Bruce Silver Associates
Workflow, Document, and Image Management

Industry Trend Reports
April 1997

Bringing Paper to Life

Adobe Acrobat Capture Unlocks Corporate Memory

Executive Summary

Adobe's Portable Document Format (PDF) has become a de facto standard on the Web, embraced as a preferred alternative to HTML for Web delivery of many kinds of documents. The rush to unlock hidden corporate "knowledge" is forcing a convergence of Web publishing with document management and making PDF a strategic technology in business. With Acrobat Capture 2.0, Adobe now allows paper documents to be included in Web-based document repositories – the last frontier – via unique image-to-PDF conversion technology. Acrobat Capture software's new programming interface, enhanced page recognition, and marketing partnerships with leading scanning and document management vendors at last meet the requirements of Web publishing.

Introduction

In today's economy, successful businesses recognize that ready access to capital, labor, and natural resources is no longer enough for competitive advantage. "Knowledge is replacing matter and energy as the primary generator of wealth," says Thomas Stewart of *Fortune* magazine's Board of Editors. McKinsey calls it "the knowledge economy." Even before we've figured out how to measure information assets on the balance sheet, *knowledge management* has burst forth as a strategic imperative in business.

Leveraging this corporate knowledge asset means it must be made much more visible and accessible, but where does it reside? Most of it exists in the form of documents, paper and electronic, unmanaged, squirreled away in private repositories throughout the organization. The drive to unlock this "corporate memory" has suddenly drawn together two software technologies – Internet publishing and document management – previously moving in independent orbits. Now, their intersection in the form of Web-based knowledge management is transforming the document repository from a departmental niche solution to an explosive growth business.

Overnight the Web, intranets, and extranets have become synonymous with easy, universal access to information, based on pervasive, virtually free, cross-platform browsers and industry-standard protocols. Information, or "knowledge," need no longer be locked inside specific applications, but can be made available to any authorized user over a global network. Web technology has in fact made access so easy that managing the content already looms as the bigger problem.

Document management vendors, long focused on access control and version management of critical "work product" documents, have only recently come to appreciate their own value in Web publishing, which historically has served up more marketing-oriented content than "knowledge." With the $700 million market value of Web content management software already twice that of traditional document management, the convergence of these software categories is now spawning a market tornado, pulling a new and diverse set of vendors and products into what was once an industry niche.

Bruce Silver Associates
260 Glen Road Weston, MA 02193
Tel: 617.237.6879 Fax: 617.237.1641 E-mail: brsilver@mindspring.com

As it did for desktop publishing a decade ago, Adobe Systems is supplying a critical enabling technology for this transformation. Adobe's Portable Document Format (PDF), long used in both document management and Web publishing, has emerged as a key catalyst for their convergence. In Acrobat 3.0 last November, Adobe seized upon the specific needs of both groups of vendors and refocused PDF technology squarely on Web-based document management. Now, with its Acrobat Capture 2.0 announcements, the company promises to extend knowledge management across the last untamed frontier of corporate memory – paper documents.

Acrobat and Web Publishing

Before we describe Acrobat Capture software, it's worth examining the role Adobe Acrobat and PDF play in the Web publishing phenomenon. Acrobat software was originally designed to solve the problem of distributing electronic documents to unknown desktop environments. Why might those desktops be incompatible with that of the author? Take the simple problem of distributing a Word document authored in Office97. The recipient might be on a Mac, PC, or UNIX; might have an earlier version of Word; might use Lotus WordPro; or might not have the fonts of the authored file. Acrobat was designed to be cross-platform and both font- and application-independent.

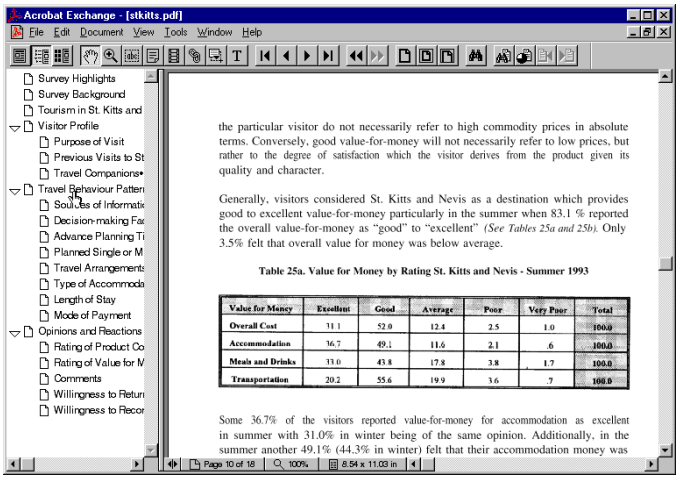

Figure 1. PDF allows Web content to be formatted exactly as the author intended, navigated via hyperlinks and thumbnails, and maintains full text searchability from popular Web search engines.

