

# Recent Advances in Document Image Segmentation, Compression, and Encoding

Luc Vincent

*Advanced Systems Development Laboratory  
Xerox Palo Alto Research Center*

# Outline

- DIR: decomposition/segmentation of scanned color documents into layers
  - Use of multiple resolutions and compression schemes
  - Encapsulation of DIR representation in a TIFF (XIFF) or PDF container
  - Use of technology in *Pagis Pro* scanning suite
- DigiPaper: symbol-based compression of binary scans or binary layers of above decompositions
- DataGlyphs for highly efficient, highly reliable encoding of data on paper
  - Applications in “turnaround documents”, automated document factories, security, etc.

# Representing Documents as Images

- **Benefits:**
  - Image-based document interchange supports scanned or electronic documents
  - New and “legacy” sources
  - Guarantee appearance and layout
  - Quick rendering for viewing and printing
- **Traditionally limited by:**
  - large size of image files
  - unstructured bits – doesn’t support easy interaction
- **Solution: *DIR representation + JBIG2 compression***

# Background on DIR

- Color or grayscale scanned document pages cannot be compressed well using standard methods
  - E.g., JPEG compressed document images remain very large, and the text areas are not well preserved
- Known compression schemes work better on some page elements than on others. For example,
  - *JPEG* compression is adequate for images
  - *MMR compression (CCITT Group 4)* only works well on binary text
- Resolution needs are a function of the page element being compressed:
  - Text requires high-resolution to remain readable and OCR'able
  - Images usually look good at much lower resolutions

# Principles of DIR multilayer representation

- Use resolution - and color depth - that is most appropriate for the data being compressed
  - Typically use 300 to 600 dpi for text and 100dpi or less for contones
- Use different compression schemes for different page elements
  - Requires prior *segmentation* of the page
- Decompose text and graphics regions into high-resolution binary plane, and low-resolution color plane
  - Use *multi-layer representation*
- Approach is ideal for “Mixed Raster Contents” (MRC) pages
  - Today, a large proportion of business type documents are MRC, and can greatly benefit from this MRC approach to compression

# Sample magazine page scan

## FYI

### Easy Sale

**T**he last thing you want, after spending hours with a fussy customer choosing the specs of a customized product, is to find that those options are unavailable. There goes the sale—and probably the customer as well.

SalesVision, Client Server Technologies Inc.'s product configuration software tool, helps to eliminate such situations. Designed for remote sales forces, value-added resellers and telemarketers, SalesVision provides immediate and valid product configuration information.



The Windows-based product enables a salesperson to enter product options such as make, model, color and style into a laptop computer; it then validates the information and verifies that the order can be filled. Once verification occurs, SalesVision sends the information via modem to the factory for manufacture.

SalesVision can be directly into a company's credit, accounting and financial records, so remote workers can send or receive data for credit approvals and financing during the product selection and configuration stage. In addition, SalesVision can be customized to access data from a company's marketing and engineering departments, giving salespeople access to explanations of features and analyses of competing products.

Pricing can be obtained by contacting Client Server Technologies in Schaumburg, Ill., at 708-397-7300.

### Cost-Effective Conferencing

**M**embers and soft-spoken executives with a tight telecommunications budget don't need to spend an arm and a leg for an audioconferencing unit that will pick up their voices.

Cohesent Communications Systems Corp. says its hands-free audio conferencing unit, ConferenceMaster Elite, will offer exceptional performance on even the poorest of telephone lines for less than the traditional audioconferencing duplex unit. Cohesent, based in Leesburg, Va., uses a new state-of-the-art 16-bit codec for improved sound clarity and 360-degree sound coverage for call participants as far as 15 feet away from the speaker. Automatic gain control allows every participant, even those being away from the ConferenceMaster Elite, the ability to speak naturally and still be heard clearly from anywhere in a room.

Where conventional conference systems would distort voices or break up a signal when speakers raise their voices or laugh loudly (causing a higher than normal transmission signal level), the ConferenceMaster Elite offers additional signal capacity to handle the call.

ConferenceMaster Elite is a portable, ergonomically designed unit that plugs into a standard analog telephone jack and electrical outlet. ConferenceMaster Elite is available for \$995. For more information, call 800-443-0726.



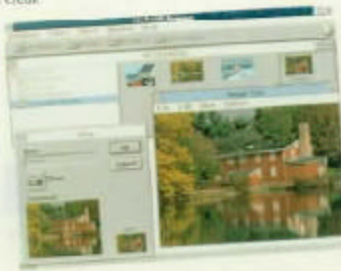
### Multimedia Management

**E**very organizational unit knows that one of the biggest problems with the surge in multimedia is how to store all those CDs, digitized LPs, videotapes, old home movies, books on tape and more. Corporations are also coming to grips with the organizational nightmare that threatens to overwhelm their multimedia operations. Enter MediaWay Inc.'s family of multimedia database-management and cataloging applications called MediaDB.

Organizations such as newspapers, real estate brokerages and high-tech marketing firms must deal with the technical challenges of storing and archiving video clips, news clips, sound clips, photos and other multimedia materials that require managing massive

amounts of memory-hungry information and unstructured relationships between items. MediaDB, a database management system that can modify multimedia information in an object-oriented environment as easily as it can traditional text and numbers, is designed from the ground up specifically to tackle multimedia organizational challenges.

MediaDB DBMS version 2.0 ranges in price from \$2,500 for a five-user license on Novell NetWare to \$200,000 for a 1,000-user license on Unix servers. For more information, call MediaWay in Sausalito, Calif., at 415-748-7400.



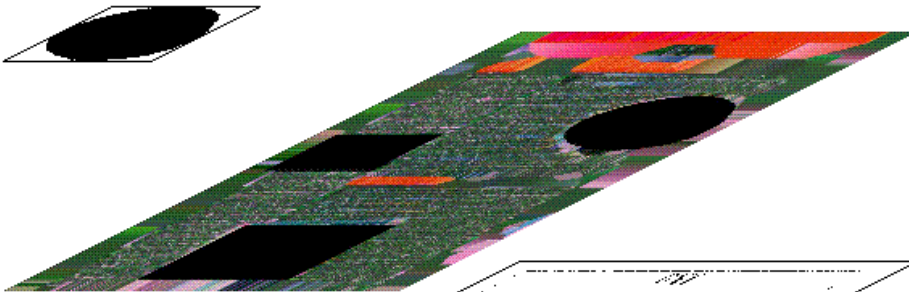
at Palo Alto Research Center  
**parc**



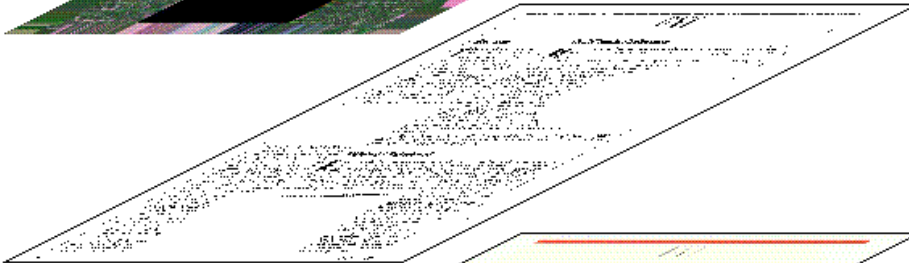
# DIR Segmentation of magazine page



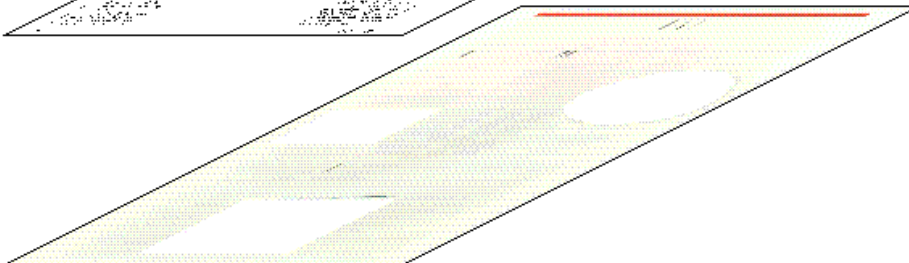
Layers #4 and #5: images  
(100dpi, JPEG and CCITT G4)



Layer #3: text tint  
(50 dpi, JPEG)



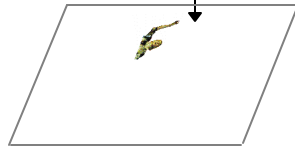
Layer #2: text  
(300dpi, G4 or token)



