

I T H A K A

JSTOR | PORTICO | ITHAKA S+R

File Format Considerations in the Preservation of e- Books

Sheila Morrissey

Senior Research Developer, Portico

NISO Webinar: Heritage Lost?

Ensuring the Preservation of E-books

May 23, 2012



PORTICO

Portico is among the largest community-supported digital archives in the world.

Working with libraries, publishers, and funders, we preserve e-journals, e-books, and other electronic scholarly content to ensure researchers and students will have access to it in the future.





PORTICO

Over 2,000 societies, and associations have committed content to Portico through 147 publishers agreements.

Committed Content

»	E-journal titles	13,675
»	E-book titles	129,781
»	D-collections	46





PORTICO

Preserved Content

»	E-journal titles	9,568
»	E-book titles	16,861
»	D-collections	12
»	Archival Units	19,433,869
»	Preserved Files	319,737,011



Portico - Audit and Certification

In 2010, Portico became the first digital preservation service to be independently audited by the Center for Research Libraries (CRL) and subsequently certified as a trusted, reliable digital preservation solution that serves the needs of the library community.

Trustworthy Repositories
Audit & Certification:
Criteria and Checklist



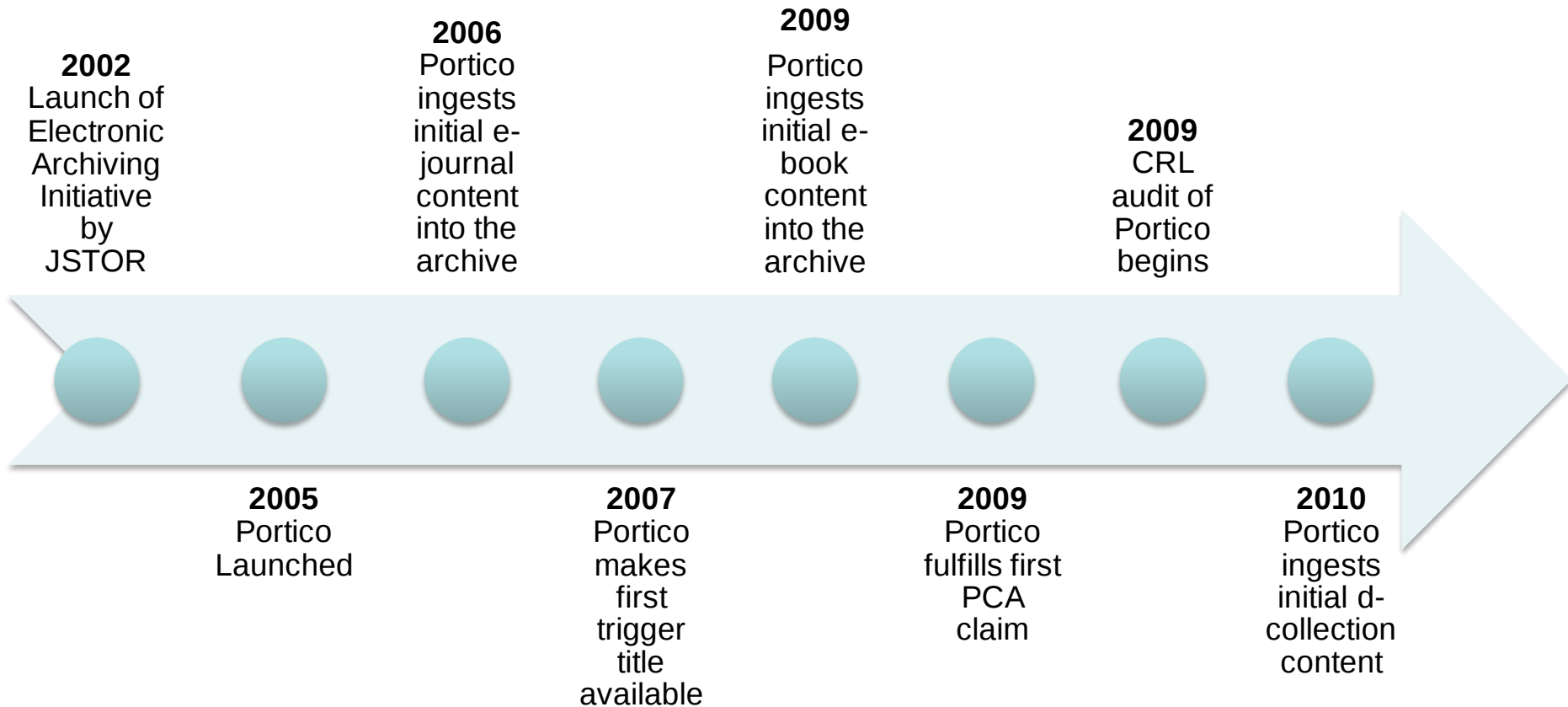
preservation repository CRL specifications certification
criteria RLG Programs oclc audit digital object management
NARA trustworthy metadata preservation repository
CRL specifications certification criteria RLG Programs
oclc audit digital object management NARA trustwor-
thy metadata preservation repository CRL specifications
certification criteria RLG Programs oclc audit digital
object management NARA trustworthy metadata preser-
vation repository CRL specifications certification criteria
RLG Programs oclc audit digital object management NARA
trustworthy metadata preservation repository CRL
specifications certification criteria RLG Programs oclc au-
dit digital object management NARA trustworthy meta-
data

Contents:

Introduction
Establishing Audit and Certification Criteria
Towards an International Audit & Certification Process
Using this Checklist for Audit & Certification
Applicability of Criteria
Relevant Standards, Best Practices & Controls
Terminology
Audit and Certification Criteria
Organizational Infrastructure
Digital Object Management
Technologies, Technical Infrastructure & Security
Audit Checklist
Glossary
Appendices

Version 1.0
February 2007

Portico - History



Digital preservation is the series of management policies and activities necessary to ensure the enduring usability, authenticity, discoverability, and accessibility of content over the very long-term. The key goals of digital preservation include:

Usability

- the intellectual content of the item must remain usable via the delivery mechanism of current technology

Authenticity

- the provenance of the content must be proven and the content an authentic replica of the original

Discoverability

- the content must have logical bibliographic metadata so that it can be found by end users through time

