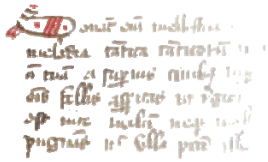


Historisch
Kulturwissenschaftliche
Informationsverarbeitung

Characterizing with a Goal in Mind: The XCL approach

Manfred Thaller, Universität zu Köln

Tools and Trends, The Hague, November 1st/2nd 2007

Why characterize?

Create technical metadata as required by organizational models for long term preservation.
Create a more abstract model of information.

Create an abstraction to achieve a specific purpose.

Why characterize?

How do we make sure, a digital object – image, text, multimedia – is the same, after it has been migrated into a new format?

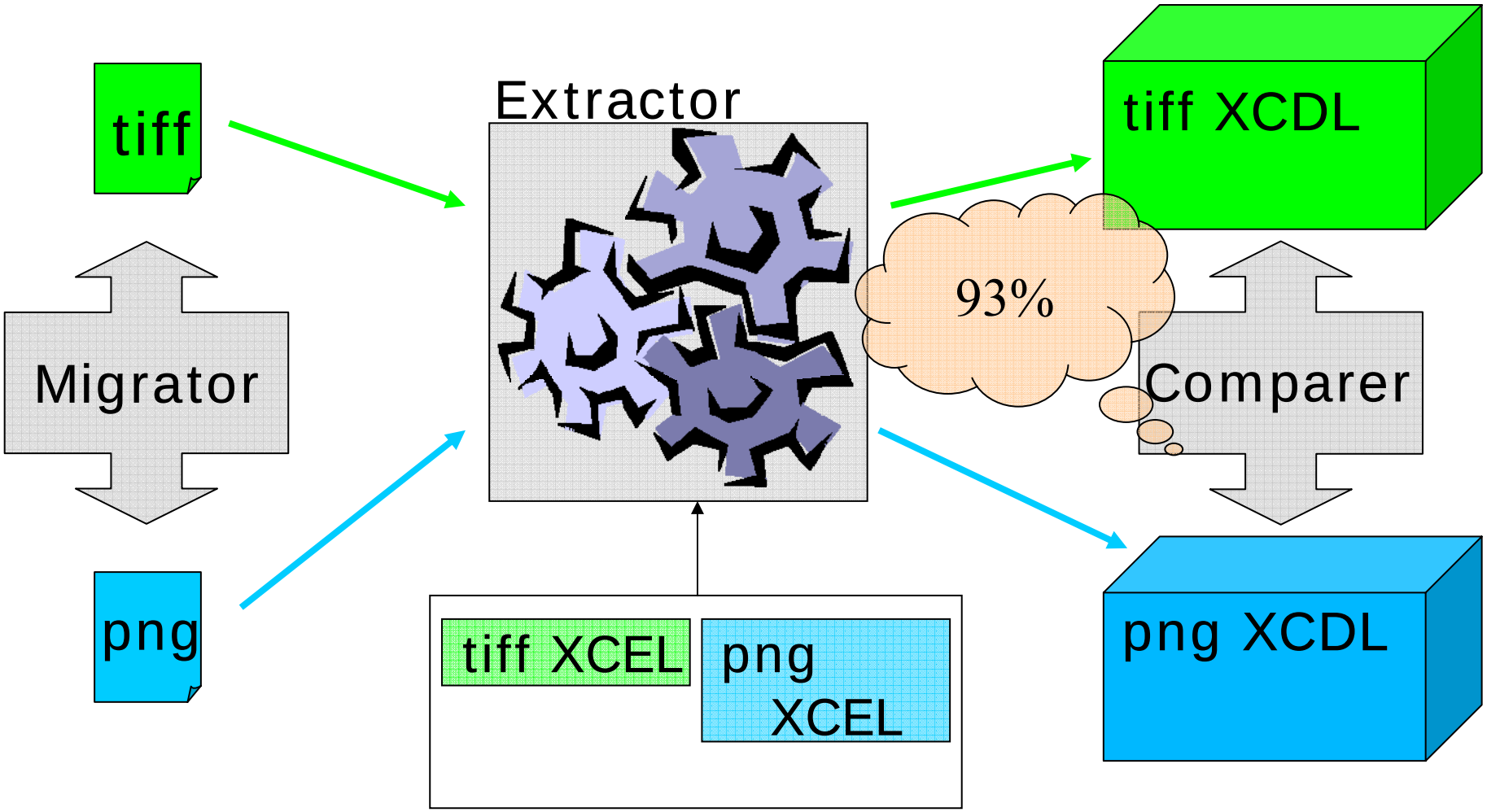
Why characterize?

