LIBER QUARTERLY, ISSN 1435-5205

K.G. Saur, Munich. Printed in Germany

JSTOR: Large Scale Digitization of Journals in the United States

by KEVIN M. GUTHRIE

> „What a terrific service you offer! Thanks for the wonderful tool that makes scholarly research so much more accessible."

Nicole C. Raeburn, Graduate Student, Dept. of Sociology, Ohio State University

Such an expression of enthusiasm was more than even we could have hoped for when the idea for JSTOR was conceived a little more than three years ago. In fact, the notion of increasing access to information was a secondary objective of the JSTOR project in its early days. Instead, our initial goal was to test whether it would be possible to save library shelf space by converting older, little-used materials to electronic formats and storing them in a central database. It seemed a simple concept but has proved to be enormously complicated in practice.

Much has transpired in a short period of time. The JSTOR database now includes well over 2 million pages from 61 important journals in 13 academic disciplines. Additional journal content is being digitized at a rate of more than 100,000 pages per month. More than 320 libraries in the United States and Canada have become participating institutions, providing support for the creation, maintenance and growth of this database. Outside of North America, we have established a mirror site in the United Kingdom. Through a novel collaborative relationship with the Joint Information Systems Committee, the JSTOR database is now being made available to over 20 higher education institutions in England, Scotland, Wales and Northern Ireland from a mirror site at the University of Manchester. In addition, plans are underway to establish a second overseas mirror site in Budapest, Hungary to serve institutions in Eastern Europe and Russia. As each day passes, new opportunities are presented to us to extend the reach of this enterprise. It is an exciting and challenging time.

Since many of the participants in the LIBER conference are from outside of the United States and may not be familiar with JSTOR, the purpose of this paper will be to provide an overview of the project and an update on where things stand.

BACKGROUND

Originally a project of The Andrew W. Mellon Foundation[1], JSTOR is now an independent not-for-profit organization with a mission to help the scholarly community take advantage of advances in information technologies. JSTOR's initial objective is to develop a trusted archive of core scholarly journal literature, with an emphasis on the retrospective conversion of the entire backfiles of key academic journals. One reason older journals offer a compelling test case of the future digital library is that they occupy large amounts of shelf space in academic libraries, representing a real and on-going cost to the system. Recent developments in information and networking technology now make it possible to store information digitally in one or a few sites and distribute it widely, potentially reducing the long-term systemwide costs associated with duplicative storage and maintenance.

A second reason for focusing on older journal literature is that these materials are relatively little-used in their present paper and microfilm formats and generate little or no revenue for publishers. While it may be that for some fields older research is of considerably less value than more recent published articles, it is certainly not the case for all fields and disciplines. In its present format, some valuable older research risks being „lost" because it is difficult to locate and inconvenient to retrieve. Thus, JSTOR offers an unusual opportunity to capitalize on new technologies in a way that benefits all participants in the scholarly communication process. Librarians can reduce their long-term costs while providing enhanced services for their patrons. Publishers can enhance the usefulness of their titles and develop a platform on which to build future electronic publishing initiatives at no cost to them. And scholars and students will have renewed access to important - albeit older - literature that had previously not been readily accessible. Approaching the problem with this systemwide perspective is an important component of JSTOR's mission to serve the scholarly community.

THE JSTOR SYSTEM

The JSTOR database is comprised of three primary elements: page images, corresponding text files, and a table of contents file. JSTOR is committed to providing a faithful archival replication of the original publication, so all pages in each journal, from the very first issue published to the most recent issue allowed by the publisher[2], are scanned at 600 dpi resolution. The material includes not only full-length articles, but also book reviews, advertisements, front matter and back matter; in short, the entire published record. In addition to the images, a text file is generated using optical character recognition software and is then manually corrected to 99.95 percent accuracy. This text file is not displayed to users but is used for facilitating full-text searches. The third element of the database is an electronic table of contents file, which is keyed to 99.995 percent accuracy and includes key bibliographic information such as article title and author, as well as abstracts and keywords for those articles that include them. For a retrospective collection like JSTOR, this database structure offers users the advantage of images (perfect fidelity to the original) without sacrificing the chief benefit of text files (allowing full-text searches). The resulting system is a powerful research and teaching tool.

WHY DIGITIZE THESE JOURNALS?

In recent years there has been a good deal of discussion and speculation about the feasibility and usefulness of digitizing large quantities of printed books and journals in library stacks. There are some who believe that it is technically and economically feasible and intellectually desirable to digitize „everything“, and there are others who think that the value of retrospective materials is limited and would not justify the high cost of digitizing and maintaining them. JSTOR pursues a middle path.

JSTOR was founded on the conviction that it is intellectually desirable and economically feasible to digitize, maintain and distribute a carefully selected body of core journals and other retrospective resources provided that the costs are shared among a very large number of libraries. It is also assumed that, over the long run, the cost savings to participating libraries will more than offset the fees that they pay to JSTOR.

Enhanced Access

Although the initial goal was to preserve important journals and save library shelf space, technological advances have made it possible to provide an entirely new form of convenient access to the older JSTOR materials. In this easily accessible form, these core journals acquire an importance they did not have as bound volumes in library stacks. For libraries that do not hold these journals, the ability to gain access to an enormous quantity of information through connections to computer networks offers possibilities previously unimaginable. With desktop access, users can mine the contents of these journals in ways and to an extent that simply was impossible with the paper copies. In addition, the JSTOR archive offers the complete run of the journal, which is rarely found on any single library's shelves because of missing volumes, defaced or damaged pages, or inconsistent historical collections policies.

