

# file format features and significant properties

Andreas Aschenbrenner, ERPANET

0010010 10 111010 1001

0 01 . .

1 0

01

1

0

11010 01

0

1

1

..

# overview

---

1. file format features
2. significant properties
3. file formats in digital preservation

0010010 10 111010 1001

0 01 . .

1 0

01

1

0

11010 01

0

1

1

..

# What is a file format?

A format is a fixed, byte-serialized encoding of an information model.

An information model is a formal expression of exchangeable knowledge.

( Abrams, Seaman: Towards a global digital format registry. IFLA 2003 )

0010010 10 111010 1001

0 01 . .

1 0

01

1

0

11010 01

0

1

1

..

# format examples

```
<FONT SIZE=3 STYLE="font-size: 13pt">Chinese-European Workshop
on Digital Preservation</FONT></H1><H1 LANG="en-GB"
CLASS="western" ALIGN=CENTER STYLE="margin-left: -0.64cm"><FONT
SIZE=3 STYLE="font-size: 13pt">Beijing, July 14&ndash;16,
2004</FONT><P LANG="en-GB" CLASS="western" ALIGN=CENTER
STYLE="margin-left: -0.64cm; margin-bottom: 0cm">Preliminary
Agenda, Version 1.0</P><FONT SIZE=2 STYLE="font-size:
11pt"><B>Day 1 (14 July 2004)<P LANG="en-GB" CLASS="western"
ALIGN=LEFT STYLE="margin-bottom: "><B>Introduction</B></P><P
LANG="de-AT" CLASS="western" ALIGN=LEFT><SPAN LANG="en-
GB">Moderator:<I>Xiaolin Zhang</I>,
<I>CSDL/CAS</I></SPAN></P><SPAN LANG="en-GB">Welcome and
Introduction</SPAN></P><P LANG="de-AT" CLASS="western"
ALIGN=LEFT><I><SPAN LANG="en-GB">Longji Dai,
CALIS.</SPAN></I></P></TD></TR><TR VALIGN=TOP><TD WIDTH=47
ALIGN=LEFT><SPAN LANG="en-GB">9:15-9:30</SPAN></P></TD><TD
WIDTH=576 BGCOLOR="#f3f3f3"><P LANG="en-GB" CLASS="western"
ALIGN=LEFT STYLE="margin-bottom: 0cm">Opening Remarks</P>
duction0Longji Dai, CALIS.009:15-9:300Opening Remarks0Haibo Y
```

# categorisation

**Document** - DOC, HTML

**Page description language** - PDF, Postscript

**Raster Images** - TIFF, PNG, JPEG

**Structured graphics** - CAD, VSD, QXD

**Audio** - WAV, MP3, MIDI

**Video** - MPEG, AVI

**Spreadsheets** - XSL

**Databases** - DBF, MDB

**Configurations, metadata** - CSS

**Raw data**

**Collections** - tar, zip

**Program-supporting** - TTF, game saves

**Object code** - EXE, COM

( wikipedia.org; Clausen, May 2004 )

0010010 10 111010 1001

11010 01

0 01 . . 1 0

1 0 01

0  
1  
1  
..

# format features



This is our document format ...

# format features



```
<doc>  
  This is our document format ...  
</doc>
```

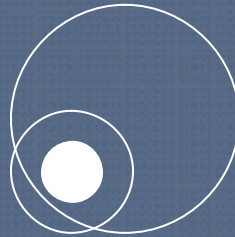
# format features



```
<doc>  
  This is our <b>document</b> format ...  
</doc>
```

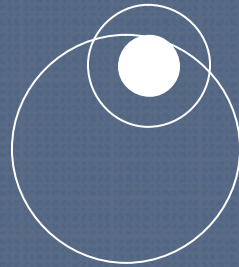


# format features



```
<doc>  
  This is our <b>document</b> format ...  
  <table><raw> with tables </raw></table>  
</doc>
```

