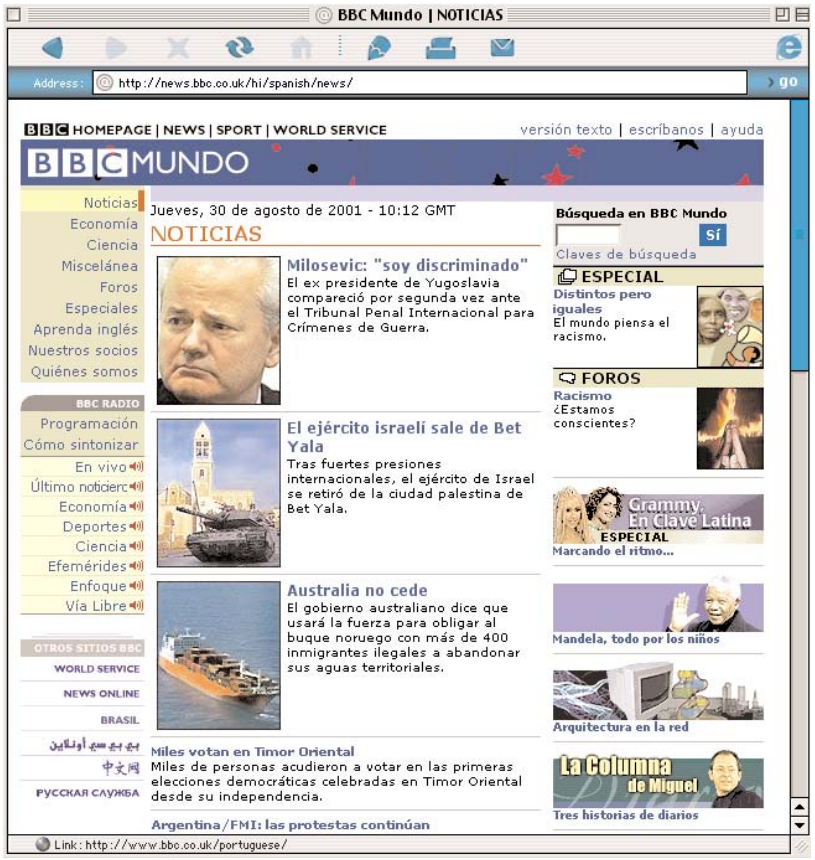
Managing Multilingual Web Sites

By Paul Ruffle

Translating your text files is the first hurdle, and it's a doozy. You also have to watch for text within your graphics and tags. And, of course, you have to update all the language versions whenever content is revised. Software can help with some of these problems.

Visit the Web sites of large multinational corporations and you will often see options for language preference or country of residence. If non-English readers see intelligible, localized material that respects their cultural sensibilities, then those companies may gain new customers. Fail that test and potential customers will be pointing their browsers at the local competition.

With around 80 percent of the Web's content in English but at least 40 percent of Web users not reading English, there is certainly a case for publishing in multiple languages. But just how difficult is it to generate and maintain effective, localized, multilingual Web sites? This article will analyze the challenges of such an endeavor, provide an overview of some of the available tools and offer examples of their use in deploying multilingual sites.



Spanish home page for BBC News. The BBC's World Service publishes online news in 43 languages and plans to progressively migrate all of them to Apple's WebObjects.

Translation on every level

Anyone who has not produced multilingual publications might be forgiven for thinking it an easy process. Just find someone to translate your text, cut and paste it into XPress or Dreamweaver, and there you go—what could be easier?

Well, for a start, just how good are your translators? Do they know your company and products well enough to localize your carefully crafted marketing copy? What measures do you have in place to assess the quality and accuracy of their translation? Do you have a local manager who can ensure that you do not unwittingly alienate customers with badly translated text or some other cultural *faux pas*?

For example, just because the managing director's secretary in your Paris office is French, is no guarantee that he is competent to translate or edit the French section of your Web site. However, try telling him this and you will begin to think that Esperanto was not such a bad idea after all. As veteran multilingual Production Director Tim Cowan, of Bell Design (www.belldesign.co.uk) in London, put it, "A little knowledge on the part of local-office staff is very dangerous when it comes to checking and approving foreign versions of a publication." Add to this the politics—head-office control versus local-office autonomy—and you are between a rock and a hard place. As one frustrated colleague once said to me, "Let's just teach everyone in the world to read English; it's got to be easier!"

Print is harder. We shall digress briefly to note that printed brochures and manuals pose additional problems for international companies. Cost considerations often dictate the use of common color, with language changes accommodated on the black plate only. This means that your layout has to take into account that translated copy can grow by 20–30 percent, or even get shorter in some languages. Tight designs require constant browbeating of translators to keep the word count down, and aggressive editing—assuming you can find editors with the appropriate language skills. Right-to-left languages such as Hebrew and Arabic require you to mirror your page layouts, and as to the likes of Chinese or Japanese . . . need I say more?