Layer #1: background  
(75 dpi, wavelet)

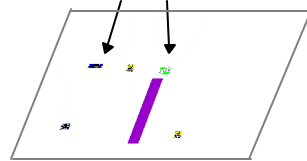
# DIR multi-layer representation example

Overlay image (+ transparency mask – not shown)

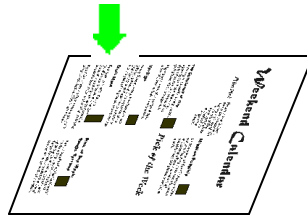


5<sup>th</sup> Layer (Includes an Overlay Picture)

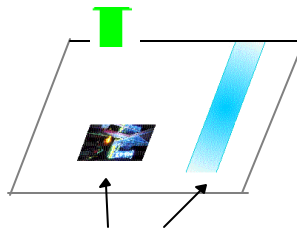
FG line art & color regions



Foreground layer



Selector layer



Background Layer

BG pictures



Composed Page

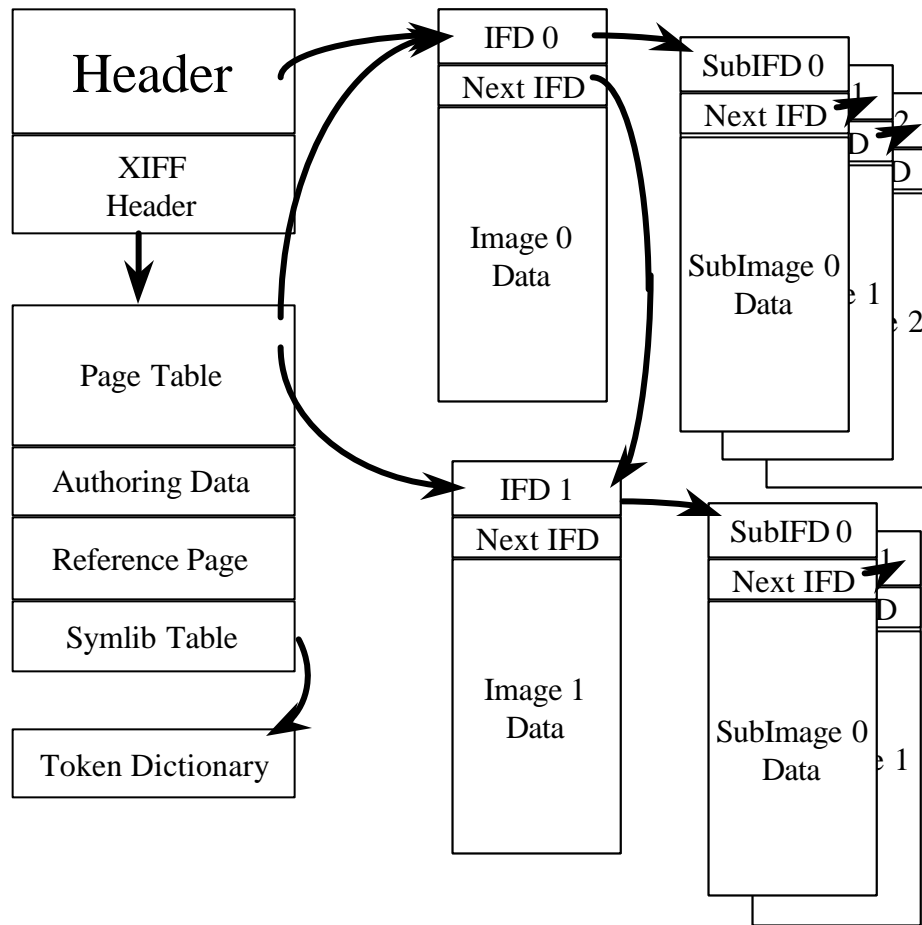
# Advantages of representation

- Compression ratios in excess of 200:1
- High quality scanned documents
- Small file size ideal for viewing, printing, archiving, and for the web
- Flexibility of file format: layer structure can be adapted to application
- No error-prone OCR required to obtain small file size from scanned color documents
  - OCR can still be used for indexing
- TIFF encapsulation: **XIFF**

# TIFF encapsulation: XIFF

- Until recently, no standard raster image file format supported multi-layer, multi-resolution, multi-compression scheme model
- XIFF designed as an eXtension of TIFF
  - Multiple layers
  - Support for new symbol based compression (JBIG2)
- XIFF actively promoted as *de facto* standard (in Pagis Pro software package) and *de jure* standard
  - ➔ Mixed Raster Content (MRC) model officially approved by ITU (International Telecommunications Union). Now being used in color fax machines
  - ➔ TIFF-FX, a variant of XIFF, is now officially recognized by the IETF (Internet Engineering Task Force)
  - ➔ Symbol based compression scheme used in XIFF is becoming new JBIG2 compression mode.

# Structure of a XIFF File

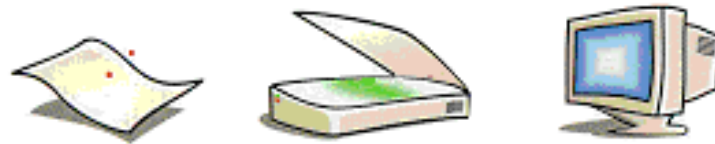


# Additional XIFF Features

- Benefits from name recognition associated with TIFF in the document world
- Support for annotations
  - Standard is open at the moment
  - Can be used for: thumbnails, “sticky notes”, highlighting, URLs, etc.
- Support for an arbitrary number of layers
- Direct page access via page table
  - Efficient browsing of large documents
- Support for token dictionaries used in symbol-based compression
- Built-in support for “stream based” application such as faxing
- Ideally suited for Web applications
  - Plug-ins and viewers available

# Pagis Pro software package

From the Pagis Web page at <http://www.pagis.com>  
or <http://www.xerox.com/xis/pagis/> :

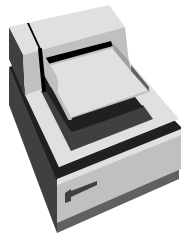


- The Best Way to Scan, Organize and Use Color Documents
- "Together with Office 97... and Pagis Pro... you've got the ultimate business application suite."
  - » 7/97 PC Computing - The Ultimate Office
- Pagis Pro97 is a *fully-featured scanning application* that allows you to scan documents into your Windows desktop. With a color, gray scale or binary scanner, you can easily *scan documents into your PC, then file, copy, print, send or use them* with your application.

# More on Pagis Pro

- Windows 95/98 or Windows NT 4.0 only
- Pagis Pro 1.0 released in late 1996, 2.0 in 98, and 3.0 in mid-99
- Bundled with TextBridge Pro OCR
- Underlying core technologies:
  - Document segmentation
  - Compression
  - Document image processing
  - Additionally: OCR (for indexing)
- Cost: between \$49 and \$99, including TextBridge Pro.
- Support most scanners

