A Retrospective Look at Greenstone: Lessons from the First Decade

Ian H. Witten and David Bainbridge
Department of Computer Science
University of Waikato
Hamilton, New Zealand
+64 7 838-4246
{ihw, davidb}@cs.waikato.ac.nz

ABSTRACT
The Greenstone Digital Library Software has helped spread the practical impact of digital library technology throughout the world, with particular emphasis on developing countries. As Greenstone enters its second decade, this article takes a retrospective look at its development, the challenges that have been faced, and the lessons that have been learned in developing and deploying a comprehensive open-source system for the construction of digital libraries internationally. Not surprisingly, the most difficult challenges have been political, educational, and sociological, echoing that old programmers' blessing "may all your problems be technical ones."

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *collection, dissemination, standards, systems issues*.

General Terms: Design, Human Factors, Standardization.

Keywords: Greenstone, architecture, internationalization.

1. INTRODUCTION
It is ten years since the name Greenstone was adopted for what was then the New Zealand Digital Library Software, and the decision was made to distribute it under the GNU General Public License. Today its user base hails from 70 countries and the reader's interface has been translated into 45 languages. Downloads from SourceForge have risen from a steady (for many years) 4,500 times a month to 6,500 over the last two years.

Greenstone is a suite of software for building and distributing digital library collections. It is not a digital library but a tool for building digital libraries. It provides a new way of organizing information and publishing it on the Internet in the form of a fully-searchable, metadata-driven digital library. It has been developed and distributed in cooperation with UNESCO and the Human Info NGO in Belgium. It runs on all popular operating systems (even the iPod). For more details see Witten and Bainbridge's book *How to build a digital library* and the website *http://www.greenstone.org*.

JCDL '07, June 18–23, 2007, Vancouver, B.C., Canada.

Many papers have been presented at JCDL (and elsewhere) on technical aspects of Greenstone: what facilities it offers and how it works. The present article takes a retrospective look at its development. How did this software project and the team behind it reach this point? What challenges were faced along the way? What lessons can be learned from the experience? They say that those who ignore history are doomed to relive it: we hope that sharing our experience will give heart to others, and also help prevent them from making the same mistakes.

2. HISTORY OF GREENSTONE
We briefly recount the history of the Greenstone project, summarized in Table 1. Serendipitous events have determined many of the significant directions in which the software has evolved, with its emphasis on stand-alone collections, humanitarian applications, multilingual collections and interfaces, broad interoperability, extensive documentation, New Zealand branding, and an international program of training courses.

2007	• Greenstone distributed with IITE's course *Digital Libraries in Education*
2006	• Finalist for the Stockholm Challenge
	• Greenstone Support Group for South Asia launched
2005	• Initial release of Greenstone3
	• Greenstone distributed with FAO's Information Management Resource Kit
2004	• IFIP Namur award
2002	• DL Consulting incorporated
	• Begin developing the Translator's Interface
2002	• Began development of Greenstone 3
	• Official opening of the Niupepa collection
	• Begin developing the *Librarian Interface*
	• First UNESCO Greenstone CD-ROM
2001	• Development of the *Collector*
2000	• Begin to distribute software on SourceForge
	• *Toki* presented to the NZ Digital Library project on behalf of the entire Māori people
	• Formally established cooperative effort with UNESCO and Human Info NGO
	• Greenstone mailing list started
1999	• BBC collection established
1998	• Greenstone.org website established
	• First CD-ROM collection released: *Humanity Development Library*
1997	• Decision to use the GPL; "Greenstone" adopted as name of software
	• Began work with Human Info NGO to produce humanitarian CD-ROMs
1995	• Digital library of Computer Science Technical Reports

Table 1 Significant events in the history of Greenstone

This *toki* (adze) was a gift from the Māori people of New Zealand in recognition of our project's contributions to indigenous language preservation, and resides in the project laboratory at the University of Waikato. In Māori culture there are several kinds of *toki*, with different purposes. This one is a ceremonial adze, *toki pou tangata*, a symbol of chieftainship. The *rau* (blade) is sharp, hard, and made of *pounamu* or greenstone—hence the Greenstone software, at the cutting edge of digital library technology. There are three figures carved into the *toki*. The forward-looking one looks out to where the *rau* is pointing to ensure that the *toki* is appropriately targeted. The backward-looking one at the top is a sentinel that guards where the *rau* can't see. There is a third head at the bottom of the handle which makes sure that the chief's decisions—to which the *toki* lends authority—are properly grounded in reality. The name of this *taonga*, or art-treasure, is *Toki Pou Hinengaro*, which translates roughly as "the adze that shapes the excellence of thought."

Figure 1. The Greenstone *toki*

In the beginning

The project grew out of research on text compression and, later, index compression. Around this time we heard of digital libraries, and pointed out the potential advantages of compression at the first-ever digital library conference (1994). The New Zealand Digital Library Project was established in 1995, beginning with a collection of 50,000 computer science technical reports downloaded from the Internet. At the time several research groups in computer science departments were harvesting technical reports and making them available on the web: our main contribution was the use of full-text indexing for effective search. We were assisted by equipment funding from the NZ Lotteries Board and operating funding from the NZ Foundation for Research, Science and Technology (1996–1998 and 2002–2007).

Humanitarian collections

In 1997 we began to work with Human Info NGO to help them produce fully searchable CD-ROM collections of humanitarian information. The CD-ROMs were the vision of a Belgian medical doctor who had worked in Africa, seen the pressing need for such information in developing countries, and hit upon electronic distribution as the solution. Unfortunately, however, he had encountered difficulties in developing the necessary software and subsequently exhausted his funds. To bring our software into line with his needs we had to make our server (and in particular the full-text search engine it used), which had been developed under Linux, run on Windows machines—including the early Windows 3.1 and 3.11 because, although by then obsolete, they were still prevalent in developing countries. This was demanding but largely uninteresting technically: we had to develop expertise in long-forgotten software systems, and it was hard to find suitable compilers (eventually we obtained a "second-hand" one from a software auction).

The first publicly available CD-ROM, the *Humanity Development Library*, was issued in April 1998. A French collection, UNESCO's *Sahel point Doc*, appeared a year later: all the documents, along with the entire interface, help text, and full-text search mechanism, were in French. The first multilingual collection soon followed: a Spanish/English *Biblioteca Virtual de Desastres*/*Virtual Disaster Collection*. Since then about 40 humanitarian CD-ROM collections have been published, listed in Table 2. They are produced by Human Info's office in Romania, which incorporates an in-house OCR production line. We wrote the software and were heavily involved in preparing the first few CD-ROMs; then transferred the technology so that they could proceed independently. At this point we realized that we did not aspire to be a digital library site ourselves, but rather to develop software that others could use for their own digital libraries.

