

Characterisation

Digital Preservation Planning: Principles,
Examples and the Future with Planets.

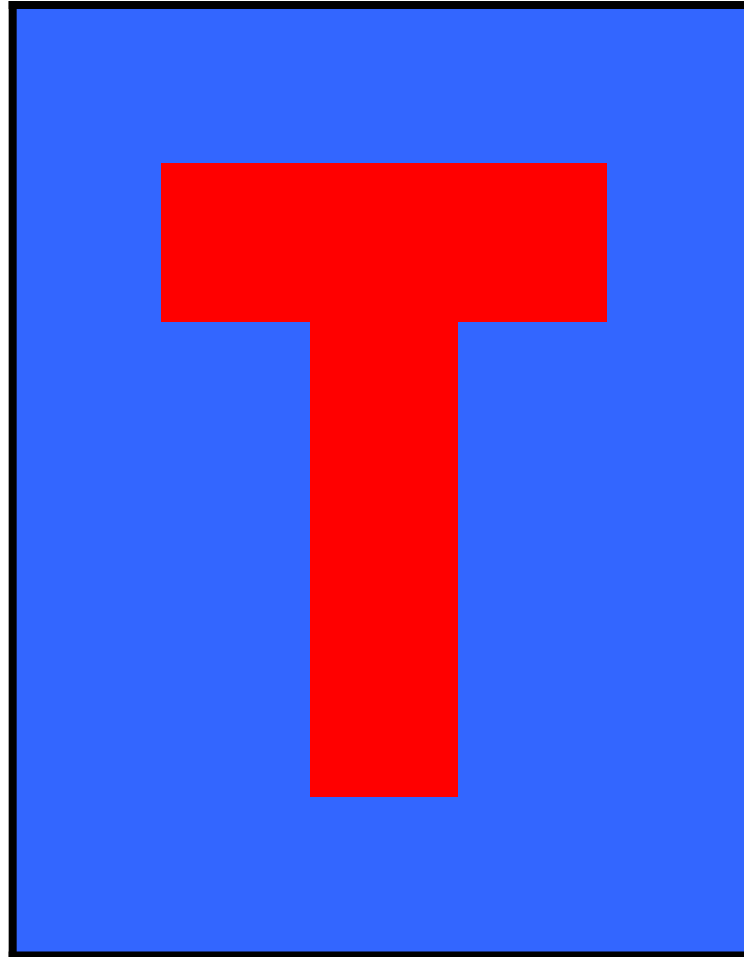
July 29th, 2008

Manfred Thaller
Universität zu* Köln
manfred.thaller@uni-koeln.de

* University at, NOT of Cologne

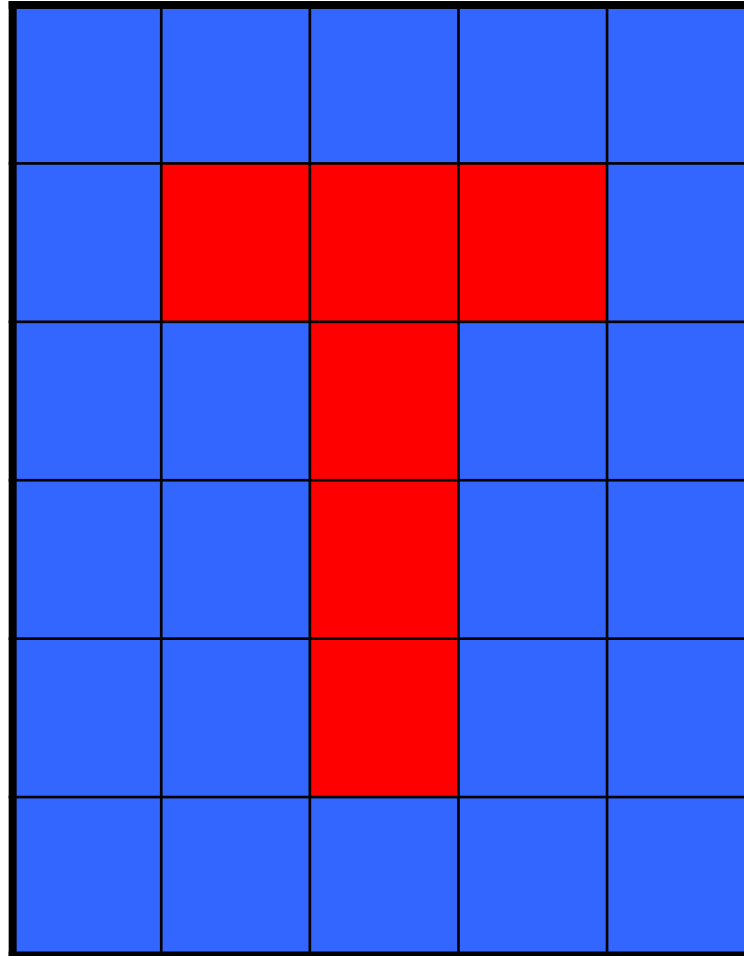
I - What is (in) a format?

