A Preservation Infrastructure for Digital Stewardship at CUL

Bill Kehoe and Adam Smith

May 23, 2008

the library's legacy of leadership

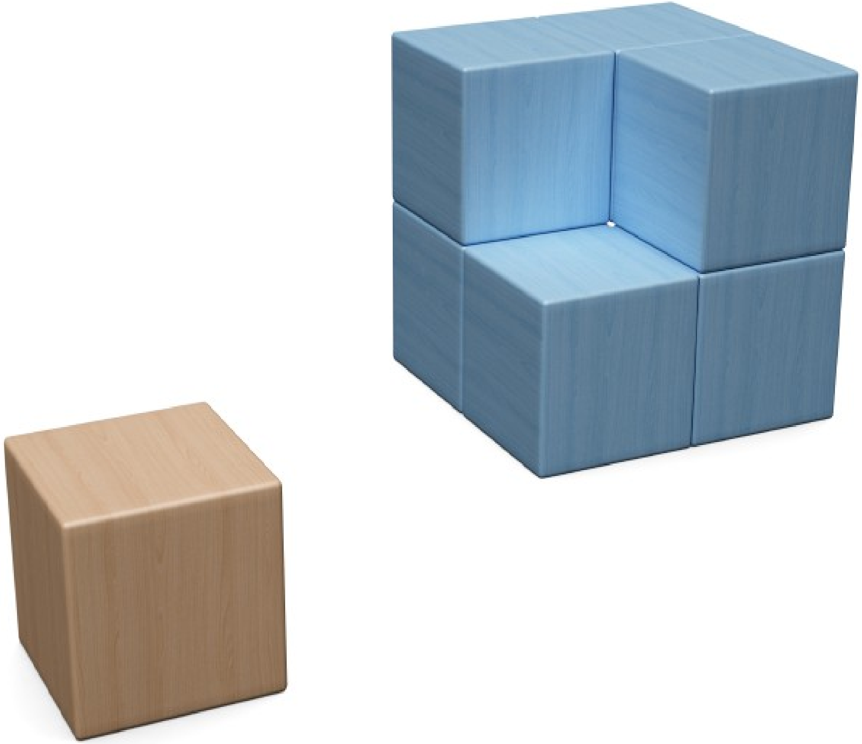
we can lead again

our current infrastructure is fragmented

a more cohesive infrastructure
is within our reach

the missing ingredient: organizational commitment

what do we already have that we can use?



collections

the technology of content

the technology of disclosure

identity
and
description

handles
and
collection registry

handles

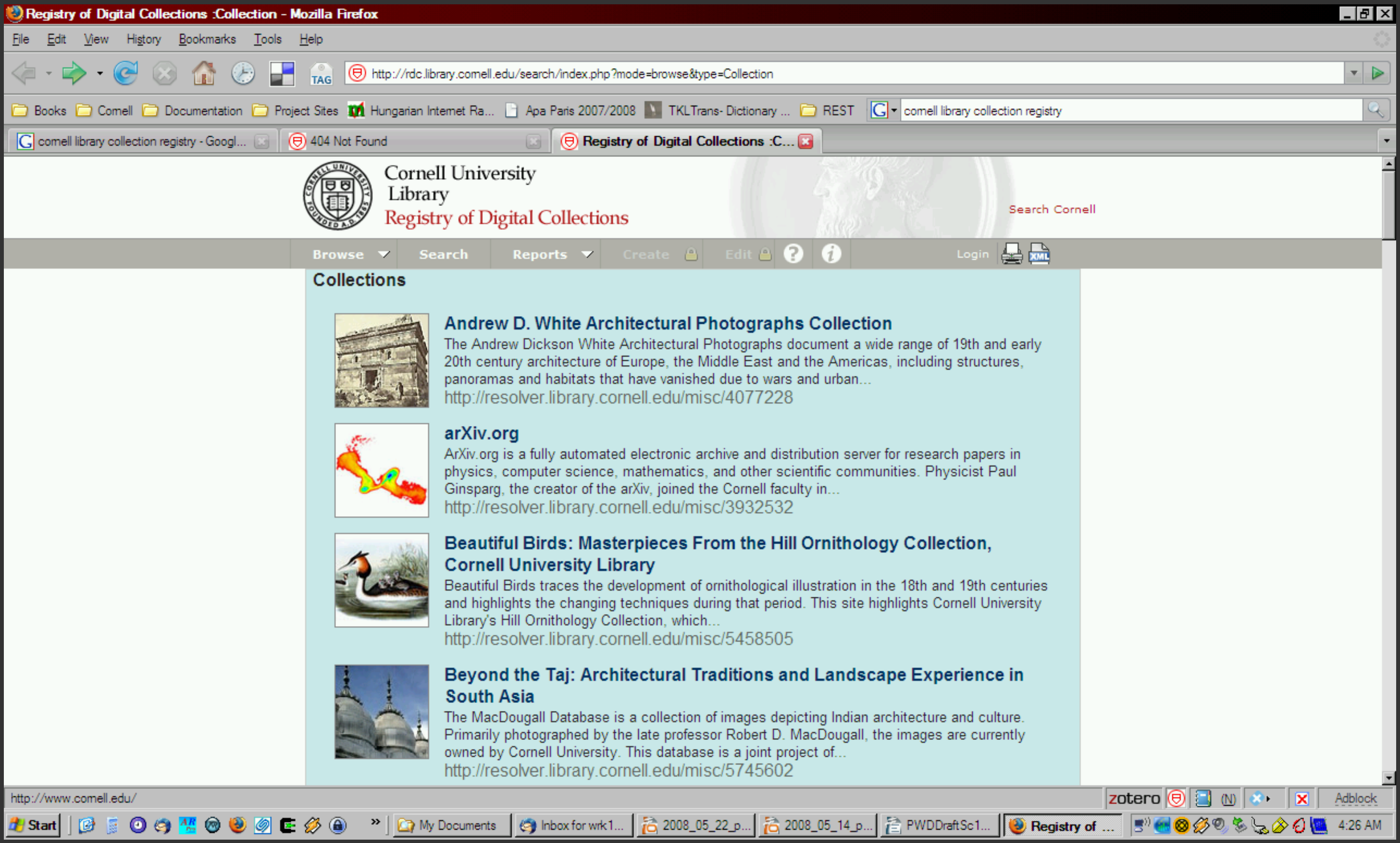
1803.01/134a74d60

100: some_site.library.cornell.edu/resource

200: cul-oais.library.cornell.edu/archived

300: de9f2c7f

registry of digital collections



Registry of Digital Collections :Collection - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://rdc.library.cornell.edu/search/index.php?mode=browse&type=Collection
Books Cornell Documentation Project Sites Hungarian Internet Ra... Apa Paris 2007/2008 TKLTrans- Dictionary ... REST
cornell library collection registry
cornell library collection registry - Googl...
404 Not Found
Registry of Digital Collections :C...
Cornell University Library
Registry of Digital Collections
Search Cornell
Browse
Search
Reports
Create
Edit
Login
http://www.cornell.edu/
zotero
Adblock
Start
My Documents
Inbox for wrk1...
2008_05_22_p...
2008_05_14_p...
PWDDraftSc1...
Registry of ...
4:26 AM

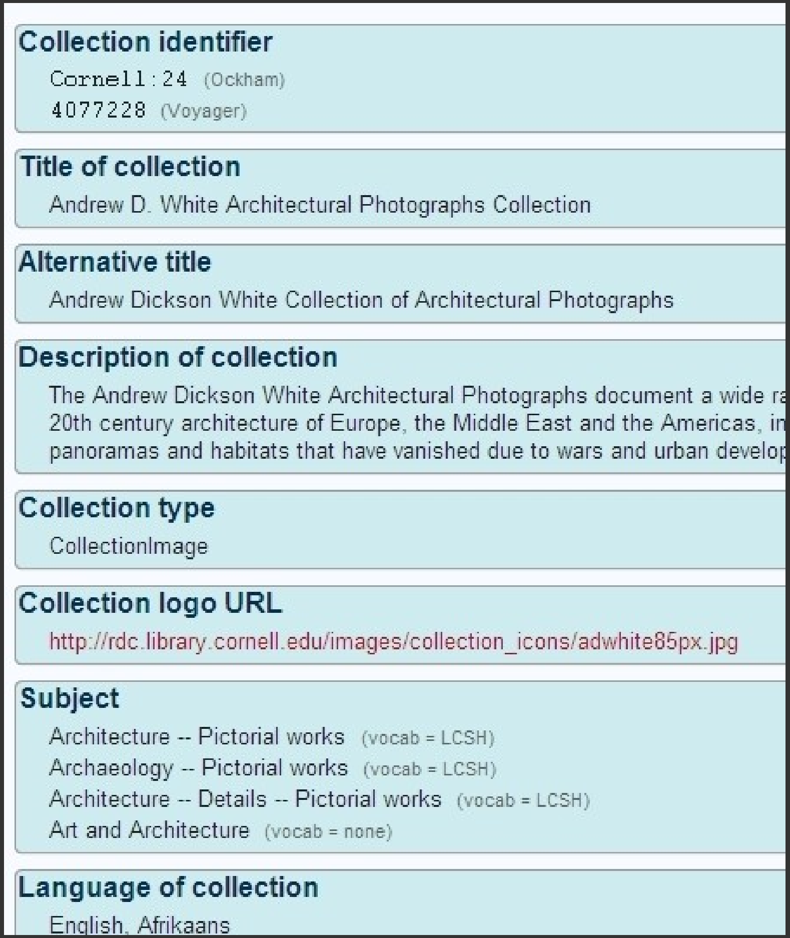
Collections

Andrew D. White Architectural Photographs Collection
The Andrew Dickson White Architectural Photographs document a wide range of 19th and early 20th century architecture of Europe, the Middle East and the Americas, including structures, panoramas and habitats that have vanished due to wars and urban...
http://resolver.library.cornell.edu/misc/4077228

arXiv.org
ArXiv.org is a fully automated electronic archive and distribution server for research papers in physics, computer science, mathematics, and other scientific communities. Physicist Paul Ginsparg, the creator of the arXiv, joined the Cornell faculty in...
http://resolver.library.cornell.edu/misc/3932532

Beautiful Birds: Masterpieces From the Hill Ornithology Collection, Cornell University Library
Beautiful Birds traces the development of ornithological illustration in the 18th and 19th centuries and highlights the changing techniques during that period. This site highlights Cornell University Library's Hill Ornithology Collection, which...
http://resolver.library.cornell.edu/misc/5458505

Beyond the Taj: Architectural Traditions and Landscape Experience in South Asia
The MacDougall Database is a collection of images depicting Indian architecture and culture. Primarily photographed by the late professor Robert D. MacDougall, the images are currently owned by Cornell University. This database is a joint project of...
http://resolver.library.cornell.edu/misc/5745602



Collection identifier

Cornell:24 (Ockham)
4077228 (Voyager)

Title of collection

Andrew D. White Architectural Photographs Collection

Alternative title

Andrew Dickson White Collection of Architectural Photographs

Description of collection

The Andrew Dickson White Architectural Photographs document a wide ra
20th century architecture of Europe, the Middle East and the Americas, in
panoramas and habitats that have vanished due to wars and urban develop

Collection type

CollectionImage

Collection logo URL

http://rdc.library.cornell.edu/images/collection_icons/adwhite85px.jpg

Subject

Architecture -- Pictorial works (vocab = LCSH)
Archaeology -- Pictorial works (vocab = LCSH)
Architecture -- Details -- Pictorial works (vocab = LCSH)
Art and Architecture (vocab = none)

Language of collection

English, Afrikaans

```
<iesr:iesrDescription>
–
      <iesr:Collection>
<dc:identifier>Cornell:24</dc:identifier
>
–
      <dc:title>
Andrew D. White Architectural
Photographs Collection
</dc:title>
<dc:subject>Architecture -- Pictorial
works</dc:subject>
<dc:subject>Archaeology -- Pictorial
works</dc:subject>
<dc:subject>Architecture -- Details --
Pictorial works</dc:subject>
<dc:subject>Art and
Architecture</dc:subject>
</iesr:Collection>
</iesr:iesrDescription>
```

preservation repository

CUL-OAIS

a METS wrapper describes and transports the digital object

storing the wrapper

the
XMLtape

storing the datastreams

the ARCfile

how we will integrate these components?

oai-pmh

asking the right questions

Identify?

ListIdentifiers?

ListMetadataFormats?

ListSets?

ListRecords?

GetRecord?

