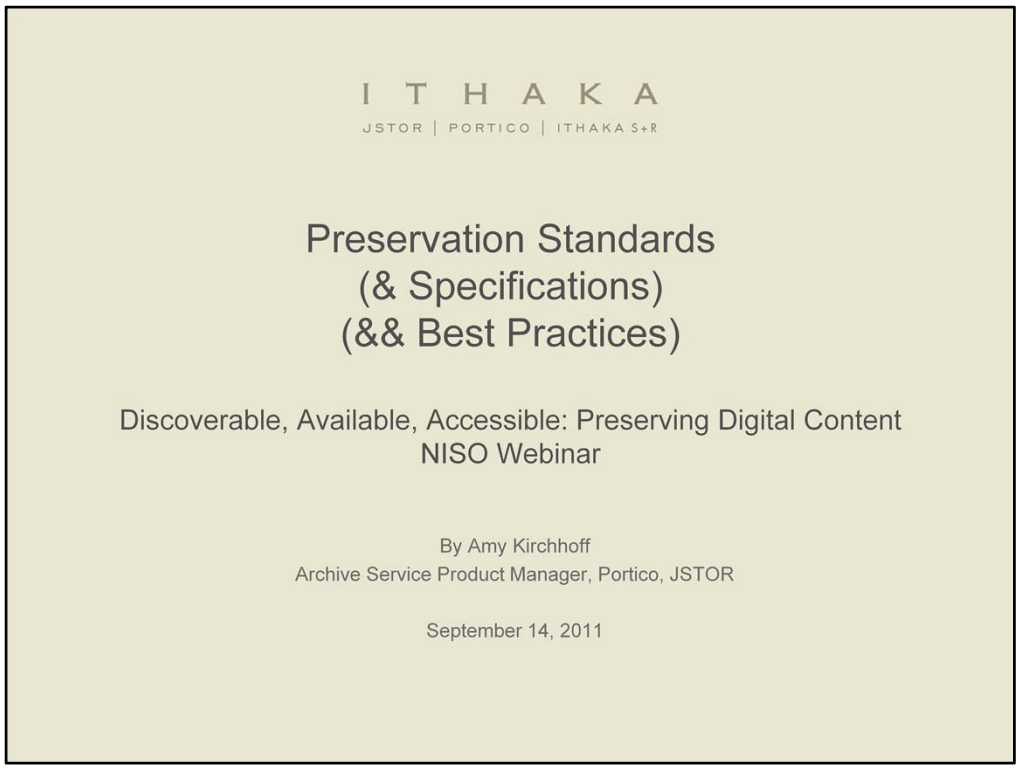


Join 15 minutes early.

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.

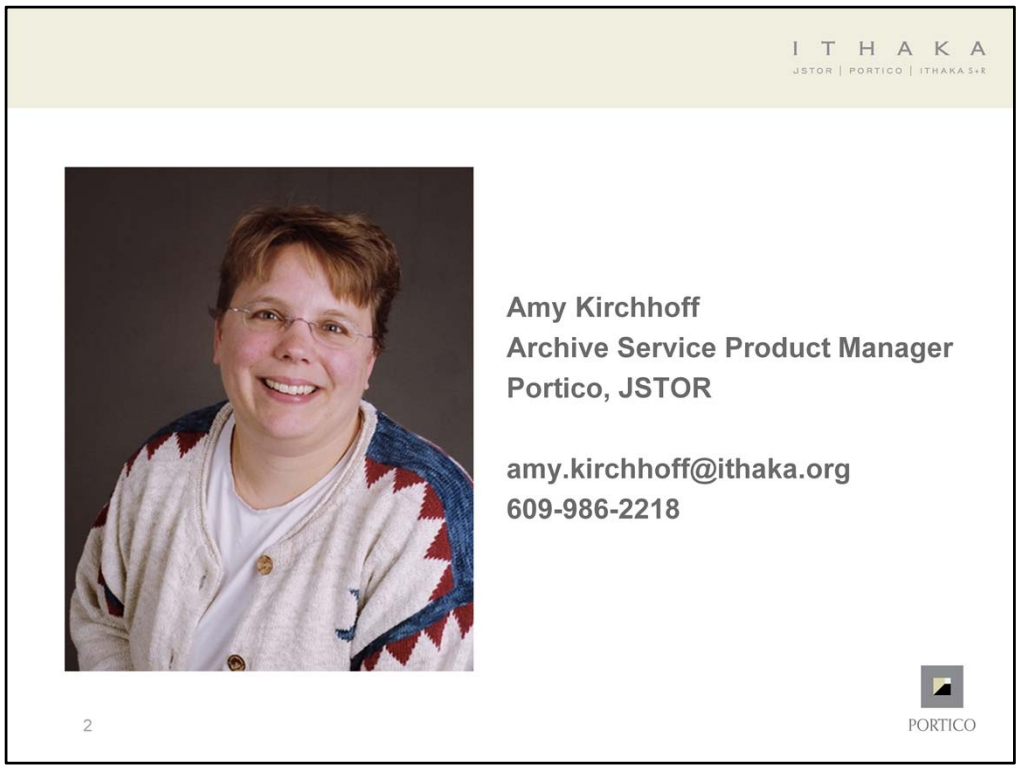
NiSO starts the broadcast. Until then, the users will see the Go To Webinar entrance screen and hear silence.

The meeting will be recorded.

--
Thank you Todd and good afternoon everyone.

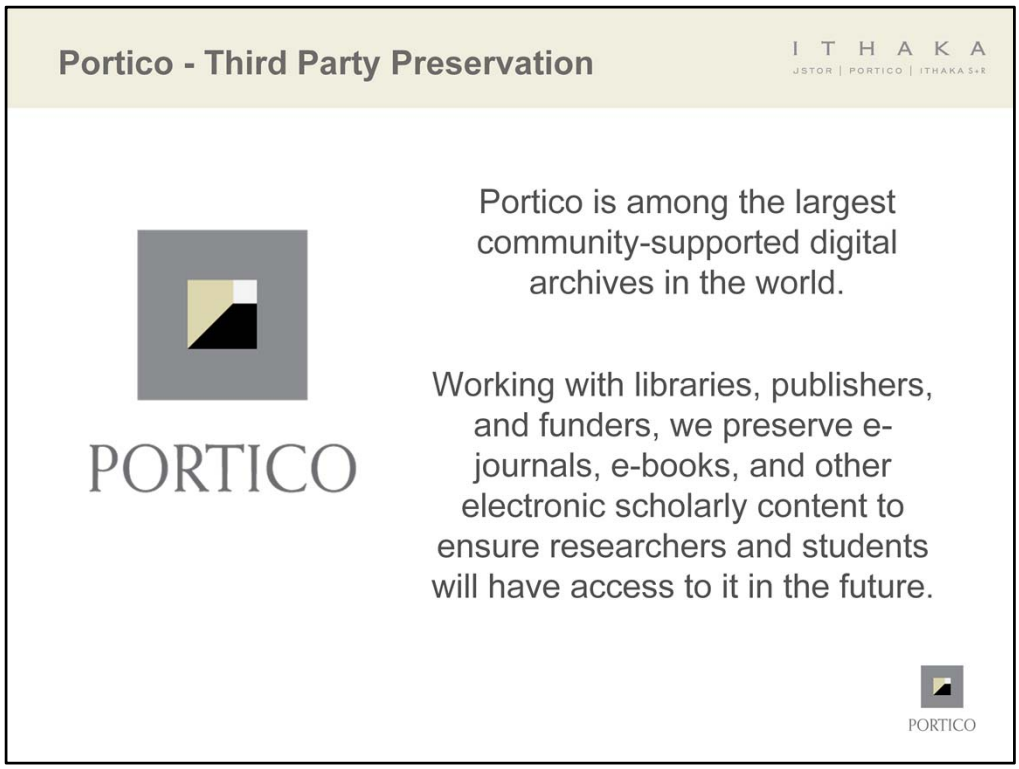
I'm excited by this opportunity to discuss digital preservation and standards, and specifications and best practices. I added in the parenthenticals when I was compling the list of "standards" I wanted to discuss and realized that not all of them are actually standards.

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.



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).



My obligatory couple of slides on Portico.

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

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.

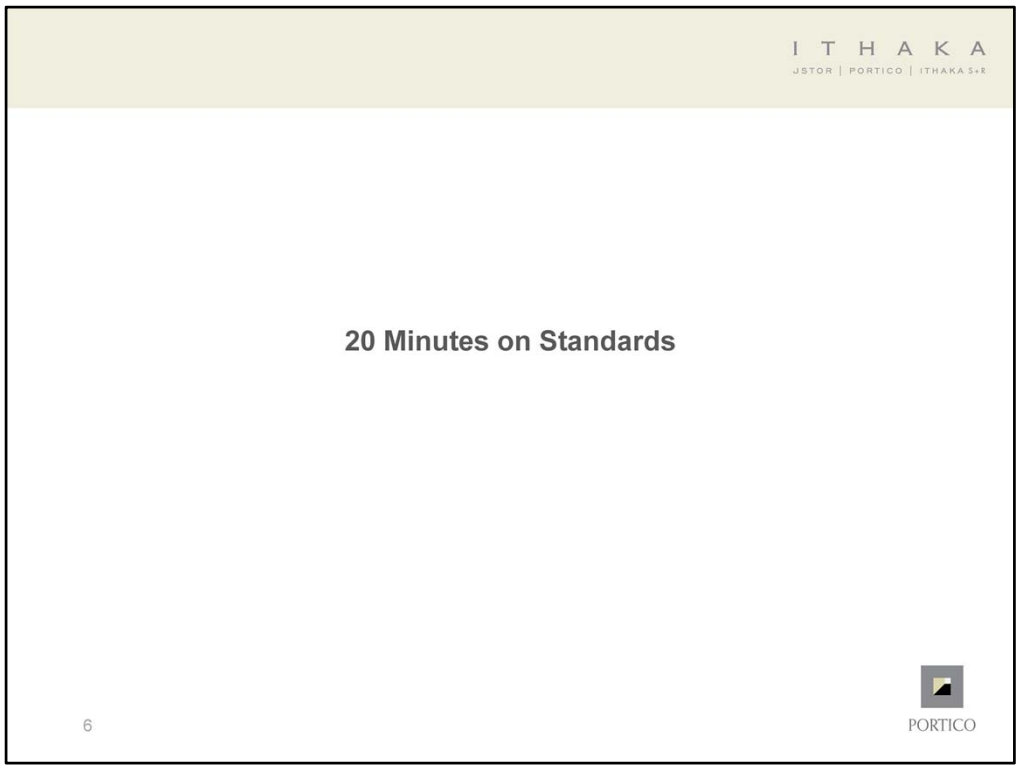
Standards are important. Standards let me plug my laptop, ipad, and cell phone into the wall.

