

**Historisch**

**Kulturwissenschaftliche**

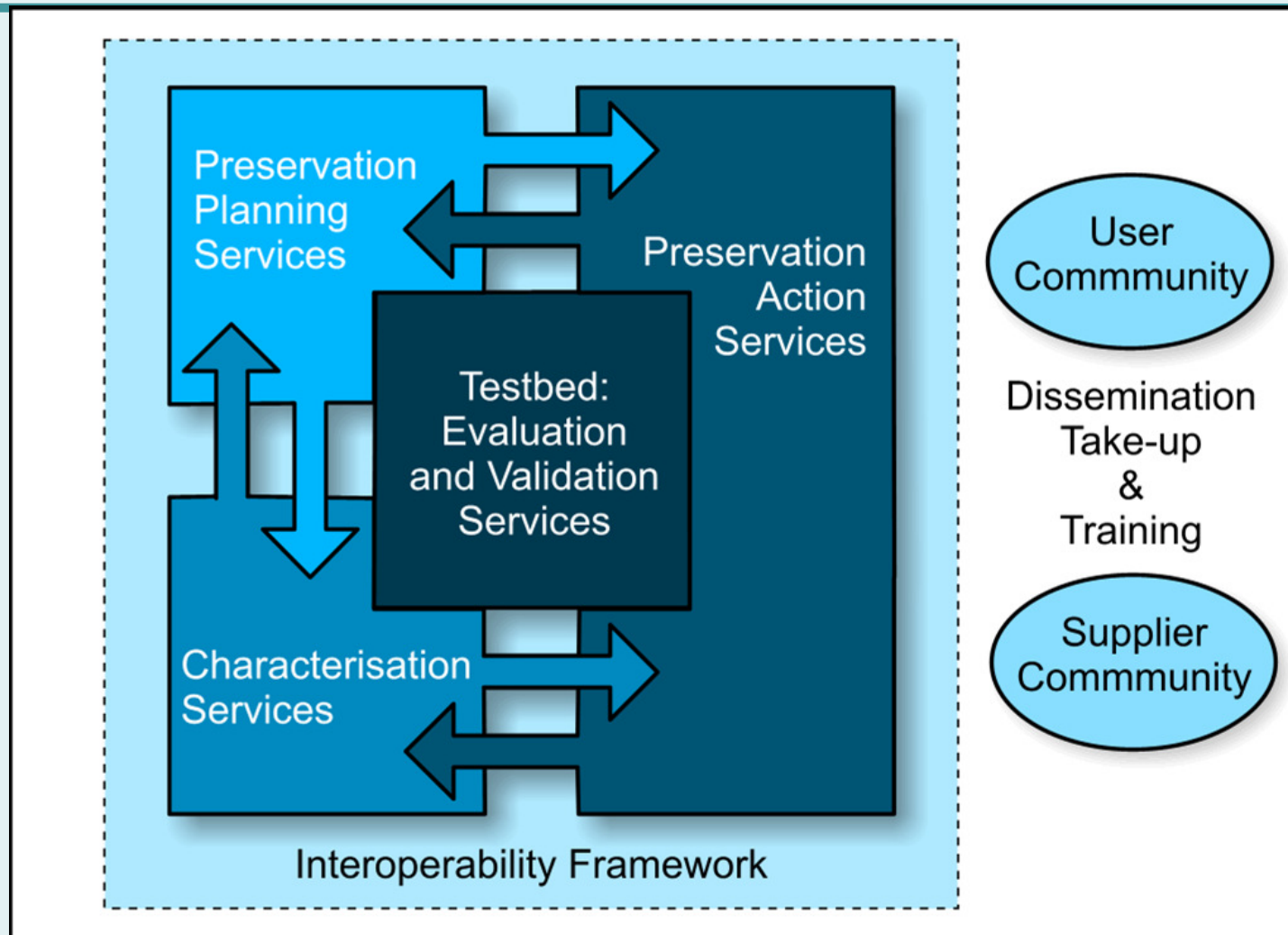
**Informationsverarbeitung**

# Tools: How to understand files!

**Jan Schnasse**

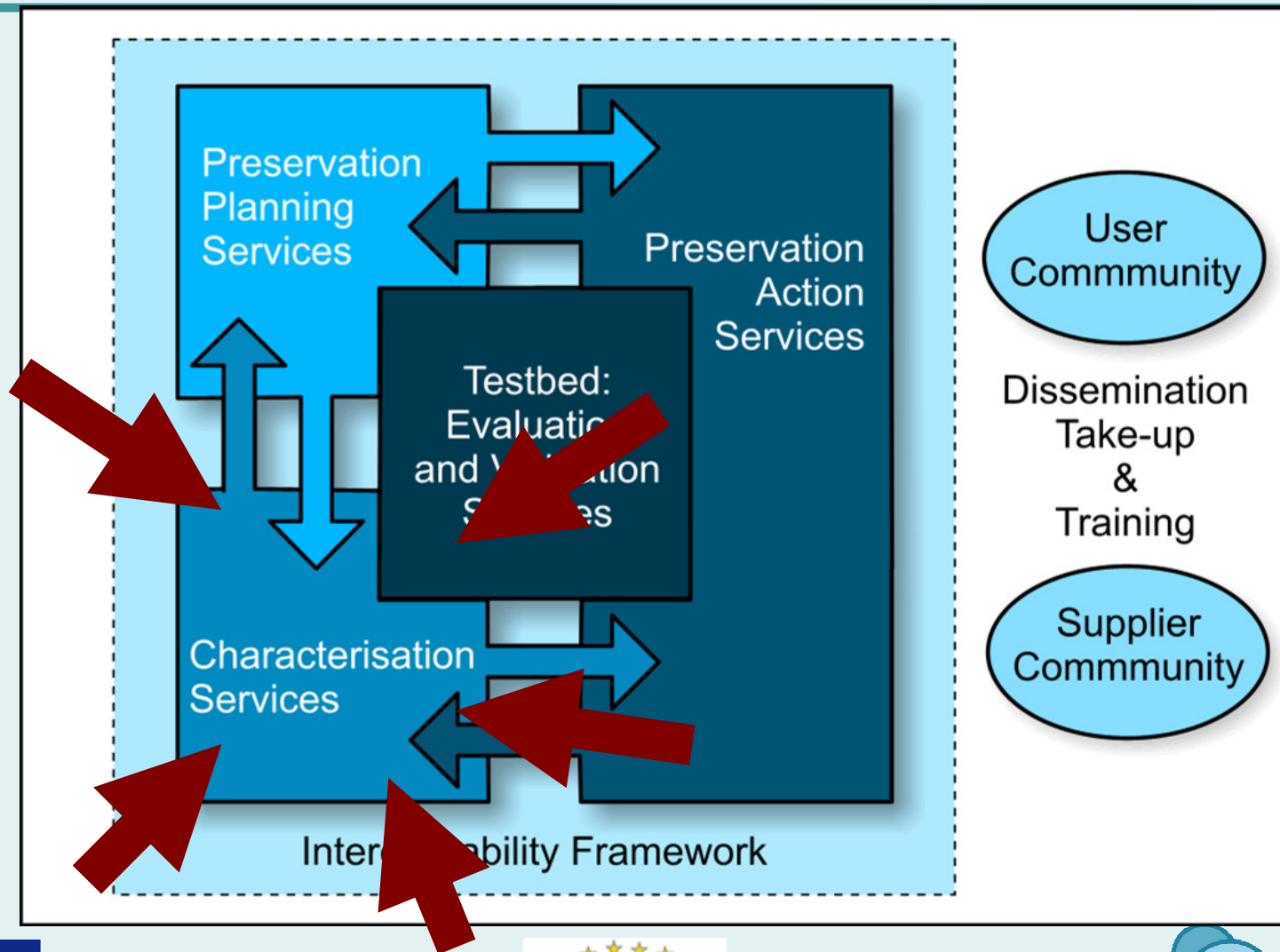


# Zoom In





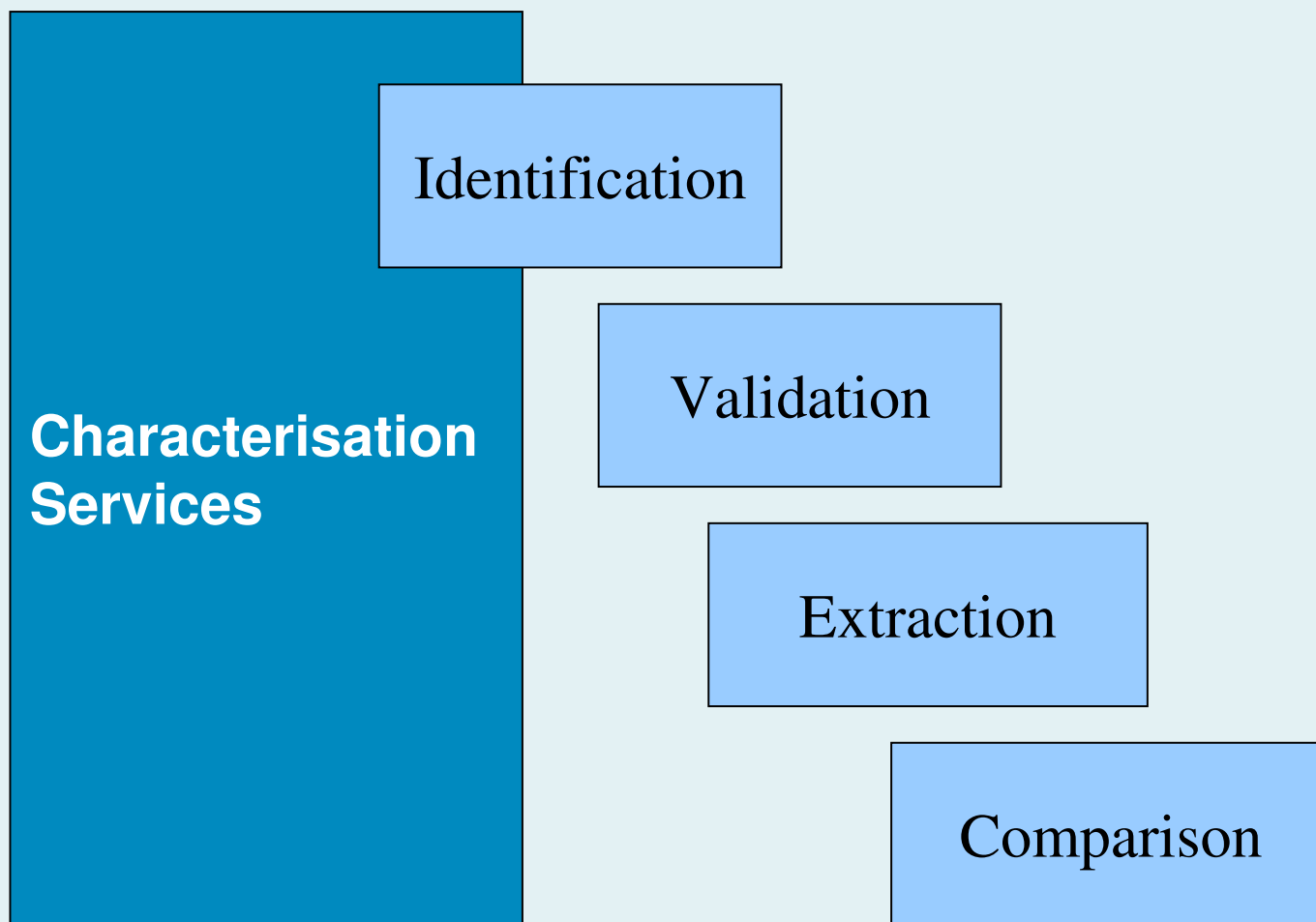
# Zoom In





# Zoom In

---





# Focus

---

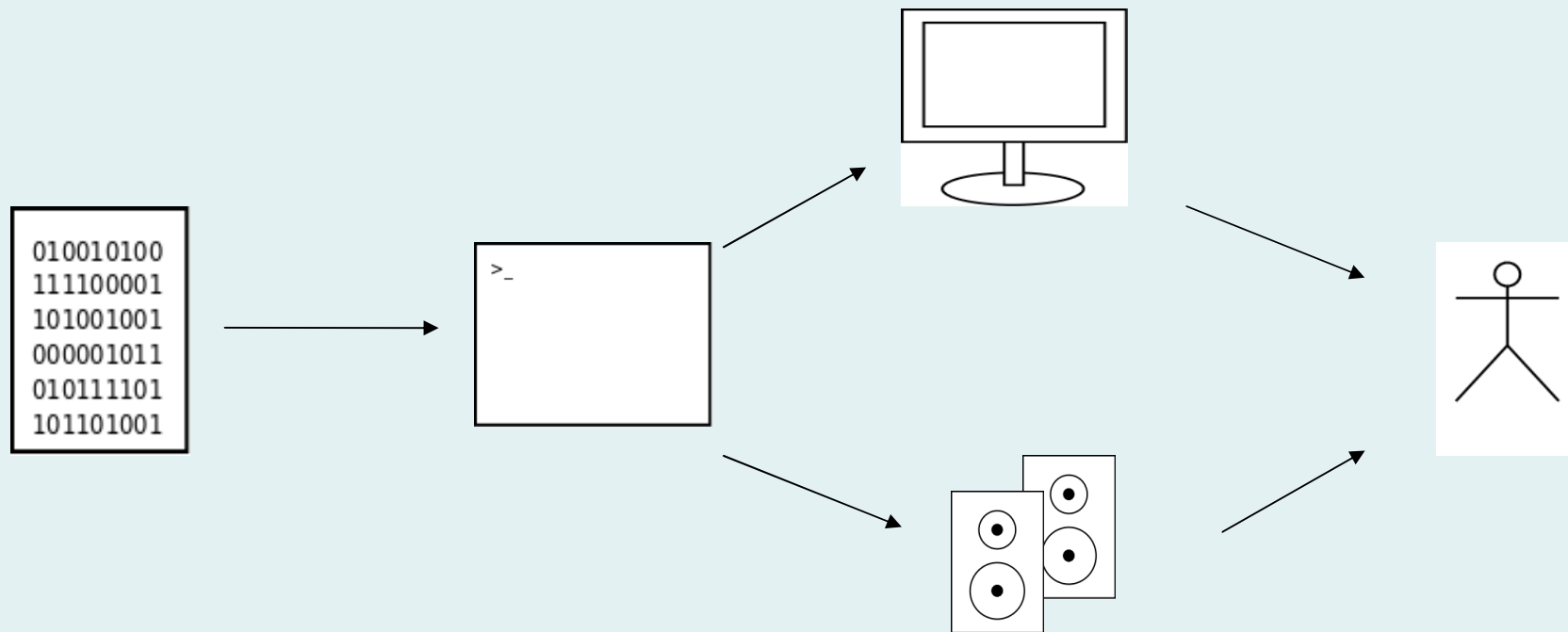
Extraction

Comparison



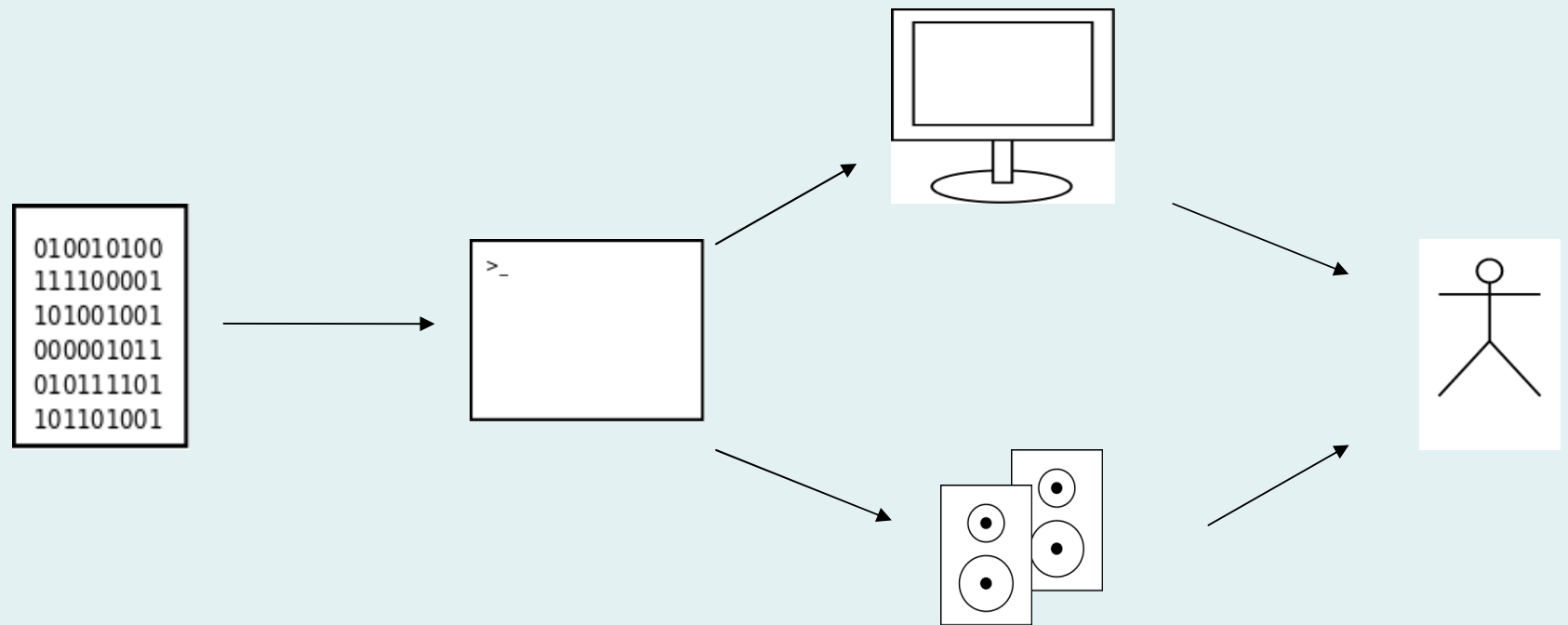


# Data, Perception, Information





# Data, Perception, Information



Data  
Representation

Processing

Presentation