Exposing descriptive metadata

OAI_DC for description

METS for preservation

federated search

the aDORe federation software presents a facade

discovery
and access

a simple data model for all repositories

a facade for any type of repository

benefits

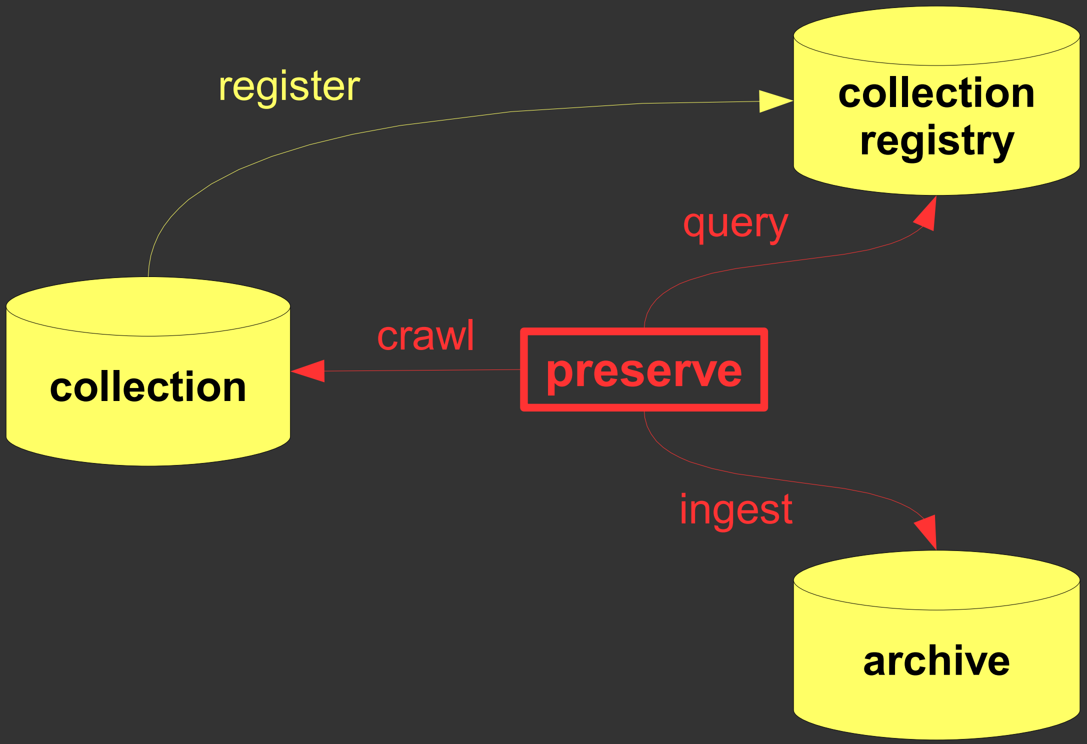
end-to-end stewardship

from creation

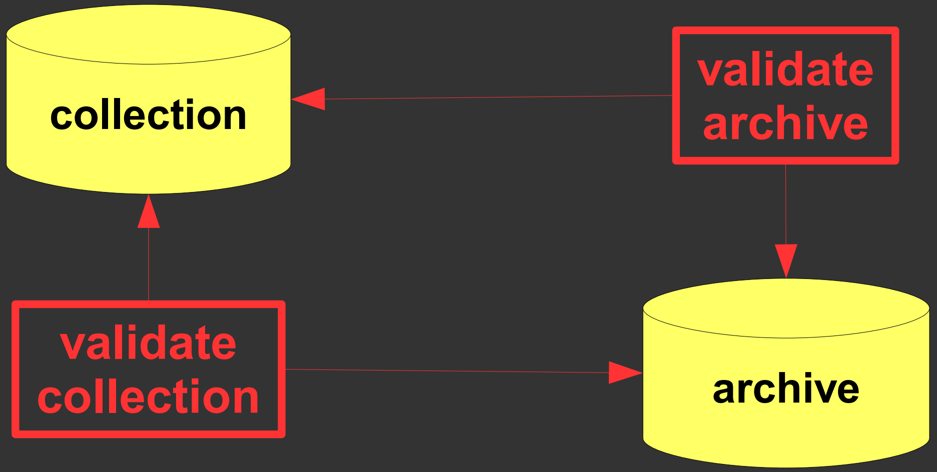
to preservation

to re-use

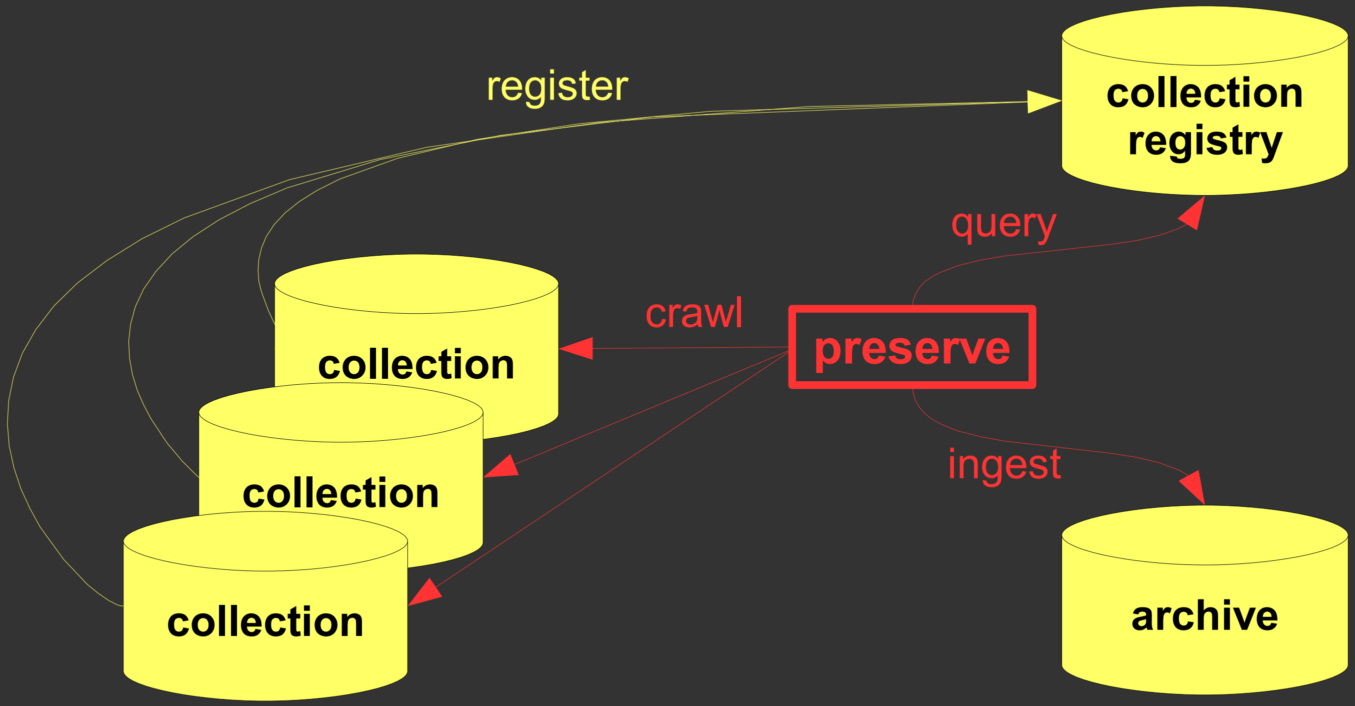
automatic preservation



self healing
archives and collections



a cohesive infrastructure creates economies of scale



maximize existing investments with minimal effort

collections become obsolete,
assets do not

collections become skinnable

assets become re-usable

lower costs
for better patron service

meet library strategic goal #1

> *“Expedite access to scholarly resources at the point and place of need.”*
>
> *~ Library Strategic Goals 2007-2010*

meet library strategic goal #2

> *“Provide cutting-edge facilities and services to support research, teaching, learning, and scholarly communication across disciplines.”*
>
> *~ Library Strategic Goals 2007-2010*

meet library strategic goal #3

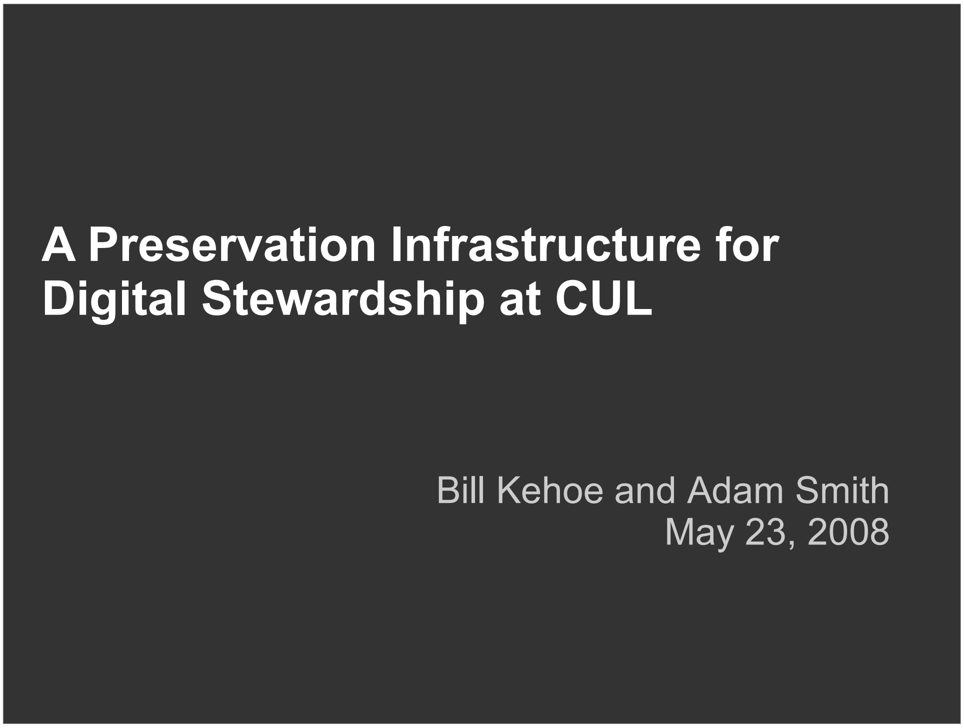
“Ensure stewardship of the University’s intellectual assets.”
~ Library Strategic Goals 2007-2010

the digital library
is not becoming less important

Cornell should
lead the way again

Now is the time to act

questions?



A Preservation Infrastructure for Digital Stewardship at CUL

Bill Kehoe and Adam Smith
May 23, 2008

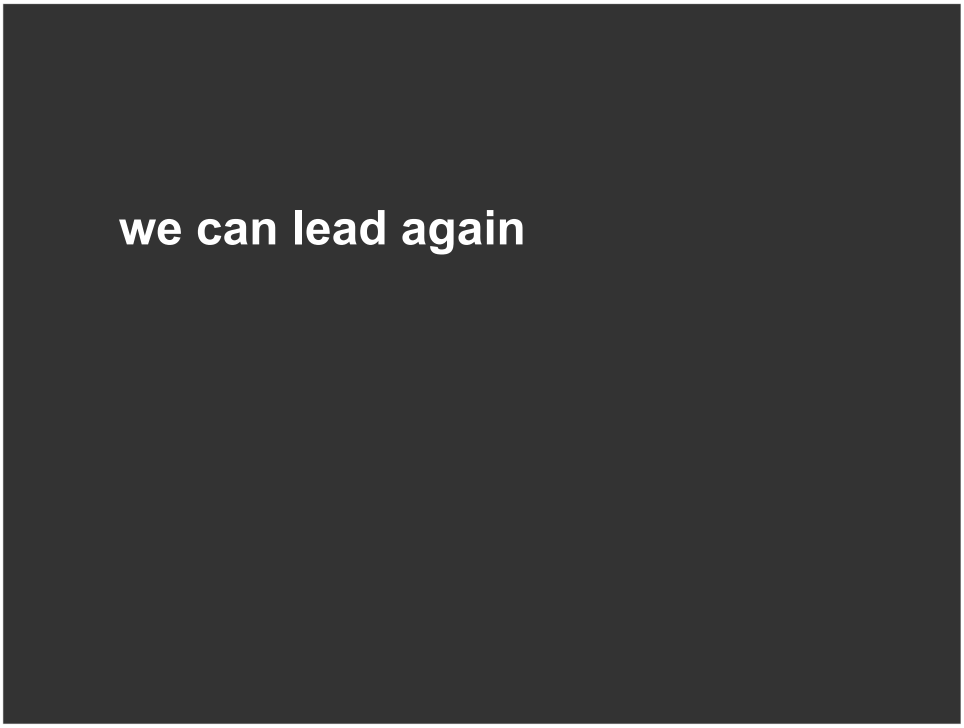
the library's legacy of leadership

CUL has been a pioneer of digital collections and services.

* but by focusing on innovation, we sacrificed developing a cohesive digital library infrastructure of common services that meet organizational goals.

