[*Translating and the Computer 22*: Proceedings of the Twenty-second international conference…16-17 November 2000. (London: Aslib, 2000)]

XML and the Localization Process

Daniel W. Dube
John D. Rice
Lighthouse Solutions, Inc.

Overview

There has been much publicity in recent months regarding developments in the extensible Markup Language (XML) area, and the revolutionary changes that this standard will be responsible for in virtually every industry. Much has been written to describe the impact of XML with regards to e-commerce and electronic delivery of content. However, XML is also ideally suited to provide tremendous benefits to companies that have a need to author, manage and produce content and/or publications in multiple languages. This paper will describe how some companies have used existing technology based on XML (and its predecessor, SGML) to realize impressive results, and will provide some of our observations as to where XML-based technology is heading in the next 24 months with regards to increasing the efficiency of the localization process.

"Typical" Localization Process

Today, most companies with a need to deliver information to a global audience still rely on a manual, paper-based process for localizing their content. Human intervention is required at all stages, from the origination of source language data through review, quality assurance, and final delivery of localized information. In most cases, authors are working with tools such as MS-Word and FrameMaker and sending complete document instances to a translation vendor. Because these files are stored and managed at the document level, authors have no easy way to identify only the pieces of content that have changed between revision cycles. As a result, authors continue to send entire documents to the translation vendor for localization, even if only a small portion of the content has been changed.

There are inherent inefficiencies to this process, which lead to the following typical results:

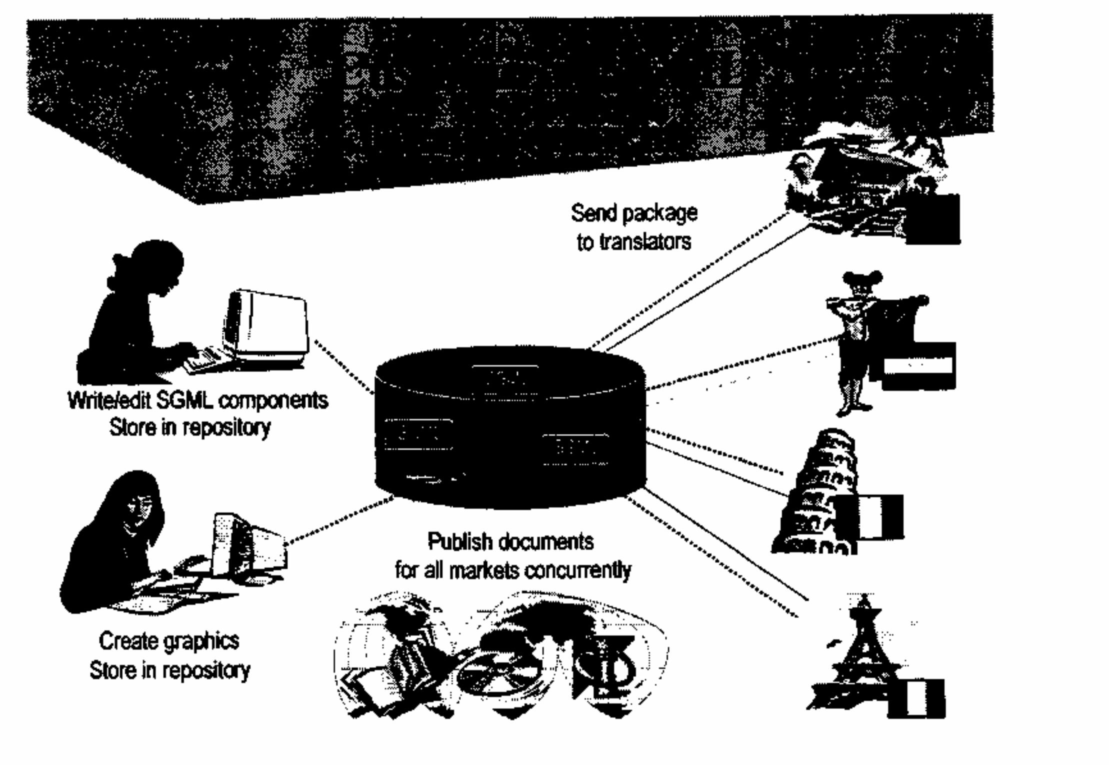
- Longer translation turnaround time
- Higher localization costs
- Lost revenue opportunities