Perception

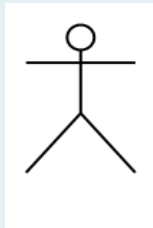




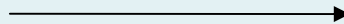
# What to preserve?

---

Perception



must be preserved





# What to preserve?

---

What do you see?

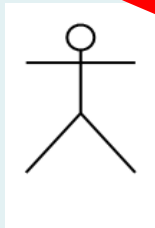




# What to preserve?

---

Perception



must be preserved



**unpreservable**

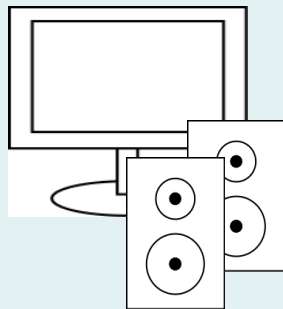




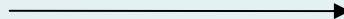
# What to preserve?

---

## Presentation



must be preserved

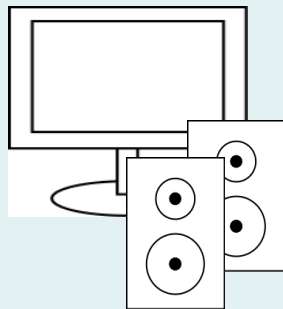




# What to preserve?

---

## Presentation



must be preserved



adequate  
accurate  
authentic  
original  
significant

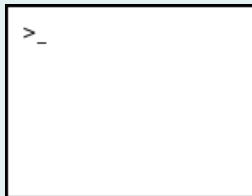




# What to preserve?

---

Processing



must be preserved  
→

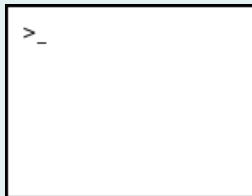




# What to preserve?

---

Processing



must be preserved  
→

sufficient





# What to preserve?

---

## Data Representation

010010100
111100001
101001001
000001011
010111101
101101001

must be preserved

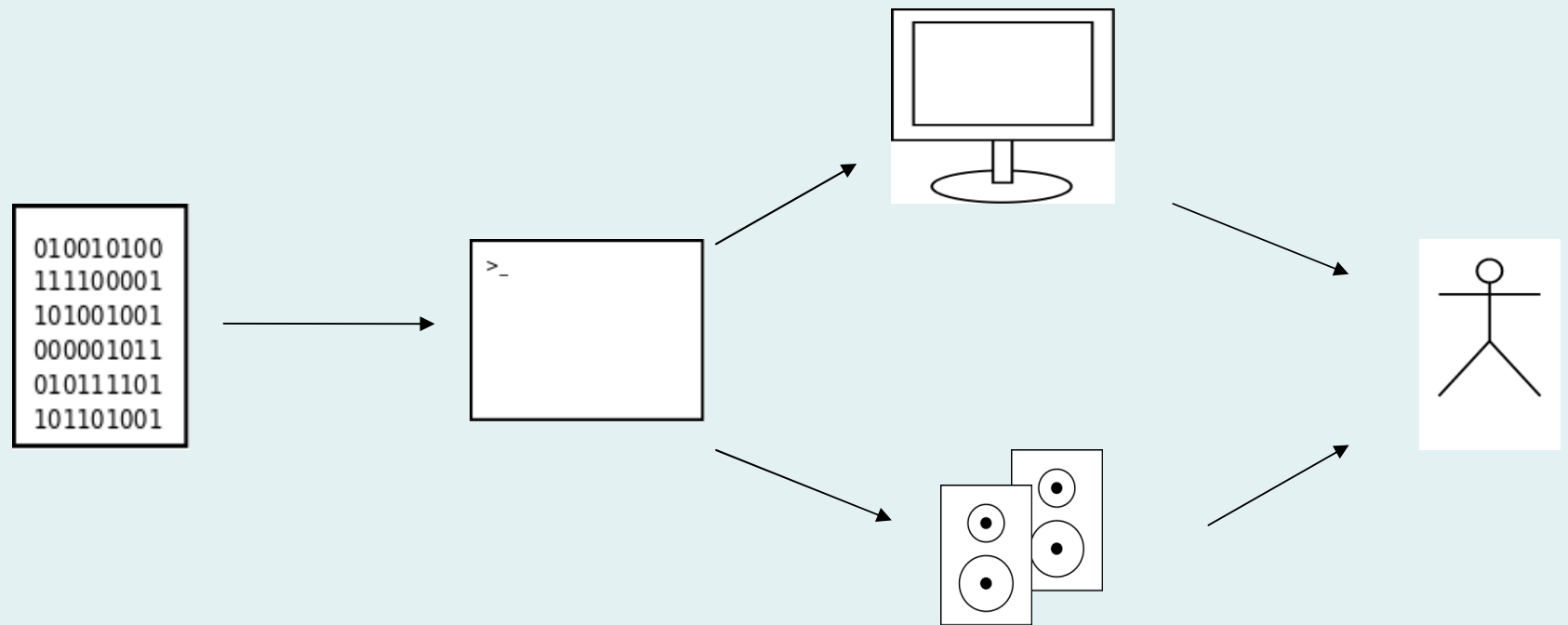


arbitrary  
but  
complete!