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

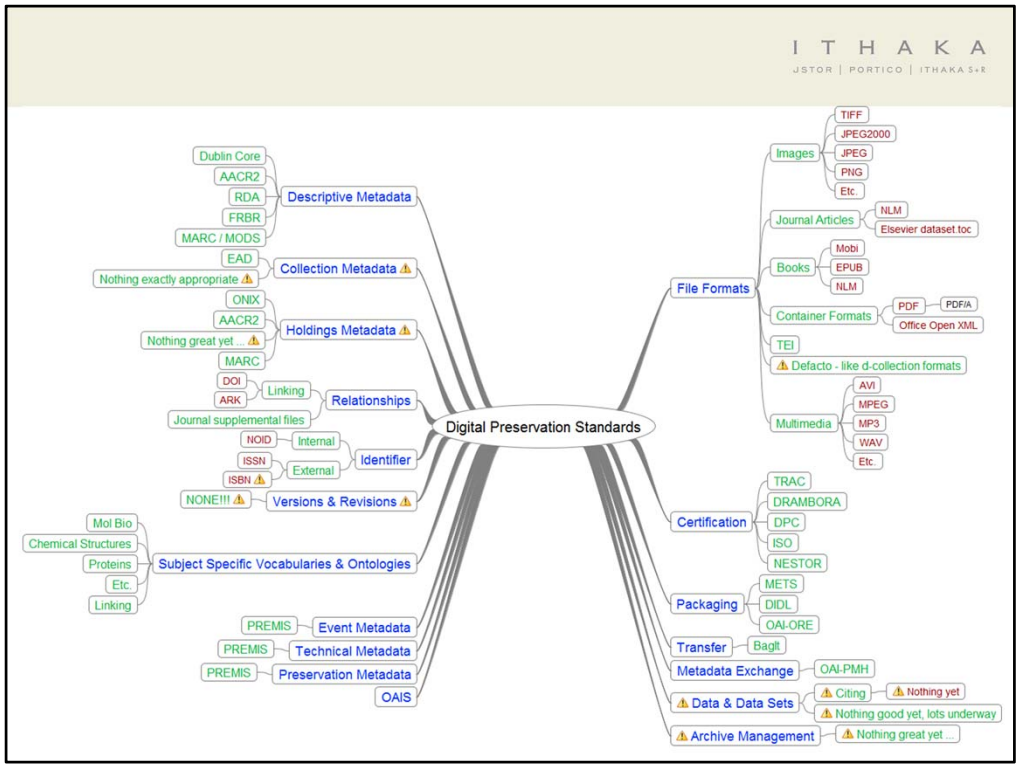
A coulpe of interesting things to point out … <click>

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.

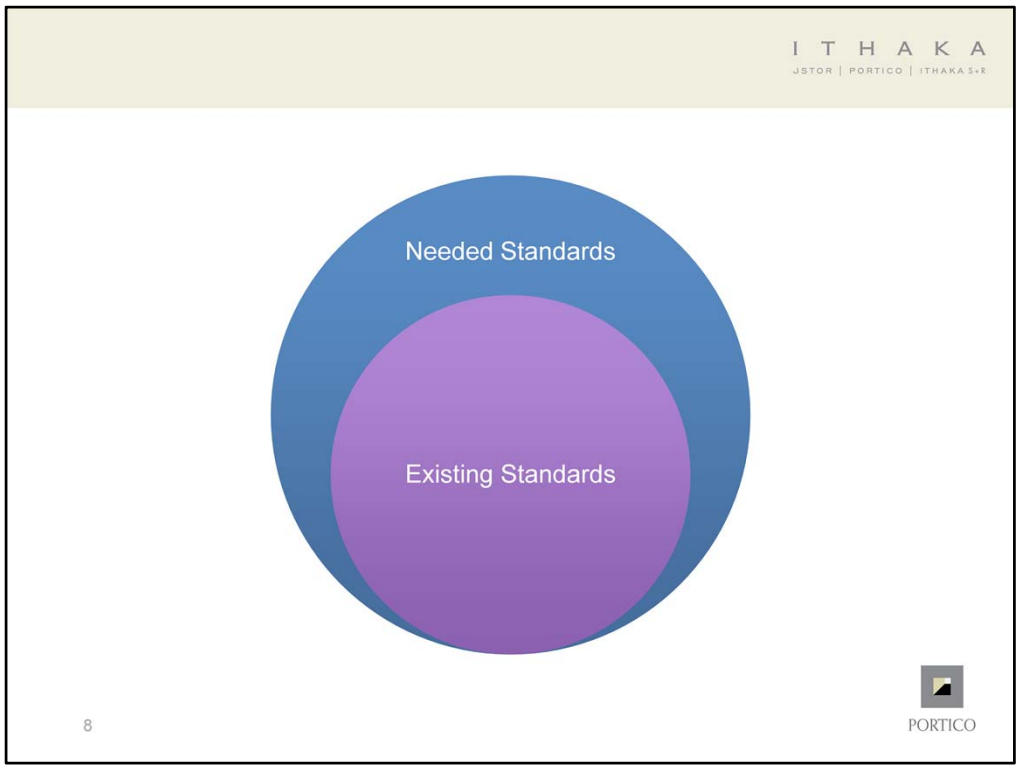


So, when Nettie first asked me to do this I thought … 20 minutes on standards … can I talk for 20 minutes on standards?



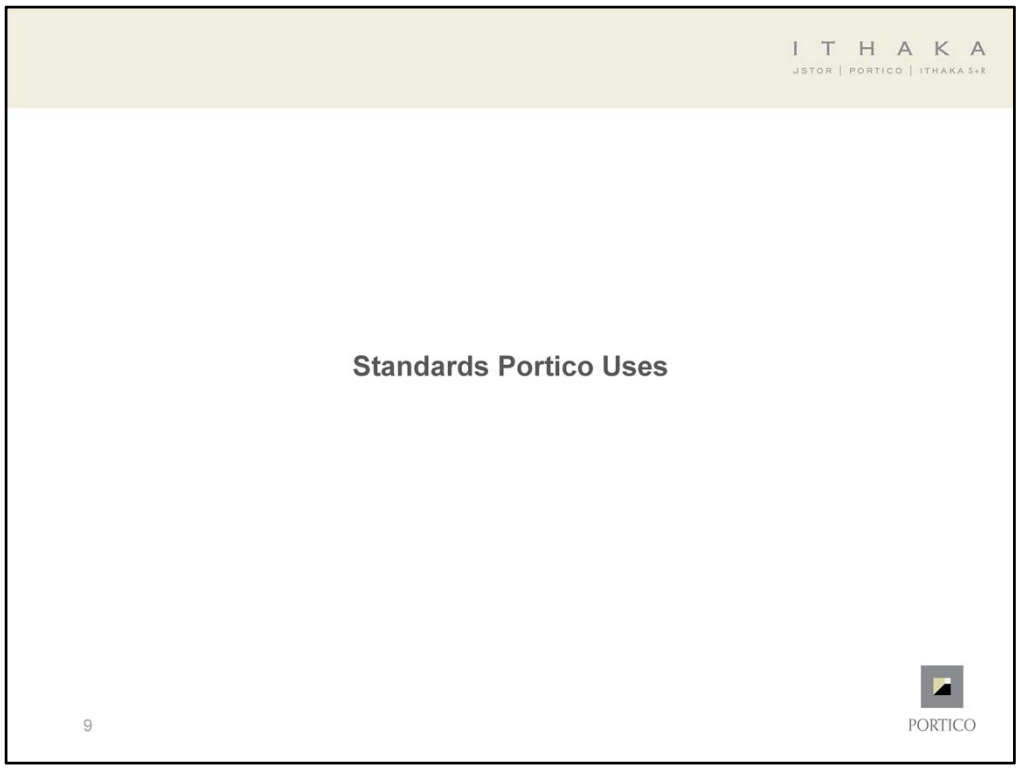
So, I started brainstorming about standards that impact the work we do in digital preservation…

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

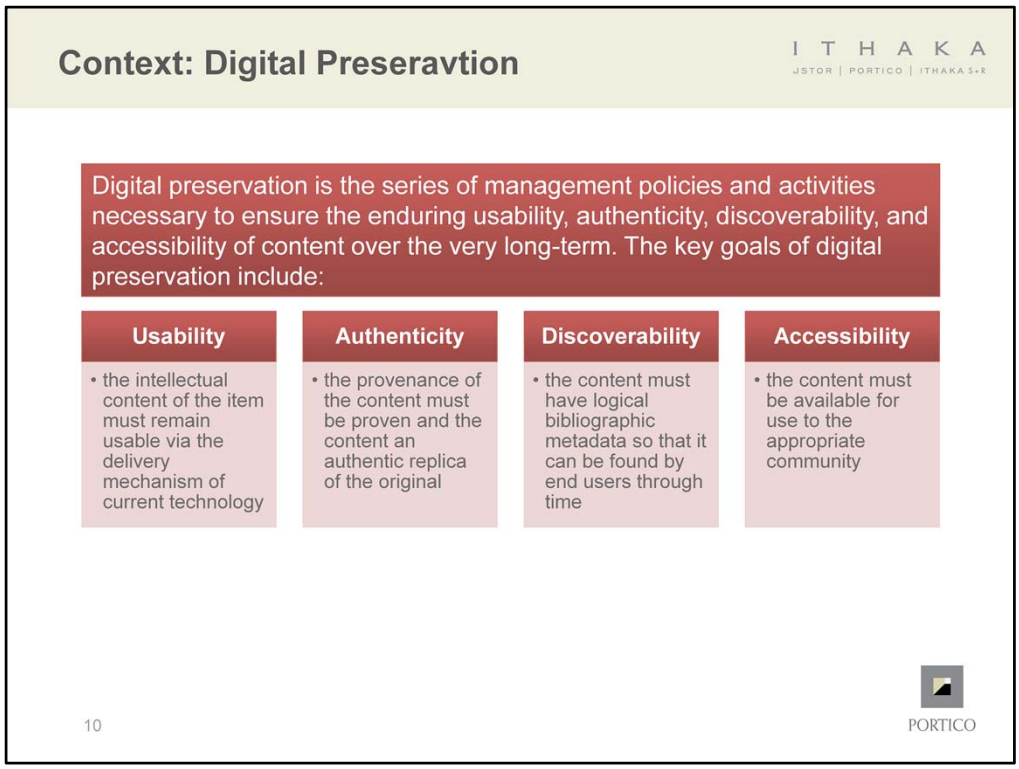


I considered using this as a platform to advocate for some missing standards, or specifications, or even best practices.