* Now we are at a point where focus on infrastructure is urgently needed to:

o meet organizational goals, such as digital preservation
o fully realize past investments in our existing digital collections, and
o continue innovating in this space to better serve our patrons.



we can lead again

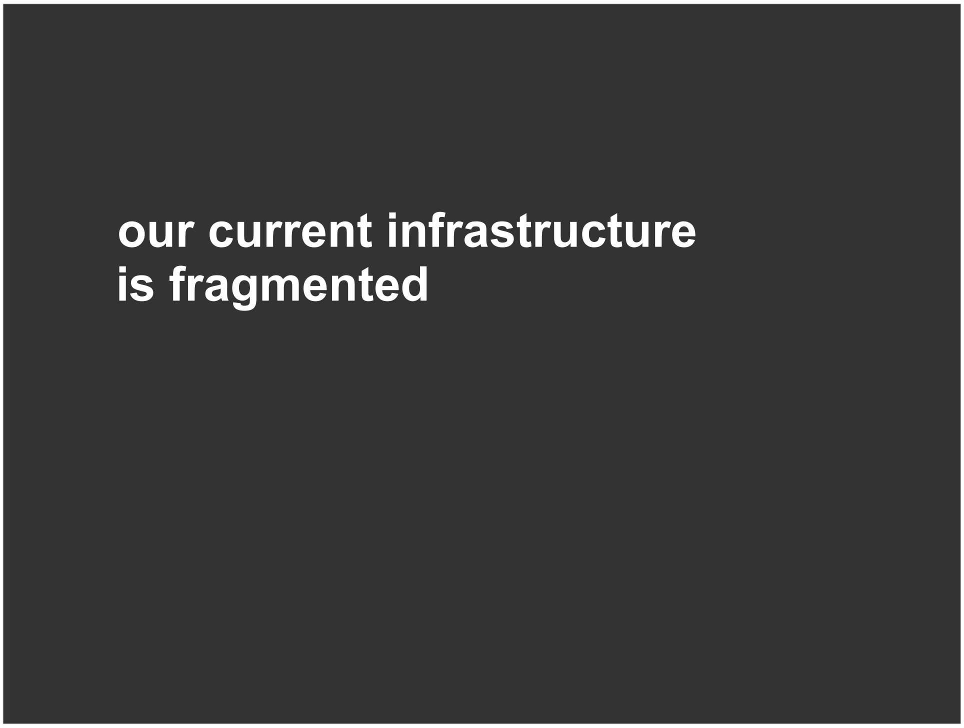
CUL can show the larger library community how to build this cohesive infrastructure...

...and we can begin by **focusing on the true digital assets** (such as image, monographs, etc.) contained within each collection and service instead of thinking in terms of the collections themselves, which are only the particular **delivery mechanisms** for those assets.

We can **integrate existing collections**

* long term, end-to-end stewardship of digital assets.
* allowing for re-use of assets
* more easily developing innovative services to patrons based on those assets

...and we can do the same **automatically** for assets in new collections.



our current infrastructure
is fragmented

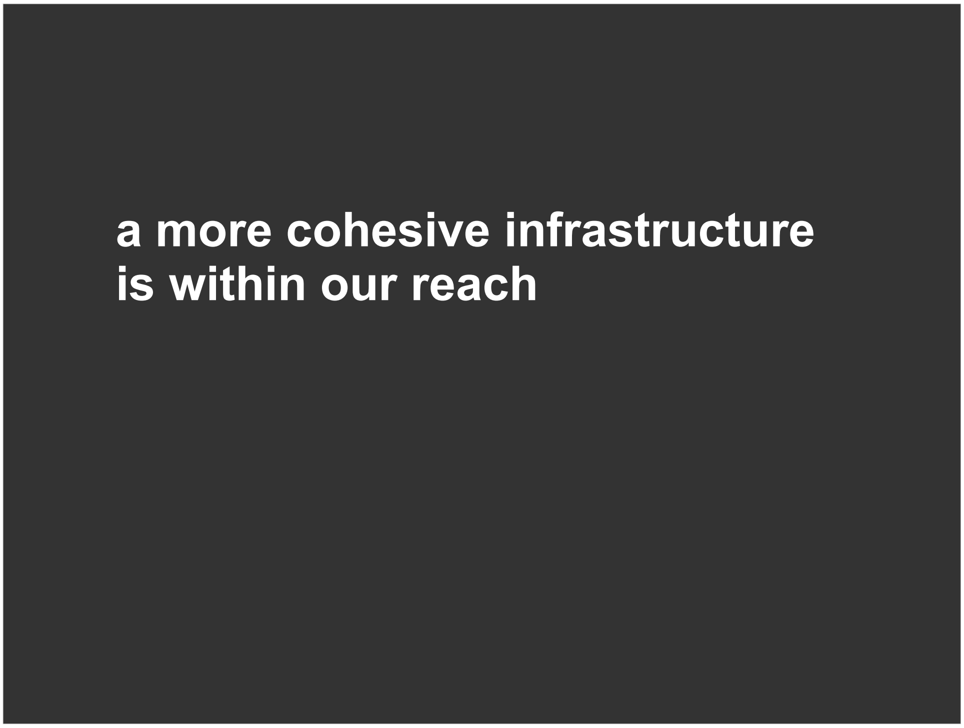
Project based grant funding--a key to our innovation--has created independent "**silos**" of assets not easily accessible outside their original collections/services.

We have **multiple, but finite platforms** for our collections and services that include:

* Dspace,
* DPubs,
* DLXS, and
* Fedora.

The answer is **not to reject grants**, and the innovation they promote, but to build a robust, cohesive infrastructure that those projects can be built upon.
...By easily participating in common services that meet organizational goals, such as long term digital preservation.

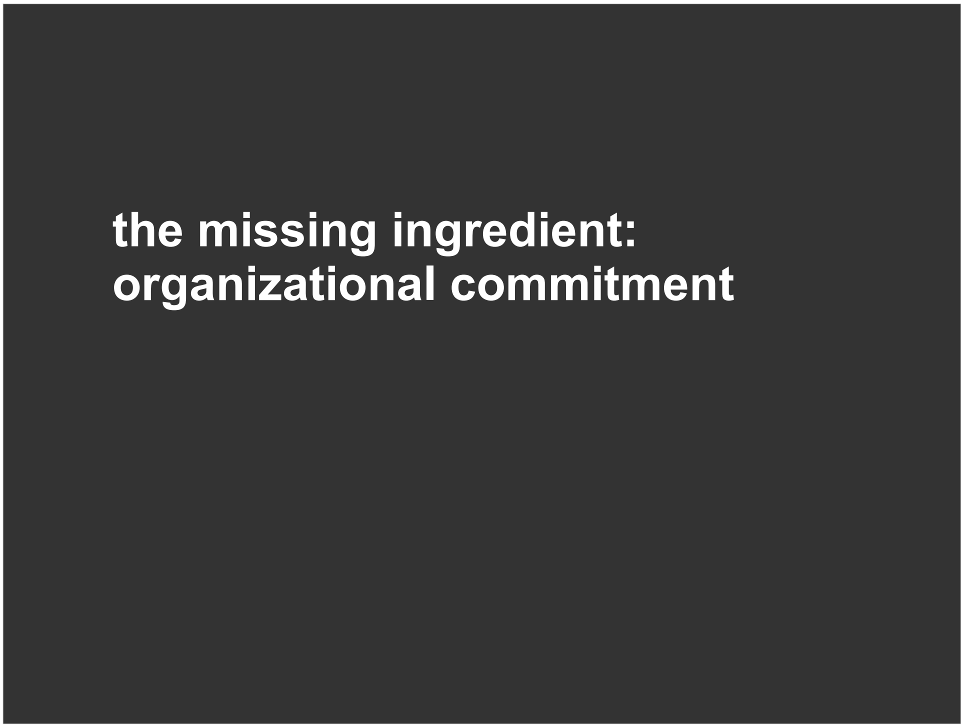
What we need is a **standard way** for these components to **discover** each other and the services they offer, along with standard protocols that allow **automated software to manage communications** between them.



a more cohesive infrastructure
is within our reach

Some of the necessary standards and components needed to build a cohesive infrastructure **already exist**.

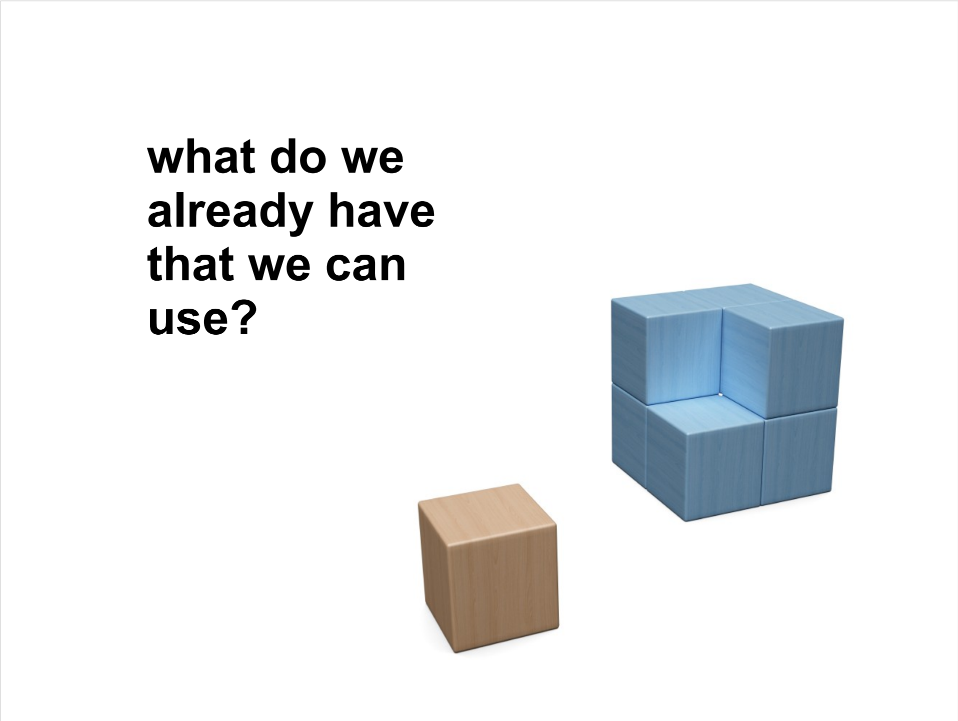
* the **collection registry** can be used for automated discovery of collections and the services they provide
* the **Handle Server** not only provides persistent identifier assignment and resolution, but can be used as a basic asset registry
* **OAI-PMH, along with the MathArc protocol**, provides a standard mechanism for harvesting both metadata and assets from collections
* **aDORe** is a massively scalable preservation repository; CUL is an early adopter and has helped influence its direction



the missing ingredient: organizational commitment

* So, we have many of the necessary tools already,
* We have the technical expertise to fully implement these components and get them working together.
* We have strong financial and mission based incentives
* What we lack is a clear organizational commitment to complete this work

CUL does not have a strong **culture of action**, possibly because of a culture of risk aversion
...but we will show that in the long run, not acting on this vision of the future of our digital collections will be the greatest **risk** of all.



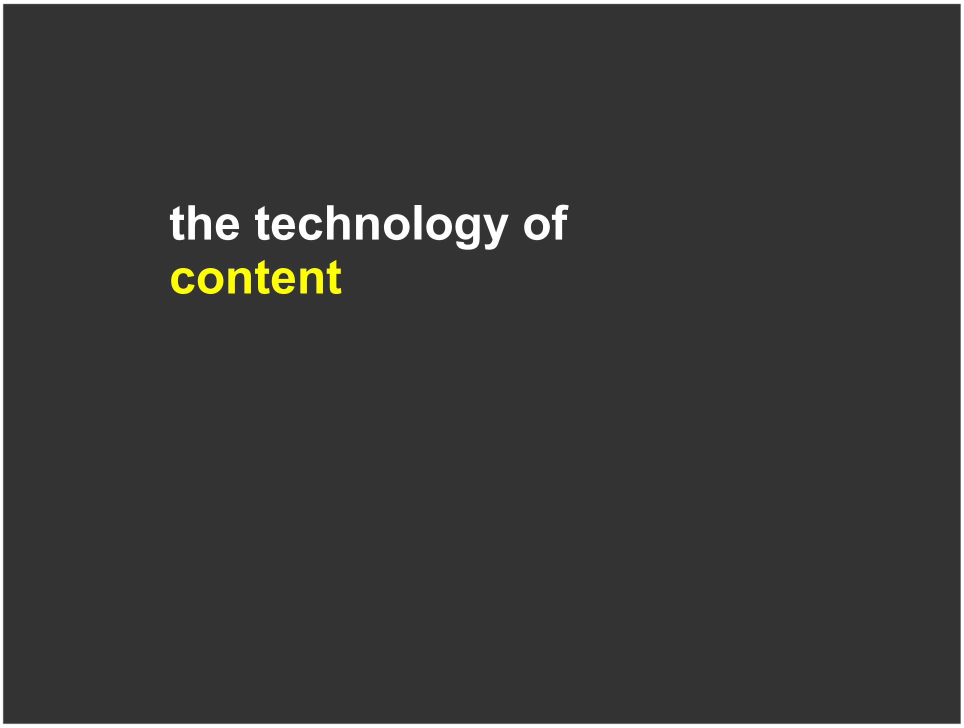
what do we already have that we can use?

