

The Role of File Formats in Digital Preservation: Opportunities and Threats

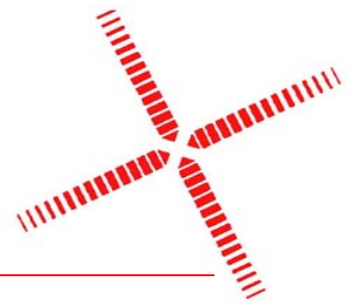
ErpaTraining on File Formats for Preservation
Vienna, May 10-11, 2004

Frank Moehle
Swiss Federal Archives
Project ARELDA

Frank.Moehle@bar.admin.ch

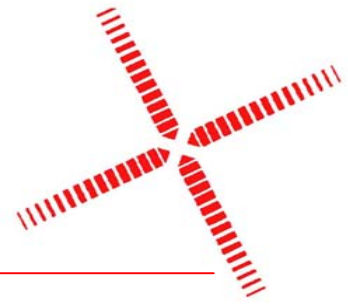
www.bundesarchiv.ch

Agenda

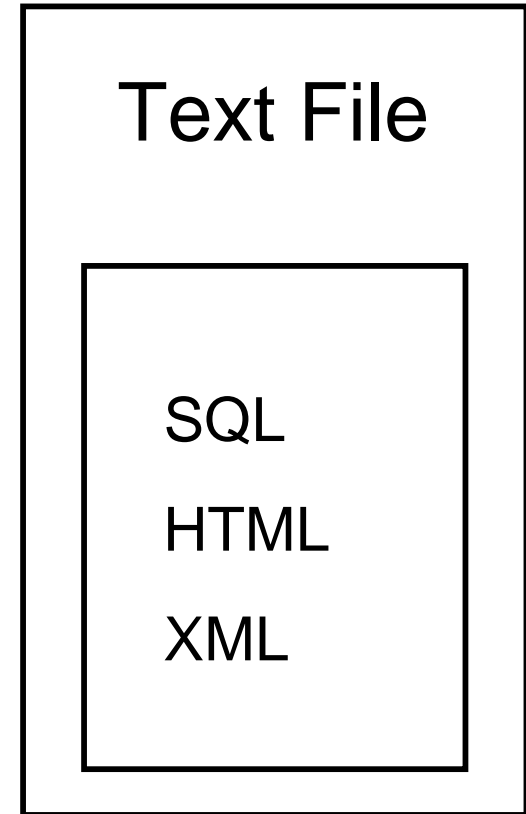


1. Introduction to File Formats
2. Text like Documents
3. Images and Graphics
4. Audio and Video
5. Conclusions

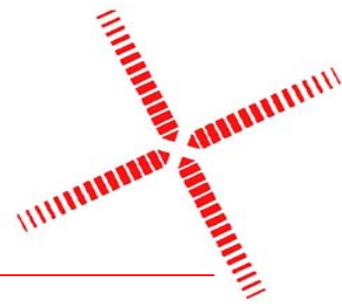
Data vs. File Format



- Some popular data formats:
HTML, SGML, XML,
SQL, etc.
- But all of them can
be stored in the
same file format...

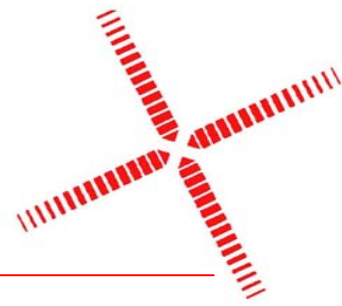


What a File Format is



- A **file** is nothing but a sequence of bits
- A file **format** defines how to interpret the contents of a file.
- Typical file formats have a file header (metadata), followed by actual (primary) data

Text vs. binary files



P1

8 8

```
0 0 1 1 0 0 1 1
1 1 0 0 1 1 0 0
0 0 1 1 0 0 1 1
1 1 0 0 1 1 0 0
0 0 1 1 0 0 1 1
1 1 0 0 1 1 0 0
0 0 1 1 0 0 1 1
1 1 0 0 1 1 0 0
```