And we will come back to that at the end.

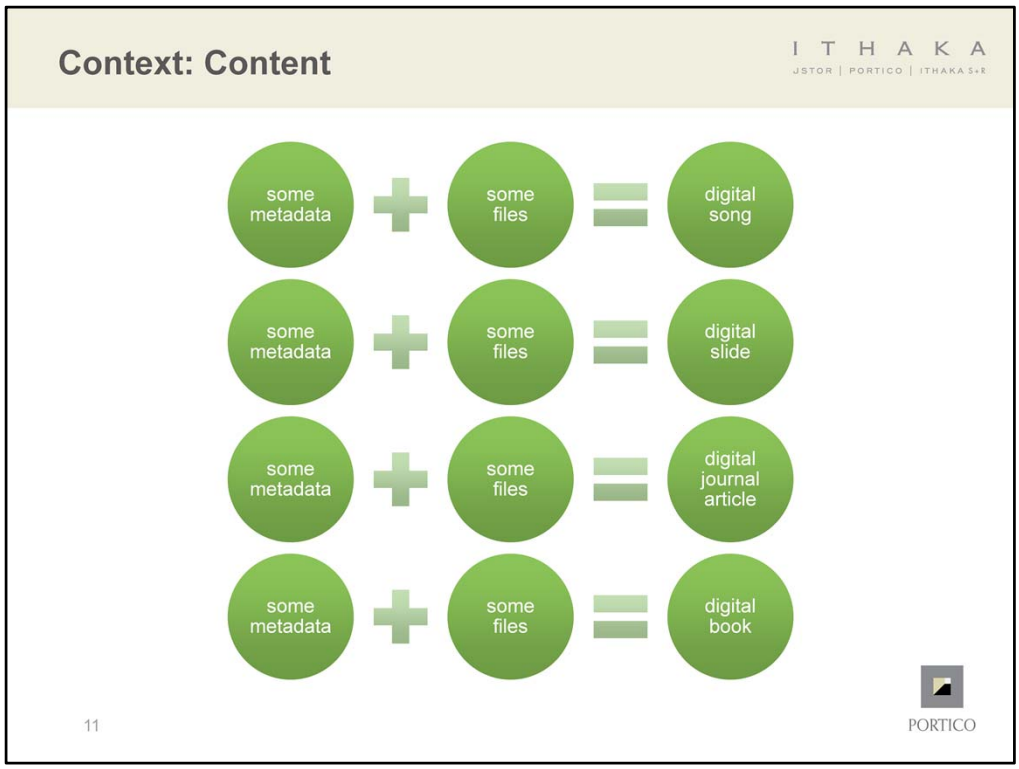


Instead I settled on a discussion of some of the standards Portico uses in its digital preservation work (you'll find out a bit later that I'm a bit free with the word "use" here -- adapters, remember adapters ☺).



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.



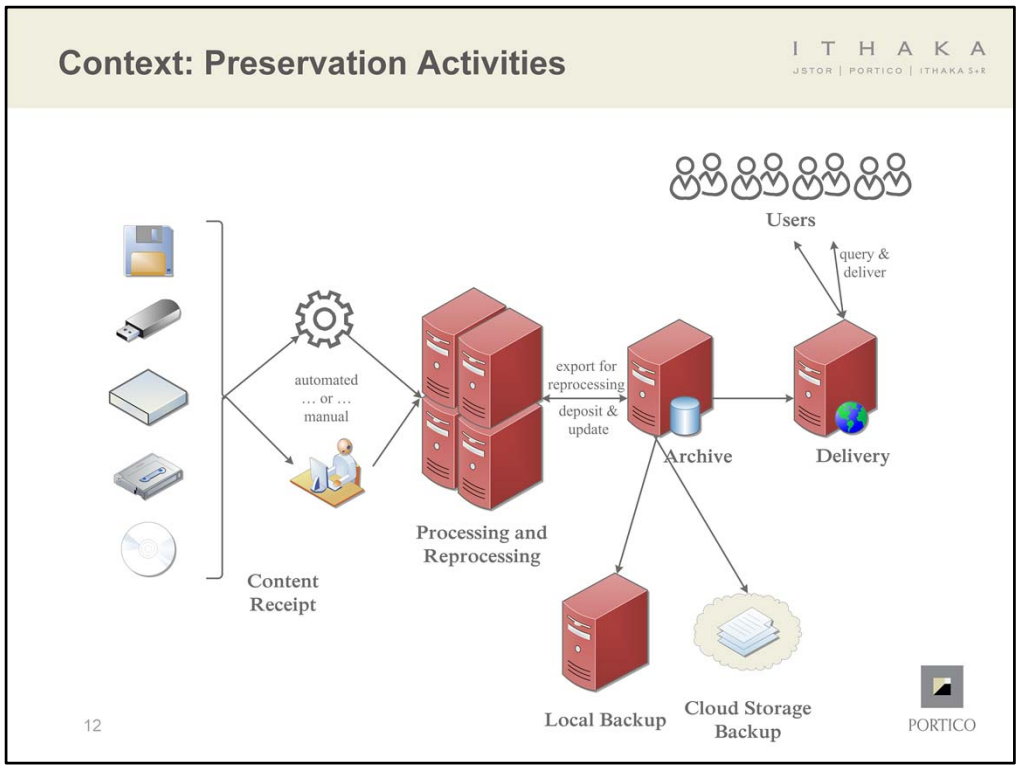
Content that is not a “database product” (like Google maps or dynamic e-reference books), content that has a clear beginning point and ending point … from a content management and preservation perspective that content all looks very similar.

I need to preserver 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.

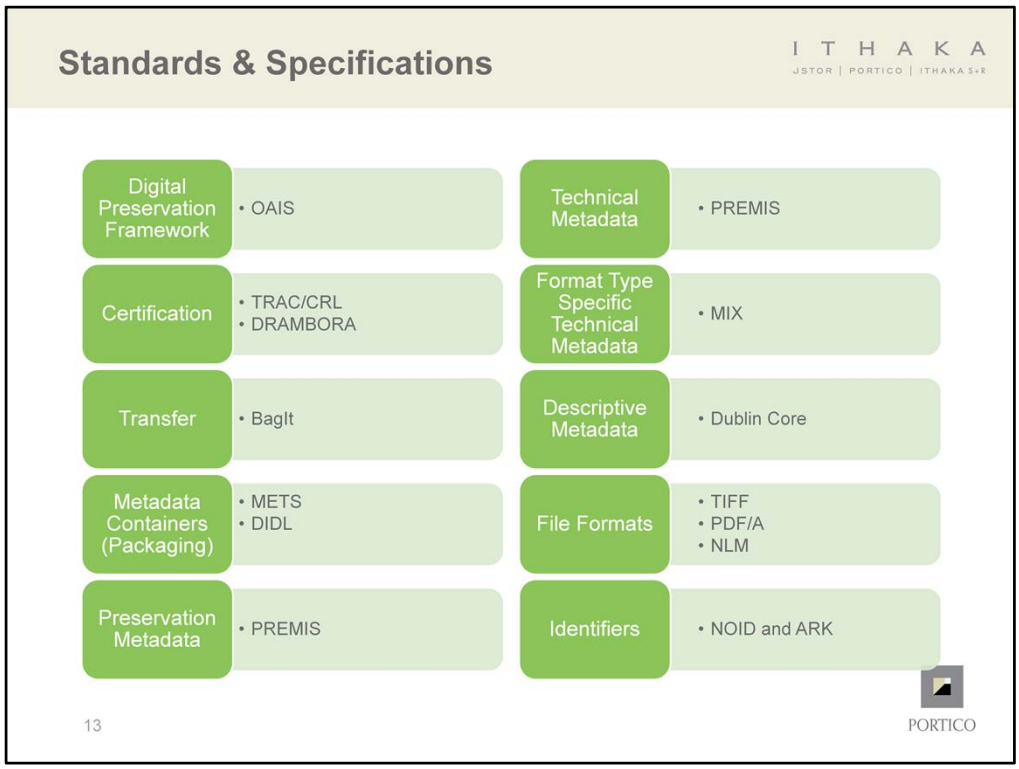
Important later on because there are standards around the metadata and around how these units are packaged that are key elements of digital preservation.



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



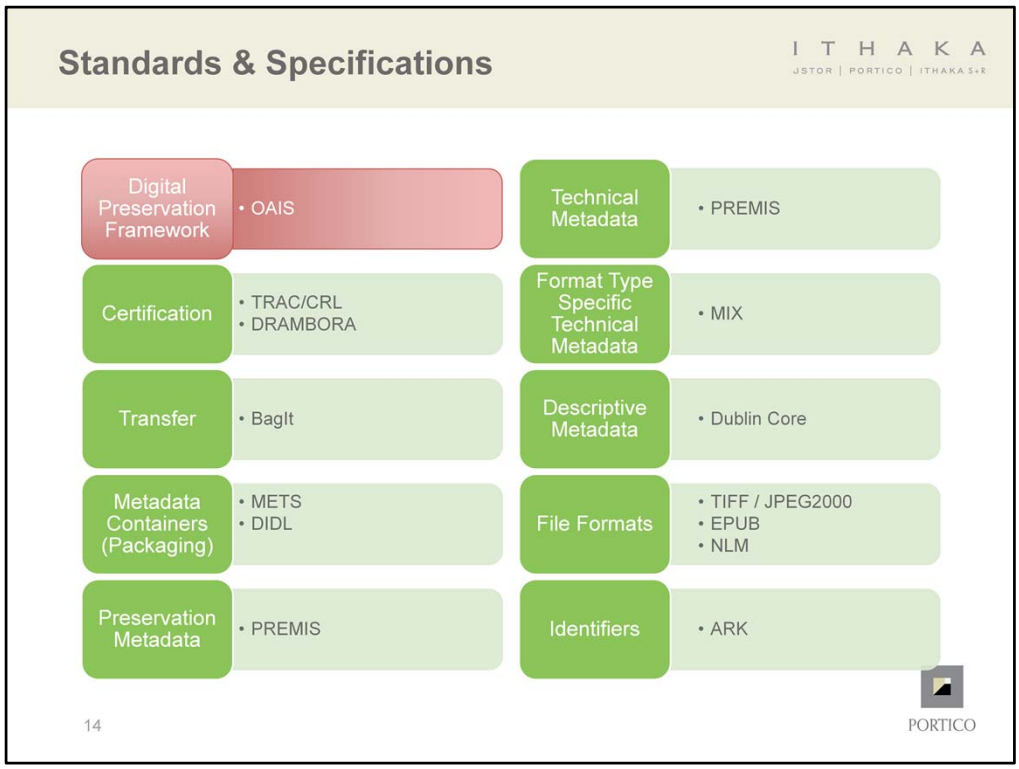
So this takes us to standards & specifications

