

Join 15 minutes early.

Close everything that might ding: yammer, AIM, rescue time, Outlook, etc.  
Undock the Chat and Questions and move them over to the left hand screen.  
Mute my phone and turn cell to alarm and BB off.

NISO starts the broadcast. Until then, the users will see the Go To Webinar entrance screen and hear silence.

The meeting will be recorded.

--

Thank you Todd and good afternoon everyone.

I'm excited by this opportunity to discuss digital preservation and standards, and specifications and best practices. I added in the parentheticals when I was compiling the list of "standards" I wanted to discuss and realized that not all of them are actually standards.

Questions ... happy to take questions. Tricky to keep an eye on both. If I miss one while going through, I'll catch it at the end. Also happy to take additional questions by email or phone after the fact.



**Amy Kirchhoff**  
**Archive Service Product Manager**  
**Portico, JSTOR**


[amy.kirchhoff@ithaka.org](mailto:amy.kirchhoff@ithaka.org)  
609-986-2218



Archive Service product manager for Portico and JSTOR as a whole.

I like to describe what I do as problem solving ... from relatively high level archival policy setting to help figuring out what to do with particular files (or rather, teaching folks how to think about what to do with particular files).


# Portico - Third Party Preservation



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

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

My obligatory couple of slides on Portico.

Portico provides preservation services to the academic community, both libraries and publishers.

Portico – Preserved Content

ITHAKA  
JSTOR | PORTICO | ITHAKA S+R



PORTICO

Preserved Content

» E-journal titles	9,190
» E-book titles	12,733
» D-collections	12
» E-journal files	223,993,405
» E-book files	869,888
» D-collection files	83,178,138
» Total Archive	308,729,560



PORTICO

We currently have three preservation services : the ej-ournal preservation service, the e-book preservation service, nad the de-collection presrevation service.



I can't accept credit for this slide, I borrowed it from a presentation on standards that Evan Owens our former CTO gave a couple of years back.

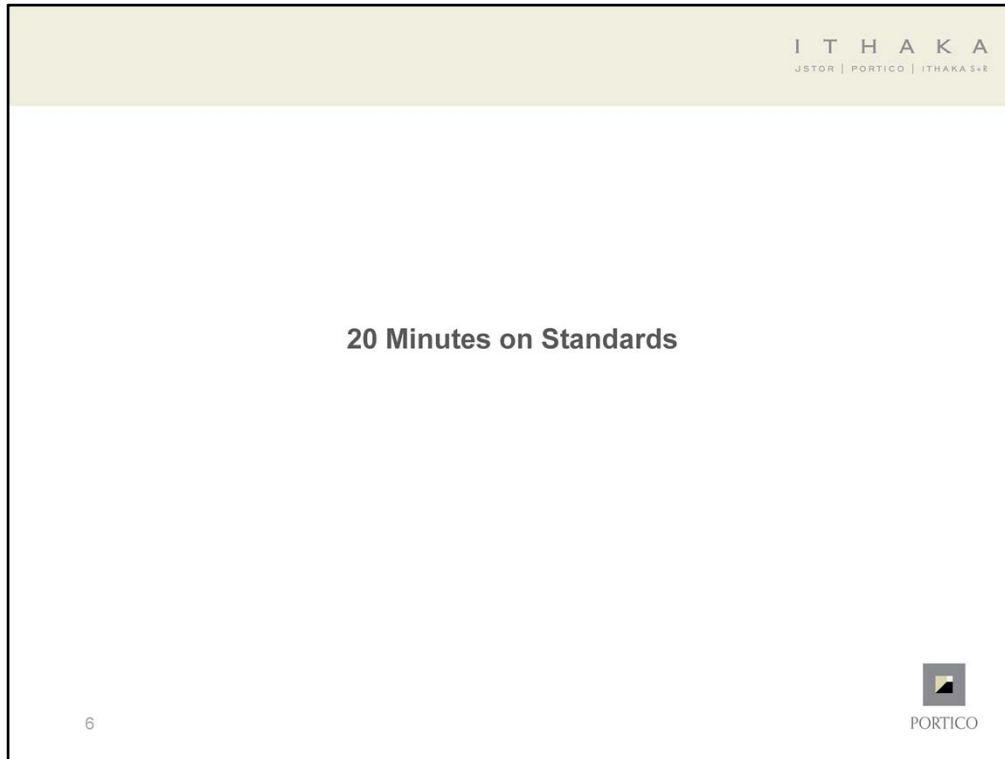
Standards are important. Standards let me plug my laptop, ipad, and cell phone into the wall.

However, they don't let me plug my laptop, ipad and cell phone into the wall in china.

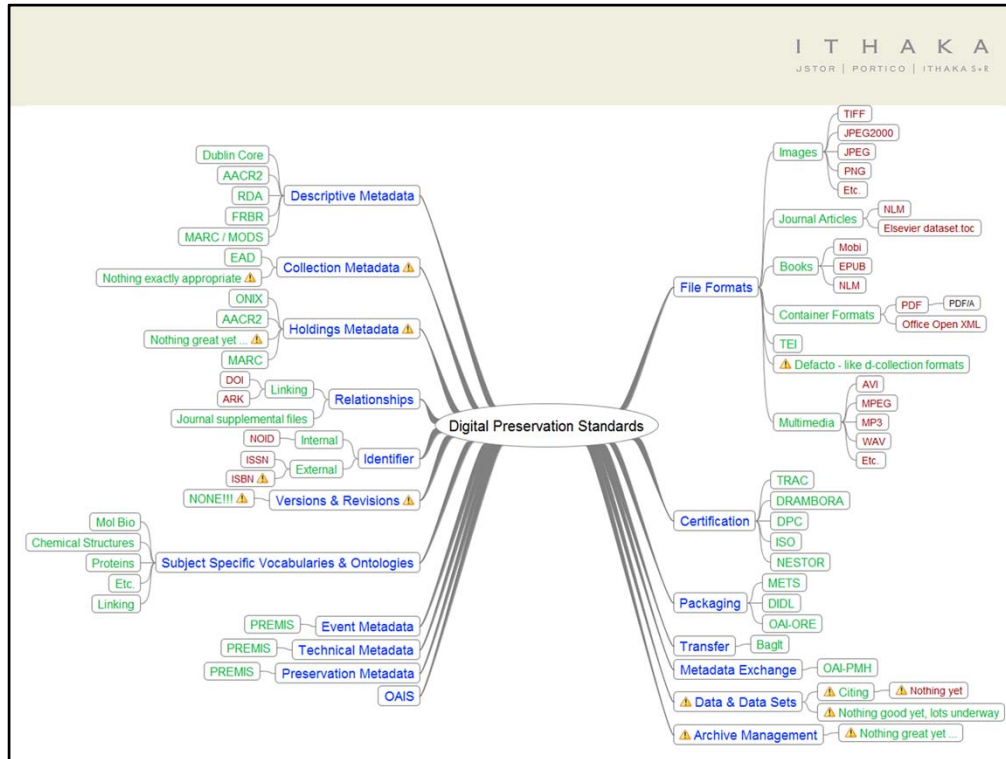
A coulpe of interesting things to point out ... <click>

These two ar e the same standard. The one on the top left has been installed upside down (standards can be misintprpreted). The one on the bottom center comes with a feature that provides for 360 degree rotating outlets. Sadly, the one on the bottom left looks like might be built to the standard of the two that are circled, but it is not.