# format features



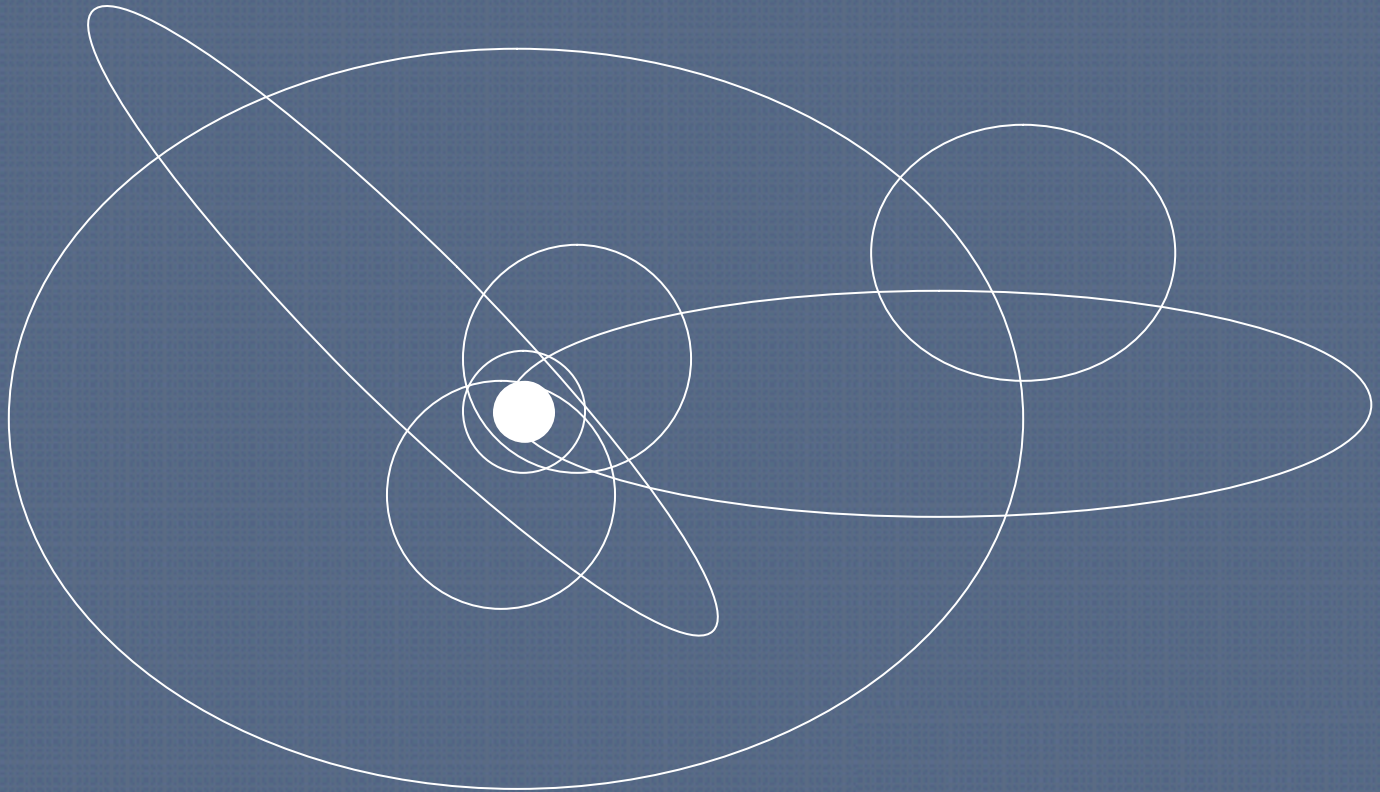
```
<doc>
```

```
  This is our <b>document</b> format ...
```

```
  or <img>images <source>image/filename</source></img>
```

```
</doc>
```

# format features



-> document formats <-



# format features

digital watermark  
password  
...  
...  
...  
flow chart  
rotation  
text  
sound  
...  
animations  
dynamic inlays  
...  
...  
layers  
transparency  
clipart  
colour spaces  
...  
revision history  
interactive graph  
...  
...

-> **image formats** <-

# feature classification

- appearance
- structure
- behaviour  
( ? A.Rauber presentation)

- readability
- comprehensibility
- appearance
- functionality
- look + feel  
( Clausen, May 2004 )

0010010 10 111010 1001

0 01 . .

1 0

01

1

11010 01

0  
1  
1

..

# lesson learnt

the **simpler** the object (**fewer** features),  
the **more manageable** the preservation challenge

0010010 10 111010 1001

0 01 . .

