Digital Libraries in Asian Languages – A TCL Initiative

Md Maruf Hasan, Kazuhiro Takeuchi, Hitoshi Isahara and Virach Sornlertlamvanich

Thai Computational Linguistics Laboratory
Communications Research Laboratory
112 Paholyothin Road, Klong 1, Klong Luang, Pathumthani 12120, Thailand
mmhasan@acm.org, {kazuh,isahara}@crl.go.jp, virach@crl-asia.org

Abstract. The Greenstone Digital Library (GSDL) system, developed by the New Zealand Digital Library (NZDL) Consortium at the University of Waikato is a suite of open-source software for building and distributing digital library collections. At the Thai Computational Linguistic (TCL) Laboratory of CRL Asia Research Center, we plan to implement and host digital libraries in several major Asian languages. In this paper, we describe our experiences in implementing Thai and Japanese digital libraries using Greenstone.

1.0 Introduction

In October 2002, CRL Asia Research Center - the Asia Pacific headquarters of Japanese Communications Research Laboratory (CRL) is established in Thailand. The Thai Computational Linguistics (TCL) Laboratory is one of the research laboratories of CRL Asia Research Center across Asia. With the aim of becoming the foothold of collaborative research on computational linguistics in Asia, TCL carries out research in Human Language Technologies, Intelligent Information Infrastructure, and Open Source Software related to language and knowledge processing. At TCL, we initiated a project on hosting digital libraries in major Asian languages.

As defined by the Digital Library Federation [1], 'digital libraries are organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities'. We need efficient digital library (DL) systems which can manage multilingual and multimedia information to achieve the above goals. We choose Greenstone Digital Library (GSDL) suite [2], developed by the New Zealand Digital Library (NZDL) Consortium [3] at the University of Waikato.

GSDL is a suite of open-source software for building and distributing digital library collections [4]. The main reasons behind choosing GSDL suite are (1) it uses Unicode [5] and XML-compliant format internally, and (2) it supports indexing of large collection of information including multimedia [6]. We developed interfaces in Thai and Japanese for Greenstone. We also aim at developing Greenstone interfaces for other Asian languages in the near future.

Unlike English, in written Japanese and Thai, words are not delimited with explicit boundaries. Japanese and Thai also have complex morphology and other unique linguistic properties. We are in the process of developing new tools for and integrating existing ones with the GSDL suite for effective processing of information in Asian languages.

We hope to build collections of electronic information in major Asian languages, which have cultural and historical values as well as collections of technical reports and thesis available in the universities and research centers written in the local languages. Our digital libraries will also host multilingual language resources such as parallel aligned corpora. Upon completion of the projects, such digital libraries will become an invaluable one-stop source of digital information ubiquitously available over the Internet for general public and researchers equally. We hope that this initiative will also circumvent digital divide in the Asian region [7].

2.0 The Greenstone Digital Library System

We choose Greenstone as the digital library suite for many reasons. Some of which are listed below:

- Greenstone runs on almost all popular computing platforms: Windows, Macintosh and popular Linux/Unix platforms. Greenstone can also be easily integrated with the two most popular Web servers: Apache and Microsoft IIS.
- The entire system is well-documented in terms of User's Guide, Installation Guide, Developer's Guide and two active mailing lists for users and developers.
- Greenstone is an open-source software suite which makes intensive use of many open source software behind the scene: for example Apache Web server, GNU database manager, *gdbm*, open-source indexing and retrieval system, *mg* [8], and several other plug-ins and utilities.
- Greenstone is capable of handling multilingual information using Unicode.
- Greenstone uses XML-compliant internal representation.
- Greenstone is capable of handling multilingual information and scalable [9].

The overview of a general greenstone system as described in the GSDL Developer's Guide [10] is summarized as follows:

Two components are central to the design of the GSDL system: "receptionists" and "collection servers." From a user's point of view, a receptionist is the point of contact with the digital library. It accepts user input, typically in the form of keyboard entry and mouse clicks; analyzes it; and then dispatches a request to the appropriate collection server (or servers). This locates the requested piece of information and returns it to the receptionist for presentation to the user. Collection servers act as an abstract mechanism that handle the content of the collection, while receptionists are responsible for the user interface.

