Preservation of image documents

Theory & Case Study

Chinese-European Workshop on Digital Preservation

Beijing July 14 -16 2004

rene.van.horik@niwi.knaw.nl

niwi

Issues covered in presentation

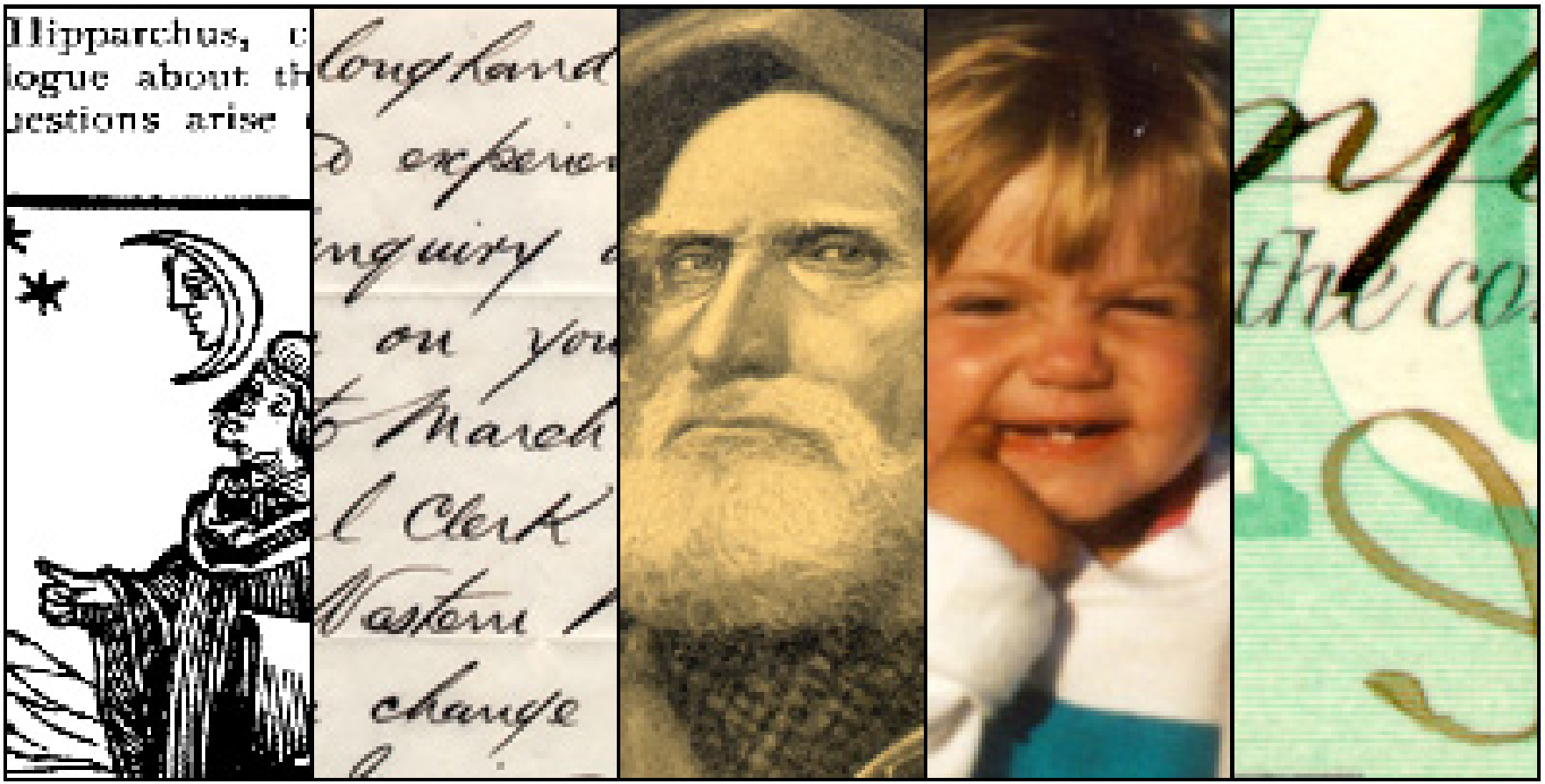
- What are image documents?
- Theories on preservation of digital image documents
- Practices to preserve image documents

What are image documents?