P4

8 8

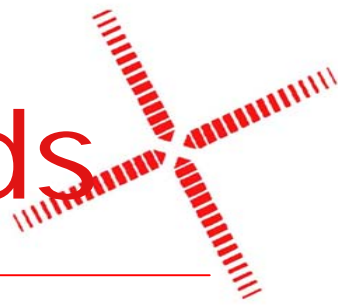
aFaFaFaF

Proprietary vs. Open Formats



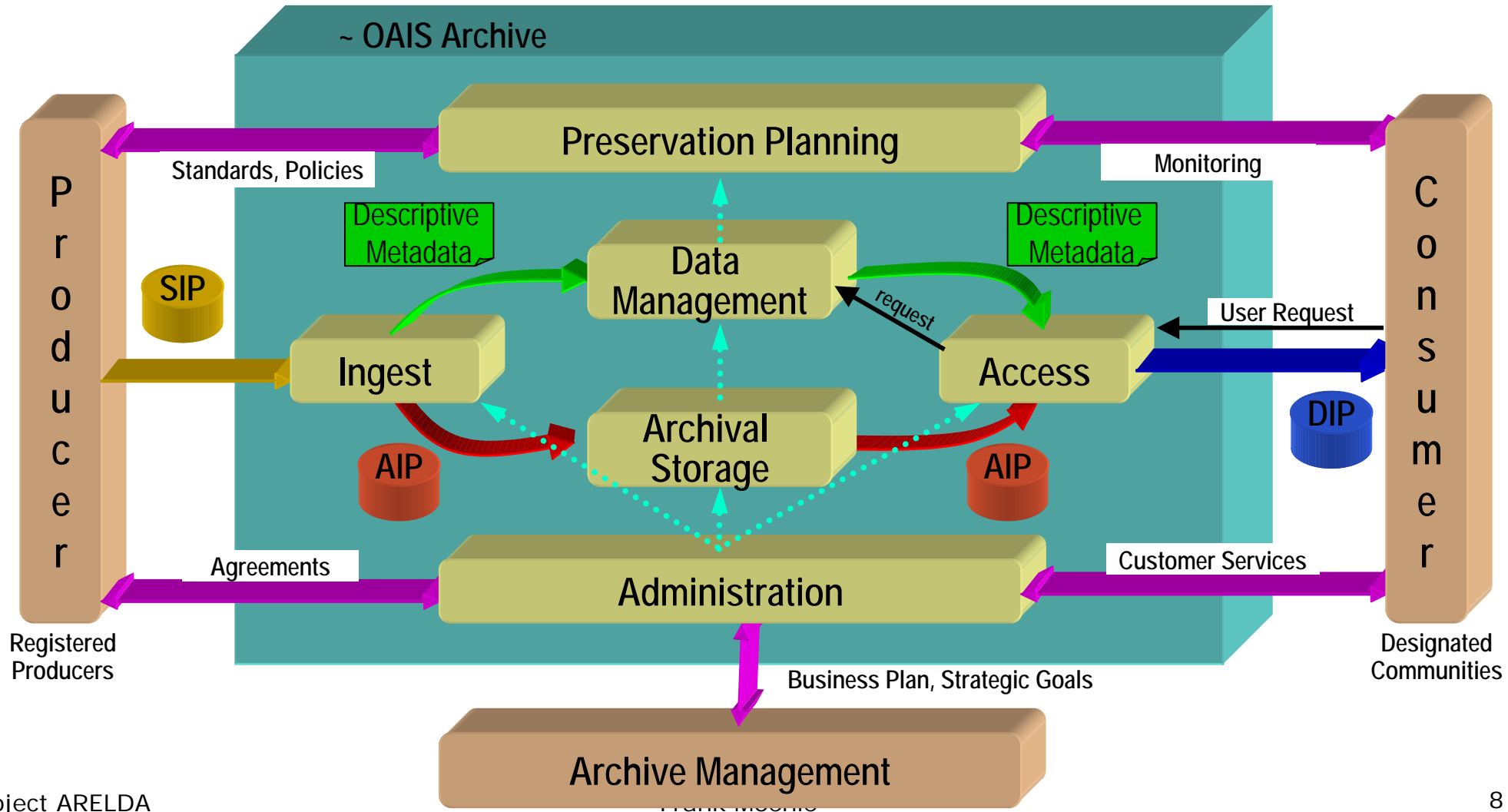
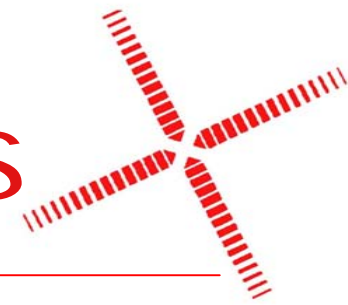
- Proprietary
 - Full Documentation not always available
 - License and patent rules may apply
 - License agreements subject to change
 - Restrictions for use and modifications may apply
- Open
 - Unlimited use
 - No license fees
 - No patent owners
 - Full Documentation Available
 - Open for self-made modifications

Digital Preservation Needs

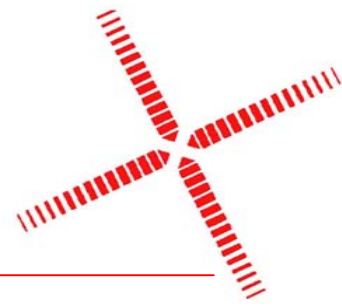


- **Integrity** uncorrupted, safe storage
- **Understandability** Intellectual understandability of content and context
- **Originality** Data structure and appearance
- **Authenticity** Authentication of author and authentication of provenance and evidence
- **Accessibility** Readable and usable

The Preservation Process

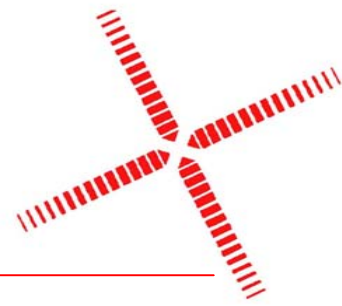


The Ingest Process



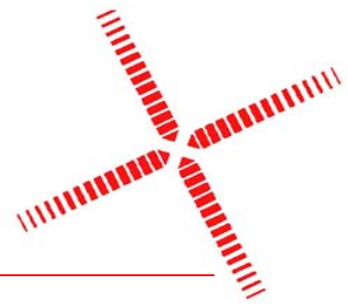
- The ingest process needs
 - Unambiguous file format specs
 - Validation tools for quality assurance
 - File format converters
- And is supported by
 - Standardization organizations (ISO, ...)
 - Registries (PRONOM)

Kind of File Formats



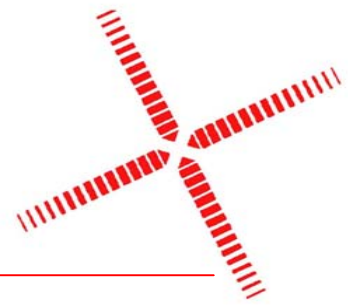
- The diversity of file types requires a whole set of file format specs for
 - Text documents
 - Data bases
 - Still and animated graphics
 - Audio recordings
 - Video sequences
 - Etc.

