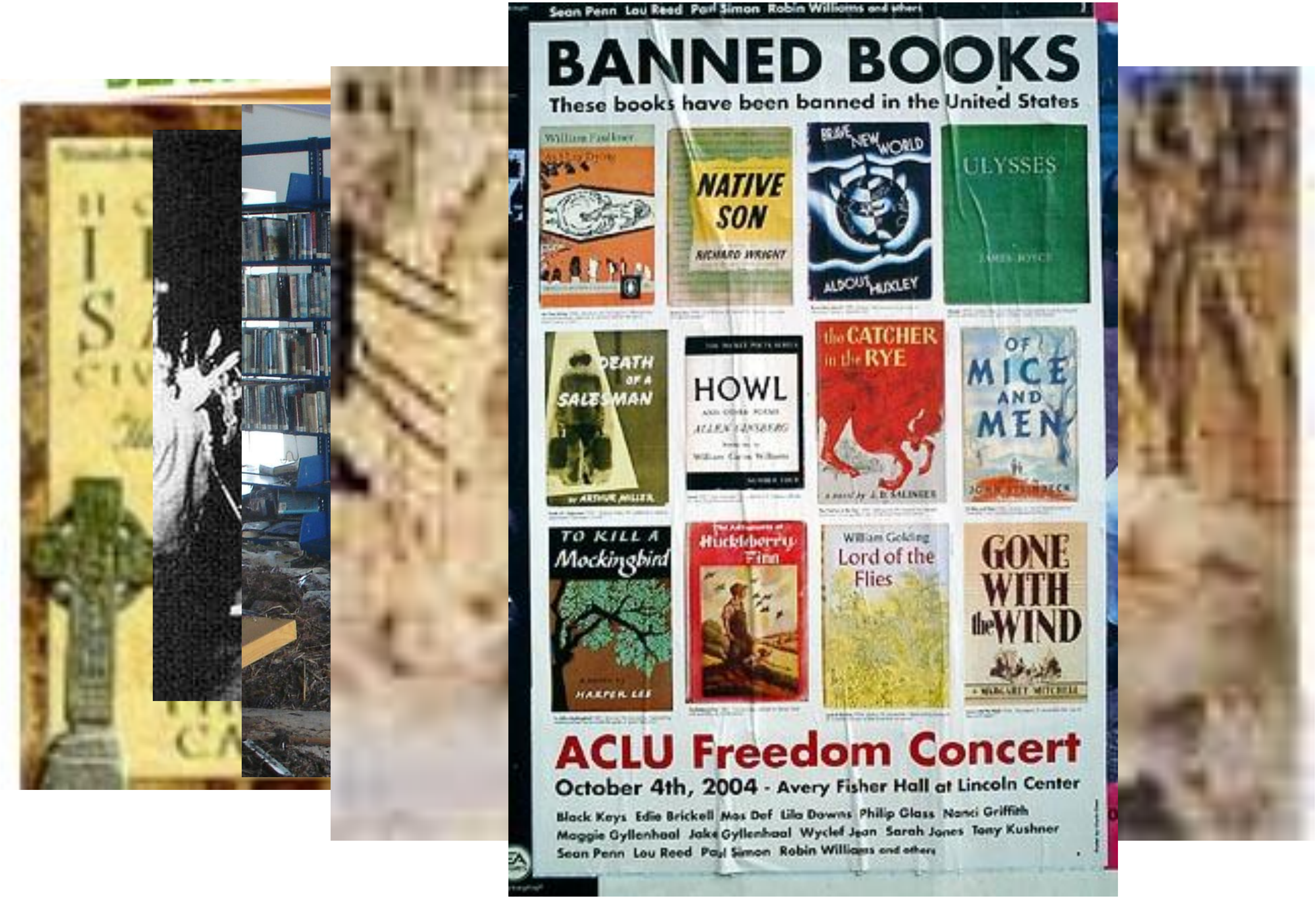
Preservation
Physical, Born-again, and Born Digital

