The Gilbane Report

White Paper

Topic-Oriented Information Development and Its Role in Globalization

The Case for the Darwin Information Typing Architecture (DITA)

Bill Trippe
December, 2004

idiom

Executive Summary

Globalization is a critical issue for any company interested in expanding its markets. For the company that markets sophisticated products, globalization is both more difficult and more critical because of the rich content that is needed to support these products.

Product document localization may well be the most difficult aspect of globalization. Documents often are long, with a mixture of text, tables, charts, and graphics. Moreover, the documentation must be produced in different forms—print, online Help sets, HTML. Translating such documents into multiple languages can be challenge.

Single-source publishing has matured as a method for producing complex documents in many formats. XML in particular has become the preferred format for single-sourcing, enabling companies to both repurpose their content into different formats and reuse content modules in different content types. Thus, a procedure that appears in one document can be stored once, edited once, reused in many different documents and repurposed into many different formats.

For all of its upside, XML-based single-source publishing has proven to be expensive and complicated to implement. XML-based single sourcing requires significant tool development, data conversion, and system integration prior to realizing the benefits of repurposing and reuse. To mitigate this, some vertical industries have developed their own XML tag sets. While successful on their own, these vertical industry efforts have not been extensible to other industries.

A new XML-based approach to information development is the Darwin Information Typing Architecture (DITA). DITA is a topic-centric architecture that provides a core Document Type Definition (DTD) and schema for developing documentation typical of many kinds of products. Conceived over several years at IBM, the extensible DITA architecture is now being managed by a technical committee at OASIS.

We looked at one organization, software developer Information Builders, Inc. (IBI), and their implementation of DITA for managing a large set of documentation that is translated into many languages. IBI made a strategic decision to adopt DITA, has implemented it, and is already realizing benefits from the decision.

This paper is sponsored by Idiom Technologies, Inc. Idiom provided the solution for IBI: WorldServer Global Electronic Publishing along with the recently introduced DITA option, WorldServer OpenTopic.[1] We think the new Idiom solution is significant for combining the traditional functions of an XML Content Management System (CMS), a Globalization Management System (GMS), and a commercial DITA solution. In doing so, Idiom seems to be taking advantage of an intriguing nexus where single-source publishing, XML encoding, and globalization meet.

[1] Product, technology, and service names in any of our publications are trademarks or service names of their respective owners.

Introduction

The term "globalization" has broader meanings within the professional localization industry. The Localization Industry Standards Association (LISA) describes globalization as "the process of making all of the necessary technical, financial, managerial, personnel, marketing, and other enterprise decisions necessary to facilitate localization.[2]" Localization, then, is "process of modifying products or services to account for differences in distinct markets."

We use the term "globalization" in this white paper as it applies to content. Such product content must be localized for each market.

Global expansion is increasingly important for growing, successful businesses:

- Global markets are an important growth area for companies well established in domestic markets.

- Companies that are already marketing and selling their products overseas face all the profitability challenges they do in domestic markets—and a number of challenges that are unique to addressing those global markets effectively and efficiently.

For industries with sophisticated products—such as industrial manufacturing, consumer electronics, and software—the challenges are even more complex. These challenges include consistency in branding and messaging, and challenges related to delivering current and complete information. These issues and others all tie back to how and what information is communicated in the process of marketing, selling, and supporting products—and the need to provide that information in multiple languages.

As a result, many companies with a global focus have identified product globalization as an essential goal. Depending on the products, globalization can be a significant effort; various combinations of Web sites, documents, and other materials need to be translated. In the case of software, the user interface mechanisms and error messages need to be localized as well. The more information that is connected to the product, the more daunting—and expensive—the globalization challenge.

Companies faced with the need for product globalization also recognize that such efforts must be done efficiently but with a close attention to quality. Extending a brand—and marketing complex products—into multiple new languages requires a kind of consistency that can best be supported through automation.

