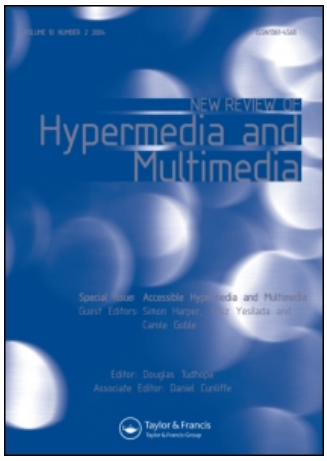
This article was downloaded by:[University of Waikato Library]
[University of Waikato Library]
On: 9 July 2007
Access Details: [subscription number 778559846]
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954
Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



New Review of Hypermedia and Multimedia
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713599880

Digital libraries and minority languages

David M. Nichols [a]; Ian H. Witten [a]; Te Taka Keegan [a]; David Bainbridge [a]; Michael Dewsnip [a]
[a] Department of Computer Science, University of Waikato. Hamilton. New Zealand

Online Publication Date: 01 December 2005
To cite this Article: Nichols, David M., Witten, Ian H., Keegan, Te Taka, Bainbridge, David and Dewsnip, Michael , (2005) 'Digital libraries and minority languages', New Review of Hypermedia and Multimedia, 11:2, 139 - 155
To link to this article: DOI: 10.1080/13614560500351071
URL: http://dx.doi.org/10.1080/13614560500351071

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.informaworld.com/terms-and-conditions-of-access.pdf

This article maybe used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

© Taylor and Francis 2007

New Review of Hypermedia and Multimedia,
Vol. 11, No. 2, December 2005, 139–155

Taylor & Francis
Taylor & Francis Group

Downloaded By: [University of Waikato Library] At: 00:53 9 July 2007

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

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 licence, 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 pivotal role to play in the preservation and maintenance of the world's culture. The mandate of museums and libraries is to collect and protect literature and artefacts, and use them to disseminate knowledge and understanding of different societies and their activities. The digital medium opens up new possibilities. Along with passive artefacts, we can now store flexible and coherent multimedia objects. Information can be presented in multiple ways and made accessible anywhere. Everything will be fully searchable, and conveniently browsable in multiple dimensions.

Digital libraries have another important application: they provide an excellent way to disseminate information in developing countries, where books are scarce. For example, many collections in various areas of human development have been created and published as digital libraries, both on the Web and on removable media—for Internet access is severely restricted in many parts of the world (Witten *et al*. 2002). A key advantage of digital technology is that it makes it cheap and easy to build and disseminate large-scale information collections, compared with the traditional world of book publishing, museums, and libraries.

In both scenarios, digital libraries lower the bar for participation by information producers and collectors. This opens up new opportunities for minority cultures and developing countries to participate actively in our

*Corresponding author. Email: dmn@ss.waikato.ac.nz

New Review of Hypermedia and Multimedia
ISSN 1361-4568 print/ISSN 1740-7842 online © 2005 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/13614560500351071

Downloaded By: [University of Waikato Library] At: 00:53 9 July 2007

information society, rather than observing it from outside. Although it is undeniably useful to disseminate information collections built in the developed world, as present digital libraries tend to do, a better strategy for sustained long-term development is to disseminate the capability to create information collections rather than the collections themselves. Effective human development blossoms from empowerment rather than gifting. Digital libraries enable indigenous people to participate actively in preserving and disseminating their own culture.

Language is the vehicle of thought and communication, and an important manifestation of cultural identity. If digital libraries are to be used to preserve and revitalize culture, and to assist in human development in general, they must operate in local languages. This will strengthen individual cultures, promote diversity, and reduce the overwhelming dominance of English and other majority languages in the global information infrastructure.

This article describes some of our work on enabling people to build and access 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—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—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 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,

Downloaded By: [University of Waikato Library] At: 00:53 9 July 2007

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 prerequisites 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 Web sites, although the techniques for building and maintaining such sites are still evolving (Cunliffe *et al.* 2002). The complexities of designing and maintaining Web sites 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 (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. Miyashita and Moll 1999, Lu *et al.* 2004), 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 localized 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,

Downloaded By: [University of Waikato Library] At: 00:53 9 July 2007

given the complexities and rapid change of modern software, it is infeasible to expect software developers to maintain localized 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, 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 sizeable 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 localized and customized to support different language communities.