Through a cross-platform print driver, Acrobat creates a universal, content-searchable visual replica of the authored document – the PDF file – that can be displayed with a free Acrobat Reader on any desktop platform or in a Web browser. Acrobat relies on unique Adobe multiple master font technology that mimics the look, size, and spacing of any font, preserving the layout of the original for viewing or printing. Then Acrobat adds to that a rich security model, thumbnail page navigation, hyperlinks, annotations, and index fields, creating an ideal format for final form electronic document distribution on the Web.

Acrobat 3.0 further enhanced PDF specifically for use in Web publishing. Besides the ability to run Acrobat 3.0 or Acrobat Reader as a Netscape plug-in or ActiveX control, Adobe added page-on-demand downloading, progressive page rendering, compression and caching of embedded fonts and graphics, cross-document hyperlinks, speed Web access to PDF documents. Lest there be any doubt that PDF has become a de facto standard for Web document delivery, the Adobe Web site for downloading the free Acrobat Reader trails only Netscape Navigator, Microsoft Internet Explorer, and Real Audio in daily volume, averaging 40,000 downloads per day.

PDF vs. HTML

But wait a minute, you say. The Web already has a standard universal format for document publishing. It's called HTML. It's also free, cross-platform, rich in functionality, and it doesn't need a special plug-in to work in a browser. True, and for most content on the Web, HTML is just fine. But HTML has characteristics that make PDF a better publishing choice for specific classes of documents:

1. ***Documents with complex or critical formatting.*** Fundamentally, the way an HTML page looks in a browser depends on the user's installed browser software, fonts, and browser setup options. The document author cannot precisely control it. HTML lacks the subtlety and fine control over text sizing and spacing, graphics placement, forms and table layout that is available in authoring programs like Word, or in PDF.

> *"The fastest growing collection of documents in any corporation may be the documents on their Web sites… HTML isn't rich enough for long-term storage, and PDF best meets this obvious requirement for a reliable format."*
> Larry Warnock, Director of Market Development, Documentum

2. ***Documents intended to be printed or saved.*** An HTML Web "page" bears no relationship to a printed page. Authors cannot control page breaks, sections, headers, and footers for optimal printed output. A user viewing HTML can't easily navigate to what would be the top of page 4 in the printed version. On the other hand, PDF layout is oriented to the printed page, like the Page Layout view in Word. Also, if you download and save a PDF document, you get the whole thing in one file, not separate files for each embedded graphic as you do with HTML.

3. ***Long documents.*** HTML offers publishers a tough choice for long documents: Either make them long, single Web "pages," easy to access and print, but slow to load and hard to navigate; or break them up into multiple short pages, each quick to load and view, but painful for the user who wants to save or print the whole document. Acrobat creates a single PDF file for easy access to the whole document, but adds features that streamline loading and navigation. PDF supports page-on-demand downloading, which, combined with page thumbnails and table of contents links, allows the user to go directly to the section or page of interest. This PDF capability relies on the standard HTTP byte-serving extension, a native feature of most major Web servers, including those from Netscape and Microsoft.

4. ***Predominantly graphical documents, such as presentations.*** In HTML, each slide in a presentation file must be converted to a GIF bitmap image, forcing a tradeoff between resolution and file size, and making the content unsearchable and virtually unprintable. With PDF, presentation graphics are resolution-independent and print beautifully, while the content remains searchable.

5. ***Documents where reader operations should be restricted.*** The text and graphics source of any retrieved HTML document can be saved and edited or reused without the owner's permission. PDF files can be encrypted and allow the author to restrict specific document operations on retrieved documents, such as copying, printing, or modifying.

