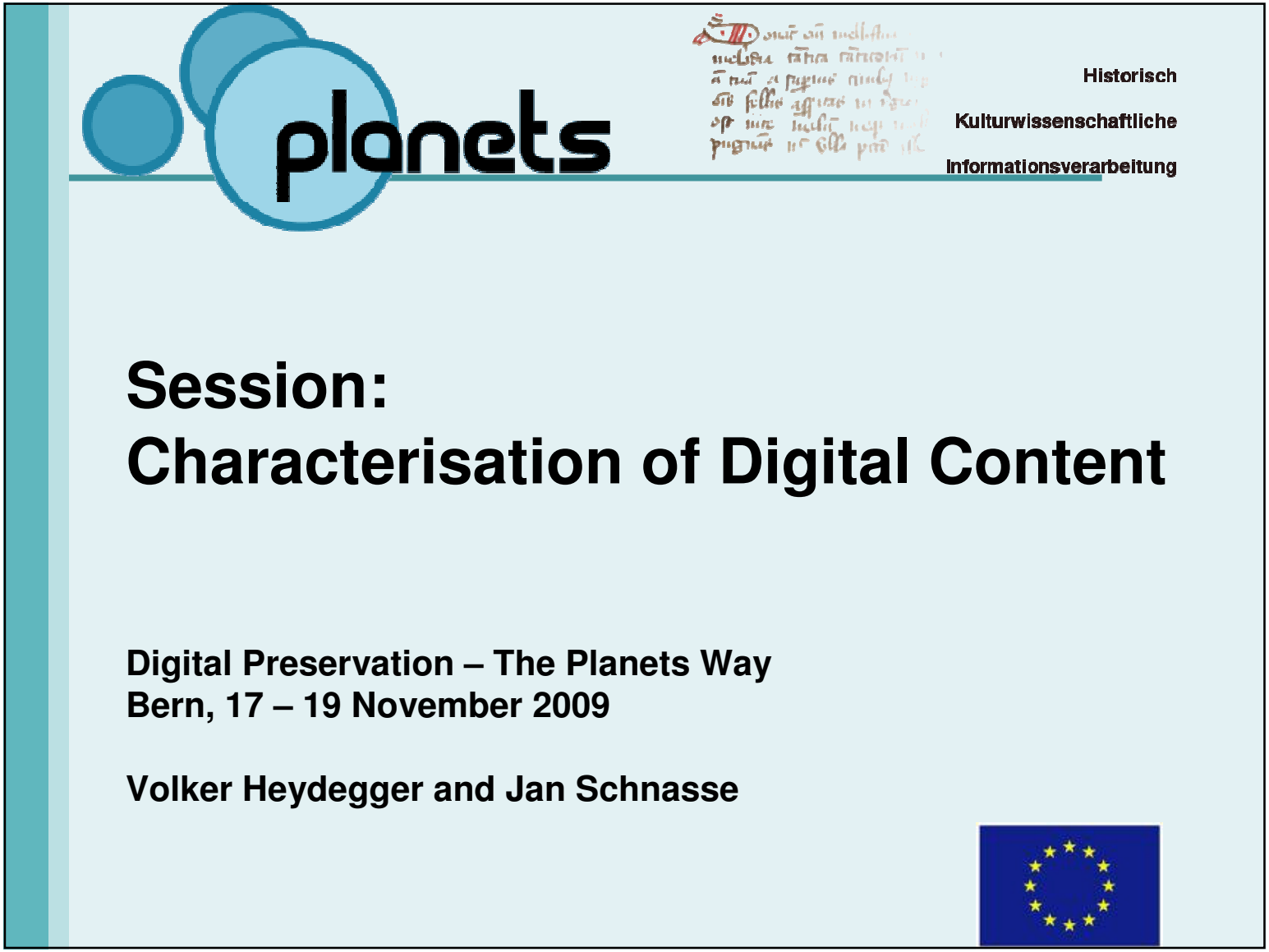
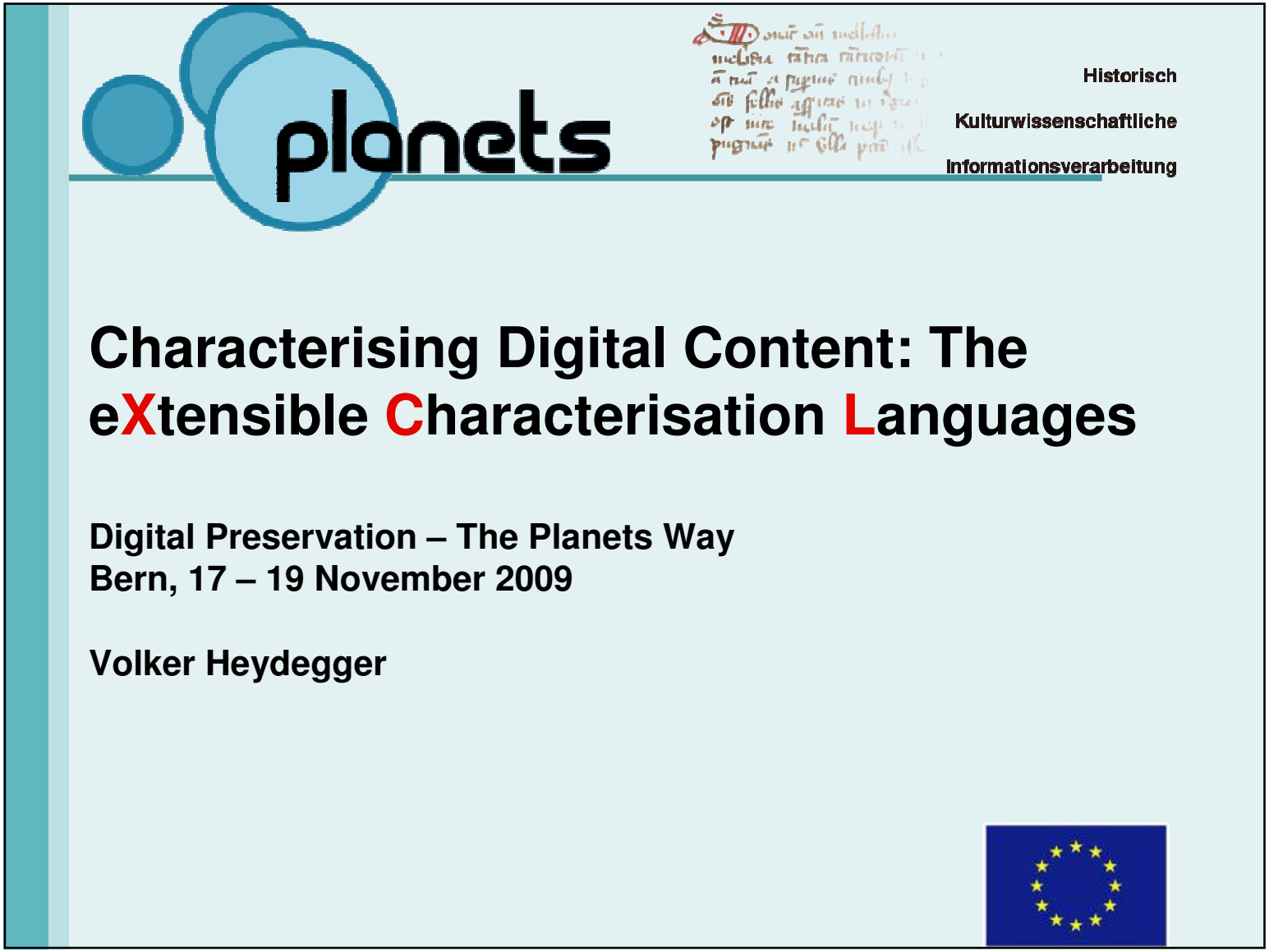

