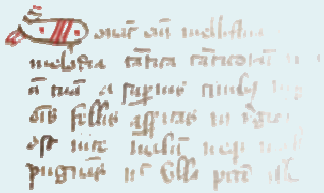


Historisch

Kulturwissenschaftliche

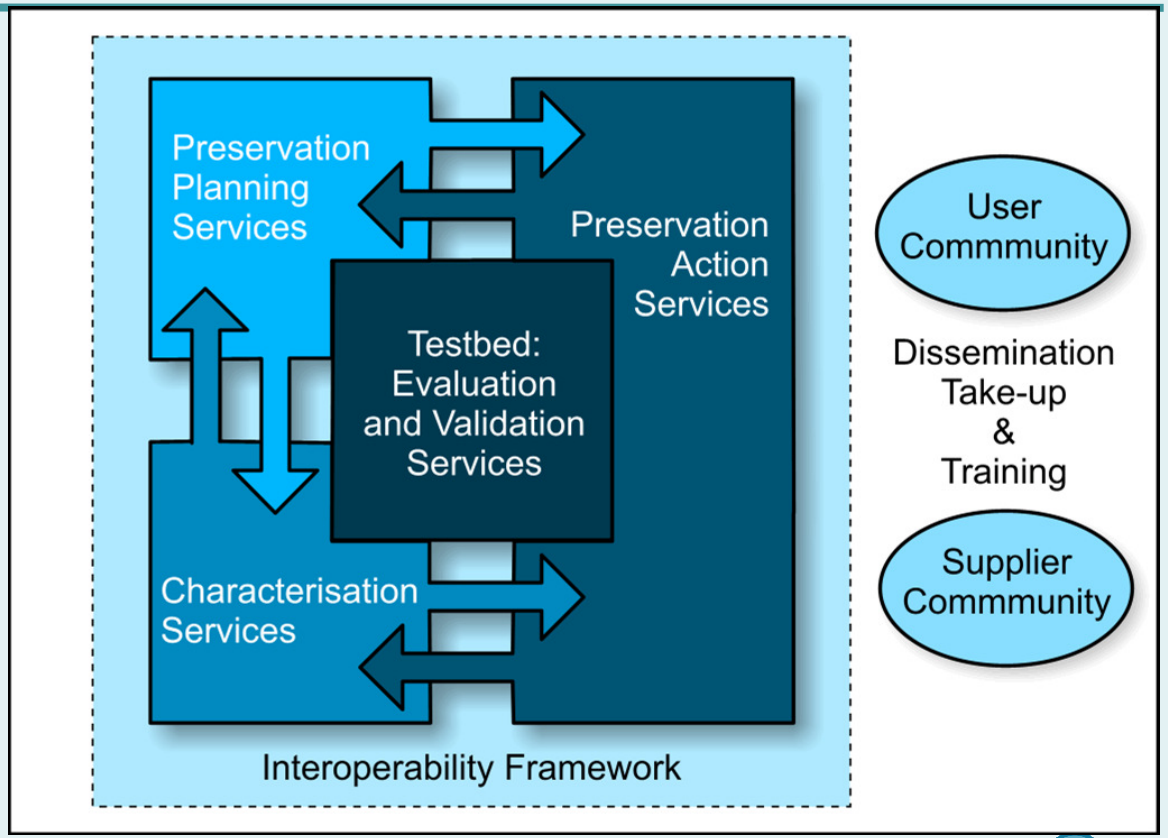
Informationsverarbeitung

Tools: How to understand files.

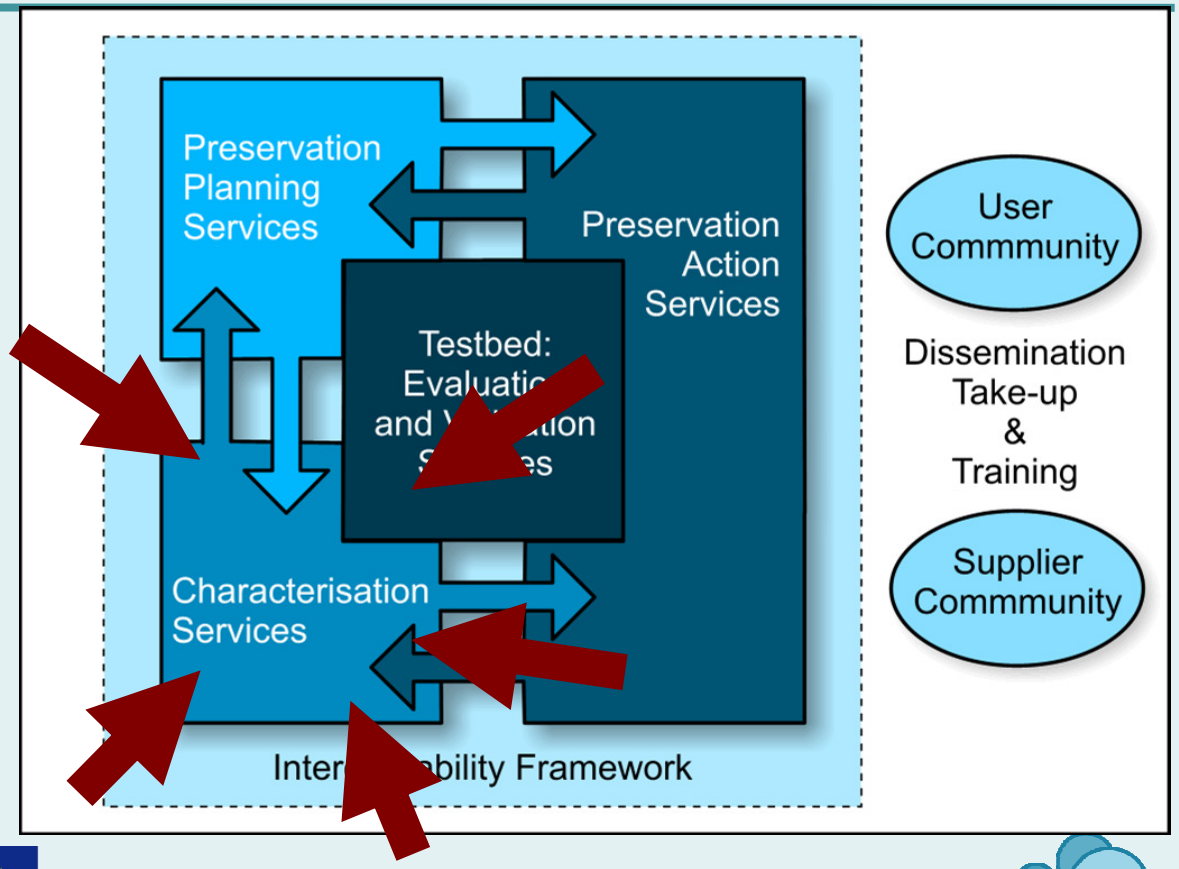
planetarium.hki.uni-koeln.de

Manfred Thaller, February 9th, London

Zoom In


Preservation Planning Services
Preservation Action Services
Testbed: Evaluation and Validation Services
Characterisation Services
Interoperability Framework
User Commmunity
Dissemination Take-up & Training
Supplier Commmunity

Zoom In

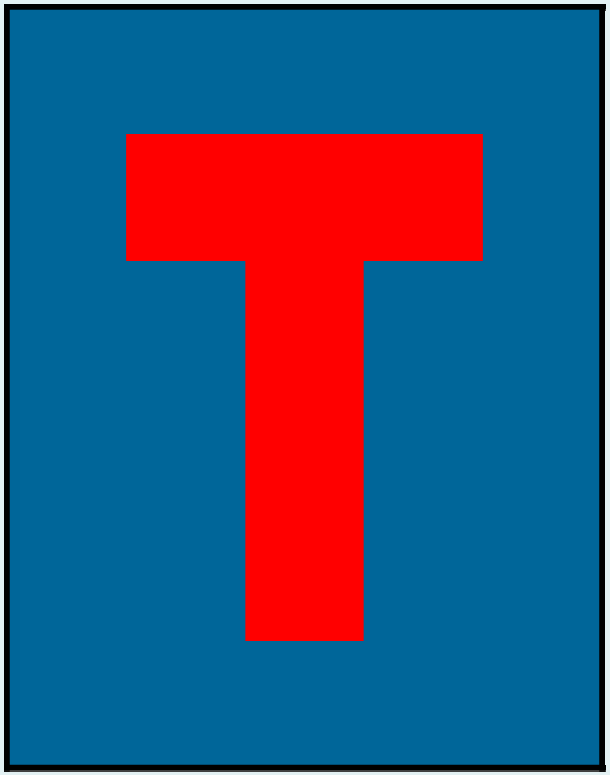

Preservation Planning Services
Preservation Action Services
Testbed: Evaluatic and ion
es
Characterisation Services
Inter bility Framework
User Commmunity
Dissemination Take-up & Training
Supplier Commmunity