CS 431 – April 23, 2007
Carl Lagoze – Cornell University

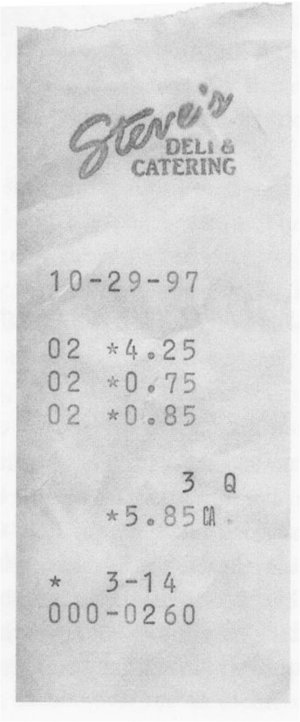
Acknowledgements:
James Cheney
Anne Kenney
Nancy McGovern
Vicky Reich

Why is preservation important and essential


Sean Penn Lou Reed Paul Simon Robin Williams and others
BANNED BOOKS
These books have been banned in the United States
William Faulkner
NATIVE SON
RICHARD WRIGHT
BRAVE NEW WORLD
ALDOUS HUXLEY
ULYSSES
JAMES JOYCE
DEATH OF A SALESMAN
by ARTHUR MILLER
HOWL
AND OTHER POEMS
ALLEN GINSBERG
the CATCHER in the RYE
a novel by J. D. SALINGER
OF MICE AND MEN
JOHN STEINBECK
TO KILL A Mockingbird
HARPER LEE
The Adventures of Huckleberry Finn
William Golding
Lord of the Flies
GONE WITH the WIND
MARGARET MITCHELL
ACLU Freedom Concert
October 4th, 2004 - Avery Fisher Hall at Lincoln Center
Black Keys Edie Brickell Mos Def Lila Downs Philip Glass Nanci Griffith
Maggie Gyllenhaal Jake Gyllenhaal Wyclef Jean Sarah Jones Tony Kushner
Sean Penn Lou Reed Paul Simon Robin Williams and others

What is information? What should be preserved?

- Paul Duguid – “The Social Life of Information
- David Levy – “Scrolling Forward”