If that was not enough, you then have the problem of character sets, code pages, font layouts, etc. The standard fonts on your Mac or PC will cope with most West European languages, but go eastward and you are in uncharted glyphic territory. Copy and paste some text from a Hungarian Word document into XPress and you are likely to end up with gobbledegook, unless you first map the hex codes used by Word to those of the font you are using in XPress. If your source data is hosted on legacy systems, *e.g.*, an IBM mainframe using EBCDIC code pages, you will have still more difficulties. Unicode was supposed to be the answer to all this but, unfortunately, many applications do not yet support it.

Delivering multilingual content to the Web

Viewing content via your favorite browser actually solves some of the problems mentioned above. Assuming your Web-page layouts are flexible as to copy depth, increases in word count will not be a problem. However, languages such as German tend to join several short words together to make really long words. So avoid narrow column widths that would be forced wider by these words; remember that browsers do not have H&J facilities.

The really good thing about Web browsers is that you can use the content-type metatag's charset parameter to specify the character encoding that your Web pages will use, both for displaying text and receiving input from the user. This means that you do not have to worry about the OS, keyboard or fonts being used by visitors to your site. All the special characters are defined in HTML, so you only have to map these known codes to whatever coding you are using in your back-end systems.

This is so liberating that many vendors of content management systems are abandoning proprietary client APIs and switching to using browser-based clients for accessing and updating their databases. Regardless of where in the world your local managers or translators sit, as long as they have an IP connection, they can log into your central repository and update or approve the language assets they are responsible for.

When it comes to using graphics files for logos, display headings and navigation icons, things are less straightforward. English seems to lend itself to short abbreviations and idiomatic expressions, but translate those short phrases in your menu bar and you end up with whole sentences. Patient explanation and negotiation with your translator and local office will get you something more succinct, but be prepared to compromise.

Better still, get your designer to take these issues into account when creating the look and feel of your site. You may have to rely on clever use of HTML text, color and tables, rather than generating copious quantities of GIF files. Use Cascading Style Sheets (CSS) and specify type sizes in pixels, not points. Remember also that there will be text that needs to be translated in img alt tags and JavaScript mouse-over statements.

Translation Tips

Whether you are writing for machine or human translation:

- Create simple and straightforward copy.
- Write clear, focused, simple sentences.
- Limit sentences to 15–20 words.
- Do not omit necessary grammatical elements.
- Avoid idioms and slang.
- Use correct punctuation and spelling.
- Do not omit necessary words such as relative pronouns, prepositions and parts of verbs.
- Consistently use the same word or phrase for the same object or actions.
- Allow text to wrap naturally; hard returns can imply the end of a sentence.
- If your text is highly colorful and expressive, use human translation.
- Create glossaries for technical jargon and specially defined terms.

An extra dimension. Managing multilingual content can be likened to going from a two-dimensional space to 3D. For example, a database of product information often includes attributes such as description, color, size and price. In a multilingual system, these attributes have alternate properties for the product description and look-up-table pointers for descriptors such as color or facilities. Even numeric attributes such as size and price can have multiple properties because of different metrics or currencies.

Language versioning can also be based on a dependence hierarchy with (say) English being the master language and French, German and Danish being dependent on English, and then Swedish and Norwegian being dependent on Danish. This dependency can drive the updating process, with appropriate changes being reflected down the hierarchy.

All of this multilingual content is just data, but when it comes to effective management and updating of these assets, an appropriate metaphor for the sort of interdependencies described above must be implemented in any client interface. If your content is large amounts of unstructured text, such as legal references or technical documentation, you will need some form of linked meta-tagging to allow your visitors to find what they are looking for in a particular language.

¡Hola! Yahoo's Spanish online news service for the Mexican market.

