[*Translating and the Computer 21*. Proceedings… 10-11 November 1999 (London: Aslib)]

Multilingual Component Management: Trends and Implications for Translation

Stuart Sklair, Senior Consultant, Multilingual Technology Ltd.

Introduction

This paper looks at how Internet and web publishing technologies are changing the way we think about publishing documents and information. It describes how what was previously the realm of technical documentation specialists in industries such as automotive and aerospace will become a pervasive part of creating information, with major implications for those providing translation services.

The WWW is not all hype

There is much hype about the Internet and the World Wide Web. It has been said that companies not doing business on the web in 5 years will not be doing business at all. The estimates for e-business carried out over the web from Gartner, Forrester, IDC, Nua and the like are in numbers that can scarcely be comprehended. Nua (www.nua.ie) claims that the global online population in 1999 is estimated to be 171 million. A vendor of website content management software estimates that the global e-commerce market will reach $1.2 trillion by 2001. Of course, these are predictions made by those with a vested interest in generating the hype. However, whatever the actual overall figures, the web and web technology presents some interesting challenges and opportunities for the language industry.

The uptake of Internet technology makes it the fastest technology growth area ever. It is estimated that it took 50 years for the telephone to reach 50 million people, whereas it took only 5 years for the Internet to do the same. This makes the web truly global, available even, for example, in the mountain-top village in Nepal from where my brother sent me an e-postcard. Despite the current English focus of the web, around 50% of web users are non-native speakers of English. Furthermore, Nua estimates that 35% of all web users do not understand English.

The web marketeers have recognised that "content is king". In order to get people to visit a site and then keep them coming back, the website has to provide more than just cool graphics or a catalogue of items to sell. It has to offer information relevant and even personalised to the end-user. For example, MyNetscape syndicates content from a variety of different sources. You can customise your default start page on the web to contain relevant aspects of the syndicated material. You can customise it to manage your personal portfolio of stocks and shares, to check your horoscope, or to view weather reports from the locations of your next business trip, and read news items on particular topics you are interested in. If you add in the dimension of language, this means that websites should also allow the user to customise the information to their own language, be it French, German, Japanese, or Chinese. This is reinforced by research carried out by Forrester, which suggests that business users on the web are three times more likely to purchase when addressed in their native language.

Another means to ensure repeat visits is to regularly update the content of the website. If it appears in multiple languages, then this should be rippled to all language versions of a site. The fact that anyone anywhere can access a website in any language means that the nirvana of fast turnaround of simultaneous launches in multiple languages, so often the topic in software localisation circles, is even more important when it comes to the web. A company's claim to a global image could take a tumble if non-English speakers find a message: "Sorry that page is not yet available in <yourlanguage>" (with the message not even localised).

As businesses start to use the web for growth, to increase the market for their products, reduce the cost of sales and customer support, reduce inventory, gather information on their customer base, and provide a customised one-to-one service, the website becomes an intrinsic part of their product offering. The content of the website and the language used then

becomes a high-profile product issue, making the web a compelling medium for those providing language services.

Before looking at how websites are evolving to build business value and the implications for translation, here is a brief look at the evolution of the technology behind website publishing. The word "publish" is appropriate because in the same sense as you publish a Word or FrameMaker document to paper, you publish web documents on the Internet.

A whirlwind introduction to SGML, HTML and XML

HTML (HyperText Markup Language) has its roots in SGML (Standard Generalised Markup Language), so that would seem a good place to start. SGML is a method to define and design the structure of documents, as opposed to how the documents are presented. For example, SGML might define a memo like this (the line numbers are there to help describe what is going on and are not part of the SGML):

```
[1]   <memo>
[2]   <date>21/ll/99</date>
[3]   <title>Translating and the Computer</title>
[4]   <from>Nicole Adamides</from>
[5]   <to>Conference Speakers</to>
[6]   <body>Make sure you use plenty of examples</body>
[7]   </memo>
```

fig 1. A memo encoded in SGML

This SGML fragment shows how such a document would be marked up. Lines [1] and [7] contain tags identifying the start and end of the memo. The memo contains five "elements", <date>, <title>, <from>, <to>, and a <body> element. Each element contains text. This says nothing about what the memo might look like. That is down to a style definition markup language that is too complex to go into here. Using the style definition language you can define the memo to look however you might want it to look, for publishing on paper of various sizes (e.g. Legal for US, A4 for Europe), or for publishing electronically. Another advantage of SGML is that it is independent of the mechanism that created it. As long as you have a tool that can recognise SGML, you can read any document encoded in it. Quite an advantage when you think about the compatibility issues between various versions even of the same software vendor's own applications!