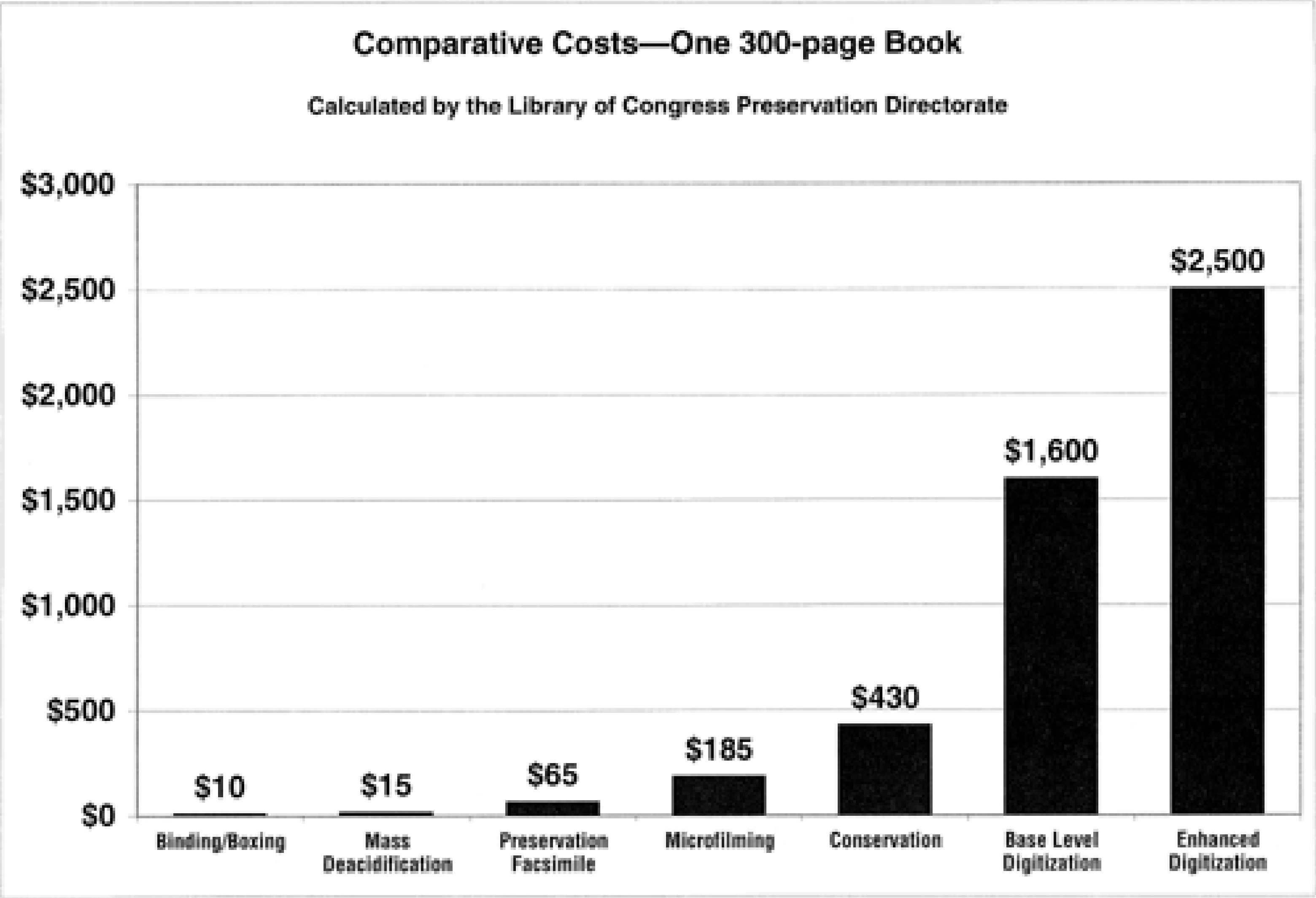


Steve's
DELI &
CATERING

```
10-29-97

02 *4.25
02 *0.75
02 *0.85

       3 Q
   *5.85CA

*  3-14
000-0260
```

Preservation of physical artifacts

- Environmental Control
- Brittle Books
  - Acidification is byproduct of paper production in 1850's to 1980's
    - Bleach for whitening
    - Alum for sizing (fixity of ink)
    - Tanning for leather tanning
  - 35-75% of paper based artifacts from this period are in danger
  - Newspapers and paperbacks especially vulnerable
  - ANSI standard Z39.48-1992 for "permanent" paper.



Comparative Costs—One 300-page Book

Calculated by the Library of Congress Preservation Directorate

Deacidification of Brittle Books

- Raise the pH level of treated paper to the acceptable range of 6.8 to 10.4pH
- extending the useful life of paper (measured by fold endurance after accelerated aging) by over 300%.
- Environmental treatment using magnesium oxide (MgO)
- Expense requires careful selection process

Digitization through Scanning
"Born-again" digital

- Alternative to deacidification
- Advantages
  - Universal access
  - Reduction in shelf costs
  - OCR (full-text access)
- Disadvantages
  - Quality reduction
  - Cost
  - Not original syndrome
  - Destruction of source (debinding)
  - A new preservation problem

Failures of Microfilm

- Popular preservation approach before digitization
- Severe problems
  - Quality of filming
  - Color to bi-tonal
  - Usability issues
  - Self-destruction of film

- “Double Fold” Nicholson Baker

Scanning

- Electronic snapshots taken of a scene or scanned from documents
- samples and mapped as a grid of dots or picture elements (pixels)
- pixel assigned a tonal value (black, white, grays, colors), represented in binary code
- code stored or reduced (compressed)
- read and interpreted to create analog version

Why Rich Digital Masters?

- Preservation
  - Original may only withstand one scan
  - Maintenance of digital files
- Cost
  - One scan may be all that is affordable
  - Conversion costs dwarfed by other costs
- Access
  - Many from one
  - The richer the file, the better the derivative in terms of quality and processibility

How to determine what’s good enough?

