2-1-1999

The University of Virginia/University of Pittsburgh Japanese Text Initiative

Sachie Noguchi

Follow this and additional works at: https://scholarsarchive.byu.edu/jeal

BYU ScholarsArchive Citation
Available at: https://scholarsarchive.byu.edu/jeal/vol1999/iss117/5

This Article is brought to you for free and open access by the All Journals at BYU ScholarsArchive. It has been accepted for inclusion in Journal of East Asian Libraries by an authorized editor of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
THE UNIVERSITY OF VIRGINIA / UNIVERSITY OF PITTSBURGH JAPANESE TEXT INITIATIVE*

Sachié Noguchi  University of Pittsburg

1. Introduction

On December 5, 1995, the University of Virginia (UVa) Library Electronic Text Center and the University of Pittsburgh (Pitt) East Asian Library jointly announced on the Internet the opening of their “Japanese Text Initiative (JTI)” and the first text of this initiative, Ogura Hyakunin Isshu available on the World Wide Web (WWW or Web) at http://etext.lib.virginia.edu/japanese/). In December 1997 seventeen new titles (thirteen of them Noh plays) and in November 1998 eleven new titles were added, making a total of twenty nine titles available on the Web. Since its official opening in 1995, use of the site has grown steadily: the total of 32,687 hits from 14,643 accesses by 2,432 unique hosts per day during January 1998 was nearly double the 16,906 hits (requests to the server) on JTI during the month of January 1997.

This paper discusses JTI, a collaborative effort to make texts of classical Japanese literature available on the Web by two ARL (Association of Research Libraries) member university libraries, focusing on its development and technical issues.

2. The Electronic Text Center and the Japanese Text Initiative

UVa Library opened its Electronic Text Center as a humanities computer service in August 1992. The Electronic Text Center has on the Web over 20,000 texts in English, French, German, Latin, Japanese, Chinese, Russian, and other languages. All of the texts are tagged in Standard Generalized Markup Language (SGML) according to Text Encoding Initiative (TEI) guidelines. For Web display, PERL (Practical Extraction and Report Language) filters convert the SGML tags to Hypertext Markup Language (HTML).

The Japanese Text Initiative is a part of the online library of the Electronic Text Center at UVa Library. The Japanese texts selected for the JTI are added to the Center’s service for research based on the same principles of the Center which are applied to other language materials. For instance, Japanese Noh texts are marked up by SGML-TEI and are displayed in Japanese, can be searched and analyzed, and can serve other research purposes in the same way as works by William Shakespeare.

In spring 1995, at Pitt, the University Library System had a plan to implement its own electronic text project and the Electronic Text Project Team was established. The Team, of which the author was a member, visited the UVa Library Electronic Text Center to learn from UVa’s

*This article is a revised version of a paper originally presented in Washington, D.C. on March 26, 1998, at the meeting of the Committee on Library Technology of the Council on East Asian Libraries, Association for Asian Studies.
endeavors and experiences. During this visit the idea of the Japanese Text Initiative, a collaborative effort, was proposed.

3. **Scope and Goals**

3.1 **Scope**

The JTI intends to put online on the Web texts of classical Japanese literature in Japanese characters. Its primary audience is English-speaking scholars and students. Where possible, the Japanese texts will be accompanied by English translations. All JTI texts will be tagged in SGML, according to TEI standards, and converted to HTML for display on the Web. An important purpose is to make JTI texts in both Japanese and English searchable, both individually and as a group.

Generally, the electronic texts at the Electronic Text Center site are not intended to be substitutes for authoritative printed editions. Copyright texts are not put on the Web, unless permission is granted from the copyright holders. With some exceptions, the texts of current scholarly editions of literature in English, French, German, Japanese, and other languages are not freely available for sites such as JTI. Because of this, texts without copyright constraints, such as editions for which copyright has expired, are employed. These electronic texts are therefore useful as supplements to current scholarly editions. What the electronic texts add to those editions is the capability of rapid searching for words and characters, which would take considerably longer in printed texts.

3.2 **Goals**

In the short term, most or all of the Twenty Classical Works in J. Thomas Rimer's *A Reader's Guide to Japanese Literature* (New York: Kodansha, 1988) are intended for online inclusion. A longer-range goal is to add pre-twentieth-century works such as *Shin Kokinshu* and the other anthologies of the *Hachidaishu*. It is also expected that twentieth-century literature that does not have copyright restrictions will be added.

