Issues of Content and Structure for a Multilingual Web Site

Shihong Huang
Department of Computer Science
University of California, Riverside
shihong@cs.ucr.edu

Scott Tilley
Department of Computer Science
University of California, Riverside
stilley@cs.ucr.edu

ABSTRACT
Most content on the Web today is in English, but the majority of the Earth's peoples speak languages other than English. To reach a wider audience, future Web sites will have to be multilingual, changing a Web site from one that is American-centric and single-language to one that is globally-oriented and multilingual. While the challenges in creating and maintaining a high-quality Web site in a single language are considerable, working with multiple languages simultaneously creates special challenges, both culturally and technically. This paper outlines issues related to two important aspects of the problem: content and structure. Several representative Web sites are examined to illustrate some of these considerations.

Categories and Subject Descriptors
D.2.7 [**Software Engineering**]: Distribution, Maintenance, and Enhancement – Documentation

General Terms
Design, Documentation, Human Factors, Standardization

Keywords
Multilingual, structure, content management, software engineering, Web site

1. INTRODUCTION

The beginning of the new millennium provides a singular opportunity to view Web sites in a new perspective: as a vehicle for truly global communication. One of the original goals of the Web was to create a new medium for communication that was truly universal. All users, regardless of their background and origin, should be able to experience the Web with equal results in terms of content impact without the limits of specialized equipment.

In early 2001, approximately 80% of Web content is in English, but only 45% of Web surfers have English as their first language [1]. It is projected that the surge in online users in the coming years – from China in particular – will have a dramatic impact on the very nature of the Web itself. Therefore, it seems clear that English-only Web sites will soon become a limiting factor in the intellectual exchange of ideas and information, and that this problem will quickly become an important long-term consideration for successful Web sites.

For example, consider non-English speaking people living and working in the Pacific Rim. For them, opportunities for cultural exchange and technological advancement are often limited to sources of information available in their native language. For example, news stories, recent developments in computing, and social impacts of world-wide events can only be experienced by Chinese-speaking people if the content is made available in Chinese. The translation from English to Chinese can be a long-term affair, resulting in the delivery of timely information by traditional means, such as printed books, that is "stale" by the time it reaches the consumer. By making the sources of this information available in Chinese as well as English over the Web, the delay in getting this information from producer to consumer can be eliminated.

Similarly, for content developers in China, their market is currently limited primarily to the Pacific Rim nations. If they were able to make their results more widely available, in English over the Web, then they could reach a much broader set of potential contacts and customers. By making the process of creating and maintaining multilingual Web sites easier for non-computer scientists, the opportunities for global communication are increased dramatically. This too addresses the original goal of the Web, as stated above.

SIGDOC '01, October 21-24, 2001, Santa Fe, New Mexico, USA.

The simplest multilingual Web site is a bilingual one, with (potentially) two versions of all content on the Web. When the two languages are relatively similar, such as English and French, the issues are not as pronounced. However, when the languages are quite dissimilar, such as German and Chinese, the issues are more difficult to solve in a consistent manner. When the site expands and goes from bilingual to multilingual, relatively simple issues such as what naming convention to use for the Web site (e.g., fr.yahoo.com or www.amazon.de or www.ibm.com/cn) become even more important.

Multilingual Web sites are fraught with both cultural issues and technical challenges. If the languages (and, by implication, the users' culture) are very dissimilar, then pre-conceived notions of what constitutes universal design criteria must be reexamined [5]. There may also be social conventions and linguistic or religious laws that must be adhered to. Although such cultural issues are extremely important, for the most part they are beyond the scope of this paper.

The technical challenges related to multilingual Web sites are themselves too broad to cover here in their entirety. The process of designing, developing, and deploying a multilingual Web site can benefit from adopting some of the same sound software engineering principles used in other application domains.

This paper outlines some of the issues related to both content and structure in developing and maintaining a multilingual Web site. The next section describes two important issues of content: management and localization. Section 3 discusses two important issues of structure: naming and organization. Section 4 illustrates aspects of these issues using a technical magazine Web site published in both English and Chinese. Finally, Section 5 summarizes the issues and offers some suggestions of managing the complexity inherent in content-rich multilingual Web sites.

