconvenient to users of Web HuTime because of the cross domain restriction, which restricts users to downloading XML files only from the same domain that downloaded the web page. There are two ways to convert data files for Desktop HuTime into JSON format for Web HuTime (Fig. 4). When a data provider uploads XML files for Desktop HuTime onto his/her own server, users can get JSON files using a conversion service on a HuTime server (Fig. 4a). On the other hand, if the data provider creates a JSON file, then users can get the JSON file directly from the provider’s server (Fig. 4b).

Web HuTime can be used as part of other web applications. When a database returns a JSON file for Web HuTime as a search result, the results can be displayed on Web HuTime in chronological order. Studies on embedding of Web HuTime into web applications such as database are ongoing (Fig. 5).

![Fig. 5](example_url) Example of web application using Web HuTime. A database about smallpox vaccine records in Japan (Kawaguchi 2009, 2014) uses Web HuTime to show search results in chronological order (experimental version). This database is a subsystem of the DANJURO system for population studies in the 17th-19th centuries in Japan (Kawaguchi 2009).

![Fig. 6](example_url) SIMILE Timeline is freeware developed by the SIMILE Project of the Massachusetts Institute of Technology. This sample shows the assassination of John F. Kennedy and related events. This is a typical simple timeline application that shows only text. Timescale and distribution of events are shown in the lower panel. (http://www.simile-widgets.org/timeline/)

### 3. Similar Applications

Various kinds of web applications that create timelines and charts are used on the internet. As timeline applications such as blogs and SNS have become popular, such content has increased. Users previously had to display their content and related information in chronological order using timeline applications published as Flash applications. However, applications based on HTML5 + JavaScript, which are easy to customize, have recently become popular.

#### 3-1. Timeline application

SIMILE Timeline is a pioneer of timeline web applications (Fig. 6). Although this application only displays textual data on a timeline, it has many users because the system is open source, simple, and flexible. For example, nihuINT, which is an integrated retrieval system from the National Institutes for the Humanities in Japan, uses SIMILE Timeline as an interface for
temporal retrieval condition (Fig. 7). Timeglider is a popular simple timeline web application as well (Fig. 7). The application has a function to display photos on a separated panel.

![Fig. 7](image7.png)

Because SIMILE Timeline is open source software licensed under the BSD license, users can customize the functions of the application. This figure is an example of customized usage of the application. The National Institutes for the Humanities uses SIMILE Timeline as an interface to input retrieval condition in its resource sharing system (nihuINT). Users can select temporal range for graphical retrieval using functions of SIMILE Timeline (http://int.nihu.jp/e/).

There are timeline applications that enable the displaying of various kinds of content on timelines, whereas simple timeline apps such as SIMILE Timeline and Timeglider can display only text and icons on the timeline. The dipity web application can display photos, videos, and text, and share them on the web (Fig. 8). These data that are displayed on the dipity is similar to the detail information window of other applications. Similarly, Tilo Toki is an application that displays various kinds of information directly on the timeline (Fig. 9). This application has two types of display modes: 2D and 3D. The time axis is placed from the back to the front of the display in 3D mode, and images are displayed on it.

![Fig. 8](image8.png)

The Timeglider web application is a simple timeline that was developed by a private company. It is free for students, but more than 5 USD per month (according to plan) is required for other users. This sample shows the history of Idaho, and related photos. The right lower popup window is for the keyword search function (http://timeglider.com/).

![Fig. 9](image9.png)

Various kinds of data, including photos and movies, are displayed on a timeline in the dipity web application. This application is provided from Underlying, Inc., and is free for personal use. As the main purpose of the application is data sharing, it is easy to display and link SNS content on the timeline (http://www.dipity.com/).

![Fig. 10](image10.png)

Tilo Toki from Webalon Ltd, UK allows users to display images on a timeline. Windows displaying an image and its explanation are placed from back to front of the display in 3D mode (lower panel), while they are placed from left to right in 2D mode (upper panel). Tilo Toki has a desktop version as well as the web version, and can be used free. However, users cannot embed the timeline in their web pages. More than 7.5 USD a month are required to embed the timeline into a user’s web page. Users who pay a fee can additionally use multiple timelines (http://www.tiki-toki.com/).
Some applications use timelines as interfaces to select main contents. Timeline JS has a main panel that displays photos and explanations, in addition to a timeline under the main panel (Fig. 11). Users can specify a photo to display on the main panel, and select thumbnails to display on the timeline. TimeRime has similar functions to Timeline JS (Fig. 12). When a user selects a record on the timeline, an enlarged photo and detail information about the photo image are displayed.

Fig. 11 Timeline JS is open source software developed in Knight Laboratory, Northwestern University. This figure shows “Mandela: A Life of Purpose,” edited by TIME Inc., as an example (http://world.time.com/2013/12/05/nelson-mandelas-extraordinary-life-an-interactive-timeline/). When a user selects a thumbnail on the timeline, a photo and detail explanation is shown on the upper panel (http://timeline.knightlab.com).

Some web applications try to link timelines with geographical maps. myHistoro has a Google Maps panel as well as a timeline (Fig. 13). Records are shown on the timeline and the map simultaneously.

Many review articles on the web introduce various kinds of timeline applications. This paper is only able to introduce typical timeline applications. For example, “Digital Timelines comparison matrix” in “10 Digital Timelines” compares ten timeline applications in terms of functions, platforms, data, formats, and available content type (http://www.ipg.ugent.be/10-digital-timelines).

Table 1 is a summary of data display in web applications. The place where images and explanations are provided vary among the applications. However, event titles are displayed on the main timeline panel in all web applications. Tilo Toki and dipity can display images on their timelines. Although thumbnails are shown on the timeline in Timeline JS and TimeRime, the other software cannot display images on the timeline. Most of the applications display explanations about events in a separated window. Users choose the application according to the type of his/her dataset, especially the existence of images.

Data constructions also vary among the web applications. XML data are used by Web HuTime and SIMILE Timeline, and should be edited by users themselves. The other applications have data editing systems such as content management system (CMS), and therefore users can edit the data in their web browser, and save them on the server provided by each web application. Although these data editing services are useful, most web application services restrict the number of events or timelines that a user can save to, according to the user’s account type (i.e., payment).

Fig. 12 TimeRime is a Flash application that, like Timeline JS (Fig. 10), uses its timeline as an interface to select photos. When a user selects a record on the timeline an enlarged photo and explanation about the photo are displayed. A fee of 7.85 USD is required to use one timeline on the supplier’s website, and 199 EUR is required to embed the timeline in the user’s web page. This application service will be terminated at the end of 2016 (http://www.timerime.com/)..

Fig. 13 myHistoro is a timeline web application linked with a geographical map. This figure is a usage example edited by Kristjan that shows events associated with the French revolution (http://www.myhistro.com/story/the-french-revolution/30635/0/0/0/1). When the user selects an event on the timeline, a related place on the map is focused on, and detail information is shown on the map. Conversely, when the user selects a place on the map, timeline is scrolled to display related event. This application is provided by Clarigato Inc., USA, and is free (http://www.myhistro.com/).
Table 1 Comparison of data display among web applications.

<table>
<thead>
<tr>
<th>Application</th>
<th>Event title</th>
<th>Place where item is displayed</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web HuTime</td>
<td>timeline</td>
<td>separated window</td>
<td>separated window</td>
</tr>
<tr>
<td>SIMILE Timeline</td>
<td>timeline</td>
<td>separated window</td>
<td>separated window</td>
</tr>
<tr>
<td>Timeglider</td>
<td>timeline</td>
<td>separated panel</td>
<td>separated window</td>
</tr>
<tr>
<td>dipity</td>
<td>timeline</td>
<td>timeline</td>
<td>separated window</td>
</tr>
<tr>
<td>Tilo Toki</td>
<td>popup on timeline</td>
<td>popup on timeline</td>
<td>popup on timeline</td>
</tr>
<tr>
<td>Timeline JS</td>
<td>timeline</td>
<td>timeline (thumbnail)</td>
<td>separated panel</td>
</tr>
<tr>
<td>TimeRime</td>
<td>timeline</td>
<td>separated panel (thumbnail)</td>
<td>popup on timeline</td>
</tr>
</tbody>
</table>

3-2. Charting applications

Charting applications are often used to embed charts into users’ web pages. Some of them are specialized for time series data. SIMILE Timeplot is a charting application originating from SIMILE Timeline (Fig. 14). It inherits basic functions and data formats from SIMILE Timeline. The web application dygraphs is a popular charting application (Fig. 15). Because its data format is simple, many official natural science organizations (such as NASA and NOAA) use it to show scientific data on the web. Highstock JS is a multifunctional charting application that enables users to stack charts like Web HuTime (Fig. 16). This application can make more than ten kinds of charts, while Timeplot and dygraphs can make only line charts.

Fig. 14 SIMILE Timeplot is a charting web application developed by the SIMILE Project of the Massachusetts Institute of Technology. It uses the same data format as SIMILE Timeline. This application is open source software licensed under the BSD license as well (http://www.simile-widgets.org/timeplot/).

Fig. 15 The web application dygraphs is an open source application for charting that can be embedded in the user’s web page. When the mouse cursor moves onto a plot on the chart, date and value is shown in the upper left of the chart. This application is utilizes a simple data format (e.g., “2009/07/12,100,200” meaning values 1 and 2 are 100 and 200 on 12th September 2009) (http://dygraphs.com/).

Fig. 16 Highstock JS is a charting web application developed by Highsoft AS, Norway. This application is available free only for non-commercial use, although it is commercial software. This application supports various kinds of charts, such as candle stick, OHCL, and area charts, as well as line and bar charts. In addition, users can stack charts on the same temporal axis (http://www.highcharts.com/products/highstock).
4. Discussion

4-1. Comparison between Web HuTime and the other applications

The most unique feature of Web HuTime is a function to stack timelines and charts on the same temporal axis. Only a few applications for timeline and charting enable users to stack timelines or charts. However, no web application enables users to display timelines and charts simultaneously except for Web HuTime.

Another advantage of Web HuTime is cooperation with data and tools for temporal information. As stated above, Web HuTime was developed as a time information system tool in the HuTime Project. The project's objective is to construct an integrated research environment for temporal information that consists of time information system and basic data (e.g., calendar conversion) (Sekino 2015). Therefore, users can make data, use basic data, and analyze the data using Desktop HuTime, although Web HuTime only has a display function. As Web HuTime cooperates dynamically with these tools and basic data through the internet, then the application operates as a true time information system on the web. Only Web HuTime has such an integrated environment for temporal information.

This study revealed that Web HuTime has not only advantages but also disadvantages compared to other applications. The most important issue is a lack of flexibility of functions. For example, SIMILE Timeline can be customized and used as an input interface, but Web HuTime has only a fixed function (i.e., data display). In addition, placing users' texts, icons, and images on timelines are difficult in Web HuTime, although users often use them to explain data. Another disadvantage of Web HuTime is its input data format complexity. Web HuTime uses gts, gtm, and data xml files for compatibility with Desktop HuTime, but it is an obstacle to data construction. Simple data formats such as dygraphs (Fig. 15) will be required for usability.

4-2. Future work

The HuTime Project is trying to develop Web APIs for Web HuTime. Because the current version is designed with a function base system, it is difficult for users to customize any actions in the application. Therefore, the project is reconstructing Web HuTime into an application of class base system, and is adding APIs to enable users to customize it easily.

Second, the project will develop functions to link Web HuTime with basic project data. Basic data about the date in some kinds of calendar (Gregorian, Julian, Japanese, and the other calendars) are provided from the project as linked data (http://datetime.hutime.org) (Sekino 2014). Moreover, these data in a calendar can be converted into the format used by other calendar (Sekino and Yamada 2013). These functions and data will be accessed through web APIs. HuTime will display the date of each calendar using this API.

4-3. Advantages of Web HuTime for historical study

In historical studies, researchers often display various kinds of information in chronological order, and compare them to each other. Comparisons between statistic values and texts data are not rare. The relationship between changes in population and political events is a good example—it requires both the timeline and the charting functions. Web HuTime enables users to display timelines and charts on the same temporal axis, and is appropriate for historical studies. Furthermore, the integrated analysis environment for temporal information provided from the project will be linked with Web HuTime in the near future. In particular, the link with calendar data is important for historical study, and will contribute to data construction, visualization, and analysis of historical data.

These features of HuTime are unique, and impossible to replicate in other web applications. This suggests that Web HuTime has the ability to satisfy requirements from historical study. It is expected that these tools and data will bring new developments for historical studies.

Acknowledgements

This study is supported by JSPS KAKENHI “Semantic chronology—Construction of infrastructure for visualization and utilization of knowledge along temporal axis” (15H01723), JSPS KAKENHI “Development of the Data Analysis System for Smallpox Mortality in the 19th Century” (25280123), Center for Integrated Area Studies, Kyoto University, Area Informatics Projects “Practical application of spatiotemporal information to area study,” and JSPS “Topic-Setting Program to Advance Cutting-Edge Humanities and Social Sciences Research: Global Initiatives.”

References


Kawaguchi, Hiroshi 2014. "Data Analysis Programs for the
Village Vaccination Reports in 1875”, *IPSJ Symposium Series*, 2014(3) pp. 81-86. (in Japanese)


Crowd-Sourcing Scoping Study

Engaging the Crowd with Humanities Research

Stuart Dunn and Mark Hedges

Project Administrator: Angela Waplington

Centre for e-Research, Department of Digital Humanities
King’s College London

www.humanitiescrowds.org

All URLs cited work as of 12th November 2012

A project of the AHRC Connected Communities Theme
Executive Summary

This project sought to establish a credible definition for, and the current state of the art of, crowd-sourcing in the humanities. The questions included what the humanities have learned from other research domains, where crowd-sourcing is being exploited, what the results are, why academics are motivated to undertake such activities, and why members of the public are willing to give up their time, effort and knowledge for free. We conducted a survey, supplemented by a set of follow-up interviews, of contributors’ motivations, which received 59 detailed responses with qualitative and quantitative information about why people contribute to humanities (see Appendix A). The project identified and reviewed 54 academic publications of direct relevance to the field, and a further 51 individual projects, activities and websites which document or present some application of humanities scholarship making use of crowd-sourcing (see Appendix B). Two workshops were held, one for academics making use of crowd-sourcing, and one for contributors to those projects.

Academics in the humanities undertake crowd-sourcing projects for a variety of reasons: to digitize content, to create or process content, to provide editorial or processing interventions, and so on. Judging the current value of crowd-sourcing in the humanities is therefore extremely difficult, even before issues of trust, reliability and academic rigour are accounted for. However, one common factor is that humanities crowd-sourcing succeeds where vibrant and interacting communities of contributors are created. Whilst the motivations of crowd-sourcing contributors are every bit as diverse as those of academics, passion for the subject (a characteristic shared with academics) is the dominant factor which draws them together into communities. These communities develop and perpetuate internal dynamics, self-correct, provide mutual support, and form their own relationships with the academic world. Despite the great diversity of humanities crowd-sourcing, it is possible to observe patterns in which such communities thrive: these patterns are dependent on the correct combinations of asset type (the content or data forming the subject of the activity), process type (what is done with that content) task type (how it is done), and the output type (the thing produced) desired. In this report, we propose a high-level typology which describes different instances of each of these, and identifies the combinations that are, on present evidence, most successful in achieving projects’ aims.
1. Introduction

1.1 Background

Crowd-sourcing\(^1\), the process of leveraging public participation in or contributions to projects and activities, is relatively new to the academy, and even newer to the humanities. However, at a time when the web is simultaneously transforming the way in which people collaborate and communicate, and merging the spaces which the academic and non-academic communities inhabit, it has never been more important to consider the role which public communities - connected or otherwise - have come to play in academic humanities research. The purpose of this report is to present a review of literature of crowd-sourcing applications in the academic humanities domains, to assess its impact and development, to consider the motivations and aspects of community of those who choose to participate, and to present a typology which captures the different approaches which have emerged.

It should be emphasized that this report, the result of a nine-month study, does not claim to be comprehensive: there are bound to be important projects, publications, individuals and activities that we have missed. Just as inevitably, there is a strong UK and Anglophone focus on the activities studied. Non-inclusion in this document does not reflect any lack of importance or interest in an activity. We have come across many fascinating crowd-sourcing projects beyond Western Europe and the US, and hope that this report will act as a catalyst to developing a broader community of discussion and application beyond the projects and activities that we have become aware of while preparing it.

The study consisted of four components: a literature review of academic research in the humanities which has drawn on crowd-sourcing, as well as papers detailing research into crowd-sourcing itself as a method; two workshops held at King's College London in May and October 2012 facilitating discussion between, respectively, academics in the humanities who use crowd-sourcing, and members of the public with records of contributing to such projects; a set of interviews with both academics and contributors, and an online survey of contributors exploring their backgrounds, histories, and motivations for participating. We also conducted an extensive web crawl, identifying projects using crowd-sourcing that may not yet have produced a tangible academic outcome; tools that facilitate crowd-sourcing, and relevant blogs and Twitter feeds. The projects and other activities identified from this are presented in Appendix B.

In order to explore the issues fully, we considered it necessary to conduct not just a review of research but also of research activity. The aim was to identify not just the kinds of humanities research that are engaging with wider communities, and how they are engaging, but also to look at how communities of contributors form (or are formed) and develop in the course of participation, and how notions of community matter both to them and to the academics running such projects.

---

\(^1\) In this report we follow the convention of hyphenating ‘crowd-sourcing’; other authors use ‘crowdsourcing’ or ‘crowd sourcing’. In quotations, we preserve the original form.
This is a complex and wide-ranging area, and it is necessary at the outset to define terms and the boundaries of the review. Public involvement in the humanities can take many forms – transcribing handwritten text into digital form; tagging photographs to facilitate discovery and preservation; entering structured or semi-structured data; commenting on content or participating in discussions, or recording one's own experiences and memories in the form of oral history – and the relationship between the public and the humanities is convoluted and poorly understood.

This diversity presents two immediate challenges for a review of crowd-sourcing as a research method. Firstly, in purely semantic terms, where should the boundaries of what is considered to be crowd-sourcing lie? And secondly, since humanities crowd-sourcing is in its very early stages, there is relatively little academic literature dealing with its application and outcomes to allow any firm judgements to be made about its potential to produce academically credible knowledge. Given this lack of evidence, we therefore do not seek to make value judgements on any individual cases, and we stress that equally this report does not seek to evangelize or promote crowd-sourcing as a method. It simply seeks to identify what, on present evidence, seems to work and what does not. Moreover, this underlines the need to examine other, less formal, sources of information, such as blogs and interviews, and emphasises that at this early stage, it is just as important to consider the academic validity of processes as well as outcomes.

