wePreserve

Historisch
Kulturwissenschaftliche
Informationsverarbeitung

Significant characteristics in Planets

Manfred Thaller
Universität zu* Köln

***University *at* not *of* Cologne**

Information Society

digital preservation Europe

planets

CASPAR

What are “significant characteristics”?

Those properties of a digital file which have to be known to enable the processing of the file within a specific setup.

Why extract them by software?

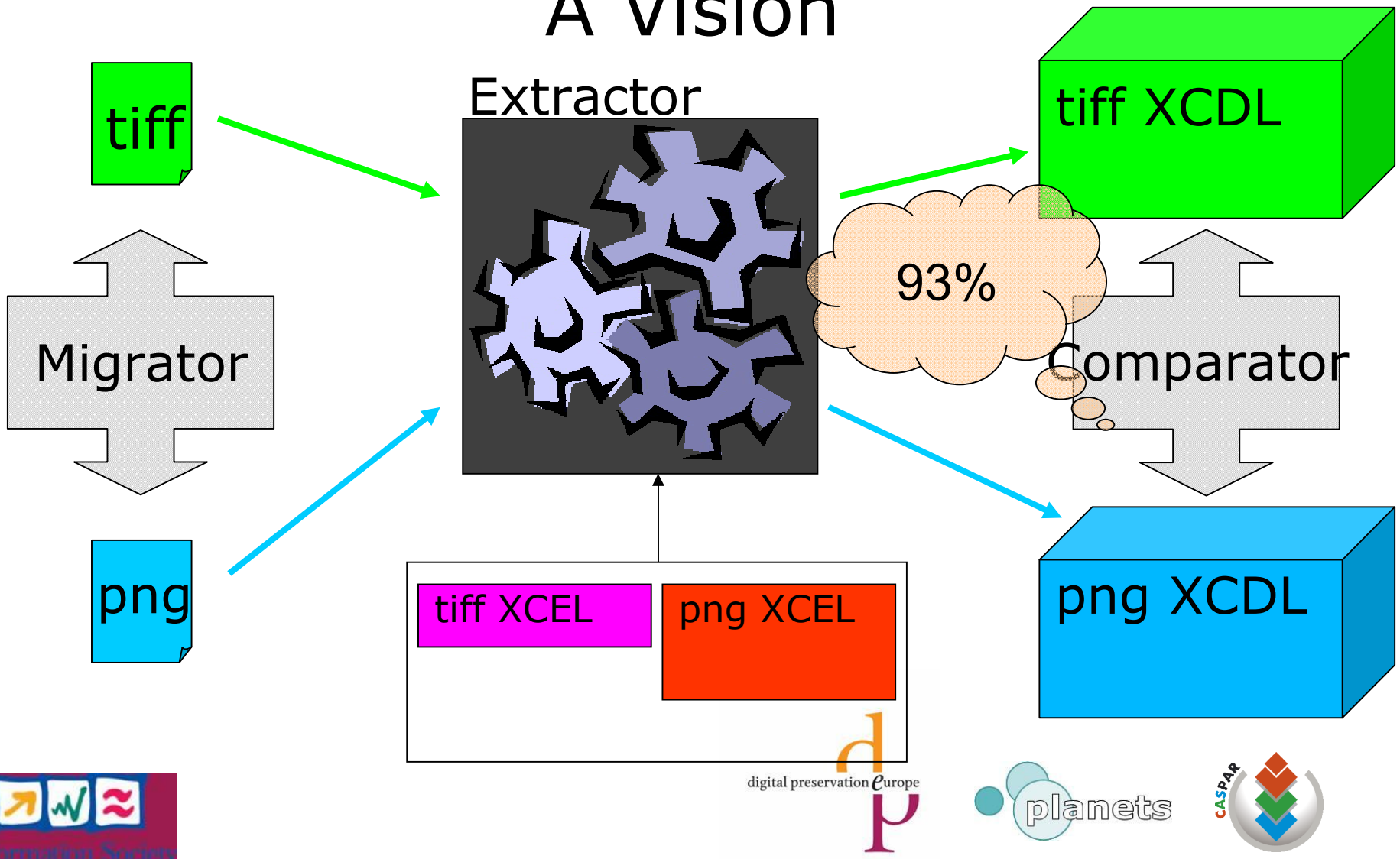
To create technical metadata as required by organizational models for long term preservation. (NLNZ)