# Data, Perception, Information



Data  
Representation

Processing

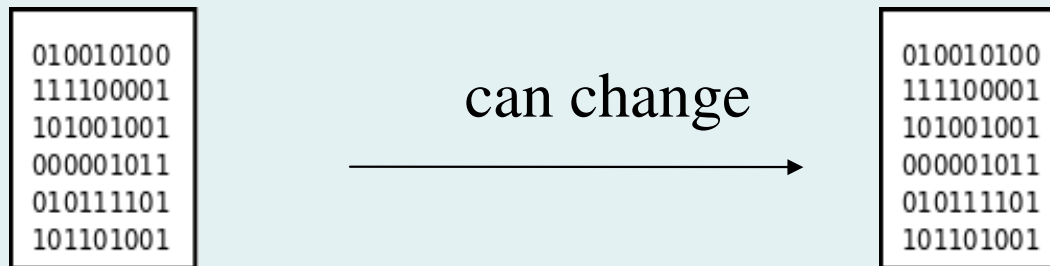
Presentation

Perception





# Preservation Scenario: Format Conversion





# Preservation Scenario: Format Conversion

```
010010100
111100001
101001001
000001011
010111101
101101001
```

can change

```
010010100
111100001
101001001
000001011
010111101
101101001
```

```
>_

```

can change

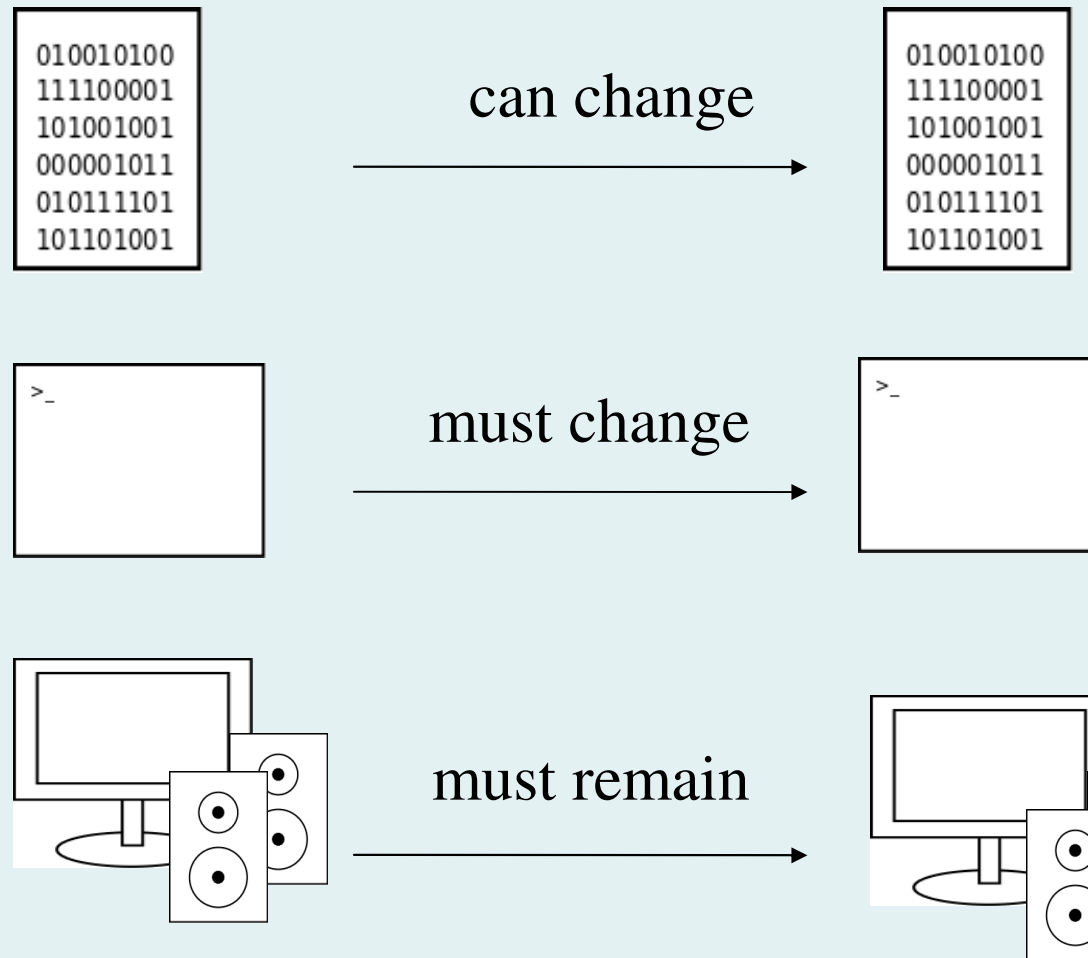
```
>_

```





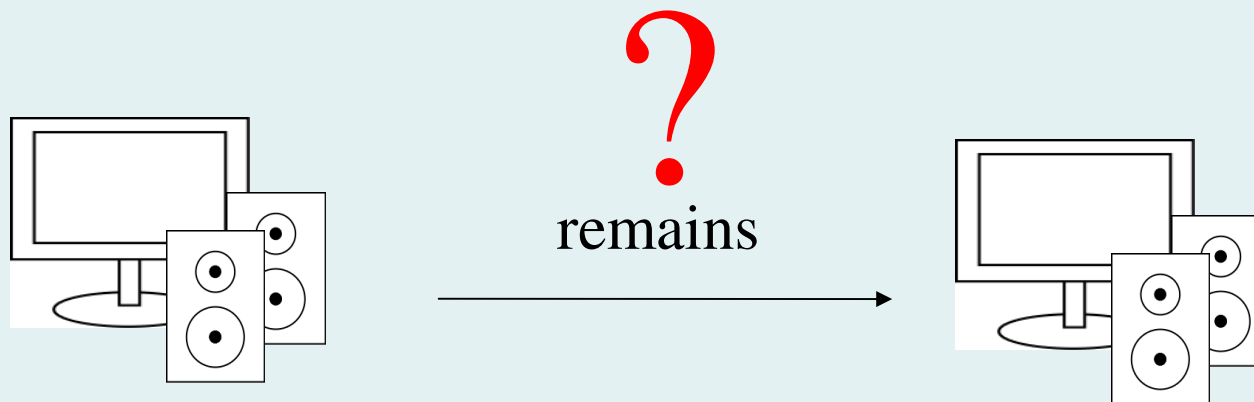
# Preservation Scenario: Format Conversion





