



Harvard University Library

---

*International Conference on Preservation of Digital Objects*  
Göttingen, Germany, November 15-16, 2005

## Digital Formats and Preservation

Stephen L. Abrams  
*Digital Library Program Manager*  
*Harvard University, USA*



## Digital formats and digital preservation

- Agenda
  - What is digital preservation?
  - What is a format?
  - OAIS format dependencies
  - Format identification, validation, and characterization
  - Format risk analysis
  - Format registries
  - Case study
  - Summary



## Digital formats and digital preservation

- Agenda
  - What is digital preservation?
  - What is a format?
  - OAIS format dependencies
  - Format identification, validation, and characterization
  - Format risk analysis
  - Format registries
  - Case study
  - Summary



## What is digital preservation?

- A set of policies, services, and practices intended to “provide reliable, long-term access to managed digital resources to ... designated communities, now and in the future”

[ RLG/OCLC, *Trusted Digital Repositories*, 2002 ]



## What is digital preservation?

- A set of policies, services, and practices intended to “provide reliable, long-term **access** to managed digital resources to ... **designated communities**, now and in the future”

[ RLG/OCLC, *Trusted Digital Repositories*, 2002 ]



## What is digital preservation?

- A set of policies, services, and practices intended to “provide reliable, long-term **access** to managed digital resources to ... **designated communities**, now and in the future”  
[ RLG/OCLC, *Trusted Digital Repositories*, 2002 ]
- In other words, preservation is for use



## Requirements for use

- Data is retrievable from its media
- Data is unchanged from its original state



## Requirements for use

- Data is retrievable from its media
- Data is unchanged from its original state
  - “Bit-level preservation”
  - Well understood technical solutions
    - Redundant storage
    - Checksums, error correcting codes
    - Periodic media refresh





## Requirements for use

- Data is retrievable from its media
- Data is unchanged from its original state
  - “Bit-level preservation”
  - Well understood technical solutions
    - Redundant storage
    - Checksums, error correcting codes
    - Periodic media refresh
- This results in the preservation of the digital object as an opaque artifact, not as a piece of usable content



## Requirements for use

- Data is retrievable from its media
- Data is unchanged from its original state
- And...



## Requirements for use

- Data is retrievable from its media
- Data is unchanged from its original state
- And...
- Data is interpretable
  - Semantic understanding of syntactic structures



## Requirements for use

- Data is retrievable from its media
- Data is unchanged from its original state
- And...
- Data is interpretable
  - Semantic understanding of syntactic structures
- Data is renderable
  - Conversion of the digital representation to a humanly-sensible form



## Requirements for use

- Data is retrievable from its media
- Data is unchanged from its original state
- And...
- **Data is interpretable**
  - Semantic understanding of syntactic structures
- Data is renderable
  - Conversion of the digital representation to a humanly-sensible form



## Preservation characterization

- Effective preservation requires proper characterization of the digital object
  - A formal description of the object's properties that are significant with respect to potential preservation activities
  - Technical, administrative, and curatorial



## Preservation characterization

- Effective preservation requires proper characterization of the digital object
  - A formal description of the object's properties that are significant with respect to potential preservation activities
  - Technical, administrative, and curatorial
- For ensuring the usability of the object, the fundamental characterization property is format



## Without format typing, all content is opaque

```
Ffd8ffe000104a46494600010201
008300830000ffed0fb050686f74
6f73686f7020332e30003842494d
03e90a5072696e7420496e666f00
00000078000000000004800480000
000002f40240ffeeffee03060252
0347052803fc0002000000480048
0000000002d80228000100000064
000000010003030300000001270f
0001000100000000000000000000
0000600800190190000000000000
0000000000000000000000000000
000000000000000000000000003842
494d03ed0a5265736f6c7574696f
6e0000000010008313a300020001
008313a3000200013842494d04 ...
```





## Without format typing, all content is opaque

Ffd8ffe000104a46494600010201	SOI
008300830000ffed0fb050686f74	APP0 JFIF 1.2
6f73686f7020332e30003842494d	APP13 IPTC
03e90a5072696e7420496e666f00	APP2 ICC
0000007800000000000480048000	DQT
000002f40240ffeeffee03060252	SOF0 183x512
0347052803fc0002000000480048	DRI
0000000002d80228000100000064	DHT
000000010003030300000001270f	SOS
0001000100000000000000000000	ECS0
0000600800190190000000000000	RST0
0000000000000000000000000000	ECS1
000000000000000000000000003842	RST1
494d03ed0a5265736f6c7574696f	ECS2
6e0000000010008313a300020001	RST2
008313a3000200013842494d04 ...	...



# Without format typing, all content is opaque

```
Ffd8ffe000104a46494600010201
008300830000ffed0fb050686f74
6f73686f7020332e30003842494d
03e90a5072696e7420496e666f00
00000078000000000004800480000
000002f40240ffeeffee03060252
0347052803fc0002000000480048
0000000002d80228000100000064
000000010003030300000001270f
0001000100000000000000000000
0000600800190190000000000000
0000000000000000000000000000
00000000000000000000000003842
494d03ed0a5265736f6c7574696f
6e0000000010008313a300020001
008313a3000200013842494d04 ...
```

```
SOI
APP0 JFIF 1.2
APP13 IPTC
APP2 ICC
DQT
SOF0 183x512
DRI
DHT
SOS
ECS0
RST0
ECS1
RST1
ECS2
RST2
...
```





## Digital formats and digital preservation

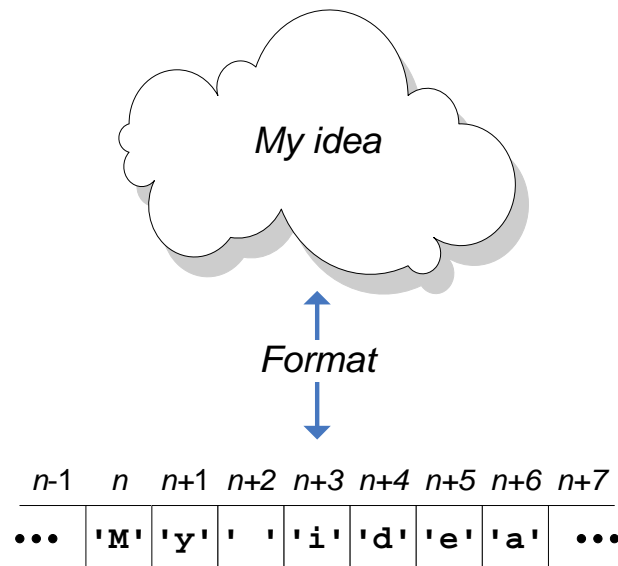
- Agenda

- ✓ *What is digital preservation?*
- What is a format?
- OAIS format dependencies
- Identification, validation, and characterization
- Format risk analysis
- Format registries
- Case study
- Summary



## What is a format?

- A byte-serialized encoding of an abstract information model
  - A set of syntactic and semantic rules that
    - Map abstract content to a sequence of bytes
    - Map a sequence of bytes back to that abstract content





## Almost anything is a format

- ASCII, JPEG, PDF, RealAudio, Word, XML



## Almost anything is a format

- ASCII, JPEG, PDF, RealAudio, Word, XML
- IEEE 754 Standard for Binary Floating-Point Arithmetic



## Almost anything is a format

- ASCII, JPEG, PDF, RealAudio, Word, XML
- IEEE 754 Standard for Binary Floating-Point Arithmetic
- LZW compression



## Almost anything is a format

- ASCII, JPEG, PDF, RealAudio, Word, XML
- IEEE 754 Standard for Binary Floating-Point Arithmetic
- LZW compression
- XML schema (and XML Schema)





## Almost anything is a format

- ASCII, JPEG, PDF, RealAudio, Word, XML
- IEEE 754 Standard for Binary Floating-Point Arithmetic
- LZW compression
- XML schema (and XML Schema)
- Windows Portable Executable (\*.exe)



## Almost anything is a format

- ASCII, JPEG, PDF, RealAudio, Word, XML
- IEEE 754 Standard for Binary Floating-Point Arithmetic
- LZW compression
- XML schema (and XML Schema)
- Windows Portable Executable (\*.exe)
- Tar archive



