File Formats and Significant Properties

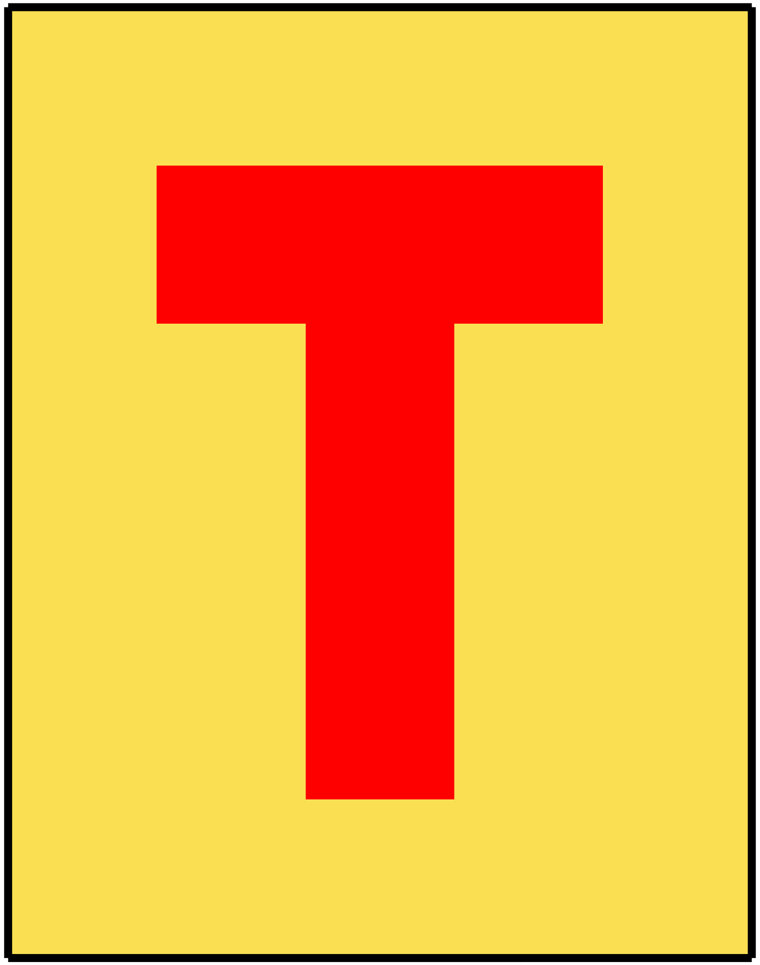
Manfred Thaller

Universität zu* Köln

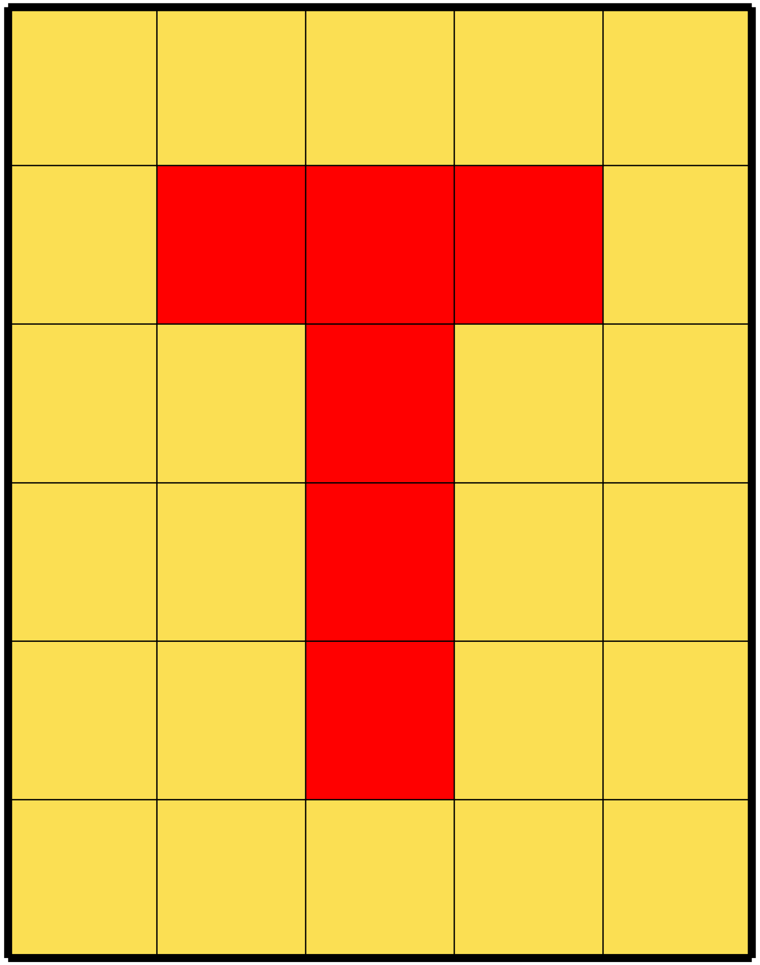
March 24th, 2009

***University _**at**_ not _**of**_ Cologne**

I – What is (in) a format?