2 Terminology and typologies

2.1 Analysis of prior research on terminology and typologies

The term crowd-sourcing is frequently used as a convenient label for a diverse range of activities. It was originally coined in 2006 in an article in *Wired* by Jeff Howe entitled *The Rise of Crowdsourcing* (Howe 2006). In this article, Howe draws a parallel between businesses farming out labour to cheaper markets in the developing world, and businesses utilising ‘the productive potential of millions of plugged-in enthusiasts’, with similar reduction in labour costs. In recent years, academics have come to use the power of the crowd to achieve research aims.

As a method of undertaking academic research, however, the term ‘crowd-sourcing’ is problematic. It is certainly less easy to define than the analogous term ‘citizen science’, which is commonly understood to refer to activities whereby members of the public undertake well-defined and (individually) small-scale tasks as part of much larger-scale scientific projects (Silvertown 2009), but which, in the past, has also been used to refer to more passive forms of participation such as making available unused CPU power of desktop machines for harvesting by research teams.

Furthermore, most discussions of crowd-sourcing treat it as being distinct from the concept of the ‘Wisdom of Crowds’ as originally advanced by James Surowiecki in 2004 in *The Wisdom of Crowds*.

---

2 The desk research for this review identified around sixty papers of potential relevance.

3 A good example is the Search for Extra-terrestrial Life project, see Anderson et. al. 2002; see also Anderson and Fedak 2006.
Crowds: Why the Many are Smarter than the Few (2004), which holds that large-scale collective decision-making can be superior to that of individuals, even experts, a thesis that lacks the elements of collaboration around activities conceived and directed for a common purpose that characterise crowd-sourcing as commonly understood. Although academic crowd-sourcing can be about decision making - and we make provision in our typology for such projects - the decisions involved are rarely as neatly packageable as those implied in the world of business, where the ‘good’ or ‘bad’ nature of a decision can be evaluated on the basis of profitability (Brabham 2008).

In their review of the field, Towards an integrated crowdsourcing definition, Estelle’s-Arola and González-Ladrón-de-Guevara (2012) identify eight characteristics, distilled from 32 distinct definitions identified in the literature: the crowd; the task at hand; the recompense obtained; the crowdsourcer or initiator of the crowdsourcing activity; what is obtained by them following the crowdsourcing process; the type of process; the call to participate; and the medium. This extremely processual definition is comprehensive in identifying stages which map easily to business processes. For the humanities, the ‘type of process’ is both more significant and more problematic, given the great diversity of processes in the creation of humanities research material. A more task-oriented approach is that of Wiggins and Crowston (2011), who construct a typology for ‘citizen science’ activities. The use of the word ‘science’ (at least in the usual Anglophone sense) confines the activities reviewed (in terms of both the methods and the content) to a particular epistemic bracket which inevitably excludes some aspects of humanities research. Wiggins and Crowston identify five areas of application: Action, Conservation, Investigation, Virtual, and Education (2011). The factors that lead to an activity being assigned to a category are multivariate; and the categories’ identification was based on whether there is an occurrence in a category or not, rather than frequency of those occurrences. The coverage is therefore extremely broad. ‘Action’, for example, covers self-organising citizen groups that use web technologies to achieve a common purpose, often to do with campaigns on local issues.

One widely-quoted set of definitions for citizen science projects was presented by Bonney et. al. in their report for Center for the Advancement of Informal Science Education (CAISE), Public Participation in Scientific Research: Defining the Field and Assessing Its Potential for Informal Science Education (Bonney et. al. 2009). This divided the field into three broad categories: Contributory projects, in which members of the public, via an open call, contribute along lines that are tightly defined and directed by scientists; Collaborative projects, which have a central design but to which members of the public contribute data, and may also help to refine project design, analyze data, or disseminate findings, and finally Co-created projects, which are designed by scientists and members of the public working together and for which at least some of the public participants are actively involved in most or all steps of the scientific process. This approach shares important characteristics of the ‘task type’ typology facet developed below, in that it is rooted in the complexity of the task being asked of the public, and the amount of effort, initiative and independent analysis required to make a contribution.

Certain subsets of the humanities disciplines have seen some efforts to develop their own typologies for crowd-sourcing, Most notable among these are the cultural heritage and
Galleries, Libraries, Archives and Museums (hereafter GLAM) sectors. This interest is hardly surprising since, unlike most humanities domains, these sectors are inherently public facing, and have long traditions of volunteerism and public engagement. Most museums, especially smaller ones, exist on competitive principles of attracting and engaging audiences to justify their funding; and ‘memory institutions’ such as national libraries and museums have formal duties to maintain access to their collections, both for scholars and the public. Against this background, approaches such as that of Tim Copeland have emphasised the importance of ‘constructivist’ approaches, where the public is encouraged to engage with the interpretation of collections, rather than ‘positivist’ approaches where they are passive recipients of knowledge organized by curators (Copeland 2004). Projects such as UCL's QRator have sought to achieve this by providing iPads as a channel for visitor feedback on the collections of the Petrie Museum (http://www.qrator.org/).

One typology for crowd-sourcing with a special focus on the GLAM sector has been suggested by Mia Ridge in a blog post (http://openobjects.blogspot.co.uk/2012/06/frequently-asked-questions-about.html). In this, Ridge proposes the categories Tagging, Debunking (i.e. correcting/reviewing content), Recording a personal story, Linking, Stating preferences, Categorizing, and Creative responses. Again, these categories imply a processual approach, and are, at least potentially, extensible across different types of online and physical-world content and collections. They are concerned with the type of task that the crowd is being requested to carry out.

An alternative typology for crowd-sourcing in the GLAM domain was developed by Oomen and Aroyo (2011). Their categories include Correction and Transcription, defined as inviting users to correct and/or transcribe outputs of digitisation processes (a category that Ridge’s Debunking’ partially, but not entirely, covers); Contextualisation, or adding contextual knowledge to objects, by constructing narratives or creating User Generated Content (UGC) with contextual data; Complementing Collections, which is the active pursuit of additional objects to be included in a collection; Classification, defined as the gathering of descriptive metadata related to objects in a collection (Ridge’s ‘tagging’ would be a sub-set of this); Co-curation, which is using inspiration/expertise of non-professional curators to create (Web) exhibits (somewhat analogous to Bonney et. al’s co-created projects category, but more task-oriented); and Crowdfunding, or the collective cooperation of people who pool their money and other resources together to support efforts initiated by others (see, e.g. BBC: Flag Fen hosts ‘crowdsourced’ Bronze Age Archaeology dig: http://www.bbc.co.uk/news/science-environment-19192220). Ridge (op. cit.) explicitly rejects crowdfunding as a component of crowd-sourcing.

2.2 Characteristics of crowd-sourcing identified by the review

In the workshop for academics using crowd-sourcing in their work, organised by this review in May 2012 (http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/09/workshop_report1.pdf), four factors were identified that characterise, very broadly speaking, crowd-sourcing for the humanities. These characteristics are not exhaustive, but they usefully highlight commonalities between some of the activities we have observed:
a) The existence of a clearly-defined humanities direction and/or research question. The question could be posed/designed by an academic team, or by an individual with particular knowledge and/or interests. This seems to preclude some categories identified by some authors, such as the ‘Action’ category of Wiggins and Crowston (2011); and will preclude elements of others, such as the ‘Co-created’ projects of Bonney et. al. (2009). It is suggested here that this characteristic is especially significant, since the academic component of academic humanities crowd-sourcing implies some form of professional rigour. However, we do not assume that the source of that rigour must necessarily originate from a Higher Education Institution.

b) The potential for a group with open membership to transform or add value to primary material or the interpretation of this material. However, a distinction has been made elsewhere between ‘community sourcing’ and crowd-sourcing, with the latter typically dealing in open calls for participation (a key factor for Brabham 2008); and the former being more closed (see, e.g. http://millennialmarketing.com/2010/09/crowdsourcing-vs-community-sourcing).

c) There needs to be a definable task, or some meaningful and replicable way of breaking the workflow down into sets of definable tasks.

d) The activity should be scalable, both to different volumes of data and different levels of participation.

In the light of these characteristics, crowd-sourcing is considered to be distinct from the production of general user-generated content (UGC) on platforms such as Google Earth, as there is no clearly-defined direction or question, although such platforms could be components of crowd-sourcing projects if such a direction were present. Equally, the harvesting and analysis of so-called transactional data, that is information about people’s (usually online) activities, is not considered here to constitute crowd-sourcing, as whatever additional value is added to the data does not result from public participation.

It has also been suggested that crowd-sourcing is distinct from ‘crowd-funding’, where large groups of people are invited to fund projects by individually contributing small sums of money, via sites such as kickstarter. However, it has become apparent that there are likely to be some crossovers where people who contribute financially to such projects are also offered the opportunity to get involved in some way. An intellectual contribution and a financial contribution are not mutually exclusive, and accordingly we have accommodated this aspect in our typology.

While (a) serves to distinguish crowd-sourcing from collaborative activities whose common purpose is primarily social or campaigning, it should be noted that these are not entirely disjoint categories. In particular, recent years have seen the emergence of several SMEs dedicated both to social goals and to the forms of principle behind humanities crowd-sourcing that we have uncovered in this review. An example of this is HistoryPin (http://www.historypin.com), a
website that allows people to georeference historical photographs on a modern map (see below).

3. Levels of participation

Rose Holley, an authority on the use of crowd-sourcing in mass digitization and archives, and formerly project manager of the TROVE project which used the crowd to correct and validate Optical Character Recognition (OCR)-derived digitisations of Australian newspapers (http://trove.nla.gov.au/newspaper), identifies a distinction between crowd-sourcing and ‘social engagement’. She states:

Social engagement is about giving the public the ability to communicate with us and each other; to add value to existing library data by tagging, commenting, rating, reviewing, text correcting; and to create and upload content to add to our collections. This type of engagement is usually undertaken by individuals for themselves and their own purposes ... Crowdsourcing uses social engagement techniques to help a group of people achieve a shared, usually significant, and large goal by working collaboratively together as a group (Holley 2010).

She also notes that crowd-sourcing is likely to involve more effort, and by implication to require greater personal commitment to volunteer time for free, than social engagement - which, after all, is an extension of the kinds of online activities - Tweeting, providing content to Facebook etc - that millions do on a daily basis anyway. In a sense, this aligns crowd-sourcing with so-called ‘citizen science’, and implies a level of commitment and participation which goes beyond simple casual interest. Wiggins and Crowston (2011) develop this theme by highlighting a distinction between citizen science and community science, and stating as a key ingredient of the former that it is not self-organizing. As they state:

Citizen science does not represent peer production in the same sense as seen in prior work because the power structure of these projects is usually hierarchical. Furthermore, citizen science is not necessarily “open science,” a term that refers to open source-like practices in formal scientific research settings. Many citizen science projects share data, but may not make the full research process publicly viewable for comment and discussion (Wiggins and Crowston 2011)

A fundamental aspect of citizen science, therefore, is that the research goal is defined by a particular person or group (almost always as part of a professional academic undertaking), with participants recruited through an open call providing some significant effort towards achieving that goal or goals. The motivations that push people to contribute that effort is therefore critical, and this is a crucial distinction between ‘citizen’ projects in the sciences and the humanities. It is certainly true that the two different intellectual traditions embrace, and are embraced by, different kinds of non-academic community. This is particularly so in the domains of the
humanities related to the Cultural Heritage sectors where, as noted above, there are existing processes of interaction between the academy and the public. As Trevor Owens has written:

Most successful crowdsourcing projects are not about large anonymous masses of people. They are not about crowds. They are about inviting participation from interested and engaged members of the public. These projects can continue a long standing tradition of volunteerism and involvement of citizens in the creation and continued development of public goods (http://www.trevorowens.org/2012/05/the-crowd-and-the-library).

The term ‘participation’ itself has connotations of community and interaction. It tends to exclude notions of passivity and serendipity and, most importantly, implies a motivation that stems from interest in the subject and, by extension, community based discussion and exchange around that subject or issue.

A crowd-sourcing project should therefore have the capacity to allow large numbers of people to be involved, even if only a very small number of contributors end up being actively engaged (which is often the case). Indeed, most of the humanities crowd-sourcing projects represented at the May meeting reported that a very small number of contributors generally do a very large percentage of the work. The point is that the body of contributors is self-organising and self-selecting, and there is not be a central(ised) recruitment process.

4. Contributor motivations and engagement

4.1 Communities of crowd-sourcing

The foregoing discussion makes it clear that there can be no analysis of the role of crowd-sourcing in the humanities without detailed consideration of the motivations of those who participate in crowd-sourcing projects. This is intimately linked with notions of community, and the sense of community felt by participants. Much of the following section is derived from discussion of the workshop for crowd-sourcing contributors held at King’s College London on 18/10/2012. These contributors were a mixture of those identified by academic colleagues running crowd-sourcing projects (especially Old Weather and the British Library Georeferencer, to whom we record our thanks), and respondents from the online survey, who had ticked the box indicating they were happy to participate further in project activities.

4.2 Previous research

Most studies conclude that most crowd-sourcing contributors do not have a single motivation; the survey conducted for the current project indicated overwhelmingly (79%) that highly active contributors of the kind who responded have both personal and extrinsic motivations; that they do it both for themselves and for others. However in many cases it is possible to identify a single, dominant motivating factor, which is almost always concerned directly with the project or activity’s subject area. In an analysis of 207 forum posts and interview responses for example,
Galaxy Zoo found that the top motivation was an interest in astronomy (39%), a desire to contribute (13%) and a concern with the vastness of the universe (11%) (Raddik et. al. 2010). The first two of these align with our survey’s findings that motivations are both personal and extrinsic. This trend can be found reflected in far more niche areas. A similar study of volunteers on the Florida Fish and Wildlife Conservation Commission (FWC)’s Nesting Beach Survey project found that concern for sea turtle conservation was the overwhelming factor motivating volunteers (Bradford and Israel 2004). Moreover, studies of the motivations of the contributors to academic crowd-sourcing projects have focused on personal interest in the subject area concerned; and opportunities that projects provide to exercise that interest, and to engage with people who share it, without material benefit. Such interest is usually concerned with the outcome, but it can also be in the process, or some combination of both. For example, in her 2009 assessment of the motivation of volunteers to the TROVE project, Holley notes:

We noticed in our communication with text correctors that a large proportion was family history researchers. These people are highly motivated to learn new skills in order to get the information they need. They also have a sense of responsibility towards other genealogists to help not only themselves but other people where possible (Holley 2009).

In general therefore, it may be said that research into crowd-sourcing motivations suggest a clear primary, although certainly not exclusive focus on the subject or activity area, and that this focus can be altruistic, extrinsic or intrinsic. Our workshop of crowd-sourcing contributors also suggested that there is a distinction to be made between abstract interest in a subject area, such as mapping, and highly focused, or even obsessive, interest in a subset of that subject, e.g. maps of a particular period or area, often deriving from a personal or family connection.

4.3 Academic versus commercial crowd-sourcing

There is an obvious distinction to be made between motivations for crowd-sourcing and motivations driven by market economics, which suggest that people will only contribute effort or submit to regulation in return for some benefit, usually material. Academic participants in projects in universities have relatively clear motivations, including, but not limited to, the fact they are materially rewarded by salaries and grants, professional recognition in their field, career advancement, and publication. Most crowd-sourcing projects however do not reward their contributors in material or professional ways (at least not directly), and members of crowds who contribute to crowd-sourcing projects are not subject to discipline (in either sense of the term) or sanction in the way that members of conventionally configured research organizations are. That said, contributors may receive “social” rewards, for example through rankings, increased standing in the crowd-sourcing community, or (in the case of GalaxyZoo) being credited and named in publications. Similarly, contributors may be subjected to social sanctions, such as banning (e.g. removal of pages or blocking of accounts on Wikipedia), which can adversely affect their reputation and enjoyment, and may even in rare cases reflect on their professional standing. However, it is clear that the motivations of academic crowd-sourcing participants are more intrinsic to the activity.
This is a (further) important distinction from business models of crowd-sourcing, which offer either small-scale material recompense for input, or the prospect of larger rewards if a contributed design etc. is chosen for production. For example, in his review of business-oriented crowd-sourcing models, Brabham (2008) singles out the fact that participation in open source projects does not lead to material reward as evidence that OS does not provide a compelling model for crowd-sourcing as a business process. He states:

These questions cast some doubt on the open source model as a supreme model for product development. Crowdsourcing, however, overcomes these limitations in the open source model by providing a clear format for compensating contributors, a hybrid model that blends the transparent and democratizing elements of open source into a feasible model for doing profitable business, all facilitated through the web (Brabham 2008).

Again, though, we should not ignore indirect benefits. The knowledge of specific open source software products that one gains from contributing code to them on a volunteer basis may provide a significant advantage in the employment market. Moreover, some open source products have social structures built up around them, providing extrinsic, social motivations analogous to those noted above for crowd-sourcing projects, in addition to the intrinsic motivation of contributing to the software.

4.4 Gamification

Other approaches in the literature have emphasized the importance of tasks being enjoyable, and have focused on the development of games for crowd-sourcing of different kinds. Prestnopnik and Crowston (2011) discuss the role of games, and in particular possible approaches to creating an application for crowd-sourced natural history taxonomy classification using design science. They also note that ‘gamification’ has the potential to act as a disincentive to contributors who have expert knowledge or deep interest in the subject. The Bodiam Castle project provides a good use case of the power and the potential for use of games in the context of archaeological analysis of extant buildings, although this had a greater emphasis on visualisation than on competition (http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/04/Masinton.pdf).

However, gamification can also be a barrier for users who simply want to engage with the assets or processes in question; and furthermore it can trivialise the process of acquiring or processing data (see http://blog.tommorris.org/post/3216687621/im-not-an-experience-seeking-user-im-a for a combative assertion of this position). In their analysis of impact of the Cornell Ornithology Lab’s The Bird Network (TBN) project, where members of the public were given the opportunity to set up bird boxes near their homes on in their neighbourhoods and gather data about the use by birds which was then shared with the scientific team, Brossard et al (2005) note that participants’ interest in ornithology was likely to overshadow awareness of scientific
process; and was in turn likely to stymie efforts by the Lab to contribute to general scientific awareness and education in the US population (Tumbrull et. al. 1999).

