University of Kentucky
UKnowledge

Library Presentations University of Kentucky Libraries

7-2009

File Formats 101

Kathryn Lybarger
University of Kentucky, kathryn.lybarger@uky.edu

Follow this and additional works at: https://uknowledge.uky.edu/libraries_present

Part of the Library and Information Science Commons

Right click to open a feedback form in a new tab to let us know how this document benefits you.

Repository Citation
Lybarger, Kathryn, "File Formats 101" (2009). *Library Presentations*. 21.
https://uknowledge.uky.edu/libraries_present/21

This Presentation is brought to you for free and open access by the University of Kentucky Libraries at UKnowledge. It has been accepted for inclusion in Library Presentations by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.

File Formats 101

Kathryn Lybarger

Paul Revere's Ride

Listen my children and
you shall hear
Of the midnight ride of
Paul Revere,
On the eighteenth of
April, in Seventy-five;
Hardly a man is
now alive
Who remembers that
famous day and year.

Paul Revere's Specification

… If the British march

By land or sea from the town to-night,

Hang a lantern aloft in the belfry arch

Of the North Church tower, as a signal light, --

One, if by land, and two, if by sea

THE BRITISH ARE COMING BY SEA

A better signal

How many signals?

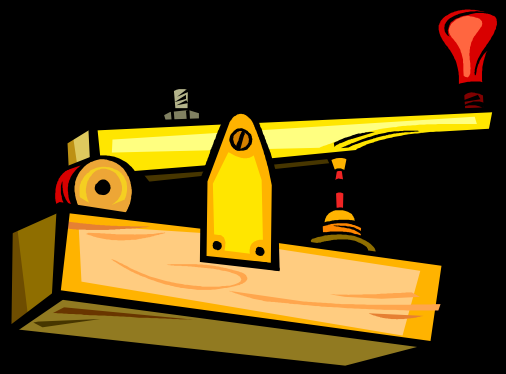
- The British are not coming (yet).
- The British are coming by land.
- The British are coming by sea.

More options

- The British are coming in some other way – look out!

- There is some other problem – come see.

Western Union “92 code” (1859)



- 1 Wait a minute.
- 7 Are you ready?
- 27 Priority, very important.
- 73 Best Regards.
- 88 Love and kisses.

More than one tower?

- (0 0 0) The British are not coming (yet).
- (0 0 1) The British are coming by land.
- (0 1 0) The British are coming by sea.
- (0 1 1) The British are coming!!
- (1 0 0) Love and kisses.
- (1 0 1) We are out of tea.
- (1 1 0) We are out of milk.
- (1 1 1) We are out of lanterns.

Binary numbers

- Each position represents a power of two:

  128 64 32 16 8 4 2 1

- 7 = 4 + 2 + 1 → 00000111

- 20 = 16 + 4 → 00010100

Binary is compact

- All numbers between 0 and 255 can be represented using 8 bits (one byte).

- 255 = 128 + 64 + 32 + 16 + 8 + 4 + 2 + 1 =

  11111111

- 128 = 128 + 0 + 0 + 0 + 0 + 0 + 0 + 0 =

  10000000

Binary is flexible

- 0, 1 written as text
- negative/positive polarity on magnetic media
- low voltage / high voltage on a wire
- lanterns not lit / lanterns lit in towers

File formats

A **file format** is a specification for interpreting a bitstream as meaningful data.

Examples:

- 0 = black, 1 = white (bitmap image)
- Group as binary numbers -> letters (ASCII)
- “Executable” code

File formats are interpreted by software.