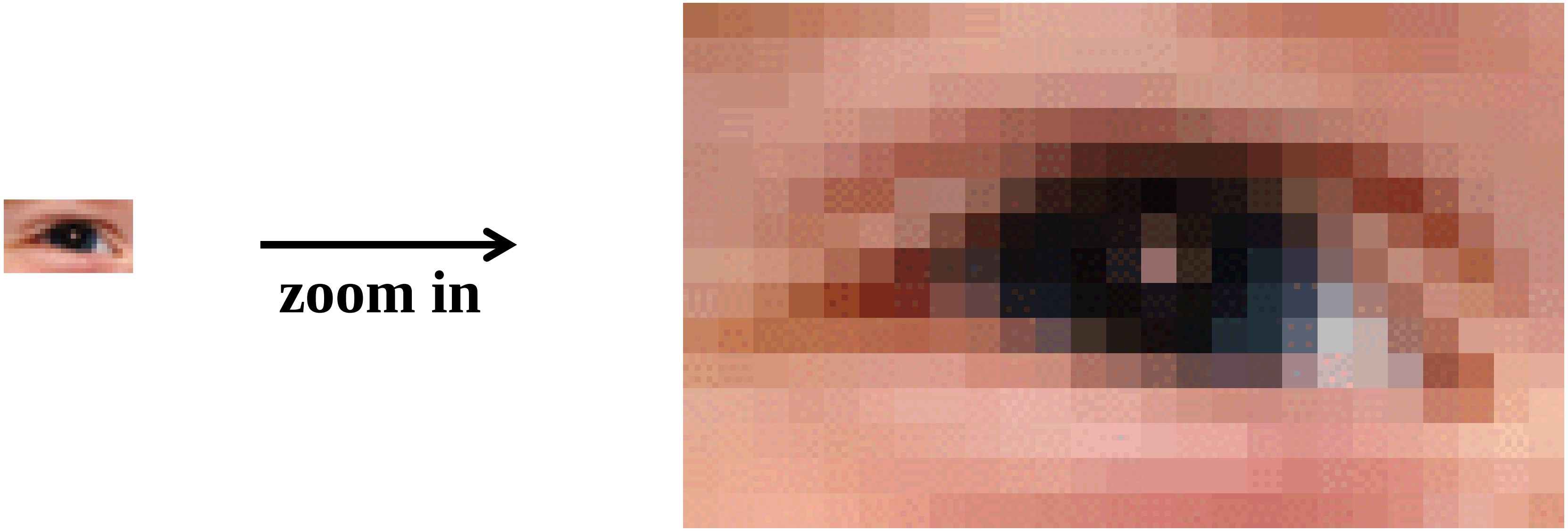
- Connoisseurship of document attributes
  - Identify key information content
  - Objectively characterize or measure attributes: size, detail, tone, and color
- Appreciate imaging factors affecting quality and cost
- Translate between analog and digital
  - Equate measurements to digital equivalencies and corresponding metrics, e.g., detail size → resolution

Digital Image Quality is Governed By:

- resolution and threshold
- bit depth
- color management
- image enhancement
- compression and file format

Resolution

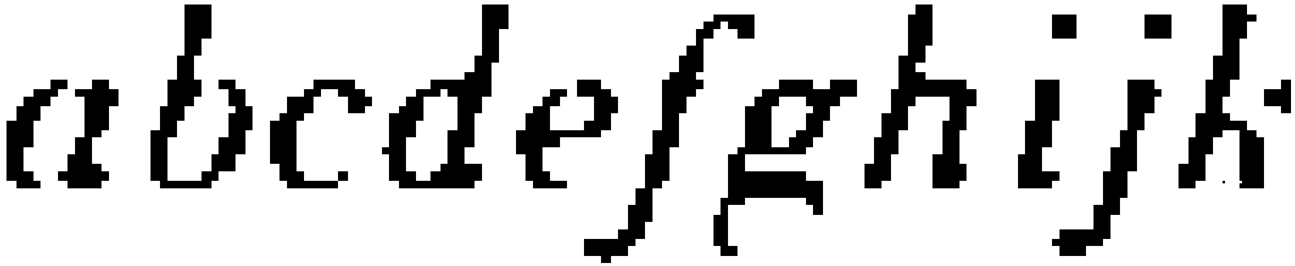
- Determined by number of pixels used to represent the image
- Increasing resolution increases level of detail captured and geometrically increases file size



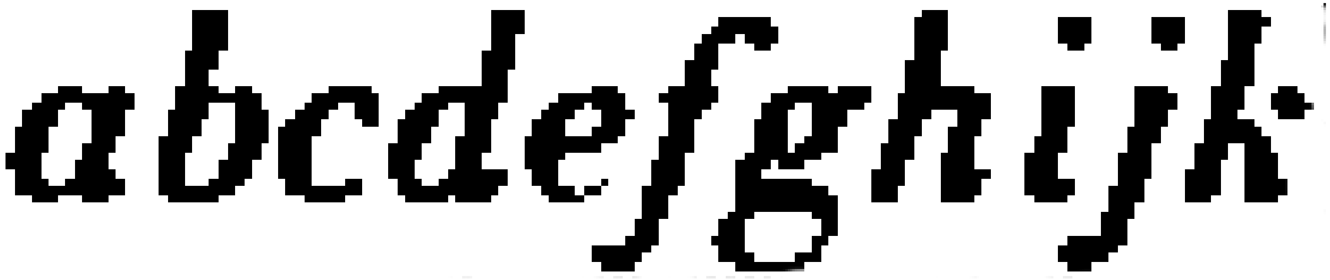
Threshold Setting in Bitonal Scanning

defines the point on a scale from 0 to 255 at which gray values will be interpreted either as black or white

