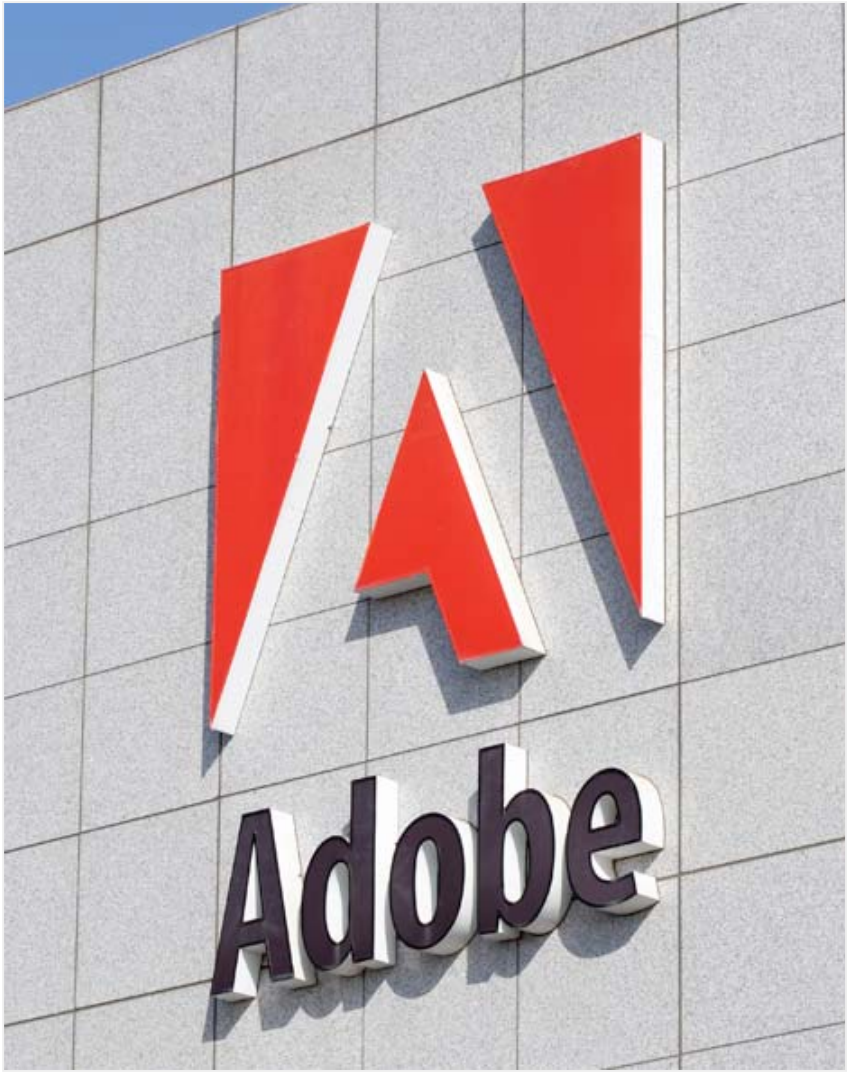
Document Formats and Image Formats

James C. King
PDF Architect/Senior Principal Scientist
Advanced Technology Laboratory
Adobe Systems Incorporated



Adobe

Outline

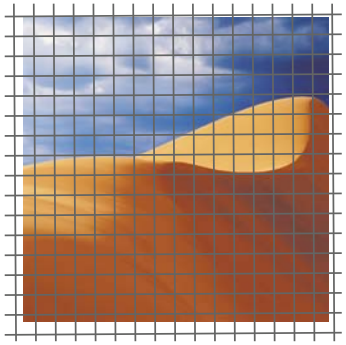
- Some Fundamentals
- PDF Documents
- PDF Pages
- Synthesized Pages versus Scanned Pages
- PDF and JPEG2000
- PDF and ISO Standards