6. ***Documents originating in paper.*** Software providing optical character recognition (OCR) conversion of paper documents to HTML is readily available, but Acrobat Capture makes PDF a better solution where a large volume of paper must be published. OCR-to-HTML requires time-consuming correction of recognition and formatting errors, while Acrobat Capture does not, and preserves both visual fidelity and content searchability.

Published vs. Processed Paper

Acrobat Capture is the component that scans and converts images into PDF, allowing paper to be digitally published, searched and retrieved, annotated and managed identically to electronically authored documents. Adobe describes Acrobat Capture as its *Paper-to-Web Solution*, but to simply leave it at that suggests the product is just another new wrinkle on document imaging software. While Acrobat Capture incorporates bits of imaging technology, it is ideally positioned as a *document management* product rather than as a document imaging product. The difference is in the kind of paper involved.

Acrobat and document management applications most often deal with *published documents*, information meant to be disseminated and read – knowledge. These documents – reports and

manuals, policies and procedures, rules and regulations, best practices, marketing literature, competitive analysis – are typically many pages in length, periodically revised, and indexed by their text content. Repositories of these kinds of documents have long been the domain of the document management vendors, like Documentum, PC DOCS, OpenText, and Saros, and partnership with these vendors is critical to Adobe's success with Acrobat Capture.

Document imaging applications, on the other hand, more often deal with a different kind of paper: *processed documents*, records and forms, often meant to trigger a business transaction, saved as legal evidence that a particular event took place at some date and time. These documents, such as invoices, loan applications, birth certificates, or insurance claims, are usually short and are scanned to be processed and archived rather than to be published. Document processing involves extracting document information into a business application, using key entry or form-based OCR, and linking the document image to an application data record. Preserving an *exact* facsimile of the original paper is critical, so these documents are typically stored as TIF bitmap images rather than in a visually approximate but content-indexable format such as PDF. While image documents may be retrieved on occasion from a Web-accessible digital archive for customer service or audit purposes, it would be a stretch to call that "publishing."

PDF vs. TIF

While they are not Adobe's immediate focus, document imaging vendors like Wang/Kodak, FileNet, or Plexus, who manage this kind of processed paper, could also benefit from integrating Acrobat Capture. At roughly 10 Kbytes per page, PDF file size is about a quarter that of a Group 4 compressed TIF. PDF also allows viewing on 13 platforms, content searchability, text printing at maximum device resolution, and a common presentation for images and electronic documents. Besides, in addition to PDF-Normal files that include layout, images, and searchable text, Acrobat Capture can create pure TIF images and surround them with a thin PDF wrapper. This retains the above PDF advantages and offers optional content searchability provided via hidden OCR text. *With these inherent benefits, integration with processed paper solutions will likely represent a growing future market for Acrobat Capture.*

Acrobat Capture: How It Works

Technically, Acrobat Capture is at heart a special kind of OCR. Its patented special "trick" is that, in addition to recognizing characters from bitmap images, it maps their precise size, shape, and location on the page. This information is used in combination with Adobe's unique multiple master font technology to create a PDF layout that is a near-exact visual replica of the original, while maintaining text searchability. Because it can invent a font that takes up exactly the prescribed space, it doesn't have to reflow the text and mess up the page layout.

Like conventional OCR, the recognizer avoids embedded graphics and does not try to convert them to text, but unlike regular OCR, it substitutes bitmap snippets from the original document image for unrecognized or suspect characters, providing a correct visual output. Because Acrobat Capture does such a good job of mimicking recognized machine print fonts and locating them on the page, the bitmap suspects are hard to distinguish from the surrounding text in the PDF output unless the document is highly magnified (Figure 2).

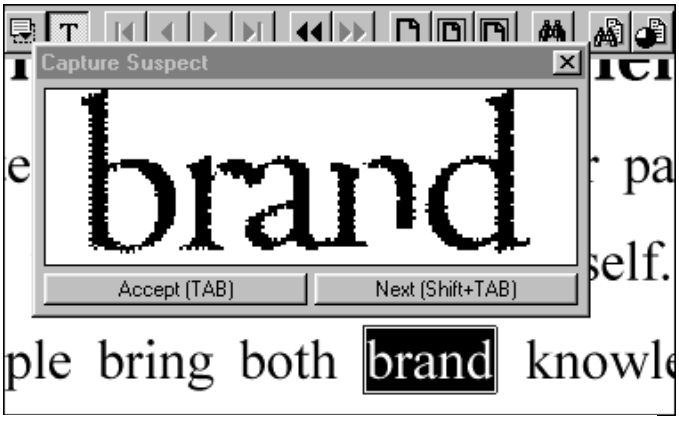

Figure 2. Magnified Acrobat Capture output shows flagged bitmap suspects along with recognized fonts in the PDF document.