planets
Historisch
Kulturwissenschaftliche
Informationsverarbeitung
Session:
Characterisation of Digital Content
Digital Preservation – The Planets Way
Bern, 17 – 19 November 2009
Volker Heydegger and Jan Schnasse

Overview

- Part 1: Characterising Digital Content: The eXtensible Characterisation Languages
- Part 2: Demonstration of XCL Tools: Evaluation of Format Conversion


planets
Historisch
Kulturwissenschaftliche
Informationsverarbeitung
Characterising Digital Content: The eXtensible Characterisation Languages
Digital Preservation – The Planets Way
Bern, 17 – 19 November 2009
Volker Heydegger

Overview

- Characterisation: General issues
- About File Formats
- XCL overview
- XCL by Example

Characterisation: General issues

Why characterisation?

"Characterisation is an essential precursor to preservation. It provides the information required to make preservation planning decisions about digital objects, and to validate the results of preservation actions. "

(A. Brown: Developing Practical Approaches to Active Preservation, IJDC, 2007)

Characterisation: General issues

Why characterisation?

"Characterisation is an essential precursor to preservation. It provides the information required to make preservation planning decisions about digital objects, and to validate the results of preservation actions. "

(A. Brown: Developing Practical Approaches to Active Preservation, IJDC, 2007)

