Digital Libraries and Minority Languages

DAVID M. NICHOLS, IAN H. WITTEN,
TE TAKA KEEGAN, DAVID BAINBRIDGE AND MICHAEL DEWSNIP

Department of Computer Science, University of Waikato, Private Bag 3105, Hamilton, New Zealand

Abstract

Digital libraries have a pivotal role to play in the preservation and maintenance of international cultures in general and minority languages in particular. This paper outlines a software tool for building digital libraries that is well adapted for creating and distributing local information collections in minority languages, and describes some contexts in which it is used. The system can make multilingual documents available in structured collections, and allows them to be accessed via multilingual interfaces. It is issued under a free open source license, which encourages participatory design of the software, and an end-user interface allows community-based localization of the various language interfaces—of which there are many.

Keywords: digital libraries, translation, participatory design, localization

1. Introduction

Digital libraries have a crucial role to play in the preservation and maintenance of international cultures in general and minority languages in particular. Libraries and their close relatives, museums, have always been involved in preserving culture. These institutions collect literature and artefacts, and use them to disseminate knowledge and understanding of different times and cultures. Digital libraries open up the possibility of flexible and coherent multimedia collections that are both fully searchable, and browsable in multiple dimensions—and they permits active participation by indigenous people in preserving and disseminating their own culture.

Because language is the vehicle of thought, communication, and cultural identity, a crucial feature of digital libraries for culture preservation and revitalization is the ability to work in local languages. This strengthens individual cultures, promotes diversity, and reduces the overwhelming dominance of English and other majority languages in the global information infrastructure. Another crucial feature is to put the power to create and disseminate information collections into the hands of local people rather than external philanthropists. We learned this when working with digital libraries in developing countries, where we observed that effective human development blossoms from empowerment rather than gifting (Witten *et al*., 2002; Witten, Bainbridge and Boddie, 2001). Disseminating information originating in the developed world, as the Web tends to do, is very useful for developing countries. But for sustained long-term development the most effective strategy is to disseminate the capability to create information collections rather than the

collections themselves. This allows developing countries to participate actively in our information society rather than observing it from outside, and avoid becoming read-only societies in the information revolution. It will stimulate the creation of new industry. And it will help ensure that intellectual property remains where it belongs—in the hands of those who produce it.

This article describes some of our work on digital libraries in minority languages. We have constructed a software tool for building digital libraries that empowers *non computer experts* to create, organize, and distribute large collections of information. Called Greenstone, it is distributed widely under a free open source licence. It allows participatory design of information collections by indigenous people. Our work began locally, with the Māori language of indigenous New Zealanders, and then spread to a broader international context. Greenstone is widely used in the developed world, with many sites at major institutions in the US, for example. But it has also been widely adopted in the developing world. For example, volunteers have contributed interfaces in almost 40 languages[1]—a testament to the enthusiasm with which people embrace the opportunity to see libraries presented in their own languages.

We begin by reviewing other work on technology and minority languages, and outline the relevance of tools for creating digital libraries. Section 3 describes the Greenstone software and how it is used to create and access collections. The elements that enable the participatory localization of the user interfaces are explained in Section 4. Finally, we outline of how the system is used and extended in two separate minority language communities who have enthusiastically adopted it to serve collections in their language through an appropriate language interface.

2. Digital libraries and languages

Globally networked computer technology is both a threat and an opportunity for minority languages (Cazden, 2003). It can emphasize dominant languages, such as English, yet also connect dispersed language groups in new ways (Almasude, 1999; Crystal, 2000; Nettle and Romaine, 2000). Previous work on language maintenance has identified many factors that are associated with successful language projects. Crystal (2000) condenses the common factors into six pre-requisites for language revitalization. His sixth factor is particularly relevant to this paper:

> An endangered language will progress if its speakers can make use of electronic technology. (Crystal, 2000)

The Web has lowered barriers to publishing for both individuals and organizations and as browser technology improves more and more languages have appeared online. Countries with more than one official language often provide multilingual official websites, although the techniques for building and maintaining such sites are still evolving (Cunliffe *et al.*, 2002). The complexities of designing and maintaining websites lead many organizations toward content management systems that emphasize structural approaches to managing large amounts of data. The problems addressed by these systems are the same problems that librarians have dealt with for years: organizing and making accessible large amounts of information. Digital libraries represent a solution to this problem, and are practical tools for preserving and revitalizing minority languages (Lu *et al*. 2004).

As production of digital documents increases, people often want to preserve their documents and disseminate them to a wider audience. The transition from local *ad hoc* solutions to organized digital libraries has been described in the case of children's creation of bilingual digital books using *Fabula* project software

[1]Arabic, Armenian, Bengali, Bosnian, Catalan, Croatian, Chinese (Traditional), Chinese (Simplified), Czech, Dutch, English, Farsi, Finnish, French, Galician, Georgian, German, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Kannada, Kazakh, Kyrgyz, Latvian, Māori, Mongolian, Polish, Portuguese (Brazil), Portuguese (Portugal), Russian, Serbian, Spanish, Thai, Turkish, Ukranian, and Vietnamese.

(Edwards *et al.* 2002). Two major threads in the use of digital libraries can be identified: a top-down approach that preserves and distributes documents (e.g. Lu *et al.* 2004, Miyashita and Moll 1999), and a bottom-up one that aims to provide minority language groups with multilingual tools that they can use in whatever manner they choose.

Valiquette (1998) notes that using technology can "often involve handing over control to technical experts"; which can become a short term 'technofix' rather than an effective long term strategy. Eisenlohr (2004) describes the social and political implications of external experts making decisions about which artefacts and resources to include or exclude in digital archives. As Crystal (2000) notes, it is important that the language speakers *themselves* make use of the technology. Consequently the systems and tools that are used should support both the top-down and bottom-up approaches to computer-based language projects.