Requirements



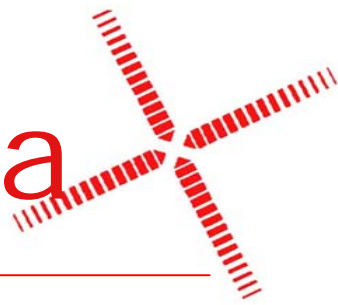
- Syntactical and semantical specs are public
- Standardized by an recognized organization like ISO, ANSI, ITEF, ...
- Widely used and accepted
- Free of patent and license fees

Requirements (cont'd)



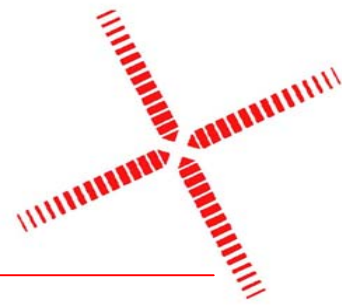
- Free of any cryptographical techniques (risk of loosing keys)
- No compression techniques (loss of reduncancies)
- Specification should be self-contained
- File format should be media-independent

Migration of Archival Data



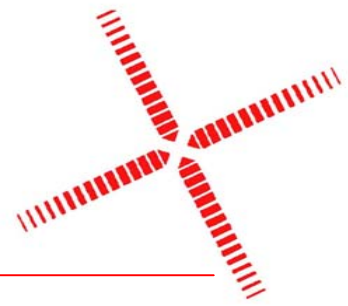
- Migration is needed due to technological obsolescences
- Migration is
The periodic transfer from one HW and/or SW configuration to another, or from one generation of computer technology to a subsequent generation.

Migration Pros and Cons



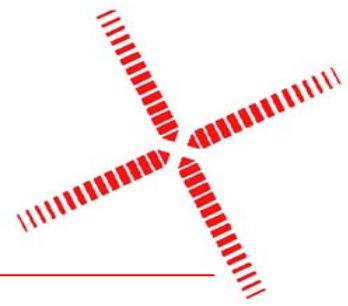
- Migration is good to
 - Preserve integrity of digital data
 - Retain ability to retrieve/access/use data
 - Exploit technological progress
- **But ...**
 - Exact digital copying is not always possible
 - Maintaining compatibility not guaranteed
 - Migration is time-consuming, costly, complex and error-prone

Text-like Documents (1)



