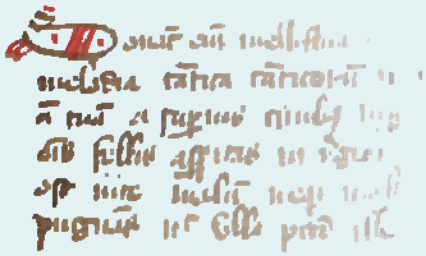


Historisch

Kulturwissenschaftliche

Informationsverarbeitung

Introduction to XCDL

Digital Preservation – The Planets Way
Copenhagen, 22 – 24 June 2009

Volker Heydegger

Overview

- XCDL in the overall context
- Application scenario
- Formalizing content: Basic concepts and elements

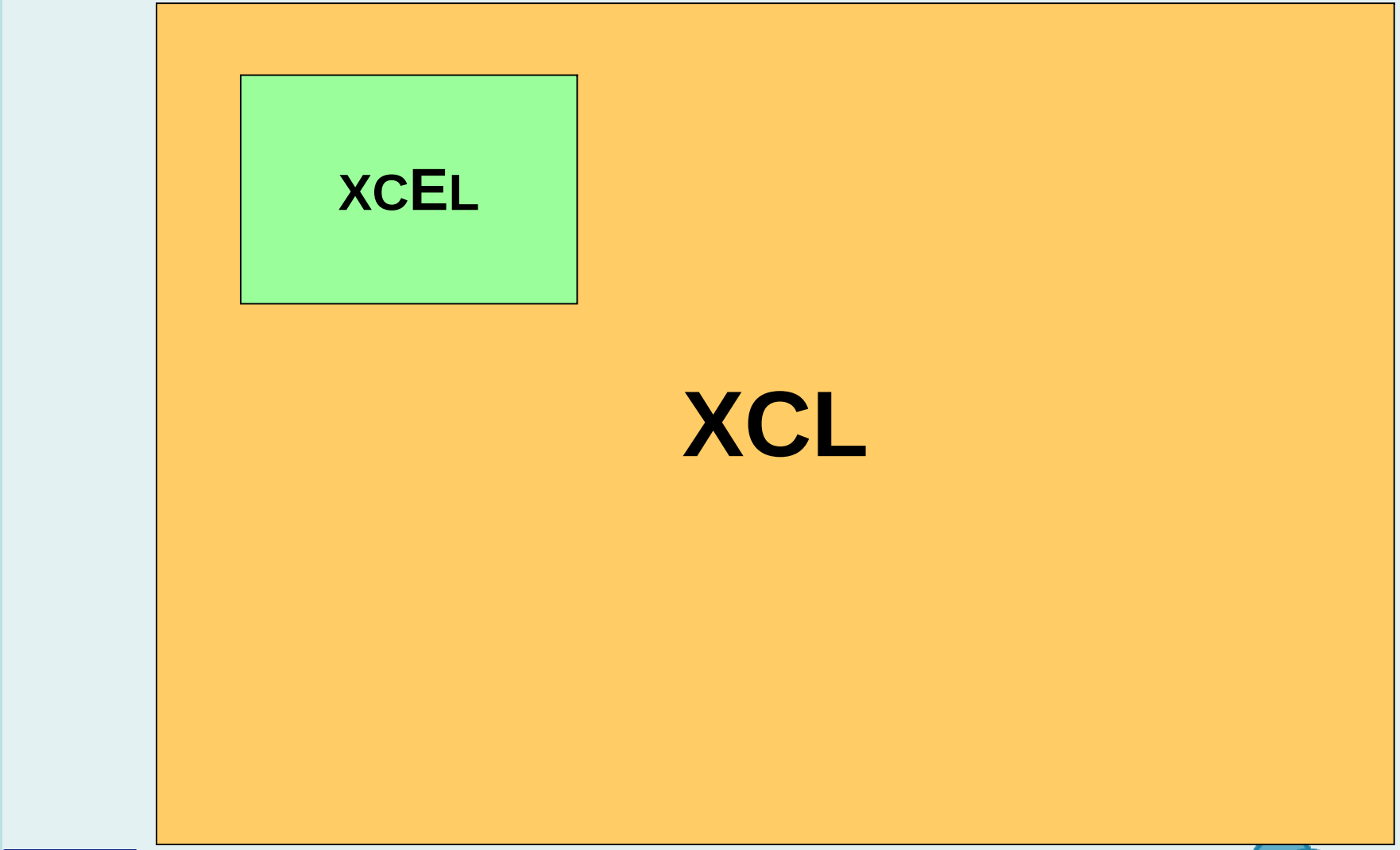
Background

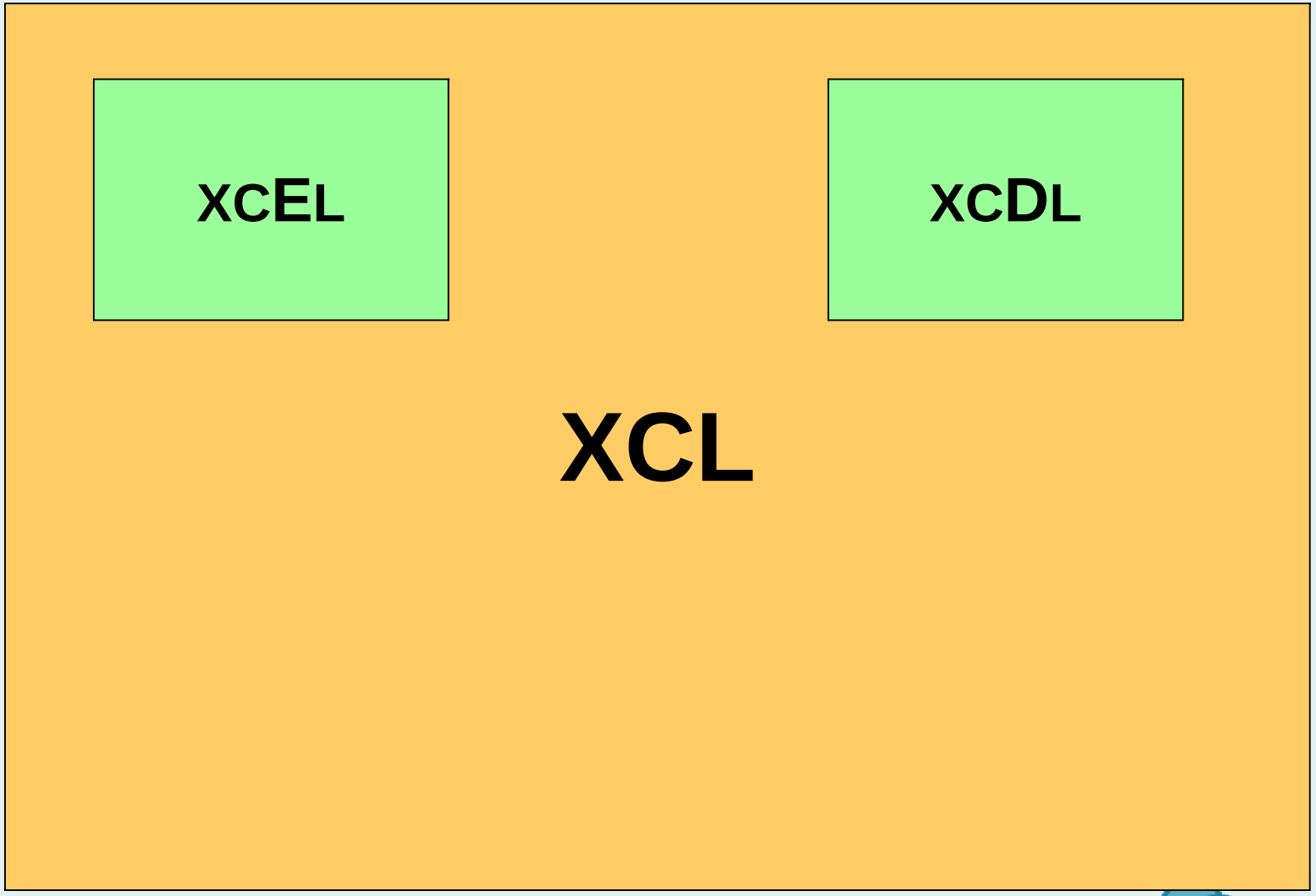
- Planets project: Preservation Characterisation sub-project (PC2, strategies development)
- Tasks:

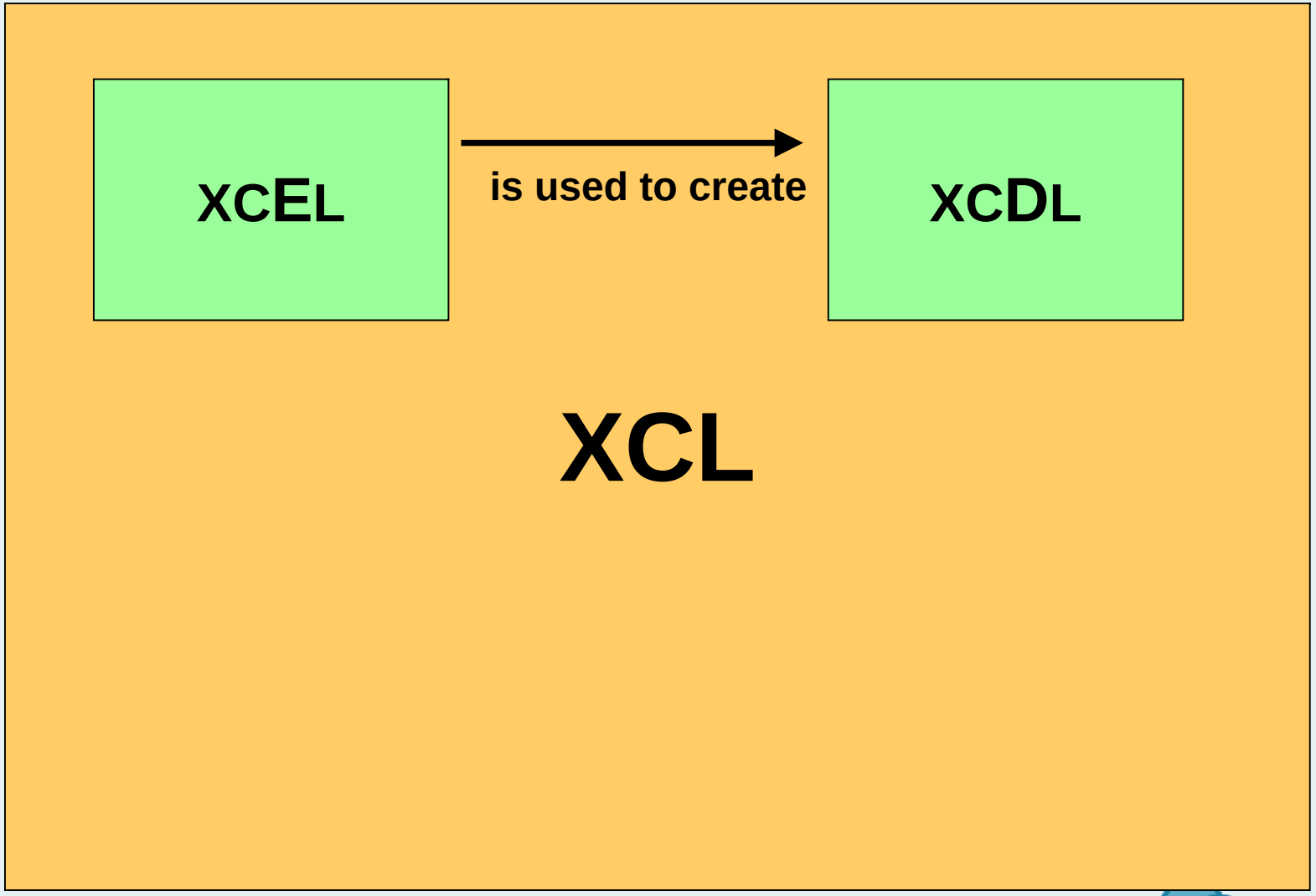
  - Develop an „eXtensible Characterisation Definition Language“ (XCDL), able to describe the *content* of digital objects (=1 + n more files).

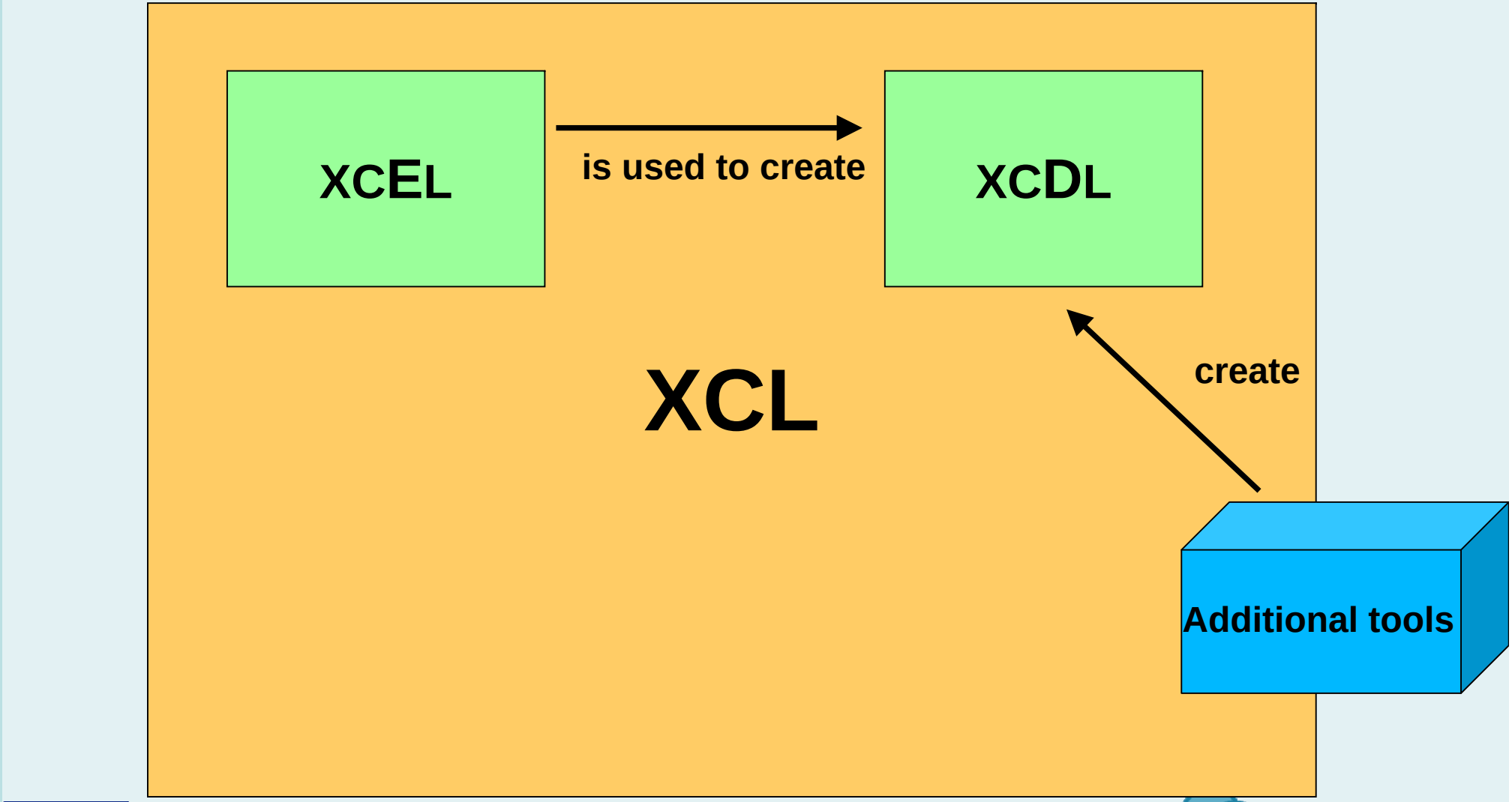
  - Develop an „eXtensible Characterisation Extraction Language“ (XCEL), able to describe any machine readable format in a formal language, processible by a software tool for extraction of content as XCDL.