# The Pagis Solution



Scanner

Facsimile

Email, including attachments

Web documents

Digital cameras

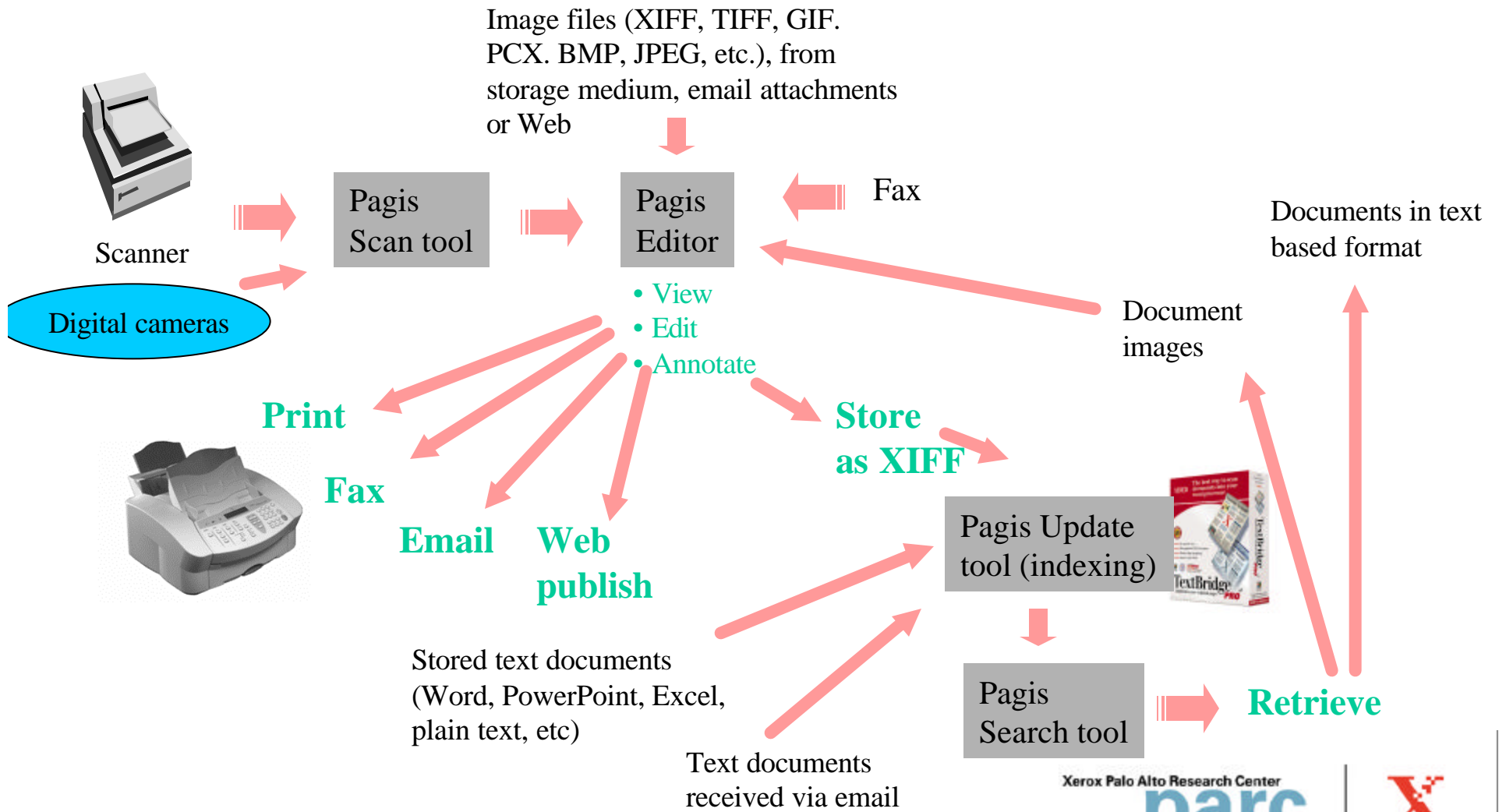


- Scan tool
- Editor
- Indexing engine
- Search tool



- Store
- Search and retrieve
- Edit
- Annotate
- Email
- Fax
- Print
- Copy

# Typical Workflow with Pagis



# Pagis use scenario #1

## *Scan and distribute*

- Use **Pagis scan tool** to scan magazine article into a XIFF file (typical size < 150k)
- Use **Pagis Editor** to cleanup and annotate XIFF file (add “sticky notes”, etc.)
- Send XIFF file as an email attachment. If recipient does not have Pagis, include URL of Pagis Viewer to message:  
[www.pagis.com/pagisviewer.htm](http://www.pagis.com/pagisviewer.htm)
- Recipients double-click on attached XIFF file to view it with **Pagis Editor** or **Free Pagis Viewer**.

# Pagis use scenario #2

## *Scan and Publish on Web*

- Use **Pagis scan tool** to scan any black and white or color document.
- **Drag and Drop** resulting XIFF file into Web editor such as Netscape Composer, FrontPage, etc.
- Document is automatically OCR'ed by **TextBridge Pro** and converted to editable HTML text with graphics.

# Pagis use scenario #3

## *Scan, Store, Retrieve, Use*

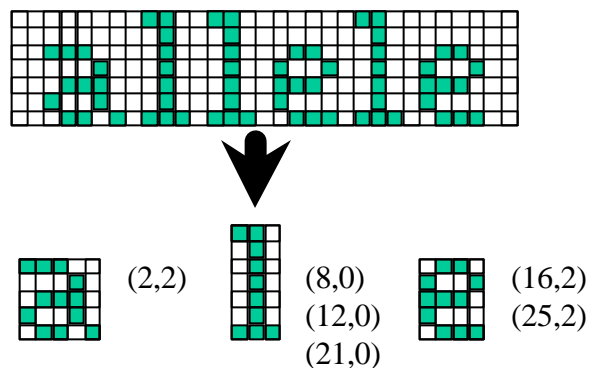
- Use **Pagis scan tool** to scan any interesting document you come across
- If needed, use **Pagis Editor** to clean-up and annotate resulting XIFF files
- Let **Pagis Update Tool** automatically index XIFF files, as well as Word, WordPerfect, PowerPoint, TIFF, etc. documents in selected folders
- Use **Pagis Search Tool** to retrieve documents from your disk
- View, print, email, OCR, or web-publish retrieved XIFF documents

# Symbol-based Compression

- Symbol-based Compression, also known as token-based compression, or “tokenization”
- Earliest token-based compression efforts date back to the seventies at AT&T
- *Lossy compression* technology that typically compresses 3 to 7 times better than CCITT Group-4
- **Principle:** repeating images (symbols) are stored in a Symbol Dictionary; the dictionary IDs and page positions are smartly encoded

# Principle of symbol-based compression

- Use *pattern matching* and clustering techniques to find classes of shapes (i.e., tokens, or characters)
- Compress page as:
  - **token dictionary**: list of shapes in the page
  - **position block**: where each token is found in the page



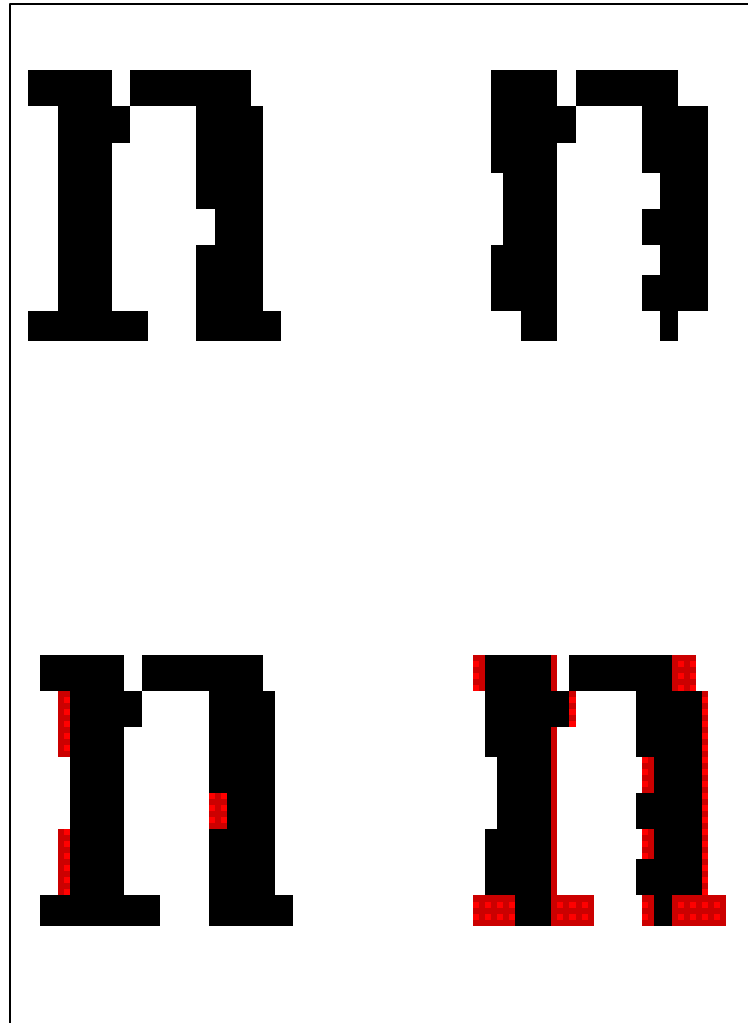
the computations, it also allows us to solve the accuracy problems encountered by most of the algorithms reviewed in Section II-D. First, the labeling of the catchment basins automatically avoids such traps as that of Fig. 7. Now, in order to get perfectly located watershed arcs, the successive geodesic SKIZ involved in the process have to be as good as possible. The first thing to notice is that, according to the discrete distance associated with the underlying grid, the set of pixels which are equidistant to two given connected components may well not be a line, but a very thick area. This is illustrated by Fig. 9. (Recall that the distance between two pixels is equal to the minimal number of grid edges to cross to go from one to the other.) Consequently, some simplistic rules in the computation of the geodesic SKIZ's could result in unwanted thick watershed areas. More precisely, suppose that the plateaus at elevation  $h$  are currently being flooded,

# Symbol Matching

- Based on Hausdorff distance:
  - candidate symbol compared to dictionary symbols
  - candidate dilated and aligned with dictionary symbol
  - dictionary symbol dilated and aligned with candidate
  - bit differences analyzed
- Optimized to avoid character substitution:
  - bit differences scrutinized about symbol interior, exterior more tolerant
- Designed to be tolerant of scanner noise (jaggedness) at the periphery of characters

# Basic Idea of Hausdorff Matching

*Characters to compare:*



*Alignment, dilation,  
analysis of peripheral  
pixels:*

# Symbol Compression Example



Original  
Binary  
Image

Repeating  
Symbols

Non-repeating  
Image  
Fragments

First  
Occurrence  
of Repeating  
Symbols

# Symbol Compression Components

- Symbol Dictionaries stored in *Symbol Libraries*, containing:
  - Symbol images of similar height grouped together in sorted width
  - Each height class is CCITT Group 4 compressed
  - Height class delta Huffman encoded.
  - Each dictionary symbol has unique ID
- Symbol location data, or *position blocks*:
  - List of X,Y locations with associated token ID
  - Symbol positions of lower left-hand corner, grouped in raster-scan order
  - X- and Y-positions delta Huffman encoded.