A further consideration for language revitalization projects is the localization of software. Localization refers to the adaptation of a product to suit a target language and culture (Crystal 2000, p.143). Warschauer (1998) describes how the Hawaiian language community developed their own software systems because they could not find localised versions appropriate to their needs. The importance of the notion of localization has steadily grown, to the extent that it is now regarded as an industry in itself. However, given the complexities and rapid change of modern software it is infeasible to expect software developers to maintain localised versions of all their products (Edwards *et al.* 2002). Purvis *et al.* (2001) note that greater attention is now being paid to software architectures that make it easier to adapt to different language environments. A distinction is made between *internationalization* of the architecture and the specific *localization* work necessary to adapt software to a specific language and culture.

Most proprietary software has restrictions that prevent users from adapting and extending it to suit their local circumstances—restrictions that may be legal, in the form of licenses that prohibit changing the software, or technical, as with systems whose source code unavailability effectively precludes language localization even when it is in principle permitted. The end result is that language communities have less power (Eisenlohr, 2004) and are forced to become software developers (Warschauer, 1998).

Buszard-Welcher (2001) notes that where technological expertise is concentrated in a few members of a community then there is a risk of burnout for the people doing all the work; the corollary of this that the tools need to be widely available and easy to use to spread the workload. However, the complexity of many tools for distributing sizable collections over the Web can lead to undesirable concentrations of expertise and consequent risks to successful tool use in information dissemination.

In the next section we describe a multi-lingual digital library tool, Greenstone, that can be used to distribute collections of documents over the Web and is flexible enough to be localised and customized to support different language communities.

3. The Greenstone digital library software

Greenstone is a suite of software for building and distributing digital library collections (Witten and Bainbridge, 2003). 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 collection. It is open-source, multilingual software, issued under the terms of the GNU General Public License. Collections built with Greenstone automatically include effective full-text searching and metadata-based browsing facilities that are attractive and easy to use. They are easily maintainable and can be rebuilt entirely automatically. Different indexes can be constructed (including metadata indexes). Browsing utilizes hierarchical structures that are created automatically from metadata associated with the source documents. Collections can include text, pictures, audio, and video, and the interface to collections can be extensively customized.

Most digital libraries are accessed over the web, and the interface to Greenstone uses a web browser accordingly. However, in many developing country environments, internet access is not as pervasive as in developed ones and it is often preferable to run the server locally. Furthermore, if people are to build and control their own libraries it is convenient for them if the software runs standalone on their own computers. Thus digital library software intended for broad access should run on a wide variety of computer systems, particularly low-end ones. The Greenstone server runs on any Windows, Unix, and MacOS/X system, and incorporates its own web serving software that can be used locally even on a standalone machine.

Greenstone is international software, and employs the Unicode character set throughout. Documents in any language and character encoding can be imported. Example collections in Arabic, Chinese, Cyrillic, English, French, Spanish, German, Hindi, and Māori are publicly available at the New Zealand Digital Library website (http://www.nzdl.org). The Greenstone web site (http://www.greenstone.org) links to sites that contain further examples, built locally in languages such as Chinese, Croatian, Dutch, Hawaiian, Hindi, Italian, Kannada, Kyrgyz, Portuguese, Russian, Spanish, Vietnamese, and Welsh. It makes little sense (and is sometimes distasteful) to have a collection whose content is in Chinese or Hindi, but whose supporting text—instructions, navigation buttons, labels, images, help text, and so on—can only be seen in English. Consequently, the entire Greenstone interface has been translated into a range of languages, and the interface language can be changed by the user as they browse from the *Preferences* page. Currently, interfaces are available in almost 40 languages.

In an international cooperative effort established in August 2000 with UNESCO and the Belgium-based Human Info NGO, Greenstone is being distributed widely in developing countries with the aim of empowering users, particularly in universities, libraries, and other public service institutions, to build their own digital libraries. UNESCO recognizes that digital libraries are radically reforming how information is acquired and disseminated in its partner communities and institutions in the fields of education, science and culture around the world, particularly in developing countries, and hopes that this software will encourage the effective deployment of digital libraries to share information and place it in the public domain.

3.1 The reader's interface

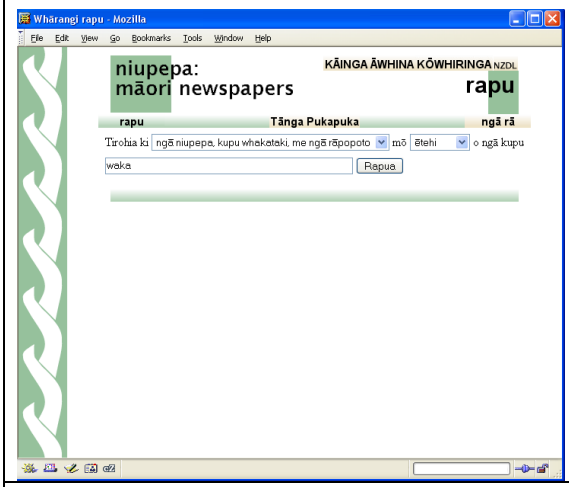
Greenstone collections can be presented on the Web or published as standalone libraries on removable media such as CD-ROM or DVD. Any Greenstone collection can be converted into a self-contained Windows CD-ROM/DVD that includes the Greenstone server software itself and an integrated installation package. The installation procedure has been thoroughly honed to ensure that only the most basic of computer skills are needed to install and run a collection under Windows. These CD-ROMs run on all Windows systems, right down to the antiquated Windows 3.1.

We illustrate the reader's interface to Greenstone using Niupepa, a collection of great local interest in our own environment. Māori are the indigenous people of New Zealand, and Niupepa is a collection of historic newspapers published primarily for a Māori audience between 1842 and 1932 (Apperley *et al*., 2002). This fascinating collection covers the period of European colonization (New Zealand, being remote, was discovered rather late by Europeans.) The newspapers can be searched (full text), browsed (by series) or accessed by date. The collection has been made available by the New Zealand Digital Library Project in the Department of Computer Science, University of Waikato.