"Graphics files can be considered as files that store any type of persistent graphics data (as opposed to text, spreadsheet, or numerical data, for example), and that are intended for eventual rendering and display."

(Murray & van Ryper, *Encyclopedia of graphics file formats* (O'Reilly) 1994)

niwi



Source: Kenney, A. and O. Rieger, *Moving theory into practice. Digital imaging for libraries and archives* (Cornell University Library, 2000.

Graphic File Formats

... - PCX – PBM – TGA – TIFF – GIF – JPEG – PSD – DXF - CGM – PNG – SVG – RAW – WPG – FITS – BMP – PCD – RAS – TGA – BPS – EPS – PDF – PCT – WBM – FITS – XBM – VFF – RIB – PCX – DMP – AVS – IMG – ICO – JFIF – IFF – WMF - ...

niwi

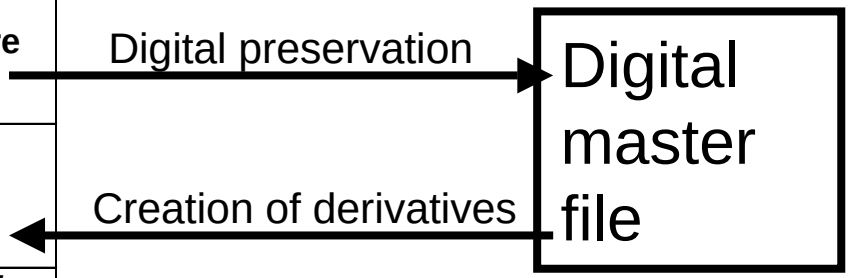
Why are there so many different graphic file formats?

- There are a number of fundamental different types of graphical data
  - raster data (sampled values)
  - geometry data (mathematical description of space)
  - latent image data (data transformed into useful images by some algorithmic process)
- To prevent usage beyond control of the developer (Who remembers KodakPhoto CD?)
- Wide range of design principles (Mainly ‘speed’ and ‘memory’)

niwi

Digital master images

	Conventional photography	Digital photography
step 1. Capture	**'Snapshot'**	**Scanning / digital capture**
step 2. Process	**Chemical film processing**	**Image processing**
step 3. Output	**Imprint**	**Screen view / Printer output**

