

Four Tips for Managing Success in Large-Scale Digitization Projects

Executive Summary

While the Internet has always offered the promise of allowing users to access content anytime, anywhere, much of the printed material created prior to 1996 – and even vast amounts created afterwards – still exists primarily in physical form. But that is changing as many organizations embark on massive digitization initiatives, following in the footsteps of libraries and national archives that first began digitizing content in the early 1980s as a way of preserving, distributing and making knowledge available to a wide range of users.

Today's digitization efforts have a similar goal, but a different motive. Commercial digitization efforts offer publishers new revenue opportunities, including subscriptions, pay-per-view, online advertising and syndication. Digitizing content can also enable publishers to broaden their reach to international markets, provide content previews or inexpensive trial subscriptions and enhance brand awareness.

While digitization offers many benefits – when done right – the process itself can appear daunting. Large-scale projects typically involve digitizing thousands, or even millions, of documents. Complexities include handling crumbling and handwritten documents, multiple column formats, different size documents and various types of content such as photos, tables, charts and graphics. Add to that the need for publishers to ensure that the content is indexed and marked-up so it can be searched by users and repurposed for other, primarily online, delivery channels.

Faced with these challenges, it's no surprise that many organizations recognize the need to work with a proven digitization partner. But the search for the right partner should entail more than finding a low-cost provider equipped with the latest scanning and conversion technology. Companies would be better served by searching for a partner with proven methodologies for handling the myriad of documents that might be involved, along with the experience to generate digitized content that can also be repurposed for a wider set of delivery channels.

In this white paper, we'll discuss these issues and also provide:

- Background on the economic reasons for embarking on large-scale digitization projects
- Guidelines to help organizations evaluate the capabilities of different outsourcing partners
- Real world examples of how publishing firms and other organizations are taking advantage of tested digitization techniques to deliver value-added content to their readers and achieve business goals

Background

Early Digitization Efforts Preserved Texts, Made them freely Accessible and Improved Searchability

Ever since Johannes Gutenberg invented the printing press in 1450, people have sought to distribute content more widely. On July 4, 1971, Project Gutenberg, the first book digitization venture, extended this mission to the earliest precursors to the Internet when someone typed in the U.S. Declaration of Independence and sent it to everyone on a computer network. Since then, the project's 20,000 volunteers have scanned or typed in about 20,000 out-of-copyright books.

The concept of large-scale digitization projects gathered steam in the mid-1990s as libraries, publishers and database companies began to see the benefits of bringing extensive volumes of printed text on-line. Efforts included the Library of Congress' American Memory Project, piloted in 1994, and the subsequent National Digital Library Program, the Million Book Project funded by the National Science Foundation (NSF) in 2002, as well as similar projects at most major libraries.

Sponsored by national funding agencies or private donations, these digitization projects were driven by the desire to preserve aging texts, make knowledge and literature freely accessible to all and improve find-ability.

Publishers and Search Engines Initiate Digitization Projects

Over the past decade, numerous commercial publishers, including newspapers, magazines, and publications have introduced online versions, while recent efforts by book resellers and search engines – led most notably by Amazon and Google, as well as Yahoo! and MSN – to bring digitized books to the Web have increased the scope and awareness of digitization projects.

Amazon started in 2001 with Look Inside the Book, followed in 2003 by Search Inside! In 2004, Google entered the book digitization arena with its announcement of Google Print, later renamed Google Book Search. In keeping with Google's mission to make information universally accessible, Google claims that the initiative will enable users' search results to include "snippets" of text from copyrighted books that contain a match for the search terms as well as the full text for out-of-copyright works.

The content comes from partnerships with publishers, who have given Google permission to make extracts of copyrighted books available through the service, as well as the plan to digitize 15 million books from five major libraries: Stanford University, Harvard University and the University of Michigan, the Public Library in New York City and the Oxford University Library in the U.K.

Google's initial efforts to scan library books have aroused considerable controversy, particularly from publishers and university presses that claim Google is infringing on copyrights. Google claims to protect copyright holders by providing only a card catalog-style entry to furnish basic information about the book and no more than two or three sentences of text surrounding the search term to help users decide if they've found the right item. Even so, several associations representing publishers and university presses have threatened to take Google to court and one association filed suit in late 2005.

To address such copyright concerns, Yahoo!, in October of 2005, joined forces with Adobe Systems, the European Archive, HP Labs, the U.K. National Archives, O'Reilly Media Inc., Prelinger Archives, the University of California, and the University of Toronto to found the Open Content Alliance (OCA). This global consortium is focused on providing open access to cultural, historical, and technological digitized print and multimedia content from libraries, archives, and publications.

