VIRTUAL LIBRARY COLLECTION WITH LINKED DATA

E. Lynette Rayle Linked Data for Libraries Cornell University



February 23,2015

MOTIVATION FOR VIRTUAL COLLECTIONS

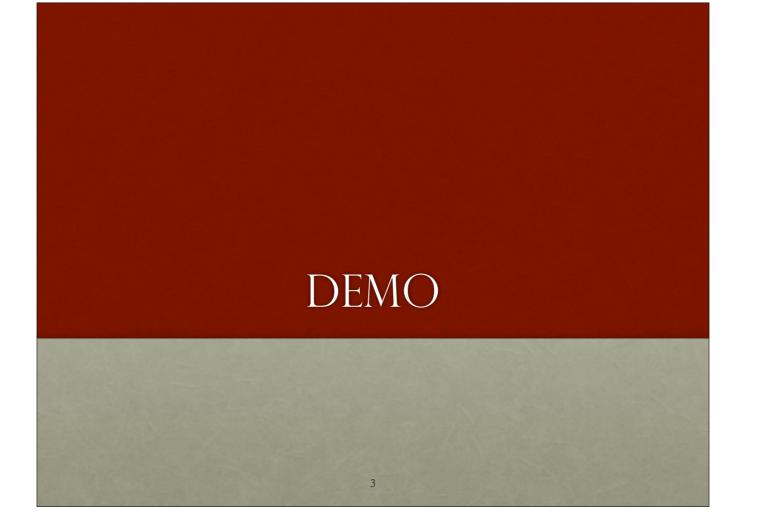
- Large collections for the library community
 - Virtual Libraries
- Small collections defined by users or small groups
 - Course reading list
 - Research reference material
 - Personal organization of resources

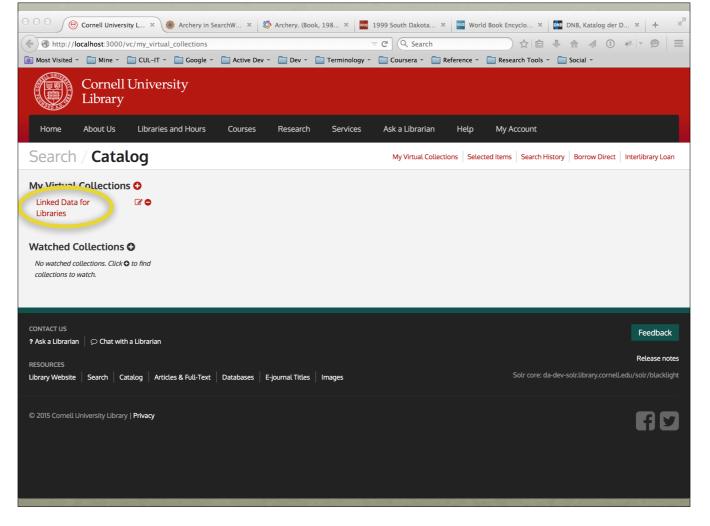
Our primary motivation is to create a system for managing large scale library collections with hundreds of thousands of resources in the collection. We want the creation tools and exploration to be integrated with the library Access & Discovery system. We want it to be able to draw resources from multiple institutions and use linked data to augment our metadata. We want to be able to share the collections we build through our internal websites and also to make them available through linked data queries.

2

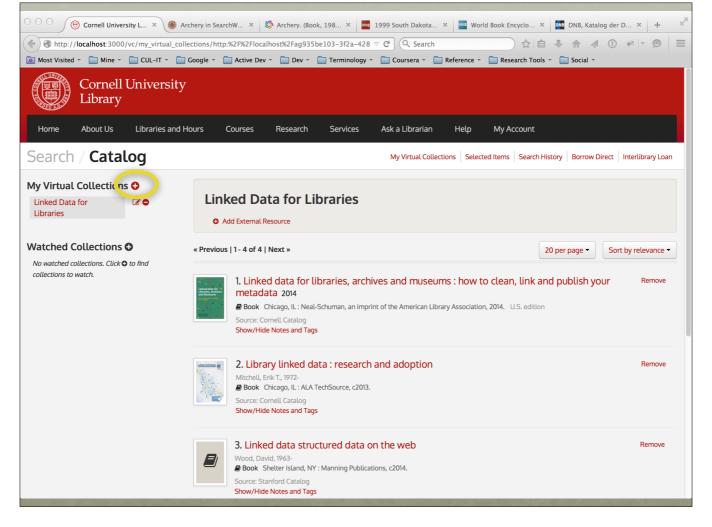
We want to be able to present users with options like 'you placed these resources in your collection, those same resources exist in another collection that may be of interest to you'.

