Digital Resources of Japanese Texts: from a Viewpoint of Digital Humanities

Kiyonori Nagasaki
International Institute for Digital Humanities, Tokyo
and University of Tokyo
Self-introduction:

• Studied Buddhist philosophy/philology in graduate school in University of Tsukuba.
• Experienced network administrator of several Internet providers and universities during graduate school and post-doc
• Taught computer science and ethics in a local university as lecturer/associate professor during 7 years
• Resigned the university and established an independent research institute in 2009 in order to manage a Buddhist database and to form DH community in Japan.
Self-introduction:

• Studied Buddhist philosophy/philology in graduate school in University of Tsukuba.

• Experienced network administrator of several Internet providers and universities during graduate school and post-doc

• Taught computer science and ethics in a local university as lecturer/associate professor during 7 years

• Resigned the university and established an independent research institute in 2009 in order to manage a Buddhist database and to form DH community in Japan.
Range of my recent activities

• Universities and Institutes
  • Senior fellow in International Institute for Digital Humanities
  • Project associate professor of DH in University of Tokyo
  • Leader of a collaborative research group of DH in Kyoto University
  • Researcher in Japanese National Diet Library
  • A member of digitization committee in National Institute of Japanese Literature
  • Co-researcher in Okinawa Prefectural University of Arts

• Academic communities
  • Secretary of Japanese Association for Digital Humanities
  • A member of steering committee of Japanese Association for Indian and Buddhist Studies
  • A member of steering committee of SIG-Computers and Humanities in Information processing Society of Japan
  • Editor of Digital Humanities Monthly (e-newsletter)
  • A member of nominating committee of DH awards

• Research projects
  • A member of SAT Daizōkyō Text Database Committee
  • A member of Indo-Tibetan Lexical Resources project
  • A member of Digital Dickens Lexicon project

• And so on...
Large Scale Digital Resources of Japanese Texts

- Freely browsable on Web

  - National Diet Library Digital Collections
    - National Diet Library
    - Over 500,000 items
    - http://dl.ndl.go.jp/
  - CiNii
    - National Institute for Informatics
    - Over 4000000 articles.
    - http://ci.nii.ac.jp/
  - NIHU
    - National Institute for Humanities
    - Integrating over 100 research databases
    - http://int.nihu.jp/

- Not freely “available”!
Middle or Small Scale Digital Resources of Japanese Texts (1)

• National Institute for Japanese Language and Linguistics
  • Balanced Corpus of Contemporary Written Japanese (BCCWJ) (over 100,000,000 words)
  • Taiyō Corpus
  • Bibliographical database of this field
  • ...various databases
  • http://www.ninjal.ac.jp/english/products/

• National Institute of Japanese Literature
  • Bibliographical database of this field
  • ...various databases
  • https://www.nijl.ac.jp/pages/database/
Middle or Small Scale Digital Resources of Japanese Texts (2)

• Various databases in universities and institutes
  • University of Tokyo
    • Historiographical Institute (Japanese History)
      • [http://wwwap.hi.u-tokyo.ac.jp/ships/db-e.html](http://wwwap.hi.u-tokyo.ac.jp/ships/db-e.html)
    • Center for Evolving Humanities (Buddhist study and DH)
    • ... various institutes and faculties
  • Art Research Center, Ritsumeikan University
    • [http://www.arc.ritsumei.ac.jp/search_arc.html](http://www.arc.ritsumei.ac.jp/search_arc.html)
  • ...
Tendencies:

• Large scale resources:
  • Managed by general organizations
  • Reliable sustainability
  • Simple search function
  • Providing image files, not text data

• Middle or Small scale resources:
  • Managed by research groups or specialized organizations
  • Unreliable sustainability
  • Specific search function
  • Providing not only image file but also tag, specific metadata, and text data in some cases.
  • Providing other specific functions
  • Strong relationship with DH (Digital Humanities)
A typical DH model

• By Willard McCarty and Harold Short
A viewpoint of digital humanities

• “DH” which has been formed by researchers who had been addressing digitization of the humanities since several decades ago.

• Humanities researchers should be expected to develop the potential for cultural resources.
  Useful resources for humanities researchers and students.

• DH researchers should be expected to develop the potential for digital cultural resources.
  Useful digital resources for DH researchers and students.
History: the dawn in Japan

1950’s

The Mathematical Linguistic Society of Japan was established.

1960’s

National Institute for Japanese Language and Linguistics started to treat Kana and Han characters in computer.