# Token-based Compression in XIFF

- Compression data consists of:
  - Position and index information
  - Token dictionaries
  - Residual
- Dictionaries can be shared across multiple pages
  - Reduced file size for multi-page documents
- Support for multiple dictionaries
  - Efficient handling of font changes in long documents
  - Concatenation of token-based documents does not require decompression-compression
- Representation is new **JBIG2** standard

# Easier and Better Printing

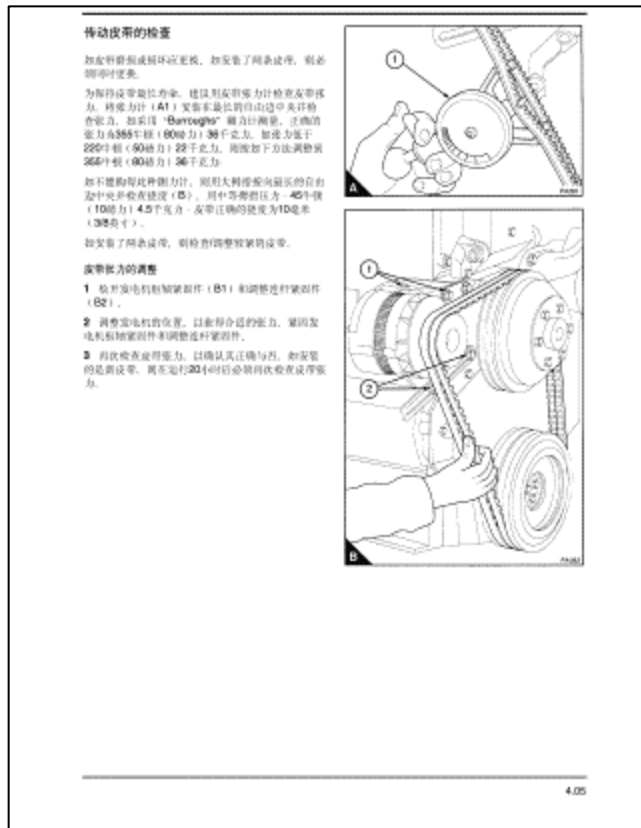
- Print what you view, quickly!
- High end – 180+ ppm at 600+ dpi
  - monochrome or “highlight” color
  - over 100 million pixels per second!
- Commodity desktop – low unit cost
  - less processing in the printer, but higher print engine speeds require faster rendering
  - using the faster desktop processor not viable due to data rates - without compression

# Example: A Scanned Article



- 33 page document scanned at 300 dpi binary
  - 3.5MB TIFF
  - 670KB DigiPaper/XIFF
- For web viewing, scale to 100dpi 2bit gray
  - 2.5MB TIFF
  - 413 KB DigiPaper/XIFF

# Example: Chinese Document



- 60 pages with many figures
  - minutes to RIP PostScript – large embedded fonts
  - prints at 180PPM with native DigiPaper/XIF decomposer
  - 67MB for PostScript source (7MB gzip compressed)
  - DigiPaper/XIF is 1MB at 300 dpi, 2MB at 600 dpi

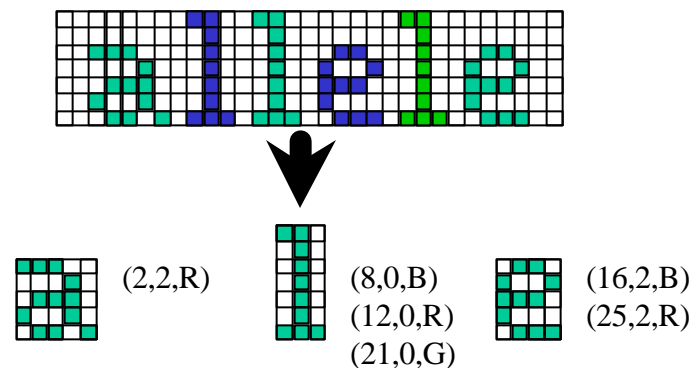
# Monochrome Performance Data

- 1320 reports scanned at 600 DPI
  - 3-341 pages long
    - 45KB-27MB (G4 TIFF)
  - 7x compression vs. G4 (between 1x and 30x)
    - 500MB vs. 3.5GB
  - encode at 2.3sec/page (from 0.3 to 10.6)
  - decode at .037sec/page (from .014 to .27 )

- 310 reports RIPped from PS at 600 dpi
  - 3-429 pages long
    - 43KB-45MB (PS2)
  - 2.3x compression vs. PS (between .55x and 26x)
    - 171MB vs. 385MB
  - encode at .34sec/page
    - “same” as RIP (ghostview)
  - decode at .016sec/page (from .008 to .09)

# Color Tokens

- Efficiently handle business graphics or colored text by tagging each instance with its color:
  - tag can specify a color or an image to mask
  - run-length code the Huffman coded color tags: often requires only a few bytes per page!
  - Consistent with the DIR multi-layer decomposition
  - share position block in mask and foreground



# MRC and JBIG2 are Frameworks

- Primarily specify what information must be stored
  - not algorithms for creating the information
- Data structures serialized to byte streams
  - can be embedded in standard file formats
    - e.g., TIFF
  - more broadly, same information can be stored in other forms without decoding to image
    - e.g., embeddings in PostScript or PDF rather than image formats