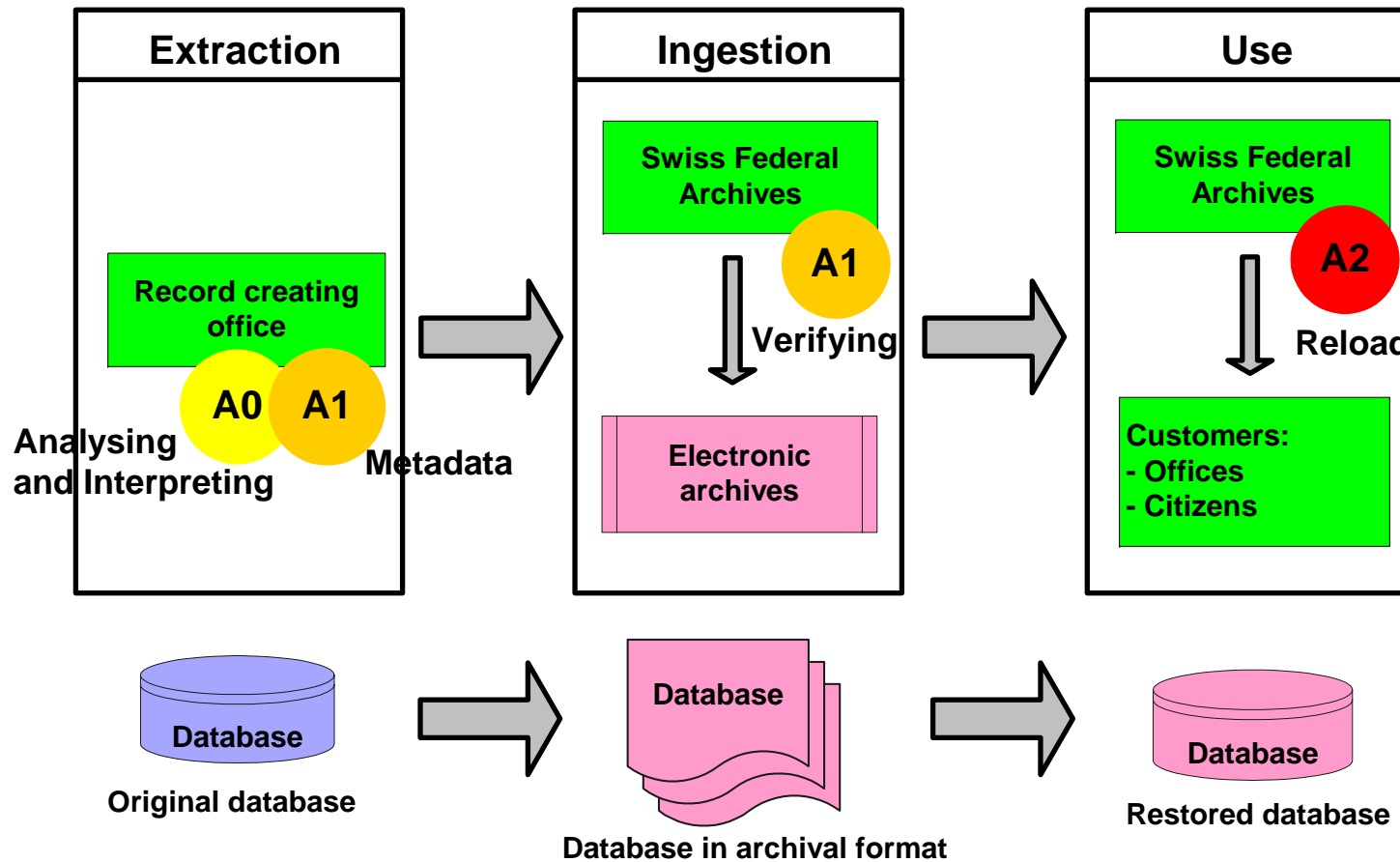
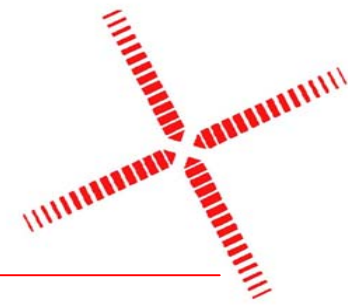
- Office documents (often MS Word or Excel Files) can be preserved as
 - PostScript, PDF, DSSSL, RTF, ASCII, SGML, TIFF, CGM, PNG