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

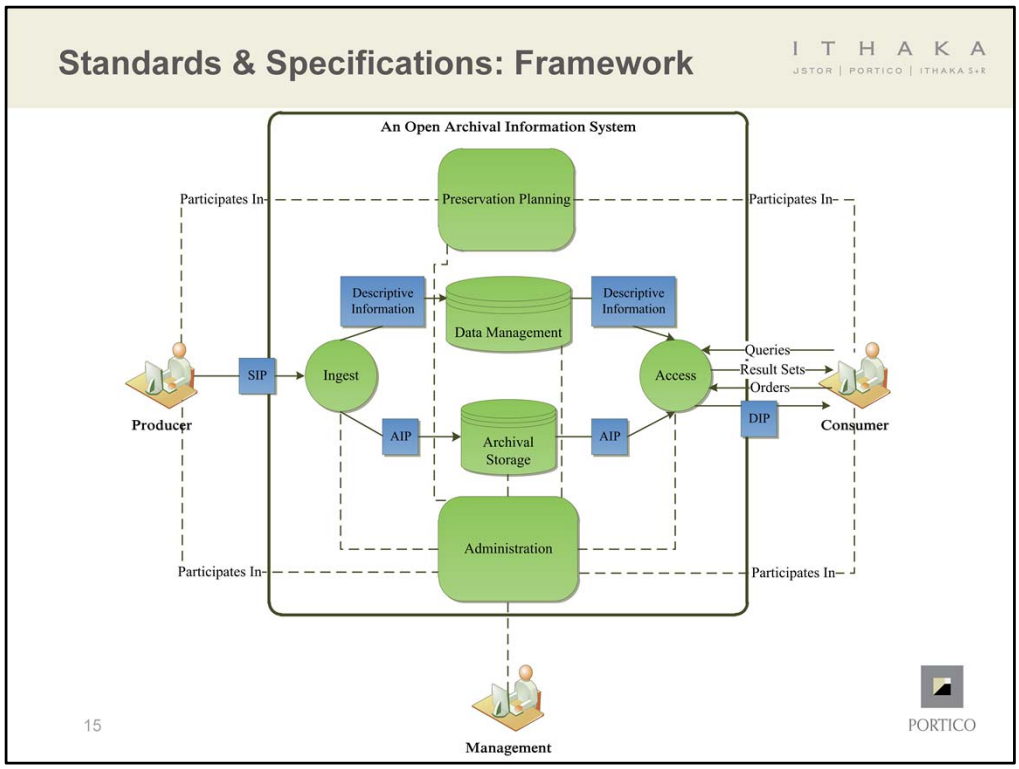
If anyone wants a cheat sheet of some standards, feel free to give me a ping and I can pull something together.

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

Talk about each, eh, maybe not ….



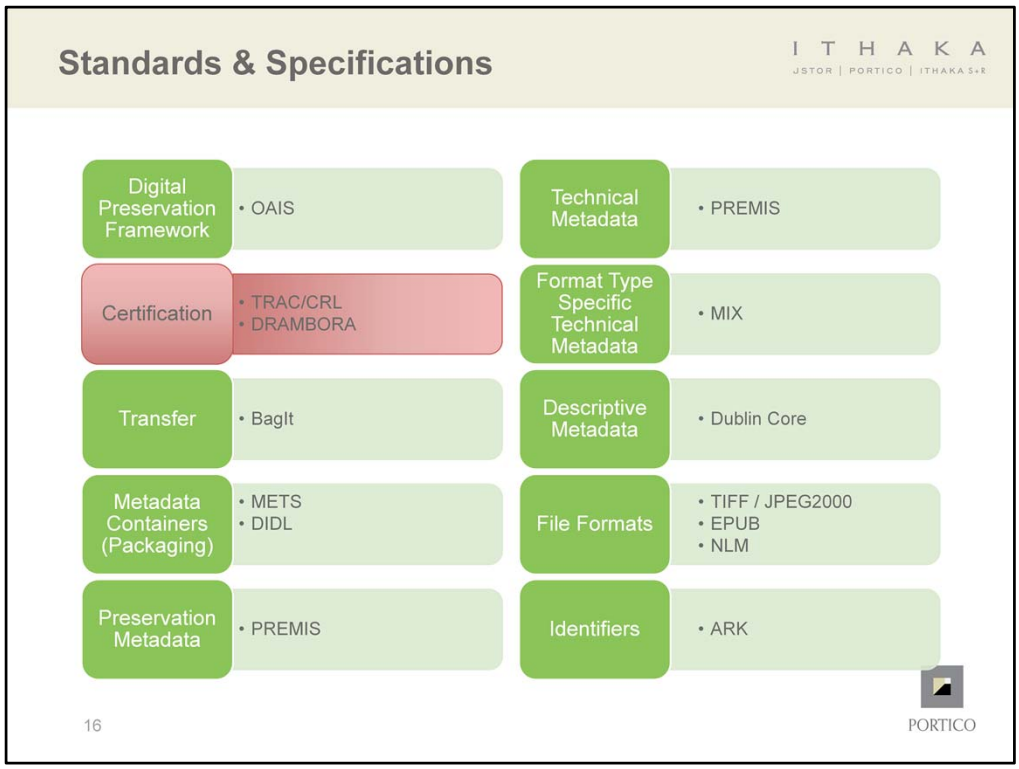
We are going to start with OAIS, as every discussion of digital preservation does …



- 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

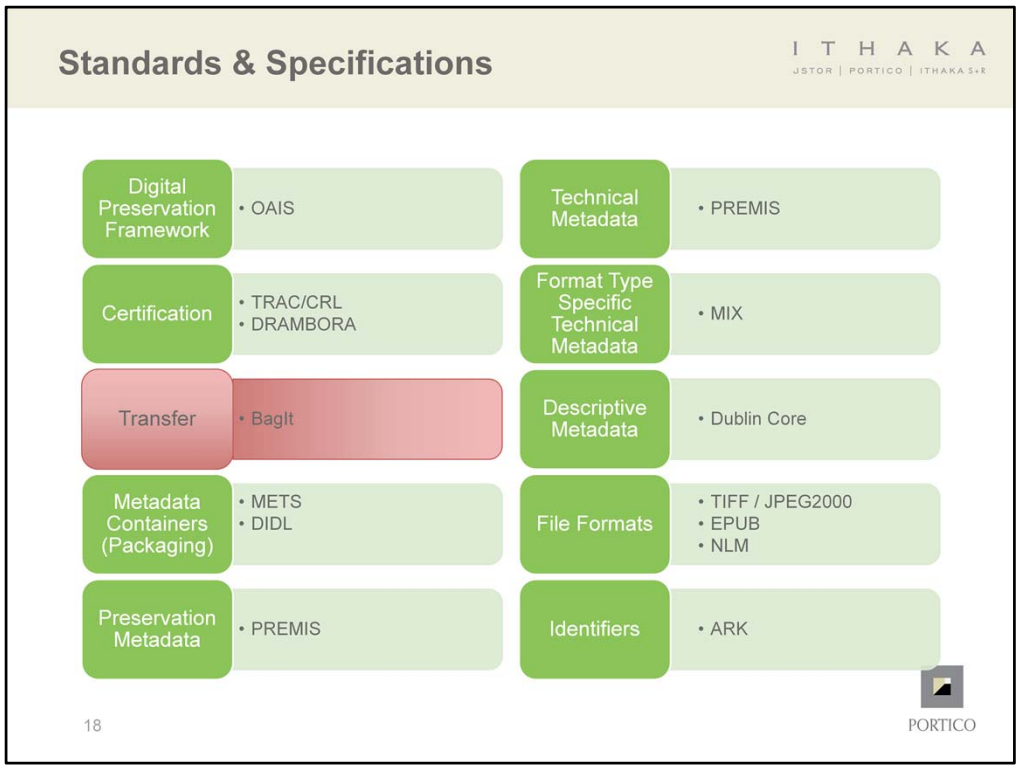
- It's a big document … my take on the 7 highlights …

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 …

Marie is going to speak later on certification, so I'll just mention that there are a number of standards and specifications around assessing and certifiying preservation services as trustworthy.