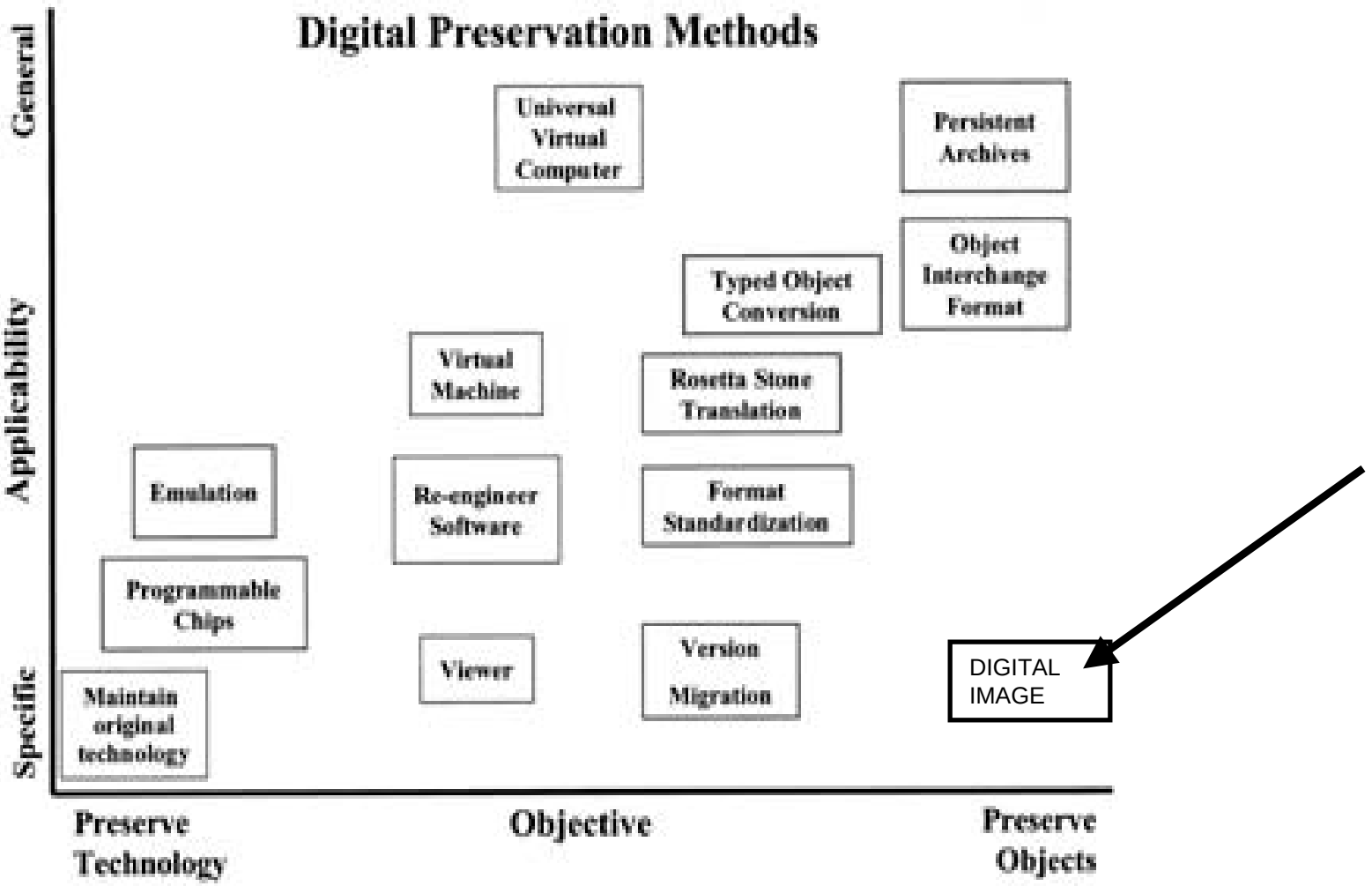


Three steps in the photographic process

How to create digital master images? See "Guides to quality in visual resource imaging" http://www.rlg.org/visguides

Theories on digital preservation

- Based on assumptions such as:
  - XML is the only durable storage format
  - Metadata is essential
  - Standards will do the job
  - Data storage media is robust
  - Registries are essential
  - Etc.
- Only the future can judge which assumptions were right...



T. Thibodeau 'Overview of technological approaches to digital preservation and challenges in coming years' (CLIR report)
<http://www.clir.org/pubs/reports/pub107/pub107.pdf>

‘Building blocks’ for the long term preservation of digital images

(Building blocks: procedures, tools, standards, specifications and guidelines available to realize the long term access of digital images)

1. Standard graphics file formats
2. Bitstream preservation (by using XML data format)
3. Preservation metadata

niwi

Assumptions

- Standards are durable, e.g. image file format standards
- Digital data encoded in the XML data format is durable data
- Metadata on digital objects is essential in order to understand and process digital objects in the future

Features of standard image file formats

1. Used by large community during a considerable period of time
2. Specifications must be in the public domain or published by SDO (standards developing organization)
3. Wide range of systems has to support the format
4. No data compression (loss of quality / higher risk)
5. Must contain facilities to store preservation metadata
6. Must enable coding of all significant characteristics of analogue original

Durability requirements and raster file formats

	Raster file requirements	**TIFF**	**JPEG**	**GIF**	**PNG**
1	Used by a large community over a long time	+	+	+	-
2	File format specification is published	+	+	+	+
3	Supported by a wide range of applications	+	+	+	+
4	Supports un-compressed / single page images	+	-	-	-
5	Facilities for preservation metadata	+	-	-	+
6	Enables “full informational capture”	+	-	-	+

File formats described in Murray & vanRyper, *Encyclopedia of graphics file formats* (O’Reilly) published in 1994 and still used in 2004.