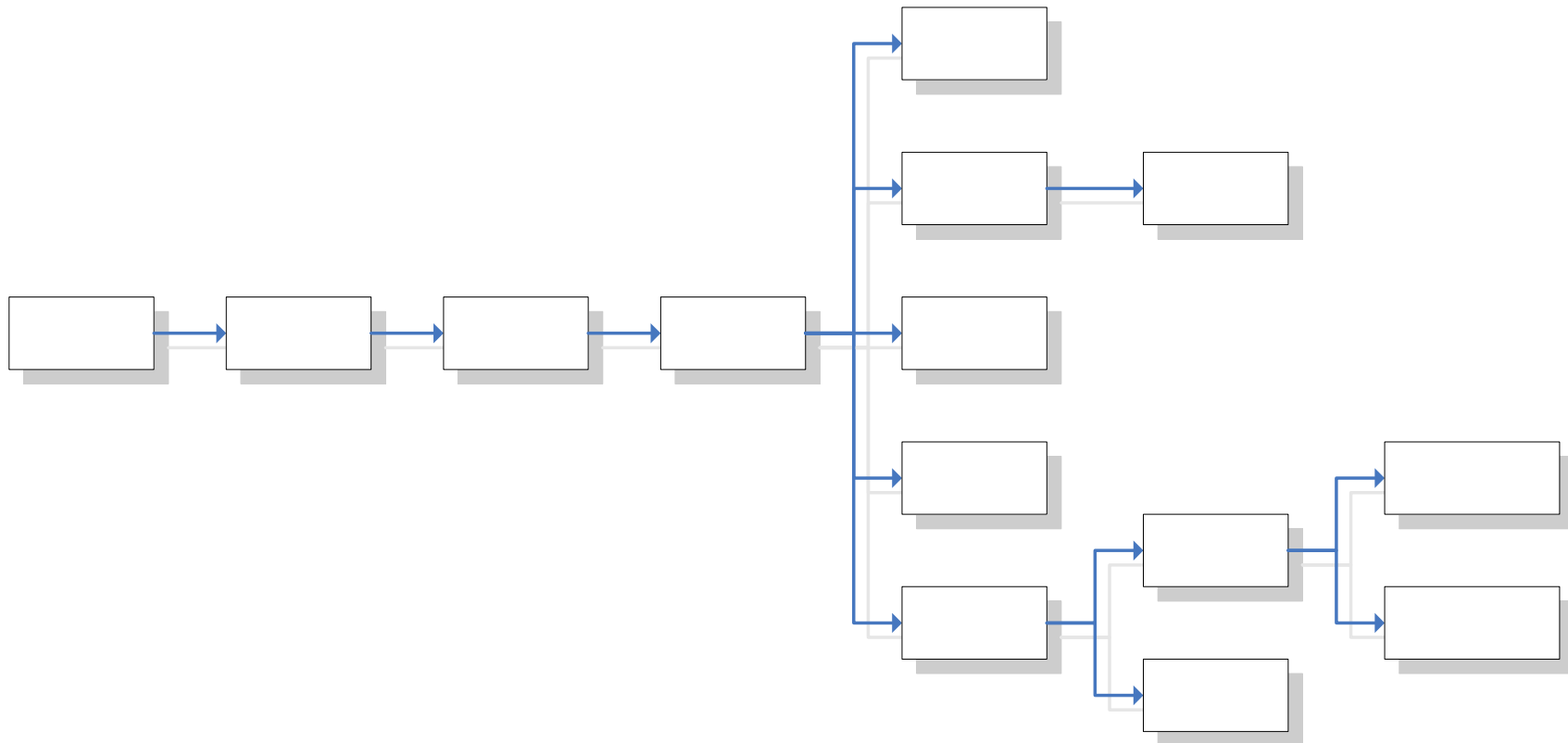
## Almost anything is a format

- ASCII, JPEG, PDF, RealAudio, Word, XML
- IEEE 754 Standard for Binary Floating-Point Arithmetic
- LZW compression
- XML schema (and XML Schema)
- Windows Portable Executable (\*.exe)
- Tar archive
- UFS file system



## Format families

- Formats often exist in well-defined relationships to each other





## Format classification

- Ontological CLASSES, abstract *families*, concrete formats, and **relationships**

BYTESTREAM

IMAGE

STILL

RASTER

**GIF**

GIF87a

GIF89a

**new-version-of**

GIF87a

**JPEG**

ISO 10918-1

JFIF

**subtype-of**

ISO 10918-1

**TIFF**

TIFF 4.0

TIFF 5.0

**new-version-of**

TIFF 4.0

TIFF 6.0

**new-version-of**

TIFF 5.0

TIFF/EP

**subtype-of**

TIFF 6.0

TIFF/IT

**subtype-of**

TIFF 6.0

TIFF/IT/CT

**subtype-of**

TIFF/IT

TIFF/IT/CT/P1

**subtype-of**

TIFF/IT/CT



## Format relationships

- Subtype
  - US-ASCII is a subtype of UTF-8
- Version
  - PDF 1.0 – 1.6
- Encapsulation
  - WAVE can contain A-law and  $\mu$ -law audio content streams
  - Tar archive can contain anything
- Affinity
  - ISO 10918-1 (JPEG) vs. ISO 10918-3 (SPIFF) vs. ISO 14495 (JPEG-LS)



## Format subtyping

- Substitutability
  - Can the subtype be substituted for its parent in all contexts without detection or loss of function?
    - All METS files are XML, but not all XML files are METS
- Arbitrary granularity of subtype
  - TIFF 6.0 →
    - Baseline bitonal (Class B) →
      - DLF Benchmark for Faithful Digital Reproductions of Monographs →
        - Harvard Open Collection Program specifications →
          - ...



## Formatted object granularity

- In general, 1 digital object = 1 file = 1 format
- But not always...





## Formatted object granularity

- In general, 1 digital object = 1 file = 1 format
- But not always...
  - TIFF with embedded IPTC and XMP metadata
    - 1 digital object = 1 file = 3 formats



## Formatted object granularity

- In general, 1 digital object = 1 file = 1 format
- But not always...
  - TIFF with embedded IPTC and XMP metadata
    - 1 digital object = 1 file = 3 formats
  - JPEG 2000 JPX profile with file fragmentation
    - 1 digital object =  $n$  files = 1 format



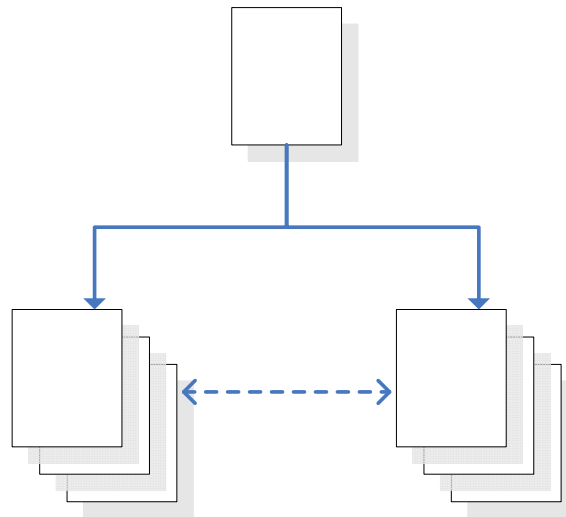
## Formatted object granularity

- In general, 1 digital object = 1 file = 1 format
- But not always...
  - TIFF with embedded IPTC and XMP metadata
    - 1 digital object = 1 file = 3 formats
  - JPEG 2000 JPX profile with file fragmentation
    - 1 digital object =  $n$  files = 1 format
  - ESRI Shapefile
    - 1 digital object = 3 files = 3 formats
      - Main file (\*.shp)
      - Index file (\*.shx)
      - dBASE table (\*.dbf)



## Formatted object granularity

- Formats apply not only to bitstreams...
  - And possibly subsets of bitstreams
- But also to objects made up of multiple bitstreams existing in specific relationships to each other
  - Fedora content model
  - METS profile





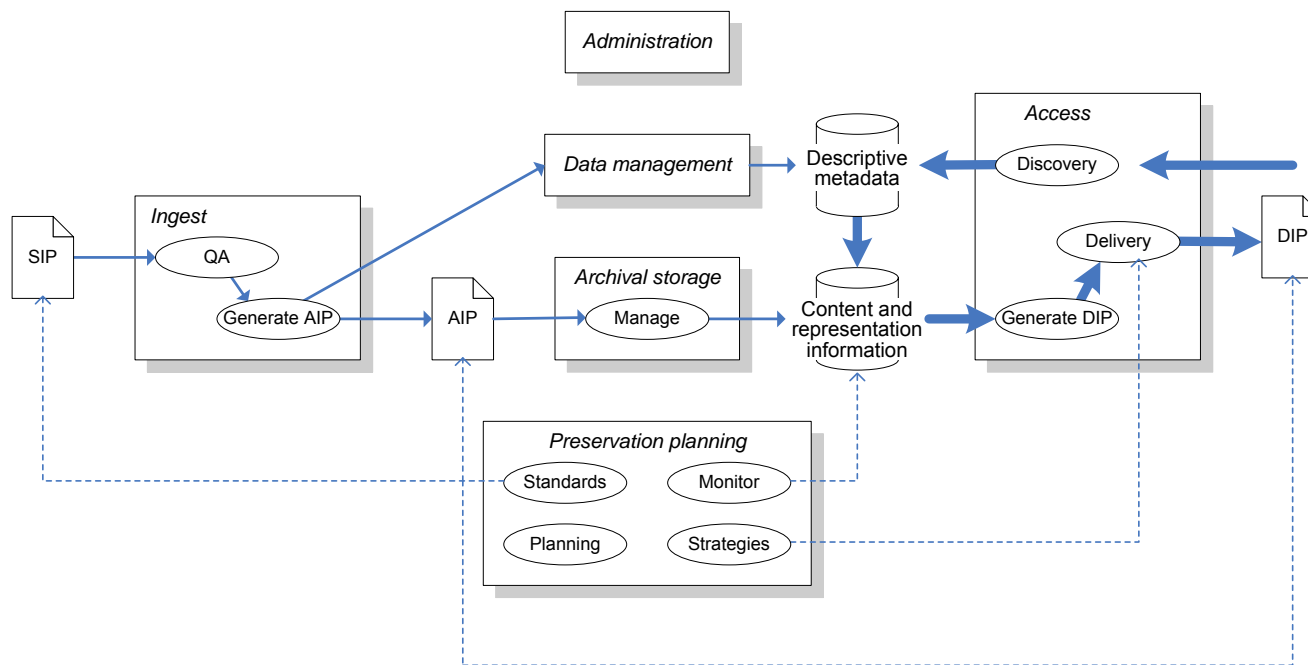
## Digital formats and digital preservation