Human vs. machine translation

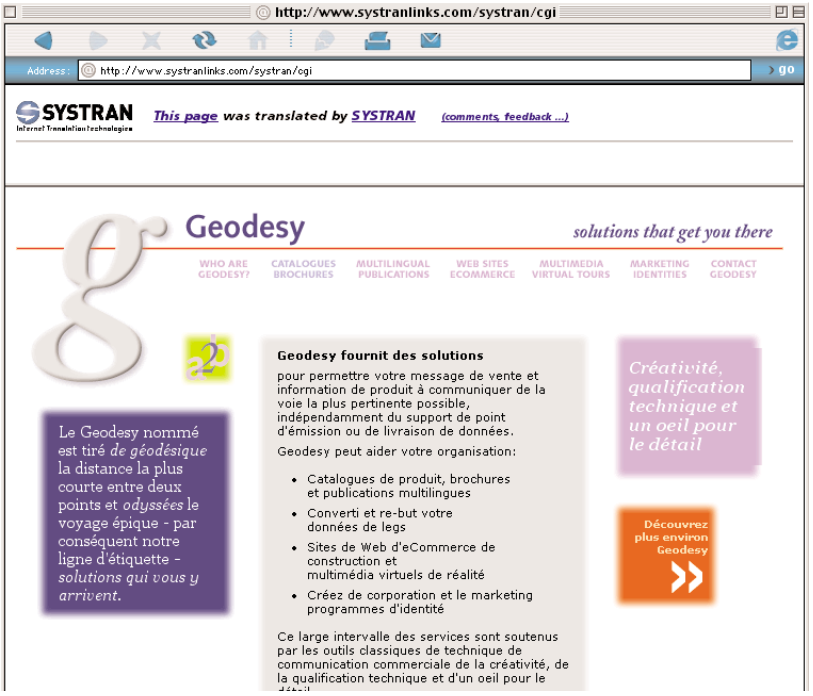
Ask any experienced translator about machine translation (MT) and you will probably be told that it is rubbish. However, the existence of several successful MT vendors is testimony that it does have a place in delivering multilingual content. Today's advanced MT engines use sophisticated language models to deliver translations that provide the gist of the original document, though the translation will not be perfectly accurate or idiomatic.

MT is often used when there is no preexisting translation available (it's better than nothing) or for company intranets where there is less concern about branding and image awareness. Content of a highly transitory nature is also a candidate for MT.

For critical texts such as marketing copy, only human translation can deliver the accuracy and localization needed. But if you have huge amounts of material and a limited budget, MT may be the only affordable route open to you.

A practical alternative may be a combination of human and machine translation. MT can be used for the first draft, followed by human editing to correct terminology, grammar, punctuation and idioms as needed. The quality of translation can be improved further by making sure that source text is well written *(see the 'Translation Tips' sidebar)*.

Partial success. Geodesy's home page has been machine-translated into French by Systran's free online service. Note that GIF text images remain in English.



Yahoo's not impressed. Even so, MT is not the solution for everyone with huge amounts of content. "We have looked into machine translation," says Arthine Cossey van Duyne, senior international producer at Yahoo, "but have never found a suitable engine that could deliver the quality and localization that we require. For example, when localizing our Web applications such as Yahoo Games and Yahoo Auctions, we had to find ways of expressing terms which were totally new concepts for some languages and cultures."

Yahoo has developed its own proprietary software tools to build more than 300 localized versions of its products and services. According to van Duyne, "We don't have a need for machine translation to build content-heavy products such as Yahoo News and Yahoo Sports, because our local teams aggregate the best content from market-specific partners."

Back in 1996, Yahoo had just five international Web sites. This number has now grown to 24 sites in Europe, Asia and Latin America encompassing 12 different languages including regional variations for Chinese and Spanish.

The mantra for generating quality translation for local markets is familiar: Find good translators that are natives of, and live in, the country in question; hire good linguistic editors who know your business inside out; and work with your local people at the coal face.

Machine translators in use

To get a feel for how machine translation works, I visited the Web site of France-based Systran (www.systransoft.com) and tried out its free online service on my own company's home page. Well, the results certainly looked French, but a closer examination revealed the limitations of MT. The translation was literal, so that "repurpose your legacy data" became "re-but vos données de legs", which does not quite hit the mark. However, we should point out that the use of customized dictionaries (which a public service obviously cannot offer) can considerably improve the accuracy of MT.

Looking further afield, IBM's WebSphere Translation Server has recently been released from the

R&D labs and, according to Brian Garr, program manager for advanced technologies, is being used for nine languages on IBM's own corporate site and also for IBM's software-support pages. Garr also highlights another important issue. "You have to decide if your MT engine is going to translate entire Web pages or data objects that are subsequently built into pages on the fly."

WebSphere Translation Server, which is also being piloted by DeutschBank, consists of MT engines for translating text from one language into another, dictionary tools that allow specific terms or phrases to be interpreted correctly according to context, a translation server that makes the engines available for Web page translation, and HTTP server support.

Gist-in-Time. Interestingly, both IBM and Systran are strategic partners with Montreal, Canada-based Alis Technologies (www.alis.com). Alis provides an integrated one-stop-shop consultancy service and automatic language translation. The latter is based on Gist-in-Time, which is a sort of MT operating system that allows various language-pair translation engines from different suppliers to be used, depending on the type of content to be translated.