Name and license

During 1997 the name *Greenstone* was adopted: "New Zealand Digital Library Software" not only seemed clumsy but impeded international acceptance. "Greenstone" turned out to be an inspired choice: snappy, memorable, and un-nationalistic but with strong national connotations within New Zealand. A form of nephrite jade, greenstone is a hallowed substance for Māori, valued more highly than gold. Moreover, it is easy to spell and pronounce. Our earlier *Weka* (think *mecca*) machine learning workbench, an acronym that in Māori spells the name of a flightless native bird, suffers from being mispronounced *weaka* by some. And the word Greenstone is not overly common—today we are the number one Google hit.

The decision to issue the software as open source, and to use the GNU General Public License, was made around the same time. We did not discuss this with University of Waikato authorities—New Zealand universities are obsessed with commercialization and we would have been forced into an endless round of deliberations on commercial licensing—but simply began to release under GPL. Since it had grown as a research tool, we had ourselves benefited from open source software. Early releases were posted on the website *greenstone.org* (registered on 13 Aug 1998), and in Nov 2000 we moved to the SourceForge site for distribution (largely due to the per-megabyte charging scheme that our university levied for both outgoing and incoming web traffic). Our employers were not particularly happy when our licensing *fait accompli* became apparent years later, but have grown to accept the status quo because of our evident international success. Table 3 lists the public releases of the production version of Greenstone, called Greenstone 2, since the year 2000.

Niupepa: the Māori newspapers

An early in-house project utilizing Greenstone was the Niupepa collection of Māori-language newspapers. We began the work of OCRing 20,000 page images and made an initial demonstration collection in 1998. In 2000–2001 we received (retrospective!) funding from the Ministry of Education to continue the work. Virtually the entire Niupepa was online early in 2001, but the collection was not officially launched until March 2002 at the Annual General meeting of Te Rūnanga o Ngā Kura Kaupapa Māori (the controlling body of Māori medium/theology schools). Niupepa is still the largest collection of on-line Māori-language documents, and is extensively used. On 13 Nov 2000, in a moving ceremony, the Māori people presented our project with a ceremonial *toki* (adze) as a gift in recognition of our contributions to indigenous language preservation (Figure 1).

BBC collection

In 1999 the BBC in London were concerned about the threat of Y2K bugs on their database of one million lengthy metadata records for radio and television programmes. They decided to augment their heavy-duty mainframe database with a fully searchable Greenstone system that could run on ordinary desktop machines. A Greenstone collection was duly built and delivered (within two days of receiving the full dataset). We tried to get them to the point where they could maintain it themselves, but they were not interested: instead they preferred to contract us to update it regularly for them. They eventually moved to different technology in early 2006, in order to make the metadata (and ultimately the programme content) publicly available online in a way that resembles what Amazon does for books—something that we think requires a tailor-made portal rather than a general-purpose digital library system.

The UNESCO connection

We became acquainted with UNESCO through Human Info's long-term relationship with them. Although UNESCO supported Human Info's goal of producing humanitarian CD-ROMs and distributing them in developing countries, they were really interested in *sustainable* development, which requires empowering people in those countries to produce and distribute their own digital library collections—following that old Chinese proverb about giving a man fish versus teaching him to fish.[1] We had by then transferred our collection-building technology to Human Info, and tried (though without success) to transfer it to the BBC, but this was a completely different proposition: to put the power to build collections into the hands of those other than IT specialists, typically librarians.

We began by packaging our PERL scripts and documenting them so that others could use them, and slowly, painfully, came to terms with the fact that operating at this level is anathema for librarians. In 2001 we produced a web-based system called the *Collector* that was announced in a paper whose title proudly proclaimed "Power to the people: end-user building of digital library collections." However, this was never a great success: web-based submission to repository systems (including Greenstone collections) is commonplace today, but we were trying (using the more limited web technologies available seven years ago) to allow users to design and configure digital library collections over the web as well as populate them. The next year we began a Java development that became known as the Greenstone Librarian Interface, which grew over the years into a comprehensive system for designing and building collections and includes its own metadata editor.

CD-ROM distributions

From the outset, UNESCO's goal was to produce CD-ROMs containing the entire Greenstone software (not just individual collections plus the run-time system, as in Human Info's products), so that it could be used by people in developing countries who did not have ready access to the Internet.[2] These were the tangible outcomes of a series of small contracts with UNESCO. However, we felt that they were more of symbolic than actual significance because they rapidly became outdated by frequent new releases of the software appearing on the Internet (Table 3). They were produced annually from 2002 to 2006.

When we and others started to give workshops, tutorials, and courses on Greenstone we adopted a policy of putting all instructional material—PowerPoint slides, exercises, sample files for projects—on a workshop CD-ROM, and began to include this auxiliary material on the UNESCO distributions. This ultimately led to their downfall, for the company producing the CD-ROMs began to question the provenance of some of the sample files they contained, and ultimately demanded explicit proof of permission to reproduce all the information and software. Although everything was, in principle, either open source or clearly covered under fair use, so much had to be stripped out that the 2006 CD-ROM distribution was seriously emasculated. CD-ROM distributions continue to be produced for workshops, however.

Multilingual documentation

Good documentation was seen by UNESCO as crucial. They were keen to make the Greenstone technology available in Spanish, French, and Russian (Arabic and Chinese are also official UNESCO languages, but for some reason never figured in these discussions). We already had versions of the interface in these (and many other) languages, but UNESCO wanted *everything* to be translated—not just the documentation, which was extensive (four substantial manuals) but all the installation instructions, README files, example collections, warning messages from PERL scripts, etc. We might have demurred had we realized the extent to which such a massive translation effort would threaten to hobble the potential for future development, and have since suffered mightily in getting everything—including last-minute interface tweaks—translated for each upcoming UNESCO CD-ROM release.

The cumbersome process of maintaining up-to-date translations in the face of continual evolution of the software—which is, of course, to be expected in open source systems—led us to devise a scheme for maintaining all language fragments in a version control system so that the system could tell what needed updating. This resulted in the Greenstone Translator's Interface, a web portal where officially registered translators can examine the status of the language interface for which they are responsible, and update it. Today the interface has been translated into many languages (see Table 8 below), most of which have a designated volunteer maintainer.