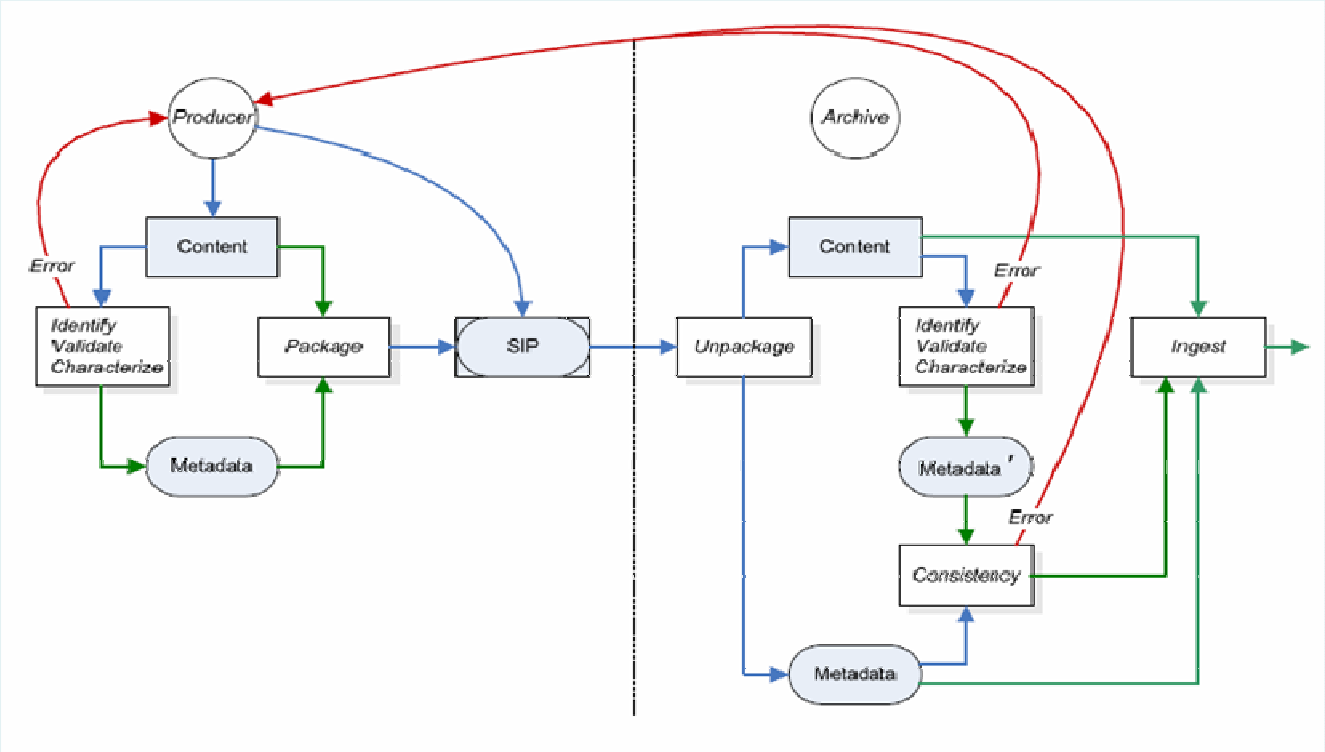
Characterisation: General issues

Why characterisation?

"Characterisation is an essential precursor to preservation. It provides the information required to make preservation planning decisions about digital objects, and to validate the results of preservation actions. "

(A. Brown: Developing Practical Approaches to Active Preservation, IJDC, 2007)

Why characterisation?



Source: S. Abrams: Automated Characterization in Preservation Workflows. *Tools and Trends*: *International Conference on Digital Preservation* Koninklijke Bibliotheek, 1-2 November 2007

Characterisation: General issues

What is subject to characterisation?

"One essential process in digital preservation is to perform format characterization to extract technical metadata associated with each digital object in the preservation archival collection. The technical metadata are important attributes for understanding and managing the digital archival collections, especially for format monitoring and researching format transformation procedures."

(C.C.H. Chou: Format Identification, Validation, Characterization and Transformation in DAITSS, [?2007])

Characterisation: General issues

What is subject to characterisation?

"One essential process in digital preservation is to perform format characterization to extract technical metadata associated with each digital object in the preservation archival collection. The technical metadata are important attributes for understanding and managing the digital archival collections, especially for format monitoring and researching format transformation procedures."

(C.C.H. Chou: Format Identification, Validation, Characterization and Transformation in DAITSS, [?2007])

Categories of characteristics

Non-technical characteristics ("associated metadata", contextual)

What's the name of the object?
Which software created the object?
Who holds the intellectual rights for the object?
When was the object modified for the last time?
Which collection does the object belong to?
Where is the object located in our repository?
…

Categories of characteristics

Technical characteristics

Information concerning
- structure of the object
- internal representation of the object
- instructions on how to represent the object

e.g. dimensions of an image, image data, compression, font information

Characterisation: General issues

Where do we find characteristics? How are they described?

"One essential process in digital preservation is to perform format characterization to extract technical metadata associated with each digital object in the preservation archival collection. The technical metadata are important attributes for understanding and managing the digital archival collections, especially for format monitoring and researching format transformation procedures."

(C.C.H. Chou: Format Identification, Validation, Characterization and Transformation in DAITSS, [?2007])

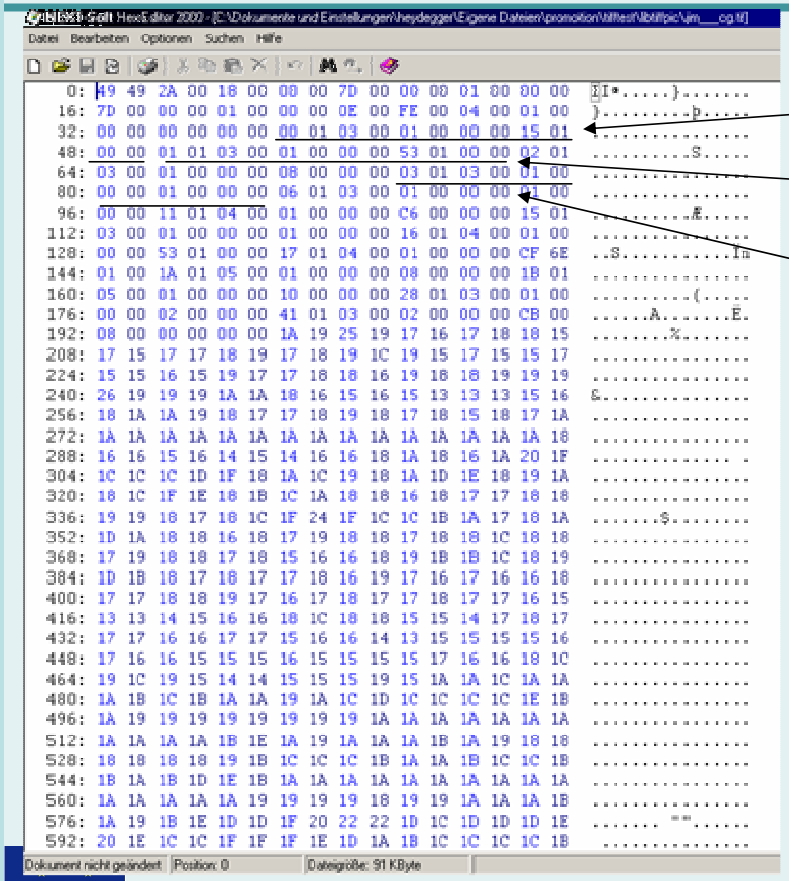


Image width: 277

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.