Prologue

1. What is in a file?

An image

6 rows,
5 columns

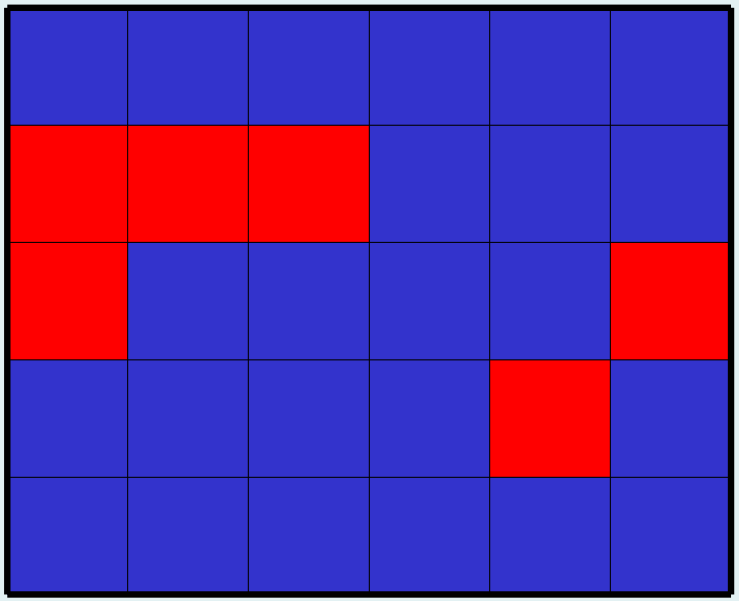


An image

6,5,1,1,1,1,1,1, 0,0,0,1,1,1,0,1,1,1,1,0,1,1,1,1,0,1,1,1,1,1,1,1

An image

5 rows,
6 columns



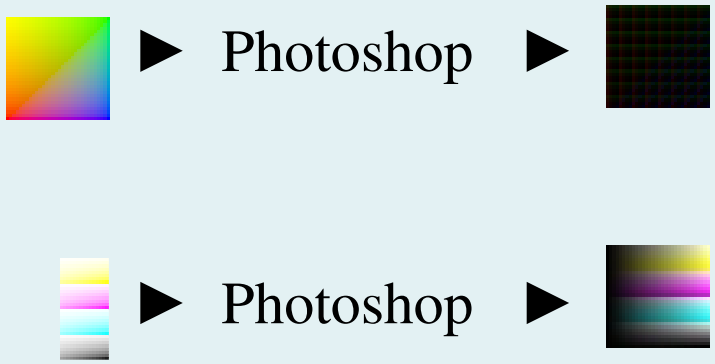
An image

6,5,1,1,1,1,1,1,1, 0,0,0,1,1,1,0,1,1,1,1,0,1,1,1,1,0,1,1,1,1,1,1,1,1

Characteristics of a file: *What properties does a specific file have?*

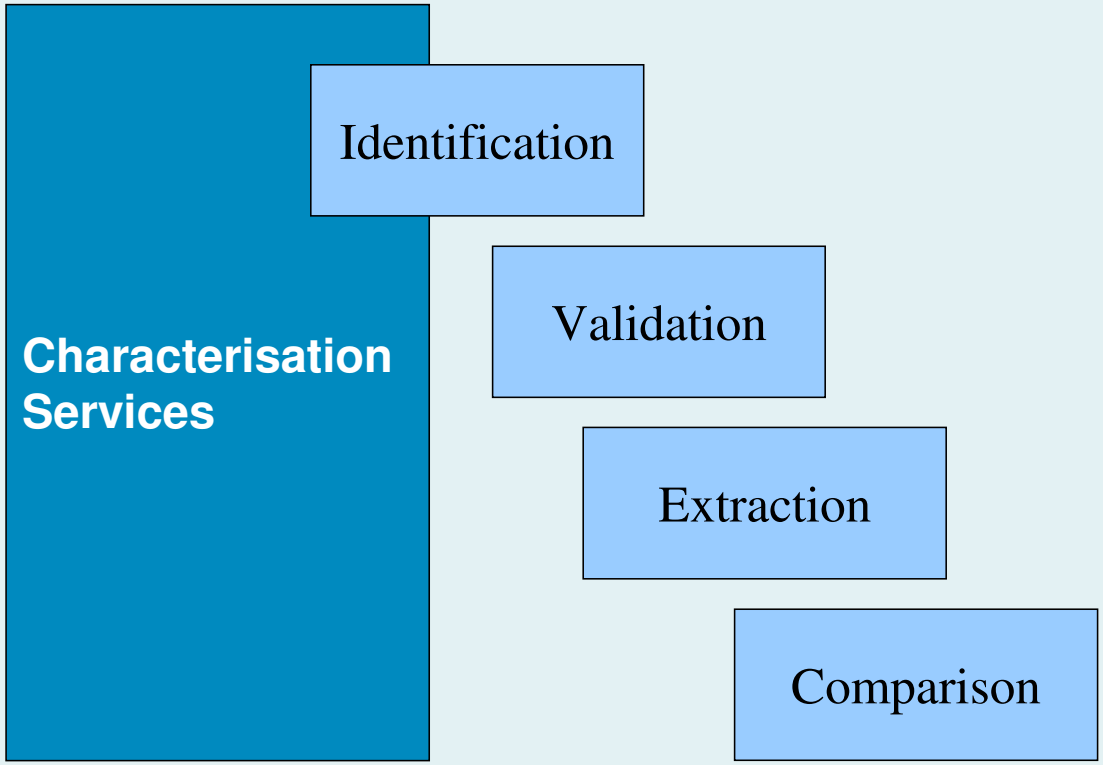
Format of a file: *What set of rules is used to encode them?*

Reality ...