In the process of solving these challenges at scale for the library community, we will have put in place the **infrastructure** that empowers individual users to create collections as well. I list a few possibilities for users and small collaborative groups. There are any number of other possibilities that users will come up with once the tools are in their hands.





From My Virtual Collection page, view the items in Linked Data for Libraries virtual collection.

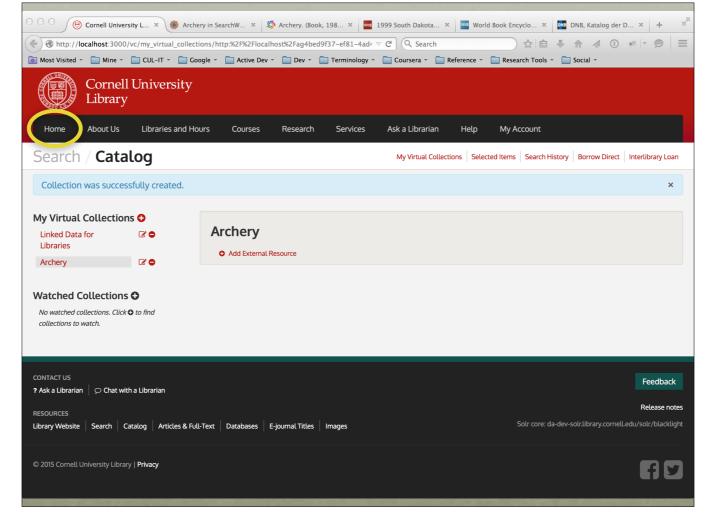


Click + beside My Virtual Collections header in left side menu to create a new virtual collection.

| OOO (Cornell University L × 🛞 Archer | y in SearchW × 🦃 Archery. (Book, 198) | | | | M ₂₁ |
|--|---|-----------------------------------|----------------|--------------------------------|-----------------|
| Attp://localhost:3000/vc/my_virtual_collection | ins | ▼ C Q Search |)☆ 自 ♣ | ^ | Ξ |
| Most Visited - Mine - CUL-IT - Goo | gle 👻 🧰 Active Dev 👻 📄 Dev 👻 📄 Termin | nology 🔻 📄 Coursera 👻 📄 Reference | Research Tools | Social - | |
| Cornell University Library Home About Us Libraries and Hou | | | | | |
| Search / Catalog | | My Virtual Collections S | | Borrow Direct Interlibrary Loa | an |
| My Virtual Collections Linked Data for Libraries Watched Collections No watched collections. Click to find collections to watch. | New Collection Title Archery Description Techniques and history of archery. | | | | |
| CONTACT US ? Ask a Librarian 🛛 📿 Chat with a Librarian RESOURCES Library Website Search Catalog Articles & Full- | | Close Save char | _ | | |
| © 2015 Cornell University Library Privacy | | | | | |

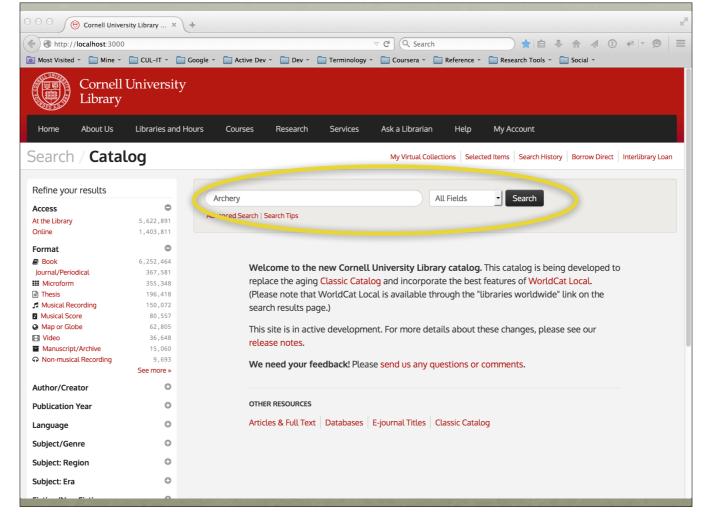
Set metadata for the new collection on Archery.

Each new collection is saved as an aggregation in the triplestore using triples as defined in the ORE ontology.