How do we make sure, which of two copies of a digital object – image, text, multimedia – is the correct one, after one of them has suffered some damage?

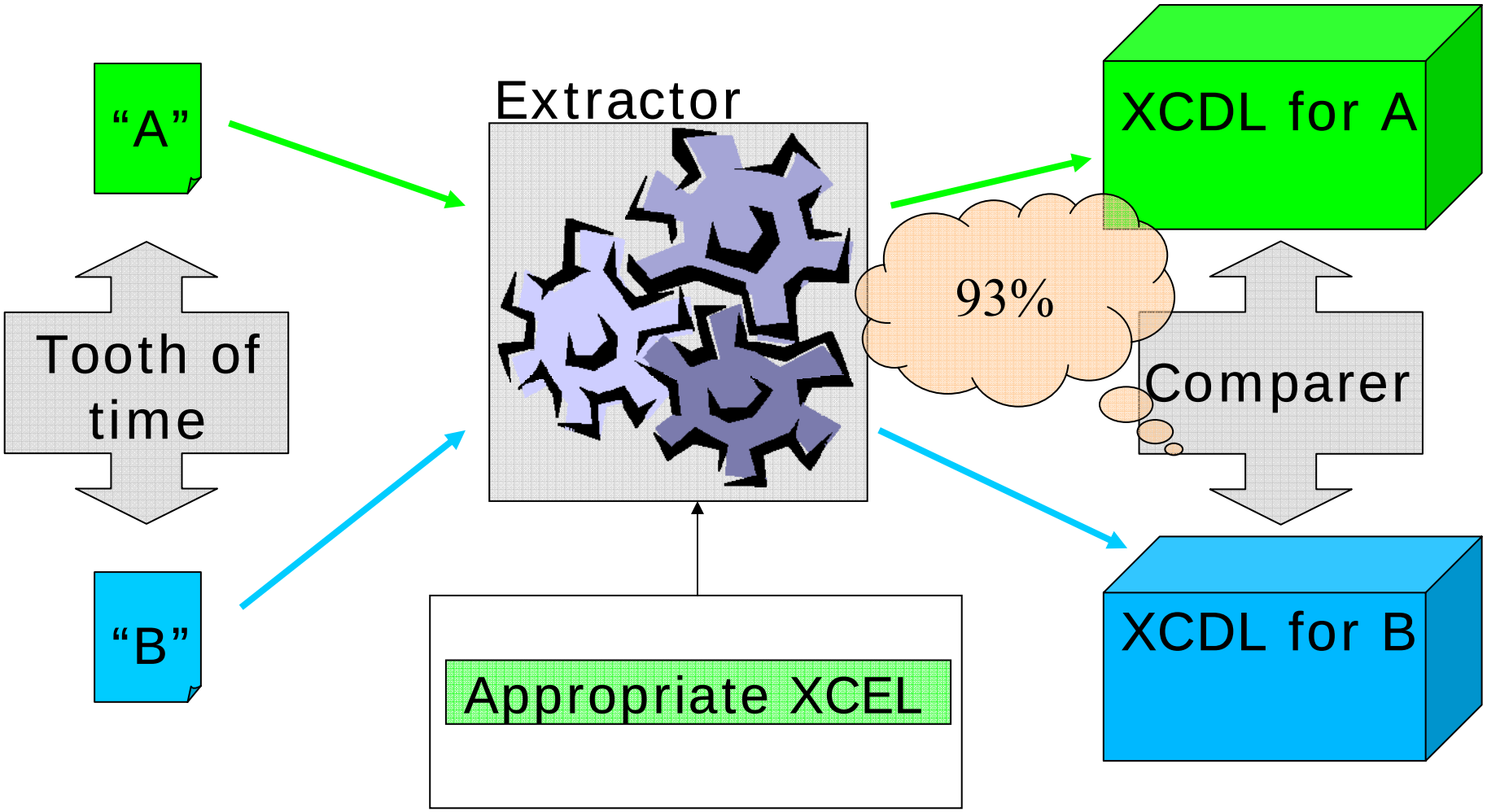
Why characterize?