As Figure 1 shows, receptionists communicate with collection servers using a defined protocol. The implementation of this protocol depends on the computer configuration on which the digital library system is running. The most common case, and the simplest one, is when there is one receptionist and one collection server, and both run on the same computer. However, CORBA based distributed configuration is also possible [11]. Another notable advantage of using Greenstone is its capability of empowering end-users with the power of building collection over the Internet [12]. Such a feature could be very useful in building linguistic resources collaboratively.

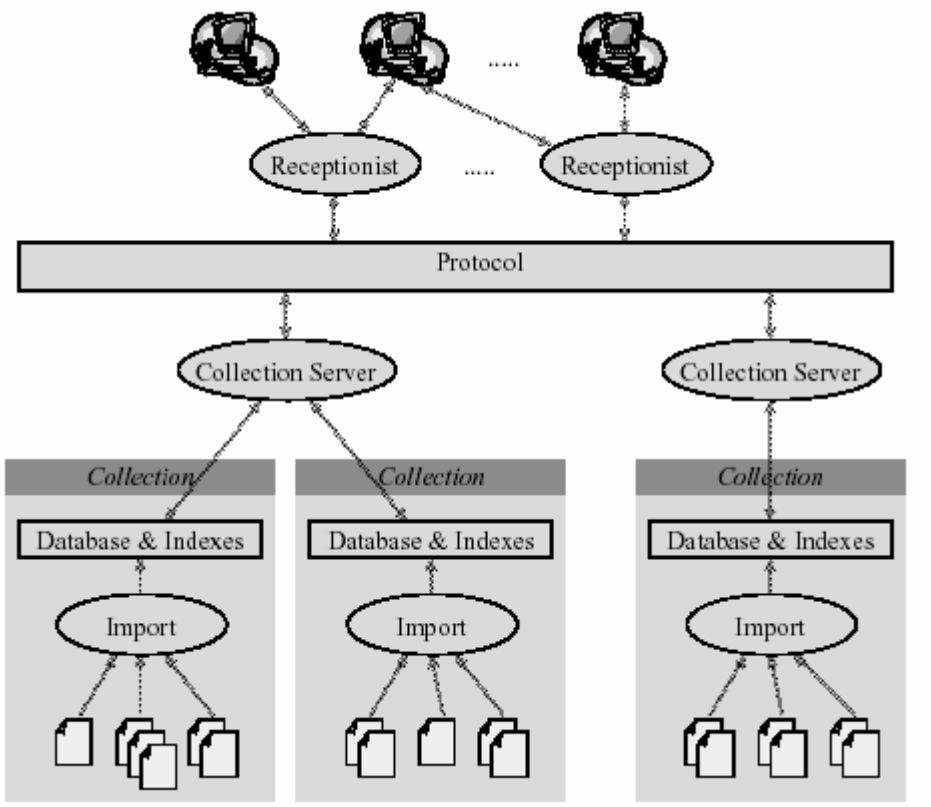

Fig. 1. Overview of a general Greenstone system *(Source: Greenstone Developer's Guide)*

It should also be mentioned here that Greenstone uses a series of open-source software within the above framework. Such software include pre-processing Plug-Ins (e.g., HTML, Image, PDF and MS-

Word Plug-Ins) and a set of other programs and scripts (e.g., collection builder, *buildcol.pl* Perl script, *mg* Indexer, *gdbm* GNU database manger and *wget* Internet crawler, etc.

In the following section, we will describe our framework for using Greenstone effectively in processing Asian language information in the context of digital libraries

3.0 Digital Libraries in Asian Language – the TCL Initiative and Framework

The Greenstone framework is a flexible digital library framework that offers plenty of freedom and advantage to work with multilingual and multimedia information. However, to use Greenstone for Asian language digital libraries we must at least address two major issues.

Firstly, we must develop the interfaces (c.f., Figure 1, *Receptionists*) for the respective languages. Secondly, we must identify the language specific issues which may interfere with effective indexing and retrieval. For some Asian languages such as Japanese and Thai, where there are no explicit word boundaries, and the morphological structures are complex, we need to develop (or integrate) respective modules to boost the accuracy of indexing and retrieval (c.f. Figure 1, *Collection Server/Indexing*).