Within Planets …

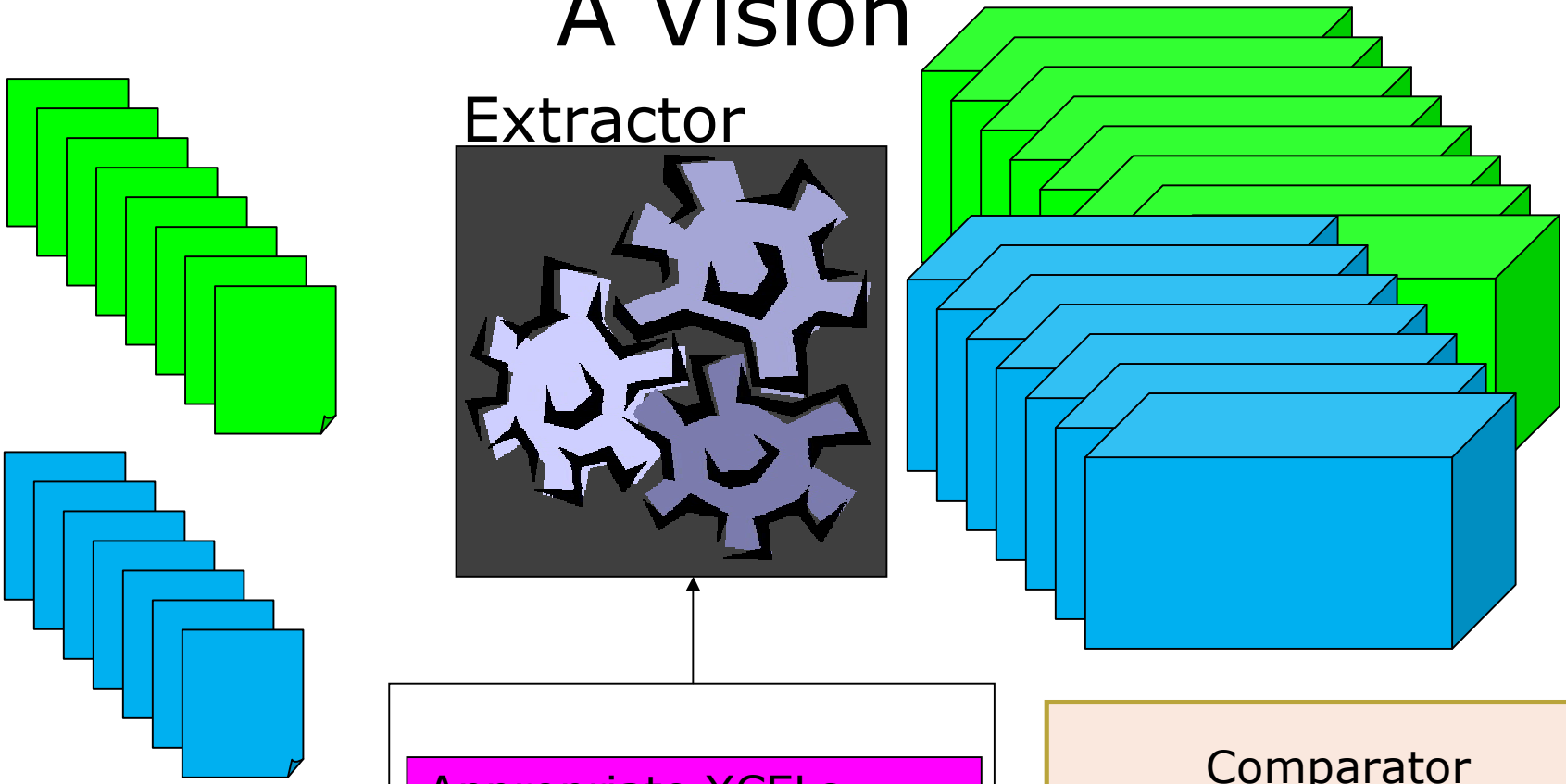
… served by solutions to *identify formats*:
formats registry / PRONOM / DROID.