2. CONTENT ISSUES

Issues related to the content of a multilingual Web site are numerous and inter-related. Content management focuses on the engineering side of the equation, maintaining consistency of translation across different versions of the Web site. Content localization is more related to the user experience, ensuring that a site has the appropriate level of native-language material.

2.1 Management

With a multilingual Web site, maintaining content consistency across different localized versions is time consuming and error prone. There are two primary approaches to content management: manual and semi-automatic. The manual approach is still the most common. Many international Web sites are still manually translated and manually maintained. Some of the Web sites separate their international content management systems from their main U.S. Web sites. Given the current rate of Web site evolution, this approach is very costly and time consuming, and makes it difficult to keep the site's contents synchronized in a timely manner [1].

There is a variant on manual content management that relies on the Web server. For example, the Apache Web server [1] can be configured to handle different content based on settings that reflect localization by using the mod-negotiation and Accept-Language modules, type maps, and character set directives. Language requests are made by the user's browser through an HTTP request header. URIs are used to map language-specific content to the language indicator when serving pages.

The second approach to content management is semi-automatic and relies on the use of globalization and localization software and services. Some of the better known global and localization service providers include Berlize [1], Bowne [2], GlobalSight [4], Idiom [6], and Uniscape [17].

This approach can also benefit by leveraging workflow management software. Used with some of the translation services mentioned above, this procedure is also known as translation workflow [1]. This approach takes much of the administrative work out of the process by performing site translation automatically (with human oversight) and synchronizing the content across all localized sites.

A popular translation workflow system is Lionbridge from LionBridge Technologies [8]. LionBridge is an automatic multilingual Web site content management system consisting of four main components: LionAccess, LionPath, LionLinguist, and LionView. LionAccess provides database connectivity by selecting, extracting, and routing content for localization based on user-defined rules. LionPath is the workflow automation management system. LionLinguist is the language management system that tracks previously translated content segments across multiple file formats and includes terminology

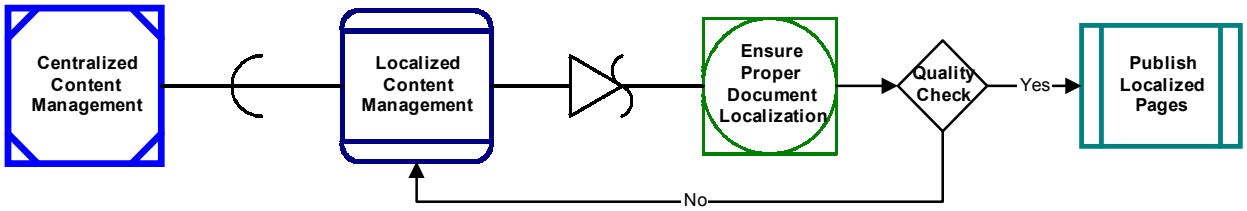

Figure 1: Content Management & Translation Workflow

management tools. LionView acts as a translation portal, capturing and managing project knowledge.

Which ever content management approach is used, it takes more than automated translation to create a multilingual Web site 0. As mentioned in Section 1, the textual content need to be translated properly, but other issues need to be addressed as well, such as graphics, logos, photos, local sell prices, payment methods, currency issues. The localization should also comply with the cultures and laws of the countries you are entering.

Although automatic translations are faster, they are often not accurate because of software can not handle idiom and some of the nuances of word usages among languages. The best approach is to combine automatic translation with the personal touch. To ease future evolution, the combination of content management and translation workflow is a better solution. This approach is illustrated in Figure .

2.2 Localization

Another important issue related to multilingual Web sites is content localization. This includes the language used for the site itself and tailored presentation styles. When a user visits a language- or country-specific version of a global corporation's Web site, the user typically expects that the local site contains content written in the user's native language. Unfortunately, this is not always the case.