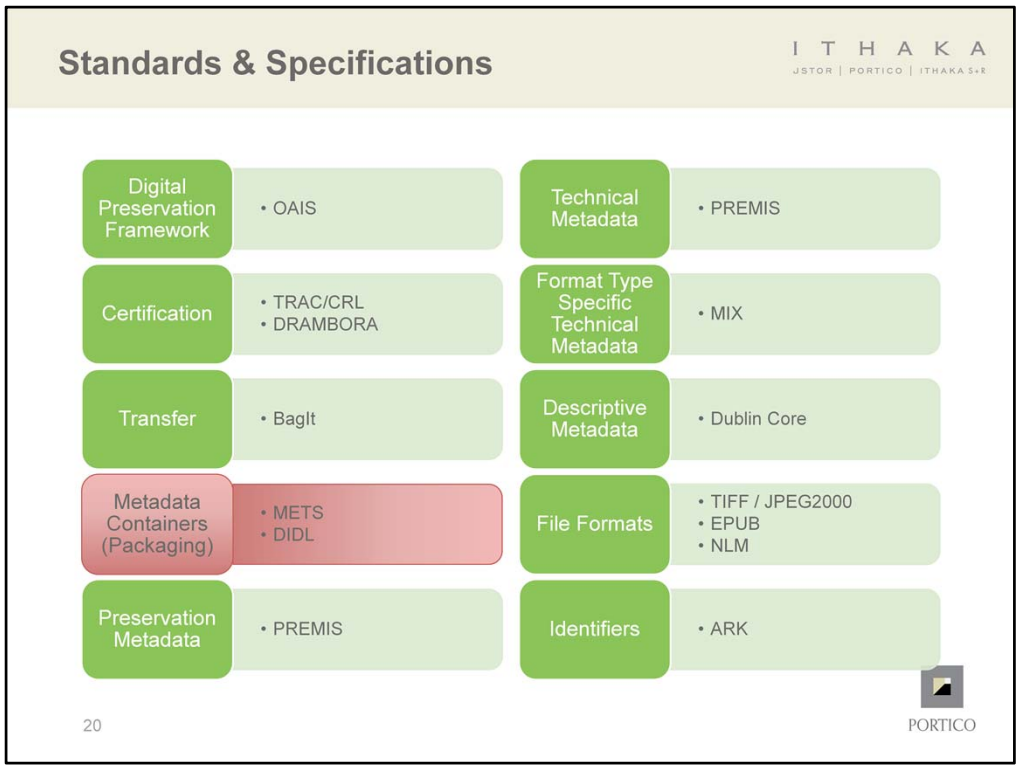


Every one preserving content will need to move that content at some point … enter standards about transfering content

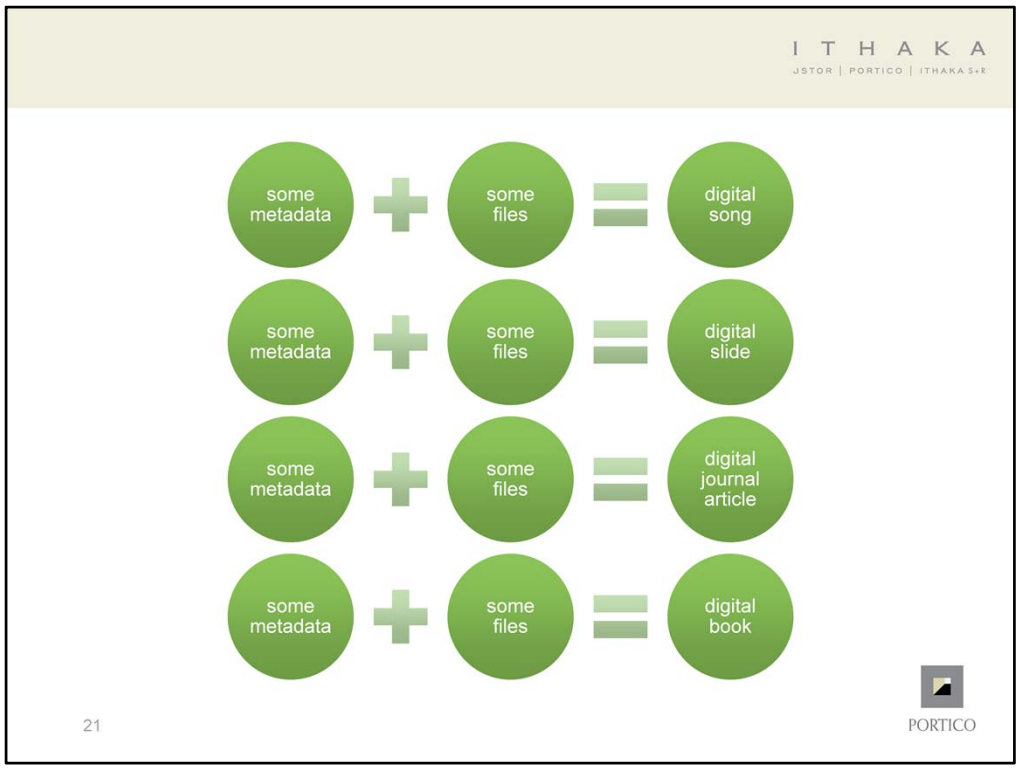
<click>

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.

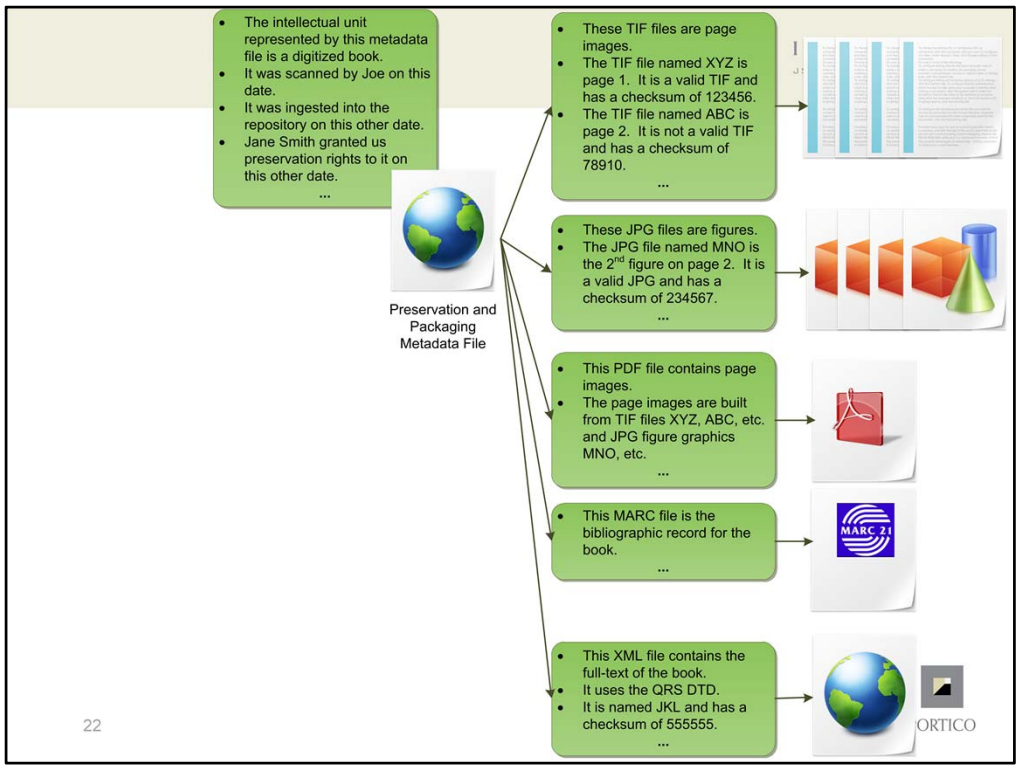


This is actually the spot where I'm going to chat at you the longest …



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.

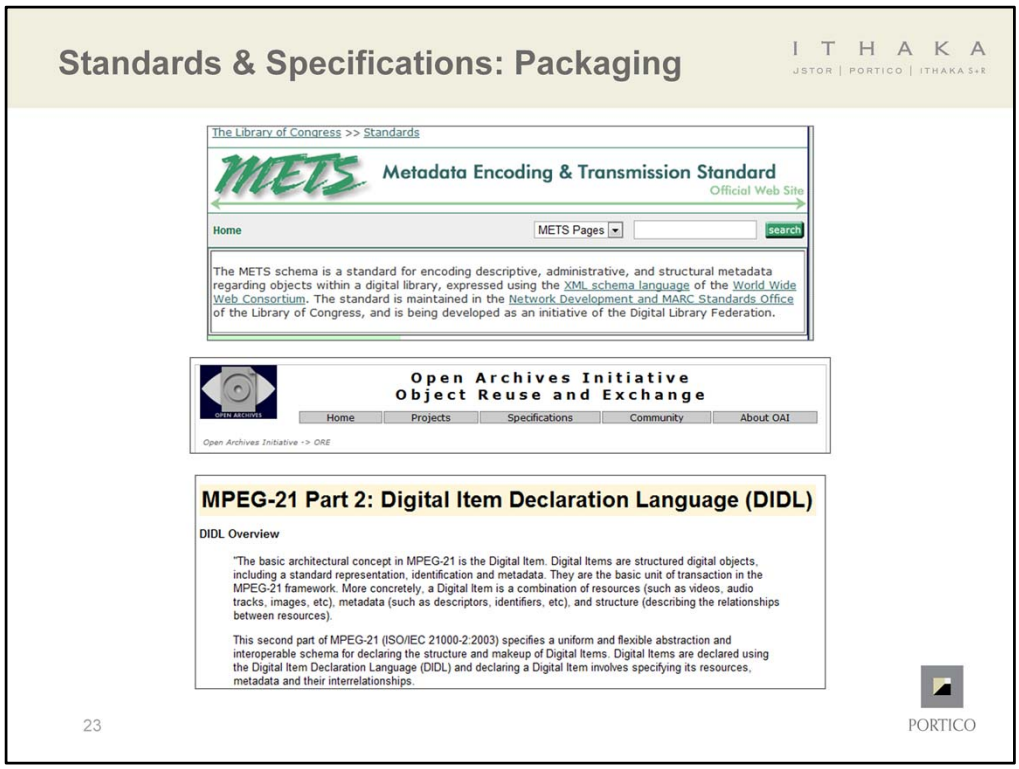


As we said <click back> every unit has files and metadata ..