Companies that are already marketing globally have learned that globalization affects both the top line and the bottom line:

- Globalization affects the top line by allowing companies to enter new markets. By extension, effective globalization enables companies to enter new markets in a

[2] *The Localization Industry Primer*, Second Edition, by Deborah Fry, updated by Arle Lommel, 2003, LISA, The Localization Industry Standards Association, http://www.lisa.org.

timely fashion and to deliver new versions of their products faster than their competitors

- Globalization affects the bottom line because, without automation, globalization can consist of time-consuming and error-prone processes that do not scale and are not cost efficient. Conversely, the right technology can make globalization processes much more efficient and less prone to error while significantly reducing cost

Technologies have emerged to support globalization—first, computer aided translation (CAT) tools, running on PCs, that helped translators be more efficient, and later, more sophisticated and comprehensive Globalization Management Systems, running on servers (to help global enterprises be more efficient). Very recently, companies have begun tying content management systems more closely to globalization solutions. This way, the content that supports products—documents, Web sites, catalogs—can be more readily translated and localized.

Some products have substantial content that is key to the operation, maintenance, and support of that product. When these products are marketed globally, the localization of the content is an integral part of the effort. This paper takes a closer look at one aspect of product globalization, namely document localization. How can this localization be done in an efficient, timely, and high-quality fashion? And what role does technology play?

We then look at the unique role that XML can play in product globalization, and in particular an important XML initiative, the Darwin Information Typing Architecture (DITA). DITA promises to be a significant new tool in developing content for product support, and may also have an important role in product globalization.

Documentation and Product Support

Sophisticated products are supported by a great deal of content—user manuals, maintenance manuals, catalogs, marketing materials, and on and on. Some documentation anecdotes are widely shared and perhaps even apocryphal—the documentation for the latest Boeing aircraft being of greater mass than the airplane itself. But there is no doubt that complex products—large manufactured platforms, enterprise software and hardware, medical equipment, consumer electronics, prescription drugs—have equally complex content that is intrinsic to the product.

When any of these products are introduced to new global markets, the globalization of that content is key to successfully selling and supporting them:

- An automobile manufactured in United States must have its manuals and other supporting documents translated to the local language. This even includes the glovebox owner's manual, which needs to be in the local language and also needs to reflect instrumentation and other features that may be unique to that country or language.

- Enterprise software is typically published with supporting user manuals, reference manuals, programming guides, and various kinds of online Help. The supporting content needs to be translated and localized. Indeed, as enterprise software grows in complexity, so too does the supporting information.

- Pharmaceutical companies have identified time to market as a key element of profitability in launching a new drug in a new marketplace. To introduce a new prescription drug into global markets, pharmaceutical companies need to translate and localize everything from marketing materials to product labeling and clinical data that supports the regulatory approval process in the locale.

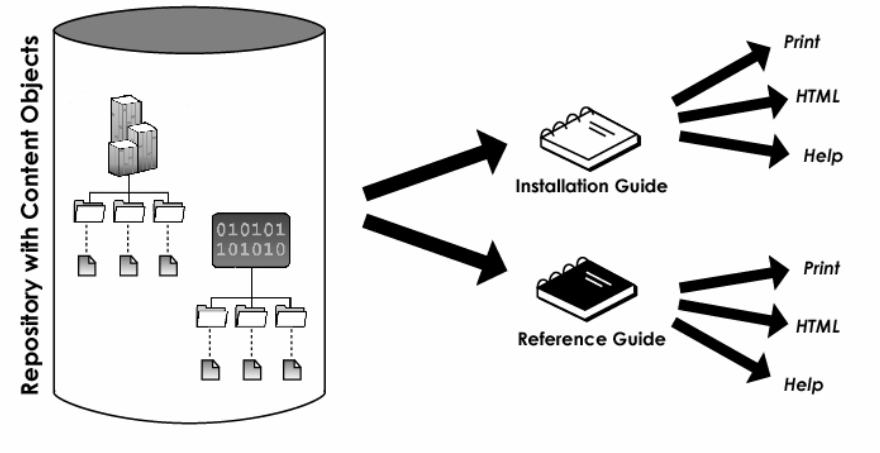
Multichannel Publishing and Reuse

In each of the above examples, the content is voluminous and has to be distributed in multiple channels—print, online, CD-ROM, HTML. As organizations have seen the content grow, and the need for multichannel publishing grows, they have invested in tools and technology to automate these processes more. Organizations talk about the "multiples problem"—multiple products, supported by multiple content types that need to be provided in multiple languages and formats. Organizations have begun to tackle this multichannel challenge with single-source publishing, where the source content is maintained in a display-neutral format, such as in a database or in the eXtensible Markup Language (XML). The single source of content can be updated once, and the various formats of the content can be produced from the central source.