International training

Training is a bottleneck for widespread adoption of any digital library software. With UNESCO's encouragement and sponsorship we have worked to enable developing countries to take advantage of digital library technology by running hands-on workshops. Many Greenstone workshops have been given in developing countries, ranging from half a day to 6 days. Table 4 lists ones given by people closely associated with the project; there have been many others. This activity has enabled team

[1] In New Zealand, by the way, they say "give a man a fish and he'll eat for a day; teach a man to fish and he'll sit in a boat and drink beer for the rest of his life."

[2] Incidentally, UNESCO refused to use our *toki* logo on the CD-ROMs because they feel that in some developing countries axes are irrevocably linked to genocide. Our protests that this object is clearly ceremonial fell on deaf ears. Dealing with international agencies can be very frustrating.

2006	• Appropriate Technology Knowledge Collection	En
2005	• Gender and HIV/AIDS Electronic Library	En
	• Textes de Base sur L'Environment au Senegal	Fr
	• Educational Aids/Lehr- und Lernmittel/ Moyens didactiques/Material didáctico v3.0	En/De/Fr/Es
2004	• Africa Collection for Transition: From Relief to Development v1.01	En
	• UNECE Committee for Trade, Industry and Enterprise Development	En/Fr/Ru
	• INEE Technical Kit on Education in Emergencies and Early Recovery	En
	• Educational Aids/Lehr- und Lernmittel/ Moyens didactiques/Material didáctico v2.0	En/De/Fr/Es
2003	• Education, Work and the Future/ Education Travail et Avenir v2.0	En/Fr
	• Revised Curricula for Technical Colleges	En
	• UNAIDS Library v2.0	En/Fr/Es/Ru
	• Biblioteca Virtual de Salud para des Desastres/ Health Library for Disasters v2.0	Sp/En
	• Food and Nutrition Library v2.2	En
	• Educational Aids/Lehr- und Lernmittel/ Moyens didactiques/Material didáctico v1.0	En/De/Fr/Es
	• ICT Training Kit and Digital Library for Africa	En
2002	• Community Development Library for Sustainable Development and Basic Human Needs v2.1	En
	• Food and Nutrition Library v2.0	En
	• UNDP Energy for Sustainable Development Library	En
2001	• UNAIDS Library v1.1	En/Fr/Es/Ru
	• East African Development Library	En
	• Safe Motherhood Strategies	En/Fr/Es
	• Researching Education Development	En
	• Biblioteca Virtual de Salud para des Desastres	Es/En
	• WHO Medicines Bookshelf	En
	• Africa Collection for Transition	En
2000	• World Environmental Library v1.1	En
	• Sahel point Doc v2.0	Fr
	• Food and Nutrition Library v1.0	En
1999	• Medical and Health Library v1.0	En
	• Bibliothèque pour le Développement Durable et des Besoins Essentials v1.0	Fr
	• Biblioteca Virtual de Desastres	Es/En
	• UNU Collection on Critical Global Issues v2.0	En
	• Sahel point Doc	Fr
	• Humanity Development Library v2.0	En
1998	• UNU Collection on Critical Global Issues v1.0	En
	• Humanity Development Library v1.3	En

Table 2 Humanitarian CD-ROMs

2006	Dec	2.72
	Oct	2.71
	Mar	2.70
	Jan	2.63
2005	Jun	2.62
	Apr	2.60
	Mar	2.53
2004	Oct	2.52
	Jun	2.51
	Feb	2.50
2003	Dec	2.41
	Jun	2.40
	Mar	2.39
2002	Jan	2.38
2001	Oct	2.37
	Jun	2.36
	May	2.35
	Apr	2.33
	Feb	2.31
	Feb	2.30
2000	Dec	2.30
	Sep	2.27
	Jul	2.25
	Jun	2.23
	Jun	2.22
	Apr	2.21
	Feb	2.12

Table 3 Greenstone releases

2007	May	• Trinidad and Tobago National Library
	Mar	• Colombo, Sri Lanka
	Feb	• Vellore, India
2006	Dec	• Calcutta, India
	Dec	• New Delhi, India
	Nov–Dec	• Kozhikode, India
	Oct	• Vladimir, Russia
	Aug	• Tirunelvelli, India
	Mar–Apr	• Madras, India
	Mar	• Durban, South Africa
	Feb	• Bangkok, Thailand
2005	Nov	• Cape Town, South Africa
	Nov–Dec	• Arusha, Tanzania
	Sep	• Suva, Fiji
	Aug	• Bangalore, India
	May	• Ho Chi Minh City, Vietnam
	May	• Kozhikode, India
2004	Dec	• Bombay, India
	Oct	• Havana, Cuba
	Sep	• Trirandom, India
	Aug–Sep	• Windhoek, Namibia
	Jul	• Suva, Fiji
	Jun	• Cape Town, South Africa
	Mar	• Dakar, Senegal
	Mar	• Cape Town, South Africa
	Feb	• Gaborone, Botswana
	Feb	• Almaty, Kazakhstan
2003	Nov	• Dakar, Senegal
	Nov	• Suva, Fiji
	May	• Bangalore, India (IISC)

Table 4 Greenstone workshops in developing countries

members to travel to many interesting places. In what other area might a computer science professor get the opportunity to spend a week giving a course at the UN International Criminal Tribunal for Rwanda in Arusha, Tanzania, at the foot of Mount Kilimanjaro—or in Havana, Cuba?

The United Nations Food and Agricultural Organization (FAO) and UNESCO's Institute for Information Technology in Education have also produced training material on Greenstone. Furthermore, we have been active in conducting Greenstone tutorials at all major digital library conferences—JCDL, ECDL, ICADL, ICDL (on several occasions in each case)—and library conferences such as LITA, DLF, and the ALA Annual Conference. The Payson Institute of International Development at Tulane University has run courses that use Greenstone collections as a resource in dozens of locations in Africa (Burkina Faso, Cameroon, Cote d'Ivoire, Democratic Republic of Congo, Ghana, Rwanda, Senegal, Sierra Leone, Togo) and Latin America (Argentina, Bolivia, Colombia, Ecuador, Guatemala).

