



A generic XML language for characterising
objects to support digital preservation

Christoph Becker, Andreas Rauber, Volker Heydegger,
Jan Schnasse, Manfred Thaller

23rd ACM Symposium on Applied Computing (SAC2008)

Fortaleza, Brasil, March 16-20 2008

Outline

- ❑ Digital Preservation
- ❑ Preservation Planning
 - Evaluation of potential actions
 - Essential characteristics of digital objects
- ❑ The eXtensible Characterisation Languages (XCL)
 - The description language XCDL
 - The extraction language XCEL
 - Comparator
- ❑ Current and future work



The Longevity of Digital Objects

- ❑ Digital objects are the dominant way we exchange information
- ❑ Heterogeneity and complexity of file formats and speed of technological change make long-term access a challenge
- ❑ Digital preservation: Long-term storage and access to digital objects
- ❑ Digital objects need technical environment to “function”
- ❑ Dominant types of preservation actions:
 - Migration
 - Emulation

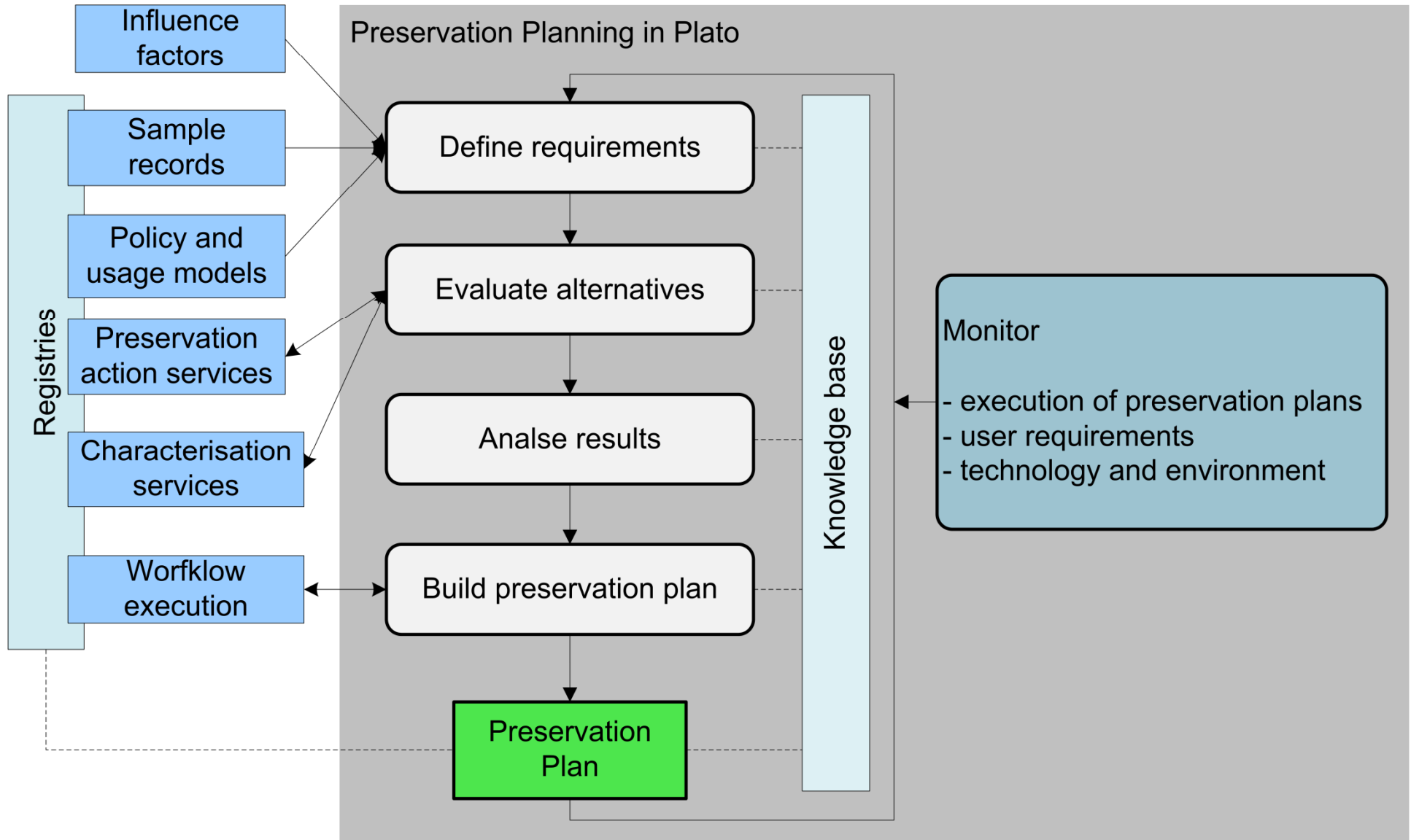


Evaluating preservation strategies

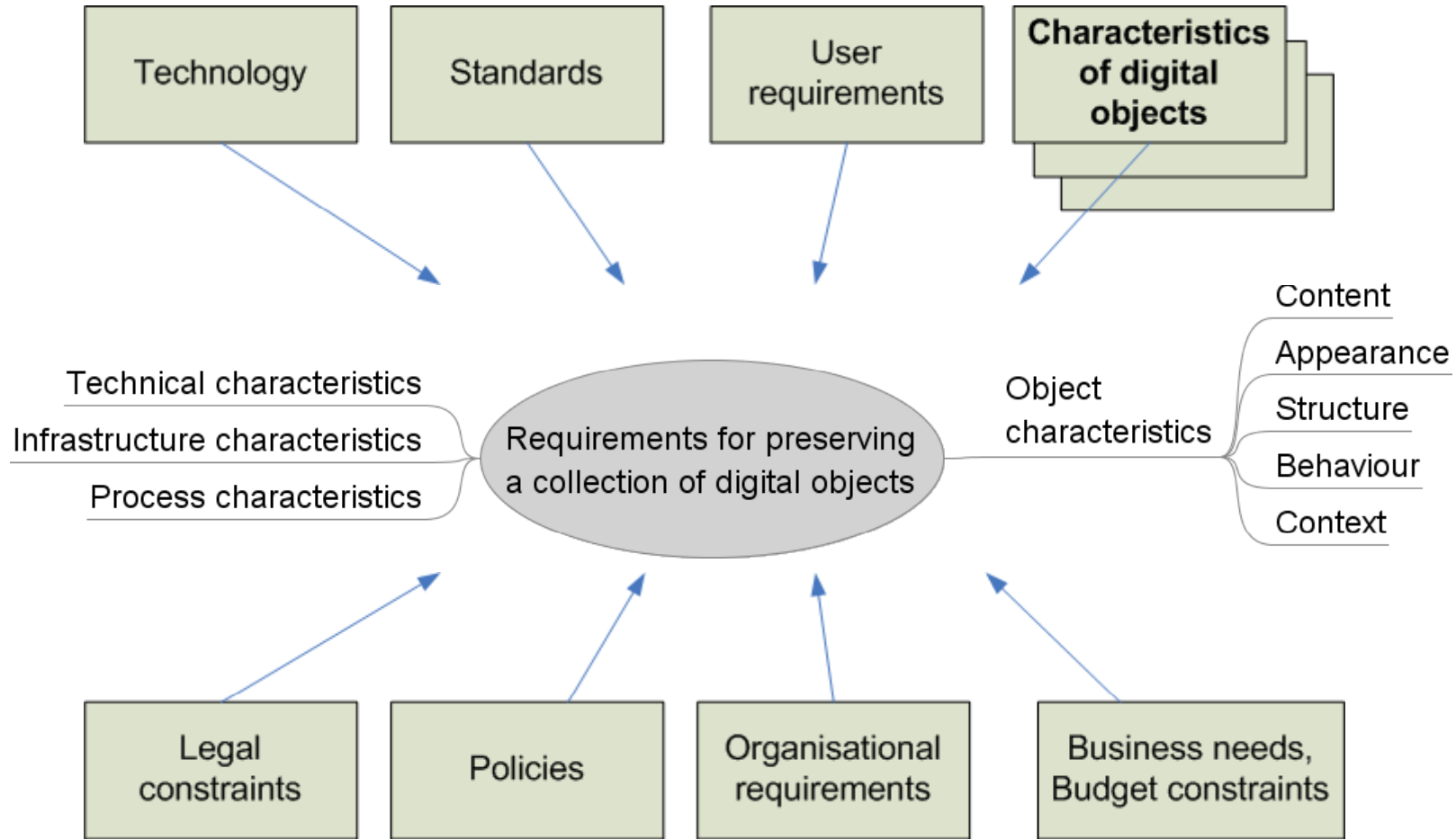
- ❑ Variety of solutions and tools exist
 - ❑ Each strategy has unique strengths and weaknesses
 - ❑ Requirements vary across settings
 - ❑ Decision on which solution to adopt is complex
 - ❑ Documentation and accountability is essential
-
- ❑ Preservation planning assists in decision making
 - ❑ Evaluating preservation strategies on representative samples according to specific requirements and criteria