Adobe

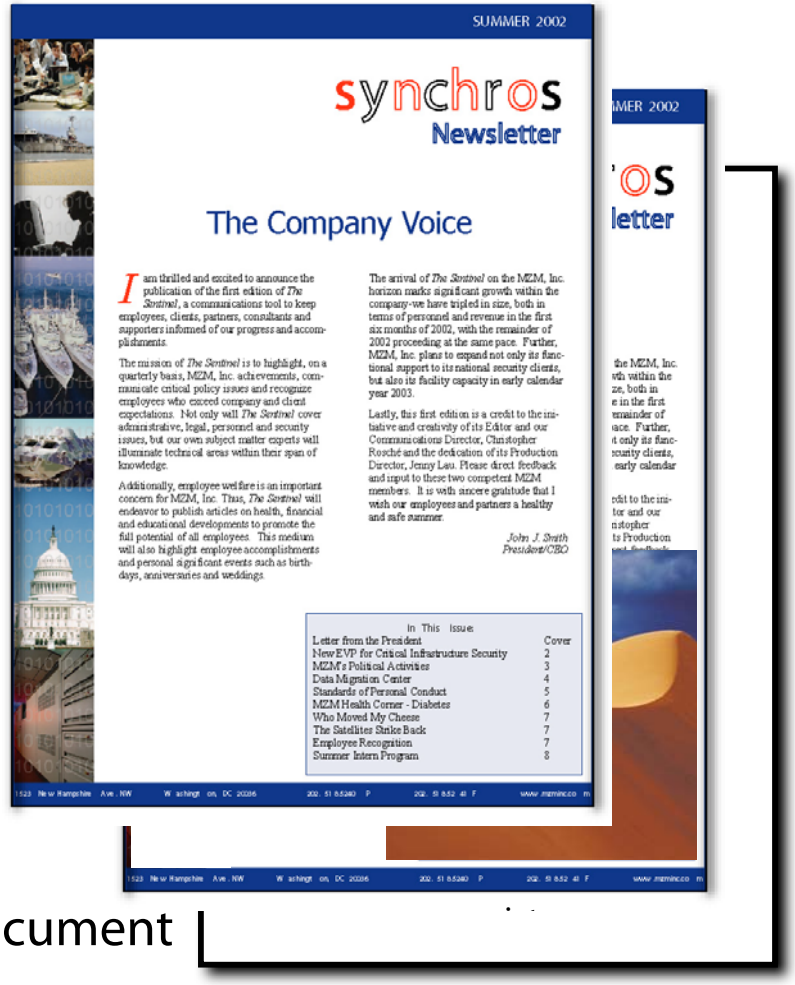
Some Fundamentals

Adobe

Image Formats versus Document Formats



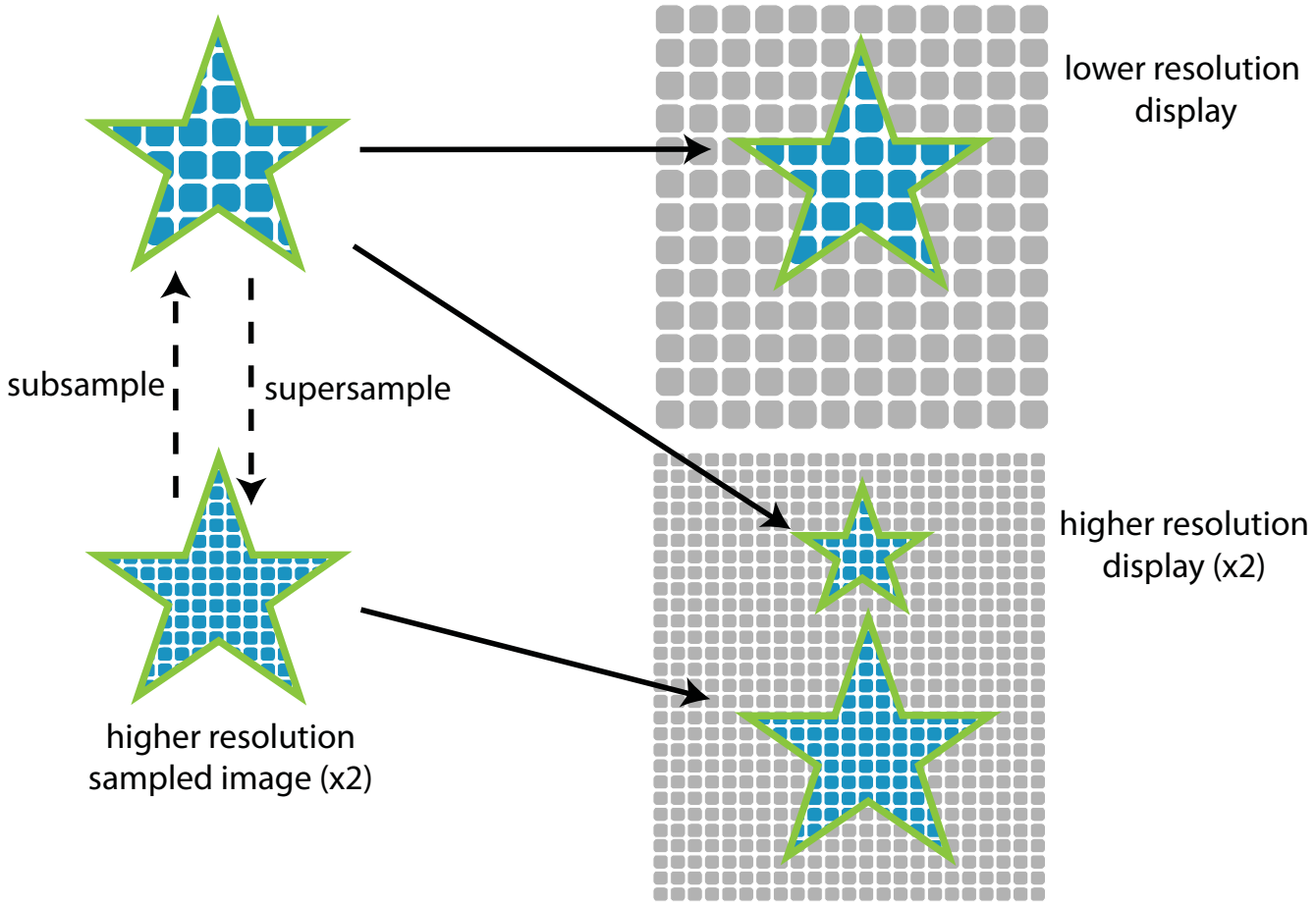
"Sampled" Image
(e.g., JPEG2000)



Multi-page Compound Document
(e.g., PDF)