However, for a document to be "valid" SGML it must follow a defined tagging structure. This is defined in a DTD (Document Type Definition). For example, the DTD for the memo above might be defined as having the following structure (for simplicity this example does not use the actual DTD syntax)

```
[1]  <memo>
[2]  | ----- <date>
[3]  |------------<title>
[4]  |-------------  <from>
[5]  |--------------------  <to>+
[6]  |---------------------------  <body>
```

fig 2. The structure of a memo

It looks much like the memo itself without the content. The memo is an instance of the document type whereas the DTD contains the rules for the document. For example, the + against the <to> field indicates that there can be one or more <to> fields in the memo.

Various industry standard DTD's have been developed, most notably in the Aerospace and Automotive industries. This allows the different subcontractors, for example, to develop documentation for their section of the product. They may use whatever SGML authoring tool they prefer as long as they follow the DTD. This enables whoever pulls all the documentation together from the various subcontractors to do so with little difficulty, and publish in whichever format is required.

One such DTD is HTML. - the Hypertext Markup Language DTD used for the web. HTML looks like this.

```
[1]  <html>
[2]  <title>Memo</title>
[3]  <body>
[4]  <hl>Translating and the Computer</hl>
[5]  <h2>from: Nicole Adamides</h2>
[6]  <h2>to: Conference Speakers</h2>
[7]  <p>Make sure you use plenty of examples/p>
[8]  </body>
[9]  </html>
```

fig 3. Some HTML

The start and end markers indicate that this is an HTML document. It contains a <title>, followed by the <body> of the HTML. The <body> contains <h1> and <h2> elements followed by a <p> element. Again, this says nothing about what the document looks like. In fact, the browser you use understands the HTML and presents it according to style rules contained within it. So that, for example, the <title> tag appears in the title bar of the browser window,

and <h1> appears as **Translating and the Computer** in Times 24pt Bold. The DTD is flexible in that the elements can appear in virtually any order. But as it stands it does not allow you to be adventurous in the way the text is presented. It was adequate for academics to post papers on the Internet, but to do anything more than simple text layout and plain graphics, it did not go far enough.

HTML, therefore, now has various features to make up for these inadequacies. For example, there is a <font> tag, allowing the author to specify a different font face from the default. There is also a
 tag to create line breaks, and a <b> tag for making text appear in **bold** type. The "purist" SGML approach of presenting structure and not format has, therefore, been eroded in the HTML DTD. There have been further extensions to HTML with every competing release of Microsoft Internet Explorer and Netscape Navigator, making it more and more difficult to design web pages that display correctly with both types of browser and in all commonly used versions of the same browser.

There are also content-related features in SGML that are lost in the HTML DTD, such as the ability to define the semantics of an element. The <from> tag in the memo DTD presented earlier indicates that the text following the tag represents the sender of the memo. In the HTML DTD this can be represented only as a formatting tag such as <h2>; in a book DTD a <chapter> tag would have a distinct structural meaning. However, it can be represented only as a <h1> or <h2> tag in HTML. For this reason, the <meta> tag was added to enable meta-information to describe the content of a page to be included in the HTML without being displayed. Search engines, for example, make use of the information contained in <meta> tags to improve search results.

With constant revisions, additions, and competing versions being released, HTML has become messy and inadequate for tne new applications for which the web is now being used. Which is where XML (extensible Markup Language) comes in. XML could be described as SGML Light - SGML without some of the more complex options. The presentation is handled by XSL (extensible Style Language). In an XML-aware browser, the XML and the XSL are combined to present the text on screen. With XML, web documents become "intelligent". For example, in an on-line shopping application, one of the many computers for sale might be coded as follows:

```
[1]<item>Computer</item>
[2]        <make>Valley</make>
[3]        <specification>P200</specification>
[4]         <price>£299</price>
```

fig 4. XML fragment

What the XML enables a user to do is, for example, tailor the view of all the information in the shop to show only those <item>s with a <price> between £200 and £500 in a way which would not be possible with HTML.