Example: TIFF file

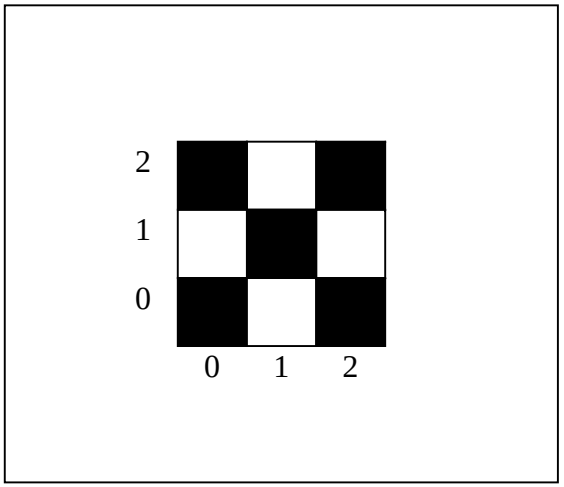
```
00208a.tiff:
Magic: 0x4949 <little-endian> Version: 0x2a
Directory 0: offset 132610346 (0x7e7792a) next 0 (0)
ImageWidth (256) SHORT (3) 1<9681>
ImageLength (257) SHORT (3) 1<6849>
BitsPerSample (258) SHORT (3) 1<16>
Compression (259) SHORT (3) 1<1>
Photometric (262) SHORT (3) 1<1>
DocumentName (269) ASCII (2) 17<ppprs/00208a.tif\0>
StripOffsets (273) LONG (4) 2283<8 58094 116180 174266 232352 290438
348524 406610 464696 522782 580868 638954 697040 755126 813212 871298
929384 987470 1045556 1103642 1161728 1219814 1277900 1335986 ...>
Orientation (274) SHORT (3) 1<1>
SamplesPerPixel (277) SHORT (3) 1<1>
RowsPerStrip (278) SHORT (3) 1<3>
StripByteCounts (279) LONG (4) 2283<58086 58086 58086 58086 58086 58086
58086 58086 58086 58086 58086 58086 58086 58086 58086 58086 58086 58086
58086 58086 58086 58086 58086 58086 ...>
XResolution (282) RATIONAL (5) 1<1425>
YResolution (283) RATIONAL (5) 1<1425>
PlanarConfig (284) SHORT (3) 1<1>
ResolutionUnit (296) SHORT (3) 1<2>
DateTime (306) ASCII (2) 20<2000:03:21 21:36:29\0>
Artist (315) ASCII (2) 20<Library of Congress\0>
```

XML: eXtensible Markup Language

- Information interchange format
- Standard, developed by World Wide Web consortium (http://www.w3c.org/xml)
- Application independent
- No pre-defined markup tags (extensible)
- Both human and machine understandable