Regional support groups

Recognizing that devolution is essential for sustainability, we are now striving to distribute Greenstone training, maintenance and support by establishing regional Support Groups. User groups for Spanish and French users have existed for some time, and in April 2006 a comprehensive Greenstone Support Group for South Asia was launched, centered in Kerala, India. This very active group operates its own email help desk and has run several courses and workshops in the region. In 2005 a study was undertaken, with UNESCO support, of the feasibility of setting up a Greenstone Support Organization for Africa [1], based on a survey questionnaire that was widely circulated to African professionals; a new project focusing on promoting digital library usage in Africa is beginning this year.

Interoperability

Many early digital library projects focused on interoperability. Although this is clearly an important issue, we felt that this attention was premature—we well remember a digital library conference where interest was so strong that there were two panel discussions on interoperability, the only catch being that they were parallel sessions, which permitted no … er … interoperability. We adopted the informal motto "first operability, then interoperability"; and focused on other issues such as ingesting documents and metadata in a wide variety of formats. More recently we have added many interoperability features, which, as we had expected, were not hard to retrofit.

Software evolution

We continually struggle with the conflict between stability and evolution. We place great emphasis on backward compatibility: it is rare for new Greenstone releases to have any effect at all on existing collections, and then only in minor respects. Only recently have we made a concession to hardware obsolescence by making alterations that no longer allow standard Greenstone collections to be served on Windows 3.1/3.11.

To take advantage of new developments in software technology we began a new project, Greenstone 3, which is a complete redesign and reimplementation of the original digital library software (Greenstone 2). It incorporates all features of the existing system, and is backward compatible: that is, it can build and run existing collections. It is structured as a network of independent modules that communicate using XML: thus it runs in a distributed fashion and can be spread across different servers as necessary. This modular design also increases flexibility and extensibility. However, although initial versions of Greenstone 3 have been released, continual demands from users for further development of Greenstone 2 have delayed progress on the new version.

Greenstone 3 was originally envisaged purely as a research framework: backward compatibility would be possible but required IT skills. Attention was focused on the future and how best to allow an ever changing heterogeneous environment of software components (including novel techniques) to mesh with a digital library infrastructure. For the most part we have achieved this aim: it is now much easier for others, such as graduate and undergraduate project students, to build upon the digital library core. However, we have found that it is beyond our resources to maintain two independent versions of Greenstone—in particular, to ensure backward compatibility when new and enhanced features are added to Greenstone 2. Consequently we have committed to a new vision: to develop Greenstone 3 to the point that, by default, its installation and operation is, to the user, indistinguishable from Greenstone 2. This work is included in a recent release of Greenstone 3 (3.02).

3. CURRENT STATE

Here is a capsule summary of some salient features of Greenstone and its user population.

Platforms. Greenstone runs on all versions of Windows, and Unix, and Mac OS-X. It is very easy to install. For the default Windows installation absolutely no configuration is necessary, and end users routinely install Greenstone on their personal laptops or workstations. Institutional users run it on their main web server, where it interoperates with standard web server software (e.g. Apache).

Interfaces. Greenstone has two separate interactive interfaces, the Reader interface and the Librarian interface. End users access the digital library through the Reader interface, which operates within a web browser. The Librarian interface is a Java-based graphical user interface (also available as an applet) that makes it easy to gather material for a collection (downloading it from the web where necessary), enrich it by adding metadata, design the searching and browsing facilities that the collection will offer the user, and build and serve the collection.

Standards. Greenstone is strongly standards-based. It incorporates a server that can serve any collection over the Open Archives Protocol for Metadata Harvesting (OAI-PMH), Z39.50 and SRW, and Greenstone can harvest documents over any of these protocols and include them in a collection. Collections can be exported to METS (in the Greenstone METS Profile, approved by the METS Editorial Board), and Greenstone can ingest documents in METS form. Any collection can be exported to DSpace ready for DSpace's batch import program, and any DSpace collection can be imported into Greenstone.

Formats. Table 5 shows the formats of metadata and documents that Greenstone works with. Four predefined metadata sets are provided with the software; new metadata sets can be created interactively within the Librarian interface using Greenstone's Metadata Set Editor.

Metadata editor. The Librarian interface includes a metadata editor for adding metadata to documents. However, where externally-prepared metadata is available it can be ingested using "plugins." These exist for about 10 widely used standard metadata formats (there are, in addition, some plugins for non-standard metadata such as the BBC collections mentioned earlier.)

Ingesting documents. Plugins are also used to ingest documents. There are plugins for most comment formats of textual documents, listed in Table 6, including PowerPoint and Excel documents. There are also plugins for multimedia image, audio, there are plugins for common image and audio formats. There is also a generic plugin that can be configured for other multimedia formats such as MPEG, MIDI, etc.

User base. As with most open source projects, the user base for Greenstone is unknown. It is distributed on SourceForge, a leading distribution centre for open source software. Table 7 shows relevant download statistics. It also shows the number of people who contribute to the Greenstone mailing lists, and the

Predefined metadata sets
Dublin Core (qualified and unqualified)
RFC 1807
NZGLS (New Zealand Government Locator Service)
AGLS (Australian Government Locator Service)
Metadata plugins
XML, MARC, CDS/ISIS, ProCite, BibTex, Refer, OAI, DSpace, METS
Document plugins
PDF, PostScript, Word, RTF, HTML, Plain text, Latex, ZIP archives, Excel, PPT, Email (various formats), source code
Multimedia plugins
Images (any format, including GIF, JIF, JPEG, TIFF), MP3 audio, Ogg Vorbis audio
Generic plugin

Table 5 Metadata and document formats

Arafura Digital Archive for East Timor
Argentina Secretary of Human Rights
Association of Indian Labour Historians, Delhi
Balearic Islands Scientific Library
British Columbia Indian Chiefs Union
Charles Darwin University, Australia
Council of Independent Colleges, Washington DC
Gresham College, London
Hawaiian Electronic Library
iArchives, Utah
Ionian University, Greece
Indian Institute of Management, Khozikode
Indian Institute of Science, Bangalore
Kazakhstan Human Rights Commission
Kyrgyz Republic National Library
Latin America and Caribbean Network of Social Science
Mari El Republic, Russia
Marshall Foundation, Virginia
Netherlands Institute for Scientific Information Services
New York Botanical Garden
Pacific Archive for Learning and Education
Peking University Digital Library
Philippine Education and Government Information Network
Slavonski Brod Public Library, Slovenia
State Library of Tasmania
Stuttgart University of Applied Sciences
Sudanese Association of Libraries and Information
UNESCO
United Nations in Pakistan
Universities of Auburn (Al), Chicago, Detroit, Illinois, Illinois Wesleyan, Iowa, Lehigh, North Carolina, Tulane, Yale
University of Namibia
Vietnam National University
Vimercate Public Library, Milan, Italy
Washington Research Library Consortium
Welsh Books Council

