An Attempt at Crowd-sourced Transcription in Japan

Although at least a decade has passed since emergence of the digital environment as an infrastructure, aside from a small handful of exceptional projects (such as the Aozora-Bunko,[1] BCCWJ,[2] the SAT project,[3] and so on) Japanese textual materials have not yet been encoded, having only been digitized to the stage of electronic facsimile. Most digital facsimiles still await further treatment from someone who takes a particular interest in them, but even in this case, the method is often nothing more than a search for rough metadata. OCR is efficient in transcribing recent publications, but it doesn't work well in the case of typesetting in the out-of-copyright publications, which, mainly being published before World War II, contain many traditional forms of Chinese and Japanese characters, and are composed in old typefaces. In addition, it is generally too difficult to distribute Japanese TEI files.

Under this circumstance, it is necessary to transcribe such resources in order to search and analyze Japanese materials as texts, which have been accumulated in the modern era when Japanese publishers started printing using typography. Digital facsimiles of such texts, numbering more than 400,000, are already available on the web site of the National Diet Library (NDL) of Japan. The NDL has already been attempting to transcribe them experimentally using OCR because a partial revision of the copyright law has now allowed the NDL use OCR, even though OCR is prohibited except for private use in Japan. But these attempts have not been sufficiently effective due to the complexity of the Chinese ideographs contained in the type.

Therefore, it seems that it would be better to manually transcribe them through a kind of crowd sourcing. The JADH began to organize this kind of activity under the project name "Transcribe JP" with the next-generation laboratory of the NDL. The International Institute for Digital Humanities serves as the office of the JADH, and the NDL will provide server space, software, and some technical support for the management of the project. This project has been supported by the National Institute of Informatics since April. It plans to cooperate with the Crowd4u project to correct OCR errors.

---

[1] Aozora-Bunko distributes over 10,000 texts which were transcribed by volunteers in the public domain like Gutenberg project. http://www.aozora.gr.jp/

[2] BCCWJ (Balanced Corpus of Contemporary Written Japanese) is a balanced corpus which consists of one hundred million words of contemporary written Japanese and is encoded in an original XML format. http://www.ninjal.ac.jp/english/products/bccwj/

[3] The SAT project manages a Web service which utilizes a series of Buddhist scriptures including about 100,000,000 Chinese characters in order to support readers of Buddhist texts.

One possible approach would be that of separating the system into two parts. One would be a crowd-sourcing site that enables everyone to transcribe the resources. The other is a set of Web pages that could be scraped various by Web crawlers such as Googlebot and the NDL's own search system. The former system would include not only manual transcription but also an OCR effort that would presuppose correction by a kind of micro-crowdsourcing with the support of the Crowd4u project. Once a text was transcribed at the level of book or a chapter in the former system, it would then be sent to the latter system. The judgment as to whether or not it should be sent to the latter system would be made by a volunteer group. The benefit for this project is not only the opportunity contribute to digital scholarship by making digital resources, but also to enhance its visibility.

From this standpoint, the Transcribe JP project has opened a web site “Hondigi 2014”[4] ("Hondigi" means "transcribe texts as digital resources") since late March 2014. This web site enables not only the transcription of digital facsimiles of the NDL by crowd-sourcing, but also to automatically generate an HTML version including linking to each facsimile and TEI version, following the Best Practices for TEI in Libraries Level 1. The system will utilize Omeka, Scripto and Mediawiki. We have had to customize the programs because this system must use external image files instead of the internal files due to its target, that is, image files of the NDL Web site, part of Japanese classical characters which consist of 4-bytes in UTF-8 demands to change setting in MySQL which is the backend of the Omeka, and adding to the result of transcription so that it can generate the HTML form and the TEI form. It has been somewhat difficult for us to change the target image files because of the time required to understand the methods of customizing the programs of Omeka and Scripto. There might be better solution, but we decided to modify not only the plugin Scripto, but also the function of handling images in Omeka itself. It was easy to enable the treatment of 4-byte UTF-8 characters in Omeka, simply by changing the value of the encoding scheme from UTF8 to UTF8mb4. As the transcribed texts are archived in Mediawiki, the program of generating the HTML and TEI files uses the Mediawiki API. Since a part of the URLs of each book are derived from an NDL permanent ID embedded in each book, the program gathers all pages of a book by use of the ID and forms them into HTML and TEI style.

As this project is quite primitive, only a few volunteers have participated in it up to now. But we aim to widen it by starting micro-crowdsourcing. It will useful to enrich the TEI files and the digital humanities.

---

[4] http://lab.kn.ndl.go.jp/dhii/

At the conference we will describe the detail of the system and discuss the possibilities of using TEI in crowd-sourcing.

(An example of transcribing a book)

This work was supported by National Institute of Informatics and National Diet Library in Japan.