- Agenda

- ✓ *What is digital preservation?*
- ✓ *What is a format?*
- OAIS format dependencies
  - Identification, validation, and characterization
  - Format risk analysis
  - Format registries
  - Case study
  - Summary



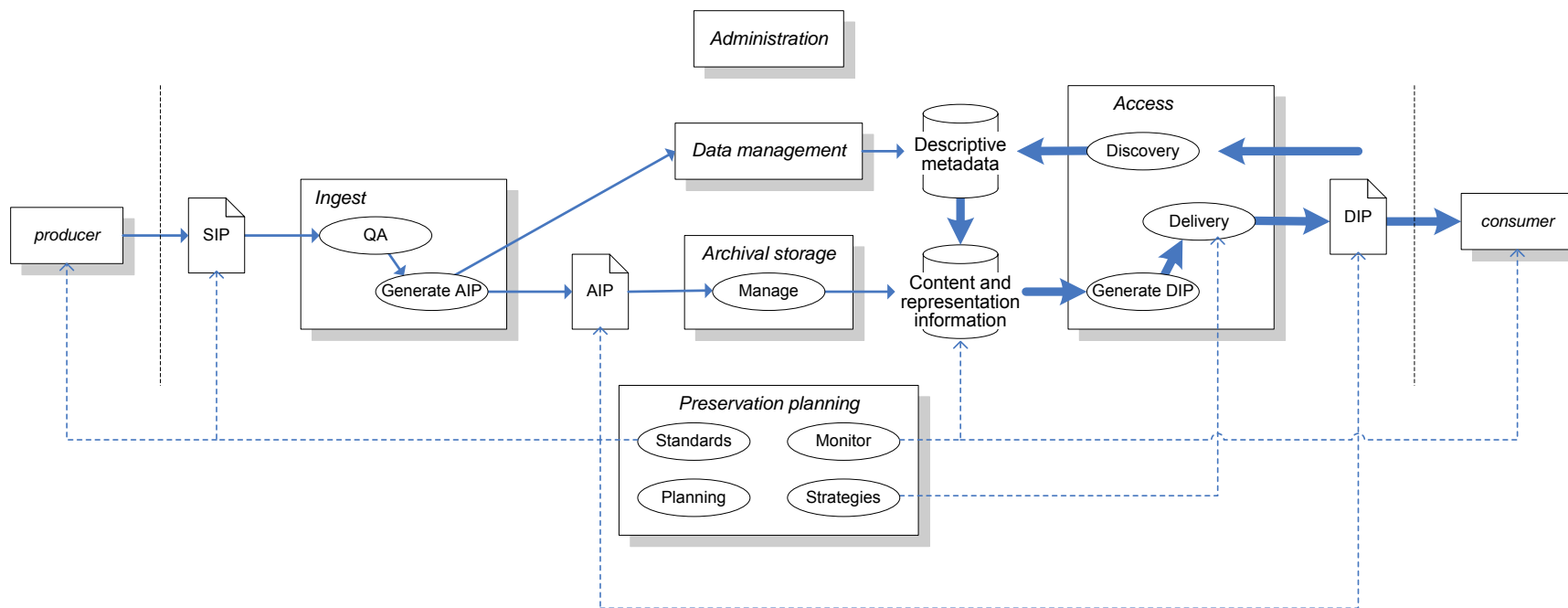
## OAIS format dependencies



[ Adapted from ISO 14721:2003, *Space data and information transfer systems—Open archival information system—Reference model* ]



## OAIS format dependencies



[ Adapted from ISO 14721:2003, *Space data and information transfer systems—Open archival information system—Reference model* ]



## Format use cases

- Selection
  - “I have abstract content; in what format can I (best) represent it by a digital object?”
- Identification
  - “I have a digital object; what format is it?”
- Validation
  - “I have an object purportedly of format  $F$ ; is it?”
- Characterization
  - I have an object of format  $F$ ; what are its salient properties?”
- Assessment
  - “I have an object of format  $F$ ; is it at risk of obsolescence?”
- Processing
  - “I have an object of format  $F$ ; how can I perform operation  $X$  on it?”





## Object validation

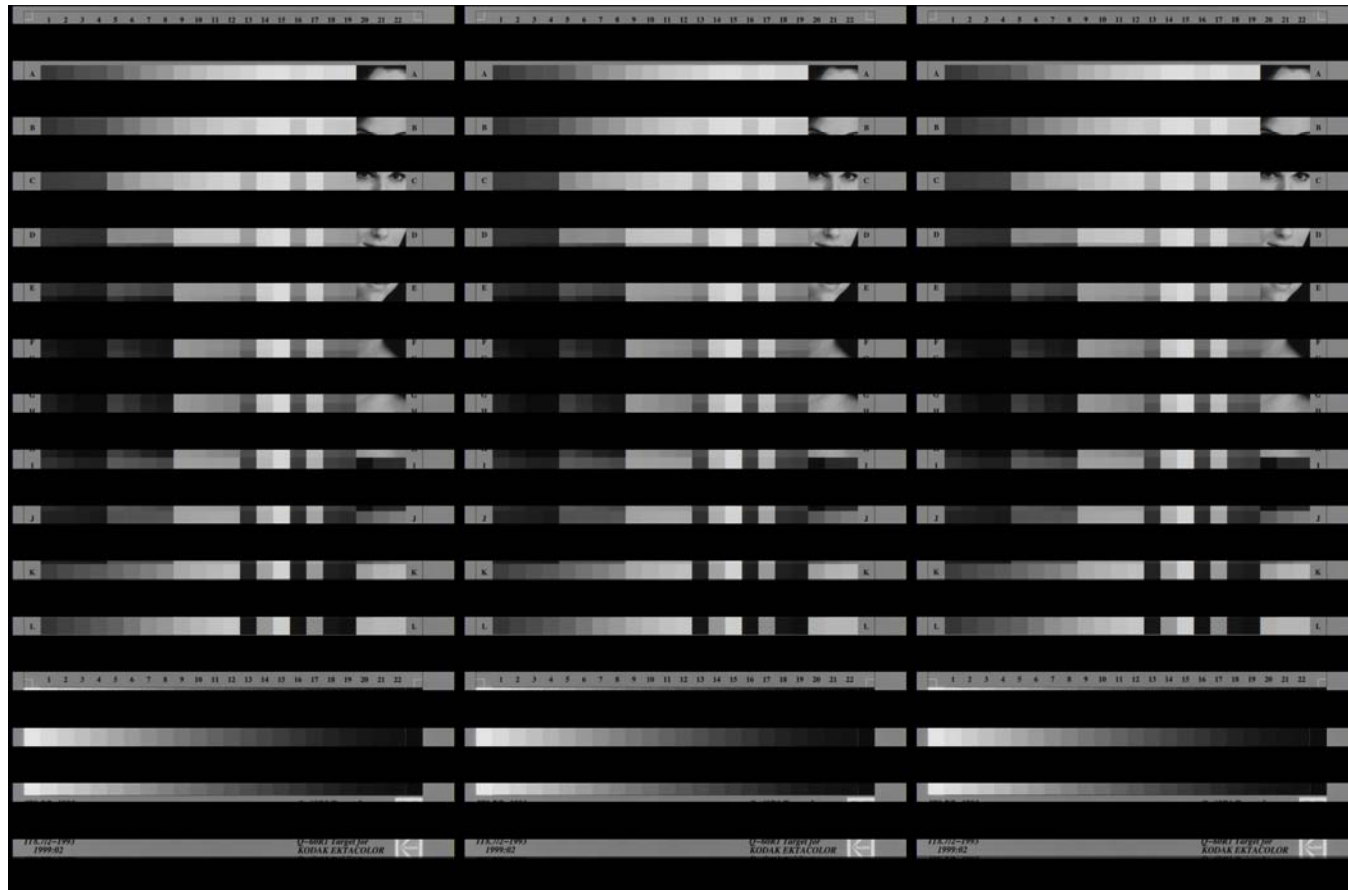
- Retrospective validation check of 1.1 million objects in the Harvard Digital Repository Service (DRS) in 2004
  - 4% of XML, 1.4% of TIFF, 0.1% of JPEG found to be invalid
  - These objects were created by technically sophisticated reformatting vendors, according to well-known specifications
  - Some violations were trivial, others were significant

[ Stephen Abrams and Gary McGath, “Format Dependencies in Repository Operation,” *DLF Fall Forum*, October 25-27, 2004 ]

- Recent deposits are still showing a high error rate for non-critical, though technical violations of the TIFF standard



# Constraint violation





## Pointer out-of-bounds

THE COLLEGE

75

As compared with the table of last year, this table shows an increase of twenty-six men over the number admitted last year from other colleges.

The Freshman Class of 1914-15 showed an unexpected increase; it contained probably the largest number of genuine Freshmen — that is, men just admitted — in the history of the College. The Freshman Class of 1915-16 shows a total decrease of twenty-one. The decrease in number admitted by examination in 1915, however, is but twelve.

Sixty-three students won a place in the First Group of Scholars. Of these twenty-three hold honorary scholarships; forty, stipendiary. Only once since the establishment of the three groups of scholars has this number been exceeded; this was in 1903-04, when seventy men won a position in this Group. Last year forty-seven students won a place in the First Group. Of these twenty-two received honorary scholarships; twenty-five, scholarships with stipend. One of the most difficult tasks of the Committee this year was to award the Jacob Wendell Scholarship; this is given, irrespective of financial need, to that member of the Freshman Class whose record is on the whole the most distinguished. In the Class of 1918 there were seven men who had grade A in at least five full courses, or their equivalent.

