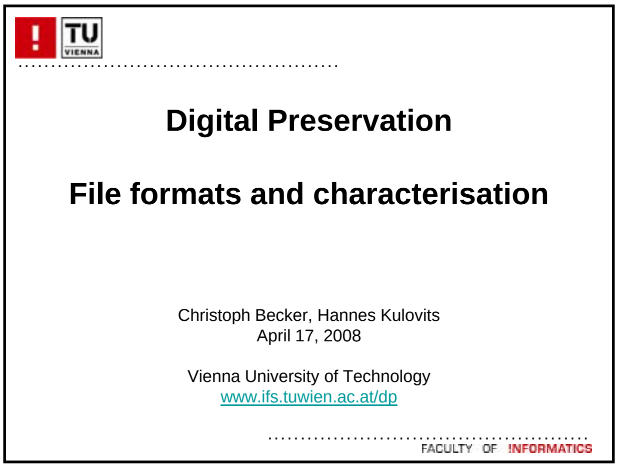
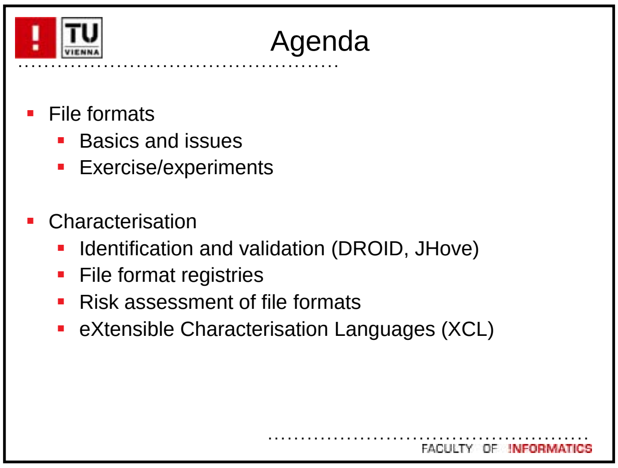


# Digital Preservation

# File formats and characterisation

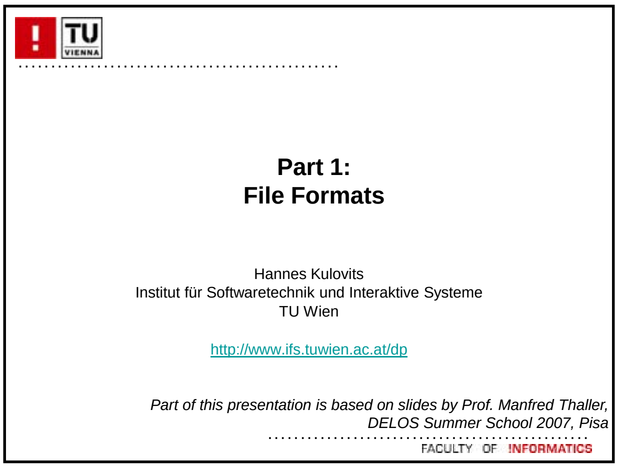
Christoph Becker, Hannes Kulovits
April 17, 2008

Vienna University of Technology
www.ifs.tuwien.ac.at/dp



## Agenda

- File formats
  - Basics and issues
  - Exercise/experiments
- Characterisation
  - Identification and validation (DROID, JHove)
  - File format registries
  - Risk assessment of file formats
  - eXtensible Characterisation Languages (XCL)

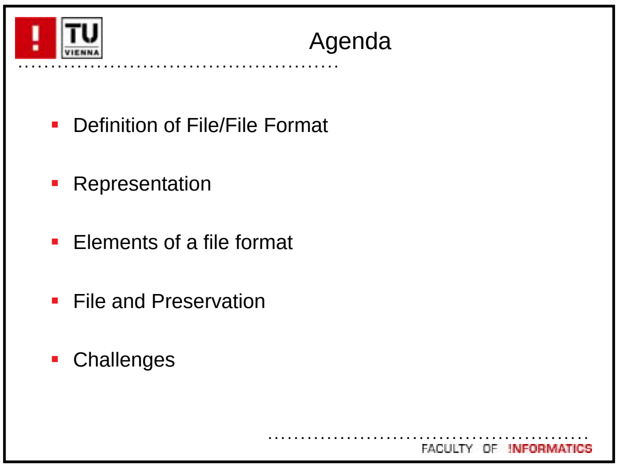


## Part 1:
## File Formats

Hannes Kulovits
Institut für Softwaretechnik und Interaktive Systeme
TU Wien

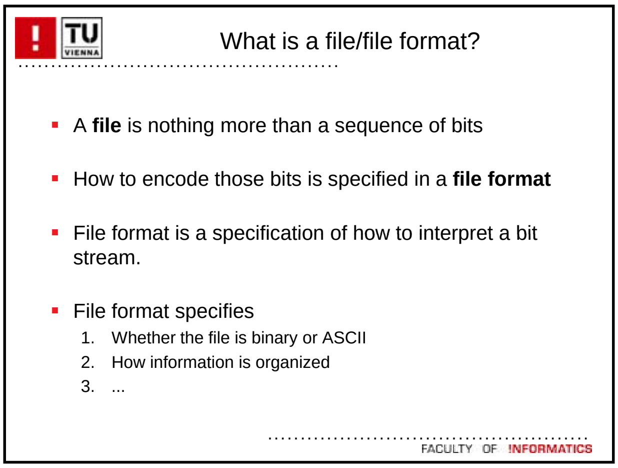
http://www.ifs.tuwien.ac.at/dp

*Part of this presentation is based on slides by Prof. Manfred Thaller, DELOS Summer School 2007, Pisa*



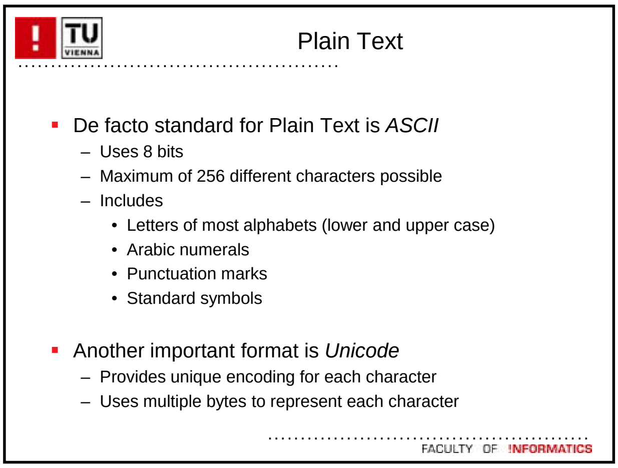
## Agenda

- 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: XHTML 1.1
- In HTML plain text must obey certain rules
  - se of tags
  - type sizes
  - color

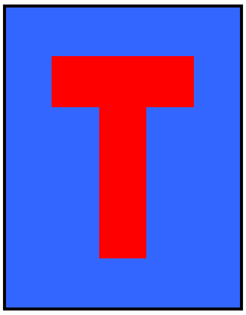
## Different types of File Formats

- 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
  - ...

## Different types of File Formats (2)

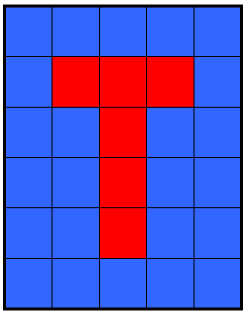
- Three-character file extension of DOS and Windows. (Neither standardised nor unique.)
- Unix ‚magic numbers‘
- Macintosh data-forks
- MIME type, also not unique
- None of them is really satisfying
  - Better solution: PRONOM with Pronom Unique Identifier

## An image



## An image

6 rows
5 columns



## An image

5 rows
6 columns

| | | | | | |
|---|---|---|---|---|---|
| 🟦 | 🟦 | 🟦 | 🟦 | 🟦 | 🟦 |
| 🟥 | 🟥 | 🟥 | 🟦 | 🟦 | 🟦 |
| 🟥 | 🟦 | 🟦 | 🟦 | 🟦 | 🟥 |
| 🟦 | 🟦 | 🟦 | 🟦 | 🟥 | 🟦 |
| 🟦 | 🟦 | 🟦 | 🟦 | 🟦 | 🟦 |

## An image

1 == blue
0 == red

| | | | | |
|---|---|---|---|---|
| 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

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 |

## 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 |

## An image

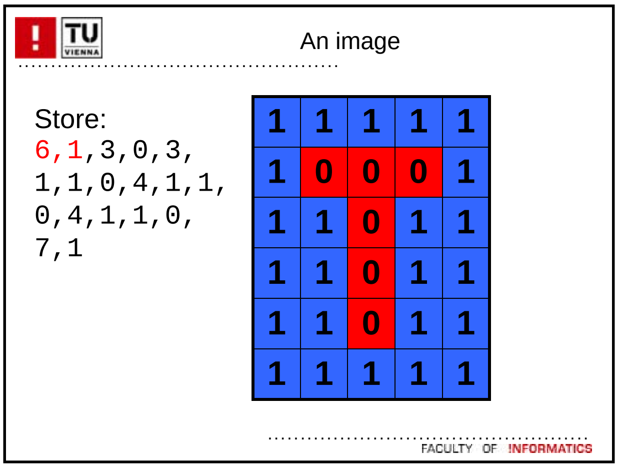
Store:
6,1,3,0,3,
1,1,0,4,1,1,
0,4,1,1,0,
7,1

| | | | | |
|---|---|---|---|---|
| 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 |