The fact that unrecognized characters don't have to be manually corrected before the output is publishable is a key benefit of Acrobat Capture. A word containing an unrecognized character may be missed in a full-text query, but the rendered PDF file – unlike OCR-to-HTML – will not display erroneous characters. With Acrobat Capture, manual validation and correction of suspect characters can be limited to just those document areas critical to document searching, or omitted altogether. For a typical paper repository, this can easily be the critical cost factor that determines whether mass conversion for Web access is economically justifiable.

In its Acrobat Capture 2.0 announcements, Adobe has enhanced the software and repackaged it for two classes of users. For personal or low-volume workgroup use, an Acrobat Capture plug-in is bundled as part of Acrobat 3.0. Acrobat includes a Scan menu pick that supports TWAIN, ISIS, or Visioneer desktop scanners. The Acrobat Capture plug-in converts selected image pages that have been scanned or imported into Acrobat. It allows suspect characters, rendered as bitmaps, to be flagged with red rectangles for manual editing and provides touch-up editing of PDF text and fonts.

For volume conversion, Adobe offers Acrobat Capture 2.0 as a separate product, significantly enhanced for Web publishing applications. The most important enhancement is a programming interface, which now allows the PDF conversion engine to be integrated with third-party subsystems for high-volume scanning and document management. The recognition that production-level implementations require tight integration with leading third-party products marks Acrobat Capture software's evolution from a stand-alone, shrink-wrapped technology to a serious component of an intranet publishing solution. On the scanning side, an API is critical because the PDF conversion process, which can take 10 to 60 seconds per page, is much slower than high-volume scanning. Production image capture subsystems like Cornerstone's Input*Accel* allow a fast scanner to distribute images to multiple PDF conversion engines and reintegrate their individual output streams to achieve the necessary page throughput. Through the API, these scanning subsystems can export both PDF documents and index data directly into a document management system, such as Documentum's EDMS.

Another enhancement is improved page recognition, including fonts, forms, and tables. Acrobat Capture 2.0 also supports text recognition in eight languages, and can handle color and grayscale images as well as black-and-white. Where reviewing of converted PDF files is important, Acrobat Capture 2.0 includes the ability to change text fonts and colors, import and place images, and review batch files before final PDF output. And in addition to PDF output, Acrobat Capture 2.0 can also OCR paper to HTML, ASCII, Rich Text Format (RTF), Word, and WordPerfect formats.

Usage Scenarios

When people think of Web publishing, they usually think of documents they author and edit themselves. Certainly those have been a major focus of Adobe for a long time, with products like Photoshop, Adobe Illustrator, and PageMill for Web page authoring, PageMaker and FrameMaker for complex document assembly, SiteMill for Web site management, and Acrobat for Web document delivery. But where does Acrobat Capture come in? In other words, where is the paper that businesses want to manage and publish in the same way that they do for internally authored documents?

First and foremost, there is *legacy paper*, documents for which the electronic source is no longer available, but which need to be managed consistently with new documents authored and published electronically.

> "Since implementing our electronic document distribution system using the Internet and PDF, we have realized cost savings of up to 90 percent over traditional paper distribution."
> David Roller, Manager, Strategic Planning, Lucent Technology Customer Information Center

These include reports and technical manuals, contracts and proposals, litigation discovery documents, or outbound correspondence, among others. For example, MCI is creating an intranet library with past and present product manuals. Lucent Technology has reduced the cost of publishing critical customer service documents by 90 percent using PDF and Acrobat Capture. The federal government's Office of Technology Assessment, which went out of business last year, leaves to posterity 25 years of technical reports in PDF format, with Web-accessible print on demand.

> *"We have thousands of pages of corporate information,… insurance forms and benefits materials. With Acrobat, it takes just minutes to output documents to PDF and make them available to employees over the intranet."*
> Joe Small, Advanced Technology Group Lead, Westinghouse Source W

With Acrobat Capture, Adobe's ability to unify the indexing, presentation, annotation, and delivery of legacy paper with electronically generated documents is unique, avoiding the dichotomy that usually occurs when marrying imaging with document management. With the API, Acrobat Capture 2.0 is now well suited to the needs of service bureaus performing large backfile conversion and integrating the output into enterprise document management systems.