Adobe

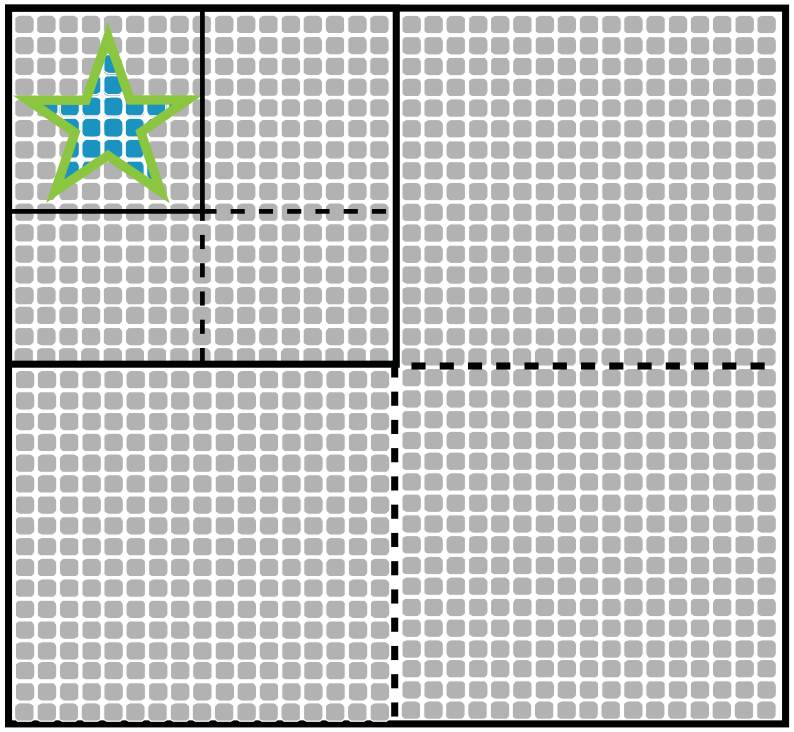
Image Resolution and Size



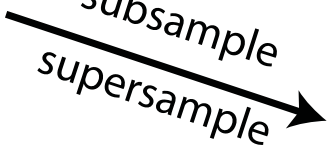
Adobe

Image Sampling (JPEG2000)

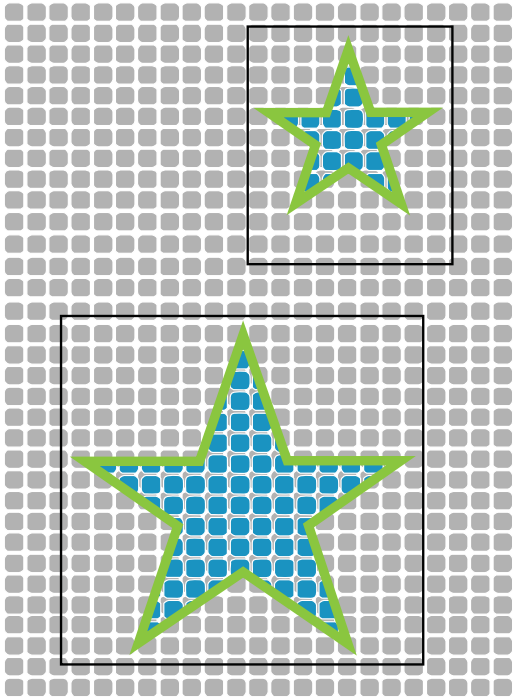
- Sub and Super Sampling Tools needed
  - Size and resolution are different things



JPEG2000 Image
(multiple resolutions)



display or page
(arbitrary image size)



Adobe

PDF Documents

Adobe

PDF: Multi-page Compound Documents

- A Comprehensive Format for Representing Documents and Forms
  - Page contents
  - Images
  - Graphics
  - Fonts
  - Colorspaces
  - Metadata
  - Annotations
  - Links
  - Digital signatures
  - <and more>
- Not an image format like TIFF or JPEG
- High fidelity, high precision text layout and graphics features
- Platform and device independent definition
- Selective compression to reduce file size (e.g., image formats)
- Color Management (ICC support)

PDF 1.0 in 1993 … PDF 1.7 in 2006. Many enhancements!

Adobe

Composite Documents

SUMMER 2002

synchros
Newsletter

The Company Voice

I am thrilled and excited to announce the publication of the first edition of *The Sentinel*, a communications tool to keep employees, clients, partners, consultants and supporters informed of our progress and accomplishments.

The mission of *The Sentinel* is to highlight, on a quarterly basis, MZM, Inc. achievements, communicate critical policy issues and recognize employees who exceed company and client expectations. Not only will *The Sentinel* cover administrative, legal, personnel and security issues, but our own subject matter experts will illuminate technical areas within their span of knowledge.

Additionally, employee welfare is an important concern for MZM, Inc. Thus, *The Sentinel* will endeavor to publish articles on health, financial and educational developments to promote the full potential of all employees. This medium will also highlight employee accomplishments and personal significant events such as birthdays, anniversaries and weddings.

The arrival of *The Sentinel* on the MZM, Inc. horizon marks significant growth within the company-we have tripled in size, both in terms of personnel and revenue in the first six months of 2002, with the remainder of 2002 proceeding at the same pace. Further, MZM, Inc. plans to expand not only its functional support to its national security clients, but also its facility capacity in early calendar year 2003.

