Digital Preservation

File formats and registries Characterisation

Hannes Kulovits
Vienna University of Technology
www.ifs.tuwien.ac.at/~kulovits
kulovits@ifs.tuwien.ac.at
www.ifs.tuwien.ac.at/dp

Agenda

- File formats and issues
- File format identification
- Registries
- Characterisation tools
  - DROID (Digital Record Object Identification)
  - JHove (JSTOR/Harvard Object Validation Environment)
  - XCL (eXtensible Characterisation Language)
  - FITS (File Information Tool Set)

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, IETF, ...)
- 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://library.harvard.edu/preservation/digital-preservation_gdfr.html

- Unified Digital Format Registry (UDFR)

http://www.udfr.org/

- Sustainability of Digital Formats Planning for Library of Congress Collections

http://www.digitalpreservation.gov

- FileExt

http://filext.com

PRONOM

The technical registry
PRONOM

Details: File format summary

? Help : detailed report on file format

Details for: Microsoft Word for Windows Document 97-2003

Save as... XML | CSV Print

Go to: Summary | Documentation > | Signatures > | Compression > | Character encoding > | Rights > | Reference files > Properties >

Summary

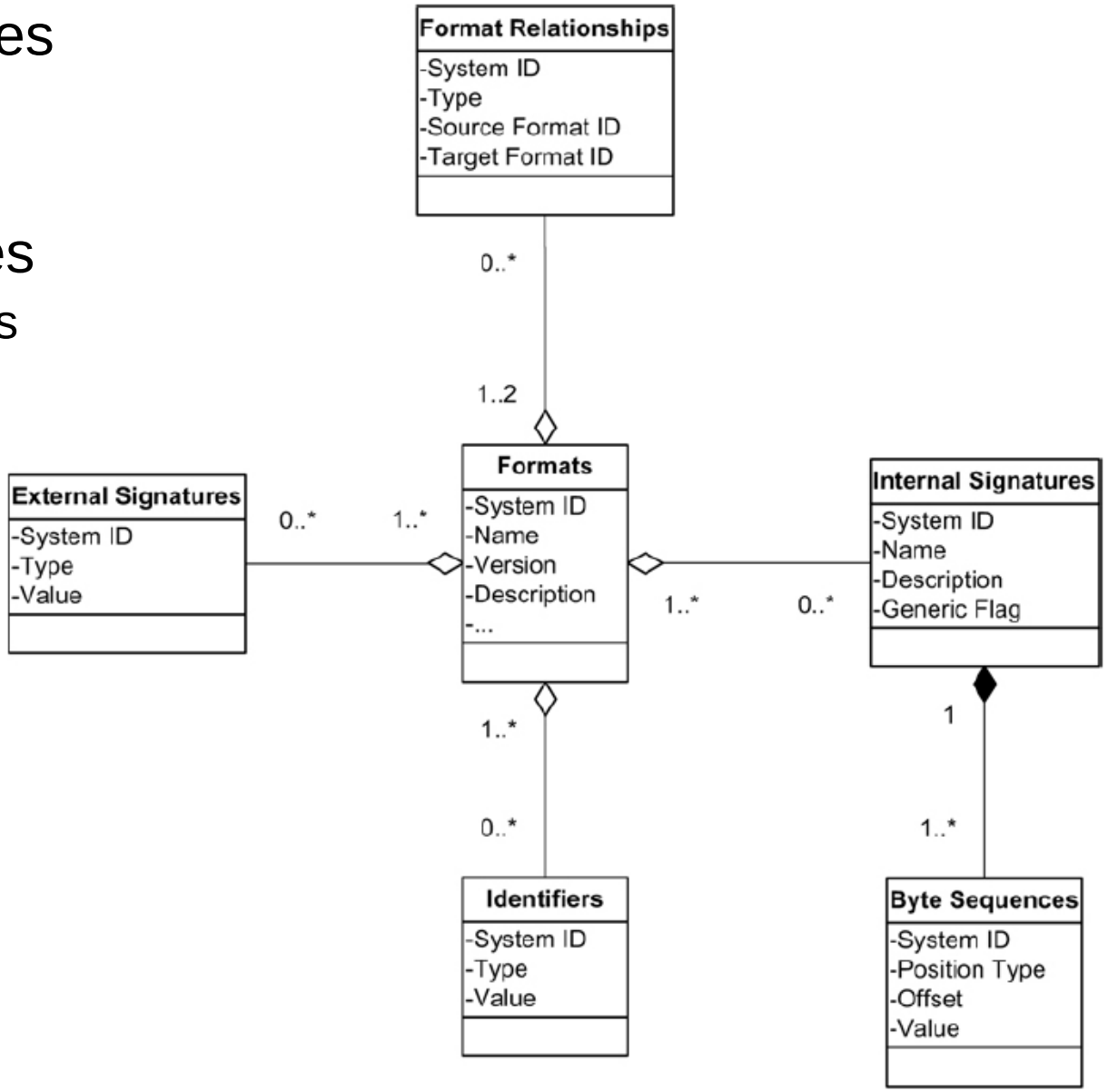
Name	Microsoft Word for Windows Document
Version	97-2003
Other names	Microsoft Word for Windows Document (97-XP)
Identifiers	MIME: application/msword Apple Uniform Type Identifier: com.microsoft.word.doc PUID: fmt/40
Family	
Classification	Text (Wordprocessed)
Disclosure	None
Description	With the release 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 Is subsequent version of Microsoft Word for Windows Document (6.0/95) 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