As an example, consider the Web sites of Cisco Systems, the successful networking company. Cisco Systems' main Web site (www.cisco.com) has over 60 country-specific Web sites, in several different languages. Although Cisco manages the site's back-end system centrally, it relies on its many local offices around the world to maintain language-specific pages and ensure that documents are relevant within countries. Offices in each division also publish their own pages to support local customers, such as Events and Seminars, Trainings, News, and Contact Information, and so on.

Figure shows the main page of Cisco Israel; Figure shows the main page of Cisco Russia. As can be seen in the figures, the localized Web sites use the same image at the top of the first page: the navigation bar, guest bar and country-language specification. All unique items in the top navigation bar in different Web sites, such as Solutions, Productions, and Ordering, link to the same location. All these Web sites use a standard template for different language site development. This standard template includes the top navigation bar and left-side navigation bar.

Some sites, such as Cisco United Arab Emirates, just point to Cisco Middle East region, which is entirely in English. However, most localized versions of Cisco's Web sites use different native languages. For example, Cisco Israel is in Hebrew and Cisco Russia is in Russian. However, since all the multilingual Web sites use the same top navigation bars (a standard English-only image used in all the different localization Web sites), the top navigation bar is in English no matter what language the rest of the Web site is written. This "consistency" of the top navigation bar makes some of the site's content unavailable to non-English speakers.

Most users are interested in information related to a particular organization that is global in nature. However, the fact that they are accessing a localized site likely indicates an interest in local content as well (besides the desire to view the site in their native language). A good example of a provider of localized content is the Yahoo! News portal. Figure shows a French-Canadian version of Yahoo!. It has news that is international, but also news that is specific to Canada (e.g., the Toronto stock exchange's main indices) and content that is specific to Québec (e.g., news about Air Transat). Interestingly, Yahoo! also makes use of non-standard country codes. In this case, .cf for Canada-French.

3. STRUCTURE ISSUES

There is no doubt that content management plays an important role in user expectations of a multilingual Web site. From the content provider perspective, there are equally important structural issues that must be addressed.

Figure 2: Cisco Israel

Figure 3: Cisco Russia

Naming of the site influences the user experience, but it also influences the organization of the site itself.

3.1 Naming

One seemingly mundane but important structural consideration of a multilingual Web site is its name. For the different language versions of the sites, should they keep similar domain name pattern among these different language sites, or use a new domain name pattern according to the local culture and customs? Users' expectation of a domain name refers to what types of implicit information is conveyed by the domain name, such as language and locality.

There are several choices possible when it comes to domain names for multilingual Web sites. One option is to use the country code to connote the language, such as .fr for French. But this choice is not foolproof. For countries with multiple official languages, like Belgium (.be), the country code does not necessarily reflect the spoken or written language. Using the country code also implies that the site is registered and (possibly) physically located in that country, which also need not be the case.

Consider the automotive giant BMW Group, a international global company with a global presence, headquartered in Munich, Germany. The main BMW Web site (www.bmw.com) has entrances to other language Web sites. BMW has 45 different Web sites all over the world according to different geographical regions, such as Europe, America, Middle East, Asia, Africa, Australia and New Zealand. Most of these Web sites are written in regional languages.

BMW's multilingual Web sites' domain names reflect the languages in which the Web site is written and the locality (services) which the Web site is hosted. For example, www.bmw.fr is in French and located in France, www.bmw.de is in German and located in Germany, www.bmw.co.uk is in English and located in the U.K., www.bmwusa.com is in English and located in the U.S., and www.bmw.com.cn is in Chinese and located in China. However, some of BMW's Web sites domain name only reflect the locality of the Web site, but not the languages in which the Web site is wrttien. For example, BMW Thailand (www.bmw.co.th) is in English rather than in Thai, and BMW Saudi Arabia (www.bmw.com.sa) is in English.

