

Document Imaging Report

Business Trends on Converting Paper Processes to Electronic Format

4003 Wood Street ● Erie, PA 16509 ● PH (814) 866-2247 ● FX (800) 507-8981 ● <http://www.documentimagingreport.com>

January 7, 2005

THIS JUST IN!

PEGASUS COMPLETES ACQUISITION

Pegasus Imaging's acquisition of the Component Product Technology division of **TMSSequoia** has closed. The deal between the two tool vendors was initially announced this summer [see *DIR* 9/3/04]. The final price of \$2.25 million was approximately \$650,000 less than originally announced.

TMS, which also recently received \$250,000 from **Measurement Incorporated** for its share in a joint venture involving test scoring, has now been dissolved. It had been a player in our industry since 1981 but recently cited Sarbanes-Oxley as making it too expensive to operate as a small (<\$3 million) public company. Proceeds from the recent sales will be distributed to shareholders.

Pegasus, meanwhile, seems excited about adding the well respected TMS tools to its product line. These include tools for image and forms processing, as well as viewing. "We merged two research-driven imaging technology powerhouses," said Jaime Cornelison, marketing manager for Pegasus. "All former TMS products' pricing just dropped to be more affordable. Plans are also underway to merge the best technology of both companies. The future is bright, solid, and fiscally healthy at Pegasus."

In other recent M&A news, **Verity** recently acquired BPM partner **Dralasoft** for \$8 million in cash. Dralasoft was already an OEM partner of Verity's—its application is embedded in the *LiquidOffice* e-forms line that Verity acquired along with Cardiff last year [see *DIR* 2/20/04]. Dralasoft CTO Suresh Ramaswamy has been named Verity's VP of BPM strategy and will report to former Cardiff CTO Mark Seamans, who is now Verity's senior VP of R&D. More on this next issue.

For more information:

<http://www.pegasusimaging.com>, <http://www.verity.com> **DIR**

2005 Buzzwords: Mailroom, Desktop, Compliance, & Books

Welcome to the new year! We hope 2004 brought you increased sales and/or successful document imaging projects and installations. From our viewpoint, 2004 was one of the most exciting years this industry has seen since the introduction of digital imaging back in the mid-1990s. Concerns about regulatory compliance and best practices in information management finally drove document management out of the backroom and into the mainstream.

Yes, traditional imaging strongholds like insurance, government, and financial services continue to be on the cutting edge of implementation innovation. However, there also has been an upsurge in ham-and-egg-type installations in the SMB market and plenty of new business in the conservative healthcare arena. These emerging markets were driven by an increasing demand for records management brought on by a combination of government regulations and fear of litigation and the related discovery. Of course, these concerns also apply to larger companies that have had to deal with the likes of Sarbanes-Oxley and the fear that wider reaching regulatory controls could be on the way.

After briefly setting the landscape for you, without further adieu, here are four key market trends we expect to see in 2005:

1. **The "digital mailroom" is not a hoax** – Yes, we realize there have been precious few implementations of this much-hyped technology that we first wrote about back in 2002 [see *DIR* 7/5/02]. However, we also realize what a potentially great application of imaging it represents. Being able to automatically sort and submit to electronic workflow an unprocessed pile of incoming mail is a very attractive value proposition. And we have seen a boatload of evolving IDR (intelligent document recognition) and auto-classification technology that will eventually enable this to happen.

The biggest problem with the digital mailroom so far, (in addition to some yet-to-be thoroughly proven technology), is the way it has been marketed. For the most part, it has been presented as a brand new implementation, rather than an expansion of an old one. In 2004, digital mailroom technology had the most traction with existing imaging customers who wanted to get more out of their installations. As more and more documents are enveloped (excuse the pun) into these applications, they will become the initial "digital mailroom" sites.

2. Compliance and records management concerns continue to drive SMB adoption – I have first-hand exposure to this trend through a series of seminars I have been delivering for a state non-profit (<http://www.ebizitpa.org>). My monthly classes are filled with small business people desperate to get a handle on their documents. Over the past 20 years, we have witnessed the PC revolution, and along with that has come a proliferation of desktop applications that make it very easy to create documents—not the least of which is e-mail. The weight of these proliferating documents is bowing the infrastructure of smaller businesses. Because of their greater masses of documents, larger businesses were the first to recognize the advantages of electronic document management. Smaller businesses, meanwhile, continued to make do until the recent publicity regarding regulatory compliance, legal ramifications, and disaster recovery finally woke them up to DM.