Using suitable test material, Alis determines the best engine to use and develops customized dictionaries to further improve the accuracy of translation. Master text is usually in English, but language pivots also exist for French and German. Pre-editing tools are also available for establishing elements of text that should not be translated, *e.g.,* proper names, dates, geographic references, etc. Alis also provides training so that writers can produce source material that is better-suited to MT.

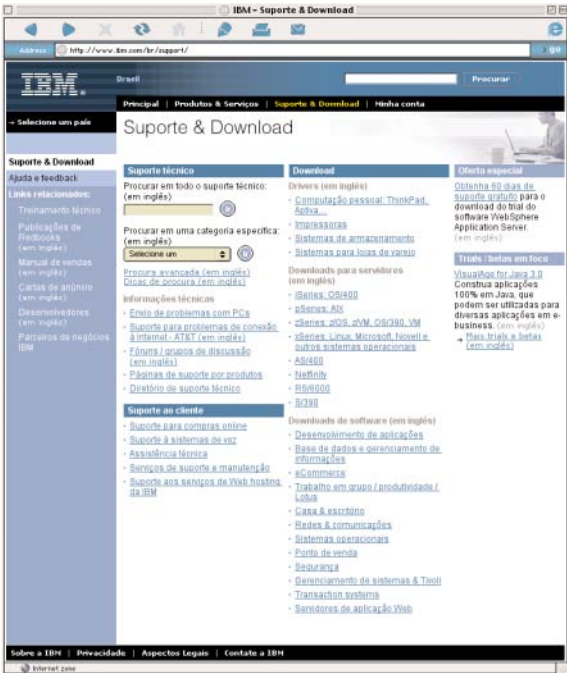
The Gist-in-Time server can be integrated with existing Web- and application-server technology, depending on where in the process the machine translation takes place, *i.e.,* before or after HTML page generation. For improved efficiency, static content in each supported language can be cached.

"Don't overestimate the quality of MT," says Thierry Gauthier, senior director of marketing and communications at Alis. "Test and find out where it works well and where it does not." However, many companies have colossal amounts of content and MT is often the only cost- and time-effective way of communicating to the variety of markets that they may want to reach.

Content-management tools

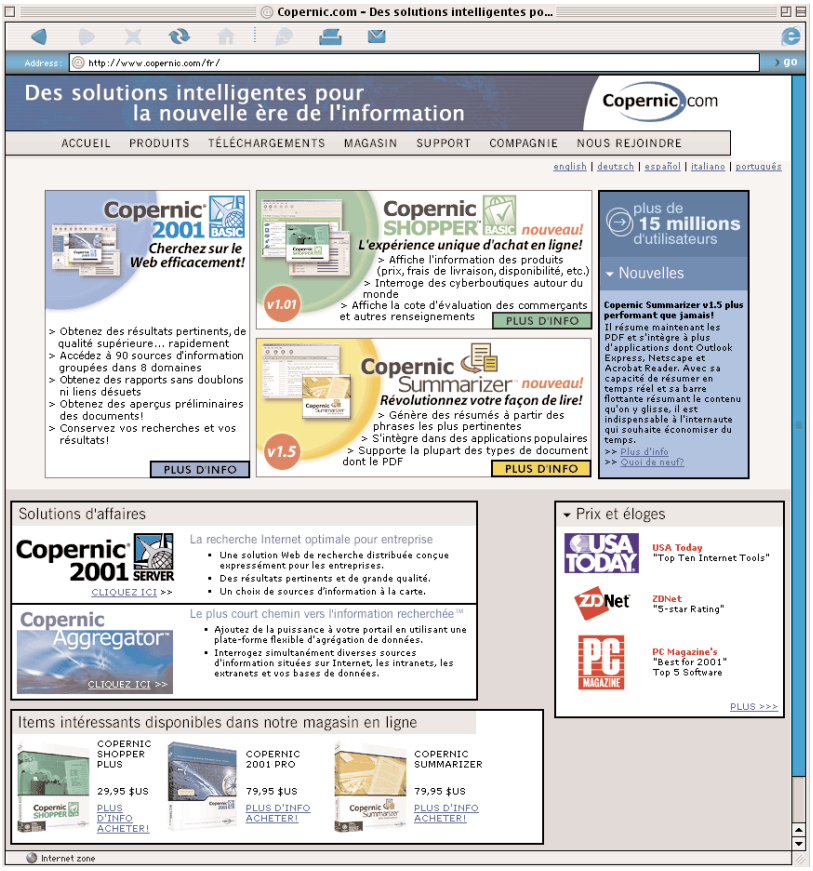
If you require enterprise-level multilingual content management (and have the budget to go with it) then big players like Vignette might meet your needs. Timeshare-exchange giant RCI uses StoryServer for its membership-based Web site, serving around 25 million pages per month. Colin Wynd, VP for Internet strategy, considers Vignette "a great choice" compared to other content-management systems. "Creating six language versions of RCI.com is generating a lot of data," says Wynd, "and StoryServer is dealing with this additional content in a structured way."