XML looks very much like a representation of a database. Not surprisingly perhaps, one advantage of XML is that data can be stored in databases, and XML documents published from the database on demand rather than all being hard-coded in HTML pages. This makes updating very much easier.

Imagine that an on-line Computer warehouse wanted to reduce its prices by 5% in the run-up to Christmas across the entire Valley range of PCs. A query of the XML database for the <make>Valley and a decrease in the <price> by 5% would do the job. Whereas, if the information had been stored as many individual static HTML pages, the webmaster would have to search all HTML pages where that make of PC appeared and update each occurrence of the price. If the website appeared in 10 languages, the information would have to be updated on each language version of the site *as* well. The added advantage of XML is that with XSL, the same information can be presented in different ways, for example as a table of data or as a bar chart. This ability is why Dell, with one of the most successful e-commerce sites, is now moving over to XML from HTML after a successful trial period. The advantage of the XML is that Dell can use the intelligence of the XML at the back end of their system, but easily convert the XML to HTML, including doing things like currency conversion for publishing on the web, transparently to the end user.

XML DTDs are being specified for a range of business areas for the transfer of information, in finance, and in the language industry (for terminology and translation memory exchange). The announcement by Microsoft that it will adopt XML as the format behind its office products must mean that it is going to become the norm for representing information of many different types.

XML is the future for the web and for all types of publishing. The web may be the first to exploit XML, in e-commerce applications, for example, but it will very quickly spread. This has great significance for translation suppliers as we shall see.

Building value on the web

There are significant challenges in providing multilingual web support. Not least the fact that web technology is evolving exceedingly quickly. For example, the specifications for the Java language and XML itself are still evolving. Browsers are constantly being upgraded to keep pace and to outdo each other. Things are changing so quickly, web software engineers have to write different versions of code and check which browser is being used first before running the code. However, we can begin to generalise a pattern of how the web is being used to build value for companies. Nua has developed a 10 Step Model that shows how companies can and do use the web to support their business. This model can be used to review the evolution and different types of websites, the technology used, and implications for translation suppliers.

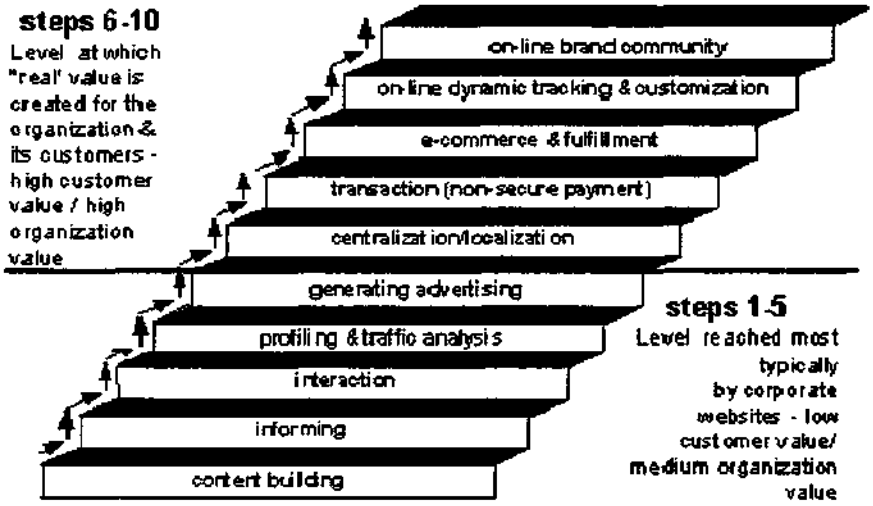
Low value websites

Steps 1-3 (Content Building, Informing, Interaction) of the Nua model are common among the first generation websites and are probably still in the majority. These websites respond to a company's immediate requirement for a web presence. Their content is purely informational and may include some low level of interaction with the user such as a basic form to collect data from site visitors.