The moral of this slide ... bring an adapter when you travel or insure you can import and export content to standards, even if what is in between the import and export does not match anything.



So, when Nettie first asked me to do this I thought ... 20 minutes on standards ... can I talk for 20 minutes on standards?



So, I started brainstorming about standards that impact the work we do in digital preservation...

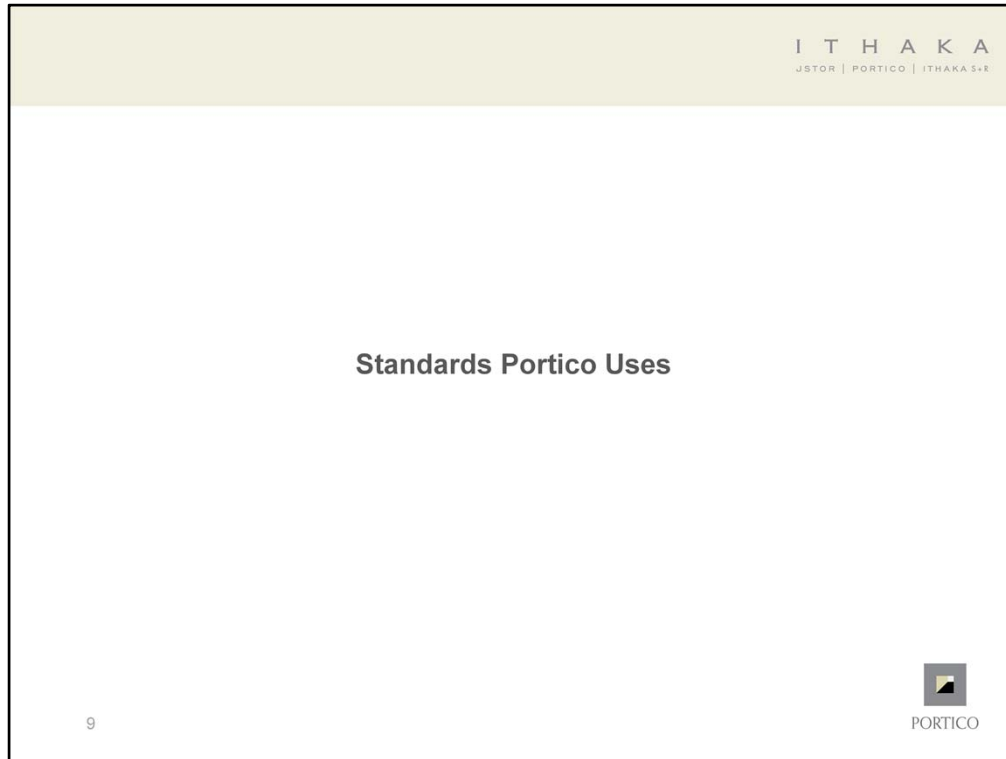
And then I thought, 20 minutes! I could teach a class an hour a week for a semester on standards in digital preservation ...



I considered using this as a platform to advocate for some missing standards, or specifications, or even best practices.

And we will come back to that at the end.





Instead I settled on a discussion of some of the standards Portico uses in its digital preservation work (you'll find out a bit later that I'm a bit free with the word "use" here -- adapters, remember adapters 😊).

## Context: Digital Preservation

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

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

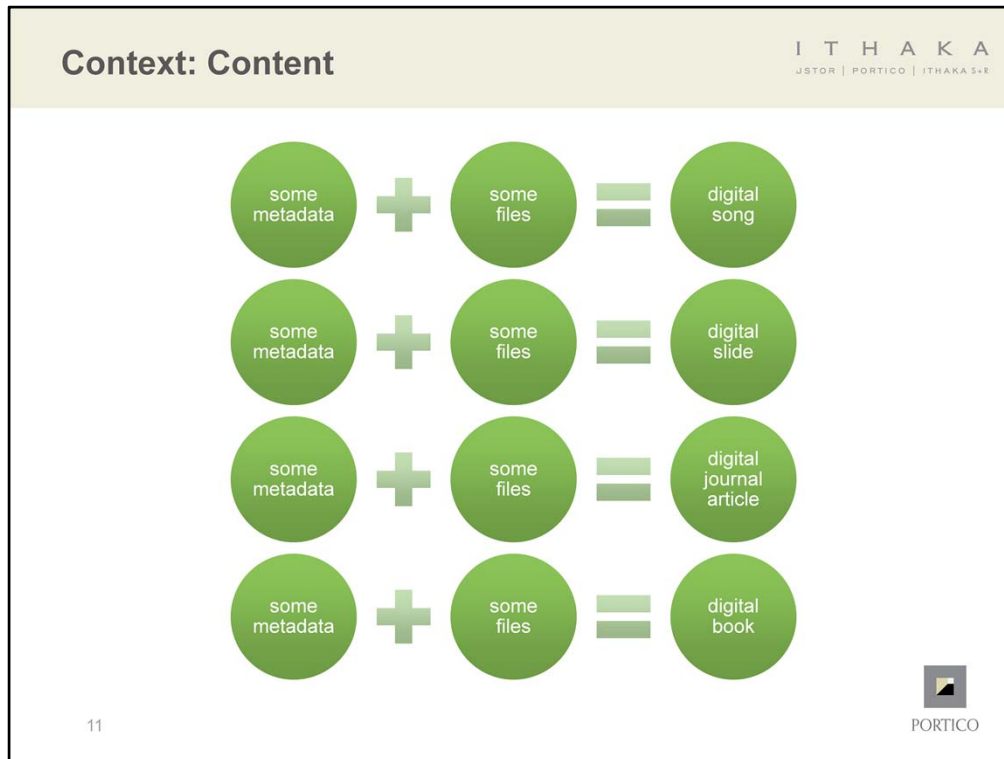
- the content must have logical bibliographic metadata so that it can be found by end users through time

### Accessibility

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

To give some context to digital preservation ....

This is the definition Portico uses, it lives on my white board and I actually teach classes on it to Portico, JSTOR and other ITHAKA staff.