How do we make sure, whether a specific software tool is able to handle a specific set of files?

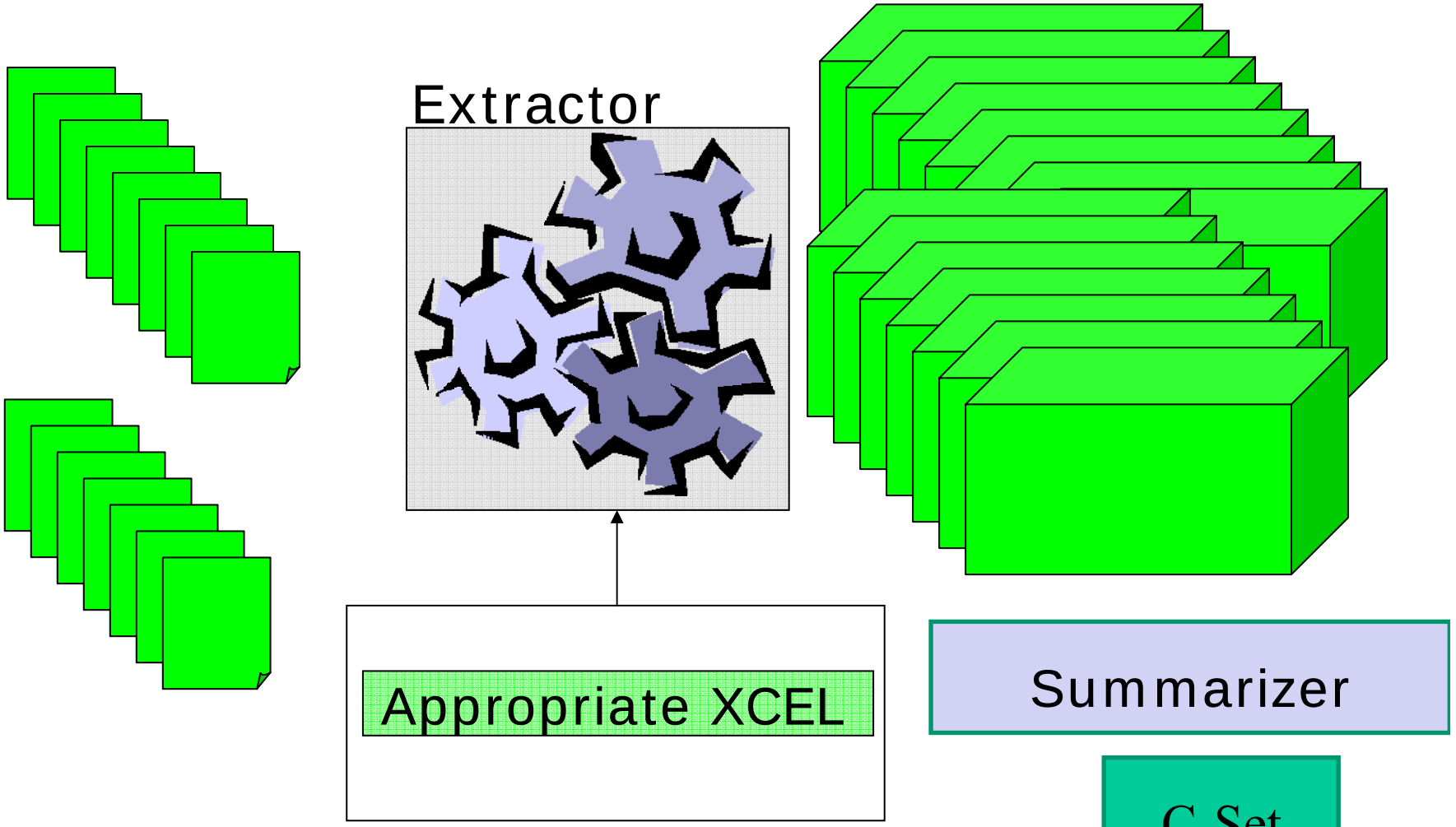
A vision I


tiff
Extractor
tiff XCDL
93%
Migrator
Comparer
png
tiff XCEL
png XCEL
png XCDL

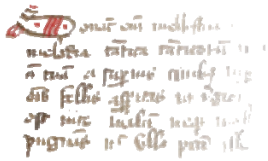
A vision II


"A"
Extractor
XCDL for A
93%
Tooth of time
Comparer
"B"
Appropriate XCEL
XCDL for B

A vision III



XCL approach



Historisch
Kulturwissenschaftliche
Informationsverarbeitung

Four building blocks:

(a)Make format specifications (traditionally directed at a human programmer) directly interpretable by generalized software.

Provide a “language” which allows to define file formats. (XCEL – eXtensible Characterisation *Extraction* Language)

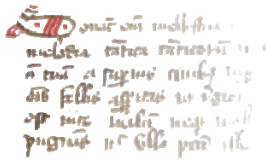
XCL approach

Historisch
Kulturwissenschaftliche
Informationsverarbeitung