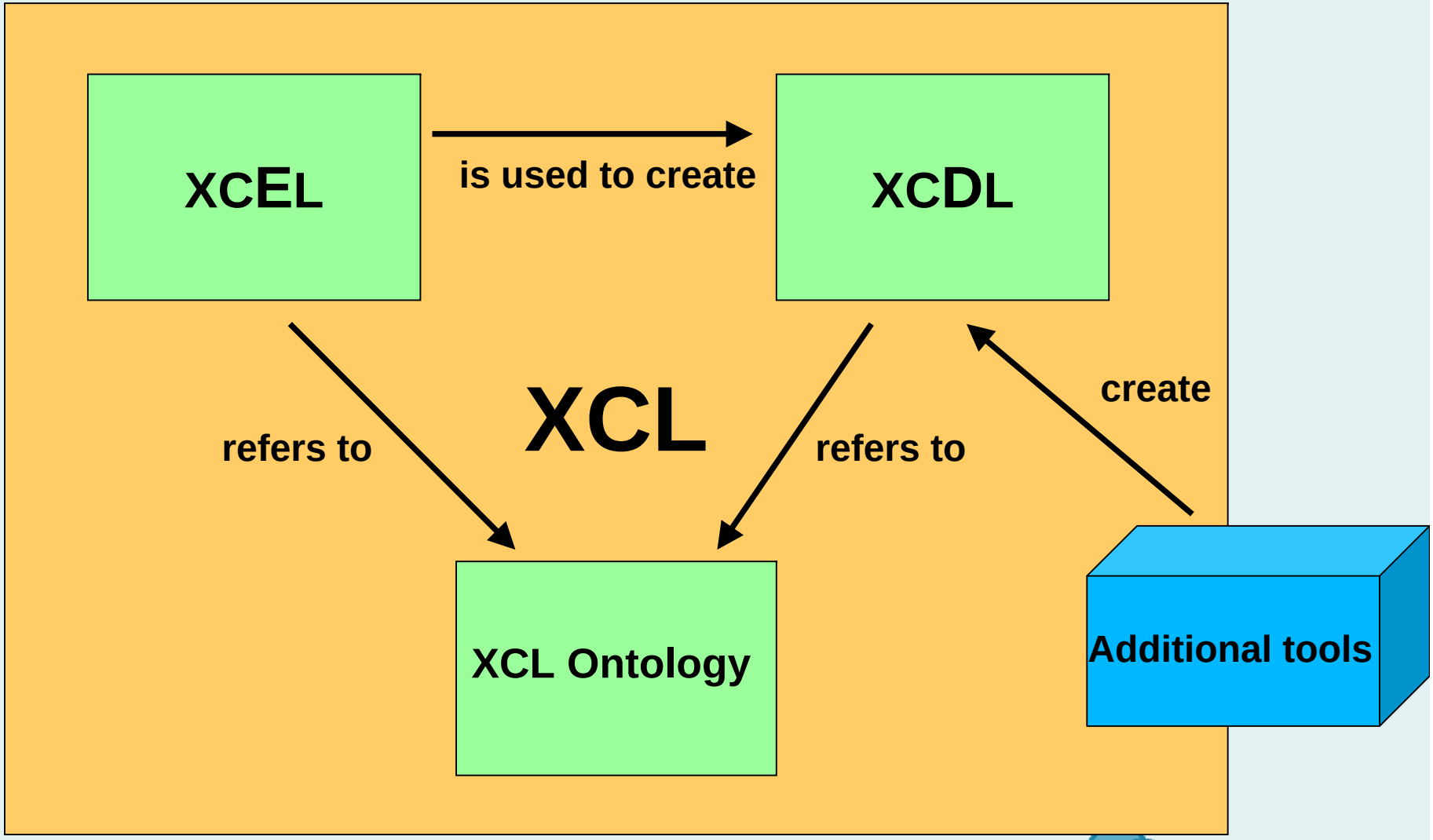
XCL overview


XCEL
XCL

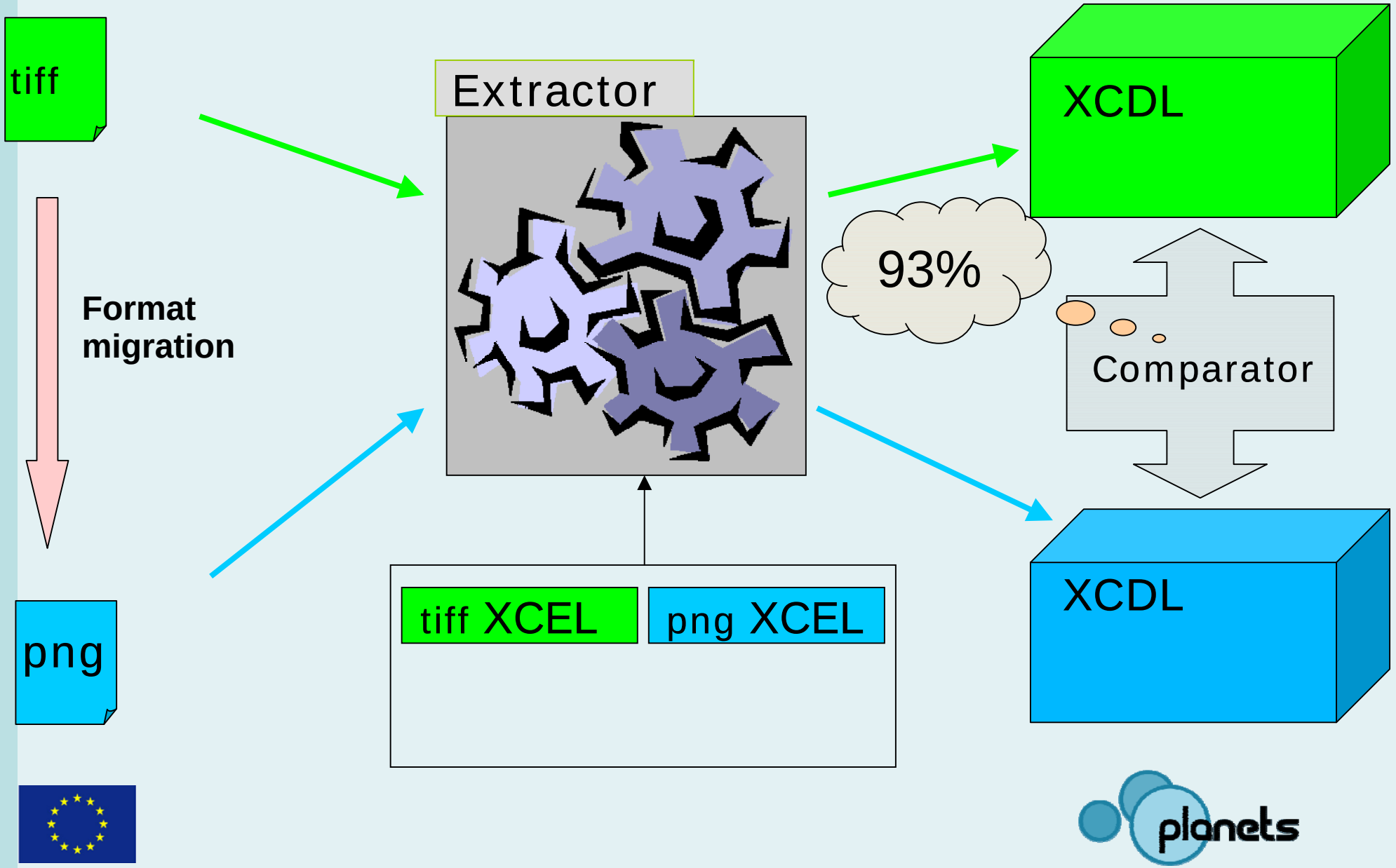

XCEL
XCDL
XCL

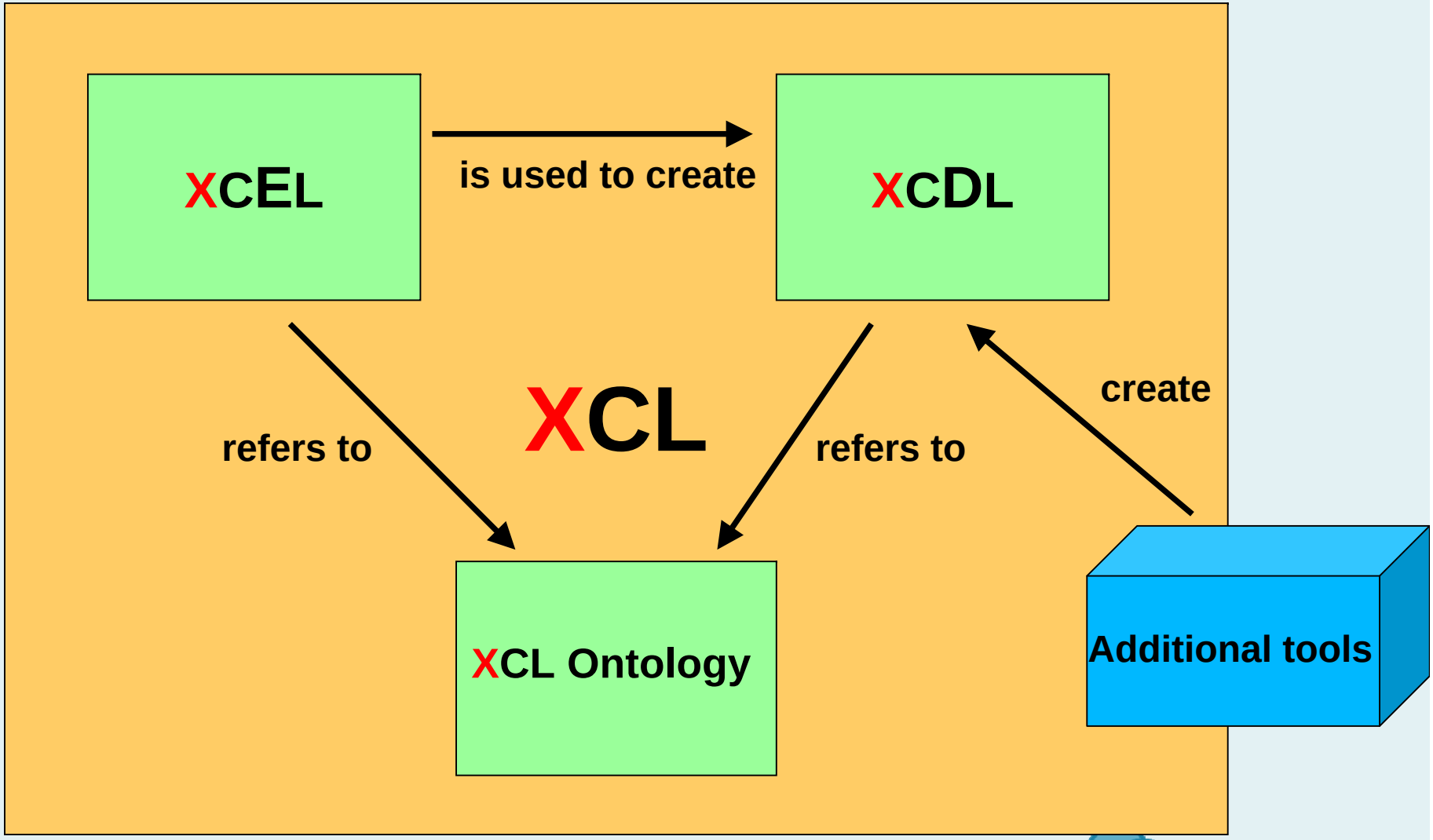

XCEL
is used to create
XCDL
XCL


XCEL
is used to create
XCDL
XCL
create
Additional tools


XCEL
is used to create
XCDL
XCL
refers to
refers to
create
XCL Ontology
Additional tools

Application scenario




XCEL
is used to create
XCDL
XCL
refers to
refers to
create
XCL Ontology
Additional tools

XCDL: Main purposes

In theory:

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

XCDL: Main purposes

In practice:

- Enable the comparison of file content in order to evaluate format migration.
- Not: Establishing a new format for archival.
- Nevertheless:

  Refinement of XCDL (and XCL as a whole) for other purposes besides Planets project can be considered

The XCDL Way of Formalizing File Content

Basic Concepts and Elements

How can we categorize file content in general?

We distinguish between:

- ‚Raw' information, e.g. pixel data of an image, characters in a text; it is additionally represented in a consistent form

  → NormData (= normalised data)

Element <normData>

- Normalised data is an abstraction of the specific format internal representation of content.
- Wraps the source data in a context-free representation (normalised to a standard representation)
- E.g., all byte sequences which appear in an encoded representation are decoded to the standard representation
- Conventions for standard representations: character data is represented as UTF-8 encodings, binary data as hexadecimals

```
• This is a text.

<data id="1">
{\rtf1\ansi\ansicpg1252\deff0\deflang1031{\fonttbl{\f0\fswiss\f
charset0 Arial;}}\viewkind4\uc1\pard\f0\fs20 \bullet  This is a
\b text\b0 .\par}
</data>

<normData id="1" type="text">
 • This is a text.
</normData>
```

Why normalisation?