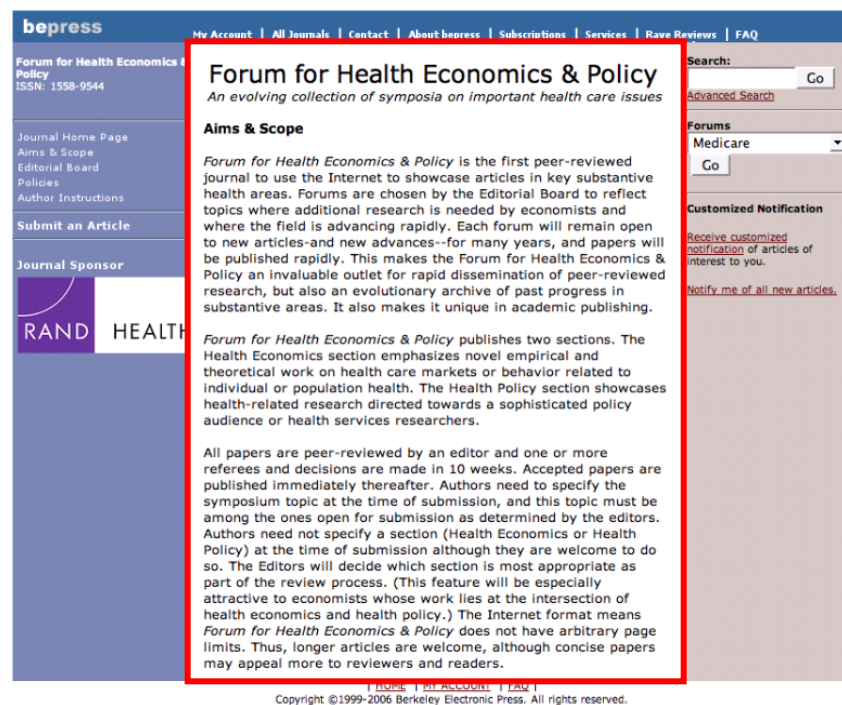
Accessibility

- the content must be available for use to the appropriate community

Legal right to preserve content

- » Not always the same as access rights
- » Specified in contracts
- » Includes embedded or supplemental files, such as images
- » DRM removed

Usability - Preserve Intellectual Content



The screenshot shows the website for the Forum for Health Economics & Policy, hosted on bepress. The page has a blue header with navigation links: Home Account, All Journals, Contact, About bepress, Subscriptions, Services, Rawe Reviews, and FAQ. The left sidebar contains links for the Forum for Health Economics & Policy (ISSN: 1558-9544), Journal Home Page, Aims & Scope, Editorial Board, Policies, Author Instructions, Submit an Article, and Journal Sponsor (RAND HEALTH). The main content area, highlighted by a red box, features the title "Forum for Health Economics & Policy" with the subtitle "An evolving collection of symposia on important health care issues". Below this is the "Aims & Scope" section, which describes the forum as a peer-reviewed journal using the Internet to showcase articles in key substantive health areas. It mentions that forums are chosen by the Editorial Board to reflect topics where additional research is needed by economists and where the field is advancing rapidly. Each forum remains open to new articles and new advances for many years, and papers will be published rapidly. This makes the forum an invaluable outlet for rapid dissemination of peer-reviewed research, but also an evolutionary archive of past progress in substantive areas. It also makes it unique in academic publishing. The text continues by stating that the forum publishes two sections: the Health Economics section, which emphasizes novel empirical and theoretical work on health care markets or behavior related to individual or population health, and the Health Policy section, which showcases health-related research directed towards a sophisticated policy audience or health services researchers. Finally, it notes that all papers are peer-reviewed by an editor and one or more referees, and decisions are made in 10 weeks. Accepted papers are published immediately thereafter. Authors need to specify the symposium topic at the time of submission, and this topic must be among the ones open for submission as determined by the editors. Authors need not specify a section (Health Economics or Health Policy) at the time of submission although they are welcome to do so. The Editors will decide which section is most appropriate as part of the review process. (This feature will be especially attractive to economists whose work lies at the intersection of health economics and health policy.) The Internet format means that the forum does not have arbitrary page limits. Thus, longer articles are welcome, although concise papers may appeal more to reviewers and readers. To the right of the main content area, there is a search bar with a "Go" button, a link to "Advanced Search", a "Forums" section with a dropdown menu set to "Medicare" and a "Go" button, and a "Customized Notification" section with a link to "Receive customized notification of articles of interest to you." and a link to "Notify me of all new articles." At the bottom of the page, there are links for "HOME", "MY ACCOUNT", and "FAQ", and a copyright notice: "Copyright ©1999-2006 Berkeley Electronic Press. All rights reserved."

Forum for Health Economics & Policy

An evolving collection of symposia on important health care issues

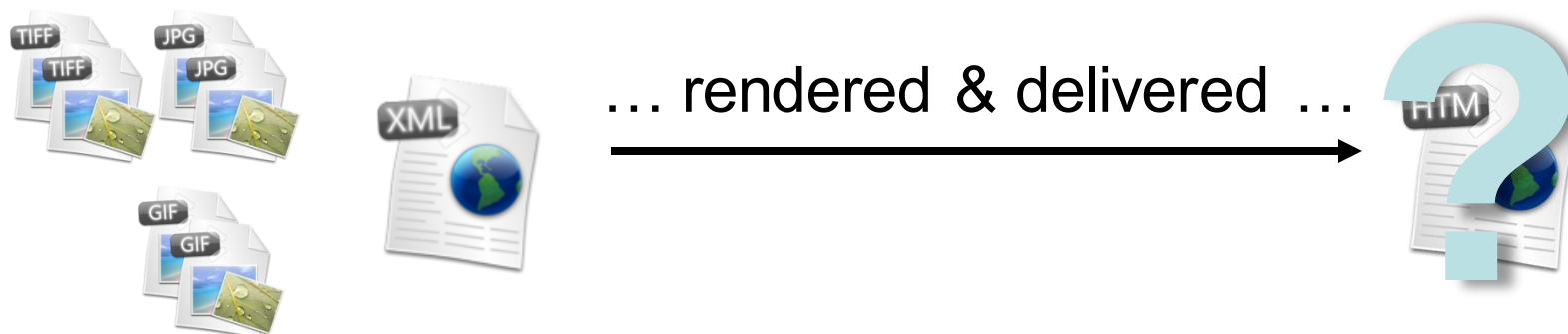
Aims & Scope

Forum for Health Economics & Policy is the first peer-reviewed journal to use the Internet to showcase articles in key substantive health areas. Forums are chosen by the Editorial Board to reflect topics where additional research is needed by economists and where the field is advancing rapidly. Each forum will remain open to new articles-and new advances--for many years, and papers will be published rapidly. This makes the *Forum for Health Economics & Policy* an invaluable outlet for rapid dissemination of peer-reviewed research, but also an evolutionary archive of past progress in substantive areas. It also makes it unique in academic publishing.

Forum for Health Economics & Policy publishes two sections. The Health Economics section emphasizes novel empirical and theoretical work on health care markets or behavior related to individual or population health. The Health Policy section showcases health-related research directed towards a sophisticated policy audience or health services researchers.