# Summary on Token Based Compression

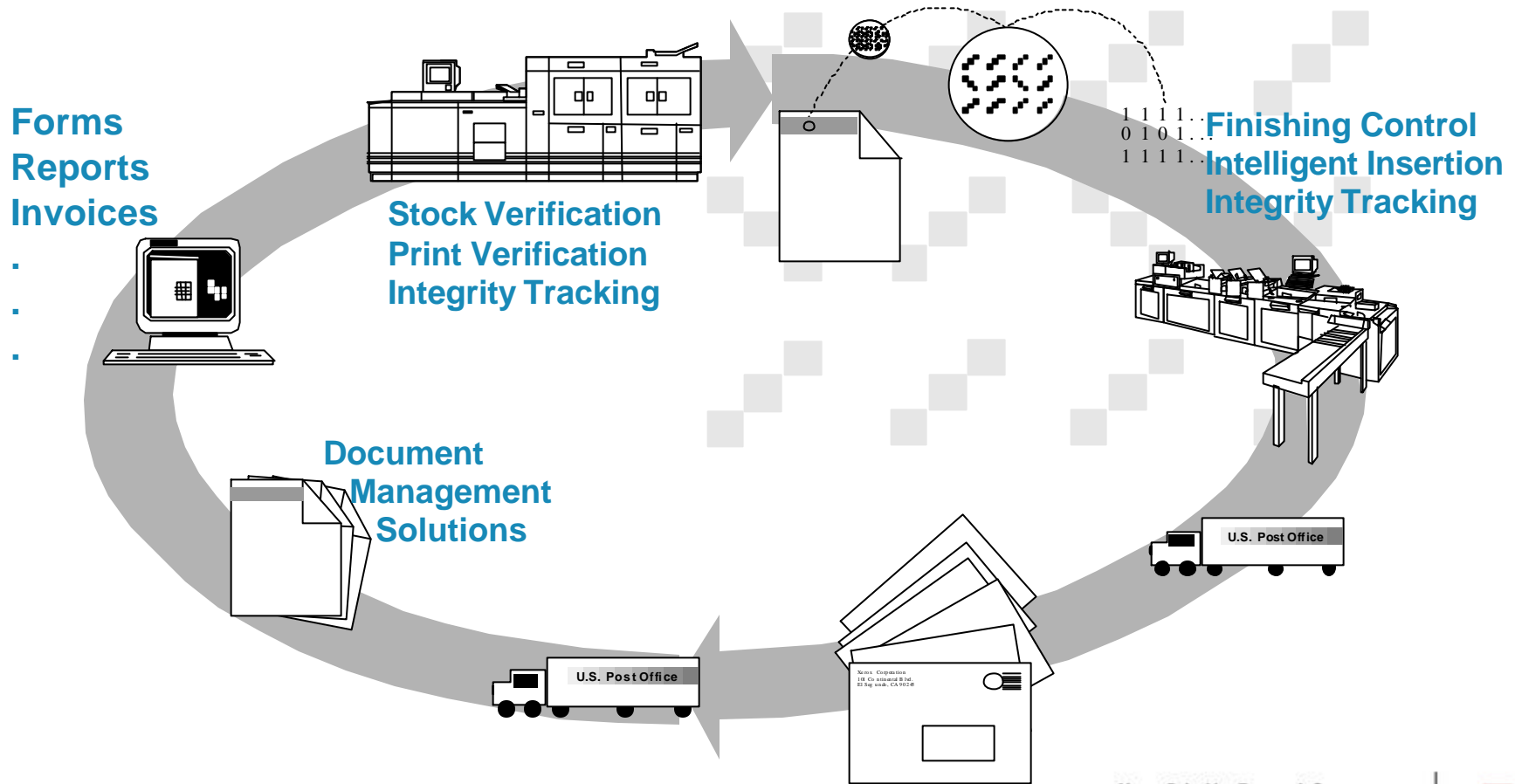
- Very high compression
- High efficiency viewing and printing
- Works for binary as well as color text/graphics
- No need to understand or recognize the symbols to compress a page
  - Works for any type of language or symbology

# Marking the Document

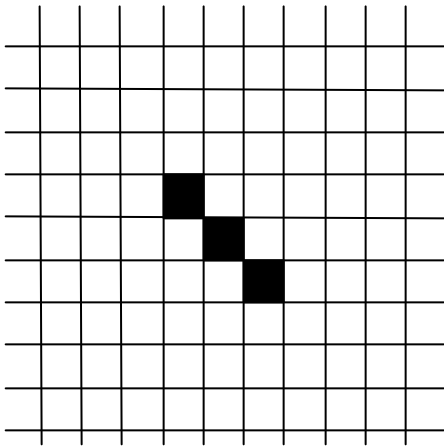
- Connect Paper to Computers
- Robustness
- Attractive marking
- High performance encoding/decoding
- Large capacity

=> DataGlyphs!

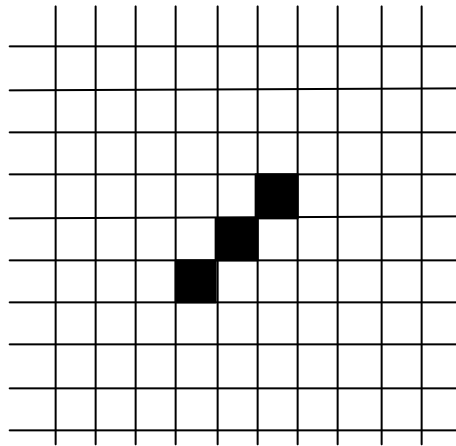
# Tracking the Document



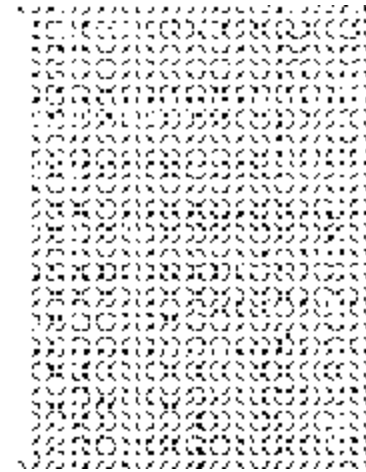
# DataGlyph Symbology



“0”

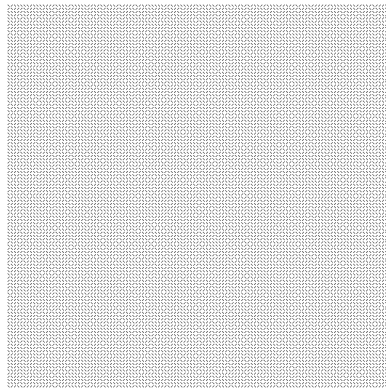


“1”



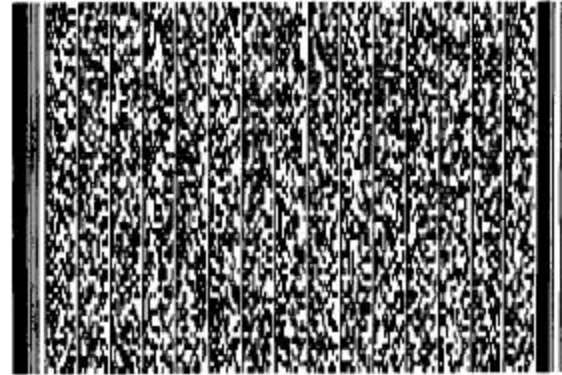
# Attractive and unobtrusive

2" x 2"  
5 x 5 Symbols  
300 DPI  
10% Error Correction



DataGlyph

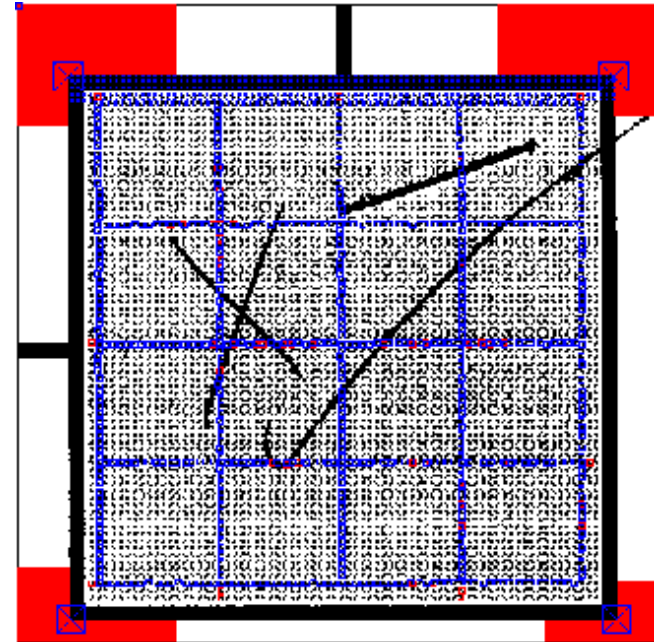
2" x 2.75"  
PDF - 417  
Nominal Symbol Size  
No Error Correction



2D Barcode

# Robustness

- Use of synchronization lines: self-clocking
- Reed-Solomon error correction built-in
- Data bits folded and interleaved
- Typical bit error rate: 1/10,000
  - Page error 1/100,000
- Tolerant to image degradations:
  - printer/scanner noise
  - damage marks
  - Coffee stains, streaks, drop outs
  - Faxing (even low-resolution) and copying



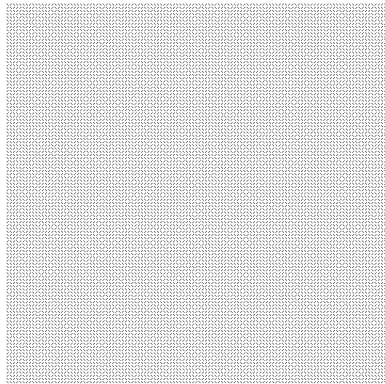
# High Capacity

## Lincoln's Address at Gettysburg, 1863

2" x 2"  
5 x 5 Symbols  
600 DPI  
10% Error Correction



2" x 2"  
5 x 5 Symbols  
300 DPI  
10% Error Correction



Fourscore and seven years ago our fathers brought forth on this continent a new nation, conceived in liberty and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation or any nation so conceived and so dedicated can long endure. We are met on a great battle field of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this. But, in a larger sense, we can not dedicate - we can not consecrate - we can not hallow - this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember, what we say here, but it can never forget what they did here. It is for us the living, rather, to be here dedicated to the great task remaining before us - that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion - that we here highly resolve that these dead shall not have died in vain - that this nation, under God, shall have a new birth of freedom - and that government of the people, by the people, for the people, shall not perish from the earth.

