Digital Preservation

File Formats and Characterisation

Hannes Kulovits
Institut für Softwaretechnik und Interaktive Systeme
TU Wien

http://www.ifs.tuwien.ac.at/dp

Agenda

- Definition of File/File Format
- Representation
- Elements of a file format
- File and Preservation
- Challenges

What is a file/file format?

- A **file** is nothing more than a sequence of bits
- How to encode those bits is specified in a **file format**
- File format is a specification of how to interpret a bit stream.
- File format specifies
  1. Whether the file is binary or ASCII
  2. How information is organized
  3. ...

Plain Text

- De facto standard for Plain Text is *ASCII*
  - Uses 8 bits
  - Maximum of 256 different characters possible
  - Includes
    - Letters of most alphabets (lower and upper case)
    - Arabic numerals
    - Punctuation marks
    - Standard symbols
- Another important format is *Unicode*
  - Provides unique encoding for each character
  - Uses multiple bytes to represent each character

Proprietary vs. Open

- Proprietary
  - Documentation mostly not available
  - License and patent rules
  - License agreements subject to change
  - Restrictions for use and modifications may apply

- Open
  - Documentation available!
  - Unlimited use
  - No license fee
  - Open for modifications
  - No patent owners

File formats based on plain text

- For example: HTML
- In HTML plain text must obey certain rules (use of tags, type sizes, color)
- ...

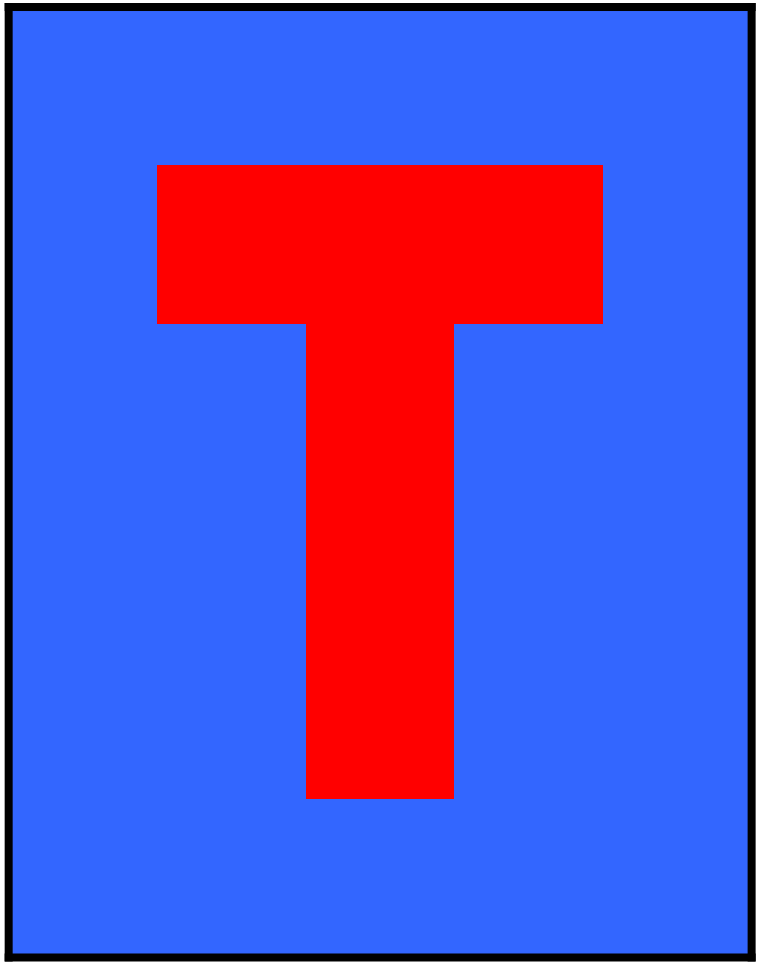
Different types of File Formats

- Different kinds of formats for different kinds of information
  [Rothenberg, 1995, Ensuring the Longevity of Digital Documents]
- Official categorisation of file formats is the IANA MIME type
  - Text documents
  - Databases
  - Still and moving images
  - Audio
  - Multipart
  - Application
  - ...

Different types of File Formats (2)

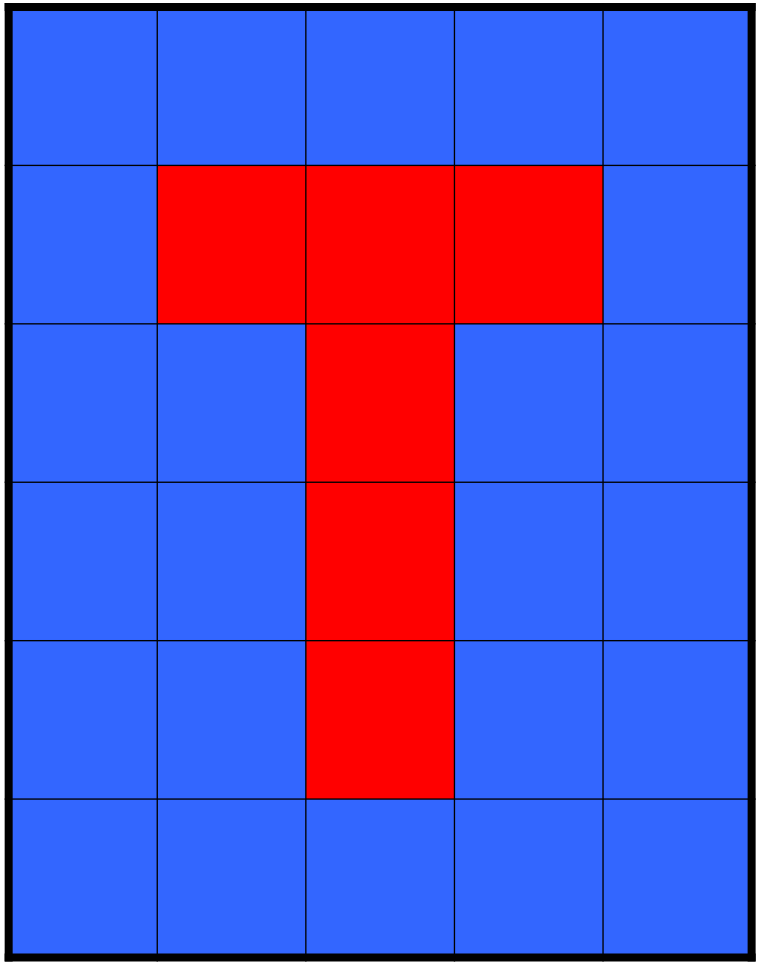
- Three-character file extension of DOS and Windows. (Neither standardised nor unique.)
- Unix ‚magic numbers‘
- Macintosh data-forks
- MIME type, also not unique
- None of them is really satisfying
  - Better solution: PRONOM with Pronom Unique Identifier