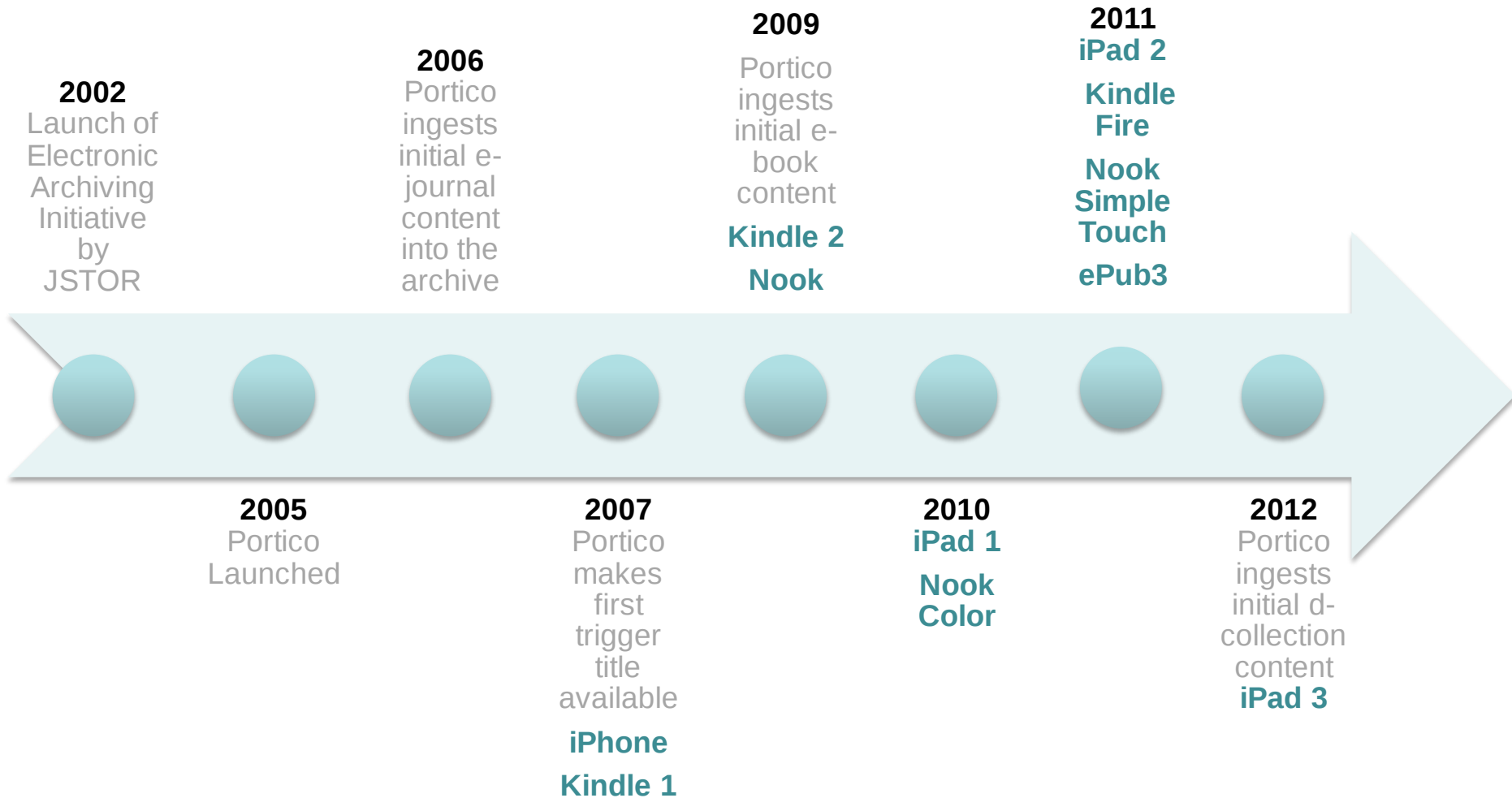
All papers are peer-reviewed by an editor and one or more referees and decisions are made in 10 weeks. Accepted papers are published immediately thereafter. Authors need to specify the symposium topic at the time of submission, and this topic must be among the ones open for submission as determined by the editors. Authors need not specify a section (Health Economics or Health Policy) at the time of submission although they are welcome to do so. The Editors will decide which section is most appropriate as part of the review process. (This feature will be especially attractive to economists whose work lies at the intersection of health economics and health policy.) The Internet format means *Forum for Health Economics & Policy* does not have arbitrary page limits. Thus, longer articles are welcome, although concise papers may appeal more to reviewers and readers.

Content is rendered to support current delivery platform, i.e. web browser.



Rendition engine can be modified to meet new technology requirements.

Portico – Another Look at the History



Usability: Anticipated usage ...

EPUB Content Documents 3.0



Recommended Specification 11 October 2011

THIS VERSION

<http://www.idpf.org/epub/30/spec/epub30-contentdocs-20111011.html>

LATEST VERSION

<http://www.idpf.org/epub/30/spec/epub30-contentdocs.html>

PREVIOUS VERSION

<http://www.idpf.org/epub/30/spec/epub30-contentdocs-20110908.html>

A diff of changes from the previous draft is available at [this link](#).

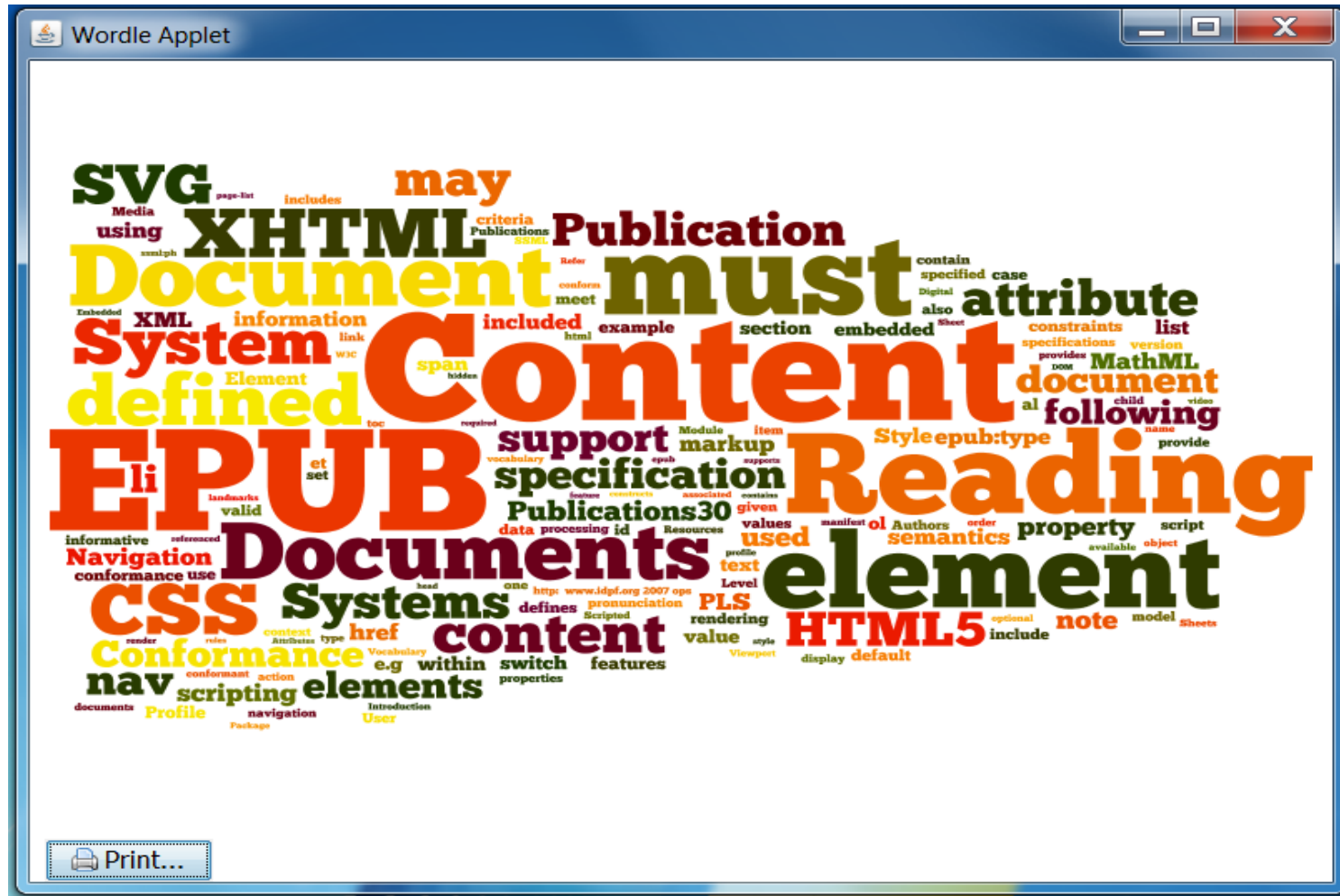
Please refer to the [errata](#) for this document, which may include some normative corrections.