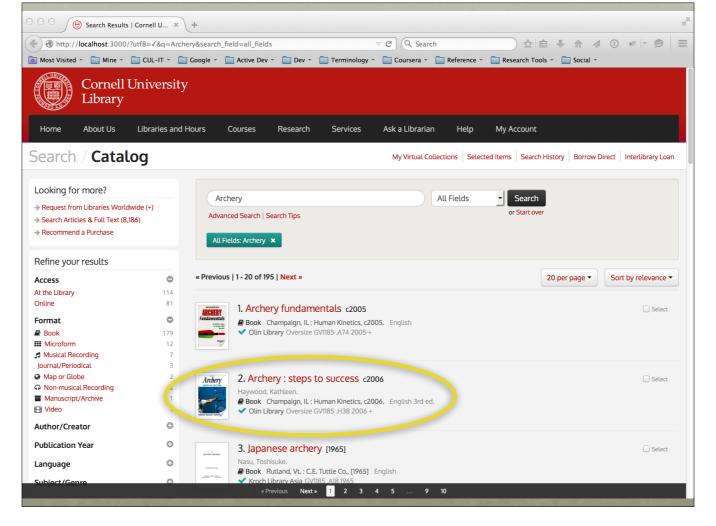


New Archery collection is selected and has no items.

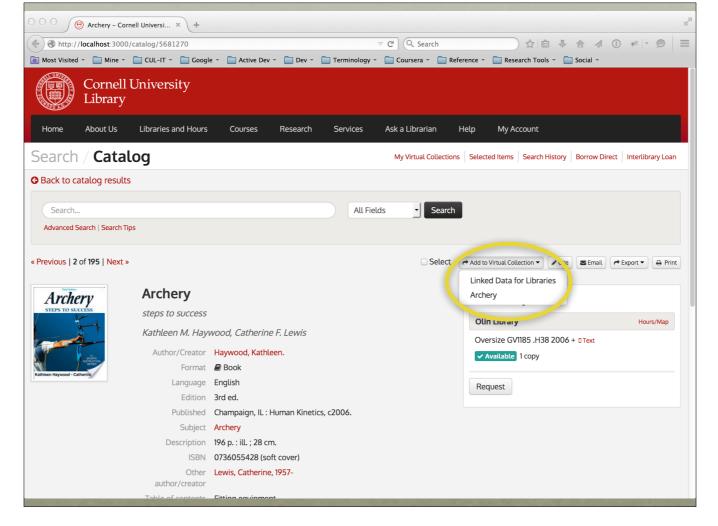
Click Home to search the Cornell Catalog for items to add to the collection.



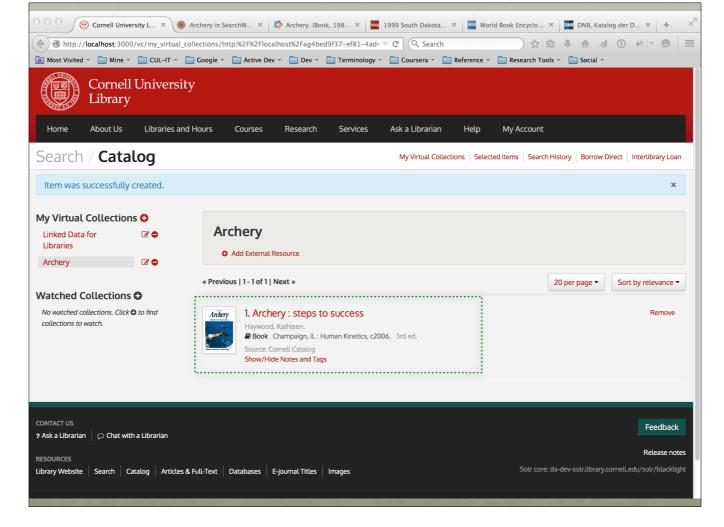
Search Cornell catalog for 'Archery'



Select one of the resources related to archery.



From the 'Add to Virtual Collection' drop list, select Archery.