One hundred and sixty-five students won a position in the Second Group. Of these sixty-five received honorary scholarships; one hundred, stipendiary. Last year one hundred and sixty-seven students were in this Group, sixty-eight holding honorary scholarships, and ninety-nine scholarships with stipend.

In the First Group are thirty Seniors, nineteen Juniors, fourteen Sophomores; in the Second Group, fifty-one Seniors, fifty-two Juniors, sixty-one Sophomores, and one candidate for a degree out of course, giving as totals for the three classes eighty-one Seniors, seventy-one Juniors, and seventy-five Sophomores.

The members of the Administrative Board for the year 1914-15 were Professors Willson, Ward, Parker, Chase, Greenough, the Assistant Dean, and the Dean. During the year, the Board kept a diary. Only once since the establishment of the three groups of scholars has this number been exceeded; this was in 1903-04, when seventy men won a position in this Group. Last year forty-seven students won a place in the First Group. Of these twenty-two received honorary scholarships; twenty-five, scholarships with stipend. One of the most difficult tasks of the Committee this year was to award the Jacob Wendell Scholarship; this is given, irrespective of financial need, to that member of the Freshman Class whose record is on the whole the most distinguished. In the Class of 1918 there were seven men who had grade A in at least five full courses, or their equivalent.

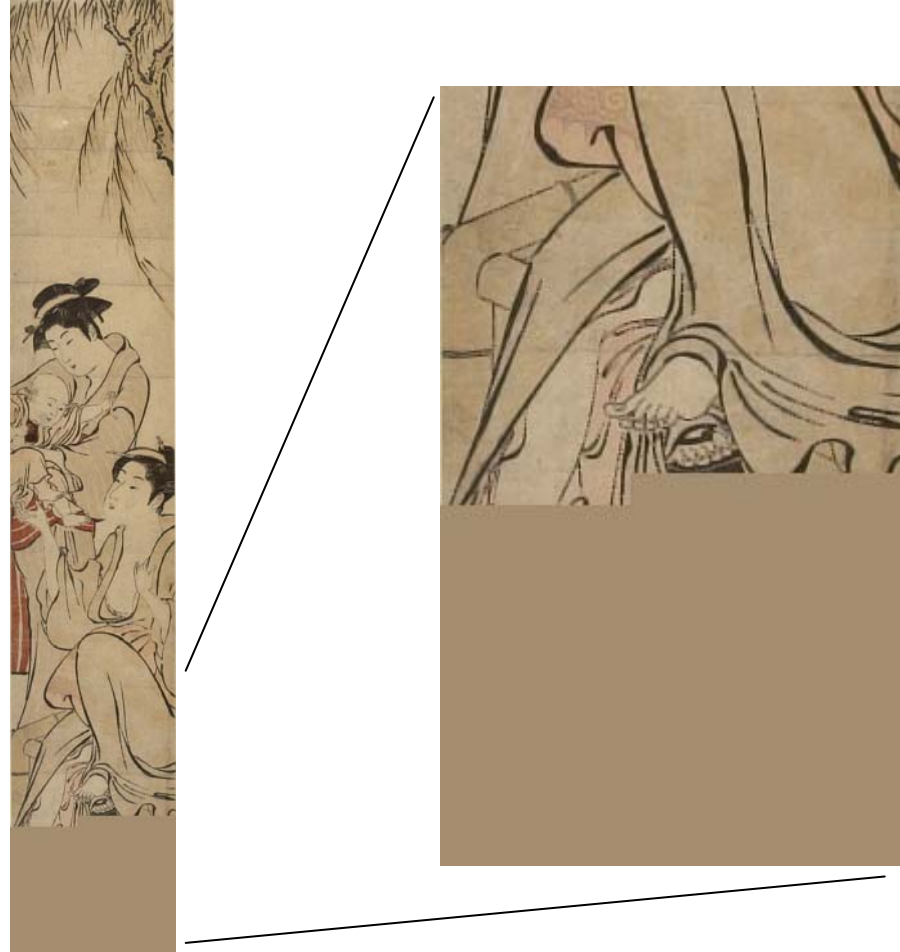
One hundred and sixty-five students won a position in the Second Group. Of these sixty-five received honorary scholarships; one hundred, stipendiary. Last year one hundred and sixty-seven students were in this Group, sixty-eight holding honorary scholar-

The members of the Administrative Board for the year 1914-15 were Professors Willson, Ward, Parker, Chase, Greenough, the Assistant Dean, and the Dean. During the year, the Board kept a diary. Only once since the establishment of the three groups of scholars has this number been exceeded; this was in 1903-04, when seventy men won a position in this Group. Last year forty-seven students won a place in the First Group. Of these twenty-two received honorary scholarships; twenty-five, scholarships with stipend. One of the most difficult tasks of the Committee this year was to award the Jacob Wendell Scholarship; this is given, irrespective of financial need, to that member of the Freshman Class whose record is on the whole the most distinguished. In the Class of 1918 there were seven men who had grade A in at least five full courses, or their equivalent.

One hundred and sixty-five students won a position in the Second Group. Of these sixty-five received honorary scholarships; one hundred, stipendiary. Last year one hundred and sixty-seven students were in this Group, sixty-eight holding honorary scholar-



## Unexpected End-of-File





## Digital formats and digital preservation

- Agenda

- ✓ *What is digital preservation?*
- ✓ *What is a format?*
- ✓ *OAIS format dependencies*
- **Format identification, validation, and characterization**
- Format risk analysis
- Format registries
- Case study
- Summary



## Format identification

- “What is the format?”
  - Of particular importance in institutional and web archives
    - Obligation to accept objects of unknown provenance
  - Unix file(1) command
    - Based on magic numbers
  - National Archives (UK) DROID (Digital Record Object Identification)
    - Internal and external signatures drawn from PRONOM
      - [ Adrian Brown, “Automating Preservation: New Developments in the PRONOM Service,” *RLG DigiNews* 9.2 (April 15, 2005) ]
  - Signature matching may need to be “fuzzy”
    - Acrobat accepts PDF header in the first 1024 bytes, and trailer in the last 1024 bytes



## Format validation

- “Is the purported format actually correct?”
  - We often make an (inappropriate) assumption that the systems and tools that produce digital objects, produce them correctly
  - HTTP Content-type headers, filename extensions, and magic numbers are insufficient indicators of validity
  - Acrobat PDF pre-flight
  - W3C HTML validator



## Format characterization

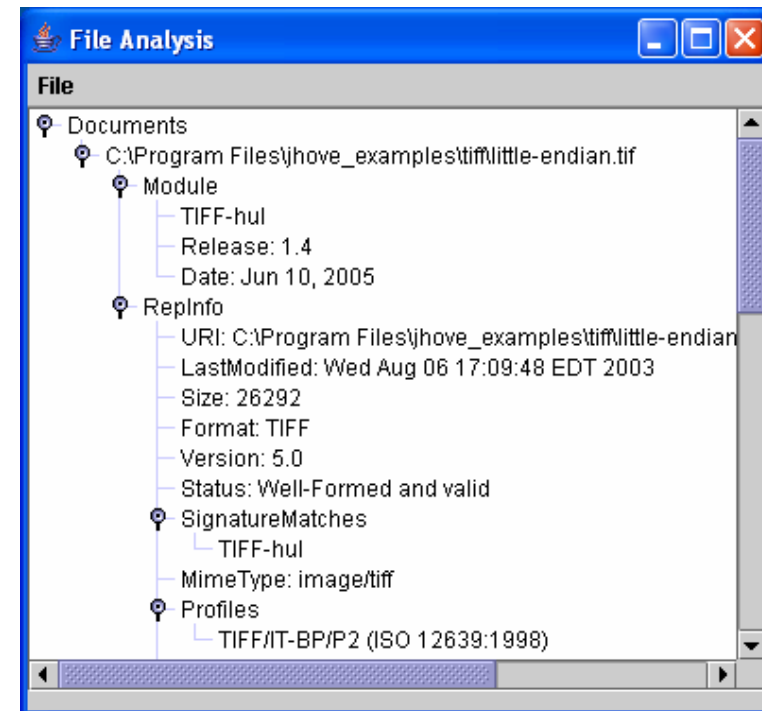
- “What are the salient technical properties of the object?”
  - Proper characterization is necessary for preservation
  - Special purposes libraries (libtiff) and tools (ImageMagick) for specific formats or format classes
  - National Library of New Zealand (NLNZ) Preservation Metadata Extraction Tool
    - Adaptors for BMP, GIF, HTML, JPEG, MS Office, OpenOffice, PDF, TIFF, WAVE, WordPerfect
    - Short-listed for 2004 Pilgrim Trust Digital Preservation Award





## JHOVE

- JSTOR/Harvard Object Validation Environment
- Format-specific object identification, validation, and characterization





## JHOVE

- Extensible, plug-in architecture with modules for
  - AIFF, AIFF-C
  - ASCII
  - GIF 87a, 89a
  - HTML 3.2, 4.0, 4.01, XHTML 1.0, 1.1
  - JPEG, JFIF, SPIFF, JTIP, JPEG-LS, Exif 2.0, 2.1, 2.2
  - JPEG 2000 JP2, JPX
  - PDF 1.0 – 1.6, PDF/X-1, -1a, -2, -3, PDF/A, Tagged PDF, Linearized PDF
  - TIFF 4.0 – 6.0, Class B, G, P, R, Y, F, RFC 1314, TIFF/EP, TIFF/IT (CT, LW, HC, MP, BP, BL, FP, and P1, P2), GeoTIFF, TIFF-FX, Exif 2.0, 2.1, 2.2, DNG
  - UTF-8
  - WAVE, BWF
  - XML