Copyright © 2010, 2011 International Digital Publishing Forum™



PORTICO

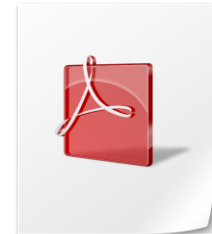
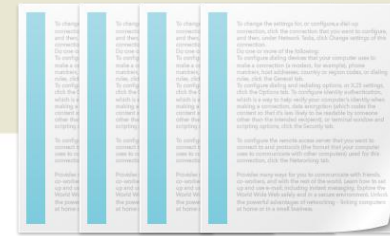
Usability: ... and new usage



Authenticity, Discoverability: Preservation Context



Preservation and
Packaging
Metadata File



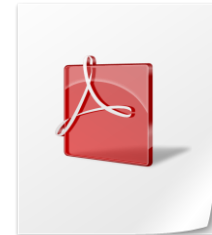
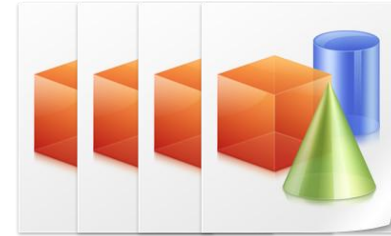
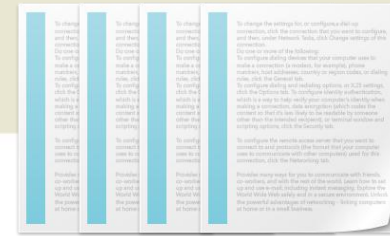
Context

- The intellectual unit represented by this metadata file is a digitized book.
- It was scanned by Joe on this date.
- It was ingested into the repository on this other date.
- Jane Smith granted us preservation rights to it on this other date.

...



Preservation and
Packaging
Metadata File



Context

- The intellectual unit represented by this metadata file is a digitized book.
- It was scanned by Joe on this date.
- It was ingested into the repository on this other date.
- Jane Smith granted us preservation rights to it on this other date.

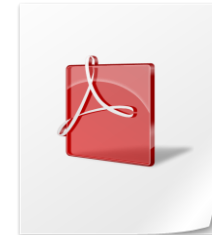
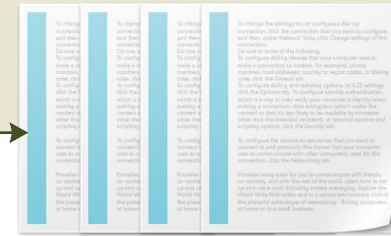
...



Preservation and Packaging Metadata File

- These TIF files are page images.
- The TIF file named XYZ is page 1. It is a valid TIF and has a checksum of 123456.
- The TIF file named ABC is page 2. It is not a valid TIF and has a checksum of 78910.

...



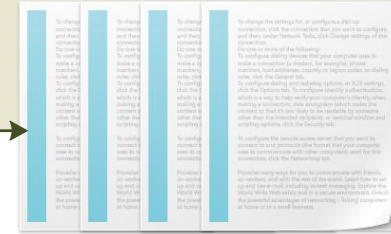
Context

- The intellectual unit represented by this metadata file is a digitized book.
- It was scanned by Joe on this date.
- It was ingested into the repository on this other date.
- Jane Smith granted us preservation rights to it on this other date.
- ...

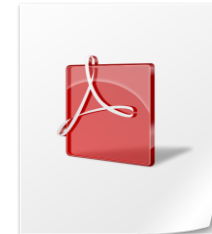


Preservation and Packaging Metadata File

- These TIF files are page images.
- The TIF file named XYZ is page 1. It is a valid TIF and has a checksum of 123456.
- The TIF file named ABC is page 2. It is not a valid TIF and has a checksum of 78910.
- ...



- These JPG files are figures.
- The JPG file named MNO is the 2nd figure on page 2. It is a valid JPG and has a checksum of 234567.
- ...



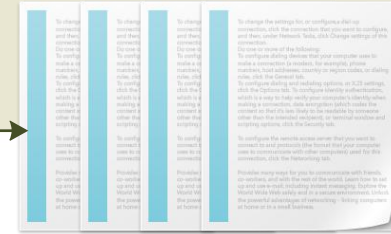
Context

- The intellectual unit represented by this metadata file is a digitized book.
- It was scanned by Joe on this date.
- It was ingested into the repository on this other date.
- Jane Smith granted us preservation rights to it on this other date.
- ...



Preservation and Packaging Metadata File

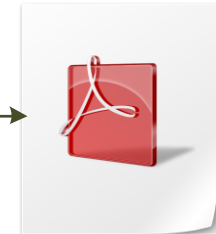
- These TIF files are page images.
- The TIF file named XYZ is page 1. It is a valid TIF and has a checksum of 123456.
- The TIF file named ABC is page 2. It is not a valid TIF and has a checksum of 78910.
- ...



- These JPG files are figures.
- The JPG file named MNO is the 2nd figure on page 2. It is a valid JPG and has a checksum of 234567.
- ...