„Extract, within a PDF, the value assigned to ‚documentAuthor‘ “

```
<processing type="pullXCEL„
  xcelRef="LiteralString">
    <processingMethod name="setName">
       <param value="documentAuthor"/>
    </processingMethod>
  </processing>
```

XCL approach



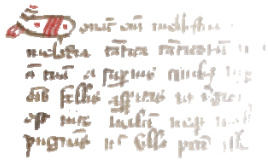
Historisch
Kulturwissenschaftliche
Informationsverarbeitung

XCEL designed to be able to allow the expression of *all* existing file formats.

4 years may be a bit short to translate all 16.000 of them …

… or even all of the approx. 2.600 pages of the PDF specification.

XCL approach



Historisch
Kulturwissenschaftliche
Informationsverarbeitung

Four building blocks:

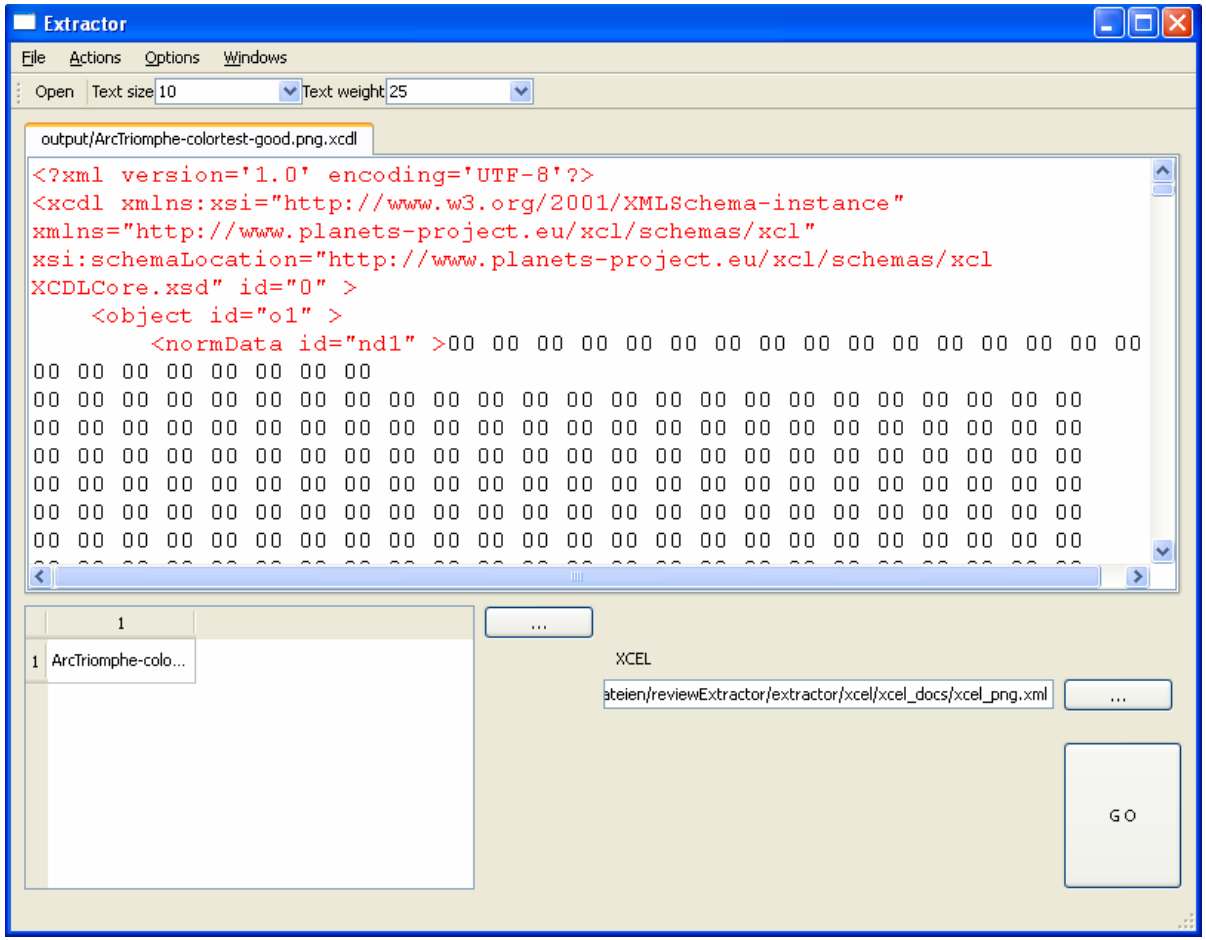
(b) Produce an “extractor” program, which uses such a specification to extract the data described by the format, expressed in XCEL, from a file.