## Potential enhancements for JHOVE V2

- Generic platform for format-related preservation activities
- Iterate over set of objects
  - FileSystemIterator performs traversal over directories and files
  - WebSiteIterator performs crawl over web sites
- For each object execute a sequence of processes
  - Processes for copy, identification, validation, characterization, display, risk assessment, transformation, ...
- Support for multi-file objects and multi-format files
- Standardized support for format profiles



## Digital formats and digital preservation

- Agenda

- ✓ *What is digital preservation?*
- ✓ *What is a format?*
- ✓ *OAIS format dependencies*
- ✓ *Format identification, validation, and characterization*
- **Format risk analysis**
- Format registries
- Case study
- Summary



## Library of Congress assessment model

- Sustainability factors
  - Disclosure
  - Adoption
  - Transparency
  - Self-documentation
  - External dependencies
  - Impact of patents
  - Technical protection mechanisms
- Functionality and quality factors
  - Specific to content genre

[ Caroline Arms and Carl Fleischhauer, “Digital formats: factors for sustainability, functionality, and quality,” *IS&T Archiving Conference*, Washington, April 26-29, 2005 ]



## Online Computer Library Center (OCLC) INFORM

- Classes of risk
  - Format
  - Software
  - Hardware
  - Associated organizations
  - Digital archive
  - Preservation plans

[ Andreas Stanescu, "Assessing the durability of formats in a digital preservation environment: The INFORM methodology," *OCLC Systems and Services* 21.1 (2005): 61-81 ]