- This PDF file contains page images.
- The page images are built from TIF files XYZ, ABC, etc. and JPG figure graphics MNO, etc.
- ...



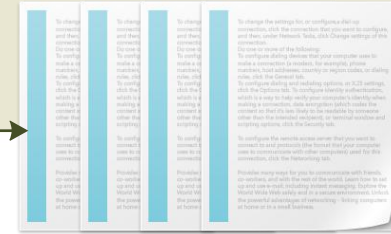
Context

- The intellectual unit represented by this metadata file is a digitized book.
- It was scanned by Joe on this date.
- It was ingested into the repository on this other date.
- Jane Smith granted us preservation rights to it on this other date.
- ...



Preservation and Packaging Metadata File

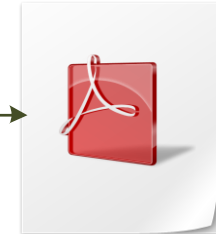
- These TIF files are page images.
- The TIF file named XYZ is page 1. It is a valid TIF and has a checksum of 123456.
- The TIF file named ABC is page 2. It is not a valid TIF and has a checksum of 78910.
- ...



- These JPG files are figures.
- The JPG file named MNO is the 2nd figure on page 2. It is a valid JPG and has a checksum of 234567.
- ...



- This PDF file contains page images.
- The page images are built from TIF files XYZ, ABC, etc. and JPG figure graphics MNO, etc.
- ...



- This MARC file is the bibliographic record for the book.
- ...



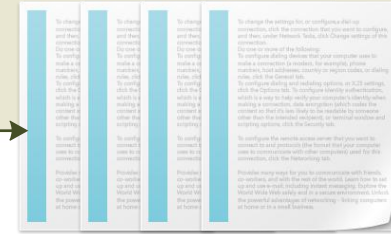
Context

- The intellectual unit represented by this metadata file is a digitized book.
- It was scanned by Joe on this date.
- It was ingested into the repository on this other date.
- Jane Smith granted us preservation rights to it on this other date.
- ...



Preservation and Packaging Metadata File

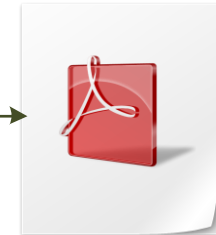
- These TIF files are page images.
- The TIF file named XYZ is page 1. It is a valid TIF and has a checksum of 123456.
- The TIF file named ABC is page 2. It is not a valid TIF and has a checksum of 78910.
- ...



- These JPG files are figures.
- The JPG file named MNO is the 2nd figure on page 2. It is a valid JPG and has a checksum of 234567.
- ...



- This PDF file contains page images.
- The page images are built from TIF files XYZ, ABC, etc. and JPG figure graphics MNO, etc.
- ...



- This MARC file is the bibliographic record for the book.
- ...



- This XML file contains the full-text of the book.
- It uses the QRS DTD.
- It is named JKL and has a checksum of 555555.
- ...



- ```

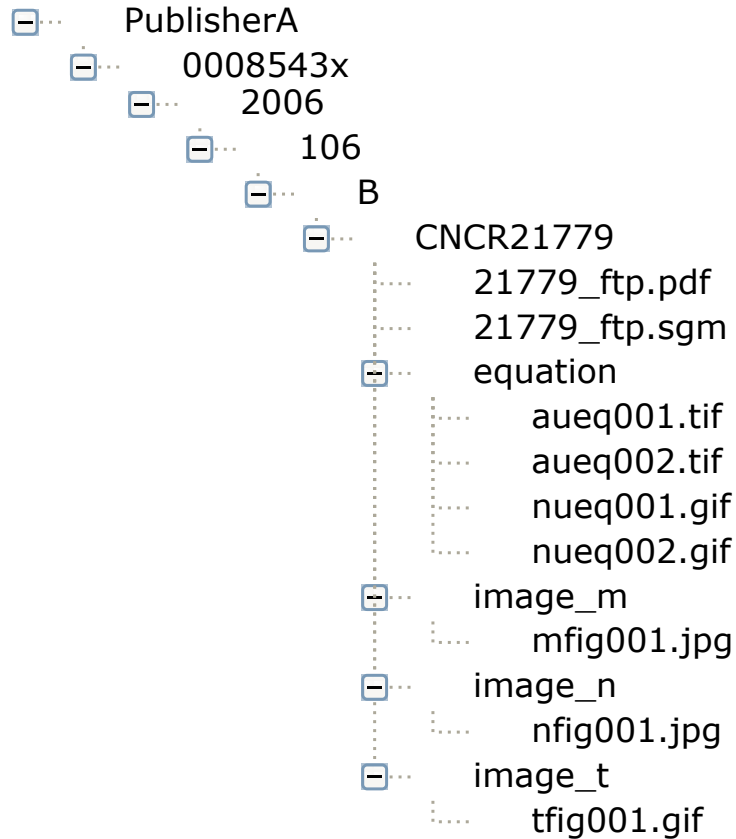
└─ ArchivalUnit "ark:/27927/pbb45qh1t"
 └─ ProcessingRecord "ark:/27927/pbb5rnpnq" 2010-07-15T13:45:42.553-04:00
 └─ Timestamp 2010-07-15T13:45:42.553-04:00
 └─ Rationale Processing of New Content
 └─ SuppliedFiles "ark:/27927/pzb5nbsf8"
 └─ System "ConPrep"
 └─ Workflow "E-Journal Workflow 2.0"
 └─ Batch "ark:/27927/pgf18mnt35"
 └─ Profile "Springer E-Book Profile"
 └─ RuntimeEnv "Java:Sun Microsystems Inc.:1.6.0_18"
 └─ ToolRegistry "2.64"
 └─ ProcessingRecord "ark:/27927/pbb60h0wg" 2010-07-16T13:54:18.517-04:00
 └─ Timestamp 2010-07-16T13:54:18.517-04:00
 └─ Rationale Processing of New Content
 └─ System "Archive"
 └─ RuntimeEnv "Java:Sun Microsystems Inc.:1.6.0_18"
 └─ ToolRegistry "2.64"
 └─ EventSet "ark:/27927/pbb5rnn9f"
 └─ Event "ark:/27927/pbb5rnnbx" 2010-07-15T13:45:36.784-04:00
 └─ Timestamp 2010-07-15T13:45:36.784-04:00
 └─ InputList
 └─ Output "ark:/27927/pbb5rnn8z"
 └─ ToolWrapper "DmdExtractTool:1.0:2007-08-03"
 └─ ToolComponentList
 └─ DescriptiveMetadata "ark:/27927/pbb5rnn8z"
 └─ DC "http://www.ithaka.org/standards/Public/XML/schema/IthakaDublinCoreMetadata/1.0"
 └─ title "http://www.ithaka.org/standards/Public/XML/schema/dc-elements/1.1/" Polyat
 └─ contributor "http://www.ithaka.org/standards/Public/XML/schema/dc-elements/1.1/"
 └─ identifier "http://www.ithaka.org/standards/Public/XML/schema/dc-elements/1.1/" IS
 └─ identifier "http://www.ithaka.org/standards/Public/XML/schema/dc-elements/1.1/" IS
 └─ publisher "http://www.ithaka.org/standards/Public/XML/schema/dc-elements/1.1/" S
 └─ rights "http://www.ithaka.org/standards/Public/XML/schema/dc-elements/1.1/" 1995
 └─ language "http://www.ithaka.org/standards/Public/XML/schema/dc-elements/1.1/" E
 └─ type "http://www.ithaka.org/standards/Public/XML/schema/dc-elements/1.1/" Text
 └─ identifier "http://www.ithaka.org/standards/Public/XML/schema/dc-elements/1.1/" ct
 └─ identifier "http://www.ithaka.org/standards/Public/XML/schema/dc-elements/1.1/" W
 └─ ContentUnit "ark:/27927/pbb45qh29"
 └─ EventSet "ark:/27927/pbb48c4cq"
 └─ EventSet "ark:/27927/pbb60h0th"
 └─ DescriptiveMetadata "ark:/27927/pbb48c4b7"
 └─ FunctionalUnit "ark:/27927/pbb45qh3s"
 └─ StorageUnit "ark:/27927/pbb44f7q6"
 └─ EventSet "ark:/27927/pbb44f7rp"
 └─ TechnicalMetadata "ark:/27927/pbb46gr4w"
 └─ File "false"
 └─ StorageUnit "ark:/27927/pbb44f7x3"
 └─ StorageUnit "ark:/27927/pbb44f84f"
 └─ StorageUnit "ark:/27927/pbb44f8bb"
 └─ StorageUnit "ark:/27927/pbb44f8j7"
 └─
 └─
 └─
 └─ FunctionalUnit "ark:/27927/pbb45rpx1"
 └─ MetadataStorageUnit "ark:/27927/pbb5rjgsq"
 └─ File "pbb5rjgsq.pmd"