An image



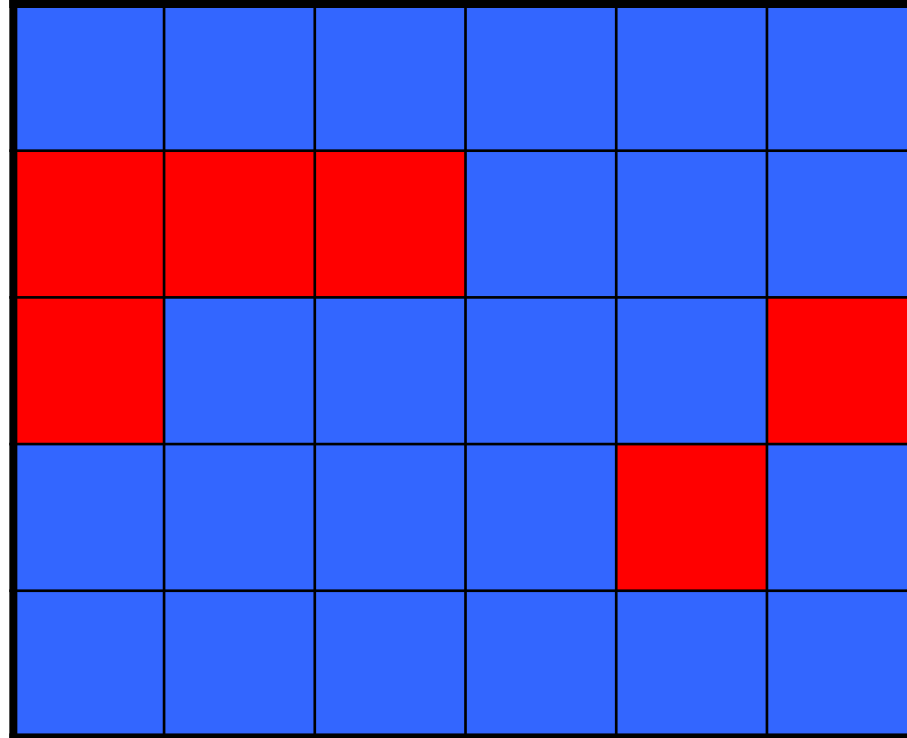
An image

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

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

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

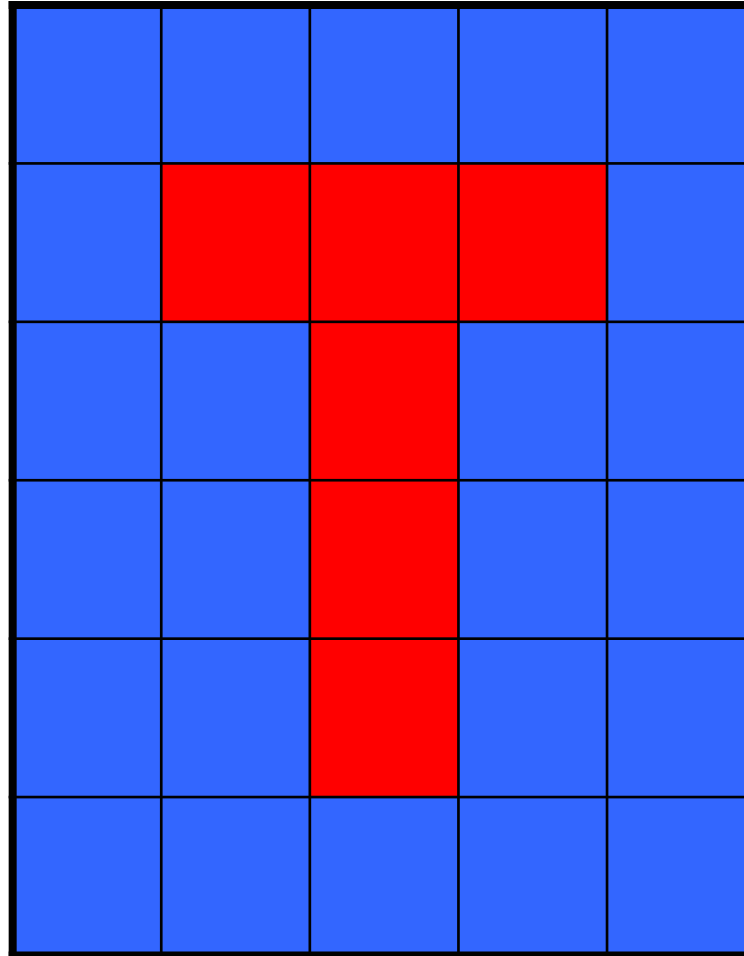
1	1	1	1	1
1	0	0	0	1
1	1	0	1	1
1	1	0	1	1
1	1	0	1	1
1	1	1	1	1