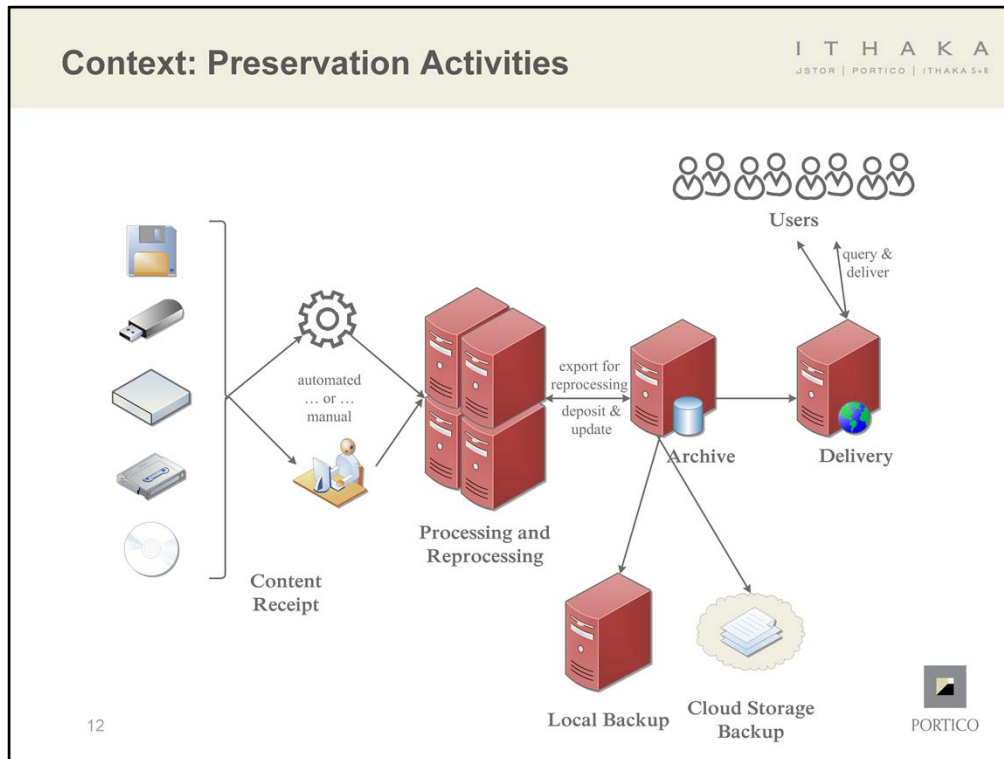
Content that is not a “database product” (like Google maps or dynamic e-reference books), content that has a clear beginning point and ending point ... from a content management and preservation perspective that content all looks very similar.

I need to preserve a digital song, well I have some files that make up said song, and I have some metadata about the song, the files, the context, the rights, etc. Together that metadata and those files must be preserved to preserve the unit that is the song.

Same is true for journal articles ... files ... metadata ... article for preservation.

Makes up the unit of preservation.

Important later on because there are standards around the metadata and around how these units are packaged that are key elements of digital preservation.



Relatively standard flow to preservation activities and we'll see a standard about this shortly ...

Content come in, content is processed and ingested, content is preserved, content is perhaps reprocessed, content is exported and delivered.

Standards around receipt

Standards around metadata to add/subtract/modify during ingest

Standards around files during ingest

Standards around archive management



So this takes us to standards & specifications

We are actually going to fly pretty fast here and just get an overview.

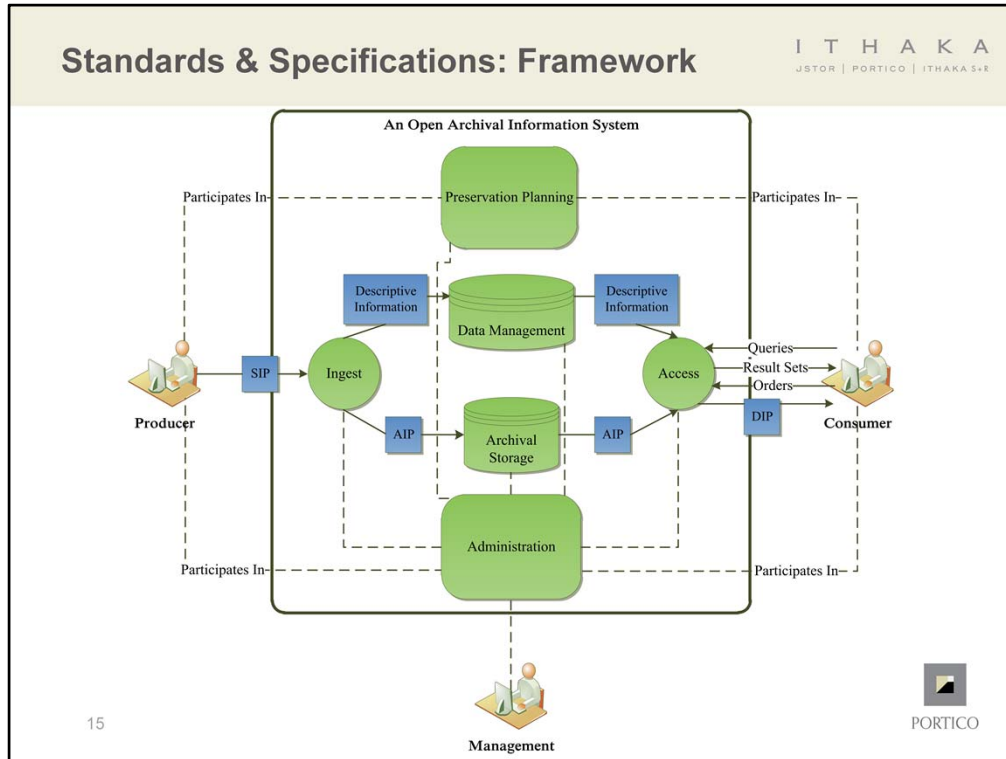
If anyone wants a cheat sheet of some standards, feel free to give me a ping and I can pull something together.

When I was considering standards, this is some of the characterization I came up with...

Talk about each, eh, maybe not ....



We are going to start with OAIS, as every discussion of digital preservation does ...



- OAIS is a standard. It describes, at a very high level, what a digital preservation service should look like.
- Similar flow to what we saw a couple of slides ago
- It's a big document ... my take on the 7 highlights ...
  1. there must be preservation planning
  2. the archive requires ongoing administration
  3. content in the form of files and metadata about the files must come into the archive
  4. content in the form of files and metadata about the files must go out of the archive
  5. the archive is not the hardware or software—rather the hardware and software are elements of the archive, which includes activities performed by people
  6. the original producer of the content and the eventual users of the content must have input into the ongoing management of the archive
  7. at its very base, the preserved information consists of the content to be archived and metadata about the content—both of which are required for long-term preservation.



Wide World of standards, I'm going to stick with the ones we use or have seriously considered ...