## 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 |

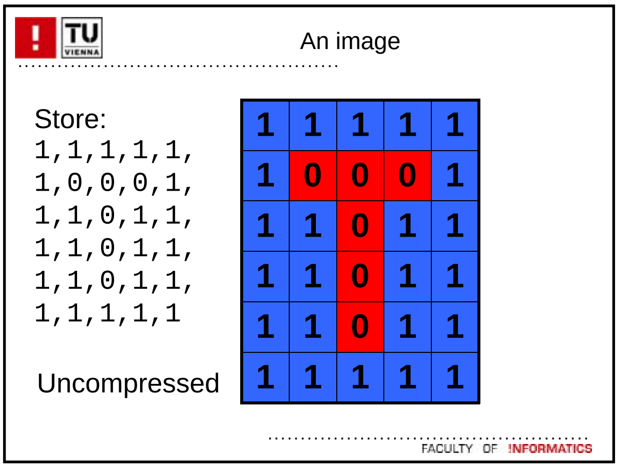


## An image

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 !NFORMATICS



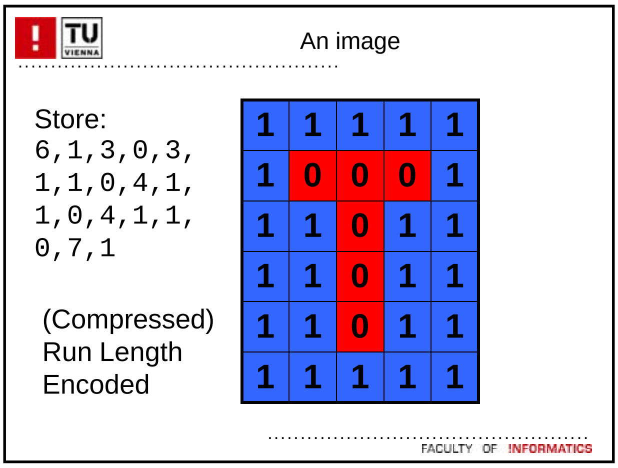
## An image

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 !NFORMATICS



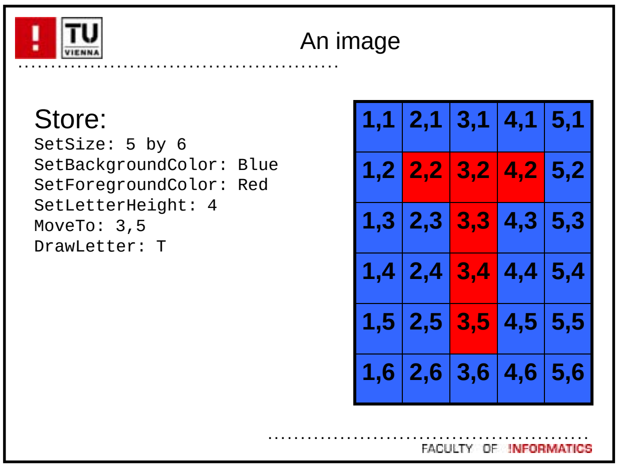
## An image

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 !NFORMATICS



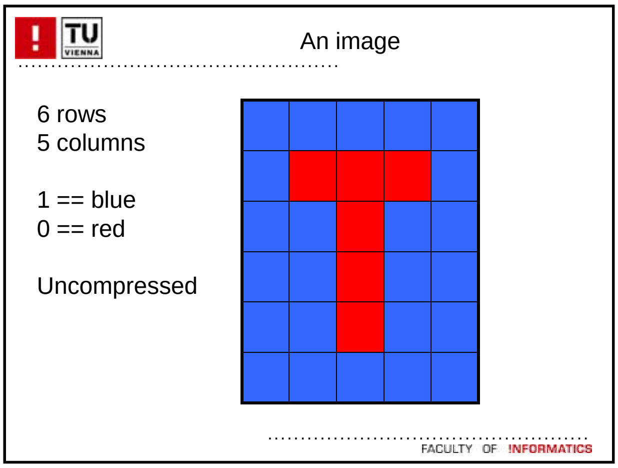
## An image

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 !NFORMATICS



## An image

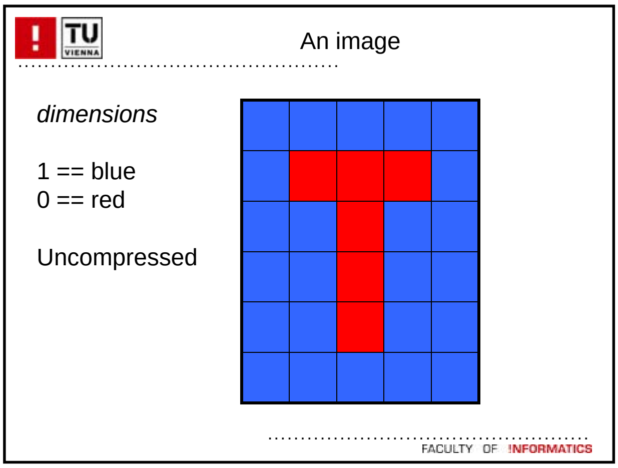
6 rows
5 columns

1 == blue
0 == red

Uncompressed

| | | | | |
|---|---|---|---|---|
| 🟦 | 🟦 | 🟦 | 🟦 | 🟦 |
| 🟦 | 🟥 | 🟥 | 🟥 | 🟦 |
| 🟦 | 🟦 | 🟥 | 🟦 | 🟦 |
| 🟦 | 🟦 | 🟥 | 🟦 | 🟦 |
| 🟦 | 🟦 | 🟥 | 🟦 | 🟦 |
| 🟦 | 🟦 | 🟦 | 🟦 | 🟦 |

FACULTY OF !NFORMATICS



## An image

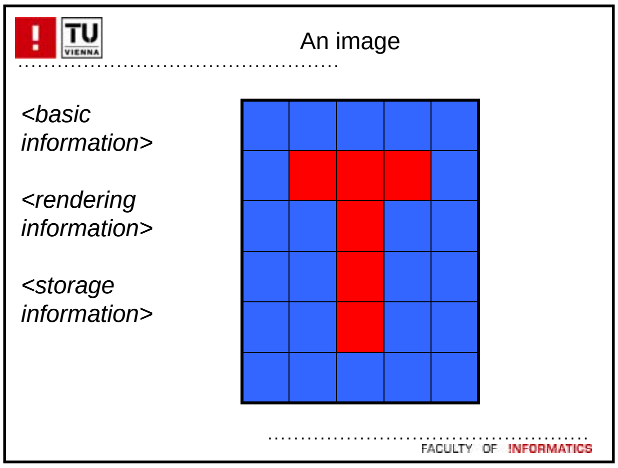
*dimensions*

1 == blue
0 == red

Uncompressed

| | | | | |
|---|---|---|---|---|
| 🟦 | 🟦 | 🟦 | 🟦 | 🟦 |
| 🟦 | 🟥 | 🟥 | 🟥 | 🟦 |
| 🟦 | 🟦 | 🟥 | 🟦 | 🟦 |
| 🟦 | 🟦 | 🟥 | 🟦 | 🟦 |
| 🟦 | 🟦 | 🟥 | 🟦 | 🟦 |
| 🟦 | 🟦 | 🟦 | 🟦 | 🟦 |