Normalisation of raw information enables to compare data extracted from different formats!

How can we categorize file content in general?

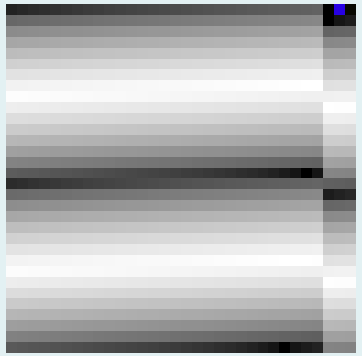
We distinguish between:

- ‚Raw' information, e.g. pixel data of an image, characters in a text; it is additionally represented in a consistent form

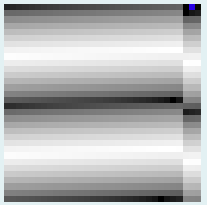
  → NormData (= normalised data)
- Description of raw information, e.g. how to process it generally (color depth) , how to handle it in a specific context (fonts to be used for rendering)

  → Properties

```
<?xml version='1.0' encoding='UTF-8'?>
<xcdl xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xmlns="http://www.planets-project.eu/xcl/schemas/xcl"
 xsi:schemaLocation="http://www.planets-project.eu/xcl/schemas/xcl
                     ../res/xcl/xcdl/XCDLCore.xsd" id="0" >
   <object id="o1" >
      <normData type="image" id="nd1">
      00 01 02 03 04 05 06 07 08 09 0a 0b 0c 0d 0e 0f 10 11 12 13 14 15 16
      ...
      </normData>
      <property id="p5">
         <name id="id30" >imageWidth</name>
         <valueSet id="i_i1_s4" >
            <labValue>
               <val>32</val>
               <type>int</type>
            </labValue>
         </valueSet>
       </property>
      ...
```



```
<?xml version='1.0' encoding='UTF-8'?>
<xcdl xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xmlns="http://www.planets-project.eu/xcl/schemas/xcl"
 xsi:schemaLocation="http://www.planets-project.eu/xcl/schemas/xcl
                     ../res/xcl/xcdl/XCDLCore.xsd" id="0" >
   <object id="o1" >
      <normData type="image" id="nd1">
      00 01 02 03 04 05 06 07 08 09 0a 0b 0c 0d 0e 0f 10 11 12 13 14 15 16
      ...
      </normData>
      <property id="p5">
          <name id="id30" >imageWidth</name>
          <valueSet id="i_i1_s4" >
             <labValue>
                <val>32</val>
                <type>int</type>
             </labValue>
          </valueSet>
       </property>
      ...
```



```
<?xml version='1.0' encoding='UTF-8'?>
<xcdl xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xmlns="http://www.planets-project.eu/xcl/schemas/xcl"
 xsi:schemaLocation="http://www.planets-project.eu/xcl/schemas/xcl
                     ../res/xcl/xcdl/XCDLCore.xsd" id="0" >
   <object id="o1" >
      <normData type="image" id="nd1">
      00 01 02 03 04 05 06 07 08 09 0a 0b 0c 0d 0e 0f 10 11 12 13 14 15 16
      ...
      </normData>
      <property id="p5">
         <name id="id30" >imageWidth</name>
         <valueSet id="i_i1_s4" >
            <labValue>
               <val>32</val>
               <type>int</type>
            </labValue>
         </valueSet>
       </property>
      ...
```

```
<property id=“p5" source="raw">
       <name id=“id30”>imageWidth</name>
...
</property>
```

The corresponding entry in ‘XCLImageNamesLib.xsd:

```
<xs:enumeration value="imageHeight">
   <xs:annotation>
      <xs:documentation>
Width of an image measured in  pixel. Corresponds to
the vertical dimension of an image (x- axis).
      </xs:documentation>
   </xs:annotation>
</xs:enumeration>
```

```
<property id=“p5" source="raw">
       <name id=“id30”>imageWidth</name>
...
</property>
```

The corresponding entry in ‘XCLImageNamesLib.xsd:

```
<xs:enumeration value="imageHeight">
   <xs:annotation>
      <xs:documentation>
Height of an image measured in  pixel. Corresponds to
the vertical dimension of an image (x- axis).[TIFF 6.0,
PNG 1.2]
      </xs:documentation>
   </xs:annotation>
</xs:enumeration>
```

TIFF: imageWidth

„ The number of columns in the image, i.e., the number of pixels per scanline. ...“