•What do we already have that we can use?

•Our digital collections,
•We also have also developed and built **mechanisms for identifying and describing** the objects and the collections:
•We have a format-agnostic **digital preservation repository**, currently called CUL-OAIS, although not formally named..

collections

Even though we often think of them as resources, collections can also be thought of as technical components in the infrastructure.
The digital objects that make up the collections – digital articles, monographs, and images – are the real resources.
The collections are merely containers for the **real** resources.



the technology of
content

The technology of content
Complex digital objects include files of many formats. The display format is the one we see, but frequently we make the decision to preserver all the files that were components of the object.
Examples: PDF, HTML, postscript, TeX, jpeg, tiff images.
The metadata about object structure, description, technical characteristics, and provenance must be preserved.
Page structure, OCR, Dublin Core descriptions, NISO Z39.87 image metadata.
Structural and descriptive metadata about collections, too

the technology of
disclosure

Exposing content
The access systems into our collections, most usually a web browser, exposes content for **users**.
Collections that expose **content to machine interfaces**, **API's**, enable themselves to be used automatically by other systems, such as a preservation system.

identity
and
description

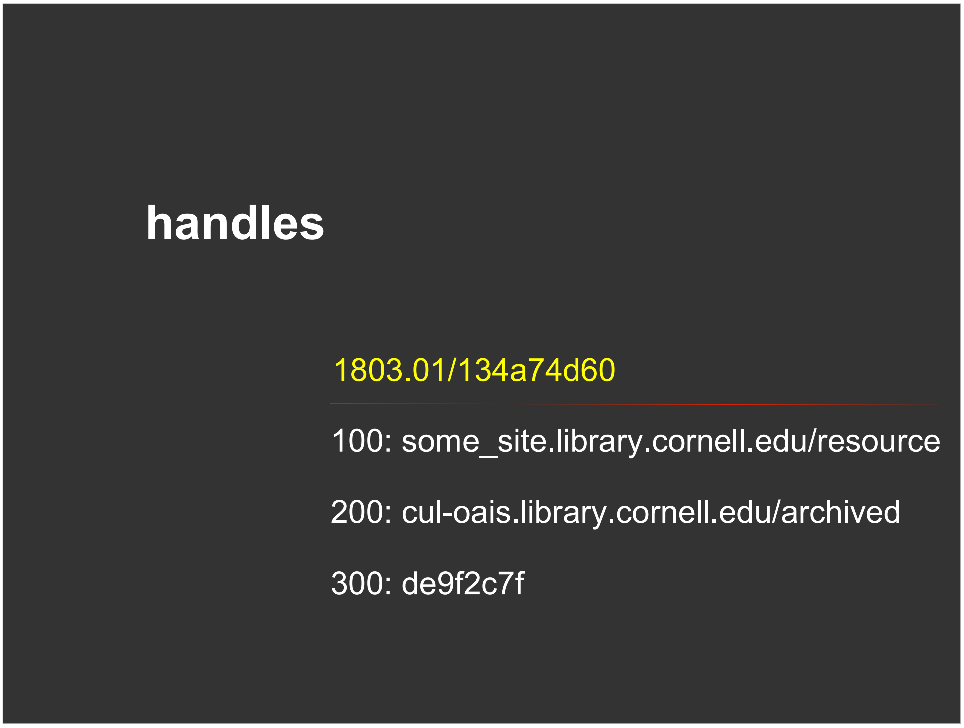
handles
and
collection registry

Identity and Description:
Handles and the Collection Registry

Identity and description are the ways we know what our resources are.

URL's are titles are the two most obvious ways we do this. Human-readable URLs sometimes lose their meanings if we change the structure or location of an object. Descriptions can change, too, over time.

The best way to minimize the disruption and cost of change in this case is to use persistent identifiers and registries of descriptive information., in CUL's case, our Handle Server and our Collection Registry.



handles

1803.01/134a74d60

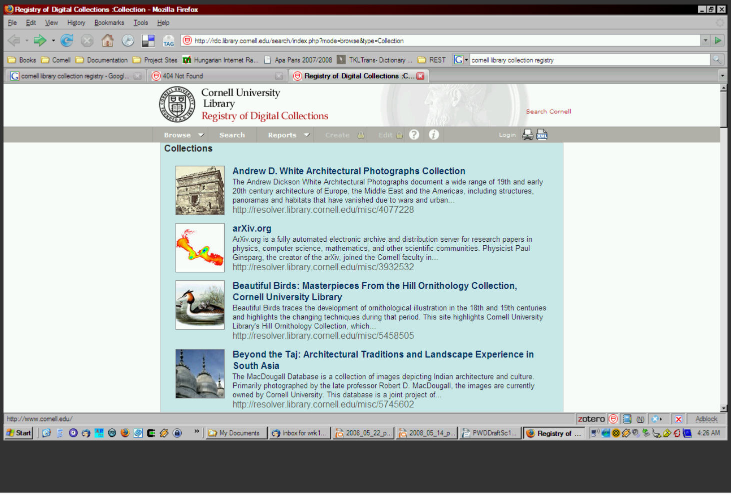
100: some_site.library.cornell.edu/resource

200: cul-oais.library.cornell.edu/archived

300: de9f2c7f

Handles
A Handle is an unchanging address to a container of information that includes the URL of an object.
The object pointed to can be a file, a complex object, or a collection of objects, or even a machine.

registry of digital collections



Registry of Digital Collections :Collection - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://rdc.library.cornell.edu/search/index.php?mode=browse&type=Collection
Books Cornell Documentation Project Sites Hungarian Internet Ra... Apa Paris 2007/2008 TKLTrans- Dictionary ... REST
cornell library collection registry
cornell library collection registry - Googl...
404 Not Found
Registry of Digital Collections :C...
Browse
Search
Reports
Create
Edit
Login
XML
Search Cornell
http://www.cornell.edu/
zotero
Adblock
Start
My Documents
Inbox for wrk1...
2008_05_22_p...
2008_05_14_p...
PWDDraftSc1...
Registry of ...
4:26 AM

Cornell University Library
Registry of Digital Collections

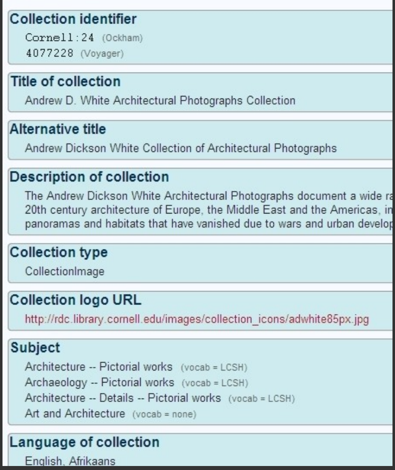
Collections

Andrew D. White Architectural Photographs Collection
The Andrew Dickson White Architectural Photographs document a wide range of 19th and early 20th century architecture of Europe, the Middle East and the Americas, including structures, panoramas and habitats that have vanished due to wars and urban...
http://resolver.library.cornell.edu/misc/4077228

arXiv.org
ArXiv.org is a fully automated electronic archive and distribution server for research papers in physics, computer science, mathematics, and other scientific communities. Physicist Paul Ginsparg, the creator of the arXiv, joined the Cornell faculty in...
http://resolver.library.cornell.edu/misc/3932532

Beautiful Birds: Masterpieces From the Hill Ornithology Collection, Cornell University Library
Beautiful Birds traces the development of ornithological illustration in the 18th and 19th centuries and highlights the changing techniques during that period. This site highlights Cornell University Library's Hill Ornithology Collection, which...
http://resolver.library.cornell.edu/misc/5458505

Beyond the Taj: Architectural Traditions and Landscape Experience in South Asia
The MacDougall Database is a collection of images depicting Indian architecture and culture. Primarily photographed by the late professor Robert D. MacDougall, the images are currently owned by Cornell University. This database is a joint project of...
http://resolver.library.cornell.edu/misc/5745602



Collection identifier
Cornell:24 (Ockham)
4077228 (Voyager)

Title of collection
Andrew D. White Architectural Photographs Collection

Alternative title
Andrew Dickson White Collection of Architectural Photographs

Description of collection
The Andrew Dickson White Architectural Photographs document a wide ra
20th century architecture of Europe, the Middle East and the Americas, in
panoramas and habitats that have vanished due to wars and urban develop

Collection type
CollectionImage

Collection logo URL
http://rdc.library.cornell.edu/images/collection_icons/adwhite85px.jpg

Subject
Architecture -- Pictorial works (vocab = LCSH)
Archaeology -- Pictorial works (vocab = LCSH)
Architecture -- Details -- Pictorial works (vocab = LCSH)
Art and Architecture (vocab = none)

Language of collection
English, Afrikaans

```
<iesr:iesrDescription>
–
        <iesr:Collection>
<dc:identifier>Cornell:24</dc:identifier
>
–
        <dc:title>
Andrew D. White Architectural
Photographs Collection
</dc:title>
<dc:subject>Architecture -- Pictorial
works</dc:subject>
<dc:subject>Archaeology -- Pictorial
works</dc:subject>
<dc:subject>Architecture -- Details --
Pictorial works</dc:subject>
<dc:subject>Art and
Architecture</dc:subject>
</iesr:Collection>
</iesr:iesrDescription>
```

preservation repository

CUL-OAIS

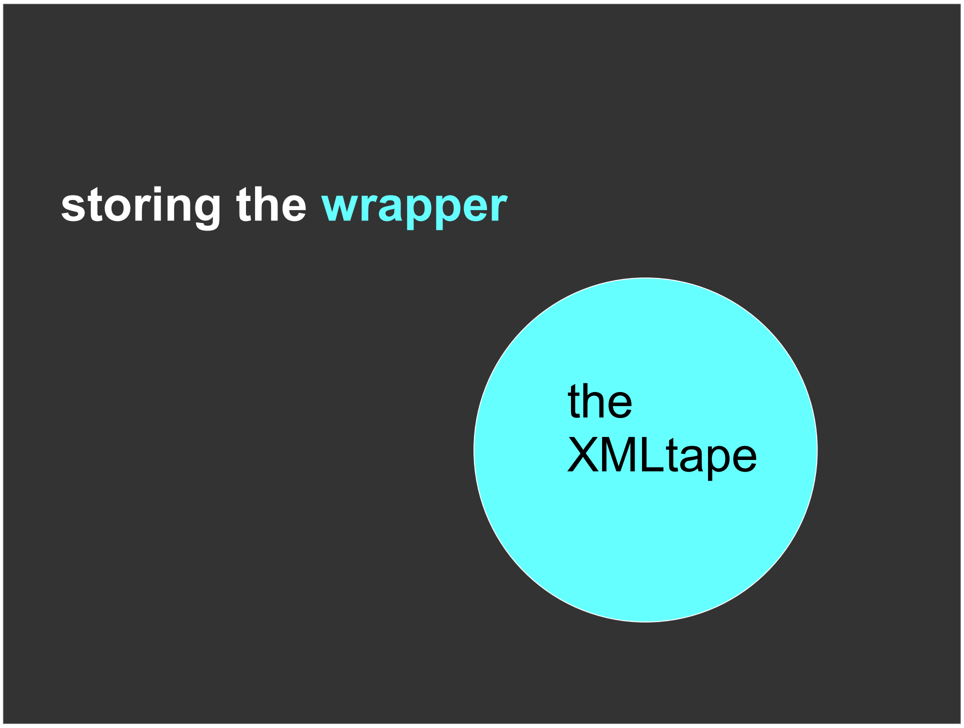
Lineage:
MathArc project
CUL Goal – A preservation system
LSDI Project

Built for Massively scalable archival storage

File-format agnostic

a METS **wrapper** describes and transports the digital object

- All the metadata goes here.
-
- For every file in the object.
-
- Metadata:
- Descriptive
- Technical
- Provenance
- Administrative
- Rights
- Structure