1970’s

Large-scale computing in several universities

National Institute of Japanese Literature

National Museum of Ethnology
History: emerging communities

1980’s

• Accelerated by IBM PC
• **SIG-CH** (Special Interest Group-Computer and the Humanities)
  • Under the Information Processing Society of Japan (Mainly aims IT)
• Japan Society of Information and Knowledge
• Japan Art Documentation Society
History: emerging communities of certain field

- Accelerated by the Internet

1990’s

- Japan Association for English Corpus Studies
- Japan Information-Culturology Society
- Japan Society for Archaeological Information

2000’s

- Japan Association for East Asian Text Processing
- Japan Society for Information and Media Studies
- The Japan Society for Archival Science
- Japan Association for Contents History Studies
History: Globalization

2010’s

- Japanese Association for Digital Humanities (JADH)
  - Established almost by humanities researchers who have been engaged in DH.
  - Became a constituent organization of ADHO
  - Holding international conference every year
    - Next: 9/19-21 in University of Tsukuba

- International Symposium series of Culture and Computing
  - Managed by IT researchers who are interested in culture
  - Publishing proceedings from Springer or IEEE.
  - Holding intl. conference in Kyoto or East Asia.

- Most of related communities remain
History: large grants by government

• Approximately 1 million dollars per year

1995-1999

• “Computer and the Humanities”
  • Constituted by leading researchers in several universities

1998-2002

• “Restructuring classical studies”
  • A group of digitization played an important role.

2001-2006

• “Grammatological Informatics based on Corpora of Asian Scripts”
  • Carried by Tokyo University of Foreign Studies
History: “Center Of Excellence” grant

- Approximately 1 million dollars per year

2002-

- “21th century COE”
  - Kyoto University
  - Tokyo Institute of Technology
  - Tokyo University of Foreign Studies
  - Ritsumeikan University ...

2007-

- “Global COE”
  - Ritsumeikan University
  - Tokyo University of Foreign Studies
Recent DH-like JSPS-funded projects

- JSPS-funded projects of DH-like studies including making digital cultural resources for the humanities:

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of projects</th>
<th>Total (in JPY)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>99</td>
<td>251,862,000</td>
</tr>
<tr>
<td>2010</td>
<td>126</td>
<td>385,300,000</td>
</tr>
<tr>
<td>2011</td>
<td>172</td>
<td>505,390,000</td>
</tr>
<tr>
<td>2012</td>
<td>203</td>
<td>654,160,000</td>
</tr>
</tbody>
</table>
SIG-CH presentations

• 1989-present, quarterly workshop
• Over 800 presentations in total
• Registered by around 300 researchers
• Non-refereed papers
• Allowed anyone to make a presentation
• Strongly IT-oriented
  • Under the auspices of a kind of IT society
• Started by IT researchers, but recently humanities scholars increasing

• Note: They don’t reflect entire situation of this field.
  • There are many other communities.
• It is suitable for an observation from a stable viewpoint.
1994-1998

Presenters in the quarterly workshop of SIG-Computers and Humanities
(Assisted by Asuka Ota)
1999-2003

Presenters in the quarterly workshop of SIG-Computers and Humanities
(Assisted by Asuka Ota)
2004-2008

Presenters in the quarterly workshop of SIGComputers and Humanities
(Assisted by Asuka Ota)
Presenters in the quarterly workshop of SIG-Computers and Humanities
(Assisted by Asuka Ota)
Each target field of the presentations in SIG-CH: Top 10

Each target field of the Presentations

<table>
<thead>
<tr>
<th>Year</th>
<th>Percentage of fields of presentations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1989-</td>
<td>0.0</td>
</tr>
<tr>
<td>1994-</td>
<td>5.0</td>
</tr>
<tr>
<td>1999-</td>
<td>10.0</td>
</tr>
<tr>
<td>2004-</td>
<td>15.0</td>
</tr>
<tr>
<td>2009-</td>
<td>20.0</td>
</tr>
</tbody>
</table>
Method adopted in the presentations in SIG-CH

- Analysis
- Digitization
- Structuralization
- Suggestion
- Tool development
- System development
- Teaching
Types of DH-like research

• Archiving
  • Making digital cultural resources
  • Providing digital cultural resources

• Analyzing
  • Analyzing digital cultural resources
  • Analyzing human (or research ) activities in digital age

• Representing
  • Making exhibition using digital cultural resources
Our responsibility for you:

• Providing:

  • useful digital cultural resources which deserve to be adopted in academic research.

  • Information of reliable digital methods to analyze Japanese digital cultural resources.

  • ...while dealing with them as our research results.

  • But...
Our difficulties in general:

• Getting budget, such as research grants
  • Budget stopped, gradually everything disappears
  • Accountability is seriously required by founders and administrators.