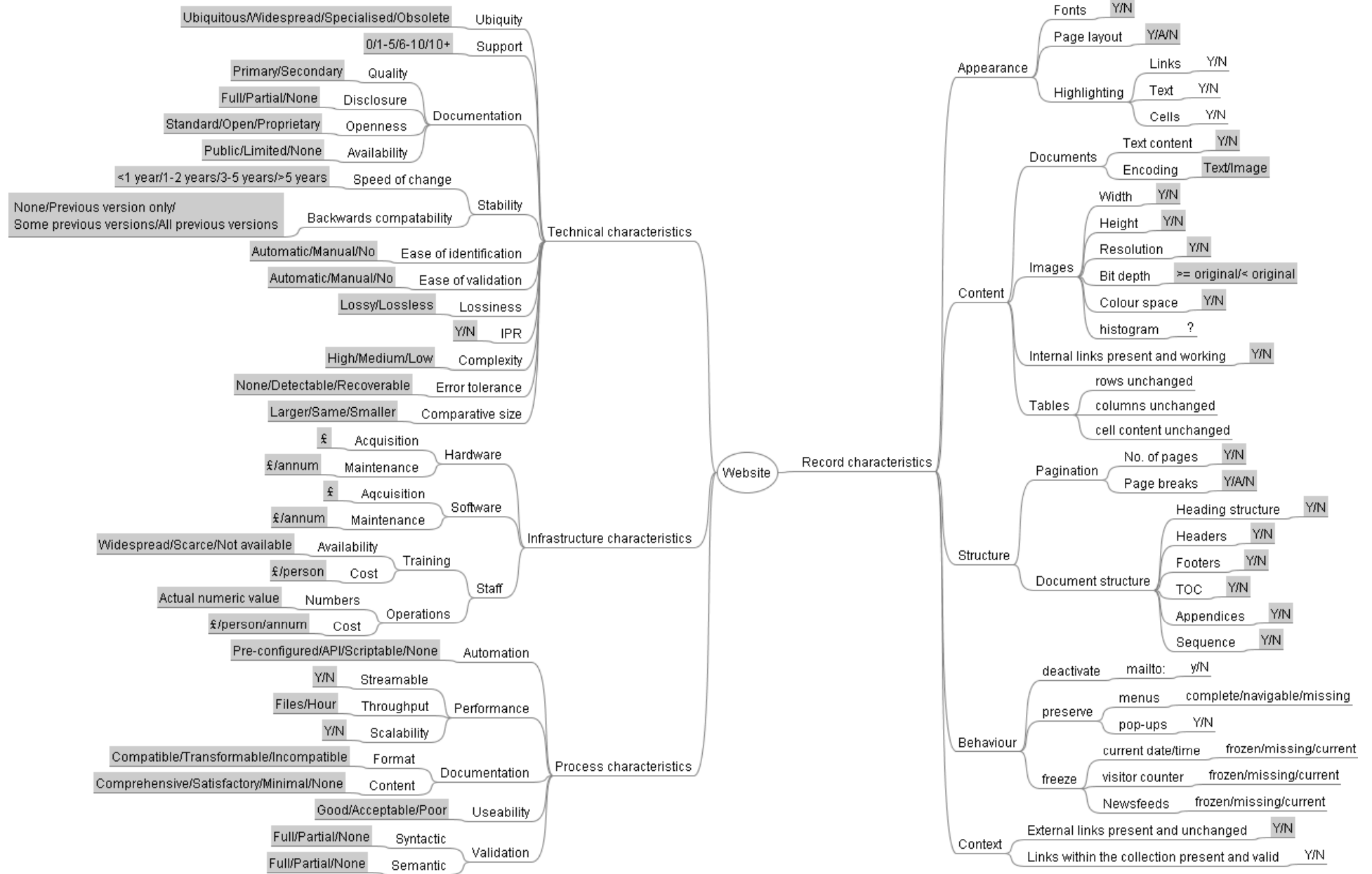
Planets Preservation Planning Workflow



Influence Factors



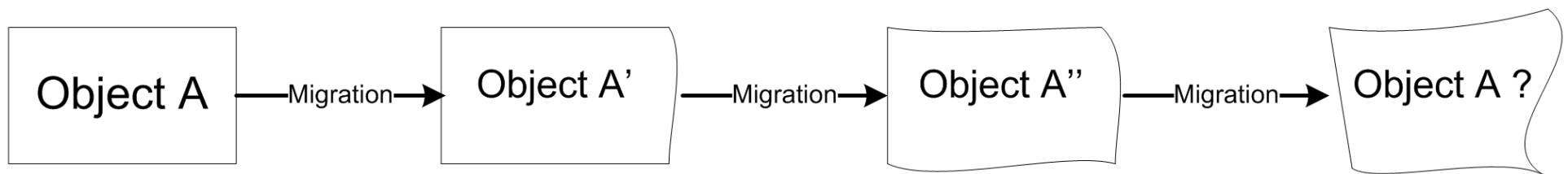
An Objective Tree



Core requirement: Keep object intact

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

- ❑ Image width in Pixel
- ❑ Color depth in bit



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

```

HexEditor 2000 - [C:\Dokumente und Einstellungen\heydegger\Eigene Dateien\promotion\lntest\lntest\pic\ym_..._cg.tif]
Datei Bearbeiten Optionen Suchen Hilfe
[Icons] [Icons] [Icons] [Icons] [Icons] [Icons] [Icons] [Icons] [Icons] [Icons] [Icons] [Icons] [Icons] [Icons] [Icons] [Icons]
0: 49 49 2A 00 18 00 80 00 7D 00 00 00 01 00 00 00  [I]*.....}.....
16: 7D 00 00 00 01 00 80 00 0E 00 FE 00 04 00 01 00  ).....p.....
32: 00 00 00 00 00 00 80 01 03 00 01 00 00 00 15 01  ←.....
48: 00 00 01 01 03 00 81 00 00 00 53 01 00 00 02 01  ←.....S.....
64: 03 00 01 00 00 00 88 00 00 00 03 01 03 00 01 00  ←.....E.....
80: 00 00 01 00 00 00 86 01 03 00 01 00 00 00 01 00  ←.....
