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

Manfred Thaller
A vision II

Tooth of time

"A"

"B"

Extractor

93%

Comparer

XCDL for A

XCDL for B

Appropriate XCEL

A vision III

Extractor

Appropriate XCEL

Summarizer

C-Set

Manfred Thaller

XCL approach

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

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

<pre>
<processing type="pullXCEL"
xcelRef="LiteralString">  
  <processingMethod name="setName">    
    <param value="documentAuthor"/>
  </processingMethod>
</processing>
</pre>
XCL approach

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

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
XCL approach
XCL approach

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

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

XCDL is built upon abstract models (X schemas) of

• Image
• Text
• Sound
• 3D
• ...

Manfred Thaller

Achievements: XCL

<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>
<symbol identifier="ID01_I01_I01_S04"
originalName="colourType">
  <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">
  <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 id="normAll />
      <propRel />
      <valueSet />
    </property>
  </property>
  <property id="p2" source="raw" cat="descr">
    <name>height</name>
    <valueSet id="i_i1_s3">
      <rawValue>0 0 1 ad</rawValue>
      <labValue>429</labValue>
      <dataRef id="normAll />
      <propRel />
      <valueSet />
    </property>
  </property>
  <property id="p3" source="raw" cat="descr">
    <name>imageType</name>
  </property>
</object>

XCL approach

- 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”.

Manfred Thaller
XCL approach

But why do we extract the actual data?

“Characteristics” are supposed to be akin to metadata?
XCL approach

Four building blocks:

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

**Compare tiffsuitBenchmark/output testpngBenchmark/output**

<table>
<thead>
<tr>
<th>XCDL</th>
<th>XCDL</th>
<th>RefTool</th>
<th>RefTool</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>bas0g01.xcdl</td>
<td>bas0g01.xcdl</td>
<td>b1</td>
<td>b2</td>
<td>failed</td>
</tr>
<tr>
<td>bas0g02.xcdl</td>
<td>bas0g02.xcdl</td>
<td>b1</td>
<td>b2</td>
<td>ok</td>
</tr>
<tr>
<td>bas0g04.xcdl</td>
<td>bas0g04.xcdl</td>
<td>b1</td>
<td>b2</td>
<td>ok</td>
</tr>
<tr>
<td>bas0g08.xcdl</td>
<td>bas0g08.xcdl</td>
<td>b1</td>
<td>b2</td>
<td>ok</td>
</tr>
<tr>
<td>bas0g16.xcdl</td>
<td>bas0g16.xcdl</td>
<td>b1</td>
<td>b2</td>
<td>ok</td>
</tr>
<tr>
<td>bas2e08.xcdl</td>
<td>bas2e08.xcdl</td>
<td>b1</td>
<td>b2</td>
<td>failed</td>
</tr>
<tr>
<td>bas2e16.xcdl</td>
<td>bas2e16.xcdl</td>
<td>b1</td>
<td>b2</td>
<td>ok</td>
</tr>
<tr>
<td>bas3p01.xcdl</td>
<td>bas3p01.xcdl</td>
<td>b1</td>
<td>b2</td>
<td>ok</td>
</tr>
<tr>
<td>bas3p02.xcdl</td>
<td>bas3p02.xcdl</td>
<td>b1</td>
<td>b2</td>
<td>ok</td>
</tr>
<tr>
<td>bas3p04.xcdl</td>
<td>bas3p04.xcdl</td>
<td>b1</td>
<td>b2</td>
<td>ok</td>
</tr>
<tr>
<td>bas3p08.xcdl</td>
<td>bas3p08.xcdl</td>
<td>b1</td>
<td>b2</td>
<td>ok</td>
</tr>
<tr>
<td>bas4a08.xcdl</td>
<td>bas4a08.xcdl</td>
<td>b1</td>
<td>b2</td>
<td>failed</td>
</tr>
<tr>
<td>bas6a08.xcdl</td>
<td>bas6a08.xcdl</td>
<td>b1</td>
<td>b2</td>
<td>failed</td>
</tr>
<tr>
<td>basn0g01.xcdl</td>
<td>basn0g01.xcdl</td>
<td>b1</td>
<td>b2</td>
<td>ok</td>
</tr>
<tr>
<td>basn0g02.xcdl</td>
<td>basn0g02.xcdl</td>
<td>b1</td>
<td>b2</td>
<td>ok</td>
</tr>
<tr>
<td>basn0g04.xcdl</td>
<td>basn0g04.xcdl</td>
<td>b1</td>
<td>b2</td>
<td>ok</td>
</tr>
<tr>
<td>basn0g08.xcdl</td>
<td>basn0g08.xcdl</td>
<td>b1</td>
<td>b2</td>
<td>ok</td>
</tr>
<tr>
<td>basn0g16.xcdl</td>
<td>basn0g16.xcdl</td>
<td>b1</td>
<td>b2</td>
<td>failed</td>
</tr>
<tr>
<td>basm2c08.xcdl</td>
<td>basm2c08.xcdl</td>
<td>b1</td>
<td>b2</td>
<td>ok</td>
</tr>
</tbody>
</table>
XCL approach

► Photoshop ►

Manfred Thaller
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.
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?

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

\[ \bullet + \bullet\bullet\bullet = \square \]
What is a model of information?

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

\[
\bullet + \bullet\bullet\bullet\bullet = \bullet\bullet\boxempty
\]

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

\[
\bullet + \bullet\bullet\bullet\bullet = \bullet\bullet\boxempty
\]
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>
    </valueSet>
    <dataRef ind="normAll" />
    <propRel/>
</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>
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>
</rawVal>
</property>
<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.”
Motto

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

*In the ground, in case it is muddy.

Manfred Thaller

Thank you!