An image



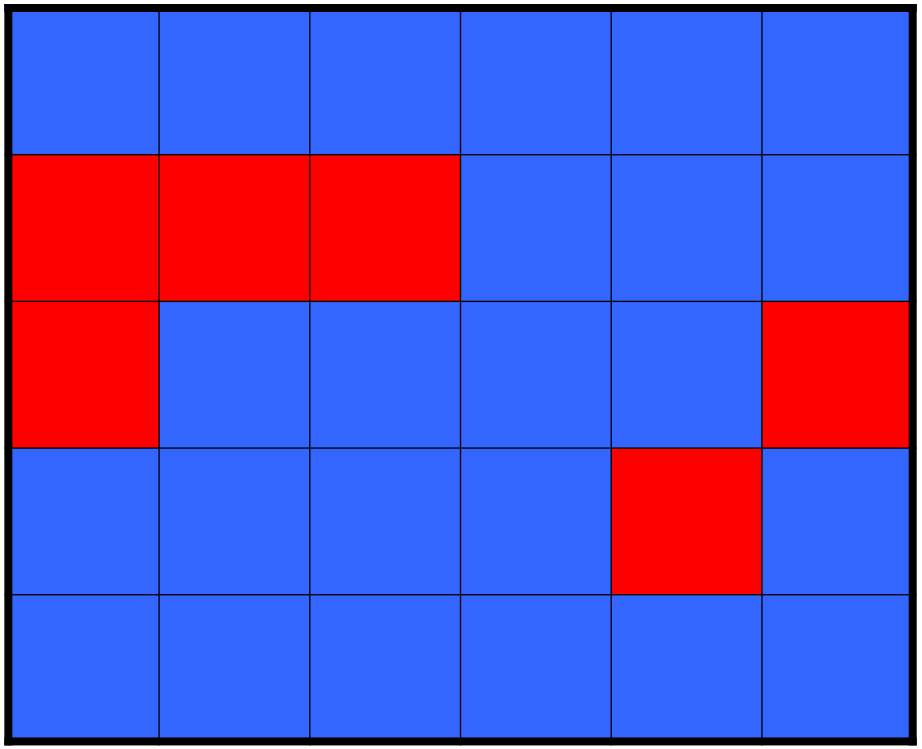
An image

6 rows
5 columns



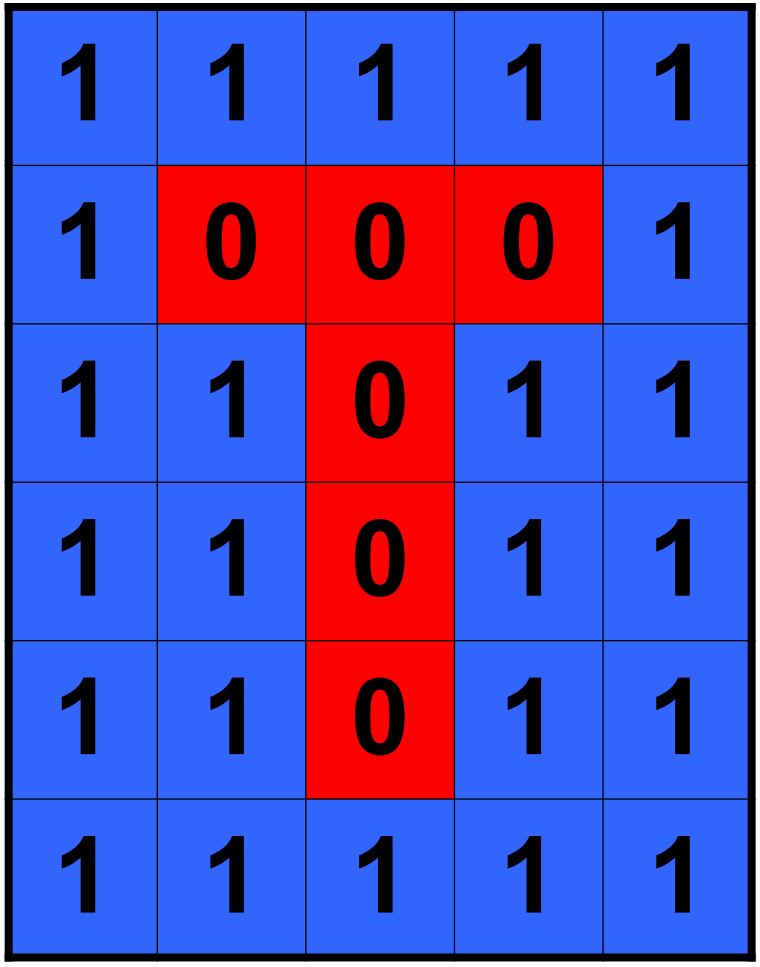
An image

5 rows
6 columns



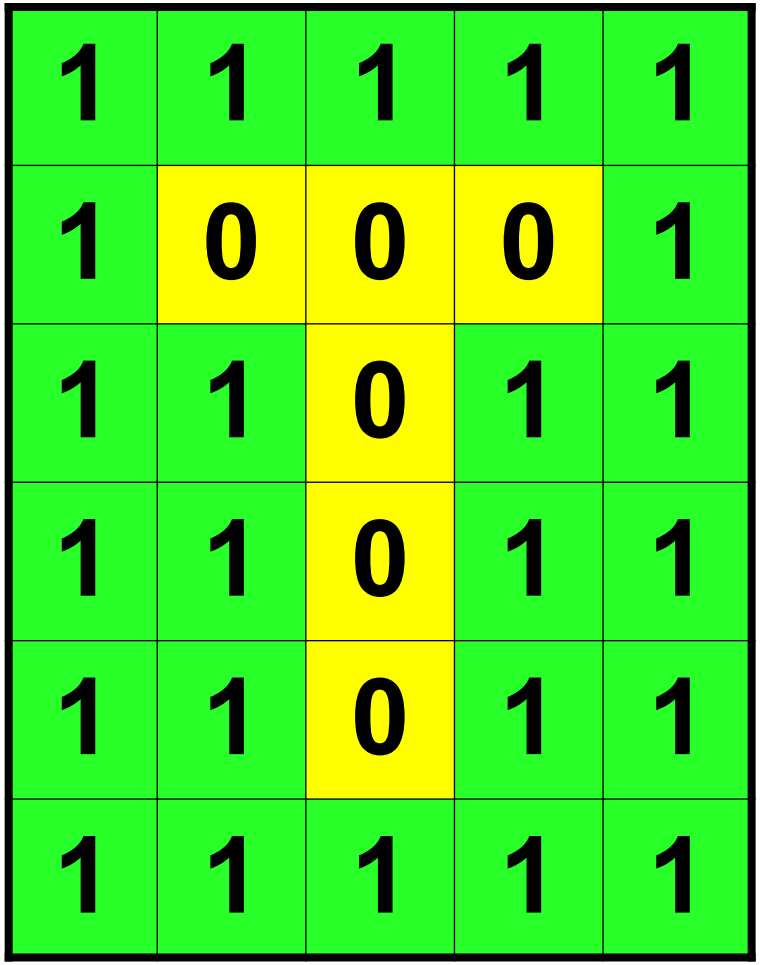
An image

1 == blue