3. Desktop market comes to forefront – A couple years ago, the term "distributed capture" was all the rage. And with the recent triple-digit growth in workgroup (sub-\$2,000) document scanner sales, you'd expect the distributed capture revolution was well underway. Although distributed capture software sales are growing, they haven't nearly kept pace with the astronomical explosion in "distributed" scanner sales. Why's that? A lot of these lower-volume scanners are simply going into lower-volume centralized applications.

As document imaging continues to move downstream, the demand for lower-volume, desktop-level document management software will obviously increase. While this may not seem like a particularly exciting proposition for document imaging software vendors who want higher margins than what can be achieved at the desktop, there is an important connection that needs to be considered: Businesses, especially smaller ones, don't operate as islands. Chances are their documents are being shared with partners, vendors, and/or customers. Thus, these desktop applications can serve as important feeders to larger ECM applications. In many cases, they can be the distributed capture application! This appears to be the path **Kofax** is headed down with **Capio**. We think it's a good one, as long as Kofax resellers can stomach the retail margins traditionally associated with the desktop. The recent convergence of the fax server and document imaging markets is another indication of this trend.

CONTINUED ON PAGE 7

Document Imaging Report

Business Trends On Converting Paper Processes To Electronic Format

DIR is the leading executive report on managing documents for e-business.

Areas we cover include:

1. Scanning
2. Forms Processing/OCR/ICR
3. Integrated Document Management
4. Content Management/XML
5. Document Output
6. Storage
7. E-Commerce

DIR brings you the inside story behind the deals and decisions that affect your business.

Vol. 15, No. 1

Editor: Ralph Gammon
4003 Wood Street
Erie, PA 16509
PH (814) 866-2247
FX (800) 507-8981
ralphg@documentimagingreport.com



Managing Editor:

Rick Morgan
PH (814) 866-1146
rickm@scandcr.com

Publisher and Circulation Manager

Larry Roberts
RMG Enterprises, Inc.
5905 Beacon Hill Lane
Erie, PA 16509
PH (412) 480-5116
FX (412) 291-1352
larry@rmgenterprises.com

DIR is published 24x per year, on the 1st & 3rd Fridays of the month, by:

RMG Enterprises, Inc.
5905 Beacon Hill Lane
Erie, PA 16509
PH (412) 480-5116
<http://www.documentimagingreport.com>

Copyright © 2004 by RMG Enterprises, Inc. Federal copyright law prohibits unauthorized reproduction by any means including photocopying or facsimile distribution of this copyrighted newsletter. Such copyright infringement is subject to fines of up to \$25,000. Because subscriptions are our main source of income, newsletter publishers take copyright violations seriously. Some publishers have prosecuted and won enormous settlements for infringement. To encourage you to adhere to this law, we make multiple-copy subscriptions available at a substantially reduced price.

Subscriptions: \$597 (electronic) or \$670 (paper) per year.

Special Pull-Out Section:

Document Imaging Report

Annual Article Index For 2004

The next four pages of this issue are devoted to our annual index of articles from the previous year. We realize this takes up valuable editorial space, but think it serves as a useful tool for those that save their issues of the *Document Imaging Report* for reference. Personally, I've been the editor of this publication for more than six years now, and know that I find tons of information on current topics by poring over my back issues. Highlights of 2004 include several introductions of both high- and lower-volume scanners, the continuing evolution of intelligent document recognition and digital mailroom technology, an uptick in mid-market sales, and a continued stream of consolidation activity.

We'd like to note some improvements to the print version of the index. This includes adding "Main Topics" and entities "Featured" columns. Of course, a fully searchable *Excel* version, which contains more complete data, is available to subscribers online at <http://www.documentimagingreport.com>. Give me a holler or shoot me an e-mail if you need your password, which unlocks this feature. Back issues are similarly available. If you want, we'll even package the 2004 issues and index on a CD and send it to you through the traditional mail. We hope 2004 treated you well and are looking forward to enjoying a great 2005 - together!