Archiving

Although storing materials in digital form offers unprecedented opportunities for distribution and access, it also presents new challenges for long-term preservation and archiving. When compared with paper, the media on which digital information is stored is relatively unstable. Perhaps more importantly, digital information must be interpreted by software before it can be presented in a format that is understandable. Words on paper have the wonderful and simple benefit of being directly discernible by the human eye. Digital bits on electronic media require computer intervention. Because the software programs that perform this interpretation are constantly evolving, systems and data must be migrated to new platforms to insure long-term availability. There is no ready solution to the challenge posed by this constant evolution. Because archiving electronic material is central and fundamental to JSTOR's reason for being, however, all aspects of its strategy, from the selection of technological formats to the way it invests its resources, are oriented to insuring that these materials will be accessible in the future.

JSTOR Today

After a pilot period during which several test site libraries used the earliest prototype of the JSTOR system with a small number of journals, it was evident that the concept held great promise. In August 1995, JSTOR was established as an independent organization separate from the Mellon Foundation.

Progress on building Phase I of the JSTOR database, which is slated to include a minimum of 100 journals before the year 2000, is ahead of schedule and continues at a steady pace. JSTOR now has signed agreements with the publishers of 98 journals in 15 academic disciplines. As previously mentioned, there are well over two million pages in the database and new content is being added at an average rate of 100,000 pages per month. In response to the general enthusiasm for the project and repeated requests from librarians to provide more journals, JSTOR production capacity will be expanded significantly during 1998. It is expected that the JSTOR production rate will near 300,000 pages per month by the end of the year.

JSTOR was first made available to libraries in the United States and Canada on January 1, 1997 and the response to-date has been extraordinary. Even though budgets at libraries remain tight, over 340 libraries around the world have elected to support the project. These libraries represent every type of academic institution, from the large research university to the small liberal arts college. What motivates the libraries to participate? There appear to be a variety of reasons. For some librarians, their reason for participation is to provide increased access to these materials. For others, JSTOR allows them to move journals to off-site storage, freeing up much-needed shelf space. Still others are motivated by JSTOR's commitment to archiving, and have indicated that, in addition to gaining a new scholarly resource, they regard JSTOR participation as a form of research and development on the future of the library in an electronic world.

USAGE

Scholars and students at participating institutions are demonstrating their enthusiasm for JSTOR through their use of the resource. Increases in usage since the beginning of the fall 1997 academic year have been dramatic and growth continues unabated. During the fall term, the total number of pages viewed, searches performed and articles printed from the database increased 340 percent. Usage in September 1998 was the most ever and exceeded the previous monthly high by more than 20 percent. Approximately 43,000 articles were downloaded for printing during the month (nearly 1,500 per day) and over 146,000 searches were performed.

It is clear that the database is being used not only for research, but for teaching as well. Professors are assigning articles they find in the JSTOR database to their classes and students are making use of the database to research and write papers. Professors report that JSTOR provides historical depth to

the materials found by students who are increasingly relying on electronic resources and the Web to find information. In addition, with important journals in a variety of fields available through a single interface and search engine, JSTOR encourages cross-disciplinary inquiry. Access to this resource is also enabling new forms of research that previously would not have been possible. One noteworthy example is the research of Fred Shapiro, a librarian and Lecturer in Legal Research at Yale University in the United States. Using JSTOR, Shapiro reports that he has been able to retrieve occurrences of important terms antedating the earliest evidence for those terms recorded in the Oxford English Dictionary.

CONCLUSION

Through its pioneering position employing technology to help scholars, publishers and libraries, JSTOR is blazing a new trail. The initial reaction among early participants in this collaborative effort has been extraordinarily positive. As JSTOR takes its first steps beyond the bounds of its initially defined objective - to provide access to 100 journal titles to academic institutions in the United States and Canada - there will be many challenges. First will be how to make the resource accessible to scholars and researchers in other parts of the world. A key question will revolve around the technological infrastructure and the available bandwidth for accessing JSTOR content. It is not possible for JSTOR to continue to establish mirror sites around the world and evidence is accumulating that bandwidth is becoming less of a problem. A second question will be defining appropriate contribution levels for non-U.S. institutions. It would not be appropriate for access overseas to be subsidized by the constrained budgets of libraries in the U.S. Establishing fees that match appropriately the amount libraries pay to the value they receive from participation will be the challenge. Where possible we will seek out other relationships like the one we have with the Joint Information Systems Committee in Great Britain to help facilitate access for groups of libraries in a particular country or region.

Another set of challenges revolves around future content for the database. Identifying journals to include in later phases, addressing the problems of including non-english and other non-U.S. content, and providing mechanisms to link JSTOR articles to other resources are just a few of the complicated issues that will have to be addressed. In seeking to rise to these challenges, JSTOR will remain focused on its mission, making adjustments to its plans as required to keep making progress in its continuing commitment to serve scholarship and the global academic community.

REFERENCES

1 The Andrew W. Mellon Foundation was established to „aid and promote such religious, charitable, scientific, literary, and educational purposes as may be in the furtherance of the public welfare or tend to promote the well-doing or well-being of mankind.“ Under this broad charter, the Foundation currently makes grants on a selective basis to institutions in higher education; in cultural affairs and the performing arts; in population; in conservation and the environment; and in public affairs. More information is available at the Mellon Foundation's website at <http://www.mellon.org>.

2 In order not to put the revenue from current issues of participating publishers at risk, JSTOR establishes a fixed lag period between the most recently published issue and the last volume included in the JSTOR database. We call this lag the „moving wall“ and it varies by publisher. Most of the moving walls for the titles in the database are either 3 or 5 years.

Kevin M. Guthrie
President, JSTOR
jstor-info@umich.edu