XCL approach

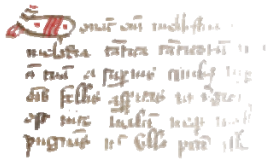
Historisch
Kulturwissenschaftliche
Informationsverarbeitung

XCL approach

Historisch
Kulturwissenschaftliche
Informationsverarbeitung



XCL approach



Historisch
Kulturwissenschaftliche
Informationsverarbeitung

Extractor designed to be useful in real life applications.

Bit of arithmetic:

1 million files, each processed within one second:

1,000,000 / 3600 = 277.7 hours = 11.5 days

XCL approach

Historisch
Kulturwissenschaftliche
Informationsverarbeitung

Four building blocks:

(c) Provide a generalized model of information contained within files.

Provide a language which expresses the content of a file. (XCDL – eXtensible Characterisation *Definition* Language)

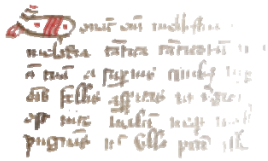
XCL approach

Historisch Kulturwissenschaftliche Informationsverarbeitung

XCDL is built upon abstract models (X schemas) of

- Image
- Text
- Sound
- 3D
- ...

Achievements: XCL



Historisch
Kulturwissenschaftliche
Informationsverarbeitung

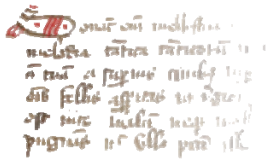
<XCELDocument...> ...

```
<formatDescription>....
<symbol identifier="ID01_I01_I01_S02" originalName="height"
        interpretation="uint32">
   <range><startposition xsi:type="sequential"> </startposition>
     <length xsi:type="fixed">4</length></range>
    <name>height</name>
</symbol>
<symbol identifier="ID01_I01_I01_S04"
        originalName="colourType">
  <range>
    <startposition xsi:type="sequential"> </startposition>
    <length xsi:type="fixed">1</length></range>
  <valueInterpretation>
       <valueLabel>greyscale</valueLabel>
       <value>0</value></valueinterpretation>
   <name>imageType</name>
</symbol>
<symbol identifier="ID01_I01_I01_S05"
        originalName="compressionMethod">
  <range>
     <startposition xsi:type="sequential"> </startposition>
      <length xsi:type="fixed">1</length></range>
  <valueInterpretation>
    <valueLabel>zlibDeflateInflate</valueLabel>
    <value>0</value></valueInterpretation>
  <name>compression</name>
</symbol>...
```