A key to the acronyms used in this index: ECM (Enterprise Content Management); M&A (Mergers and Acquisitions); IP (Image Processing) IDR (Intelligent Document Recognition); BPM (Business Process Management); EDM (Electronic Document Management); DAM (Digital Asset Management), RM (Records Management); WCM (Web Content Management)

Issue Date	Pg.	Name of Article	Main Topics	Featured
Jan 9, 04	1	Group Banding Together To Challenge Patent Claims	Patents	Millennium, L.P
Jan 9, 04	1	IBML Introduces ImageTrac III	Scanners	IBML
Jan 9, 04	3	Annual Article Index For 2003		
Jan 23, 04	1	Stellent Picks Up Optika	M&A	Stellent, Optika
Jan 23, 04	1	FileNET Beefs Up Manufacturing Efforts	Verticals	FileNET
Jan 23, 04	3	Kodak Continues Assault On Lower-Volume Markets	Scanners	Kodak
Jan 23, 04	4	Kofax Signs Federal Intelligence Contract	Capture	Kofax
Jan 23, 04	5	Keynotes Highlight AIIM 2004 Agenda	Trade Assoc.	Advanstar
Jan 23, 04	6	Real-Time Data Entry A Growing Services Trend	Services	Fidesic, EasyLink
Feb 6, 04	1	Vignette, Tower Combine Forces	M&A	Vignette, Tower
Feb 6, 04	1	Captiva Strengthens Healthcare Practice	M&A	Captiva, Context
Feb 6, 04	1	CDIA To Host Awards Event	Trade Assoc.	AIIM, CompTIA
Feb 6, 04	3	ECM Movers and Shakers	Financials	Several
Feb 6, 04	3	IBM Continues Aggressive ECM Push	M&A, EDM, DAM	IBM
Feb 6, 04	4	Next Generation Forms Processing Leverages Two Kinds Of PDF	Forms Processing	Dataintro
Feb 6, 04	6	Cardiff Fortifies <i>LiquidOffice</i> Channel	E-Forms	Cardiff
Feb 6, 04	7	Panasonic Introduces Document Distribution Package For MFPS	Digital Copiers, Capture, Workflow	Panasonic
Feb 6, 04	7	<i>Digital Mailroom</i> Released To GA	Capture	Captiva
Feb 6, 04	8	TiS For Sale?	Forms Processing	Top Image
Feb 20, 04	1	Cardiff Throws In With Verity	M&A	Cardiff, Verity
Feb 20, 04	1	eiStream Keeps Acquisition Train Rolling	M&A	eiStream,
Feb 20, 04	3	Cardiff Competitors Share Their Views	M&A	Captiva, Dicom
Feb 20, 04	4	Kodak Launches New Mid-Volume Line	Scanners	Kodak
Feb 20, 04	6	Kofax Zeroing In On Forms Space	Forms Processing	Kofax
Feb 20, 04	Web	Lemelson Verdict Big News In Patent Wars	Patents	