niwi

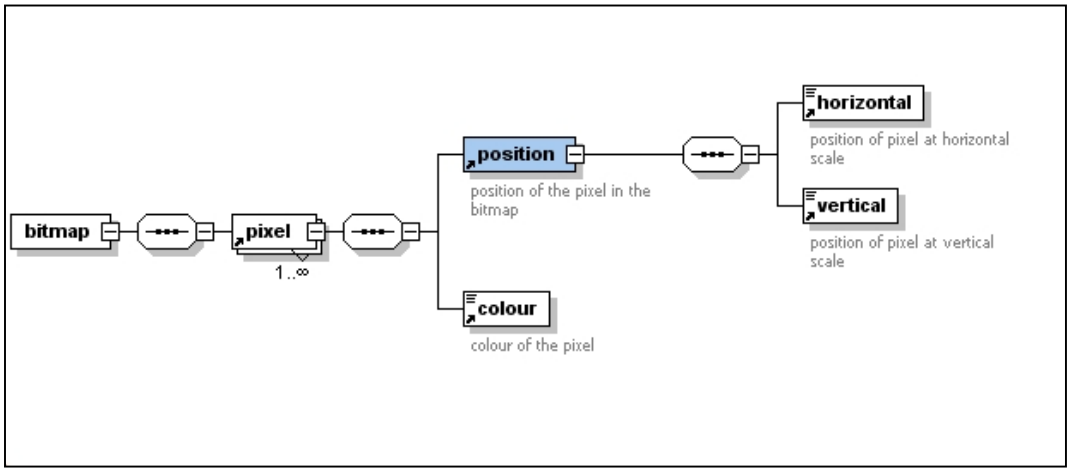
Durable encoding of the bitstream



Bi-tonal bitmap consisting of 9 pixels

```
<bitmap>
   <pixel>
     <position>
        <horizontal>0</horizontal>
        <vertical>0</vertical>
   </position>
        <colour>black</colour>
   </pixel>
   <pixel>
     <position>
        <horizontal>0</horizontal>
        <vertical>1</vertical>
    </position>
        <colour>white</colour>
    </pixel>
    ...
</bitmap>
```

Bitmap expressed in XML

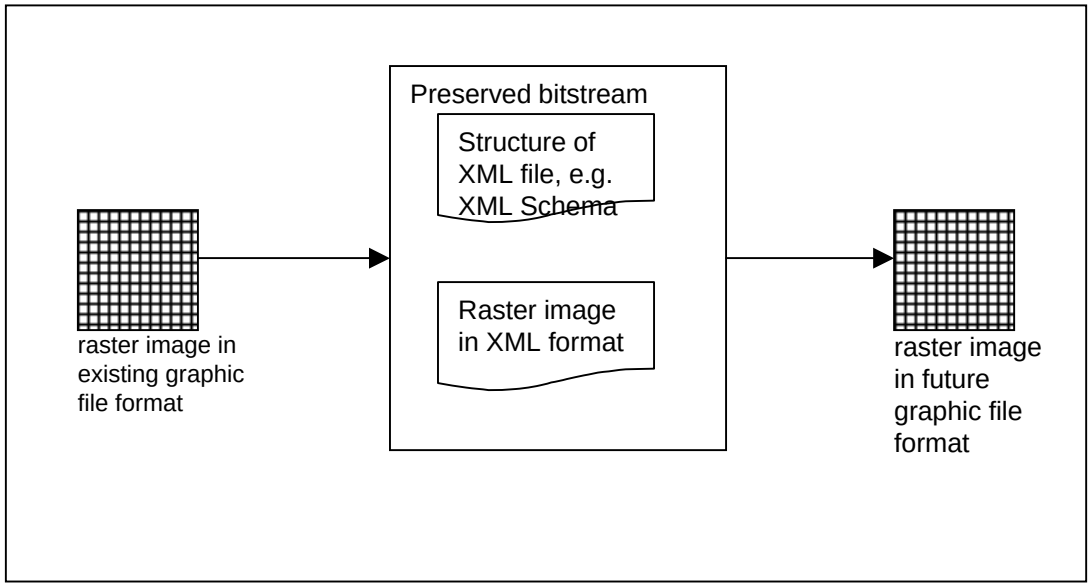


Digital image expressed in XML

- Expression of content model in XML
  - Elements and attributes that are part of the bitstream, e.g. standardized color coding of pixels
- Binary to XML conversion
  - Conversion of image format (e.g. TIFF) into XML
- XML to binary conversion
  - In the future