ITHAKA  
JSTOR | PORTICO | ITHAKA S+R

PORTICO



Every one preserving content will need to move that content at some point ... enter standards about transferring content

<click>

# Standards & Specifications: Transfer

ITHAKA

JSTOR | PORTICO | ITHAKA S+R

[\[Docs\]](#)
[\[txt\]](#)
[\[pdf\]](#)
[\[Tracker\]](#)
[\[Email\]](#)
[\[Diff1\]](#)
[\[Diff2\]](#)
[\[Nits\]](#)

Versions: 00 01 02 03 04 05 06

Network Working Group

Internet-Draft

Expires: October 17, 2011

A. Boyko

J. Kunze

California Digital Library

J. Littman

L. Madden

Library of Congress

B. Vargas

April 18, 2011

The BagIt File Packaging Format (V0.97)

<http://www.ietf.org/internet-drafts/draft-kunze-bagit-06.txt>

Abstract

This document specifies BagIt, a hierarchical file packaging format for storage and transfer of arbitrary digital content. A "bag" has just enough structure to enclose descriptive "tags" and a "payload" but does not require knowledge of the payload's internal semantics. This BagIt format should be suitable for disk-based or network-based storage and transfer.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of [RFC 78](#) and [RFC 79](#).

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on October 17, 2011.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to [RFC 78](#) and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of

19

Boyko, et al.

Expires October 17, 2011

[Page 1]

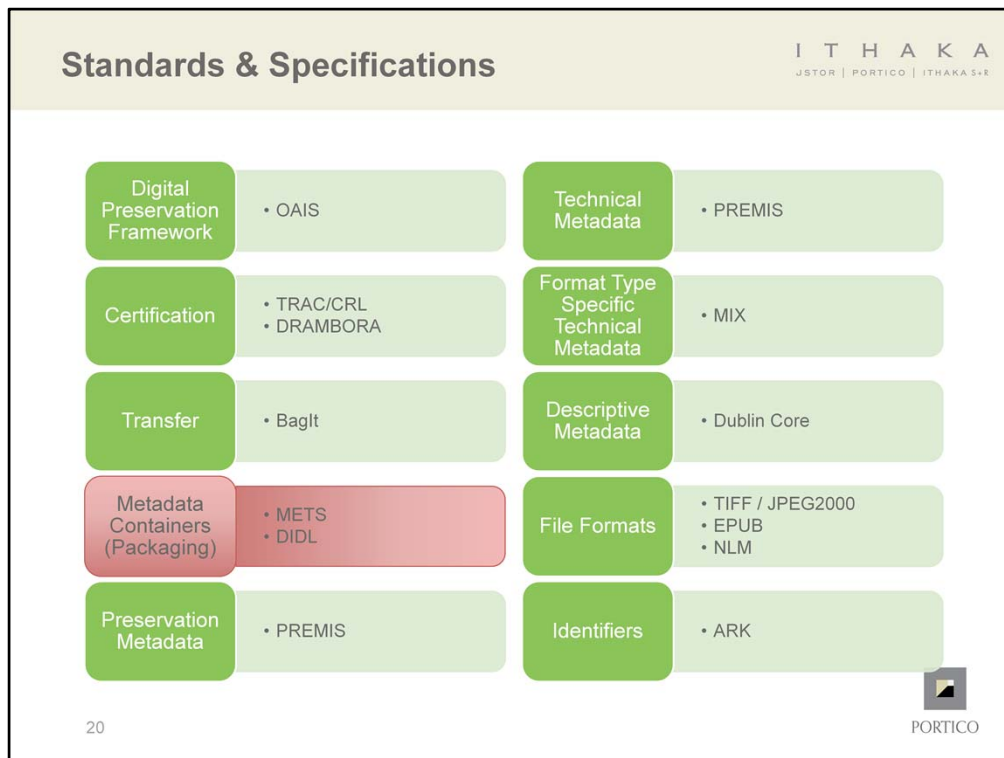


PORTICO

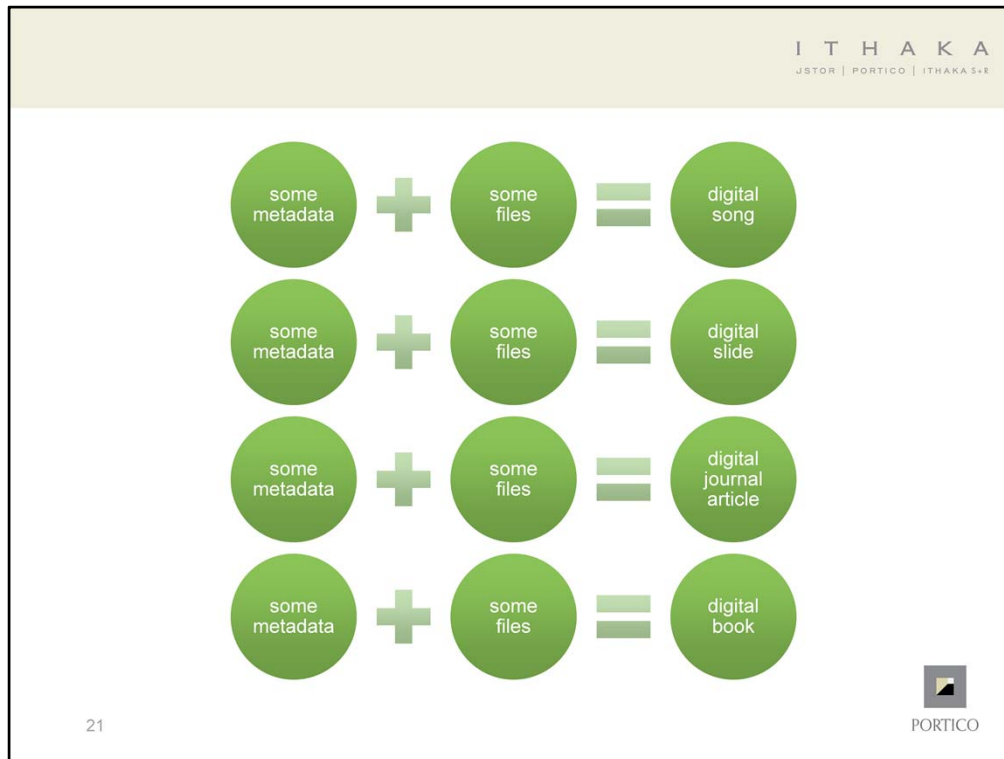
CDL, Stanford & LC

It describes a structure for how to wrap disparate content up into a "bag" that another agency can open. The key element is the required manifest file listing all content in the bag and the checksums of that content.