FACULTY OF !NFORMATICS



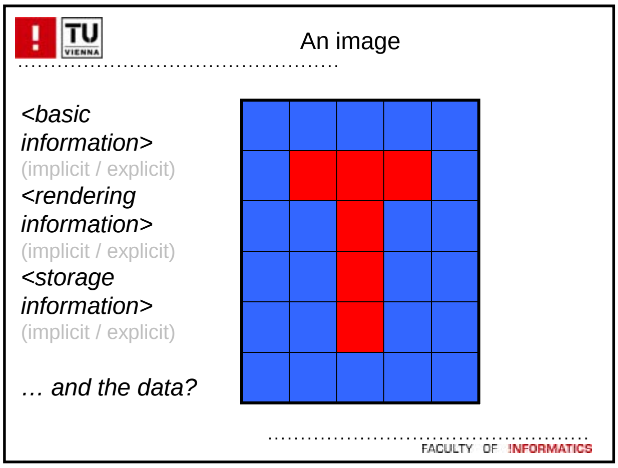
## An image

*<basic information>*

*<rendering information>*

*<storage information>*

🟦🟦🟦🟦🟦
🟦🟥🟥🟥🟦
🟦🟦🟥🟦🟦
🟦🟦🟥🟦🟦
🟦🟦🟥🟦🟦
🟦🟦🟦🟦🟦

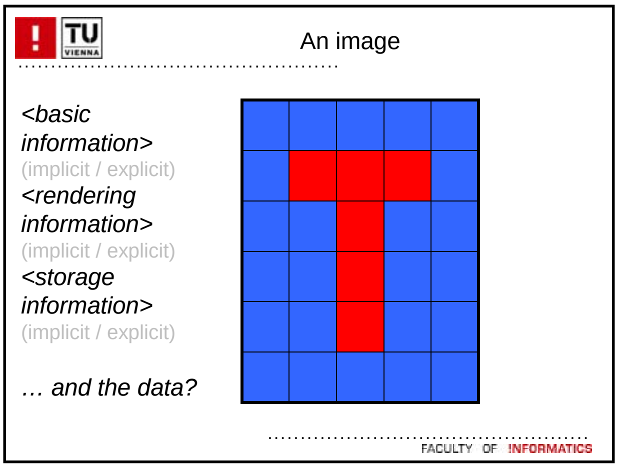


## An image

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

*… and the data?*

🟦🟦🟦🟦🟦
🟦🟥🟥🟥🟦
🟦🟦🟥🟦🟦
🟦🟦🟥🟦🟦
🟦🟦🟥🟦🟦
🟦🟦🟦🟦🟦

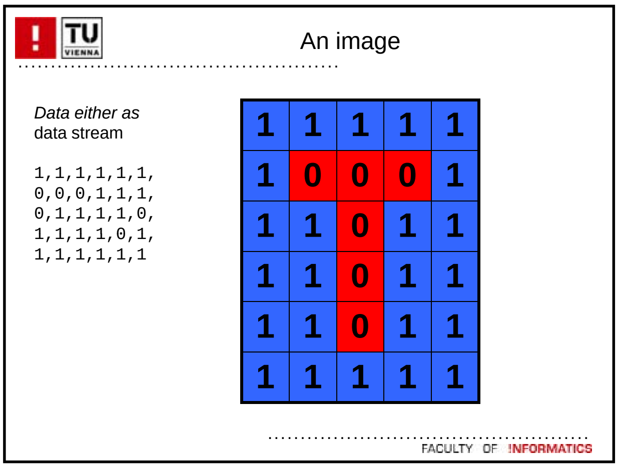


## An image

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

*… and the data?*

🟦🟦🟦🟦🟦
🟦🟥🟥🟥🟦
🟦🟦🟥🟦🟦
🟦🟦🟥🟦🟦
🟦🟦🟥🟦🟦
🟦🟦🟦🟦🟦

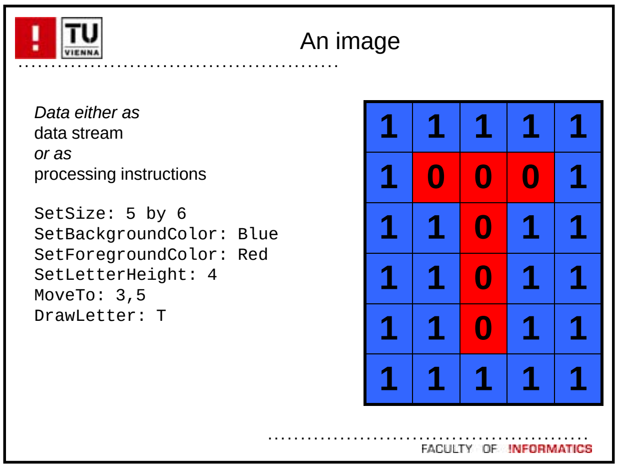


## An image

*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
```

| | | | | |
|---|---|---|---|---|
| 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 |

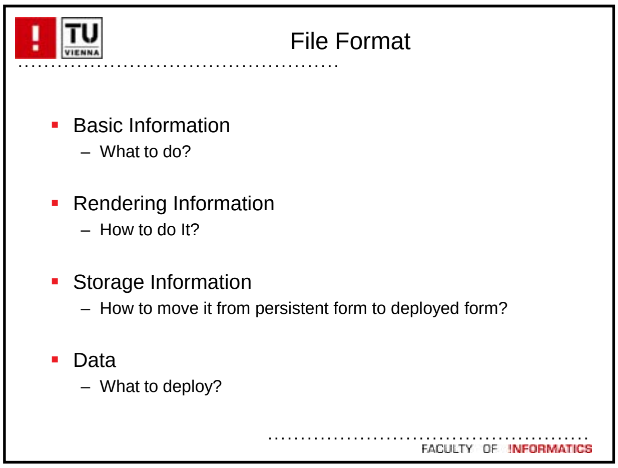


## An image

*Data either as*
data stream
*or as*
processing instructions

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

| | | | | |
|---|---|---|---|---|
| 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 |



## 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

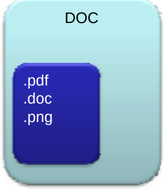
- A clearer definition of the term file form format:

  ‚[...] 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'

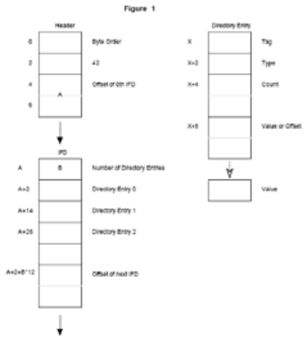
  *Brown, A. (2006). Digital Preservation Technical Paper 2.*

## 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



- Image File Header
- Image File Directory
  - Information about image
  - Pointers to actual data
- IFD Entry
  - TIFF Tag
  - Value
- Custom tags possible

## 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
```