On a small scale, this kind of multichannel publishing can be done with desktop tools. Adobe FrameMaker, for example, can produce print, HTML, online Help and other formats natively or with the support of add-on products and helper tools. Microsoft Word can even be used for authoring, with plug-ins and custom tools providing the multichannel publishing. But true single-source publishing assumes some kind of format-neutral encoding, and, increasingly, organizations are looking at XML as the encoding mechanism.

Organizations that have solved the multichannel distribution problem with single-source publishing also have often identified content reuse as a primary goal. Whereas multichannel publishing involves repurposing content into other formats, reuse occurs when a single content object is used in multiple content types. For example, when a procedure for lubricating a certain part is used in maintenance manuals for several automobiles, or when a task description for a certain software application is used in the user's manual, reference manual, programming guide, and online Help.

Indeed, there is great potential when content reuse is combined with content repurposing. Consider the following illustration, where several content modules are reused in several content types, and the products themselves are then repurposed into several formats.



How typical is it that organizations have achieved this level of reuse? While we do not have any quantitative data, our contact with various industries suggests that the largest organizations are capitalizing on reuse after investing in the appropriate supporting technologies, single-source encoding, and deployment of the tools necessary for multichannel publishing. This is especially true in certain vertical industries, where standard XML vocabularies were promulgated and, in some cases, enforced. In military contracting, for example, contracts for major weapons systems developed for the U.S. Navy often require the documentation to be delivered in SGML or XML files that follow certain standard Document Type Definitions (DTDs).

In a later section, we will discuss a specific example of multichannel publishing and content reuse at an enterprise software company, Information Builders.

XML and Reuse

How does XML support reuse? Reuse is when logical components or objects of content are separately created, edited, managed, assembled, and reassembled into different content types. In our software documentation example, this could be separate tasks in a software application. XML-encoding by itself does not provide the mechanisms for reuse; rather, the XML-encoding can provide the hierarchical structure—the clear encoding of the tasks that allows the encoded content to be machine controlled for creating, editing, storage, and assembly. Typically, content management technologies provide the mechanisms for reuse:

- Repository functions that support the storage of the content in modular fashion, and core repository services such as check-in, check-out, and revision control.

- Application integration functions that enable multiple content objects to be collected, concatenated, and presented to applications such as XML editors, publishing tools, and converters.

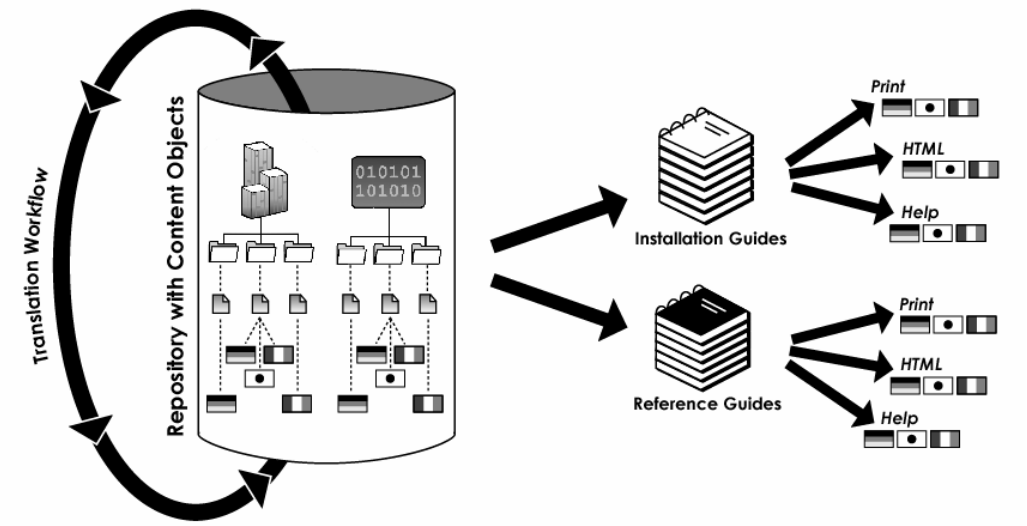
- Distribution functions that enable the content objects to be published into multiple formats and through channels such as syndication.