PNG: width

„Width and height give the image dimensions in pixels. ..“

```
<?xml version='1.0' encoding='UTF-8'?>
<xcdl xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xmlns="http://www.planets-project.eu/xcl/schemas/xcl"
 xsi:schemaLocation="http://www.planets-project.eu/xcl/schemas/xcl
                     ../res/xcl/xcdl/XCDLCore.xsd" id="0" >
   <object id="o1" >
      <normData type="image" id="nd1">
      00 01 02 03 04 05 06 07 08 09 0a 0b 0c 0d 0e 0f 10 11 12 13 14 15 16
      ...
      </normData>
      <property id="p5">
          <name id="id30" >imageWidth</name>
          <valueSet id="i_i1_s4" >
             <labValue>
                <val>32</val>
                <type>int</type>
             </labValue>
          </valueSet>
       </property>
      ...
```

- Labelled value: Interpretation of the “raw value” according to predefined types
- Two elements:
  - <val>: picks up the labelled value
  - <type>: indicates the (data)type of the value
- Types are defined in XCL schema “XCLBasicDataTypesLib.xsd”

What else do we need to formalize file content?

```
...
<object id="o1" >
      <normData type=„text" id="nd1">
          This is a text.
      </normData>
      <property id="p5">
          <name id="id39" >fontSize</name>
          <valueSet id="i_i1_s4" >
              <labValue>
                  <val>12</val>
                  <type>int</type>
              </labValue>
          </valueSet>
      </property>
      ...
</object>

<object id="o2" >   ...
</object>
```

Recursive content: The „footnote problem“

This is only a short1 text.

1: short is another term for "not long".

```
…
<object id="o1">
      <normData id="nd1" type="text">This is only a short1 text.</normData>
      <property id="p1">
            <name id="id170">footnote</name>
            <valueSet id="vs1">
                  <objectRef>.:02</objectRef>
            </valueSet>
      </property >
</object>
<object id="o2">
      <normData id="nd2" type="text">1: short is another term for "not long".
      </normData>
      <property> …
</object>
…
```

What about properties that relate to a specific part of data ?

This is a sentence with *a few italic* words.

```
<object id="o1">
    <normData id="nd1" type="text">This is a sentence with a few italic words.
    </normData>
    <property id="p1">
        <name id="id162">italic</name>
        <valueSet id="vs1">
            <labValue>
                <val>default</val>
                <type>XCLLabel</type>
            </labValue>
            <dataRef ind=”normSpecific” propertySetId=”ps1” />
        </valueSet>
    </property>
    <propertySet id="ps1">
        <valueSetRelations>
            <ref valueSetId=”vs1” />
        </valueSetRelations>
        <dataRef>
            <ref begin=”24” end=”35” id=”nd1” />
        </dataRef>
    </propertySet>
</object>
```

This is a sentence with *a few italic* words.

```
<object id="o1">
    <normData id="nd1" type="text">This is a sentence with a few italic words.
    </normData>
    <property id="p1">
        <name id="id162">italic</name>
        <valueSet id="vs1">
            <labValue>
                <val>default</val>
                <type>XCLLabel</type>
            </labValue>
            <dataRef ind=”normSpecific” propertySetId=”ps1” />
        </valueSet>
    </property>
    <propertySet id="ps1">
        <valueSetRelations>
            <ref valueSetId=”vs1” />
        </valueSetRelations>
        <dataRef>
            <ref begin=”24” end=”35” id=”nd1” />
        </dataRef>
    </propertySet>
</object>
```

This is a sentence with *a few italic* words.

```
<object id="o1">
    <normData id="nd1" type="text">This is a sentence with a few italic words.
    </normData>
    <property id="p1">
        <name id="id162">italic</name>
        <valueSet id="vs1">
            <labValue>
                <val>default</val>
                <type>XCLLabel</type>
            </labValue>
            <dataRef ind=”normSpecific” propertySetId=”ps1” />
        </valueSet>
    </property>
    <propertySet id="ps1">
        <valueSetRelations>
            <ref valueSetId=”vs1” />
        </valueSetRelations>
        <dataRef>
            <ref begin=”25” end=”34” id=”nd1” />
        </dataRef>
    </propertySet>
</object>
```

What about properties that relate to a specific part of data ?

This is a sentence with ***italic and bold*** words.

```
<object id="o1">
    <normData id="nd1" type="text">This is a sentence with italic and bold words.
    </normData>
    <property id="p1">
        <name id="id162">italic</name>
        <valueSet id="vs1">
            …
            <dataRef ind=”normSpecific” propertySetId=”ps1” />
        </valueSet>
    </property>
    <property id="p2">
        <name id="id162">italic</name>
        <valueSet id="vs2">
            …
            <dataRef ind=”normSpecific” propertySetId=”ps1” />
        </valueSet>
    </property>
    <propertySet id="ps1">
        <valueSetRelations>
            <ref valueSetId=”vs1” />
            <ref valueSetId=”vs2” />
        </valueSetRelations>
        <dataRef>
            <ref begin=”25” end=”39” id=”nd1” />
        </dataRef>
    </propertySet>
</object>
```

Thank you!

Any Questions?