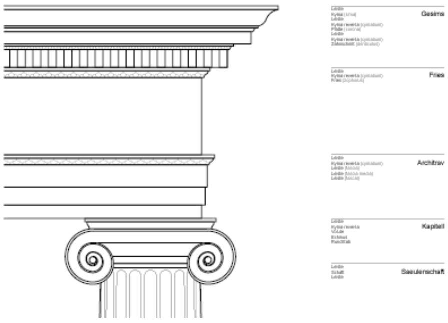
## File format: PDF

```
3 0 obj
<< /Length 4605 /Filter /FlateDecode >>
stream
H‰„WÛŽÛÈ}×Wô#Œ4jR”¨`±Àø ™Í"  ¶(²5j›"¹lräý‘|oêÖ-j
—‹udTÙÂ…fPnˆ¿ìþ>Ó›Ež ²Ý ÕË½âä”uª2 i*<<v ú[Óžk9Q‰¼‡x»X TP{
‹±/[i²½Ö)}ÔÏö&ªÙH;<Cµ

… and about 4000 bytes more

ŠøL"È÷Û’Æ ¬JYØÂm]j¥ÝqõÏºº¥Õ™·²ôÒ·Ûº¤–÷.u-kP0
4“øTxM<éï§¼9uôøˆòLi¦Øo TÖ m–;Ç¯ ÷¤ÿlÕºvéU—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)

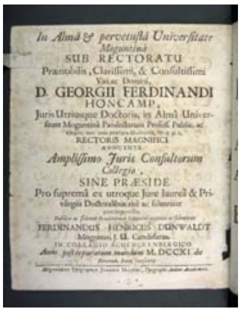


## Files and Preservation

1. Bit rot.
2. Obscolescence of software.

## Bit rot

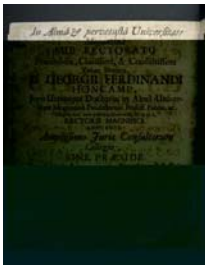
An Image file before ….



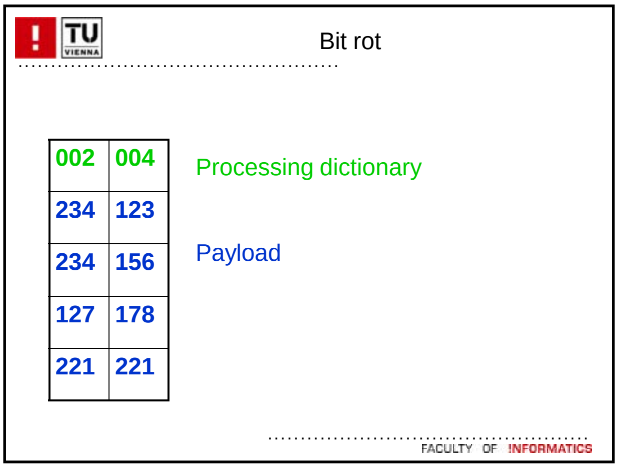
In Almâ & perveteuſtâ Univerſitate
Moguntinâ
SUB RECTORATU
Prænobilis, Clariſſimi, & Conſultiſſimi
[illegible]
D. GEORGII FERDINANDI
HONCAMP,
Juris Utriusque Doctoris, in Almâ Univer-
ſitate Moguntinâ [illegible]
[illegible]
RECTORIS MAGNIFICI
[illegible]
Ampliſſimo Juris Conſultorum
Collegio,
SINE PRÆSIDE
Pro ſupremâ ex utroque Jure laureâ & Pri-
vilegiis [illegible]
[illegible]
FERDINANDUS HENRICUS DUNWALDT
[illegible]
IN COLLEGIO [illegible]
Anno [illegible] M. DCCXI. [illegible]
[illegible]

## Bit rot

… and after *one* byte is changed.



Undetectable by software.

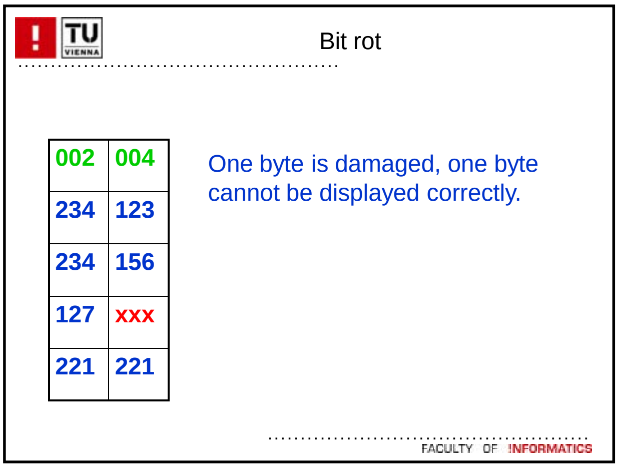


## Bit rot

| | |
|---|---|
| 002 | 004 |
| 234 | 123 |
| 234 | 156 |
| 127 | 178 |
| 221 | 221 |

Processing dictionary

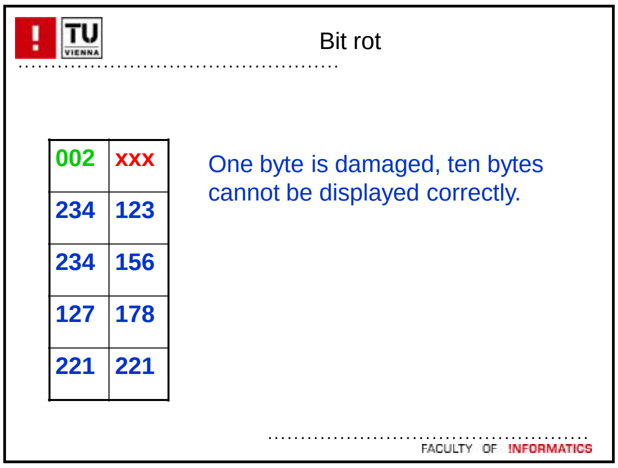
Payload



## Bit rot

| | |
|---|---|
| 002 | 004 |
| 234 | 123 |
| 234 | 156 |
| 127 | xxx |
| 221 | 221 |

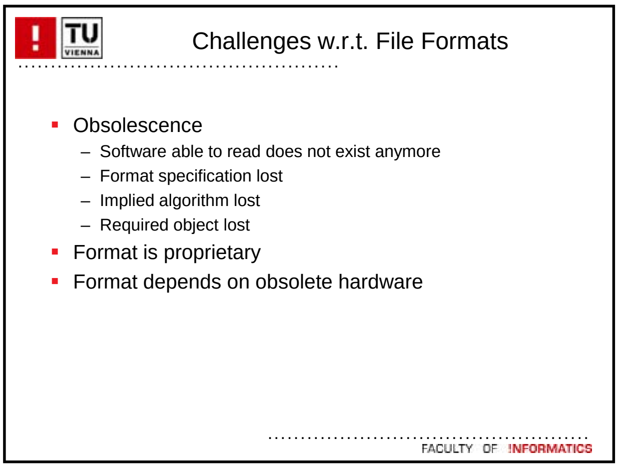
One byte is damaged, one byte cannot be displayed correctly.



## Bit rot

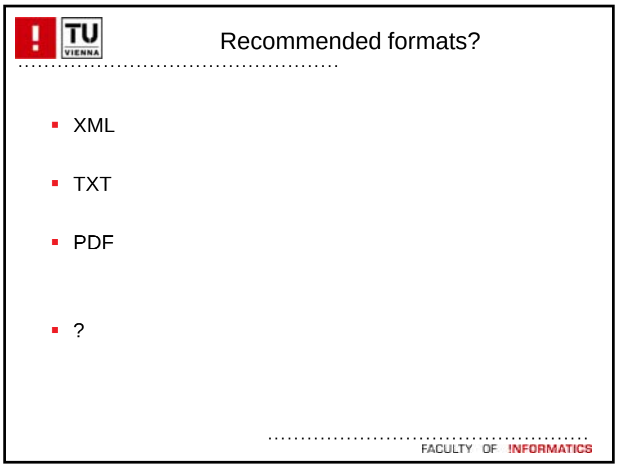
| | |
|---|---|
| 002 | xxx |
| 234 | 123 |
| 234 | 156 |
| 127 | 178 |
| 221 | 221 |

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



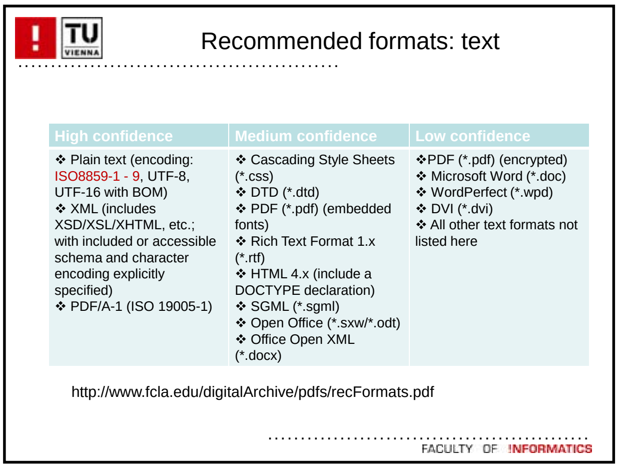
## Challenges w.r.t. File Formats

- 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?

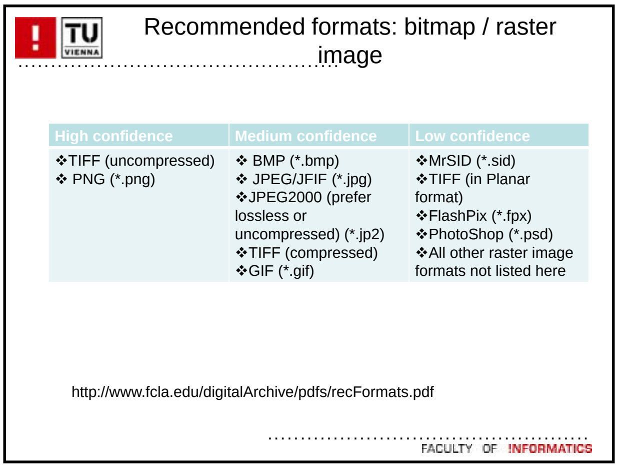
- XML
- TXT
- PDF
- ?



## Recommended formats: text

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

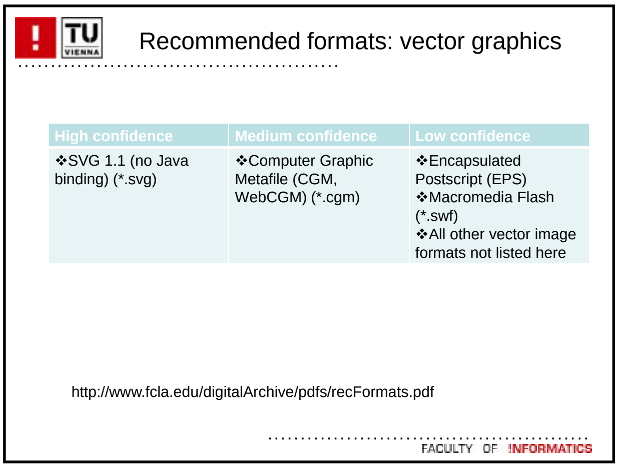
http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf



## Recommended formats: bitmap / raster image

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

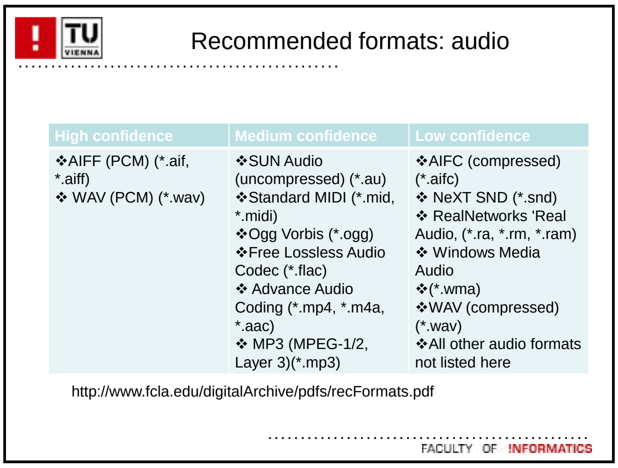
http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf



## 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)<br>❖Macromedia Flash (*.swf)<br>❖All other vector image formats not listed here |

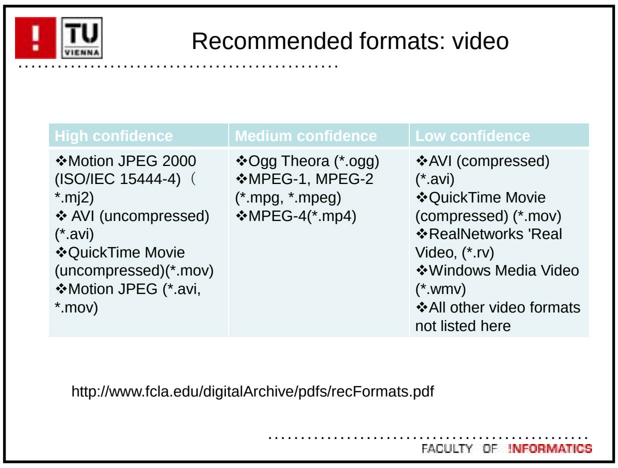
http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf



## Recommended formats: audio

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

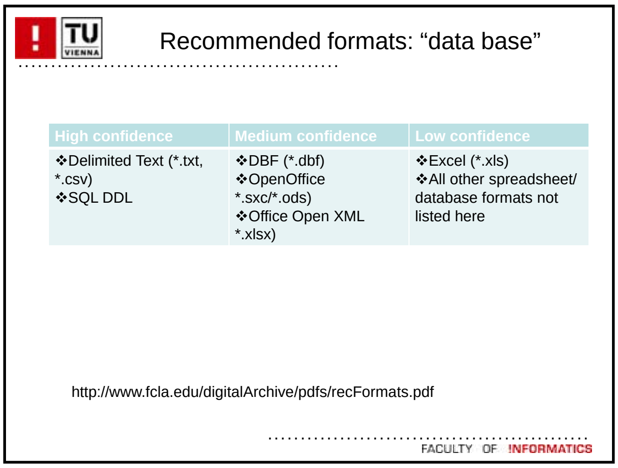
http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf



## Recommended formats: video

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

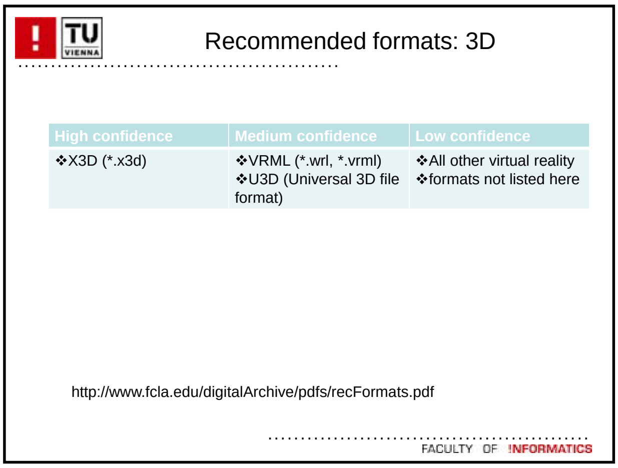
http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf



## Recommended formats: “data base”

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

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



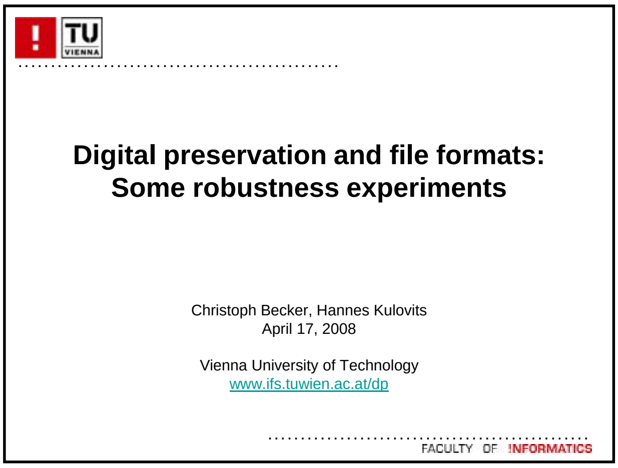
## Recommended formats: 3D

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

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

Thank you very much for your attention!

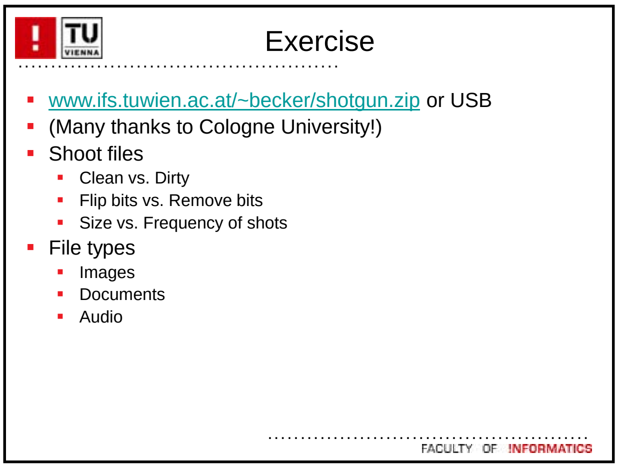
Questions?



## Digital preservation and file formats: Some robustness experiments

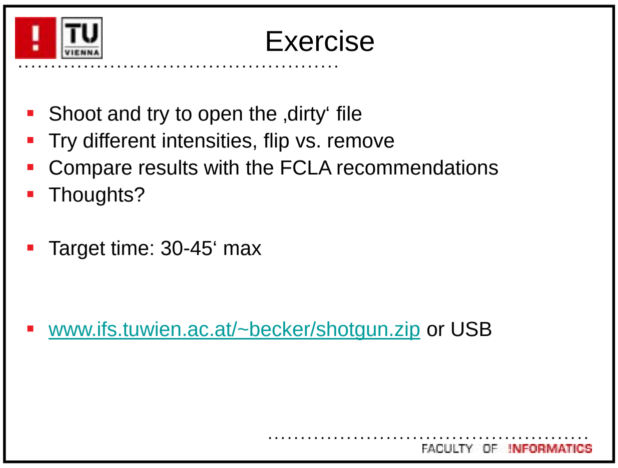
Christoph Becker, Hannes Kulovits
April 17, 2008

Vienna University of Technology
www.ifs.tuwien.ac.at/dp



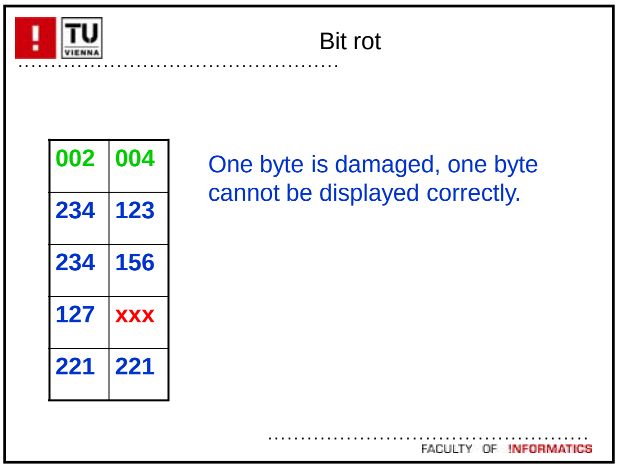
## Exercise

- www.ifs.tuwien.ac.at/~becker/shotgun.zip or USB
- (Many thanks to Cologne University!)
- Shoot files
  - Clean vs. Dirty
  - Flip bits vs. Remove bits
  - Size vs. Frequency of shots
- File types
  - Images
  - Documents
  - Audio



## Exercise

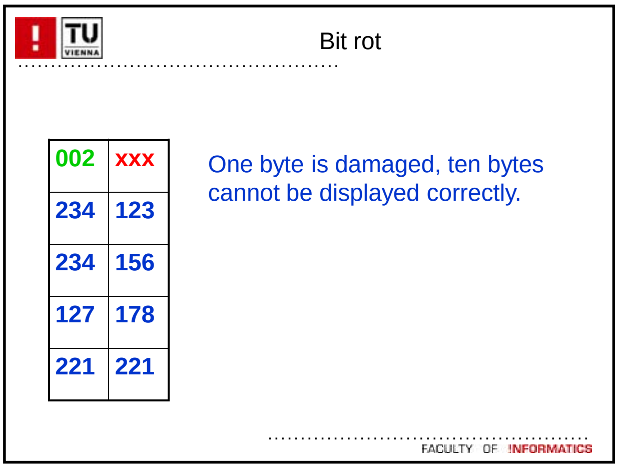
- Shoot and try to open the ‚dirty‘ file
- Try different intensities, flip vs. remove
- Compare results with the FCLA recommendations
- Thoughts?
- Target time: 30-45‘ max
- www.ifs.tuwien.ac.at/~becker/shotgun.zip or USB



## Bit rot

| | |
|---|---|
| 002 | 004 |
| 234 | 123 |
| 234 | 156 |
| 127 | xxx |
| 221 | 221 |

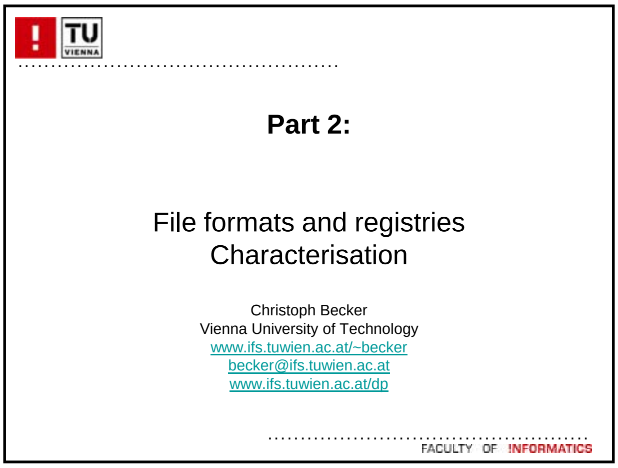
One byte is damaged, one byte cannot be displayed correctly.



## Bit rot

| | |
|---|---|
| 002 | xxx |
| 234 | 123 |
| 234 | 156 |
| 127 | 178 |
| 221 | 221 |

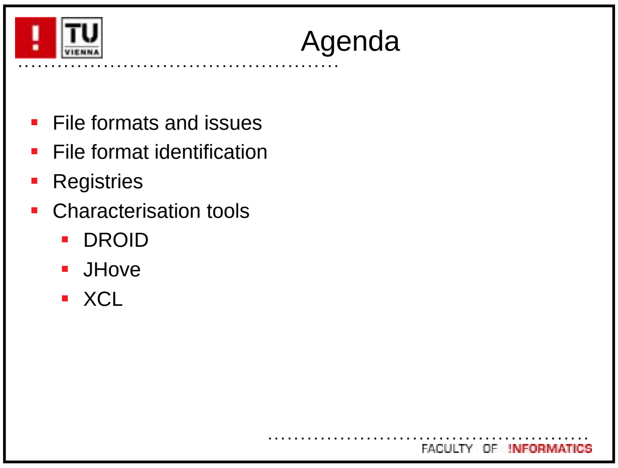
One byte is damaged, ten bytes cannot be displayed correctly.



## Part 2:

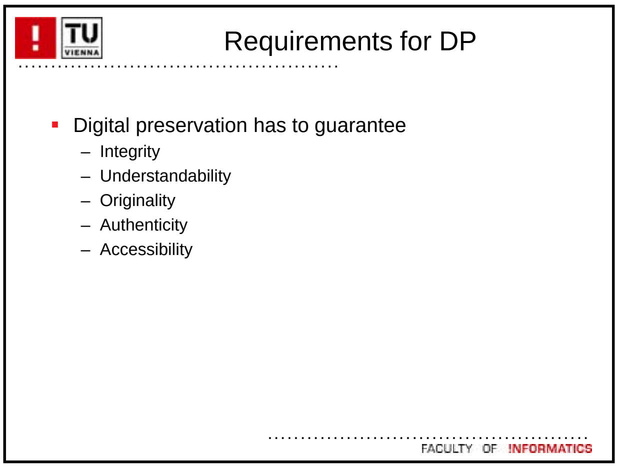
## File formats and registries
## Characterisation

Christoph Becker
Vienna University of Technology
www.ifs.tuwien.ac.at/~becker
becker@ifs.tuwien.ac.at
www.ifs.tuwien.ac.at/dp



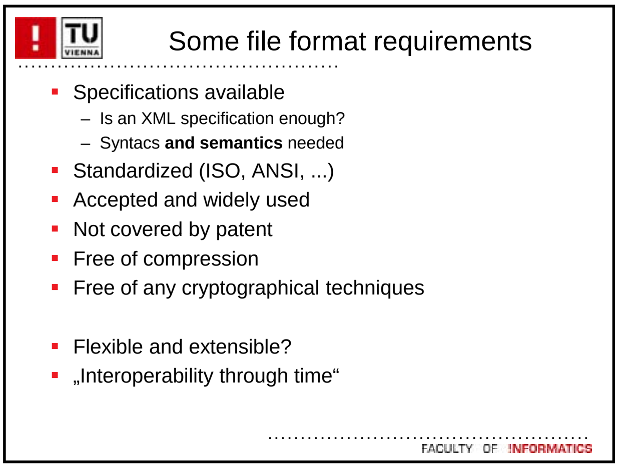
## Agenda

- File formats and issues
- File format identification
- Registries
- Characterisation tools
  - DROID