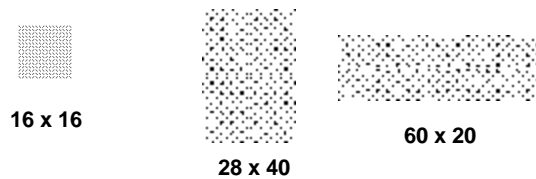
- Assuming 5-pixel glyph mark at 600dpi and 20% error correction, data density is about 1250 bytes per square inch
- Almost double the maximal data density of PDF-417
- An 8.5"x 11" page can hold about 100 Kbytes of data

# Performance

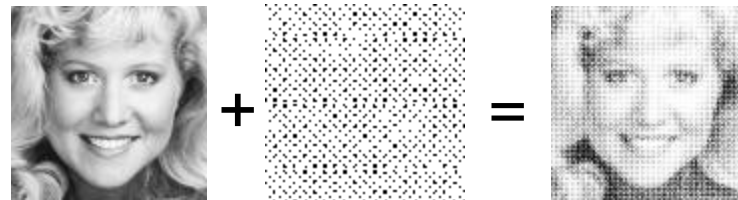
- High Rate Scanning
  - Significant skew tolerance
  - Get data from scanner, camera, handheld device, etc.
  - FormScan (a Xerox partner) scans and decodes 660+ glyphs zones per minute
- High Rate Printing
  - Glyph fonts
  - Xerox DocuPrint 180 prints glyphs at 180 ppm

# Beyond simple glyph rectangles

## Rectangular



## Halftone



## Logos



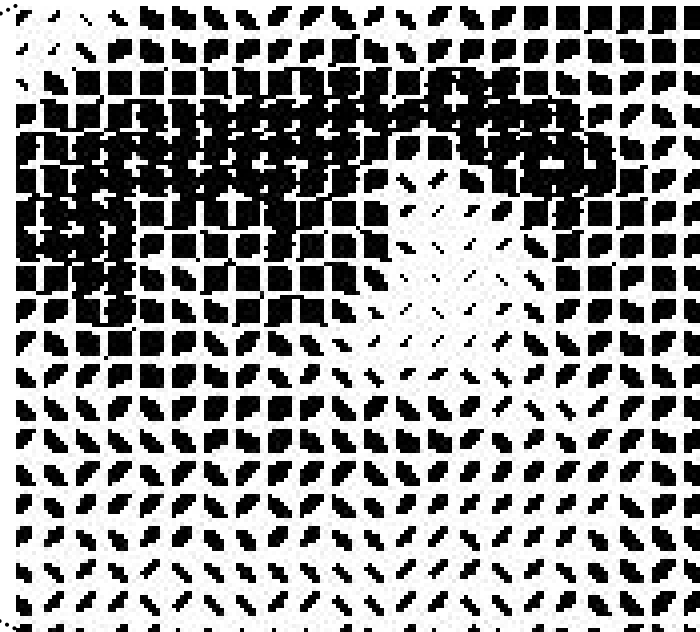
# GlyphTones

Principle: represent a grayscale image by varying the weight of the glyph marks.

Very light and very dark areas need special consideration.



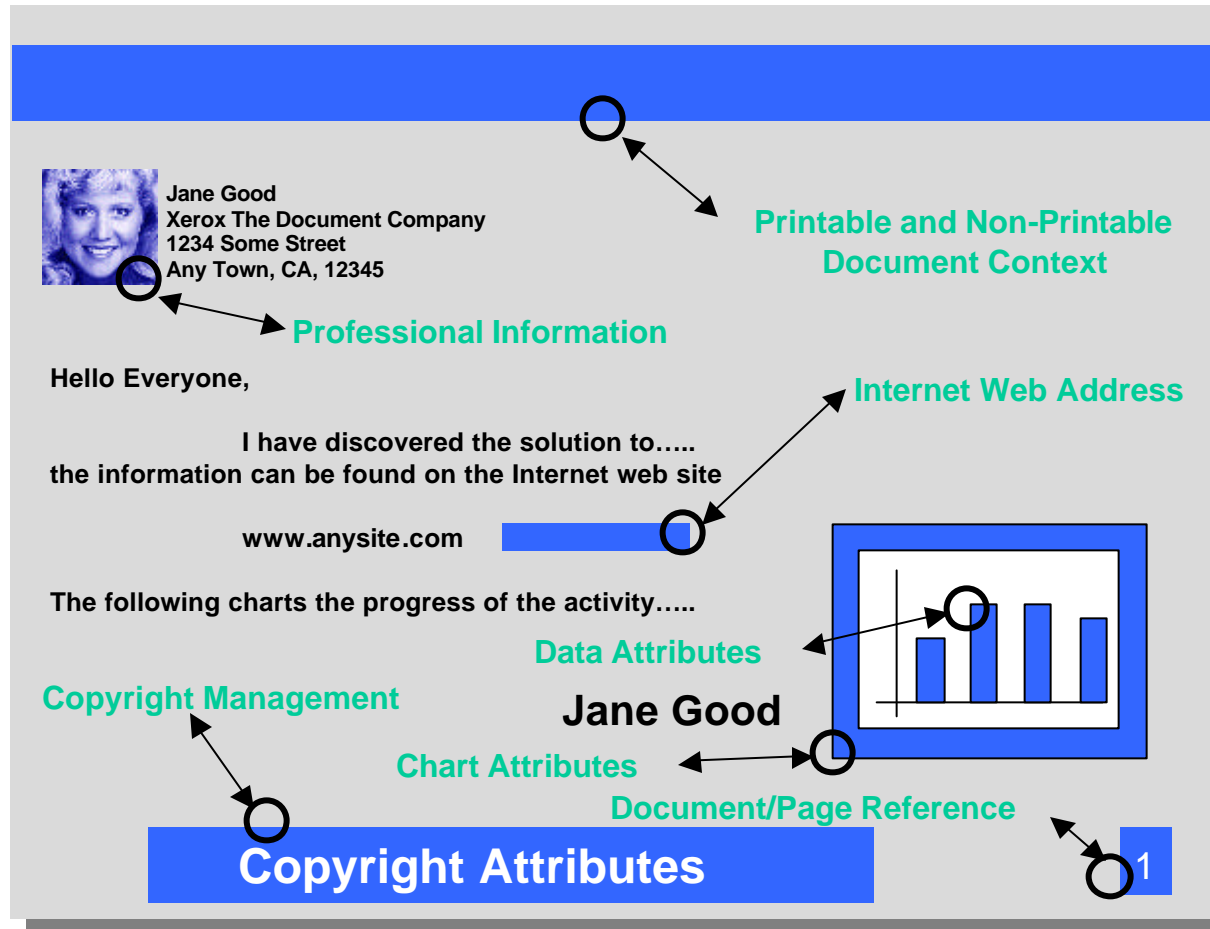
# Other Glyphtone example



# DataGlyph™ Applications

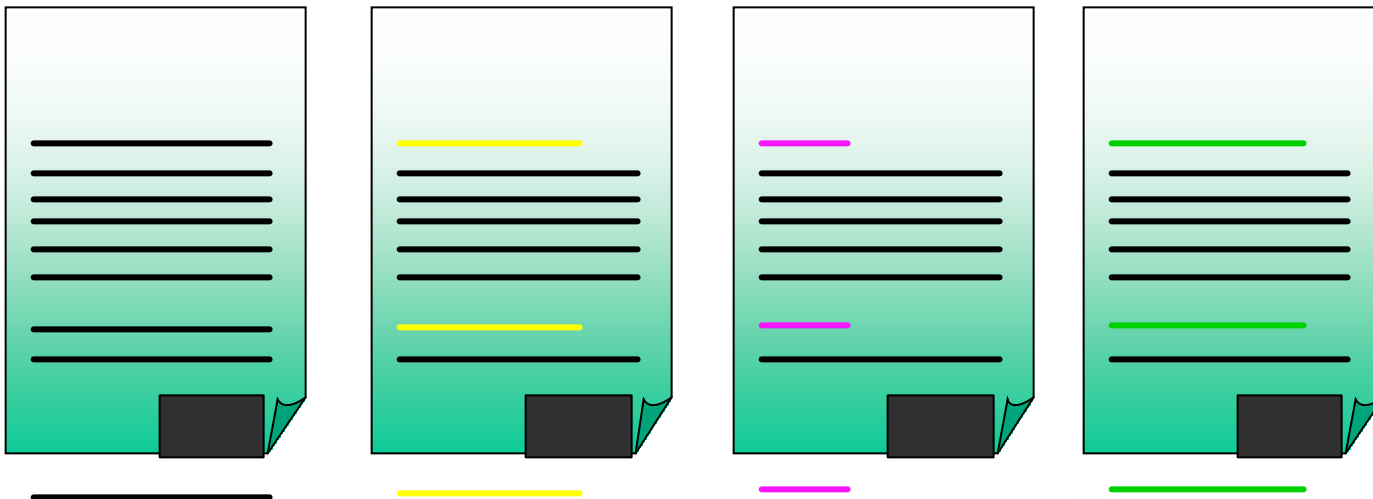
- Unobtrusive marking technology
- X-ray linking to patient report
- Parts tracking (aircraft, rolling stock, heavy machinery, etc.)
- Paper stock verification
- Effectiveness of mailers
- Auto processing of mailers (turn-around document)
- Demographics
- Remittance application
- PC Franking
- Document tracking/processing validation
- Cover sheet: describe job, shipping information
- Promotional Mail application
- Page number to validate document integrity/page ordering
- Use to control finishing equipment
- Cover-sheet for copier/scanner-based scan-and-ship