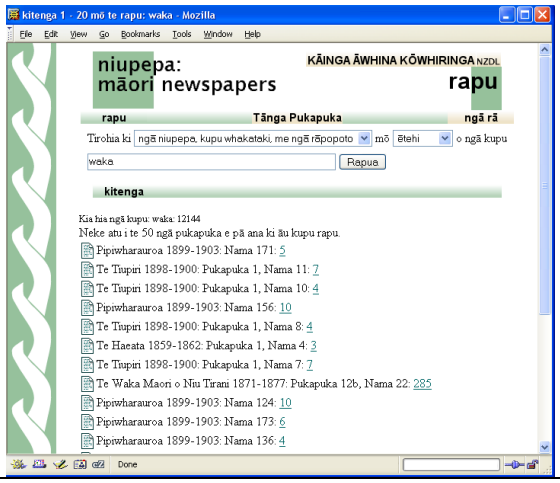
The *Niupepa* collection (Niupepa 2005) contains over 17,000 newspaper pages taken from 34 separate periodicals, some of which were government sponsored, others initiated by Māori, and the remainder by religious groups. The collection is based on Niupepa 1842-1933, a microfiche collection produced by the Alexander Turnbull Library in New Zealand. Most (70%) of the collection is written solely in Māori, some (27%) is bilingual and a small proportion (3%) is in English only. The digital library collection has

repositioned these Māori newspapers from extremely restricted access using microfiche readers in particular libraries to global availability from any Internet terminal. In addition, and equally importantly, new access mechanisms like full text searching have been added. These changes provide a baseline that educators, historians, and researchers can exploit to design educational activities involving this—the largest single body of machine-readable Māori text.

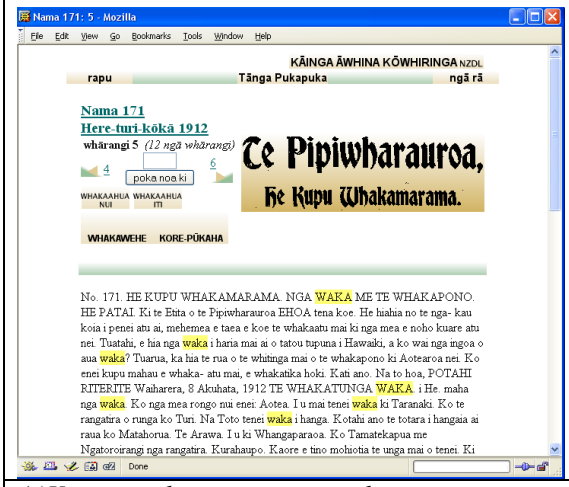
Figure 1 shows various snapshots of the collection in use. The user begins by viewing the search page to learn what is in the collection (Figure 1a). Next they search the full text of the newspapers for occurrences of the word *waka*—Māori for *canoe* (Figure 1b). Scanning down the list of matching documents, they select the first item. Clicking on this brings up an initial view of the document (Figure 1c), a view of the extracted text with the search term highlighted. From here various views are possible, including the facsimile image of the newspaper's first page (Figure 1d). In Figure 1e the user is browsing by series title, first looking at the newspaper series, which also shows how many issues each series has, then (Figure 1f) expanding the bookshelf for *Anglo Maori Warder* to see the individual items.



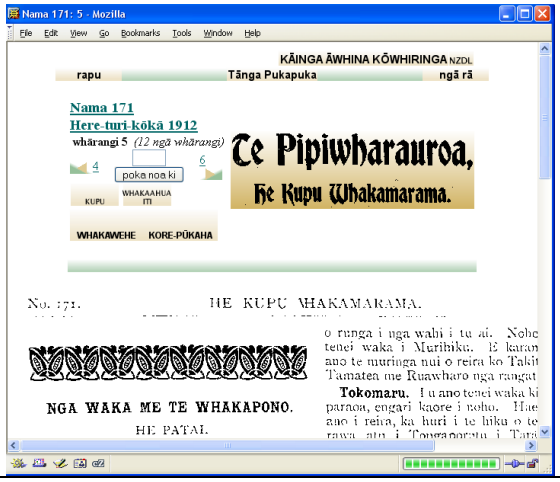
(a) Searching the collection for waka



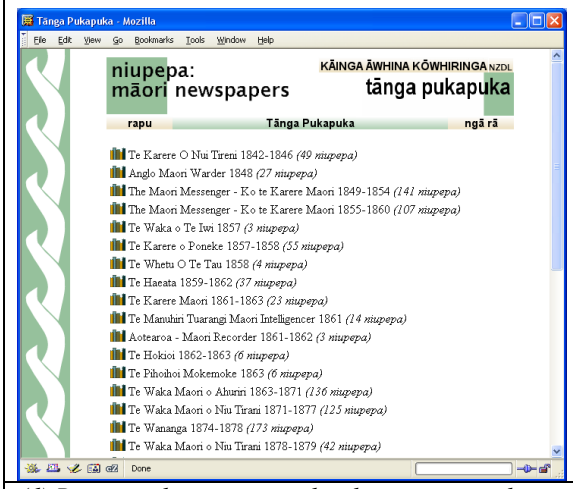
(b) Search results



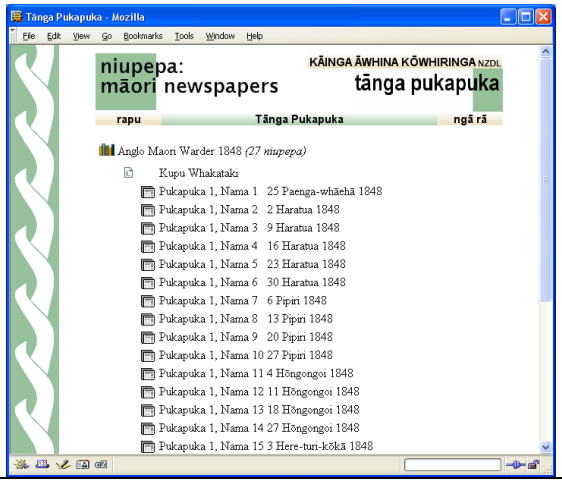
(c)Viewing a document: extracted text



(d) Viewing a document: scanned image



(d) Browsing by series: top-level newspaper titles



(e) Browsing by series: Anglo Maori Warder 1848

Figure 1. Snapshots of the *Niupepa* collection

3.2 The librarian's interface

Users whose skills are those of librarians rather than computer specialists can use Greenstone to build and distribute their own digital library collections. Figure 2 shows the librarian's interface, in which users can gather together a set of files (downloading them from the Web by mirroring parts of external web sites if necessary); manually augment these source documents with textual metadata if desired; perform a collection design step that determines its appearance and the access facilities it will support; build the collection including all data structures necessary for searching, browsing, and document access; and preview it in their web browser. From here, a couple of clicks can produce a self-installing CD-ROM version of the collection.