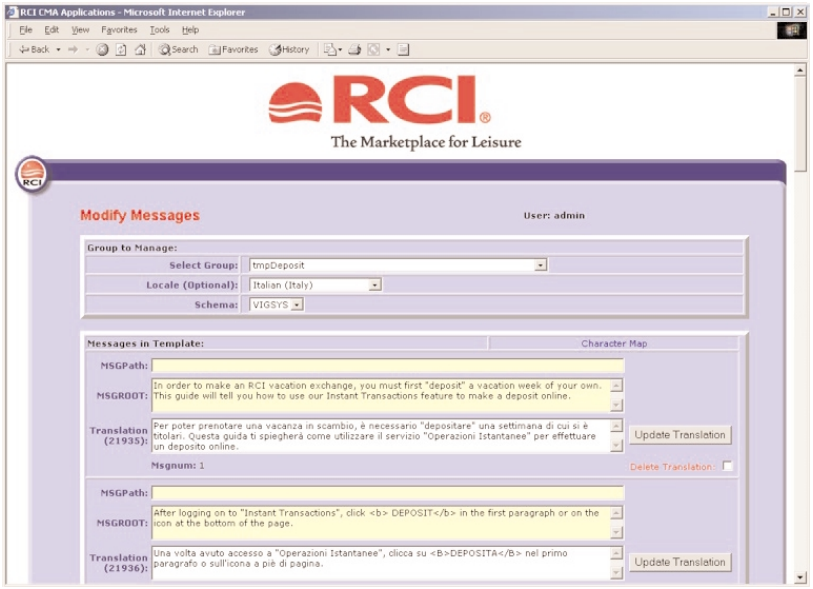
According to Louise Wojnicki, RCI's European Internet-development manager, translated content is



WebSphere Translation Server. IBM uses its translation engine on its own Brazilian Portuguese software-support pages.

OS for MT. Alis Technologies' Gist-in-Time is used at Copernic.com, a localized search-engine portal that works in several languages, including this example in French.





RCI translation. Vignette's Content-Management Application browser interface for RCI allows editors to enter the Italian equivalent of each page component into a text-entry box underneath a display of the original English text.

dealt with by the AST Message Catalog; page elements are broken down into logical pieces that may be fully translated chunks of text or lookup-table attributes. These elements can be created and updated remotely through the Content Management Application browser interface. The system also allows for dealing with the extra length that translated copy takes by creating suitable page templates.

Vignette's Integrated Workflow Management system automatically sends e-mail notices to remote editors when a change is made in the master texts (usually English). There is a review-and-approval process before new or changed content goes live. Other features include delivery of personalized pages and use of alternate images because of cultural issues (less bare flesh, for example).

Lockside's GlobalEdit system for Sun. The German text of one section of a document is entered alongside a display of the original English text.

Process management. Success in developing and deploying a global Web presence also depends on ensuring that changes to content and infrastructure are separate processes. EGrail's Enterprise Content Server puts the emphasis on IT productivity; a browser-based interface localized into each native language lets Webmasters track down problems on any site, in any language. Access privileges allow remote users to accomplish their tasks in their specific language and country, but only designated users have the ability to change global aspects of the site.

Enterprise Content Server emphasizes reusability, implementing templates that inherit characteristics and formats from other templates, thereby minimizing changes that need to be made to individual pages. For example, navigation code need be written only once, then all instances on the site inherit that code.

Since content is the main element that changes from language to language, these templates insulate content providers from the technical details of each page. However, remote content providers and Web programmers do have the flexibility to customize the look, feel and content to match the local language and customs without compromising the global branding of the site.

Content publishing tools

Bridging the gap between enterprise systems and catalog publishing is GlobalEdit from U.K.-based Lockside Software (www.lockside.co.uk). Described as "just-in-time publishing," by Managing Director Peter Ritchie, GlobalEdit gives remote users the ability to edit copy in their own language while viewing the base language side by side.

The system can automatically generate HTML, XML, RTF, PDF or XPress pages in their final form, ready for deployment. A publishing center can be established where remote users can edit content while the designers determine the look and layout of the final pages using GlobalEdit's document-definition tools. Character-conversion maps handle the internal translation of foreign characters for the appropriate output medium.