# Paper-Electronic Connectivity



# Simple Variable Data

- Mailings
- Multiple Choice Tests
- Personalized Documents
  - Insurance contracts, Sales contracts
- Bills, Invoices and Statements
- Numbering and Sequencing
  - Lottery tickets, Financial Instruments



# Example of Paper UI Form

**Cover Sheet** fabops  
Use this Cover Sheet to give the copier instructions. Jul 22 17:13:19 PDT 1999

---

## Burlap Cover Sheet v1.2

---

**How do I use this cover sheet?**  
Enter your 4-digit phone extension into the checkboxes below. Then, select the template you wish to use. Place this cover sheet on top of your document, place them in the feeder, and choose the FABRIC template on the Hodaka console.

---

**Who am I?**  
(That's a good question. Please select the checkboxes corresponding to your 4-digit phone extension; we will use it to determine your userid)

0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

---

**Where will my scan go?**  
The document you scan will be placed in your home directory in a Burlap container called 'bdmdocs'. For example, if your userid is jdoe, the container will be /fide/jdoe/bdmdocs.  
NOTE: This 'bdmdocs' directory must be made world-writable for the upload to work (for now, anyway).

---

**Check the box(es) of addition operations you would like performed.**

- Create text (OCR it).
- Create a summary.
- Make it PUBLICALLY searchable.



---

**How should I be notified? (optional)**

- Print a confirmation sheet (may take a few minutes)

(To see your document, you can go to <http://burlap.parc.xerox.com> and log in to burlap.)

---

 Re-fetch      Help

# Merryl Lynch example

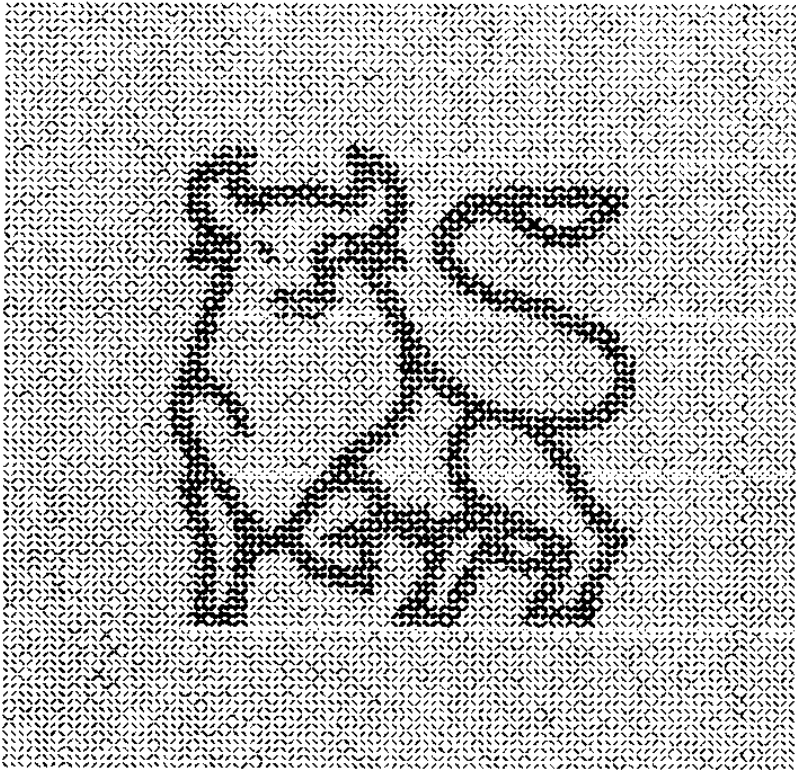


+

NEW YORK, Dec. 1 -- Merrill Lynch today launched an expansive online investing website as the latest component in one of the world's most complete packages of personal financial services for U.S. investors. "We're tremendously excited to be launching Merrill Lynch DirectSM," said David H. Komansky, Chairman and Chief Executive Officer. "Backed by the full global resources of Merrill Lynch, Merrill Lynch Direct combines content, intelligence and innovation to create the smartest place for the self-directed client to invest online. We are dedicated to the proposition that when our clients succeed, we succeed, and Merrill Lynch Direct completes our platform of choice for clients. No matter how you may wish to approach the market - whether by working with a professional Financial Consultant or self-directing a financial portfolio online - you can do it at Merrill Lynch."

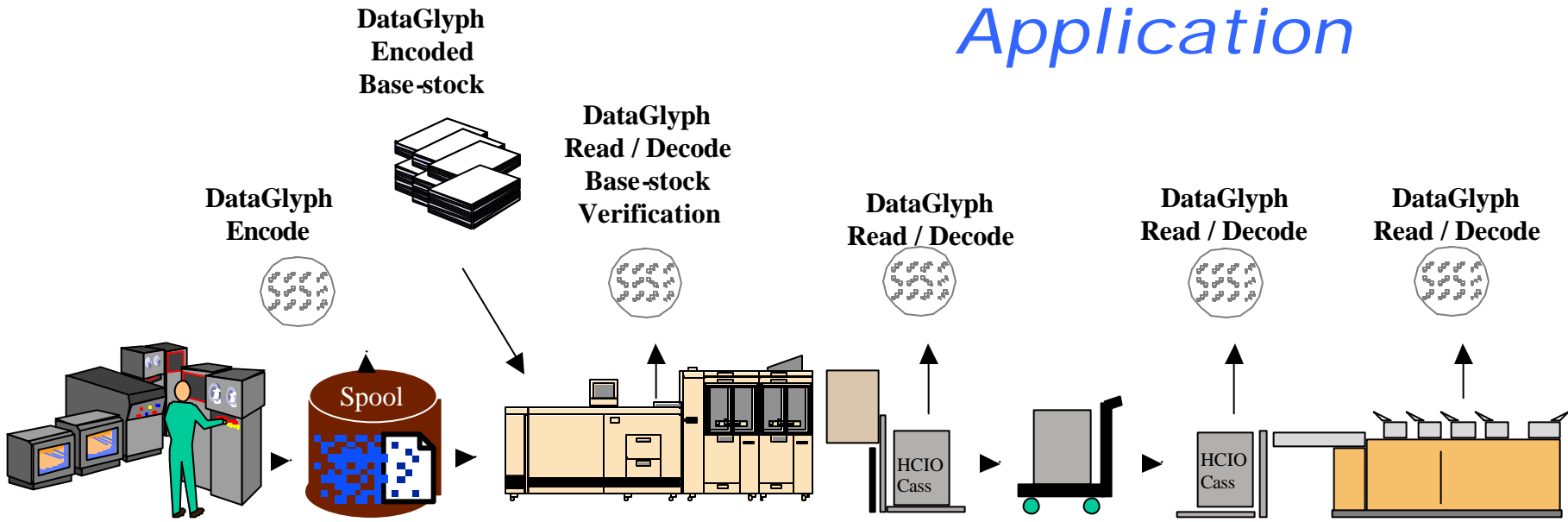


# Merryl Lynch DataGlyph Logo



# The Automated Document Factory

## A DataGlyph™ Application




**Achieve  
99.9999%  
Reliability  
& Job  
Integrity**

**Enable  
1:1  
Marketing**

# DocuStamps™ for Paperware™

## Distribution Services



**Bridget Wu**  
 Member Technical Staff  
 Xerox / ADP/EC  
 3490 Hillview Ave.  
 Palo Alto, CA 94304  
 Phone: 650.813.7904  
 Fax: 650.813.7160  
 http://parc.xerox.com