An image

dimensions

*photogrammetric
interpretation*

compression



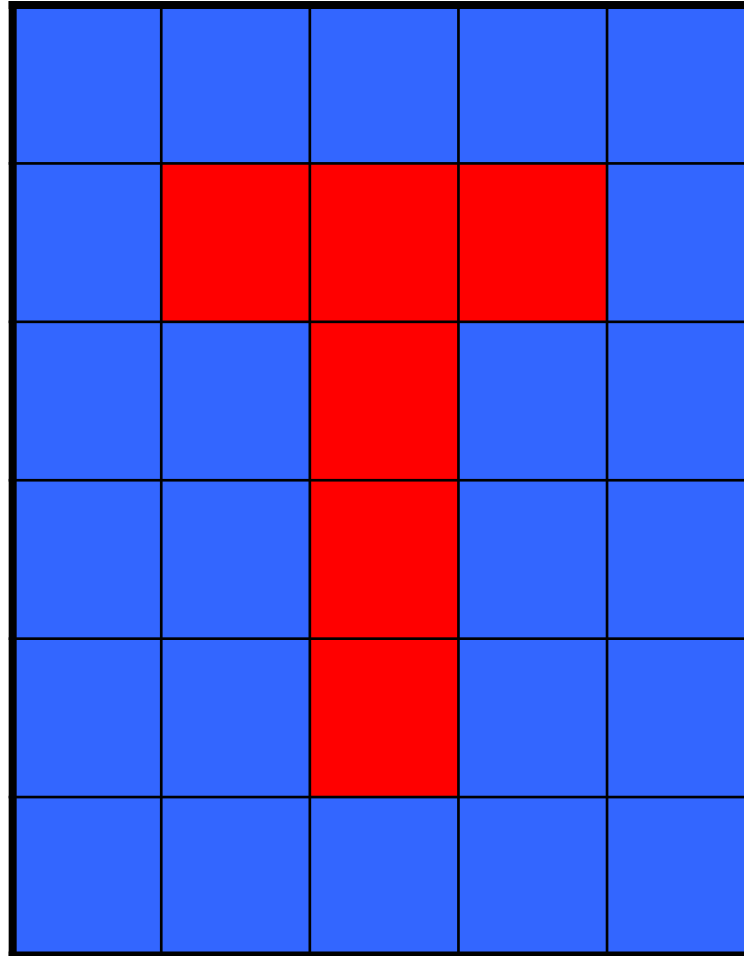
An image

<basic
information>

<rendering
information>

<storage
information>

<data>



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>

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?

III - Which format to choose?

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>

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>

Recommended formats: vector araphics

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
<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>

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>

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>

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>

Doctoral thesis on robustness of file formats:

Volker Heydegger, University at Cologne.

herrmanv@uni-koeln.de

IV - How to we identify a format?

What kind of file is this?

Two ways to identify a file:

(a) By extension.

„Each file ending with *.doc is a MS Word document“

What kind of file is this?

Two ways to identify a file:

(b) By internal characteristics („magic number“, „signature“).

A TIFF file begins with ...

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

“II” (4949.H) / “MM” (4D4D.H)

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

File format registries - URLs

PRONOM:

<http://www.nationalarchives.gov.uk/pronom/>
(does not only rely on extensions)

Global Digital Format Registry:

<http://hul.harvard.edu/gdfr>
(predominantly project description)

FileExt:

<http://filext.com>
(predominantly links to software)

V - What's a file characteristic, than?

Technical metadata →

A high proportion of the preservation metadata will be in narrative format and will require manual entry by Library staff. A significant subset of the data however, relating to technical file characteristics, can be automatically extracted from the digital object by reading the file header details. This successful extraction of preservation metadata has been proved in a previous National Library proof of concept project. The automated capture of this information will significantly reduce the amount of manual data entry required from Library staff.

→ file characteristics.

Why automate?

1 million objects: use one second for each.

== 16666.7 minutes == 277.8 hours

== 11.57 working days of a computer

== 34.7 8-hour days for a Human

== 7 working weeks

Why automate?

1 million objects: use five minutes for each.

== 416 666.7 hours

== 52 803.4 8-hour days for a Human

== way too much for anything

Formats in PLANETS:

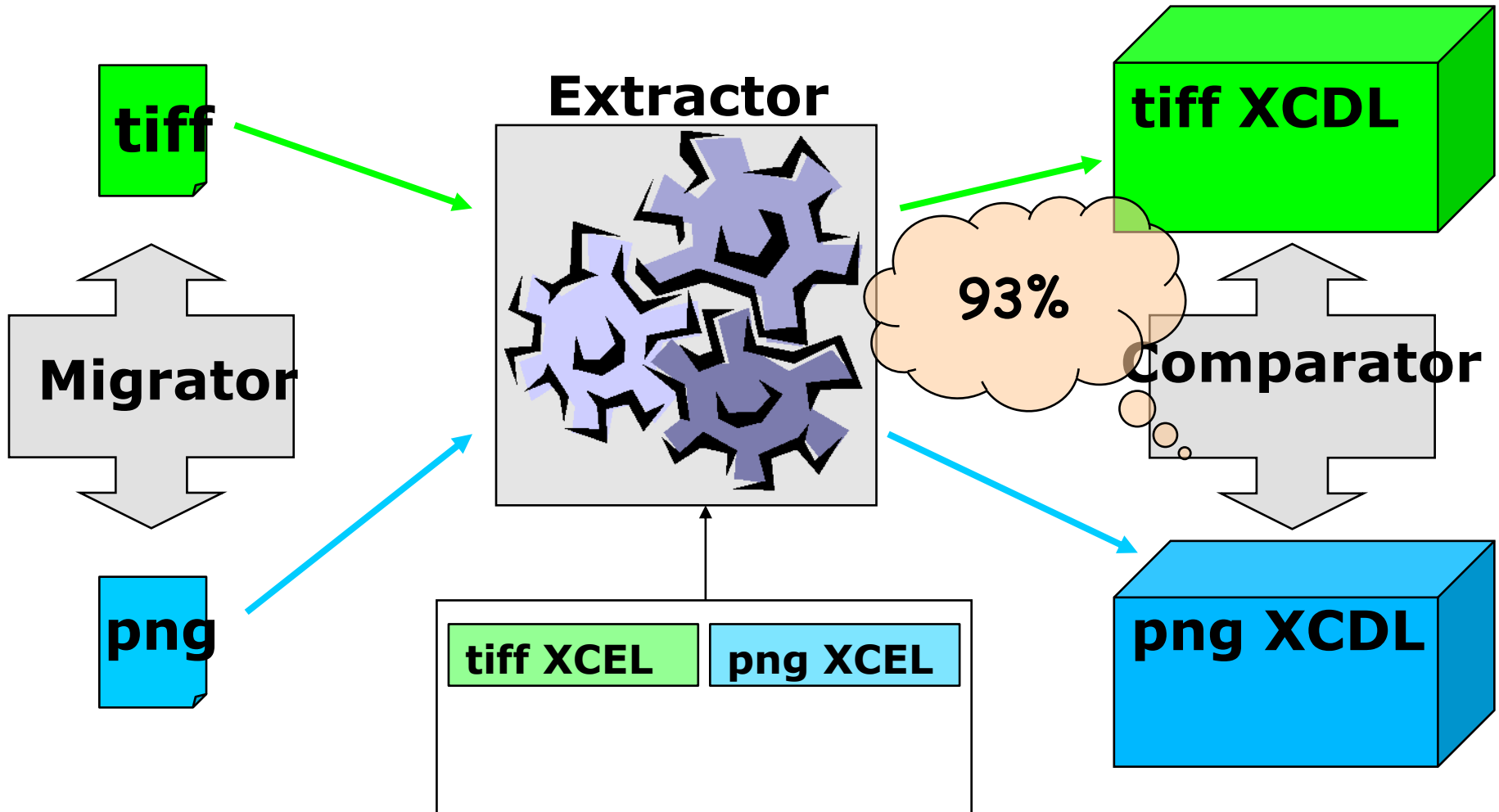
File characteristics

Based on two formal languages:

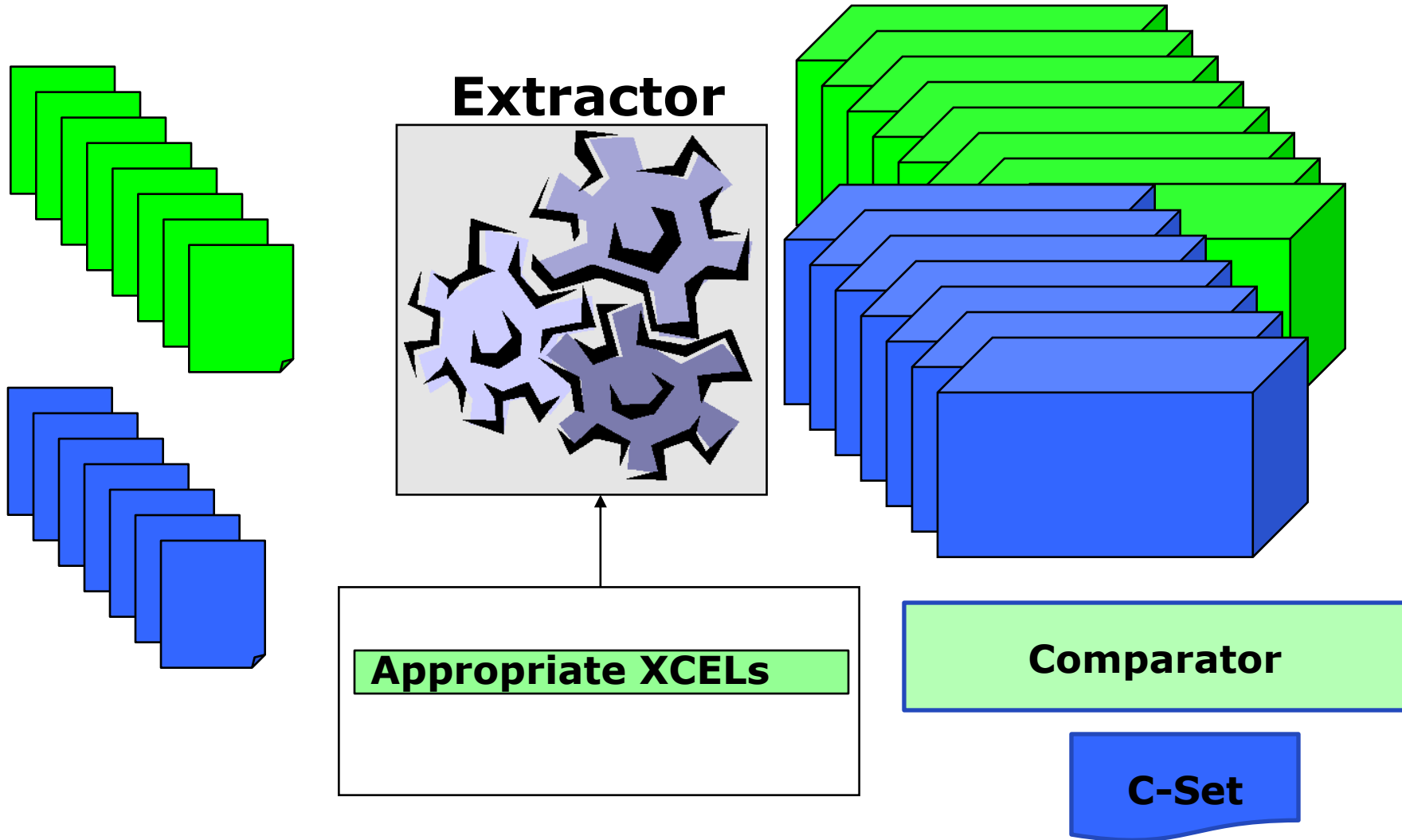
(1)eXtensible Characterisation
Extraction Language (= XCEL)

