**• What Societies and Associations Are Doing with Their Back Issues •**
**• How to Perform a Legacy Conversion with Allen Press •**

*The Newsletter for Journal Publishers*

JP

*Allen Press, Inc.*
*Year 2005 No. 2*

# Legacy Content Conversion: Generating citations, revenue and goodwill from your publication history

*"Relevancy is driven by content, not publication date"*

By Duff Johnson
Founder, President, Principal Consultant, Document Solutions
Contributing Editor, Business and Government, Planet PDF

*This issue of **JP** explains, in a detailed but non-technical way, the ins and outs of assessing, planning, and executing the conversion of a publication's back-catalogue to electronic format. Written by a veteran service provider and consultant, and including detailed case studies of legacy content usage published for the first time anywhere, this article delivers information designed to provoke reflection on the considerations, opportunities, and pitfalls of converting legacy content for use online or on disk. The services outlined on this newsletter are available from Allen Press, Inc.*

JP, THE NEWSLETTER FOR JOURNAL PUBLISHERS, is published four or more times a year by Allen Press, Inc. JP is dedicated to providing comprehensive management and technological information and guidance to journal publishers.

**Allen Press, Inc.**
810 E. 10th Street
Lawrence, KS 66044
Phone: 785-843-1234 or 800-627-0326
Fax: 785-843-1226
www.allenpress.com
Rand Allen, Chief Executive Officer
e-mail: rallen@allenpress.com

**Printing Division**
Guy Dresser, Vice President, ext. 123
e-mail: gdresser@allenpress.com

**Publishing Division**
*Alliance Communications Group*
Theresa Pickel, Director, ext. 263
e-mail: tpickel@acgpublishing.com
www.acgpublishing.com

**Web Publishing Division**
Ted Freeman, Director, ext. 170
e-mail: tfreeman@allenpress.com

**Association Management Division**
*Allen Marketing and Management*
Frank Cherry, Director, ext. 255
e-mail: fcherry@allenpress.com

**SUBSCRIPTIONS:**
Newsletter subscriptions are available at no charge to qualified subscribers for a limited time.

**SUBMISSIONS:**
Article queries, press releases, news tips, and comments are welcome. Please mail, fax, or e-mail material to the editor.

**EDITOR:**
Gene Kean, Executive to the CEO
e-mail: gkean@allenpress.com

**ASSOCIATE EDITOR:**
Martha Chapin
e-mail: mchapin@allenpress.com



## Back Issues Online Can Be a Valuable Resource!

# Societies are Putting Back Issues Online for Current and Future Researchers, and You Can Too!

Dear Society Directors, Editors, and Managers in STM Publishing:

The two most valuable resources for any scientific or medical society are its members and its journal. Nothing happens without members, of course, and the journals, both current and past issues, represent the accumulated knowledge of the discipline.

Many societies already place current journal content online. Historical content, however, often receives less attention. While legacy content is often of great value to researchers and practitioners worldwide, if users cannot get easy access to it, then the journal cannot offer its full potential to current and future researchers.

This newsletter describes some of the ways Allen Press has organized and simplified the placement of recent and back issues online. We will be happy to provide these services for your society or organization.

Providing legacy content conversion is only one of the many services that we provide for our clients. Four divisions of Allen Press provide more than 25 different services in printing, publishing, and association management. We are the only U.S. printer that provides ALL the services associations and publishers need.

Allen Press is more than just the leader in printing quality for scientific, medical, technical, and special interest journals and magazines. More than 400 journals and magazines in a variety of academic and trade fields have entrusted Allen Press with their printing and publishing support services. It is our commitment to the STM communities that has enabled our company to be a world leader in scientific, technical, and medical printing and publishing. In 2004, our client publications won the **TOP FOUR GOLD INK AWARDS** (the most prestigious and challenging print production competition in the business) for journal web printing.

Allen Press serves clients coast to coast. More than 150 national association leaders and editors attended our April, 2005 Emerging Trends Seminar at the National Press Club in Washington, D.C. This annual seminar provides state-of-the-art information on new and emerging challenges facing the scholarly community.

We invite you to contact us regarding your interest in any Allen Press printing or publishing services.

Sincerely,

Rand Allen

RAND ALLEN
Chief Executive Officer

## Introduction

Many scholarly and scientific societies and associations have discovered that digitizing and publishing their journals' legacy data online or on CD-ROM not only generates goodwill with their members and constituents, but also increases the visibility of their content by connecting it to current research, raising its impact factor, and increasing revenue. For too many STM publishers, however, journal content more than four or five years old now exists solely on dusty shelves, languishes on unusable diskettes, or worst of all, lies unread in the printer's warehouse, at considerable expense.

How relevant is your legacy research? In fields such as bio-medicine and particle physics, the very latest research is unquestionably the most relevant. In many other fields, such as geophysics, integrative biology, and social science, however, research from the 90s, 80s, 70s, 60s, and even earlier remains relevant and continues to be cited.

Regardless of the scientific field, the half-life of research not found online declines at a much faster rate simply as a result of unavailability. That is because for most researchers these days, particularly the younger generation, research that cannot be found online simply does not exist. Imagine you are a young researcher. You can barely remember the time when Google was unavailable. You seek out the biggest and the best collections of online content. Relevancy to you is determined by content, not publication date.

In this article, I describe the benefits to publishers of digitizing legacy content and offer case studies of three organizations who have done so and profited. I discuss and detail the issues, considerations, and background information that publishers will need to understand their options, including action plans to get the process started.

## Top Five Reasons to Digitize Back Issues

From a simple scanned image to searchable PDFs or richly tagged XML, the right conversion process blends the need for a near-term return on the investment with an eye to the future. There are many approaches to freeing a publication's history from the bonds of paper and putting it to work on Web sites or in disk-based collections.

**1. More online presence, more usage.** Placing more content online increases the volume of search words and phrases available to search engines. More searches find more hits, so more researchers download more articles for reference, use, and citation.

**2. Increases citations.** It is becoming clear that journal usage and the citations resulting from it increase in part as a simple function of availability. Previous, current, and future readers and authors are predisposed to an interest in the history of their publication of choice. In many disciplines, highly specific studies relevant to current research are often found in older material.

When older journal issues go online, they should ideally become available to CrossRef and other citation linking systems as "targets" for links from later issues and other journals, increasing readership as researchers "drill down" into references from current articles to older material.

A subtler benefit arises from renewed awareness of existing research. When scientists can easily check entire publication histories for relevant content, they can quickly determine if similar research has ever occurred, thus saving countless hours in locating, replicating, or describing work already done by others. This effect is unaddressed in a journal's Impact Factor, but is an important contribution nonetheless, in many cases helping refine current ideas or making costly additional research unnecessary.

**Since they began publishing online in 1996, Annual Reviews saw an increase of 21% in the aggregated Impact Factors of 26 of their titles, spanning all disciplines.**

Providing journal legacy content online or on disk makes it far easier for new researchers to mine content and citations deep within the publication history. While material from the 1950s or 1960s might be considered of little relevance by some, evidence suggests that search engine users clearly want access to that information.