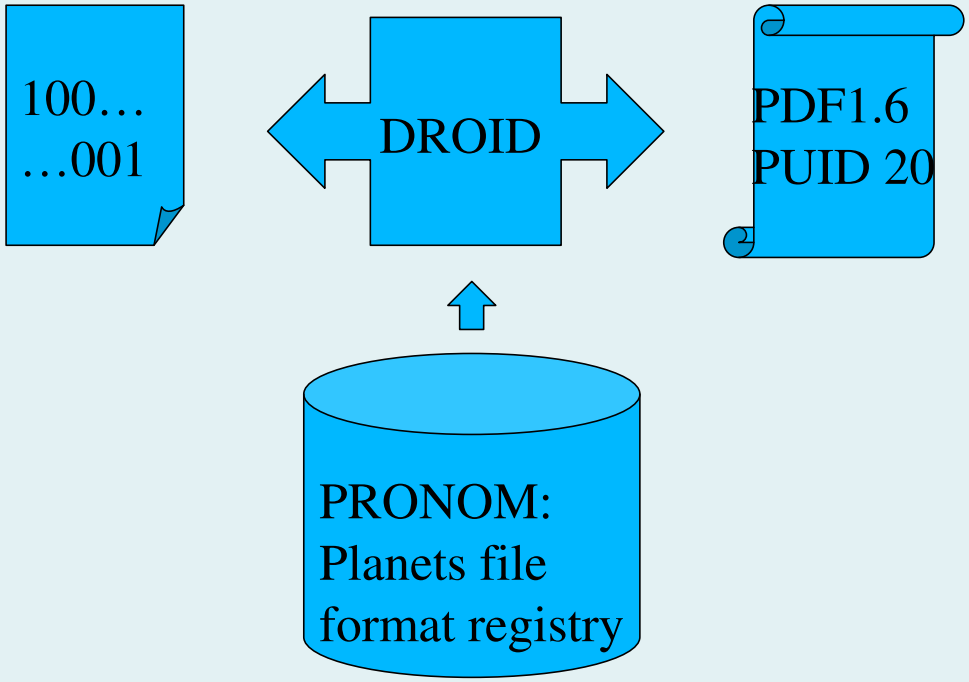


2. How can Planets help?

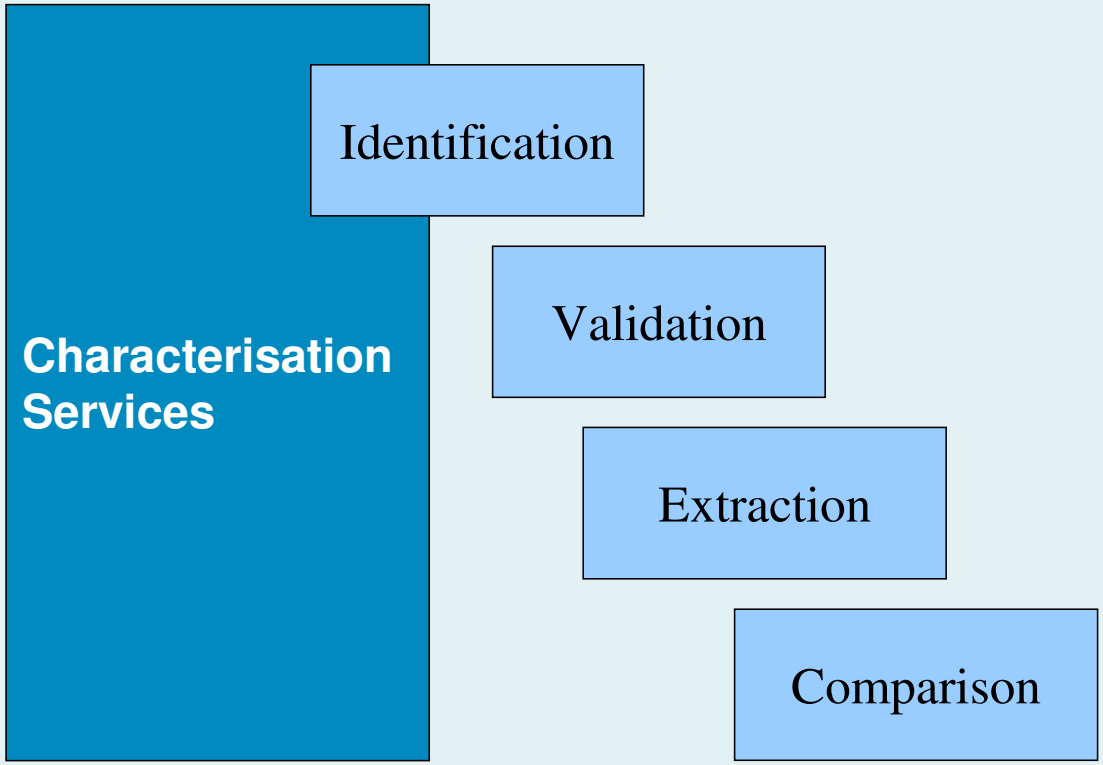
File related services



Identification: What kind of file is it?



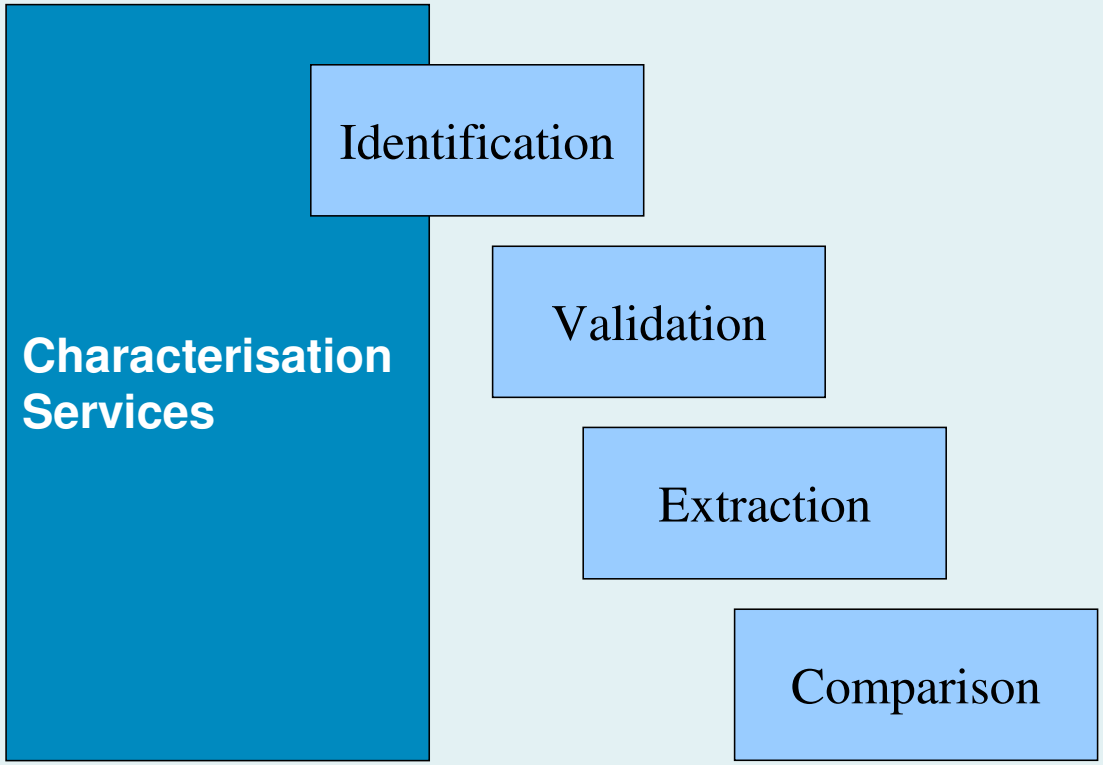
File related services



Validation: Will we be able to process it in twenty years time?

- No „Planets borne“ validation services …
- … but validation tools like JHOVE available via the interoperability framework.
- CAVEAT 1: Validation is a surprisingly complex topic.
- CAVEAT 2: Reality occasionally supersedes theory …
- Validation by proving ability to process by multiple tools?