(2)eXtensible Characterisation
Description Language (= XCDL)

The comparator



The comparator



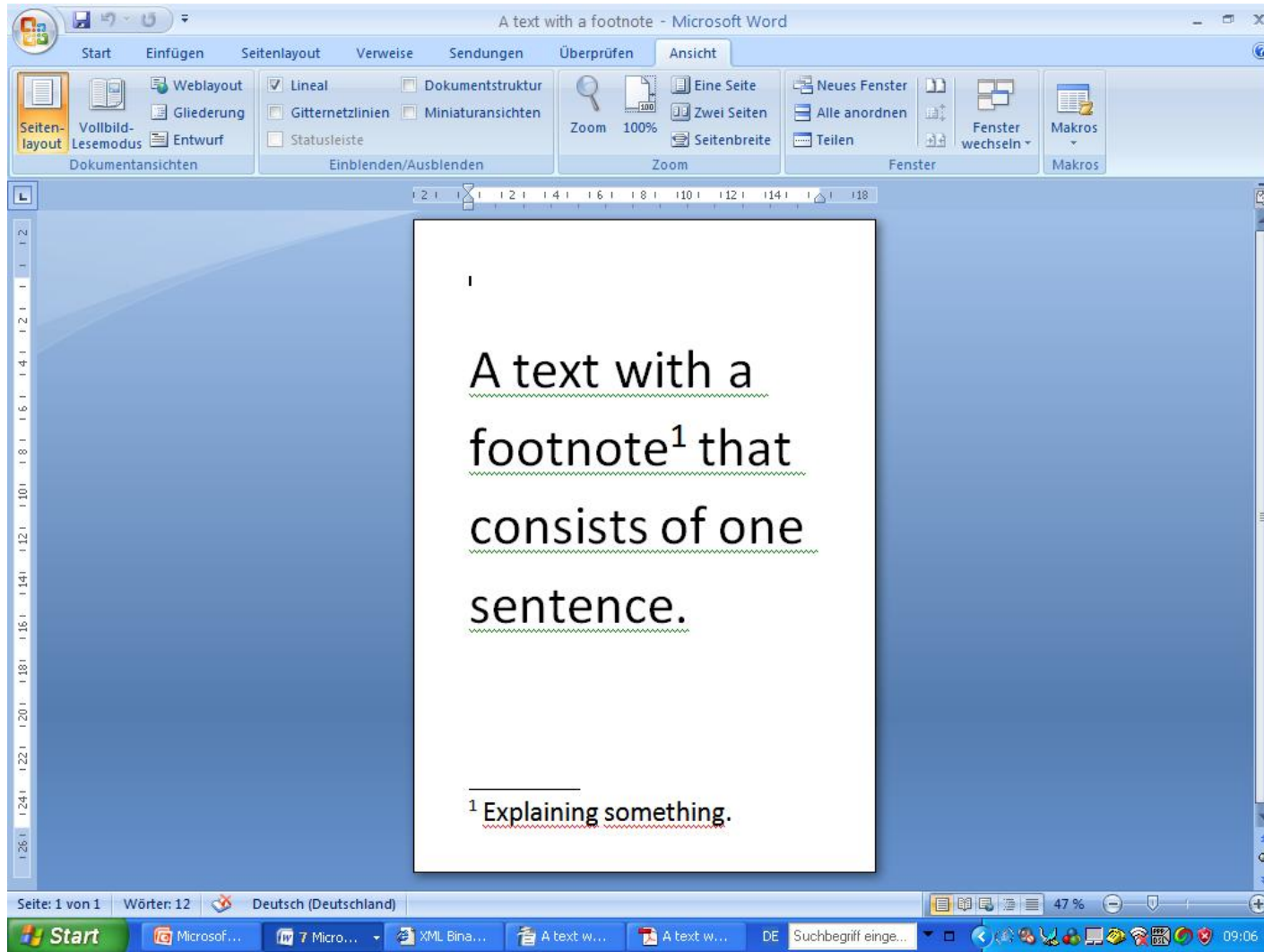
Why data?