SGML/XML and Document Management Technology Have Helped

Over the last several years, some forward-thinking companies have made a strategic investment to create and manage their data in a structured, neutral data format: SGML (Standard Generalized Markup Language, an ISO standard) and, more recently, XML (extensible Markup Language, a W3C standard). Through a combination of standard products and extensive customization, these companies have created production environments with these characteristics:

- Author and manage information in "chunks" as information objects in a database (rather than entire documents), which enable the reuse of information across multiple documents (e.g., a "warning" or a "task" can be written once and shared by many documents)
- Write the content once and produce multiple end deliverables (e.g., paper, web, and CD-ROM)
- Track only the information fragments that change between revision cycles and their associated target markets, and only send out changed objects for translation to a specific language/market

The following diagram depicts the workflow of this type of environment:



Consider the following statistics from companies that have successfully put an environment like this into production:

- Cummins Engine realized a dramatic 70% *reduction in translation costs*
- Tweddle Litho Company, in a project to produce owners literature for a major American automotive manufacturer, was able to cut the time to translate the manuals into 30 languages *from six months down to two weeks*

Note: For more detailed information on these two case studies, please refer to my paper from the ASLIB Translating and the Computer 21 conference (1999).

SGML/XML and Document Management:
Still Room for Improvement

There are solutions available from traditional SGML/XML systems vendors that attempt to replicate the environment depicted in the diagram above. The most well-known of these solutions are Lingua (produced by Chrystal Software as an add-on to their Astoria content management system) and Parlance Ambassador (produced by XyEnterprise as an adjunct to their Parlance Content Manager product). These systems are sold as a "toolkit", requiring a purchase of base product technology and significant customization services to model the solution to the specifics of your production environment. While these solutions certainly add efficiencies to the localization process, there are still some inherent limitations:

- *These products typically only support content that is marked up in XML or SGML.* There is no advertised capability to support legacy data or unstructured data in other formats, such as FrameMaker, Interleaf, or MS-Word.
- *These products only support XML/SGML content that is stored or managed within their repository.*
  For example, Chrystal's Lingua solution will only work with structured content stored in the Astoria repository. It will **not** work with FrameMaker files stored in Chrystal's Canterbury repository system.
- *These solutions can potentially lock a customer into a proprietary vendor solution.* This is possibly the most disturbing issue associated with these solutions. Even though they are based on the open standards of SGML and XML, it is very difficult to migrate information from one of these repositories to another system in the event you ever want to upgrade to another product in the future. While the core SGML/XML data may be easily exported from the repository, it is often very difficult to migrate information about *links, metadata, workflow, and version/history information.* As a result, you risk losing much of the value investment of the content during the migration process.
- *It is difficult to integrate associated applications and data sets with these products.* For example, it may be desirable to connect these content repositories to your translation memory tool, terminology database, or a translation web portal maintained by your localization service provider. This would most likely be an expensive customization to these technologies.

The Next Wave: XML Portals

These problems will be addressed by a new generation of open tools, based on the concept of XML portals. In today's world, the localization process involves collaboration of information stored in many related, yet disparate, applications, including:

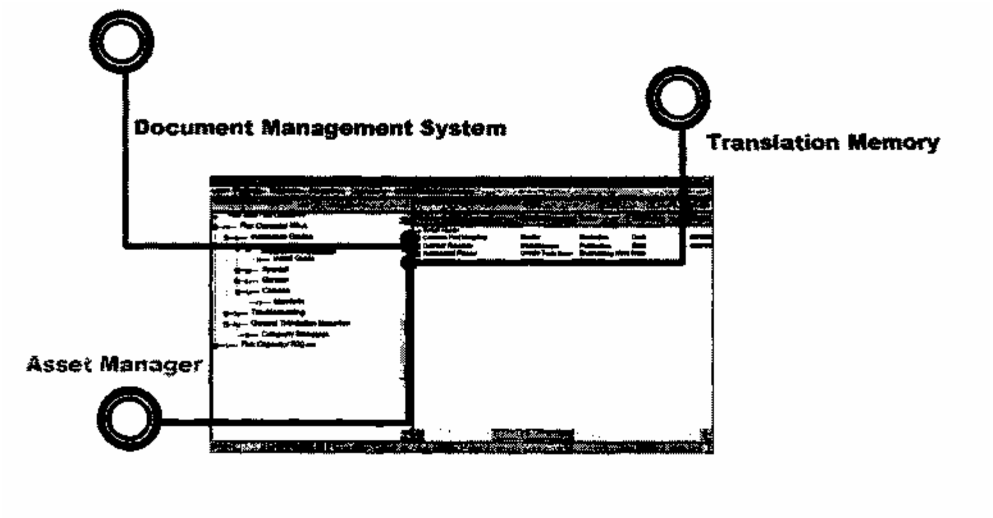
- Content management systems (e.g., Documentum, Astoria)
- Translation memory tools (e.g., Trados, STAR Transit, SDLX)
- Machine translation tools (e.g., LOGOS)
- Terminology databases and glossaries
- Digital Asset Management (DAM) systems