Table 6 Sample collections (URLs at *www.greenstone.org*)

Distributed via SourceForge since:	Nov 2000
Average downloads since then:	4500/month
Currently running at:	6500/month
Proportion of downloads that are documentation:	44%
Proportion of downloads that are software:	56%
Of these, 82% are Windows binaries	
10% are Linux binaries	
4% are MacOS binaries	
4% are source	
Number of people on Greenstone email lists:	600
Number of countries represented:	70
Number of messages (excluding spam):	150/month

Table 7 Download and mailing list statistics

Reader's Interface is available in:
Arabic, Armenian, Bengali, Catalan, Chinese (both simplified and traditional), Croatian, Czech, Dari, Dutch, English, Farsi, Finnish, French, Gaelic, Galician, Georgian, German, Greek, Hebrew, Hindi, Hungarian, Indonesian, Italian, Japanese, Kannada, Kazakh, Kirghiz, Latvian, Maori, Marathi, Mongolian, Polish, Portuguese (both European and Brazilian versions), Pushto, Romanian, Russian, Serbian, Slovak, Spanish, Thai, Turkish, Ukrainian, Vietnamese
Other languages in progress:
Amharic, Azeri, Bulgarian, Burmese, Gujarati, Khmer (Cambodian), Malayalam, Oneida (Iroquoian), Samoan, Tamil, Telugu, Urdu
Librarian Interface available in:
Arabic, English, French, Marathi, Spanish, Romanian, Russian Chinese (simplified), Latvian, Vietnamese in progress
Full documentation (four manuals) available in:
English, French, Spanish, Russian (three of the four also in Kazakh, Vietnamese)

Table 8 Languages of Greenstone

volume of traffic. The website *http://www.greenstone.org* points to a representative selection of examples of public Greenstone collections; Table 6 shows the institutions they belong to. A survey of Greenstone users was undertaken in 2004-2005 [2].

Educational usage. Greenstone forms a popular basis for training purposes in Library and Information Science programs, particularly in the US. Although we know of several institutions that employ it for this purpose, we have no definitive list. Some evidence for its influence comes from the fact that Witten and Bainbridge's book *How to build a digital library*, which contains extensive material on Greenstone, is the most frequently assigned text in US digital library courses [3].

Languages. One of Greenstone's unique strengths is its multilingual nature. The reader's interface is available in the 45 languages shown in Table 8, with another 9 in progress. The Librarian interface and the full Greenstone documentation (which is extensive) are available in several languages including English, French, Spanish, and Russian.

4. THE WIDER PICTURE

There are many alternatives to Greenstone. EPrints was created in 2000 as an open source software package for building open access repositories [4]. Now into its third generation [5], an account of its core abilities can be found in [6]. Robert Tansley, the lead programmer for EPrints in its early days, went on to produce DSpace [7], a well-known institutional repository system from MIT and HP Labs; an interesting account exists of its deployment [8]. Fedora [9,10] is a general architecture for digital asset management, intended as a foundation for many types of digital library, while Fez [11] is a configurable digital repository and workflow management system based on top of it. Cheshire II [12] is used to implement full-text and fielded searching of bibliographic information for the University of California Berkeley Digital Library Initiative, with Cheshire III [13] its successor. Koha [14] is an open source web-based integrated library system. In addition, there are several commercial systems that allow collection building, such as CONTENTdm [15].

This is not the place to attempt a comparison between digital library systems, which is notoriously difficult to do (see [16], [17], [18] for some attempts, and [19] for a usage survey of such systems in India). Goh *et al.* [20] developed a checklist consisting of 12 categories of items and used it to evaluate several open source digital library packages. They judged that Greenstone was the only software package that consistently fulfilled the majority of the criteria in many of the checklist categories, and concluded that it was the best performer overall, followed by CDSware/ Invenio, Fedora and EPrints.

However, while it is natural to seek such comparisons, they tend to be invidious, self-serving, and generate more heat than light. In truth one is comparing apples and oranges—rather complex ones that require intensive and detailed study—and the fruit is constantly changing and ripening over time. While there is certainly much overlap between their capabilities, these systems

have quite different goals and strengths. Quoting from what is probably the only paper with joint authorship from more than one of these projects [21], the key points that Greenstone makes its core business to support include:

- Design and construction of collections
- Distribution on the web and/or removable media
- Customized structure depending on available metadata
- End-user collection-building interface for librarians
- Reader and librarian interfaces in many languages
- Multiplatform operation;

whereas those for DSpace support include:

- Repositories at an institutional level
- Self-deposit of digital assets by faculty
- End-user interface for depositors
- Assets made available for searching and browsing
- Data retrievable many years in the future
- Institutional commitment to ensure the continued availability of certain named formats.

5. LESSONS LEARNED

What has been learned from Greenstone's first decade?

Project directions

Purely serendipitous events have had a far-reaching impact on the Greenstone technology. Our most satisfying rewards have come from having an open mind and seizing interesting opportunities as they arrive. The humanitarian connection, UNESCO support, Niupepa's local indigenous flavor, and a strong international emphasis—probably stemming from our small population and geographical isolation—have materially affected the software.

Funding

Funding on the scale necessary to run a serious software project based in New Zealand is simply not available for open source efforts—even ones whose worth is widely recognized. UNESCO applauds our work but has an explicit policy of not funding software development, particularly in developed countries, which should, it believes, be able to muster sufficient resources nationally. New Zealand applauds our work but denies having any such sources. Research organizations like to see outcomes being put to practical use but only fund new research. Our universities have no internal funds for this kind of project. The library community can muster support for tailoring software to specific needs, but not for general development. Philanthropists prefer to support projects that alleviate visible problems like health and living conditions. It is difficult for those based elsewhere to gain support from charitable foundations in the US and Europe.

How do other open source DL groups fare? JISC funds a sizable portfolio of open source projects in the UK (e.g. EPrints). Like UNESCO, they disfavor isolated software development, these projects are driven by other criteria and outcomes, such as content creation and community building. The US National Science and Mellon Foundations have a more liberal attitude to software development: the latter currently funds 14 prominent open source projects.