In Figure 2 the user is developing a collection of Georgian documents, which in this case are in Microsoft Word. Unfortunately in this case they cannot work in Georgian, because unlike the reader's interface, which is available in nearly 40 languages (including Georgian), the librarian's interface is currently only available in four: English, Spanish, French, and Russian. Instead they have chosen to switch the Librarian interface into Russian. They begin by creating a new collection using the file menu (Figure 2a), and fill out general information about the collection. In Figure 2b, for example, they have opted to copy the design of an existing collection, and are selecting from a menu of collections. Then a series of panels guides the user through the processes required to build the collection. The left-hand pane of the panel in Figure 2c shows the file system, and the right-hand one represents the contents of the collection, initially empty, which the user populates by dragging and dropping files. Then the user switches to another panel to add textual metadata (typically titles, authors, dates, keywords) to the selected documents, shown in Figure 2d.

Normally, at this stage the user would switch to a further panel to design the collection by selecting what full-text indexes and browsing facilities to add. For example, one might have a full-text index of the contents, and another of abstracts, and another of titles; and perhaps alphabetical browsers by title and author metadata, and another date browser. In this case, Greenstone's default settings are used and this stage is elided. Finally the user commands the collection to be built. This is shown in Figure 2e, and a scrolling log of program output is produced—also in the Russian language. Finally the user clicks a *Preview* button (shown in Figure 2e) to examine the collection from the reader's point of view as shown in Figure 2f, which illustrates an alphabetical title browser.

The interface explicitly supports four levels of user: Library Assistants, who can add documents and metadata to collections, and create new ones whose structure mirrors that of existing collections; Librarians, who can, in addition, design new collections, but cannot use specialist IT features (e.g. regular expressions); Library Systems Specialists, who can use all design features, but cannot perform troubleshooting tasks (e.g. interpreting debugging output from Perl scripts); and Experts, who can perform all functions. For example, the work in Figure 2 is appropriate for Library Assistant mode because although documents and metadata were added, no collection design was required.

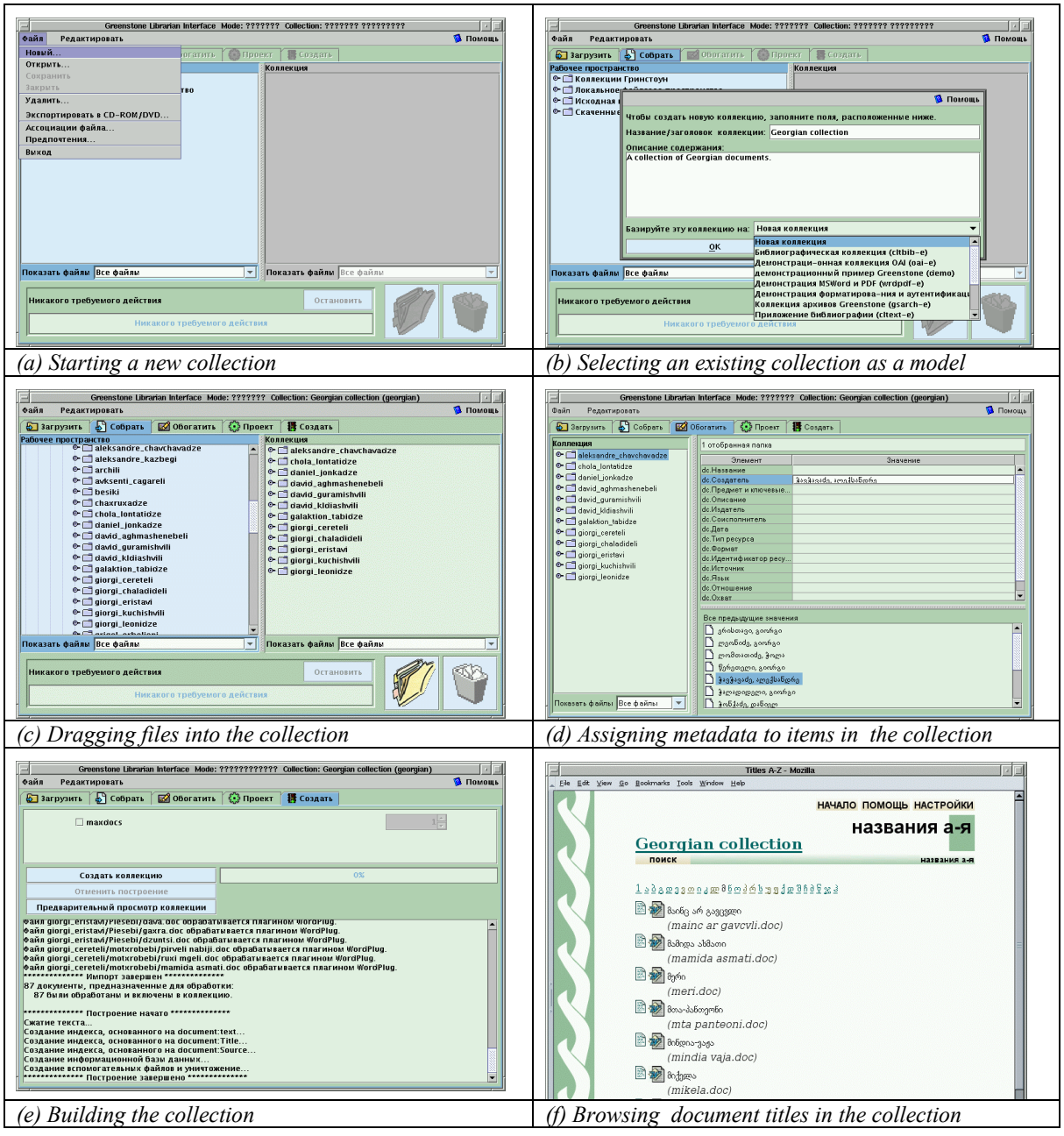


Figure 2. Building a collection of Georgian language documents

4. Managing multiple interface languages

The problem of maintaining an evolving multilingual digital library software system is severe—particularly when the software is open source. No single person knows all interface languages; no single person knows about all modifications to the software—indeed there is likely no overlap at all between those who translate the interface and those who develop the software. Currently, Greenstone has about 40 interface languages and there are around 750 linguistic fragments in each interface, ranging from single words like *search*, through short phrases like *search for*, *which contain*, *of the words*, to sentences like *More than ...*