4.5 Models of engagement

Even crowd-sourcing projects that are quite similar in terms of characteristics such as process, content or task type, can involve quite different models of collaboration and engagement with participants. Consider the following examples, each of which addresses the digitisation (via transcription) of handwritten, textual material that could not be processed by computer alone:

- MarineLives (ML) aims to create an academically respectable digital edition of 17th Century manuscripts originating in the High Court of Admiralty, London. Rather than recruiting volunteers widely via an open call, ML targets the subset of potential participants who are sufficiently dedicated to learn the required skills, commit their time, and persist when they encounter problems. Volunteers in ML make an explicit ‘deal’ with the project organisers, who require participants to commit three hours a week for fourteen weeks; in turn the project regards itself as responsible for investing in processes to sustain the contributors’ motivation and engagement. The project thus aims to recruit people who will be able to make a significant contribution from the start.

- Old Weather (OW), on the other hand, takes a quite different approach to engagement, based on a ‘long tail’ type of model. It recruits many more volunteers, but a great proportion of the work is done by a small number of them, with a far greater number transcribing only a single page.

This difference in engagement models is to some extent driven by a difference in the nature of the work being carried out. In OW, the work can be split up into small components that can be carried out independently and do not individually require a lot of expertise. Each individual transcription is performed by 3 people and cross-checked for accuracy. This is not the case for ML, which aims to create an academically respectable digital edition. The British Library Georeferencer (BLG) is another example of a project taking the ‘long tail’ approach, although this has an emphasis on spatial metadata creation, which is a task requiring more specialist expertise.

Note however that in some projects it is possible for participants to move on to a different level, and to carry out more complex tasks; examples of this are the ships’ histories created in OW, or the biographies of notable botanists in Herbaria@Home.

4.6 Roles

It is commonly noted in crowd-sourcing projects that the roles played by contributors develop as their experience increases; however, the kind of model followed by a project influences the kinds of role that emerge. In OW, a “captain” can see every page that gets transcribed, which gives an overview of the resource. There is a strong element of competition among ‘super-contributors’ (and it should be noted that this particular motivation it is likely to be particularly strong among super-contributors, such as those present at the second workshop, and less so in the long tail component of the community). For example there is a single transcriber in OW who undertakes massive quantities of work, and participants are aware that if she ‘joins’ the ship
they are currently working on, their status is likely to be reduced by comparison. As this would mean that ‘their’ ship would be completed more quickly, it clearly indicates the level of personal competition among contributors.

ML has a different model in that it actively promotes collaboration, and provides a great deal of support via the team facilitators, who are identified individuals with clearly-defined roles. In OW there is a great deal of support via the forum (and the forum moderators are also volunteers, rather than project staff). It was noted that in OW there are tasks that require different skill sets, and conversely people who possess specific skills, and there is a degree of self-organisation whereby people match themselves to tasks. For example, some transcribers are particularly good at deciphering difficult handwriting.

In the BLG, the work was accomplished so swiftly, mainly by the super users, that a discernible collaboration model or infrastructure did not have time to emerge.

A sense of community is important. One participant in the second workshop considered joining OW but decided not to because she was concerned at being ‘just a number’ among the thousands of people involved; that the lack of a defined and developable role was an off-putting factor. This also contributed to a sense that her input would be more valuable elsewhere. It was noted that OW does not provide individual feedback from project staff, in contrast to Transcribe Bentham (TB) which does; and TB transcribers consider this to be important (this was also picked up in the interview with Contributor B, 10/10/2012).

Roles are also important in the development of specialised knowledge. OW in particular provides good examples of transcribers becoming expert in specialised areas of naval history. Indeed the development of roles is allied to both the development and the pre-existence of specialist knowledge and interest. This means it is necessary to nuance what is meant by subject area. For example, only a relatively small number of people are interested in maps of Leith in the eighteenth century and therefore in georeferencing only that type of map, but many more people are interested in maps and georeferencing generally. This means it is necessary to distinguish between interest in a question and interest in a subject area or particular kind of activity. At least two participants we spoke to compared crowd-sourcing work with crosswords as an activity that was pleasurable for its own sake. Contributor E, for example (interview 23/10/2012), discovered the Old Weather project via a piece on the radio. Whilst her initial interest was piqued by the climatology side of the project, her interests broadened as a result of her involvement, and she became interested also in synthesizing and editing the histories of particular ships. Some aspects of the material had personal geographic relevance to her, for example the WWI German Fleet being taken to the Firth of Forth after its capture. The editing part of the project gives a feeling of ‘having a whole ship to yourself’, and being able to ‘tell the story of a whole ship’.

A sense of support, either from a community or from the project is critical in recruiting and maintaining contributors. Two key questions therefore are how can sustainable support structures be set up for any given project, and how do communities develop around projects?
One participant in the second workshop observed the formation of ‘a transcription community, and a researcher community, with some overlap between the two’. In building infrastructure for crowd-sourcing projects, ‘the crowd’ is a difficult thing to predict, in terms of what kind of community will be formed, and what the group will achieve. It is hard to know how much support/input will materialise, making it difficult to formulate requirements. A good example is the BLG project, where 750 maps were transcribed in four days. This completely outstripped all infrastructural and support mechanisms the BL had put in place.

4.7 The role of competition

Competition is one possible motivation for people to participate in crowd-sourcing projects, although it is worth noting that very few participants studied in the qualitative research cited above admit to being motivated by competition with each other. For many projects it is possible to track individual participants’ contributions, and acquire statistics on who contributed the most etc (although not all; see below). Where this is the case, projects can establish ‘leader boards’, indicating which participants have made the biggest contributions (in whatever terms the project is working with). In the case of the BLG project for example, it displayed the handles of the users who processed the most maps. The ‘winner’ was invited to meet the BL’s head of cartography, and this kind of contact with a prestigious institution was itself highly valued by the participant community. BL staff also felt that the project made the participants feel that they had a stake in the BL itself, and were part of the community it represents. However, in order for competition to be a significant factor in encouraging and sustaining crowd-sourcing, the tasks, and the result/outcome of accomplishing a particular task must be easily quantifiable and definable so that it can be compared to the outcomes of tasks completed by others.

The nature of the material however means this can become complicated, for example where the kinds of content are not consistent. For example, in the BLG project, some maps are more complex than others, so the team felt there was relatively little meaning in comparing the effort needed to georeference different ones. Another general possibility is that different aspects of the same project could have different leader boards, thus reflecting a model of ‘diffused competition’ (see further discussion in the report of the May workshop, http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/09/workshop_report1.pdf). As a solution however, the notion of encouraging competition should be qualified by the need not to exclude potential participants who are not, by nature, competitive people, yet may have valuable knowledge or effort to bring. Another qualification with using competition as a means of encouraging participation is the extent to which it encourages speed and volume at the expense of quality and care. One participant in the second workshop asked ‘Is there a prejudicial (or factual) connection between “free labour” and “low quality”? Even when data are vetted by software or by editors, can this idea still jeopardise the (actual or perceived) reliability of the outcome? There is also an issue of how conflicts in participants’ contributions are to be handled. This is likely to be especially so where creative/interpretive outputs are being generated. This kind of output is also less likely to lend itself to the leader board type approach outlined above.
In the second workshop, competition featured strongly as a motivating factor for participating; but this should be contextualised/qualified by the fact that all of those present were ‘super contributors’, who are likely to feel a sense of competition more keenly than those in the ‘long tail’ part of ‘the crowd’. Competition can be defined in various ways: in the quantitative sense, e.g. number of pages or other asset unit processed by individuals that can be displayed on a leader board. Others, however, are concerned with producing high-quality work in a more qualitative sense. ML for example has a volunteer (who works in a museum) whose main aim is to produce pages of excellent transcription with minimal help requests. This is not, however, incompatible with a sense of common purpose, e.g. in OW ‘you feel part of the ship’ you are working on.

4.8 Other motivations

It is possible for motivations to change over time, for example many OW volunteers are initially interested by the possibility to contribute to climate change research, but become interested in maritime history as they are exposed to the project’s content. Motivations can also change with the kind of task type. One participant remarked that ‘palaeography is only fun when you can’t do it’.

Different tasks also attract the motivation of simple curiosity, and the desire to locate new knowledge. One example given was ‘investigative’ crowd-sourcing, for example researching the corpus of MPs’ expenses published in the UK in 2009 (see http://www.niemanlab.org/2009/06/four-crowdsourcing-lessons-from-the-guardians-spectacular-expenses-scandal-experiment/)

Many participants in the second workshop felt that the crowd-sourcing activity they are engaged with is highly addictive. It was also considered important for recruitment that participants can do work on a project at any time (and put it down at any time); much like a computer game.

Personal factors are sometimes involved in the decision to engage in crowd-sourcing, for example redundancy frees up time to get involved and generates a need to keep the mind active, also displacement activity following bereavement has been cited. In this case crowd-sourcing formed part of a ‘restorative process’.

4.9 What initiates interest in a crowd-sourcing project?

Most super contributors at the second workshop discovered their projects serendipitously, for example via Twitter. The Zooniverse community has been very successful in generating interest across related projects, since it presents the projects in dashboard style and encourages cross-fertilization.

In many cases discussed, mass media exposure leads to a spike in uptake. For example in the Transcribe Bentham project a New York Times article about the project was published on 28th
December 2010⁴. This was, however, problematic. When the article appeared, many enquiries were made but the project team was not available due to the Christmas break. BBC Radio 4’s PM programme mentioned OW, which also led to a big spike.

4.10 Reward

Reward may be considered in terms of satisfying/fulfilling interest in the subject; or as a more tangible quid pro quo for being involved. This can take the form of status, some personal benefit such as training and experience (how can this be qualified and made provable?), or in instant/gamified gratification, or deferred gratification.

In many projects, the feedback loop, affirming that the contributions made were correct and valuable, is a very important component of the reward for engagement, and conversely lack of feedback can be very frustrating and discouraging.

Contributors mentioned a number of skills gained; these included general practical IT competencies, such as learning to edit wikis and to use Skype for distributed collaboration, as well as specialised skills such as XML encoding (e.g. through Transcribe Bentham). As noted above, many contributors gain domain knowledge; for example the material and opportunity to edit ships’ histories in OW for the Naval Histories website. One participant in the second workshop is now actively involved in editing a set of historical documents (ship's history), which would not have happened without been involved in the OW project. This can contrast with the interests of the project team: in this case, for example, the team is interested in weather history, whereas some contributors are interested in ships (http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/04/Brohan.pdf). Other Old Weather participants highlighted multiple motivations, including the ability to learn about history and address the issue of global warming (Interview with Contributor A, 7/9/2012). Contributor F’s interest (interview, 29/10/2012) was aroused by a piece on the BBC news site. As with others, her initial interest centred on concern about climate change, but her motivations broadened as her familiarity with the material increased. She has ‘captained’ two ships, but noted that lack of time meant it was hard work staying captain, especially when other contributors have more free time. She also became interested in synthesizing and editing the histories of individual ships, and stressed her interest in exploring niches of history that had been hitherto unexplored. For her, interaction with other participants via the forum is extremely important, both for ‘exchanging chit chat’, and for discussing the practical and technical problems that the transcription process presents. She also noted that other Zooniverse projects had not piqued her interest in the same way; this is simply because she is more interested in the area of history."

Participants can also pick up a basic grounding in research methods of collation, synthesis and analysis in the area of interest to them – contra observations made about some ‘citizen science’ projects, where a focus on the content often negates that on background method. See e.g. the Cornell Lab ornithology experiment (Trumbull et al 2009).

Where tasks develop in this way, people may gain the experience of working collaboratively, in contrast to situations where the tasks remain purely mechanical (e.g. anonymous marking up/transcribing etc. of records, as in FamilySearch).

5. Motivations of academics and other project organisers

At least part of the success of GalaxyZoo and other Zooniverse projects is that they catered to clear and present academic needs. In the case of Galaxy Zoo itself, the assets – photographs of galaxies from the Sloan Digital Sky Survey – were far too numerous to be examined individually by any research team, and the task – the classification of those galaxies’ morphology – was not one that could be performed by computer software, although for the most part could be carried out by a person without specialist expertise. In 2008, they reported that $4 \times 10^7$ individual classifications had been made by around $10^5$ participants (Lintott et. al. 2008). Quite simply, this is work that could not have been carried without large-scale public engagement and participation.

Since humanities crowd-sourcing is an emergent area, identifying the motivations of academics engaging with such work is not always so straightforward. In most cases, there is an academic research question or a need for a particular resource. For example, the Transcribe Bentham project was motivated by the fact that 40,000 folios of Bentham’s work was untranscribed, and that consequently (as the project states explicitly at [http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/04/Causer.pdf](http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/04/Causer.pdf)) [o]ur understanding of Bentham and his thought – of importance to anyone studying the eighteenth or nineteenth centuries – is incomplete. In short, the Bentham Papers are a source of enormous historical and philosophical importance, yet much of the collection remains unknown, let alone adequately studied.

The British Library’s Georeferencer project is another example of this. The Library is able to make its map collections more searchable, and therefore more exploitable by using crowd-sourcing to georeference them. Again, this is not a task that a computer could do, and that a team tasked with metadata creation could only do over a very long period of time, and with prohibitive cost.

The OldWeather project provides an example where the motivations of the researchers involved are quite clear, namely to be able to use the information contained within the assets to explore historic weather patterns (although these motivations may not necessarily be shared by the participants):

Climate researchers need the millions of historical weather records archived in Royal Navy ship’s logbooks from decades and centuries ago, marine historians want the day-to-day records of people, ships and places in the same documents (see [http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/04/Brohan.pdf](http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/04/Brohan.pdf)).
Other researchers, particularly those in the GLAM sector, see crowd-sourcing as a means of filling gaps in cultural heritage institutions’ coverage, as noted by Terras (2010). This can be an effective way of obtaining information about assets (or obtaining the assets themselves) to which only certain members of the public have access, for example through personal or family connections. However, in order to be usable for academic purposes, the ‘circulation of knowledge’ (see Section 6.1) must be curated in some way, and this may involve expert input.

Finally, crowd-sourcing is sometimes seen as a means of community building, and thus increasing the wider impact of academic research activities. Moyle et al. 2012 note:

[The project is] helping to stimulate public engagement with scholarly archives and manuscript transcription – always a challenge, but perhaps carrying extra significance at a time when, in the UK at least, humanities research units and related services are under intensified pressure to quantify their social impact. It will also help to open up the thought of Jeremy Bentham to new audiences.

The motives for undertaking crowd-sourcing projects in the humanities are therefore less clear-cut that those for undertaking analogous ‘citizen science’ activities. In order to achieve a useful and/or usable academic outcome, it is necessary to correctly match the processes to be undertaken to the asset in question, to select the right kind of task by which those processes will be carried out, and to be clear about what the intended output is to be.

6. Outcomes of humanities crowd-sourcing

There was extensive discussion among participants in the second workshop concerning the sort of outcomes, whether anticipated or not, that arise from humanities crowd-sourcing projects. These outcomes fall broadly into two categories: creation of knowledge and creation of communities (and this is reflected in the typology).

6.1 Creation of knowledge

The following kinds of new knowledge, or new information that can lead to new knowledge, were identified in the discussion as potential outcomes of crowd-sourcing projects:

- Everyday ephemera that would not otherwise be accessible.
- Information that would normally only be accessible to local history groups.
- Specific kinds of information that is ‘locked away’ in projects’ assets. For example, the ships’ logs used in OW are a unique record of burials at sea, as such burials are not recorded anywhere else.
- Personal histories, for example from diaries. It is possible to source data about individuals and groups that are under-represented in official records and archives; such information can contribute a quite different angle on historical research.
- Personal family links to historical processes and events (e.g. ship histories, commonwealth war graves).
• Information relating to personal or family histories can be obtained indirectly as a result of a crowd-sourcing project rather than from the assets on which the project is working, for through connections that would not otherwise have been made. For example, one OW participant received a forum private message from someone whose grandfather died on the ship Odin, and who was able to relate some of the history around it.

• Identifying links between information or between physical objects.

• Sources such as working class diaries, and other social history resources. An example of research based on such diaries could be an exploration of changes in personal mobility changes brought about by the advent of the railway, a topic otherwise unexplored in railway history.

• Providing summaries of datasets to make them more discoverable (this may be regarded as a form of descriptive metadata).

• New syntheses of existing data.

• Knowledge of how to conduct collaborative research.

• Access to knowledge and skills that other individuals have but have not shared.

• Recording knowledge before it disappears, for example records of finds made with metal detectors (e.g. via the Portable Antiquities Scheme), or of damage to the Thames foreshore. There is a perceived value among participants in plugging holes in existing datasets.

In addition, participants can also learn practical skills; these can include ICT-related skills, but also the skills required for collaboration, as crowd-sourcing gives people the opportunity to work in large groups.

A key point raised is that crowd-sourcing can improve the circulation of knowledge. The internet, and especially social media, privileges a certain kind of knowledge circulation, which stems from very simply designed and widely understood tasks: applying hashtags, commenting, ‘liking’, conveying ephemeral information in text slang, and so on. These are not necessarily the best kind of channels for the discovery and development of humanities data. Crowd-sourcing can redress this imbalance through intelligent task design. The more clearly defined a task is, the more likely the crowd will be to participate. This needs to be accompanied by some direct value that will accrue, either to the individual or to the crowd.

It was also felt strongly that knowledge created through crowd-sourcing should be available on an open access basis. There were two aspects to this. Firstly, the ability to access the final results of a project to which the volunteer has freely contributed their efforts, and to show these results to friends or colleagues, can give a great level of satisfaction to the volunteer, and forms part of a project’s reward structure. Secondly, it was considered that the results of crowd-sourcing projects should be open for ethical reasons; if they are not open, this devalues the contributions of the volunteers.

6.2 Creation of communities
One characteristic of crowd-sourcing in the humanities is that it often leads to the creation of a new community around a humanities subject, question or topic, and in turn this community can carry out some interesting or valuable work that may go beyond what was envisaged by the project. This can be done in multiple ways, by selection, self-selection, tapping into existing communities. For this to be possible, there needs to be wide, if not total, distribution of the project, usually via the internet. There needs to be a shared purpose in carrying out the activity both for the academic (or other organiser) and for the participants.