These sites tend to be "static". That is, they follow a typical paper publishing model, using HTML as the mechanism to create and publish, concentrating on the layout of the information. The sites can be likened to many single page Word documents with page sizes optimised for the screen and some graphics and links between the pages. The site may even have been created in MS Word or a DTP package designed for HTML output such Netscape Composer. The documents are edited or replaced as required. There tends to be little formal

configuration control. As these sites start off relatively small, updates are done manually as required, directly in the HTML pages.

Fig 5. Nua's 10 step model of building value through internet presence



This type of website has certain advantages. It works well for marketing sites or "brochure-ware." It is a quick way for a company to develop a web presence. Even where there is a multilingual requirement, there are certain advantages, particularly in marketing, since translators in the local markets will have better local knowledge of style, preferences, and how to present the site to the end customer.

While the site remains small, this is probably adequate for a company's purposes. There are, however, many disadvantages. In multinational organisations, with many different business divisions, these sites spring up individually at the regional or local country level or for each business division. There is a lack of visibility of the total cost of the company's web presence, including the cost of translation; there is no corporate control over image; there is a lack of consistency of style; updating the information across all sites and languages is difficult if not impossible. Once you start adding the language dimension, what is a relatively small single language site and can be managed more or less adequately, can suddenly become a major budget item and a configuration control logistical nightmare.

Those providing the translation service have a considerable task. They must manage the configuration control of all language versions, work out the differences in the source language between the old and new versions of a website, retrieve previous translations from the appropriate translation memory, and then carry out DTP work on the HTML to fix any display issues of translated website.

For some sites, this approach is sufficient. But as a company's website or websites evolve, questions of cost, consistency, and value are raised. For example, Shell found itself with over 40 different websites world-wide, each with a different look and feel, content, objectives etc. Their 40 websites raised issues of cost and corporate consistency and were completely re-engineered to provide a single point of access for all visitors to the Shell website.

Medium value websites

Centralisation, however, is only one way in which a website can bring increased value to customer and organisation, as the next steps of the Nua model show. Steps 4-7 (Proofing and Traffic Analysis, Generating Advertising, Centralization/Localization) can be grouped together and considered to be the nature of a second-generation website. To build further value, a second generation website must enable the company to analyse what the cost/benefit of the site is to the company, where it is working and where it is not working by looking at which areas of the site are visited most. The company must be able to answer and respond to such questions as: how do changes to the site design affect the way visitors navigate the site?

What advertising works on the site in terms of bringing visitors in to the different areas? What advertising space on the website can be sold to third parties?

These sites tend to be integrated with one or more key business processes and deliver a measurable business benefit. For example, a computer manufacturer may put up a Frequently Asked Questions (FAQ) section of the website to provide first level support to end-users. This enables them to troubleshoot common problems themselves without having to contact the call centre. The costs of maintaining a call centre for a computer manufacturer can put 10% on the price of a PC, so the savings to manufacturer and consumer by developing a self-help website can be considerable. An average number of calls per product per day to a call centre for a computer manufacturer is in the region of 1000. The average length of a call is around 10 minutes. That is 10,000 minutes of call time per day. The staffing costs to support this would be around £10,000 per day, or £500,000 per year. Whereas, if many of the calls could be answered by visiting the website, where the cost to the computer manufacture would be less than 1% of that for answering a call, a large percentage of call centre costs could be avoided.

Providing these benefits usually requires the website to be centralised. Central control enables budgets to be managed and consistency to be maintained. Design and maintenance tends to be carried out centrally with updates being reviewed and released by a central function. This tends to free up budget for additional bells and whistles such as animations, sound, video, making use of Java applets, JavaScript, and Dynamic HTML. If the website is multilingual, the various language versions can hang from the corporate version and be managed centrally. This often coincides with centralising the translation function. A single group within the organisation is made responsible for tracking the changes in the source language, and for commissioning the translations. This would typically be outsourced to a medium to large translation supplier with offices or freelance contacts in each of the language countries into which translation is required.