documents matched the query, to complete paragraphs like those in the on-line help text. Maintaining the interface in many different languages (30,000 fragments), which is done by volunteers all around the world, is a logistic nightmare.

In following sections we describe the software tool we have developed to cope with this challenge. We do this by showing some examples of the translation system. The language fragments and how they fit together to form a page is managed in Greenstone by a *macro language* facility. This is at the heart of the system, and we make reference to it in the description below. The technique is simple yet surprising powerful. A macro consists of a name and its definition. Optionally it can take parameters, such as (*l=en*) to specify which language the definition corresponds too. Embedded within a macro definition there can be references to other macros (denoted with an underscore on either side, like *_this_*). The translation facility we have devised, however, is not specific to macro files: it can be adapted to any language management technique—such as one based on Java resource bundles—that records for each language and each item of text to be displayed a pair comprising a language independent label and a language dependent value. This approach combines two of the three aspects that Hogan *et al.* (2004) describes as key activities for interface internationalization: "externalization of UI [user interface] strings" and the "maintenance of a string database." The third activity, preparing the text for translation, is constrained by the structure of the interface and as internationalization has become more important in the Greenstone project greater care has been taken over selecting new interface elements.

The Greenstone translation facility helps users to perform three kinds of task:

- translate the interface into a new language,
- update an existing language interface to reflect new Greenstone facilities, and
- refine an existing language interface by correcting errors.

To enter the translation facility a user selects two languages: a “base language” and the target language they are translating into. The base language is usually English, as this is used to develop Greenstone and so is the most up-to-date representation of the interface. However, users are free to select other base languages if they prefer. Having selected these two languages, the user enters the main section of the tool as shown in Figure 3 (in this case the base language is English and the target language is Māori).

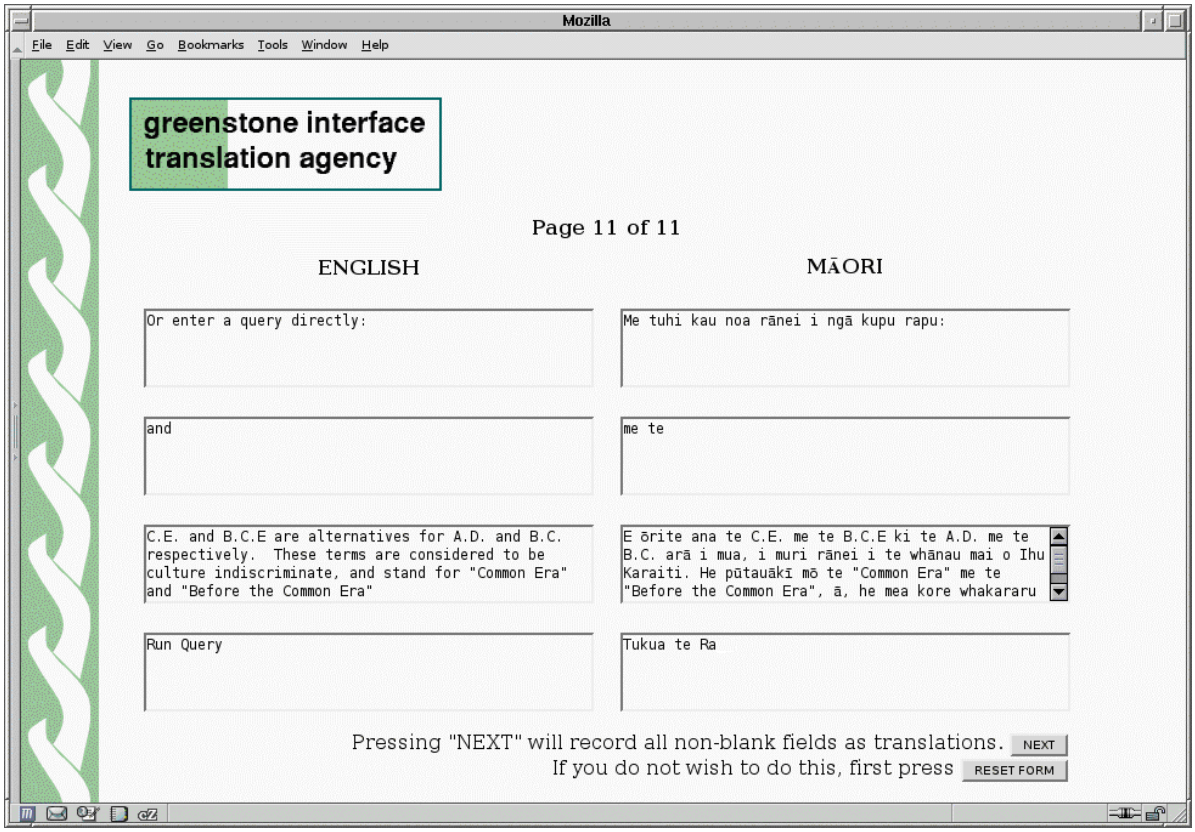


Figure 3 Updating a section of the Māori interface

4.1 Updating an existing language interface

In Figure 3 the user has begun to update the Māori language interface, using English as the base language. On the left are the base language phrases, and on the right are boxes into which translated versions can be entered. Two kinds of phrase appear: ones that are missing from the Māori version, and ones whose Māori translation is outdated because the English version has been edited more recently. In the latter case the outdated translation appears as a visual cue. In Figure 3 the user has methodically worked through the phrases and is in the process of entering text into the final box. After completing the translation the changes are committed back to the translation server.