 - PostScript, PDF, RTF are proprietary
 - DSSSL, SGML not (yet?) widely used
 - CGM has multiple variants in use

Text-like Documents (2)

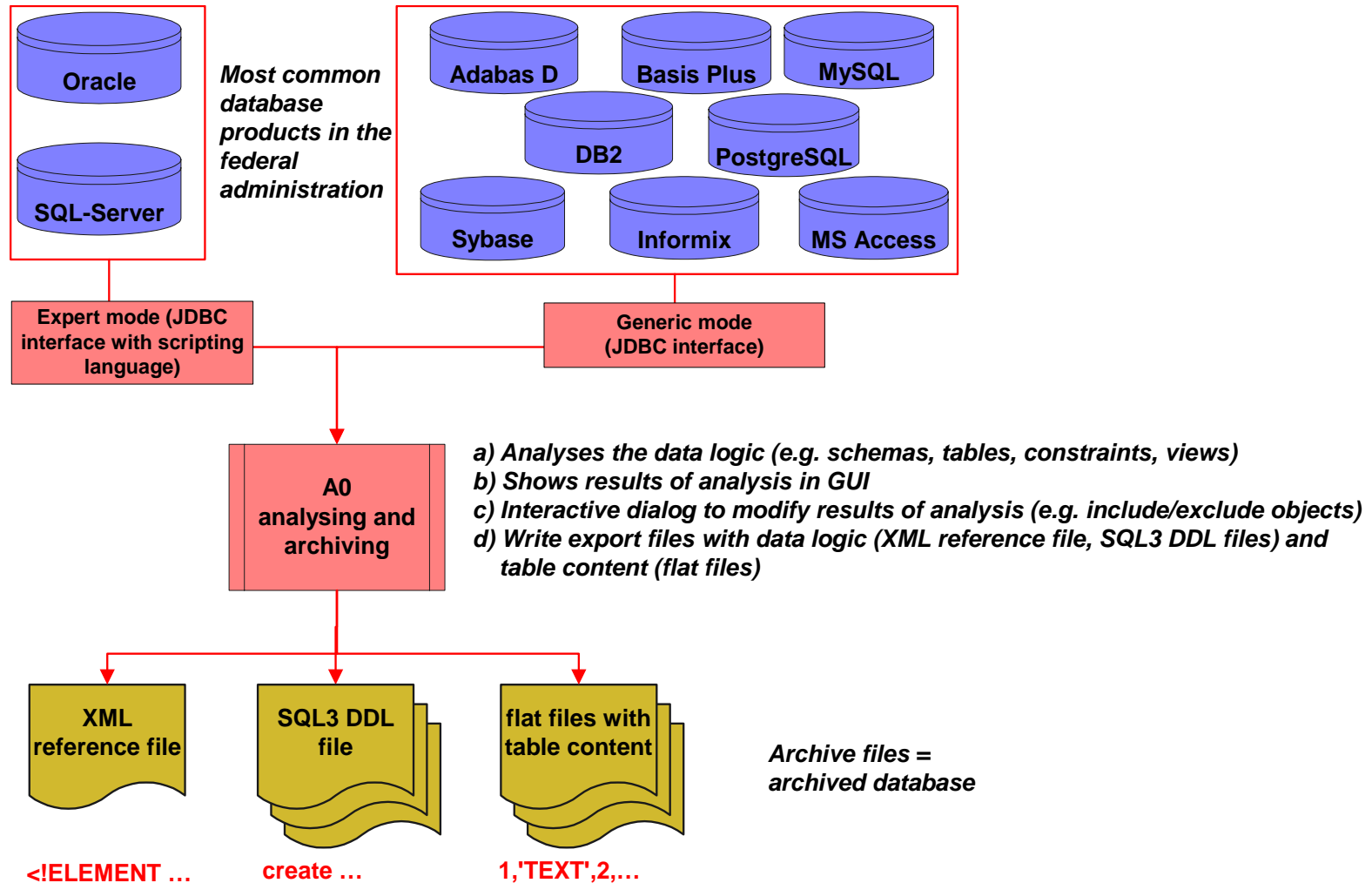
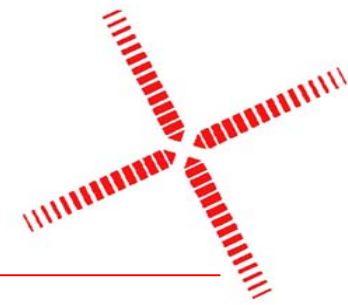