This approach has certain advantages. It enables the translation budget to be planned and managed. A consistent corporate image can be maintained not just in the source language, but in all target language versions of the site as well. There is also greater opportunity for the owner of the website to ensure that they benefit from existing translations, rather than leave it as an opportunity for translation suppliers to increase their margin.

However, more often than not, centralising the website means that the home country of the multinational tends to takes precedence over the language versions. For example, the home page is in English, and a pull-down menu lists the language versions (not localised) which are available. Even worse, the "<mycompany> worldwide" link is buried away at the bottom of the link list. What tends to happen is that the language version websites are brought together under the corporate site but without any further integration. They are still separate sites for each language. So tracking changes to the source language pages and identifying what needs to change on the target language pages of the site is made even more difficult on the larger, now centralised, site.

To localise these medium-size sites, often loaded with the latest web effects, can be a very time-consuming, expensive, and tricky business. In many cases, the site is being translated for the first time and little or no consideration has been given to localisation issues in the design. It is likely that some level of communication with the original developers is necessary. Any prospective translation supplier will need to be web-aware to cope with the many different methods and tools used to create the various effects and be familiar with the various computer languages and scripts used on the site. Often, however, re-coding the site on a large scale is not an option and translation becomes a fraught and expensive activity.

High value websites

Once companies have gained an awareness of how the web is building value as a marketing tool with perhaps some basic selling infrastructure, they may look at how they might use the web to build further value to the organisation. The third generation website is equivalent to Steps 8-10 in the Nua model (e-commerce and fulfillment, online dynamic tracking and

customisation, and online brand community). In these steps the website becomes an intrinsic part of a company's main business process.

Companies such as Amazon.com deliver their service directly over the web. They may be selling books, toys, financial services, or whatever, but the web enables them to have a much closer "relationship" with their customer. Making use of web technology, they can glean considerable information from visitors to and purchasers from their website. Using the information about their customers and visitors they can carry out a lifestyle profile to predict future buying habits.

Using the technology, companies such as Amazon can track a user's visit to a website, register preferences and customise the information specifically to that user's interests. For example, someone buying summer baby clothes size 3-6 months might be presented with a special offer for 6-9 month winter clothes. In a multinational website, an analysis of traffic may show that a particular background colour attracted more visitors in one region, whereas another background colour was more effective in another. So the website could be customised appropriately per region. This is taking the store loyalty card concept to the web; in this case, every customer, not just those with the loyalty card, can be tracked. Of course, it can go wrong. For example, a highly politically motivated vegetarian kept being sent special offers on meat. The customer relationship management software program used to analyse the purchases on her supermarket's loyalty card had noted that she did not buy meat there and assumed she must be buying it somewhere else. It took several angry letters before the "50p off best mince" vouchers stopped arriving.

The ultimate in using the web to build value is to create an "online brand community". For example, an online store selling car parts might also run chatrooms for enthusiasts of various makes of car. If that site becomes the *de facto* place to meet other car enthusiasts, then it is likely to be the *de facto* place to buy car spares as well.

Because this environment is tailored to individual requirements and draws on various different resources to create the correct "experience" for the end-user, by its very nature the architecture of this approach must be distributed - across different systems and locations.

A job for XML

To provide this level of customisation across a distributed architecture requires the integration of various technologies, both new web and legacy, that form the company's complete business process. This is where XML comes in. XML can be used to provide the glue that binds together the transfer of information from the different sources, for example, stock control, pricing, and credit card transactions.

Because the web is extending beyond the PC screen to handheld computers, the mobile phone, and webTV, the presentation of the site will need to vary depending upon the device being used to view it. Again, XML provides the answer, defining the structure of the information, with XSL determining how that information will look on whatever device is being used to access it.

