

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

.....
FACULTY OF **INFORMATICS**


 **File formats based on plain text**

- For example: HTML

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


- ...

.....
FACULTY OF **INFORMATICS**

 **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
 - ...

.....
FACULTY OF **INFORMATICS**

 **Different types of File Formats (2)**

- Three-character file extension of DOS and Windows. (Neither standardised nor unique.)

- Unix ,magic numbers'

- Macintosh data-forks

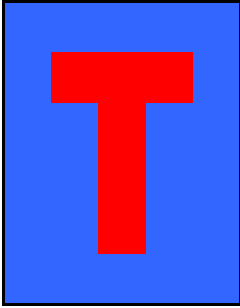
- MIME type, also not unique

- None of them is really satisfying
 - Better solution: PRONOM with Pronom Unique Identifier

.....
FACULTY OF **INFORMATICS**

TU
VIENNA

An image

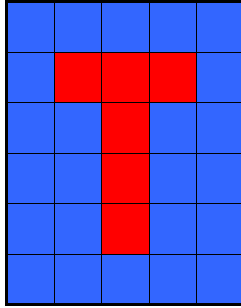


FACULTY OF INFORMATICS

TU
VIENNA

An image

6 rows
5 columns

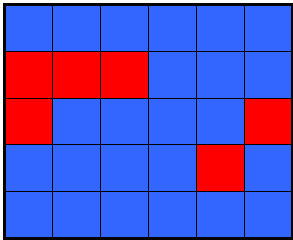


FACULTY OF INFORMATICS

TU
VIENNA

An image

5 rows
6 columns



FACULTY OF INFORMATICS

TU
VIENNA

An image

1 == blue
0 == red

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

FACULTY OF INFORMATICS

TU VIENNA

An image

1 == green
0 == yellow

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

FACULTY OF INFORMATICS

TU VIENNA

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

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

FACULTY OF INFORMATICS

TU VIENNA

An image

Store:

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

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

FACULTY OF INFORMATICS

TU VIENNA

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

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

FACULTY OF INFORMATICS

TU VIENNA 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

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

FACULTY OF INFORMATICS

TU VIENNA An image

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

1,1	2,1	3,1	4,1	5,1
1,2	2,2	3,2	4,2	5,2
1,3	2,3	3,3	4,3	5,3
1,4	2,4	3,4	4,4	5,4
1,5	2,5	3,5	4,5	5,5
1,6	2,6	3,6	4,6	5,6

FACULTY OF INFORMATICS

TU VIENNA An image

<basic
 information>

<rendering
 information>

<storage
 information>

FACULTY OF INFORMATICS

TU VIENNA An image

<basic
 information>
 (implicit / explicit)

<rendering
 information>
 (implicit / explicit)

<storage
 information>
 (implicit / explicit)

... and the data?

FACULTY OF INFORMATICS

TU VIENNA

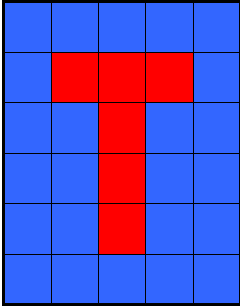
An image

<basic information>
(implicit / explicit)

<rendering information>
(implicit / explicit)

<storage information>
(implicit / explicit)

... and the data?



FACULTY OF INFORMATICS

TU VIENNA

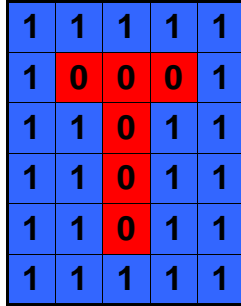
An image

Data either as data stream

```

1,1,1,1,1,1,
0,0,0,1,1,1,
0,1,1,1,1,0,
1,1,1,0,1,1,
1,1,1,0,1,1,
1,1,1,1,1,1

```



FACULTY OF INFORMATICS

TU VIENNA

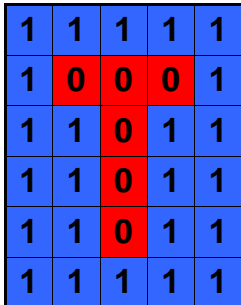
An image

Data either as data stream
or as processing instructions

```

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

```



FACULTY OF INFORMATICS

TU VIENNA

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?

FACULTY OF INFORMATICS

TU VIENNA File Format (2)

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

FACULTY OF INFORMATICS

TU VIENNA 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.

FACULTY OF INFORMATICS

TU VIENNA 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!

FACULTY OF INFORMATICS

TU VIENNA File format: TIFF

Figure 1

FACULTY OF INFORMATICS

TU VIENNA 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

```

FACULTY OF INFORMATICS

TU VIENNA 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

```

FACULTY OF INFORMATICS

TU VIENNA File format: PDF

```

3 0 obj
<< /Length 4605 /Filter /FlateDecode >>
stream
H&„WŰŽŰĚ}xWô#G4jR“`±Åø ƵÍ“ ¶(²5j>“1lräý'|oêô-j
-<udTŰĀ...fPn^žip>ô>Ež²YŌĚ%ââ“u²2i²<<v ū[Ōžk9Q&ž+x>XTP{
<±/[i²%Ō)}Ōİđ&³ŰH;<Cp

... and about 4000 bytes more

ŠøL"Ē+Ű'Ē-JYŌĀm]jYŶqŌYŶĪ°°Ō™.²đŌ·Ű°q-+.u-kP0
4°øTxM<éi$ž9uđø^đLi|ŌoTŌ m-;Ç~+qýlŌ°véU-Ē
±oLm°gŶ^ulĀëu5l3~'çŌ %đĒTfŭ7?iNdh
endstream
endobj

```

FACULTY OF INFORMATICS

TU VIENNA File format: XML (SVG)