96: 00 00 11 01 04 00 81 00 00 00 C6 00 00 00 15 01  ..S.....In
112: 03 00 01 00 00 00 81 00 00 00 16 01 04 00 01 00  ..S.....
128: 00 00 53 01 00 00 17 01 04 00 01 00 00 00 CF 6E  ..S.....
144: 01 00 1A 01 05 00 81 00 00 00 08 00 00 00 1B 01  .....(.....
160: 05 00 01 00 00 00 10 00 00 00 28 01 03 00 01 00  .....A.....E.
176: 00 00 02 00 00 00 41 01 03 00 02 00 00 00 CB 00  .....x.....
192: 08 00 00 00 00 00 1A 19 25 19 17 16 17 18 18 15  .....
208: 17 15 17 17 18 19 17 18 19 1C 19 15 17 15 15 17  .....
224: 15 15 16 15 19 17 17 18 18 16 19 18 18 19 19 19  .....
240: 26 19 19 19 1A 1A 18 16 15 16 15 13 13 13 15 16  &.....
256: 18 1A 1A 19 18 17 17 18 19 18 17 18 15 18 17 1A  .....
272: 1A 1A 1A 1A 1A 1A 1A 1A 1A 1A 1A 1A 1A 1A 1A 18  .....
288: 16 16 15 16 14 15 14 16 16 18 1A 18 16 1A 20 1F  .....
304: 1C 1C 1C 1D 1F 18 1A 1C 19 18 1A 1D 1E 18 19 1A  .....
320: 18 1C 1F 1E 18 1B 1C 1A 18 18 16 18 17 17 18 18  .....
336: 19 19 18 17 18 1C 1F 24 1F 1C 1C 1B 1A 17 18 1A  .....$.