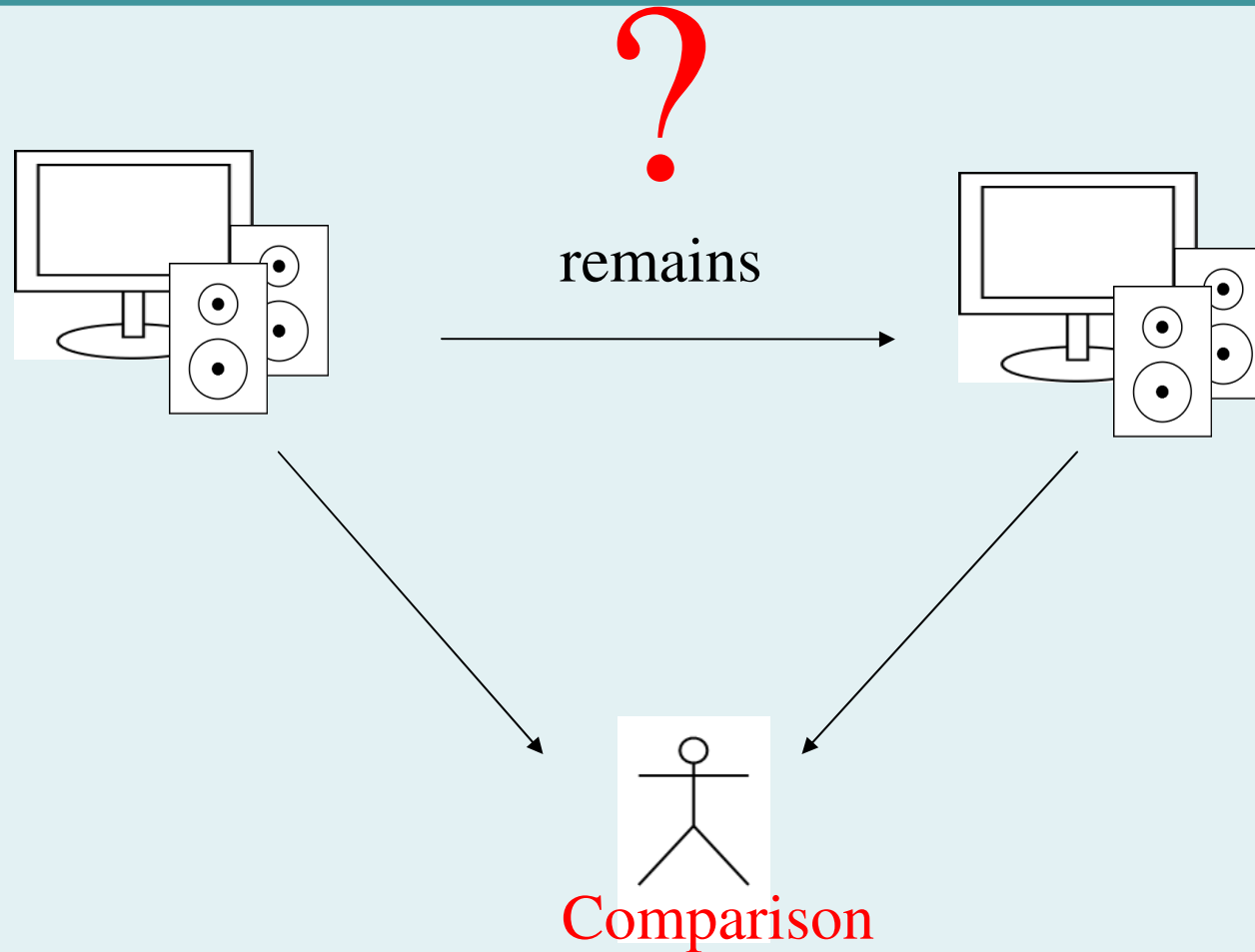
# Preservation Scenario: Format Conversion

---





# Evaluation of Format Conversion





# Evaluation of Format Conversion

---

## Why automate?

1 million objects: use five minutes for each.