Do not trust file name extensions

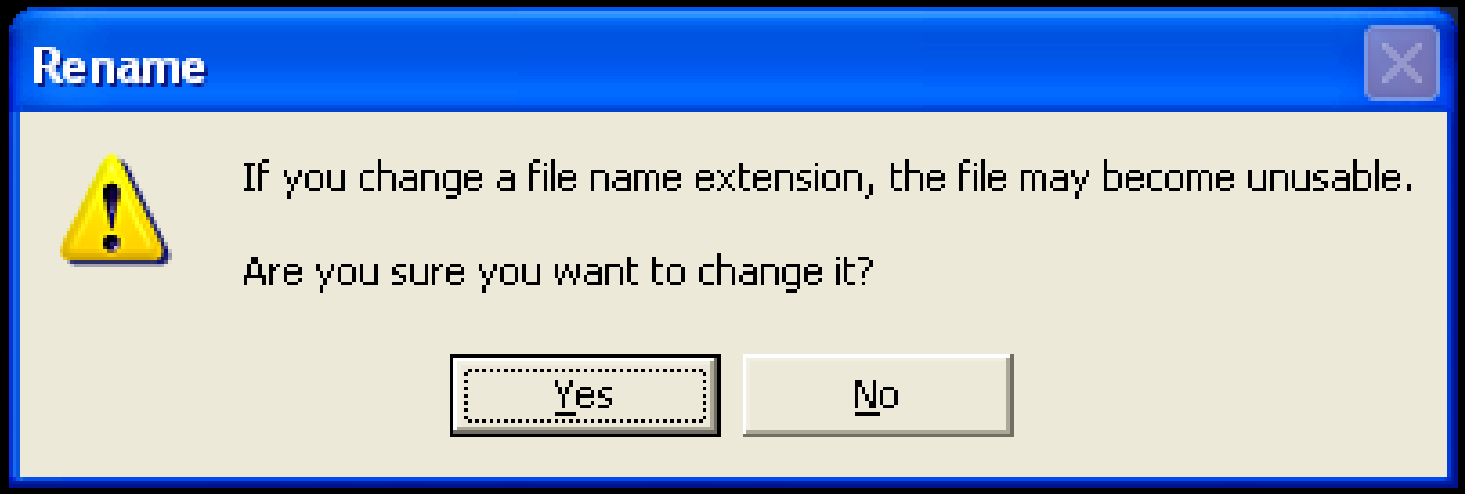


photo.jpg

photo.mp3

Preservation file formats

A **preservation file format** is a file format which stores data in a way such that it can be faithfully rendered by computer systems now and in the future.

The same file format forever?

- Example: Project Gutenberg (1970's)
- Now allows XHTML, images, audio
- Insists on plain ASCII copy

Format migration

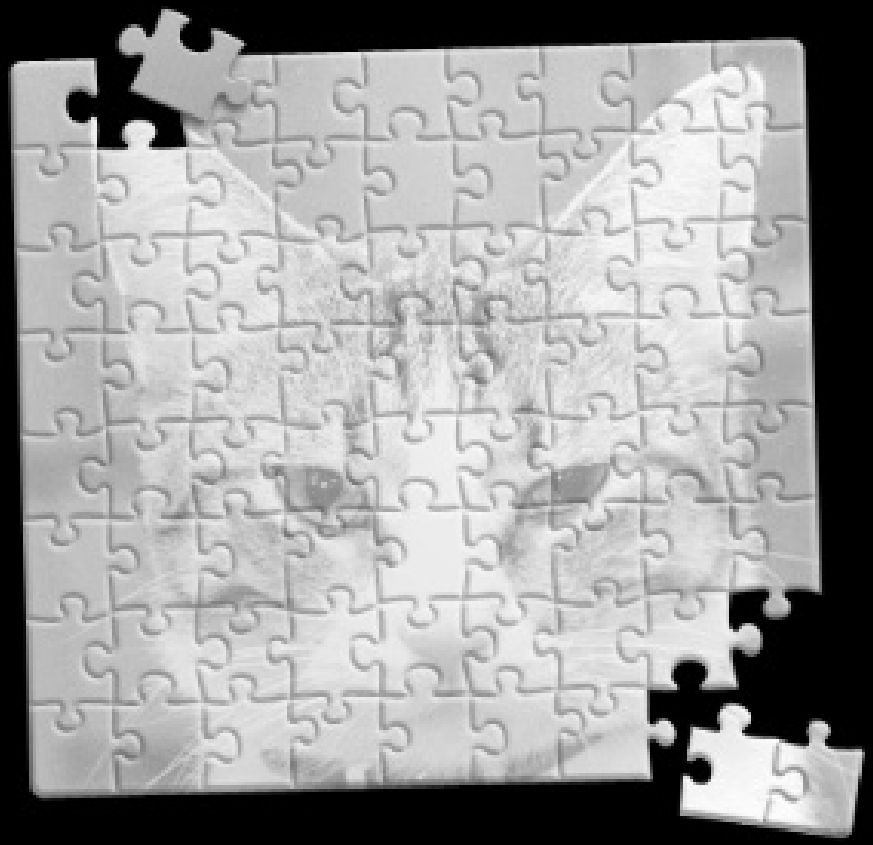
- You need not use the same file format forever
- Must have sufficient data and context to migrate data to other formats
- Those formats should similarly be preservation file formats

Preservation file formats should be lossless

- All analog to digital conversions are lossy.
- A **lossless** format is one such that conversion of digital data into this format loses no more data.