1 0

01

11010 01

0  
1  
1

..

1 0

format

functional  
technological



# significant properties

*essence of an object*

*properties of  
conceptual objects*

significant properties

*essential characteristics*

*core features*

*aspects of preservation quality*

0010010 10 111010 1001

0 01 . .

1 0

1

11010 01

0  
1  
1

..

01

# significant properties - the intention

important

- to fulfil a business function
- to adequately purport meaning

depends on a specific task  
in a specific business environment

0010010 10 111010 1001

0 01 . .

1 0

01

11010 01

0  
1  
1

..

# significant properties - the medium

<b>this</b>					<b>take</b>			<b>this take</b>
<b>If</b>	<b>is</b>		<b>shall</b>		<b>really</b>	<b>to</b>		<b>If is shall really to</b>
	<b>flying I</b>			<b>never</b>		<b>it.</b>		<b>flying I never it.</b>

Piglet getting bounced along by Kanga,  
*Winnie The Pooh*, A. A. Milne, p. 103

... transcribed as linear text

text

formatting

page format, pagination

headlines, footnotes

metadata

Jeff Rothenberg (May 2001):  
Emulation vs. Migration for  
Preserving Software-Dependent  
Artefacts.

0010010 10 111010 1001

11010 01

0 01 . . 1 0

1 0

01

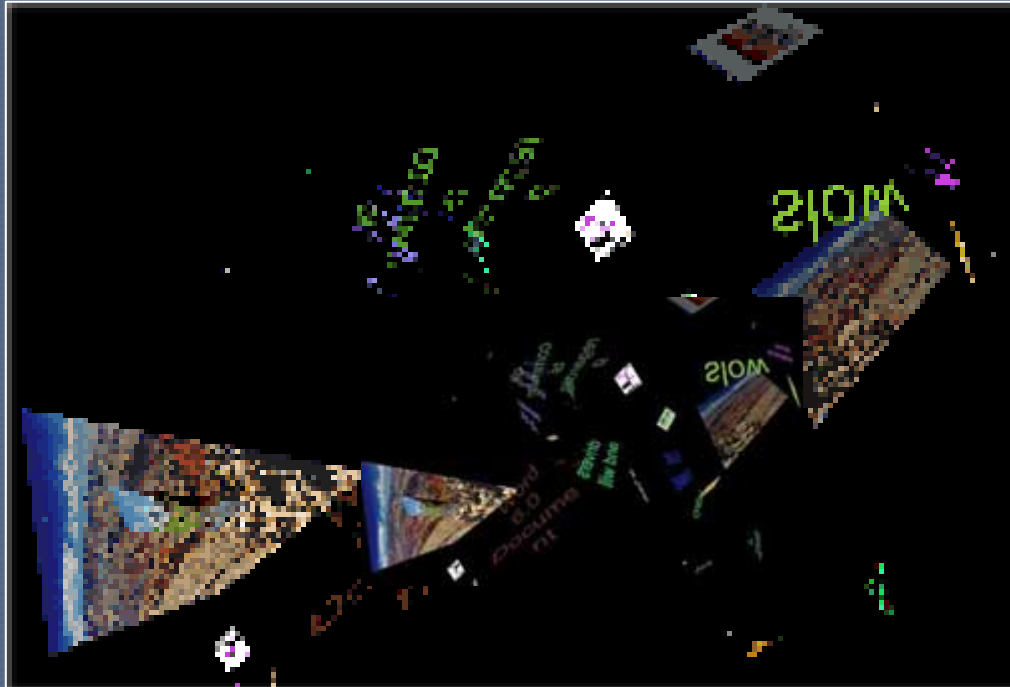
..

1

1

0

# significant properties - a comparison



Mary Flanagan: [phage]

[www.maryflanigan.com/virus.htm](http://www.maryflanigan.com/virus.htm)

[www.variablemedia.net/e/seeingdouble/home.html](http://www.variablemedia.net/e/seeingdouble/home.html)

0010010 10 111010 1001

11010 01

0 01 . .

1 0

1 0 01

0  
1  
1

..

# significant properties

fundamental purpose, intellectual content

preserve the idea? or preserve the performance?