19

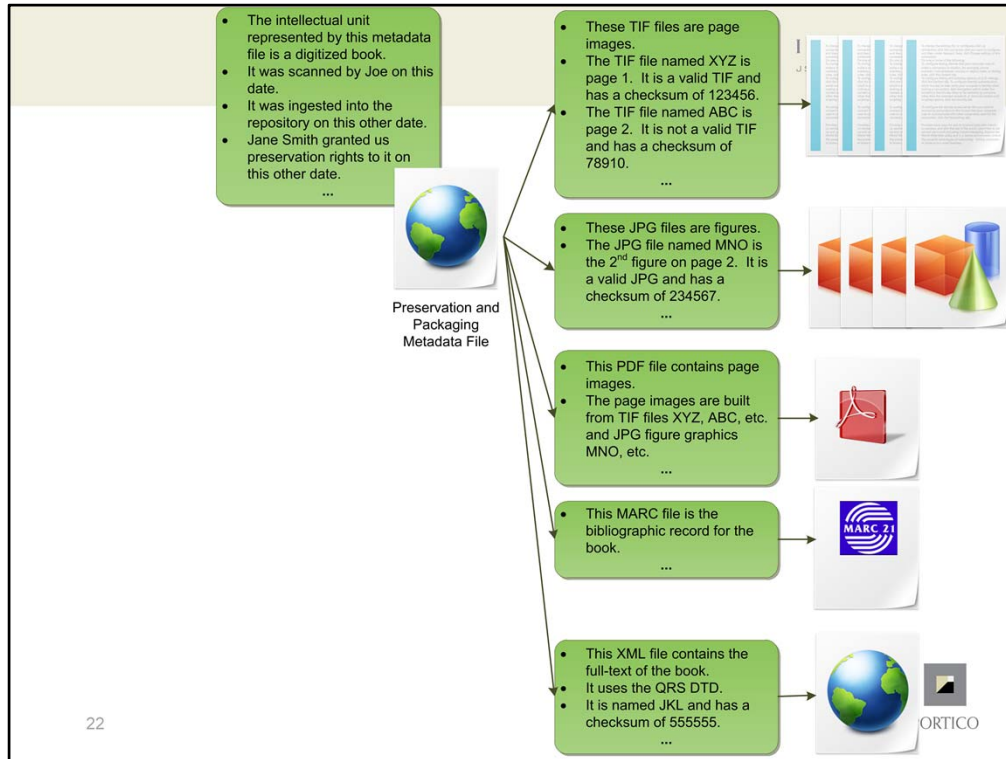


This is actually the spot where I'm going to chat at you the longest ...



If we come back to our content slide ...

Somehow, we need a structure into which to put the metadata and files for each of those units and this is where packaging comes in.



As we said <click back> every unit has files and metadata ..

<click forward>

The packgaing lets us wrap all of that up and say things like ...

- <read first>
- <read next>
- So on and so forth.

**Standards & Specifications: Packaging**
I T H A K A  
JSTOR | PORTICO | ITHAKA S+R

The Library of Congress >> Standards

## Metadata Encoding & Transmission Standard

Official Web Site

Home
METS Pages ▾
search

The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the XML schema language of the World Wide Web Consortium. The standard is maintained in the Network Development and MARC Standards Office of the Library of Congress, and is being developed as an initiative of the Digital Library Federation.

## Open Archives Initiative Object Reuse and Exchange

Home
Projects
Specifications
Community
About OAI

Open Archives Initiative -> ORE

### MPEG-21 Part 2: Digital Item Declaration Language (DIDL)

DIDL Overview

"The basic architectural concept in MPEG-21 is the Digital Item. Digital Items are structured digital objects, including a standard representation, identification and metadata. They are the basic unit of transaction in the MPEG-21 framework. More concretely, a Digital Item is a combination of resources (such as videos, audio tracks, images, etc), metadata (such as descriptors, identifiers, etc), and structure (describing the relationships between resources).

This second part of MPEG-21 (ISO/IEC 21000-2:2003) specifies a uniform and flexible abstraction and interoperable schema for declaring the structure and makeup of Digital Items. Digital Items are declared using the Digital Item Declaration Language (DIDL) and declaring a Digital Item involves specifying its resources, metadata and their interrelationships.

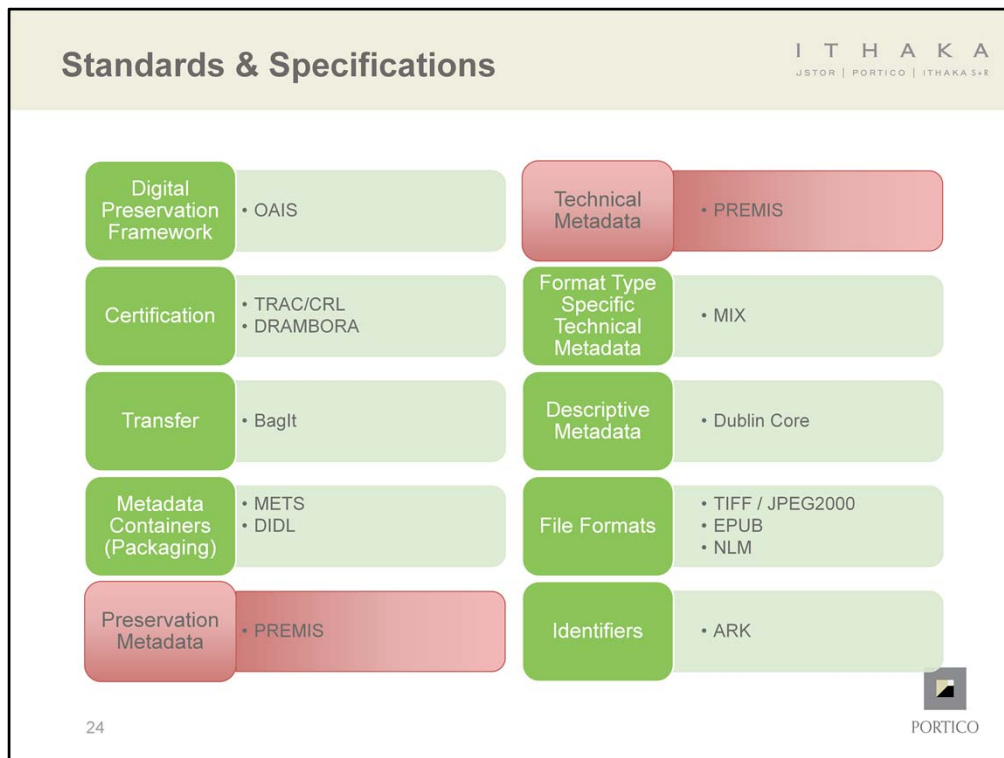
23

A couple of standards in use ...

METS and DIDL are big player ... METS more than DIDL ..