== 416 666.7 hours

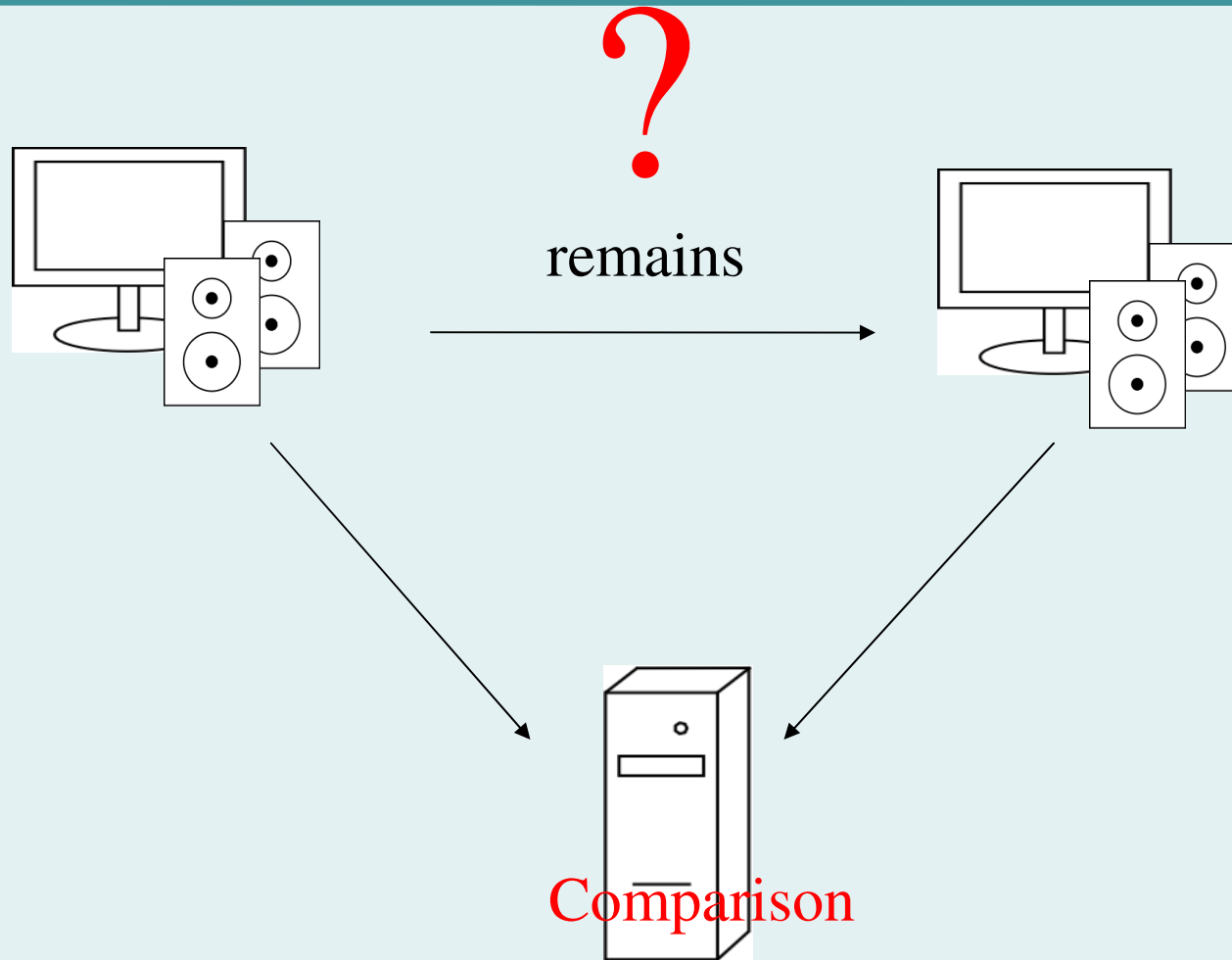
== 52 803.4 8-hour days for a Human

== way too much for anything



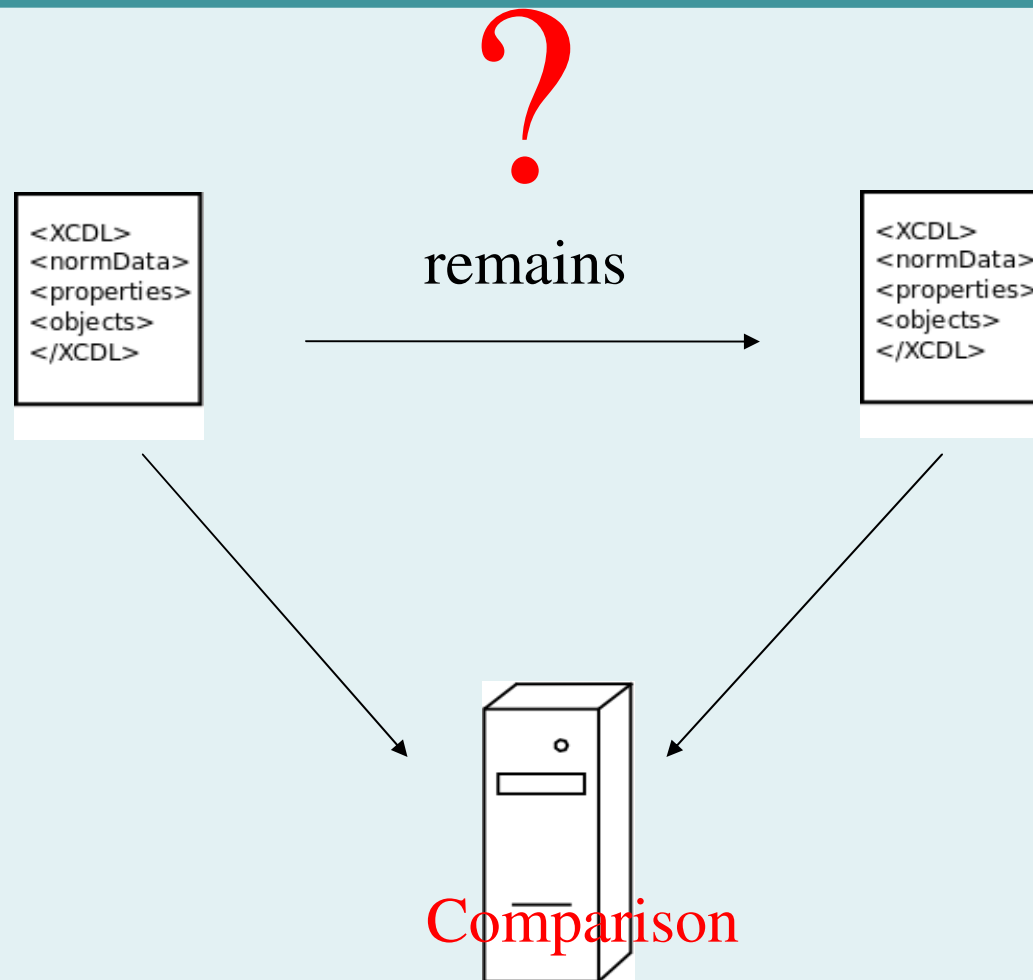


# Evaluation of Migration



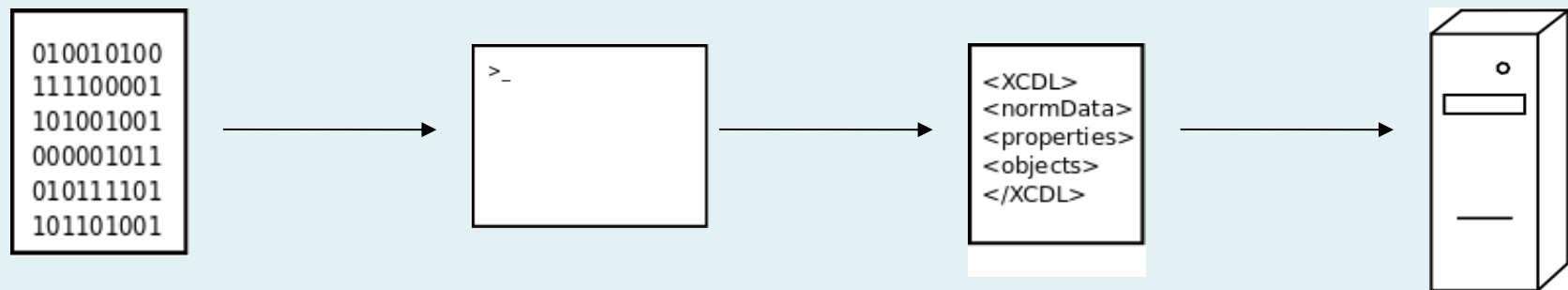


# Evaluation of Format Conversion





# Evaluation of Format Conversion



Data  
Representation

Processing

Machine-readable  
Presentation

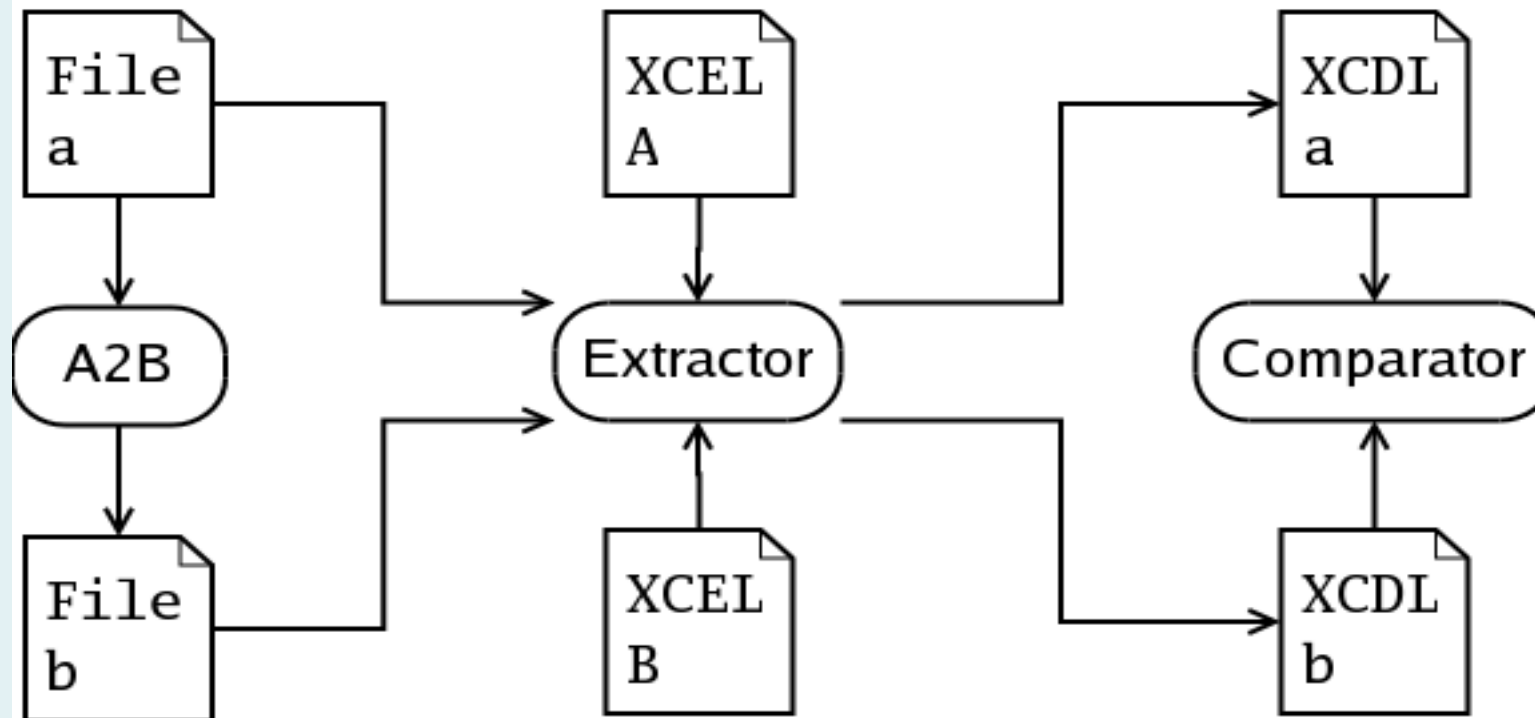
Calculation





# The Planets XCL Approach - Overview

## Evaluation of Format Conversion





# The Planets XCL Approach – Describing File Formats

## 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/>)





# The Planets XCL Approach - Extractor

The screenshot displays the Planets XCL Extractor application window. The menu bar includes File, Actions, Options, and Windows. Below the menu is a toolbar with an 'Open' button, a 'Text size' dropdown set to 12, and a 'Text weight' dropdown set to 25. The main text area shows the XML output for 'output/FontTest1.pdf.xcdl'. The XML structure includes a root element with namespaces, an object element, a normData element containing the text 'Font Font Font Font' (with the fourth 'Font' in blue), and two property elements. The first property describes a font change, and the second describes the font name 'NimbusRomNo9L-Medi'. At the bottom left, a table lists the processed file: '1 FontTest1.pdf'. At the bottom right, there is an 'XCEL' section with a text input field containing 'xcl/xcel/xcel\_pdf.xml' and a 'GO' button.

```
<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 ../xcl/xcdl/XCDLCore.xsd" id="0" >
  <object id="o1" >

    <normData type="text" id="nd1" >Font Font Font Font</normData>

    <property id="p115" source="raw" cat="descr" >
      <name id="id381" >fontChange</name>
      <valueSet id="i_i1_i52_s1" >
        <labValue>
          <val></val>
          <type>string</type>
        </labValue>
        <dataRef ind="normSpecific" propertySetId="id_0" />
        <dataRef ind="normSpecific" propertySetId="id_1" />
        <dataRef ind="normSpecific" propertySetId="id_2" />
      </valueSet>
    </property>

    <property id="p93" source="raw" cat="descr" >
      <name id="id159" >fontName</name>
      <valueSet id="i_i1_i49_i4_i3" >
        <labValue>
          <val>NimbusRomNo9L-Medi</val>
          <type>string</type>
        </labValue>
      </valueSet>
    </property>
  </object>
</xcdl>
```