The consortium says it will respect copyrights by securing permission from copyright holders and allowing all content contributors to specify use restrictions. Yahoo! will power the search engine on the OCA engine and all content will be available through Yahoo! Search.

Later that same month, Microsoft announced its intention to launch MSN Book Search to make content from books, academic materials, periodicals and other print resources available through MSN Search. MSN also joined the OCA to support its efforts to scan and digitize publicly available print materials as well as work with copyright owners to legally scan protected materials.

While these initiatives have garnered the lion's share of attention in the news media, they only represent one arena, digitizing recently printed books as well as those that are either out of print or hard to find. Another front in the effort to digitize existing content is being driven by newspaper and magazine publishers that are seeking to digitize their own archives or backfiles.

Economic Benefits of Digitizing Content

Unlike earlier non-profit projects, commercial digitization programs are designed to take advantage of new business opportunities. Common models for driving revenues from digitized content include:

- **Subscriptions**—these services charge for ongoing access to online content.
- **Pay per view**—allows users to purchase individual pieces of content, such as an article.
- **Online advertising**—consists of advertisements shown alongside text and other aspects of the site in formats that include banner ads, pop-ups, skyscrapers, buttons and leader boards. These ads can be shown in context through purchasing history or behavioral targeting, in which the site monitors a user's clickstream. One of the most popular forms of online advertising is search advertising.
- **Syndication**—where publishers deliver content to a third party and receive royalties.

Digitized content can also support publishers' business objectives in ways that indirectly drive revenues. For example:

- Enabling publishers to broaden their subscriber base by providing content to an international audience quickly and cost effectively.
- Permitting inexpensive trial subscriptions to new audiences
- Offering searchable archives, either as a free service to attract readers and enhance awareness of the site or as a pay-per-article

Four Tips for Selecting a Proven Digitization Partner

While digitization offers tangible benefits for publishers, the process itself can be daunting, frequently involving volumes of publications with thousands, if not millions, of pages that include such complex challenges as:

- Handwritten pages requiring transcription
- Pages that may be crumbling with age
- Multiple page sizes and column formats
- Graphic elements such as photos, tables, charts and graphics

Beyond that, publishers also face the additional challenge of adding features that will simplify search and support repurposing. This means publishers need to:

- Isolate content instances to simplify searching as well as index and link content
- Have sufficient conceptual XML expertise to enable content repurposing

That's one reason many publishers choose to work with a partner that specializes in large-scale digitization projects. Yet finding the right partner requires an understanding of what is necessary to tackle challenging projects. While many digitization partners typically emphasize costs, publishers also need to consider additional factors, such as quality, scalability, proven expertise and leadership in technology.

Quality

Issues related to quality at any point during the digitization process can rapidly snowball, causing potentially nightmarish results. That's why it is imperative to work with a partner that views quality as a pervasive and fundamental part of the entire digitization process.

To guarantee a high level of accuracy, for example, you need to work with a partner that maintains an independent quality certification group that will audit all work prior to delivery, ensuring it meets agreed-upon quality and accuracy levels. That means the partner should have adequate quality reviews in place to check for proper pagination and image quality and to ensure that documents are searchable.

The partner also needs to have trained staff who can not only identify deficiencies in source material, but also have the experience to offer ways to resolve the problem.

Scalability

Your digitization partner should offer the proven capacity to support large, complex digitization projects and to quickly ramp up to support your digitization requirements in a timely fashion. The need to respond to shifting deadlines or emerging opportunities can require digitization projects to be completed faster than planned. Working with a partner that can scale rapidly will enable you to get materials online sooner, add more subscribers and potentially gain a competitive advantage.

Therefore, you need to ensure that the partner's production facilities:

- Have adequate, well equipped hardware and infrastructure.
- Have sufficient numbers of well-trained, highly skilled staff members to handle the project.
- Provide geographic diversity. Having facilities in different geographic locations allows work to progress in the event of an emergency or natural disaster and also means that they may be able to shift additional work to meet a pressing deadline.

Proven Expertise

By the same logic, a partner with a proven track record supports rapid ramp up for even the largest digitization project. So when selecting a partner, be sure to check their track record on large-scale projects by asking the following questions:

- Can they provide referenceable clients?
- Can they handle complex materials, such as handwritten notes, microfiche/film and other difficult materials?
- How many people are on staff for each part of the digitization process?
- Do they use employees or outsource?
- What is the level of staff turnover?