<xcdl>

```
<object id="o1" >
  <normData id="nd1" > ... </normData>
  <property id="p1" source="raw" cat="descr" >
    <name>compression</name>
    <valueSet id="i_i1_s6" >
      <rawValue>0 </rawValue>
      <labValue>...</labValue>
      <dataRef ind="normAll" />
      <propRel/>
    </valueSet>
  </property>
  <property id="p2" source="raw" cat="descr" >
    <name>height</name>
    <valueSet id="i_i1_s3" >
      <rawValue>0 0 1 ad </rawValue>
      <labValue>
        <val>429</val>
        <type>uint32</type>
      </labValue>
      <dataRef ind="normAll" />
      <propRel/>
    </valueSet>
  </property>
  <property id="p3" source="raw" cat="descr" >
    <name>imageType</name>
.....
```

XCL approach



Historisch
Kulturwissenschaftliche
Informationsverarbeitung

- XCDL provides abstract language to represent (potentially) *full* content of file.
- “characteristics” ➔ “format independent representation”.
- “extraction = interpretation”; execute, e.g., decompression, palette lookup etc.

XCL approach

Is the compression used within a file a characteristic of the file?

For a librarian probably “no” ...

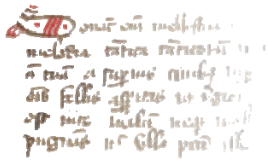
... for an archivist possibly “yes”.

XCL approach

But why do we extract the actual *data*?

"Characteristics" are supposed to be akin to metadata?

XCL approach



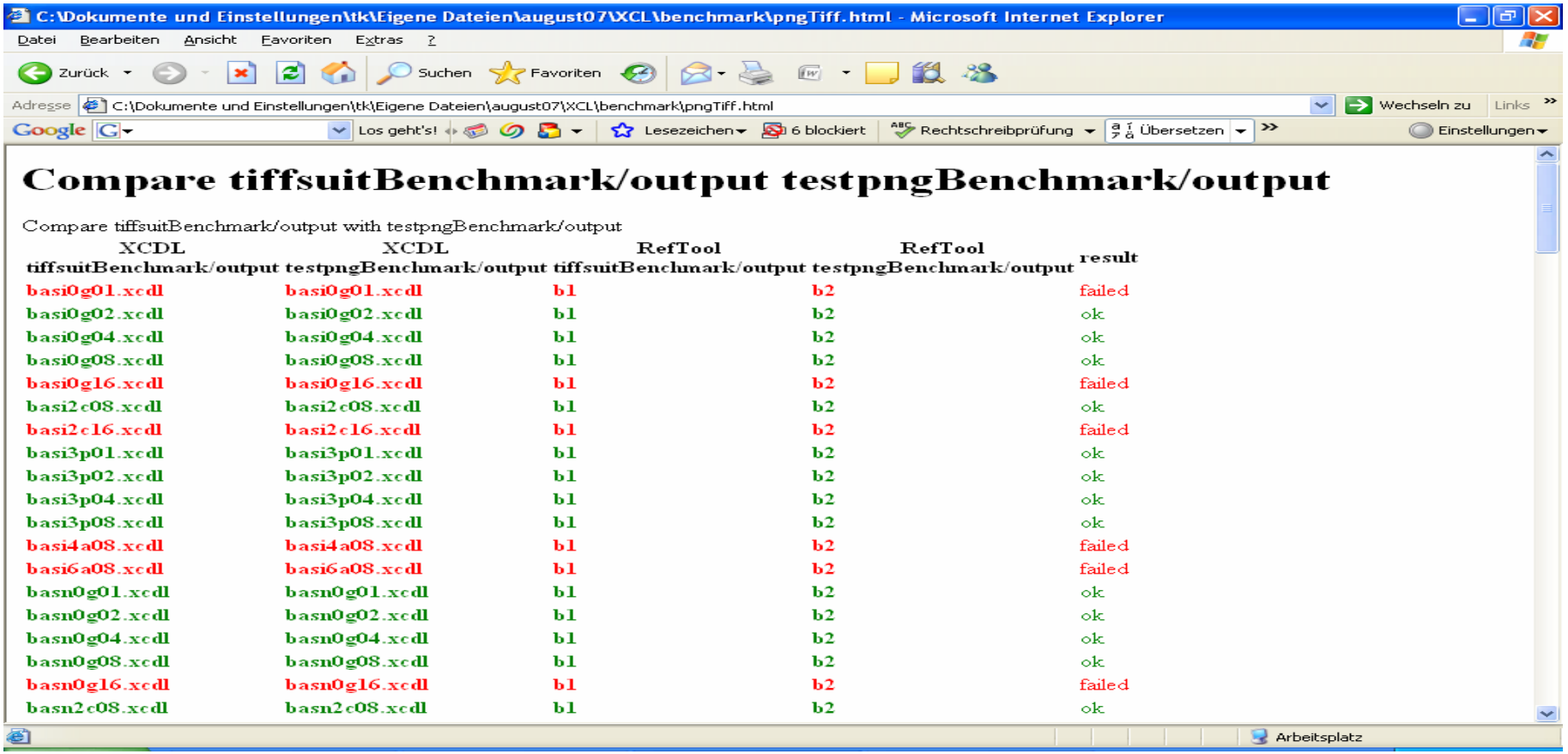
Historisch
Kulturwissenschaftliche
Informationsverarbeitung

