

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, 1012

Portico - Third Party Preservation





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

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



Portico - Participating Content





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

Committed Content

>>	E-journal titles	13,675
		10,07

» E-book titles 129,781

» D-collections 46



Portico – Preserved Content





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 oct. audit digital object management NARA trustworthy metadata preservation repository CRL specifications Certification criteria RLG Programs oct. audit digital object management NARA trustworthy metadata preservation repository CRL specifications certification criteria RLG Programs oct. audit digital object management NARA trustworthy metadata preservation repository CRL specifications certification criteria RLG Programs oct. audit digital object management NARA trustworthy metadata preservation repository CRL specifications certification criteria RLG Programs oct. audit digital object management NARA trustworthy metadata preservation repository CRL specifications Certification criteria RLG Programs oct. audit digital object management NARA trustworthy metadata preservation repository CRL specifications certification criteria RLG Programs oct. audit digital object management NARA trustworthy metadata

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



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

Portico - History

2002
Launch of
Electronic
Archiving
Initiative
by
JSTOR

2006
Portico ingests initial e-journal content into the archive

Portico ingests initial e-book content into the archive

2009

2009 CRL audit of Portico begins

















2005 Portico Launched 2007
Portico
makes
first
trigger
title
available

2009
Portico
fulfills first
PCA
claim

2010
Portico
ingests
initial dcollection
content



Digital Preservation



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



Preservation: Legal aspects

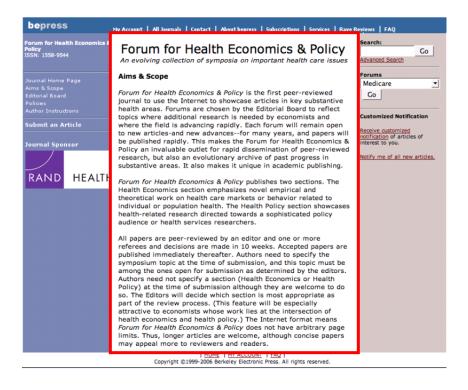


Legal right to preserve content

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



Usability - Preserve Intellectual Content





Usability - Preserve Intellectual Content

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.



Usability: Rendition and Delivery

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





... rendered & delivered ...



Rendition engine can be modified to meet new technology requirements.



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

Portico – Another Look at the History

2002

Launch of Electronic Archiving Initiative by JSTOR

2006

Portico ingests initial e-journal content into the archive

2009

Portico ingests initial ebook content

Kindle 2

Nook

2011

iPad 2

Kindle Fire

Nook Simple Touch

ePub3

















2005

Portico Launched

2007

Portico makes first trigger title available

iPhone

Kindle 1

2010

iPad 1

Nook Color

2012

Portico ingests initial d-collection content

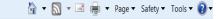
iPad 3



Usability: Anticipated usage ...



EPUB Content Documents 3.0



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

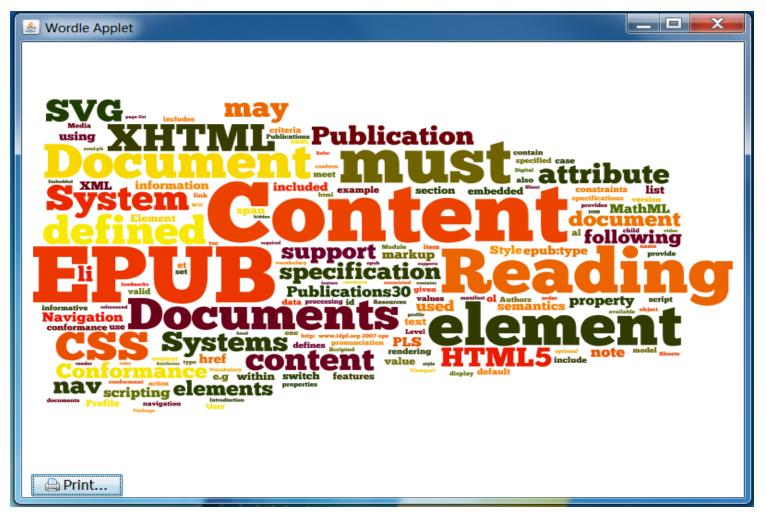
http://www.idpf.org/epub/30/spec/epub30-contentdocs-20110908.html

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™







Authenticity, Discoverability: Preservation Context



Preservation and Packaging Metadata File











- 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











- 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.