3. 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

Downloaded By: [University of Waikato Library] At: 00:53 9 July 2007

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 systems. 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 Web site (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 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

Downloaded By: [University of Waikato Library] At: 00:53 9 July 2007

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 selecting the search page to initiate a search (figure 1a). The full text of the newspapers is searched for occurrences of the word waka (Māori for canoe), and the results are displayed (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.

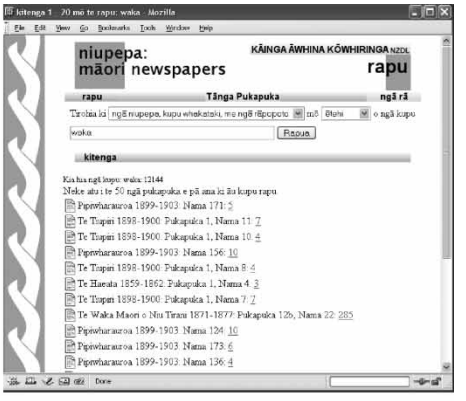
3.2 Librarian's interface

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.

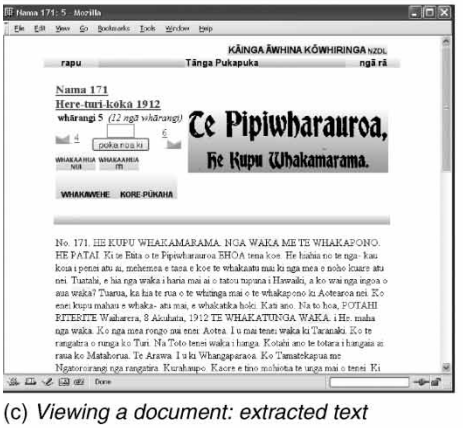
In figure 2, the user is developing a collection of Georgian documents, which in this case are in Microsoft Word. Unfortunately, in this case, they cannot work in Georgian, because unlike the reader's interface, which is