niwi

Components of preserved bitstream in XML format



Components of preserved bitstream in XML format

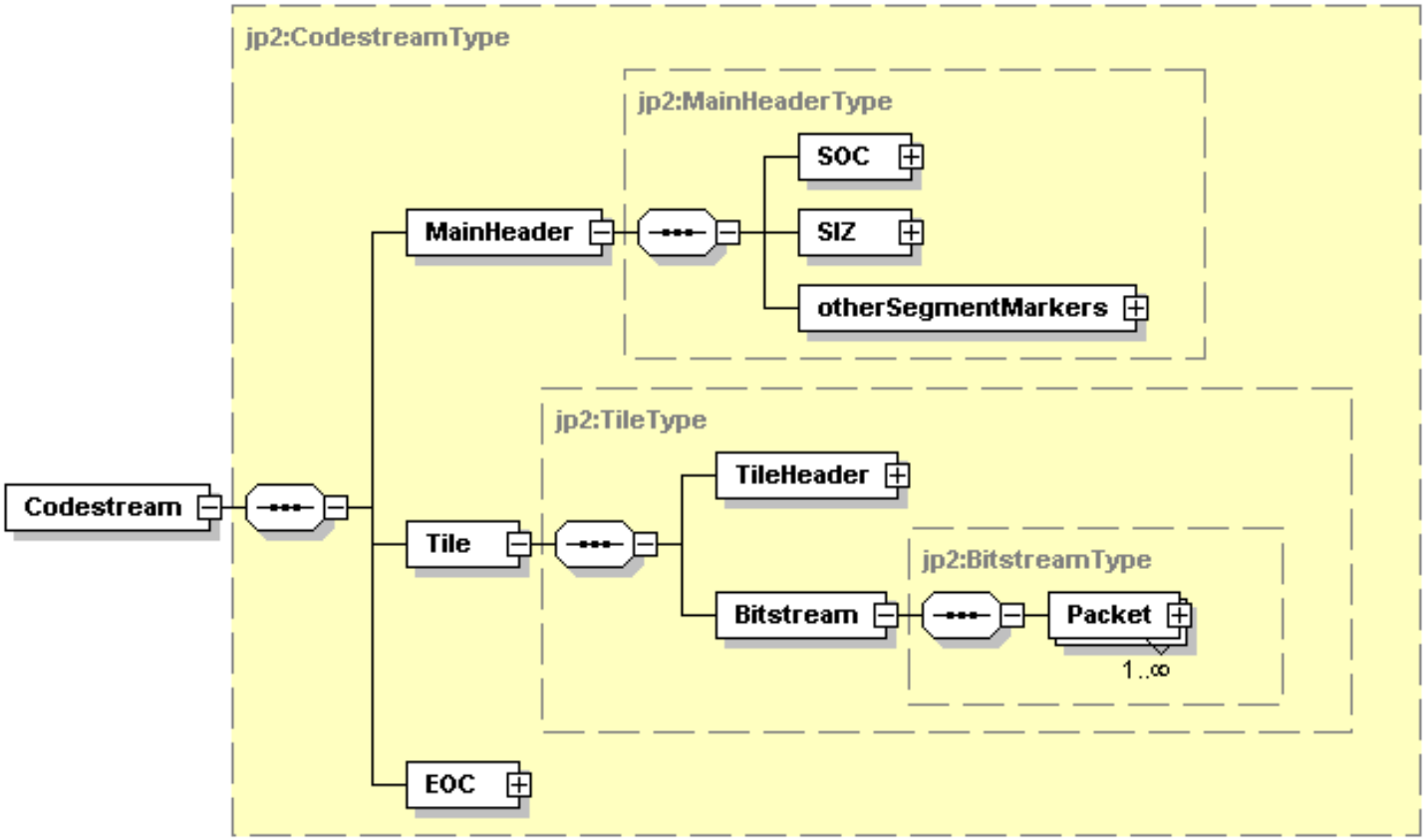
Methods available to express image in XML format

- Bit stream syntax description language (BSDL)
- Universal Virtual Computer (UVC)
- Formal language for audio-visual object representation (Flavor / Xflavor)

Bitstream syntax description language (BSDL)

- XML-based language which forms part of a methodology designed to provide a generic solution at bitstream level to adopt a single media element
- Described high level structure of a bitstream. Each format requires specific content model
- Thorough knowledge required on the way the bits are organized
- Absence of “binary to XML” and “XML to binary functionality

niwi



Example: BSDL Schema of a JPEG2000 image

Universal virtual computer (UVC)