6 rows
5 columns



🟨	🟨	🟨	🟨	🟨
🟨	🟥	🟥	🟥	🟨
🟨	🟨	🟥	🟨	🟨
🟨	🟨	🟥	🟨	🟨
🟨	🟨	🟥	🟨	🟨
🟨	🟨	🟨	🟨	🟨

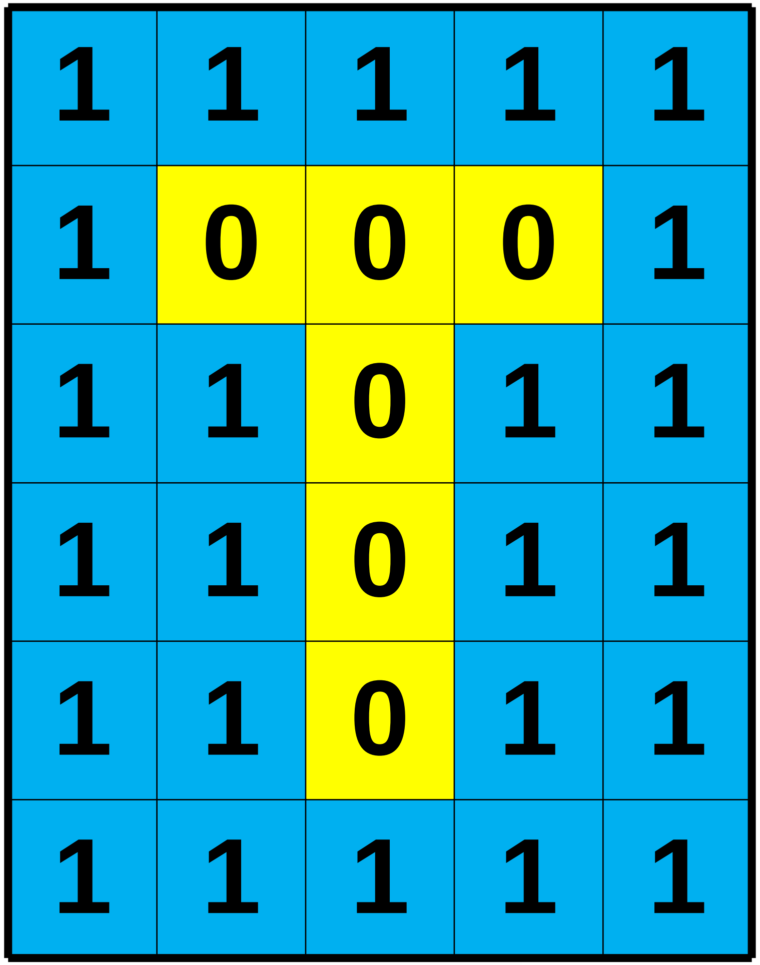
5 rows
6 columns

🟨	🟨	🟨	🟨	🟨	🟨
🟥	🟥	🟥	🟨	🟨	🟨
🟥	🟨	🟨	🟨	🟨	🟥
🟨	🟨	🟨	🟨	🟥	🟨
🟨	🟨	🟨	🟨	🟨	🟨

1 == ochre
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

1 == blue
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