- ? intention of the author
- ? mission of preserver
- ? requirements of future users

0010010 10 111010 1001

0 01 . .

1 0

01

11010 01

0  
1  
1

..

1 0

# lesson learnt

the **simpler** the object (**fewer** features),  
the **more handable** the preservation challenge

preserve as **simple** objects as possible,  
but as **rich** as necessary to adequately  
carry their purpose into the future

0010010 10 111010 1001

0 01 . .

1 0

01

1

0

11010 01

0

1

1

..

business

environment

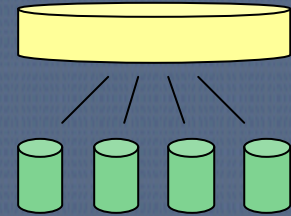
task

significant properties  
format features

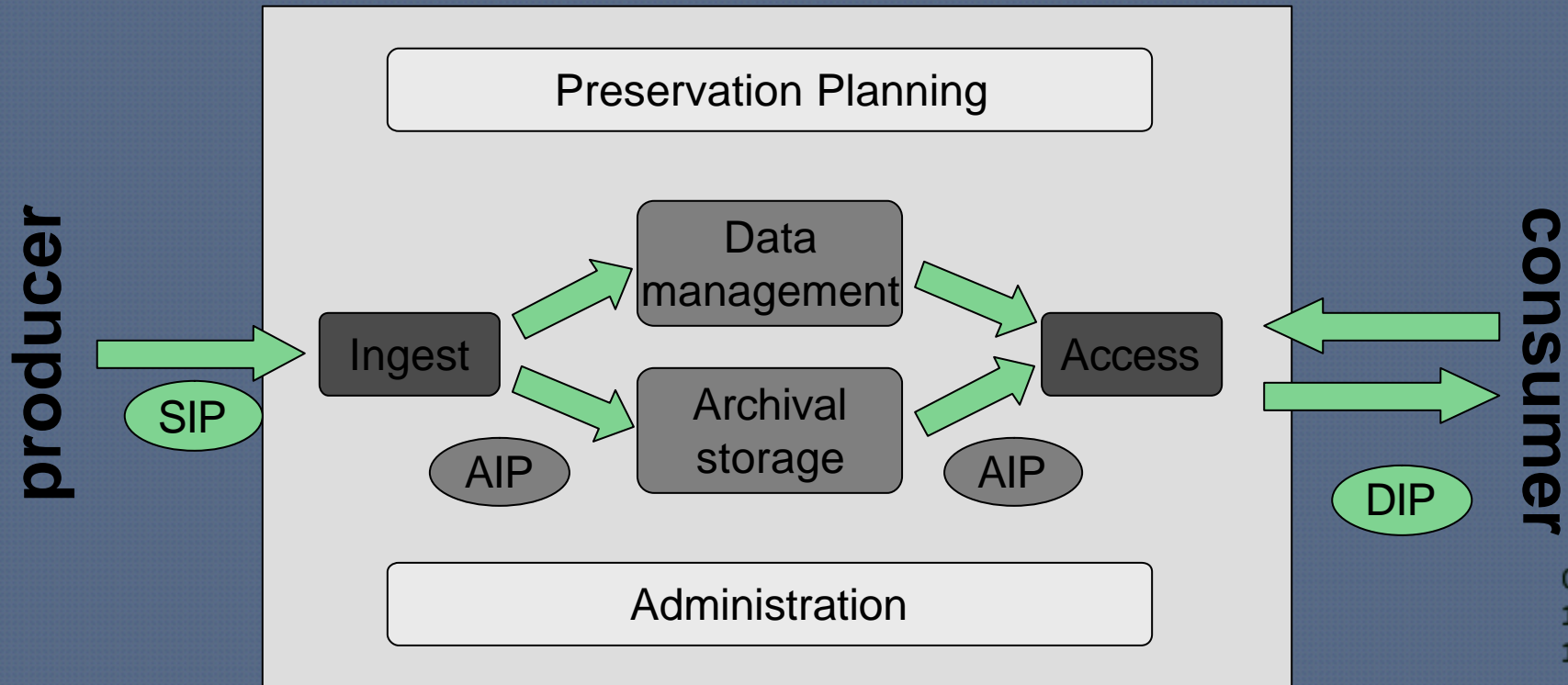
format

functional

technological



# OAIS model



- SIP** Submission Information Package
- AIP** Archival Information Package
- DIP** Dissemination Information Package

0010010 10 111010 1001

0 01 . .