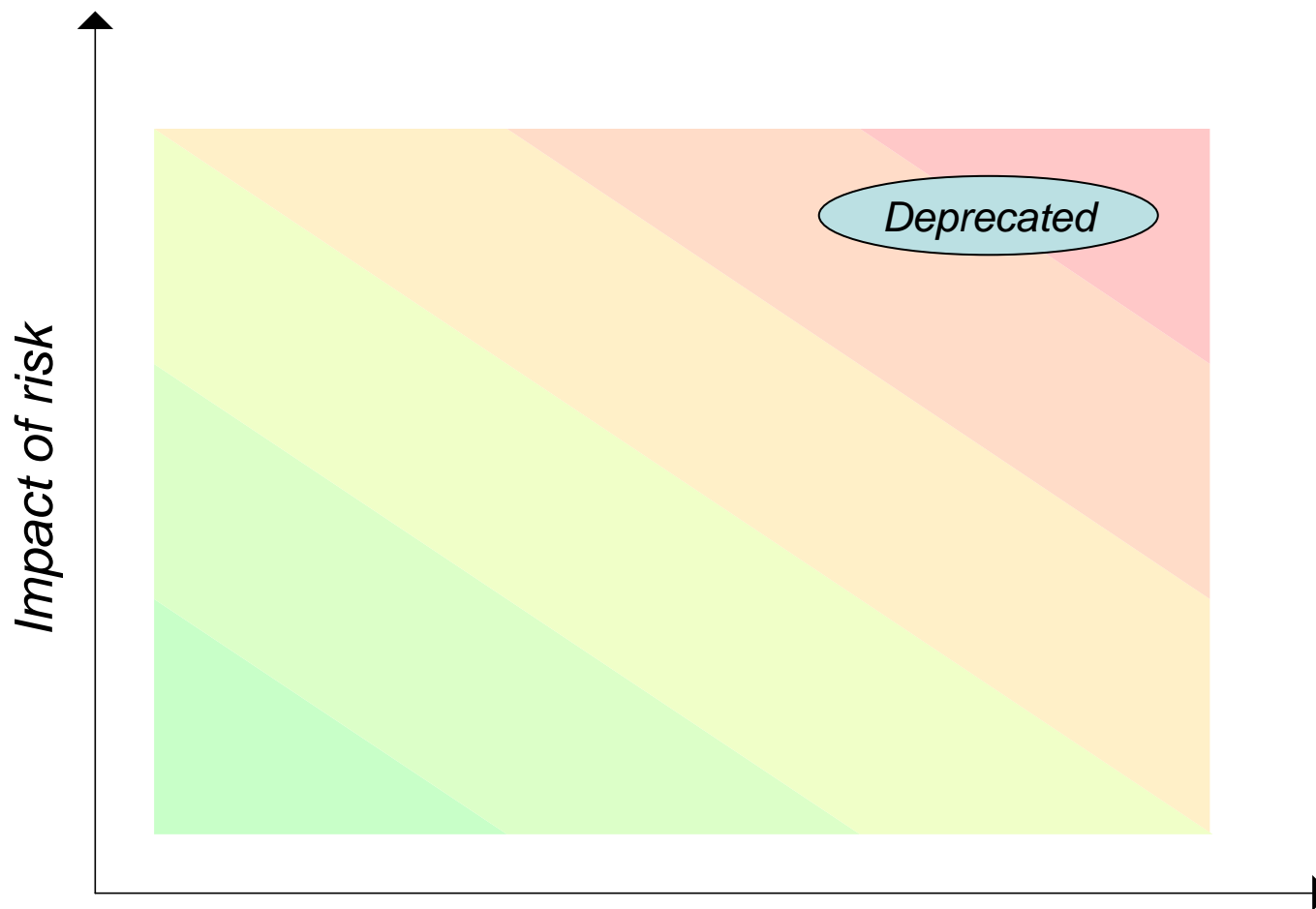


## Format risk assessment





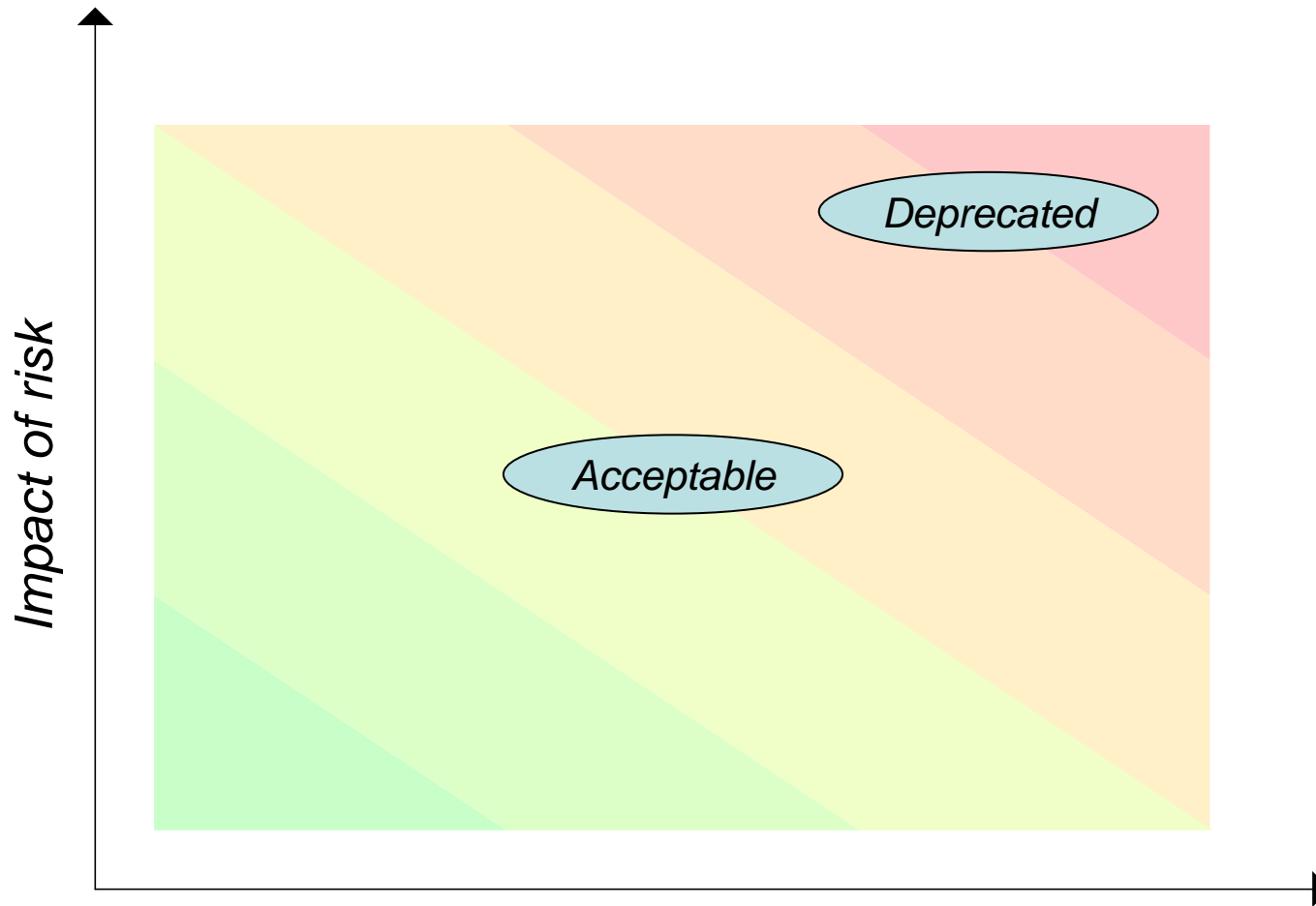
## Format risk assessment





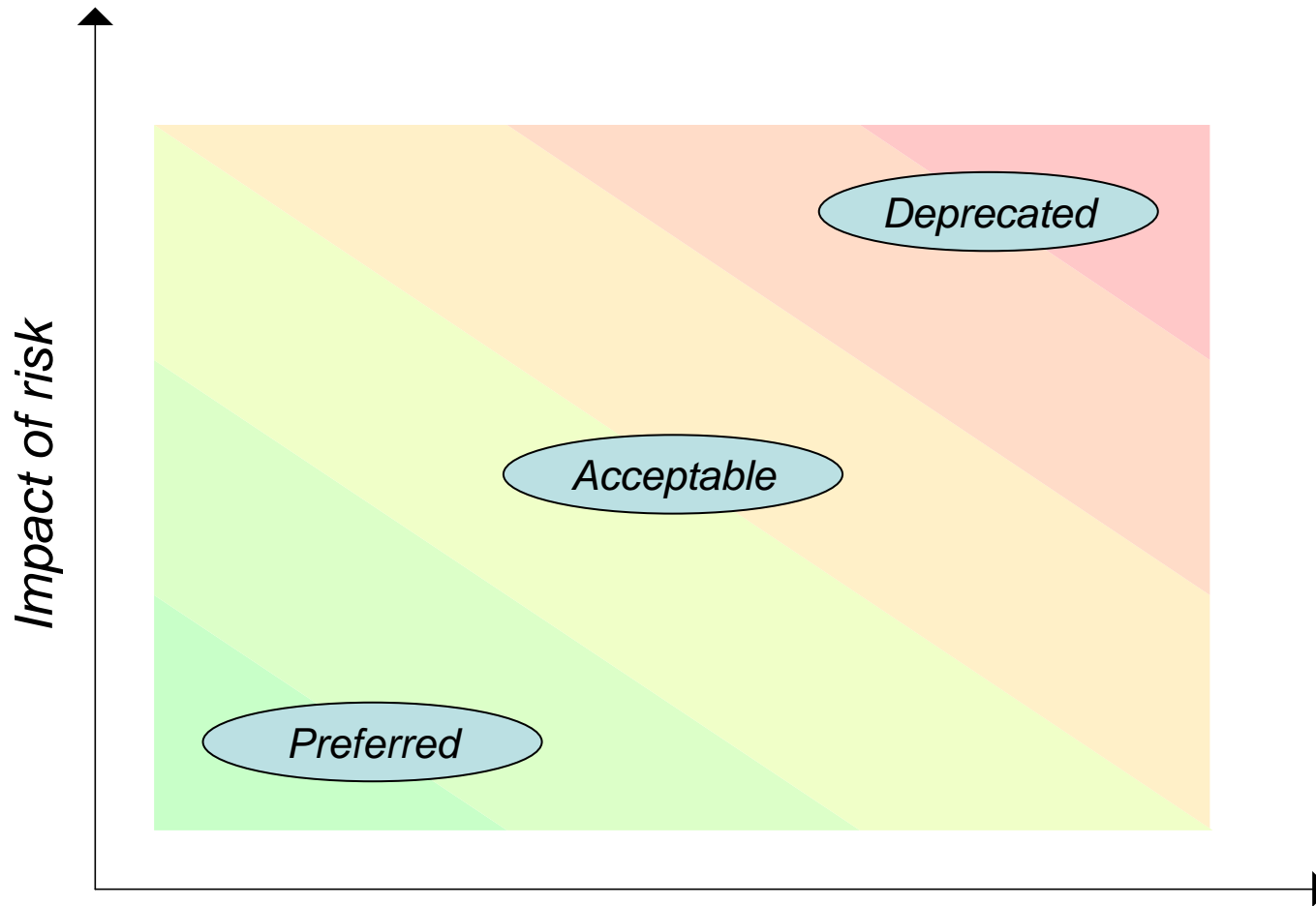


## Format risk assessment





# Format risk assessment





## A new preservation format

- PDF/A (ISO 19005-1:2005)
- Emphasis on preserving unambiguous and reliable reproduction of the original static visual appearance
  - No encryption
  - No LZW compression
  - No external or optional content
  - No executable code, e.g. PostScript, JavaScript
  - No audio or video content
  - Device-independent color spaces
  - Embedded fonts
  - XMP metadata
  - Consistency with other standards, e.g., PDF/X, PDF/E, PDF/UA
  - And for Level A conformance:
    - Unicode character mapping
    - Structural tagging



## Digital formats and digital preservation

- Agenda

- ✓ *What is digital preservation?*
- ✓ *What is a format?*
- ✓ *OAIS format dependencies*
- ✓ *Format identification, validation, and characterization*
- ✓ *Format risk analysis*
- **Format registries**
- Case study
- Summary



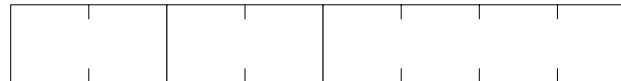
## Format representation information

- Information that maps formatted content to more meaningful concepts

[ ISO 14721:2003, *Space data and information transfer systems—Open archival information system—Reference model* ]

### – Syntax

- A TIFF header is composed of a two byte string, “II” or “MM”; a two byte string, 0x2A00 or 0x002A; and an unsigned 32 bit integer



### – Semantics

- “II” indicates big-endian byte order; “MM”, little-endian
- The two byte string is the decimal value 42 in correct byte order
- The integer is the byte offset of the first IFD structure



## Format representation information

- There are many sources for format representation information, but ...



## Format representation information

- There are many sources for format representation information, but ...
  - For the most part they are informal, incomplete, or ephemeral



## Format registries

- IANA MIME registry
  - 641 formats
- Wotsit.org
  - 918 formats
- Ace.net.nz
  - “Almost every file format in the world”
  - 598 formats
- WhatIs.com
  - “Every file format in the world”
  - 3,189 formats





## Format registries

- IANA MIME registry
  - 641 formats
- Wotsit.org
  - 918 formats
- Ace.net.nz
  - “Almost every file format [*extension*] in the world”
  - 598 formats
- WhatIs.com
  - “Every file format [*extension*] in the world”
  - 3,189 formats



## Diffuse web site c/o Internet Archives

**DIFFUSE:**  
Dissemination  
of InFormal  
and Formal

**Standards and Specifications List**

**APPLICATION SPECIFIC**

**Electronic Commerce**

- [Architectures](#)
- [Business semantics](#)
- [Electronic data interchange \(EDI\)](#)
- [Information security](#)
- [Payment](#)
- [Product data](#)

**Sectorial Data Interchange**

- [Geographic information](#)
- [Medical informatics](#)
- [Museum information](#)
- [Scientific information](#)
- [Electronic learning](#)

**GENERAL PURPOSE**

**Information Management**

- [Data classification](#)
- [Metadata interchange](#)
- [Directories](#)
- [Archiving](#)
- [Library information](#)

**Data Representation**

- [Character sets](#)
- [Text-based documents](#)
- [Multimedia/hypermedia](#)
- [Audio](#)
- [Video](#)
- [Raster graphics](#)
- [Vector graphics](#)
- [Colour information](#)

**Communications**

- [File transfer](#)
- [Electronic mail and newsgroups](#)
- [Electronic conferencing](#)
- [Mobile data communication](#)
- [Webcasting](#)
- [Digital television](#)

Thank you for using this service. The Diffuse project, which built on one of the first online services launched by the European Commission in 1995, concluded on 31st January 2003. No decision regarding maintenance of the contents on this website has yet been made.

For details of knowledge technologies developments, and funding opportunities under the IST programme of FP6, visit [http://www.ist-fp6.org](#)

**What's New**

**Reference**

- [Business Guides](#)
- [Standards List](#)
- [Standards Fora List](#)
- [RTD Project List](#)

**News**

- [Electronic Commerce](#)
- [Information Management](#)
- [Information Society RTD](#)
- [Standards Conferences](#)
- [Diffuse Conferences](#)

**User Support**

- [Index](#)
- [Search](#)
- [Help Desk](#)

**Background**

- [About IST](#)
- [About Diffuse](#)
- [Diffuse FAQ](#)
- [RTD Initiatives](#)
- [IPR Statement](#)
- [Disclaimer](#)



## Digital formats for Library of Congress collections

**Digital Formats for Library of Congress Collections**

[Introduction](#) | [Sustainability Factors](#) | [Content Categories](#) | [Format Descriptions](#) | [Contact](#)

The Digital Formats Web site provides information about digital content formats. An initial offering is being compiled during 2004, and the analyses and resources presented here will increase and be updated throughout the year. The compilers, Caroline R. Arms and Carl Fleischhauer, invite [feedback](#) on the content.