In contrast to BMW, the popular online travel service Expedia has a very consistent international Web site naming convention. Besides the company's main Web site (www.expedia.com), it has three international sites: Expedia in Germany (www.expedia.de), Expedia in the U.K. (www.expedia.co.uk) and Expedia in Canada (www.expedia.ca). All four Web sites have similar layouts and the domain names reflect users' expectation both in language perspectives and locality perspectives. By reading the domain name, the user (correctly) surmises that www.expedia.de is in Deutsch and hosted in Germany, www.expedia.co.uk and www.expedia.ca are in English and hosted in UK and Canada, respectively. Besides the different languages, the sites also serve as a local travel

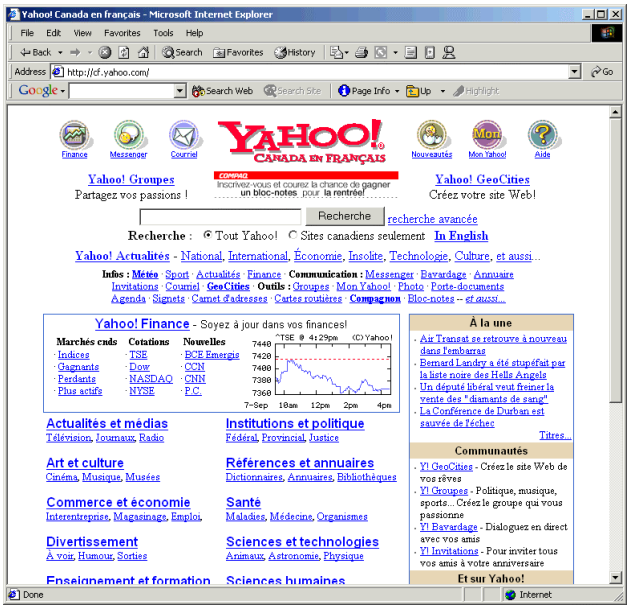

Figure 4: Yahoo! French Canadian

agent for Expedia: the local site only deals with sending tickets to their local people. For example, if you find and order an air ticket on the Expedia Germany site (www.expedia.de), they will only send your ticket to a German address.

Another naming convention choice is to use a "virtual" domain, such as fr.yahoo.com. In this case, the country code is used as an indicator to the user that the language of this site is French, even though the site itself may not be physically hosted in France. Sometimes nonstandard country code indicators are used to reflect the differences between countries and languages. For example, chinse.yahoo.com is in traditional Chinese and cn.yahoo.com is in Simplified Chinese.

If a virtual domain name approach is chosen, a potential issue arises related to name collision. Consider the Yahoo! "news" section. In English, the page is reachable at news.yahoo.com. But the French page is accessed at fr.dir.yahoo.com/Actualites_et_medias. This is clearly not consistent with the main English site, nor is it obvious to the user. The use of a virtual sub-domain further complicates matters.

Still a third choice is to embed the language selector as a subdirectory of the main site, such as www.chessclub.com/swedish shown in Figure . This choice offers flexibility on the part of the engineer, but it is not

Figure 5: Chess Club -- Swedish Version

idiomatic; most users would not know where to find the Swedish content on this site, and if the content is located below the root at /se (the country code) or /swedish (the language). In terms of consistency, this particular site is also lacking in the ability for its users to switch languages: all supported languages are visible from the main Web site (which is in English), but for language-specific sites only a subset of the links to other languages are available – and these subsets differ from language to language.

There is a relatively new development related to naming conventions: internationalized domain names using multilingual URLs [1]. There are currently 39 different writing systems used in more in 350 different languages and dialects. This means that sites containing non-English content could choose a domain name made up of non-ASCII characters, such as all Kanji for a Chinese site. While this approach may attract users with computers setup for typing in Kanji characters, it will deter most people from ever visiting these sites, if only because of the inability to type in the URL at a regular keyboard (without special software).

3.2 Organization

Web site organization is an engineering problem that requires careful attention in the context of multilingual content. A site's physical structure does not have to reflect the logical structure that the user experiences, although that sometimes makes sense for single-language sites. For multiple language sites, the physical structure quickly becomes too complicated to the mirror the logical structure.