**In the 21st century, the only thing worse than not showing up as a search-engine hit is not being available on the Web in the first place.**

**3. Provides a complete and permanent solution for back-issue requests.** Imaging legacy content to archive-quality PDF files pro-

*Continued on page 4*

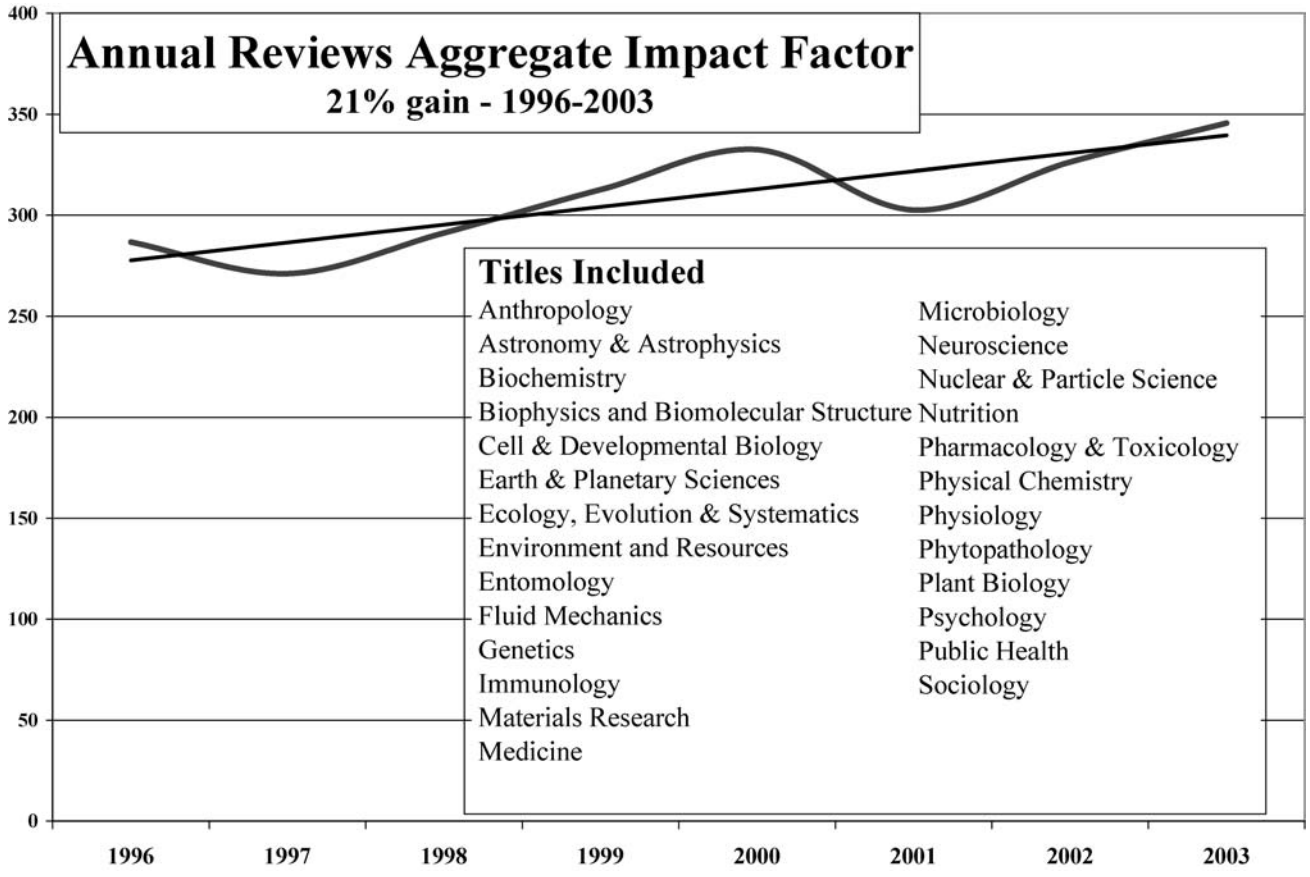
vides a permanent solution to the problem—and opportunity—of back issue requests. Digitized articles and issues in PDF form are permanently available, and may be delivered to end-users via Web site or e-mail for authorized reprints at the user's own expense. Legacy content collections may be used in disk products, delivering some or all back-issues as a valuable collection and reference.

**4. Be of service to the membership.** Converting the publication history is always appreciated by association members, who tend to regard this activity as in the best interests of the institution. In reviving and promoting legacy content, senior members may be honored, anniversaries celebrated, and the community will appreciate the sense that the entirety of the intellectual effort associated with the publication will be preserved and available forever.

**5. Facilitate a variety of distribution options.** Electronic publication histories may be deployed in various applications that may be of real value to members. Many scientists and researchers appreciate the offline access benefits of CD or DVD-ROMs. These formats are a valuable way to offer "the last ten years" collections, complete historical archives, subscription premiums, celebrate publication anniversaries, and market the publication to new readers. Supplementary materials, such as very large image files, movies, databases, and other ancillary materials may also be distributed via disk, often in conjunction with current or legacy content.

Some associations sell disk collections simply to cover the cost of conversion, then post the content online alongside current issues via "Legacy Content" subscriptions. Others simply add value to the current subscription model by providing legacy content access without additional charge.

Others sell individual articles or full issues as electronic reprints, or use their PDFs as an easy, low-cost replacement to existing back-issue services, shifting the cost of "printing" each reprint directly to the end-user.



# What Other Societies Are Doing: 3 Case Studies

### American Society of Agronomy, Crop Science Society of America, and the Soil Science Society of America

Together, these three societies have published a wealth of comprehensive agronomy research spanning almost a century.

The tri-societies were out of copies for many of their frequently-requested back issues. Responding to subscribers' requests to make legacy content available for general access, the societies contracted to convert their entire publication history—over 180,000 pages including over 8,000 images. To offset the conversion costs, they produced four sets of CD-ROM collections, indexed by title, author and date. Each disk may be used in a stand-alone fashion, or installed to a hard-drive for full-text search across the entire collection at once.

At the end of the conversion and disk-development process, the societies had created valuable new products to add to their bookstore offerings. Of course, once converted, the same material became available for deployment online, as well as on disk.

### Agronomy Journal/JNRLSE CD-ROM Collection (7 CD Set)

- *Agronomy Journal* (1907–2001)
- *Journal of Production Agriculture*, (1988–1999)
- *Journal of Natural Resources and Environmental Education* (1972–2001)

Personal/Single-user - $225,
Institutional/Multiple-user - $800

### Crop Science CD-ROM Collection (6 CD set)

- *Crop Science* (1961–2001)

Personal/Single-user - $225,
Institutional/Multiple-user - $800

### Journal of Environmental Quality CD-ROM Collection (3 CD set)

- *Journal of Environment Quality* (1972–2001)

Personal/Single-user - $225,
Institutional/Multiple-user - $800

## Seven Key Indicators Favoring Legacy Conversion

- Historical content is manifestly useful to current researchers.
- The membership is asking for increased access to historical content.
- Libraries would value the expansion of your current online subscription offering to include "Legacy" content.
- You think the membership would enjoy access to disks including current, legacy, or ancillary content.
- An organizational anniversary or celebration inspires the urge to preserve the past.
- You would like to reduce or eliminate your back-issue stock and begin providing back issues past a certain date in electronic-only form.
- You have run out of back issues.