**Introduction**  
Background information and overview: What is a format? How shall we evaluate formats? What projects in other organizations are addressing these questions? >>  
[Overview](#) | [Formats, Evaluation Factors, and Relationships](#) | [Papers and Presentations](#) | [Related Resources](#)

**Sustainability Factors**  
What affects the ability of the Library to preserve a given format? These sustainability factors apply to all formats. >>  
[Disclosure](#) | [Adoption](#) | [Transparency](#) | [Self-documentation](#) | [External Dependencies](#) | [Impact of Patents](#) | [Technical Protection Mechanisms](#)

**Content Categories**  
The evaluation of formats must take into account quality and functionality. These factors vary according to the type of content under consideration. The initial offering of four content types will be expanded during 2004. >>  
[Still Image](#) | [Sound](#) | [Textual](#) | [Video](#)

**Format Descriptions**  
Documents with more information about specific formats. >>  
[Browse categories](#) | [Browse alphabetical list](#)

Last updated Friday, 29-Oct-2004 11:58:39 EDT

[NDIIPP Home](#) | [Digital Formats Home](#)



## Digital formats for Library of Congress collections



## What's wrong with MIME types?

- Level of granularity
  - image/tiff vs. TIFF/EP, TIFF/IT, GeoTIFF, Exif, DNG, ...
  - The distinctions between profiles may be significant with regard to preservation workflows
- Level of disclosure
- Level of detail
- Non-actionable



## What's wrong with MIME types?

**MIME TYPE NAME:** application

**MIME SUBTYPE NAME:** msword

**REQUIRED PARAMETERS:** none

**OPTIONAL PARAMETERS:** An optional version parameter can be specified. Some of the more common versions are: 4 : Microsoft Word 4.0 for the Macintosh. 5 : Microsoft Word 5.0 and 5.1 for the Macintosh. 2w : Microsoft Word for Windows 2.0 6 : Microsoft Word 6 for Windows and Macintosh platform independent format (coming soon)

**ENCODING CONSIDERATIONS:** Microsoft word files are in a binary format. Some encoding will be necessary for MIME mailers as in application/octet-stream. Microsoft Word files for the Macintosh are encoded in the data fork of a macintosh file. The type creator is MSWD, the file type is WDBN. Microsoft Word files that contain external data references such as publish & subscribe services are explicitly not allowed.

**SECURITY CONSIDERATIONS:** None known.

**PUBLISHED SPECIFICATION:** Specification by example: From any microsoft word application select "Save As..." from the "File" menu. Enter a filename, make sure that "Normal" is specified for the file type, and click "Save".



## What's wrong with MIME types?

MIME TYPE NAME: application

MIME SUBTYPE NAME: msword

REQUIRED PARAMETERS: none

OPTIONAL PARAMETERS: An optional version parameter can be specified. Some of the more common versions are: 4 : Microsoft Word 4.0 for the Macintosh. 5 : Microsoft Word 5.0 and 5.1 for the Macintosh. 2w : Microsoft Word for Windows 2.0 6 : Microsoft Word 6 for Windows and Macintosh platform independent format (coming soon)

ENCODING CONSIDERATIONS: Microsoft word files are in a binary format. Some encoding will be necessary for MIME mailers as in application/octet-stream. Microsoft Word files for the Macintosh are encoded in the data fork of a macintosh file. The type creator is MSWD, the file type is WDBN. Microsoft Word files that contain external data references such as publish & subscribe services are explicitly not allowed.

SECURITY CONSIDERATIONS: None known.

PUBLISHED SPECIFICATION: Specification by example: From any microsoft word application select "Save As..." from the "File" menu. Enter a filename, make sure that "Normal" is specified for the file type, and click "Save".



## Characteristics of a format registry

- Predictable data
- Arbitrary granularity
- Inclusive
- Machine actionable discovery
- Interoperable
- Trustworthy
  - Authoritative
  - “Honest broker” with regard to proprietary information
- Informative, not evaluative





## The Harvard format registry





## National Archives (UK) PRONOM

The National Archives | PRONOM | Welcome - Mozilla

File Edit View Go Bookmarks Tools Window Help


Back Forward Reload Stop <http://www.nationalarchives.gov.uk/PRONOM/> Search Print

Home Bookmarks mozilla.org mozillaZine mozdev.org

**A** the national archives [Contact us >](#) [Help >](#) [A to Z index >](#) [Site search >](#) Sunday 31 October

Home About us Visit us Exhibitions & Learning online Getting started Search our collections Search other archives Services for professionals News Shop Business services

You are here: [Home](#) > [Services for professionals](#) > [Preservation](#) > PRONOM

 **PRONOM**  
The file format registry

Welcome : About Add an entry  
Search ? Help

**Welcome to PRONOM**

An online source for information about file formats and software products. PRONOM is a resource for anyone requiring impartial and definitive technical information about the file formats used to store electronic records, and the software products that are required to create, render, or migrate these formats.

**Search PRONOM**  
PRONOM stores information on software products used to create or view electronic records. [Search Pronom now >](#)

**Want to contribute?**  
We actively invite and encourage the submission of new information for inclusion on PRONOM, via our online [submission form >](#)

**New to PRONOM?**

- [What is PRONOM? >](#)
- [How do I search PRONOM? >](#)
- [Who is PRONOM for? >](#)
- [How do I find out more? >](#)

Find out more [about PRONOM's creators](#).

**Terms of use | Copyright | Privacy | Top of page ^**  
The National Archives, Kew, Richmond, Surrey, TW9 4DU email: [enquiry@nationalarchives.gov.uk](mailto:enquiry@nationalarchives.gov.uk) tel: +44 (0) 20 8876 3444



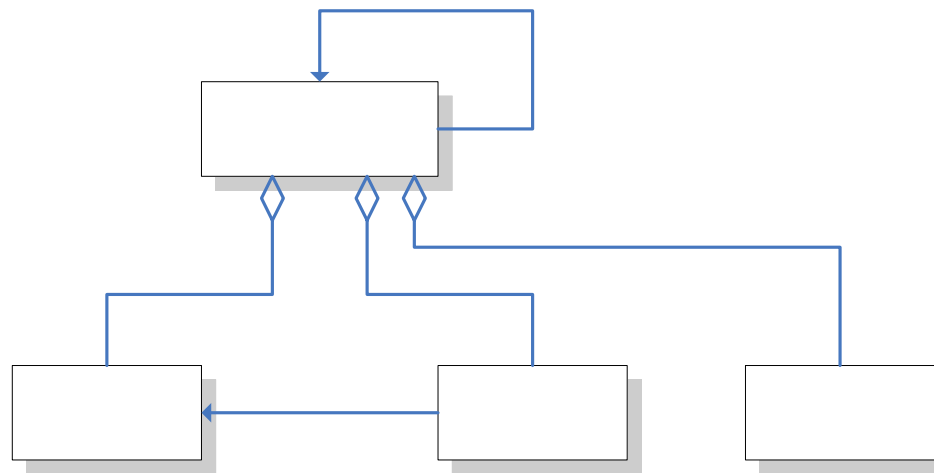
## PRONOM

- Technical format information
  - Currently, information on 546 formats
- Product life-cycle information
- Work underway on:
  - Migration pathway generation
  - Content variance analysis
    - Quantify loss/change introduced by migration



## DCC Representation Registry/Repository

- JISC Digital Curation Centre
- OAIS (ISO 14721) representation net
  - Representation information is itself data that needs representation information for its proper interpretation



[ David Giaretta, *DCC Approach to Digital Curation*, May 28, 2005

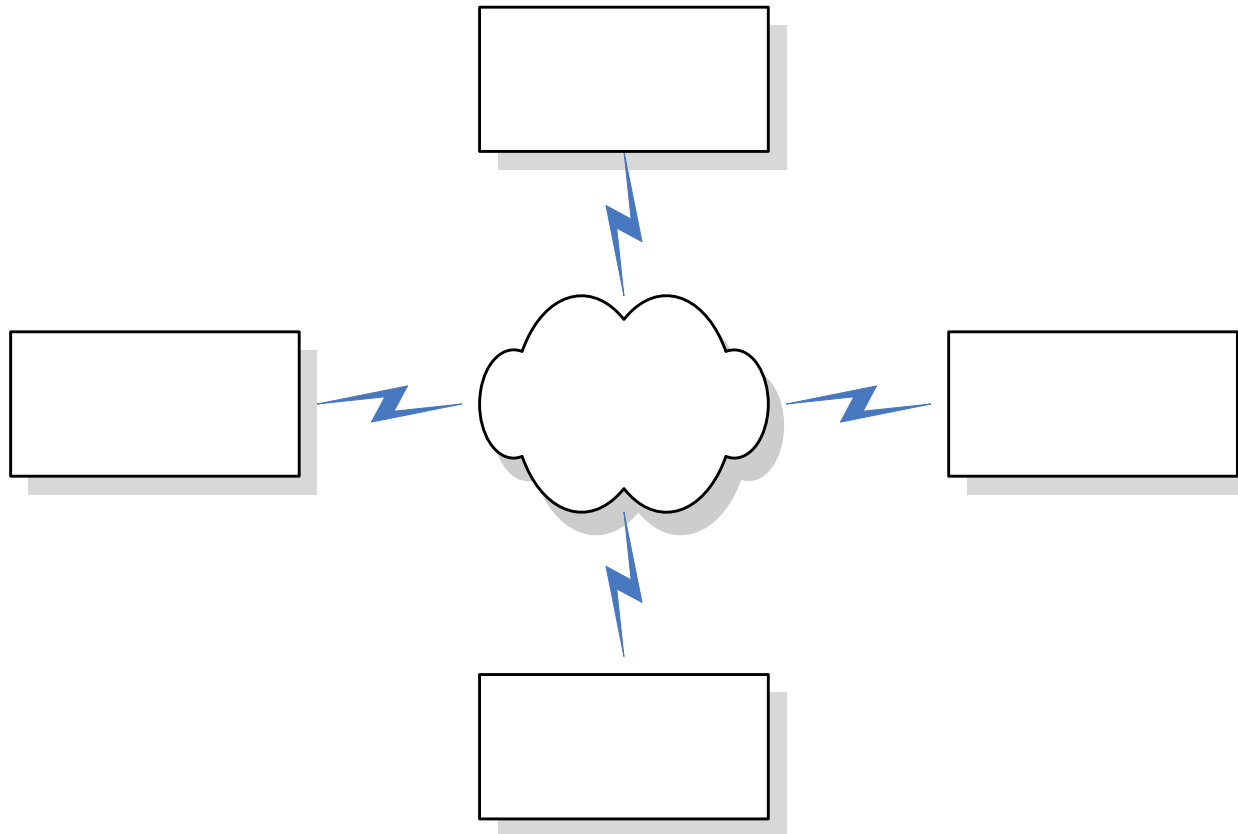


## Global Digital Format Registry (GDFR)

- The Digital Library Federation (DLF) funded two invitational workshops in 2002 to investigate issues surrounding the establishment of a GDFR
  - Bibliothèque nationale de France
  - California Digital Library
  - Digital Library Federation
  - Harvard University
  - Internet Engineering Task Force
  - JISC
  - JSTOR
  - Library of Congress
  - MIT
  - National Archives, UK
  - NARA
  - National Archives of Canada
  - New York University
  - NIST
  - Online Computer Library Center
  - Research Libraries Group
  - Stanford University
  - University of Pennsylvania