U.K. consultant Beechwood (www.beechwood.uk.com) helped in the specification and development of GlobalEdit to create an online publishing system for Sun Microsystems' EMEA Education and Training services. Sun produces 21 different versions of its 120-page course catalog. Although the master text is in English, language versions vary due to local market demands.

Eleven languages (including East European) required support, which posed a problem due to the mixture of PC, Mac and Unix clients in use by local agents. Of the companies contacted by Beechwood, only Lockside was prepared to meet Sun's needs.

The resultant system enables Sun's 21 marketing managers and the support staff in each market to edit their own content, which they alone can sign off. Their edits automatically update a 4D database (on a Mac server at Beechwood) that in turn automatically updates the final styled output. Sun's editorial staff works in conjunction with Italian translation agency Logos (www.logos.it), which uses a Sun server and Oracle database to manage the creation of initial translations. Interestingly, Chris Caffyn, Beechwood's managing director, points out that "Europe may be a single market, but don't forget that it is also a multicultural and multilingual market."

Apple success story. Apple's WebObjects is a good example of how you can deploy various types of multilingual site with an inexpensive, general-purpose application toolset. Used for the AppleStore and other parts of Apple's corporate site, it delivers text that has been translated into multiple languages by external agencies and uses lookup tables for other types of data, such as product codes and prices.

Four years ago, the BBC (www.bbc.co.uk/worldservice) started construction of its online news service, which was built initially with support from Apple's consulting group. According to software-development manager Kevin Hinde, "WebObjects was chosen as a robust and scalable technology to automate site production, and allow the BBC to invest in journalists rather than developers. Its frameworks provide an open and flexible toolset with default behaviors that we can customize as required."

Content structure is defined in an Oracle database, which has grown more complex as more news stories have accreted. As well as delivering output to the Web, WAP, PDAs and e-mail, news feeds are also routed to interactive television and to portals like Moreover.com. WebObjects is now delivering the BBC World Service in English, Spanish, Russian, Chinese and Arabic. Altogether, the BBC publishes online news in 43 languages and plans to progressively migrate all of them to WebObjects.

However, creating multilingual sites was not without its difficulties. The team had to work with different and subtly incompatible methods of character-set encoding in the Visual Basic client, Oracle database and WebObjects publishing system. "We had to give our foreign-language contributors a different version of Windows NT that used the appropriate HTML charset coding for each language," says Hinde. "This text is posted to the database such that no conversions are made when it is fetched for page creation. WebObjects' frameworks allowed us to achieve this easily by extending the basic frameworks with additional classes and methods."

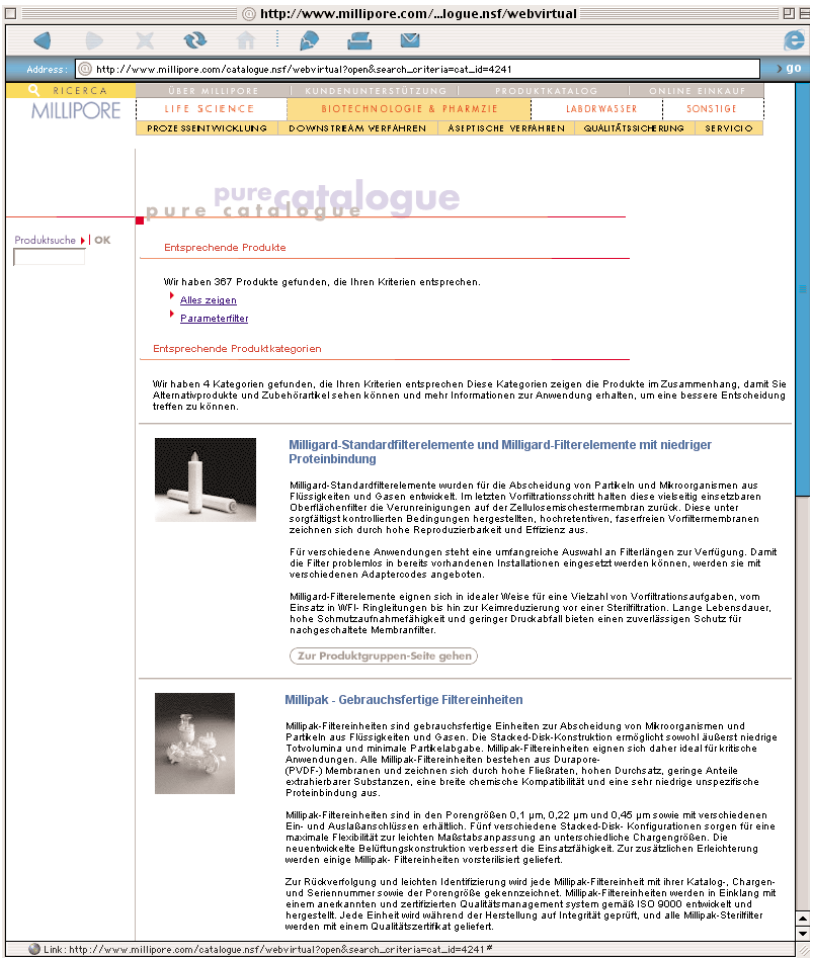
Apple claims that the latest version of WebObjects is 100-percent Java-compliant, so standardized character-set handling should require even less integration work. It can be developed under OS X or Windows 2000 and, being Java-based, is highly portable when it comes to deployment. The price of a development and deployment license has been reduced to just £479 ($700) per server, regardless of size or number of processors, so it could be described as a bargain, especially if you have Java skills in-house.