352: 1D 1A 18 18 16 18 17 19 18 18 17 18 18 1C 18 18  .....
368: 17 19 18 18 17 18 15 16 16 18 19 1B 1B 1C 18 19  .....
384: 1D 1B 18 17 18 17 17 18 16 19 17 16 17 16 16 18  .....
400: 17 17 18 18 19 17 16 17 18 17 17 18 17 17 16 15  .....
416: 13 13 14 15 16 16 18 1C 18 18 15 15 14 17 18 17  .....
432: 17 17 16 16 17 17 15 16 16 14 13 15 15 15 15 16  .....
448: 17 16 16 15 15 15 16 15 15 15 15 17 16 16 18 1C  .....
464: 19 1C 19 15 14 14 15 15 15 19 15 1A 1A 1C 1A 1A  .....
480: 1A 1B 1C 1B 1A 1A 19 1A 1C 1D 1C 1C 1C 1C 1E 1B  .....
496: 1A 19 19 19 19 19 19 19 19 1A 1A 1A 1A 1A 1A 1A  .....
512: 1A 1A 1A 1A 1B 1E 1A 19 1A 1A 1A 1B 1A 19 18 18  .....
528: 18 18 18 18 19 1B 1C 1C 1C 1B 1A 1A 1B 1C 1C 1B  .....
544: 1B 1A 1B 1D 1E 1B 1A 1A 1A 1A 1A 1A 1A 1A 1A 1A  .....
560: 1A 1A 1A 1A 1A 19 19 19 19 18 19 19 1A 1A 1A 1B  .....
576: 1A 19 1B 1E 1D 1D 1F 20 22 22 1D 1C 1D 1D 1D 1E  ....."".....
592: 20 1E 1C 1C 1F 1F 1F 1E 1D 1A 1B 1C 1C 1C 1C 1B
  
```

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)