- Bitstream representing the data is stored together with the logical view of the data
- Also specification to process data on a future platform is archived
- Processing specification based on UVC (= interpreter independent of computer architecture)

niwi

Formal language for audio-visual object representation (Flavor / Xflavor)

- Developed for the description of binary multimedia objects
- Xflavor: application of XML in order to simplify interoperability among different applications

Comparison of 3 methods

	Binary to XML conversion	Content model in XML	XML to Binary conversion
BSDL	-	++	-
UVC	++	+	++
XFlavor	+	+	+

- \- task is available in system design
- \+ task can be performed with the method, but adjustments are required to enable the processing of digital master images
- ++ task can be performed by the method

niwi

Metadata

- Three ways to store metadata on digital images:

1. As part of the image (e.g. File header)
2. In separate database
3. In file system (/images/thumbnails/2003/05/…)

Application of preservation metadata

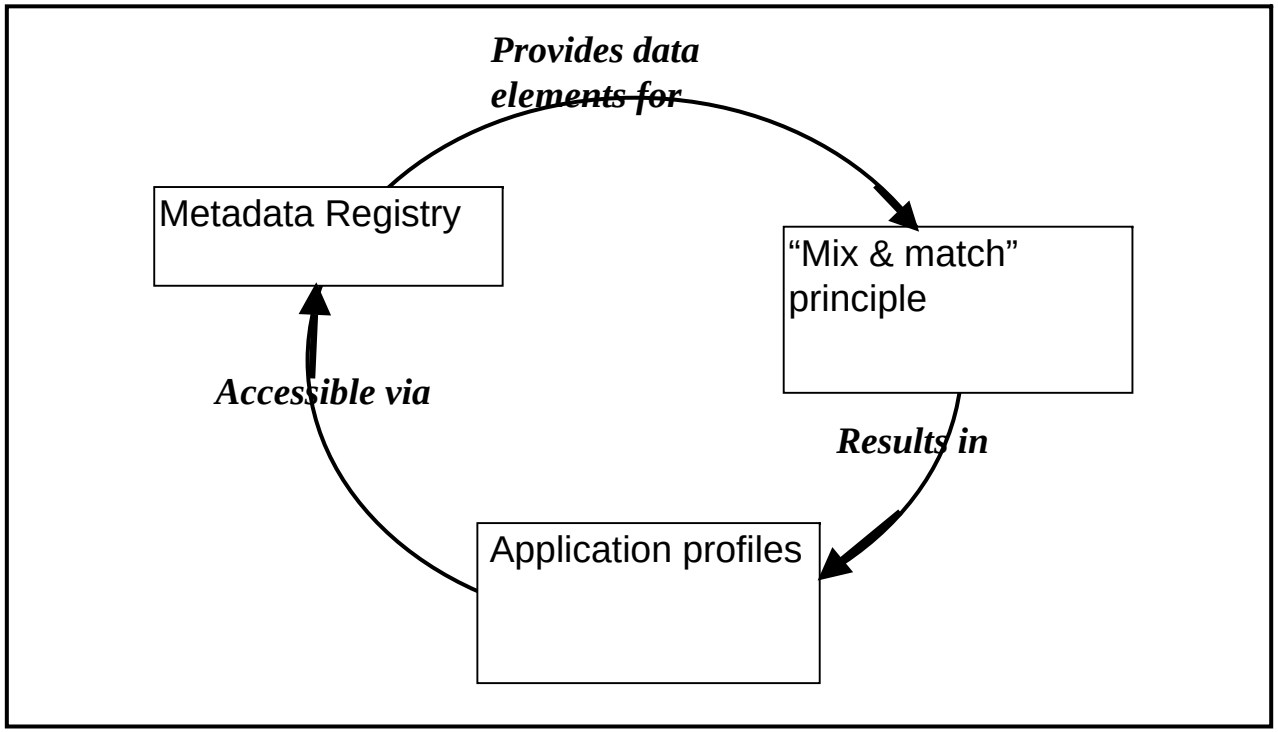
- Two methods:
  - Create from scratch
  - (Re)use existing data elements