0 == red



An image

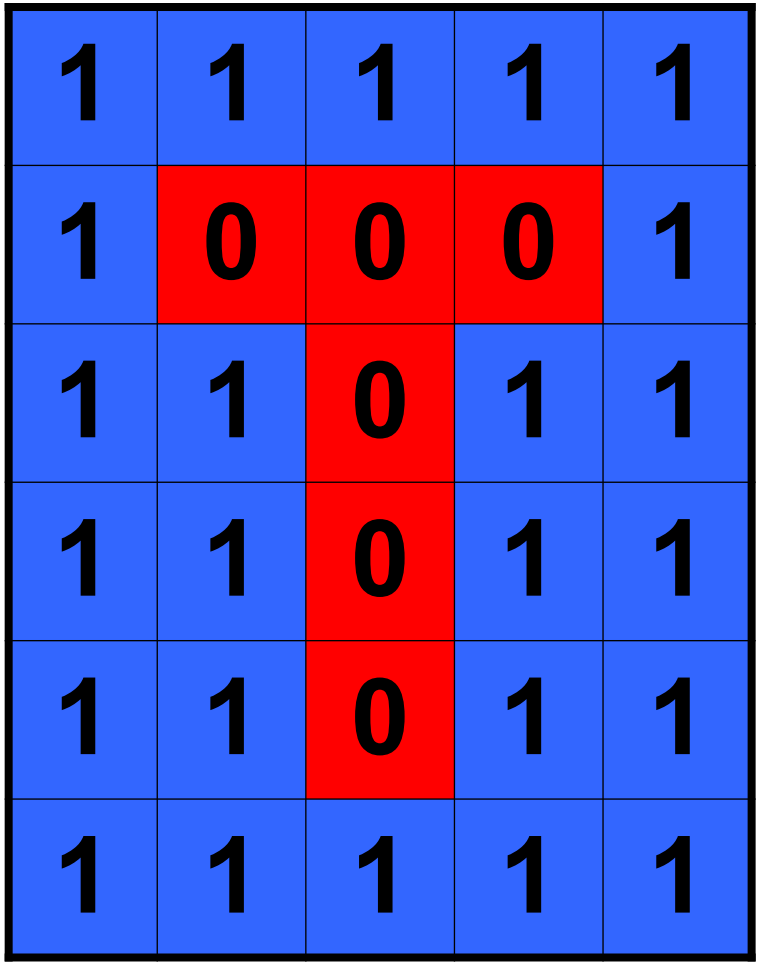
1 == green
0 == yellow



An image

Store:

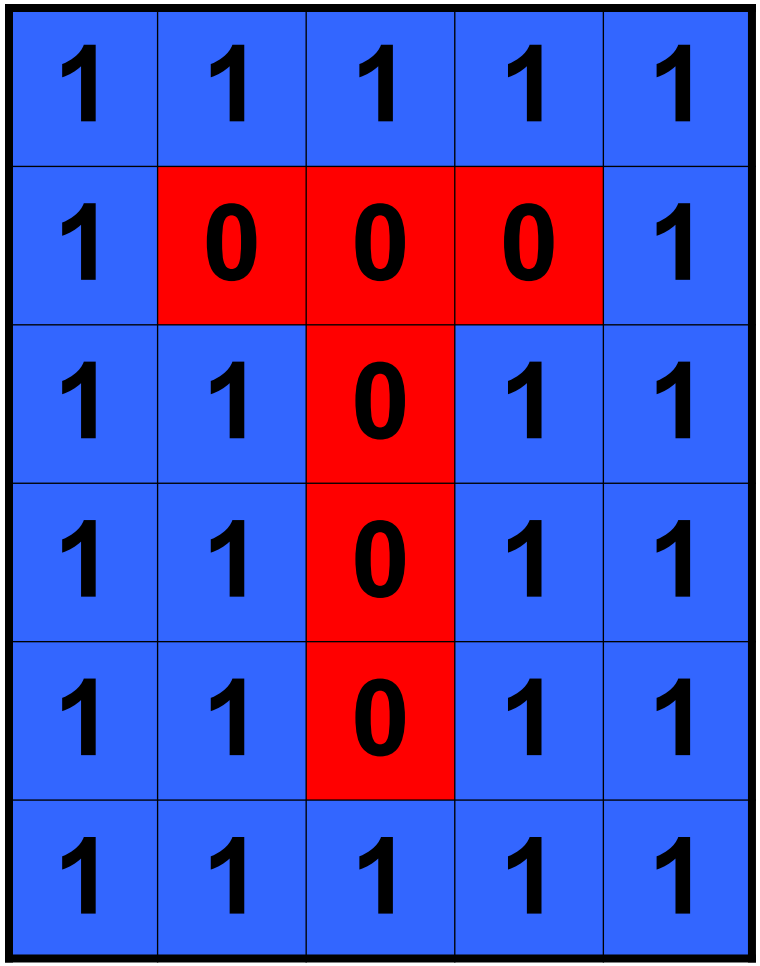
```
1,1,1,1,1,
1,0,0,0,1,
1,1,0,1,1,
1,1,0,1,1,
1,1,0,1,1,
1,1,1,1,1
```