Issue Date	Pg.	Name of Article	Main Topics	Featured
Sep 17, 04	1	Dicom To Acquire Communications Specialist	Fax, Capture, M&A	Dicom, Topcall
Sep 17, 04	1	KVS Gets 10x Revenue From Veritas	M&A, E-Mail	KVS, Veritas
Sep 17, 04	4	LuraTech Moving Forward After Management Buyout	IP, Color	LuraTech
Sep 17, 04	6	Interwoven Latest to Add Records Management	ECM, RM, M&A	Interwoven
Sep 17, 04	7	Imaging Momentum Driving Wholesale Lockbox Success	Check Scanning	DMP
Sep 17, 04	8	German Market Demands Signatures For Imaging	Capture	AuthentiDate
Oct 8, 04	1	Visioneer Sharpens Its Channel Focus	Scanners, Channel	Visioneer
Oct 8, 04	1	Parascript Subsidiary Has Capture App For Note Taking	OCR/ICR, M&A	Parascript
Oct 8, 04	3	Latest Reorg Combines Document Input And Output	Financials	Eastman Kodak
Oct 8, 04	4	What's In An Acronym?	M&A	Kofax, IDC
Oct 8, 04	5	ReadSoft Rebounding	Forms Processing	ReadSoft
Oct 8, 04	6	Xerox Upgrades Capture App	Capture, MFPs	Xerox
Oct 8, 04	7	Forms Capabilities Added To ABBYY Toolkit	OCR/ICR	ABBYY USA
Oct 8, 04	8	Verity Releases Extraction Application	Capture, Search	Verity
Oct 22, 04	1	Conversion Services Caught In BPO Wave	Services, Capture, OCR/ICR	Data Dimens'ns, 3SG
Oct 22, 04	1	Captiva Waiting On Big Deals	Financials, Capture	Captiva, Kirtas
Oct 22, 04	4	Patents Help Kodak Recoup Imaging Losses	Patents, M&A	Kodak, Wang
Oct 22, 04	5	IBM Focuses On Records App For The Enterprise	RM, ECM	IBM
Oct 22, 04	7	German Claims Processor Deploys Innovative Solution	Forms Processing	TiS, BKK
Oct 22, 04	7	Asaca Introduces High-Density Optical Library	Storage	Asaca, Sony
Nov 5, 04	1	Acquisition Beefs Up Kofax Forms Play	M&A	Kofax, Neurascript
Nov 5, 04	1	New ImageRunners, MEAP Apps Highlight Canon Event	Digital Copiers	Canon, eCopy
Nov 5, 04	4	Fax Server Leader Buys IMR	M&A	Captaris, IMR
Nov 5, 04	6	IKON Revamps Imaging Strategy - Again	Channel, Digital Copiers	IKON, EMC Documentum
Nov 5, 04	8	Reduced Set-Up Time Key To New Invoice Solution	IDR	SER Solutions
Nov 19, 04	1	Digital Mailroom Part Of Big Retail Win	Capture, Verticals	Captiva, Wal-Mart
Nov 19, 04	1	Thanks Herb	Trade Assoc.	TAWPI
Nov 19, 04	3	Fujitsu Defends Its Turf	Scanners	Fujitsu, Kofax
Nov 19, 04	5	ScanSoft, Adobe Strengthen Desktop Apps	PDF	Adobe, ScanSoft
Nov 19, 04	5	Captiva Gets Serious About Channel	Capture, Channel	Captiva, Kofax
Nov 19, 04	7	I.R.I.S. Upgrades PDF Batch Scanning App	OCR/ICR, PDF	I.R.I.S., Kodak
Nov 19, 04	8	IXOS Suite <i>NetWeaver</i> Certified	ECM, ERP	Open Text, SAP
Dec 3, 04	1	Panasonic Turns Up Heat In The LVP Space	Scanners, IP	Panasonic
Dec 3, 04	1	Xerox Driving Capture From Mobile Phones	Capture, IP	Xerox
Dec 3, 04	3	NSi Establishing Itself In Capture Space	Digital Copiers	NSi; HP, IKON, Kyocera Mita
Dec 3, 04	5	CDIA+ Continues To Evolve	Trade Assoc.	CompTIA
Dec 3, 04	5	Visioneer Ready To Go Toe-To-Toe With Fujitsu	Scanners, IP	Visioneer
Dec 3, 04	6	Softheon Lands Deal With Large Mortgage Lender	BPM, Verticals	Softheon, DSS
Dec 3, 04	7	Optika Visionary Reflects On Retail	Capture	Stellent, SER
Dec 17, 04	1	Hyland Puts IPO On Hold	Financials, ECM	Hyland Software
Dec 17, 04	1	Toy Giant Considers Imaging	IDR	Santa, Inc.
Dec 17, 04	4	Publishers Waking Up To MRC	IP, Color	LizardTech
Dec 17, 04	6	Fax Over IP Gaining Traction	Fax	Venali, Inc.
Dec 17, 04	7	InfoPath Introducing Users To XML	XML	Vasont
Dec 17, 04	8	Captiva Improves <i>Invoices</i> App	IDR	Captiva, SAP