It should be mentioned here that a straight forward indexing which may be suitable for English like European languages may cause potentially poor indexing and retrieval results for Asian languages. Our experiences with Thai and Japanese digital libraries show degraded performance in terms of precision and recall when we did not add any special measures. However, using Thai and Japanese segmentation systems in the preprocessing steps of the digital collection, did overcome some of the indexing problems. We assume that integration of proper NLP tools with GSDL core system may further improve the retrieval accuracies.

After considering the above issues, the Asian Language Digital Library Framework at TCL is therefore designed around Greenstone core system with auxiliary modules around the core (c.f., Figure 2). From the overview of a general Greenstone system as explained in Section 2.0, it is imperative that in order to use Greenstone effectively in building digital libraries in Asian languages, we can enhance the system in two stages: (1) adding the native interface for the presentations and interactions: *Receptionists*, and (2) adding linguistic processing modules for effective indexing: *Collection Servers* (c.f. Figure 1 & 2).

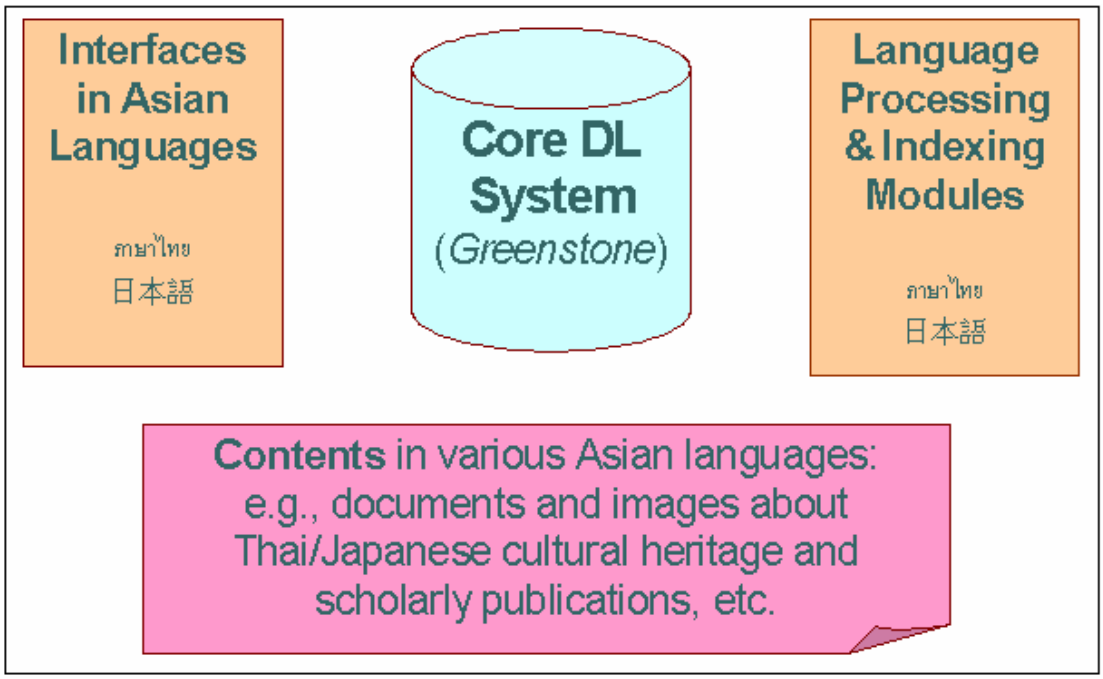

Fig. 2. Overview of TCL digital library framework for Asian languages

We invite the reader's attention to another non-trivial issue of automatic metadata extraction. In order to create rich digital library contents, metadata play crucial roles. At the moment, we have not yet initiated any work in this area. However, we foresee that automatic extraction of metadata in each language and collection is a crucial issue in building large digital libraries for that particular language

or collection. It is almost impossible to employ manual labor to create and maintain a large scale digital library using manually annotated metadata. The above framework can essentially accommodate seamless integration of such metadata extraction tools.

4.0 Current Status and Future Work

At TCL, we developed Thai and Japanese interfaces for the Greenstone suite using the macro language explained in Greenstone Developer's Guide. Figure 3 shows the screenshot of Japanese Digital Library Interface.