Sustainability

Because open source projects depend on the interest and efforts of individuals it is hard to guarantee that they will continue in the indefinite future. Despite strenuous efforts, we have been unable to develop a reliable sustainability model for Greenstone. The best we have been able to manage is the establishment of many local centers of expertise—there are now small pockets of Greenstone expertise in universities in several countries—and locally organized support groups such as that for South Asia.

A sadly unsuccessful approach has been a "Friend of Greenstone" scheme. For a modest annual sum the benefits include prioritized responses to messages sent to the Greenstone discussion list; a small amount of programming work on a specific coding issue; and a Greenstone memento. Just a few subscribers would enable us to secure the salary of one of our programmer, yet we have been unable to gain any leverage from this program. Projects such as DSpace and Fedora have probably been more successful. Their core adopters are universities and other large-scale institutions in the West, which can afford subscription costs and paid educational meetings for staff.

Callback feature

Greenstone does not incorporate any callback mechanism to automatically register installations and collections. However, it would be of enormous help both for funding purposes and for technical feedback to have a record of the user base. Although we considered an automatic callback feature we rejected it because users of open source software surely have a right to assume anonymity by default. Alternatives such as popup warnings were also rejected on various grounds. More recently, the Directory of Open Access Repositories (*www.opendoar.org*) provides an independent route for registration (albeit with an institutional focus) that goes some way towards addressing this problem.

We are constantly astonished to learn of new places where the software is used (e.g. at a recent conference it transpired that a Greenstone course has been held in Pyongyang, North Korea). We are surprised to receive no feedback even when instructors use Greenstone for courses: we stumble upon evaluations, done by students as coursework, that provide extremely useful feedback. We want to "know our users" and to hear their problems: many issues that we do hear about are completely trivial to solve.

Start-up cost

Most open source digital library software is complex to install. Installation of essential components such as databases is often left as the user's problem—and it can be a very difficult problem, particularly because issues arise from subtle interactions between the chosen components and the core system. Traffic on discussion groups and forums shows that considerable effort is dedicated to helping people solve the myriad of problems encountered.

From the very beginning it was impressed on us that the humanitarian CD-ROMs must be trivial for anyone to install—as easy as opening a book—and very robust. They had to work on any Windows computer, in any environment, no matter how broken its software (computers are often recycled from the developed to the developing world, with software configured in a way that is entirely inappropriate to the current operating environment). We retained this philosophy when moving from distributing individual collections to distributing the ability to build them on popular computer platforms—a far more complex prospect but, armed with the right development tools, achievable.

In our opinion, every effort should be made to ensure that DL software is simple to install. We know of organizations—even

ones with dedicated IT teams—who tried to undertake a comparative study of possible options but simply could not get particular DL products to run. Not only does this rule out these systems from consideration; it gives the field a bad name.

Even when complex installation seems unavoidable, we advocate a turnkey approach that provides a workable system out of the box, perhaps with the option of installing a more capable back-end later if required. For example, a DL could be shipped with a Java implementation of a relational database (e.g. Apache's Derby), but use JDBC (Java Database Connectivity) so that a more powerful system could be substituted. This would allow organizations to trial the system and workshop participants to receive a copy, both of which we have found very useful.

Institutional vs. individual users

The ease of acquiring and installing a software project has a direct impact on the users it attracts, and consequently—in the open source world—on the extent and nature of contributions that users make to the project. While greatly appreciated by individual users, it is less relevant to institutions with their own software support personnel. Indeed, a subtle corollary is that tricky installation procedures give IT departments the opportunity to exercise their skills and demonstrate their value to the organization. From their point of view, it is actually counter-productive to deskill the installation process to the extent that anyone can install the software on their laptop or desktop workstation. In practice, Greenstone has encountered far more opposition from large institutional libraries than from individual users.

Now that we are aware of this obstacle, we can imagine having the best of both worlds. As the above example illustrates, easy installation combined with the ability to swap components lets in low-end users, permits easy evaluation, yet allows specialist IT staff to configure more complex installations.

Impact of user base on open source projects

A greater proportion of individual rather than institutional users has a further corollary for open source software. End users of DLs are not themselves software specialists, whereas end users of many other open source projects—compilers, editors, version control systems—are themselves programmers. There is generally no way that they can fix or rectify any bugs or shortcomings they encounter. In an institutional setting the problem can be referred in the first instance to the IT department, which is likely to solve it and contribute the fix back to the software developers. Open source systems whose users are predominantly non-programming individuals gain less technical leverage from the user community.

Imposing metadata standards

Life is far easier for the developers of DL systems with fixed metadata schema! Insisting on a certain schema greatly simplifies many software decisions. Internal data structures can be tailored to a particular schema; more importantly, fixed forms can be used for input. However, we have found that many users have made substantial investments in existing metadata that is not standard. The BBC had a million records in an idiosyncratic format that had served them well and they could not reasonably abandon. Users of UNESCO's CDS-ISIS software (which is widespread in the developing world) are used to developing their own metadata schema and cannot see why new technology should remove this advantage. The applicability of Greenstone is greatly enhanced by its catholic approach to metadata, but at the expense of ease of use for those who design and build collections.

Motivation of project personnel

Most people who work on open source software projects are highly motivated. Furthermore the international, developing world, humanitarian philosophy of Greenstone is particularly effective in motivating project personnel. Staff run Greenstone workshops in attractive places (Fiji, Hawaii) and meet end users—less-skilled users from developing countries rather than institutional IT personnel. This provides an enormous sense of satisfaction, and greatly helps in understanding user problems. Motivation is important because the Greenstone project has been seriously under-resourced throughout its duration.

Informing the research community

Having a fully operating high-functioning software system for digital libraries has proven to be a surprising disadvantage in publicizing our work in the research community. The problem was neatly summed up by the reply to a question at JCDL when a presenter was asked what he knew about Greenstone's (existing, operational, fully-deployed) solution to a problem he had been struggling with: "Oh Greenstone, that's more of a production system, isn't it?" Most DL researchers have heard of Greenstone and think they know all about it, despite the fact that huge and novel advances are made in the software every year. We have frequently had papers rejected for insufficient novelty ("yet another Greenstone paper"), only to listen to other presenters years later describe their solutions to the very same problems. Despite pleas from referees (including those for this paper) for more cross-platform comparisons, submissions jointly authored by different groups (e.g. [21]) have been summarily rejected by prominent DL conferences as being insufficiently interesting.