Four building blocks:

(d) A software “comparator” able to make a meaningful numerical estimate whether two files contain the same information.

XCL approach

Historisch
Kulturwissenschaftliche
Informationsverarbeitung



Compare tiffsuitBenchmark/output testpngBenchmark/output

Compare tiffsuitBenchmark/output with testpngBenchmark/output

XCDL tiffsuitBenchmark/output	XCDL testpngBenchmark/output	RefTool tiffsuitBenchmark/output	RefTool testpngBenchmark/output	result
basi0g01.xcdl	basi0g01.xcdl	b1	b2	failed
basi0g02.xcdl	basi0g02.xcdl	b1	b2	ok
basi0g04.xcdl	basi0g04.xcdl	b1	b2	ok
basi0g08.xcdl	basi0g08.xcdl	b1	b2	ok
basi0g16.xcdl	basi0g16.xcdl	b1	b2	failed
basi2c08.xcdl	basi2c08.xcdl	b1	b2	ok
basi2c16.xcdl	basi2c16.xcdl	b1	b2	failed
basi3p01.xcdl	basi3p01.xcdl	b1	b2	ok
basi3p02.xcdl	basi3p02.xcdl	b1	b2	ok
basi3p04.xcdl	basi3p04.xcdl	b1	b2	ok
basi3p08.xcdl	basi3p08.xcdl	b1	b2	ok
basi4a08.xcdl	basi4a08.xcdl	b1	b2	failed
basi6a08.xcdl	basi6a08.xcdl	b1	b2	failed
basn0g01.xcdl	basn0g01.xcdl	b1	b2	ok
basn0g02.xcdl	basn0g02.xcdl	b1	b2	ok
basn0g04.xcdl	basn0g04.xcdl	b1	b2	ok
basn0g08.xcdl	basn0g08.xcdl	b1	b2	ok
basn0g16.xcdl	basn0g16.xcdl	b1	b2	failed
basn2c08.xcdl	basn2c08.xcdl	b1	b2	ok

XCL approach

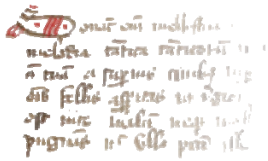
Historisch
Kulturwissenschaftliche
Informationsverarbeitung


► Photoshop ►



► Photoshop ►


Squaring circles? - I



Historisch
Kulturwissenschaftliche
Informationsverarbeitung

1. Just about everything in a file, including the "data", may be needed to evaluate its status.
2. A "not-storage-optimized" format, however, will make explode the storage space needed by at least one order of magnitude.
3. So, the most useful representation for long term storage, is the least useful for practical handling.