… and a solution for extracting and processing such characteristics: XCL.

A Vision


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

A Vision



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

Why automate?

Assumption: Preservation is only feasible, if the content of two digital objects can be compared without human intervention, giving a numerical estimate of their degree of similarity.

wePreserve

Historisch
Kulturwissenschaftliche
Informationsverarbeitung

Demo

Information Society

digital preservation Europe

planets

CASPAR

Abstract solution I

(1) Language to represent the complete content of a digital object.
XCDL

(2) Language to describe any machine readable format in a formal language.
XCEL

(3) Software to extract the content of a file based upon a description as under (2) and express it in the language as specified under (1).
"extractor"

(4) Software to compare two such content descriptions.
"comparator"

<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>
```

```
<request2>
    <measurementRequest>
          <source name="XCDL1.xml"/>
          <target name="XCDL2.xml"/>
       <property id="45" name="rgbPalette">
           <metric id="10"
 name="hammingDistance"/>
 </property>
       <property id="300" name="normData">
           <metric id="10"
 name="hammingDistance"/>
           <metric id="50" name="RMSE"/>
       </property>
        <property id="2"
 name="imageHeight" unit="pixel">
 <metric id="200" name="equal"/>
 <metric id="201" name="intDiff"/>
 <metric id="210" name="percDev"/>
        </property>
       <property id="30" name="imageWidth"
 unit="pixel">
           <metric id="200" name="equal"/>
           <metric id="201" name="intDiff"/>
           <metric id="210" name="percDev"/>
       </property>
```