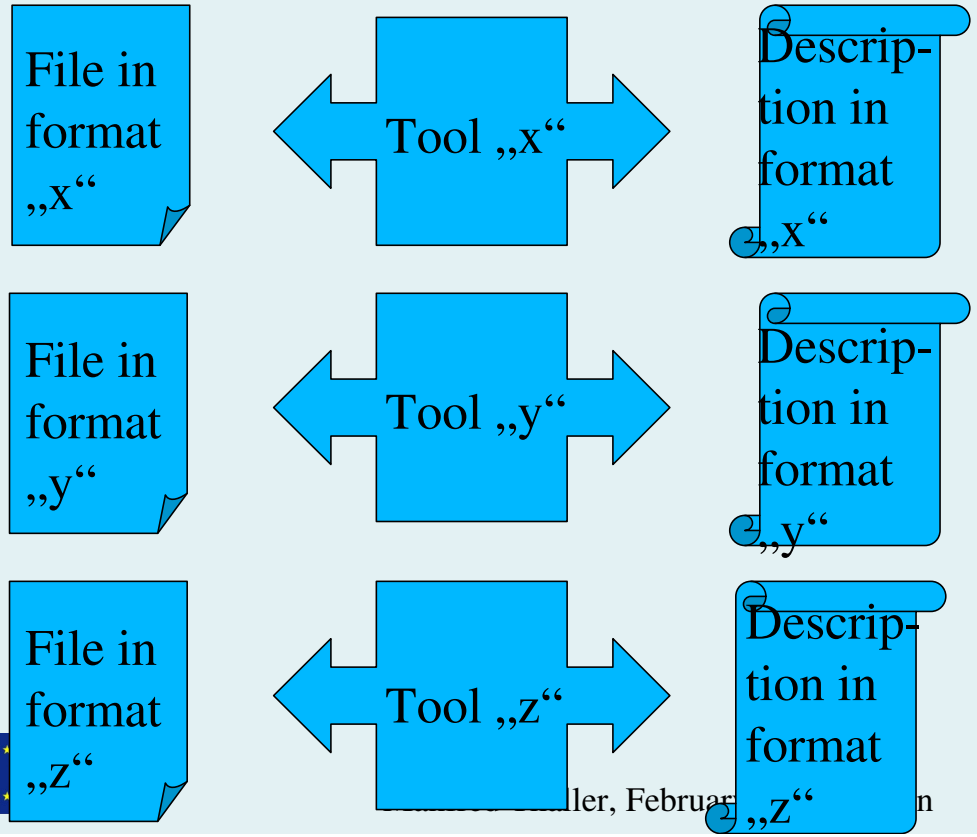
File related services



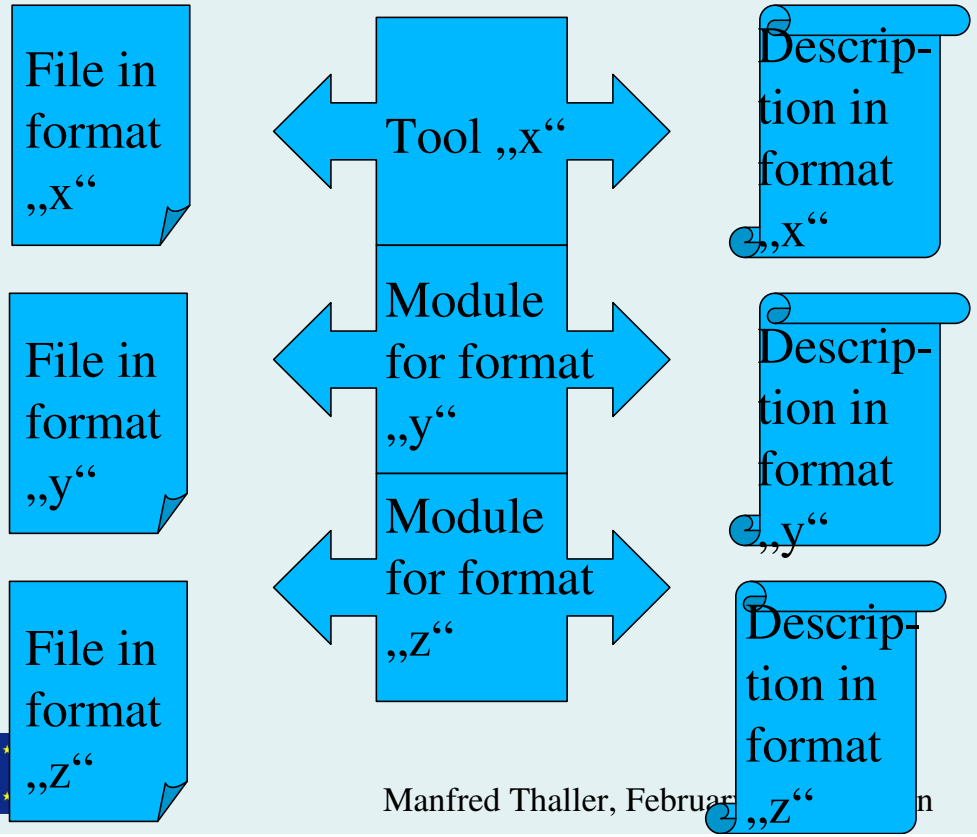
Extraction: How can we examine, whats really in a file?

- Many services which extract *some* characteristics from a file available via interoperability framework.
- “Planets borne”: The XCL approach.

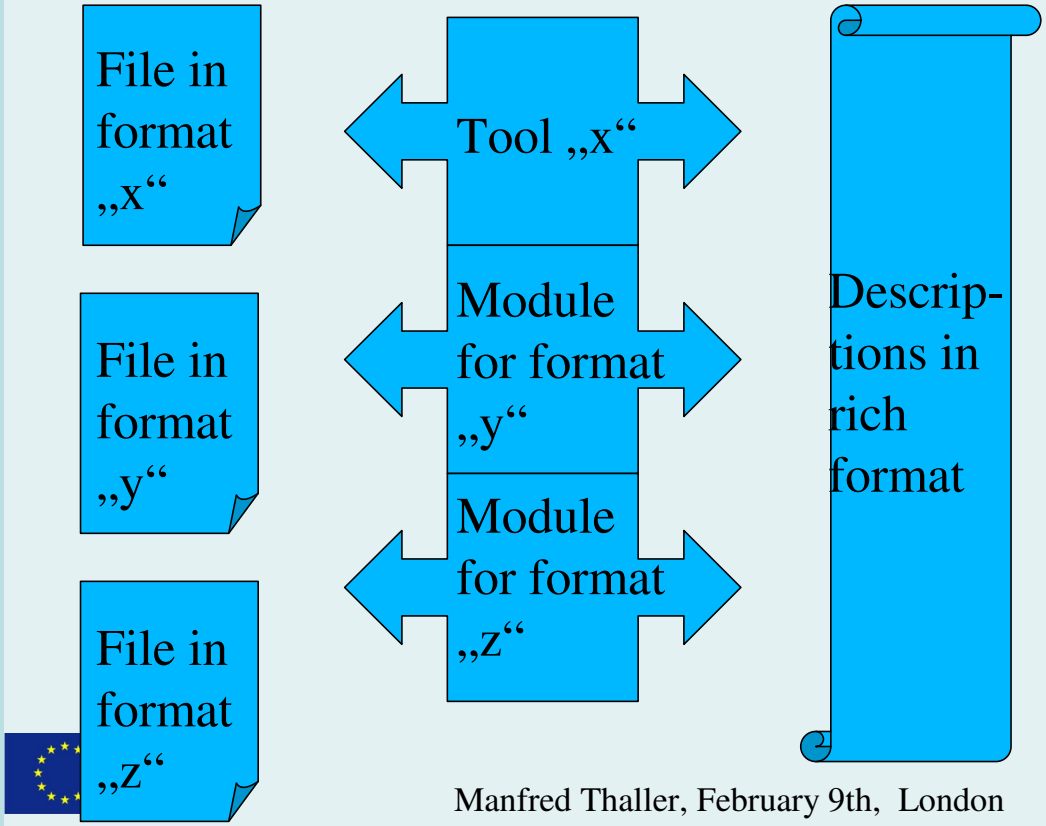
Traditional approach I:



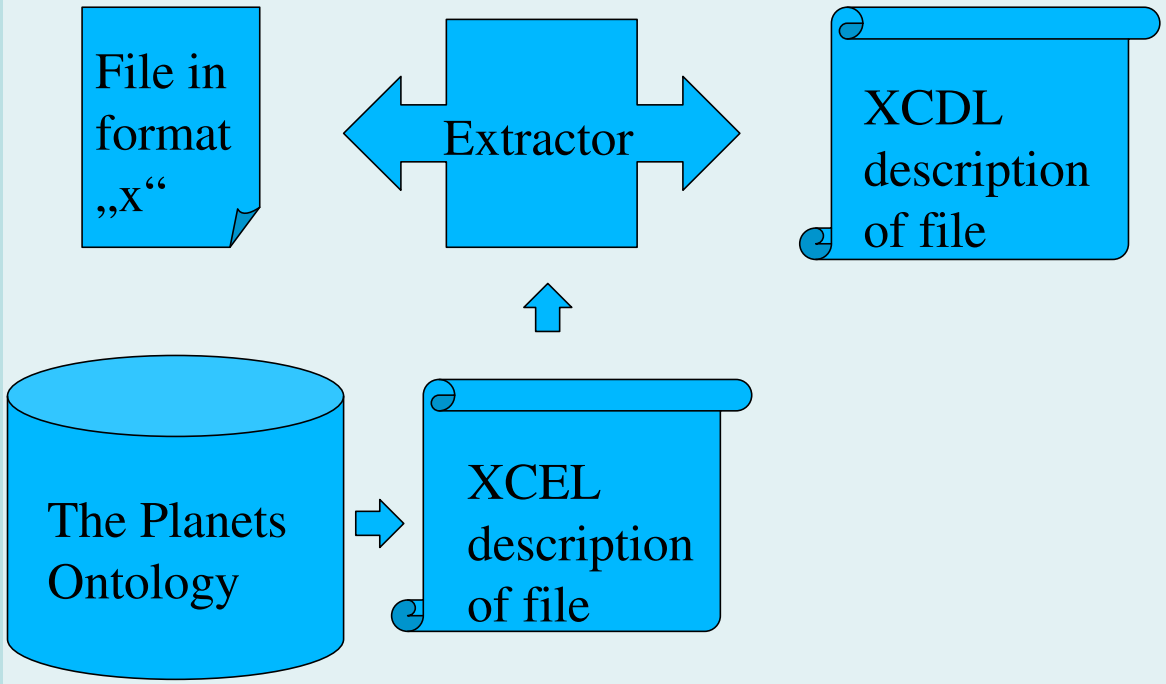
Traditional approach II:



Traditional approach III:



XCL approach



XCEL: The eXtensible Characterisation Extraction Language

XCEL (Extensible Characterisation Extraction Language)

```
<symbol value="137 80 78 71 13 10 26 10"/>

<symbol interpretation="uint32" length="4"/>

<symbol value="IHDR" interpretation="ASCII">

 <symbol interpretation="uint32"
      name="imageWidth" length="4"/>
```

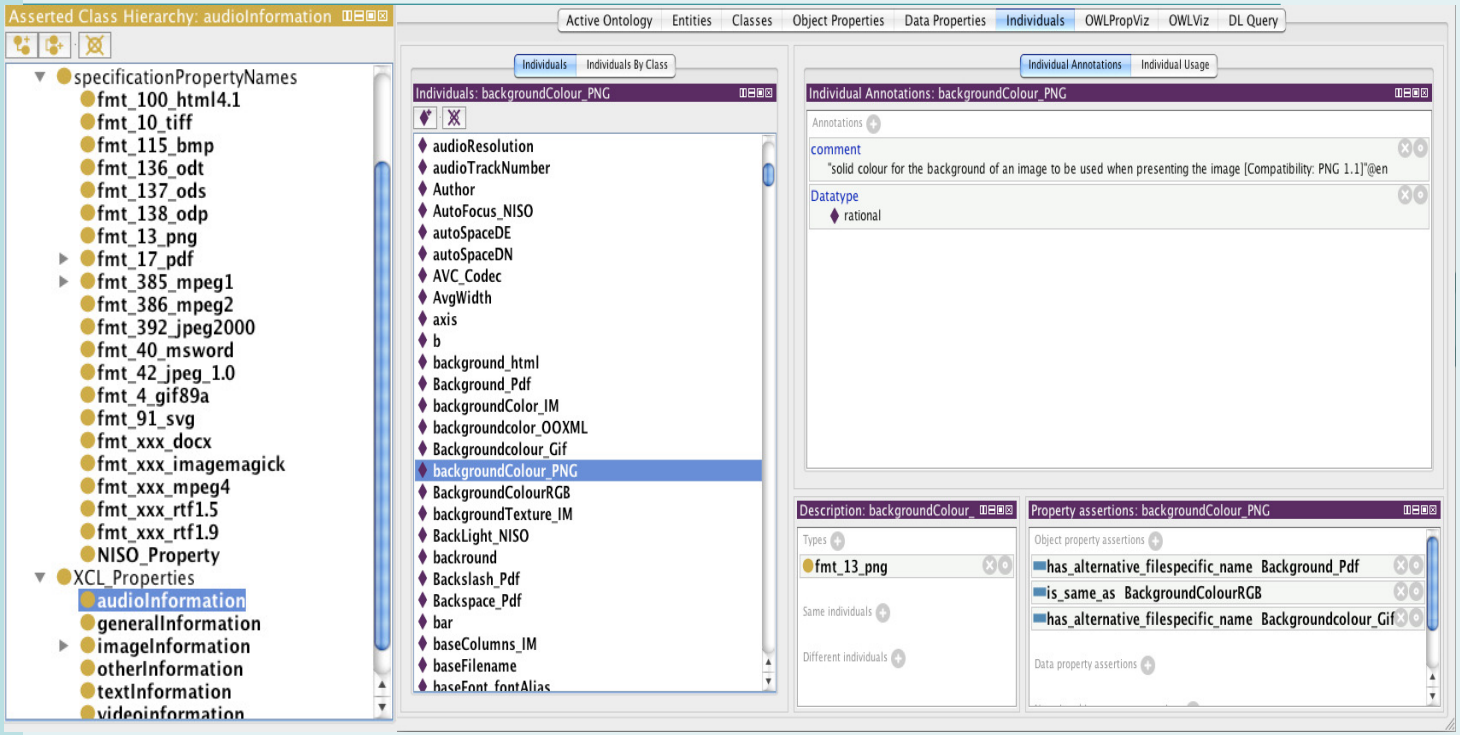
Natural Language

„The first eight bytes of a PNG datastream always contain the following (decimal) values: 137 80 78 71 13 10 26 10 [...] The four-byte chunk type field contains the decimal values 73 72 68 82[.] The IHDR chunk shall be the first chunk in the PNG datastream. It contains: Width 4 bytes [...] Width and height give the image dimensions in pixels. They are PNG four-byte unsigned integers. Zero is an invalid value.“ (http://www.w3.org/TR/PNG/)

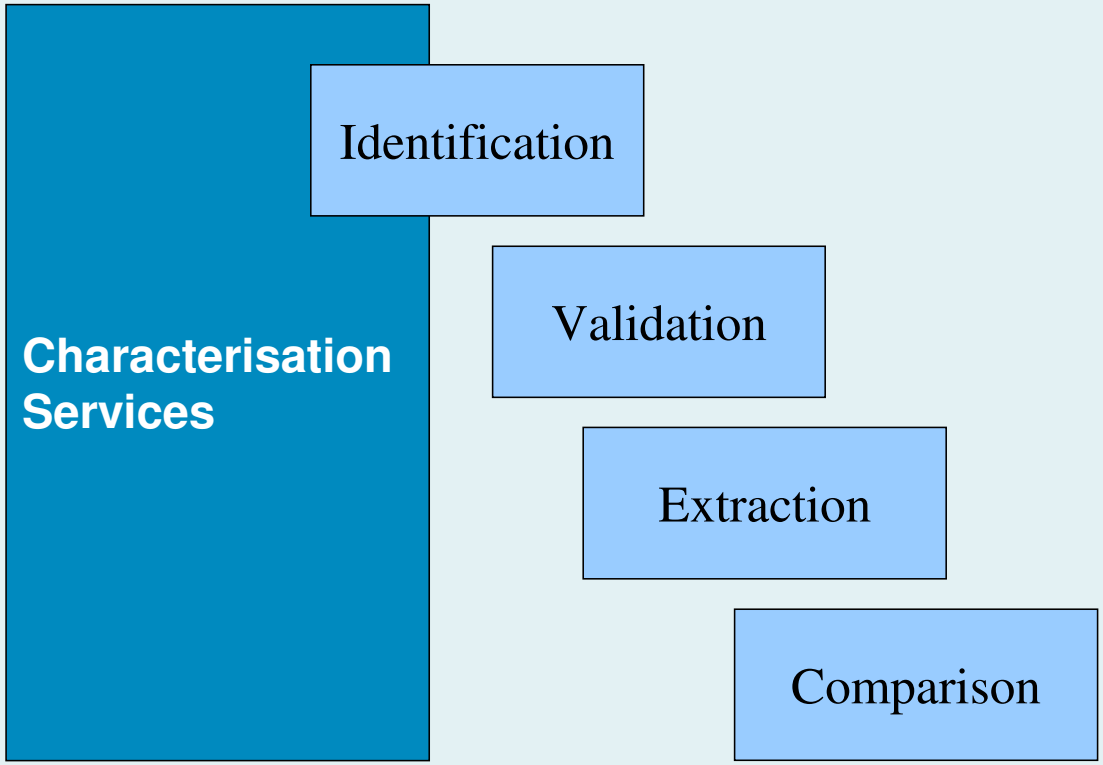
XCDL: The eXtensible Characterisation Definition Language

```
    <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 17 18 19 1a 1b 1c 1d ...
        </normData>
        <property id="p13" source="raw" cat="descr" >
            <name id="id2" >imageHeight</name>
            <valueSet id="i_i1_s10" >
                <labValue>
                    <val>32</val>
                    <type>int</type>
                </labValue>
            </valueSet>
        </property>
        <property id="p14" source="raw" cat="descr" >
            <name id="id30" >imageWidth</name>
            <valueSet id="i_i1_s8" >
                <labValue>
                    <val>32</val>
                    <type>int</type>
                </labValue>
            </valueSet>
        </property> ...
    </object>
```