So while XML on its own does not provide reuse of content, it provides an excellent industry standard, machine-readable mechanism for supporting reuse. As a result, many organizations that are creating systems for content reuse are doing it with XML-encoded content.

XML, Reuse, and Globalization

For companies with a need to globalize complex products that require substantial supporting content, there is an intriguing nexus where XML, reuse, and globalization technologies meet. Consider our earlier illustration showing how content reuse and multichannel publishing can work together. Content objects can be reused and reassembled into different content types, and then those content types can be repurposed into different formats. Consider this same process, but now with the need to translate and localize the content.

One approach might be to manage the single-source content so that the single source is also the language source. For a US-based company, this could mean managing the single source as English. When one of the content objects is updated, a workflow is triggered that flags the object for translation. After the translation is complete, the newly translated object—and the existing unchanged objects—are assembled into the finished manual. The finished manual is then converted into print and the other necessary formats.



In this kind of publishing model, the benefits of having the content in reusable components extends to globalization as well as reuse and multichannel publishing. By being able to manage translation and localization of content in such a discrete, controlled, and automated manner, companies enjoy significant efficiencies and benefits of scale.

XML, Content Development, and DITA

XML supports multichannel publishing, reuse, and automation that can make globalization more efficient and effective. So, why isn't everyone using it? The simple answer is that XML-based publishing is hard to implement.

Before XML there was SGML, the Standard Generalized Markup Language, but which some people jokingly referred to as "Sounds Good, Maybe Later." In other words, it was sometimes seen as a great but impractical idea. That said, some organizations have been successful with SGML-based publishing, and more recently with XML-based publishing.

Increasingly, content management implementations have XML in some part of the workflow—either some of the content or metadata is encoded in XML, or the system produces XML as an output. Even personal blogging tools like Movable Type, for example, produce an XML-encoded RSS feed. It's notable that RSS, which stands for Really Simple Syndication, is widely used. Virtually all news feeds on the Internet and all blogs are available through an RSS feed. (If you Google RSS, you get a staggering number of hits—more than 90,000,000. This means that the phrase RSS appears on over 1% of the 8 billion pages currently indexed by Google.)

Clearly organizations have found ways to use XML in content management and publishing. But providing an RSS feed of a blog or news stories is one thing, and encoding product documentation in XML is quite another. Technical documentation represents a much more complex document type, hence a much more complex Document Type Definition (DTD) and much more complex encoding of source files.

Technical documents have many complex text elements that require unique encoding—tables, bulleted and numbered lists, figures. Moreover, technical documents are typically lengthy and multi-level, so the XML encoding must reflect and enforce the hierarchical structure. The complexity and volume of the markup can be expensive to implement, and may require dedicated technology and more user training and support than productivity tools such as Microsoft Word or desktop publishing tools.

To overcome this complexity, many vertical industries developed their own vertical DTDs, with a goal of promulgating tools, technologies, and best practices among a user community. Different organizations within the military have DTDs, the automotive industry has an initiative called J2008, and the aviation industry has its own initiative, ATA 2000. As a result, some of the XML technology vendors have developed packaged solutions for these DTDs that help customers get up and running faster.

A broader, cross-industry solution is still needed.

DITA and Topic Architecture

The Darwin Information Typing Architecture (DITA) is a horizontal industry effort, initially spearheaded by IBM Corporation, now developed within the OASIS open standards consortium, to provide an XML framework for developing product documentation. While there have been other efforts to provide a "standard DTD" for technical documentation (the DocBook initiative, for example), DITA takes a more extensible approach.

The Gilbane Report is a long-time supporter of industry standards like SGML and XML. There is great value in industry-wide adoption of vendor-neutral approaches to core business

problems like information development. Of course, merely declaring a standard is not sufficient. Other things must follow for a standard to be successful—a demand from a critical mass of users, an open process for honing and maintaining the standard, and support from both the vendors and from open source developers.

DITA seems to have all of these things.

- Judging from mailing lists, newsletters, and other publications, there is a high level of interest in DITA from the user community.

- IBM developed and promulgated all of the initial work on DITA, but the advancement and governance of DITA has been handed over to the OASIS open standards consortium, which has formed the OASIS DITA Technical Committee (http://www.oasis-open.org/committees/dita). Members include folks from IBM (Don Day of IBM is the chair), Intel, Sun, Lucent, Nokia, ArborText, Blast Radius, BMC Software, US Department of Defense, and Innodata.