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

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

Uncompressed

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

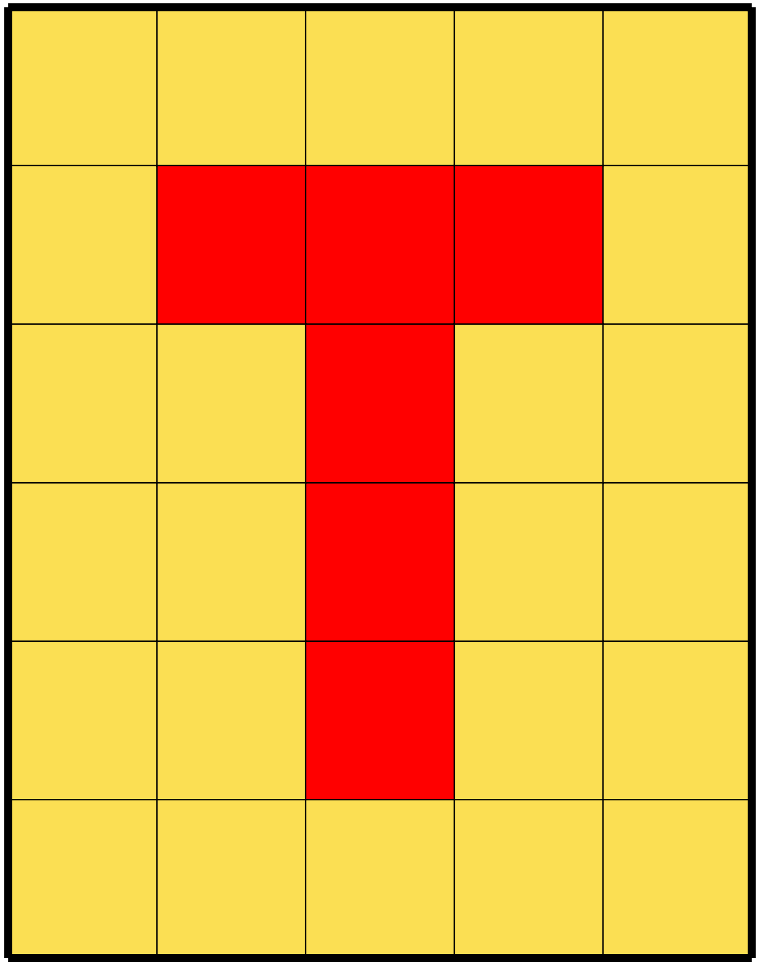
(Compressed)
Run Length
Encoded

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

6 rows
5 columns

1 == ochre
0 == red

Uncompressed