- 'Best practice' in the Swiss Federal Archive:
 - PDF and TIFF
 - No proprietary extensions allowed
 - Private fields and values will be ignored
 - Multitpage TIFF allowed with restrictions
 - Pages belong to the same document or
 - Pages display the same content with different resolutions

Database Preservation



Database Extraction



Raster Image Example

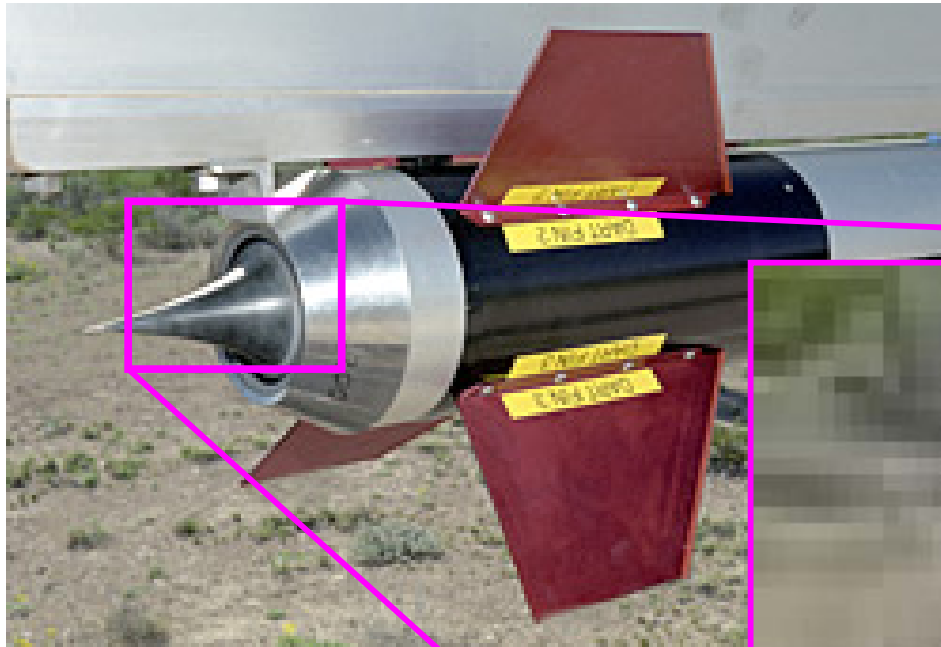
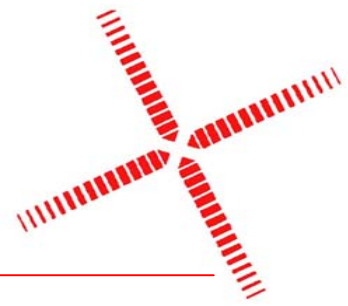
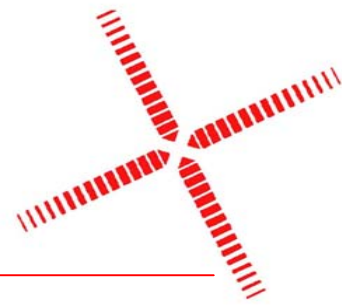
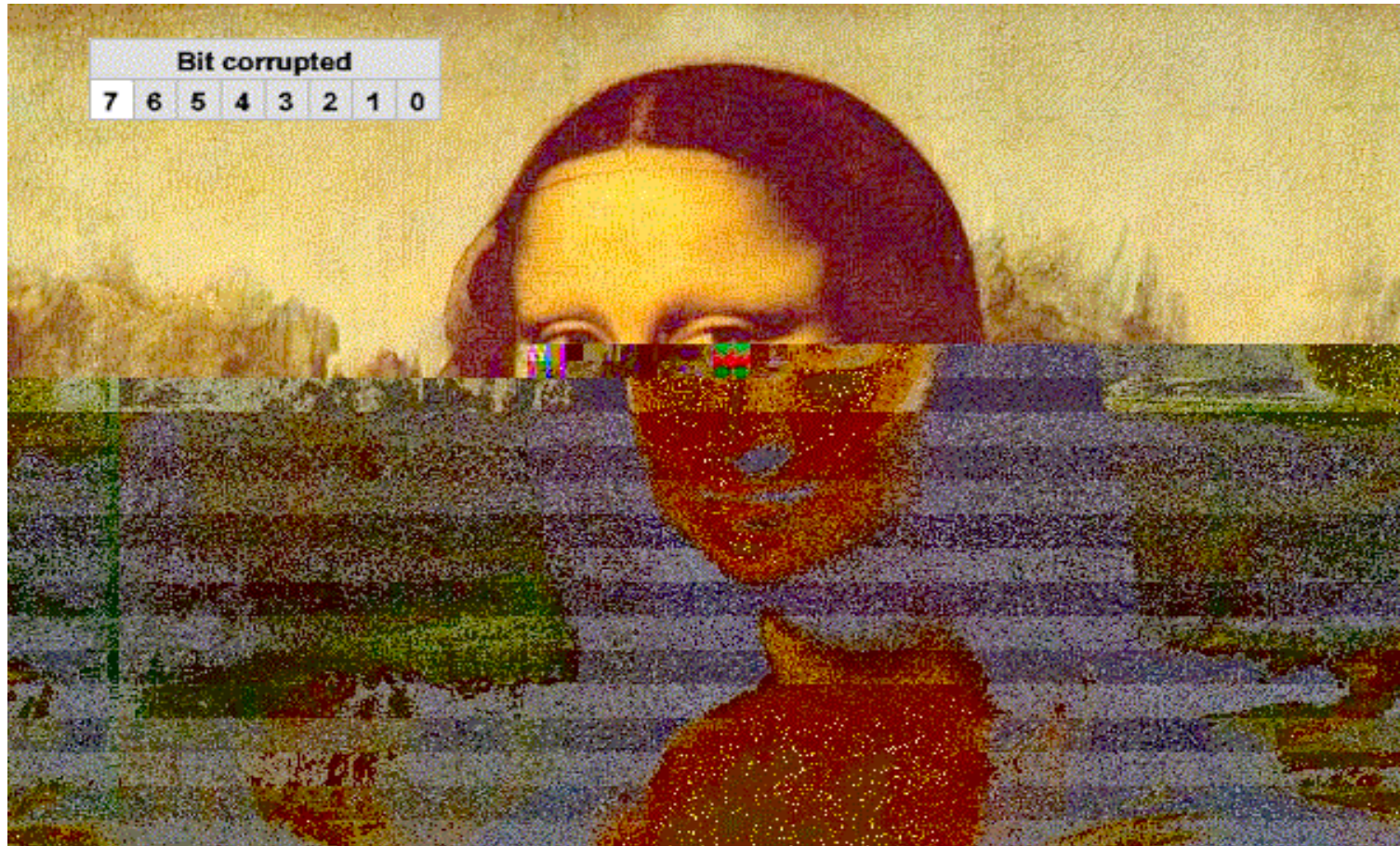
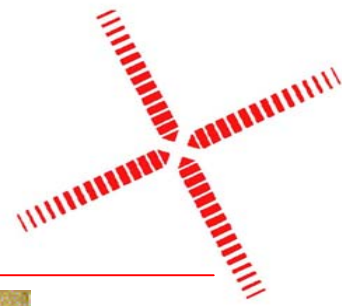


Image Compression



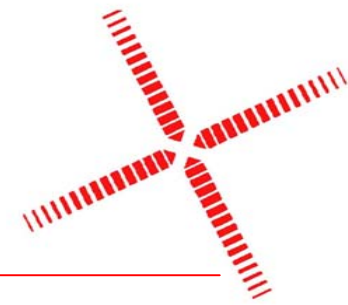
- Lossy compression: Loss of Quality
 - JPEG, JPEG2000
- Lossless compression
 - TIFF, PBM, PNG, etc.