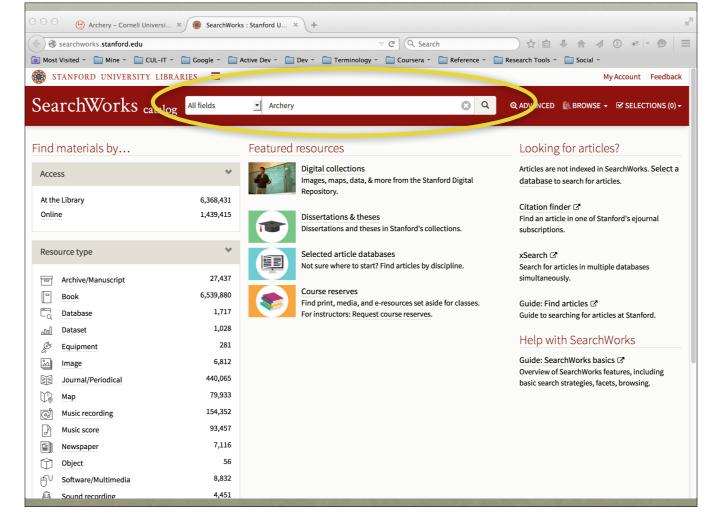
The book is added to the Archery virtual collection.

The host 'localhost: 3000' is registered, so follow the Cornell catalog process.

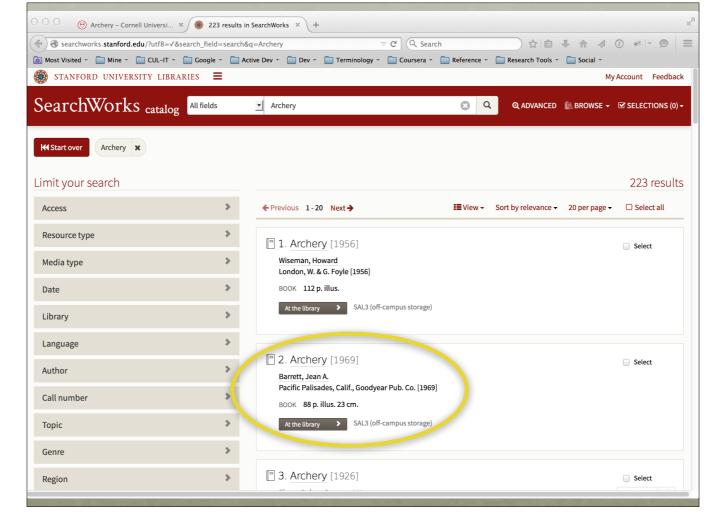
Process to add:

- * use content negotiation to get application/marcxml+xml
- * convert marcxml to bibframe triples
- * extract display metadata from bibframe triples

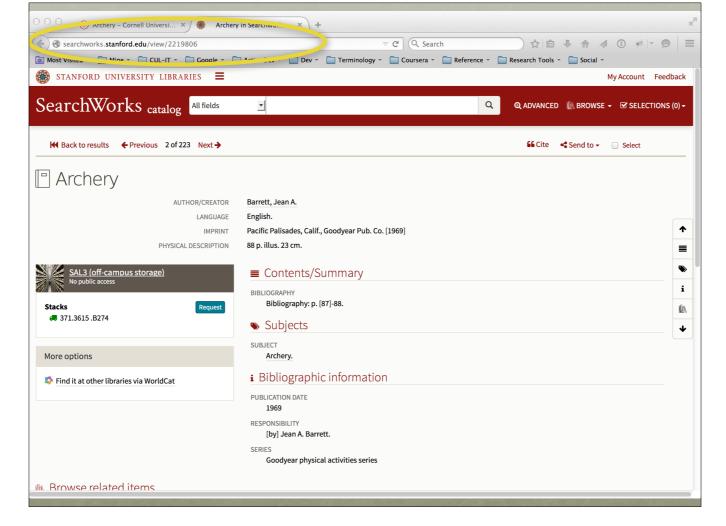
* add triples to add the URI to the aggregates list for the Archery aggregation and add the triples for a proxy resource according to the ORE ontology



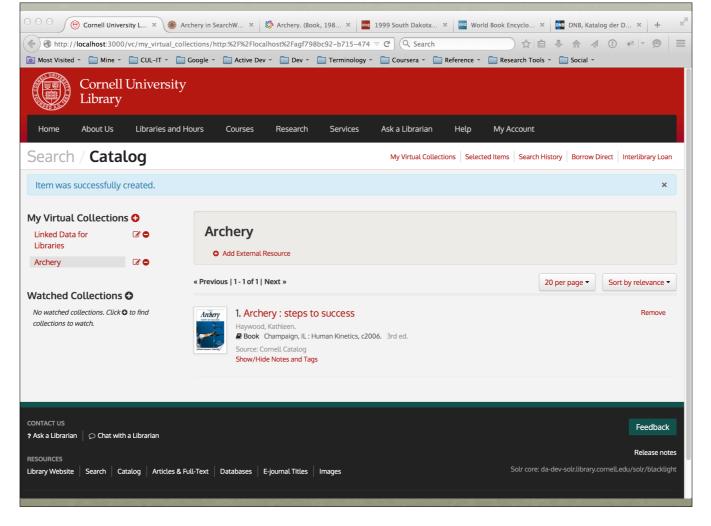
Search Stanford's searchworks for 'Archery'.