### SSSAJ CD-ROM Collection (6 CD set)

- *Soil Science Society of America Journal* (1936–2000)
- *American Soil Survey Association Bulletins* (1921–1936)

Personal/Single-user - $225,
Institutional/Multiple-user - $800

Since offering the CD-ROM collections in early 2004, the Society cleared their conversion and disk development costs **in less than 12 months**.

**Annual Reviews**

Many publishers believe that while there might be continuing interest in articles going back a decade or two, few, if any, users would find *Bio-Medical Reviews* from before the 1970s of sufficient interest to justify the investment in a wholesale conversion of the legacy content. They could be wrong.

*Annual Reviews*, of Palo Alto, California, has published comprehensive collections of critical reviews written by leading scientists in 29 distinct disciplines, with many titles dating to 1950 and earlier. A nonprofit publisher, *Annual Reviews'* mission is to provide the worldwide scientific community with a synthesis of primary research literature. Their publications are among the most highly cited in science, ranking within the top ten publications for their respective disciplines.

In 2002, *Annual Reviews* decided to move the entire corpus of their publication history online, a collection spanning 70 years, 475,000 pages, and 5,400 color and grayscale images, as a new subscription option for online subscribers.

Using the PDF/MultiResolution conversion process to maximize image quality while keeping 30-page chapters to the smallest possible file size, the project was largely completed within three months. Once imaged, pages were checked extensively for quality, corrected for orientation, and OCRed. As the bulk of the legacy content was processed, an inventory database generated by the contractor assisted *Annual Reviews* staff in identifying and locating missing volumes. An *Annual Reviews* logo and hyperlink to the *Annual Reviews* Web site was "hard-wired" to each PDF page before delivery.

In six months, all data processing was complete, and the entire *Annual Reviews* publication history was ready to go online.

Based on filtered server logs, *Annual Reviews* was able to determine that during 2004, their legacy project resulted in over 741,000 downloads of legacy (pre 1996) content. In fact, adding the legacy material online increased total PDF downloads for all *Annual Reviews* chapters by 38%.

The following charts show the effect on downloads of adding legacy content subscriptions to 64% of *Annual Reviews* online subscribers in 2004. Had all online subscribers possessed legacy subscriptions, total PDF downloads of legacy material would have been even greater. The legacy product has been an immediate success for the publisher and its customers. Priced modestly, access to the legacy material was quickly subscribed.

**American Meteorological Society**

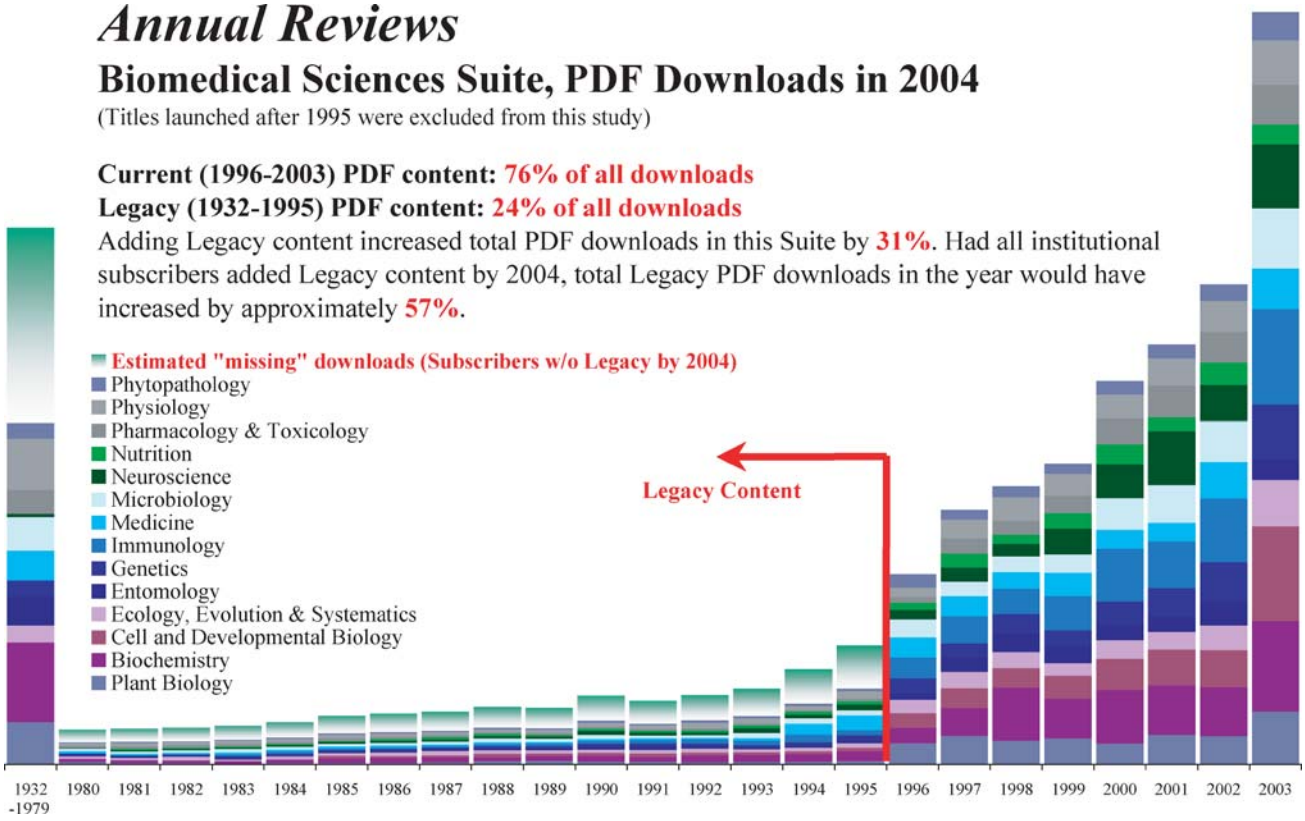
AMS Director of Publications, Ken Heideman, knew that the Society's legacy content was valuable to the membership. According to Heideman, the half-life of citations in atmospheric science journals is about 10 years. This means that the number of citations listed in 2005 AMS journals to material published before 1995 will be roughly equal to the number of citations in 2005 to content published between 1995 and 2004, and is an indication of the enduring relevance of many of the articles contained in the legacy collection.