The "developing country" stigma

Another lesson we have learned is the PR danger of making your research accessible to people in the developing world, and having it adopted there. Like the "free software" stigma ("it can't be any good if they have to give it away"), many people dismiss our work with backhanded compliments—"jolly good for developing countries, but we need something more professional." In fact, Greenstone is second to none in functionality and usability. Developing country usage is *more demanding* than Western usage because there is a greater need for functionality and usability. The idea that second-rate software is good enough for these people smacks of 19th century colonialism.

Multilingual software

With UNESCO's encouragement and support multilinguality is a key strength of Greenstone, built in from the outset. Translations are willingly provided by our non-programming user base. However, more is needed than just the technical infrastructure for displaying multilingual text: it is necessary to manage the translation process in the face of continual change in the software. This motivated a line of research that was released as open source and published at ECDL in 2003.

Many of the projects mentioned in Section 4 provide multi-language interfaces, albeit mostly retrofitted. However, there is typically little organized coordination of the translation process over time; whole-hearted commitment is lacking (help text, for example, may defined to be outside the official core translation);

and translators must use *ad hoc* methods—for example, they are left to generate images containing language-specific text themselves. We are disappointed that the solutions we described have not been adopted; nor has our open source code.

Build bridges rather than islands

The language translation work is a clear example of open source software that like-minded DL projects could benefit from, but have not; and as noted above we know of only one paper whose authors span more than one of these projects. Other parts of the open source world are more accommodating: Lucene is a popular choice for full-text indexing, MySQL or PostgreSQL for a relational database, and so on. Why is it that we insist on rolling our own digital libraries, on building islands rather than bridges?

We have built some bridges. Greenstone can export collections to DSpace, Fedora, and XMLMARC, in addition to making them available over OAI. It can import from DSpace, MARC, SRW and OAI (even the last, surprisingly, is unusual for a DL system); we have pointed out how our importing interface can be used to facilitate document exchange and interoperability between various systems: for example, going from DSpace to Fedora and back again—without involving Greenstone at all!

We suspect the reason is social rather than technical. The idea that we put out the fruit of our labor for others to use does not seem to translate into effective redeployment by other DL developers. If this is indeed a social phenomenon then perhaps we need tutorials and workshops on open source DL software *in general*. When at JCDL 2005 we mounted a tutorial on "Practical digital library interoperability standards using open source software" which drew upon and demonstrated examples that combined Cheshire, Fedora, DSpace and Greenstone the room was full.

Standing on the shoulders of others?

A problem endemic to all digital library research is that we do not stand on our predecessor's shoulders the way other scientists do—more likely, we stand on their toes. Regrettably, we cannot claim that Greenstone's development has materially benefited from digital library research elsewhere, despite originating in the heady days of the first NSF Digital Libraries Initiative. Nor, equally regrettably, can we credibly claim that the papers we have published have materially affected the development of other digital library systems, or indeed the digital library field in general—except insofar as it has been influenced by the existence of the Greenstone software itself. We find this rather depressing, but feel it is time that the digital library research community faced the fact that though all our work is valuable it does not really give a strong sense of cumulative scientific progress.

6. CONCLUSIONS

University research projects rarely produce software that others can use, let alone systems that are widely deployed on an international scale. Greenstone is a notable exception: we find it refreshing. However, the birds-eye retrospective synopsis in this paper reveals some of the challenges we have faced in making this happen. On the one hand it has been immensely satisfying. On the other, it seems likely that Greenstone's practical deployment has make it harder to get our work recognized for its innovative contributions than if we had pursued laboratory-based research.

One original motivation for Greenstone from ten years ago was to produce a platform that allowed us to showcase our research and have it used in practice. Greenstone does contain a few subsystems that arose out of research projects on information retrieval—for example, heuristic acronym extraction, automatic key-phrase extraction, and innovative phrase browsing and collage-based image browsing techniques. In other cases it has been impossible to embed novel techniques into a publicly distributed system because of licensing restrictions. Our work on Chinese text segmentation is a case in point, for it employs a machine learning method based on a large hand-segmented corpus, which cannot be distributed except for research purposes.

Most people are surprised by the small size of the Greenstone team. Historically, for most of the duration of the project we have employed 1–2 programmers. Several faculty involved in aspects of digital library research are associated with the project, but only two view the Greenstone software as their main interest—partly because the research outputs are of questionable value in the university evaluation and promotion process. Although several graduate students work in areas cognate to digital libraries they rarely contribute to the code base directly because we insist upon retaining the production-level code quality and programming conventions painstakingly acquired over many years. Our external users tend to be librarians rather than software specialists and we have received few major contributions or bug fixes from them. To summarize, the Greenstone digital library software has been created by a couple of skilled people working over a 10-year period—and along the way there have been several changes of personnel. It's amazing what excellent programmers can do.

We would like to underscore—from our own personal experience—the enormous importance of digital libraries for the developing world. Most digital library research is conducted in libraries whose purpose is scholarship, and from the perspective of other people, libraries often seem esoteric. But they are not necessarily so. Digital libraries are the killer app for information technology in developing countries. Priorities here include health, agriculture, nutrition, hygiene, sanitation, and safe drinking water. Computers are not a priority, but simple, reliable access to targeted information meeting these basic needs certainly is—as is low-cost technology for wide distribution of organized information throughout the vast Internet-challenged regions of the world. In comparison, digital libraries are relatively unimportant in developed countries, because there are so many alternative sources of information.

Sustainability is one of the most difficult challenges for open-source projects with large user populations—particularly when the users are not programmers; particularly when much usage is in poor countries. Despite extensive efforts, we have found no real solution to this problem. It will be interesting, we hope, for others to learn about the success of our ongoing efforts to set up a sustainable Greenstone infrastructure in the next ten-year Greenstone retrospective.

Finally, it's been fun.

ACKNOWLEGEMENTS

We would like to acknowledge the entire New Zealand Digital Library Project team for their unstinting work in providing an environment that makes this kind of research meaningful—and enjoyable. And we would like to thank our users for making it all worthwhile. Finally, we gratefully acknowledge the referees, whose penetrating criticisms helped us to significantly improve this paper.

REFERENCES

[1] Peters, D.P. (2006) *Feasibility study on the establishment of a Greenstone Support Organization for Africa*. Digital Imaging South Africa, University of KwaZulu-Natal.