(data element: unit of data for which the definition, identification, and permissible values are specified by means of as set of attributes (ISO/IEC 11179))

Some Metadata elements sets

- NISO Z39.87-2002/AIIM 20-2002, *Data Dictionary – Technical Metadata for Digital still images*, 2002 <http://www.niso.org/standards/resources/Z39_87_trial_use.pdf>
- EXIF 2.2, *Exchangeable image file format for digital still cameras*, April 2002 <www.exif.org>
- SepiaDES (Sepia Description Element Set): metadata element set for historical photographic collections <http://www.knaw.nl/ecpa/sepia.html>
- Etc.

Metadata registries



Ref. R. Heery and M. Patel, 'Application profiles: mixing and matching metadata schemas' in: *Ariadne*, vol. 5 (2000) <http://www.ariadne.ac.uk/issue25/app-profiles/>

Conclusions

- TIFF image file format often used as format for digital master image (Adobe Systems Incorporated, *TIFF revision 6.0*, Final – June 3, 1992 <http://partners.adobe.com/asn/developers/pdfs/tn/TIFF6.pdf>)
- Bitstream in XML format: more research required
- Preservation metadata: application profiles & registries help to ‘discriminate exactly what we know vaguely’

"Theory without practice is empty.
Practice without theory is blind"

John Dewey

Digital preservation solutions

- Format registry (e.g. GDFR, PRONOM)
- Format identification (e.g. Jhove)
- Digital archiving (e.g.VERS)
- Distributed storage (e.g. OAI, Lockss)
- Emulation (e.g. UVC)
- Etc.

Digital preservation practices

- Several organisations committed themselves to preserve digital objects.
- Relatively recently started (but scientific data archives, holding datasets, exist for more than 25 years!)
- Examples:
  - NARA: Transfer of permanent E-records
  - KB the Netherlands: e-Depot
  - Harvard University: DRS
  - Etc.
- In common: commitment!

Practices

- Usage of microfilm as archival medium!
- Risk management (G. Lawrence, R. Kehoe, O. Rieger, W. Walters, and A. Kenney, *Risk management of digital information: A file format investigation* (Washington, DC: CLIR, 2000) <http://www.clir.org/pubs/reports/pub93/contents.html>)
- Practice depends on project (characteristics of originals, budget, skills, purpose, etc.)

(Example)
Digital image of historical photograph + Metadata

This digital reference image is created by London Metropolitan Archives (LMA). The image is a derivative of a digital master file stored on CD-ROM and archived by LMA. The original photograph on which this image is based is stored under inventory number SC/PHL/02107976/1167248.0X73/73. The name of the image is L13071AR. The image is stored in the “jpeg” format. The copyright is owned by LMA. The image has the following title: “Open Spaces Committee on a visit”. The original photograph is created in 1884. The reference image is 400 pixels wide in the horizontal dimension. The original photo is taken by an unknown photographer employed by the Greater London Council. A reproduction of this image on basic paper (100gms paper printed @ 360 dpi) costs £2.60. A photographic print (from negative as per LMA reprint service) costs from £12.75. Etc. etc.

(Example cont.)
Metadata in XML format according to DC syntax

```
<metadata>
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    <dc:title> lang="eng"
        Open Spaces Committee on a visit
    </dc:title>
    <dc:description>
        Metropolitan Board of Works: Parks, Commons and Open Spaces
        Committee on a visit
    </dc:description>
    <dc:date>
        1884
    </dc:date>
    <dc:creator>
        Greater London Council
    </dc:creator>
    <dc:identifier>
        http://www.lma.uk/data/images/L13071AR
    </dc:identifier>
    <dc:publisher>
        London Metropolitan Archives
    </dc:publisher>
    <dc:keywords>
        Parks
    </dc:keywords>
</metadata>
```

Example (cont.) Using distributed architecture as part of digital archiving solution for digitised historical photographs.