Issue Date	Pg.	Name of Article	Main Topics	Featured
Jun 18, 04	1	Captiva Details Growth Plans	Capture, Financials	Captiva, IAC
Jun 18, 04	1	Clerke Leaves Verity	Forms Processing	Cardiff, Verity
Jun 18, 04	3	Kodak Introduces Workgroup Models	Scanners	Kodak
Jun 18, 04	5	Kofax Creates Innovative On-Line Store	Capture, Channel	Kofax
Jun 18, 04	7	EMC Keeps Building On Centera	Storage	EMC, Pegasus
Jun 18, 04	Web extra	Iron Mountain Implements <i>EmailXtender</i>	E-Mail	Iron Mountain, EMC
Jul 2, 04	1	BPM Offers New Opportunity For Workflow Vendors	Workflow, BPM, ECM, M&A	Several
Jul 2, 04	1	IBM Channel Embracing <i>CM Express</i>	ECM, Channel	IBM
Jul 2, 04	4	Paper Forms Specialist Making Transition To Imaging	Document Imaging	Altec
Jul 2, 04	5	Latest <i>TaskMaster</i> Release Reflects CEO's Vision	Forms Processing	Datacap
Jul 2, 04	6	Softheon Targets Insurance BPM Applications	BPM, Verticals	Softheon
Jul 2, 04	7	Wausau Beefs Up Document Technology	Workflow	Wausau
Jul 2, 04	8	Fujitsu Recognizes Distributed App	Scanners	FCPA, Kofax
Jul 23, 04	1	New Strategy & Financing Have Lason Poised For Turnaround	Services, M&A	Lason, Charterhouse
Jul 23, 04	1	New Captovation App Images Electronic Documents	Capture	Captovation
Jul 23, 04	4	Japanese Acquisition Fits With TiS Growth Strategy	M&A	TiS
Jul 23, 04	4	ScanSoft Emerging As PDF Power	PDF, Capture	ScanSoft, Adobe
Jul 23, 04	6	Canon Updates Scan-To-CD Appliance	Scanners	Canon
Jul 23, 04	6	TMS Identifying Emerging Opportunities	Viewing, IP	TMSSequoia
Aug 6, 04	1	DocuPen: Breaththrough In Mobile Document Imaging	Scanners	Planon
Aug 6, 04	1	Front Office Imaging Not New To Jetsoft	Scanners	Jetsoft
Aug 6, 04	3	Financial Applications Giant Enters Imaging Market	M&A	John Harland, Mitek
Aug 6, 04	4	Patent Protects Revolutionary Screen Scraping Technique	Document Imaging	
Aug 6, 04	5	<i>TeleForm</i> Becomes More Enterprise Friendly	Forms Processing	Verity
Aug 6, 04	6	Silas Exploring OEM Possibilities	Capture, M&A	Silas
Aug 6, 04	7	Book Scanner Adds Software	Scanners, IP	Kirtas
Aug 6, 04	8	Storage Briefs	Storage	EMC, KOM
Aug 20, 04	1	2010 Census Forms To Be Scanned In Color	Color, OCR/ICR, Archiving	Census Bureau, ADI, NARA
Aug 20, 04	1	Pegasus Buys TMS' Tools Business	M&A, Tools	TMSSequoia, Pegasus Imag.
Aug 20, 04	5	New Sheetfed Model Shakes Up Low-Volume Segment	Scanners	Fujitsu
Aug 20, 04	5	The Low-Volume Production Landscape	Scanners	Several
Aug 20, 04	6	New Features On An Old Workhorse	Scanners	Canon
Aug 20, 04	7	AIIM Finally Gives Resellers Their Due	Trade Assoc.	AIIM
Aug 20, 04	8	LizardTech Releases New Version Of DjVu Software	IP, Color	LizardTech
Sep 3, 04	1	Title App Developer Scanning More Than 1 Billion Images	Scanners, Microfilm	Zenodata, nextScan
Sep 3, 04	1	IBM Acquires <i>Venice Bridge</i>	ECM, M&A	IBM, Venetica
Sep 3, 04	3	Böwe To Complete Buyout By 2009	Scanning, M&A	Böwe Systec
Sep 3, 04	4	Pegasus Beefs Up Document Imaging Portfolio	IP, Tools, M&A, Color	Pegasus Imag., TMSSequoia
Sep 3, 04	5	Daeja Introduces Java-Based Universal Viewer For COLD	Viewing, COLD	Daeja, FileNET
Sep 3, 04	6	Vignette Gives Facelift To Tower Products	Document Imaging	Vignette, Tower
Sep 3, 04	7	Documentum Introduces Search Aggregator	ECM, Search	Documentum
Sep 3, 04	7	Municipal Gov't Expanding Scope Of Imaging Apps	Verticals	LaserFiche
Sep 3, 04	8	NARA Open To TIFFs Under Right Circumstances	Archiving	NARA