XML portal technology shows the promise of being able to provide content owners with the ability to link information stored in these distributed, heterogeneous environments. A portal can act as a client to retrieve resources and provide a virtual view of a collection of information. The goals of this technology will be to:

- Connect information collections
- Layer new business rules over existing applications
- Streamline the localization process with automated operations

Connect information Collections

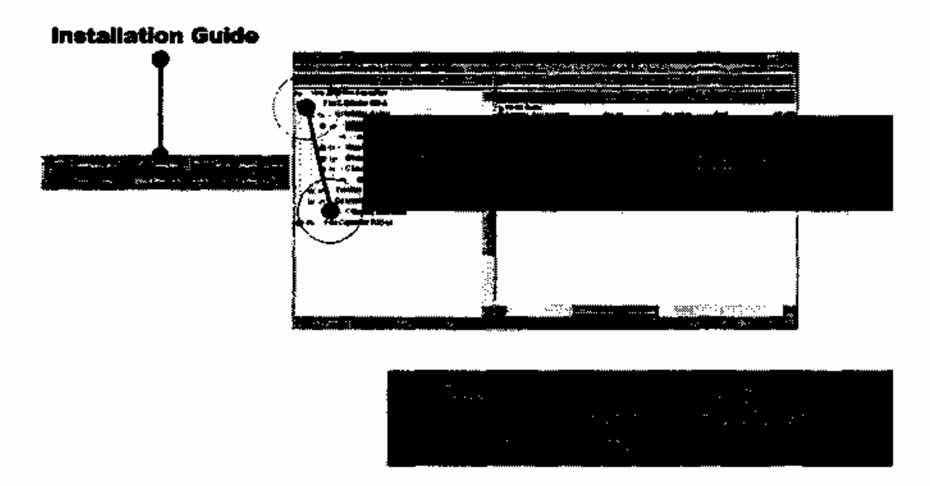
Such tools will enable a user to create a customized view of information that is relevant for his/her work, regardless of where the information may physically reside. The following diagram shows a localization example of this paradigm:



In this example, a localization project manager is provided with a view of information that is relevant for a current translation project. While the user interface may give the appearance of a normal Windows Explorer directory view of information in a folder, the actual information resources physically reside in different applications: a document resides in a content management system, an illustration is stored in an asset manager database, and a translation memory is managed by a TM processor.

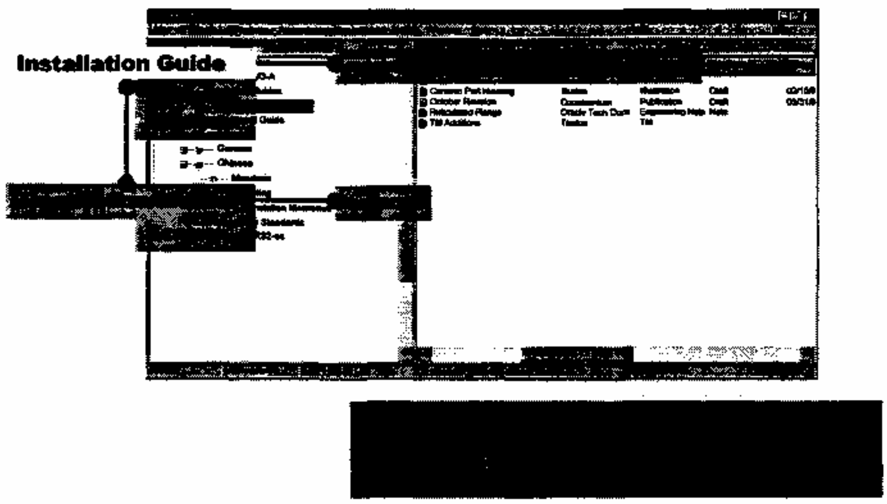
Layer New Business Rules Over Existing Applications