```

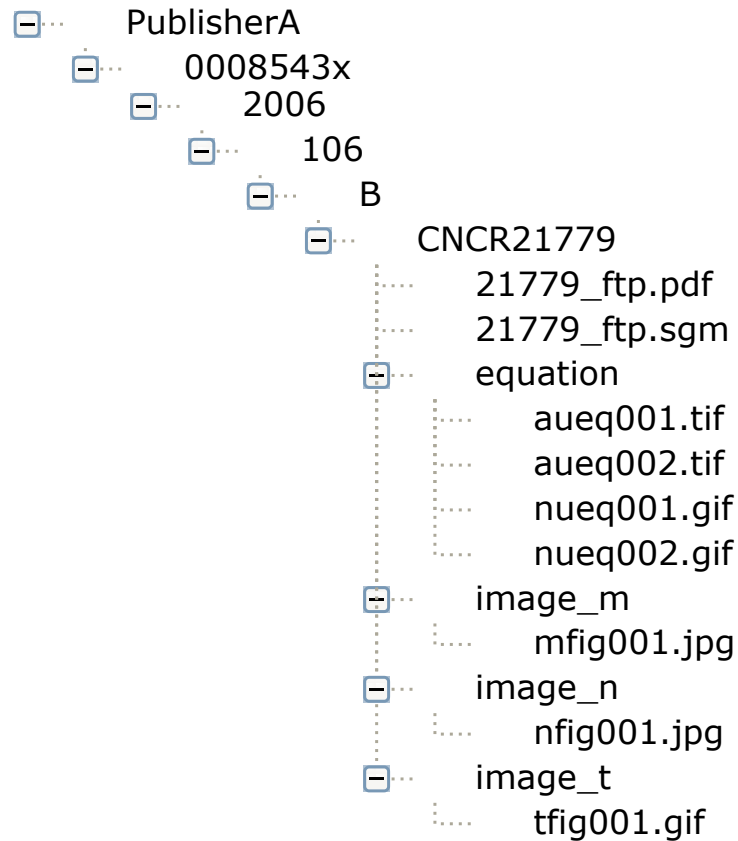


# Formats: Packages

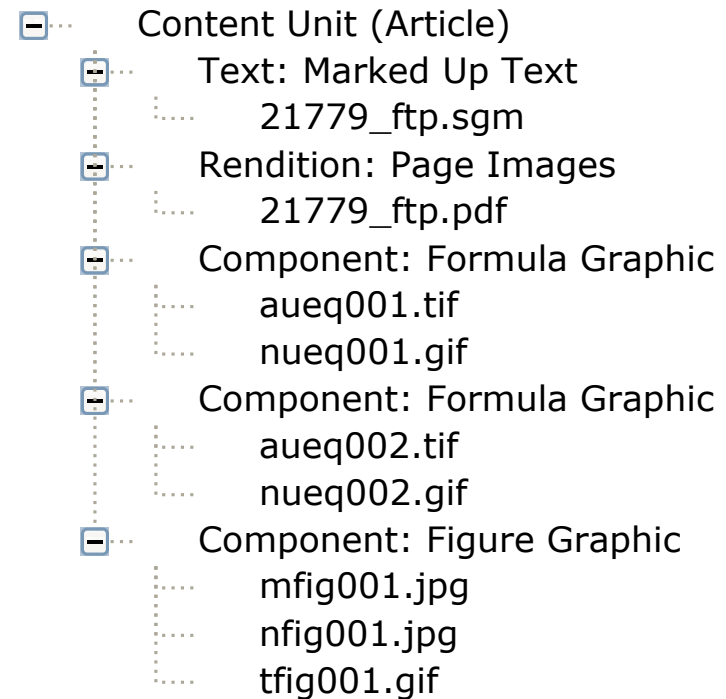
## Incoming File System



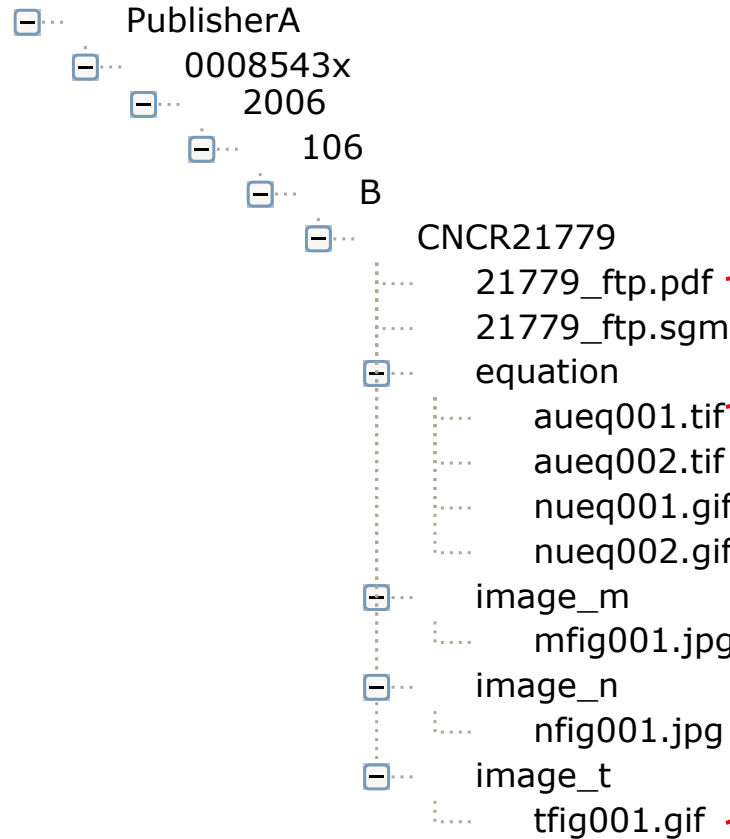
## Incoming File System



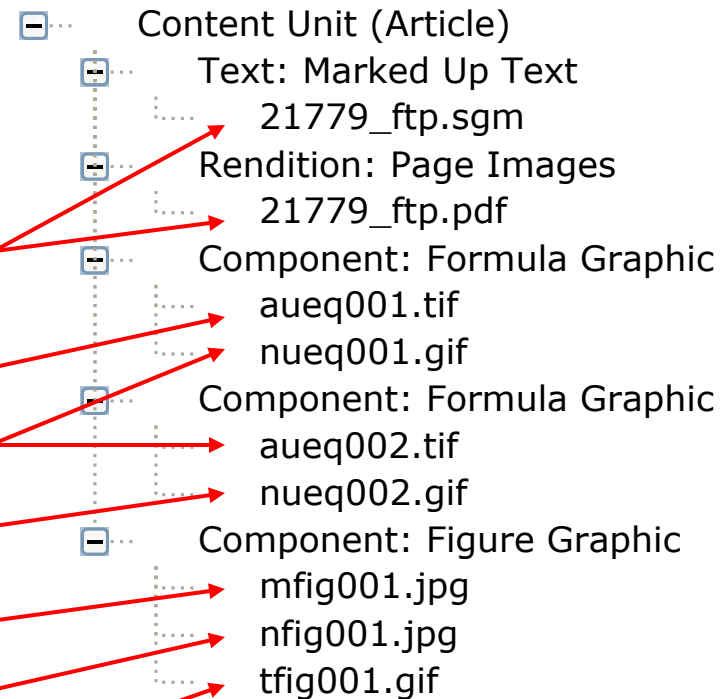
## Resulting Content Model



## Incoming File System



## Resulting Content Model



## Flat directory

- » ONIX xml file with bibliographic metadata, one PDF file per book
  - Front Cover image JPG files

## TAR file (multiple books per file)

- » XML manifest file
- » One directory for each book,
  - Proprietary XML file (3 possible versions of XML) with bibliographic metadata,
  - Subdirectory with files for front matter “chapters” (XML, PDF, OCR of PDF)
  - Subdirectory with files for regular “chapters” (XML, PDF, OCR of PDF) front
  - Subdirectory with files for back matter “chapters” (XML, PDF, OCR of PDF)
  - Subdirectory with TIFF file for cover image of book

## ZIP file (sometimes one book per file, sometime multiple books)

- » Sometimes flat (all books at one level)
- » Sometimes one directory for each book,
  - Sometimes cover images (JPG or TIFF)
  - Sometimes one PDF for entire book in addition to PDF for each chapter
- » Sometimes a manifest

# Hello, World!!



BT

/H2 <</MCID 0 >>BDC

/CS0 cs 0.31 0.506

0.741 scn

/TT0 1 Tf

-0.004 Tc 0.006 Tw

12.96 0 0 12.96 72

697.68 Tm

[ (H) -4 (e) -1 (l) -1 (l) -

11 (o, ) -3 ( W) -15 (or) -

6 (l) -11 (d!) -12 (!) ]TJ

0 Tc 0 Tw 6.481 0 Td

( )Tj

EMC

ET

# Hello, World!!



# Formats: Text Content

```
<html>
<head>
<style type="text/css">
<!--
 p { color: #4F81BD;
font-family: serif;
font-weight: bold;
font-size: 13pt; }
 -->
</style>
</head>
<body><p>Hello,
World!!</p></body>
</html>
```

# Hello, World!!



## Title (Title Style)

Author Name(s) (do not put prefixes, titles, or job description such as Dr., CEO, lab manager, etc.); Professional Affiliation; City, State/Country (Author Style)

### Abstract

Please include a brief abstract of this paper. Avoid using figures or equations in the abstract.

### Overall Document Guidelines: Head

Clear document of any fonts other than Times New Roman, Arial, and Symbol. The paper should be formatted using the tags provided in the template, i.e., title, author, head, subhead, tertiary head, body copy, eq./fig., references, etc. in a 2 column format on 8.5 x 11 inches (US letter-size paper; 21.6 cm x 27.9 cm).