An image

Store:

```
6,1,3,0,3,
1,1,0,4,1,1,
0,4,1,1,0,
7,1
```

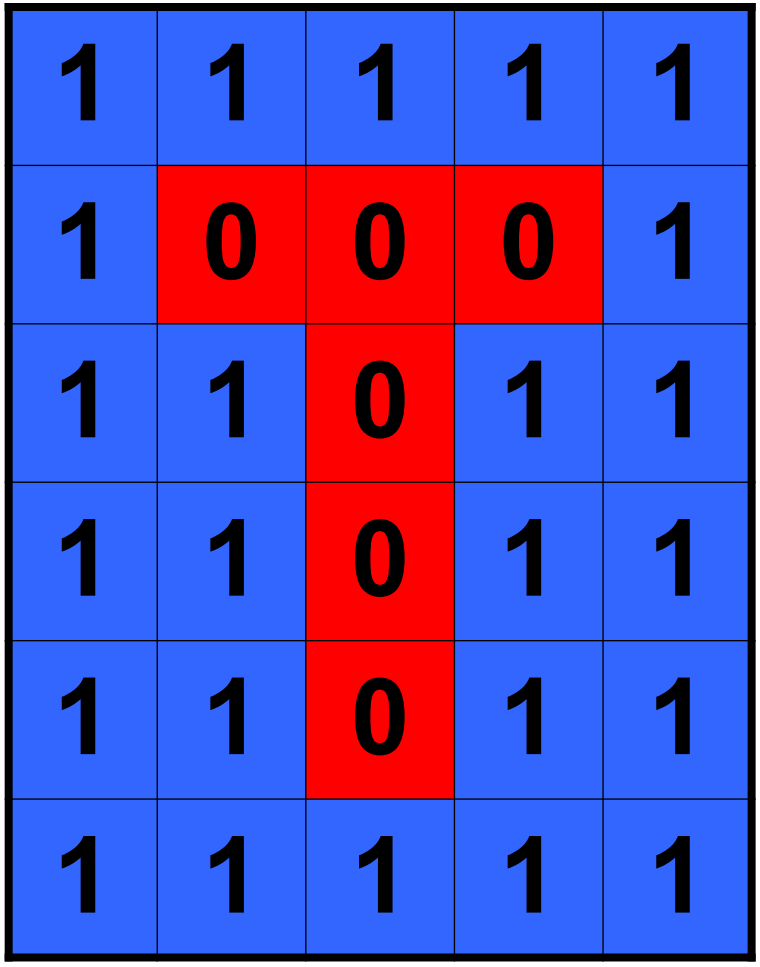


An image

Store:

```
1,1,1,1,1,
1,0,0,0,1,
1,1,0,1,1,
1,1,0,1,1,
1,1,0,1,1,
1,1,1,1,1
```

Uncompressed

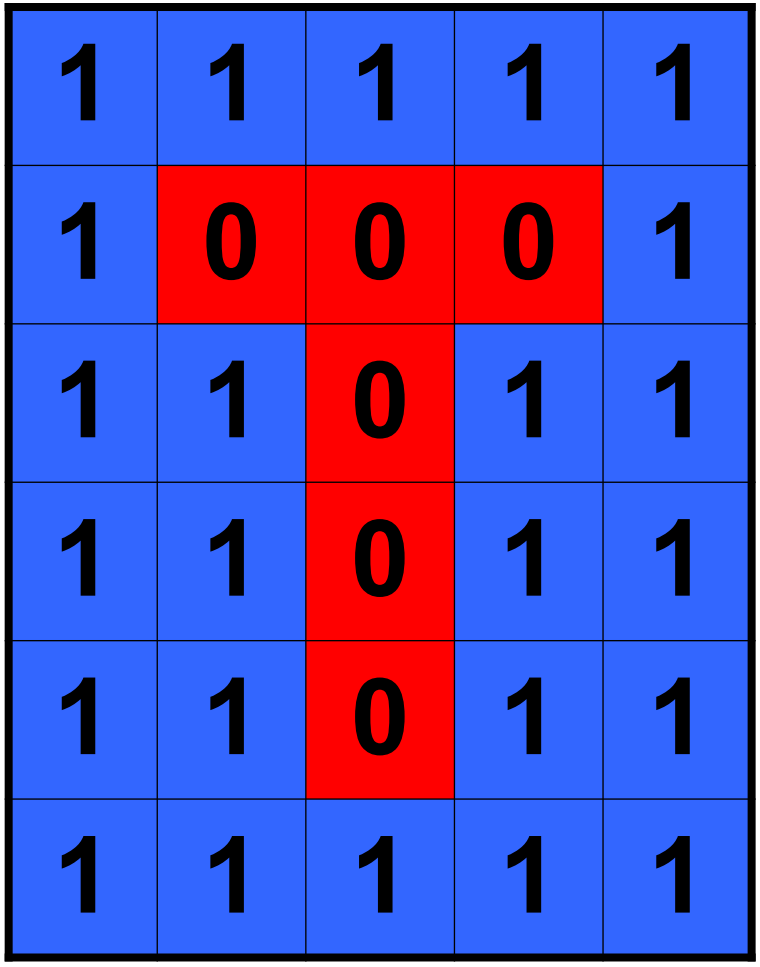


An image

Store:

6,1,3,0,3,
1,1,0,4,1,
1,0,4,1,1,
0,7,1

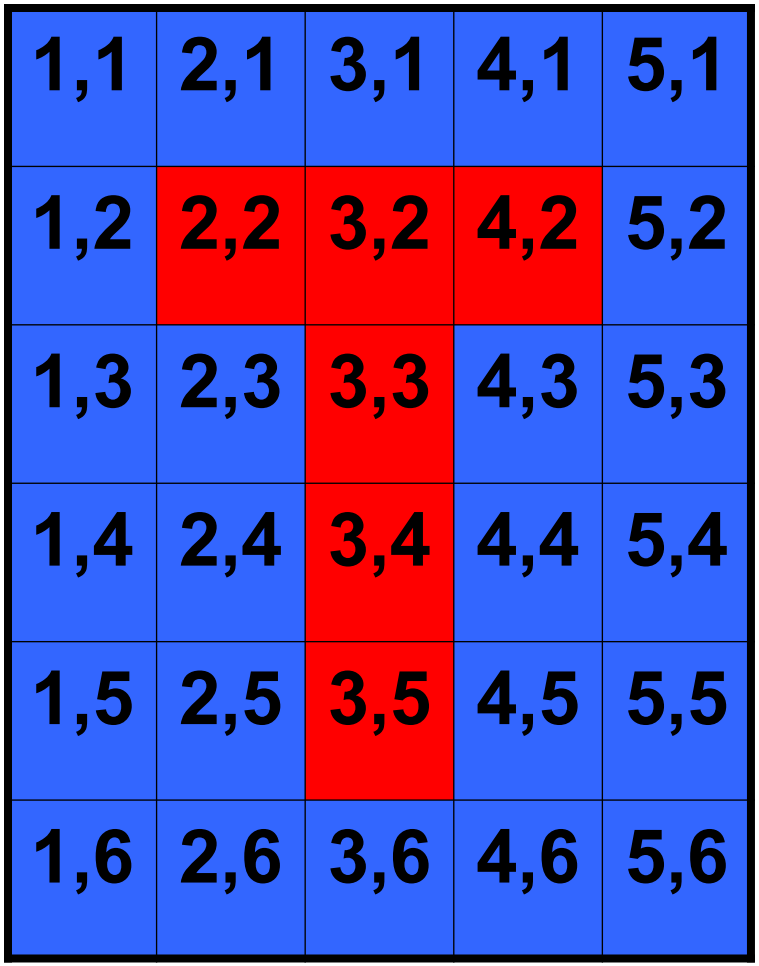
(Compressed)
Run Length
Encoded



An image

Store:

```
SetSize: 5 by 6
SetBackgroundColor: Blue
SetForegroundColor: Red
SetLetterHeight: 4
MoveTo: 3,5
DrawLetter: T
```