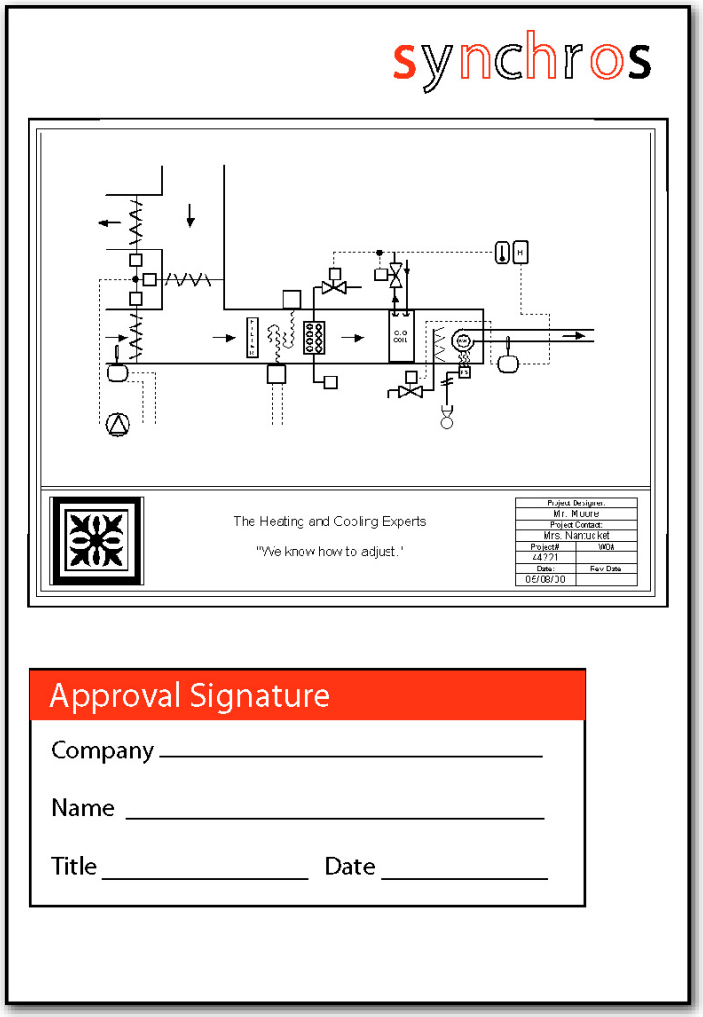
Lastly, this first edition is a credit to the initiative and creativity of its Editor and our Communications Director, Christopher Rosché and the dedication of its Production Director, Jenny Lau. Please direct feedback and input to these two competent MZM members. It is with sincere gratitude that I wish our employees and partners a healthy and safe summer.

John J. Smith
President/CEO

In This Issue:

Letter from the President	Cover
New EVP for Critical Infrastructure Security	2
MZM's Political Activities	3
Data Migration Center	4
Standards of Personal Conduct	5
MZM Health Corner - Diabetes	6
Who Moved My Cheese	7
The Satellites Strike Back	7
Employee Recognition	7
Summer Intern Program	8

1523 New Hampshire Ave., NW Washington, DC 20036 202.518.5240 P 202.518.5241 F www.mzminc.com



synchros

The Heating and Cooling Experts

"We know how to adjust."

Project Designer:	
Mr. Moore	
Project Contact:	
Mrs. Nantucket	
Project#	WO#
4771	
Date:	Rev Date
06/08/30	