```
<property id="2"
   name="imageHeight"
   unit="pixel"
   compStatus="complete">
   <values type="int">
      <src>32</src>
      <tar>32</tar>
   </values>
   <metric id="200"
   name="equal" result="true"/>
   <metric id="201"
   name="intDiff" result="0"/>
   <metric id="210"
   name="percDev"
   result="0.000000"/>
</property>
```

Abstract solution I

(1) Language to represent the complete content of a digital object.
XCDL

(2) Language to describe any machine readable format in a formal language.
XCEL

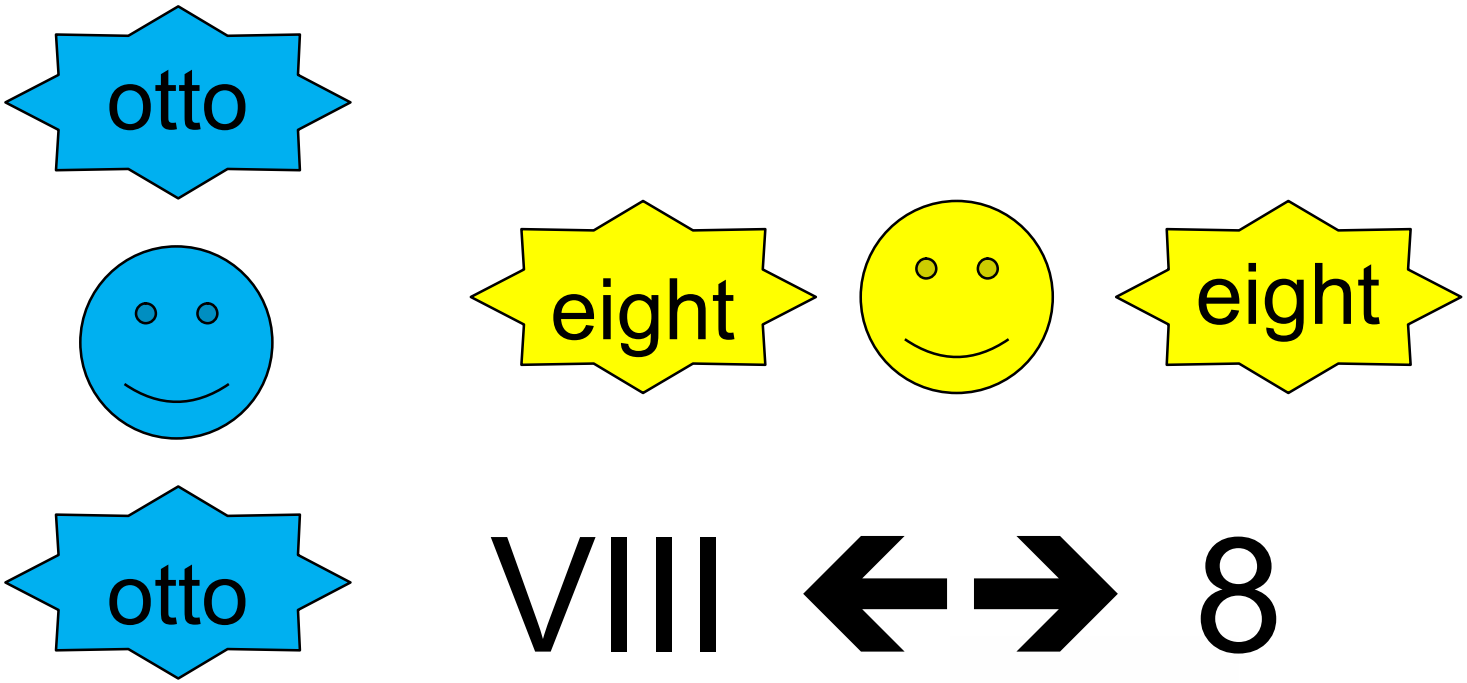
(3) Software to extract the content of a file based upon a description as under (2) and express it in the language as specified under (1).
"extractor"

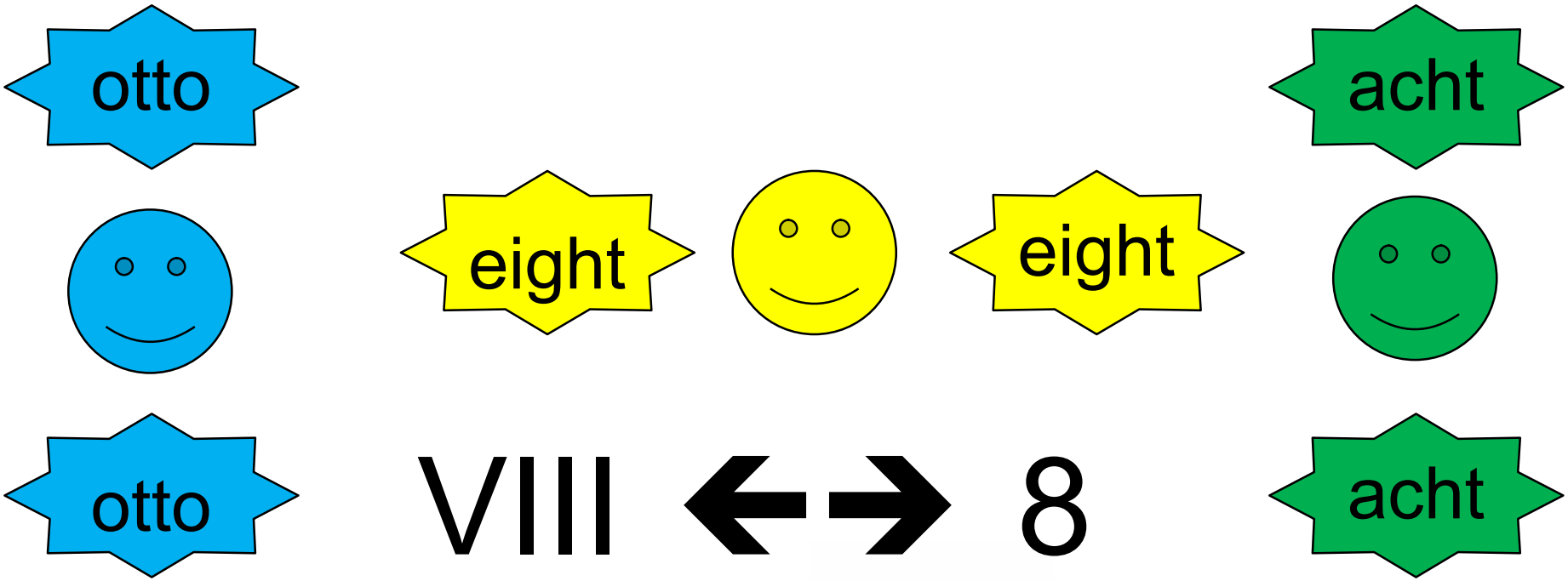