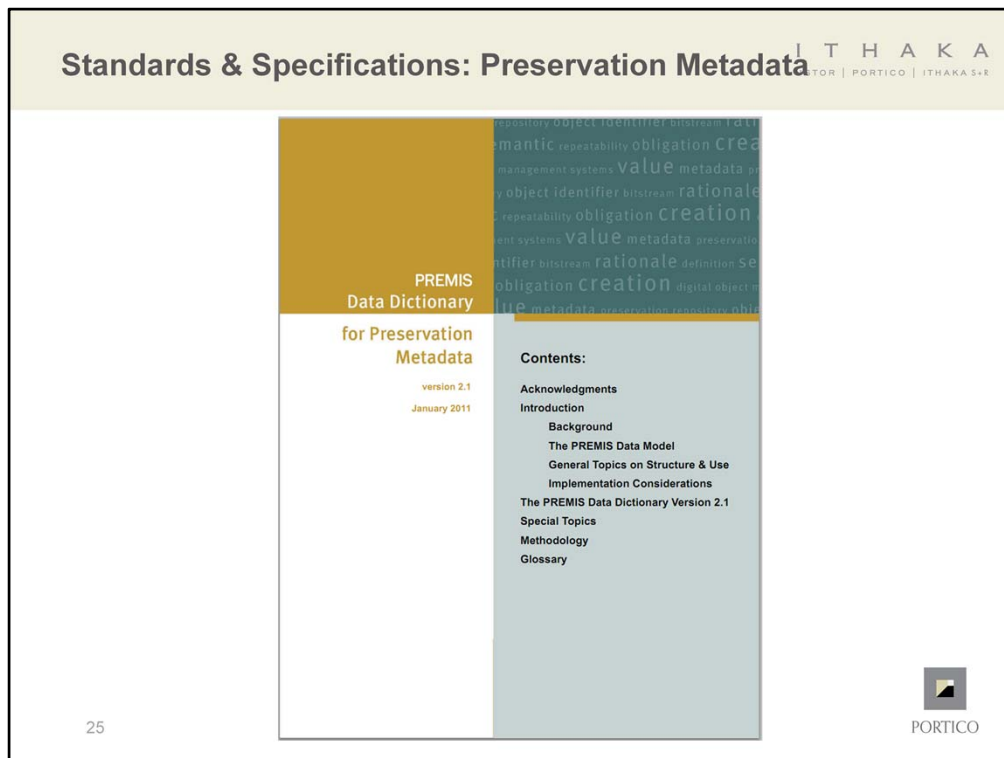
I must confess that at Portico we use neither and in fact, use a self-developed packaging scheme that is very closely tied to our internal content model. We moved to it, PMD – preservation metadata, last year. It was heavily influenced by METS, DIDL, PREMIS and our own 5 years of experience.

I should have thrown the slide of outlets back in here, or maybe an adapter. At Portico we went this route of our own packaging scheme with the policy that we must be able to export to METS or DIDL – so long as we can import and export to the standards, our internal scheme can be different.



Next is preservation metadata of which there is one big player ... <click>





PREMIS ...

It is a data dictionary and provides the semantics necessary to define preservation metadata for preserved content, including things like technical and events metadata, preservation level, and other significant properties of a preserved unit of content.



Another area of standards & specifications is format or content type specific technical metadata <click>

The Library of Congress >> Standards >> MIX

**MIX** NISO Metadata for Images in XML Schema  
Technical Metadata for Digital Still Images Standard  
Official Web Site

The Library of Congress' Network Development and MARC Standards Office, in partnership with the NISO Technical Metadata for Digital Still Images Standards Committee and other interested experts, is developing **MIX Version 2.0** is now the current version of MIX. The schema provides a format for interchange and/or storage of the data specified in the [Data Dictionary - Technical Metadata for Digital Still Images \(ANSI/NISO Z39.87-2006\)](#). This schema is currently referred to as 'NISO Metadata for Images in XML (NISO MIX)'. MIX is expressed using the [XML schema language of the World Wide Web Consortium](#). MIX is maintained for NISO by the [Network Development and MARC Standards Office](#) of the Library of Congress with input from users.

**MIX Schema & Documentation**

- [MIX Schema Version 2.0 \(current version\)](#)
- [MIX Schema Version 1.0 \(previous version\)](#)
- [Data Dictionary - Technical Metadata for Digital Still Images \(ANSI/NISO Z39.87-2006\)](#) [used for MIX version 1.0]
- [MIX Schema Version 0.2 \(previous version\)](#)

**MIX Example Documents**

- Example MIX XML Documents
  - [Test MIX document instance \(version 1.0\)](#)
  - [Test MIX document instance \(version 0.2\)](#)

**Tools & Utilities**

- [JHOVE - JSTOR/Harvard Object Validation Environment](#)



My example here is MIX, the NISO metadata for images in XML schema, because it is one used extensively by Portico.

There are many other examples for different file formats and even different subject areas (for example, there are specific standards for geospatial data, etc.)



And, then, there are standards around the bibliographic metadata describing a preserved unit ...

Some content will have its own, very robust and extensive descriptive metadata (at Portico we have metadata marked up in XML files for each item and that metadata is either in NLM or a proprietary DTD), however at the level of preservation metadata, there needs to be some descriptive or bibliographic metadata that can be use to informly retrieve and find content in the archive

... and for that there is ...  
<click>

## Standards & Specifications: Descriptive MD

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

The screenshot shows the Dublin Core Metadata Initiative website. The header includes the logo and navigation links: Home, Metadata Basics, DCMI Specifications, Community and Events, and About Us. Below the header is a search bar with the text "Enter keyword" and a "Search" button. The main content area is titled "Dublin Core Metadata Element Set, Version 1.1". It lists the Identifier, Replaces, Latest version, Date issued, Status, and Description of the document. The Description states: "This document provides ready reference for the Dublin Core Metadata Element Set, Version 1.1. For more detailed document: documentation and links to historical versioning information, see the document 'DCMI Metadata Terms'." Below this is an "Introduction" section. The Introduction explains that the Dublin Core Metadata Element Set is a vocabulary of fifteen properties for use in resource description. It mentions that the name "Dublin" is due to its origin at a 1995 institutional workshop in Dublin, Ohio, and "Core" because its elements are broad and generic. It also states that the fifteen element "Dublin Core" described in this standard is part of a larger set of metadata vocabularies and technical specifications maintained by the Dublin Core Metadata Initiative (DCMI). The full set of vocabularies, DCMI Metadata Terms (DCMI-TERMS), also includes sets of resource classes (including the DCMI Type Vocabulary (DCMI-TYPE), vocabulary encoding schemes, and syntax encoding schemes). The terms in DCMI vocabularies are intended to be used in combination with terms from other, compatible vocabularies in the context of application profiles and on the basis of the DCMI abstract model (DCMI-AM).

All changes made to terms of the Dublin Core Metadata Element Set since 2001 have been reviewed by a DCMI Usage Board in the context of a DCMI Namespace Policy (DCMI-NAMESPACE). The namespace policy describes how DCMI terms are assigned Uniform Resource Identifiers (URIs) and sets limits on the range of editorial changes that may ultimately be made to the labels, abbreviations, and usage comments associated with existing DCMI terms.

This document, an excerpt from the more comprehensive document "DCMI Metadata Terms" (DCMI-TERMS) provides an abbreviated reference version of the fifteen element descriptions that have been formally endorsed in the following standards:

- ISO Standard 15926-2009 of February 2009 [ISO15926]
- ANSI/NISO Standard Z39.85-2007 of May 2007 [ANSI/NISOZ39.85]
- IETF RFC 5013 of August 2007 [RFC5013]

Since 1995, when these fifteen elements entered into a standardization track, notions of best practice in the Semantic Web have evolved to include the assignment of formal domains and ranges in addition to definitions in natural language. Domains and ranges specify what kind of described resources and value resources are associated with a given property. Domains and ranges express the meaning implicit in natural-language definitions in an explicit form that is usable for the automatic processing of logical inferences. When a given property is encountered, an inferring application may use information about the domains and ranges assigned to a property in order to make inferences about the resources described thereby.