Approval Signature

Company _____

Name _____

Title _____ Date _____

Adobe

PDF Pages

Page Content Objects

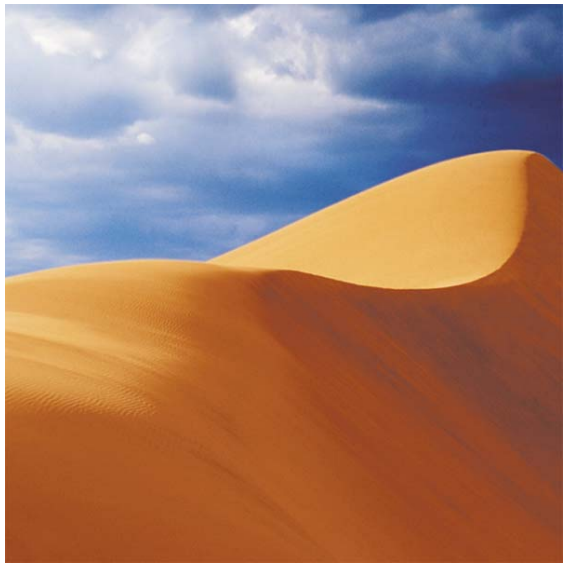
Adobe

Text, Graphics and Image

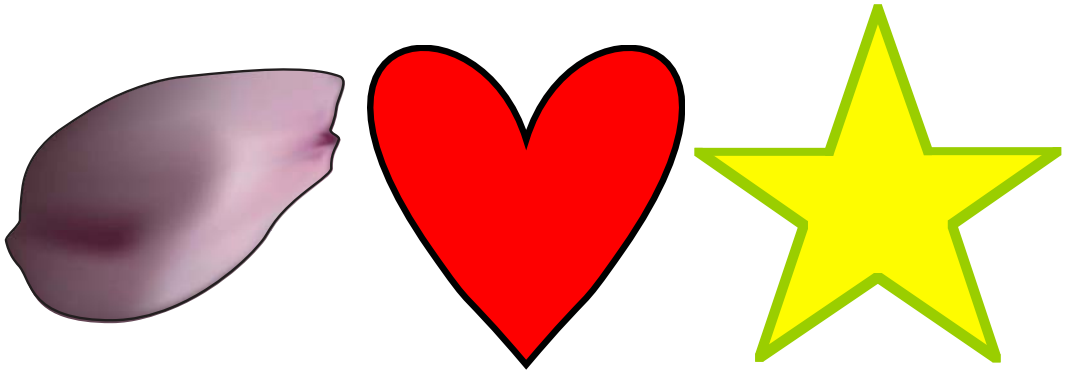
Typographic Text

Typographic Text

Sampled Images



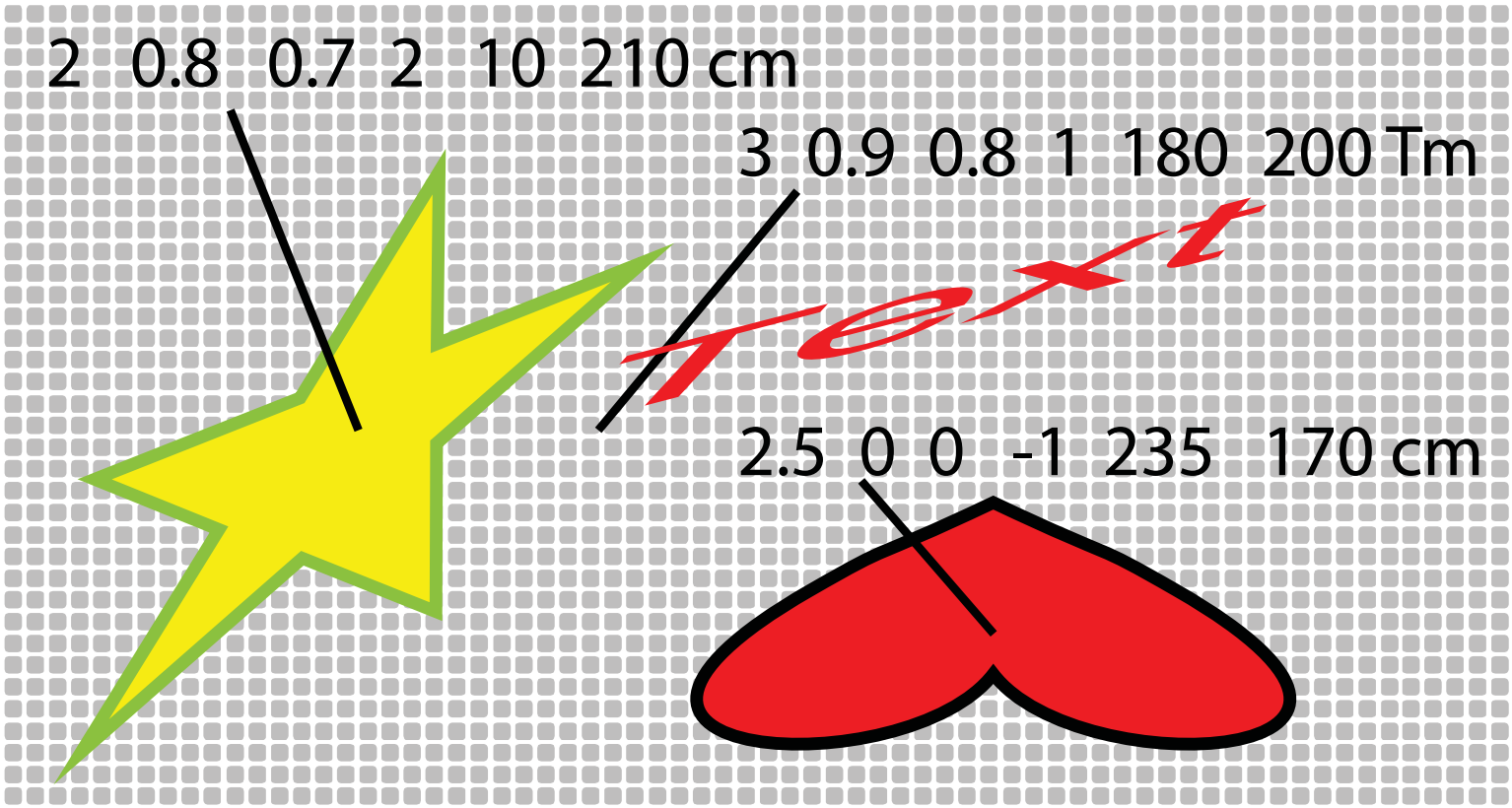
Vector Graphics



Adobe

Coordinate Transforms

- x-scale, rotate/skew, rotate/skew, y-scale, x-pos, y-pos



Adobe

Clipping and Masking

Picture

Mask

Clip to path (star)

Mask off sky

Adobe

Text as Text

The JPEG2000 image compression technique has been cited by experts as a new archiving format for digital images. It is both a preservation and delivery format, and has been seen as a possible alternative to the TIFF format which most institutions use as a long-term archiving standard. Produced by both imaging experts and the Joint Photographic Experts Group, it is now a recognised ISO standard. The standard JPEG file format which is so widely in use is not yet an ISO standard.