Effects of Threshold



threshold = 60

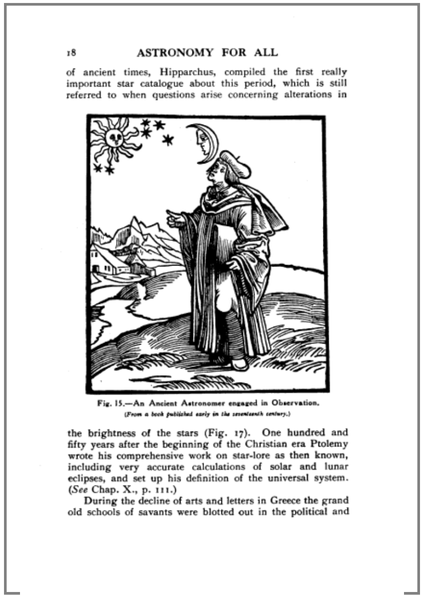


threshold = 100

Bit Depth

- Determined by the number of binary digits (bits) used to represent each pixel

1-bit



18 ASTRONOMY FOR ALL

of ancient times, Hipparchus, compiled the first really important star catalogue about this period, which is still referred to when questions arise concerning alterations in

Fig. 15.—An Ancient Astronomer engaged in Observation.

(From a book published early in the seventeenth century.)

the brightness of the stars (Fig. 17). One hundred and fifty years after the beginning of the Christian era Ptolemy wrote his comprehensive work on star-lore as then known, including very accurate calculations of solar and lunar eclipses, and set up his definition of the universal system. (*See* Chap. X., p. 111.)

During the decline of arts and letters in Greece the grand old schools of savants were blotted out in the political and