Some facts about file formats

How many file formats are there?

- PRONOM: ~ 550
- www.wotsit.org: ~ 900
- www.fileformat.info: 567
- www.fileinfo.com: > 3000 (file extensions)

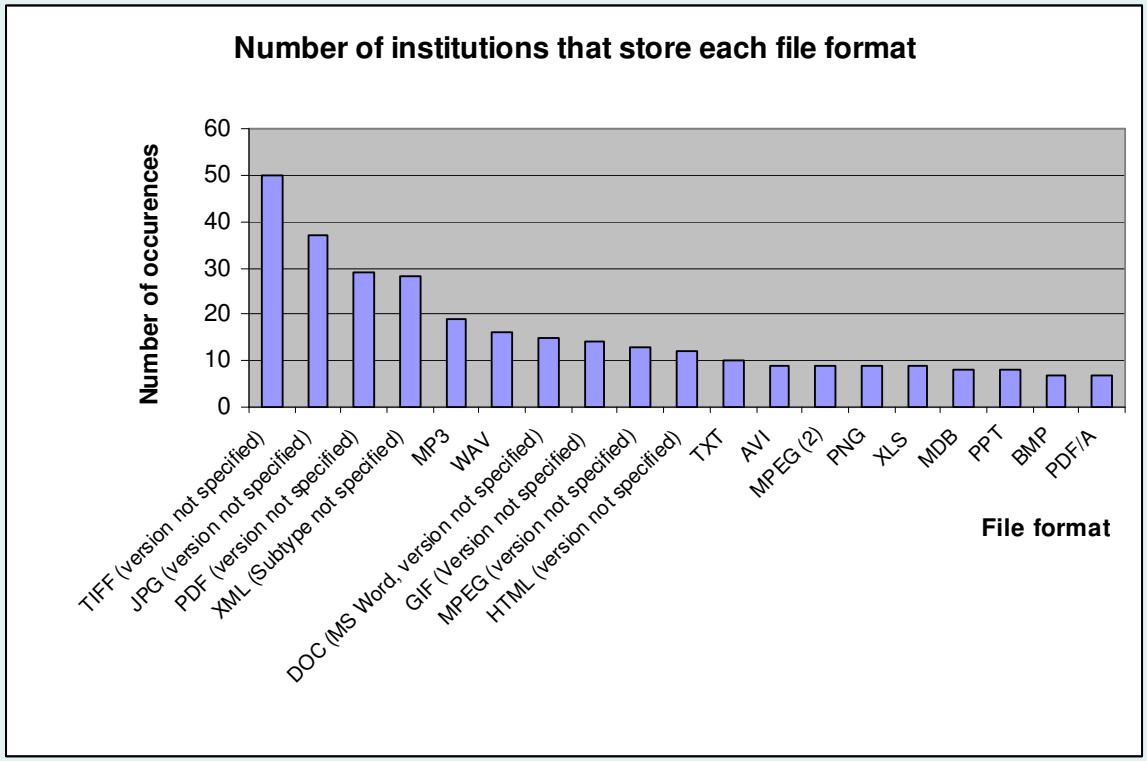
How many file formats can we find in institutions?

Planets internal study: “Gap analysis in tool provision”

- 76 institutions from 13 countries
- 124 different file formats

Source: Planets internal report:Gap analysis in tool provision (third version).

How many file formats are used more often?



Source: Planets internal report:
Gap analysis in tool provision (third version).

Suitability of formats for preservation (1)

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

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

Suitability of formats for preservation (2)

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

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

Suitability of formats for preservation (3)

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

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

Criteria for suitability

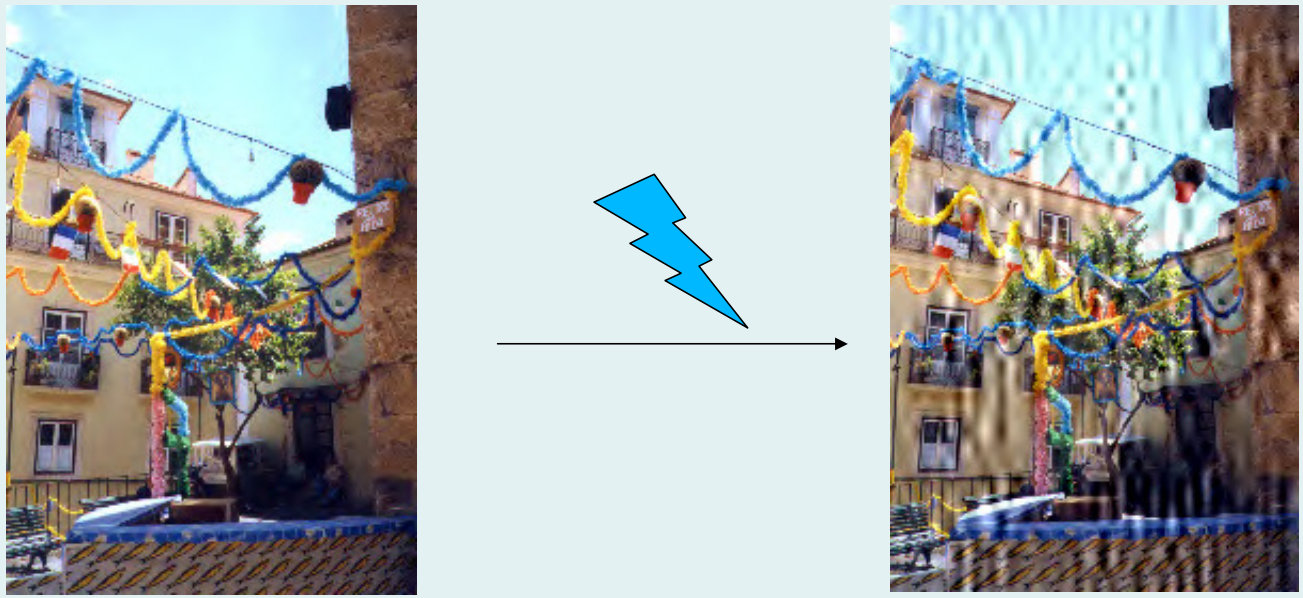
- Openess
- Adoption
- Complexity
- Technical protection mechanism
- Self-documentation
- Robustness
- Dependencies