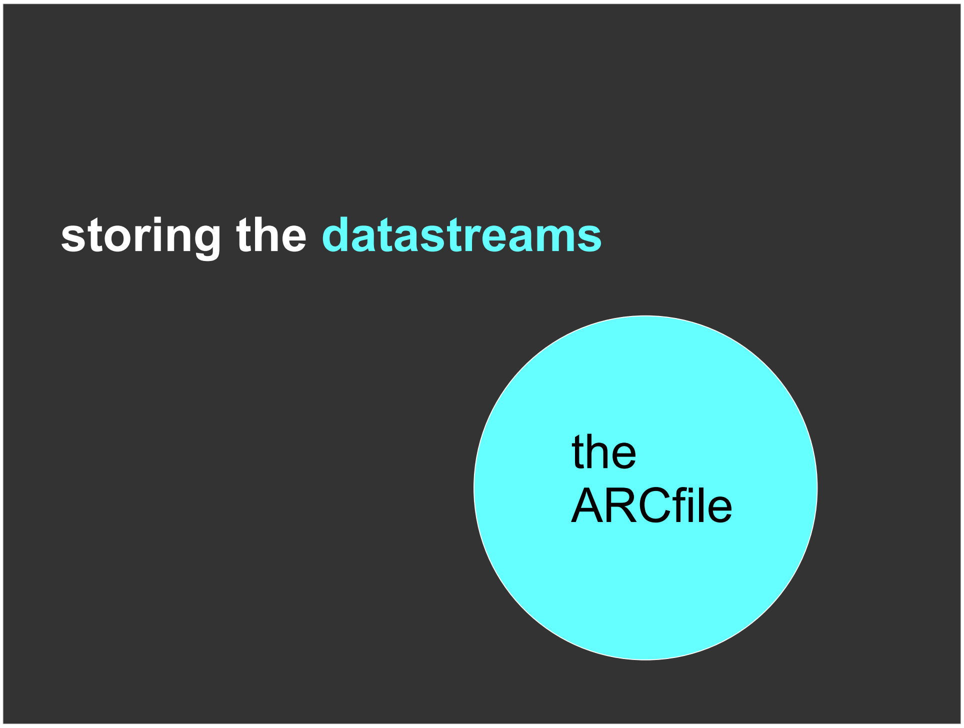


storing the wrapper

the XMLtape

All the metadata is stored in the oddly-named XMLtape file.

Metadata is indexed in a separate file.



storing the datastreams

the ARCfile

All component files, binary or text, are aggregated in the ARCfile.
Internet Archive format.

1600 files per book are combined into one data file and one index file.

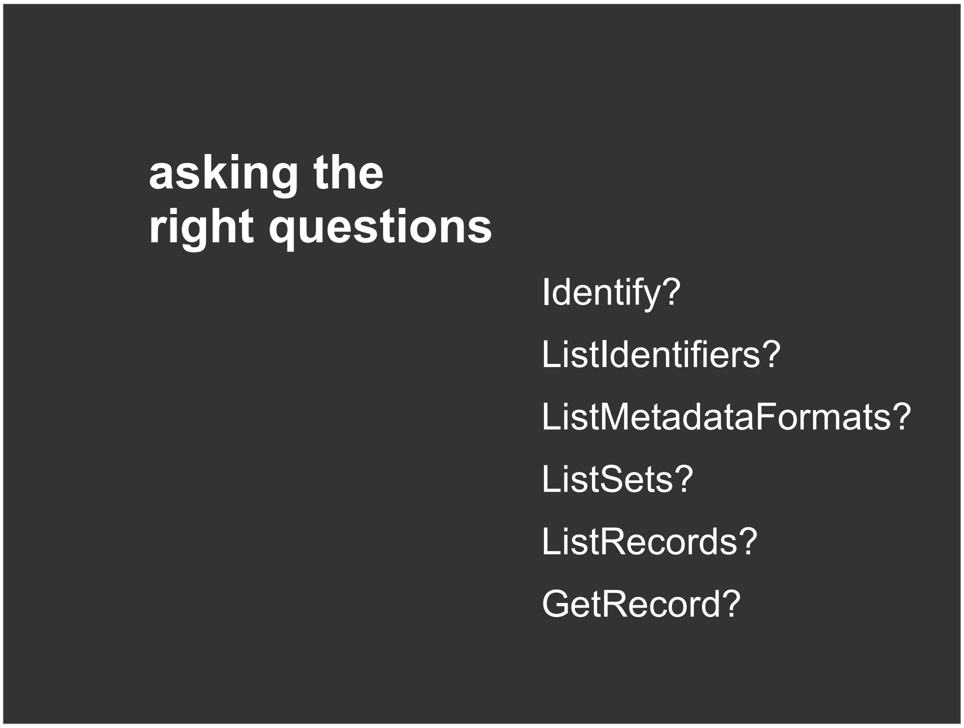
ARCfile and XMLtape are cross-indexed.

how we will integrate these components?

oai-pmh

Open Archives Initiative-Preservation Metadata Harvesting format

the protocol used for messaging, asking for information about resources and collections



asking the right questions

Identify?

ListIdentifiers?

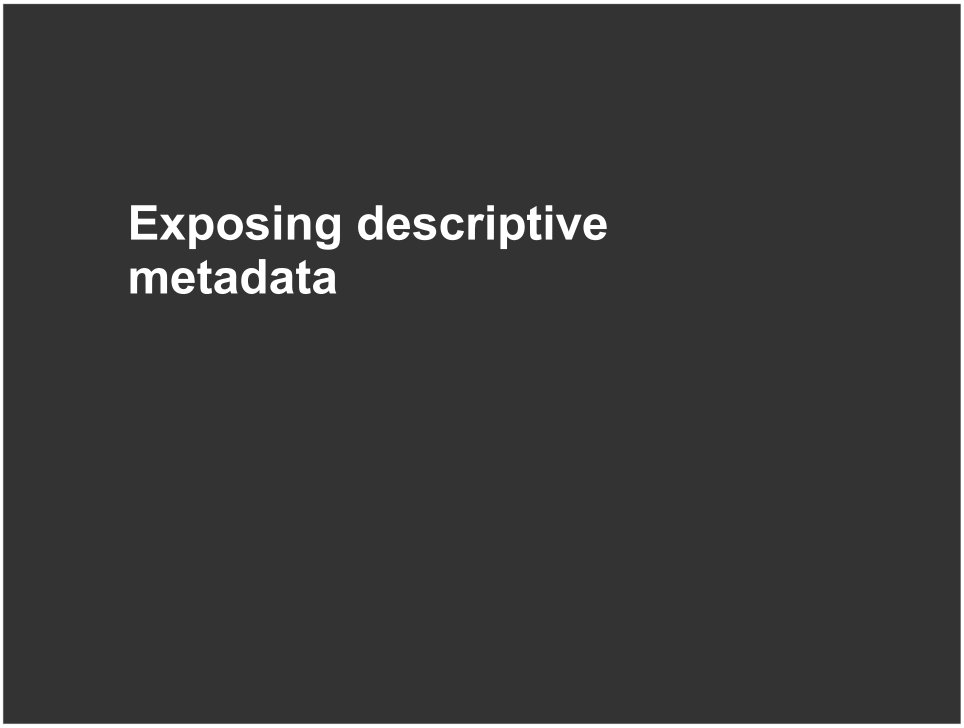
ListMetadataFormats?

ListSets?

ListRecords?

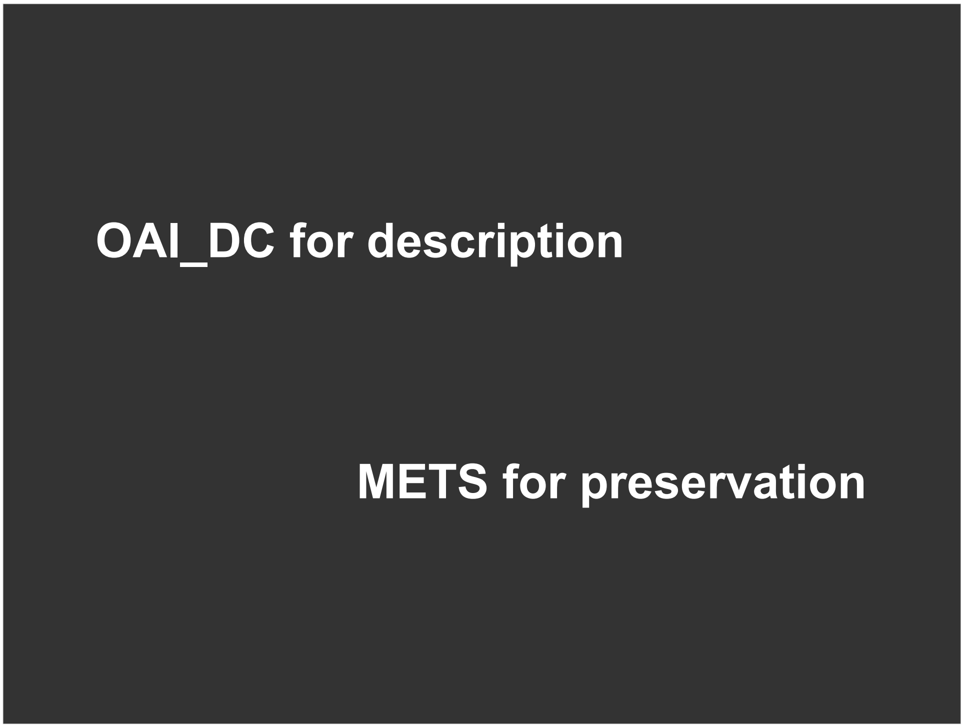
GetRecord?

These are the questions that can be asked.



Exposing descriptive metadata

The preservation repository can ask the Collection Registry for information about a particular collection; the returned metadata can be used to describe an archived object automatically, with no need for human input.

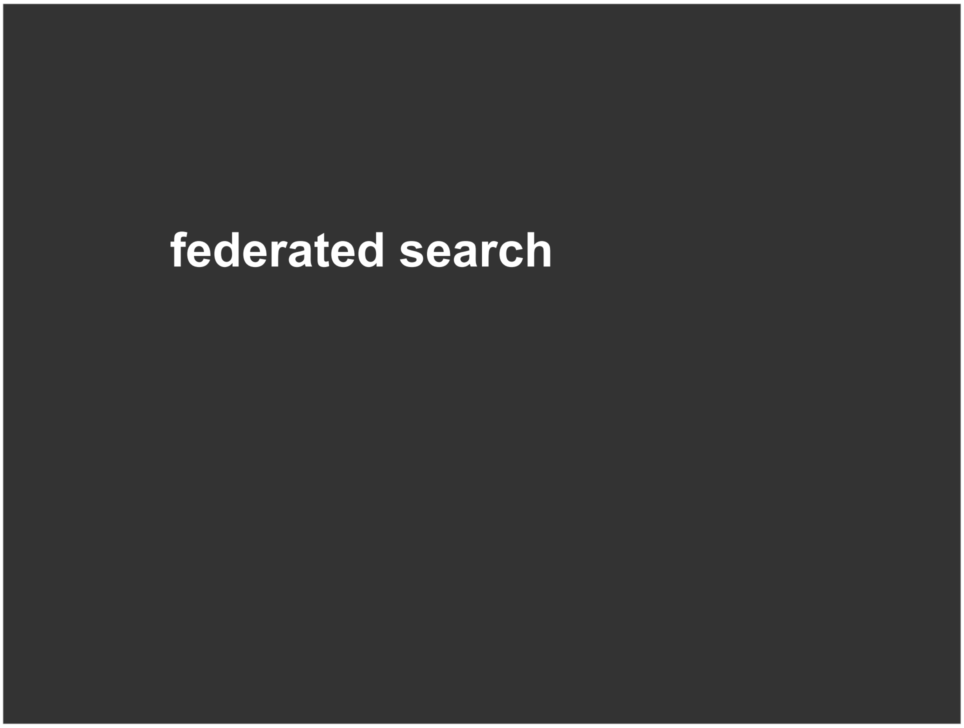


OAI_DC for description

METS for preservation

OAI Dublin Core is the default metadata format for OAI-PMH.
Best for simple descriptive metadata.