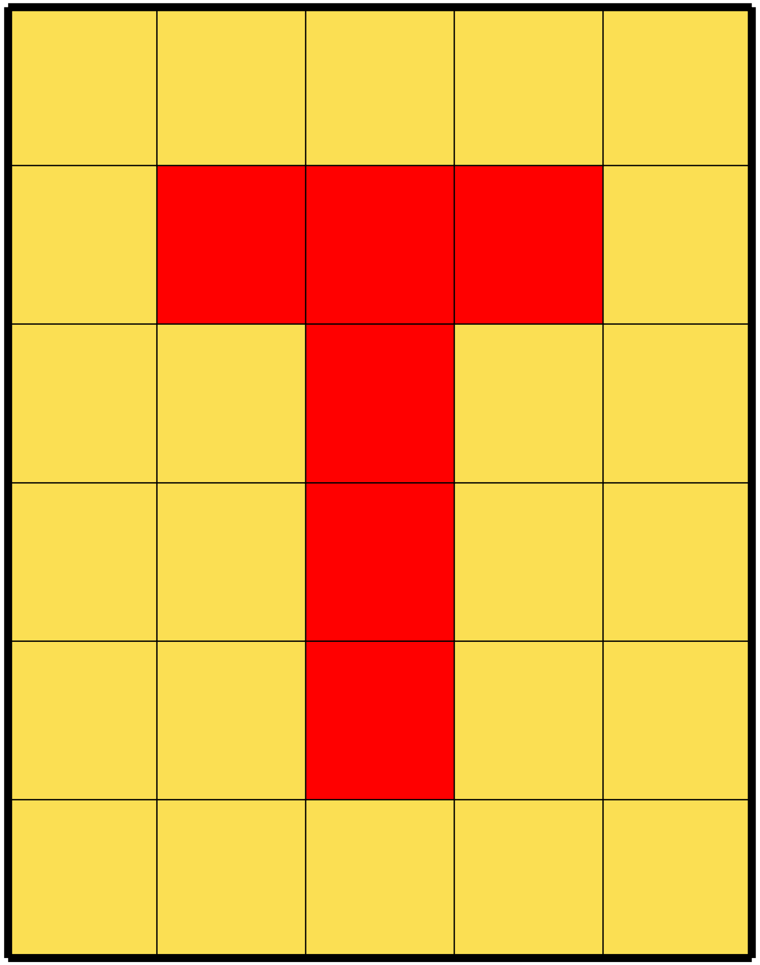


🟨	🟨	🟨	🟨	🟨
🟨	🟥	🟥	🟥	🟨
🟨	🟨	🟥	🟨	🟨
🟨	🟨	🟥	🟨	🟨
🟨	🟨	🟥	🟨	🟨
🟨	🟨	🟨	🟨	🟨

dimensions

1 == ochre
0 == red

Uncompressed

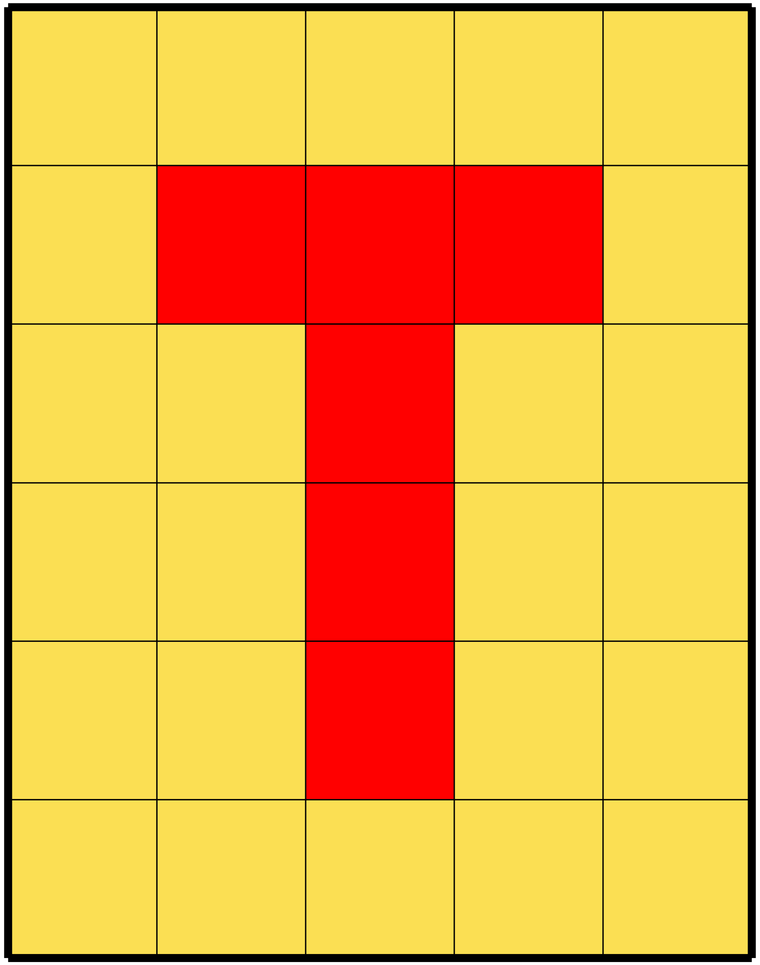


🟨	🟨	🟨	🟨	🟨
🟨	🟥	🟥	🟥	🟨
🟨	🟨	🟥	🟨	🟨
🟨	🟨	🟥	🟨	🟨
🟨	🟨	🟥	🟨	🟨
🟨	🟨	🟨	🟨	🟨

dimensions

photogrammetric interpretation

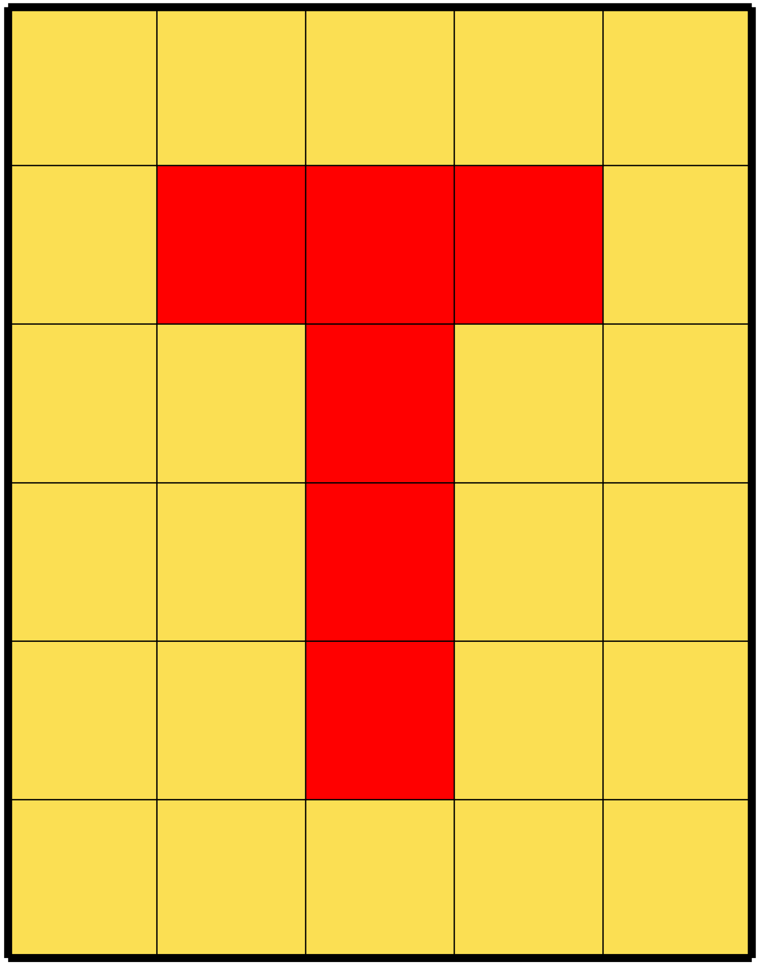