There also needs to be some element of peer review/quality control to reassure the community that their efforts create academically viable outputs; this may help to address the issue raised by Rose Holley (2010), who distinguishes between crowd-sourcing and social engagement. Quality control can be of collective outputs (e.g. quality assuring a mass produced dataset); or individual ones (e.g. keeping page images next to transcribed text to ensure correct transcription). In any case, a project should identify/specify what its quality requirements and processes are (and adhere to them), and also make the QA process evident. It probably is not necessary to aim for an “ideal” output, e.g. the comprehensive TEI markup of a text; rather crowd-sourcing can get a larger amount of material into the public domain, rather than a small amount of content in very great detail. For example, one of the shortcomings of Wikipedia is the lack of clearly-defined peer review roles, which results in a significant quantity of poorly-written and inaccurate material being published, as well as malicious or frivolous edits.

An interface enabling users to (for example) annotate/tag and suggest links without focus is not crowdsourcing; the focus on a shared task or purpose is critical. This relates to an observation that the more closely defined the task is, the more successful it will be. The more successful tasks are those where it is ‘easier not to be wrong’, and the outcome immediately verifiable or checkable. If the task is ambiguously or subjectively designed, then contributors worry about being wrong, or producing inaccurate information. This applies mainly to asset types which are empirical, e.g. dates, or weather observations: items that are either transcribed correctly or not. It does not apply to asset types which are more conjectural or creative in their nature. It depends on the purpose of the outcome type: In some cases the output is intended to digitise resources in order to answer a question (e.g weather readings to reconstruct historic weather); in others the focus can be on digitising material to make it accessible (e.g. an archive); or simply to record ephemera to make it available to posterity. In these cases, the questions are formed post hoc.

7. Typology

7.1 Introduction

This typology seeks both to bring together the research cited above in Section 2.1 and to reflect the experiences and processes uncovered elsewhere in this research review. It does not seek to provide an alternative set of categories or labels specifically for humanities crowd-sourcing; rather it recognises that there are a set of fluid and interchangeable categories within four key typological areas: asset type, process type, task type, and output type. Table A-1 fed in to its
development by demonstrating how participant-defined descriptions of crowd-sourcing activity can be mapped onto this structure.

It is the main conclusion of this research review that crowd-sourcing projects in the humanities – including the motivations of the participating communities and individuals – can best be understood by analysing them in terms of these four ‘primitive’ facets and of the relationships between them, and in particular by observing how the categories applicable in one facet are dependent on those in others.

Of course, not all projects will map straightforwardly onto single categories under of the four facets. HistoryPin (http://www.historypin.com), for example, is involved with georeferencing, images, metadata, impact, engagement and recording. While it operates outside the academic sector, it has developed strong links with the GLAM sector by providing a set of tools to allow embedding of HistoryPin content in cultural collections (see http://wearewhatwedo.org/press-releases/historypin-unleashes-new-tools). Such examples constantly challenge this typology, and provide the impetus that will guide its future evolution.

7.2 Processes

A process is a sequence of tasks (see Section 7.4), through which an output is produced by operating on an asset. It is conditioned by the kind of asset involved, and by the questions that are of interest to project stakeholders (both organisers and volunteers) and can be answered, or at least addressed, using information contained in the asset.

<table>
<thead>
<tr>
<th>PROCESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collaborative tagging</td>
</tr>
<tr>
<td>Linking</td>
</tr>
<tr>
<td>Correcting/modifying content</td>
</tr>
<tr>
<td>Transcribing⁵</td>
</tr>
<tr>
<td>Recording and creating content</td>
</tr>
<tr>
<td>Commenting, critical responses and stating preferences</td>
</tr>
<tr>
<td>Categorising</td>
</tr>
<tr>
<td>Cataloguing</td>
</tr>
<tr>
<td>Contextualisation</td>
</tr>
<tr>
<td>Mapping</td>
</tr>
<tr>
<td>Georeferencing</td>
</tr>
<tr>
<td>Translating</td>
</tr>
</tbody>
</table>

**Table 1: Process Types**

**COLLABORATIVE TAGGING**

⁵ This category also includes marked-up transcriptions, e.g. as TEI XML.
Collaborative tagging may be thought of as crowd-sourcing the organisation of information assets by allowing users to attach tags to those assets. Tags can be based on existing controlled vocabularies, but are more usually derived from free text supplied by the users themselves. Such ‘folksonomies’ are distinguished from deliberately designed knowledge organisation systems by the fact that they are self-organising, evolving and growing as the crowd adds new terms. Research has also been carried out into extracting formal data structures from folksonomies (see Lin and Davies 2010).

Collaborative tagging can result in two concrete outcomes: it can make a corpus of information assets searchable using keywords applied by the user pool, and it can highlight assets that have particular significance, as evidenced by the number of repeat tags they are accorded by the pool. Previous research in this area has examined the patterns and information that can be extracted from folksonomies. Golder (2006) found that patterns generated by collaborative tagging are, on the whole, extremely stable, meaning that minority opinions can be preserved alongside more highly replicated, and therefore mainstream, concentrations of tags. Other research (Trant, Bearman and Chun 2007; Trant 2009) has shown that user-assigned tags in museums may be quite different from vocabulary terms assigned by curators, and that relating tags to controlled vocabularies can be very problematic, although it could be argued that this allows works to be addressed from a different perspective than that of the museum’s formal documentation. In any case, such approaches to knowledge organisation are likely to play a significant part in the organisation of humanities data in the future.

A good example of this is the BBC’s YourPaintings project, developed in collaboration with the Public Catalogue Foundation, which has amassed a collection of photographs of all paintings in public ownership in the UK. The public is invited to apply tags to these, which makes them searchable by keyword. See http://www.bbc.co.uk/arts/yourpaintings/.

Case study: the Prism Project
http://www.scholarslab.org/category/praxis-program/

One key assumption underlying the process of collaborative tagging is that the assets being tagged are themselves stable and clearly identifiable as distinct objects. This has been the case for most research focusing on this as a method, e.g. tagging pictures or museum objects. However, the rich TEL/XML-based markup used in many digital humanities projects has led to exploration of users defining semantically significant areas of text and then applying tags to them.

The Prism project has developed a tool which allows readers of texts to create collective interpretations of them by a combined process of tagging and highlighting. Users can highlight a section of text and then associate the highlighted section with a tag from a controlled vocabulary. This differs from conventional text markup, in which the text is marked up by a single editor in conformance to an XML schema; rather, multiple takes on the same passage can be created and overlaid. It also
differs from other text-focused crowd-sourcing projects in which texts are processed in a single way, for example transcribing handwriting to machine-readable assets. A colour-coding system is used, making it possible to visualise different contributors’ interpretations over time. Key to the design of the project is the relationship between instructor and students; the tagging, text selection and visualisation are an appropriate means of capturing debates in a classroom, and most of the discussions about the interpretations were undertaken offline.

A cross-section of different kinds of text was selected for the first release of the project, with the aim of representing different genres, including prose, poetry and archival material. The texts are:

*The Sneetches* by Dr. Seuss
*The Raven* by Edgar Allen Poe
*Notes on the State of Virginia* by Thomas Jefferson (excerpt)
*Portrait of the Artist as a Young Man* by James Joyce (excerpt).

The project raises interesting questions of professionalism versus amateurism. Whose interpretation is being presented? It could be a professional reader (e.g. a scholar) or an amateur enthusiast. The process of generating the interpretation is the same in each case, but there is no way of knowing if an interpreter is an emeritus professor or a member of the general public. The interpretations are also of sociological interest, to determine the reception of a text by its readership; there are therefore parallels with the Year of Shakespeare project ([http://www.yearofshakespeare.com/](http://www.yearofshakespeare.com/)), where reviews are commissioned from professional reviewers, and the public are able to comment on them.

Many people like to read texts, and this forms a core motivating factor for undertaking an activity such as this. There is also likely to be an intrinsic interest among groups of researchers in synthesising their interpretations of texts of common interest. User motivation is likely to increase when, in future releases, users can upload a text to Prism and pose their own questions, for example asking for a community’s interpretation of it. Giving people this agency to frame discussion and interpretation would be important an important motivating factor, and it reflects discussion in the May workshop on the role of crowd-sourcing contributors versus research teams in the design of both projects and research questions (see [http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/09/workshop_report1.pdf](http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/09/workshop_report1.pdf)).

The project team would like to be able to capture interpretations at a particular moment in time, e.g. by providing a function that allows a snapshot of the text, highlights and tags to be downloaded, navigated and visualised. They are also interested in extending the project to support collective interpretation of other media, for example segmenting video or annotating images.
A possible research question, which could lead to ‘conventional’ research outputs, would be to investigate whether there is a relationship between different kinds of texts, their structures, their stylistics/stylometrics, and people’s interpretations of different parts of the texts. Many ‘mainstream’ humanities crowd-sourcing activities, such as transcribing texts according to well-defined procedures, have identifiable completions, whereas interpretations can go on indefinitely. Furthermore, there is no right or wrong interpretation; the questions posed are therefore very open. The broader goal of Prism is not to see whether people interpreted a text correctly or not (in contrast, say, to whether they transcribed a text correctly or not), but rather to see how, in the aggregate, people read and give meaning to a text.

Source: Interview with Jeremy Boggs, 9/8/2012

LINKING

Linking can take the form of identifying and documenting links of a specified type between individual assets, or, far more commonly, of linking via semantic tags (where in this case the tags describe binary relationships); it is thus included in the typology as a subset of collaborative tagging.

CORRECTING/MODIFYING CONTENT

While digital content is increasingly ‘born digital’, projects for digitising existing analogue material abound. Many of the technologies for digitising research content on a large scale, such as Optical Character Recognition (OCR) and speech recognition, are generally error-prone, and factoring in quality control and error correction is essential for any such enterprise. Crowd-sourcing can be used for such error correction.

The TROVE project, which produced OCR-ed scans of newspapers from 1803 onwards held in the Australian National Archives, is an excellent example of this (Holley 2009; 2011). In this case, the volume of digitised material would simply have been too great for the library to undertake the corrections using its own staff, and if only page images were produced then there would have been little or no capability for searching the text, significantly reducing the benefits of having the material in digital form at all (Holley 2009).

There have also been attempts to support crowd-sourced correction and modification of content using automation. An example of this is COoperative eNgine for Correction of ExtRacted Text, or CONCERT (Karnin et. al. 2010), which tries to match tasks to contributors’ skills, and aims to implement robust quality assurance mechanisms.

TRANSCRIPTIONING

Transcribing is closely linked to correction and modification, and is currently one of the most high-profile areas of humanities crowd-sourcing, as it addresses directly one of the most
fundamental problems with OCR: that handwriting, especially complex and/or difficult to read scripts, cannot be automatically rendered into machine-readable form using current technology. It can only be transcribed manually with the human eye and, in many cases, with human interpretation.

Two projects have contributed significantly to the prominence of transcription among crowd-sourcing projects: Old Weather and Transcribe Bentham. The latter “invites the public to play a part in academic research and attempts to break down traditional barriers” (Causer et al. 2012; see also Moyle et. al. 2012). The aim of Transcribe Bentham, funded by the AHRC during 2010/11, was to encourage volunteers to transcribe and engage with unpublished manuscripts by Jeremy Bentham (1748–1832), the philosopher and reformer, by rendering them into text marked up using TEI XML. These volunteer-transcribed manuscripts will contribute to the production of the Collected Works of Jeremy Bentham, and will be uploaded to UCL’s digital repository in order to make the collection searchable. However, as the project has stated, during this period the rate of volunteer transcription did not compare favourably with that of professional scholar-transcribers. 1009 manuscripts were transcribed in the six month testing period (8 Sept 2010 – 8 March 2011); in contrast to this, project staff estimated that the project’s research assistants could have transcribed 5000 manuscripts in a twelve month period (Causer et. al. 2012). Several explanations are given for this: that Bentham’s handwriting can be extremely difficult to decipher, and that the material is extremely complex. It has also been noted that (as of May 2012) 304 (19%) registered users transcribed material, and around two-thirds of these have worked on only one manuscript (op. cit.). There is also an extremely high moderation overhead, with significant staff time needed to process the outputs, determine whether they are good enough to be locked for further editing and, crucially, to provide feedback to the contributors.

Since then, the volunteer transcription rate has improved. As of 2 November 2012, 4,612 manuscripts had been entirely or partially transcribed, of which 94% were of sufficient quality for editorial work and uploading to the digital repository. Between 8 September 2010 and 2 November 2012, an average of 41 manuscripts (c. 20,000 words) were worked on each week; however, between 28 January 2012 to 2 November 2012, the average was 51 manuscripts (c. 25,500 words) per week, which means that collectively the volunteers are now transcribing at a faster rate than a full-time researcher, so there is potential for avoiding significant costs in the future (see Causer and Wallace, 2012)6.

An interview with a contributor to Transcribe Bentham (interview with Contributor B 10/10/2012) stressed that getting feedback from the project on what was being done correctly and what needed improvement was essential to maintaining motivation and a sense that their contribution was being valued. The project also recognises this (op. cit.). Such a sense of

---

6 The project has received additional funding from the Andrew W. Mellon Foundation for the period 1 October 2012 to 30 September 2014, and it is hoped that improving the transcription interface (e.g. by developing a WYSIWYG interface) will increase the transcription rate and reduce staff time spent validating submissions (significant effort is spent correcting XML).
community is therefore critical in a project where the transcription involves the recognition of complex information.

*Old Weather* is also based on the need to digitise content that could not be digitised by computer programme, in this case the logs of ships of the British Royal Navy, which were transcribed in order to acquire and analyse the weather observations they contain. This information is of major scientific significance, as it allows researchers to reconstruct historical weather trends in a way, and to a level of detail, that could not be extrapolated from observations of the current physical environment. About 350,000 of pages of log-books held by The National Archives, covering the period 1914-23, have been photographed and are being transcribed by 12,000 volunteers, each page being transcribed three times independently to ensure accuracy (Brohan et. al. 2009). As with Transcribe Bentham, a sense of community is extremely important in this project, with the project’s discussion forum providing a critical channel of communication, both between participants and with project staff (*Interview with Contributor A, 7/9/2012*; see [http://forum.oldweather.org/](http://forum.oldweather.org)). Exchanges on technical aspects of the transcription process are very common on the forum, and roles are assigned to contributors based on their level of experience (i.e. the number of pages transcribed).

One question posed by a user with the role ‘Newbie’ is:

*I am currently working on the HMS Mantis which is an insect class river gunboat. I have been following the logs from summer of 1918 and I am now up to summer of 1919. Everything was going well until the ship went back to England for a month's leave. Now that the logs have started back up they switched to what looks like blue ink. It is so faded that I can barely make out anything more than date and most of the weather entries. Most of the time I cannot make out the location (they are not using lat/lon location just listing a port or writing in to and from info) and I definitely cannot make out any events.*

The first response, from a highly experienced ‘Hero Member’ is:

*Welcome to the forum. Sorry I haven't got better news for you on your first visit here, but ... if you can grit your teeth for a couple of weeks of variable clarity, your logkeeper fills his pen, or gets a new bottle of ink, on 1st August.*

*I managed to get all the weather readings, which is the important part, and your eyes are bound to be younger than mine. Some of our more savvy transcribers reckon that loading the page into Photoshop or a superior version of Word will allow you to adjust contrast/background and other stuff to improve legibility, but that is beyond my abilities.*

*If you really are struggling, then abandon ship. There are others around to tax your ingenuity in different ways.*
Remember, it's acceptable for transcribing to be a challenge but don't let it become a chore.

Good Luck, whatever you decide.


Such exchanges among community members are indicative of a high degree of collaborative, communal working in addressing problems that arise during the process.

The collaborative model needed for successful crowd-sourced transcription in the humanities varies according to the complexity of the material to be transcribed. Complex material, such as the two cases cited here, requires a high level of support, whether from a project team or from the participant’s peers in a forum; or more usually a combination of the two. Less complex material is likely to require less support, or at least less support in the form of contact with other users or project staff. For example, when transcribing the more structured data found in family records (e.g. http://www.familysearch.org), the information (text or integers) to be transcribed is presented to the user in small segments – names, dates, addresses etc. on birth certificates – and it requires different cognitive processes, which in turn are less dependent on the mode of community represented by feedback and engagement from ‘experts’.

**RECORDING AND CREATING CONTENT**

Typically processes in this category are concerned with recording ephemera and intangible culture, such as oral history or reminiscence. These frequently take the form of a GLAM or other cultural institution soliciting memories from the communities it serves, for example the Tenbury Wells Regal Cinema’s Memory Reel project (http://www.regaltenbury.org.uk/memory-reel/). Such processes can incorporate a form of editorial control or post hoc digital curation, and their outputs can be edited into more formal publications, or analysed/explored using methods such as sentiment analysis (see below).

This category is more likely to conform to the ‘social engagement’ model, in terms of Holley’s distinction of social engagement from crowd-sourcing (2010).

**COMMENTING, CRITICAL RESPONSES AND STATING PREFERENCES**

These processes are only likely to fall within the definitions of crowd-sourcing as set out in this review if there is some specific purpose for which people come together. One example of this is capturing audience response to the 2012 World Shakespeare Festival, which is the aim of the Shakespeare’s Global Communities project (www.yearofshakespeare.com). A key question of this project was ‘How do new social networking technologies reshape the ways in which diverse global communities connect with one another around a figure such as Shakespeare?’ (http://crowds.cerch.kcl.ac.uk/wp-uploads/2012/09/workshop_report1.pdf). The question itself provides a focus for the activity, and although in and of itself it does not produce a verifiable
academic output, it provides a dataset for looking at research questions on the modern reception of Shakespeare. In some cases, appropriately presented blogging software provides a platform for focused scholarly interaction. For example, a review by Sonia Massia from King’s College London of *King Lear* on the Year of Shakespeare site attracted controversial responses, leading to an exchange about critical methods as well as content⁷. What differentiates such exchanges from amateur blogging is the scholarly focus and context provided by the project, and its proactive directing of the content creation. The project thus provides a tangible link between the crowd and the subject.

There is much potential in mining corpora of comments and critical responses, using techniques such as sentiment analysis (e.g. Chmiel et. al. 2007), which attaches quantitative weights to the positivity or negativity of units of text, such as tweets or commentaries, and allows collective analysis of these.

**CATEGORISING AND**

Categorising involves the placing of assets into predefined categories; it differs from collaborative tagging in that the latter is unconstrained.

**CATALOGUING**

Cataloguing – or, more expansively, the creation of structured, descriptive metadata (e.g. for cultural objects) – is a more open-ended process than categorising, but it is nevertheless constrained to follow accepted metadata standards and approaches. It frequently includes categorising as a sub-activity, e.g. according to LoC subject headings.