- Vendors have already stepped up to support DITA with products. These include Idiom (the sponsor of this paper), Arbortext, Blast Radius / Ixiasoft, X-Hive, Syntext, and Altova.

- IBM's unique role in launching and now supporting DITA is noteworthy. In addition to the initial technical development, IBM has continued to support DITA through a series of workshops and by providing an extremely useful DITA toolkit. The toolkit includes a comprehensive DITA Language Reference along with the latest DTDs, schemas, and example documents. The example documents and demonstrations are very useful as they help to illustrate some of the critical DITA concepts such as specialization.

We like DITA for all of these technical and standards-related reasons, but we also like DITA because of its promise for promoting best practices in information development—and in the development of content management systems and tools to support information development.

DITA does in fact provide a core DTD and schema, which define a "topic." In DITA terminology, the topic is the basic architectural unit. As the architects of the DITA framework have explained, "a topic is a unit of information that describes a single task, concept, or reference item."[3] Put more simply, a topic is "a chunk of information organized around a single subject."[4] Topic types already accounted for in the DITA architecture are "concept," "task," and "reference." These types make a lot of sense for technical documentation, especially complex product documentation where much of the material falls logically into one of these types of information.

Topic-oriented writing is also widely understood, practiced, and promulgated in the industry, beginning with product companies like IBM and extending through the work of consultants such as Ann Rockley, JoAnn Hackos, and the folks at Information Mapping.

---

[3] *Introduction to the Darwin Information Typing Architecture: Toward Portable Documentation*, by Don R. Day, Michael Priestley, and David R. Schell, IBM Corporation, March 1, 2001 (updated October 7, 2003), online at http://www-106.ibm.com/developerworks/xml/library/x-dita1/.

[4] *Frequently Asked Questions About the Darwin Information Typing Architecture*, Don R. Day, Michael Priestley, and Gretchen Hargis, IBM Corporation, March 1, 2001 (updated November 1, 2004), online at http://www-106.ibm.com/developerworks/xml/library/x-dita3/.

Structurally, a topic starts with a title element followed by a mix of text and images. Topics are organized into sections, and they can also nest. These core modules, the hierarchical depth, and the ability to nest the modules provides the kind of flexibility that would cover a broad range of technical documentation needs.

More importantly, DITA provides a mechanism for extending the core schema/DTD and framework through something called specialization. In short, specialization involves creating extensions, or deltas, to the existing DTDs or schemas. This way the specialized DTDs and schemas can be tracked more closely with existing tools such as style sheets and transformations.

Specialization is a useful concept, and it ties directly to another useful DITA concept, delivery contexts. Delivery contexts are things like the Help set, the printed form, and the Web site or portal; these are enabled in DITA through something called the "DITA map" referencing structure. As organizations use more and more XML for data storage, the XML needs to be tied to its delivery contexts through mechanisms such as stylesheets and transformations. Specialization provides a better, more overt, and more manageable mechanism for managing both the data—and the mechanisms for transforming the data.

DITA at Work: Information Builders

Information Builders, Inc. (http://www.informationbuilders.com) is a New York City-based developer of business reporting and intelligence software for enterprise applications. They have a broad product line, including Web-based, client-server, and mainframe applications. iWay, an IBI subsidiary, develops and markets adapters for integrating with enterprise applications such as those from SAP, PeopleSoft, and Oracle. The iWay business is an important part of IBI's growth, as it targets the market for service-oriented architectures (SOAs) and service-oriented software development.

The adapter business presents a challenge—and an opportunity—for the information developers at IBI. The adapters reuse core IBI software components across many operating systems and many target OEM platforms such as PeopleSoft and SAP. In documenting the adapters, there is a big opportunity for reuse. When you consider the number of content types that the information developers need to produce—print, online, help—you have an opportunity for reuse and repurposing together. Finally, overseas markets are important for IBI, so they translate the product information into 17 languages.

IBI has been a long-time consumer of translation services and technologies, and has been moving toward an XML-based single-source publishing solution over a number of years. In fact, product documentation is only one of the content types that IBI needs to translate. The others include text files, error messages, InstallShield files, embedded comments in libraries and executables, and more. As a result, IBI has built up significant translation memory over the years using a client-side translation tool called Catalyst from Alchemy Software Development.