3.3 **Advisors**

Advisors for the Japanese Text Initiative are Professors J. Thomas Rimer and Mae J. Smethurst of the University of Pittsburgh and Professor Lewis Cook of Queens College.

4. **Japanese Texts in the JTI**

By spring of 1993, UVA acquired the Japanese texts *Ogura Hyakunin Isshu* and *Genji Monogatari*. Because of its literary importance, cultural permeation in Japanese society, and the length of the text, *Ogura Hyakunin Isshu* was selected as the first Japanese text of the JTI. Currently there are twenty nine texts: *Man' yoshu; Kokinshu; Ise Monogatari, Sarashina Nikki, Hojoki, Ogura Hyakunin Isshu, Tsurezuregusa*, thirteen Noh plays, Ihara Saikaku, *Koshoku Gonin Onna; Saikaku, Koshoku Ichidai Onna; Matsuo Basho, Oku no Hosomichi; Chikamatsu, Sonezaki Shinju; Chikamatsu, Shinju Ten no Amijima; Ueda Akinari, Ugetsu Monogatari; Kobayashi Issa, Ora ga Haru; Mori Ogai, Gan; Yosano Akiko's Midaregami.*
4.1 Authoritative Texts and Copyright

When possible, it is better for the electronic version of the text to be based on some authoritative printed editions or manuscripts, and copyright law must be observed. In working on *Hyakumin Isshu*, it was not an easy task to find pre-1920 Japanese literature publications in libraries of North America. Even when the copies of the publications are found in libraries, they are not always available through interlibrary loan, for example because of loss or physical fragility of the book. In the case of Noh plays, two versions, *Yokyoku Hyoshaku* (Chosha, Owada Tateki. Tokyo : Hakubunkan, Meiji 40 [1907]) and *Yokyoku Sosho* (Haga Yaichi, Sasaki Nobutsuna kochu. Tokyo : Hakubunkan, Taisho 3-5 [1914-1916] 3 vols.) were examined, and the former was chosen. To get all the texts of thirteen titles selected from *Yokyoku Hyoshaku* for the JTI, interlibrary loans had to be made from the collections of the University of California at Los Angeles, Columbia University, and the University of Washington, Seattle.

The use of older texts causes a number of problems such as language in the text: antiquated expressions (especially in the use of *kana*), different fonts, unavailable characters, old publication formats, etc. One can argue that using an old text itself provides an educational opportunity and has merit but for modern readers it is more difficult to use compared with copyrighted newer publications which are modernized and much easier to read.

4.2 Character Sets and Fonts

Ideally, electronic texts should be put on the Web faithfully following the original authoritative texts. However, the printed texts which were published before 1920 include many characters and fonts which are not available in today’s character sets and fonts employed by Japanese word processors. For example, in *Yokyoku Hyoshaku*, the character “ao = blue” appears in a different font from the one which is used today. It is necessary to compromise by substituting an available font and/or modernized characters in such cases. In addition to the problem of font, there are characters missing or unavailable. In the thirteen Noh texts, there was only one character that is absolutely not available in Japanese character sets, and *hiragana* was substituted for it. UNICODE promises to standardize and thus ultimately to simplify the encoding of multiple languages at the character level although some scholars claim that UNICODE does not have sufficient space to include all characters necessary for classical Japanese texts. The issue of missing characters is a major problem for electronic texts.

4.3 Readings, Furigana, English Translation

The intended primary audience of the JTI is English-speaking scholars and students. To benefit these audiences, where possible, the JTI provides readings of Chinese characters and modern transliteration of archaic expressions: a transliteration of Noh plays in *hiragana* and *Ogura Hyakunin Isshu* in romanization. Romanization is based on the revised Hepburn system and word division employed by the Library of Congress. Macrons were not used in *Ogura Hyakunin Isshu* because of display and search problems. For transliteration and reading, various reference sources, such as *Nihon Koten Bungaku Daijiten* and *Nihon Bungaku Taikei* (both published by Iwanamai Shoten) and *Kokushi Daijiten* (Yoshikawa Kobunkan) were consulted.
In addition to transliteration and reading, where possible, the Japanese texts are accompanied by English translations. *Ogura Hyakunin Isshu* is accompanied by a modernized version of MacCauley's translation. Noh plays incorporate *furigana*, together with one or more English translations of each play by authors ranging from Pound and Waley to Tyler, Brazell, and other recent translators thanks to generous permission granted by Columbia University Press and the Cornell East Asian Program. Each play is displayed in frames with parallel Japanese and English. Markers in the texts allow users to locate a passage in the Japanese version, click on a marker, and display the English translation of the same passage in the other frames. For *Hagoromo*, for example, users can compare the translations of Pound, Waley, and Tyler with the original Japanese on the same Web page. For *Ogura Hyakunin Isshu*, the Japanese text, romanization, and English translation of each poem can be displayed on the same Web page.