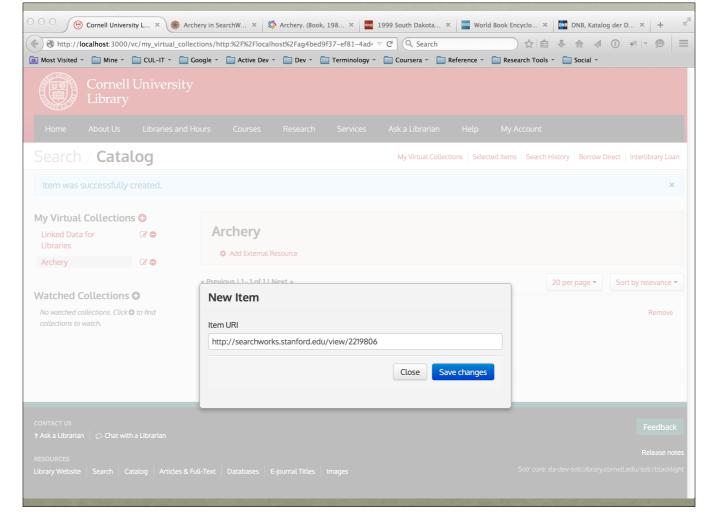
Select one of the resources related to archery.



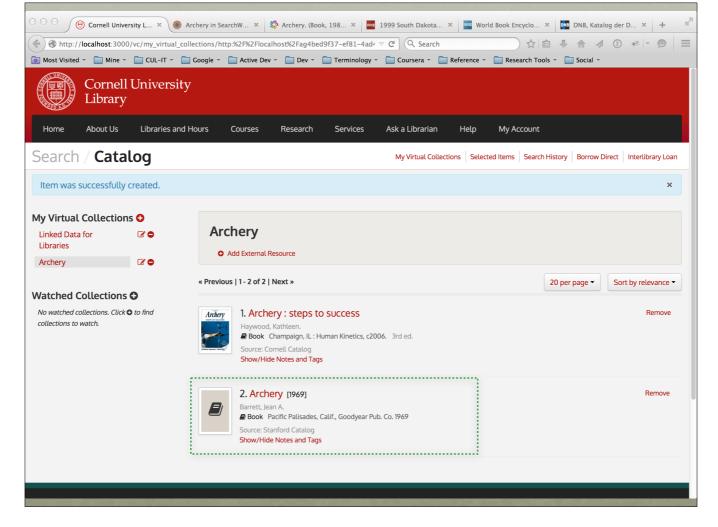
For Stanford's searchworks, the URI is the URL. Copy the URL from the browser address bar.



Click + Add External Resource under the virtual collection title Archery in the header of the main content area of the page.



Paste in the URI and Save changes.

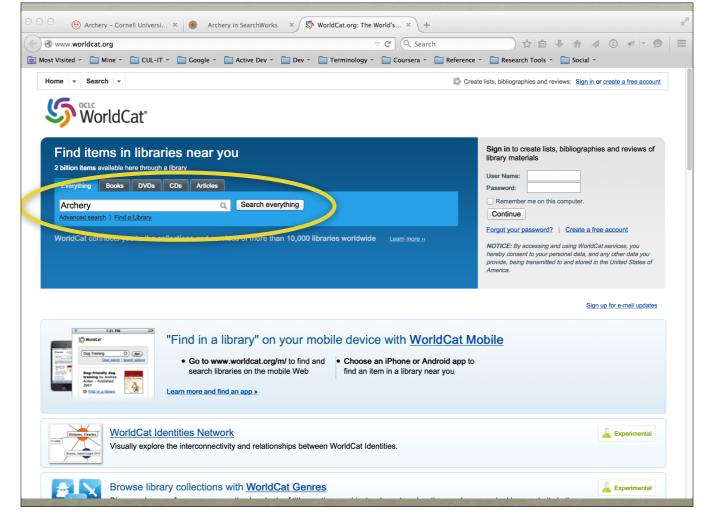


The book is added to the Archery virtual collection.