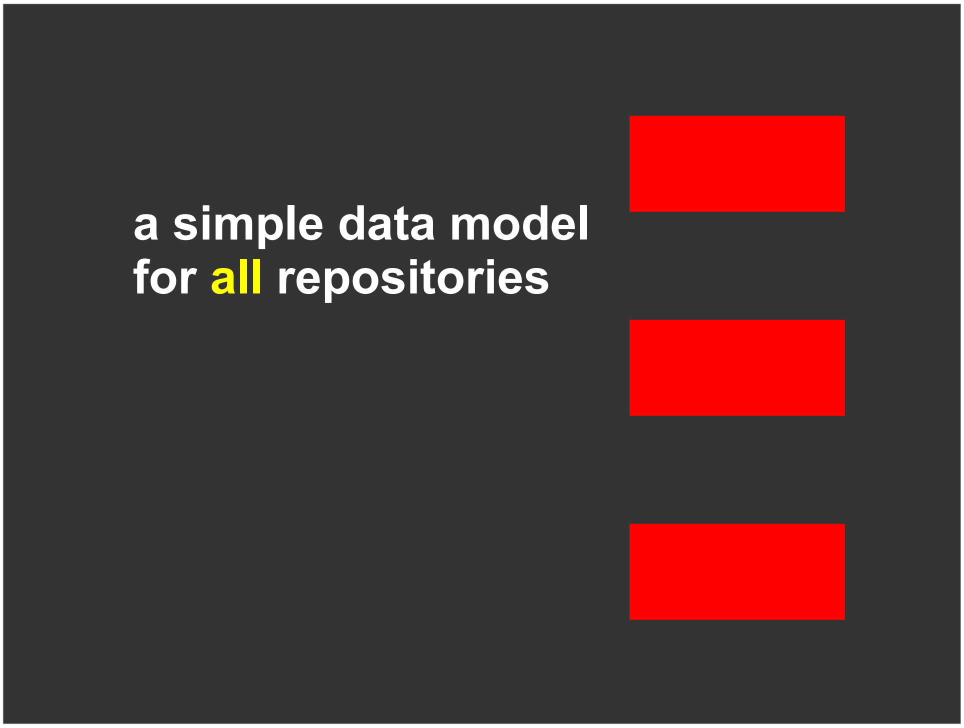
Our system uses METS to transport many kinds of metadata about a digital object or its components.



federated search

Federated search could unite CUL's many different collections and repositories.

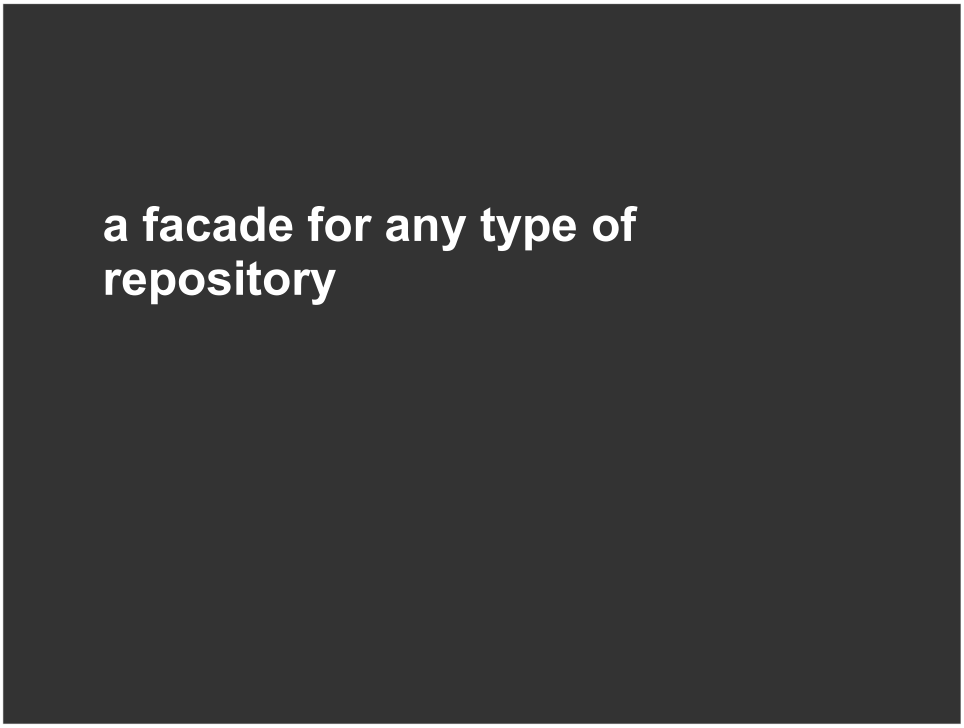
The new aDORe Federation software



a simple data model
for **all** repositories

Three objects only:

- Digital Objects – the whole resource
- Datastreams – the component files
- Surrogates – the metadata that describes and structures the object



a facade for any type of repository

The federation software from LANL can be used to search and browse *any* repository that understands the federation protocol.

The queries look similar to OAI-PMH requests.

Fedora and Dspace would have to be empowered to understand the protocol if federated search is to become a reality.

benefits

Integrating components achieves organization goals such as:

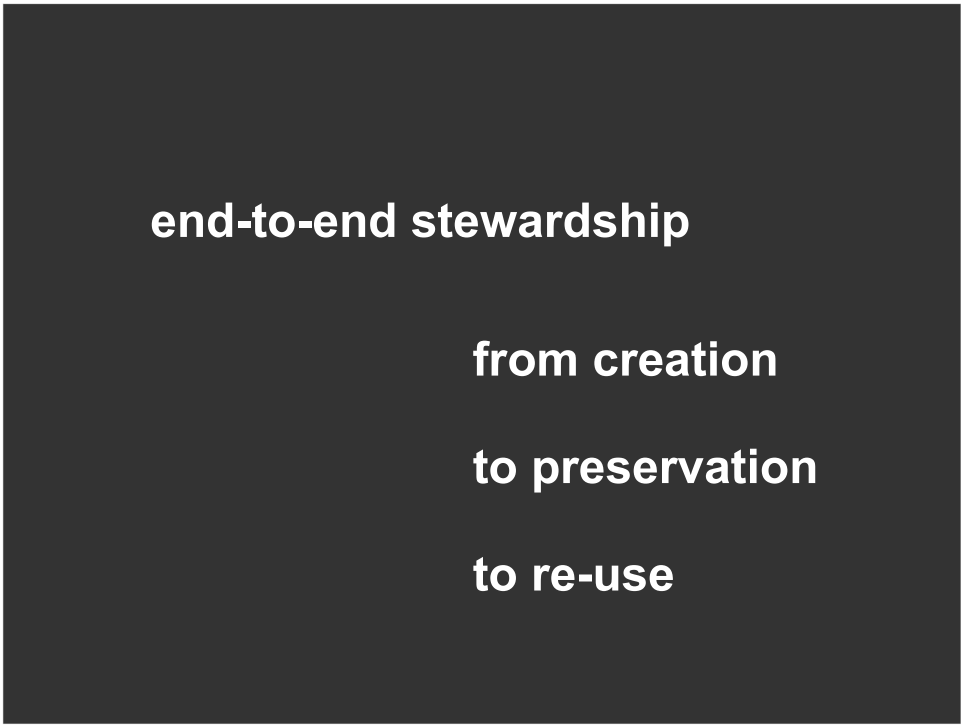
* long term preservation of assets
* persistent identification/resolution of assets
* automated integrity checking and "self healing" collections

* We know we should do these things, but they are typically
    * invisible
    * indirectly serve patrons

* So, the benefits aren't obvious, they don't have the same sense of urgency as access, its hard to get excited about them.

* But as we have seen, a cohesive infrastructure can also greatly facilitate Web 2.0 "**mash-ups**" or new services based on assets that can span multiple collections
* this directly impacts users by allowing us to more easily provide more innovative services.

* MONEY: a cohesive infrastructure maximizes the **re-use of assets**, which gives us a greater **return on our initial investment** in those assets
* We can **lower the cost** of new development and ongoing maintenance
* ... by lowering the cost of change and minimizing the amount of human intervention required to make the digital library function.

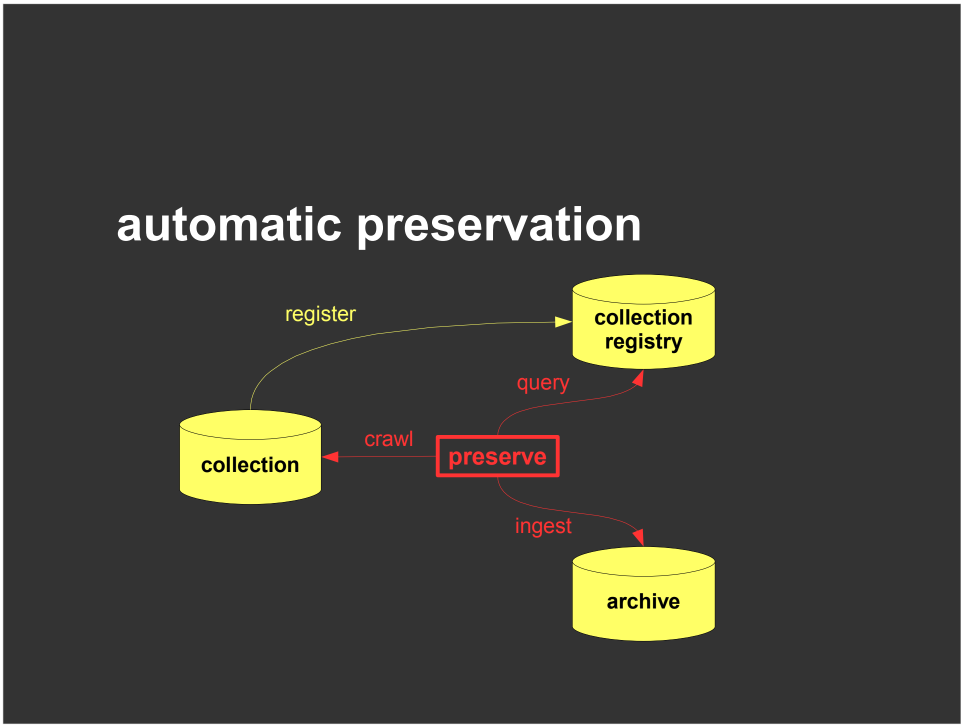


end-to-end stewardship

from creation

to preservation

to re-use



automatic preservation

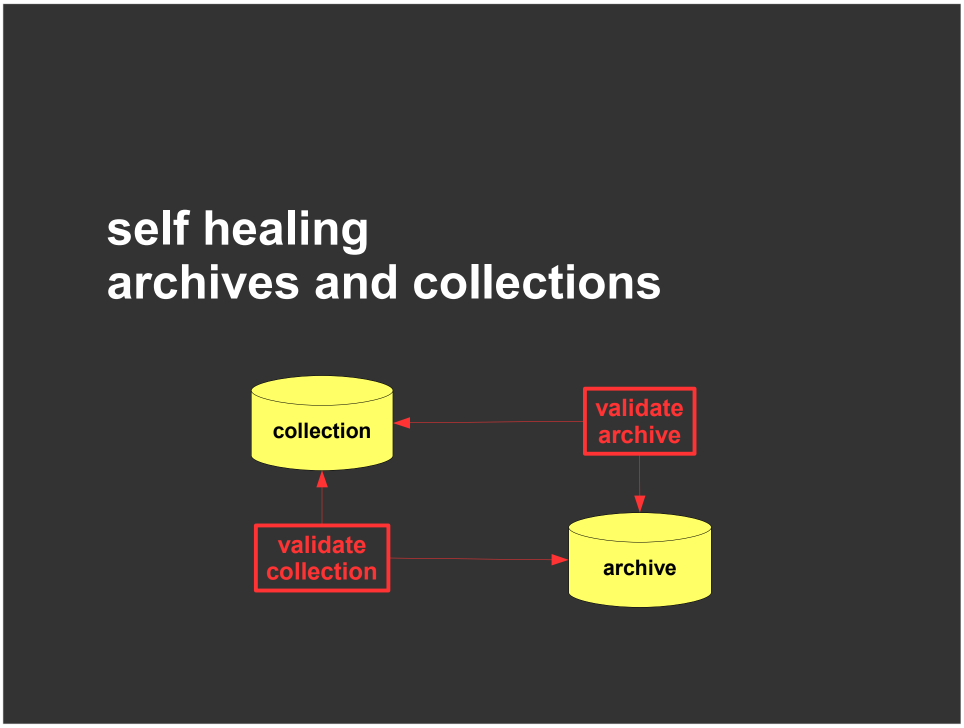
Any collection **exposing an OAI-PMH service** can be configured to be automatically preserved by implementing what we internally call the "MathArc protocol"

The core of the **MathArc protocol** is the use of METS and Premis metadata along with DC on top of OAI-PMH:

* so that both metadata and assets can be harvested, preserved in an archive,

* information about rights, level of commitment, provenance, versioning and more are available for software to make sophisticated decisions about what to do with the assets.

A preservation subsystem can query the collection registry for available collections and the address of their OAI-PMH interfaces, **automatically** starting the preservation process with minimal impact on the collection itself.



self healing
archives and collections

collection
validate archive
validate collection
archive

The MathArc protocol is robust enough to be used for other digital library functions.

* Assets can be continuously harvested from the preservation repository itself and validated to check the integrity of the archival copies.
* A failure would prompt a just in time update of that asset from the originating collection.