## Distributed network of cooperating registries

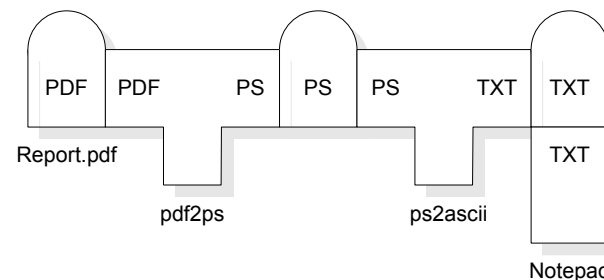


Root  
GDFR ca



## Representation information

- Canonical and variant names
- Signatures
  - Internal and external
- Specifications
- Authors, rights holders, maintenance agencies
- Ontological classification and relationships
- Systems, services, and tools
  - Support for transformative processing chains



[ Adapted from Christensen, "Towards format repositories for web archives,"  
*IWAW04: 4th International Web Archiving Workshop*, Bath, September 16, 2004 ]



## Work so far

- Provisional data and service models
  - Alignment with PRONOM data model
- Proposal for 2 year development project under review
  - Common good services are difficult to fund and sustain
- Demonstration project
  - Built on top of Typed Object Model (TOM) infrastructure





## FRED – A Format REgistry Demonstration

The screenshot shows a Mozilla browser window with the following content:

**Fred: A format registry demonstration** -- [Release 0.10](#)  
To add or change information on this site, you must [log in](#)

---

### Format: info:gdf/fred/f/jpeg

[See permissions](#) -- [See edit history](#) -- [Discuss](#) -- [Show XML](#)

<b>Canonical identifier</b>	info:gdf/fred/f/jpeg		
<b>Description</b>	Joint Photographic Experts Group (JPEG) image format		
<b>Legal or recognized owner</b>	<b>Name</b>	Joint Photographic Experts Group	
	<b>Organization type</b>	Non-profit entity	
	<b>Web site</b>	<a href="http://www.jpeg.org/">http://www.jpeg.org/</a>	
	<b>Note</b>	A joint committee of ISO/IEC and ITU-T	
	<b>Description last modified</b>	7:03:47 PM, on September 29, 2004 (GMT)	
<b>Specification</b>	<b>Document title</b>	ISO/IEC 10918-1:1994, Information technology -- Digital compression and coding of continuous-tone still images: Requirements and guidelines	
	<b>Document type</b>	Standard	
	<b>Publication date</b>	February 15, 1994	
	<b>Access regime</b>	Unrestricted access	
	<b>Identifier</b>	<b>Type</b>	URI: Uniform resource identifier
		<b>Value</b>	<a href="http://www.w3.org/Graphics/JPEG/itu-t81.pdf">http://www.w3.org/Graphics/JPEG/itu-t81.pdf</a>
<b>Note</b>		This URI is for the 1993 recommendation, which might not be identical to the 1994 published standard	
<b>Last modified</b>		7:03:47 PM, on September 29, 2004 (GMT)	



## What are the benefits of GDFR?

- The GDFR is an enabling technology underlying digital repository operations and preservation activities
  - Enables the typing of digital objects at an appropriate level of granularity
  - Enables the future recovery of the syntax and semantics associated with typed digital objects
  - A means to pool and redistribute the expertise of the digital preservation community



## Digital formats and digital preservation

- Agenda

- ✓ *What is digital preservation?*
- ✓ *What is a format?*
- ✓ *OAIS format dependencies*
- ✓ *Format identification, validation, and characterization*
- ✓ *Format risk analysis*
- ✓ *Format registries*
- **Case study**
- Summary



## Archive and Ingest Handling Test (AIHT)

- Test corpus of 57,000 thematically related files (13 GB)
  - No technical metadata
    - Checksum
    - MIME type (not reliable)
- Ingest phase
  - Deposit into Harvard's Digital Repository Service (DRS)
    - Requirements for accompanying technical metadata
  - JHOVE-based Submission Information Package (SIP) packaging tool
    - 97% of all files were in 9 formats
      - AIFF, ASCII, GIF, HTML, JPEG, PDF, TIFF, WAVE, XML



## AIHT ingest phase

	<i>By MIME type</i>	<i>By extension</i>	<i>JHOVE</i>
application/octet-stream	1,141	1,578	3,618
application/pdf	1,663	1,664	1,659
audio/x-aiff	162	151	162
audio/x-wave	2,015	2,016	2,015
image/gif	1,337	1,320	1,339
image/jpeg	12,752	12,763	12,576
image/tiff	1,538	1,533	1,537
text/html	16,677	16,184	3,649
text/plain	20,207	20,822	30,887
text/xml	0	1	8
	57,492	57,492	57,450



## AIHT handling phase

- GIF, JPEG, TIFF-to-JPEG 2000 conversion
- The 15,452 source files were categorized into 25 sub-populations
  - Based on source format, color space, compression, image size
- Aware JPEG 2000 codec
  - No support for GIF as a source format
  - All GIF images were first transformed to RGB TIFF
- Automated and manual QC
  - In the case of RGB-to-RGB transforms (TIFF-to-JP2) the conversion was numerically lossless
  - In the case of YCbCr-to-RGB transforms (JPEG-to-JP2) the conversion resulted in some numerical round-off error ( $\sigma = 0.02$ )
  - In all cases the conversion was “visually” lossless



## Digital formats and digital preservation

- Agenda

- ✓ *What is digital preservation?*
- ✓ *What is a format?*
- ✓ *OAIS format dependencies*
- ✓ *Format identification, validation, and characterization*
- ✓ *Format risk analysis*
- ✓ *Format registries*
- ✓ *Case study*
- Summary



## Summary

- Preservation is for use
- To ensure use, it is necessary that data are retrievable, unchanged, interpretable, and renderable
- To facilitate preservation activities it is important that data are well characterized
- The fundamental characterization property is format
- Format is the byte-serialized encoding of an abstract information model
- Formats can be usefully categorized into classes and families
- A digital object may be composed of more than one formatted bit stream, and may be manifest in more than one file
- Format dependencies exist throughout the OAIS reference model
- Mature tools for format identification, validation, and characterization are emerging
- Format assessment must weigh the probability of risk and its impact
- The digital preservation community needs a sustainable format registry in order to carry out its activities effectively
- Automated conversion and QC of image formats is possible





## More Information

RLG/OCLC <i>Trusted Digital Repositories</i>	<a href="http://www.rlg.org/en/page.php?Page_ID=583"> &lt;www.rlg.org/en/page.php?Page_ID=583&gt;</a>
OAIS/ISO 14721	<a href="http://www.ccsds.org/CCSDS/documents/650x0b1.pdf"> &lt;www.ccsds.org/CCSDS/documents/650x0b1.pdf&gt;</a>
Library of Congress	<a href="http://www.digitalpreservation.gov/formats"> &lt;www.digitalpreservation.gov/formats&gt;</a>
OCLC INFORM	<a href="http://www.oclc.org/services/preservation"> &lt;www.oclc.org/services/preservation&gt;</a>
NLNZ	<a href="http://www.natlib.govt.nz/en/whatsnew/4initiatives.html#extraction"> &lt;www.natlib.govt.nz/en/whatsnew/4initiatives.html#extraction&gt;</a>
JHOVE	<a href="http://hul.harvard.edu/jhove/"> &lt;hul.harvard.edu/jhove/&gt;</a>
Diffuse	<a href="http://web.archive.org/web/20030128052128/http://www.diffuse.org/"> &lt;web.archive.org/web/20030128052128/http://www.diffuse.org/&gt;</a>
IANA MIME registry	<a href="http://www.iana.org/assignments/media-types/"> &lt;www.iana.org/assignments/media-types/&gt;</a>
PRONOM	<a href="http://www.nationalarchives.gov.uk/pronom/"> &lt;www.nationalarchives.gov.uk/pronom/&gt;</a>
DCC	<a href="http://dev.dcc.ac.uk/dccrrt/"> &lt;http://dev.dcc.ac.uk/dccrrt/&gt;</a>
OASIS/ebXML	<a href="http://www.oasis-open.org/committees/regrep/"> &lt;http://www.oasis-open.org/committees/regrep/&gt;</a>
GDFR	<a href="http://hul.harvard.edu/gdfr/"> &lt;hul.harvard.edu/gdfr/&gt;</a>
FRED / TOM	<a href="http://tom.library.upenn.edu/fred/"> &lt;tom.library.upenn.edu/fred/&gt;</a>
AIHT	<a href="http://www.digitalpreservation.gov/about/pr_060904.html"> &lt;http://www.digitalpreservation.gov/about/pr_060904.html&gt;</a>