Changes to the user's interface take place immediately: users can see their new translations in context by accessing (or reloading) appropriate pages in Greenstone. However, these changes are not made automatically to the public Greenstone site, nor are they automatically committed to the master software repository. Instead, to guard against error and misuse, they take effect in a special replica of the Greenstone site used for translation. If Greenstone encounters any phrases that have not been translated, the fallback strategy is to render them in the default language for the site, usually English. If desired, users can reset the page to its initial settings, or proceed to the next page without committing any changes. When satisfied with the entire translation, users notify the central Greenstone repository's administrator of the change through email. Then, issuing a single command fully integrates the changes into the officially released version of the software.

Because each page of translated text strings is saved when it is submitted, a user need not translate all phrases in one sitting. Moreover, when they return to the service the sequence of pages is regenerated, which means that only the outstanding phrases are shown. Greenstone distinguishes between phrases that are used in the main system—for instance, search, browsing and help pages—and phrases in less-frequently-used subsystems—for instance the site administration pages through which usage statistics and logs are viewed, and the translator service itself—for this too needs translating! For well-maintained language interfaces such as Spanish, French and Russian, only one or two pages of new translation requests are generated when new features are added. However, some less-used language interfaces contain translations only for the core phrases that appear in the main system.

4.2 Adding new languages

New languages are added in the same way that existing ones are updated, except that no existing translations appear in the right-hand column. A total of 75 pages of translations are generated, averaging 10 phrases each (750 in total). About 60% of these (450 phrases) pertain to the core Greenstone system, which every language interface covers; the remainder are for the less-used "auxiliary" parts of the interface. Of the existing language interfaces, 15 are for the complete interface and the remaining 23 cover just the core parts.

4.3 Character encoding issues

Because of the multilingual nature of Greenstone, careful attention must be paid to issues of character encoding. There are many different character encoding schemes in use today—as an example, the code 253 means "small letter Y with an acute accent" (ý) in the standard Western character set (International Organization for Standardization (ISO) standard 8859-1) while the same code corresponds to "a dot-less lower-case i" (ı) in the standard Turkish character set. All text in Greenstone, including the macro files, is handled internally using Unicode (UTF-8) (Unicode Consortium, 2004).

Unicode is an ISO standard providing every character in every language with a unique number. For example, in Unicode a Western "y" with an acute accent has the code 253, while a Western dot-less "i" has the code 305. Greenstone supports any character set that can be mapped onto Unicode, which includes the majority of sets currently in use world-wide. Modern browsers allow Unicode text to be entered into web forms. Unfortunately there is no standard way of forcing a browser to upload information in Unicode—or even to check what character set it uses to submit text fields. Some browsers always submit forms encoded in the user's default character set. However, major browsers generally submit forms using the same character set that is used for the current page, so in practice if pages are sent out with Unicode specified, returned text is usually encoded the same way.

4.4 Refining a language interface

Sometimes phrases in an existing language interface need to be refined. For example, a typographical error may have been overlooked when entering a phrase, or seeing a phrase in its actual interface context may suggest a better form of expression. To accommodate this requirement, users need to be able to locate an existing phrase and update its translation. One solution is to provide a complete list of all phrases in the language, not just the empty or outdated ones presented in Figure 3. The user could scan through this list to locate the desired phrase and correct or revise it. However, given the number of pages and phrases involved this would be a tedious and impractical task.

A more effective strategy is to allow users to locate a given phrase by searching for it, and then receive from the system a page that translates that one phrase. Interestingly, this idea can be implemented by making

the set of phrases into a multilingual digital library collection. Within Greenstone, this special collection is designed as follows. Treat each language as a document and each phrase as a section within it. For each document, store its language as metadata and use this as a discriminator to form sub-collections. Finally, set the collection to be a "private" (rather than a "public") one to prevent it from appearing on the site's home page. It can still be accessed by a URL that includes the collection name explicitly.

5. Usage examples

In this section we describe two examples of Greenstone's ability to support localization, in Māori and Hawaiian. The ability to create interfaces in multiple languages allows collection maintainers to alter the appearance of their collections. In the case of *Niupepa* we describe the effects of switching interfaces, while for the Hawaiian language collections of *Ulukau* we can see how extensibility can add valuable functionality.

5.1 Usage of Niupepa, a digital library in Māori

We collected 2004 usage statistics of the Niupepa collection and analysed them to determine usage by indigenous language speakers (Keegan and Cunningham 2005a, 2005b). We removed inappropriate activity (e.g. from web robots, local testing and incorrectly recorded data) and by using cookies were able to split the activity into sessions that accessed pages and/or undertook searching on the Niupepa website. There were 1370 sessions in which the user interface was set to Māori, 3649 in English, and 364 in which the user alternated between the two. We conclude that despite the overwhelming dominance of English in our local culture, digital libraries that make information available in the indigenous language are indeed utilized in that language: 25% of sessions were conducted wholly in Māori with a further 10% partially in Māori.

The session analysis showed some significant usage differences. The Māori language sessions were twice as likely to access newspaper pages by browsing the collection than the English ones, which preferred access by full text searching. The Māori language sessions are four times as likely to download full size images, presumably for serious reading. The Niupepa collection is written mostly in Māori so it seems plausible that a user in the medium of Māori would find it easier to browse the documents than one in the medium of English, and would also be more interested in downloading the full size image for screen viewing.

The Māori language sessions had a clear preference for accessing actual newspaper pages (which are in Māori), while English ones have a clear preference for accessing the commentary information (which is in English). This seems logical: users seek information that is available in the language of the interface. While full text search is an important method of accessing information for both kinds of session, it is utilised 10% more in English ones, presumably to compensate for users' inability to browse Māori language newspapers. A third session type comprises users who spend a significant amount of a single session in both languages. Generally these had characteristics that fell between those of the monolingual sessions.