Downloaded By: [University of Waikato Library] At: 00:53 9 July 2007

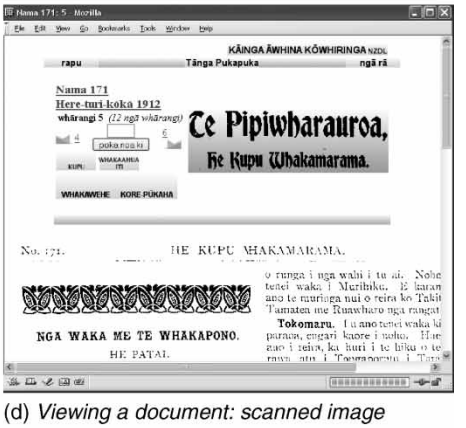
(a) *Searching the collection for* waka



(b) *Search results*

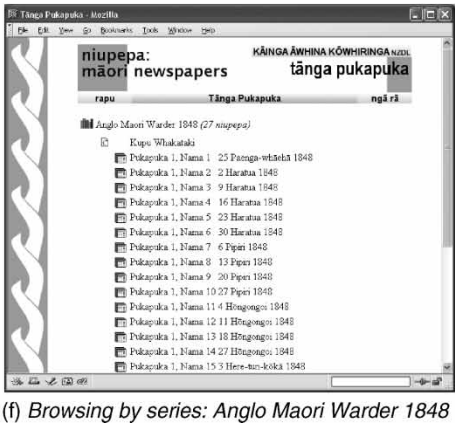


(c) *Viewing a document: extracted text*



(d) *Viewing a document: scanned image*

(e) *Browsing by series: top-level newspaper titles*

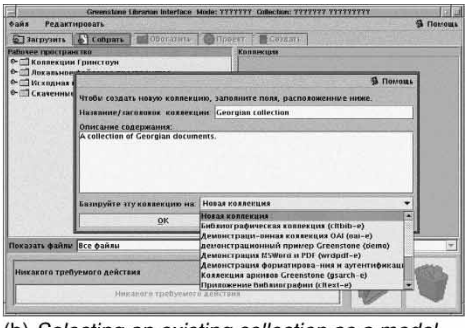


(f) *Browsing by series: Anglo Maori Warder 1848*

Figure 1. Snapshots of the *Niupepa* collection. (a) Searching the collection for *waka*. (b) Search results. (c) Viewing a document: extracted text. (d) Viewing a document: scanned image. (e) Browsing by series: top-level newspaper titles. (f) Browsing by series: Anglo Maori Warder 1848.

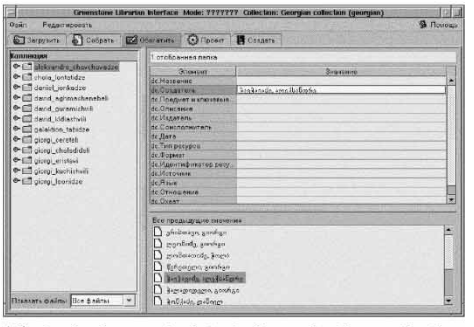
Downloaded By: [University of Waikato Library] At: 00:53 9 July 2007

(a) *Starting a new collection*

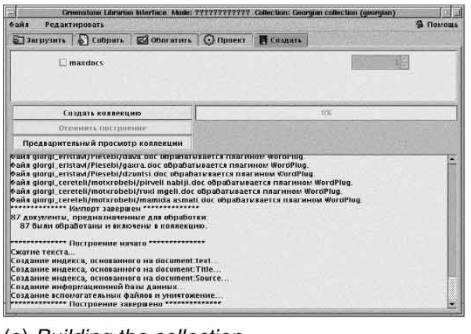


(b) *Selecting an existing collection as a model*

(c) *Dragging files into the collection*



(d) *Assigning metadata to items in the collection*



(e) *Building the collection*

(f) *Browsing document titles in the collection*

Figure 2. Building a collection of Georgian language documents. (a) Starting a new collection. (b) Selecting an existing collection as a model. (c) Dragging files into the collection. (d) Assigning metadata to items in the collection. (e) Building the collection. (f) Browsing document titles in the collection.

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

Downloaded By: [University of Waikato Library] At: 00:53 9 July 2007

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 pane 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.

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 online 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 (Bainbridge *et al*. 2003).

Downloaded By: [University of Waikato Library] At: 00:53 9 July 2007

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 web page are 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 surprisingly 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 to. 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) describe 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:

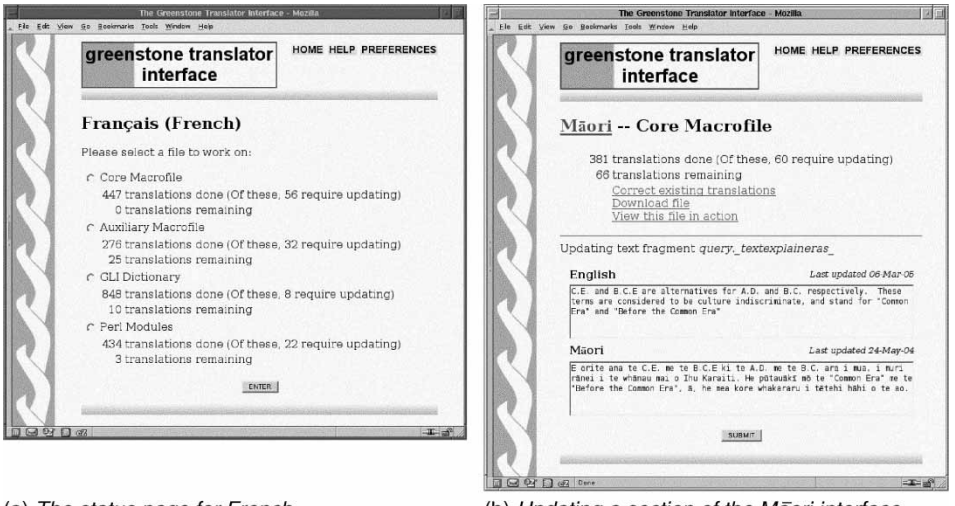
- translate the interface into a new language;
- update an existing language interface to reflect new Greenstone facilities; and
- refine an existing language interface by correcting errors.