(J. Rog, C. van Wijk: Evaluating File Formats for Long-term Preservation, iPres 2007)

Example: Robustness of Formats

Robustness

::= resilience of file formats against bit-stream corruption



See: ‚Just one Bit in a Million: On the Effects of Data Corruption in Files'. http://planets-project.eu/publications/?l=j

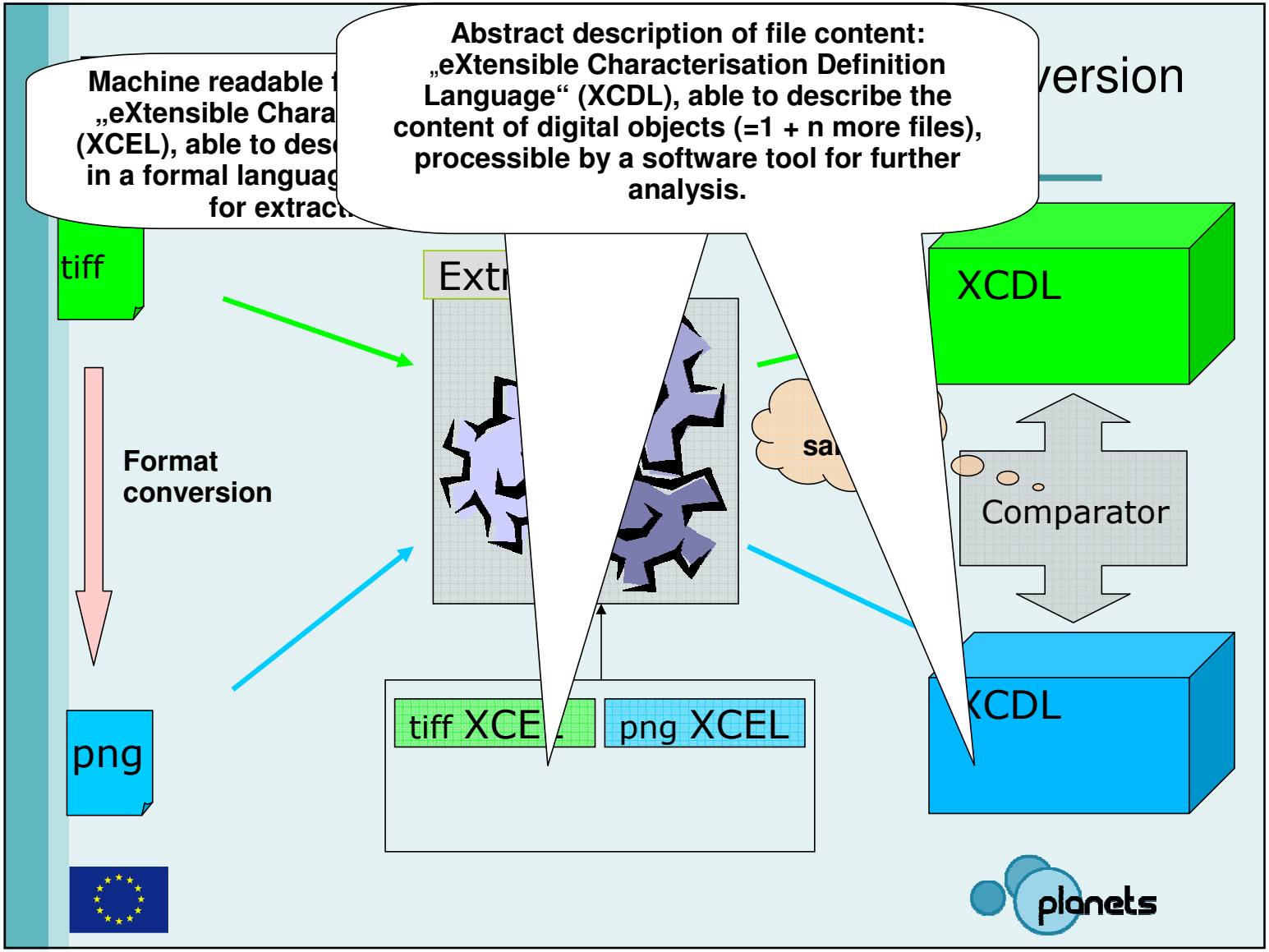
In a nutshell…

- Characterisation is an essential part within an overall preservation framework.
- File Format is the central concept for representation of digital content, i.e. the characteristics of objects.
- There is a large number of formats which characterise objects in its own different way, but only a couple of them are actually suitable for preservation.