Cataloguing is a particularly time- and resource-consuming process for many GLAM institutions, and crowd-sourcing this activity has been explored as a cost-effective means of increasing access. For example the *What’s the Score* project at the Bodleian Library, whose principal aim is to investigate a cost-effective approach to increasing access to music scores from the Bodleian’s collections, to be achieved by a combination of rapid digitisation and the crowd-sourcing of descriptive metadata (see the crowd-sourcing site at [http://www.whats-the-score.org](http://www.whats-the-score.org), and the delivery site at [http://scores.bodleian.ox.ac.uk](http://scores.bodleian.ox.ac.uk)).

The cataloguing process type is linked to contextualisation, as ordering, arraying and describing assets will also make explicit some of their context.

**CONTEXTUALIZING**

Contextualization is typically a more broadly-conceived activity. It involves enriching an asset of a particular type by adding to it or associating with it other relevant information or content.

**MAPPING**

In the sense in which it is used here, *mapping* refers to the process of creating a spatial representation of some information asset(s). This could involve the creation of map data from scratch, as in the OpenStreetMap initiative ([http://www.openstreetmap.org/](http://www.openstreetmap.org/)), but it could also be applied to the visual or spatial mapping of concepts (as in a ‘mind map’). The precise sense will be highly dependent on the *asset type* to which mapping is being applied.

Note that *mapping* should not be confused with *georeferencing*.

**GEOREFERENCING**

Georeferencing refers to the process of establishing the location of un-referenced geographical information in terms of a modern real-world coordinate system (such as latitude and longitude). Georeferencing can be used to enrich significantly geographical datasets that do not include such information, but which could or should do, and there has been significant activity in this area in terms of crowd-sourcing and user engagement.

**Case study: the British Library Georeferencer (BLG) Project**

Georeferencing has been used in the UK in the BLG project (see also Section 3). BLG is based on the Georeferencer application, along with the Old Maps Online and TEMAP projects. The tool allows ‘free, crowd-sourced collaborative online georeferencing of map images from a number of libraries’, and has been applied to four other map collections in addition to those of the British Library itself (Fleet et. al. forthcoming). The purpose of its application by the British Library was to ‘geo-enable’ historic maps by asking participants to assign spatial coordinates to digitised images of maps. This task would have been too labour-intensive for BL staff to undertake themselves, so it was exposed to crowd-sourcing. Once digitised and georeferenced, the maps can be viewed using online geographic technologies, and are geographically searchable due to the inclusion of latitude and longitude coordinates in the metadata.

For participants, there was an element of instant gratification as they could see the results of their work immediately. The project had a ‘citizens’ forum’ tab, which proved important for generating a sense of community among the participants. 725 maps were georeferenced between 13 and 18 February 2012 by around 90 participants, with very high data quality. Social media was considered to be a key factor in this astonishing success, as also was the institutional association with British Library, and the credibility that this gave to the processes and outcomes.

BLG therefore corresponds to the output type of *metadata* (see below) and, as no new map data was created, it is distinct from the process type of *mapping*. However, the researchers behind BLG have noted that
‘while the advantages that online georeferencing offers to cartographic collections are considerable, these very specialised benefits may be eclipsed by the broader ability to expose and share collections with the public in a new and much more engaging way than was before possible”. (Fleet et. al. forthcoming).

It is therefore evident that, as well as metadata, the more qualitative output types of engagement and knowledge/awareness also apply.

**TRANSLATING**

This process type covers the translation of content from one language to another. In many cases, a crowd-sourced translation will require a strongly collaborative element if it is to be successful, given the semantic interconnections and interdependencies between that can occur between different parts of a text. However, in cases where a large text can be broken up naturally into smaller pieces, a more independent mode of work may be possible. An example of this is the Suda On-Line project ([http://www.stoa.org/sol/](http://www.stoa.org/sol/)), which is (among other things) translating the entries in a 10th Century Byzantine lexicon/encyclopaedia. A more modern, although non-academic, example is provided by the phenomenon of ‘fansubbing’, where enthusiasts provide subtitles for television shows and other audiovisual material (Cintas 2006).

**7.3 Asset type**

The Asset in a crowd-sourcing project refers to the kind of content which is, in some way, transformed as a result of exposure to the crowd.

<table>
<thead>
<tr>
<th>ASSET</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geospatial</td>
</tr>
<tr>
<td>Text</td>
</tr>
<tr>
<td>Numerical or statistical information</td>
</tr>
<tr>
<td>Sound</td>
</tr>
<tr>
<td>Image</td>
</tr>
<tr>
<td>Video</td>
</tr>
<tr>
<td>Ephemera and intangible cultural heritage (e.g. oral history)</td>
</tr>
</tbody>
</table>

**Table 2: Asset Types**

**GEOSPATIAL**

A geospatial asset is a dataset or text that includes components that identify or refer to locations on the earth’s surface. A map is an obvious example, but the category also covers assets such as travelogues or gazetteers. The asset may refer to location absolutely (e.g. by using coordinates that refer to known locations on the earth’s surface), or relatively (e.g. by referring to an identifiable location or feature, such as ‘Birmingham, England’ or ‘the River Severn’), or
indeterminately, referring to spatial features or types of features in a way that is not transferrable directly to the earth’s surface (e.g. ambiguous place-names, or a reference to ‘rivers’).

There is an important distinction to be made in this context between geospatial assets created by expert organisations, such as the Ordnance Survey or the US Geological Survey, and those created by community-based initiatives. The former may have the authority of a governmental imprimatur, and the distinction of being officially endorsed for a variety of reasons (such as emergency response or guiding planning processes). However, the recent emergence of crowd-sourced geospatial assets – a product of the recent global growth in the ownership of hand-held devices with the ability to record location using GPS (Goodchild 2007) – has led to the emergence of resources such as Open Street Map, which in turn has led to a discussion about the reliability of such resources. In general, it has been found that Open Street Map in particular is extremely reliable (Haklay and Weber 2008, Haklay 2010), but that the specifications for such resources must be carefully defined (Brando and Bucher 2010). The impact of Open Street Map on the cartographic community generally has been noted (Chilton 2009). The importance of mapping as a means of convening spatial significance means that this kind of asset is particularly open to different discourses, and possibly conflicting narratives. The digital realm, with the potential for many users to contribute many kinds of information, holds great potential for this area, but the issues raised are still emergent (see Fink 2011, Graham 2010).

**TEXT**

Our review suggests that history, especially social history, is currently by far the most common subject area for crowd-sourcing projects in the humanities, and consequently text is by far the most common asset type to be engaged with by humanities crowd-sourcing projects.

It should be noted that, as asset type describes the material that is transformed by crowd-sourcing, the text category refers to text that already exists, rather than the creation of new text, for example by recording personal experiences (this would fall under the asset type ephemera and intangible cultural heritage). This category of asset is likely to result in one or more of the output types transcribed text, corrected text or enhanced text.

**NUMERICAL OR STATISTICAL INFORMATION**

In humanities crowd-sourcing, the line between text and statistical/numerical information is not always clear, not only because source documents for statistical/numerical information frequently contain text as well, but also in terms of the organisation and operation of a project. Old Weather is an example of this. The logbooks contain both numerical information (e.g. weather observations) and free text, and the stated aim of the project was to extract the weather data for subsequent machine processing, yet much textual data has also been transcribed at the same time, and used for synthesis and interpretation of the histories of individual ships.

**SOUND**
The principal application to date of crowd-sourcing involving this asset type has been the gathering of audio recordings for various purposes. The British Library’s Sound Map, for example, has gathered examples of regional accents and dialects, wildlife sounds, environmental sounds, and other ‘soundscapes’ from around the UK (http://sounds.bl.uk/sound-maps/).

An emergent aspect of this area is the use of software tools to correct transcriptions of audio recordings. Automated transcription of recorded speech is currently highly error-prone, with error rates of 30% or more (Wald 2011). Application of crowd-sourcing techniques for processes of correcting/modifyng content of recorded speech via transcription is likely to be very useful for a variety of humanities-oriented projects in the future.

**IMAGE**

Images are important asset type for humanities crowd-sourcing, and most current applications are concerned with the production of metadata for the purposes of enhancing search and discovery.

One key example is the Flickr Commons project, a collaboration between Flickr and the Library of Congress (Springer et. al. 2008) launched in 2008, which allows participating cultural heritage institutions to expose their images (if they have "no known copyright restrictions") to the Web via Flickr, and users of Flickr then to tag them. This can improve access to and knowledge of these collections, and also allows the public to contribute aggregated knowledge. This can have unexpected extrinsic benefits; for example, in our interview with Contributor C, it emerged that he had tagged photos from the Netherlands National Library on Flickr Commons as part of his efforts to learn Dutch.

Another example is the YourPaintings project, developed by the Public Catalogue Foundation for the BBC (http://www.bbc.co.uk/arts/yourpaintings). This project has put online images of 200,000 oil paintings in public ownership in the UK, and exposed this super-collection to the public for them to tag, again to increase searchability and accessibility.

It should be noted that the ubiquity of the Web, and access to content creation and digitisation technologies, has led to the creation of non-professionally curated online archives, which are independent of conventional library and archive-derived curatorial narratives. These have a clear role to play in enriching and augmenting collections produced by ‘memory institutions’, and distributed via collaborative platforms such as Flickr Commons. As Melissa Terras has noted:

Enthusiastic digitization by amateurs, a phenomenon previously ignored by information professionals, is providing a rich source of online cultural heritage content which often documents areas not covered via traditional institutions. Indeed, ephemera and popular culture materials are often better served by the pro-amateur community than memory institutions. The energy and zeal displayed by amateur digitizers is worthy of further consideration, as amateur collections often
complement existing collections, providing an alternative free discussion space for enthusiasts (Terras 2010).

**VIDEO**

Video as an asset type has to date demonstrated its potential for crowd-sourcing to a lesser degree than other kinds listed here.

However, YouTube has brought about a revolution in the communication of video media, and comments on YouTube videos provide an extensive and diverse source for the opinion and attitudes of (a subset of) the public. These comments have proved to be a rich seam for network analysis, where links between comments sharing the same characteristics can be mapped, and for sentiment analysis (Thelwall et. al. 2012), where the positivity or negativity of comments’ sentiments can be quantitatively measured and analysed (op. cit.). It is also possible to discern the occurrence of discrete events depicted in the comment streams (Steiner et. al. 2011).

In these examples the content created by the crowd provides raw material for the academic research, but its creation does not itself form part of an academic crowd-sourcing activity. However, it suggests that one avenue for video as an object of crowd-sourcing is afforded by the processes of (a) **mapping** and (b) **commenting**, **critical responses and stating preferences**, and that this will become particularly interesting when crowds interact around them.

Studies in this area have highlighted the positive connection between the attention the crowd gives to an individual user’s videos, and the productivity of that user, as measured by the numbers of videos they upload (Huberman et. al. 2009). For such media to be of significant interest in humanities crowd-sourcing therefore, it is likely that they will have to be closely associated with human interaction, especially under the **Commenting, critical responses and stating preferences** process type.

**EPHEMERA AND INTANGIBLE CULTURAL HERITAGE (E.G. ORAL HISTORY)**

Ephemera and intangible cultural heritage form potentially the most productive category of asset for humanities crowd-sourcing, and yet also the most underdeveloped from a methodological point of view. Intangible cultural heritage covers any cultural manifestation that does not exist in tangible form; typically, crowd-sourcing is used to document such heritage through a set of processes and tasks, resulting in some form of tangible output. The importance of preserving intangible cultural heritage has been recognised by the UN (Kurin 2004), and the ways in which this can be documented and curated by distributed groups – indeed by ‘the crowd’ – is an important area for future research.

By emphemera, we understand cultural objects that are tangible, but are at risk of loss because of their transitory nature, for example home videos or personal photographs⁸. There are a number of project addressing such assets, for example the Europeana 1914-1918 project

---

⁸ This usage differs from the standard usage of the term by museums.
The Scottish Words and Place-names project was a JISC-funded activity to gather words in Scots and, ultimately, to offer selected words for inclusion in the word collection of the Scottish Language Dictionaries resource. Scots is not a distinct language in the way that, for example, Gaelic is, but it has a community of people who believe strongly in its cultural and linguistic importance. The aim of the SWAP project was to see which Scots words are in current use, and where/how they are used.

The interface was relatively straightforward: a user could enter a word into a box on the website, and this was then harvested into a backend database. There was no moderation or checking done at this point, but words that the project felt were suitable were passed to the DSL (Dictionary of the Scots Language), where they were scrutinised by expert lexicographers.

Words were also gathered via Facebook and (to a lesser extent) Twitter. The Facebook page was an important venue for developing conversations around the forms and meanings of individual words. Contributions via Facebook were less formal and structured, taking the form of discussions with project staff in comment threads. This required more of an overhead in the time spent processing and extracting them. However, it generated valuable material, and allowed the project to target specific areas and questions that were of interest, such as local dialects and rhyming slang. For example:

“Thanks for all your place-name, alcohol and miscellaneous Scots info so far! This week we'd like to know about your local dialect rhyming slang. Some examples we already know about are: mammy mine - wine (Glasgow), Mick Jagger - lager (Gl), Lee Van Cleef) - deif (Edinburgh), broon breid - deid (Ed & Gl), Mars bar - scar (Gl), Oscar (Slater) - later (Ed). Do you use these? Do you know any others? Please let us know!” (see http://www.facebook.com/scotswap/posts/230076623689593).

This generated responses such as:

“I remember Caroline Macafee's book on Glasgow: Language Varieties around the World collected some of these: askits (= shoulders) askit pooders, shooders); benny (lynch, cinch), etc. I never heard them used, though my godmother in Ayur [sic] still addresses me as 'china.'”

This highlights the importance of active engagement with (potentially small) interest communities for gathering contemporary linguistic content. In general, the project found that they received more useful information when they asked about specific words rather than framing open questions. For the same reason, acquiring information about specific
places was easier, as these are associated with a more concrete sense of belonging: ‘crowdsourcing informal place-names was relatively successful, perhaps because of a greater sense of involvement from participants, and social media does show good potential for further research of this sort’ (Hough et. al. 2011).

Information on Scottish place-names was also gathered by the project. Information about a Scots place-name element (e.g. 'liggat' meaning gate) would be provided from the Glossary of Scots Place-name Elements, with examples of place-names using this element, then other examples would be solicited from the public. These could then be added to the examples field of the Glossary (http://swap.nesc.gla.ac.uk/database).

The project was extremely successful in building up a community of followers and contributors among schoolchildren and teachers. This is largely because it was able to utilise and capitalise on the GLOW intranet network which connects Scottish schools. There was a slight potential ethical tension between the success of using Facebook to generate discussions, and the success of generating networks using GLOW: one is not allowed to have a Facebook account if under the age of 13. In June 2012 the project ran a competition for schoolchildren, with judges including Louise Welsh. This was again a highly successful way of engaging the community, and furthering the project’s aims of both capturing and encouraging the use of Scots.

For more information on the project, see the website at (http://swap.nesc.gla.ac.uk/), and especially the final JISC report (Hough et. al. 2011).

Source: interview with Ellen Bramwell and Jean Anderson, (14/09/2012)

7.4: Task type

A task is some particular activity that a project participant undertakes in order to create, process or modify an asset (usually a digital asset). Tasks can differ significantly as regards the extent to which they require initiative and/or independent analysis on the part of the participant, and the difficulty with which they can be quantified or documented. We have identified a number of task types with the aim of categorising this complexity.

These task types are identified in Table 3, and are ordered by the extent to which they require such independent analysis from the participant. Note that this refers only to what is required of the participant – a project requiring tasks that are in themselves mechanical will typically require a great deal of structure and planning on the part of the project team. Conversely, a task that is complex from the participant’s point of view may require less structure and input from the project team. Such a categorisation is not, of course, any reflection of the task’s value.
Superficially similar tasks can correspond to quite different task types, depending on the nature of the asset involved, and the requirements of the project; for example, the tasks involved in transcribing a section of text could be either mechanical or editorial.

<table>
<thead>
<tr>
<th>TASK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mechanical</td>
</tr>
<tr>
<td>Configurational</td>
</tr>
<tr>
<td>Editorial</td>
</tr>
<tr>
<td>Synthetic</td>
</tr>
<tr>
<td>Investigative</td>
</tr>
<tr>
<td>Creative</td>
</tr>
</tbody>
</table>

**Table 3: Task Types**

**MECHANICAL**

Processing discrete and granular units of information (typically quantitative/numerical/statistical information, or very short pieces of text) from one form into another, for example transcribing ages and dates of birth from birth certificates. Such tasks require little or no initiative. The FamilySearch site provides an example of this type of task; users simply transcribe numerical data from images of records into a web form (see [https://familysearch.org](https://familysearch.org)).

**CONFIGURATIONAL**

This category covers tasks that involve identifying structural patterns or ‘configurations’ in information, rather than processing individual pieces of information. Some such tasks will require a predisposition for working with quantitative data. The Bostonography project provides examples of such tasks (see [http://googlemapsmania.blogspot.co.uk/2012/07/crowdsourcing-neighborhood-boundaries.html?spref=tw](http://googlemapsmania.blogspot.co.uk/2012/07/crowdsourcing-neighborhood-boundaries.html?spref=tw)).

**EDITORIAL**

Modifying, or improving an existing asset or assets. This requires initiative, depending on the nature of the task and on the standards or procedures laid down by the project or activity. Feedback and interaction with peers is typically required, and also acts as a significant encouraging factor. Wikipedia and Wikisource are examples of editorial crowd-sourcing tasks.

**SYNTHETIC**

Synthesising information from different sources. This requires initiative and analytical ability, and is also likely to require interaction with peers. Carrying out tasks of this type may also lead to the development of expertise in some area of the humanities. Bringing together information on individual ships (e.g. information obtained in the OldWeather project) for the Naval History site ([http://www.naval-history.net](http://www.naval-history.net)) is an example of a synthetic task.
INVESTIGATIVE

Includes elements of the configuration and synthetic types, but more specifically involves seeking out information that is latent or implicit in corpora or datasets, but which may not be immediately apparent. The Guardian’s 2009 experiment in crowd-sourcing documents relevant to the UK MPs’ expenses scandal provides examples of this, where members of the public identified documents in the corpus that were interesting enough to be investigated further (see http://www.niemanlab.org/2009/06/four-crowdsourcing-lessons-from-the-guardians-spectacular-expenses-scandal-experiment).