Issue Date	Pg.	Name of Article	Main Topics	Featured
Sep 17, 04	1	Dicom To Acquire Communications Specialist	Fax, Capture, M&A	Dicom, Topcall
Sep 17, 04	1	KVS Get 10x Revenue From Veritas	M&A, E-Mail	KVS, Veritas
Sep 17, 04	4	LuraTech Moving Forward After Management Buyout	IP, Color	LuraTech
Sep 17, 04	6	Interwoven Latest to Add Records Management	ECM, RM, M&A	Interwoven
Sep 17, 04	7	Imaging Momentum Driving Wholesale Lockbox Success	Check Scanning	DMP
Sep 17, 04	8	German Market Demands Signatures For Imaging	Capture	AuthentiDate
Oct 8, 04	1	Visioneer Sharpens Its Channel Focus	Scanners, Channel	Visioneer
Oct 8, 04	1	Parascript Subsidiary Has Capture App For Note Taking	OCR/ICR, M&A	Parascript
Oct 8, 04	3	Latest Reorg Combines Document Input And Output	Financials	Eastman Kodak
Oct 8, 04	4	What's In An Acronym?	M&A	Kofax, IDC
Oct 8, 04	5	ReadSoft Rebounding	Forms Processing	ReadSoft
Oct 8, 04	6	Xerox Upgrades Capture App	Capture, MFPs	Xerox
Oct 8, 04	7	Forms Capabilities Added To ABBYY Toolkit	OCR/ICR	ABBYY USA
Oct 8, 04	8	Verity Releases Extraction Application	Capture, Search	Verity
Oct 22, 04	1	Conversion Services Caught In BPO Wave	Services, Capture, OCR/ICR	Data Dimens'ns, 3SG
Oct 22, 04	1	Captiva Waiting On Big Deals	Financials, Capture	Captiva, Kirtas
Oct 22, 04	4	Patents Help Kodak Recoup Imaging Losses	Patents, M&A	Kodak, Wang
Oct 22, 04	5	IBM Focuses On Records App For The Enterprise	RM, ECM	IBM
Oct 22, 04	7	German Claims Processor Deploys Innovative Solution	Forms Processing	TiS, BKK
Oct 22, 04	7	Asaca Introduces High-Density Optical Library	Storage	Asaca, Sony
Nov 5, 04	1	Acquisition Beefs Up Kofax Forms Play	M&A	Kofax, Neurascript
Nov 5, 04	1	New ImageRunners, MEAP Apps Highlight Canon Event	Digital Copiers	Canon, eCopy
Nov 5, 04	4	Fax Server Leader Buys IMR	M&A	Captaris, IMR
Nov 5, 04	6	IKON Revamps Imaging Strategy - Again	Channel, Digital Copiers	IKON, EMC Documentum
Nov 5, 04	8	Reduced Set-Up Time Key To New Invoice Solution	IDR	SER Solutions
Nov 19, 04	1	Digital Mailroom Part Of Big Retail Win	Capture, Verticals	Captiva, Wal-Mart
Nov 19, 04	1	Thanks Herb	Trade Assoc.	TAWPI
Nov 19, 04	3	Fujitsu Defends Its Turf	Scanners	Fujitsu, Kofax
Nov 19, 04	5	ScanSoft, Adobe Strengthen Desktop Apps	PDF	Adobe, ScanSoft
Nov 19, 04	5	Captiva Gets Serious About Channel	Capture, Channel	Captiva, Kofax
Nov 19, 04	7	I.R.I.S. Upgrades PDF Batch Scanning App	OCR/ICR, PDF	I.R.I.S., Kodak
Nov 19, 04	8	IXOS Suite <i>NetWeaver</i> Certified	ECM, ERP	Open Text, SAP
Dec 3, 04	1	Panasonic Turns Up Heat In The LVP Space	Scanners, IP	Panasonic
Dec 3, 04	1	Xerox Driving Capture From Mobile Phones	Capture, IP	Xerox
Dec 3, 04	3	NSi Establishing Itself In Capture Space	Digital Copiers	NSi; HP, IKON, Kyocera Mita
Dec 3, 04	5	CDIA+ Continues To Evolve	Trade Assoc.	CompTIA
Dec 3, 04	5	Visioneer Ready To Go Toe-To-Toe With Fujitsu	Scanners, IP	Visioneer
Dec 3, 04	6	Softheon Lands Deal With Large Mortgage Lender	BPM, Verticals	Softheon, DSS
Dec 3, 04	7	Optika Visionary Reflects On Retail	Capture	Stellent, SER
Dec 17, 04	1	Hyland Puts IPO On Hold	Financials, ECM	Hyland Software
Dec 17, 04	1	Toy Giant Considers Imaging	IDR	Santa, Inc.
Dec 17, 04	4	Publishers Waking Up To MRC	IP, Color	LizardTech
Dec 17, 04	6	Fax Over IP Gaining Traction	Fax	Venali, Inc.
Dec 17, 04	7	InfoPath Introducing Users To XML	XML	Vasont
Dec 17, 04	8	Captiva Improves <i>Invoices</i> App	IDR	Captiva, SAP