As the product lines and documentation needs have continued to grow, IBI found that they had outgrown both the client-side translation tools and desktop publishing applications. IBI had been using FrameMaker to produce both the documentation and the Help files, but found that they needed manual workarounds to then run the source FrameMaker files through the translation tool. The workaround involved stripping out formatting tags, translating the material, and then re-entering the formatting tags to the produce the print and Help files.

Such processing is workable with a small- and perhaps even a medium-sized operation, but clearly doesn't scale. Such processing also requires the documents to be essentially unformatted and reformatted each time they are translated. Because documents change—they

are revised or modified for other OEM platforms—these steps were causing rework for the IBI information developers.

IBI is implementing a new system from Idiom that will manage the content as single-source modules, encoded in XML, using the standard DITA DTDs. IBI has begun to enjoy the following benefits from the new system:

- Elimination of all the formatting and reformatting in the existing desktop publishing workflow.
- The ability to manipulate the XML-encoded content, enabling them to parse, sort, and otherwise manipulate the content.
- The DITA topic-oriented markup reinforces the topic-oriented approach to writing. IBI sees this as a very sensible approach that aligns well to the kinds of content types that support the software, including the documentation and the Help sets.
- Significant improvements in turnaround time because they have eliminated the need to format and reformat the translated content. A typical translation effort takes 5-6 days instead of the previous 11 days.
- The single-source is maintained in XML. By concentrating on the reuse of the core English modules, IBI is also reducing the amount of editing that needs to happen on the translated side.

Interestingly, IBI made a decision to use the DITA markup structures "as is," instead of customizing the DTD/schema or developing their own. Moreover, in choosing the Idiom solution (WorldServer Global Electronic Publishing with WorldServer OpenTopic), IBI also made a decision to use Idiom publishing tools with little modification. In other words, IBI decided to mold their content and documentation to a more standard method of doing things, rather than to develop customized and proprietary content structures that would have allowed them to continue producing content as they had been. This is a decision worth noting.

It's also worth noting that IBI didn't make this decision in a vacuum. They're a customer-focused organization, and the documentation is a key part of the product. They conduct ongoing usability studies of their software and their content. As their documentation manager noted, "It is a short trip from the software to the documents. We hear a lot from our users." Beginning late in 2003, they began showing new DITA-based documents to their users; after getting some input from them, they adjust their design, and started talking with XML vendors.

About the Idiom Solution at IBI

After evaluating a number of solutions and vendors, IBI is implementing WorldServer Global Electronic Publishing from Idiom Technologies, Inc. WorldServer Global Electronic Publishing is new kind of platform that combines XML-based content management with globalization management. IBI also uses an add-on product from Idiom called WorldServer OpenTopic. OpenTopic provides a set of out-of-the-box DITA publishing tools that enable organizations to immediately tackle XML-based publishing and globalization using DITA. The publishing tools include processors for creating globalized print, Help, and HTML output automatically from the DITA-encoded modules.

Conclusions

There is great promise in XML-based content management and its ability to facilitate both reuse and repurposing. While SGML-based publishing proved to be highly effective and efficient in niche publishing situations, XML—with its broader adoption and much wider universe of tools—promises to bring these same efficiencies to a much wider base of users.

One of the barriers to entry in using XML has been complexity, and the need for organizations to do significant tool development, data conversion, and system integration prior to realizing the benefits of repurposing and reuse. One way that XML advocates have sought to accelerate adoption—and drive down complexity and cost—has been to promulgate industry vocabularies for XML. This has proven successful in vertical industries like automotive, aerospace, and telecommunications. The emergence of DITA as a horizontal, extensible architecture for information development is significant for the broader marketplace.

IBI's decision to adopt DITA more or less "as is" is instructive to other organizations. By doing this, IBI avoided the expensive upfront development that many organizations have faced. They also ensured their content is developed in a way that makes it readily interchangeable with other organizations—this will likely be a benefit to IBI as its business grows. Their decision-making process is noteworthy; they made the content development changes with ongoing input from their customers.

The true benefits and ROI for an organization such as IBI are closely connected to the volume of the content they produce, their need to reuse and repurpose the content, and the need to translate the content into so many different languages. It is precisely this combination of content management requirements and globalization management requirements that drives the IBI solution. Given the complexity of their requirements, they have chosen technology that closely meets their needs.