...











- 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.

...

images.
The TIF file named XYZ is page 1. It is a valid TIF and has a checksum of 123456.

These TIF files are page

The TIF file named ABC is page 2. It is not a valid TIF and has a checksum of 78910.

memorial control of the control of t

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.

Preservation and Packaging Metadata File







- 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.

worked by the control of the control

• 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.







- 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.

MARC 21

...



- 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.

Us cauge the control of the control

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.

...



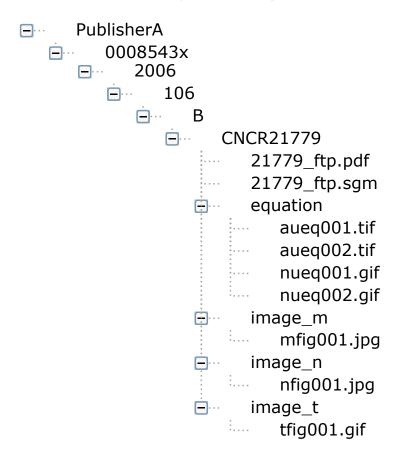
- ArchivalUnit "ark:/27927/pbb45qh1t"
 - ProcessingRecord "ark:/27927/pbb5rnpng" 2010-07-15T13:45:42.553-04:00
 - Timestamp 2010-07-15T13:45:42.553-04:00
 - Rationale Processing of New Content
 - SuppliedFiles "ark:/27927/pzb5nbsf8"
 - System "ConPrep"
 - Workflow "E-Journal Workflow 2.0"
 - Batch "ark:/27927/pgf18mnt35"
 - 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/" Polyatx
 - 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/" Sp
 - 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/pbb45gh29"
 - 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/pbb5rjgsg"
 - File "pbb5rjgsq.pmd"



Formats: Packages

Incoming File System

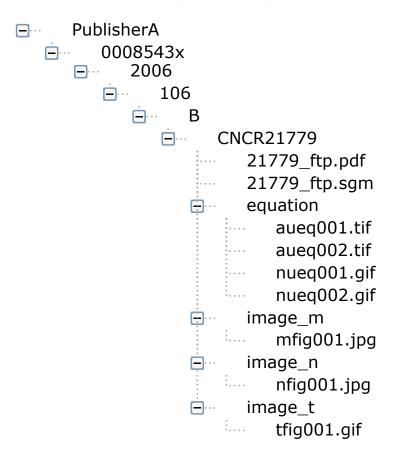




Formats: Packages



Incoming File System



Resulting Content Model

```
Content Unit (Article)
         Text: Marked Up Text
            21779_ftp.sgm
         Rendition: Page Images
            21779_ftp.pdf
         Component: Formula Graphic
            aueq001.tif
            nueq001.gif
         Component: Formula Graphic
            aueq002.tif
            nueq002.qif
         Component: Figure Graphic
            mfig001.jpg
            nfig001.jpg
            tfig001.gif
```

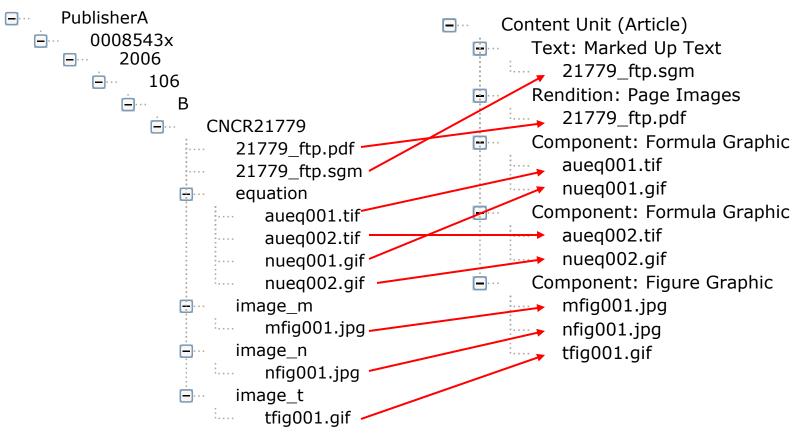


Formats: Packages



Incoming File System

Resulting Content Model





E-Book Packages in Portico Submissions



Flat directory

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



E-Book Packages in Portico Submissions



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



E-Book Packages in Portico Submissions



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







Formats: Text Content

```
BT
/H2 <</MCID 0 >>BDC
/CS0 cs 0.31 0.506
0.741 scn
/TTO 1 Tf
-0.004 Tc 0.006 Tw
12.96 0 0 12.96 72
697.68 Tm
[(H)-4(e)-1(1)-1(1)-
11(o,)-3(W)-15(or)-
6(1)-11(d!)-12(!)]TJ
0 Tc 0 Tw 6.481 0 Td
( )Tj
EMC
ET
```