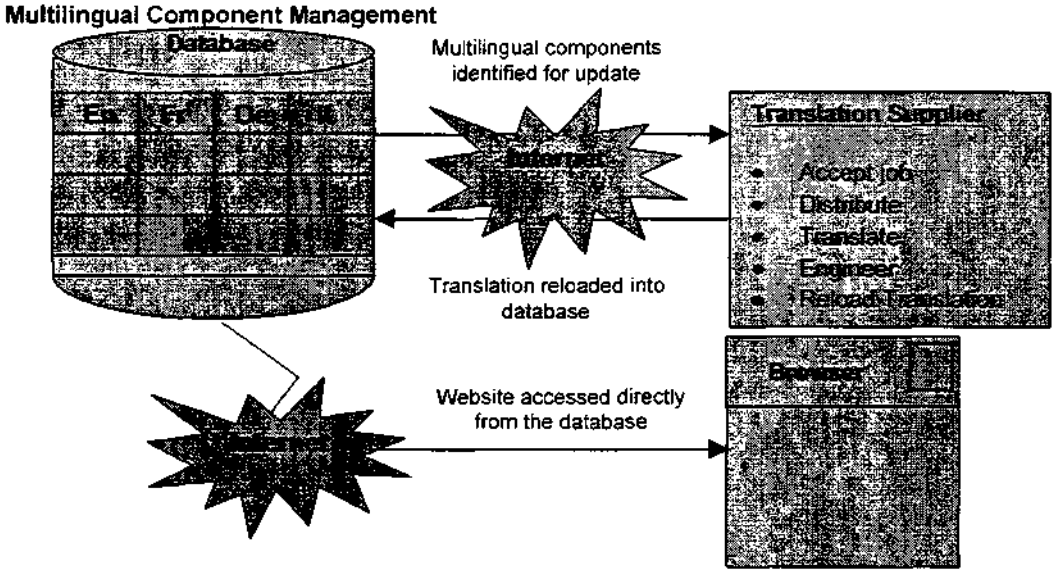
The requirement to customise content to the individual user's requirements means that the pages do not exist as complete and static HTML pages, but as chunks of information. These chunks of information are brought together as "virtual" pages when and only when they are required and requested, existing only for that user and only as long as that user requires it. XML stored in databases can be used to do this. It enables information to be componentised so that it can be re-used in the different, possibly infinite, combinations in which users might request it. On a news website, for example, the headlines of all of the current news stories might appear on the first page of the site. If a user had customised their view to see only Sport or European Politics, they would get a subset of the news items. If the user clicks on a Summary button, they might get the headline plus the first paragraph, giving an overview of the item. If the user then clicks on the Details button, they would get the news item in full. In the HTML-only world these pages would have to exist before the user came to the site. In the component world, the headline, the summary paragraph and the full story exist only once in the database. Managing information as components makes it feasible to make frequent

changes without having to completely rework the presentation. For example, adding the latest news stories, re-prioritising existing stories, and removing old ones.

Multilingual Component Management

One of the advantages of an XML world is that when XML is stored in an XML component database, the database knows exactly what is contained within every element of the XML. In the old world of DTP format documents, like Word, FrameMaker, etc. a document management system would see the document as a closed file. XML turns the document inside out, exposing its contents to applications. It is a bit like turning an orange inside-out to expose the segments. Component Management enables the user to control the individual elements of an XML document, pulling elements together, updating them, and reusing them as required.

Fig 6. Integration of a translation service with multilingual component management



Because XML describes the structure of the text, the Component Management system can manage parallel document structures. Applied to a multilingual environment, Component Management can be used to store and manage parallel language versions, so when a source language element is updated or reused, the equivalent target language version can be flagged for update or reuse. "Multilingual Component Management" gives information creators control of their data across all languages stored, thus allowing them to manage closely their translation requirements. It gives them control of 100% matched "translation memory" so that when they make minor source language updates they can identify the target language components which require translation. It ensures that the latest, validated versions of translations are stored back into the database for future use, and avoids the pitfalls of various out-of-date translation memory databases scattered among different translation suppliers. Moreover, the website can be driven directly from the database; XML can be converted into HTML "on-the-fly" as the information is requested by a browser visiting the site. The latest browsers now support XML directly and so in the near future even the conversion to HTML will not be required.

XSL can be used to tailor the layout of the information specifically for the end user, separately from the creation of the information. For example, if our computer manufacturer found that French buyers were more likely to make purchases based on computer specification, whereas English buyers were more price sensitive, the same, single source of information could be presented accordingly to French and English consumers. In the HTML world, to offer this level of customisation requires separate and differently structured HTML pages. To achieve this, the translated HTML source language page would have to be re-engineered for the target language. If other preferences were detected among other nationalities, and yet further preferences depending upon the age or sex of the customer, the re-engineering would

become unsustainable. In the XML world, the information is translated and stored once in the database, and the XSL provides the presentation rules for the different language versions, or user types.