Online catalog tools

Pindar's (www.pindarsystems.com) modular Catalog Management System (CMS) includes Active Catalog, its e-commerce platform. CMS can also be used to drive other vendors' e-commerce solutions. There are two ways of updating language content: Data Manager, an application running locally as a Windows or Mac client, and Online Translator, running within a Web browser on any client with an IP connection. Data Manager also incorporates an offline translation function, which is an RTF-based export-edit-import process.

Data Manager and Online Translator have similar functionality. When changes are made in a master language, the system automatically flags derivative versions as requiring attention. It then creates to-do lists

Stepping forward. Millipore's online bioscience products catalog is built with Stibo's Enterprise Publishing (STEP) system. One feature of the site is the ability to switch among English, French, Italian, German or Spanish within any individual page.



that alert remote product managers of the need to update their translations. When editing an object attribute such as the product description, the user is presented with a side-by-side view of the master copy and the translated text to be updated.

The offline translation function allows text to be exported in RTF format and then edited within Word without disturbing the data tags necessary for successful import back into the database. This is well suited to the large volumes of text entry involved when populating a new data set. It is also used as a means of soft proofing to distant product managers and as a way around the problems of unreliable data-links to client offices in the more remote parts of Eastern Europe or Latin America.

Data objects are stored and updated on an Oracle 8i server. This, in turn, updates the live Web server, either on an incremental or "re-import everything" basis. A staging server is used for Web proofing and approvals prior to updating the Web server. The success of this is dependant, as always, on a well structured and documented set of proofing, control and checking processes.

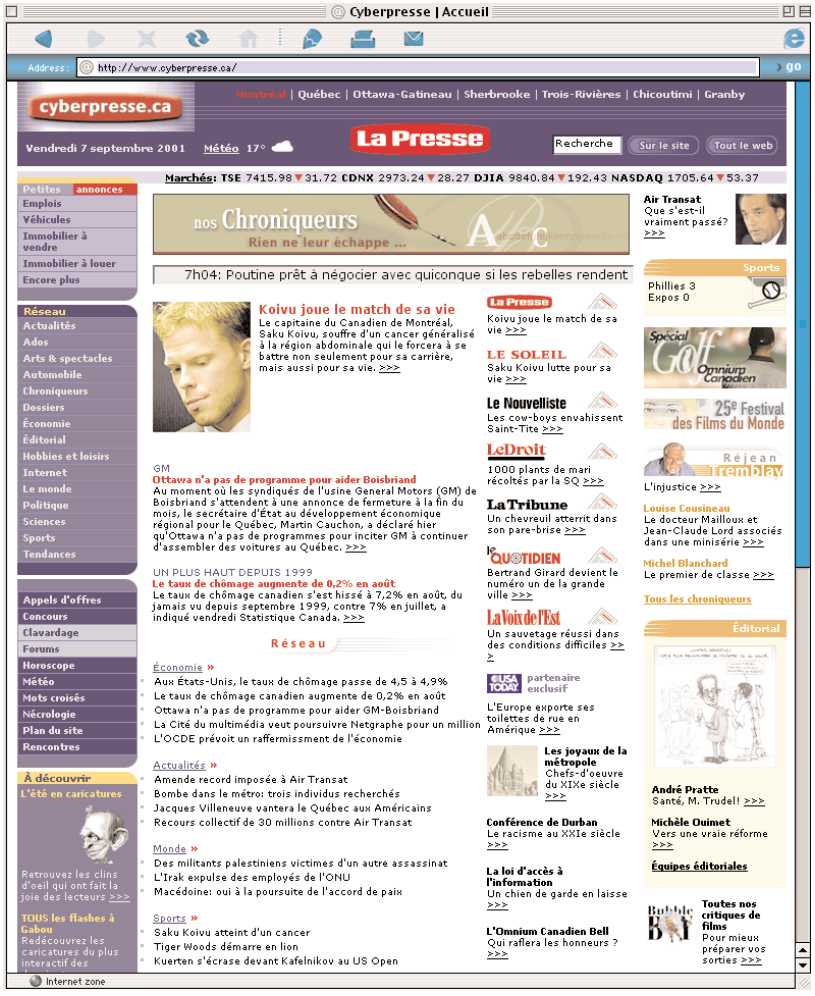
Languages as layers. Until recently, The Stibo Group provided only an outsourcing service for its clients. (This type of solution now goes under the trendy acronym of ASP.) Interestingly, Stibo, having completely remodeled its operation, now sells its system and support services for installation at the client's premises. (Well, who cares where the server sits anyway?)