Uncompressed



dimensions

photogrammetric interpretation

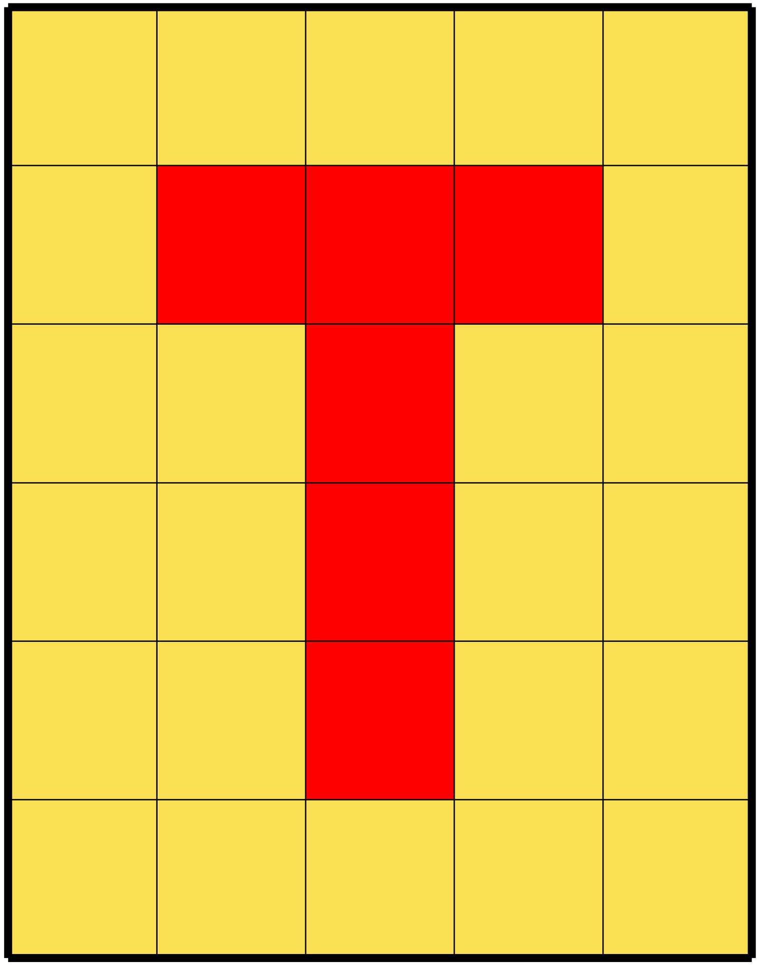
compression



<basic information>

<rendering information>

<storage information>



File format

<basic information>

What to do?

<rendering information>

How to do it?

<storage information>

How to move it from persistent to deployed form?