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.

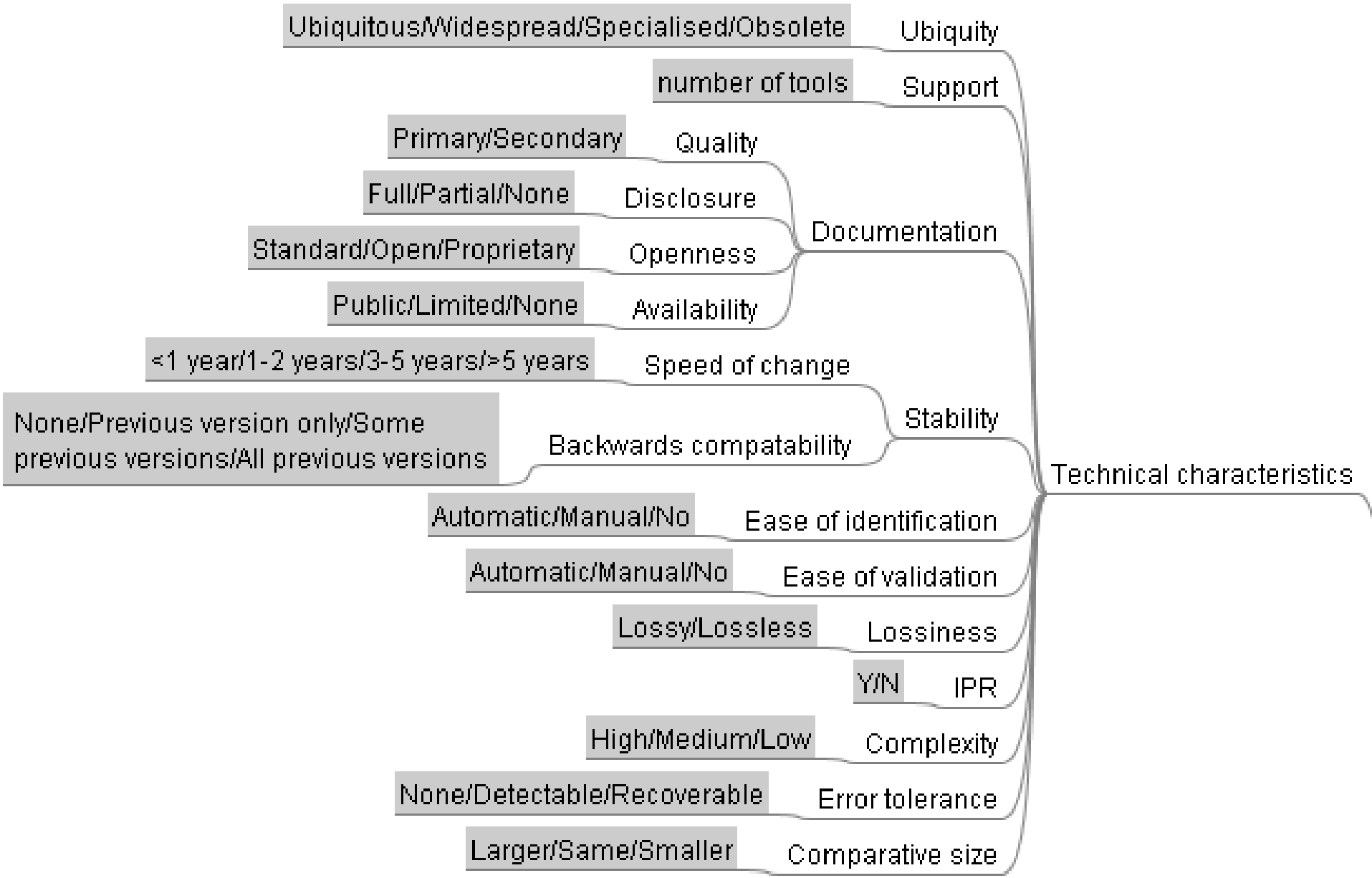
Demo: DROID, PRONOM

Registry content

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

- Risk factors

File format characteristics



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
- JHove2
  - Identification and validation
  - Feature extraction
  - Policy based assessment
  - Able to handle complex objects