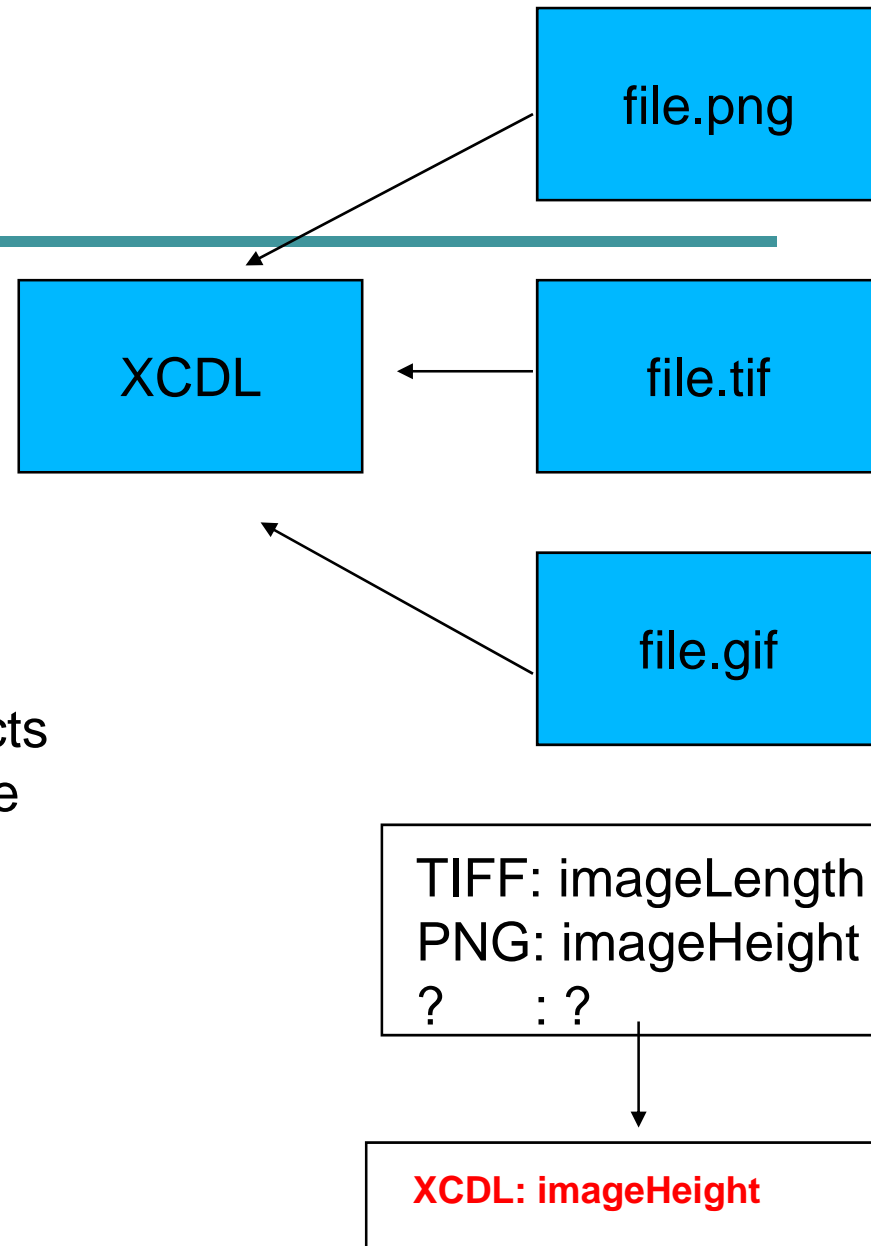
Type = SHORT or LONG

N = 1

No default. See also ImageLength.