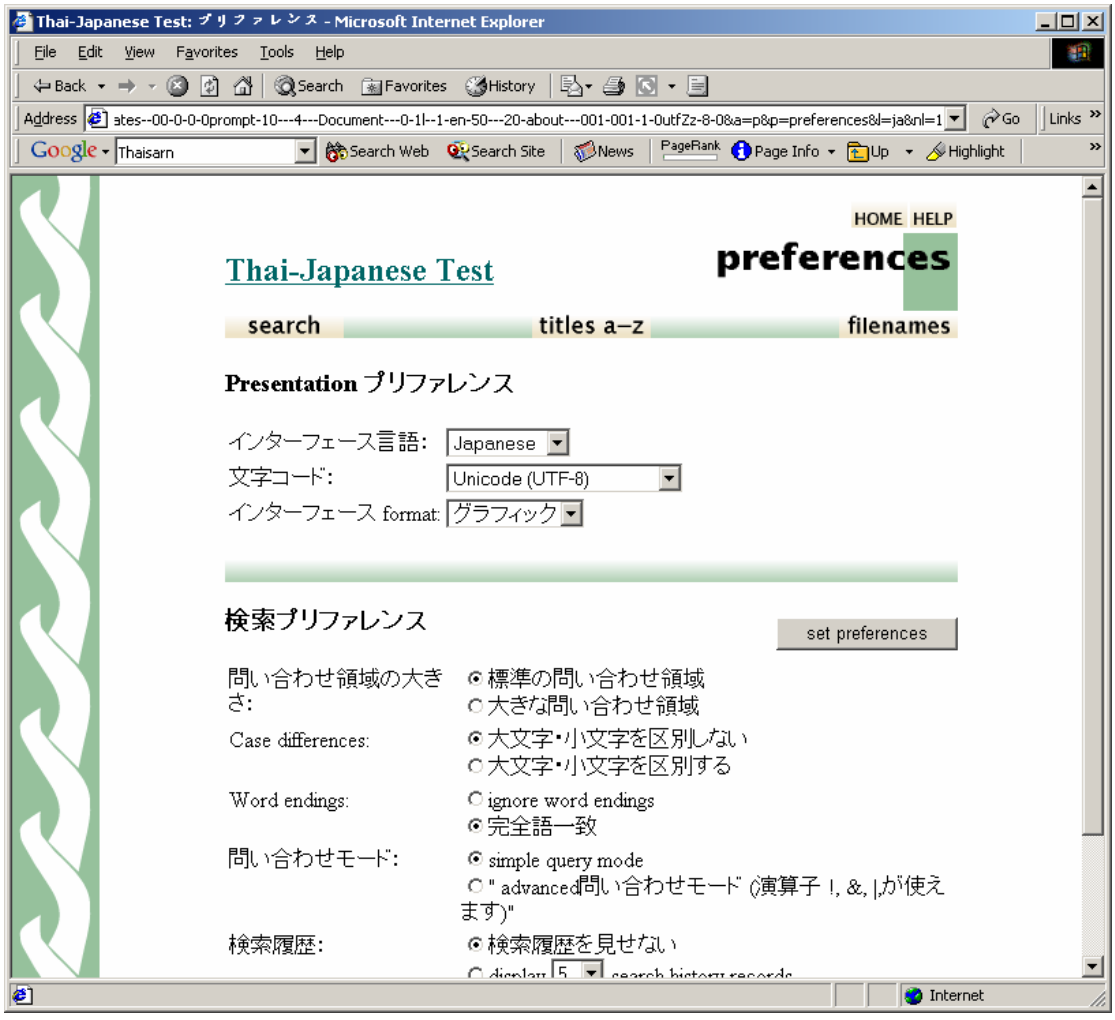

Fig. 3. Digital Library Interface for Japanese

We used un-processed Thai and Japanese texts to test the digital libraries in these languages. However, due to the absence of word delimiters, the indexing and retrieval was erroneous. We then added word delimiters using Thai and Japanese Segmentation Tools and performed indexing and retrieval to overcome such problems. We are confident that by adding language specific modules and tools with the existing Greenstone suite we will make the GSDL system effectively useable with Asian languages. We are in the process of integrating *Chasen* [13], a Japanese morphological analyzer, and the Thai Segmenter, *Swath* [14] developed at NECTEC with the core GSDL suite.