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 be 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?"

Use Case Coverage

- Identification
- Risk assessment

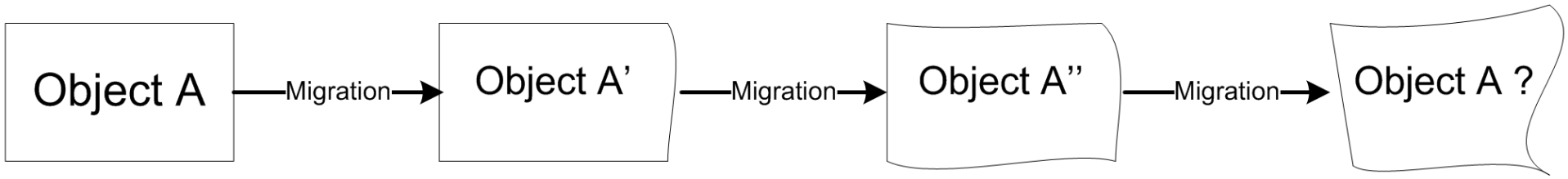
- Delivery
- Transformation

- 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

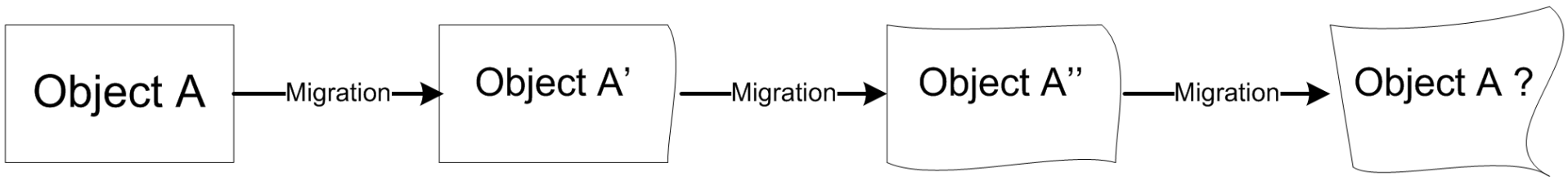


Validating a migrated image

- Yes, it's in JPEG 2000 format

- Yes, it's well-formed
- Yes, it's valid
- Yes, it still has the same dimensions
- …. But is it still the same image?