With this type of view, it is possible to create links between disparate objects and establish important relationships. In our example illustrated below, we have associated an Installation Guide with a specific translation memory. When the source language version of the Installation Guide file is updated, it will be quite easy to retrieve the appropriate translation memory that is linked to it. The following diagram depicts this scenario:



Streamline the Localization Process with XML Portals

Once you have linked relevant information objects and created associations between them with appropriate business rules, it is possible to further automate the localization process by assigning operations to linked information collections. The following diagram illustrates this concept:



In this example, a business rule has been established to do the following:

- Once the "status" of an installation guide has been updated from *Draft* to *Approved,* generate a new translation package.
- Find the linked translation memory that is associated with the installation guide and add it to the translation package.
- Export the translation package (containing both the installation guide *and* the translation memory) to the translation vendor's workflow system to be localized.

Benefits of XML Portals for Localization

The benefits of migrating to a localization solution based on XML portal technology include the following:

- Maximize reusability and repurposing of information, which will lead to lower translation costs and faster turnaround time
- Leverage the ability of the Internet to increase efficiency of remote collaboration between content creators and translation vendors
- Extend the life of existing technology infrastructure by adding new functionality and automation to current business processes
- Personalize the features of existing technologies (e.g., document management systems, translation memory tools) to the needs of specific users, such as localization project managers
- Allow for easy migration to new technologies, since the core infrastructure will be based on an accepted international web standard (XML)

Peering Into the Future: XML Content Management and Translation Memory Technology Will Merge!

As we have researched the potential for XML and its associated linking capabilities in the context of XML Portals, we have developed a very intriguing (and, we suspect, highly controversial) prediction:

We believe that XML will enable translation memory tools to dramatically evolve to a more efficient and scalable storage model.

Current translation memory applications typically work in this manner:

- A "database" is created for language pairs that are usually stored as segments (such as a sentence). Note that the "database" is often physically stored as files in a directory structure on disk, not within a true relational database.
- As new files are fed to the TM application, algorithms are utilized to compare the segments in the new file with segments in the TM database. 100% matches are flagged and substituted with the approved translation match from the TM database. "Fuzzy" matches are highlighted and the translator can utilize the suggested match to refine the localized version of the segment.

While translation memory tools have certainly provided a tremendous technical leap forward for the localization community and have unproved the efficiency and consistency of the translation process, there are also opportunities for improvement with this technology:

- Most TM tools are PC-based and are not necessarily optimized to scale to enterprise levels for large groups of people to utilize concurrently.
- TM tools require redundant storage of data. The "segments" residing in a TM database are not the actual source or target language content that authors and translators are developing.

We boldly suggest that if content is authored in XML and managed in an XML-aware repository, then **the XML data <u>is</u> the translation memory!** We predict that it should be feasible to apply TM algorithms to process the actual XML content residing in a repository. This approach would provide several significant advantages over current TM technology:

- Most XML content management systems are designed to scale to large numbers of users. Tighter integration of TM technology with content management systems will allow TM tools to be better utilized in a multi-user environment.
- Tighter integration of XML and TM technology will eliminate redundant data storage and allow for a more streamlined and efficient processing of the data. There will be no more need for importing and exporting information into TM tools, and filtering between different file formats will no longer be required.
- This approach will also allow a user to associate content with multiple translation memories (for example, if a company wants different translation results by various product lines, technical vs. user audience, etc.)

Conclusions

By now we hope the message is clear: the benefits of XML extend far beyond the e-commerce B2B scenario. The true promise of XML for localization professionals is the ability to streamline and automate the translation process, enable collaboration, and to maximize the capabilities and efficiencies of content management systems and computer-aided translation tools.

Dan Dube
President
Lighthouse Solutions, Inc.
136 Harvey Road, Suite A-102
Londonderry, NH 03053 USA
Phone: +1603 627 4090
Fax: +1 603 627 4060
E-mail: dan@lighthouse-solutions. com