(4) Software to compare two such content descriptions.
"comparator"

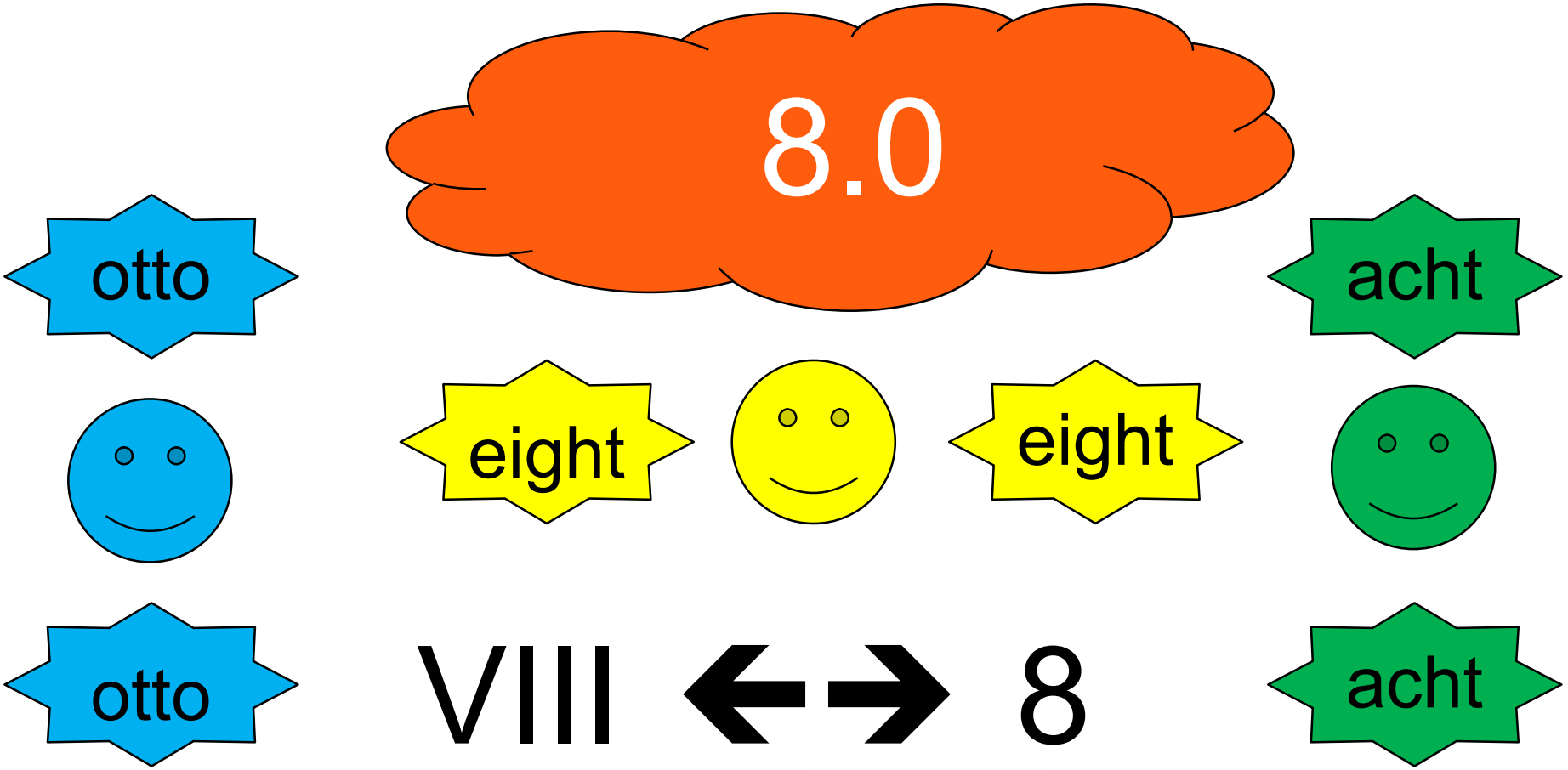
Are the following two items equal:

VIII ←→ 8

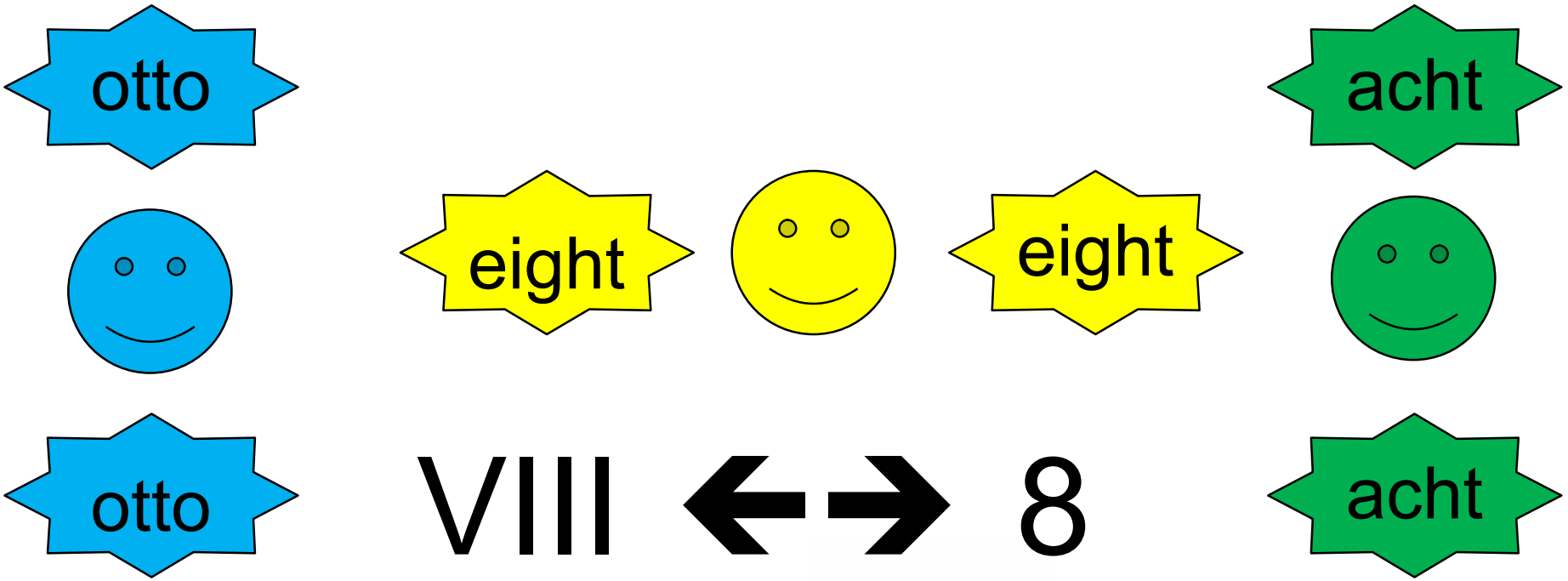
VIII ←→ 8


otto
eight
eight
otto
VIII ←→ 8


otto
acht
eight
eight
otto
VIII
8
acht


8.0
otto
acht
eight
eight
otto
VIII ←→ 8
acht

Information model: „an image“



information model: „an image“

format ontology: „what terms are used in formats to describe image properties“

VIII ←→ 8

Information model: „what is an image“

Format ontology: „what terms are used in formats to describe image properties“

Extraction language: “how to get the terms describing an image out of a file”

Abstract solution II

(1) A theoretical model of information (not: data) types – “image”, “text”, “audio” ...

(2) Ontologies, which map existing file format terminologies onto these model.

(3) A language – XCDL – which allows to express the content of files in different formats using the vocabulary of the ontologies and the “grammar” of the information model.

XCDL

eXtensible Characterisation Definition Language

Purpose: Describe the contents of a file in terms of an abstract model.

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.*

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>

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>

wePreserve

Historisch
Kulturwissenschaftliche
Informationsverarbeitung

Thank you!

Questions?

Manfred.thaller@uni-koeln.de

Information Society

digital preservation europe

planets

CASPAR