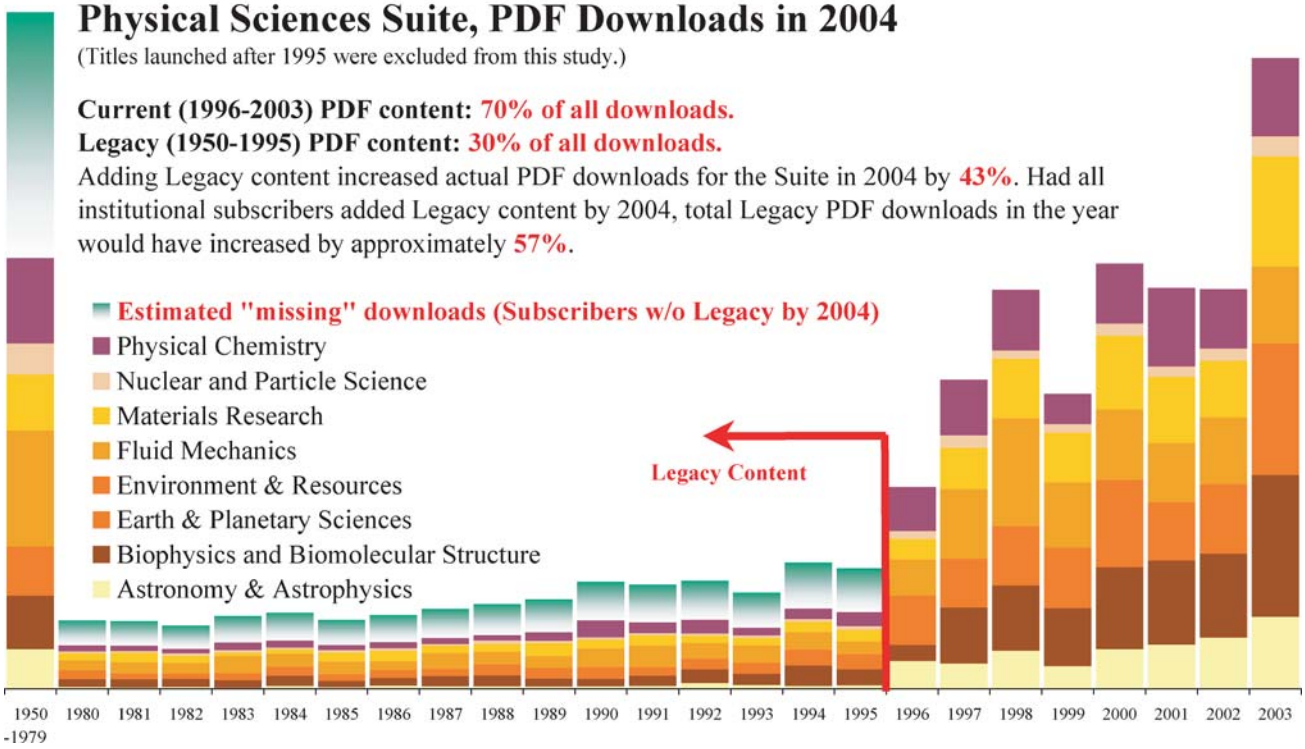
Even earlier articles gain a significant number of new citations as each year passes, representing savings in time and research budgets. It was obvious to AMS Journals that placing legacy content online was an excellent way to ease the research burden for tens of thousands of atmospheric scientists.

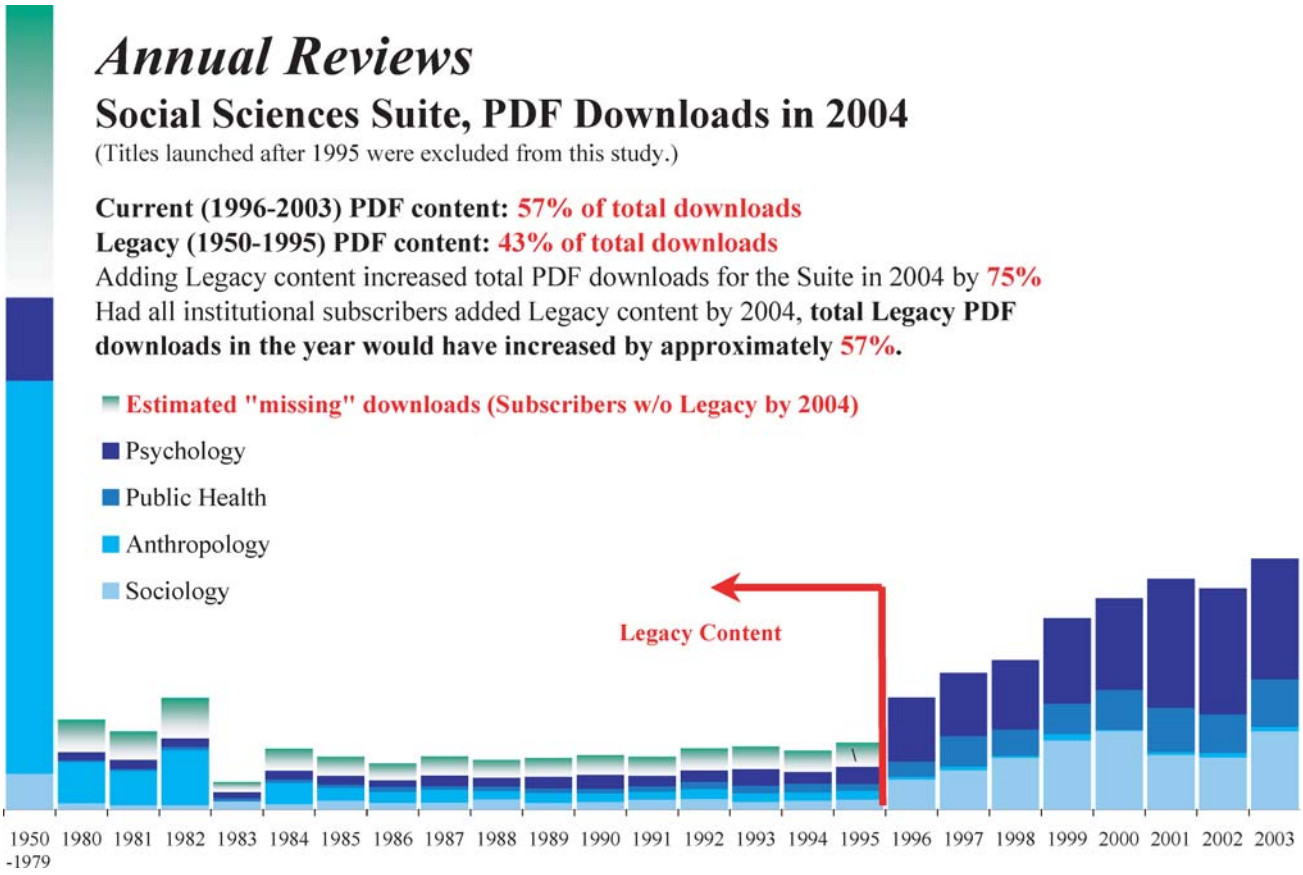
When the AMS decided to begin their legacy conversion project in 1999, both the Society membership and subscribing institutions were enthusiastic. AMS elected to support the costs of legacy content conversion by pricing access to the entire legacy collection, including articles back to the 19th century, at a one-time cost equal to the combined total price of subscriptions in 1994, 1995 and 1996. Even though AMS Publications declared their intention to make the legacy content available at no extra charge to all subscribers by January 2006, the Society recovered almost the entire conversion cost from subscribers who elected to pay the price for "early" access to the legacy content.