<click forward>

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

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

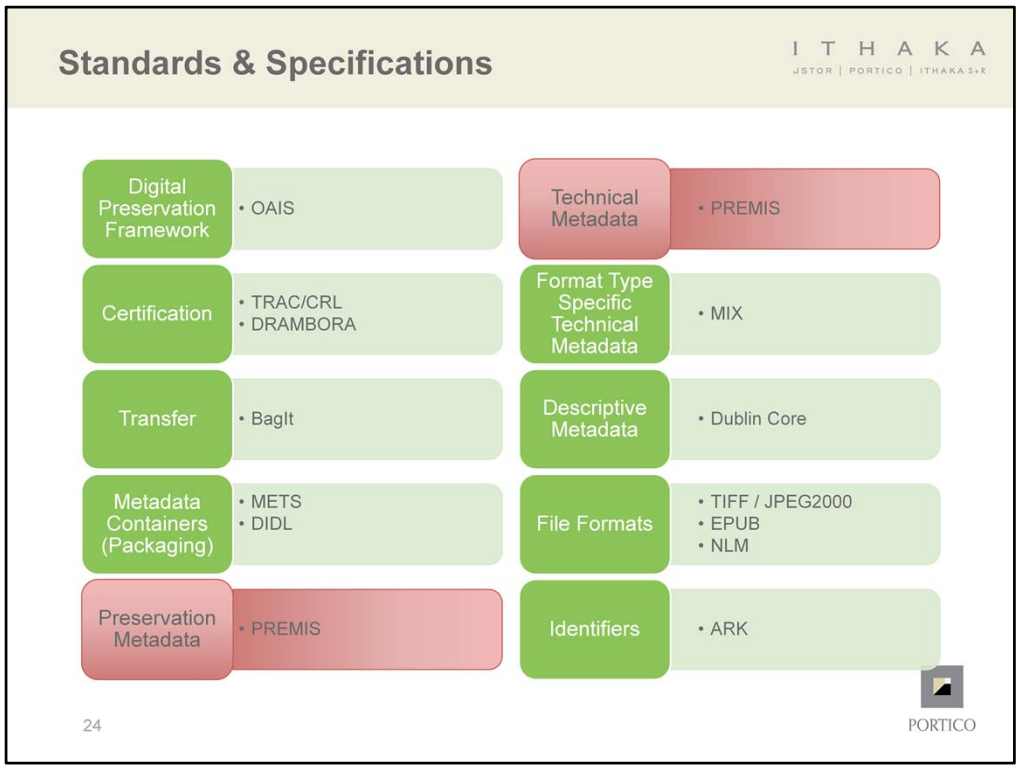


A couple of standards in use …

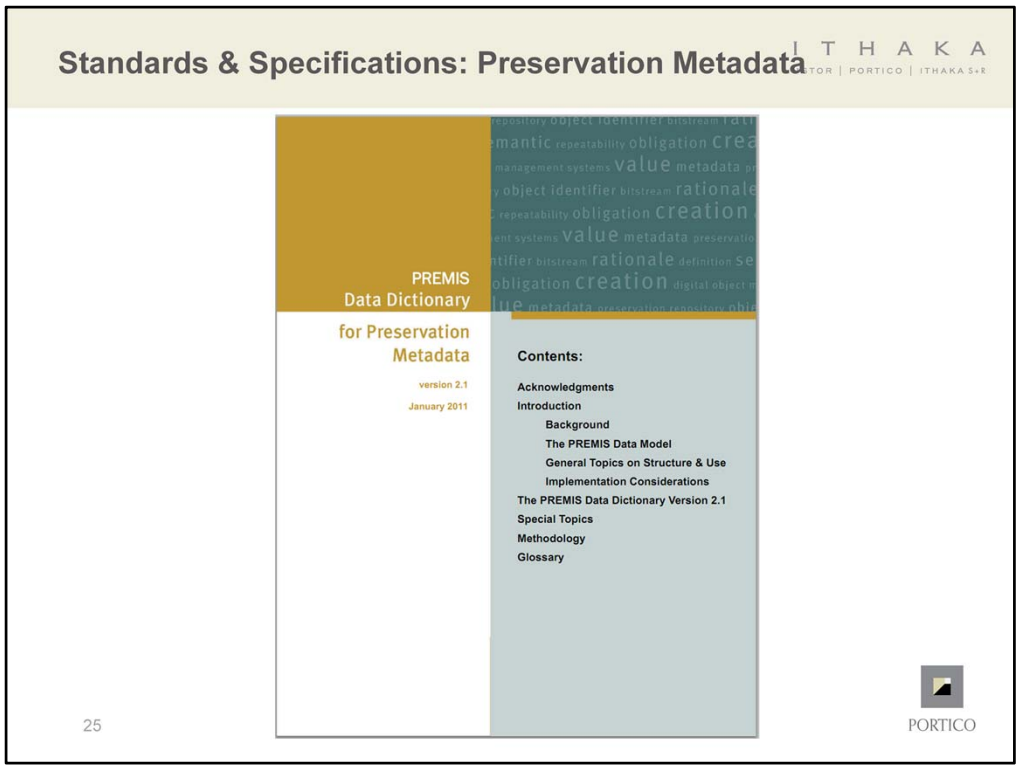
METS and DIDL are big player … METS more than DIDL ..

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.

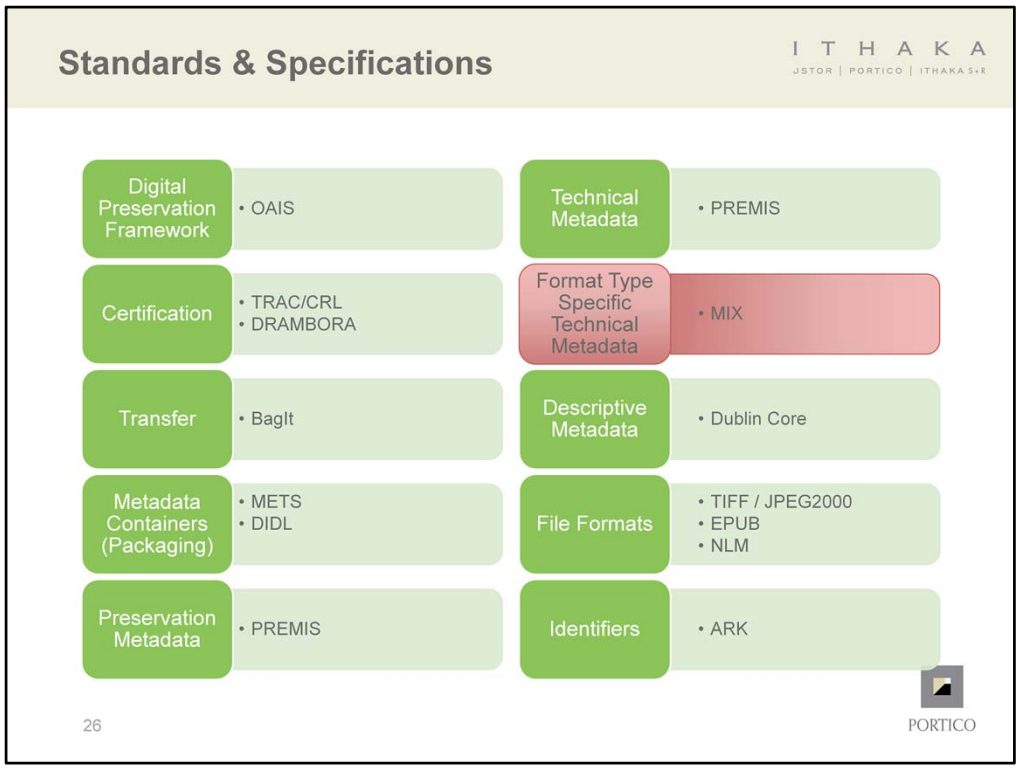


Next is preservation metadata of which there is one big player … <click>

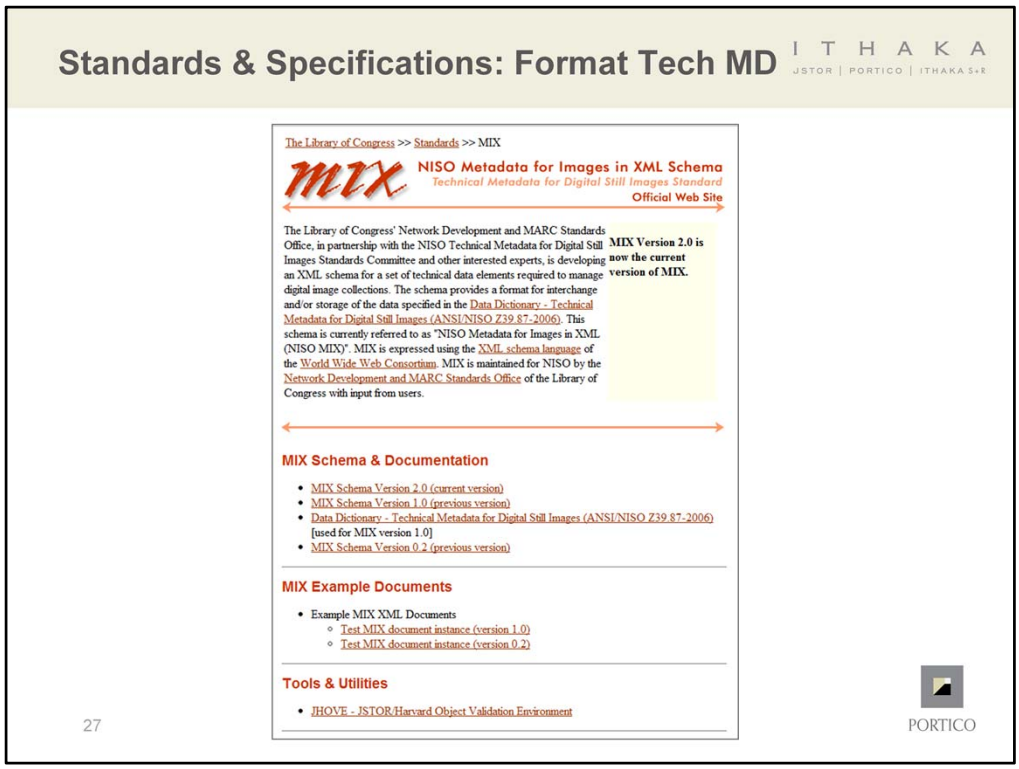


PREMIS …

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 propreties of a preserved unit of content.

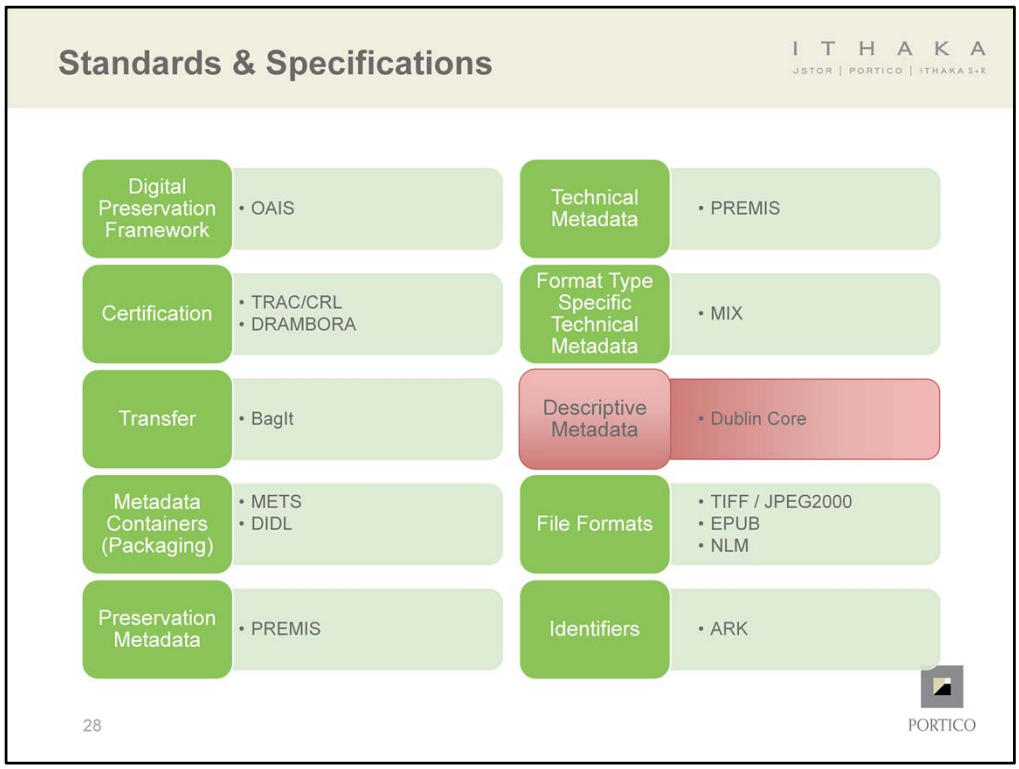


Another area of standards & specifications is format or content type specific technical metadata <click>



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.)