8-bit

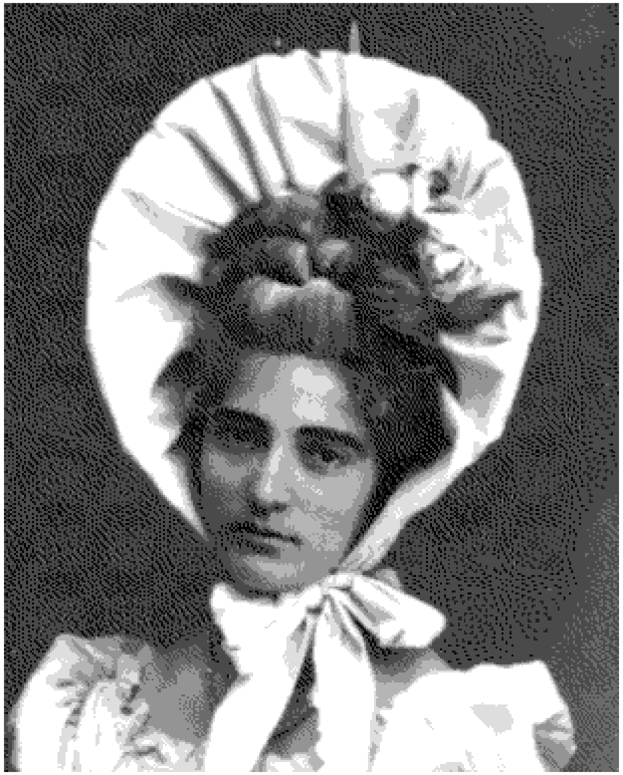
24-bit

Louis A. Fuertes

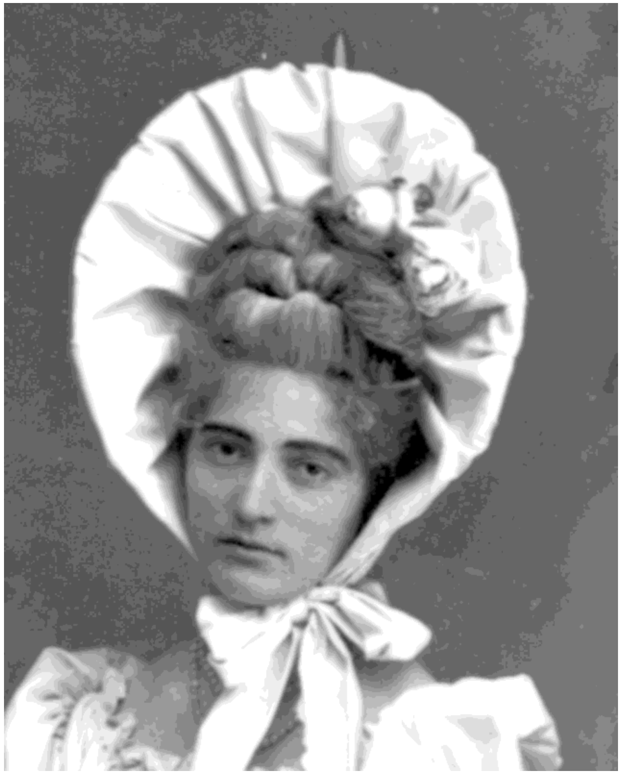
Bit Depth

- increasing bit depth increases the level of gray or color information that can be represented and arithmetically increases file size
- Bit depth, dynamic range, and color appearance