Set the left and right margins to .75 inch (1.90 cm), and the top and bottom margins to 1.0 inch (2.54 cm). Set the document to a 2-column format with column widths set at 3.38 inches (8.57 cm) and the gutter—the space between columns—at .25 inch (.635 cm).

Please do not place folios or page numbers in your paper. That information is inserted when we assemble the book.

### Graphics and Equations

Graphics and equations should fit within one column (approximately 3.38 inches wide), but full width (7") figures are also acceptable. Equations, figures and figure captions each have their own style tags. Number equations using parentheses flushed right as shown below.

$$IS\&T + members \times Confs. = Success \quad (1)$$

### Helpful Hints and Style Tags: Subhead

For a complete listing of the style tags for use in this template refer to Table 1. These are the style tags for conference proceedings; if you use the wrong template/style tags your paper will be sent back to you to be reformatted. All of these forms and templates related to the publication of conference papers are available at [1]

[www.imaging.org/conferences/guidelines.cfm](http://www.imaging.org/conferences/guidelines.cfm).

Select the specific conference and download the Authors Kit. The template may vary from one conference to another.

Style Tag Table: Table head

1. Title	18 pt Times New Roman Bold, flush left; line spacing exactly at 22 pt; 11 pt. after
2. Byline: Author/Affiliation	8 pt. Arial Bold Italic, flush left; line spacing exactly at 11 pt; 22 pt. after
3. Head	11 pt. Arial Bold, flush left; line spacing exactly at 11 pt; Spacing at 11 pt before
4. Subhead	10 pt. Arial Bold Italic, flush left; line spacing exactly at 11 pt. Spacing at 11 pt before
5. Tertiary head	10 pt. Times New Roman Bold Italic, flush left; line spacing exactly at 11 pt. Spacing at 11 pt before
6. Abstract	9 pt. Times New Roman Italic, justified; indent first line by .25"; line spacing exactly at 11 pt. No spacing before or after
7.0. Body text	9 pt Times New Roman, justified; indent first line by .25"; line spacing exactly at 11 pt; Non-justified body text may be used if justification is inappropriate.
8.0 Equation	Equation flush left, equation number flush right. Base font of equation n 9 pt.
8.1 Figure	Flush left line spacing at least 22 pt. Bold the word "figure" the figure number.
8.2 Figure caption	7 pt. Arial Italic, flush left; line spacing at exactly 11 pt;

Hello, World!!