Technology Leadership

With the right technology, a digitization partner can convert any type of document and enable your organization to take that content to the next level of profitability. These questions will help you determine whether the partner can handle your requirements:

- Do they have equipment that can handle all types of documents, including books, historical documents, microfiche/film and film transparencies?
- Does their process implement automated workflows and sophisticated review and processing?
- Do they have expertise in XML to enable content reusability and provide options for future revenue generation opportunities, such as rich data?

Digitization Projects

The following are examples of digitization customers who have benefited from partnering with a digitization expert.

Case Study

ProQuest

Challenge: When ProQuest acquired the rights to the microfilm archives of *The Washington Post* and *The New York Times*, it wanted to convert them into a format that could be preserved indefinitely to maximize the potential for reuse and repurposing. This meant converting 5.6 million pages, more than 100,000 editions and 150 years of news.

Solution: ProQuest partnered with an experienced digitization solutions provider to cost effectively and efficiently convert newspaper archives into XML files that could be repurposed for multiple formats. The conversion process involved pioneering new digitization techniques that could handle multiple page layouts, articles of varying lengths, page jumps, photos and artwork. The result was a fully searchable file that allows users to view articles in their original context.

Result: Launching the archives strengthened the publisher's competitive edge in the college library marketplace and made a strong contribution to the publisher's bottom line. ProQuest's subscription renewal rate has grown steadily since the archive was launched and revenues from these archives increased 24 percent in 2003.

American Heritage

Challenge: *American Heritage* magazine wanted to use archived content as a rich, credible base from which to launch its new Web site. The magazine, therefore, needed to digitize 28,240 hard-copy magazine pages from more than 300 back issues published over 50 years.

Solution: With a deadline looming, *American Heritage* partnered with an experienced content solutions provider who used optical character recognition (OCR) software to scan each page, then edited the resulting scans to clean and enhance the image, manually sequenced text columns and used a proprietary OCR process to recognize the text data. The partner also proofed the content and inserted XML tags for publication over the Internet.

Results: The completed archives have enabled the publication to launch a new Web site and attract new site visitors, thereby increasing subscriptions and advertising revenues.

Global Financial Organization

Challenge: A global financial organization charged with overseeing and monitoring foreign exchange rates, balance of payments and government economic models has consistently recorded its activities and viewpoints in an extensive, paper-based library. To maximize the value of this content repository, it needed to digitize its collection of 7,000 publications, which were in a variety of formats.

Solution: The partner helped the organization digitize its collection, at the lowest possible cost and at 99.95 percent accuracy. Using proprietary imaging technologies and processes, the partner processed the body text of the documents, along with the photographs, charts, and other graphic elements, providing six output files for each document. To improve searchability, the partner set up a disciplined process for extracting targeted metadata.

Results: The organization now has a virtual content treasury of its key documents that enables it to showcase its thought leadership capabilities, while also simplifying use – and its value – for researchers from other organizations.

Online Retailer

Challenge: One online retailer wanted to compete with brick and mortar stores where customers could browse by allowing customers to search for specific phrases or topics of interest in more than 500,000 books and to view selected parts of a book online. That meant the retailer needed to digitize about 150 million pages.

Solution: To keep its focus on its core business, the online retailer relied on a digitization partner's conversion expertise and ability to scale rapidly to digitize the books. The digitization partner also performed basic indexing for the converted files, creating one image and attribute file per page to help the retailer upload the digitized images to its Web site. The partner met the company's tight deadlines and quality standards.

Results: By eliminating one of the barriers to buying books over the Internet, the retailer drove higher sales. In fact, in the first five days after introducing the feature, the retailer's online book sales rose 9 percent, prompting another 37 publishers to join the companies that had signed up for the initial launch.

Conclusion

Organizations are increasingly seeking to digitize massive paper archives to extend the Internet's ability to make content available to users anytime, anywhere.

But to truly realize the preservation, distribution and revenue goals of these endeavors, organizations need assistance navigating the complexities of the digitization process.

A partner who can digitize large amounts of content more accurately and enhance searchability, all within tight budgets and timeframes, will often pay dividends beyond the digitization project itself—by providing XML-based content that supports an integrated content-driven strategy, more flexible search features and better data presentation.

As a result, publishers gain an end result that more effectively supports business goals, whether those include greater content sales, higher advertising rates, brand leadership or another business objective.