Squaring circles? - II

Historisch
Kulturwissenschaftliche
Informationsverarbeitung

3. If we save the file specifications in a way, however, that lets general purpose "extractors" apply them to old byte streams ...
4. ... we arrive at "just-in-time-characterisation-extraction".

What is a model of information?

+ ● = ● ● ●

+ ● ● = ● ● ● ●

● + ●● = 

you *do* understand Maya numerals – as you have an abstract concept of numbers, irrespective of their representation.

What is a model of information?

erefore, even a couple of hundred years later, you know,
that the following is bad arithmetic:

●● + ●●●● = ● 

What is a model of information?

erefore, even a couple of hundred years later, you know, that the following is bad arithmetic:

●● + ●●●● = 

en if you might not have known that the correct expression reads:

●● + ●●●● = 

XCDL: image model (1)

A pixel cube …
Each pixel:
MSB (channel 1), … LSB (channel 1),
…
MSB (channel n), … LSB (channel n),
MSB (aux 1), … LSB (aux 1),
…
MSB (aux m), … LSB (aux m)

XCDL: image model (2)

A pixel cube …

Accompanied by *rendering info* plus *deployment info* plus *historical info.*

XCDL: image model - example

```
<property id="p4" source="raw" cat="descr" >
        <name>imageType</name>
        <valueSet id="i_i1_s5" >
          <rawValue>2</rawValue>
          <labValue>
            <val>truecolour</val>
            <type>fixedLabel</type>
          </labValue>
          <dataRef ind="normAll" />
          <propRel/>
        </valueSet>
      </property>
```

XCDL: text model (1)

A text (= <object>) is composed of
- data (= <normData>) plus
- interpretations of data according
to the underlying format specification
(= <property>).

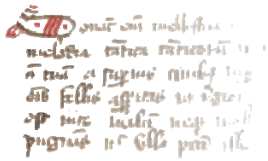
XCDL: text model (2)

Or, one level of abstraction higher, a text is composed of content carrying tokens, accompanied by *rendering info* plus *deployment info* plus *historical info*.

XCDL: text model - example

This is a text

```
<refData id="1">54 68 69 73 20 69 73 20 61 20 74 65 78 74</refData>
…
<property>
<name>fontsize</name>
<rawVal>
<val>48</val>
<type>unsignedInt8</type>
</rawVal>
<dataRef> <!-- property refers to discrete part of reference data-->
<ref id="1" start="0" end="3"/>
<ref id="1" start="10" end="12"/>
</dataRef>
</property>
```



Historisch
Kulturwissenschaftliche
Informationsverarbeitung

XCDL: text model - example

This is a text

```
<refData id="1">54 68 69 73 20 69 73 20 61 20 74 65 78 74</refData>
…
<property>
<name>fontsize</name>
<rawVal>
<val>48</val>
<type>unsignedInt8</type>
</rawVal>
<dataRef> <!-- property refers to discrete part of reference data-->
<ref id="1" start="0" end="3"/>
<ref id="1" start="10" end="12"/>
</dataRef>
</property>
```

XCDL: text model - example

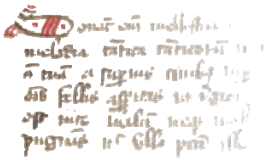
This is a text

<refData id="1">**54 68 69 73** 20 69 73 20 61 20 **74 65 78 74**</refData>
…
<property>
<name>fontsize</name>
<rawVal>
<val>48</val>
<type>unsignedInt8</type>
</rawVal>
<dataRef> <!-- property refers to discrete part of reference data-->
<ref id="1" start="0" end="3"/>
<ref id="1" start="10" end="12"/>
</dataRef>
</property>

Relationship between “file format” and “information found” in a file?

For XCL a file format is a hint at how to understand a file, but:

(i)Reality is never wrong.
(ii) People make mistakes.
➔
(a)“Partial parsing.”
(b)“Effective sub-versioning.”



Historisch
Kulturwissenschaftliche
Informationsverarbeitung

Motto

Look at the stars, but keep your feet solidly on* the ground.

**In* the ground, in case it is muddy.

Thank you!