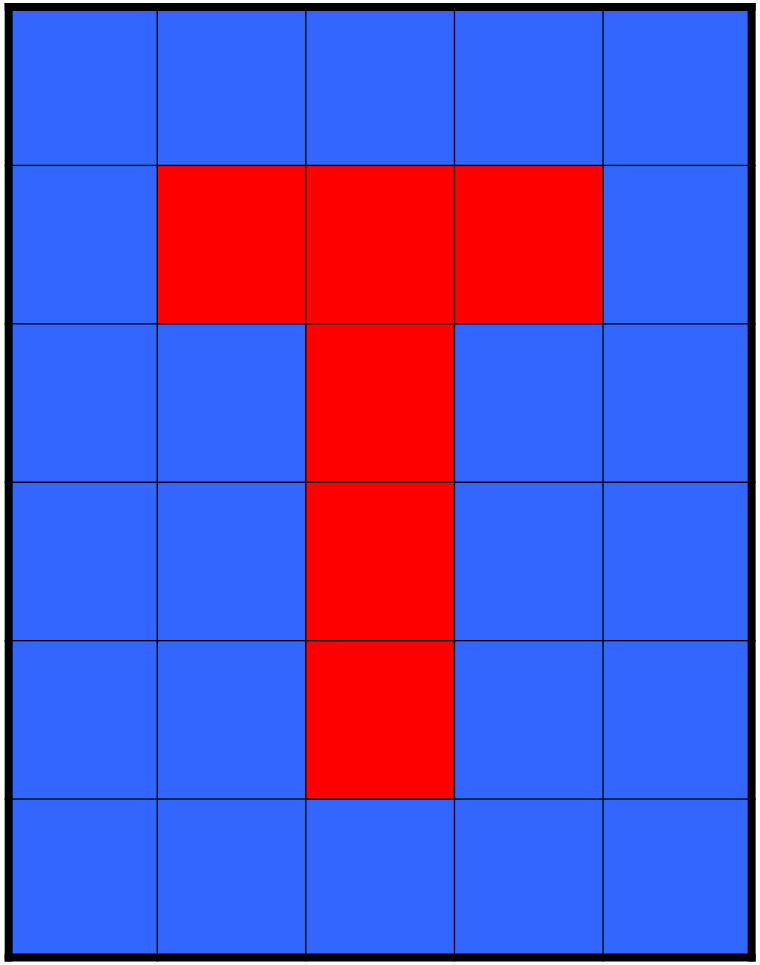


An image

<basic information>

<rendering information>

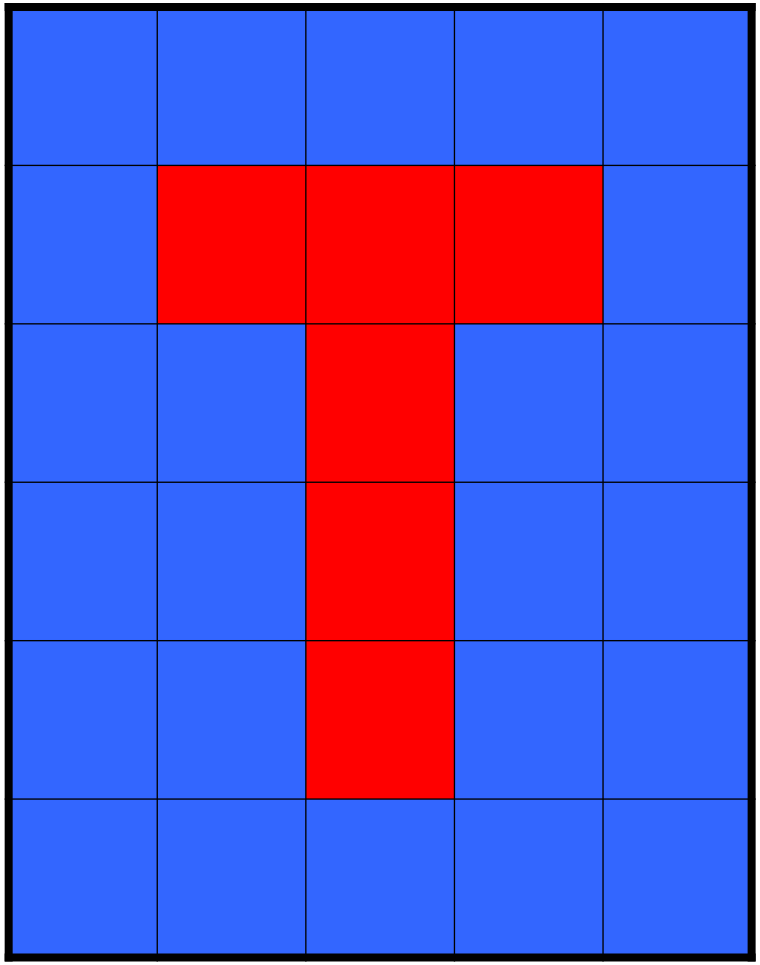
<storage information>



An image

<basic information>
(implicit / explicit)
<rendering information>
(implicit / explicit)
<storage information>
(implicit / explicit)

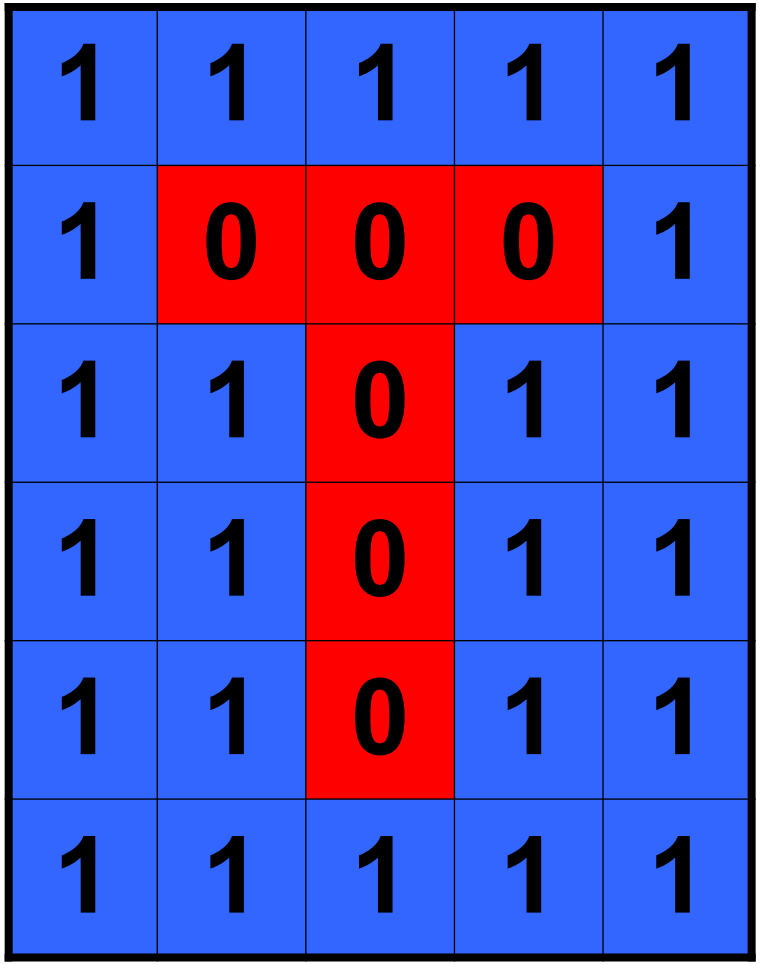
… and the data?