Implications for Translation Suppliers

By their very nature, these high value-building, componentised, and database-driven websites require frequent updating, often of relatively small amounts of text, so fast turnaround of components is required to ensure that the multilingual versions visible around the world are as up-to-date as the source language version. For example, a vendor of virus detection software will require the latest news on the latest virus circulating globally to be available simultaneously in all languages as soon as the information is released.

The material to be translated for this third generation website tends to be located within a variety of company systems, in many locations, and cover a wide spectrum of document types and language styles. There might be back-end legacy databases of product information and descriptions, there is the website user interface itself, chatrooms, background whitepapers, legal agreements between the supplier and customers, plus all of the additional aggregated content which creates the user "experience" of the particular brand.

This requires a multi-skilled translation supplier. The translation supplier must be able to link the different content types and language requirements to the appropriate translator, deal with any website engineering issues which may arise, understand the brand experience being created and be able to mirror this in the various languages. The supplier must therefore have at its disposal a world-wide network of translation and IT expertise in many subject areas, supported by the IT infrastructure to deliver the turnaround times at levels of quality required over the internet.

Having said that websites are being used more frequently to deliver value to businesses using increasingly complex technology, it is likely, of course, that there will still be hundreds of thousands of first and second generation sites which translation suppliers will still be required to localise. However, it is also likely that while HTML may remain the format in which these sites are published, XML may well be the format in which they are written, with the XML being converted to HTML for display on the web. This turns the web into just another publishing medium.

The use of XML within applications such as Microsoft Office will enable much tighter integration of document and component management techniques within everyday computing environments. This means that information developers will be able to harness this functionality, making component management and therefore multilingual component management more accessible. The future pervasiveness of XML and component management means more frequent, faster translation of smaller chunks of text will become the norm. While this model pushes the translation supplier further down the supply chain, reducing the value-added element and hence margins, the internet and web technology opens up the possibility for a new approach.

The New Translation Suppliers

Translation suppliers need to move up the value chain by providing a mix of technology and translation service to enable companies to maintain their multilingual website cost effectively. The Internet enables translation suppliers to work in close partnership with their customers, polling the contents of whatever technology is used to manage the multilingual dimension of the website. The software should detect changes in the website and automatically trigger a translation job.

Different levels of translation quality can be provided depending on the requirement for translation. Web tools can track the requests for different pages in different languages requesting human translation for the most important or popular and perhaps Machine Translation (MT) for those where gist versions would suffice. Technology such as MT could be used for chatroom sections of multinational websites.

To be able to respond to these more frequent requests for translation, the translation suppliers themselves must make use of the web technology to recognise customers at a one-to-one level and provide them with a swift service tailored to their needs. If an established legal client requires a 12 hour turnaround for a job from English into Brazilian Portuguese, the translation supplier should be able to receive the job over the Internet, automatically scan its translator database, identify an appropriate translator, calculate costs, payment details for both customer and supplier, estimate a return date, and send out the job for translation.

New pricing models are required for this "translation clearinghouse" concept, not just based on total volume, but on frequency of update, response times, and number of components for translation.

The larger translation suppliers are now beginning to develop technology, form partnerships, and make acquisitions in order to provide a mix of technology and a service offering to meet this new demand. Some suppliers are using the web as another route to their traditional translation services. Some are partnering with software suppliers offering multilingual management as a feature of their product. Others are developing their own toolsuites. Translation technology vendors are partnering with component management vendors or with web portal providers who will provide translation as a service of their website.

The future?

It is still early days to determine how the translation market will shake out. But it is clear that the massive market of business over the web will become a key area for translation suppliers. The drive for multilingual websites will come from the e-business requirement to "go global". XML is the technology that will enable these e-businesses to communicate over the Internet. As XML and component management become pervasive, the types of translation model now being developed for the web will be required in other areas of publishing. While there will still be room for the smaller niche translation suppliers, for the mainstream providers, only the those who are big enough and ready for this change will survive.