CREATIVE

Creation of new content, as for example in the Strandlines project (www.strandlines.net).

6.5: Output type

The output type refers to the thing an activity produces as the result of the application of a process, using tasks of a particular task type, to an asset. These outputs can be tangible and/or measurable in various ways, but we make allowance here also for intangible outcomes, such as awareness, knowledge etc. Our interviews and discussions with super-contributors indicated that these can be particularly important for the participants, although of course they do not lend themselves to quantification or measurement.

ORIGINAL TEXT

Text that is created as the result of a project or activity, and did not exist prior to the project or activity.

TRANSCRIBED TEXT

Text, almost always digital and machine-readable, that is created by processing other text that is not machine-readable. The transcribed text will have little or no semantic enhancement.

CORRECTED TEXT

Text that has been modified only to the extent of correcting errors in the input asset by manual intervention. Typically, this will involve a combination of mechanical and editorial task types, in a process of type correcting/modifying content.

ENHANCED TEXT

Text that has been semantically enhanced, for example by marking it up using TEI/XML.
TRANSCRIBED MUSIC

Music (in the form of musical notation, not audio) that has been transcribed into digital form and made retrievable from a digital system.

METADATA

Data about an asset, usually created to make it more discoverable, retrievable, curatable or usable, or generally to expose it via the Web more effectively. This category includes both descriptive metadata and preservation metadata, although most examples highlighted by our review are of the former.

STRUCTURED DATA

Data whose structure has been improved and/or made more explicit, for example by exposure to processes such as collaborative tagging or linking.

KNOWLEDGE/AWARENESS

Increased knowledge of a subject (including practical skills), or increased awareness of a project or topic, in the wider community.

FUNDING

Money that has been raised from open, distributed, voluntary sources. Frequently, such ‘crowd-funding’ also engenders knowledge/awareness, not only among the donors but also more generally through the ensuing publicity.

It has been shown that location is not always a significant factor in the ability of projects to attract money in this way (Agrawal et al. 2011); however in the humanities, this phenomenon has been particularly associated with participation by funding in archaeological digs (see, e.g. http://www.bbc.co.uk/news/science-environment-19192220), and in such cases there is an obvious relationship between the participant and a particular location.

SYNTHESIS

A collection of pre-existing information that has been enriched and expanded by being combined, compared and developed through editorial or structural intervention.

COMPOSITE DIGITAL COLLECTIONS WITH MULTIPLE MEANINGS

The use of interoperable standards for markup, e.g. TEI (for text) and KML (for mapping), underpins many of the asset types identified here, and is essential for their subsequent searching and exploration. However, these standards are extrinsic to the crowd-sourcing
process, and have been separately developed for the purpose of organising and standardising data.

On the other hand, the crowd-sourced collection of ordered, but largely unstructured, information such as Wikipedia can also provide a means for navigating and describing content, in addition to the content itself. In Wikipedia’s case this is DBpedia (http://dbpedia.org/About), a linked data hub that identifies the subjects discussed in Wikipedia as unique entities, and allows these to be associated with external collections. Such complex information structures as the combined Wikipedia/DBpedia have their own value, and are thus defined here as a distinct asset type. In most cases, such outputs will include components of both original/enhanced text and structured data.

This type of output sits in the context of the collaborative creation of digital objects that the Web 2.0 internet encourages. For example, the British Museum’s Wikipedia project sought to assist the creation of articles dealing with all of the Library’s most significant objects and collections. The question of how objects should be described comes up again and again in the (digital) humanities, be it XML markup, textual chunks, museum schemata for objects such as the CIDOC CRM (http://www.cidoc-crm.org) etc. One key objective of humanities crowd-sourcing, although expressed differently by different projects, is therefore to leverage this increasingly sophisticated transformation of humanities content into digital objects and to promote public interaction with it.

Case study: Wikisource

Wikisource is a sister project of Wikipedia (http://wikisource.org/) based, like Wikipedia, on the MediaWiki platform. It aims to create a crowd-sourced, wiki-based library of digital texts. Scans of texts from Google Books and the Internet Archive are read by participants, and transcribed into machine-readable form, and edited by the community. There are no strictures on what text or texts members of the Wikisource community work on; thus the subject interest of individuals is easily catered for. There are relatively few ‘power users’ who do a great deal of work, and the Wikisource community in individual countries, such as Italy, is relatively small. A key source of traffic and interest is the fact that Wikisource resources are embedded in Wikipedia pages – e.g. the ‘rilettura del mese’ (rereading of the month) in the bottom left of the page at http://it.wikipedia.org/wiki/Pagina_principale – which allows the project to ‘exploit the power of the Wikipedia user base’.

A ranking is applied to each unedited page of each resource, with an orange square indicating that the page is formatted but not proofread, and a green one indicating that it is in a publishable condition, both corrected and validated. Wikisource provides a set of templates for structuring and ordering texts according to good editorial practice. It is noted however that it can be complex for users without programming knowledge or the ability to manipulate the code underlying the
templates. Although it has the same kind of cross-fertilising community functionality as the Zooniverse projects, it is not as easy to use as the Zooniverse interface.

**Source: interview with Andrea Zanni, 22/10/2012**

Such work is contextualised by the findings of Kittur et al. (2007), who note that, while platforms such as Wikipedia generally start by being driven by large numbers of ‘super users’, both in terms of content created and of edits made, over time much of the content creation and modification is taken over by larger numbers of users with fewer time and effort commitments. This suggests that project-based funding models, which are inherently time limited, do not lend themselves to creating successful composite crowd-sourced archives of information.

### 8. Conclusion

The overall conclusion of this research review is that research involving humanities crowd-sourcing can be divided into the four facets of process, asset, task type and output type. Any robust set of replicable methodologies for creating or processing information by or for humanistic scholarship must be framed in these terms.

Depending on the project or activity, and what it aims to do, some categories, or indeed some facets, will have primacy. Outputs might be original knowledge, or they might be more ephemeral and difficult to identify: however, considering the processes of both knowledge and resource creation as comprising of these four facets gives a meaningful context to every piece of research, publication and activity we have uncovered in the course of this review. We hope the lessons and good practice we have identified here will, along with this typology, contribute to the development of new kinds of humanities crowd-sourcing in the future.

Most significantly, we have found that most humanities scholars who have used crowd-sourcing in its various forms now agree that it is not simply a form of cheap labour for the creation or digitization of content; indeed in a cost-benefit sense it does not always compare well with more conventional means of digitization and processing. In this sense, it has truly left its roots, as defined by Howe (2006) behind. The creativity, enthusiasm and alternative foci that communities outside that academy can bring to academic projects is a resource which is now ripe for tapping in to, and the examples shown in this report illustrate the rich variety of forms that tapping can take.

### Acknowledgements

We are grateful to all those who have shared their knowledge and experience with us: Rebekkah Abraham, Anne Alexander, Donna Alexander, Jean Anderson, Keith Ball, Richard Blakemore, Jeremy Boggis, Ellen Bramwell, Phili Brohan, Geoff Browell, Ben Brumfield, Tim Causer, Sarah Cornell, David De Roure, Alison Dickens, Alexandra Eveleigh, Chris Fleet, Abigail Gilmore, Andrew Gray, Colin Greenstreet, Martin Holmes, Helen Julian, Kimberly Kowal, Sam Leon, Janet Lomas, Susan Major, Paola Marchionni, Anthony Masinton, Fabrizio Nevola,

References cited


Trant, J. 2009: *Tagging, Folksonomy, and Art Museums: Results of steve.museum’s research*. (online at http://conference.archimuse.com/blog/jtrant/stevemuseum_research_report_available_tagging_f o)


Appendix A: Survey

The survey (approved by King’s College London Research Ethics Committee, project number REP-H/11/12-19; see Appendix A) was open to all, and was available on the project website at http://www.humanitiescrowds.org/survey from May 21st until October 30th 2012. Although the survey was publicised on various mailing lists in humanities communities, and was promoted to project participants by colleagues running humanities crowd-sourcing projects (especially those projects represented at the May workshop), it was completely self-selecting, and is therefore likely to have attracted only those with significant interests in crowd-sourcing, either ‘super contributors’ or people helping to run projects. It thus makes no claim to being statistically representative; it is purely a means of gathering further qualitative information about contributors’ backgrounds and motivations.

The survey was conducted online using SurveyMonkey, and included the following questions and possible responses. Where no response options are indicated, free text was allowed.

1. Are you male or female?
   Male, Female
2. Which category below includes your age?
   17 or younger, 18-25, 25-35, 35-45, 45-55, 55-60, 60+
3. Which of these best describes you?
   At school, Student, In work, Self-employed, Unemployed, Pensioner, Retired
4. Your nationality
5. Your current location (country, city)
6. Do you currently consider yourself to be an active contributor to a crowd-sourcing project or projects?
7. When did you last work on a crowd-sourcing project?
   In the last three days, In the last week, In the last month, In the last year
8. Which crowd-sourcing project(s) have you worked with?
9. How did you find out about the project(s) you have worked with?
10. What is (or was) the nature of your work with the project(s)?
Transcribing, Classifying, Proofreading, Tagging, Commenting

11. Roughly how many hours a week do you (or did you) spend contributing?

Less than 1 hour, 1 to 4 hours, 5 to 10 hours, 10 to 20 hours, 15 to 20 hours, More than 20 hours

12. Typically how long does a single session last?

Less than 1 hour, 1 to 3 hours, More than 3 hours

13. Do you see your work with the project(s) as a solitary activity, or do you discuss it with others?

14. Do you use any social networks to discuss, disseminate or help your crowd-sourcing work?

15. Please describe what first interested you in working with crowd-sourcing projects.

16. What motivates you to do crowd-sourcing work?

17. Do you see your contribution as being for your own interest, or for the benefit of others?

18. What would increase your motivation to keep doing crowd-sourcing work, or encourage you to do more?

19. What would cause you to stop contributing?

Completion of project, Lack of time, Loss of interest, Other (please specify)

20. Have you ever created or edited a Wikipedia article?

59 responses were received. 58% of respondents were male, 42% female. Most were in the 35-45 age bracket (see Fig. A-2). A significant majority (59.6%) were in work, with students (22.8%) and retirees (15.8%) being the next largest categories (see Fig. A-3). Unsurprisingly, most of the respondents (25) were from the UK, followed by Italy (11), the USA (9), Ireland (2), and individual participants from Malta, Greece, Germany, Canada, Spain and Latvia. 80% were currently active as crowd-sourcing participants, with most having been active within the week before they completed the survey (see Fig. A-4).
Figure A-1: Age of respondents

Figure A-2: Employment status of respondents
Figure A-3: Time since last crowd-sourcing work

Figure A-4: Nature of crowd-sourcing work undertaken
Respondents were asked (Question 10) what kind of crowd-sourcing work they had undertaken, with options drawn from the early stages of our desk research. These categories were subsequently refined, but at that stage they were: transcribing, classifying, proofreading, tagging and commenting. These were incorporated into the typology described in Section 7 under the process type facet, with some modifications, e.g. proofreading was subsumed into Correcting/modifying content (see, e.g. Newby and Franks 2003). We also allowed participants to specify other kinds of activities, which identified some important lacunae in the typology, such as georeferencing and creating, which were subsequently incorporated. Table A-1 lists these additional responses, along with the process type to which they correspond in the typology. In some cases the fit is better than in others, and in all cases the process type needs to be qualified by the asset and task type involved (see Section 7).

<table>
<thead>
<tr>
<th>Response to Question 10</th>
<th>process type (from typology)</th>
</tr>
</thead>
<tbody>
<tr>
<td>data gathering and management</td>
<td>Linking, categorising and cataloguing</td>
</tr>
<tr>
<td>Research, online and library work.</td>
<td>Linking, categorising and cataloguing</td>
</tr>
<tr>
<td>OCR error-correction</td>
<td>Correcting/modifying content</td>
</tr>
<tr>
<td>aligning/reorienting images</td>
<td>Correcting/modifying content</td>
</tr>
<tr>
<td>Translating</td>
<td>Translating</td>
</tr>
<tr>
<td>Contributing Content (wikis)</td>
<td>Recording and creating content</td>
</tr>
<tr>
<td>Adding photos to iSpot. Writing for wikibooks.</td>
<td>Recording and creating content, linking</td>
</tr>
<tr>
<td>Editing, adding information</td>
<td>Correcting/modifying content</td>
</tr>
<tr>
<td>Technical work (Javascript etc.)</td>
<td>Recording and creating content</td>
</tr>
<tr>
<td>Sysop</td>
<td>Recording and creating content</td>
</tr>
<tr>
<td>Translating captions on flickr images</td>
<td>Translating</td>
</tr>
<tr>
<td>Creating</td>
<td>Recording and creating content</td>
</tr>
<tr>
<td>developing the infrastructure (technical, policies, etc.) of the whole project</td>
<td>Recording and creating content</td>
</tr>
<tr>
<td>Georeferencing - placing historic maps over current maps</td>
<td>Georeferencing</td>
</tr>
<tr>
<td>Georeferencing</td>
<td>Georeferencing</td>
</tr>
<tr>
<td>Georeferencing (BL) and OCR correction (Dickens)</td>
<td>Georeferencing</td>
</tr>
</tbody>
</table>

Table A-1: Nature of crowd-sourcing work undertaken

A majority of respondents (52.4%) stated that they spent less than an hour on an individual crowd-sourcing session, and 42.9% saying they spent between one and three hours. Only 4.8% of respondents who answered this question said they spent more than three hours, which possibly reflects the fact that most of the respondents were in work, and thus probably had less free time (many, but not all, of the super-contributors at the second workshop were not currently working for one reason or another), although it may also be the case that some tasks are not sufficiently absorbing to attract effort for very extended periods of time.

The question “Do you see your work with the project(s) as a solitary activity, or do you discuss it with others?” was left open, but it is important as a measure of the sense of community within
crowd-sourcing. 17 responses can be classified as indicating that crowd-sourcing is a ‘solitary’ activity, whereas 23 felt that it was not. Many responses however raised more complex issues and could not be classified so simply. Some indicated that they discussed the project with others, for example ‘Solitary effort, but contributing to a group end. Discussed with others (family / friends) as it was interesting activity, but did not discuss with others in crowdsourcing project.’ Others, reflecting comments made in the second workshop, highlighted the importance of project discussion fora: ‘[I had discussions] only on discussion pages of projects if have a problem and need a second opinion or help’. There was also reference to a sense of community arising indirectly, rather than through direct contact, for example through the visible achievement of common aims: ‘It was solitary but being able to view the stats. showing other participants progress enhanced the teamwork aspect and also added an element of competition’.

The responses to Question 15, which asked what first interested the respondent in contributing to crowd-sourcing, were analysed by classifying them under a number of headings, as shown in Table A-2. This analysis confirms that subject interest plays the greatest part in recruiting super-contributors, and that other motivations are largely altruistic in character.

<table>
<thead>
<tr>
<th>Heading</th>
<th>No. of Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject interest</td>
<td>24</td>
</tr>
<tr>
<td>Helping others learn</td>
<td>3</td>
</tr>
<tr>
<td>Want to contribute to science</td>
<td>2</td>
</tr>
<tr>
<td>Want to know how crowd-sourcing works</td>
<td>2</td>
</tr>
<tr>
<td>Wanting to be involved in voluntary work</td>
<td>1</td>
</tr>
<tr>
<td>Novelty</td>
<td>1</td>
</tr>
<tr>
<td>Unused CPU power might as well be used</td>
<td>1</td>
</tr>
</tbody>
</table>

Table A-2: reason for initial interest in crowd-sourcing

Question 16 asked respondents directly what motivates them to participate in crowd-sourcing work; the responses are listed in Table A-3 (note that spelling, etc. are as in the original responses). The responses indicate (a) that subject interest is paramount, but (b) that despite this, multiple motivations are usually in play, often involving a desire to contribute to a larger whole\(^9\).

| Continuing to make a contribution to science     |
| The possibility to use the data we create        |
| It is useful and valuable                        |
| I believe that knowledge is part of the common heritage of humanity. |
| Feel it's worthwhile and it's easy to do         |

\(^9\) See also (Causer and Wallace, 2012).
Small individual effort has significant group benefits. I was also interested in the subject matter (Astronomy, antique maps)

giving back to community
Herbaria makes available information for anyone with an interest that would otherwise be too difficult to get at on the same scale - so makes the 1000's of specimens stored by our predecessors potentially more useful than at any time in the last 100+ years... Old weather should improve climate modelling and so refine our understanding of the world's greatest challenge: Climate Change. Both projects also provide fascinating insights help to mainly late 19th century naturalists lives and OW that of the RN in the early 20th Century.
The intrinsic value of the project
Social contact/interest/altruism
I like the idea of contributing to something of value to the world
A desire to be useful, I suppose, and sometimes a wish to see from farther in (if not precisely the inner circle!) how a project has been organized and to learn more about the content.
I always do ones that are part of subjects that I am interested in, so I suppose its a combination of entertainment and learning. I gain both from doing these tasks. I've not done a human comp thing that doesn't cover those; however I might in the future - with the caveat: I don't think I'd ever do something like Mechanical Turk for example, but I would get involved in something like the google hunt for the aeroplane.
Boredom while working public service point
working on such projects is an opportunity to actively participate in digital culture in a meaningful way
Largely selfish enjoyment. I enjoy identifying other peoples photos and enjoy adding mine and having them classified.
Being part of wild interesting projects
My cultural interests.
Helping to build a better world
I don't know, I feel better when I do it
Sense of completion. "Someone is wrong on the Internet!" phenomenon.
Share my local and global interests
Passion
I'm motivated by the Wikimedia Foundation purpose: make every single human being able to share the sum of all knowledge.
help people
Reasons above. They are still the same, plus, in a way, I gained competences which gave me a job in the field and now I want to stay in contact with that world, also for professional reasons.
The opportunity to make available interesting resources
Interest in subject. Making material more accessible for others.
Appreciation for the democratic approach; the challenge (like a crossword puzzle), I have the time, and the effort does have some reasonably useful results. I find I can do it, unlike crosswords.
Providing resources for public viewing, and the possibility of research developing due to greater accessibility of materials
<table>
<thead>
<tr>
<th>Reason for initial interest in crowd-sourcing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contribution to a worthwhile project; progress up series of ranks</td>
</tr>
<tr>
<td>1. The subject matter 2. Being part of team 3. Believing the deliverable is useful &amp; will be used</td>
</tr>
<tr>
<td>Mostly the subject - the maps themselves, and the enjoyment of looking at what has changed between old and modern maps.</td>
</tr>
<tr>
<td>Having recently retired, I have spare time to fill in, and I have found crowd-sourcing work that I have an interest in to be not only stimulating but very enjoyable to do</td>
</tr>
<tr>
<td>Interest in Dickens and in maps, keen on computing work</td>
</tr>
</tbody>
</table>