<data>

What to deploy?

File format

<basic information>

Mandatory

<rendering information>

Useful

<storage information>

Historical

<data>

Mandatory

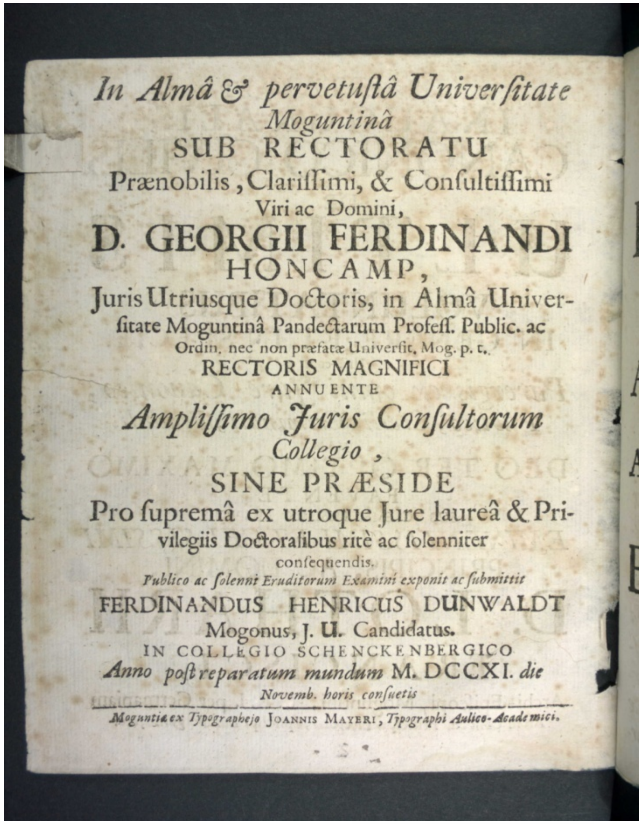
File format

A deterministic specification how the properties of a digital object can reversibly be converted into a linear bytestream (bitstream).

