

ELECTRONIC PUBLISHING

A PENNWELL PUBLICATION

VOL. 25, NO. 6

JUNE 2001



Photo courtesy of NUR America Inc. The 150x150-foot graphic was printed on a NUR Blueboard superwide printing system by Mega Art.

Features

20 **Perspective: Jim Thrush**

The president of Xitron reflects on the company's heritage in the CEPS days and reveals his plans for the future.

24 **Wide-format opportunities**

Although it can be a battle just to stay in the game, many industry players are delving into unique, uncharted territories and having a ball.
BY COURTNEY E. HOWARD

36 **Inside color management**

The latest on what color management is, how it works, and what tools are available. BY DAVID BROUDY

44 **In the Field**

- Digital proofing saves printer hours in production
- One service provider finds its niche in textile printing
- Printer helps company speed marketing material to customers

Departments & Columns

4 EDITORIAL Direct-to-trash

6 TIMELINE A colorful era

16 WHAT'S NEW BY IRA A. GOLD The shows of spring

18 FUTURE IMAGE BY ALEXIS GERARD Imaging without wires, Part III

42 WORKFLOW BY DUFF JOHNSON Capturing legacy content

48 MAKEREADY BY DAN MARGULIS Breakfast of champions

68 FRANK TALK BY FRANK J. ROMANO All the news that's fit to watch

Products

11 TECH NEWS Seybold Boston 2001

30 RESOURCE GUIDE Wide and superwide printers

54 REVIEWS CreoScitex oXYgen scanning software.
Enfocus Software's Certify PDF 1.0

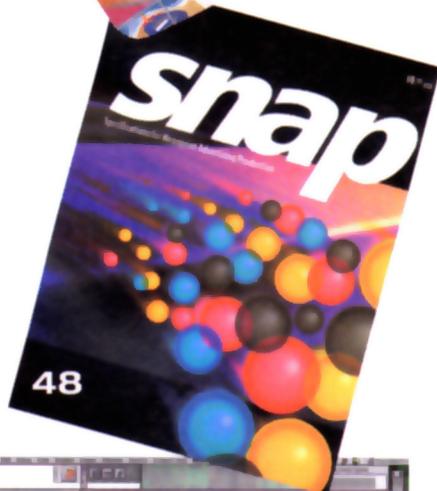
57 NEW PRODUCTS Seybold Boston 2001



24



36



48



54

Capturing legacy content: a primer for action

Just because your older content isn't digital doesn't mean it's worthless.

by Duff Johnson

At the recent Seybold 2001 show in Boston, repurposing—a fancy term for legacy conversion—was all the rage. The release of the latest version of Adobe Systems' global electronic document application, Acrobat 5.0, had the Hynes Center buzzing. During his keynote address, Adobe Systems' CEO Bruce Chizen unveiled the future of content Adobe-style: anytime, anywhere, on anything.

Great. But how are we going to get there? Until our content is finally corralled, authored, and maintained in the necessary technical harmony, we must deal with the archives of our existing work. Yesterday's content (the stuff more than, say, a couple of years old) now exists as purely peripheral value on dusty shelves or is languishing on unusable diskettes. Does this material have value in the new world of e-publishing? What do you think? Are you willing to bet against it?

Uses of legacy content

From a simple raster image to complete conversion to dense XML or fancy PDF, the right conversion blends the near-term need for a Return On Investment (ROI) with an eye for an investment in the future. The following are just a few examples of such conversion projects:

- Allen Press converted hundreds of thousands of pages of academic journals to a blend of SGML headers, multiresolution PDF/image files, and raw OCR-captured text for an online search application. The collection serves its purpose now and, taken



together, is itself a platform for further repurposing.

- American Flyers converted printed pilot training documents into a new authoring system, dramatically reducing the time required to produce, edit, and revise new editions.

PDF and the various Markup Languages can, should, and will co-exist as distinctive modes of expression.

Along the way, the company generated a specialized and unique clip-art library, ready for use in its future training materials and publications.

- Olympus America converted thousands of old technical, installation, and service manuals to PDF for use on a support intranet. Along the way, the company improved response time in serving its customers, institutionalized its knowledge base, and

improved research and reference efficiency for designers, engineers, and technicians. The next step—conversion to XML.

Preparation

The conversion from paper to electronic format starts with document preparation—a deceptively mundane process containing many of the secrets of a successful conversion. In this phase, you'll find the pages that defy the procedures you'd like to use. Is each article to be a separate file? Are there any pages that shouldn't be included? You should have a clear vision of at least the initial, baseline repurposed product to properly prepare your documents for conversion.