  - JHove
  - XCL



## Requirements for DP

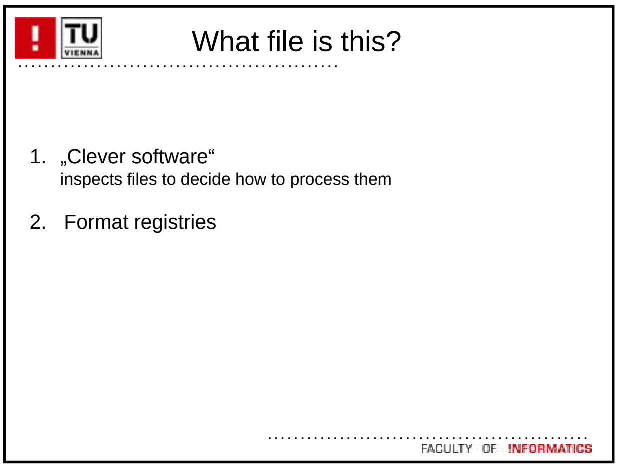
- Digital preservation has to guarantee
  - Integrity
  - Understandability
  - Originality
  - Authenticity
  - Accessibility



## Some file format requirements

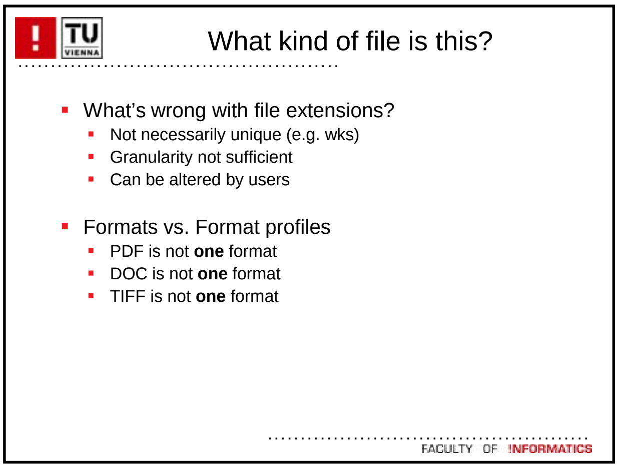
- Specifications available
  - Is an XML specification enough?
  - Syntacs **and semantics** needed
- Standardized (ISO, ANSI, ...)
- Accepted and widely used
- Not covered by patent
- Free of compression
- Free of any cryptographical techniques

- Flexible and extensible?
- „Interoperability through time“



## What file is this?

1. „Clever software“
   inspects files to decide how to process them
2. Format registries



## What kind of file is this?

- What’s wrong with file extensions?
  - Not necessarily unique (e.g. wks)
  - Granularity not sufficient
  - Can be altered by users

- Formats vs. Format profiles
  - PDF is not **one** format
  - DOC is not **one** format
  - TIFF is not **one** format

## What's Wrong with MIME Types?

- Insufficient depth of detail
  - No requirements regarding syntax and semantic description
  - No requirement for complete disclosure, especially of proprietary formats
- Insufficient granularity
  - Both tiled RGB GeoTIFF with LZW and striped bi-tonal TIFF-FX with Group 4 are typed as "image/tiff"
  - All of PDF 1.0 – 1.4, PDF/X-1, X-2, X-3, and PDF/A are typed as "application/pdf"
  - These variants might require radically different workflows

## Why Do We Need a Registry?

- Repository functions are performed on a format-specific basis
- Interpretation of otherwise opaque content streams is dependent upon knowledge of how typed content is represented
- Interchange requires mutual agreement of format syntax and semantics

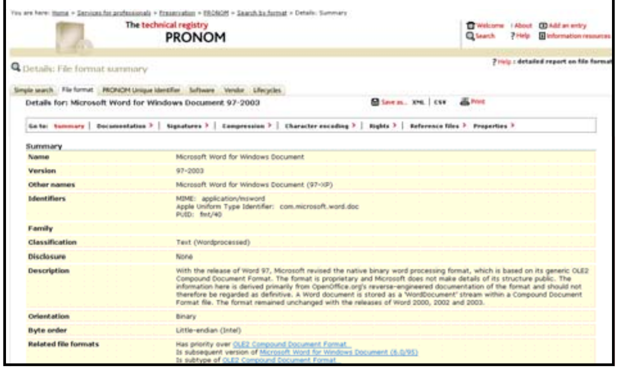
## Use Cases

- Identification
  - "I have a digital object; what format is it?"
- Validation
  - "I have an object purportedly of format *F*; is it?"
- Transformation
  - "I have an object of format *F*, but need *G*; how can I produce it?"
- Characterization
  - "I have an object of format *F*; what are its significant properties?"
- Risk assessment
  - "I have an object of format *F*; is it at risk of obsolescence?"
- Delivery
  - "I have an object of format *F*; how can I render it?"

## File format registries

- PRONOM:
  http://www.nationalarchives.gov.uk/pronom/
- Global Digital Format Registry
  http://hul.harvard.edu/gdfr
- FileExt
  http://filext.com

## PRONOM



You are here: Home » Services for professionals » Preservation » PRONOM » Search by format » Details: Summary

The technical registry
PRONOM
Welcome
About
Add an entry
Search
Help
Information resources
Help : detailed report on file format
Details: File format summary
Simple search
File format
PRONOM Unique Identifier
Software
Vendor
Lifecycles
Save as... XML | CSV
Print
Go to: Summary | Documentation | Signatures | Compression | Character encoding | Rights | Reference files | Properties

Details for: Microsoft Word for Windows Document 97-2003

Summary

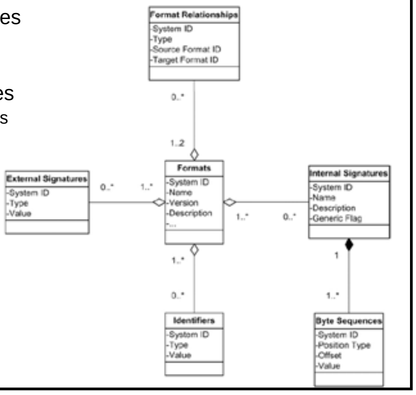
| | |
|---|---|
| Name | Microsoft Word for Windows Document |
| Version | 97-2003 |
| Other names | Microsoft Word for Windows Document (97-XP) |
| Identifiers | MIME: application/msword<br>Apple Uniform Type Identifier: com.microsoft.word.doc<br>PUID: fmt/40 |
| Family | |
| Classification | Text (Wordprocessed) |
| Disclosure | None |
| Description | With the release of Word 97, Microsoft revised the native binary word processing format, which is based on its generic OLE2 Compound Document Format. The format is proprietary and Microsoft does not make details of its structure public. The information here is derived primarily from OpenOffice.org's reverse-engineered documentation of the format and should not therefore be regarded as definitive. A Word document is stored as a 'WordDocument' stream within a Compound Document Format file. The format remained unchanged with the releases of Word 2000, 2002 and 2003. |
| Orientation | Binary |
| Byte order | Little-endian (Intel) |
| Related file formats | Has priority over OLE2 Compound Document Format<br>Is subsequent version of Microsoft Word for Windows Document (6.0/95)<br>Is subtype of OLE2 Compound Document Format |

## Tools

- DROID (Digital Record Object Identification)
  - relies on PRONOM
  - The National Archives, UK
- JHOVE
  - JSTOR/Harvard Object Validation Environment
  - Validation and characterisation
- eXtensible Characterisation Languages (XCL)
  - Two XML meta-languages
  - Goal: express complete informational content of an object in an abstract model

## Signatures in DROID

- External signatures
  - File extensions
- Internal signatures
  - Format indicators in the bitstream
  - Byte sequences



Format Relationships
-System ID
-Type
-Source Format ID
-Target Format ID

0..*

1..2

Formats
-System ID
-Name
-Version
-Description
-...

External Signatures
-System ID
-Type
-Value

0..*

1..*

1..*

0..*

Internal Signatures
-System ID
-Name
-Description
-Generic Flag