# Hello, World!!

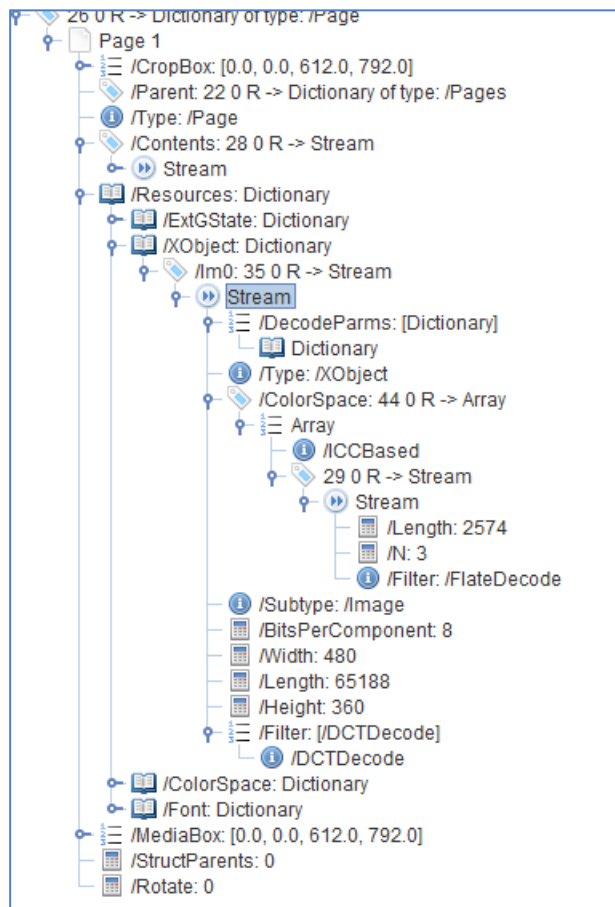


```
BT
/H2 <</MCID 0 >>BDC
/CS0 cs 0.31 0.506 0.741 scn
/TT0 1 Tf
-0.004 Tc 0.006 Tw 12.96 0 0 12.96 264
697.68 Tm
[(H)-4(e)-1(l)-2(l)-11(o,)-3(W)-15(or)-
6(l)-11(d!)-12(!)]TJ
0 Tc 0 Tw 6.481 0 Td
()Tj
EMC
/P <</MCID 1 >>BDC
/CS1 cs 0 scn
/TT1 1 Tf
11.04 0 0 11.04 72 682.08 Tm
()Tj
EMC
/P <</MCID 2 >>BDC
36.478 -24.185 Td
()Tj
EMC
ET
/Figure <</MCID 3 >>BDC
q
/GS0 gs
336 0 0 252 139.1000061 414.6812744 cm
/Im0 Do
Q
EMC
```

# Hello, World!!



## Hello, World!!



(iText RUPS)



```
<html>
<head>
<style type="text/css">
<!--
 p { color: #4F81BD;
font-family: serif;
font-weight: bold;
font-size: 13pt; }-->
</style>
</head>
<body><p>Hello, World!!

<IMG
width="447"
height="336"
src="images/Image_001.j
pg" /></p></body>
</html>
```

# Hello, World!!



# Trade-offs: Encapsulation vs. Articulation

**mydir/  
myFile.pdf**

**mydir/  
myFile.html  
images/  
Image01.jpg**



## PDF

- » One file per chapter
- » One file per book

## TIFF

- » One file per page

## JPEG

- » One file per page

## XML

- » For bibliographic metadata
- » Proprietary
- » ONIX variants
- » NLM variants



## EPUB 3 (<http://idpf.org/epub/30>)

- » “EPUB defines a means of representing, packaging and encoding structured and semantically enhanced Web content-- including HTML5, CSS, SVG, images, and other resources-- for distribution in a single-file format.”

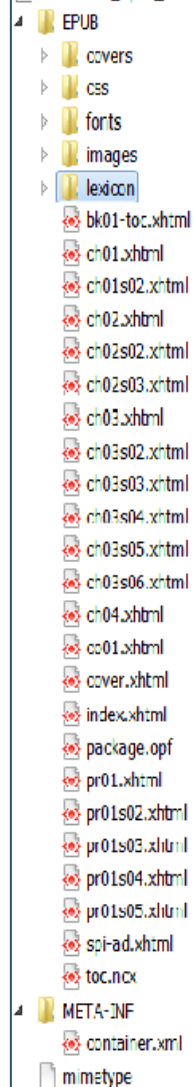


## EPUB 3

- » Web standards for key component technologies
- » Free and open specification
- » Must work in at least *some* appliance
  - Outside publisher's own workflow



accessible\_epub\_3-20120425.epub



```
<?xml version="1.0" encoding="utf-8" standalone="no"?>
<container xmlns="urn:oasis:names:tc:opendocument:xmlns:container"
version="1.0">
 <rootfiles>
 <rootfile full-path="EPUB/package.opf"
media-type="application/oebps-package+xml"/>
 </rootfiles>
</container>
```

- package "http://www.idpf.org/2007/opf"
  - ▷ ● metadata urn:isbn:9781449328030
  - ▷ ● manifest
  - ▷ ● spine "ncx"
  - ▲ ● guide
    - reference "cover.xhtml"
    - reference "bk01-toc.xhtml"



PORTICO

## **“Profiles” of standard formats for authoring content**

- » XHTML5, SVG 1.1, CSS 2.1, CSS 3
  - Constraints (extensions to HTML5, constraints on SVG)
  - Specs a “moving target”

## **Conforming readers must support rendition of certain formats**

- » Image, audio, video
  - Defined fallbacks

## **Globalization, Encoding, Fonts**



## Amazon

- » Announces it is replacing MOBI with K8

## iBooks

- » Different mimetype
- » Proprietary extension of CSS Media Queries
- » Proprietary XML namespace
- » Etc.

# Complications: *"More What You'd Call 'Guidelines' Than Actual Rules"*

I T H A K A  
JSTOR | PORTICO | ITHAKA S+R



*Pirates of the Caribbean: The Black Pearl. The Walt Disney Company (2003)*



PORTICO



PORTICO

## Questions or Comments?

Sheila Morrissey  
[sheila.morrissey@ithaka.org](mailto:sheila.morrissey@ithaka.org)  
[@sheilaMorr](https://twitter.com/sheilaMorr)  
[www.portico.org](http://www.portico.org)



PORTICO