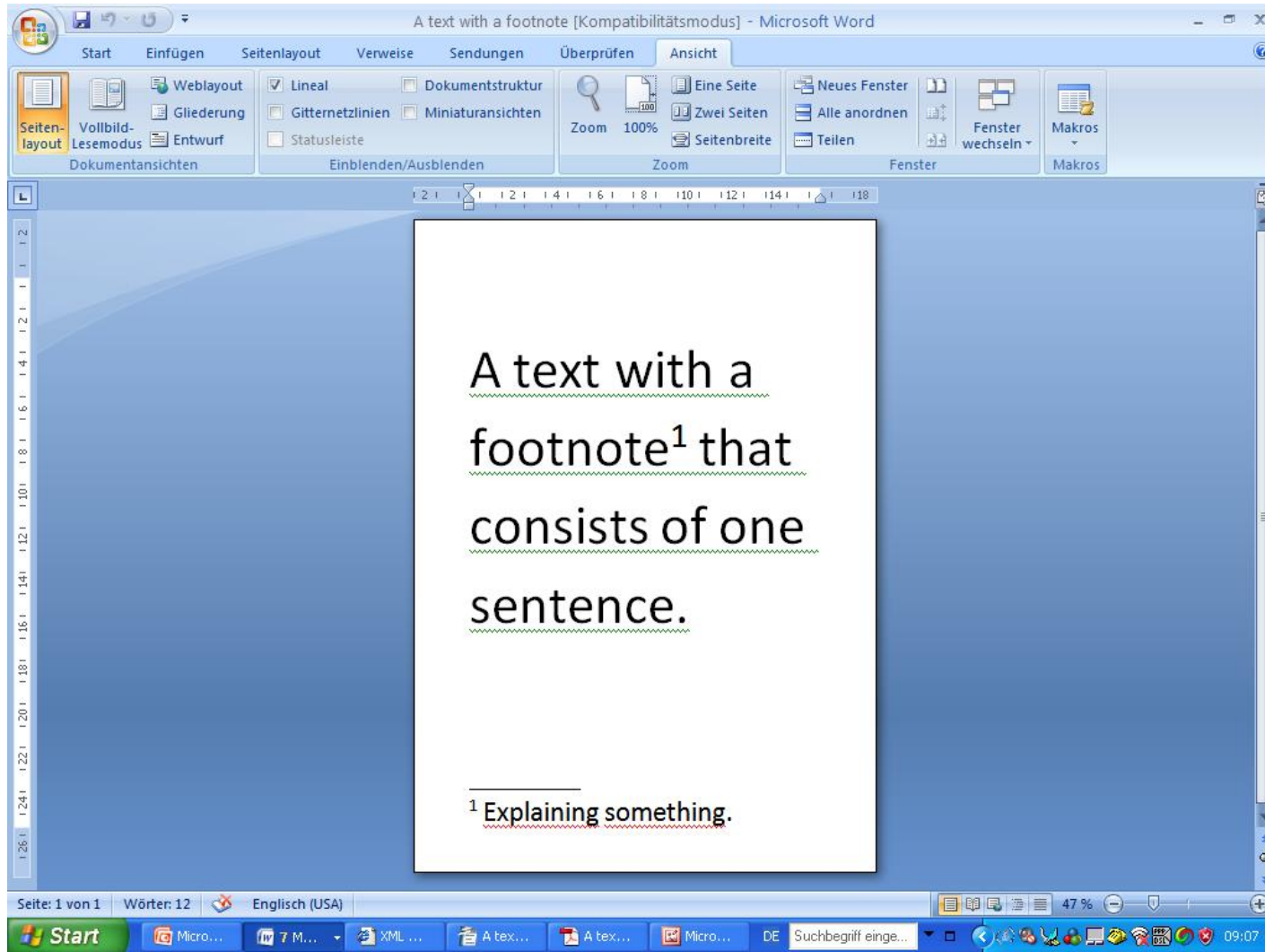
Becomes discoverable only from the actual data ...

V - What is not in a file format?