Loss of Redundancy



**Only one bit of a Byte is corrupted in this image:
No error recovery due to compression!**

Lossy Compression



2 kB JPEG

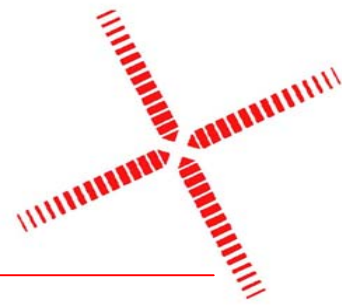


4 kB JPEG



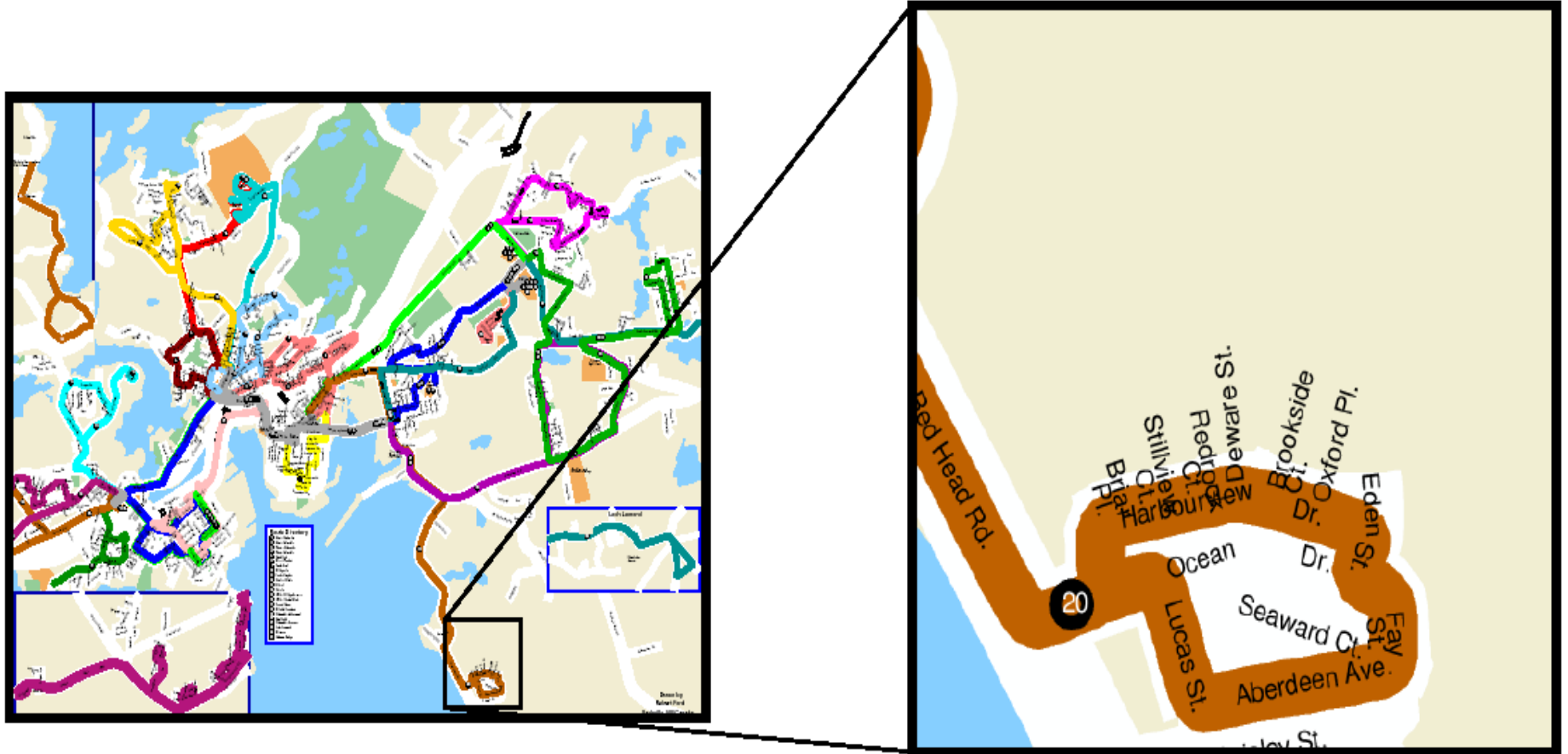
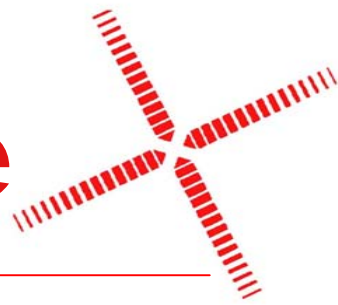
8 kB JPEG