Stibo's auto-pagination system, STEP, which can now use Framemaker as the base engine for page building (as well as a proprietary CCI system), provides interfaces into XPress and Framemaker that facilitate bidirectional updates to an underlying database. STEP can also generate XML content for third-party providers to deploy on eMarket sites, Ariba.com being one example. Web sites can also be generated directly with eSTEP, an integral part of the STEP system that delivers data dynamically to a linked Web server without the need for an XML dump.

There are three areas of the STEP system: Product Information Manager, which deals with factual attributes like color or size; Publications Manager, which deals with free-form text descriptions; and Media Asset Manager, which deals with images and other media-rich content. Support for Unicode is included.

An inheritance structure is based on the Product Information Manager attributes, and alternate languages exist as layers in the database architecture. Updates are made through a browser-based interface, and the Module Builder plug-in provides a WYSIWYG view of a publication after copy has been reviewed and approved.

According to Tom Anderson, marketing VP for Millipore Corporation, "Stibo's STEP relational-database publishing system, with its very granular and highly structured content-management foundation, has allowed Millipore to successfully publish its product offerings to its global audience through one Web site and in many languages, enabling Millipore to present an international, unified message worldwide."



Unstructured publishing. *La Presse*, the oldest and largest French newspaper in North America uses Stellent's content-management system to create its online news service.

Unstructured content tools

Stellent's Content Server (formerly Intranet Solutions' Xpedio) and Publisher is a Java-based system that can store and manage unstructured business content. The core technology converts contributor file formats (typically Word) to PDF, HTML, XML or WML. Original and converted versions of files are stored in an SQL-compliant database for subsequent reuse or repurposing. The general structure of documents may be used to generate appropriate metatags that can be used for subsequent indexing. Alternately, the structure can be explicitly defined through the use of preformatted templates.

From a content creator's viewpoint, a Stellent installation usually involves nothing more than using the correct templates for the various types of documents that are created, and then saving to appropriate hot folders. The CMS server uses the context of these folders to categorize the document and create appro-

priate metatags. It is also possible to use an explicit interface into the CMS.

Verity's Search Engine is used to do the subsequent indexing to facilitate full-text searching. Each collection of documents in a specific language is usually indexed and stored on a separate server that includes the Verity search engine. With minimal configuration, a site visitor's language preference can be identified and the search requests routed to the appropriate server. This may be done in the background so the user sees just one multilingual collection of unstructured data that they can do natural-language searches on.

The typical workflow process involves creation of master documents in English. Once reviewed and approved, a new master triggers either machine translation, manual translation or a combination of the two. MT can be integrated so that equivalent metatags are generated in the translated text. Manual translation is kick-started by automated e-mail alerts to remote offices, who then upload texts (via secure login and a Web-browser interface) to the central repository. This alert-and-access method is also used for subsequent updates and approvals. Full versioning is available with the ability to roll back to earlier editions.

Conclusion

In Britain, we have an old saying, "horses for courses", which means that you can only run the race if your horse is suited to the course. But if media-neutral publishing is today's Holy Grail, we have to demand a horse that can run on any track. Multilingual solutions that work for the Web must also deliver to other media, and especially to print, that messy (but occasionally profitable) business of smearing ink onto paper.

For the sake of clarity, I have grouped the different products described above under a number of generic headings. However on reflection, I am not at all convinced that this categorizing is either fair or appropriate. Some products do give evidence of being—or at least becoming—media neutral and, in addition, of being capable of use across a wide range of applications. Equally, though, there will always be a place for solutions that fulfill very specific multilingual requirements.

For any company wanting to sell its wares across national boundaries, multilingual publishing to the Web is a fact of global life. More and more non-English speakers are coming to the Web and demanding content they can understand and interact with. Ignore that fact at your peril. *Vive la difference!* **TSR**

About the Author

Paul Ruffle has produced more multilingual pages than he cares to remember, and is the managing director of Geodesy Limited, a provider of consultancy and production services for print and online delivery. He also lectures in graphic design at his local college. He can be reached at paulruffle@geodesy.co.uk.