An image

Data either as data stream

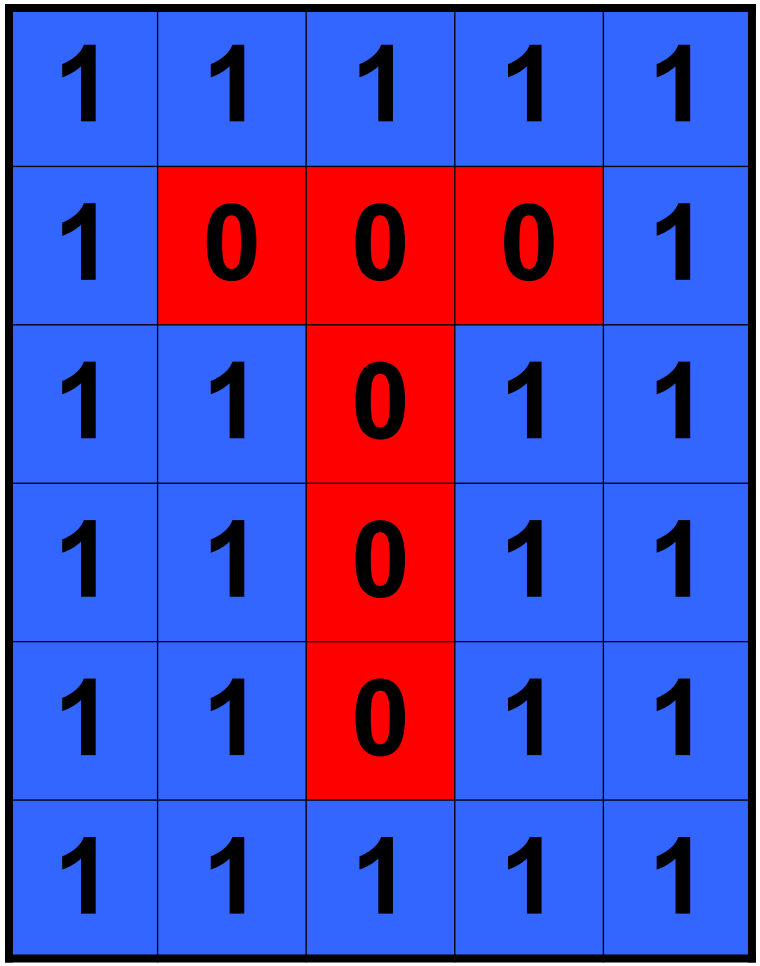
```
1,1,1,1,1,1,
0,0,0,1,1,1,
0,1,1,1,1,0,
1,1,1,1,0,1,
1,1,1,1,1,1
```



An image

Data either as
data stream
or as
processing instructions

```
SetSize: 5 by 6
SetBackgroundColor: Blue
SetForegroundColor: Red
SetLetterHeight: 4
MoveTo: 3,5
DrawLetter: T
```



File Format

- Basic Information
  - What to do?
- Rendering Information
  - How to do It?
- Storage Information
  - How to move it from persistent form to deployed form?
- Data
  - What to deploy?

File Format (2)

- Basic Information
  - Mandatory
- Rendering Information
  - Useful
- Storage Information
  - Historical
- Data
  - Mandatory

File Format - Definition

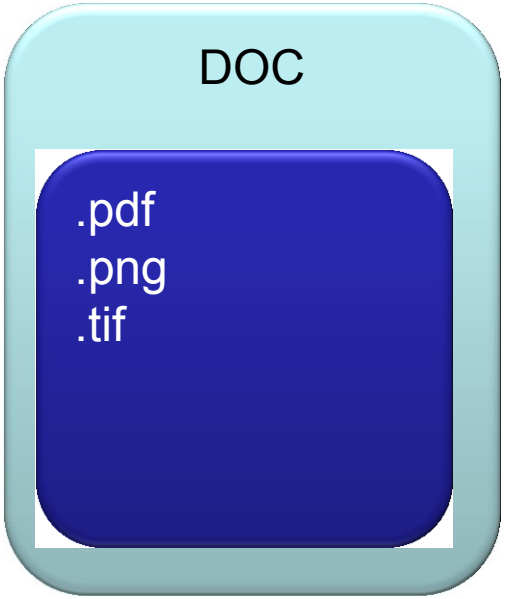
- A clearer definition of the term file form format:

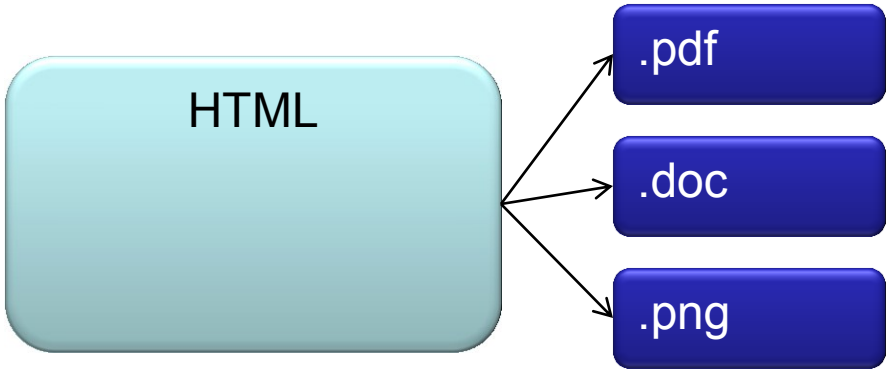
  [...] the internal structure and encoding of a digital object, which allows it to be processed, or to be rendered in human accessible form. A digital object may be a file, or a bit stream embedded within a file‘

Brown, A. (2006). Digital Preservation Technical Paper 2.

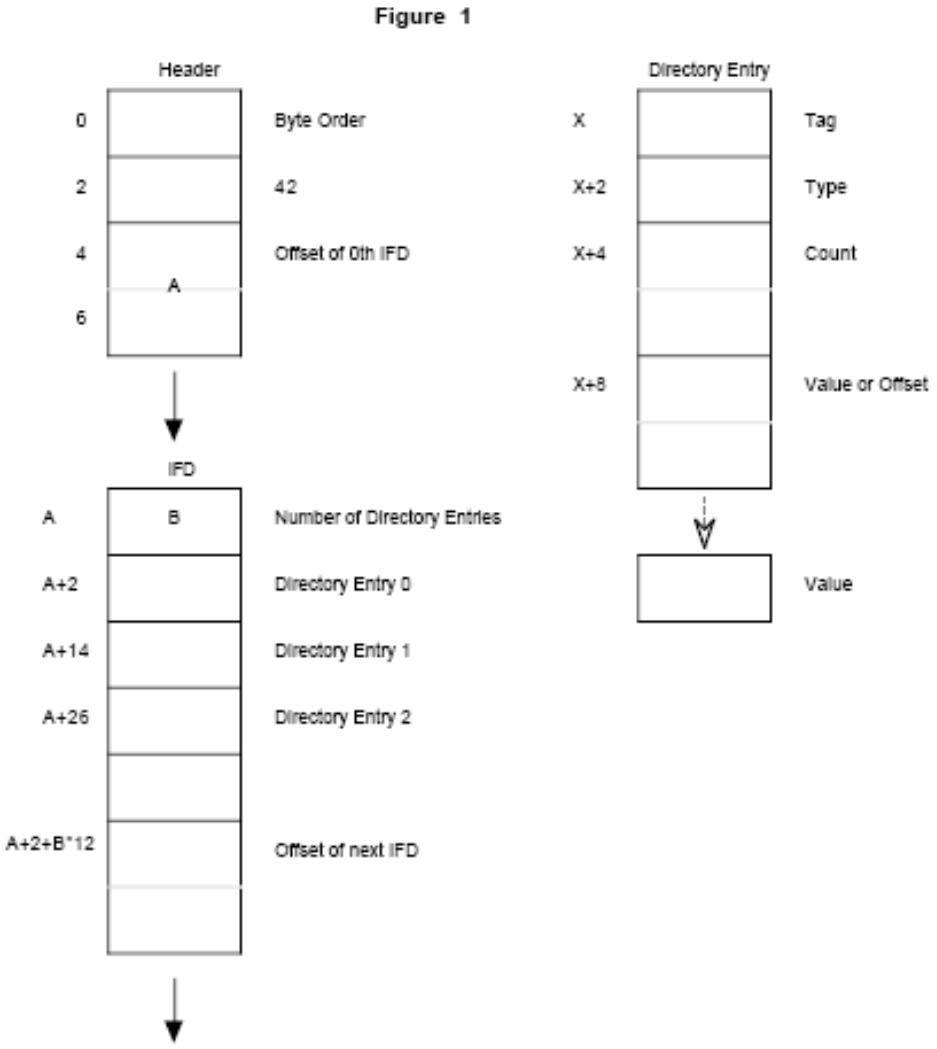
File as a composite object

- Rather popular file formats at them moment are for instance HTML, XML and PNG
- But all of them can be stored in the same file format!