We also found that the interface's default language strongly favours usage of the collection in that language—even though users could easily switch to the other language. Over a 4-week period we alternated the default language between Māori and English. We found that in either case, more sessions were conducted in the default language, i.e. the proportion of English sessions increased when the default language was set to English over what it was when the default was Māori, and vice versa. Setting the default language also produced longer sessions in that language, more pages in that language were accessed, and, if the default was Māori, a smaller reliance on searching as a means of access. These studies help to address the 'relative scarcity of published case-studies of bilingual developments' (Cunliffe *et al.*, 2002).

5.2 Ulukau, a digital library in Hawaiian

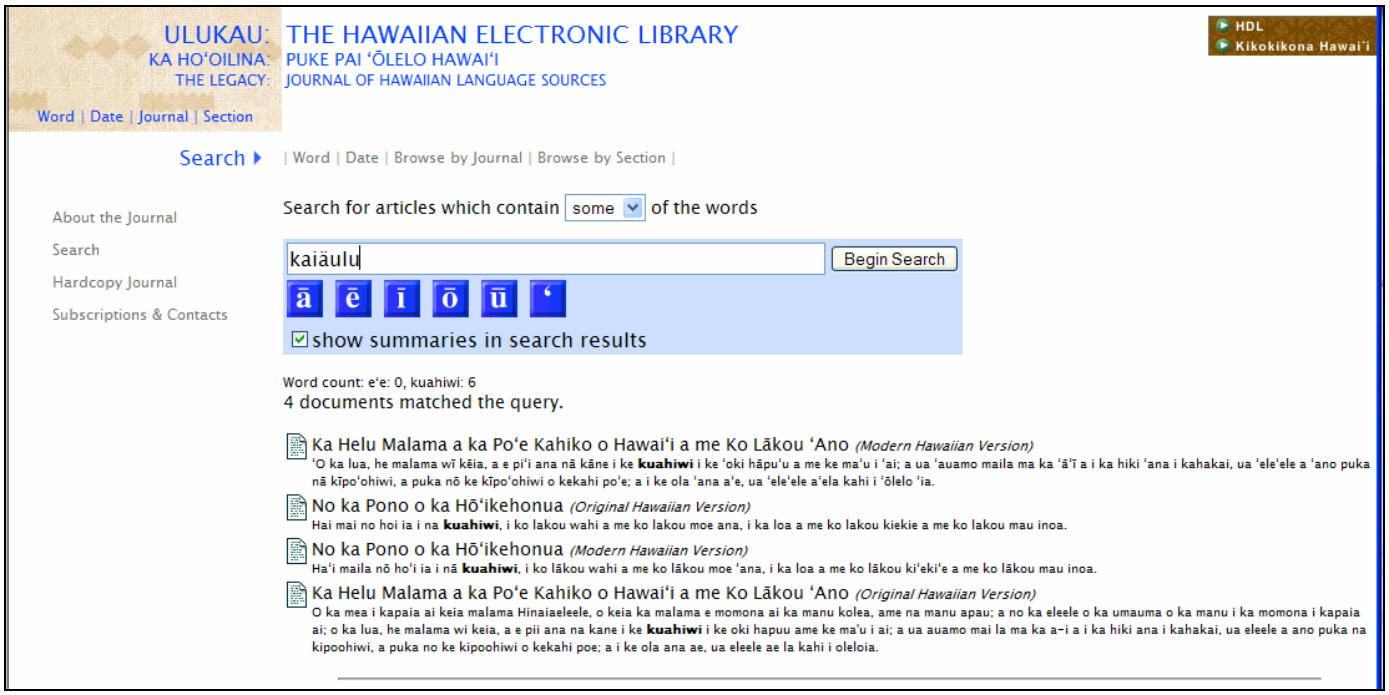


Figure 4 A search at the Greenstone-based collections at *Ulukau: the Hawaiian Electronic Library*

Figure 4 shows a screenshot of a search at *Ulukau: The Hawaiian Electronic Library* (Ulukau 2005), which uses a customized version of Greenstone. Its appearance differs from Figures 1 and 2; this customization is common in Greenstone collections and can be specified in the Librarian Interface during the design phases. However, the open source licensing of Greenstone allows for more extensive customization. Figure 4 shows a way of searching for specific Hawaiian characters, allowing users to create an accurate query without having to remember sequences of keystrokes. The open nature of Greenstone, in both licensing and technical terms, allows a high degree of localization; collections can be adapted to the specific requirements of minority language communities.

In the top right of Figure 4 is a link to switch to a full Hawaiian language interface at the current point in the user's session. Using the abstract structural models described in Cunliffe *et al.* (2002) *Ulukau* illustrates a 'direct language link' architecture from a monolingual home page (in Hawaiian). In contrast Niupepa immerses the user in a Māori-only environment once past the home page of the collection. Greenstone is flexible enough to support many different architectural customizations depending on the needs of its users.

6. Conclusion

We have shown how a tool for building digital libraries can support internalization and localization in many languages. Digital libraries are powerful vehicles for preserving and revitalizing minority languages, provided that they allow language communities to take control of the technology and customize it to meet their needs. Open source licensing provides a mechanism to enable this localization to occur, and the collaborative community approach of open source can also be extended to the translation of the interface

itself. Interesting organizational and maintenance issues arise when combining open source with localisation: no single person knows all interface languages; no single person knows about all modifications to the software—indeed there is likely no overlap at all between those who translate the interface and those who develop the software. With appropriate collaborative software, these problems can be overcome.

The provision of such software tools empowers communities to revitalize their languages in a contemporary technological environment. When the tool is free and open, it can be extensively customized to align its functionality more closely with users' needs. The examples of *Niupepa* and *Ulukau* show the power of localization to support minority language content with specialized interfaces. An interesting side-effect of the open licensing and distribution model is that we, Greenstone's developers, have little knowledge of how users are adapting the software to their local needs. We are frequently surprised to learn of imaginative new customizations and collections produced by groups around the world.