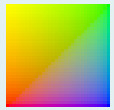
The Planets XCL Approach – The Ontology

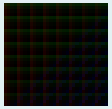


File related services



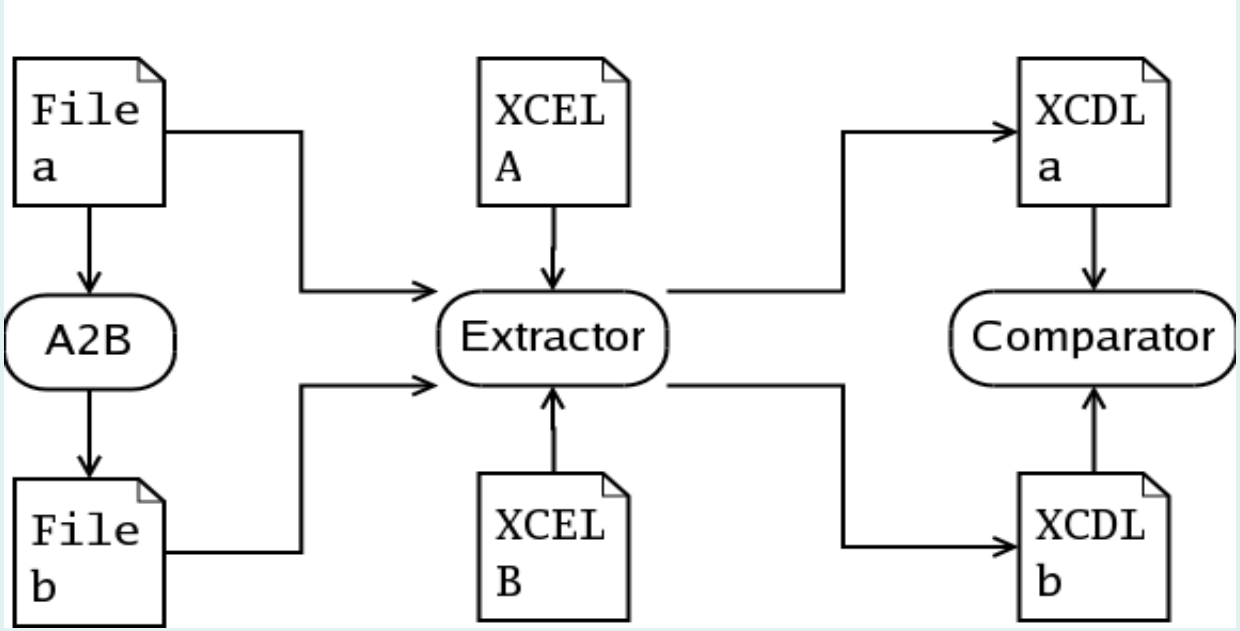
Comparison: Did a migration succeed?





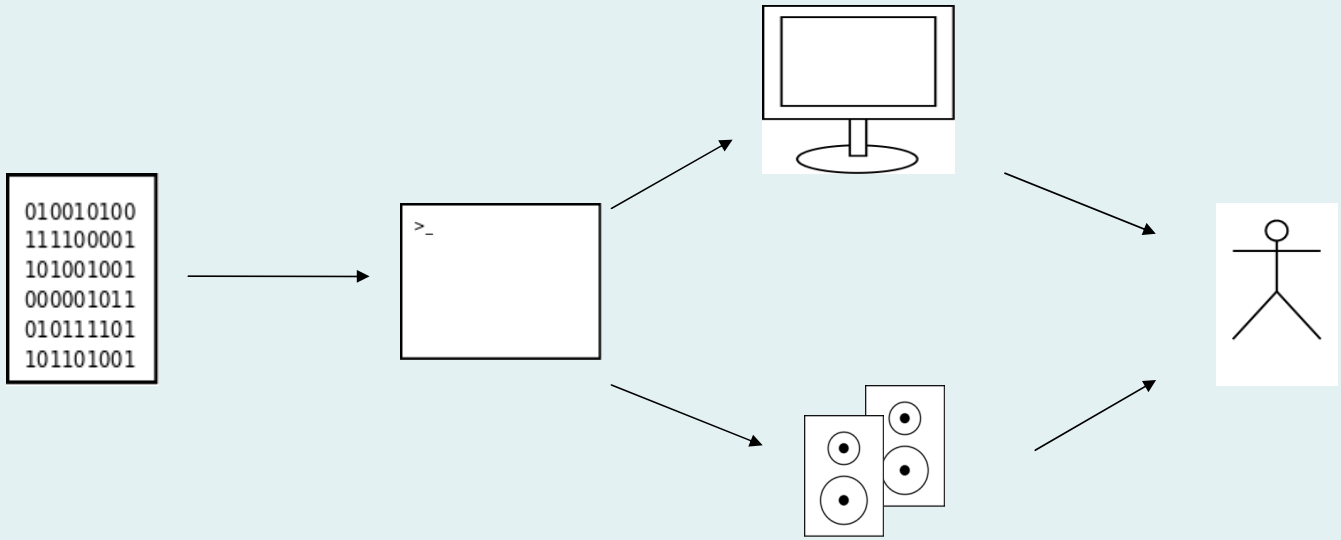
Comparison: Did a migration succeed?