II – Why would we want to know?

Bit rot

An Image file before ….

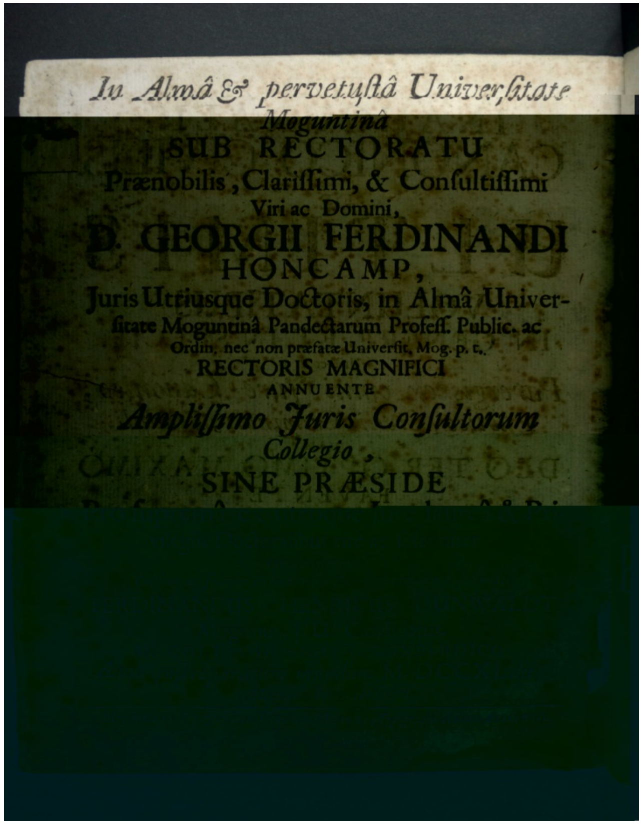


In Almâ & pervetuſtâ Univerſitate
Moguntinâ
SUB RECTORATU
Prænobilis, Clariſſimi, & Conſultiſſimi
Viri ac Domini,
D. GEORGII FERDINANDI
HONCAMP,
Juris Utriusque Doctoris, in Almâ Univerſitate Moguntinâ Pandectarum Profeſſ. Public. ac
Ordin. nec non præfatæ Univerſit. Mog. p. t.
RECTORIS MAGNIFICI
ANNUENTE
Ampliſſimo Juris Conſultorum
Collegio,
SINE PRÆSIDE
Pro ſupremâ ex utroque Jure laureâ & Privilegiis 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 Typographejo JOANNIS MAYERI, *Typographi Aulico-Academici.*

Bit rot

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



In Almâ & perpetuâ 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

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.

But 1 …

Why should I care?
Can I not just pay a technician to keep some system of checksums?

Counter-but 1 …

Do you rather buy a brand of car with a reputation of an excellent network of maintenance shops, or one with a reputation for needing little maintenance?

But 2 …

But is bit rot really *that* important?
I have read, that files most of the time get either unreadable completely, or stay completely undamaged?

Counter-but 2a …

In disaster recovery: yes!
With files on degrading storage systems / devices: no!

But 2 …

But is bit rot really *that* important?

Counter-but 2b …

Bit rot is, indeed, just *one* problem.!

We do this is just to show, that there are differences between the technical fitness for preservation between formats. Others go beyond 15 minutes.

But 3 …

But is there not a simple list in this type of problems, which I can consult easily?

Counter-but 3 …

No.

III – Which format to choose?

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 (“virtual reality”)

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

Thank you!

Exercise

<format 1>

Size 1 - Count 1 - Shoot:
Some <xxx> are not correctly displayed
Some <xxx> are not recognized

Size 1 - Count 1 - Corrupt:
Unable to open the file
File size has changed

Size 512 - Count 1 - Shoot:
<xxx> are not displayed

<format 2>

…

Our findings support / do not support the Florida recommendations for this type of content, because …