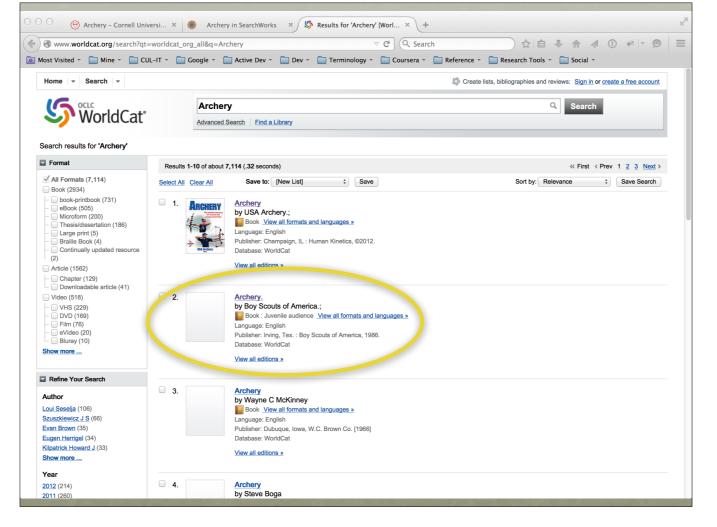
The host 'searchworks.stanford.edu' is registered, so follow the Stanford catalog process, which is the same as the Cornell catalog process. Process to add:

- * use content negotiation to get application/marcxml+xml
- * convert marcxml to bibframe triples
- * extract display metadata from bibframe triples

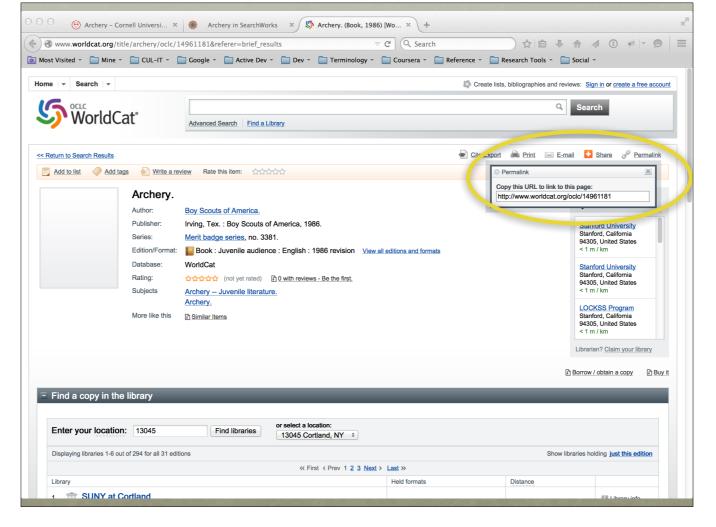
* add triples to add the URI to the aggregates list for the Archery aggregation and add the triples for a proxy resource according to the ORE ontology



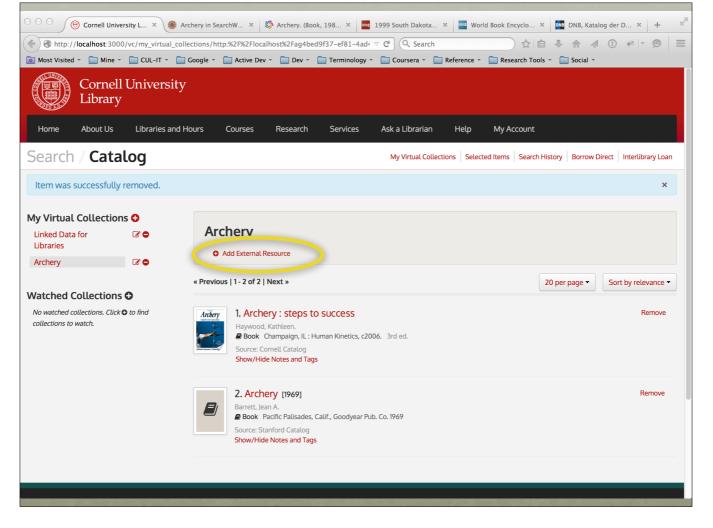
Search OCLC for 'Archery'.



Select one of the resources related to archery.