### 4.4 SGML-TEI Markup

All Japanese texts of the JTI are encoded with SGML-TEI. Noh texts are marked up according to the TEI “Drama” tag sets. The TEI guidebook is written in English and most of the examples employed in the guidebook are dramas written in western languages. To apply these guidelines derived from western dramas to markup Japanese Noh plays required adaptation to the Japanese texts. There are a number of elements in Noh plays which have not been marked up in the JTI texts; these need to be developed in the future.

The SGML tags operate both as a series of database fields - descriptive markup - that are invaluable when constructing and limiting searches, and also as a set of typographic and layout instructions for the appearance of the text on the screen. For documents of limited complexity, HTML, rather than SGML, can be an effective authoring environment; however HTML seriously limits the ways in which a more complex document or a set of documents can be used. Besides, the Web lacks inherent document management or document access capabilities. To take advantage of the complexity of the documents and yet make them available through the Web, a gateway from the Web to an indexed collection of texts in an SGML-aware text retrieval program system, PAT, was developed using the Web’s CGI (Common Gateway Interface). The texts are all held on the server in forms of SGML more sophisticated than HTML, and are converted to HTML “on the fly” using a PERL script in the gateway that is triggered each time a user browses a text or searches a database (Seaman 1996, Price-Wilkin 1994B and 1994C).

Text processing and retrieval software [behind the Web] is important for literary computing. PAT developed by Open Text is one of the examples of these software which include products that support Boolean searching, thesaurus generation, and a variety of forms of information retrieval, such as WordCruncher (Johnston & Co.), askSAM (askSAM Systems), ZyIndex (ZyLAB), Lector (Open Text), and TACT (a shareware program from the University of Toronto) (Sutton 1994).

### 4.5 Interactive Searching

Each text (Japanese, romanization, and English) is fully searchable by using its respective script. An interactive search interface lets users search online in Japanese, English, or romanization, whenever provided, for any characters or words in individual texts or in all the texts together. The interface also supports proximity searching of *Kanji* or *kana*: for example, look for the *kanji*
for “aki” within 40 characters of the kanji for "kaze," for example, and users will get 39 hits as of December 1997 in Kokinshu, Matsukaze, Semimaru, Midaregami, etc.

Interactive searching is enabled by the above mentioned Web-to-PAT Gateway which takes a four step modular approach to access text collections. The four steps are:

1. FORM Handling
   Users, with the aid of the FORM, submit a query.

2. CGI Query Handling
   The Query is received and Translated to a PAT search.

3. PAT Result Handling
   Information from PAT is transformed into lists or entries that can be selected.

4. SGML-to-HTML Filtering
   The richer SGML is transformed into HTML. (Price-Wilkin 1994C p. 8-9)

The Gateway using an HTML FORM to handle query submission supports a range of queries from simple to complex, from word and phrase searches to browsing of the document in a variety of ways, such as by chapter or other section.

4.6 Other features
4.6.1 Images

The electronic image of the source material can be reproduced, enhanced, or rotated, but not otherwise manipulated by a program (Hockey 1996). For users who do not have a Japanese language client which enables them to read Japanese text on the Web, the Japanese text displays images of Ogura Hyakunin Isshu that were provided in addition to searchable character texts. The image text itself cannot be searched or otherwise manipulated. Since images require large disk space for storage and various means to display Japanese text have been recently developed, Ogura Hyakunin Isshu is only the text which is accompanied by images.

For Ogura Hyakunin Isshu, the woodblock print of the poem in Onna Kotobuki Ogura Shikishi (Osaka: Keio 2 [1867]) is provided by clicking on the number of any poem in the text pages of the Web edition. This is a late Edo-era edition of the Hyakunin poems designed for women, with various useful cultural material, Genji seals, origami directions, poetry, and notes on morals and etiquette. These images are included with this edition not only as a delightful version of the 100 poems and cultural artifact from the nineteenth century, but also because of the portraits of the poets. These somewhat rough-hewn woodblock prints from the latter days of the ukiyo-e age show the one hundred poets in typical iconography. The same configurations appear and re-appear in illustrations in books and on playing cards of the Hyakunin Isshu.