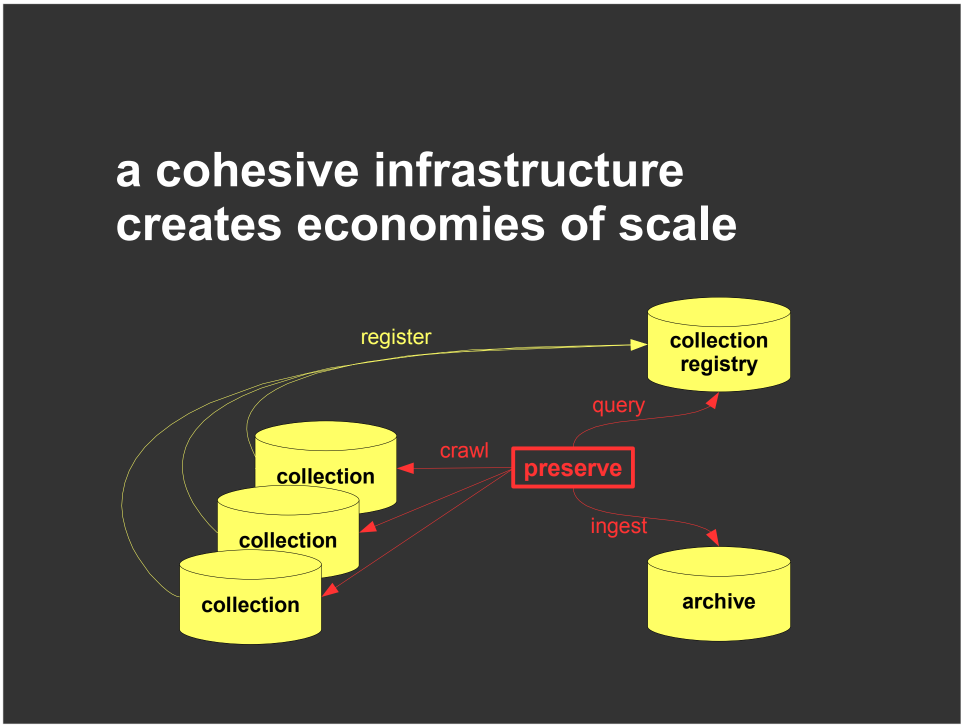
It can also work the other way...

* A similar service, running continuously in the background, can check the validity of assets in the collection to detect corruption in the access copy of an asset
* A failure would prompt a just in time recovery of the asset from the preservation archive.

The preservation archive is **not for passive storage** to be accessed only in the distant future, if ever,

It is an **active participant** in the infrastructure, directly and dramatically impact the user experience.

We preserve for the present as much as for the future.



a cohesive infrastructure creates economies of scale

Because of these standardized communication protocols between components, we can build such well defined, and automated, processes.
...and, because we use a **finite number of collection platforms**, it is cost effective to have them participate in this infrastructure and take advantage of these processes.

In terms of **human intervention**, there is little to no cost to add more collections into this infrastructure.

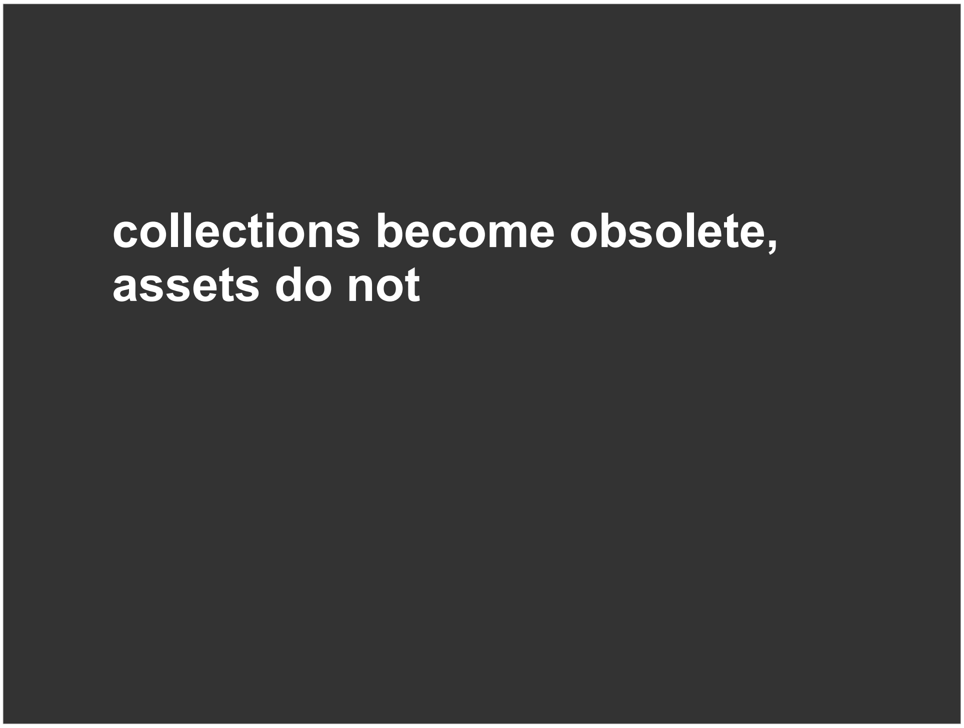
The ongoing execution of these processes is similar in many respects to **search engine robots**, except with much more intelligence and more capabilities.

**maximize existing investments
with minimal effort**

We **spend** a lot of money, time and human capital digitizing materials and creating digital services...

...we would like to stop treating that as a **cost**, and view it instead as an **investment** both by ensuring the existence of those assets long term and using those assets to their fullest now.

This means not only preserving assets, but giving asses **first class status** independent of any particular service that a patron uses to access it.
...as a result, assets become less costly to **re-purpose** in other collections and services.



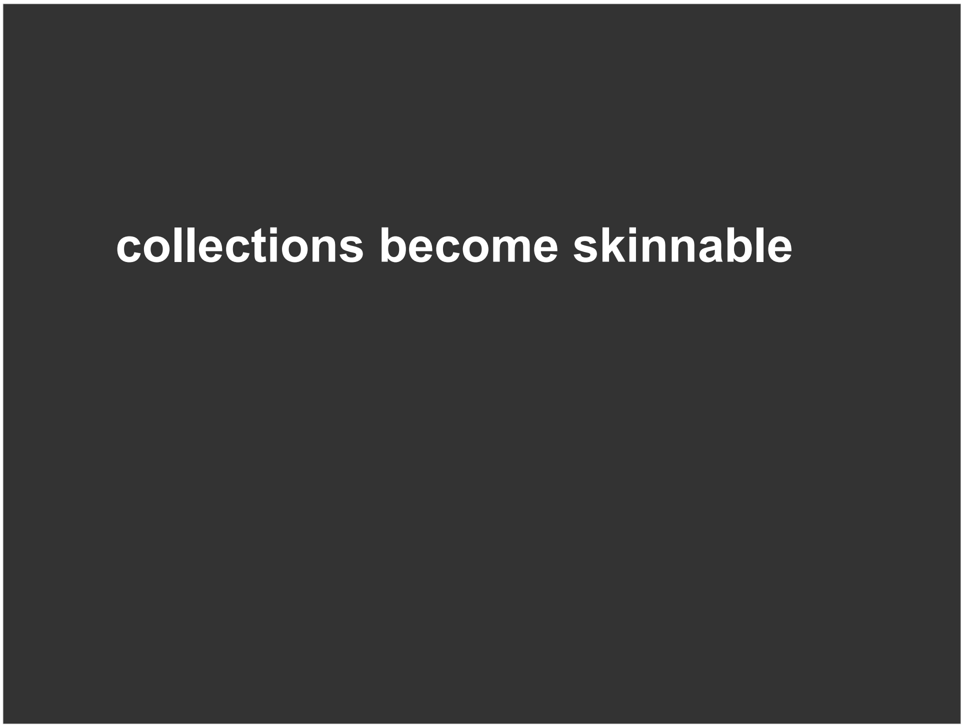
collections become obsolete,
assets do not

Every digital collection and service eventually becomes **obsolete** as it becomes **too expensive** to update the collection in the face of advancing technology.

If the collection can't be supported or migrated, it can be "**killed**" without a significant development effort to preserve the underlying assets.

Preservation is part of the infrastructure, the collection gets it for free simply by registering its OAI-PMH interface with the collection registry.

The assets will be available for use in the future.



collections become skinnable

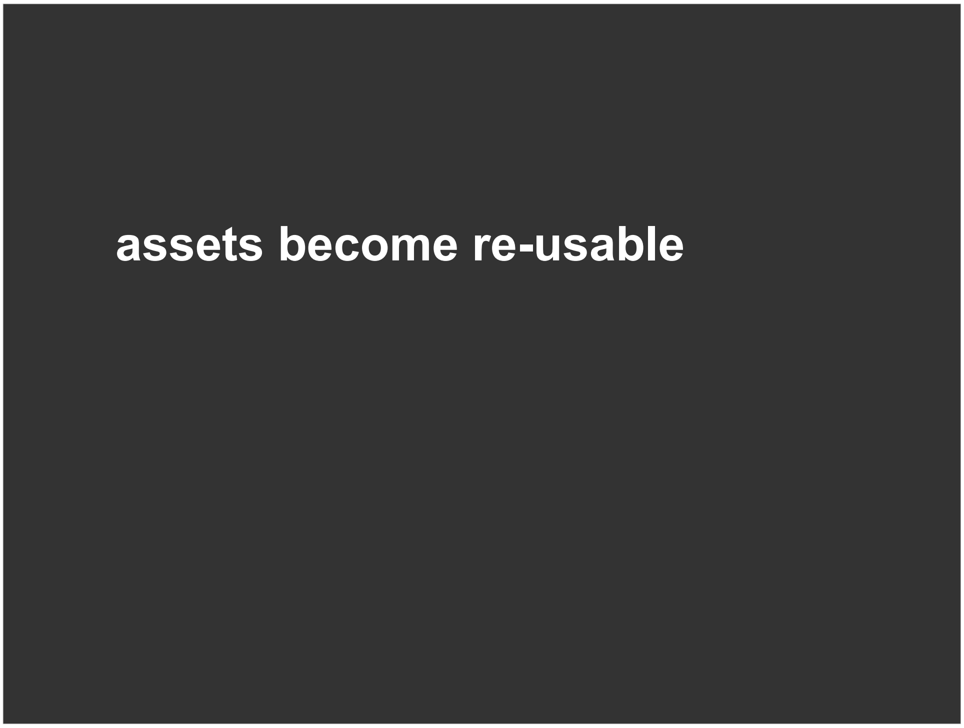
...assets within collections can be reused when the original collection is **migrated**.

Collection migrations cost less because decoupled assets can be migrated automatically.

They are not dependent upon their current delivery mechanisms. and migration amounts to putting a new face on the assets.

So, as a collection or service nears its end of life, development of a next generation service can occur in **parallel** because of uniform access to data in the repository.

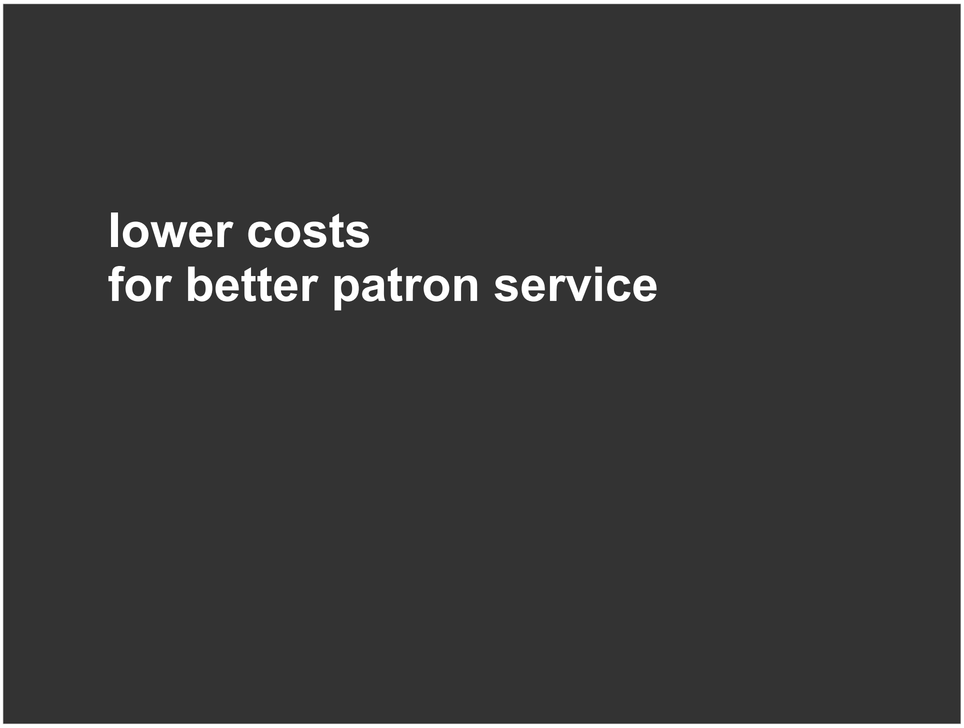
This lowers the cost of developing next generation services and further supports **better policies** around obsolete services.



assets become re-usable

Once assets are decoupled from their original collections, they are **available to new collections** and services at any time. We don't have to wait for the original collection to become obsolete before using assets in new ways.