4. **The adoption of more complex file formats** – The recent news coming out of **Google** regarding its book scanning initiative virtually makes this a no-brainer. Currently, it appears Google is set to post its images in old-fashioned JPEG. However, this is clearly not the optimal format for viewing books, and if Google doesn't recognize this soon, its competition surely will. This will lead to increased use of mixed raster content (MRC) document imaging, which can create, smaller, clearer, and easier to view document image files that are perfect for managing large, complex documents such as books. We've been writing about this technology for several years wondering when the killer app would appear. Google may have finally unearthed it. **DIR**

Google Imaging Project Biggest Ever

Google's recently announced book scanning initiative could be the largest document image archiving project in history. The well publicized undertaking involves some 15 million books from four prestigious libraries. The goal is to make the content from these books available to people doing on-line research. This appears to be both a noble and potentially lucrative endeavor.

On the nobility front, it will bring into play research materials that were created previous to the digital age. Many experts have lamented that these materials are often inadvertently ignored by modern researchers who rely on the Internet. On the lucrative front, providing access to this information will clearly differentiate Google from other search engine providers—none of which has announced a similar project, yet.

There seems to be only one issue standing between this project and its success. That involves the capture of page images from 15 million books. Can this all really be accomplished within the six-year timeframe initially being projected?

Let's take a look at some of the numbers: Based on an estimated 200 pages per book, Google will need to capture some 3 billion total images. Based on a 365-day work year, this breaks down to 6,849 books, or 1.3 million images, per day. Accounting for a 16-hour work day, we can break it down further to 85,000 images per hour, or more than 1,400 images per minute. Great, just take seven or eight high-speed document scanners and have at it, right? There's just one small problem. Using document scanners for book scanning involves ripping the pages from their bindings before inserting them into the ADFs. This works fine for books currently in print. However, because of copyright concerns and

its historical perspective, the Google project will involve a large number of out-of-print books. One key to gaining access to these books is that Google has promised to return them unharmed.

Calculating the capture costs

One early report has stated that Google is using some sort of manually intensive process to turn the pages and scan the books. According to a representative from the **University of Michigan** library, this process enables an operator to scan approximately 50 books in a workday. Based on our calculations, that means Google would need to employ 137 operators per day to complete its project in six years. Based on an average employee cost of \$50,000 per year, this comes out to a little more than \$41 million in personnel costs. This is probably not a big deal for a company that recently raised \$1.7 billion through an IPO.

At least according to these calculations, the capture portion of the equation seems doable. But are these realistic numbers, and is there a better way than what Google has planned?

Is there a better way?

Google has a history of being protective with internal information. Characteristically, it is not openly discussing the technical or financial details of its book scanning project. However, we think the 50 books per operator figure seems a little high—especially in a manual page-turning environment. According to Lofti Belkhir, CEO of **Kirtas Technologies**, even using his company's BookScan capture device, which features an automatic page-turning cradle, an operator can only be expected to capture three to four, 300-page books per hour. "The capture is not going to happen as quickly as early articles have suggested," Belkhir told *DIR*. "15 million books is a lot of scanning for many years to come."

Apparently there have already been some talks between Google and Kirtas. "We have talked with Google, and we are aware of the project," Belkhir said. "I'm not in a position to disclose anything else."