The structural and metadata-based approach at the heart of digital library software supports various different information architectures for serving collections. *Niupepa* customizes at the level of homepage, and then only provides language switching via the home page. *Ulukau* provides for cross-language linking at all levels of a collection. The internationalization-aware architecture of Greenstone also allows language-based studies, such as those in Section 5.1, to be performed easily. Thus in addition to empowering language groups to revitalize their languages using digital libraries, Greenstone can also act as a research tool for understanding the behaviour of users in multilingual environments.

References

Almasude, A., The New Mass Media and the Shaping of Amazigh Identity, in J Reyhner, G Cantoni, RNS Clair & EP Yazzie (eds), *Revitalizing Indigenous Languages*, 1999, (Center for Excellence in Education, Northern Arizona University: Flagstaff, Arizona), pp. 117-28.

Apperley, M., Keegan, T.T., Cunningham, S.J. and Witten, I.H., Delivering the Māori-language newspapers on the Internet, in J Curnow, N Hopa & J.McRae (eds), *Rere atu, taku manu! Discovering history, language and politics in the Māori-language newspapers*, 2002, (Auckland University Press: Auckland, New Zealand), pp. 211-32.

Buszard-Welcher, L., Can the Web help save my Language? in L Hinton & K Hale (eds), *The Green Book of Language Revitalization in Practice*, 2001, (Academic Press: San Diego, CA), pp. 331-48.

Cazden, C.B., Sustaining Indigenous Languages in Cyberspace, in J Reyhner, OV Trujillo, RL Carrasco & L Lockard (eds), *Nurturing Native Languages*, 2003, (Northern Arizona University: Flagstaff, AZ), pp. 53-7.

Crystal, D., *Language Death*, 2000 (Cambridge University Press: Cambridge, UK).

Cunliffe, D., Jones, H., Jarvis, M., Egan, K., Huws, R. and Munro, S., Information architecture for bilingual web sites. *Journal of the American Society for Information Science*, vol. 53, no. 10, 2002, pp. 866-73.

Bainbridge, D., Edgar, K.D., McPherson, J.R. and Witten, I.H., Managing change in a Digital Library System with Many Interface Languages. *Proceedings of the 7th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2003)*, pp. 350-61. Springer.

Edwards, V., Pemberton, L., Knight, J. and Monaghan, F., Fabula: A Bilingual Multimedia Authoring Environment for Children Exploring Minority Languages. *Language Learning & Technology*, vol. 6, no. 2, 2002, pp. 59-69.

Eisenlohr, P., Language Revitalization and New Technologies: cultures of electronic mediation and the refiguring of communities. *Annual Review of Anthropology*, vol. 33, 2004, pp. 21-45.

Hogan, J.M., Ho-Stuart, C. and Pham, B., Key challenges in software internationalisation. *Proceedings of the Australasian Workshop on Software Internationalisation (AWSI 2004). ACSW Frontiers 2004: Conferences on Research and Practice in Information Technology Volume 32*, pp. 187-94. Australian Computer Society, Inc.

Keegan, T.T. and Cunningham, S.J., Language Preference in a Bi-language Digital Library. *Proceedings of the Joint Conference on Digital Libraries (JCDL '05)*, 2005a to appear. ACM.

Keegan, T.T. and Cunningham, S.J., What happens if we switch the default language of a website? *Proceedings of the 1st International Conference on Web Information Systems and Technologies (WEBIST '05)*, 2005b to appear.

Lu, S., Liu, D., Fotouhi, F., Dong, M., Reynolds, R., Aristar, A., Ratliff, M., Nathan, G., Tan, J. and Powell, R., Language engineering for the Semantic Web: a digital library for endangered languages. *Information Research*, vol. 9, no. 3, 2004, pp. paper 176. http://informationr.net/ir/9-3/paper176.html

Miyashita, M. and Moll, L.A., Enhancing language material availability using computers, in J Reyhner, G Cantoni, RNS Clair & EP Yazzie (eds), *Revitalizing Indigenous Languages*, 1999, (Center for Excellence in Education, Northern Arizona University: Flagstaff, Arizona), pp. 113-6.

Nettle, D. and Romaine, S., *Vanishing Voices: the Extinction of the World's Languages*, 2000 (Oxford University Press: New York, NY).

Niupepa: Māori newspapers, 2005, http://nzdl.org/niupepa

Purvis, M., Hwang, P., Purvis, M., Madhavji, N. and Cranefield, S., A Practical Look at Software Internationalisation. *Transactions of the Society for Design and Process Science*, vol. 5, no. 3, 2001, pp. 79-90.

Ulukau: the Hawaiian Electronic Library, 2005, http://ulukau.org/

The Unicode Consortium, *The Unicode Standard, Version 4.0*, 2000 (Addison-Wesley: Boston, MA).

Valiquette, H.P., Community, professionals, and language preservation: First things first, in N Ostler (ed.), *Endangered languages: What role for the specialist (Proceedings of the 2nd FEL Conference)*, 1998, (Foundation for Endangered Languages: Bath, UK.), pp. 107-12.

Warschauer, M., Technology and indigenous language revitalization: Analyzing the experience of Hawai'i. *Canadian Modern Language Review*, vol. 55, no. 1, 1998, pp. 140-61.

Witten, I.H. and Bainbridge, D., *How to Build a Digital Library*, 2003 (Morgan Kaufmann: San Francisco, CA).

Witten, I.H., Bainbridge, D. and Boddie, S.J., Power to the people: end-user building of digital library collections. *Proceedings of the Joint Conference on Digital Libraries (JCDL'01)*, pp. 94-103. ACM Press.

Witten, I.H., Loots, M., Trujillo, M.F. and Bainbridge, D., The promise of digital libraries in developing countries. *The Electronic Library*, vol. 20, no. 1, 2002, pp. 7-13.

This is an electronic version of an article published in *New Review of Hypermedia and Multimedia*:

Nichols, D.M., Witten, I.H., Keegan, T.T., Bainbridge, D. and Dewsnip, M. (2005) Digital libraries and minority languages, *New Review of Hypermedia and Multimedia*, 11(2) 139-155.

http://dx.doi.org/10.1080/13614560500351071

New Review of Hypermedia and Multimedia is available online at:

http://www.informaworld.com/openurl?genre=article&issn=1361%2d4568&volume=11&issue=2&spage=139