**Table A-3: reason for initial interest in crowd-sourcing**

Question 17 asked whether participants see their contributions as being for their own interest or for the benefit of others, again a question that is concerned with the importance of a sense of community in humanities crowd-sourcing. The responses were free text, and of respondents who provided answers, only two said definitively that it was for their own benefit, whereas six said that it was for the benefit of others. Overwhelmingly (30 responses), the respondents stated that it was a combination of both, thus confirming that the motivations described in Table A-3 are, in fact, an alignment of the communal and the personal.
## Appendix B: Crowd-sourcing projects and activities

<table>
<thead>
<tr>
<th>Project</th>
<th>URL</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addressing History</td>
<td><a href="http://addressinghistory.edina.ac.uk/">http://addressinghistory.edina.ac.uk/</a></td>
<td>project</td>
</tr>
<tr>
<td>BBC Your Paintings</td>
<td><a href="http://www.bbc.co.uk/arts/yourpaintings/">http://www.bbc.co.uk/arts/yourpaintings/</a></td>
<td>project</td>
</tr>
<tr>
<td>Billion Graves</td>
<td><a href="http://billiongraves.com/">http://billiongraves.com/</a></td>
<td>project</td>
</tr>
<tr>
<td>British Library Georeferencer</td>
<td><a href="http://www.bl.uk/maps/">http://www.bl.uk/maps/</a></td>
<td>project</td>
</tr>
<tr>
<td>British Library Pin-a-tale</td>
<td><a href="http://www.bl.uk/pin-a-tale/pin-a-tale-map.aspx">http://www.bl.uk/pin-a-tale/pin-a-tale-map.aspx</a></td>
<td>project</td>
</tr>
<tr>
<td>Citizen Archivist Dashboard</td>
<td><a href="http://www.archives.gov/citizen-archivist/">http://www.archives.gov/citizen-archivist/</a></td>
<td>project</td>
</tr>
<tr>
<td>Crowdsourcing Australian Climate Change</td>
<td><a href="http://rose-holley.blogspot.co.uk/2012/02/crowdsourcing-australian-climate-change.html">http://rose-holley.blogspot.co.uk/2012/02/crowdsourcing-australian-climate-change.html</a></td>
<td>blog</td>
</tr>
<tr>
<td>Crowdsourcing Cultural Heritage: The Objectives Are Upside Down</td>
<td><a href="http://www.trevorowens.org/2012/03/crowdsourcing-cultural-heritage-the-objectives-are-upside-down/">http://www.trevorowens.org/2012/03/crowdsourcing-cultural-heritage-the-objectives-are-upside-down/</a></td>
<td>blog</td>
</tr>
<tr>
<td>Crowdsourcing Neighborhood Boundaries</td>
<td><a href="http://googlemapsmania.blogspot.co.uk/2012/07/crowdsourcing-neighborhood-boundaries.html?spref-tw">http://googlemapsmania.blogspot.co.uk/2012/07/crowdsourcing-neighborhood-boundaries.html?spref-tw</a></td>
<td>project</td>
</tr>
<tr>
<td>Crowdsourcing the AAM Annual Meeting</td>
<td><a href="http://futureofmuseums.blogspot.co.uk/2012/07/crowdsourcing-aam-annual-meeting.html">http://futureofmuseums.blogspot.co.uk/2012/07/crowdsourcing-aam-annual-meeting.html</a></td>
<td>blog</td>
</tr>
<tr>
<td>Crowdsourcing: Now With a Real Business Model!</td>
<td><a href="http://www.wired.com/business/2008/12/crowdsourcing-n/">http://www.wired.com/business/2008/12/crowdsourcing-n/</a></td>
<td>blog</td>
</tr>
<tr>
<td>DigitalKoot</td>
<td><a href="http://www.digitalkoot.fi/en/splash">http://www.digitalkoot.fi/en/splash</a></td>
<td>project</td>
</tr>
<tr>
<td>eBird</td>
<td><a href="http://ebird.org/content/ebird/">http://ebird.org/content/ebird/</a></td>
<td>project</td>
</tr>
<tr>
<td>Editathon, British Library</td>
<td><a href="http://uk.wikimedia.org/wiki/Editathon_British_Library">http://uk.wikimedia.org/wiki/Editathon_British_Library</a></td>
<td>project</td>
</tr>
<tr>
<td>Editor’s Choice: Crowdsourcing and Cultural Heritage Round-up</td>
<td><a href="http://digitalhumanitiesnow.org/2012/06/editors-choice-crowdsourcing-and-cultural-heritage-round-up/">http://digitalhumanitiesnow.org/2012/06/editors-choice-crowdsourcing-and-cultural-heritage-round-up/</a></td>
<td>blog</td>
</tr>
<tr>
<td>Even Crowdsourcing Can Get Too Expensive</td>
<td><a href="http://scholarlykitchen.sspnet.org/2011/03/14/even-crowdsourcing-can-get-too-expensive/">http://scholarlykitchen.sspnet.org/2011/03/14/even-crowdsourcing-can-get-too-expensive/</a></td>
<td>blog</td>
</tr>
<tr>
<td>First World War Poetry Digital Archive</td>
<td><a href="http://www.oucs.ox.ac.uk/ww1lit/">http://www.oucs.ox.ac.uk/ww1lit/</a></td>
<td>project</td>
</tr>
<tr>
<td><strong>Frequently Asked Questions about crowdsourcing in cultural heritage</strong></td>
<td><a href="http://openobjects.blogspot.co.uk/2012/06/frequently-asked-questions-about.html">http://openobjects.blogspot.co.uk/2012/06/frequently-asked-questions-about.html</a></td>
<td>blog</td>
</tr>
<tr>
<td><strong>Galaxy Zoo</strong></td>
<td><a href="http://www.galaxyzoo.org/">http://www.galaxyzoo.org/</a></td>
<td>project</td>
</tr>
<tr>
<td><strong>HistoryPin</strong></td>
<td><a href="http://www.historypin.com/">http://www.historypin.com/</a></td>
<td>project</td>
</tr>
<tr>
<td><strong>I’m not an experience-seeking user, I’m a meaning-seeking human person</strong></td>
<td><a href="http://blog.tomorris.org/post/3216687621/im-not-an-experience-seeking-user-im-a">http://blog.tomorris.org/post/3216687621/im-not-an-experience-seeking-user-im-a</a></td>
<td>blog</td>
</tr>
<tr>
<td><strong>iBayDigital</strong></td>
<td><a href="http://www.ibaydigital.co.uk/btbond/">http://www.ibaydigital.co.uk/btbond/</a></td>
<td>project (concluded)</td>
</tr>
<tr>
<td><strong>JISC and Crowdfunding</strong></td>
<td><a href="http://www.jisc.ac.uk/blog/crowd/?utm_source=feedburner&amp;utm_medium=feed&amp;utm_campaign=Feed%3A+JISCBlog+%28JISC+Blog%29">http://www.jisc.ac.uk/blog/crowd/?utm_source=feedburner&amp;utm_medium=feed&amp;utm_campaign=Feed%3A+JISCBlog+%28JISC+Blog%29</a></td>
<td>blog</td>
</tr>
<tr>
<td><strong>Meanification and crowdscaffolding: forget badges</strong></td>
<td><a href="http://www.playthepast.org/?p=1027">http://www.playthepast.org/?p=1027</a></td>
<td>blog</td>
</tr>
<tr>
<td><strong>Old Reel project</strong></td>
<td><a href="http://www.regaltenbury.org.uk/memory-reel/">http://www.regaltenbury.org.uk/memory-reel/</a></td>
<td>project</td>
</tr>
<tr>
<td><strong>Old Weather</strong></td>
<td><a href="http://oldweather.org/">http://oldweather.org/</a></td>
<td>project</td>
</tr>
<tr>
<td><strong>On Making, Use and Reuse in Digital Humanities</strong></td>
<td><a href="http://melissaterras.blogspot.co.uk/2012/03/on-making-use-and-reuse-in-digital.html">http://melissaterras.blogspot.co.uk/2012/03/on-making-use-and-reuse-in-digital.html</a></td>
<td>blog</td>
</tr>
<tr>
<td><strong>OurWikiBooks</strong></td>
<td><a href="http://www.jisc.ac.uk/whatwedo/programmes/digitisation/communitycontent/ourwikibooks.aspx">http://www.jisc.ac.uk/whatwedo/programmes/digitisation/communitycontent/ourwikibooks.aspx</a></td>
<td>project</td>
</tr>
<tr>
<td><strong>Portable Antiquities Scheme</strong></td>
<td><a href="http://finds.org.uk/">http://finds.org.uk/</a></td>
<td>project</td>
</tr>
<tr>
<td><strong>Prism project</strong></td>
<td><a href="http://prism.scholarslab.org/">http://prism.scholarslab.org/</a></td>
<td>project</td>
</tr>
<tr>
<td>Resource</td>
<td>URL</td>
<td>Type</td>
</tr>
<tr>
<td>-----------------------</td>
<td>----------------------------------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Qurator</td>
<td><a href="http://www.qrator.org/">http://www.qrator.org/</a></td>
<td>project</td>
</tr>
<tr>
<td>The Crowd and the Library</td>
<td><a href="http://www.trevorowens.org/2012/05/the-crowd-and-the-library/">http://www.trevorowens.org/2012/05/the-crowd-and-the-library/</a></td>
<td>blog</td>
</tr>
<tr>
<td>Transcribe Bentham</td>
<td><a href="http://www.ucl.ac.uk/transcribe-bentham/">http://www.ucl.ac.uk/transcribe-bentham/</a></td>
<td>project</td>
</tr>
<tr>
<td>Treading Water on Open Access</td>
<td><a href="http://www.dancohen.org/2012/09/25/treading-water-on-open-access/">http://www.dancohen.org/2012/09/25/treading-water-on-open-access/</a></td>
<td>blog</td>
</tr>
<tr>
<td>Ur Crowdsorce</td>
<td><a href="http://urcrowdsorce.org/omeka/">http://urcrowdsorce.org/omeka/</a></td>
<td>project</td>
</tr>
<tr>
<td>Various resources</td>
<td><a href="http://www.delicious.com/stacking/crowdsourcing">http://www.delicious.com/stacking/crowdsourcing</a></td>
<td>links list</td>
</tr>
<tr>
<td>What’s the Score at the Bodleian</td>
<td><a href="http://www.whats-the-score.org/">http://www.whats-the-score.org/</a></td>
<td>project</td>
</tr>
<tr>
<td>Wikisource</td>
<td><a href="http://wikisource.org/wiki/Main_Page">http://wikisource.org/wiki/Main_Page</a></td>
<td>project</td>
</tr>
<tr>
<td>Stranlines</td>
<td><a href="http://www.strandlines.net">http://www.strandlines.net</a></td>
<td>project</td>
</tr>
</tbody>
</table>
A Survey of Digital Approaches to the Large-scale Transcription of Pre-modern Japanese Documents

Yuta Hashimoto
National Museum of Japanese History
October 1, 2018
Abstract

The development of digital methods to transcribe a large volume of historical documents are the subject of active research in the field of digital humanities. This paper reviews the state-of-the-art technologies and projects focused on large-scale transcription of Japanese historical documents. It is estimated that of the 20 billion plus pre-modern documents preserved across the country, only a handful have been transcribed despite the long efforts of historians. For effective information retrieval from these documents, information scientists and digital humanities scholars in Japan have made various attempts, with methods grouped into two categories: (1) a crowdsourcing approach represented by projects such as Minnade Honkoku, Wikisource, Aozora Bunko, and Hondigi and (2) machine recognition approaches such as MOJIZO, DSC Search, and Kuzushiji Challenge.

After a brief description on the bibliographic nature and writing system of Japanese pre-modern documents, the author will examine these projects and their technical backgrounds.

Keywords: historical documents, Japanese studies, crowdsourcing, handwritten text recognition
1. Introduction

Transcribing historical documents that are difficult for modern readers is the most fundamental basis of historical research. Through this process, historical documents preserved in archives, museums, or private sectors become public resources available to every scholar. In the case of Japan, transcription of historical documents has become one of a historians’ duties since the end of the nineteenth century, when letterpress printing started to replace woodblock printing in Japan. Over a hundred years, scholars of Japanese history have collected historical documents from all over the country, created their catalogs, and transcribed them for academic publications.

However, despite the long-term efforts by historians and archivists, there are still a huge number of historical documents left to be transcribed. In particular, the number of manuscripts and books written or printed in the Edo period (1603–1863) was so large that it will take several centuries to transcribe, even if all historians and archivists in Japan work on them. Thus, many of those scholars put their hope on the digital methods that enable large-scale transcription of historical documents. To meet such expectations, information scientists and digital humanities scholars in Japan have made continuous efforts to develop technologies for conducting large-scale transcription of Japanese pre-modern documents. The first attempt in this field was probably a Kaken project led by Shibayama (https://kaken.nii.ac.jp/en/grant/KAKENHI-PROJECT-11410090/) that started in 1999. This project’s aim was to develop an algorithm for the recognition of kuzushiji, an old calligraphic rendering of kanji characters that were common both for publishing and handwriting in the pre-modern Japan. The project produced a number of
results, including an electric dictionary of *kuzushiji* with image search functionality,¹ a n-gram based transcription support system,² and layout analysis of Japanese historical documents.³ However, it turned out that the technologies available at that time were not powerful for the automatic recognition of *kuzushiji*.

After Shibayama’s project, various approaches including machine learning and crowdsourcing have since been attempted. Only in recent years have some of them produced good, but still limited, results. In this paper, projects with large-scale transcription of Japanese historical documents using state-of-the-art technologies will be reviewed.

2. Overview of Pre-modern Japanese Documents

This section gives a brief description of the bibliographic nature, writing system, and archival situation of pre-modern Japanese documents, which will aid us in seeing the difficulties and significance of conducting large-scale transcription of Japanese pre-modern documents.

---


2.1 Genres

Pre-modern Japanese documents are usually classified into the following three genres:

- **Komonjo** (古文書) is often used as a general term of Japanese historical documents. In an academic context, a komonjo is defined as a handwritten document that conveys the will or thought of its sender to its recipient. This includes administrative documents such as Imperial rescript and shogunate's decree. It also includes documents of more private natures such as promissory notes, land certificates, and letters between persons.

- **Kokiroku** (古記録) refers to historical documents written without any specific readers, e.g., personal diaries, organizational records written by imperial or local governments, temples, or feudal lords.

- **Kotenseki** (古典籍) is a general word that refers to books published in the pre-modern era. These include literary works such as novels, essays, and collections of poems. Kotenseki also includes various types of other publications such as scrolls and sutras. As the modern letterpress printings began only after the end of 19th century, most of kotenseki are woodblock printings.

The above genres comprise many sub-genres and that there are other historical materials with written characters that do not fall into these genres such as maps, paintings, scrolls, and calligraphic works.

How many of these types of documents are preserved? It seems that no one knows the exact amount but some scholars have made rough estimations; according to Mitsutoshi
Nakano, a scholar of early modern Japanese literature, the whole number of kotenseki published in the Edo period (1603–1868) is at least 1.5 million, of which only less than ten thousand have been transcribed by scholars or archivists.\(^4\) In addition to kotenseki, an enormous number of handwritten komonjo and kokiroku are preserved across the country. Some of them are stored in public archives, libraries, and historical museums. Others are privately owned by citizens. Hiroshi Okumura, a historian of early modern Japan who has long engaged in the preservation activities of such historical manuscripts, estimates the number of documents to be at least 20 billion from the investigations he has done in the past.\(^5\)

### 2.2 Writing System

The major difficulty of transcribing pre-modern Japanese documents – by both machines and humans – comes from the writing systems used; most pre-modern Japanese documents are written with kuzushiji, old calligraphic renderings of kanji characters, which modern Japanese find difficult to read. The current Japanese writing system consists of two types of characters: forty-six syllabic kana and thousands of logographic kanji, which are adapted Chinese characters. The modern system has been developed since the Meiji Restoration of 1868; in this process, Japanese society made obsolete some


\(^5\) Hiroshi Okumura, “Big Earthquake and Historical Heritage of Regions : The Role of Universities for Forming Regions Strong against a Disaster,” *Nagoya University Archives News*, no. 21 (March 2013): 133–64.
types of characters and writing styles that were common before the Meiji period.

First, pre-modern Japanese had the option to use different characters other than the current kana to denote a single syllable. For instance, a kana “す” (pronounced "su") had more than five alternative characters (see Figure 1). These variants of kana are called hentaigana, which were officially made obsolete by the amended education law announced in 1900. Second, in most publications and writings before the Meiji Era (1868), kanji had been written mostly in cursive forms that appear today only in calligraphy (see Figure 2). As letterpress printing had replaced woodblock printing after the Meiji period, this cursive style of writing gradually declined both in handwriting and printing. Kuzushiji is a general term (including hentaigana and cursive kanji) for characters that are not used today.

For the historical reasons above, most pre-modern documents written with kuzushiji are illegible for modern Japanese except for a small number of trained experts.
Figure 1 Five hentaigana for a syllable "ず" (su).

Figure 2 A comparison of a kuzushiji and a modern Japanese type character. Both represent the same character 前, "front" in English.
2.3 Archives, Compilations, and Databases

In many other countries, a large number of Japanese historical documents are collected by and preserved in libraries, museums, archives, and other public or private institutions. The Historiographical Institute at the University of Tokyo holds approximately 200,000 komonjo and kokiroku while the National Diet Library (NDL) approximately 280,000 kotenseki, and National Institute of Japanese Literature (NIJL) 37,000 kotenseki.