Vector Graphic Files

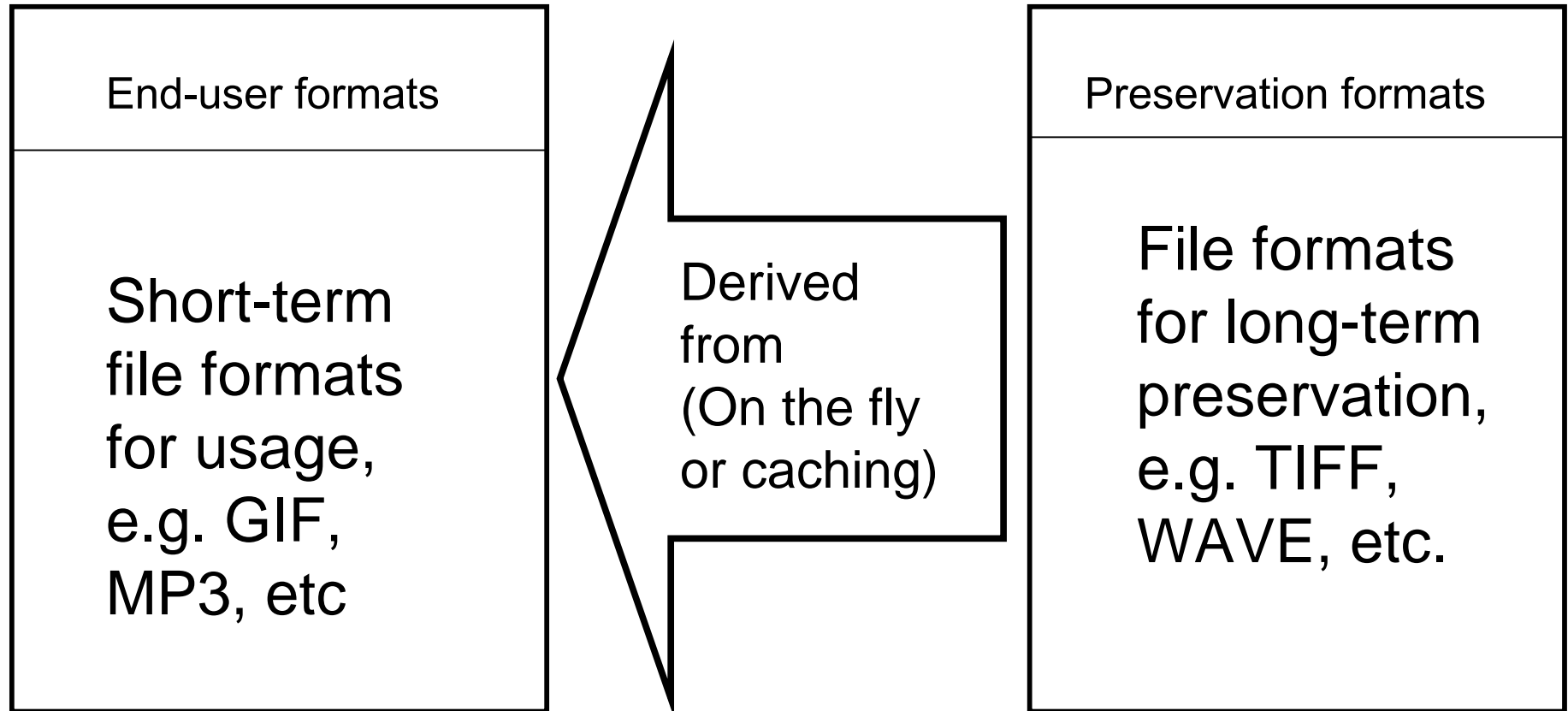
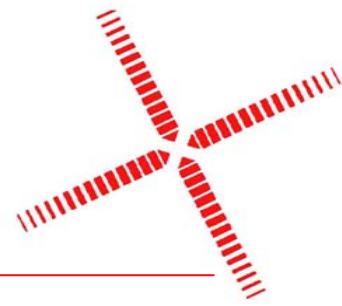


- Vector graphics are composed of geometric figures, splines, polynomials, etc.
- Thus, scaling, rotating, and other operations can be performed efficiently
- Possible formats: PS, PDF, SVG

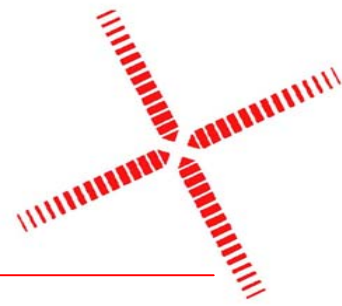
Vector Graphics Example



Preservation Strategy

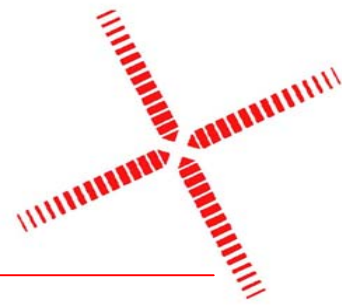


Opportunities



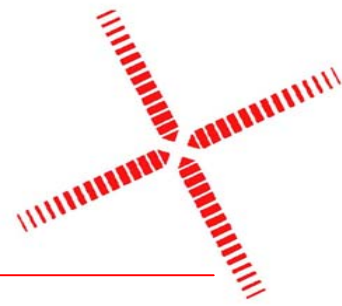
- Up-to-date documents can be read/accessed
- Workable documents can be processed with current software on current hardware
- Quality may be improved and new features may become available

Opportunities (cont'd)



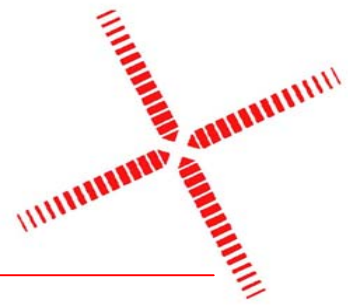
- Up-to-date file formats require less education than historical ones
- Up-to-date and widely used file formats are better supported by software producers than old, obsolete file formats
- Documents are “alive” as they have to be migrated periodically

Threats



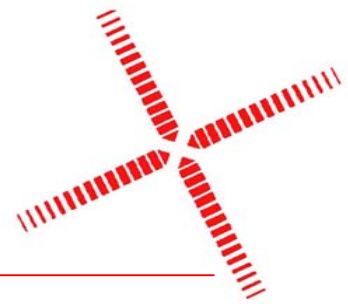
- Migration is not an established, uniform process
- Format conversion has an intrinsic risk of data corruption
- The migration process requires a high effort (e.g. quality assurance)
- Too many file formats require too much resources for tracking their development

Threats (cont'd)



- New features of new file formats may affect derivative creation
- Migration affects watermarks, signatures, or other cryptographic techniques for “fixity”
- Unpredictable migration cycles make staff planning difficult

Q & A



Q & A

Discussion