To enter the translation facility, a user first selects the target language they are translating into. The base language is always English, because this is used to develop Greenstone and so is the most up-to-date representation of the interface. Originally, we planned to allow users to select other base languages, but we removed this facility for practical reasons: we did not want to compound any errors by using a base language that was incomplete or included incorrect translations.

Having selected the target language, the user is shown a status page for that language. 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! The phrases are divided into a few sections that reflect this distinction; currently, there are four. For each one, the status page shows the number of translations that have been done and the number still left to do, as illustrated (for French) in figure 3a.

After selecting one of these sections to work on, the translator must log in to the system by presenting a username and password before beginning work.

Downloaded By: [University of Waikato Library] At: 00:53 9 July 2007



(a) *The status page for French* (b) *Updating a section of the Māori interface*

Figure 3. Using the translator's interface. (a) Status page for French. (b) Updating a section of the Māori interface.

4.1 Updating an existing language interface

In figure 3b, the user has begun to update the core macro file for the Māori language interface. A single language fragment is shown, and when this string was last updated. Also included is a progress indicator: how many fragments remain to be done. The English phrase appears at the top; below it is the box into which the translated version can be entered. Two kinds of phrase appear: phrases that are missing from the Māori version, and phrases 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 (as in figure 3b). 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. When satisfied with the entire translation, users notify the central Greenstone repository's administrator of the change through e-mail. Then, issuing a single command fully integrates the changes into the officially released version of the software.

Because each translated text string is saved when it is submitted, a user need not translate all phrases in one sitting. Moreover, when they return to the service, the system regenerates everything from scratch, which means that

Downloaded By: [University of Waikato Library] At: 00:53 9 July 2007

only the outstanding phrases are shown. For well-maintained language interfaces such as Spanish, French, and Russian, only a few new translation requests are generated when new features are added.

4.2 Adding new languages

New languages are added in the same way that existing languages are updated, except that no existing translations appear in the right-hand column. A would-be translator e-mails the system administrator to get a username and password, and the administrator manually adds the new language to the list. There are a total of about 750 phrases in the entire interface (the 'Core Macrofile' and 'Auxiliary Macrofile' parts in figure 3a). Of these, about 60% (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 worldwide. 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 in 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

Downloaded By: [University of Waikato Library] At: 00:53 9 July 2007

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. Consequently, each page of the interface contains a link to a search facility that allows the user to find all fragments that contain the specified search terms.

5. Usage examples

In this section, we describe two examples of Greenstone's ability to support localization, in Hawaiian and in Māori. The ability to create interfaces in multiple languages allows collection maintainers to alter the appearance of their collections and gives multi-lingual users the ability to use different browsing and searching strategies that occur when using different languages.

5.1 Ulukau, a digital library in Hawaiian

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.