The journals of the American Meteorological Society are richly illustrated, with critical information contained in subtle details of the included images. The initial legacy conversion included the complete canon of eight journals—over 260,000 pages and 16,000 images—covering the period from 1990 through 1998.

*Continued on page 8*


Annual Reviews
Biomedical Sciences Suite, PDF Downloads in 2004
(Titles launched after 1995 were excluded from this study)
Current (1996-2003) PDF content: 76% of all downloads
Legacy (1932-1995) PDF content: 24% of all downloads
Adding Legacy content increased total PDF downloads in this Suite by 31%. Had all institutional subscribers added Legacy content by 2004, total Legacy PDF downloads in the year would have increased by approximately 57%.
Estimated "missing" downloads (Subscribers w/o Legacy by 2004)
Phytopathology
Physiology
Pharmacology & Toxicology
Nutrition
Neuroscience
Microbiology
Medicine
Immunology
Genetics
Entomology
Ecology, Evolution & Systematics
Cell and Developmental Biology
Biochemistry
Plant Biology
Legacy Content
1932 -1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003


Annual Reviews
Physical Sciences Suite, PDF Downloads in 2004
(Titles launched after 1995 were excluded from this study.)
Current (1996-2003) PDF content: 70% of all downloads.
Legacy (1950-1995) PDF content: 30% of all downloads.
Adding Legacy content increased actual PDF downloads for the Suite in 2004 by 43%. Had all institutional subscribers added Legacy content by 2004, total Legacy PDF downloads in the year would have increased by approximately 57%.
Estimated "missing" downloads (Subscribers w/o Legacy by 2004)
Physical Chemistry
Nuclear and Particle Science
Materials Research
Fluid Mechanics
Environment & Resources
Earth & Planetary Sciences
Biophysics and Biomolecular Structure
Astronomy & Astrophysics
Legacy Content
1950 -1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003




Destined for Allen Press's Web servers, the hosts of AMS content online, the end-product of the initial conversion effort was a collection of 20,000 articles converted from paper back-issues to MultiResolution PDF/Image files, SGML header files, and OCRed text. Follow-up conversion work has reached further back in time, bringing the total AMS publications legacy volume to over 33,000 articles dating back to 1873.

The response from the membership has been exceptionally positive. As far as the Society's journals are concerned, Heideman says, the legacy conversion effort is the, "centerpiece of who we are and what we have to offer." The AMS legacy project is hosted by Allen Press and may be seen at http://ams.allenpress.com.

**Legacy Conversion: Start with PDF**

PDF, HTML, SGML, XML, JPEG, TIFF …. just tell me what I need!

Before you approach a scanner, you will need to have thought about the information to be captured in addition to article text. Most publishers find it essential to capture metadata such as ISSN number, issue, and page number, as well as key article metadata (title, author), and often other metadata as well, such as author's affiliation and keywords for use in SGML headers, PDF index fields, or for other indexing systems. In many cases, the abstract from each article is also captured, whether to text, SGML or XML, for use in enriching the information available through search engine results.

While there are many format options in the legacy conversion process, here we will cover the essentials of almost any legacy conversion project.

**No matter what else you do, you will want top-quality PDF files.**

Since Adobe Systems' original innovation in the early 1990s, PDF has become the standard

*Continued on page 9*

electronic format for viewing, printing and retention of final-form documents. As demonstrated in Annual Reviews' download logs (see bar graph on next page), PDF remains a clear preference for most users. PDF is popular due to its total reliability, fidelity to the original printed page, and for a variety of subtle features that are less well known but equally felt. Today, new journal issues are customarily loaded to association Web sites in PDF, often alongside HTML (often SGML or XML generated) versions of the article. However, users still clearly prefer PDF over HTML or XML.

**The key advantages of PDF conversion include:**

- PDF files are freely readable with the free Adobe Reader, and freely distributable.
- PDF prints just like the original page, without fail.
- PDF files deliver a consistent, predictable, and familiar presentation throughout the publication history.
- Searchable Image PDFs offer full-text search hits highlighted right on the page.
- PDF/MultiResolution gives the best-quality option for pages combining black and white with color or halftone content, and allows high-resolution images to be captured separately for use in HTML or XML.
- PDF files can contain comprehensive XML metadata at the document level via Adobe's XMP architecture, facilitating interoperability standards such as the Dublin Core. A PDF file may thereby serve as a "Rosetta Stone" —the Reference Document—for the content it contains.
- A single PDF file may contain any package of content, articles, whole journal issues, sound or movie files, and other attachments.
- PDF files present lower up-front conversion costs than XML.
- As the de-facto electronic document standard, PDF serves as an investment in, and staging platform for, an eventual piecemeal or complete XML conversion.

Some archives have required the use of uncompressed TIFF files as an ultimate long-term storage format. However, this requirement is generally restricted to a very limited set of high-value historical documents. Although Adobe Systems, a for-profit company, makes the ubiquitous (and free) Reader, PDF is an open file-format specification. The ability to create and view PDF files is not restricted to Adobe's products. PDF files can (and should) be properly future-proofed via adherence to the PDF/A (Archive) standard. Most legacy conversion projects deliver PDF of a type that is inherently compliant with PDF/A.

### Content Tagging and Conversion to XML/SGML

For electronic content, users rely on indexes, full-text search engines, and existing citations to provide a guide to the goldmine, and they are happy to download a PDF file when they find one that suits their interest.

While full text XML or SGML is usually "overkill" for most legacy conversion needs, we nonetheless strongly recommend that all imaging