One notable feature of the BookScan is that it can be used to capture color images, which could open up a wide realm of display possibilities for Google. As reported on in *DIR* last issue, several book scanning projects already underway are using rich color image viewing formats like DjVu and PDF to display books. However, Google has indicated that the library project will be an extension of its current Google Print program (<http://print.google.com>), which displays pages as JPEG images.

Titles in the Print Program can be found by running a Google search on "books on 'fill-in-the-

topic.” At the top of your results will be a “book results” section and icon. You can view a limited number of pages of in-print books (with the option to buy) and a few complete works on which the copyrights have expired. The books have all had OCR applied and words that match your search query are highlighted in yellow.

Google has employed some technology to make copying and printing the images a difficult process but no special viewer or plug-in is required. The images are fast enough to download and easy to navigate. However, typical of JPEGs, the text is a bit fuzzy on the edges.

If Google continues to go with JPEGs, it could open up an interesting competitive angle for **Yahoo!**, which has a close relationship with Adobe and could leverage PDF technology. **Microsoft**, another would-be search competitor, could also differentiate itself by possibly embedding a DjVu or JPEG 2000 viewer in its popular *IE* browser and then offering online books in the corresponding format. So far, the examples of books that we’ve seen in these MRC (mixed raster content) file formats are much higher quality than what we’ve seen of the Google Print selections.

A goldmine for ECM vendors


In addition to providing researchers with access to online material, there has been a lot of talk about the preservation potential of the Google library project. Belkhir said this may not necessarily prove to be valid. “An archival image needs to be of very high quality,” he told *DIR*. “It is not necessarily the same image that is best for on-line access. JPEGs can be very quickly created and viewed. However,

JPEG 2000, because of its lossless compression and raster qualities, is showing the most promise for archiving. It’s not uncommon for our current library customers to scan their holdings twice.”

In addition to image quality concerns, OCR application presents another huge challenge for Google. “Especially, when scanning books published before? 1937, Google is going to have some issues with older font styles,” said Dean Tang, CEO of OCR software developer **ABBYY USA**. “ABBYY has been involved in many smaller books scanning projects worldwide, including the European METAe archiving project, and has developed a special module to handle these fonts.”

Tang indicated that some of the book projects he’s seen have used manual data operators to verify full-text OCR results as well as enter meta data about books. “Obviously, with a project of this size, you want to automate as much as possible,” he said. “However, you also have to be sure you achieve a high enough level of accuracy so users’ search results are relevant.

“The Google project is quite an undertaking. An organization 10 times the size of Google would feel the strain of a project this large. There is going to be a lot of opportunity for companies in the ECM industry. This project is going to lead to scanning of other reference items like journals and periodicals. And the opportunity is not just limited to North America. This is really a potential gold mine for our industry over the next 10 to 20 years.”

For more information: <http://www.kirtas-tech.com>; <http://www.abbyyusa.com> 

Subscription Order Form for RMG Enterprises, Inc.

5905 Beacon Hill Lane • Erie, PA 16509 • Phone (412) 480-5116 • Fax (412) 291-1352 • <http://www.documentimagingreport.com>

Document Imaging Report

Business Trends on Converting Paper Processes to Electronic Format

Mass Storage News

Opportunities and Trends in Data Storage and Retrieval

SCAN: The DATA CAPTURE Report

Premier Management & Marketing Newsletter of Automatic Data Capture

1 year (24 issues)

- electronic copy @ \$597
 paper copy @ \$670

- electronic copy @ \$597
 paper copy @ \$670

- electronic copy @ \$597
 paper copy @ \$670



Name _____ Title _____

Company _____

Street _____

City _____ State _____ Zip _____

Phone (____) _____ Fax (____) _____

E-Mail _____

Please enter / renew the following subscription.
 (Add \$33 on all orders outside of Canada and the United States.)

Payment Enclosed (Remit to: RMG Enterprises, Inc., 5905 Beacon Hill Lane, Erie, PA 16509)

Charge My Credit Card (Charge will appear as RMG Enterprises.)
 _____ AmEx _____ Visa _____ MC _____ Discover _____
 _____ card number _____ expire date

Bill My Organization (Purchase order # optional.) _____