	1
1	FontTest1.pdf

XCEL

xcl/xcel/xcel\_pdf.xml

GO





# The Planets XCL Approach – General Data Representation with XCDL (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_il_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_il_s8" >
      <labValue>
        <val>32</val>
        <type>int</type>
      </labValue>
    </valueSet>
  </property> ...
</object>
```





# The Planets XCL Approach – The Ontology

The screenshot displays the Protégé ontology editor interface for the Planets XCL ontology. The left pane shows the 'Asserted Class Hierarchy: audioInformation' with a tree structure including 'specificationPropertyNames' and 'XCL\_Properties'. The 'XCL\_Properties' class is expanded, showing 'audioInformation' as a subclass. The main pane shows the 'Individuals: backgroundColour\_PNG' class, listing various properties like 'audioResolution', 'audioTrackNumber', 'Author', 'AutoFocus\_NISO', 'autoSpaceDE', 'autoSpaceDN', 'AVC\_Codec', 'AvgWidth', 'axis', 'b', 'background\_html', 'Background\_Pdf', 'backgroundcolor\_IM', 'backgroundcolor\_OOXML', 'Backgroundcolour\_Gif', 'backgroundColour\_PNG', 'BackgroundColourRGB', 'backgroundTexture\_IM', 'BackLight\_NISO', 'background', 'Backslash\_Pdf', 'Backspace\_Pdf', 'bar', 'baseColumns\_IM', 'baseFilename', and 'baseFont\_fontAlias'. The right pane shows 'Individual Annotations: backgroundColour\_PNG' with a 'comment' annotation: 'solid colour for the background of an image to be used when presenting the image [Compatibility: PNG 1.1]"@en' and a 'Datatype' annotation: 'rational'. The