## PDF/Archive

**PDF**

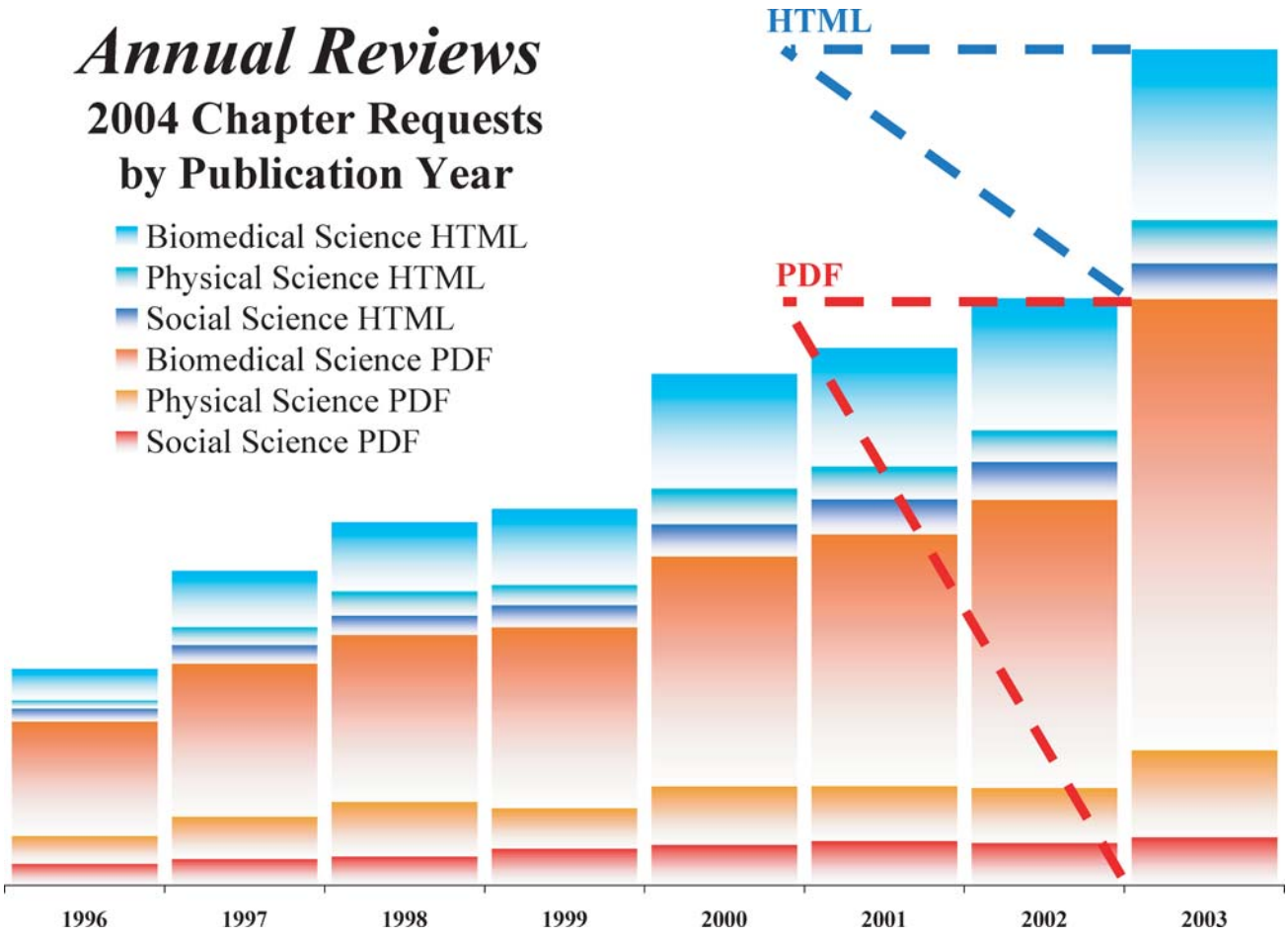
- Non-archival format
- Text, raster images, vector graphics, music, video, etc.
- Encryption and executable scripts permitted
- No type fonts included

**PDF-A**

- Archival format
- Text, raster images and vector graphics only
- Future ISO standard
- Encryption and executable scripts not permitted
- Type fonts included

More information about PDF/A available at http://www.aiim.org/standards.asp?ID=25013

*Continued on page 10*




and conversion work should be of sufficient quality to allow future conversion to XML or SGML. Technology moves on, and what we consider to be technically impossible today (such as the reliable automated conversion of scanned images to well-formed, quality-controlled and richly-tagged XML), may in ten or twenty years be a different story. The work performed to bring legacy content to present-day users should be conducted very much with an eye for the possibilities of the future.

Converting from paper to XML, SGML, TeX, or other specialized tagging formats requires complete proofreading of the original content and some manually intense tagging work. Full scale tagging is rarely cost-effective in legacy conversions of scholarly journals. Typically, article-level metadata is deployed within the PDF document information fields and XMP metadata, captured to SGML or XML header files, or otherwise made deliverable to indexing systems such as CrossRef.

Conversion of the full text into SGML or XML is significantly more expensive than conversion to PDF and rarely does not add value to the on-line experience in proportion to the added cost. Creating rich, densely tagged XML from unstructured source documents, especially paper originals, is still a specialized task. Basic PDF conversion is generally necessary, in any event, as a staging and reference point for XML conversions.

Circumstances indicating full text SGML/XML as an initial goal of conversion:

- A high degree of interoperability is required with other documents or datasets.
- The content is dynamic in nature or would clearly benefit in material ways from the advantages of XML tagging (large volumes of tabular data, for example).

# The Imaging Process

## Preparation

It may seem obvious that every article should be a separate file—but depending on the way your layout and content have changed over time, there are likely to be sections that challenge seemingly easy conversion approaches.

You will need to consider the following:

- The "front matter,"—the table of contents, advertisements, announcements, covers, and so on. Should each subsection be a separate PDF? Should some pages be dropped?
- Book reviews and letters to the editor. Are they citable? If so, and if they often run together over a few pages, you might want to break them out separately.
- Printed corrections and/or retractions. These should be identified and indexed. The ASA, as we have seen, decided to append the correction as an extra page with the PDF file of the article being corrected.
- In addition to PDFs of each article and of the front-matter and other content, perhaps you will want a cover-to-cover PDF of each issue as a self-contained "electronic back-issue" to be e-mailed on request, or placed on a Web site. Look hard at your publication and make sure all of your needs get on the table up-front.

## The Basic Black and White Scan

In legacy conversions, there is a great divide between journals with images and journals without, and the mathematics journals usually get off the easiest! In general, color and grayscale images and charts tend to increase the conversion cost because special handling is required to preserve the page content. As any printer will tell you, printing images for scholarly journals requires subtle skills to ensure fine detail is preserved. It is vital to treat these images better than is customary for conventional Web content.

First, consider the conventional black and white page. Legacy scanning (or imaging, to use the industry term) is typically performed using bi-tonal (black and white) scanner settings at a minimum resolution of 300 dpi, or dots-per-inch. (You may want to consult the glossary at the end of this article for explanations of the technical terms to follow.)

While it is more expensive, due to slower scanning speeds, larger initial file sizes, and the use of specialized quality-control workflows, 600 dpi is often recommended for initial page-scans when the following features are observed in the text:

- Significant usage of fonts under 9 points in size.
- Heavy use of italics, especially for likely search terms (e.g., Latin names of plants). Italics, especially small text italics, can suffer dramatically in 300 dpi scans. While naked-eye legibility is often unaffected, OCR accuracy, and therefore search hits, can drop precipitously.
- Journals including equations. 600 dpi guarantees a high-quality printout, without concern for legibility, even with tiny superscripts and other features common to math and engineering publications.
- Citation pages combining small text and italics. It is vital to get the best possible image quality and OCR results on these pages if you intend to extract text for use as XML or plan to add links to the PDF. On the other hand, you may wish to exclude these areas from full text search (as ASA elected to do), in order to focus full-text search results on the article text itself.
- Poor-quality or faded print, where higher resolutions can help reconstruction.

While it is often appropriate for publishing, 600 dpi scanning is frowned on by most of the imaging industry for reasons of large initial file-size—and because in most cases their fastest equipment cannot scan at 600 dpi—another reason to deal with a specialist in scholarly publishing requirements. Adobe Reader supports JBIG2 compression, which means that top-quality 600 dpi images can make for smaller files than 300 or even 200 dpi pages compressed with the standard TIFF (CCITT-G4) compression. While costs may be cut with lower resolution scanning, 600 dpi should be strongly considered if there is any question of poor OCR performance due to small text, or if italic text is likely to contain search terms of relevance to researchers.