Utilizing Sufficient Bit-Depth

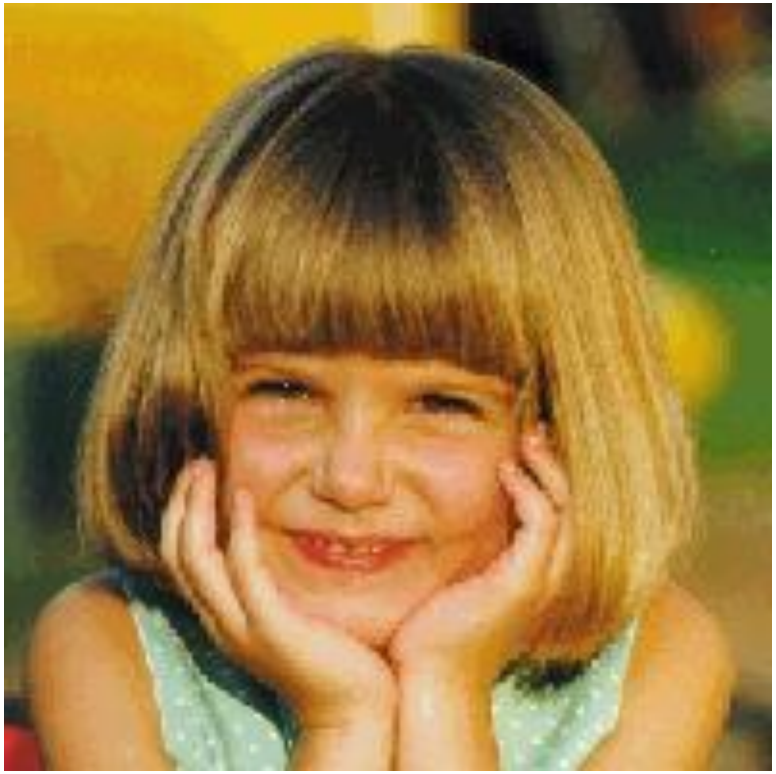


3-bit gray

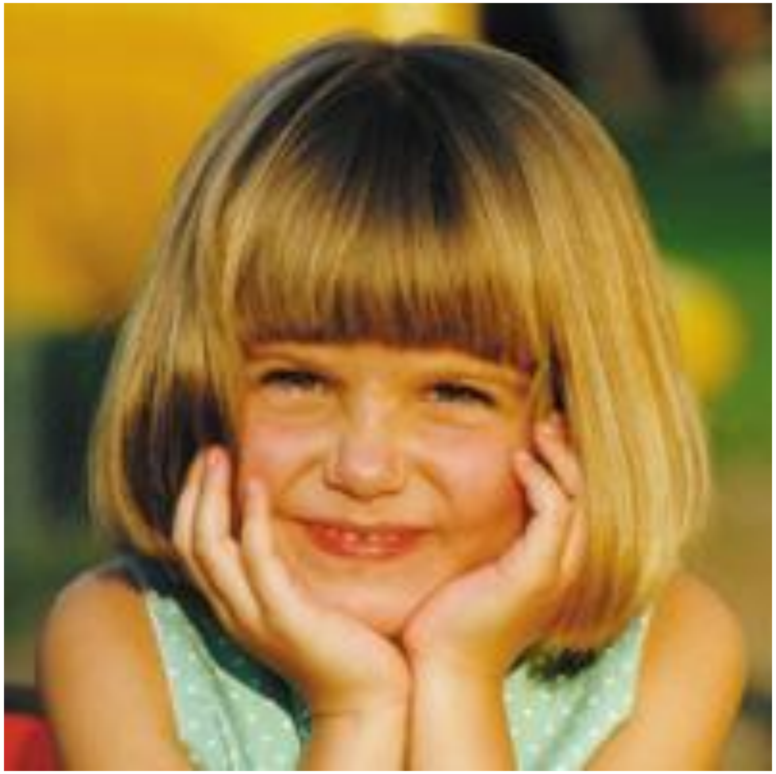


8-bit gray

Utilizing Sufficient Bit Depth



8-bit color



24-bit color

One Size Does Not Fit All!

- Different document types will require different scanning equipment and processes
- The more complex the document, the higher the conversion/access requirements
- Scan the original whenever possible
- No standards for image conversion: guidance rather than guidelines

Digital Preservation Strategies

Disclaimer: monolithic, homogeneous solutions are likely to fail, many digital preservation approaches are required

Emulation

- Preserve original “look and feel” and functionality of digital artifact
- Enable obsolete systems to be run on future unknown systems
- Notion of universal virtual machine
- Jeff Rothenberg, Raymond Lurie
- CAMiLEON Project
  - http://www.si.umich.edu/CAMILEON/about/aboutcam.html

Migration

- *File formats change over time and become extinct*
- *Issues of proprietary vs. open source formats*
- *Lossiness of formats*
- *Risk Management of Digital Information: A File Format Investigation*
    - http://www.clir.org/pubs/reports/pub93/contents.html
- CAMiLEON Project

Canonicalization

- **Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information**
  - http://www.dlib.org/dlib/september99/09lynch.html
- Tie to XML standards
  - http://www.w3.org/TR/2002/REC-xml-exc-c14n-20020718/

Trusted Repository
Centralized Storage Approach

- *Attributes of a Trusted Digital Repository* (RLG-OCLC)

  http://www.rlg.org/longterm/attributes01.pdf

  - Administrative responsibility
    - *OAIS Reference Model (CCSDS)*

      http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf
  - Organization viability
  - Financial sustainability
  - Technical suitability
  - System security
  - Procedural accountability
- National Archives and Records Administration (NARA)

LOCKSS – Decentralized Storage Approach

- Lots of Copies Keep Stuff Safe
- http://lockss.stanford.edu/

“Diffused Knowledge Immortalizes Itself”

Sir James Mackintosh 1765-1832

LOCKSS Mission

- Build tools and provide support

- **Libraries**, *so they can easily and affordably* build, preserve, and archive local e-collections
  - Own rather than lease electronic information
  - Retain traditional custodial role of scholarly information
- **Publishers**, *so they can easily and affordably* provide content for preservation and archiving
  - With minimal risk to their business model or to their publishing platforms
  - Relinquish responsibility to provide perpetual access
  - Fulfill librarians' requirements that publishers guarantee long-term access to content sold

Paper Library System

- Libraries act for their institution to
  - Acquire copies of important "stuff"
  - Keep copies on shelves
  - Give access to local readers
- Libraries cooperate to
  - Supply copies to other libraries
    - a reader can easily to find *a* copy
    - a "bad guy" has trouble finding and destroying all copies
- Libraries ensure content persists simply by supporting their local communities
- *A cooperative, affordable, decentralized, 'archive system' with LOTS OF COPIES*

LOCKSS “Library System”

- Libraries act for their institution to
  - Acquire copies of important “stuff”
  - Keep copies in transparent web caches
  - Give access to local readers
- Libraries cooperate to
  - Detect and repair damage
    - a reader can easily find *a* copy
    - a “bad guy” has trouble finding and destroying all copies
- Libraries ensure content persists simply by supporting their local communities *(Preservation is integrated with access)*
- *A cooperative, affordable, decentralized, archive system with LOTS OF COPIES*

LOCKSS Technology

- LOCKSS web caches (no delete)
- Collect HTTP delivered content
  - All file formats (PDF, HTML, JPEG, TIF, Audio, Video)
  - Collect "presentation files" of content as published
    - Subscribe to publishers for new content
  - Must have authorized access to publisher's site
- Preserve and audit content integrity
  - Independent content collection
  - Cooperate to resolve content differences
    - Continuously validate against other caches
    - Repair gaps from publisher and other caches
  - Resist attackes
    - Prevent damage from spreading
    - Isolate hostile participants
    - Reputation management
- Provide access
  - Readers access content via desktop web browser
  - Content is never "dark"