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>XCLabel</type>
```

```
</labVal>
```

```
<dataRef ind="normSpecific">
```

```
<ref id="n6" start="3" end="11">
```

```
</dataRef>
```

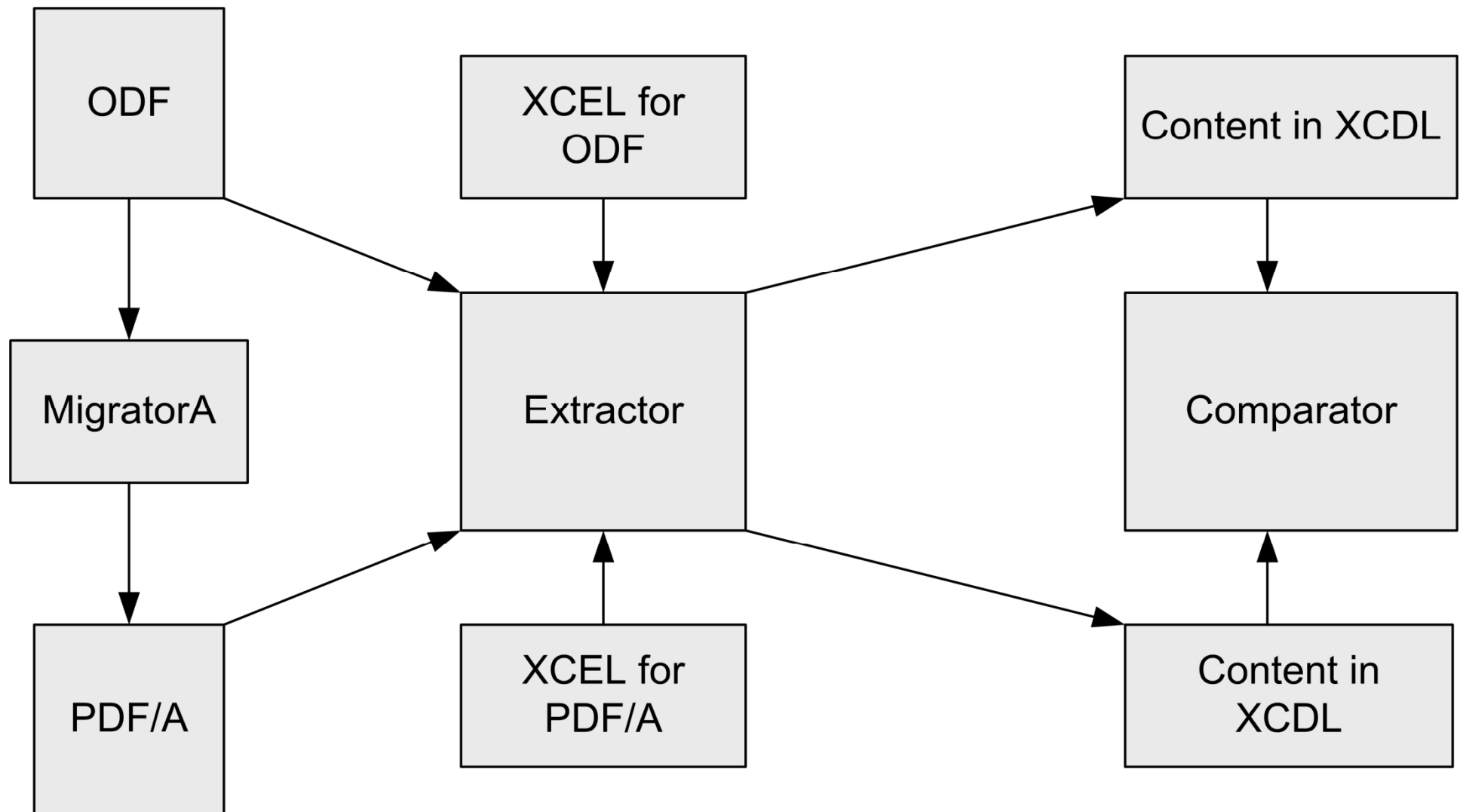
```
</valueSet>
```



.....



Comparing migrated documents











































<input checked="" type="checkbox"/>	TIFF
<input checked="" type="checkbox"/>	EPS
<input checked="" type="checkbox"/>	JPEG2000
<input checked="" type="checkbox"/>	RTF-acrobat
<input checked="" type="checkbox"/>	RTF-convertdoc
<input checked="" type="checkbox"/>	TXT

Show

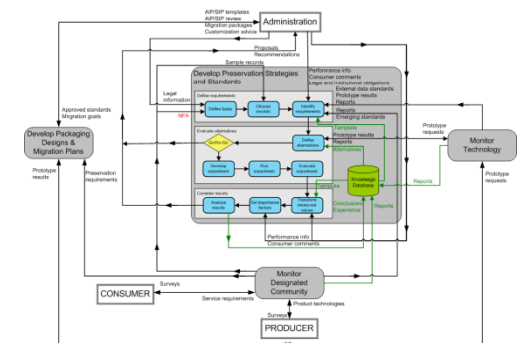
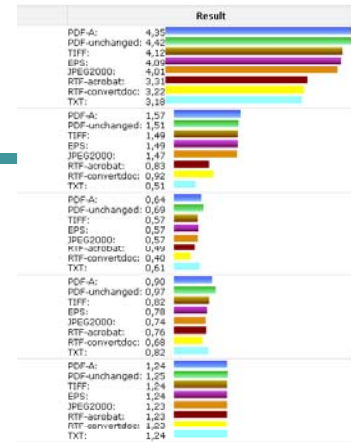
[Expand All](#) | [Collapse All](#)

ONB Master thesis > [Object characteristic](#)

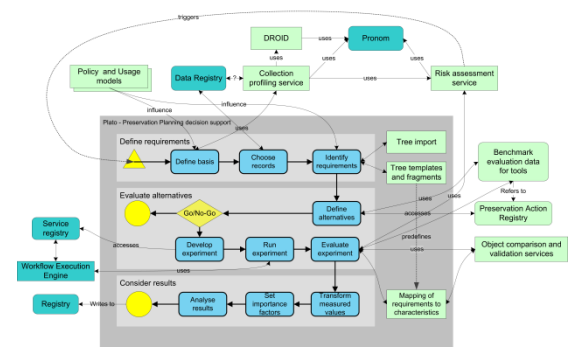
Focus	Name	Result
	▼ Object characteristic	PDF-A: 1,66  PDF-unchanged: 0,00  TIFF: 1,62  EPS: 1,62  JPEG2000: 1,60  RTF-acrobat: 0,00  RTF-convertdoc: 1,36  TXT: 0,00 