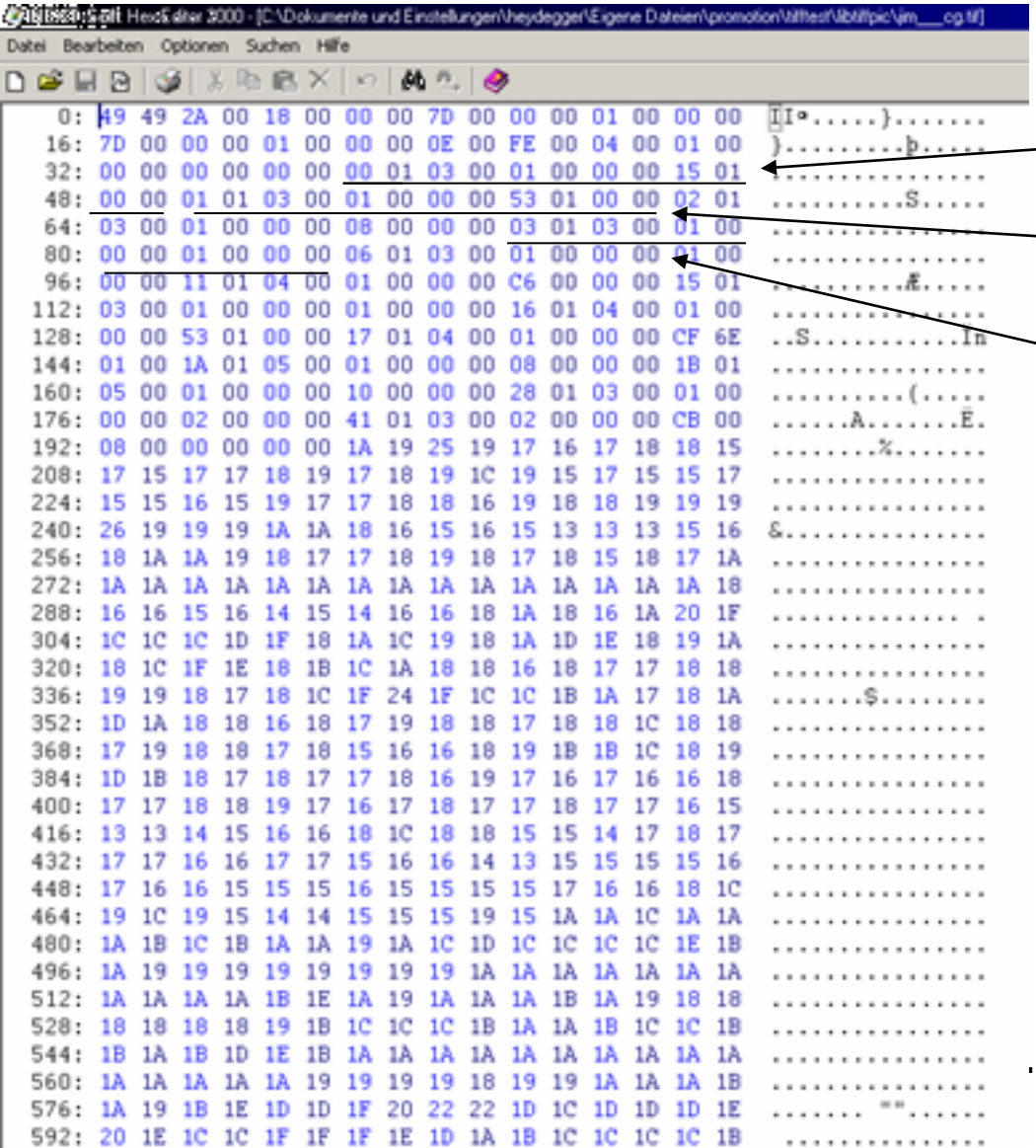
- We need more characterisation.



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



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.

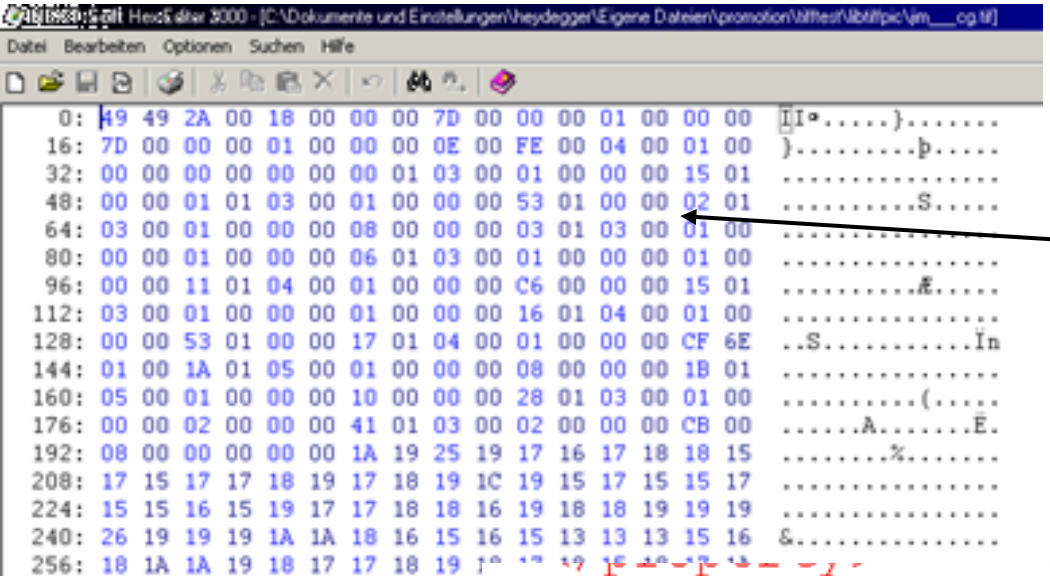
Tag = 256 (100.H)