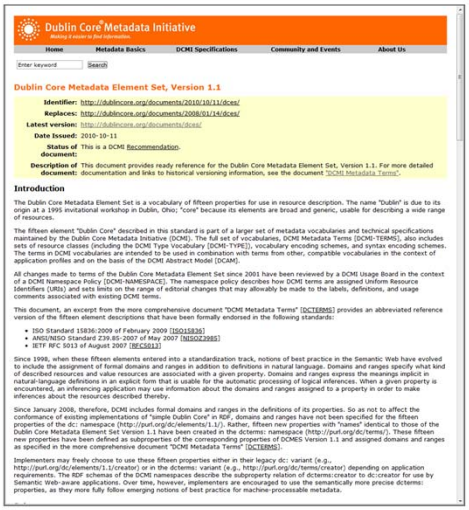
And, then, there are standards around the bibliographic metadata describing a preserved unit …

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

ITHAKA
JSTOR | PORTICO | ITHAKA S+R



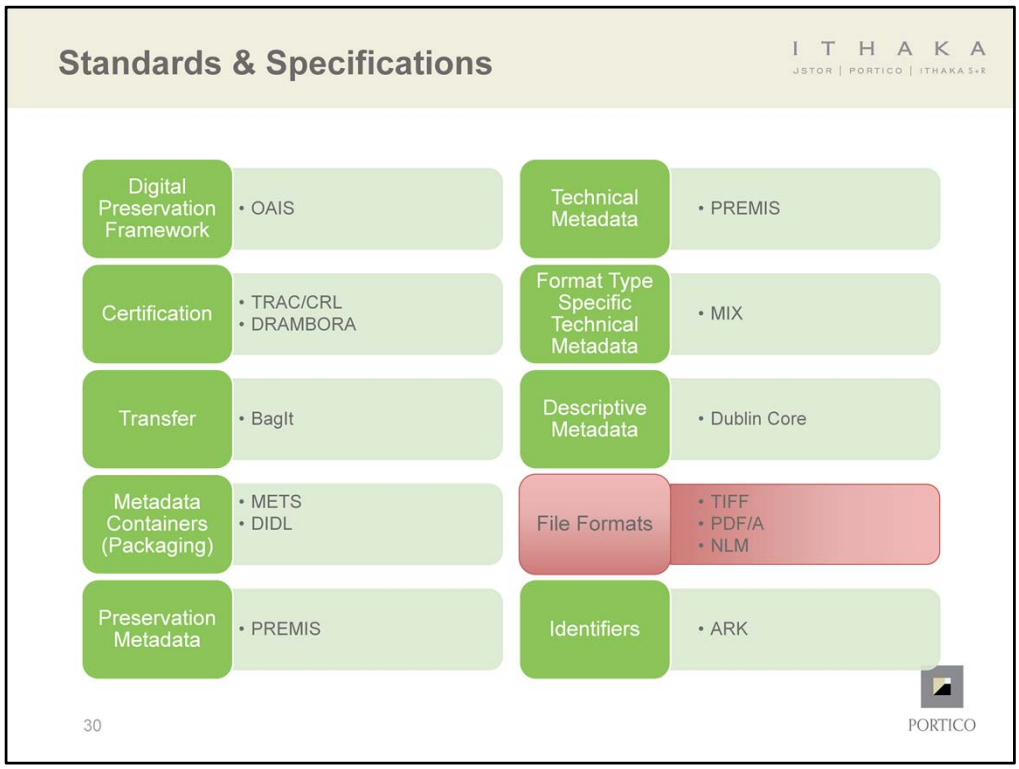
- Contributor
- Coverage
- Creator
- Date
- Description
- Format
- Identifier
- Language
- Publisher
- Relation
- Rights
- Source
- Subject
- Title
- Type

29

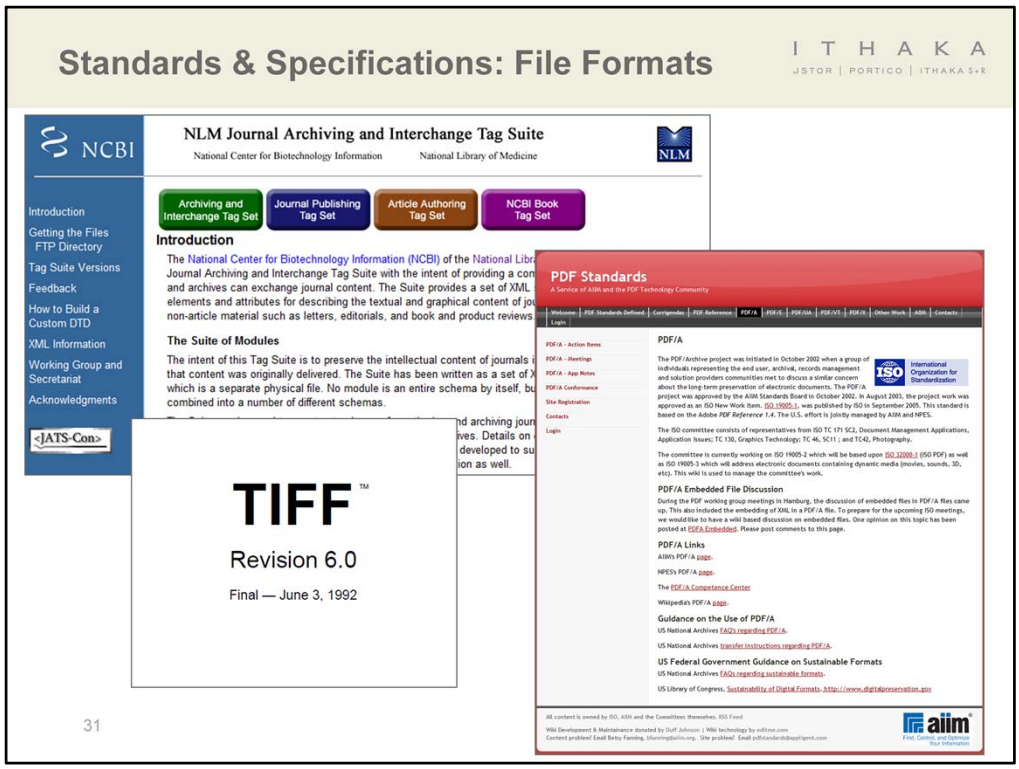
PORTICO

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.



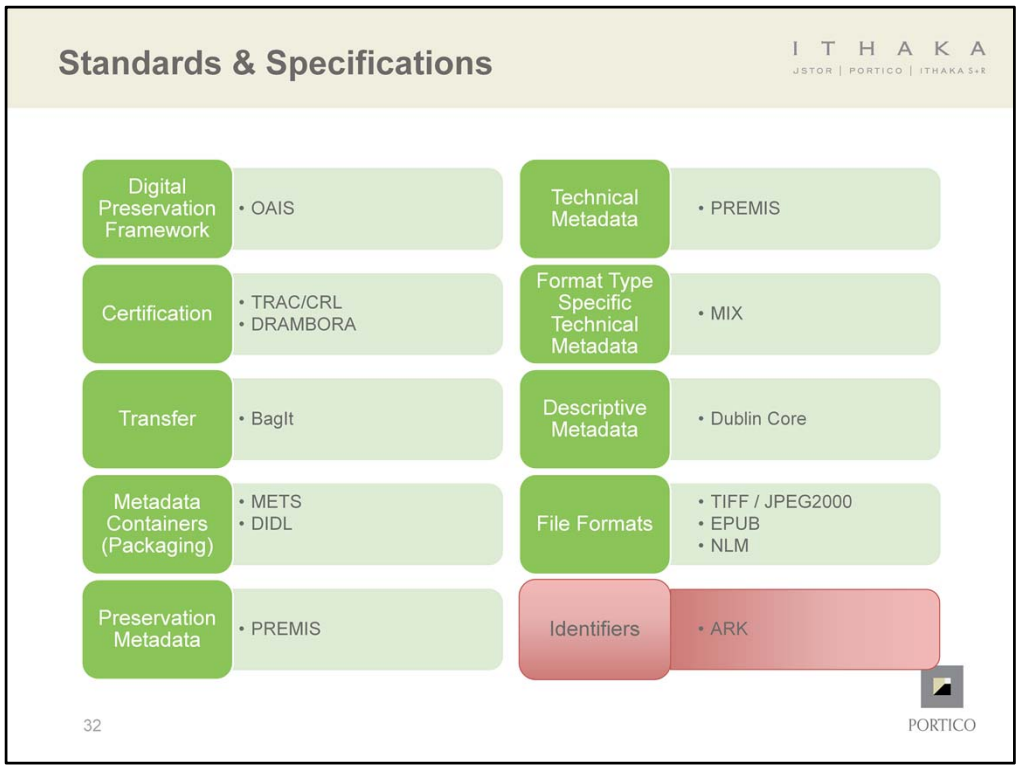
Each file that is preserved is in a format and most of the formats will be associated with standards … <click>



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.