In parallel with the collecting and archiving of historical documents, the historians and archivists at these institutions have conducted a number of compilation projects of historical documents for over a century. The Historiographical Institute at the University of Tokyo was founded in 1901 as the only public institution whose mission is the compilation of historical documents. The institute has published over 1,100 volumes of compilations, including Dai Nihon Komonjo and Dai Nihon Kokiroku, whose publications are still in progress. Such compilation projects take a long time, often decades. For instance, in 2004, the Kyoto Prefectural Library and Archives started the compilation project of Toji Hyakugo Monjo, which is a collection of medieval komonjo originally preserved in Toji, a Buddhist temple founded at Kyoto in 1796. The collection comprises nearly 25,000 documents from the 8th to 18th centuries. The Kyoto Prefectural Library and Archives has kept publishing one volume of the compilation annually since 2004. However, the total number of volumes to be published is estimated at sixty, which will not be finished by 2050 at the current pace.

In recent years, the public libraries and archives actively conduct the digitization
projects of historical documents. The NIJL, for instance, is administering a project to publish online the digital images of the 300,000 historical documents held by the NIJL and its associated institutions by 2023. At NDL’s website (http://dl.ndl.go.jp/?__lang=en), the scanned images of the 23,000 kotenseki are available. The Kyoto Prefectural Library and Archives launched a digital archive in 2014 which provides the 80,000 images of Toji Hyakugo Monjo under the CC BY 2.1 license (http://hyakugo.kyoto.jp/eng/). Moreover, some institutions such as NDL, University of Tokyo, and Kyoto University have started to provide the digital images of historical documents in the IIIF (International Image Interoperability Framework) format.

There are also text databases that offer transcribed texts of historical documents. The largest ones among them are those administered by the Historiographical Institute (http://wwwap.hi.u-tokyo.ac.jp/ships/), which offers the full-text search of the compilations published by the institute. As for kotenseki, the NIJL (http://base1.nijl.ac.jp/~nkbthdb/) provides a full-text database of the 556 classical works originally from Japanese Classic Literature Systematic Edition published by a publishing company Iwanami Shoten. Japan Knowledge (https://japanknowledge.com/library/), an online reference collection of Japanese encyclopedias, dictionaries, and compilations, also provides several text databases of pre-modern Japanese documents. For instance, the full-text of Gunsho Ruijū, a collection of classical books assembled by Hanawa Hokiichi which was published in 666 volumes from 1793 to 1819, is available on Japanese Knowledge. On the contrary, there are only a few digital archives of historical documents that provide both scanned images and transcribed texts.
2.4 Expected Outcomes of Large-scale Transcription

If it becomes possible to transcribe a large number of historical Japanese documents in a short period of time with high accuracy, what will it bring to Japanese studies? If, for instance, the full-text of all the 300,000 kotenseki provided by NIJL becomes available, what benefits will scholars receive? The immediate benefits of such a technology will be the following:

**Full-text search.** In order to retrieve information from a typical historical document on a digital archive, currently one must read through the entire document, which is written with kuzushiji and is often hundreds of pages long. The ability to conduct a full-text search across a large number of historical documents will greatly streamline information retrieval. It is easy to imagine that this will lead to the discovery of many previously unknown historical facts.

**Text analysis.** Automated text analysis is a useful tool used to discover and measure patterns of language use that can be interpreted as prevalent attitudes, concepts, or events. Since the rise of digital humanities in the 2000s, text analysis on humanities texts has become the subject of extensive research. However, the unavailability of electric text makes it difficult to apply this method to Japanese pre-modern documents with few exceptions on classical works.⁶ By making a vast number of historical Japanese

---

documents available as electronic texts, will open up a new research field in this direction.

**Interdisciplinary use of historical documents.** While descriptions written in historical documents are sometimes inaccurate and unreliable, some of them provide the information on the past events whose observation data are unavailable. There is a growing interest among scientific communities in historical documents as a means to study the past natural phenomena that occurred in the pre-modern era. Aurora4D (https://aurora4d.jp/), a joint group of physicists and humanities scholars, aims to reveal the past astronomical events from the written records. The group found from the survey of historical Japanese records that the largest-ever magnetic storm took place in 1770.\(^7\) The Historical Earthquake Study Group (http://kozisin.info/) at Kyoto University also uses written documents to study historical earthquakes and volcanic eruptions. Large-scale transcription of historical documents will accelerate these interdisciplinary studies.

3. Crowdsourcing

Machine recognition of handwritten text on historical documents is quite a difficult task, thus some scholars try to use the crowdsourcing technique for large-scale transcription of historical documents. Crowdsourcing is the process of getting work from a large group of people for their skills, ideas, and participation to generate content or help

---

facilitate the creation of content or products. Although this term originally referred to commercial services such as Amazon Mechanical Turk, \(^8\) from the late 2000s crowdsourcing started to spread in the academic world as a brand-new method that enables large-scale data analysis, collection and creation with a low cost.\(^9\)

In the humanities domain, some projects adopted crowdsourcing for the correction of OCR-scanned texts in the late 2000s. Such examples include Chinese Text Project (https://ctext.org/) run by Donald Sturgeon since 2006 and Australian Newspapers Digitization Project (https://www.nla.gov.au/content/newspaper-digitisation-program/) administered by The Australian National Library. In 2010, a number of crowdsourcing projects launched that aims to transcribe a huge volume of handwritten texts: Transcribe Bentham (http://www.transcribe-bentham.da.ulcc.ac.uk/) run by University College London, Old Weather (https://www.oldweather.org/) hosted on Zooniverse, Smithsonian Transcription Center (https://transcription.si.edu/), Europeana’s “transcribathon” (https://transcribathon.com/), NYPL’s “What’s on the menu?” (http://menus.nypl.org/), and so on.

In contrast, it had been widely believed that crowdsourcing on Japanese pre-modern documents is not possible given the difficulty of reading *kuzushiji*. However, as more and more learning materials for reading *kuzushiji* become available online, it seems that this

---


hindrance is being removed. In the following, several crowdsourced transcription projects and platforms for Japanese historical documents will be examined. For reference, I will include a few projects as such, whose subjects are the modern documents published or written after the Meiji Era (1868–1912).

3.1 Minna de Honkoku

Minna de Honkoku (https://honkoku.org/) is a crowdsourced transcription project of the pre-modern earthquake records stored in the Earthquake Research Institute (ERI) at the University of Tokyo. Most of the records are from the Edo period (1603–1868). The project is run by the members of the Historical Earthquake Study Group (HESG), a joint group of seismologists and historians, including the authors at Kyoto University, who have been studying pre-modern earthquake records for seismic research and disaster prevention. As the instrumental observation of earthquakes in Japan began only after the end of nineteenth century, transcribing written records are required for studying past earthquakes. Therefore, Japanese seismologists have developed an extensive collaboration with historians and archivists. However, the number of records to be transcribed is vast and cannot be handled by a small group of scholars. This prompted the members of HESG to think of using crowdsourcing for transcribing historical earthquake records.10

One notable feature of Minna de Honkoku is that it collaborates with KuLA (Kuzushiji Learning App), a mobile learning app for reading kuzushiji that was developed at Osaka University in 2016 and has been downloaded over 100,000 times. After completing a set of basic lessons for reading kuzushiji, the users of KuLA are invited to Minna de Honkoku as an opportunity to acquire more practical training by transcribing actual materials from pre-modern Japan. Thus, they can begin participating in the project as a continuation of their learning (see Figure 3).

![Figure 3 Screenshots of KuLA](image)

The website of Minna de Honkoku was launched on January 2017. As of May 2018, 443 books out of 472 (12,768 pages out of 13,932) from ERI have been transcribed by 4,233 registered users. A total number of 4.91 million characters. From these numbers,

---

one may call Minna de Honkoku the first successful crowdsourcing project on Japanese pre-modern documents.

3.2 Wikisource

Wikisource, a digital library based on wiki engine run by the Wikimedia foundation, also invites volunteers to historical Japanese texts. As of May 2018, the Japanese version of Wikisource (https://ja.wikisource.org/) holds thirty-three pre-modern texts under the "history" category, most of which seemed to be copied from the published compilations. Although participation from a large group of people has not occurred on Japanese Wikisource so far, the support from Wikimedia foundation is its big advantage. It is possible that Wikisource grows into a major platform for sharing the transcription of historical documents in the near future.

3.3 Aozora Bunko

Aozora Bunko (https://www.aozora.gr.jp/), or Blue Sky Library, is an online digital library similar to Project Gutenberg. The website was launched in 1997 by a group of volunteers. Since then, Aozora Bunko has provided to Internet users free access to classical Japanese novels, poems and essays including some translations from English literature. As of February 2018, the collection encompasses 14,000 out-of-copyright works by 1,000 authors. Most of the works on Aozora Bunko are usually transcribed from published books by volunteers, proof-read by other volunteers, and published online in
both plain text and HTML formats. As Aozora Bunko currently focuses on the modern books published after the Meiji Era, only few texts transcribed from the pre-modern documents are available on the website.

### 3.4 Hondigi

Hondigi ([http://lab.ndl.go.jp/dhii/omk2/](http://lab.ndl.go.jp/dhii/omk2/)) is a crowdsourced transcription project of digitized books available on the digital library of the National Diet Library (NDL). NDL publishes online scanned images of over 350,000 out-of-copyright books. However, owing to the low recognition rate of OCR programs, the entire text of each of these books is not provided. Hondigi aims to create full-text transcriptions of these books so that they will be indexed by search engines such as Google and will become more discoverable on the web. The project was planned and administered by a special interest group at the Japanese Association for Digital Humanities (JADH).

The website of Hondigi was launched in March 2014. As of May 2018, the number of items registered on Hondigi is still only twenty-one, probably due to the experimental nature of the project. Later in 2016, a derived project of Hondigi was launched on

---


Crowd4U (http://crowd4u.org/ja/) a portal of academic crowdsourcing projects.\textsuperscript{14} The Crowd4U version of Hondigi adopts “microtask” system similar to the old version of reCAPTCHA in order to reduce the labor of volunteers; it displays a small part of a document image at once, which volunteers can transcribe in a short period of time. The result of this project has not been reported yet.

4. Machine Recognition

Handwritten text recognition (HTR) is a form of computer technology to receive and interpret handwritten input from sources such as paper documents, touchscreens, photographs, among others. HTR should not be confused with optical character recognition (OCR), which has been developed for the recognition of letterpress printings. HTR, especially for a historical document, is a very difficult task regardless of the language the document is written. The recognition must be done "off line", i.e., the information of pen-tip movement which aids the character recognition is not available unlike the handwriting recognition on electric devices.\textsuperscript{15}

In addition, the recognition of kuzushiji entails additional difficulties. First, unlike alphabetical languages, the HTR for Japanese texts must be able to correctly recognize at least a few thousands of kanji characters. Second, the character segmentations of kuzushiji


are much more difficult than modern Japanese texts since the past Japanese wrote characters continuously and quickly with brushes. Third, some \textit{kazushiji} characters tend to be written in a very abbreviated manner. For instance, all the images in Figure 4 represent the same \textit{kazushiji} "候" (so-ro), an old auxiliary verb frequently put at the end of sentences. For these difficulties, some scholars develop semi-automatic methods that will work in cooperation with human transcribers, rather than full-automatic methods.

![Kazushiji Characters "候" (so-ro)](image)

\textbf{Figure 4} \textit{Kazushiji} characters "候" (so-ro), taken from the \textit{Kazushiji} database at the Historiographical Institute, the University of Tokyo.
There are two major algorithms for HTR: feature extraction and neural network. Feature extraction recognizer converts a given image to a set of properties (feature values or vectors) that well summarizes the target image, from which the recognizer identifies the character or word written in it. Neural network recognizers first learn from the initial image training set. The trained network then makes the character identifications. Each neural network uniquely learns the properties that differentiate training images. It then looks for similar properties in the target image to be identified. In recent years, recognition with multi-layered neural network, also known as deep learning, has become a major technique in this area. On the contrary, to train a deep neural network, a huge amount of the training data is required.

In the following, I will examine several products and projects developed for the HTR of Japanese historical documents.

4.1 MOJIZO

MOJIZO (http://mojizo.nabunken.go.jp/) is an image-based search engine for kuzushiji, developed by Akihito Kitadai for assisting in manual transcription of historical documents.16 Given an image of kuzushiji, MOJIZO looks for characters with similar shapes from the online kuzushiji database run by the Historiographic Institute at the

University of Tokyo (see Figure 5). The database encompasses 198,852 images of 5,963 characters, mostly taken from the medieval manuscripts the institute has digitized.

The basic algorithm behind MOJIZO is a feature extraction; it conducts pre-processing on a given image to reduce noises, normalize the character shape to remove variant factors, extracts the feature vectors, and makes identifications by the comparison of pre-calculated feature vectors generated from the kuzushiji database.
Figure 5 MOJIZO’s search results for an image of kuzushiji “最” (sai).

4.2 DSC Search

DSC Search is a word-spotting engine for handwritten manuscripts developed by an information scientist Kengo Terasawa. Word-spotting is a technology developed to bypass the difficulty of recognizing handwritten texts; unlike HTR, word-spotting does not interpret handwritten texts. Instead, it looks for all the occurrences of a given word
across the document images based on image similarities so that a human transcriber can transcribe the multiple occurrences of the word at once.\textsuperscript{17}

To conduct word-spotting, the target documents images must be indexed in advance. DSC Search indexes the document images by converting the text lines in the image into a sequence of the feature vectors. Given a query image, DSC Search looks for the similar images from the document by calculating the distances of the feature vector sequences between the query and candidate images. As the algorithm of DSC Search does not require character segmentations, it works well with Japanese historical documents written continuously with \textit{kuzushiji}.\textsuperscript{18}

Some digital humanity projects have imported DSC Search into their products. For instance, SMART-GS (https://osdn.net/projects/smart-gs/), a transcription suite for historical documents developed at Kyoto University, embeds DSC Search as an auxiliary tool to find keywords across documents.

4.3 \textit{Kuzushiji Challenge}

It is well known that deep learning is data-greedy, i.e., a large amount of training data must be provided to train a deep neural network. Most popular image datasets for deep learning encompass a large number of examples; MNIST


The Center for Open Data in the Humanities (CODH) founded in 2017 is a research institute that promotes the wide use and publishing of open data in humanities. As one of its projects, CODH provides a data set of kuzushiji images for the training of neural networks. The data set is composed of about 400,000 images of 4,000 characters (see Figure 7). Moreover, CODH launched a competition titled "Kuzushiji Challenge" (http://codh.rois.ac.jp/old-char-challenge), which provided three recognition tasks of different levels on the recognition of kuzushiji. Programs submitted by participants were evaluated by their performances such as recognition accuracy and execution speed.

Among the thirty groups who participated in the competition, the winner of the competition was a team from Tokyo University of Agriculture and Technology. Their program, which combines convolutional neural network and bidirectional long short-term memory, achieved recognition accuracy of 96.8% for single characters and 87.12% for three characters sequences. Their program also won the best paper award at the 4th International Workshop on Historical Document Imaging and Processing.19

Although the recognition rates the winner team has achieved are remarkable, it should also be noted that the recognition tasks for the Kuzushiji Challenge was only limited to kana characters whose number is less than one hundred. For the full-text transcription of Japanese historical documents, the recognition program must be able to read thousands of kanji characters. It is very unlikely that this will be achieved purely based on image-based algorithms.

5. Future Directions

The success of both the crowdsourced transcription and deep-learning-based HTR of Japanese pre-modern documents took place only in the last few years. There will be continuous efforts to elaborate these methods for the next several years. In the near future, it may be possible to conduct a full-text transcription of large digital archives such as NIJL’s kotenseki database or Toji Hyakugo Monjo. However, both methods have their own limitations. This section outlines the challenges and possibilities of these methods.

The major challenge of crowdsourced transcription comes from its scalability. In the case of Minna de Honkoku, it took fifteen months to transcribe all the 450 pre-modern earthquake records with the help of 4,000 registered users. At this pace, it will take several centuries to transcribe, for instance, the 300,000 kotenseki books in the NIJL's database, which is of course not acceptable. To administer a crowdsourced transcription on such large digital archives, a much larger participation is required.

One possible way to gain larger participation would be to make calls to local history communities, which the present crowdsourcing projects including Minna de Honkoku fail to access; there are a large number of history communities across Japan, most of which their members are retired seniors interested in history or classical literature. One can find a hundred examples of such communities on the Internet. Most communities regularly conduct group readings of historical documents at local libraries, museums, archives, and universities. If it is possible to gain cooperation from such history communities, the full transcription of large digital archives by crowdsourcing will become a real possibility.

The next challenge of HTR on historical Japanese documents is the inclusion of language models. Although deep-learning-based HTR has produced a very good result in Kuzushiji Challenge, it is very unlikely that the exact same method will work on historical documents that contain kanji. A possible solution to this is to combine image-based recognition and language models. A language model is a probabilistic model that assigns probabilities to any sequence of words. With a trained language model, it is possible to detect and correct recognition errors to some extent. This method has been long and widely used as a post process of OCR for error corrections.

The limitation of deep-learning-based HTR is that it is essentially a pattern recognition algorithm; when a trained historian comes across an illegible word in a historical document, he or she tries to read it using the background knowledge on the document: the temporal and geographical context in which the document was written, the social situations and common customs at the time, and so on. Deep learning is not a sort of algorithm capable of such knowledge-based inferences. Thus, it is likely that no matter
how the algorithms are improved, HTR on historical Japanese documents will make errors to some extent.

6. Conclusion

This paper has surveyed several projects for the large-scale transcription of Japanese historical documents. There are major approaches in this research field: crowdsourcing and machine recognition. Neither of them is a silver bullet; the full-text transcription of large digital archives will probably be possible only through the cooperation between humans and machines. In the same way, the Chinese Text Project and the Australian Newspapers Digitization Project invites a large number of people to correct OCR errors.\(^{20}\)

Text databases of historical documents are becoming new infrastructures for humanities. In contrast to Europe and North America where many full-text databases of classical literature and historical records are provided, there are only a few text databases for historical Japanese documents, mainly because letterpress printings had not prevailed until the end of the nineteenth century in Japan. Therefore, the technological development for large-scale transcription will be an important key for Japanese studies moving forward to the next stage as an academic field.

Bibliography


Kitadai Akihito, Shirai Keichiro, Inoue Satoshi, Kurushima Noriko, Baba Hajime, Watanabe Akihiro, and Nakagawa Masaki. “Character pattern retrieval in historical documents for seamless computing (Japanese).” *Proceedings of Computers and the Humanities Symposium 2012* 2012, no. 7 (November 10,


Okumura, Hiroshi. “Big Earthquake and Historical Heritage of Regions: The Role of Universities for Forming Regions Strong against a Disaster.” *Nagoya University Archives News*, no. 21 (March 2013): 133–64.