X	▶ Appearance	PDF-A: 1,50  PDF-unchanged: 1,50  TIFF: 1,50  EPS: 1,50  JPEG2000: 1,50  RTF-acrobat: 1,26  RTF-convertdoc: 1,19  TXT: 0,00 
X	▶ Structure	PDF-A: 1,27  PDF-unchanged: 1,27  TIFF: 1,27  EPS: 1,26  JPEG2000: 1,26  RTF-acrobat: 0,00  RTF-convertdoc: 1,18  TXT: 0,00 
X	▶ Content	PDF-A: 1,84  PDF-unchanged: 1,84  TIFF: 1,84  EPS: 1,84  JPEG2000: 1,84  RTF-acrobat: 0,00  RTF-convertdoc: 1,43  TXT: 0,00 
X	▶ Behaviour	PDF-A: 1,21  PDF-unchanged: 0,00  TIFF: 1,13  EPS: 1,14  JPEG2000: 1,09  RTF-acrobat: 1,18  RTF-convertdoc: 1,19  TXT: 1,19 

Current work

- ❑ Comparator engine
- ❑ Pluggable infrastructure for the automated evaluation of preservation actions
- ❑ Mapping requirements to characteristics
- ❑ Recommender systems
- ❑ Additional XCL definitions
- ❑ Case studies evaluating strategies
- ❑ Audit and Certification of Preservation planning activities
- ❑ Tool support: Service integration



Integrating Preservation Planning Decision support with Planets components
November 2007



Thank you very much for your attention.

becker@ifs.tuwien.ac.at
www.ifs.tuwien.ac.at/dp
www.planets-project.eu