Legacy document scanning is typically bi-tonal (black and white), at a minimum resolution of 300dpi. For bi-tonal scanning, many libraries prefer 600dpi for archive-grade images and enhanced readability, especially

with small characters. Scanning in color tends to increase the cost and complexity of conversion projects but, when well considered and executed, always yields a treasure-trove of widely deployable content.

Put careful thought into how you'll want to locate and group your files. Well conceived and executed indexing means the difference between finding and losing your documents.

Page-image formats

The basic reclamation effort is scanning, and scanning alone reaps significant rewards for the content repurposer in the form of simple TIFF and JPEG files. Simple page-images can also serve as goldmines of artwork and as an idea or reference bank, or can provide legal or documentation functions.

Most popular image file types are easily converted to PDF for archive. Why bother? PDF provides a familiar interface, powerful tools, and a global electronic document standard, and with Acrobat 5.0, images are retrievable from the PDF in a wide variety of formats. All the interactive functionality—bookmarks, form fields, links, and so on—are fully available in PDF/image files, as they are on all PDF file types.

Typical applications for page-image based repurposing include:

- Searchable-image PDF files with full-text search
- Research/reference materials where guaranteed fidelity to the original page image is required
- Electronic libraries
- Staging source data for further conversion, cleanup, or tagging
- Original document reprint

Full text and layout conversion

Converting directly to RTF or SGML makes sense for editing or complete reauthoring of the original. Unlike PDF files, all components of RTF files, including text, graphics, and layout, are ready for editing anew, although duplicating the original layout is challenging. RTF files are an excellent starting point for XML tagging, or development in XML-aware or other DTP applications, while SGML addresses document structure with a high degree of refinement, and is highly redeployable.

Allowing for fonts and images, PDF/Formatted Text and Graphics (FTG) files are essentially competitive with SGML and HTML in terms of file size per page. FTG files are effective when deploying high-value original

pages in low-bandwidth environments, or where the structure accessibility features of PDF are required.

Circumstances implying PDF/FTG-based repurposing:

- Delivery of original page-layout is essential or highly valued.
- Limited bandwidth considerations are paramount, such as enabling modem-based or international users, or for facilitating the downloading of longer documents for offline use.
- Applications requiring maximum on-screen appearance and print quality.
- Applications requiring maximum flexibility and accessibility.
- Advanced source data staging for tagging or further conversion.

Content tagging and XML

As of this writing, content-driven tagging of XML documents is a "day forward" process, with few tools available to production environments engaged in processing legacy content. Creating rich, densely tagged XML from unstructured source documents, especially paper originals, is still a specialist's task. PDF conversion is generally significantly less expensive and serves as a staging area and reference point for higher-order conversions.

PDF and the various Markup Languages can, should, and will co-exist as distinctive modes of expression for documents, whether original or repurposed. Indeed, the tagging and partial XML-enabling in the new PDF 1.4 specification are the first steps down the last road to the full power of the "Portable Document Format" concept. The electronic page concept currently embodied in PDF will continue to deliver value as both the original reference and live content source for as long as humans buy printers.

Circumstances indicating Markup Language-based repurposing:

- Documents are dynamic in nature.
- A high degree of interapplication operability is required.
- The application deploys with maximum flexibility and accessibility.

As with any project, your choice of

Tips to get you started

Avoid the need for a legacy conversion altogether

- Author in XML-aware applications.
- At a minimum, archive your recent and current design files as top-quality PDF files.

Are you feeling lucky?

- At least half of publishing companies have no effective design file archiving system.
- At least half of those who think they have a functional design file archive don't.

Spread the cost

- Perform the conversion over time to soften the impact on cash flow. For instance, convert one year of back issues every business quarter. This process also creates a natural stream of editorial material, perfect for content-hungry websites.
- Account in many places: Converted legacy content is an investment, a real, tangible asset, not a consumable (such as a printed magazine or a book). The freedom to electronically deploy an entire publication history can generate revenue forever and depreciates pretty slowly. Marketing, sales, and production budgets all stand to benefit from effective content reclamation.—DJ

legacy conversion method will tend to be driven by short-term needs and budgets. Happily, almost every conversion operation—page-image, text and layout conversion, and finally full tagging—is in some way additive to the overall direction of document repurposing.

Duff Johnson, founder and CEO of Document Solutions Inc., is an industry veteran, commentator, and electronic document technology expert.