bottom right pane shows 'Property assertions: backgroundColour\_PNG' with object property assertions: 'has\_alternative\_filespecific\_name Background\_Pdf', 'is\_same\_as BackgroundColourRGB', and 'has\_alternative\_filespecific\_name Backgroundcolour\_Gif'.





# The Planets XCL Approach - Comparator

The screenshot displays the XCLSuite application window. On the left, a 'File' list contains various image files with status indicators (red X, green checkmark, blue X). The file 'basi2c16' is selected. In the center, two 'Data View : basi2c16' windows show image thumbnails. On the right, a 'Comparator View : basi2c16' window displays comparison results for 'imageWidth' and 'normData'.

File	Status 1	Status 2
basi0g01	✗	✗
basi0g02	✓	✗
basi0g04	✓	✗
basi0g08	✓	✗
basi0g16	✗	✗
basi2c08	✓	✗
<b>basi2c16</b>	✗	✗
basi3p01	✓	✗
basi3p02	✓	✗
basi3p04	✓	✗
basi3p08	✓	✗
basi4a08	✗	✗
basi6a08	✗	✗
basn0g01	✓	✗
basn0g02	✓	✗
basn0g04	✓	✗
basn0g08	✓	✗
basn0g16	✗	✗
basn2c08	✓	✗
basn2c16	✗	✗
basn3p01	✓	✗

**Comparator View : basi2c16**

Property: imageWidth  
Metric: equal  
Result: **true**

Property: normData  
Metric: hammingDistance  
Result: **1288**





# How to understand files!

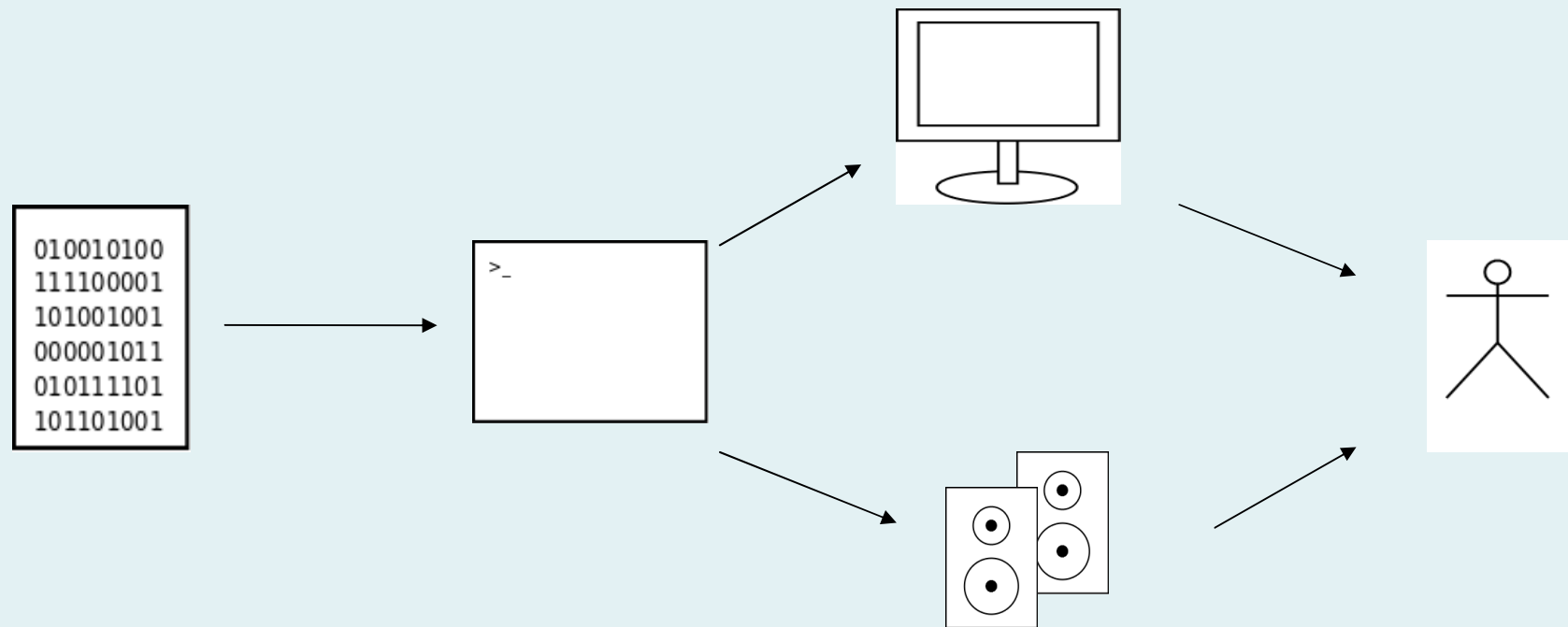
---

## What is not in a file?!





# Data, Perception, Information



Data  
Representation

Processing

Presentation

Perception





# Focus

---

Extraction

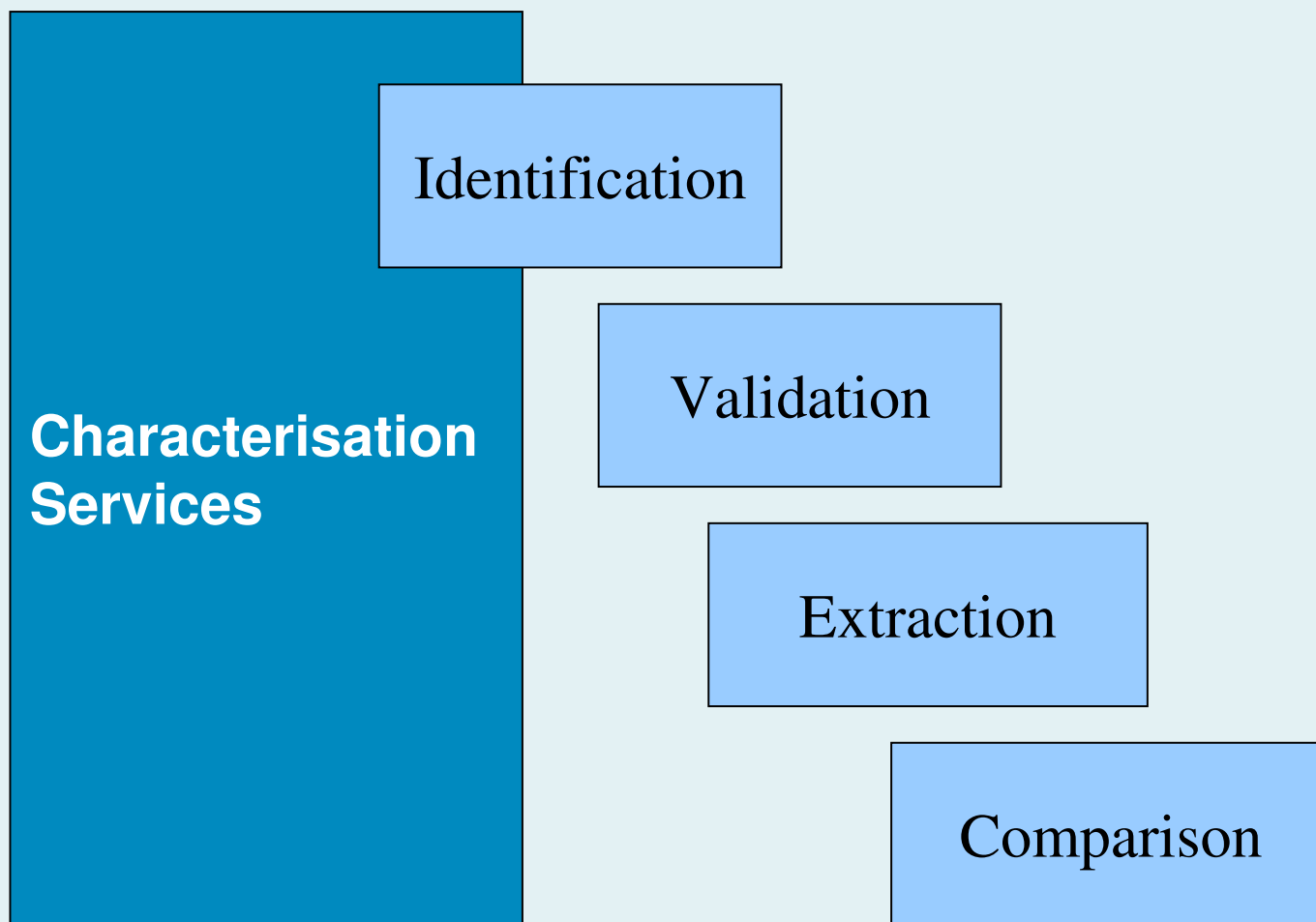
Comparison





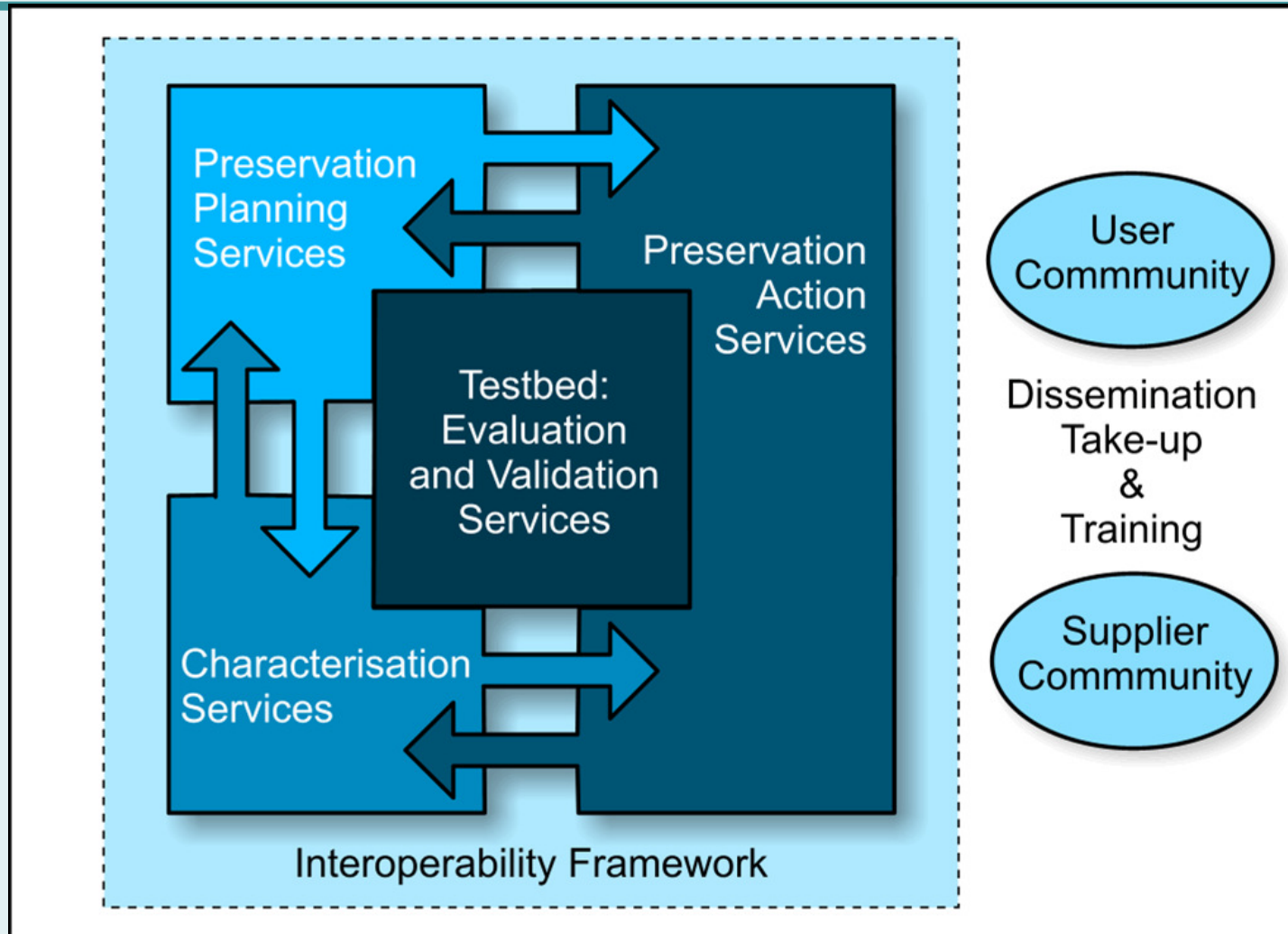
# Zoom Out

---





# Zoom Out





# How to understand files!

---

## The End