Text as text
(using outline fonts)

Text as image

Adobe

Various Resolutions for Image Text

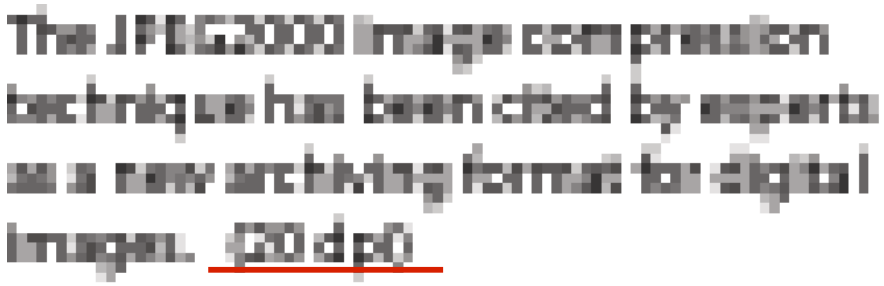
The JPEG2000 image compression technique has been cited by experts as a new archiving format for digital images. (outlines)

The JPEG2000 image compression technique has been cited by experts as a new archiving format for digital images. (96 dpi)

The JPEG2000 image compression technique has been cited by experts as a new archiving format for digital images. (300 dpi)

The JPEG2000 image compression technique has been cited by experts as a new archiving format for digital images. (60 dpi)

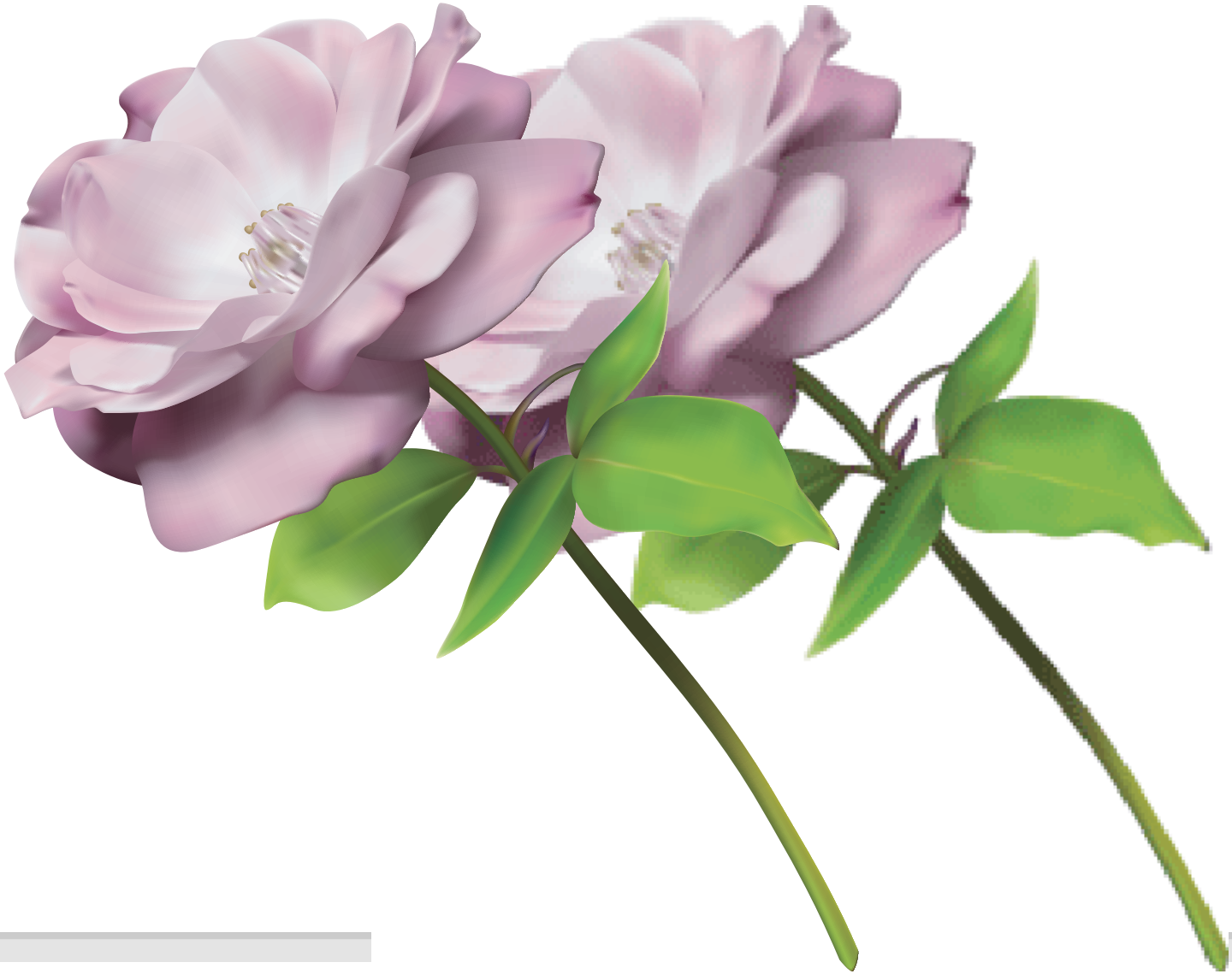
The JPEG2000 image compression technique has been cited by experts as a new archiving format for digital images. (150 dpi)