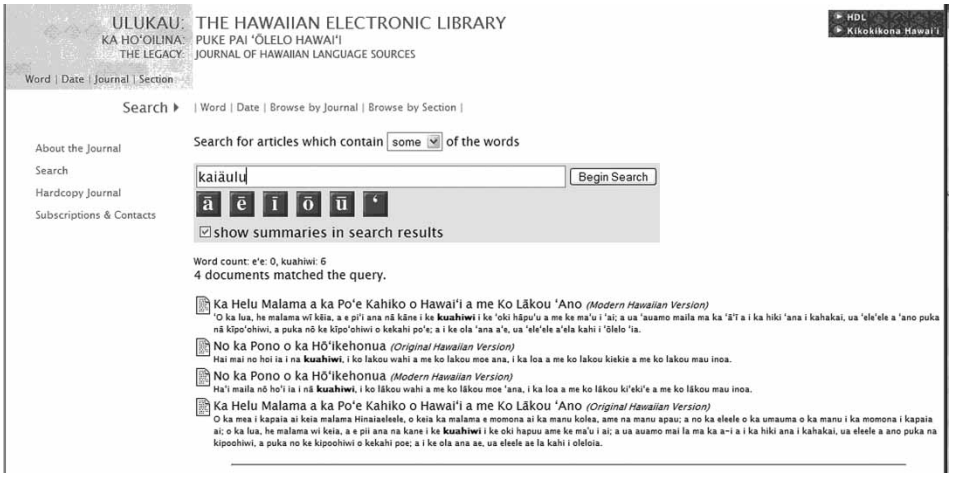


Figure 4. Search at the Greenstone-based collections at *Ulukau: the Hawaiian Electronic Library*.

Downloaded By: [University of Waikato Library] At: 00:53 9 July 2007

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.

5.2 Usage of Niupepa, a digital library in Māori

Having created interfaces in indigenous languages, we were interested to see how much they were actually used in the indigenous language when compared with English. We were also interested to determine if there were any differences in browsing strategies and searching techniques by users of an interface in an indigenous language when compared with users of the same interface in English. We analysed the 2004 Web log of the *Niupepa* collection (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* Web site, i.e. sessions where the user actually interacted with the information of the digital library. Of these sessions, 1370 (25%) were when the user interface was set to Māori, 3649 (68%) were when the user interface was set to English, and 364 (7%) were when the user switched between the two languages. We concluded 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 7% partially in Māori.

The session analysis also highlighted some differences in browsing strategies. The Māori language sessions were twice as likely to access newspaper pages by browsing the collection than the English sessions, which had the highest preference to make use of the full text-search facility. The Māori language sessions were four times as likely to download full-size images, presumably for online 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. However, even with the language of the material taken into account, the session analysis does suggest an indigenous preference for browsing as opposed to full text searching. This has implications for designers of indigenous-language digital libraries.

While undertaking the session analysis, there was another characteristic that became apparent. Aside from the home page, the *Niupepa* digital library does not offer a direct link to switch languages, yet some users (7% of the sessions) were still choosing to switch languages. These bilingual users had a

Downloaded By: [University of Waikato Library] At: 00:53 9 July 2007

more varied approach to accessing the information. They made use of both browsing and full text-searching facilities, and used both English and Māori as search terms. This further emphasizes the importance of having digital-library collections available in many languages to allow multi-lingual users to use all their languages when accessing information.

This initial analysis of the *Niupepa* Web log has shown that indigenous information stored in an indigenous language environment is accessed by people in the indigenous language. It has also highlighted some key preferences in indigenous language usage. More in-depth studies are required to confirm these initial findings and to help address the 'relative scarcity of published case-studies of bilingual developments' (Cunliffe *et al.* 2002).

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 localization: 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 home page, 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, Green-

Downloaded By: [University of Waikato Library] At: 00:53 9 July 2007

stone can also act as a research tool for understanding the behaviour of users in multilingual environments.

References

A. Almasude, "The new mass media and the shaping of Amazigh identity", in *Revitalizing Indigenous Languages*, J. Reyhner, G. Cantoni, R.N.S. Clair and E.P. Yazzie, Eds., Flagstaff, AZ: Center for Excellence in Education, Northern Arizona University, 1999, pp. 117–128.

M. Apperley, T.T. Keegan, S.J. Cunningham and I.H. Witten, "Delivering the Maori-language newspapers on the internet", in *Rere atu, taku manu! Discovering History, Language and Politics in the Māori-language Newspapers*, J. Curnow, N. Hopa and J. McRae, Eds., Auckland, New Zealand: Auckland University Press, 2002, pp. 211–232.

D. Bainbridge, K.D. Edgar, J.R. McPherson and I.H. Witten, "Managing change in a digital library system with many interface languages", in *Proceedings of the 7th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2003)*, New York: Springer, 2003, pp. 350–361.