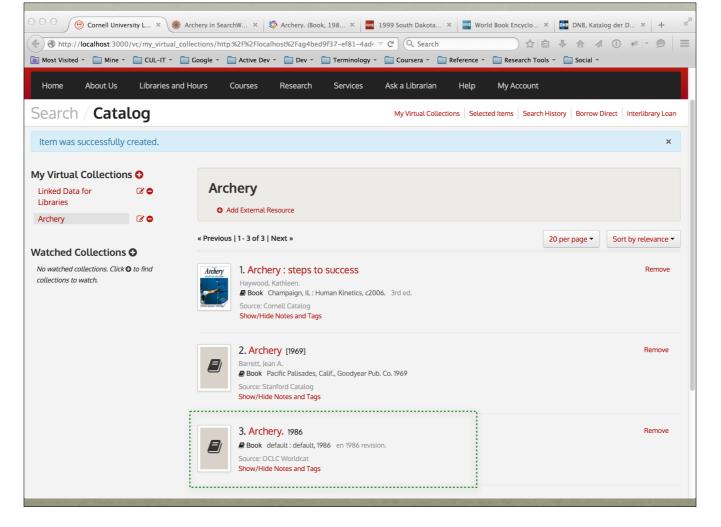
For OCLC, the URI is a permalink. Click the permalink symbol in the top right corner of the resource page and copy the URI.



Click + Add External Resource under the virtual collection title Archery in the header of the main content area of the page.

| OOO 🖲 Cornell University L × 🌸 Arche | ry in SearchW × 🦃 Archery. (Book, 198 × 🧱 1999 South Dakota × 🔤 World Book Encyclo | × 🛤 DNB, Katalog der D × 🕂 🖉 |
|---|--|--|
| O http://localhost:3000/vc/my_virtual_collection | ons/http:%2F%2Flocalhost%2Fag4bed9f37-ef81-4ad4 🗵 🤁 🔍 Search | |
| Most Visited - Mine - CUL-IT - Go | ogle 🔹 🦳 Active Dev 👻 📄 Dev 👻 🦳 Terminology 🐑 🦳 Coursera 👻 🦳 Reference 👻 🦳 Research T | iools 🔻 🛅 Social 👻 |
| Cornell University Library | | |
| Home About Us Libraries and Ho | urs Courses Research Services Ask a Librarian Help My Accou | nt |
| Search / Catalog | My Virtual Collections Selected Items See | arch History Borrow Direct Interlibrary Loan |
| Item was successfully removed. | | × |
| My Virtual Collections Linked Data for Libraries Archery | Archery Add External Resource | |
| Watched Collections | R Previous L1-2 of 21 Next » | 20 per page Sort by relevance |
| No watched collections. Click ☺ to find collections to watch. | Item URI | Remove |
| | http://www.worldcat.org/oclc/14961181 | |
| | Close Save changes | |
| | | Remove |
| | Book Pacific Palisades, Calif., Goodyear Pub. Co. 1969 Source: Stanford Catalog Show/Hide Notes and Tags | |
| | | |
| | | |

Paste in the URI and Save changes.

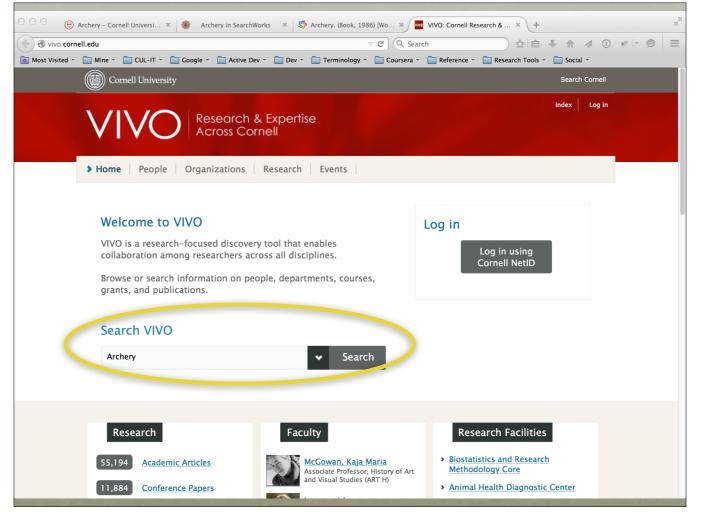


The book is added to the Archery virtual collection.