Figure 6: SIGPC -- English Version

Figure 7: SIGPC -- Chinese Version

There are several options available to content developers and Web masters regarding organizing a Web site, both in terms of logical layout and physical structure. Such options range from relatively simple decisions related to stylistic page consistency to more complex considerations related to placement of navigational aids. For multilingual Web sites, it quickly becomes much more complicated to select the right option for each problem instance.

Depending on the type of content management system used, and the degree of overlap between the two versions of the site, the physical Web may in fact be heavily cross-linked, making for potentially serious maintenance problems later.

The naming convention issues discussed in Section 3.1 directly influence the user experience and hence indirectly the site structure. For example, if the site has adopted the country-code prefix method for distinguishing localized content, then the site's physical structure may also be similarly structured.

There is one practical consideration related to site structure and that is deployment cost. Registering multiple domain names is relatively easy, but it can become quite costly. Independent of the content management synchronization issues discussed above, the choice of real, virtual, or vanity domains – and their corresponding physical structure – all have an impact on long-term Web site evolution.

4. AN EXAMPLE: SIGPC

To illustrate some of the issues related to multilingual Web sites, consider our own experiences in editing and publishing an online magazine in both English (Figure) and Chinese (Figure). SIGPC is an online publication that explores the impacts of personal computing on computer science, information technology, and software engineering [1]. It is now in its fifth year of publication and is sent to several hundred subscribers. When a new article is released (usually once a month), a short summary of the text is sent by email to all SIGPC subscribers, as well as the URL for accessing the complete story and its corresponding audio version.

4.1 Content Management

As with most of today's Web sites, the SIGPC site was English-only. We decided to migrate it to a bilingual Web site to attract a larger audience, and to better serve some of our newer subscribers – many of whom are international graduate students. For us, this meant adding Chinese language support, which includes text in Kanji and streaming audio in Mandarin.

Adding a second language to a Web site can be challenging. Adding an Asian language such as Chinese is particularly challenging. Structural issues must be addressed to incorporate the new dual-path through the site. There are also issues specific to languages such as Chinese. These include creating content (which editor to use?) and viewing content (which browser supports this?).

Perhaps most importantly, one must endeavor to create two versions of the same material that are true to one another. This challenge is present whenever there are two or more languages used to represent the same thing. We have tried to make our translation as true as possible as to the Chinese philosophy of "xindaya": the translation should be accurate, understandable, and elegant. Given the technical nature of SIGPC, this is not always an easy goal to meet.

From our preliminary investigations, we have found that creating and viewing content for a Chinese Web site is quite challenging. Keeping two versions of the same material synchronized with one another is a classic software maintenance problem. We also found that there are many small details (and software bugs) that to deal with to produce a widely accessible bilingual Web site.

For example, something that must be considered while creating Web content, no matter which editing method is used, is how the user will see the resultant Chinese text. Our first approach was to create GIF files that captured the Kanji on the screen. The main advantage of this approach is that the user doesn't need to have any special software installed to see the text.

However, there are a few drawbacks to this approach. Every time the original text is changed, the GIF file must be regenerated. Since the GIF file will become unreadable if it is rescaled, it must be created at the proper resolution from the start. This means knowing the width and height of the container frame or page that will hold the GIF. In some instances, the result is a very wide page with a relatively narrow GIF file, which produces an unpleasant browsing experience for the user. From an accessibility perspective, using GIF files precludes the use of the Web site by sight-impaired users who rely on automated reading tools.

We decided instead to save the file produced by the Microsoft Input Method Editor (IME [1]) in native HTML format and use it directly in the Web page. While this solution works fine for the Internet Explorer browser, it doesn't always work for Netscape Navigator. When Word saves files in HTML format, it relies on XML and Cascading Style Sheets to reproduce the exact look of the original Word file. This is usually rendered properly by the latest version of Netscape Navigator, but it won't work for earlier versions. We decided to implicitly require users to have at least version 4.5 of Netscape Navigator to view the Chinese version of the Web site.