### XEROX

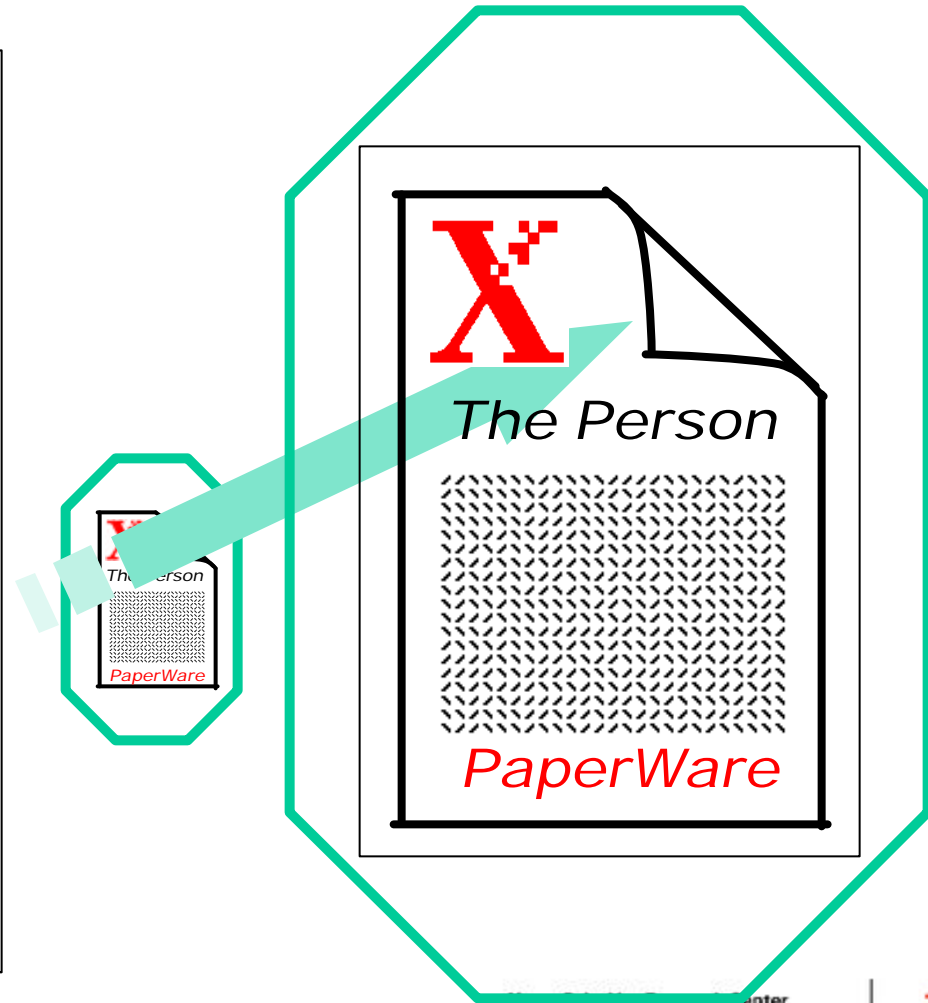
#### Providing the Digital Document Connection

Xerox provides the seamless transition, transaction and interoperability between Paper and Electronic Document domains. The "document" definition and use now extends beyond the simple separation of the "digital electronic" document and the "printed" document, such that, the digital requirement now exists in both "document" environments. Xerox provides the digital connectivity through the use of Xerox DataGlyphs®. Xerox DataGlyphs® provide the ability to encode and decode digital information that can be graphically integrated with today's business documents while exceeding data capacity of other one-dimensional or two-dimensional symbols. Originally invented at the Xerox Palo Alto Research Center (PARC), the technology was developed by Xerox's Corporate Technology Centers and is now an available product offering as a developer's toolkit, ready for implementation with customer / consumer products. As an illustration, Xerox DataGlyphs® have been implemented over a diverse range of applications including identification cards, digital x-ray film, and high-volume production print management.

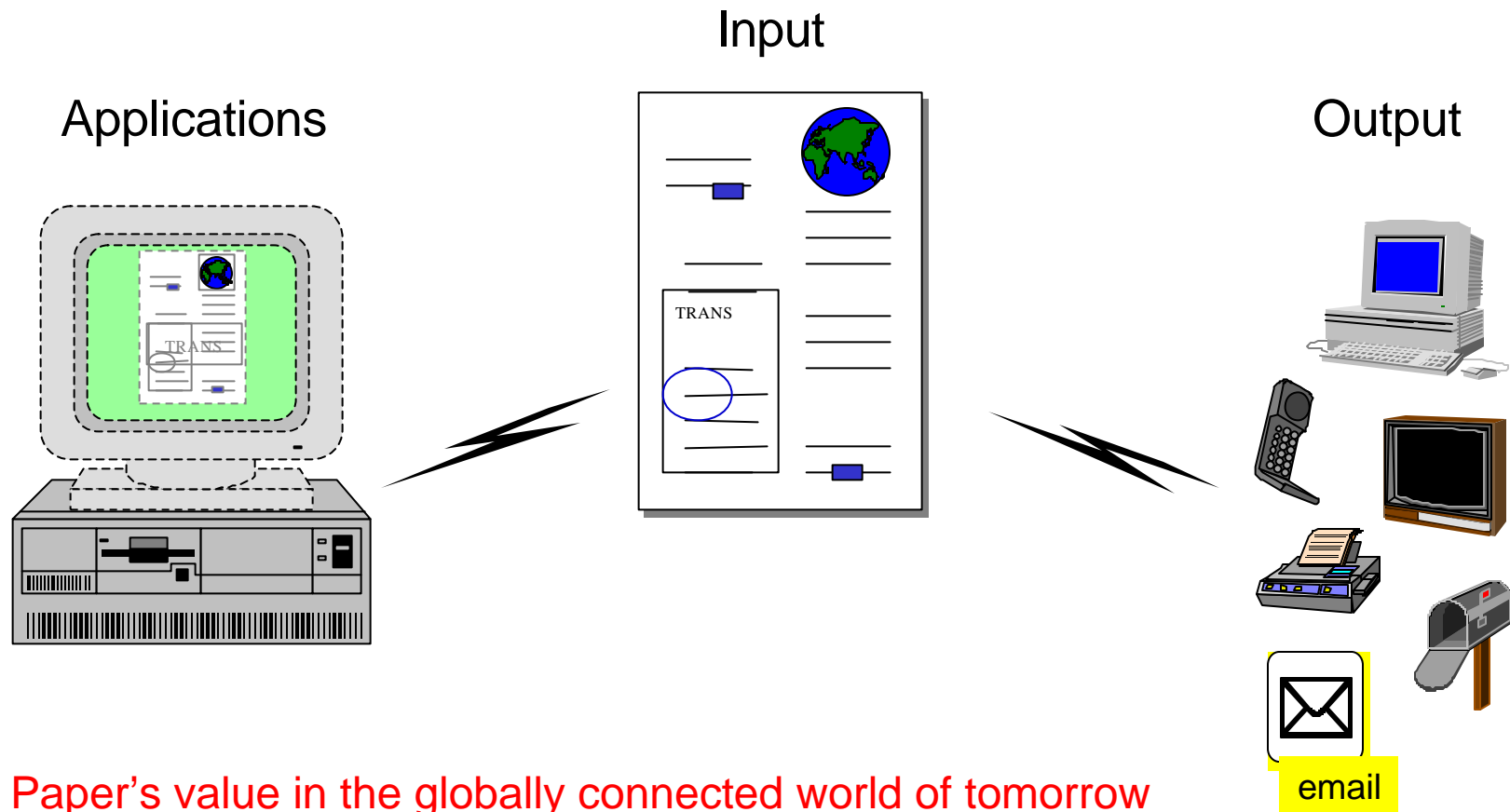
Contact : Xerox Marketing --- 1-800-xxx-xxxx


--- Xerox DataGlyph® ---  
 --- Document Application Environment ---  
 --- PaperWare ---





# Future Applications



Paper's value in the globally connected world of tomorrow

Control from paper anywhere

Invoke any application

Receive / send output to any device

# Summary

- Set of technologies that connect paper and electronic documents
- Enable a range of new applications
- Token-compression + Dataglyphs = new way to deal with security in scanned documents