Second, there is *published paper received on an ongoing basis*, such as news clippings, competitive product information, product maintenance or safety bulletins, blank forms, and HR documents. For example, pharmaceutical companies preparing all-electronic New Drug Applications to the FDA use PDF to integrate Case Report Forms faxed in from clinical trials, typically integrated with an enterprise document management system, such as Documentum. These systems automatically create a publishable PDF *rendition* of every document when it is checked in following revision. Smaller intranets for HR or sales support can also use Acrobat Capture integrated with the inexpensive document management capabilities of the Web server platforms themselves, such as Netscape's SuiteSpot, Lotus Domino, or OpenText.

Third, there are *processed paper* applications that, while not Adobe's immediate target market, benefit from PDF: cross-platform or browser access, for example, smaller file size than TIF, or full content searchability. For example, Safeco Insurance is using PDF with Saros in claims management. As the historical barriers between document imaging systems and document management systems blur and erode, this segment will be a growing market for PDF and Acrobat Capture.

> *"Our goal is to automate the process of handling claims… We use PDF as our standard format [because it] is better suited to long-term archive requirements… [and] we can support a single viewer for all documents."*
> Carl Kulgrove, Jeff Weeks, Claims Department, Safeco Insurance

Industry Partnerships: Down to Serious Business

Acrobat Reader can be downloaded from a Web site, but a serious Acrobat Capture implementation is not so simple. A production-level solution requires integration with proven scanning and document management components. What makes Adobe's new Acrobat Capture initiative credible is its enlarging list of vendor partnerships in these areas. At the low end, Fujitsu is already bundling Acrobat 3.0, including the Acrobat Capture plug-in, with its new ScanPartner 600C, a 15-ppm color desktop scanner being marketed as a Web content development tool. Also, Visioneer distributes an Acrobat Capture Link for the PaperPort desktop. These will help build awareness of paper-to-Web publishing for small intranets with modest capture requirements.

For medium-volume conversion, Xerox will link Acrobat Capture 2.0 with its 40-ppm duplex DocuImage scanner and DocuPath software, currently sold mainly as part of a print publishing solution. With PDF capture in the new version of DocuPath, Xerox will be able to leverage its publisher-oriented image processing features in the Web publishing domain. Also, Kofax will create an Acrobat Capture Release Module for its Ascent Capture software, a turnkey scanning

subsystem sold in many imaging solutions. The Release Module will hand off scanned images to Acrobat Capture and reintegrate the PDF output for export to a document management system.

For high-volume conversion, Cornerstone is using the Acrobat Capture 2.0 API to integrate PDF conversion into Input*Accel*, with support for parallel Acrobat Capture engines and direct export into Documentum's EDMS. A prototype of this integration was demonstrated at a pre-announcement analyst briefing in March. It will be sold by Cornerstone as an Input*Accel* module (Figure 3), integrated with export modules to leading document management products. Adobe continues to work actively with scanning device and subsystem vendors to make their products PDF-aware.