• Surveying global / North American trends and standards for digital cultural resources
  • General style of usage in North American libraries such as JSTOR
  • Related de-facto standards such as Unicode, TEI, Discovery services
  • Often a brand-new domestic format had been invented...

• Providing useful information in English
Our difficulties in general:

• Getting budget, such as research grants
  • Budget stopped, gradually everything disappears
  • Accountability is seriously required by founders and administrators.

• Surveying global / North American standards for digital cultural resources
  • General style of usage in North American libraries such as JSTOR
  • Related de-facto standards such as Unicode, TEI, Discovery services
  • Often a brand-new domestic format had been invented…

• Providing useful information in English

I would like to ask you: (1)

⇒ To build a system of evaluating of digital resources which built by researchers, such as award or so.

⇒ It is very helpful for our accountability
Our difficulties in general:

- Getting budget, such as research grants
  - Budget stopped, gradually everything disappears
  - Accountability is seriously required by founders and administrators.

- Surveying global / North American trends and standards for digital cultural resources
  - General style of usage in North American libraries such as JSTOR
  - Related de-facto standards such as Unicode, TEI, Discovery services
  - Often a brand-new domestic format had been invented…

- Providing useful information in English

To help to get information about such things because they need in early stage of project, as it is difficult to change anything after formulating strategy and policy.
Our difficulties in general:

• Getting budget, such as research grants
  • Budget stopped, gradually everything disappears
  • Accountability is seriously required by founders and administrators.

• Surveying global / North American standards for digital cultural resources
  • General style of usage in North American libraries such as JSTOR
  • Related de-facto standards such as Unicode, TEI, Discovery services
  • Often a brand-new domestic format had been invented…

• Providing useful information in English

I would like to ask you: (3)

To help to write useful information in English and to gather it on any convenient place such as Wiki-like site if you can.
Change the situation...?
Change the situation...?

North American Libraries

Japanese/East Asian Libraries

Japanese Practitioners

It will be better to...
An example of efforts in a project (1)

- The SAT Daizōkyō Text Database Committee
- Since 1994, led by Professor Masahiro Shimoda
- Published a Buddhist text database including over 100,000,000 Han characters on Web in 2008
- Linking other digital resources so that users can use it easily in English
  - Search Chinese/Japanese texts in English term
  - Browse English terms and aligned English sentences with Chinese texts
- Has used Web collaboration systems since 2005
- Has been incubated in both domestic and international DH communities since its early stage.
長阿含経 (No. 0001 佛陀耶念譜 筋佛念譜) in Vol. 01

12. 後秦弘始九年那揭陀共竺佛念譜

太阿含経 (No. 0001 佛陀耶念譜 筋佛念譜) in Vol. 01

12. 後秦弘始九年那揭陀共竺佛念譜

電子仏教辞典

入於涅槃

入於涅槃 to enter nirvāṇa (rù yu niépán)

BDK-SAT parallel corpus

Search results in BDK-SAT corpus:

1. 我今於中夜當入於涅槃
   → Today I will enter nirvana...
2. 佛語善男子當入於涅槃而捨於菩提之道
   → Although attaining the enlightenment...
3. 名聞善言者漸入於涅槃
   → Their fame will spread universally...
4. 我所作業當入於涅槃
   → When the time I have to do is...
An example of efforts in a project (2)

• Proposing Han characters which are not encoded in Unicode to ISO/IEC 10646 (IRG and WG2).

• Proposing variant characters of Siddham script which lacks in current Unicode to ISO/IEC 10646 (WG2).

• Attempting to form a group for eastern text encoding adopting Text Encoding Initiative Guidelines) with related DH researchers.
ISO/IEC JTC1/SC2/WG2/IRG
Ideographic Rapporteur Group

Breaking News:

IRG#41 was held in Tokyo, Japan (2013.11.18 - 2013.11.22)

(Documents for IRG#41, actively updated from 08/08/2013)

IRG Principle and Procedures (Summary/Submission Form, Blank Submission Form, IMDS update Flow Chart)
(Latest Version is IRGN 1975, confirmed at IRG41)

IRG Meetings:

<table>
<thead>
<tr>
<th>Meeting</th>
<th>Document Range</th>
<th>Location</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>42</td>
<td>IRGN1981-</td>
<td>Qingdao, China</td>
<td>2014.06.19-2013.05</td>
</tr>
<tr>
<td>41</td>
<td>IRGN1980-IRGN1980</td>
<td>Tokyo, Japan</td>
<td>2013.11.18-2013.11</td>
</tr>
<tr>
<td>40</td>
<td>IRGN1923-IRGN1949</td>
<td>Hong Kong, China</td>
<td>2013.05.20-2013.05</td>
</tr>
<tr>
<td>38</td>
<td>IRGN1810-IRGN1879</td>
<td>Gyeonggi, Korea</td>
<td>2012.06.18-2012.06</td>
</tr>
<tr>
<td>37</td>
<td>IRGN1757-IRGN1844</td>
<td>Mountain View, California, USA</td>
<td>2011.07.01-2011.12</td>
</tr>
<tr>
<td>36</td>
<td>IRGN1710-IRGN1798</td>
<td>Chongqing, China</td>
<td>2011.04.11-2011.04</td>
</tr>
<tr>
<td>35</td>
<td>IRGN1873-IRGN1748</td>
<td>Macao, China</td>
<td>2010.11.08-2010.11</td>
</tr>
<tr>
<td>34</td>
<td>IRGN1617-IRGN1710</td>
<td>Nagasaki, Japan</td>
<td>2010.06.21-2010.06</td>
</tr>
<tr>
<td>33</td>
<td>IRGN1580-IRGN1564</td>
<td>Ho Chi Minh, Vietnam</td>
<td>2009.11.23-2009.11</td>
</tr>
<tr>
<td>32</td>
<td>IRGN1433-IRGN1606</td>
<td>Hong Kong, China</td>
<td>2009.06.15-2009.06</td>
</tr>
<tr>
<td>31</td>
<td>IRGN1450-IRGN1567</td>
<td>Kunming, China</td>
<td>2008.11.10-2008.11</td>
</tr>
</tbody>
</table>
Welcome to the official home of

JTC1/SC2/WG2 - ISO/IEC 10646 - UCS

Latest update: 2014-03-18

Latest Contributions can be accessed via the WG document register or anonymous ftp ftp://std.dkuug.dk/ftp.anonymous/JTC1/SC2/WG2/docs

---

N4571 4th edition delta fdis charts
N4570 Subdivision of work for PDAM2
N4569 PDAM2
N4568 DAM1 Content
N4567 Background for power symbols document N4566
N4566 Unicode Liaison Report to WG2
N4565 Approved Disposition of Comments Amendment
N4564 Approved Disposition of Comments DIS 4th edition
N4563 Annex S Source Separation example
N4562 Feedback on Nushu N4566
N4561 Nushu ad hoc report
N4560 Siddham ad hoc report
N4559 Not Used
N4558 Request of Changing One CJK C Source and Grammatical Document
N4557 Brief summary Siddham
N4556 Reply to N4513 and N4533 Comments on Nushu
N4555
N4554 Resolutions - meeting 62
N4553 Draft Minutes - meeting 62

---

Figure 1: Differentiation of Vowel Sign U in the Mantra for Dhalsaiya (多摩伽婆底) (Liu, Tzuhsi, Taiho Vol. 78, Text No. 2499, p. 605 (left) and p. 610-611 (mid-right)).

Figure 2: Another Semantic Differentiation of Vowel Sign U in Single Document Byakuhoush (by Chouen, 13C), Taiho Vol. 56 (Tzuhsi Vol. 10), p. 310

2. Co-Usage of Vowel Sign U U Variants

Figure 3 and Figure 4 are submitted in L2/14-055. For the detailed analysis of their semantics are described in L2/14-055.
Additional information (1):

- SIG-Transcribe JP in JADH started “翻デジ2014” with support from Japanese National Diet Library to transcribe its digital images by use of Omeka so that users can search Japanese texts.
Brief summary of “翻デジ2014”

• Transcribing the image files in Japanese or Chinese texts in digital collection of the library.

• “翻デジ2014” will play a role of incubating and supporting transcribing projects:
  • Providing a function of collaborative transcription to the public.
  • Transcribed text data should be public domain in general.
  • Collaborators should transcribe it manually so far.
  • Planning to combine OCR with manual correction and semi-automated correction which will be provided by “Crowd4U” project (http://crowd4u.org/).
If you would be interested in it, please contact me!
Additional information (2):

• As you may know:
  - National Institute of Japanese Literature has gradually been starting a large digitization project of Japanese cultural materials.
    - But the institute gave up to transcribe them.
      - A typical Japanese weakness...
    - The project aims to make digital facsimiles.
    - Please attend to the project with your warm eyes.
      - And give me any advices if you would find something.
Additional information (3)

• JADH conference will be held at University of Tsukuba, Ibaraki, Japan, **19-21, September**.
  • "Bridging GLAM and Humanities through Digital Humanities“
  • Deadline of CFP is **7 May**

• Please submit your research or practice!
Additional information (4)

• If you are able to read Japanese and interested in Japanese DH situation, please see “Digital Humanities Monthly” http://www.dhii.jp/DHM/
  • Essays of researchers and practitioner of DH
  • DH-related event report
  • And so on...