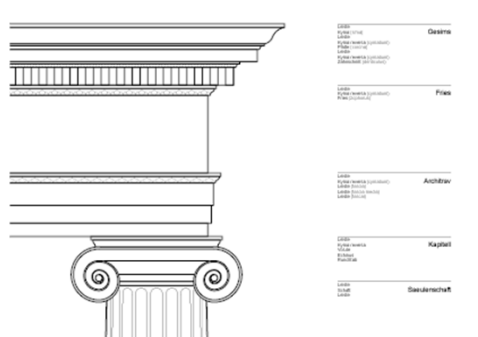
```

<?xml version="1.0" encoding="UTF-16"?>
<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">

```

FACULTY OF INFORMATICS

TU VIENNA File format: XML (SVG)



CLASS	Classica
TYPE	Column
ORDER	1
DESCRIPTION	Column
FILE	Column
DESCRIPTION	Column
ORDER	2
TYPE	Column
ORDER	1
DESCRIPTION	Column
FILE	Column
DESCRIPTION	Column
ORDER	3
TYPE	Column
ORDER	1
DESCRIPTION	Column
FILE	Column
DESCRIPTION	Column
ORDER	4
TYPE	Column
ORDER	1
DESCRIPTION	Column
FILE	Column
DESCRIPTION	Column

FACULTY OF INFORMATICS³

TU VIENNA Files and Preservation

1. Bit rot.
2. Obsolescence of software.

FACULTY OF INFORMATICS

TU VIENNA Bit rot

An Image file before



FACULTY OF INFORMATICS


TU VIENNA Bit rot

... and after one byte is changed.



Undetectable by software.

FACULTY OF INFORMATICS


 Bit rot

002	004
234	123
234	156
127	178
221	221

Processing dictionary

Payload


FACULTY OF INFORMATICS

 Bit rot

002	004
234	123
234	156
127	xxx
221	221

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


FACULTY OF INFORMATICS

 Bit rot

002	xxx
234	123
234	156
127	178
221	221

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

FACULTY OF INFORMATICS

 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

FACULTY OF INFORMATICS

TU VIENNA Recommended formats?

- XML
- TXT
- PDF
- ?

FACULTY OF INFORMATICS

TU VIENNA Recommended formats: text

High confidence	Medium confidence	Low confidence
<ul style="list-style-type: none"> ❖ 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) 	<ul style="list-style-type: none"> ❖ 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) 	<ul style="list-style-type: none"> ❖ 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>

FACULTY OF INFORMATICS

TU VIENNA Recommended formats: bitmap / raster image

High confidence	Medium confidence	Low confidence
<ul style="list-style-type: none"> ❖ TIFF (uncompressed) ❖ PNG (*.png) 	<ul style="list-style-type: none"> ❖ BMP (*.bmp) ❖ JPEG/JFIF (*.jpg) ❖ JPEG2000 (prefer lossless or uncompressed) (*.jp2) ❖ TIFF (compressed) ❖ GIF (*.gif) 	<ul style="list-style-type: none"> ❖ 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>

FACULTY OF INFORMATICS

TU VIENNA Recommended formats: vector graphics

High confidence	Medium confidence	Low confidence
<ul style="list-style-type: none"> ❖ SVG 1.1 (no Java binding) (*.svg) 	<ul style="list-style-type: none"> ❖ Computer Graphic Metafile (CGM, WebCGM) (*.cgm) 	<ul style="list-style-type: none"> ❖ Encapsulated Postscript (EPS) ❖ Macromedia Flash (*.swf) ❖ All other vector image formats not listed here

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

FACULTY OF INFORMATICS

TU VIENNA Recommended formats: audio

High confidence	Medium confidence	Low confidence
<ul style="list-style-type: none"> ❖ AIFF (PCM) (*.aif, *.aiff) ❖ WAV (PCM) (*.wav) 	<ul style="list-style-type: none"> ❖ 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) 	<ul style="list-style-type: none"> ❖ 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>

FACULTY OF **INFORMATICS**

TU VIENNA Recommended formats: video

High confidence	Medium confidence	Low confidence
<ul style="list-style-type: none"> ❖ Motion JPEG 2000 (ISO/IEC 15444-4) (*.mj2) ❖ AVI (uncompressed) (*.avi) ❖ QuickTime Movie (uncompressed) (*.mov) ❖ Motion JPEG (*.avi, *.mov) 	<ul style="list-style-type: none"> ❖ Ogg Theora (*.ogg) ❖ MPEG-1, MPEG-2 (*.mpg, *.mpeg) ❖ MPEG-4 (*.mp4) 	<ul style="list-style-type: none"> ❖ 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>

FACULTY OF **INFORMATICS**

TU VIENNA Recommended formats: "data base"

High confidence	Medium confidence	Low confidence
<ul style="list-style-type: none"> ❖ Delimited Text (*.txt, *.csv) ❖ SQL DDL 	<ul style="list-style-type: none"> ❖ DBF (*.dbf) ❖ OpenOffice *.sxc/*.ods ❖ Office Open XML *.xlsx) 	<ul style="list-style-type: none"> ❖ Excel (*.xls) ❖ All other spreadsheet/database formats not listed here

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

FACULTY OF **INFORMATICS**

TU VIENNA Recommended formats: 3D

High confidence	Medium confidence	Low confidence
<ul style="list-style-type: none"> ❖ X3D (*.x3d) 	<ul style="list-style-type: none"> ❖ VRML (*.vrl, *.vrm) ❖ U3D (Universal 3D file format) 	<ul style="list-style-type: none"> ❖ All other virtual reality formats not listed here

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

FACULTY OF **INFORMATICS**



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



Thank you very much for your attention!

Questions?