Formats: Text Content

```
<html>
<head>
<style type="text/css">
<1--
  p { color: #4F81BD;
font-family: serif;
font-weight: bold;
font-size: 13pt; }
  -->
</style>
</head>
<body>Hello,
World!!</body>
</html>
```



Trade-offs: Expressiveness vs. Simplicity

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 × 11 inches (US letter-size paper; 21.6 cm × 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.

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

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

Style	Tag	Table:	Table	head

Style Tag Table: Table head	
1. Title	18 pt Times New Roman
	Bold, flush left; line spacing
0.00	exactly at 22 pt; 11 pt. after
2. Byline:	8 pt. Arial Bold Italic, flush
Author/Affiliation	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;
0.21 iguite caption	line spacing at exactly 11 pt;
	inic spacing at exactly 11 pt,





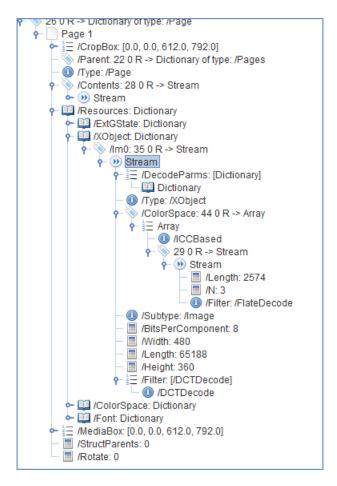




```
BT
/H2 <</MCID 0 >>BDC
/CS0 cs 0.31 0.506 0.741 scn
/TT0 1 Tf
-0.004 Tc 0.006 Tw 12.96 0 0 12.96 264
697.68 Tm
[(H)-4(e)-1(1)-2(1)-11(o,)-3(W)-15(or)-
6(1)-11(d!)-12(!)]TJ
0 Tc 0 Tw 6.481 0 Td
( )Tj
EMC
/P <</MCID 1 >>BDC
/CS1 cs 0 scn
/TT1 1 Tf
11.04 0 0 11.04 72 682.08 Tm
( )Tj
EMC
/P <</MCID 2 >>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
EMC
```







Hello, World!!



(iText RUPS)



```
<html>
<head>
<style type="text/css">
<1--
  p { color: #4F81BD;
font-family: serif;
font-weight: bold;
font-size: 13pt; }-->
</style>
</head>
<body>Hello, World!!
<br/><span><IMG
width="447"
height="336"
src="images/Image 001.j
pg"/></span></body>
</html>
```









```
mydir/
myFile.pdf
```

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



E-book formats in Portico Submissions



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



Looking ahead: EPUB 3



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."



Looking ahead: EPUB 3



EPUB 3

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



EPUB3 Packaging

```
🗸 accessible epub 3-20120425.epub
🗸 📗 EPUB
  ▶ III fonts
  images
  ▶ 🔐 lexicon
     bk01-toc.xhtml
     ch01.xhtml
     ch01s02.xhtml
     ch02.xhtml
     ch02s02.xhtml
     ch02s03.xhtml
     ch03.xhtml
     ch03s02.xhtml
     oh03s03.xhtml
     ch03s04.xhtml
     ch03s05.xhtml
     ch03s06.xhtml
     ch04.xhtml
     💀 co01.xhtml
     cover.xhtml
     index.xhtml
     package.opf
     pr01.xhtml
     pr01s02.xhtml
     pr01s03.xhtml
     pr01s04.xhtml
     pr01s05.xhtml
     spi-ad.xhtml

    toc.ncx

■ META-INF

     container.xml
  mimetype
```

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



EPUB3 Formats



"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



Complications: The New "Browser Wars"



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"



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





Questions or Comments?

Sheila Morrissey
sheila.morrissey@ithaka.org
@sheilaMorr
www.portico.org