File format: TIFF



File format: PDF

```
1 0 obj
<<
/Type /Page
/Parent 281 0 R
/Resources 2 0 R
/Contents 3 0 R
/StructParents 2
/MediaBox [ 0 0 612 792 ]
/CropBox [ 0 0 612 792 ]
/Rotate 0
>>
endobj
```

File format: PDF

```
2 0 obj
<<
/ProcSet [ /PDF /Text ]
/Font << /TT2 292 0 R /TT4 288 0 R >>
/ExtGState << /GS1 300 0 R >>
/ColorSpace << /Cs6 289 0 R >>
>>
endobj
```

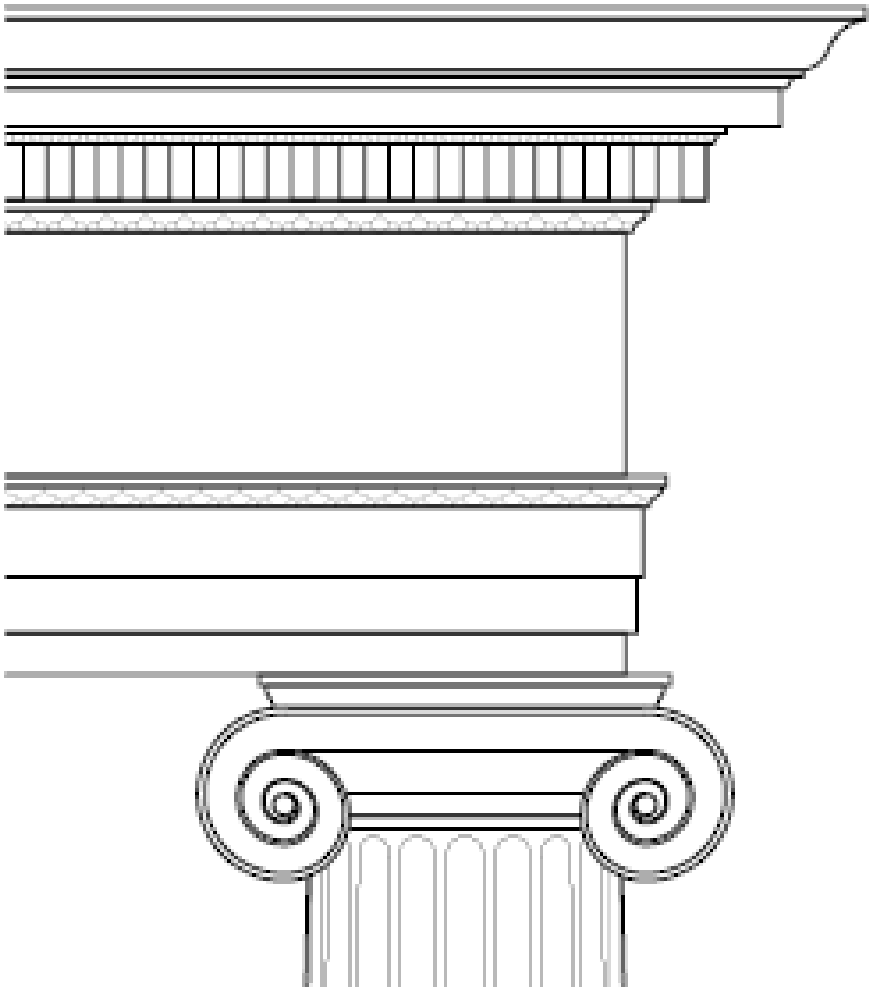
File format: PDF

```
3 0 obj
<< /Length 4605 /Filter /FlateDecode >>
stream
H‰„WÛŽÛÈ}×Wô#Œ4jR”¨`±Àø ™Í"□¶(²5j›"¹lräý'|oêÖ-j
—‹udTÙÂ…fPnˆ¿ìþ>Ó›Ež□ÝÕË½âä”uª2□*<<v ú[ÓžK9Q‰¼‡x»X□P{
‹±/[i²½Ö)}ÔÏö&ªÙH;<Cµ

… and about 4000 bytes more

ŠøL"È÷Û'Æ□JYØÂm]j¥Ýqõ¥Ï°°Õ™·²ôÒ·Û°¤–÷.u-kP0□
4"øTxM<éï§¼9uôøˆòLi¦ØoTÖ m–;Ç¯÷¤ÿlÕ°véU—Ë
±¤Lm°gŸˆu1Åëu5l3¯'¢O %òËTîü7?ìNdh
endstream
endobj
```

File format: XML (SVG)

```
<?xml version="1.0" encoding="UTF-16"?>
  <svg:svg width="800" height="1000" xmlns:svg="http://www.w3.org ...
  <svg:rect x="0" y="0" width="800" height="1000" fill="white" />
  <svg:g transform="translate(-140,0)">
    <svg:line x1="600" y1="20" x2="500" y2="20" stroke="black" …
    <svg:text x="600" y="28.8" font-size="6" fill="black" …
  </svg:g>
  <svg:g transform="translate(-140,0)">
    <svg:text x="500" y="24.4">
      <svg:tspan font-size="4" fill="black">Leiste</svg:tspan>
    </svg:text>
  </svg:g>
    <svg:defs>
      <svg:g id="halbeSaeuleLeiste0">
```

File format: XML (SVG)



Files and Preservation

1. Bit rot.

2. Obscolescence of software.

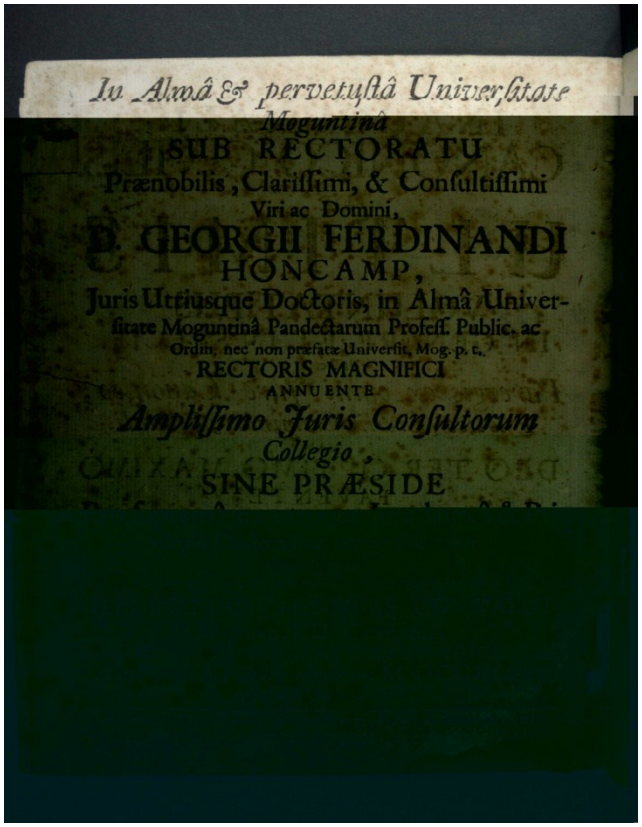
Bit rot

An Image file before ….