9.5 in x 5.3 in

Adobe

Resolution Independence



Adobe

Resolution Independence

Adobe

Synthesized Pages versus Scanned Pages

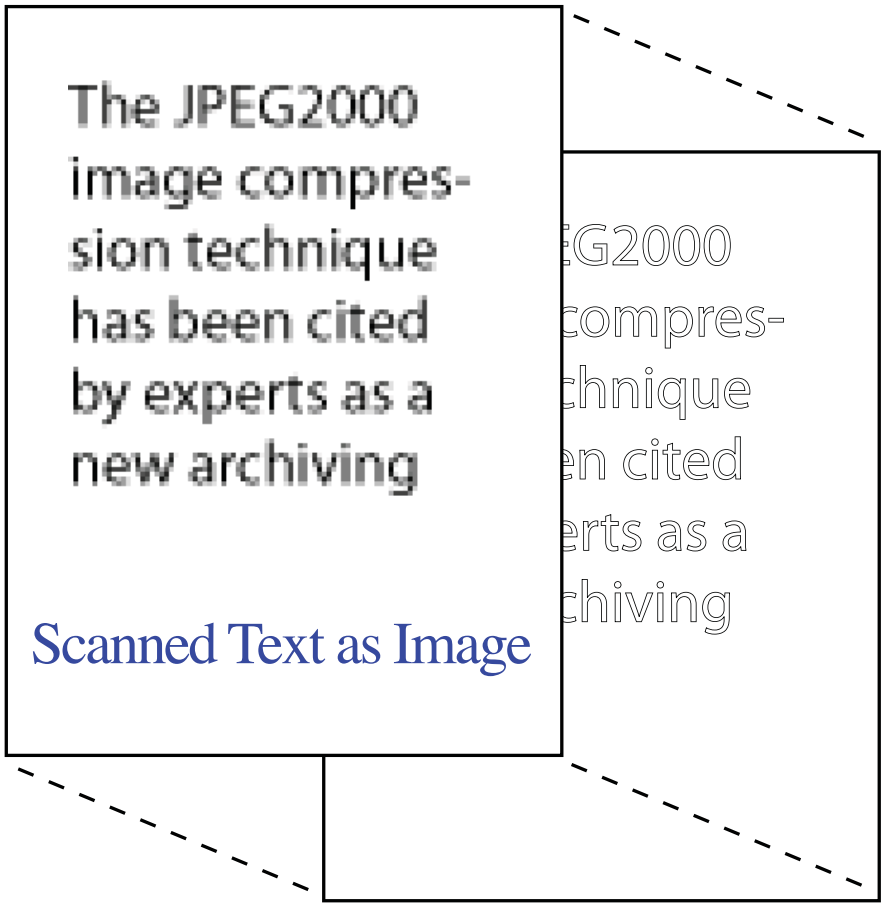
Adobe

Document Sources

- Born digital
  - More compact
  - Editable
  - Device independent/resolution independent
    - Zoom-able
- Scanned from paper
  - Bulky
  - Need to pick a sampling resolution
  - Text and image need different treatment
  - Can do OCR or DR (document recognition)
- Born digital is a luxury

Adobe

OCR'ed Text as Underlayer



A PDF Page

OCR'd Text

- underlaid
- made invisible
- may have mistakes
- used for search

Adobe

Image Text and Image Pictu Require Different Treatment

The JPEG2000 image compression technique has been cited by experts as a new archiving format for digital images.

The standard JPEG file format which is so widely in use is not yet an ISO standard.

Needs 1-bit per pixel black and white at 600 dpi

Needs 24-bit per pixel color at 150 dpi

- MRC (Mixed Raster Content)
- Both JPEG2000 and PDF support this

Adobe

PDF and JPEG2000

Adobe

PDF Support for JPEG2000

- JPEG2000 images can be included on PDF pages
  - JPX Baseline is supported
    - Enumerated color spaces 19 (CIEJab) not supported
    - Enumerated color space 12 (CMYK) is supported
  - All four progressions supported: resolution, color depth, band, location
    - Inappropriate progression will just cost time
  - One global soft mask within the JPEG2000 supported
  - JPEG2000 document features are not supported
    - PDFs document features are more general and more flexible
    - An image to display in a rectangle is obtained from the JPEG2000 stream

Adobe

Software Support

- Key to use of any image format or document format are the tools available
  - Tools for creation
    - support advanced features
  - Tools for presentation
  - Tools for incorporating with other formats
  - Ubiquity of viewing tools
  - OCR and DR capabilities