L. Buszard-Welcher, "Can the web help save my language?", in *The Green Book of Language Revitalization in Practice*, L. Hinton and K. Hale, Eds., San Diego, CA: Academic Press, 2001, pp. 331–348.

C.B. Cazden, "Sustaining indigenous languages in cyberspace", in *Nurturing Native Languages*, J. Reyhner, O.V. Trujillo, R.L. Carrasco and L. Lockard, Eds., Flagstaff, AZ: Northern Arizona University, 2003, pp. 53–57.

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

D. Cunliffe, H. Jones, M. Jarvis, K. Egan, R. Huws and S. Munro, "Information architecture for bilingual web sites", *J. Am. Soc. Inform. Sci.*, 53, pp. 866–873, 2002.

V. Edwards, L. Pemberton, J. Knight and F. Monaghan, "Fabula: A bilingual multimedia authoring environment for children exploring minority languages", *Lang. Learn. Technol.*, 6, pp. 59–69, 2002.

P. Eisenlohr, "Language revitalization and new technologies: cultures of electronic mediation and the refiguring of communities", *Annu. Rev. Anthropol.*, 33, pp. 21–45, 2004.

J.M. Hogan, C. Ho-Stuart and B. Pham, "Key challenges in software internationalisation", in *Proceedings of the Australasian Workshop on Software Internationalisation (AWSI 2004). ACSW Frontiers 2004: Conferences on Research and Practice in Information Technology Volume 32*, Sydney: Australian Computer Society, 2004, pp. 187–194.

T.T. Keegan and S.J. Cunningham, "Language preference in a bi-language digital library", in *Proceedings of the Joint Conference on Digital Libraries (JCDL '05)*, New York: ACM Press, 2005a, pp. 174–175.

T.T. Keegan and S.J. Cunningham, "What happens if we switch the default language of a website?", in *Proceedings of the 1st International Conference on Web Information Systems and Technologies (WEBIST '05)*, Setúbal, Portugal: INSTICC, 2005b, pp. 263–269. http://www.insticc.net/index.php?option=com_contact&Itemid=3

S. Lu, D. Liu, F. Fotouhi, M. Dong, R. Reynolds, A. Aristar, M. Ratliff, G. Nathan, J. Tan and R. Powell, "Language engineering for the Semantic Web: a digital library for endangered languages", *Information Research*, 9(3), paper 176, 2004. Available online at: informationr.net/ir/9-3/paper176.html (accessed 25 August 2005).

M. Miyashita and L.A. Moll, "Enhancing language material availability using computers", in *Revitalizing Indigenous Languages*, J. Reyhner, G. Cantoni, R.N.S. Clair and E.P. Yazzie, Eds., Flagstaff, AZ: Center for Excellence in Education, Northern Arizona University, 1999, pp. 113–116.

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

Niupepa: Māori newspapers, 2005. Available online at: nzdl.org/niupepa (accessed 25 August 2005).

M. Purvis, P. Hwang, M. Purvis, N. Madhavji and S. Cranefield, "A practical look at software internationalisation", *Trans. Soc. Design Process Sci.*, 5, pp. 79–90, 2001.

Ulukau: the Hawaiian Electronic Library, 2005. Available online at: ulukau.org (accessed 25 August 2005).

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

H.P. Valiquette, "Community, professionals, and language preservation: First things first", in *Endangered Languages: What Role for the Specialist? (Proceedings of the 2nd FEL Conference)*, N. Ostler, Ed., Bath, UK: Foundation for Endangered Languages, 1998, pp. 107–112.

M. Warschauer, "Technology and indigenous language revitalization: Analyzing the experience of Hawai'i", *Can. Mod. Lang. Rev.*, 55, pp. 140–161, 1998.
I.H. Witten and D. Bainbridge, *How to Build a Digital Library*, San Francisco, CA: Morgan Kaufmann, 2003.
I.H. Witten, M. Loots, M.F. Trujillo and D. Bainbridge, "The promise of digital libraries in developing countries", *The Electronic Library*, 20, pp. 7–13, 2002.

Downloaded By: [University of Waikato Library] At: 00:53 9 July 2007