We are also searching for digital contents (which can be used without violating copyrights) in Asian languages which can be placed under the DL systems at TCL, and can be made available to the general public and researchers. Once the above-mentioned integration task is complete and significant amount of digital contents are acquired, we will make the digital libraries in Asian languages available on the Internet for broader range of users.

The following figure (Figure 4) is a screenshot of the digital library where Thai digital library interface is used with bilingual (English/Japanese) query terms; keywords are highlighted in the retrieved documents.

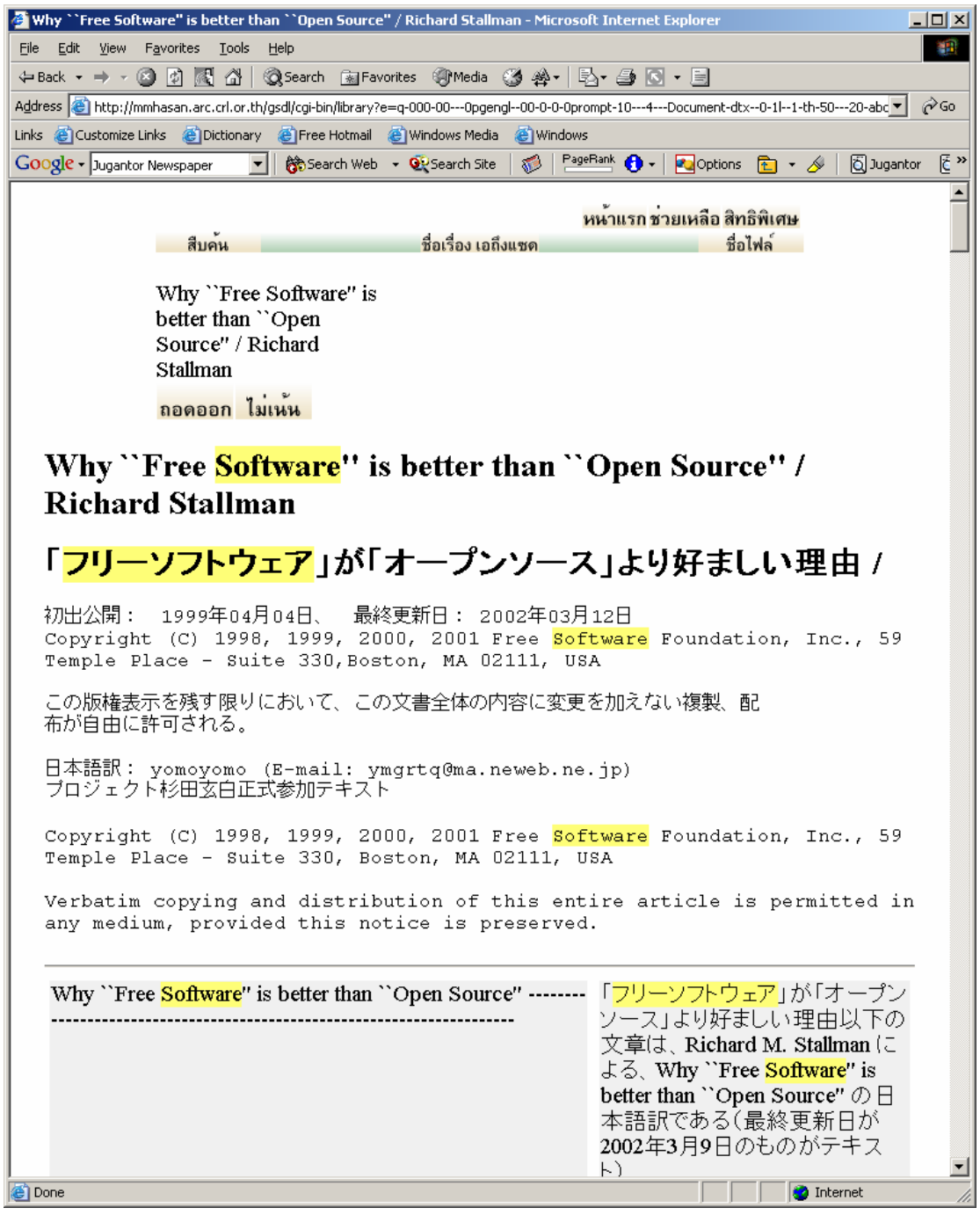

Fig. 4. Screenshot of the Thai Interface and English-Japanese bilingual content search

As for the future work, we are focusing on developing tools which can automate the process of metadata extraction in the context of digital library.