Evaluation of Format Conversion

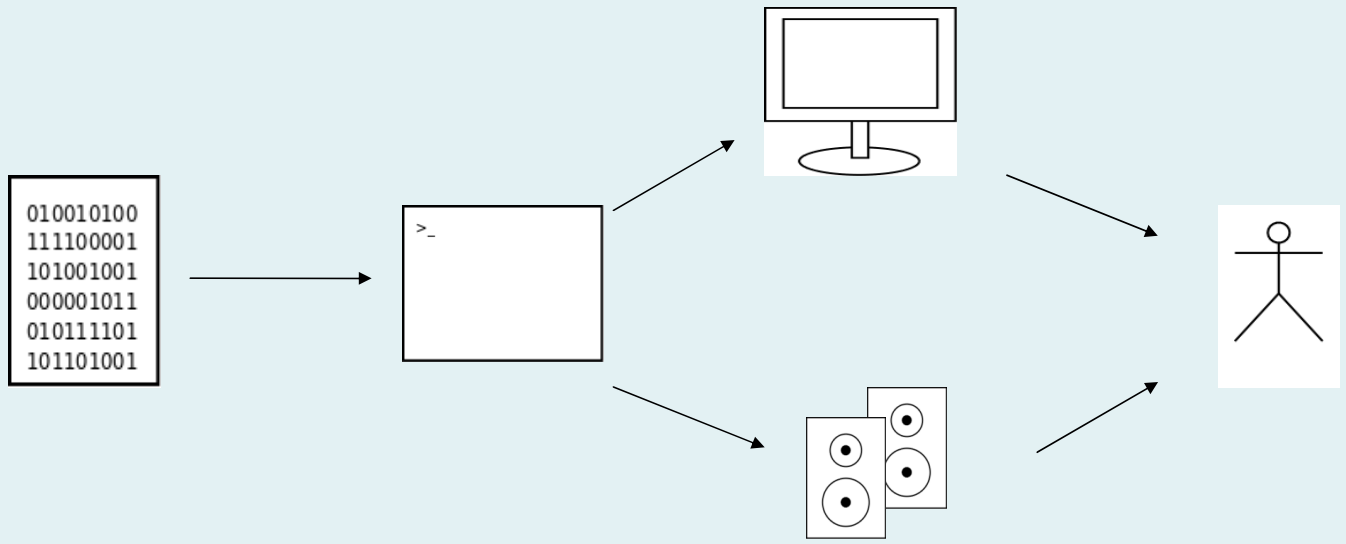


3. What´s the Philosophy behind it‘

Data, Perception, Information



Data, Perception, Information



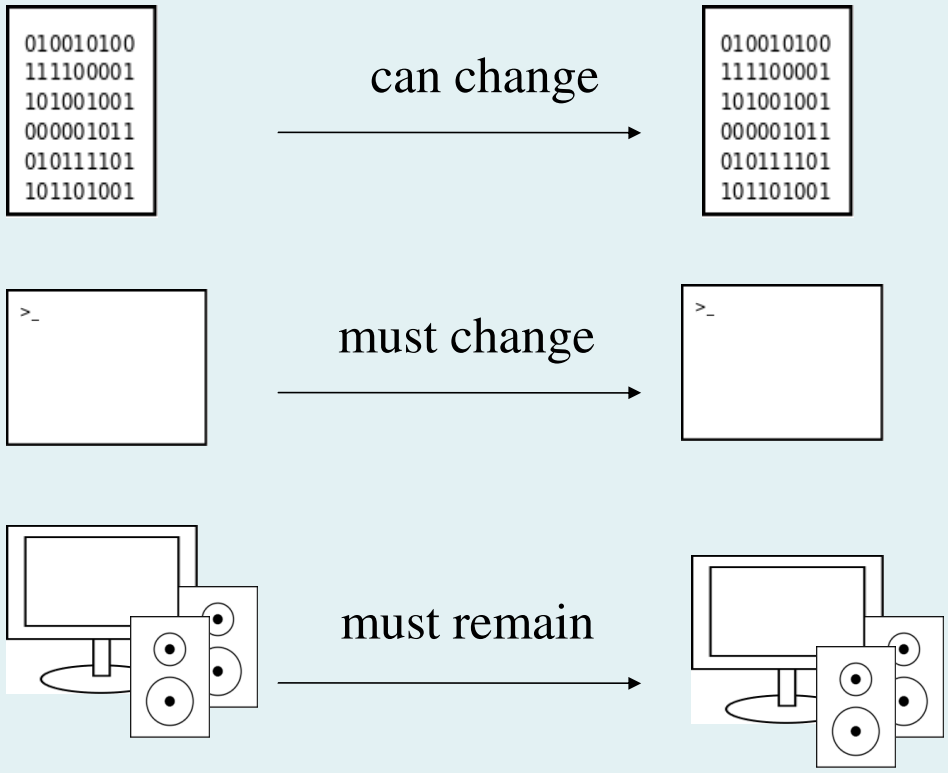
Data Representation

Processing

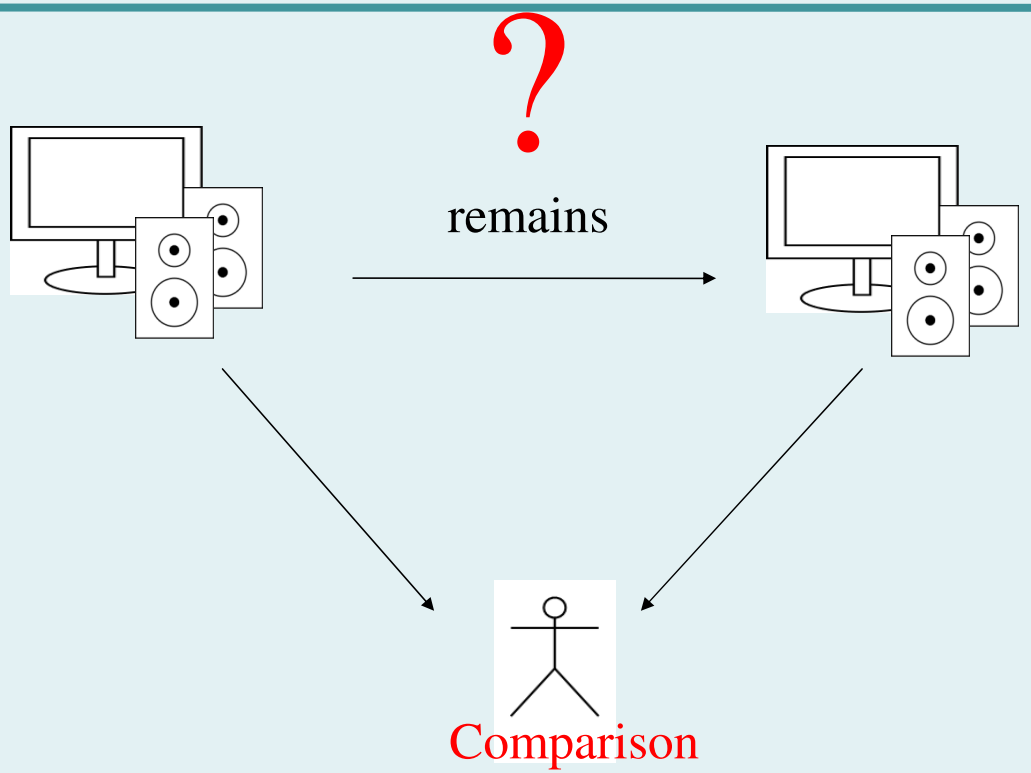
Presentation

Perception

Preservation Scenario: Format Conversion



Evaluation of Format Conversion



Evaluation of Format Conversion

Why automate?

1 million objects: use one second for each.

== 278 hours

== 35 8-hour days for a Human

== 1,8 months

Evaluation of Format Conversion

Why automate?

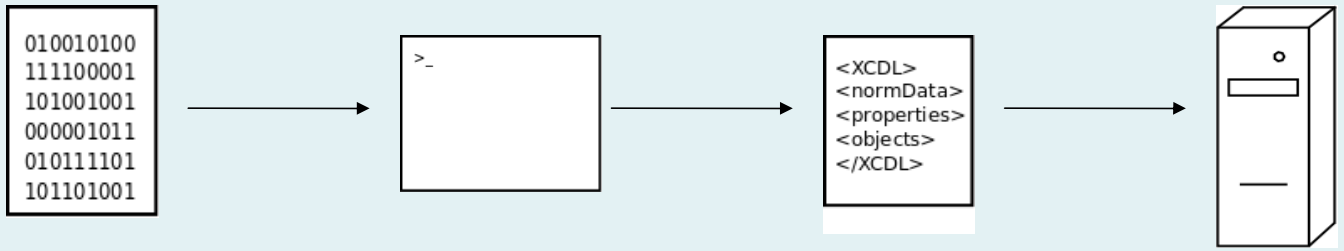
1 million objects: use five minutes for each.