1..*

0..*

Identifiers
-System ID
-Type
-Value

1

1..*

Byte Sequences
-System ID
-Position Type
-Offset
-Value

## What kind of file is this?

(a) By external characteristics (file extensions)
(b) By internal characteristics („magic number“, „signature“).

A TIFF file begins with …

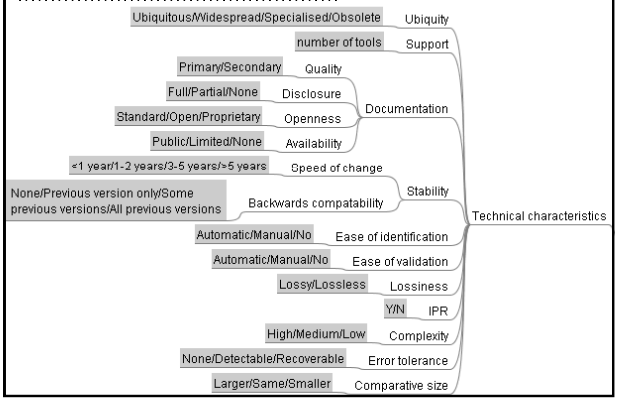
1. Bytes 0-1:
   The byte order used within the file.
   Legal values are: “II” (4949.H) / “MM” (4D4D.H)
2. Bytes 2-3:
   An arbitrary but carefully chosen number (**42**)
   that further identifies the file as a TIFF file.

## Demo: DROID, PRONOM

## Registry content

- Descriptive information
- Identifiers
  - MIME
  - Pronom Unique Identifier (PUID)
- Relationships to formats
- Technical environment
- References and links…
- Risk factors

## File format characteristics



Technical characteristics

Ubiquity: Ubiquitous/Widespread/Specialised/Obsolete

Support: number of tools

Documentation
- Quality: Primary/Secondary
- Disclosure: Full/Partial/None
- Openness: Standard/Open/Proprietary
- Availability: Public/Limited/None

Stability
- Speed of change: <1 year/1-2 years/3-5 years/>5 years
- Backwards compatability: None/Previous version only/Some previous versions/All previous versions

Ease of identification: Automatic/Manual/No

Ease of validation: Automatic/Manual/No

Lossiness: Lossy/Lossless

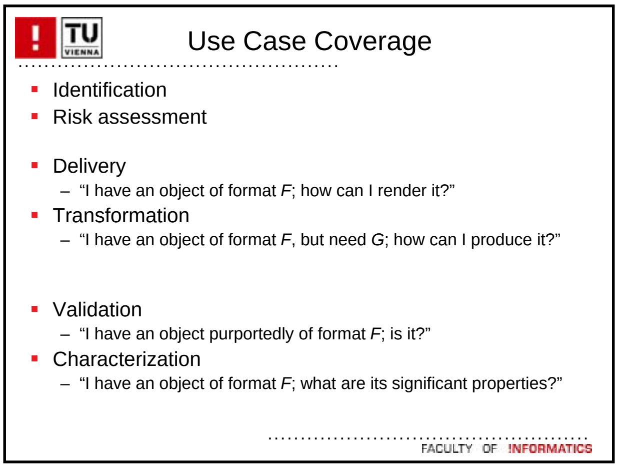
IPR: Y/N

Complexity: High/Medium/Low

Error tolerance: None/Detectable/Recoverable

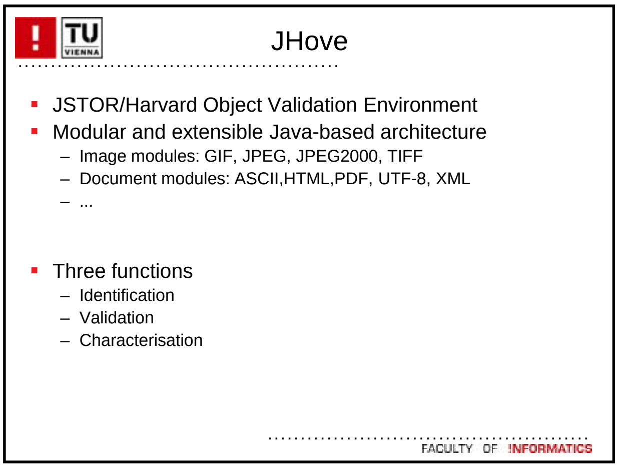
Comparative size: Larger/Same/Smaller

## Questions?



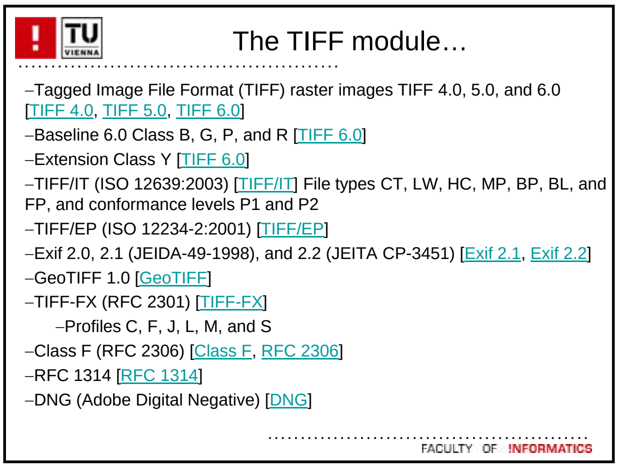
## Use Case Coverage

- Identification
- Risk assessment
- Delivery
  - “I have an object of format *F*; how can I render it?”
- Transformation
  - “I have an object of format *F*, but need *G*; how can I produce it?”
- Validation
  - “I have an object purportedly of format *F*; is it?”
- Characterization
  - “I have an object of format *F*; what are its significant properties?”



## JHove

- JSTOR/Harvard Object Validation Environment
- Modular and extensible Java-based architecture
  - Image modules: GIF, JPEG, JPEG2000, TIFF
  - Document modules: ASCII,HTML,PDF, UTF-8, XML
  - ...
- Three functions
  - Identification
  - Validation
  - Characterisation



## The TIFF module…

–Tagged Image File Format (TIFF) raster images TIFF 4.0, 5.0, and 6.0 [TIFF 4.0, TIFF 5.0, TIFF 6.0]

–Baseline 6.0 Class B, G, P, and R [TIFF 6.0]

–Extension Class Y [TIFF 6.0]

–TIFF/IT (ISO 12639:2003) [TIFF/IT] File types CT, LW, HC, MP, BP, BL, and FP, and conformance levels P1 and P2

–TIFF/EP (ISO 12234-2:2001) [TIFF/EP]

–Exif 2.0, 2.1 (JEIDA-49-1998), and 2.2 (JEITA CP-3451) [Exif 2.1, Exif 2.2]

–GeoTIFF 1.0 [GeoTIFF]

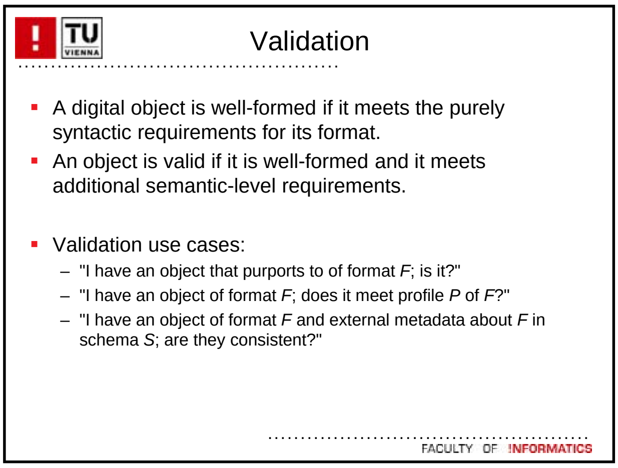
–TIFF-FX (RFC 2301) [TIFF-FX]

–Profiles C, F, J, L, M, and S

–Class F (RFC 2306) [Class F, RFC 2306]

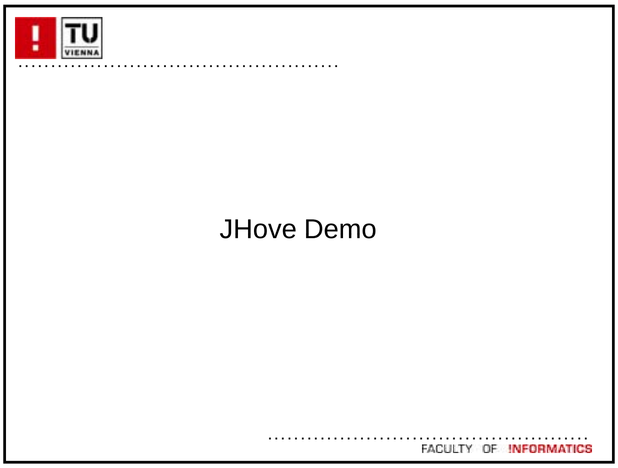
–RFC 1314 [RFC 1314]

–DNG (Adobe Digital Negative) [DNG]

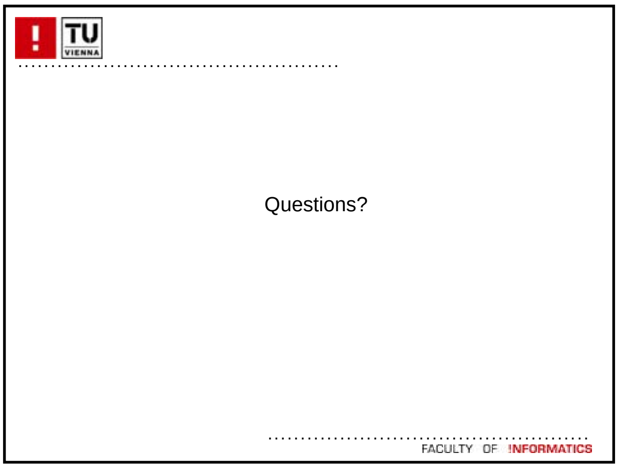


## Validation

- A digital object is well-formed if it meets the purely syntactic requirements for its format.
- An object is valid if it is well-formed and it meets additional semantic-level requirements.
- Validation use cases:
  - "I have an object that purports to of format *F*; is it?"
  - "I have an object of format *F*; does it meet profile *P* of *F*?"
  - "I have an object of format *F* and external metadata about *F* in schema *S*; are they consistent?"



JHove Demo



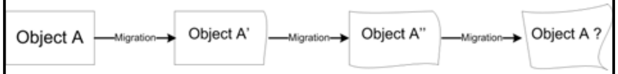
Questions?

## Use Case Coverage

- Identification
- Risk assessment
- Delivery
- Transformation
- Validation
- Characterization
  - “I have an object of format *F*; what are its significant properties?”

## Core requirement: Keep object intact

- ❑ Essential object characteristics
  - ❑ Content
  - ❑ Appearance
  - ❑ Structure
  - ❑ Behaviour
  - ❑ Context

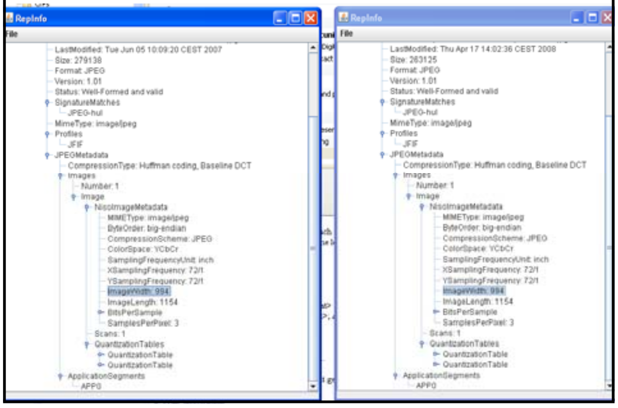


## Validating a migrated image

- ❑ Yes, it’s in JPEG 2000 format
- ❑ Yes, it’s wellformed
- ❑ Yes, it’s valid
- ❑ Yes, it still has the same dimensions
- ❑ …. But is it still the same image?

Object A
Migration
Object A'
Migration
Object A''
Migration
Object A ?

## Validating a migrated image

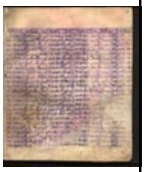


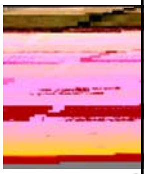
```
LastModified: Tue Jun 05 10:09:20 CEST 2007
Size: 279138
Format: JPEG
Version: 1.01
Status: Well-Formed and valid
SignatureMatches
  JPEG-hul
MimeType: image/jpeg
Profiles
  JFIF
JPEGMetadata
  CompressionType: Huffman coding, Baseline DCT
  Images
    Number: 1
    Image
      NisoImageMetadata
        MIMEType: image/jpeg
        ByteOrder: big-endian
        CompressionScheme: JPEG
        ColorSpace: YCbCr
        SamplingFrequencyUnit: inch
        XSamplingFrequency: 72/1
        YSamplingFrequency: 72/1
        ImageWidth: 994
        ImageLength: 1154
        BitsPerSample
        SamplesPerPixel: 3
      Scans: 1
      QuantizationTables
        QuantizationTable
        QuantizationTable
  ApplicationSegments
    APP0
```