This nexus of globalization management and content management raises a significant question: do organizations need separate technologies for globalization management and XML-based content management? In other words, can a single platform manage both of these processes?

Up until now, organizations with these requirements faced a marketplace where they had essentially four choices:

1. Use a globalization management system on its own or side-by side with a content management system.

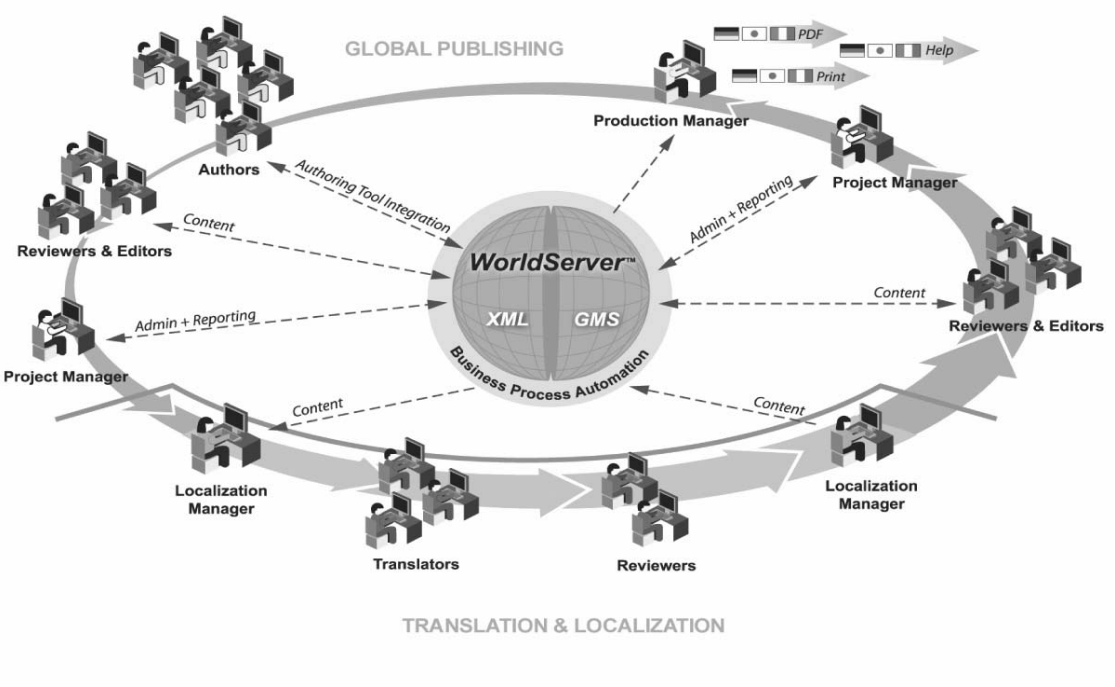
2. Use a content management system on its own (and typically outsource globalization) or side-by-side with a global management system.

3. Integrate these two systems together.

4. Develop a custom solution with best-of-breed components.

WorldServer Global Electronic Publishing system combines XML-based content management with globalization management. This new approach adds a new option for organizations that face a requirement for globalization and XML content management. We think this is potentially significant. Because there are now more options for globalization and content management—and because DITA provides the important necessary framework for best practices for globalization and content management—organizations can now think more strategically about how to manage these complex and expensive processes. We feel that solutions such as WorldServer Global Electronic Publishing warrant consideration.

Idiom WorldServer™ Global Electronic Publishing Solution



Globalization management plus XML content management enables you to create, manage, globalize, and produce information in one global content lifecycle[5].

[5] Illustration courtesy of Idiom.

Sponsoring Company Information

For more information, please contact:

IDIOM TECHNOLOGIES, INC.

NORTH AMERICAN HEADQUARTERS:
Idiom Technologies, Inc.
200 Fifth Avenue
Waltham, MA 02451 USA

Phone: +1 781.464.6000
Fax: +1 781.464.6100

EUROPEAN HEADQUARTERS:
Idiom Technologies
Quatro House
Frimley Road
Camberley
Surrey GU16 7ER UK

Phone: +44 01276 804488
Fax: +44 01276 804489

ON THE WEB:
www.idiominc.com