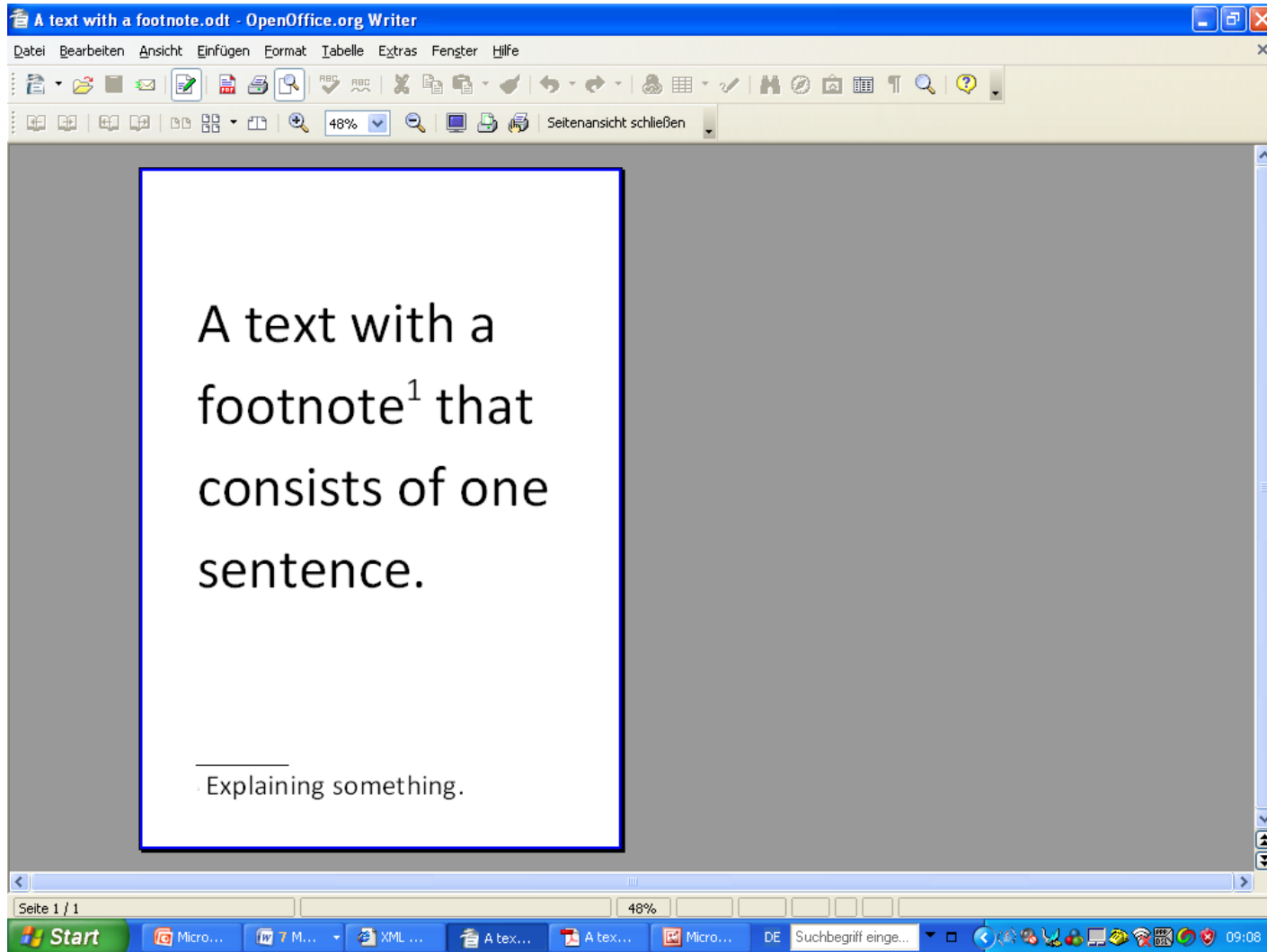
Testfile in Word 2007



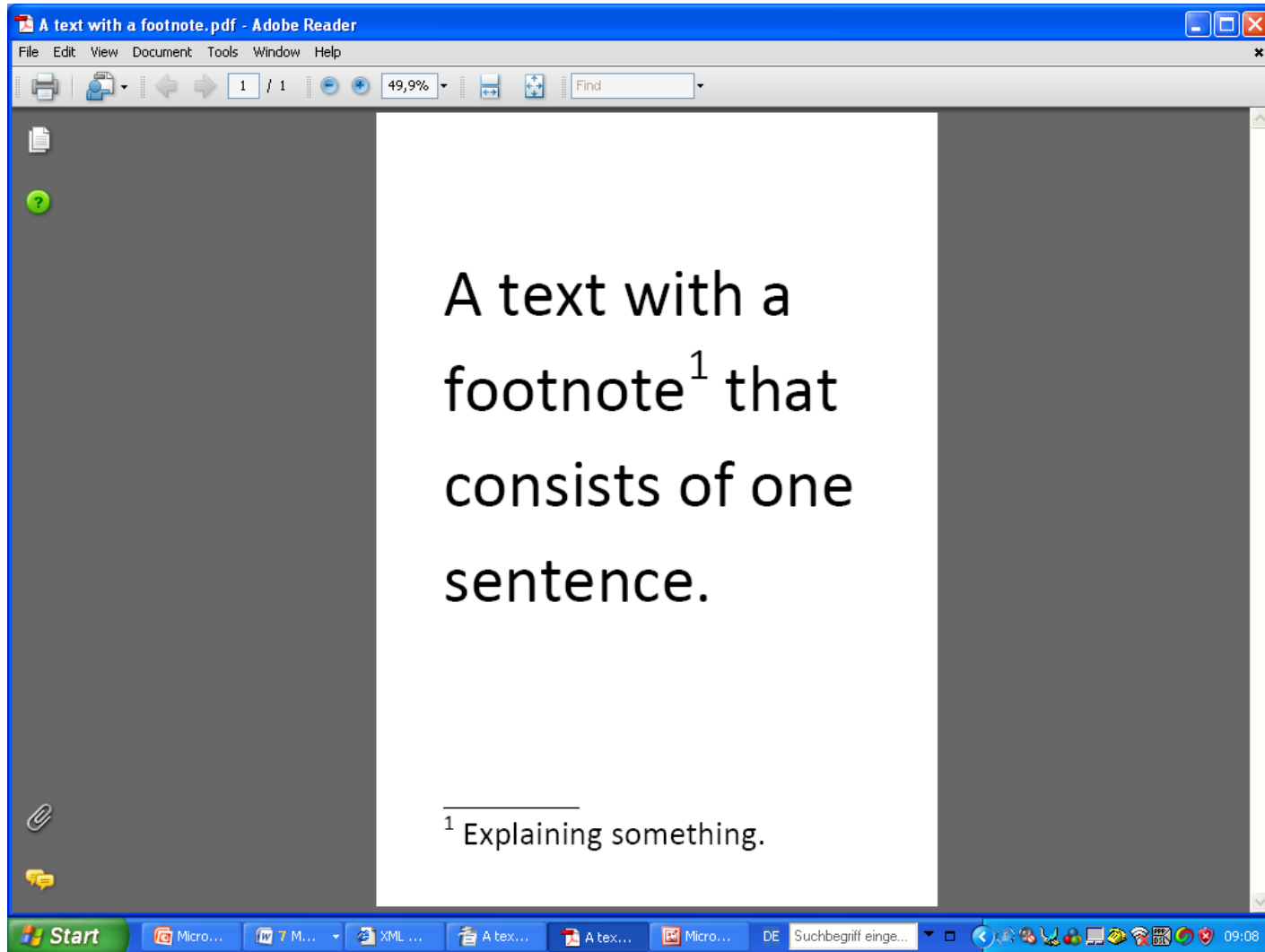
Testfile in Word 2003 (2007)