5.0 Conclusions

We warmly invite the participants of ICADL-2003 and others to join hands with us by providing digital contents in Asian languages and by jointly developing tools for processing those contents under the Digital Library Framework initiated by TCL.

We are especially interested in acquiring digital contents which reflect Asian values, culture and heritage. We are also interested in networking with the universities and research institutes in Asia which can provide us with scholarly digital documents (thesis and technical reports, multilingual corpora and NLP tools, etc.) for making them available publicly through our digital libraries.

Acknowledgements

The first author likes to acknowledge the financial supports from Communications Research Laboratory (CRL), Japan in terms of a visiting fellowship to work at TCL, Thailand. Thanks to Kazuhiro Takeuchi, Thatsanee Charoenporn, Phanicha Phavananun, Woranuch Wasinanont and Nartdanu Suttinon for their helps in translation and proof-reading in Japanese and Thai.

We also thank and acknowledge the continuing help and support we are receiving from the GSDL development team. Among them, Michael Dewsnip deserves special thanks for his timely reply of our e-mail queries with the right answers.

References

[1] Digital Library Federations:
http://www.diglib.org/

[2] Greenstone Digital Library Suite:
http://www.greenstone.org/

[3] New Zealand Digital Library Consortium:
http://www.nzdl.org/

[4] Witten, I.H., David Bainbridge and Stefan J. Boddie (2001), Open Source Digital Library Software.
http://www.dlib.org/dlib/october01/witten/10witten.html

[5] Unicode Home Page:
http://www.unicode.org/

[6] New Zealand Digital Library music library
http://nzdl2.cs.waikato.ac.nz/cgi-bin/gwmm?c=meldex&a=page&p=coltitle

[7] Witten, I.H., Loots, M., Trujillo, M.F. and Bainbridge, D. (2001) "The promise of digital libraries in developing countries." Comm. ACM, Vol. 44, No. 5, pp. 82-85.
http://www.acm.org/pubs/articles/journals/cacm/2001-44-5/p82-witten/p82-witten.pdf

[8] Witten, I.H., Moffat, A. and Bell, T.C. (1999) Managing Gigabytes: Compressing and Indexing Documents and Images. Morgan Kaufmann, San Francisco, CA.
http://www.cs.mu.oz.au/mg/

[9] Paynter, G.W., Witten, I.H., Cunningham, S.J. and Buchanan, G. (2000) "Scalable browsing for large collections: a case study." Proceedings of the Fifth ACM Conference on Digital Libraries, San Antonio, TX, pp. 215-223.
http://www.acm.org/pubs/articles/proceedings/dl/336597/p215-paynter/p215-paynter.pdf

[10] Bainbridge, D., Dana McKay and Witten I.H., Greenstone Digital Library Developer's Guide, http://flow.dl.sourceforge.net/sourceforge/greenstone/Develop-2.39-en.pdf

[11] Bainbridge, D., Witten, I.H., Buchanan, G., McPherson, J., Jones, S. and Mahoui, A. (2001) "Greenstone: A platform for distributed digital library applications." Proc. European Digital Library Conference, Darmstadt, Germany.
http://www.cs.waikato.ac.nz/~davidb/ecdl01/platform.ps

[12] Witten, I.H., Bainbridge, D. and Boddie, S.J. (2001) "Power to the people: end-user building of digital library collections." Proc Joint Conference on Digital Libraries, Roanoke, VA, pp. 94-103.
http://www.acm.org/pubs/articles/proceedings/dl/379437/p94-witten/p94-witten.pdf

[13] Japanese Morphological Analyzer: *Chasen*
http://chasen.aist-nara.ac.jp/

[14] NECTEC Thai Wordbreak Insertion Service: *Swath*
http://ntl.nectec.or.th/services/wordbreak/