There is another niggling issue related to character sets and operating systems. By default, the character set used by Word when creating HTML files is "windows-1252." This is essentially equivalent to the default character set of "Western (ISO-8859-1)" used by Netscape Navigator. When the Chinese text is included in the SIGPC file, Internet Explorer will properly render the HTML as Kanji characters, but only when the user manually sets the encoding to "Simplified Chinese (GB2312)." If this manual change is not made, the result is unreadable text. This problem seems to have been resolved in Windows 2000.

4.2 Site Structure

The main SIGPC Web site is at www.sigpc.net and the Chinese version is at cn.sigpc.net. However, both of these sites are actually located underneath the srtilley.com domain. Physically, the top-level domains related to srtilley.com are siblings underneath the root directory of /usr/www/users/srtilley on the Web presence provider's Unix machine. The situation becomes slightly complicated because not all of SIGPC is available in Chinese, so the file structure under cn.sigpc.net is not a complete copy of its English counterpart.

Nevertheless, this approach is extensible. If a third language was to be added to SIGPC, say French under fr.sigpc.net, the same directory structure could be used. By design, the user would experience the same overall page look and feel, independent of the national language used. Only the content specific to each article is localized.

5. SUMMARY

Many of the issues encountered in adding Chinese language support to the SIGPC Web site would no doubt be viewed as relatively minor to those who had done a similar task before. There were many details and software idiosyncrasies that we had to address to make the final result acceptable. But, as with most things, it's only easy in hindsight.

The SIGPC experience illustrates just some of the challenges in managing content and structure in a multilingual Web site. There are documentation packages available to ease the reuse of translated content, and there are software engineering techniques that can be used to minimize the difficulty in evolving the site's organization. However, making proper use of these programs and techniques can be quite challenging. Nevertheless, the benefits of going global are well-worth the effort.

REFERENCES

[1] Berlitz GlobalNETet. Online at www.berlitzglobalnet.com.

[2] Bowne Global Solutions. Online at www.bowneglobal.com.

[3] Cisco Systems, Inc. Online at www.cisco.com.

[4] GlobalSight. *Accelerating Global eBusiness.* Online at www.globalsight.com.

[5] Gould, E.; Zakaria, N.; and Yusof, S. "Applying Culture to Web Site Design: A Comparison of Malaysian and US Web Sites." In *Proceedings of the 18th Annual International Conference on Systems Documentation* (SIGDOC 2000: Cambridge, MA; September, 2000), pp. 161-172. New York, NY: ACM Press, 2000.

[6] Idiom Technologies, Inc. *Software and Services for Enterprise Globalization.* Globalizing e-Business Online at www.idiominc.com.

[7] Kostya Vasilyev, K. "Multilingual Web Site Development." Online at www.microsoft.com/mind/0100/internat/internat.asp.

[8] Lionbridge Technologies, Inc. *Rapid Globalization Methodology. Globalization Platform*. Online at www.lionbridge.com.

[9] Locke, N. "The Localization Industry in Montréal." *Multilingual Computing & Technology*, 12(5):41-43, July/August 2001.

[10] McCollum, T. "Foreign Affairs." *The Industry Standard*, August 7, 2000, pp. 174-176.

[11] Microsoft Corp. Input Method Editor (IME). Online at www.microsoft.com/GLOBALDEV/wrguide/WRG_ime.asp.

[12] Network Solutions. "Internationalized Domain Names." Online at global.networksolutions.com.

[13] Paulson, L. "Translation Technology Tries to Hurdle the Language Barrier." *Computer* 34(9):12-15, September 2001.

[14] SIGPC. *Exploring the Impacts of Personal Computing*. Online at www.sigpc.net.

[15] The Apache Software Foundation. Online at www.apache.org.

[16] Tilley, S. (editor). *Proceedings of the 3rd International Workshop on Web Site Evolution* (WSE 2001: (WSE 2001: Florence, Italy; November 10, 2001). Los Alamitos, CA: IEEE Computer Society Press, 2001.

[17] Uniscape, Inc. *The Globalization Infrastructure for eBusiness*. Online at www.uniscape.com

信，达，雅