In Almâ & pervetuſtâ Univerſitate
Moguntinâ
SUB RECTORATU
Prænobilis, Clariſſimi, & Conſultiſſimi
Viri ac Domini,
D. GEORGII FERDINANDI
HONCAMP,
Juris Utriusque Doctoris, in Almâ Univer-
ſitate Moguntinâ Pandectarum Profeſſ. Public. ac
Ordin. nec non præfatæ Univerſit. Mog. p. t.
RECTORIS MAGNIFICI
ANNUENTE
Ampliſſimo Juris Conſultorum
Collegio,
SINE PRÆSIDE
Pro ſupremâ ex utroque Jure laureâ & Pri-
vilegiis Doctoralibus ritè ac ſolenniter
conſequendis,
Publico ac ſolenni Eruditorum Examini exponit ac ſubmittit
FERDINANDUS HENRICUS DUNWALDT
Mogonus, J. U. Candidatus.
IN COLLEGIO SCHENCKENBERGICO
Anno poſt reparatum mundum M. DCCXI. *die*
Novemb. horis conſuetis
Moguntiæ ex Typographeo JOANNIS MAYERI, *Typographi Aulico-Academici.*

Bit rot

... and after *one* byte is changed.



Undetectable by software.

Bit rot

002	004
234	123
234	156
127	178
221	221

Processing dictionary

Payload

Bit rot

002	004
234	123
234	156
127	xxx
221	221

One byte is damaged, one byte cannot be displayed correctly.

Bit rot

002	xxx
234	123
234	156
127	178
221	221

One byte is damaged, ten bytes cannot be displayed correctly.

Challenges w.r.t. File Formats

- Obsolescence
  - Software able to read does not exist anymore
  - Format specification lost
  - Implied algorithm lost
  - Required object lost
- Format is proprietary
- Format depends on obsolete hardware

Recommended formats?

- XML
- TXT
- PDF
- ?

Recommended formats: text

High confidence	Medium confidence	Low confidence
❖ Plain text (encoding: ISO8859-1 - 9, UTF-8, UTF-16 with BOM) ❖ XML (includes XSD/XSL/XHTML, etc.; with included or accessible schema and character encoding explicitly specified) ❖ PDF/A-1 (ISO 19005-1)	❖ Cascading Style Sheets (*.css) ❖ DTD (*.dtd) ❖ PDF (*.pdf) (embedded fonts) ❖ Rich Text Format 1.x (*.rtf) ❖ HTML 4.x (include a DOCTYPE declaration) ❖ SGML (*.sgml) ❖ Open Office (*.sxw/*.odt) ❖ Office Open XML (*.docx)	❖ PDF (*.pdf) (encrypted) ❖ Microsoft Word (*.doc) ❖ WordPerfect (*.wpd) ❖ DVI (*.dvi) ❖ All other text formats not listed here

http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf

Recommended formats: bitmap / raster image

High confidence	Medium confidence	Low confidence
❖TIFF (uncompressed) ❖ PNG (*.png)	❖ BMP (*.bmp) ❖ JPEG/JFIF (*.jpg) ❖JPEG2000 (prefer lossless or uncompressed) (*.jp2) ❖TIFF (compressed) ❖GIF (*.gif)	❖MrSID (*.sid) ❖TIFF (in Planar format) ❖FlashPix (*.fpx) ❖PhotoShop (*.psd) ❖All other raster image formats not listed here

http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf

Recommended formats: vector graphics

High confidence	Medium confidence	Low confidence
❖ SVG 1.1 (no Java binding) (*.svg)	❖ Computer Graphic Metafile (CGM, WebCGM) (*.cgm)	❖ Encapsulated Postscript (EPS) ❖ Macromedia Flash (*.swf) ❖ All other vector image formats not listed here

http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf

Recommended formats: audio

High confidence	Medium confidence	Low confidence
❖ AIFF (PCM) (*.aif, *.aiff) ❖ WAV (PCM) (*.wav)	❖SUN Audio (uncompressed) (*.au) ❖Standard MIDI (*.mid, *.midi) ❖Ogg Vorbis (*.ogg) ❖Free Lossless Audio Codec (*.flac) ❖ Advance Audio Coding (*.mp4, *.m4a, *.aac) ❖ MP3 (MPEG-1/2, Layer 3)(*.mp3)	❖AIFC (compressed) (*.aifc) ❖ NeXT SND (*.snd) ❖ RealNetworks 'Real Audio, (*.ra, *.rm, *.ram) ❖ Windows Media Audio ❖(*.wma) ❖WAV (compressed) (*.wav) ❖All other audio formats not listed here

http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf

Recommended formats: video

High confidence	Medium confidence	Low confidence
❖ Motion JPEG 2000 (ISO/IEC 15444-4)（ *.mj2) ❖ AVI (uncompressed) (*.avi) ❖ QuickTime Movie (uncompressed)(*.mov) ❖ Motion JPEG (*.avi, *.mov)	❖ Ogg Theora (*.ogg) ❖ MPEG-1, MPEG-2 (*.mpg, *.mpeg) ❖ MPEG-4 (*.mp4)	❖ AVI (compressed) (*.avi) ❖ QuickTime Movie (compressed) (*.mov) ❖ RealNetworks 'Real Video, (*.rv) ❖ Windows Media Video (*.wmv) ❖ All other video formats not listed here

http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf

Recommended formats: “data base”

High confidence	Medium confidence	Low confidence
❖ Delimited Text (*.txt, *.csv) ❖ SQL DDL	❖ DBF (*.dbf) ❖ OpenOffice *.sxc/*.ods) ❖ Office Open XML *.xlsx)	❖ Excel (*.xls) ❖ All other spreadsheet/ database formats not listed here

http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf

Recommended formats: 3D

High confidence	Medium confidence	Low confidence
❖ X3D (*.x3d)	❖ VRML (*.wrl, *.vrml) ❖ U3D (Universal 3D file format)	❖ All other virtual reality ❖ formats not listed here

http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf

Requirements for DP

- Digital preservation has to guarantee
  - Integrity
  - Understandability
  - Originality
  - Authenticity
  - Accessibility

Some file format requirements

- Specifications available (syntax + semantics)
- Standardized (ISO, ANSI, ITEF, ...)
- Accepted and widely used
- Not covered by patent (license fees)
- Free of any cryptographical techniques (risk of losing keys)
- Free of compression

Questions?