Testfile in Open Office ODT



Testfile in PDF



Measuring the pages

Cut out page from rendering surface.

Scale to common dimensions: $371 \pm 1 \times 521 \pm 1$

Measure

1. The leftmost and lowest completely black pixel in the letter "A" starting the first line of the main text.
2. The leftmost and highest completely black pixel in the letter "E" starting the first line of the text in the footnote.
3. The geometrical centre of the period at the end of the main sentence.
4. The geometrical centre of the period at the end of the footnote text.

Measuring Word 2003

A text with a
footnote¹ that
consists of one
sentence.

¹ Explaining something.

$$(i) = 45 / 134;$$

$$(ii) = 57 / 470;$$

$$(iii) = 215 / 322 ;$$

$$(iv) = 254 / 483$$

Measuring Word 2007

A text with a
footnote¹ that
consists of one
sentence.

¹ Explaining something.

(i) = 45 / 134;

(ii) = 57 / 470;

(iii) = 215 / 322 ;

(iv) = 254 / 483

A text with a
footnote¹ that
consists of one
sentence.

¹ Explaining something.

$$(i) = 44 / 132;$$

$$(ii) = \underline{52} / 469;$$

$$(iii) = 214 / 320 ;$$

$$(iv) = \underline{247} / 482$$

A text with a
footnote¹ that
consists of one
sentence.

¹ Explaining something.

$$(i) = 45 / 130;$$

$$(ii) = 59 / 467;$$

$$(iii) = 215 / 317 ;$$

$$(iv) = 254 / 480$$

Summary I

The comparison of the four renderings of the example pages described above seem to indicate clearly, that a migration from the Word family of formats to PDF is a *better* way to preserve the content of the document, than a migration to the Open Office format.

A text with a
footnote¹ that
consists of one
sentence.

¹ Explaining something.

Relationship tagged
explicitly.

Text / footnote separation
clear.

Rendering / layout not
(totally) predicatble.

Footnote indicator
unpredictable.

A text with a
footnote¹ that
consists of one
sentence.

¹ Explaining something.

Relationship tagged
explicitly.

Text / footnote
separation extremely
clear.

Rendering / layout
pretty predictable.

Footnote indicator not
predictable.

A text with a
footnote¹ that
consists of one
sentence.

¹ Explaining something.

Relationship tagged
explicitly.

Text / footnote
separation extremely
clear.

Rendering / layout a
little bit predictable.

Footnote indicator
predictable.

A text with a
footnote¹ that
consists of one
sentence.

¹ Explaining something.

Relationship expressed
by layout.

Text / footnote
separation missing.

Rendering / layout very
much predictable.

Footnote indicator
predictable.



Summary II



The comparison of the four internal structures of the example pages described above seem to indicate clearly, that a migration from the Word family of formats to PDF is a *worse* way to preserve the content of the document, than a migration to the Open Office format.

Do not forget, that the whole movement started by SGML, carried into the WWW by HTML, transferred to content by the TEI and started XML as a basic empowering technology ...

... assumes that rendering is NOT particularly relevant.

|


 A text with a
footnote¹ that
consists of one
sentence



Explaining something


```

<significantPoints>
  <point x="45" y="134" />
  <point x="57" y="470" />
  <point x="215" y="322" />
  <point x="254" y="483" />
</significantPoints>
    
```