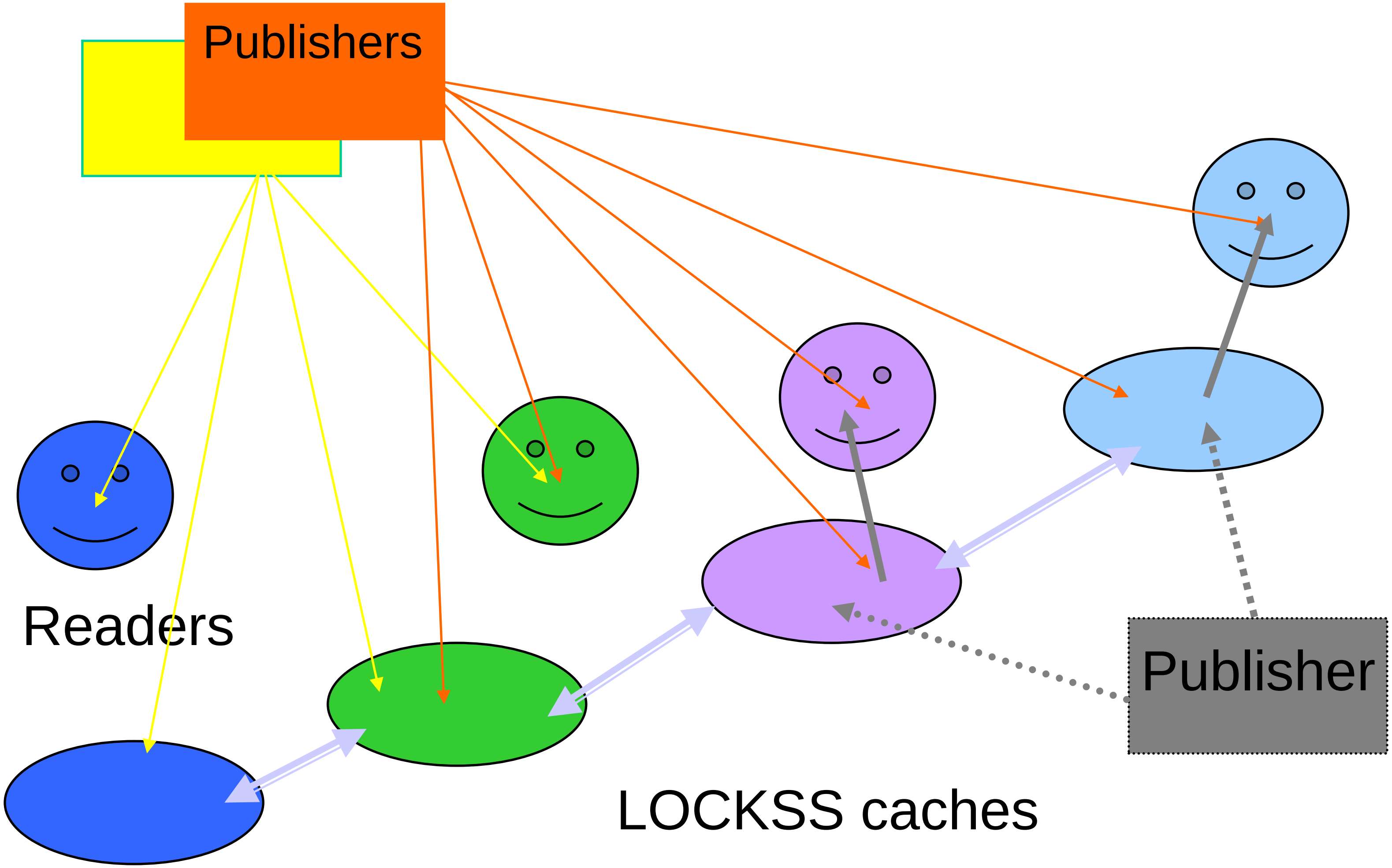


Figure 1: In this example, each LOCKSS cache (oval) collects journal content from the publisher's web site as it is published. Readers (circles) can get content from the publisher site. When the publisher's web site is not available (gray) to a local community, readers from that community get content from their local institution's cache. The caches "talk" to each other to maintain the content's integrity over time.