Since January 2008, therefore, DCMI includes formal domains and ranges in the definitions of its properties. So as not to affect the conformance of existing implementations of "Simple Dublin Core" in RDF, domains and ranges have not been specified for the fifteen properties of the dc: namespace (<http://purl.org/dc/terms/>). Rather, fifteen new properties with "name" identical to those of the Dublin Core Metadata Element Set Version 1.1 have been created in the dc: namespace (<http://purl.org/dc/terms/>). These fifteen new properties have been defined as subproperties of the corresponding properties of DCMI Version 1.1 and assigned domains and ranges as specified in the more comprehensive document "DCMI Metadata Terms" (DCMI-TERMS).

Implementers may freely choose to use these fifteen properties either in their legacy dc: variant (e.g., <http://purl.org/dc/elements/1.1/creator>) or in the dc: namespace variant (e.g., <http://purl.org/dc/terms/creator>) depending on application requirements. The RDF schema of the DCMI namespaces describe the subproperty relation of dc:terms:creator to dc:creator for use by Semantic Web-aware applications. Over time, however, implementers are encouraged to use the semantically more precise dc:terms properties, as they more fully follow emerging notions of best practice for machine-processable metadata.

- » Contributor
- » Coverage
- » Creator
- » Date
- » Description
- » Format
- » Identifier
- » Language
- » Publisher
- » Relation
- » Rights
- » Source
- » Subject
- » Title
- » Type

29



Dublin Core -- It is very highlevel and very generic, but the lowest common denominator is what is needed when managing a diverse repository of content.

At Portico we use Dublin Core extensively and populate the metadata from the more robust, content-type specific metadata formats we manage.



Each file that is preserved is in a format and most of the formats will be associated with standards ... <click>

# Standards & Specifications: File Formats

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

NCBI

National Center for Biotechnology Information

National Library of Medicine

Archiving and Interchange Tag Set

Journal Publishing Tag Set

Article Authoring Tag Set

NCBI Book Tag Set

Introduction

Getting the Files

FTP Directory

Tag Suite Versions

Feedback

How to Build a Custom DTD

XML Information

Working Group and Secretariat

Acknowledgments

**Introduction**

The National Center for Biotechnology Information (NCBI) of the National Library of Medicine (NLM) has developed the National Journal Archiving and Interchange Tag Suite with the intent of providing a common set of XML elements and attributes for describing the textual and graphical content of journal articles. The Suite provides a set of XML elements and attributes for describing the textual and graphical content of journal articles. The Suite provides a set of XML elements and attributes for describing the textual and graphical content of journal articles. The Suite provides a set of XML elements and attributes for describing the textual and graphical content of journal articles.

**The Suite of Modules**

The intent of this Tag Suite is to preserve the intellectual content of journals in a format that can be exchanged between different systems. The Suite has been written as a set of modules, each of which is a separate physical file. No module is an entire schema by itself, but they can be combined into a number of different schemas.

**TIFF**

Revision 6.0

Final — June 3, 1992

**PDF Standards**

A Service of Adobe and the PDF Technology Community

Introduction | PDF Standards Defined | Conformance | PDF Reference | PDF/A | PDF/E | PDF/UA | PDF/VT | PDF/X | Other Work | A111 | Contacts

Log in

PDF/A - Action Items

PDF/A - Meetings

PDF/A - App Notes

PDF/A - Conference

PDF/A - Registration

PDF/A - Links

PDF/A - Embedded File Discussion

PDF/A Links

AIMS PDF/A page

NPES PDF/A page

The PDF/A Competence Center

Wikipedia PDF/A page

Guidance on the Use of PDF/A

US National Archives [FAQ regarding PDF/A](#)

US National Archives [transfer instructions regarding PDF/A](#)

US Federal Government [Guidance on Sustainable Formats](#)

US Library of Congress [Sustainability of Digital Formats](#) <http://www.diglibpreservation.gov>

All content is owned by ISO, AAMI and the Committee members. ISO Fund

Web Development & Maintenance donated by Duff Johnson | Web technology by redline.com

Content provided by David Berly Panning, [berlypanning@ncbi.nlm.nih.gov](mailto:berlypanning@ncbi.nlm.nih.gov) Site provided by [pdfcenter@adobe.com](mailto:pdfcenter@adobe.com)

**aims**  
First, Connect, and Optimize Your Information

The NLM DTD for e-journal XML files and e-book XML files is one Portico uses extensively.

As it the TIFF standard and PDF standards.

This area is very broad, however, so this is just a sampling.




And, we'll end with identifiers. <click>



# Standards & Specifications: Identifiers

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



University of California  
California Digital Library  
CDL NOID  
published by Portico (last edited by Portico Willott on Jan 25, 2010) (view change)

✚ Add • ⚙ Tools •

Name

NOID

Version

0.424 (2006-04-21)

Status

Beta

Specification

NOID

Download

<http://research.cdlib.org/~jpk/noid-0.424/>

More information

[Curation home page](#)

Have you ever noticed how some of the most "mission critical" identifiers in your daily life are numbers? How often do you use

- a driver's license number,
- a social security number, or
- a bank or credit card account number

instead of your name and address, or a photo of your honest, smiling face? We use numbers because they are short, precise, and opaque. Opaque identifiers, such as numbers or random combinations of letters, are useful as long-term descriptors for information objects because they don't contain information that is at risk of becoming untrue later.

**Why Opaque Identifiers**

Non-opaque descriptors represent object properties that change over time: subject classifiers, where an object "lives", the spelling of an author's name, etc. They can also be imprecise in large collections where a keyword or title search returns too many results. Moreover, unstable or imprecise identifiers, such as a web address that worked 6 months ago but not today, are a common complaint. So it is important to have precise, stable identifiers that don't include vague or changeable properties.


To help stability, an opaque identifier doesn't contain any information related to potentially changeable properties. For instance, if an identifier contains an organizational acronym and that organization is merged with another, there is often political pressure to break with the past, which means pressure not to support previously published identifiers in which the old acronym appears. Opaque identifiers also have the advantage that they can be short; for example, using combinations of letters and digits, only four characters are needed to represent as many as 1.6 million identifiers.

While opaque object identifiers have distinct advantages, they aren't always easy to use. They contain no widely recognizable words that allow people to guess what the object is, and are hard to repeat because a typo doesn't create an obviously misspelled word.

**Nicer Opaque Identifiers**

This is where NOID rhymes with "employe'd" comes in.

The NOID software tool mint (generates) opaque identifiers and tracks information to help them remain unique, stable, and closely connected to the objects that they identify. These identifiers should be opaque enough to age and travel well, but should easily resolve (connect you) to objects and to their descriptions.