And because assets from multiple collections will be stored in, and can be queried and harvested from, the same repository, **related assets from diverse collections** can be brought together in **new ways not originally envisioned** while implementing the original collection.



lower costs
for better patron service

Collections implemented on a common, cohesive architecture get **common services** like PIDs, preservation, integrity checking and more, **virtually for free**.

The **costs** of developing and maintaining collections and services is lowered and **creating truly innovative services** becomes more practical.

Lets look at more specific examples of how this infrastructure will result in better patron services.

meet library strategic goal #1

“Expedite access to scholarly resources at the point and place of need.”

~ *Library Strategic Goals 2007-2010*

For example, if we harvest and preserve images from collections, we can then **filter** and push appropriate images out to a photo sharing site like **Flickr**, one of the more popular social networking sites frequented by students and young professionals.

Based on parsing **copyright** restrictions, and applying appropriate rules, we may push out images that have been automatically resized, **downsampled** and watermarked.

We can link to related images in Flickr and link back to the originating collection, **driving traffic from Flickr back to CUL**.

Tags, comments and other Web 2.0 user generated content can be harvested from sites like Flickr and updated in our repository.

Essentially, Flickr has developed an interface for us for free...
...and we are populating it automatically.

meet library strategic goal #2

> *“Provide cutting-edge facilities and services to support research, teaching, learning, and scholarly communication across disciplines.”*
> *~ Library Strategic Goals 2007-2010*

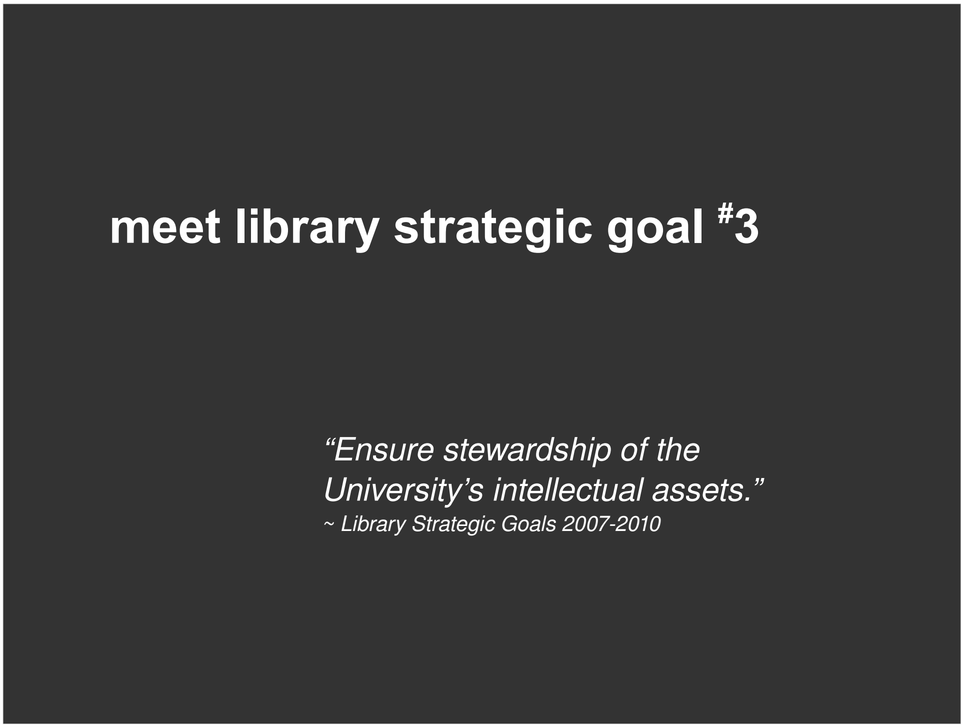
Harvested assets can be filtered and selected from diverse collections and brought together in new services.

We can think of obvious opportunities:

* Samuel J. May Anti Slavery collection
* Friend of Man abolishionist newspaper
* Digitized copy of Gettysburg address

But once we have automated means of storing and managing assets, new connections will **emerge that we couldn't have imagined.**

And the software may greatly help us see those connections by finding them automatically.



meet library strategic goal #3

> *“Ensure stewardship of the University’s intellectual assets.”*
> *~ Library Strategic Goals 2007-2010*

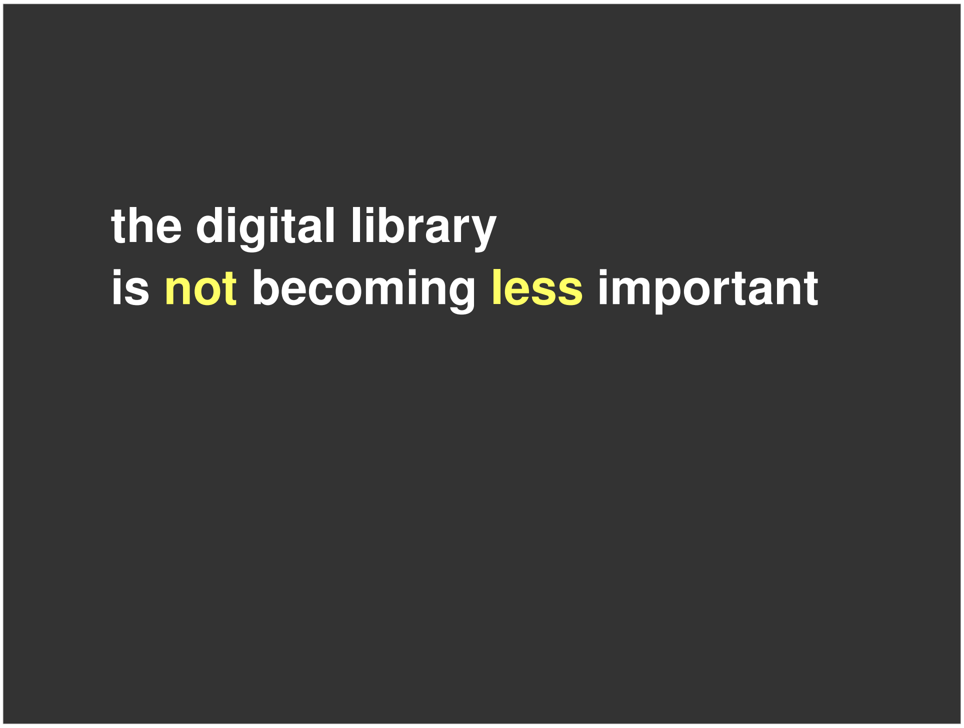
Preservation is not just something we do to benefit people years from now...
... its something we do to benefit us, now

Whole collections and individual assets in collections can be corrupted, lost.

Continuous, **automatic integrity checking** can find problems before patrons do.

Automatic recovery can fix problems before a patron encounters them.

All of this can happen with a **very low cost to individual collections**.

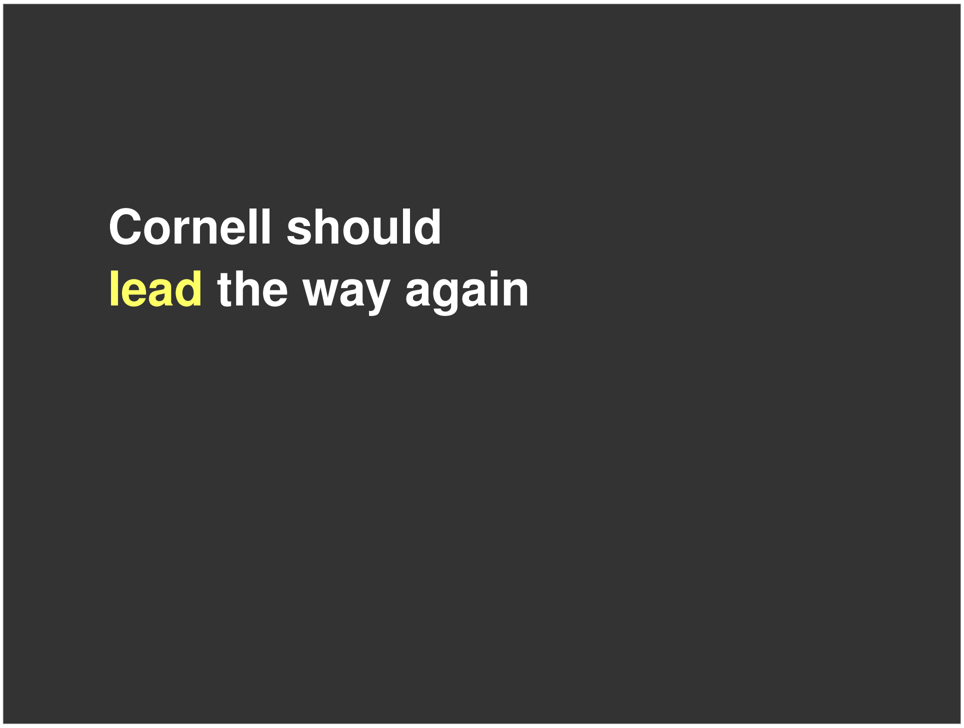


the digital library
is not becoming less important

...yet we continue to develop collections and services largely in **isolation**, not getting the most return from our investment in that development.

This approach once allowed us to be leaders in digital library development, it is now too costly and **unsustainable**.

Innovation now lies as much in building a sustainable infrastructure for digital services as in building the services themselves.

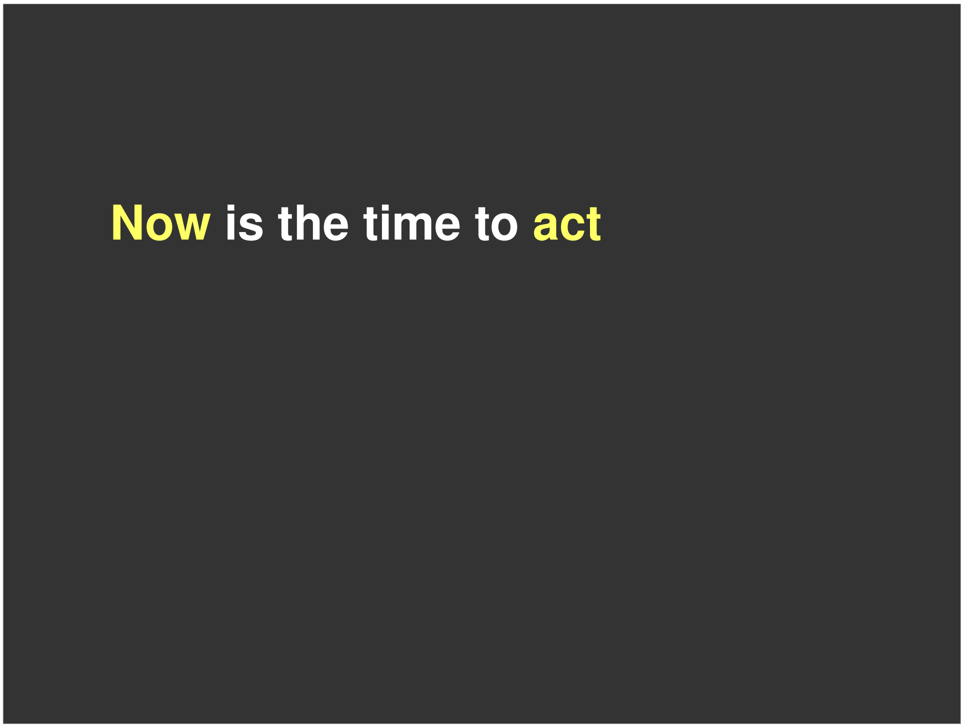


Cornell should
lead the way again

By creating a cohesive digital library infrastructure that is focussed on assets instead of collections,

* we can easily address organizational goals, such as digital preservation
* we can get more from our investment in assets by facilitating their reuse in new projects
* we can lower the cost of developing new services with these assets and encourage innovation

We can accomplish this using standard components and protocols, some of which are already partially implemented.



Now is the time to act

We have the tools and expertise to realize this vision and we can start NOW.

Yet despite the fact that our vision directly addresses 3 of the 6 library strategic goals, we lack a clear organizational commitment to complete this work.

We hope we've shown that these improvements are necessary for the future of our digital library development and that you will join us in creating that future.

questions?