Adobe

Tools for Scan to PDF

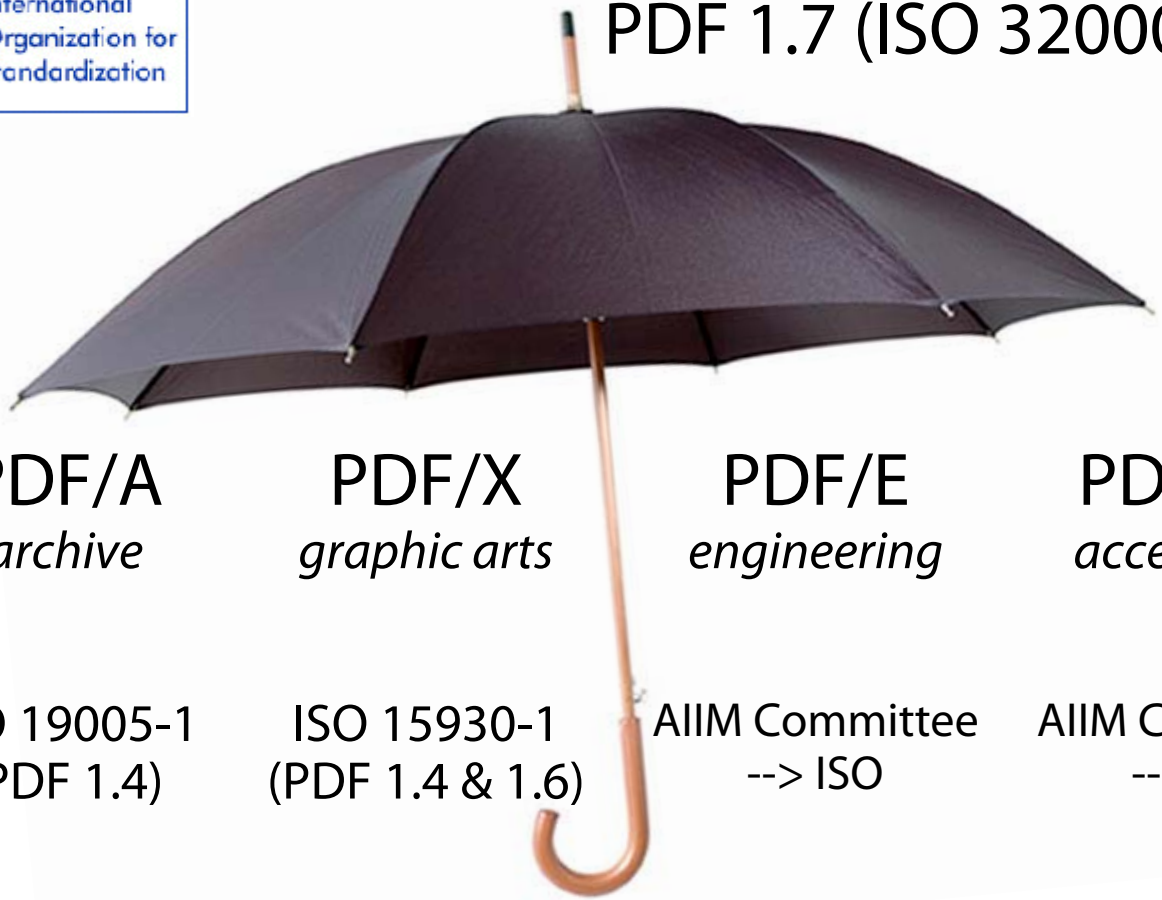
- Tools that separate image text and image pictures (MRC)
  - Adobe Professional Create from Scanner
  - Adobe PDF Scan Library 3.0 (OEM product)
  - CVision Technologies (www.cvisiontech.com) option for JPEG2000
  - Canon desktop printers and multi-functions devices (some)
  - Iris (www.irislink.com) MRC called IHQ
  - LuraTech (www.luratech.colm) MRC using JPEG2000
  - Nuance (www.thedevilincarnate.com)
  - JRAPublish (Jim Rile)
  - Spigraph (www.spigraph.fr)
  - VeryPDF (www.veryPDF.com)
- Of course the ubiquitous Adobe Reader presents them all

Adobe

PDF and ISO Standards

Adobe

Establishing the ISO PDF Umbrella



PDF 1.7 (ISO 32000 in 2008

PDF/A	PDF/X	PDF/E	PDF/UA
archive	*graphic arts*	*engineering*	*accessibility*
ISO 19005-1 (PDF 1.4)	ISO 15930-1 (PDF 1.4 & 1.6)	AIIM Committee --> ISO	AIIM Committee --> ISO

PDF/A

A PDF subset for archiving

ISO 19005-1

Adobe

Long-term Preservation Needs for Electronic Documents

- Characteristics identified as objectives for PDF/A were
  - **Device Independent** - Can be reliably and consistently rendered without regard to the hardware or software platform
  - **Self-contained** - Contains all resources necessary for rendering
  - **Self-documenting** - Contains its own description
  - **Unfettered** - Absence of technical file protection mechanisms
  - **Available** - Authoritative specification publicly available
  - **Adoption** - Widespread use may be the best deterrent against preservation risk

Adobe

PDF/A -- A PDF Subset of PDF 1.4
(Standard: ISO 19005-1)

- Some useful PDF features work against, and are incompatible with, preserving information over the long-term
- PDF/A
- PDF Subset: restricted from using some PDF features, for example
  - Anything that would alter the visual appearance over time (forms)
  - No external references or embedded files
  - Encryption
- PDF Subset: required to use some PDF features, for example
  - Accessibility features for recoverable text (tagged PDF)
  - Embed all fonts
  - Specific metadata requirements
  - Device independent color

Adobe

Uses for PDF/A

- Archival storage of electronic documents
  - Documents of record
  - Government records
  - Corporate records
- Distributing read only material
- Documents with assured accessibility (read to the blind)

Adobe

Better by Adobe™

Adobe