1 0

01

11010 01



# file formats in digital preservation

possible indicators for long-term stability

- stable, robust
- portable
- open
- ? widely used and well documented

0010010 10 111010 1001

0 01 . .

1 0

01

11010 01

0  
1  
1

..

1 0

# conversion

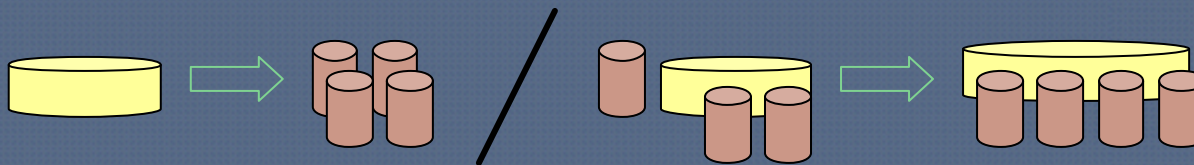
## moving between formats

\* SIP ? AIP

\* AIP ? DIP

## conversion problems:

format features; significant properties; process



0010010 10 111010 1001

0 01 . .

1 0

01

1

1

..

01

1

0

11010 01

# more format issues in preservation

- \* multiple objects, external libraries
- \* nested files
- \* XML inside? /.
- \* encryption, compression
- \* metadata

( XML alone is only syntax for information model )

0010010 10 111010 1001

0 01 . .

1 0

01

..

1

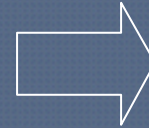
1

0

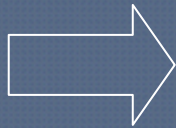
11010 01

# XML inside?

```
<?xml version="1.0" encoding="UTF-8"?>
<workshop>
  <location>beijing</location>
  <date>july 2004</date>
  <topic>digital preservation</topic>
</workshop>
```



**XML  
database**



```
^A ^E@href^V^Etitle
^Fgroups^B^F@title^Y^Dauth^A^Duse^Csup^]^Glast-
id^L^Cuid^O
^C@id^F^Dname^category^Z^Eusers^G^Ab^document^T^Dlink^
\^ApESlocation^Q^Ginclude
^U^Etopic^workshop^P^Dhref^_ @password
@last-id^C^Ddate^R^D@uid
^Dbody^W^E@name^E^Egroup^D^Bid^K^Gsection^passwo
rd^N^A^_http://www.w3.org/2001/XInclude^A^A^Bxi^A
```

0010010 10 111010 1001

0 01 . .

1

11010 01

# conclusions

- ! think about what you are doing
- ! watch what's going on
- ! check closely
- ! document it
- ! ensure you have all the information you need

0010010 10 111010 1001

0 01 . .

1 0

01

11010 01

0  
1  
1

..

1 0

# further reading

- \* ERPANET Training Seminar:  
File Formats for Preservation.  
May 2004; Vienna, Austrian National Library.  
[www.erpanet.org](http://www.erpanet.org)
- \* Lars R. Clausen: Handling File Formats. The Royal Library, Copenhagen, and The State and University Library Aarhus, Denmark. May 2004.  
<http://www.netarkivet.dk/website/publications/FileFormats-2004.pdf>
- \* Adrian Brown: Selecting File Formats for Long-Term Preservation. UK National Archives Digital Preservation Guidance Note 1. June 2003.  
[www.pro.gov.uk/about/preservation/digital/guidance/selecting-file-formats.pdf](http://www.pro.gov.uk/about/preservation/digital/guidance/selecting-file-formats.pdf)
- \* John Bennett: A framework of Data Types and Formats, and Issues affecting the long term Preservation of Digital Material. 1997.  
<http://www.ukoln.ac.uk/services/elib/papers/supporting/pdf/rept011.pdf>
- \* Lawrence et al.: Risk Management of Digital Information: A File Format Investigation. June 2000. CLIR Report, ISBN 1-887334-78-5.  
<http://www.clir.org/pubs/reports/pub93/pub93.pdf>