3

XCL overview

- Support preservation planning task
- Support a specific preservation action task: Evaluation of file format conversion
- Develop a more abstract model for extraction of characteristics (syn. properties) from files
- Develop tools which use this model in order to enable characterisation in an efficiently, i.e. in an automated way

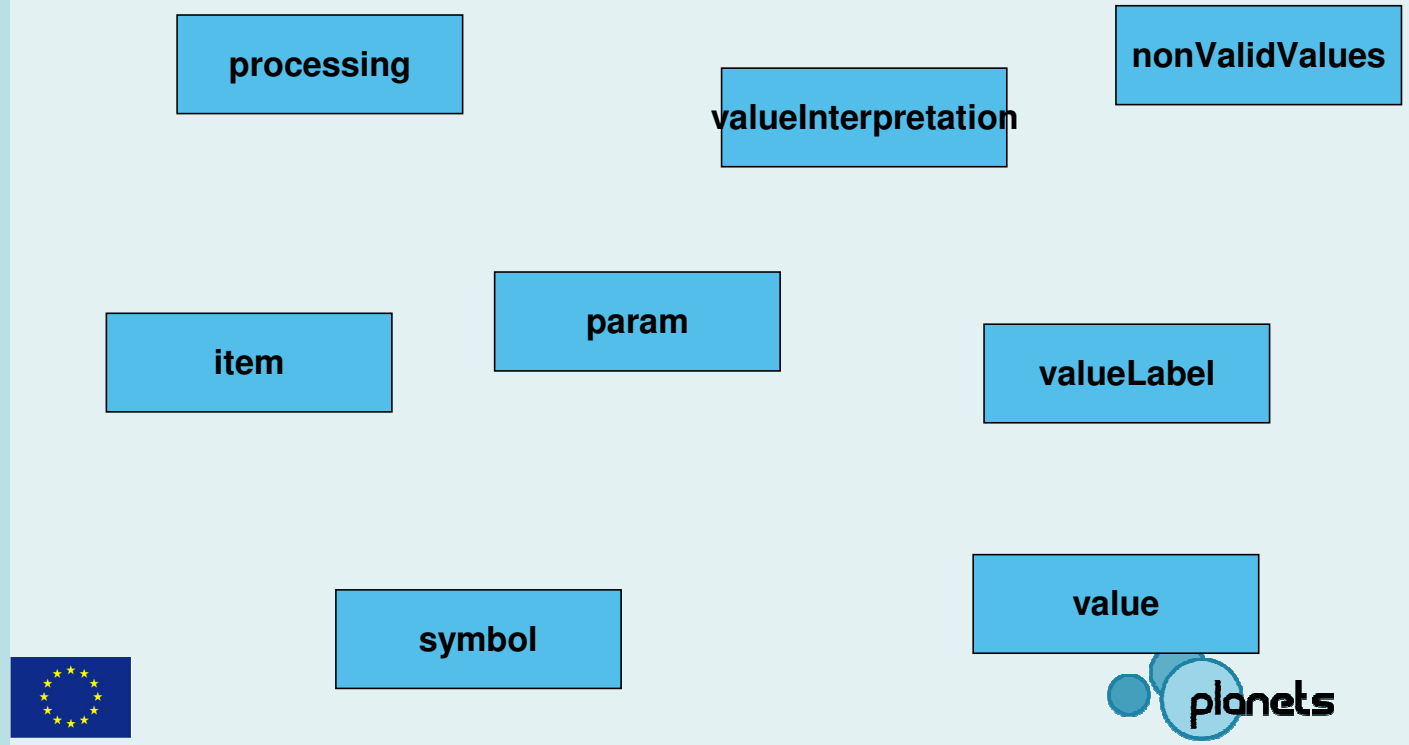

Machine readable f
„eXtensible Chara
(XCEL), able to des
in a formal languag
for extract
Abstract description of file content:
„eXtensible Characterisation Definition Language“ (XCDL), able to describe the content of digital objects (=1 + n more files), processible by a software tool for further analysis.
tiff
Extr
XCDL
Format conversion
sa
Comparator
tiff XCE
png XCEL
png
XCDL
planets

eXtensible Characterisation Extraction Language (XCEL)

- Describing how properties of digital objects are stored
- File format specification tagged in XML, according to the XCEL language definitions
- Interpretable through an XCEL interpreter (Extractor), able to extract characteristics

XCEL: Basic Structuring Elements

There are just a few elements sufficient enough to describe a file format:

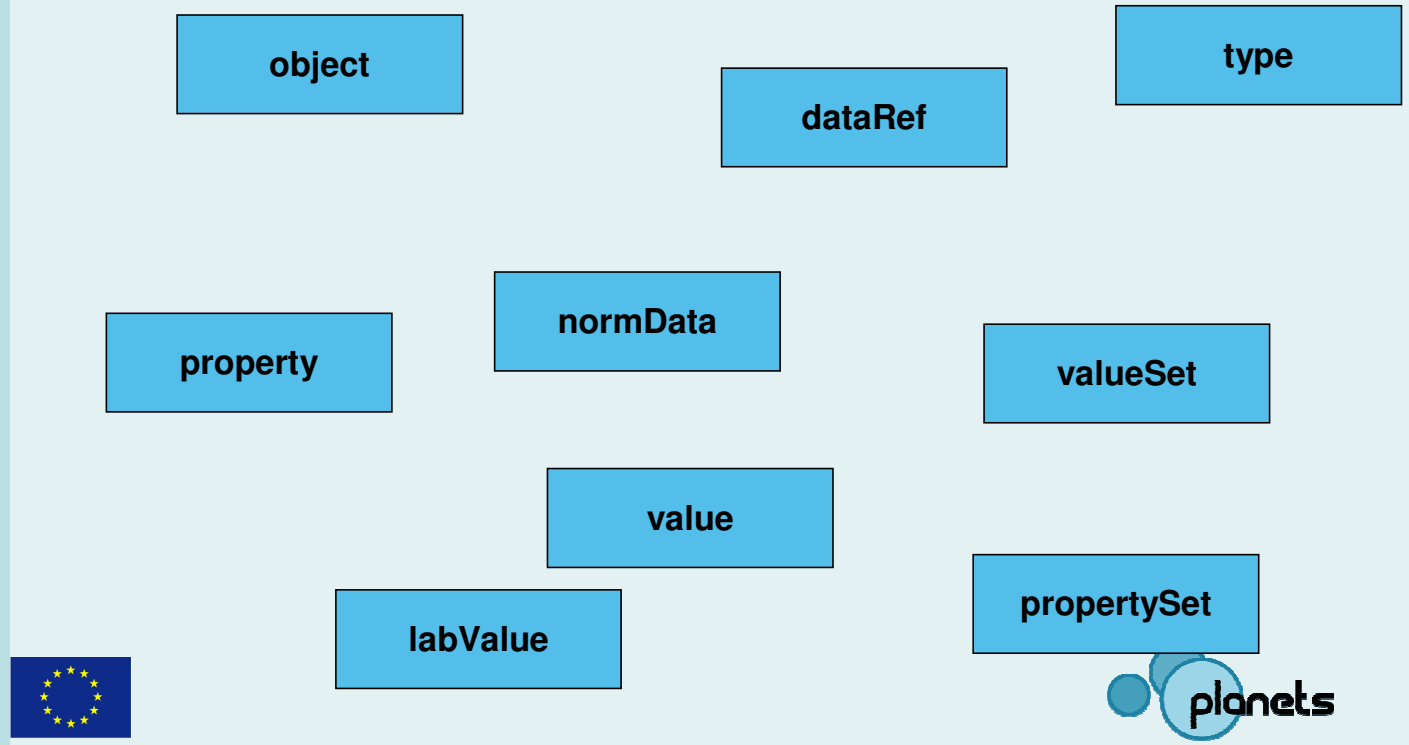


eXtensible Characterisation Definition Language (XCDL)

- Describes the content of a file /set of files in an abstract way.
- Designed for decription of the content of *any* file format.
- Designed as a means to describe only parts *or* all of the content.

XCDL: Basic Structuring Elements

Again, there are just a few elements sufficient enough to describe the content of a digital object:



Benefits of the XCL approach

- XCL is a generic solution, uses an abstract model, provides a unique vocabulary (‚XCL ontology').

→Extensible: XCL is based on XML

→XCEL provides a means for description of any file format

→ XCDL is a language with which all sort of content can be expressed

5

XCL by Example

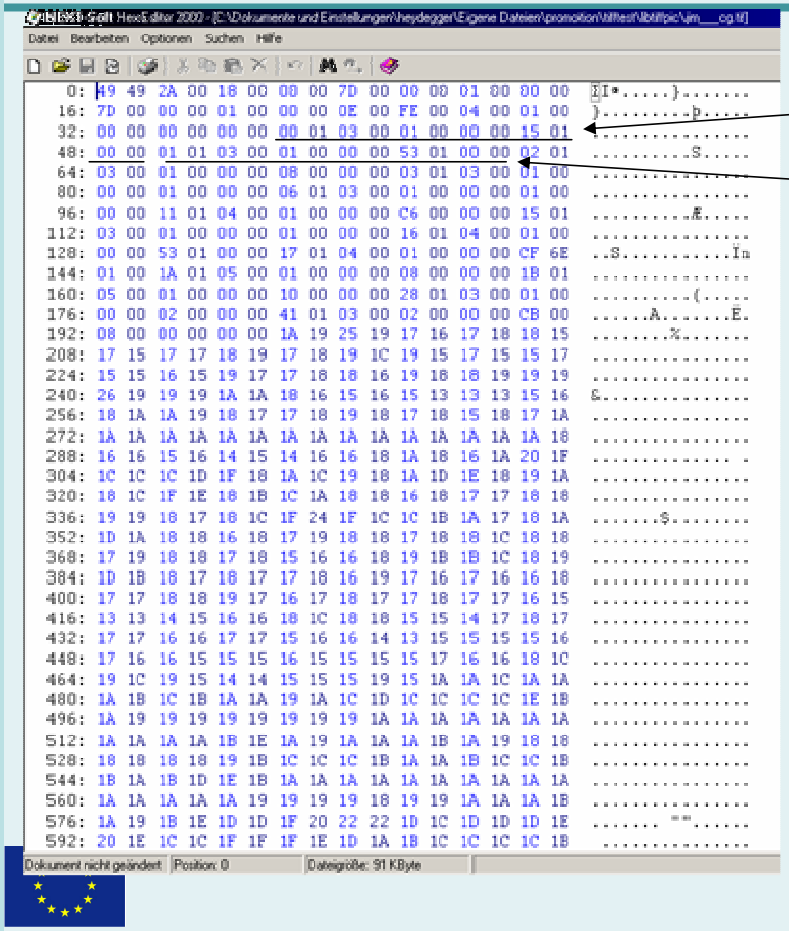


Image width: 277

Image length: 339

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.