[2] Sheble, L. (2006) *Greenstone user survey*. Wayne State University, Detroit, MI.

[3] Pomerantz, J., Oh, S., Yang, S., Fox, E.A. and Wildemuth, B.M. (2006) "Digital Library Education in Library and Information Science Programs." *D-Lib Magazine*, 12 (11).

[4] Tansley, R. and Harnad, S. (2000) "Eprints.org software for creating institutional and individual open archives." *D-Lib Magazine*, Vol. 6, No. 10.

[5] Millington, P. and Nixon, W.J. (2007) "EPrints 3 Prelaunch Briefing." *Ariadne*, Issue 50.

[6] Ashworth, S., Mackie, M. and Nixon, W.J. (2004) "The DAEDALUS project, developing institutional repositories at Glasgow University: the story so far." *Library Review*, Vol. 53, No. 5, pp. 259–264.

[7] Smith, M., Bass, M., McClella, G., Tansley, R., Barton, M., Branschofsky, M. Stuve, D. and Wakler, J (2003) "DSpace: An open source dynamic digital repository." *D-Lib Magazine*, Vol. 9, No. 1.

[8] Smith, M., Rodgers, R., Walker, J. and Tansley, R. (2004) "DSpace: A Year in the Life of an Open Source Digital Repository System," *Proc European Conf on Digital Libraries*, pp. 38-44.

[9] C. Lagoze, S. Payette, E. Shin, C. Wilper (2006) "Fedora: An architecture for complex objects and their relationships." *Int J Digital Libraries*, Vol. 6, No. 2.

[10] S. Payette and C. Lagoze (1998) "Flexible and Extensible Digital Object and Repository Architecture." *Proc European Conf on Digital Libraries*, Heraklion, Crete.

[11] Kortekaas, C. (2006) "Don't keep it under your hat!" *Fedora Users Conference*, University of Virginia, Charlottesville.

[12] Larson, R.R. and Carson, C. (1999) "Information access for a digital library: Cheshire II and the Berkeley digital library." *Proc American Soc for Information Science*, pp. 515–535.

[13] Larson, R.R. and Sanderson, R. (2005) "Grid-based digital libraries: Cheshire3 and distributed retrieval." *Joint Conf on Digital Libraries*, Denver, 112–113.

[14] Eyler, P. (2003) "Koha: A gift to libraries from New Zealand." *Linux Journal*, Issue 103, p.1.

[15] Bond, T. and Cornish, A. (2002) "Digitizing special collections using the CONTENTdm suite." *Microform and Imaging Review*, Vol. 31, No. 11, pp. 31–36.

[16] Kaczmarek, J.W., Habing, T.G. and Eke, J. (2006) "Repository software evaluation using the audit checklist for certification of trusted digital repositories." *Proc Joint Conf on Digital Libraries*, pp. 107–108.

[17] Chawner, B. (2005) "F/OSS in the library world: An exploration." *ACM SigSoft Software Engineering Notes*, Vol. 40, No. 4, pp. 1–4.

[18] Bose, R. (2006) "Geospatial repository literature review." Digital Curation Centre and School of Informatics, University of Edinburgh.

[19] Jose, S. (2007) "Adoption of open source digital library software packages: A survey." *Proc Int Convention on Automation of Libraries in Education and Research Institutions*, pp. 98–102, Punjab University, India.

[20] Goh, D. H.-L., Chua, A., Khoo, D.A., Khoo, E.B.-H., Mak, E.B.-T. and Ng, M.W.-M (2006) "A checklist for evaluating open source digital library software." *Online Information Review*, Vol. 30, No. 4, pp. 360–379.

[21] Witten, I.H., Bainbridge, D., Tansley, R., Huang, C.Y. and Don, K. (2005) "A bridge between Greenstone and DSpace." *D-Lib Magazine*, Vol. 11, No. 9.

APPENDIX Selected Greenstone Bibliography

Apperley, M., Keegan, T.T., Cunningham, S.J. and Witten, I.H. (2002) "Delivering the Maori-language newspapers on the Internet." *Rere atu, taku manu! Discovering history, language and politics in the Maori-language newspapers*, edited by J. Curnow *et al*. Auckland University Press: 211-232.

Bainbridge, D., Edgar, K.D., McPherson, J.R. and Witten, I.H. (2003) "Managing change in a digital library system with many interface languages." *Proc European Conf on Digital Libraries*, Trondheim, Norway.

Bainbridge, D., Ke, K.-Y.J. and Witten, I.H. (2006) "Document level interoperability for collection creators." *Proc Joint Conf on Digital Libraries*, pp. 105-106, Chapel Hill, NC.

Bainbridge, D., Thompson, J. and Witten, I.H. (2003) "Assembling and enriching digital library collections." *Proc Joint Conf on Digital Libraries*, Houston, Texas.

Bell, T.C., Cleary, J.G. and Witten, I.H. (1990) *Text compression*. Prentice Hall, Englewood Cliffs, NJ.

Bell, T.C., Moffat, A. and Witten, I.H. (1994) "Compressing the digital library." *Proc Digital Libraries*, pp. 41–46, College Station, Texas.

Teahan, W.J., Wen, Y, McNab, R. and Witten, I.H. (2000) "A compression-based algorithm for Chinese word segmentation." *Computational Linguistics* Vol. 26 No. 3, pp. 375–393.

Witten, I. H., Bainbridge, D. and Boddie, S.J. (2001) "Power to the people: end-user building of digital library collections." *Proc Joint Conf on Digital Libraries*, Roanoke, VA.

Witten, I.H. and Bainbridge, D. (2003) *How to build a digital library*. Morgan Kaufmann, San Francisco, CA.

Witten, I.H., Cunningham, S.J., Vallabh, M. and Bell, T.C. (1995) "A New Zealand digital library for computer science research." *Proc Digital Libraries*, pp. 25–30, Austin, Texas.

Witten, I.H., Loots, M., Trujillo, M.F. and Bainbridge, D. (2002) "The promise of digital libraries in developing countries." *The Electronic Library* Vol. 20, No. 1, pp. 7–13.

Witten, I.H., Moffat, A. and Bell, T.C. (1994) *Managing gigabytes: compressing and indexing documents and images*. Van Nostrand Reinhold, New York.

Wu, S. and Witten, I.H. (2006) "Towards a digital library for language learning." *Proc European Conf on Digital Libraries*, Alicante, Spain, pp. 341–352.