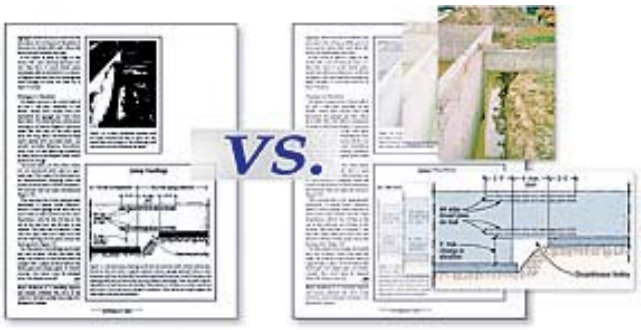
It is vital that rigorous quality control standards be applied to the black and white imaging process, as overall results are almost completely

*Continued on page 12*

dependant on successfully planning and executing the basic imaging functions with the full foresight and attention to detail required. From OCR accuracy to file-size to print and on-screen appearance, everything about your conversion project depends on getting it right, from the start. This means careful despeckle, trimming and deskew for every page, as well as the rotation of landscape pages for consistent onscreen appearance.

**Now for the Color!**

Imaging, quality-controlling, and managing color images are the single largest components of the legacy conversion expense—and the most likely source of poor or otherwise undesirable results.




Scanning in color tends to increase the cost and complexity of conversion projects but, when well considered and executed, preserves the information from the original page without grossly inflating the file-size.

The conventional way to deal with journals including color or grayscale images is to scan the entire page in color, rather than just the color sections of the page.

Black and white scanning of the entire page is undesirable due to the rather dramatic and negative effect of effectively "photocopying" color images. Color scanning of the entire page is undesirable due to the ugly "pixilation" effect of color compression on text and to the far larger file-sizes that typically result from the use of full-page color.

The best solution to this problem in terms of the quality of result is PDF/MultiResolution, the conversion process first developed in 1997 by Document Solutions. At the heart of this process is the concept that **every section of every page** should be treated in a way that is sensitive to the content contained therein. MultiResolution processing assures archive-quality files because it permits the use of higher resolutions and reduced compression (and therefore higher image quality) for both color and black and white images while keeping file-sizes lower on average than is possible with any other conversion technique.

Typically, PDF/MultiResolution pages contain black and white full page images at 300 or 600 dpi with loss-less or effectively loss-less compression. Color and grayscale image "zones" at 144 or 200 dpi with moderate compression are added only as necessary. By contrast, the use of full-page color images usually mandate lower resolutions (150 or even 96 dpi) for the entire page, along with harsher compression, in order to achieve even a moderate file-size.

Due to higher costs, MultiResolution processing is generally only for journals with occasional images, i.e., about one image every three to five pages, or fewer. With more frequent color images, a different technique combining full-page color and full-page black and white scans may be appropriate to help reduce the overall conversion cost. Make sure you ask your service provider for a specific recommendation and the logic behind it. Remember, you NEVER want to do this more than once!

**In the 21st century, the only thing worse than not showing up as a search-engine hit is not being available on the Web in the first place.**

# Evaluating Legacy Content: Seven Key Considerations

**1. Gathering Metrics**
There are several key numbers you will need to get a quotation. First, how many total pages do you expect in the entire collection? Exact numbers are not required, but costs are generally developed on a per-page basis. Another key consideration: how many color or photographic images are included in the pages? These images often represent the single largest cost, since color requires special handling to ensure that quality is not compromised.

**2. Gathering Samples**
Select three or four issues from different stages in the life of the publication. The organization, layout, design, and image content of the publication as it changed over time is a critical factor in evaluating a legacy conversion. Ideally, select at least one example issue from each decade or from each major change in layout.

**3. Preparation**
While you are selecting sample issues, examine your entire in-house collection. Can you lay hands upon two copies of every issue? Do you have even one copy of every issue? It is cheaper to scan books that have had spines removed so that the pages are loose and may be fed into a scanner. If only one copy of a given issue is available, however, it can be scanned intact, but at a greater cost.

**4. When Content Breaks: How to think about the way layouts affect current usage**
It is a good idea to spend some time examining your back issues. Differing layout and printing styles create a few options—and can affect costs as well. For example, if your articles are printed "tailed-in," that is, with the end of the preceding article and the beginning of the following article on the same page, then you will want to consider asking your service provider to copy and edit the pages to separate the articles.

In addition, you will need to decide how the secondary sections of the journal—Letters, Reviews, Corrections, and so on—should be addressed. For many journals, this is a question of what is citable in the text. In general, any content that may be cited should be broken out into a separate PDF file. Some clients also prefer to have each Correction located and that page added to the article being corrected. The variety of different publishing, layout, and printing methods and choices from days gone by should be fully thought through and discussed with your service provider.

**5. Future-Proofing the Investment**
Even though relatively few imaging service companies cater specifically to the needs of STM and academic publishers, there are many approaches, and sadly, some still get it wrong with cut-rate execution and short-sighted perspective. Exotic or small-bore technologies should be avoided. It is best to stick with industry standards such as PDF and SGML/XML. Imaging should be performed at high resolutions with plenty of scope for quality control because OCR accuracy, color-image clarity, and overall reprint quality is of paramount importance in scientific publishing. Be sure that your service provider not only understands these concerns, but is proactive in pointing them out and helping evaluate your particular publication, situation, and reuse needs.

**6. The Question of File Size**
There is no real "standard" file size, as this varies according to page size, number of words, cleanliness of the scanned image, the presence of color, and many other factors.

For some journals, there is little cause for concern with respect to file size, as typical 15-page articles with black and white text can reasonably be expected to come in at under 400 KB using modern JBIG2 compression. Add three color images, and the file size may grow to 1 MB or more. Add a few full-page color plates, and you could be looking at 4 MB for a single article. There is only room for 150 such articles on a standard CD-ROM, and when online, very large files are a barrier to many users, especially those overseas, or on slow connections.

For this reason, it is imperative to focus on a combination of smallest file size while ensuring that the minimum desirable quality is retained for color and grayscale images. Be sure to ask for samples at different quality levels, and choose some challenging pages with especially detailed images for your test.

**7. Can You Do It Yourself?**
Today, average users can readily buy capable color or black and white scanners, some inexpensive scanning and OCR software, and set about making their own electronic archives. Right?

It is true, you can do this yourself. You might decide to do it in-house, especially if your material is all black and white, and if there is not too much of it. Study your material closely, look for all the ways you would like to organize it online or elsewhere, and think through whether you want to take those issues on-board yourself. Superficially, imaging, OCR, PDF creation, indexing, SGML headers … they are all easy, but they do need to be planned, have a specific purpose, and each step must undergo rigorous quality-control in order to be successful, no matter who performs the work.

Bear in mind that your printer or hosting company may charge a processing fee to receive, check and load your content to the system, or to build disk products. If you have done the work yourself, or elsewhere, they may not want to be held responsible for the quality of the outcome. If your OCR was uneven, for example, and searches do not turn up the results you would expect, you will not be able to turn to your host or disk service provider for help.