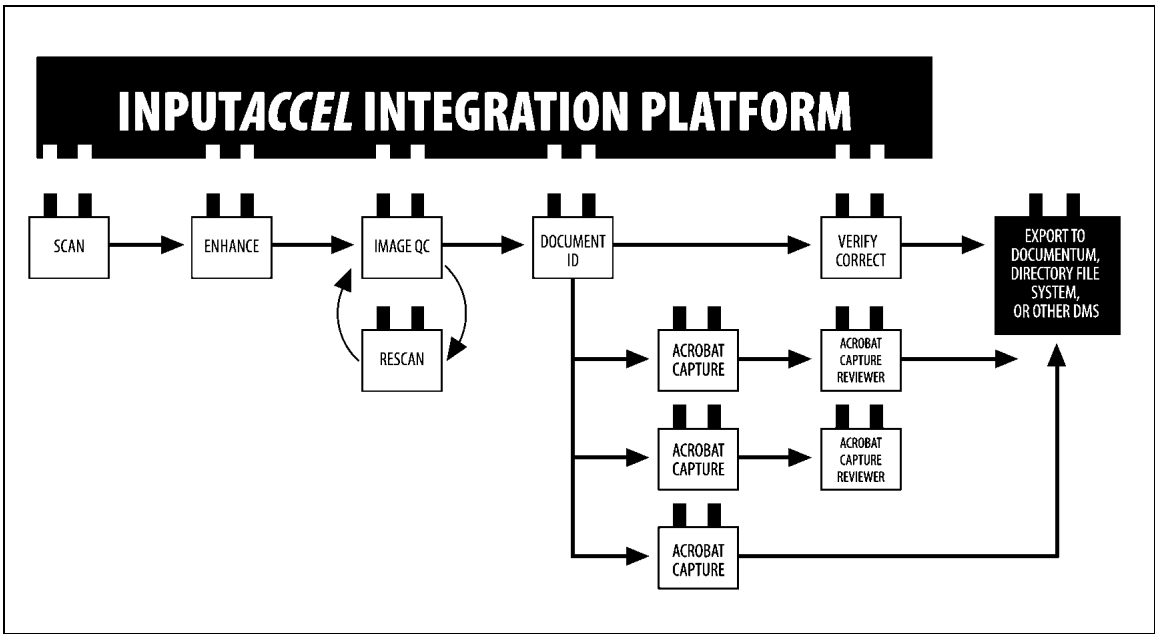

Figure 3. Using the Acrobat Capture 2.0 API, third-party scanning subsystems like Cornerstone's InputAccel provide high-volume paper-to-Web conversion for service bureaus and production-level document management systems.

Most document management vendors are jumping on the PDF bandwagon. Like Documentum, Saros creates PDF renditions for Web publishing, and other vendors will be doing so in the near future. The new object-relational *universal servers* from database vendors like Informix and Oracle are extending their mission-critical capabilities into document management as well. For example, the INFORMIX-Universal Server includes a PDF DataBlade. And groupware vendors like Lotus and Netscape have introduced new low-cost Web document management/publishing suites that incorporate PDF. Netscape's Java Web Publisher and Enterprise Server 3.0 simplify document publishing from the desktop and support page-on-demand serving of PDF documents, while the Netscape Catalog Server provides text searching of native PDF content along with HTML and Office document formats.

Bruce Silver' s Analysis

Now that PDF is established as a de facto standard for Web document delivery, and Acrobat Capture has been reconfigured as a programmable building block of real-world publishing solutions, Adobe's paper-to-Web crusade is not only appealing, it's credible. Vendor partnerships are crucial, and Adobe is pursuing them actively. The Fujitsu and Xerox partnerships are a great start, and we should expect to see the list expand in the future.

We'd like to see Adobe go even farther to stimulate paper-to-Web publishing on small intranets. An integrated solution combining Acrobat Capture with a medium-volume scanner, complementary Adobe Web authoring and site management tools, and a simple but powerful document management server like Netscape SuiteSpot or OpenText – all easily installed from a single CD – could really jump-start paper-to-Web in the marketplace.

At the high end, products like the Input*Accel* PDF Capture module are exactly what Adobe needs. Production capture solutions must be solid end to end, from scanning to recognition to review editing to document management exporting, yet offer flexible options for scanner hardware, indexing, and the range of document management systems supported. Adobe needs to make sure that in addition to Documentum, other enterprise-scale document repository vendors – already sold on PDF – integrate interfaces for production scale paper-to-Web deployment.

Adobe also needs to stay focused. Published paper, processed paper, what's the difference? Acrobat Capture has benefits for both – and Adobe wants to remain open to both. But document management systems are where the PDF is today and represent the likeliest immediate channel for paper-to-Web deployment. If Adobe can make Acrobat Capture as transparent as possible and keep the focus on Web publishing and document management, users won't get hung up on the exotic technology underneath.

The Web is the key to leveraging corporate knowledge. The convergence of the intranet publishing explosion with document management has liberated electronic documents. With the new urgency in corporations to unlock their hidden knowledge assets, PDF has become a strategic technology for business, a de facto standard for Web document delivery. Paper documents, locked up in file cabinets and dusty basement archives, now represent the final frontier. With Acrobat Capture 2.0 and its market partnerships, Adobe is at last bringing that paper to life.

Bruce Silver
April 1997