```
LastModified: Thu Apr 17 14:02:36 CEST 2008
Size: 283125
Format: JPEG
Version: 1.01
Status: Well-Formed and valid
SignatureMatches
  JPEG-hul
MimeType: image/jpeg
Profiles
  JFIF
JPEGMetadata
  CompressionType: Huffman coding, Baseline DCT
  Images
    Number: 1
    Image
      NisoImageMetadata
        MIMEType: image/jpeg
        ByteOrder: big-endian
        CompressionScheme: JPEG
        ColorSpace: YCbCr
        SamplingFrequencyUnit: inch
        XSamplingFrequency: 72/1
        YSamplingFrequency: 72/1
        ImageWidth: 994
        ImageLength: 1154
        BitsPerSample
        SamplesPerPixel: 3
      Scans: 1
      QuantizationTables
        QuantizationTable
        QuantizationTable
  ApplicationSegments
    APP0
```

## Validating a migrated image

- ❑ Yes, it’s in JPEG 2000 format
- ❑ Yes, it’s wellformed
- ❑ Yes, it’s valid
- ❑ Yes, it still has the same dimensions
- ❑ …. But is it still the same image?





- ❑ We need more characterisation.

Object A
Migration
Object A'
Migration
Object A''
Migration
Object A ?

## The XCL languages

- ❑ The eXtensible characterisation description language XCDL
  - ❑ describes properties of digital objects
- ❑ The eXtensible characterisation extraction language XCEL
  - ❑ extracts properties from files
  - ❑ Creates a mapping from a file format to XCDL

## Essential properties as described by file formats

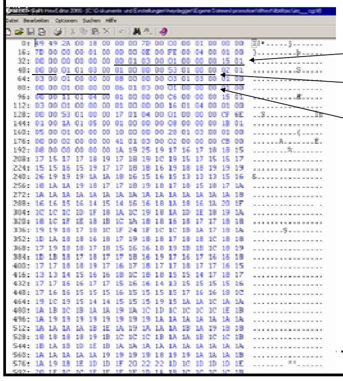


Image width: 277

Image length: 339

Compression: uncompressed

*ImageLength*

The number of rows of pixels in the image.

Tag = 257 (101.H)

Type = SHORT or LONG

N = 1

No default. See also ImageWidth.

*ImageWidth*

The number of columns in the image, i.e., the number of pixels per row.

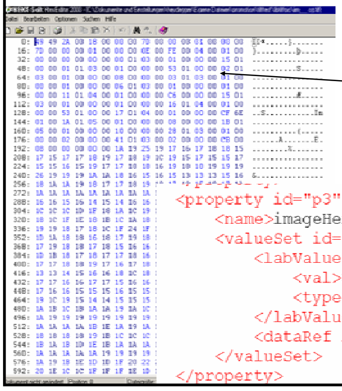
Tag = 256 (100.H)

Type = SHORT or LONG

N = 1

No default. See also ImageLength.

## Properties described by the format



*ImageLength*

The number of rows of pixels in the image.

Tag = 257 (101.H)

Type = SHORT or LONG

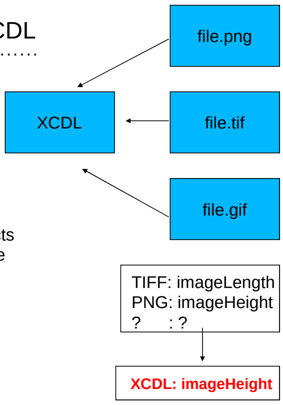
N = 1

No default. See also ImageWidth.