University of California  
California Digital Library  
CDL ARK  
published by Portico (last edited by Portico Willott on Aug 20, 2010) (view change)

✚ Add • ⚙ Tools •

Name

ARK

Version

2008-05-22

Status

Beta

Specification

The ARK Identifier Scheme: PDF TXT

Additional Information

- [Towards Electronic Persistence Using ARK Identifiers \(July 2003\)](#)
- [Curation home page](#)

**Abstract**

An ARK is a URL, created to allow persistent, long-term access to information objects. ARKs can identify objects of any type: digital documents, databases, images, software, and websites, as well as physical objects (books, bones, statues, etc.) and even intangible objects (chemicals, diseases, vocabulary terms, performances).

ARKs support persistent identification, which is necessary and useful because both the protocols used to access objects (such as http and ftp) and the sites that host the objects are subject to change. An ARK contains parts that are impervious to such changes and parts that are flexible enough to support technological changes/improvements. The idea is to create a stable "name" or reference that can be permanently associated with that specific object.

**ARK Anatomy**

**NAANS: Name Assigning Authority Numbers**

**Generating ARKs**

**ARKs in Action**

**CDL Name Assignment and Support Policy Statements**

**Related Specifications**

- [NOID: Name-to-Thing](#)
- [NOID: \(Nice Opaque Identifier\) Minting and Binding Tool](#)

33

At Portico we extensively use the CDL Curation Micro-Services NOID software and ARK specification. We assign unique internal IDs, in the form of ARKs, to everything.

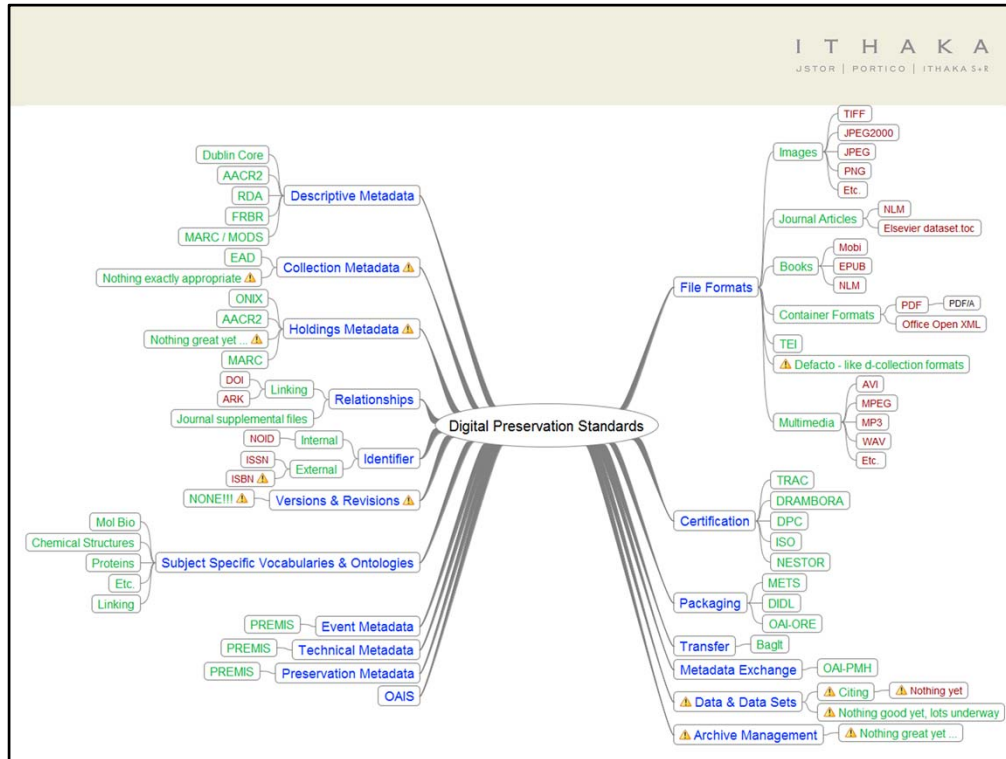
There is a whole 'nother realm related to standards around external identifiers, but we'll touch on that in a couple of slides.

33



... Being the savvy researcher that I am, when starting to sketch out this talk I Googled “digital preservation standards” and found this really cool page build by the folks at JISC. Many of the standards I talk about here and many others are listed there.

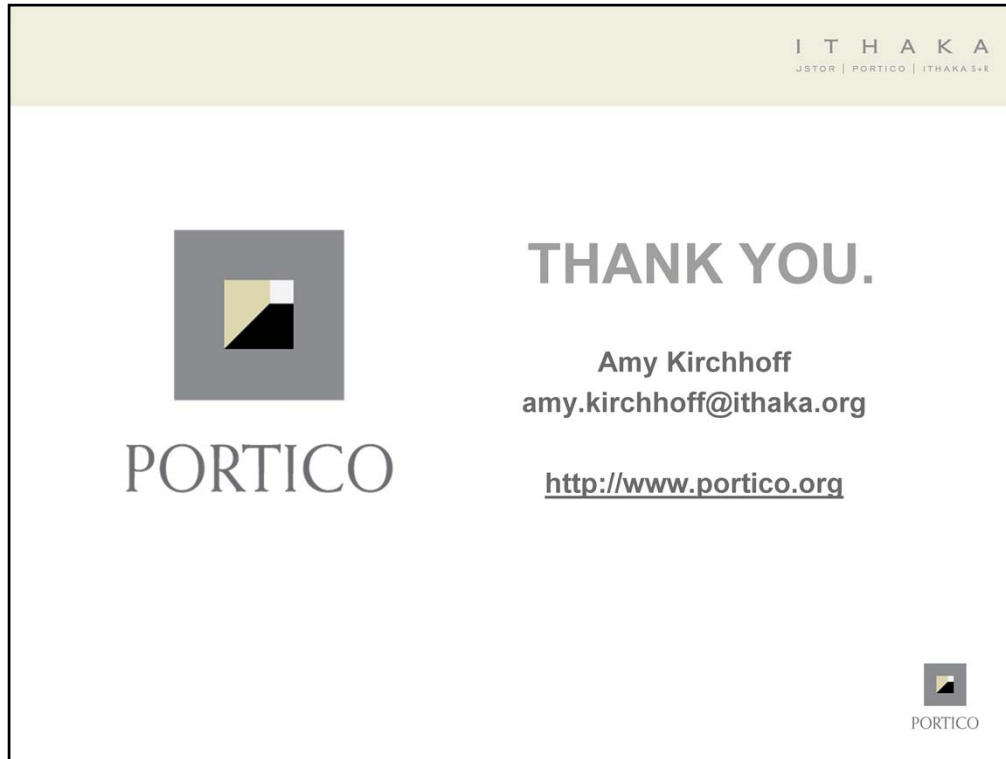
Again, if anyone wants a cheat sheet of standards we use at Portico, their purpose, and where those standards live, please just give me a ping.



So, that brings me back to my brainstorming mind map.

I wanted to pause here and point out all the little explanation marks. These are areas where standards, guidelines, and even just best practices are lacking, from what my research has shown.

There is lots more work to be done!



Stay on afterward.

No need to change presenter, Tom should be able to do that.