XCEL representation

```
<!-- Tag 256: ImageWidth (XCL: imageWidth) -->
  <item xsi:type="structuringItem" identifier="IFDE_256"
   optional="true">
   <symbol interpretation="uint16" length="2" value="256"/>
   <item xsi:type="structuringItem" order="choice">
     <item xsi:type="structuringItem" order="sequence">
       <!– Data type (value ‚3‘ means uint16)-->
       <symbol interpretation="uint16" length="2"  value="3"/>
       <!– number of values (N)->
       <symbol interpretation="uint32" length="4" value="1"/>
       <!-- the value and name of property -->
       <symbol interpretation="uint16" length="2"
       name="imageWidth"/>
       <!-- wasted space-->
       <symbol interpretation="uint16" length="2"/>
       […]
     </item>
   </item>
  </item>
```

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.

XCEL representation

```
<!-- Tag 256: ImageWidth (XCL: imageWidth) -->
  <item xsi:type="structuringItem" identifier="IFDE_256"
   optional="true">
   <symbol interpretation="uint16" length="2" value="256"/>
   <item xsi:type="structuringItem" order="choice">
     <item xsi:type="structuringItem" order="sequence">
       <!– Data type (value ‚3‘ means uint16)-->
       <symbol interpretation="uint16" length="2"  value="3"/>
       <!– number of values (N)->
       <symbol interpretation="uint32" length="4" value="1"/>
       <!-- the value and name of property -->
       <symbol interpretation="uint16" length="2"
       name="imageWidth"/>
       <!-- wasted space-->
       <symbol interpretation="uint16" length="2"/>
       […]
     </item>
   </item>
  </item>
```

ImageWidth

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

Tag = 256 (100.H)

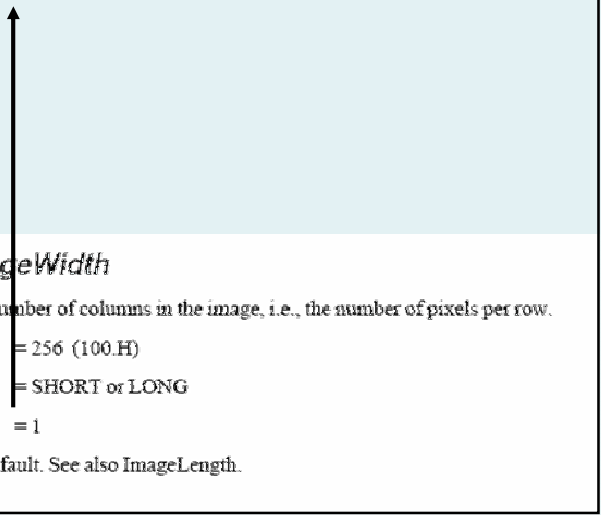
Type = SHORT or LONG

N = 1

No default. See also ImageLength.

XCEL representation

```
<!-- Tag 256: ImageWidth (XCL: imageWidth) -->
  <item xsi:type="structuringItem" identifier="IFDE_256"
   optional="true">
   <symbol interpretation="uint16" length="2" value="256"/>
   <item xsi:type="structuringItem" order="choice">
     <item xsi:type="structuringItem" order="sequence">
       <!– Data type (value ‚3‘ means uint16)-->
       <symbol interpretation="uint16" length="2"  value="3"/>
       <!– number of values (N)->
       <symbol interpretation="uint32" length="4" value="1"/>
       <!-- the value and name of property -->
       <symbol interpretation="uint16" length="2"
       name="imageWidth"/>
       <!-- wasted space-->
       <symbol interpretation="uint16" length="2"/>
       […]
     </item>
   </item>
  </item>
```



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.

XCDL representation

```
...
<property id="p5">
    <name id="id30" >imageWidth</name>
    <valueSet id="i_i1_s4" >
      <labValue>
        <val>277</val>
        <type>int</type>
      </labValue>
    </valueSet>
  </property>
...
```

XCEL entry:
<!-- the value and name of property -->
<symbol interpretation="uint16" length="2"
name="imageWidth"/>

XCDL representation

```
...
<property id="p5">
    <name id="id30" >imageWidth</name>
    <valueSet id="i_i1_s4" >
      <labValue>
        <val>277</val>
        <type>int</type>
      </labValue>
    </valueSet>
  </property>
...
```

XCEL entry:

```
<!– Data type (value ‚3‘ means uint16)-->
    <symbol interpretation="uint16"
      length="2"  value="3"/>
```

XCDL representations can now be compared…

Measure name: equal

Id: 1
Explanation: Metric 'equal' is a simple comparison of two values (A, B) of any XCL data type on equality.
Data type of input value: Any XCL data type
Data type of output value: XCL: boolean (true, false)

Example:

Value for property X of XCDL1 (src)	Value for property X of XCDL2 (tar)
<labValue> <val>32</val> <type>int</type> </labValue>	<labValue> <val>32</val> <type>int</type> </labValue>
copra output: … <property id="2" name="imageHeight" unit="pixel" state="complete"> <metrics> <metric id="1" name="equal"> <result state="ok">true</result> </metric> </metrics> </property> …	

Thank you!

More about XCL on:

http://planetarium.hki.uni-koeln.de/planets_cms/index.php