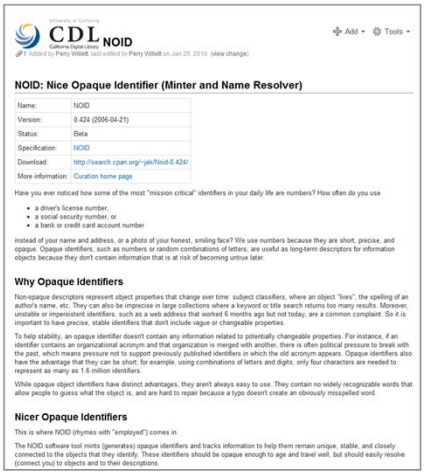
This area is very broad, however, so this is just a sampling.



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

Standards & Specifications: Identifiers

ITHAKA
JSTOR | PORTICO | ITHAKA S+R



CDL NOID

Add Tools

1 Added by Perry Willett, last edited by Perry Willett on Jan 25, 2010 (view change)

NOID: Nice Opaque Identifier (Minter and Name Resolver)

Name:	NOID
Version:	0.424 (2006-04-21)
Status:	Beta
Specification:	NOID
Download:	http://search.cpan.org/~jak/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 impersistent 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.

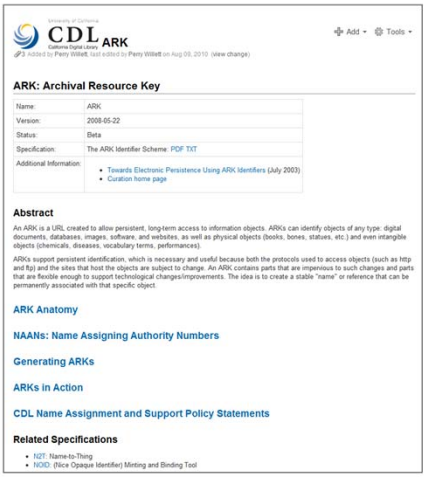
To help stability, an opaque identifier doesn't contain any information related to potentially changeable properties. For instance, if an identifier contains an organizational acronym and that organization is merged with another, there is often political pressure to break with the past, which means pressure 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 repair because a typo doesn't create an obviously misspelled word.

Nicer Opaque Identifiers

This is where NOID (rhymes with "employed") comes in.

The NOID software tool mints (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.



CDL ARK

Add Tools

3 Added by Perry Willett, last edited by Perry Willett on Aug 09, 2010 (view change)

ARK: Archival Resource Key

Name:	ARK
Version:	2006-05-22
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

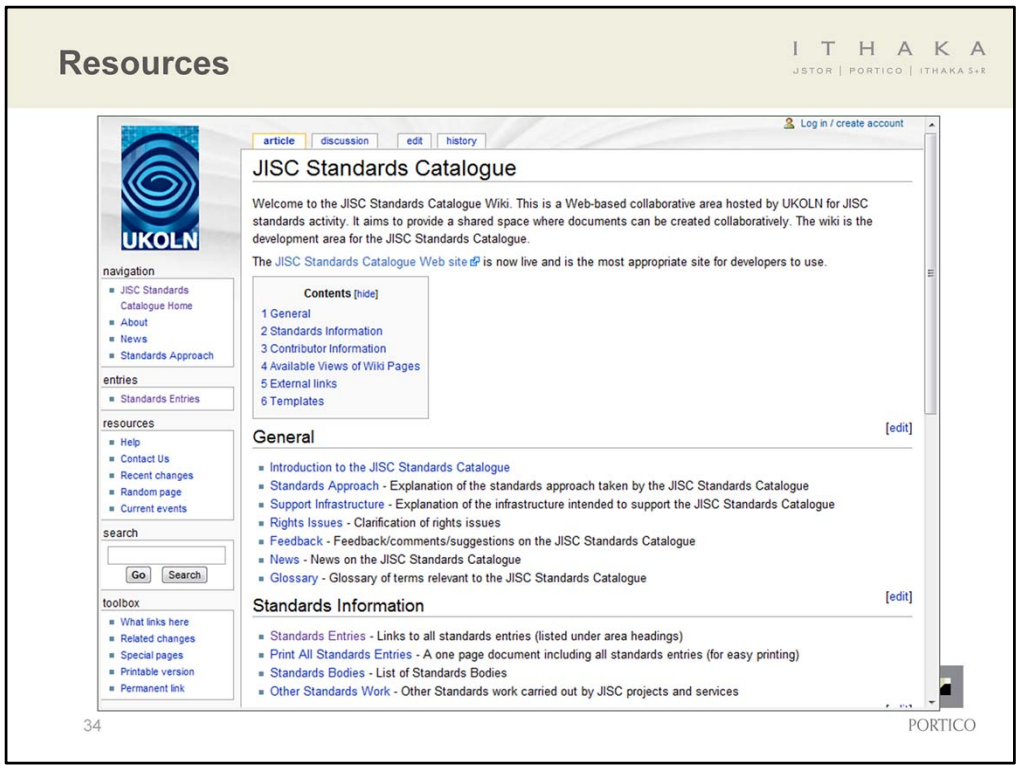
- N2T: Name-to-Thing
- NOID: (Nice Opaque Identifier) Minting and Binding Tool

33

PORTICO

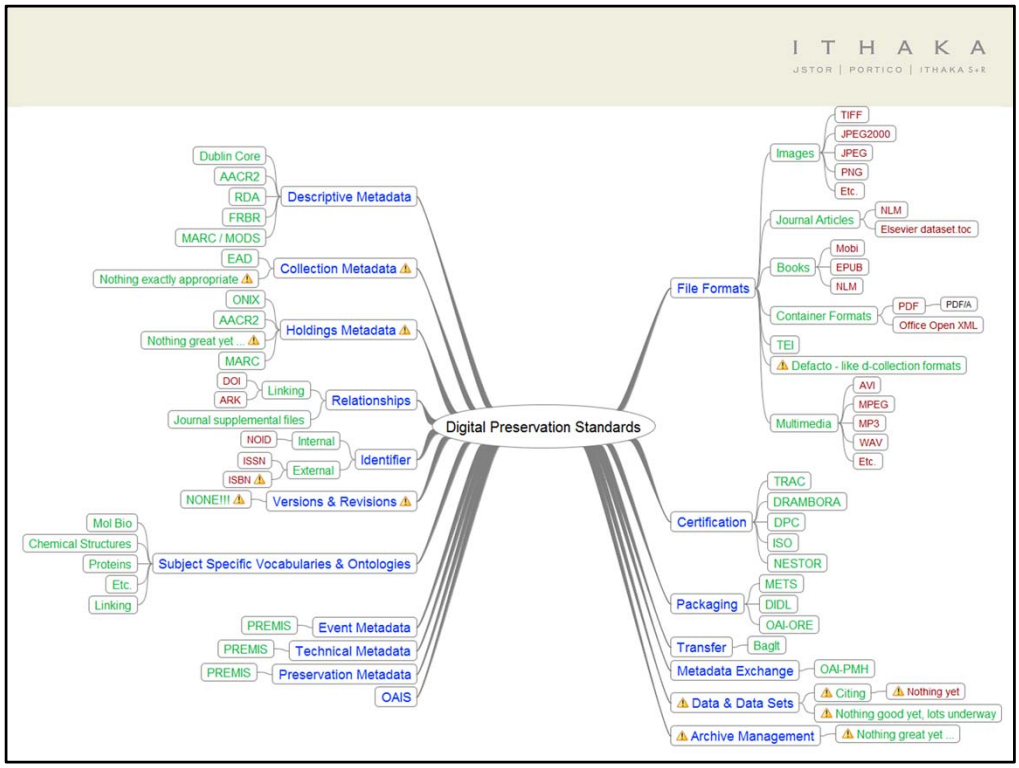
At Portico we extensively use the CDL Curation Micro-Services NOID software and ARK specification. We assign unique internal IDs, in the form of ARKs, to everything.

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



… 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.

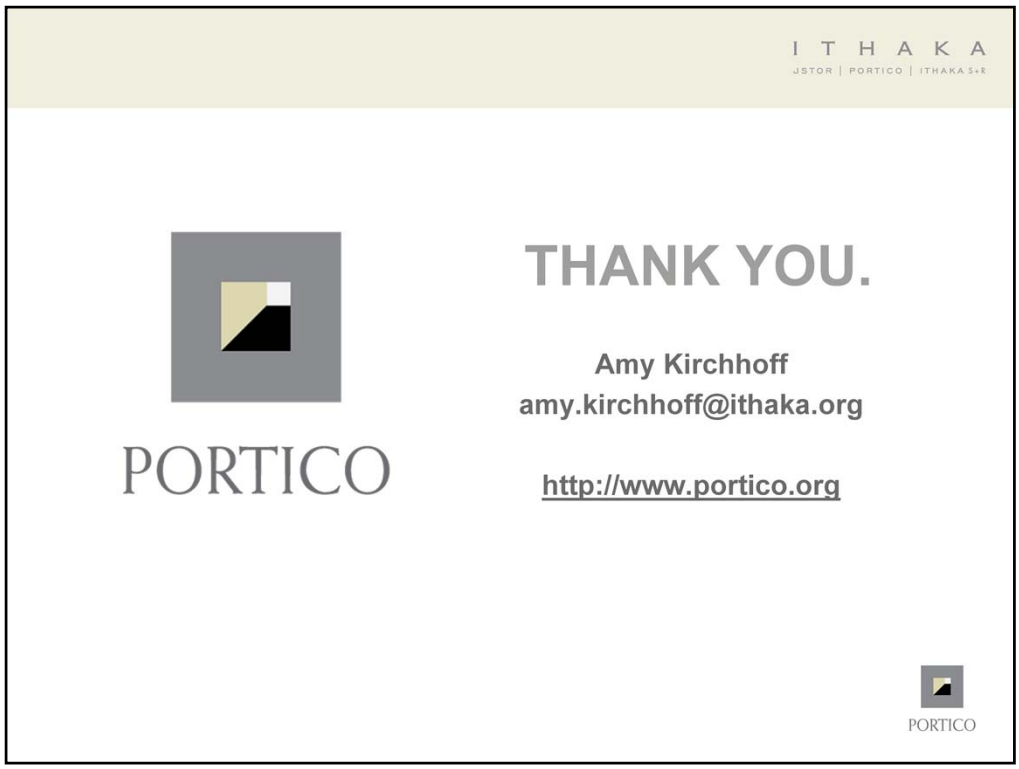
Again, if anyone wants a cheat sheet of standards we use at Portico, their purpose, and where those standards live, please just give me a ping.



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.