== 83,333 hours

== 10,416 8-hour days for a Human

== way too much for anything

Evaluation of Format Conversion



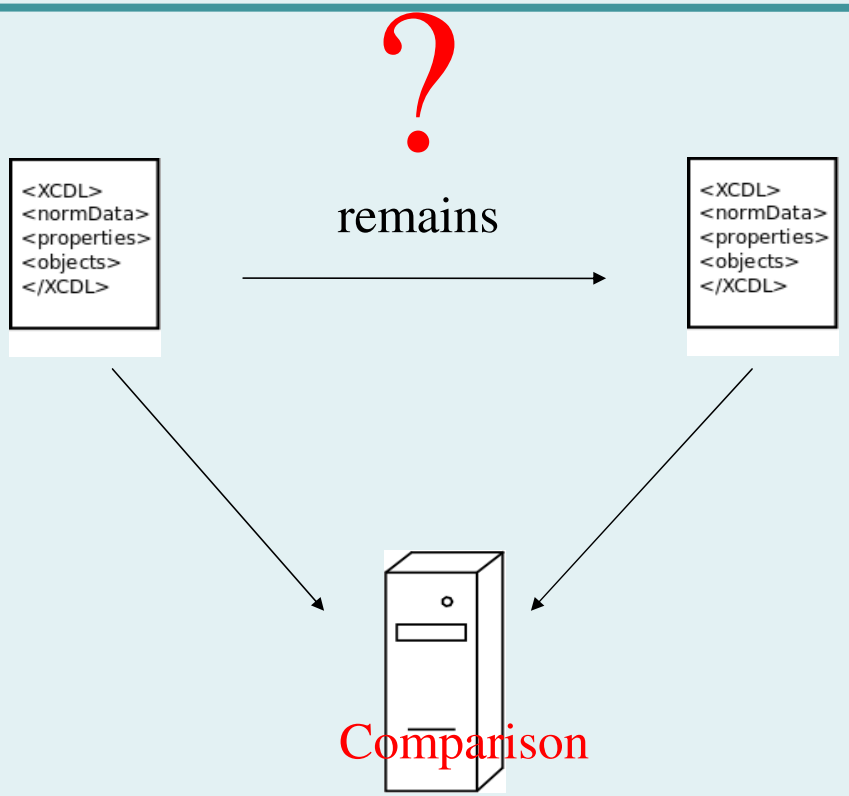
Data Representation

Processing

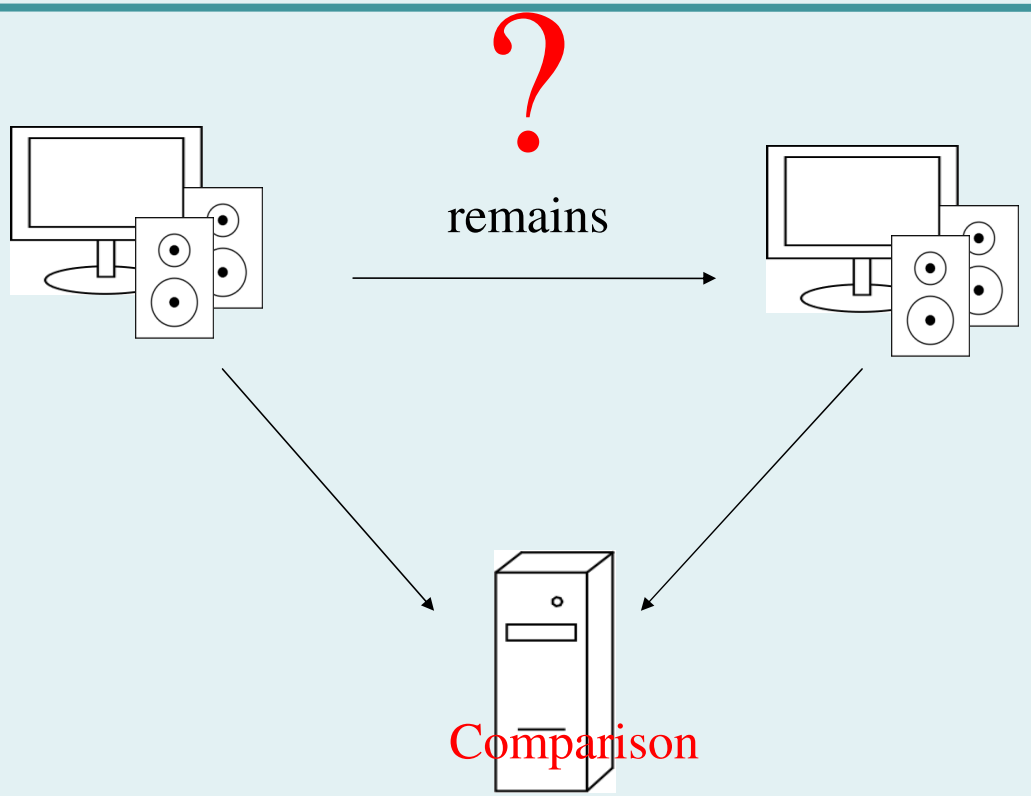
Machine-readable Presentation

Calculation

Evaluation of Format Conversion



Evaluation of Migration



4. How can Planets help? (Revisited)

How can Planets help: I

Extract the significant properties of a particular object or collection

The eXtensible Characterisation Description Language (XCDL) describes properties of digital objects such as colour and depth of an image. The eXtensible Characterisation Extraction Language (XCEL) describes how significant properties are encoded and makes it possible to extract them automatically.

The Extractor tool extracts characteristics from digital objects. The extensible framework supports third-party tools such as DROID (Digital Record Object Identification).

How can Planets help: III

Verify that actions have been successful and quality assure the outcome

The Planets Comparator compares digital objects before and after treatment to ensure that treatment has not changed the significant properties of the object and so has been successful.

5. What is not in a file?

…
… no further discussion.

2
Except, if we assume …

Foreign Ministry, The Hague

…
… no further discussion.
Except, if we assume …

2

Dutch Embassy, Washington DC

…
… no further discussion.

2
Except, if we assume …

Foreign Ministry, The Hague

It took us some time to convince the techies, but now we really see on the screen what appears on the printer.

It took us some time to convince the techies, but now we really see on the screen what appears on the printer.

…
… no further discussion.
Except, if we assume …

2

Dutch Embassy, Washington DC

…
… no further discussion.

2
Except, if we assume …

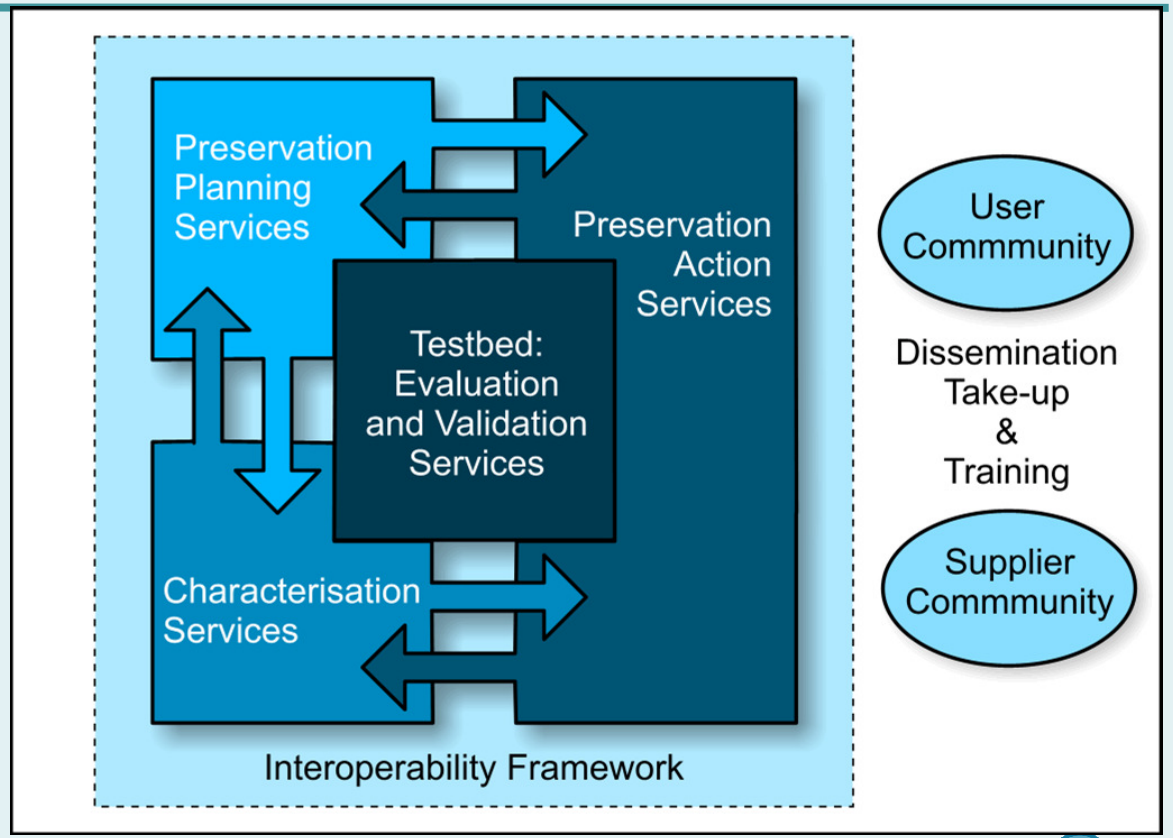
…
… no further discussion.
Except, if we assume …

2

Each party has seen this document only in its own settings.

Which view shall we preserve?

Zoom Out


Preservation Planning Services
Preservation Action Services
Testbed: Evaluation and Validation Services
Characterisation Services
Interoperability Framework
User Commmunity
Dissemination Take-up & Training
Supplier Commmunity

Further Information (on XCL, particularly):

planetarium.hki.uni-koeln.de

Thank you!