



SGML Refineries: Distilling “Docubases” for CD-ROM and Online Delivery

not long ago we were happy with a half-dozen TV channels; soon, we're told, the marvels of narrow-casting will give us 500 to choose from. We're also witnessing an explosion in CD-ROM storage capacity, as DVD takes us from 650MB CD-ROM to 4.7GB and more. Concurrent with exponentially increasing amounts of information, there is the additional requirement to publish in more media to ever-narrowing markets. Dealing with this deluge of information is difficult, but managing its creation and production is even more so.

If you are an Information Systems manager, you are likely no stranger to such document management concerns. You are also probably quite familiar with established standards such as Windows and SQL, and mainstream database suppliers such as Oracle, Sybase, and Watcom. CD-ROM drives, bundled with PCs, are common in your company and in your customers' sites. Your corporate publishing organization (with whom IS has growing relationships) has begun acquiring CD recorders. You have set up a corporate Web site and management wants more than just a feel-good home page. Your customers want you to deliver document data, masses of it. All this, and you need to generate revenue from the many investments in infrastructure, standards, and re-engineering that SGML requires.

The solution may be found in several existing tools for managing and deploying large collections of SGML objects, collectively called docubases.

ONE DOCUBASE, MANY PRODUCTS

In a traditional database, information is updated continuously and is used to satisfy many information needs via reports, queries, and various extracts. Data warehouses are being built to provide information corporate-wide to all authorized to use it.

By contrast, until SGML went mainstream, investments in documents were one-at-a-time. Very little sharing occurred, and every document was a handcrafted “one-off.” Document management consisted of whiteboard charts that tracked the

**The solution may be found in
several existing tools for
managing and deploying large
collections of SGML objects,
collectively called docubases.**

progress of items from authors through the production process. If document content was shared, it was often only because authors worked in adjacent cubicles. Inexpensive word processors and desktop publishing systems fueled the growth of documents but did little to manage them or promote integrated efforts. The explosive growth of the World Wide Web often prompted companies to handcraft a Web site, duplicating the same one-off habits employed with documents.

Soon, electronic publishing products like Adobe's Acrobat enabled easy creation of accurate digital copies of paper documents. Creating CD-ROM/online hybrids from existing desktop documents became easy, provided the paper layout was usable online and handcrafting hypertext links was acceptable.

Since the media and tools have become readily available, publishing departments have insisted on exploiting all of them, but none has solved the problem of how to manage the document products. Equally daunting is the question of how to derive multiple information products appropriate to each medium from what is often a common investment in document content.

A new genre of database-managed SGML application is emerging from the likes of Exoterica, OpenText (who has recently acquired Odesta), Electronic Book Technologies, and Inforium. Inforium—one of the earliest solution providers to offer robust implementations—has developed products that let users harness the power of relational databases to create, control, and manage SGML docbases. By using such products, experienced users of SGML can design a richly tagged SGML document collection and exploit the power of SQL relational technology to manage that data. Moreover, the built-in power of properly designed SGML will allow database managers to derive strategic competitive advantage from this document investment.

Inforium, for example, has developed LivePage, an integrated set of products that let you select your favorite SGML authoring tool, update document content online, extract various SGML products, and view the results according to styles you define (or plug into your own SGML viewers). Each document that a docbase management software program like LivePage stores in an SQL database is associated with a Document Type Definition (DTD). LivePage works with many DTDs.

Several initial questions must be asked about any document collection and manipulation tool. These questions generally consist of the following:

- Can it manage custom DTDs?
- Can it produce HTML extracts?
- Can it maintain hypertext links via SGML?
- Can it perform zoned full-text searches, restricting those searches to specific SGML elements?
- Does it provide inline graphics and multimedia OLE support?
- Does it support graphic and multimedia OLE objects?

In the case of Inforium's LivePage, the answer to each of these questions is yes. Furthermore, LivePage offers all these capabilities entirely within the comfortable confines of a document manager's own Oracle, Watcom, or Sybase database system. And LivePage also allows users to make drag-and-drop changes to Web pages, or interface to PowerBuilder, C, or Visual Basic applications.

Like most industrial-strength database systems, LivePage also lets users extract, modify, delete, or replace any section of a document online as other users are browsing or updating the same document. And of course, any number of users can update the docbase simultaneously; the system assures that no two users update the same section of any document at the same moment. One docbase can now yield different information products and deploy them on different media, all under the control of disciplined data management.

DELIVERING THE CONTENT

One advantage of choosing an SGML-capable docbase system like Inforium's LivePage is the delivery and publishing opportunities it creates. With LivePage WebMaster, users can manage integrated docbases rather than thousands of individual files. Link management is automatic, and the system provides additional navigation tools such as a dynamic table of contents.

Inforium also offers a LivePage CD Publisher that is optimized to help MIS professionals deliver their docbase content to this medium as well. Inforium offers a special version of its SGML browser so that content can be any SGML content model you choose. You can even secure the content by serializing

a key to read that docbase or collections of docbases.

By using Inforium's Browser, docbase system operators can extract and deliver SGML document content on CD-ROM in whichever view they choose. You also get limited full-text searching by SGML element. Even users who already chosen other arbitrary SGML viewers such as Panorama Pro or Electronic Book Technologies products can benefit from using docbase management tools; if you extract SGML from the docbase, it is valid, universal SGML, so you can use it with any arbitrary SGML viewers.

As Web and CD-ROM content becomes more complex, Inforium is preparing an update release to WebMaster that supports Java applets, which, although not stored in the database, still include pointers to the applets.

EARLY ADOPTERS AND OTHER PROVIDERS

Large corporations are already placing their bets on document management systems, both from traditional publishing vendors such as Interleaf and Xyvision and from object-oriented offerings such as Documentum. These products do not offer the degree of SGML integration found in Inforium's LivePage, but they do offer other critical facilities such as "check-in" and "check-out" of documents and workflow management. Moreover, many of these document management offerings are standardizing on a few text-retrieval vendors, usually Fulcrum and Verity.

As of April 1996, Inforium was investigating the addition of workflow, but at present, this is only a gleam in the company's eyes. Moreover, no integration is planned with more mainstream document management products, such as Documentum, that do offer what LivePage lacks.

As in the rest of life, there are no 100 percent complete tools and no perfect solutions to docbase management. However, management options for MIS professionals and other database jockeys are clearly multiplying.

Robert J. Boeri and Martin Hensel are regular columnists for INFORMATION INSIDER. Boeri is Advanced Systems Specialist in the Information Services Division of Factory Mutual Engineering in Norwood, Massachusetts. Hensel is founder of Martin Hensel Corporation, a consulting firm that builds SGML-based editorial and production systems for publishers, corporations, interactive services, and composers. ■