```
<property id="p3" source="raw" cat="descr" >
    <name>imageHeight</name>
    <valueSet id="i_i1_s10" >
        <labValue>
            <val>339</val>
            <type>uint16</type>
        </labValue>
        <dataRef ind="normAll" />
    </valueSet>
</property>
```

## XCDL

- Uniform description of properties and values
- Uniform structure
  - Properties of different objects are described using a single vocabulary and grammar
- eXtensible



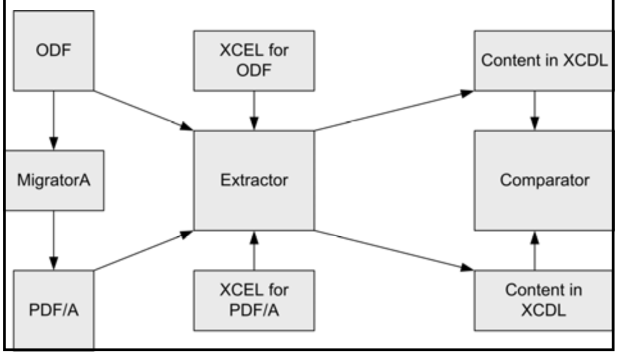
## Extracting properties: XCEL

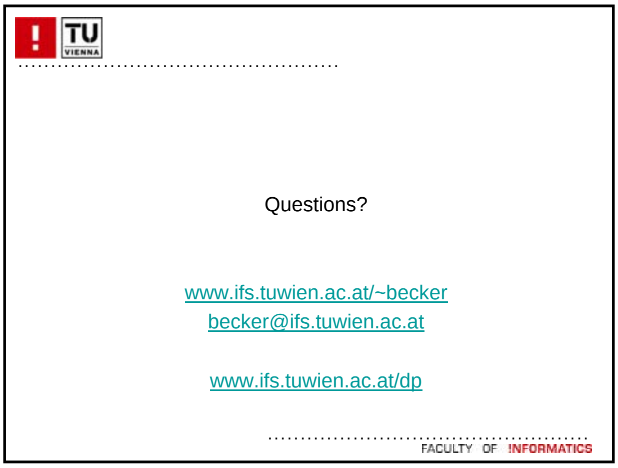
- One generic XCEL processor instead of specific extractor for every file format
- Preprocessing instructions
  - Configuration tasks
- Format description
  - Defines the structure of an object
- Templates
  - Describe recurring structures
- Postprocessing instructions
  - On the results of processing

## XCDL example: ‘An **important** word’

```
<normData id="n6">An important word</normData>

<property id="p8" source="raw">
<name>Fontname</name>
<valueSet id="v2">
  <labVal>
    <val>Times-Bold</val>
    <type>XCLLabel</type>
  </labVal>
  <dataRef ind="normSpecific">
    <ref id="n6" start="3" end="11"/>
  </dataRef>
 </valueSet>
  ...............
```

## Comparing migrated documents





Questions?

www.ifs.tuwien.ac.at/~becker
becker@ifs.tuwien.ac.at

www.ifs.tuwien.ac.at/dp

FACULTY OF INFORMATICS