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

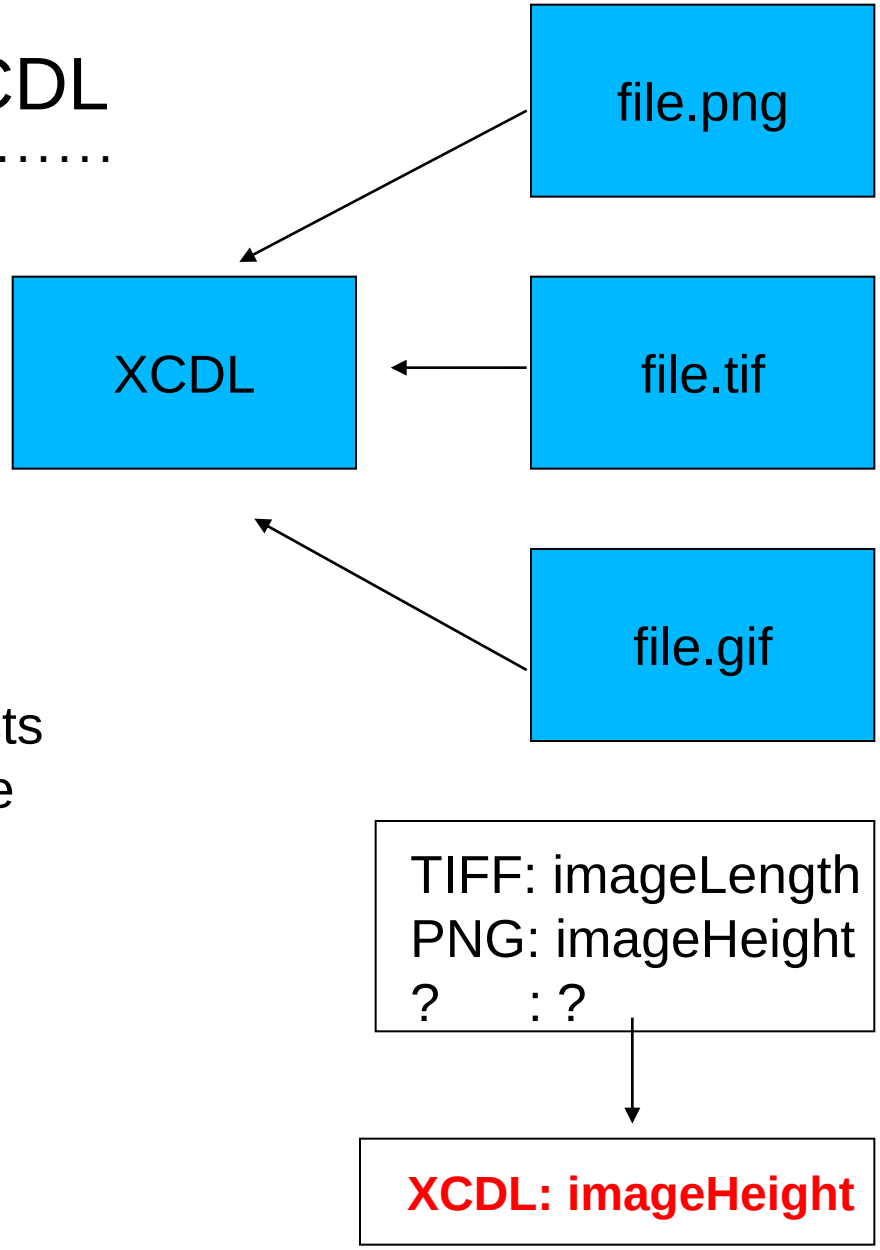
N = 1

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

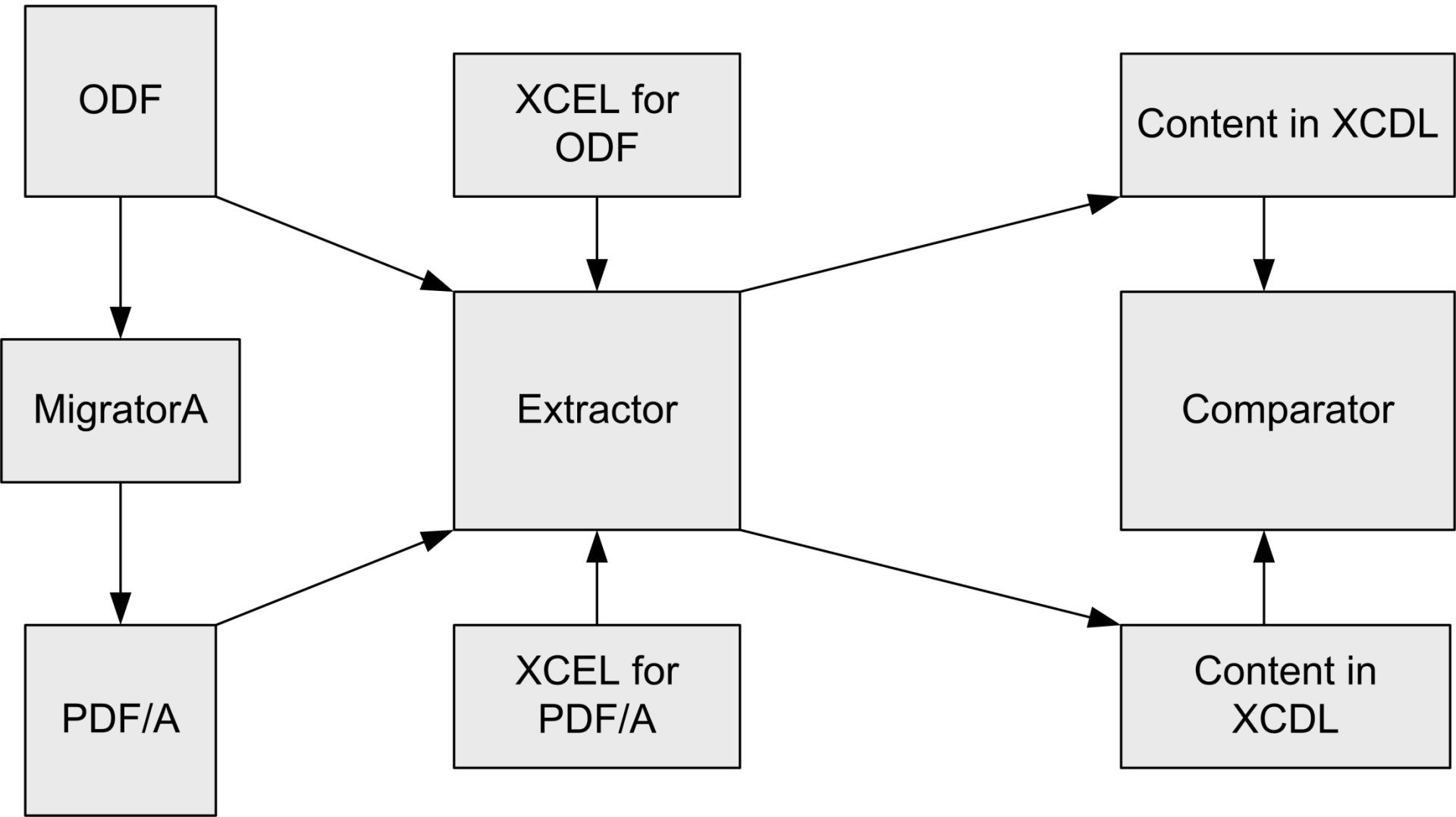
- Preprocessing instructions
  - Configuration tasks
- Format description
  - Defines the structure of an object
- Templates
  - Describe recurring structures
- Postprocessing instructions
  - On the results of processing

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



Conclusion

- 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?”

Questions?

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

www.ifs.tuwien.ac.at/dp