Lossless / lossy formats

- Files in lossy formats do not (typically) lose data when you view them
- They might if you SAVE them as you close them, even if you save in the same format



JPG → JPG → JPG → JPG …

Preservation file formats should be open

An open format is one where the mode of presentation of the data is transparent, or the format specification is publically available.

-- from openformats.org

Transparent presentation of data

HTML code:

My favorite show is Quantum Leap.

Renders as:

My **favorite** show is *Quantum Leap*.

Format specification

A TIFF file begins with an 8-byte image file header, containing the following information:

Bytes 0-1: The byte order used within the file. Legal values are:

"II" (4949.H)

"MM" (4D4D.H)

In the "II" format, byte order is always from the least significant byte to the most significant byte, for both 16-bit and 32-bit integers This is called *little-endian* byte order. In the "MM" format, byte order is always from most significant to least significant, for both 16-bit and 32-bit integers. This is called *big-endian* byte order.

Bytes 2-3 An arbitrary but carefully chosen number (42) that further identifies the file as a TIFF file.

The byte order depends on the value of Bytes 0-1.

Bytes 4-7 The offset (in bytes) of the first IFD. The directory may be at any location in the file after the header but *must begin on a word boundary*. In particular, an Image File Directory may follow the image data it describes. Readers must follow the pointers wherever they may lead.

Preservation file formats should be unencumbered

- Formats may require royalties to use the format.
- Licenses may disallow reverse-engineering
- Leads to “lock-in”

Example: LZW compression

- Used in GIF, compressed TIFF
- Subject to multiple patents (now expired)

Example: EndNote

- Academic reference manager
- An open-source alternative, Zotero, allowed importing EndNote files
- EndNote brought a lawsuit against Zotero
- Case was dismissed

Preservation file formats should be resistant to corruption

- Physical media degrades
- File systems become corrupt
- Files do not always transfer correctly

File corruption

File corruption

File corruption

Location of corruption is important

- Many file formats have a “magic number”
  - PDF %PDF
  - GIF GIF87a or GIF89a
  - Java CAFEBABE or CAFED00D
  - TIFF II or MM followed by 42 in binary
- Corrupted magic number may make a file “unrecognizeable”

Not all software handles corruption the same way

- Some may not notice it
- Some may refuse to open the file
- Some may help you salvage the file

Preservation file formats should allow embedded metadata

- File name / directory structure is insufficient
- Files may be stored in different ways
- File names are not part of files

QE^TQE^TE^TQE^TQE^T'Qlih¢(¢(¢l^DP:RNE^TQE^T§^T(sKE^TQE^TQE^TQE^TQE^TQE^TQE^TQE
TQE^TQE^TQE^TQE^TQE^TQE^TQE^TQE^TQE^TQE^TQE^TQE^TQE^TQIE-^TQE^TQE^TQE^TQE^TQE^T
E^TQE^T^AKE^TQE^TQE^TQE^Téh¢(¢(¢(¢(¢ÿ^@tÔ^Q¸ñV)h¢))h¤=)^GZu^TQE^TQE^TQE^TQE%^B(
(¢(¢ÿ^@tÔqO¢(¢(¢(¢Ý)GJZ(¢(¤=)÷@(¢(¢(¢(¢JZ(¢(¢:Ó¨¢cÓJ)h¢(¤=(¥¢J^GJZ(¢cÓJZ(¢3m^TQ
^TQE^TQE^TQE^TQE^TS^ ¥8t¥¢(¢(¢(¢(¢(¢>(¦·jZZ(¢(¢(¢(¤¥¢(¢(¢(¢(¢(¢(¢(¢(¢(¢(¢(¢(¢(¢
¢(¢(¢(¢(¢(¢(¢(¢(¢(¢(¢(¢(¢(

Preservation file formats

- Lossless
- Open
- Unencumbered
- Resilient to corruption
- Allow metadata

File formats need not be perfect

- Have a realistic view of how your data is being stored
- Respond accordingly
- Migrate when new formats are adopted

Using preservation file formats

- Not always possible
- Not sufficient to keep data safe forever
- Important part of complete preservation strategy

Any questions?