4.6.2 Uta Karuta Game

In Japan a traditional card game often played at New Year’s, Uta Karuta, is based on Ogura Hyakunin Isshu. To simulate the Uta Karuta game on the computer, a poem is provided at
random and the first three lines (the “Kami-no-ku”) displayed. Users then try to say to
themselves the last two lines of the poem (the “Shimo-no-ku”). To verify whether a user
remembered correctly, the whole poem will be provided if the “See Entire Poem” box is clicked.

5. **Merits of the Collaborative Project**

The major benefit of the JTI has been the relatively short time in which it was possible to put our
first texts on the Web. That allows anyone, within reach of the equipment, anywhere in the
world to reach Japanese texts anytime. If one of the two participating libraries, UVa without a
Japanese studies librarian and Pitt without special technique or experience of electronic texts,
had tried to do it alone, it might have taken many years before it would be launched, if the entire
effort had not been made impossible.

It is also meaningful that two ARL libraries took on such an initiative. By introducing JTI, we
expanded our opportunities of forming networks of people who are engaged and/or interested in
Japanese electronic texts, learning from them, and cooperating with them.

In addition to these institutional benefits, the author, who was a part of this project under the
guidance of the UVa, benefited enormously from the rich learning experience as an individual as
well as a professional.

6. **Conclusion**

The article “Konna ni Aruzo, Nihon Bungaku Denshi Tekisuto” (*Hon to Konputa* No. 3 [Winter
1998], published in Japan) reports on a survey of Japanese literary works available as electronic
texts on the Web. The survey found that there are more than 400 texts available from 83 sites
including the JTI. Nevertheless, reading and writing Japanese characters on the Web is still an
art, not a science. With the 4.0 versions of Netscape and Internet Explorer, reading *Kanji* and
*kana* characters has become much easier. But inputting characters to Web forms is still a tricky
and frustrating business.

In her introduction to the special issue on electronic texts of *Information Technology and
Libraries*, Katharina Klemperer rightly points out the analogy between electronic text practices
in most libraries today and OPACs (online public access catalogs) in the early 1980s: “Just as the
pioneering OPAC institutions were then starting to convert their card catalogs, the electronic text
pioneers, . . . , are now starting to build their own text collections, often “converting” and marking
up the texts on their own. . . . The OPAC movement was supported by the MARC [Machine
Readable Catalogue] standard; e-texts can rely on the Standard Generalized Markup Language
(SGML)” (Klemperer p. 6). This analogy makes it easier for us to understand the electronic text
environment. By extending this analogy, the JTI can be understood as the CJK records in
OPACs and/or bibliographic utilities.

Electronic texts have been used for scholarly research in the humanities for the past fifty years or
so ever since Father Roberto Busa began work on his *Index Thomisticus* in 1949 (Busa 1950). In
the humanities, as in other disciplines, electronic resources offer many more possibilities than
print, but, in general, libraries do not yet have any well-established practices for collecting and
handling electronic texts as they have had with print material (Hockey 1994). As David Seaman
points out, "an electronic text initiative belongs in a library because it is a textual as much as a technical endeavour, although these two are sometimes difficult to separate. Certainly, the textual, bibliographic, and educational skills needed to evaluate, prepare, and present electronic texts to users unfamiliar with such services are all found in libraries" (Ream, p. 7). I hope more librarians become aware of the equal importance of electronic texts and join in the exciting challenges they bring to the world of scholarship.

REFERENCES


<URL: http://www.dlib.org/dlib/july97/07seaman.html>


**NOTES**

1. Full access to the Japanese texts requires front end processor software that can read and input Japanese characters (a Japanese language client) in EUC code. Also required is a Web browser that can read tables and forms, such as at least Netscape 3.0 or more recent versions of Netscape or Microsoft Internet Explorer.

2. The Electronic Text Project Team approved the JTI proposal under the condition that Pitt would become the miller site of the JTI when it was established. This had not occurred as of August 1998. The Team itself was dissolved in 1995.


4. To implement the macron, “entity references” must be employed.

5. Kana which are written usually immediately to the side or above the side of the respective Chinese character to show the pronunciation or reading.