## CONCLUSIONS

Bringing historical content online is almost always an effective use of resources for any academic publishing organization. Commonly heard predictions that older content is of little or no value to present-day researchers are easy to disprove, but what is harder to predict is the profound satisfaction that comes with the act of preserving the collected intellectual effort of decades. Improvement in service to the membership, increased citations, and a reduction in costs are all reasons to undertake a legacy conversion project. At the end of the day, however, the value lies in the contribution to the professional and greater communities. Legacy content conversion is a deposit in the history bank, made against a future time when all the world's knowledge may be spread out as a map for us and for future generations.

# How to Perform a Legacy Conversion with Allen Press

First, realize that your legacy content is an investment, a tangible asset you are about to unlock. The ability to deploy the publication history electronically can generate revenue forever and depreciates slowly. Discuss online hosting options and disk-based products with your colleagues and your Allen Press service representative to understand how best to make the legacy conversion serve both the organization and its membership.

Gather samples, estimate total page and image volume, and submit for quotation.

Discuss your options with Allen Press's legacy conversion experts and decide how to handle the issues raised in the assessment and quotation process.

Locate at least one, and ideally two, copies of every issue. Box and ship to Allen Press. If Allen Press already stores your back-issues, then you can skip this step—they will locate them for you!

After an inventory, Allen Press will produce a list of missing issues while the conversion process proceeds. Allen Press recommends placing journals online from the latest material backwards as complete publication years are processed, closing any gaps in the sequence with specific efforts to locate individual missing issues.

After completing the conversion, Allen Press will upload your files to Web-servers for use online, and at your request, go to work on any disk products you have ordered for sale or use in promotions.

Duff Johnson is founder, president, and principal consultant for Document Solutions, an electronic content services and consulting organization located in Oakland, California and Boston, Massachusetts. Specializing in PDF technology since 1996, Document Solutions performs archive-grade imaging, conversion, capture, accessibility, disk development, and Web application services. The company enjoys a national reputation for developing and executing effective solutions to electronic publishing challenges and opportunities.

Duff Johnson is also contributing editor for *Business and Government* at PlanetPDF.com. He is a frequent speaker and seminar leader at industry trade shows and professional association meetings, including Seybold Seminars, AIIM, the PDF Conference, and others.

The author developed and executed all of the services discussed herein on behalf of Allen Press, their clients, and others. Annual Reviews and the Agronomic Society of America are not clients of Allen Press. If you have further questions or require more information about these services, please contact Ted Freeman at Allen Press at 1-800-627-0326 X170, or tfreeman@allenpress.com.

## Glossary of Terms

**Adobe Reader:** The free (and freely distributable) version of Acrobat, designed to allow the user to view and print PDF files. To create or change PDFs, you will want the full version of Acrobat.

**Bitonal:** (bye-tonal) means, simply, "black and white." An imaging industry term for the common scanning mode used with text pages. Bitonal images are smaller than color or grayscale images, and deliver excellent text quality when correctly handled.

**Bookmarks:** A popular feature in PDF files, bookmarks provide an easy method of navigating between subheadings within an article, or as a clickable table of contents.

**DPI:** Dots Per Inch, also referred to as "image resolution." The higher the number, the better the quality of an image. Most computer monitors can display 72 to 96 dpi, thus most Web graphics are designed at 72 dpi. For online use, we generally recommend a resolution of 200 dpi for color or grayscale images and 600 or 300 dpi for bitonal images.

**Index Data:** Title, author, publisher, volume, and other information that serves to identify the article.

**PDF Links:** PDF files may contain links that work exactly as hyperlinks do in HTML to connect text or images to the Internet.

**OCR:** Optical Character Recognition. The name given to pattern-recognition technologies used to convert the raw scanned image to computer-readable (and therefore searchable) text.

**PDF:** The acronym for Adobe Systems' Portable Document Format. PDF refers to an electronic document technology for final-form content. PDF delivers fonts, formatting, colors, and graphics to any computer or printer, with identical results. With Adobe's free Reader, PDF is now the de facto standard for final-form electronic documents.

**Resolution:** The DPI chosen when scanning and for final delivery. The higher the scan resolution, the finer the detail of an image will be—but high-resolution images require more memory (and cause slower downloads) than lower resolution images.

**SGML:** Standard Generalized Markup Language. SGML is a standard for how to specify a document markup language or tag set. SGML is not in itself a document language, but a description of how to specify one. HTML and XML are examples of SGML-based languages. SGML is used widely in scientific publishing to manage highly interrelated content, author complex text elements such as equations, and manage high-value documents that are subject to frequent revisions.

**XML:** XML (Extensible Markup Language) is another description of a document language, akin to SGML. XML documents consist of text and tags, and the tags denote a specific structure upon the document. The purpose of XML is to facilitate the development of highly specific sets of tags to address specific content handling and management needs.

# ALLEN PRESS, INC.

## We do it all, and we do it better!

## QUOTATION REQUEST FORM

**For Complete Legacy Conversion, Disk Production Service, and/or Web Journal Publishing Services**

**Current content is presently hosted by Allen Press:** ❑ **Yes** ❑ **No**

*Publication title* _____ *Issues per year* _____ *Today's date* _____

*Name* _____ *Job title* _____ *Organization* _____

*Address* _____ *City* _____ *State, Zip* _____

*Phone* _____ *Fax* _____ *E-mail* _____

### ❑ I would like a quotation and information for the following:

❑ **LEGACY CONVERSION SERVICES FOR JOURNALS AND MAGAZINES**
PLEASE PROVIDE THE FOLLOWING INFORMATION

Date range of publication _____–_____
Total number of issues _____
Avg # of articles per issue _____
Avg # of pages per article _____
Avg # of pages per issue (all pages, including covers) _____
Avg # of grayscale images per issue _____
Avg # of color images per issue _____
Approximate total images per issue _____

❑ Yes ❑ No Are you interested in having Allen Press host your legacy content online?
❑ Yes ❑ No Are you interested in having Allen Press create a CD/DVD product from your legacy content?
❑ Yes ❑ No Does Allen Press currently provide your organization with any print or electronic services?

❑ **WEB PUBLISHING SERVICES** (Visit us online at: http://www.allenpress.com)
- ❑ Full-text online journal Web sites (SGML, XML, HTML, and PDF file formats)
- ❑ **AllenTrack** online manuscript tracking and peer review software program
- ❑ **AllenTurnStyle** online pre-editing software for copy editors and editors
- ❑ Online titles and abstracts (FREE for full composition clients via APT • Online)
- ❑ Online abstract submission and meeting planning system
- ❑ Digital conversion of hard copy back issues

❑ **DISK PRODUCTION SERVICES** (Journals, Magazines, Newsletters, Directories, and Books)
- Production of annual, historical, proceedings, ancillary content, and anniversary CD and DVD-ROM

**Submit your proposal request with a recent copy of your publication to:**

**Ted Freeman, Director of Web Publishing**
**Allen Press, Inc.**
**810 East 10th Street • Lawrence, KS 66044**
**Ph: 800-627-0326, ext. 170 • Fax: 785-843-6153 • E-mail: tfreeman@allenpress.com**


Print



**www.allenpress.com**