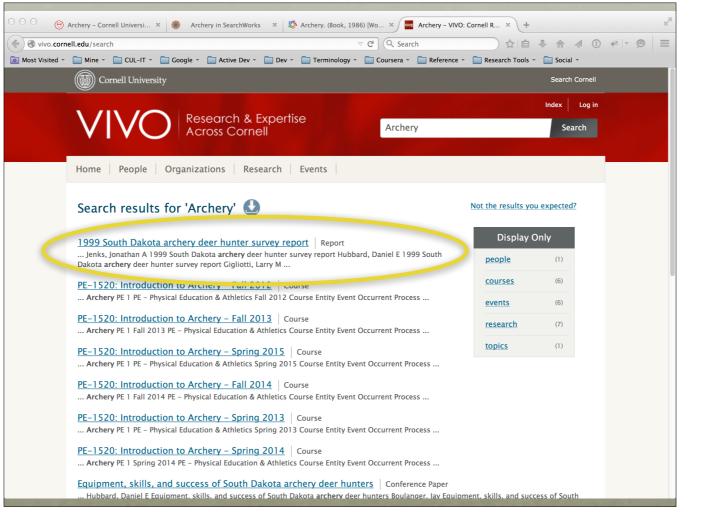
The host 'www.worldcat.org' is registered, so follow the OCLC process.

Process to add:

- * use content negotiation to get application/rdf+xml
- * extract display metadata from schema triples
- * add triples to add the URI to the aggregates list for the Archery aggregation and add the triples for a proxy resource according to the ORE ontology



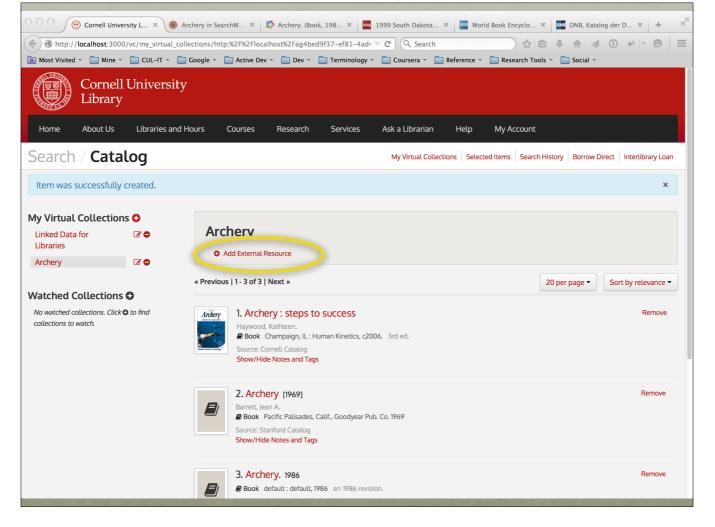
Search Cornell VIVO for 'Archery'.



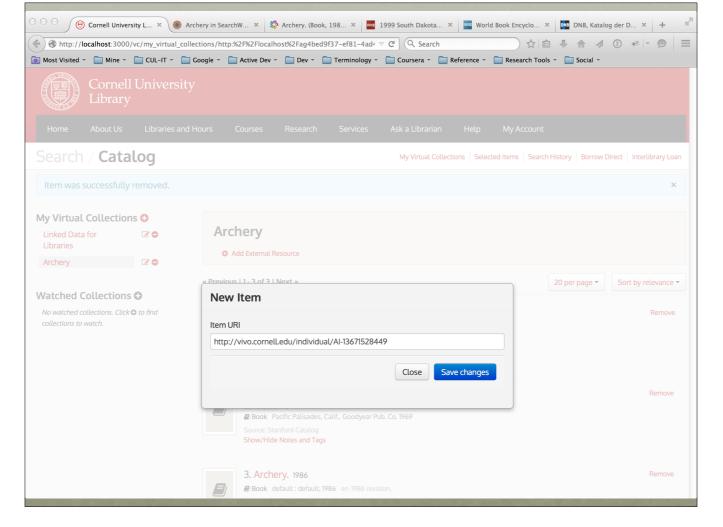
Select one of the resources related to archery.

| 000 🖲 🗚 | Archery - Cornell Universi 🗴 🌸 Archery in SearchWorks 🗴 🏷 Archery. (Book, 1986) [Wo 🗴 🚾 1999 South Dakota archer 🗴 🕂 | R _M |
|----------------|---|----------------|
| 🔶 🎯 vivo.corne | rell.edu/display/Al-13671528449 🗸 🕻 🖉 🖉 🖉 | |
| Most Visited - | Mine 🕆 🗋 CUL-IT 🕆 🦳 Google 👻 🦳 Active Dev 🕆 🛄 Dev 👻 🦳 Terminology 🐑 🦳 Coursera 🔻 🋄 Reference 🔻 🋄 Research Tools 👻 🦳 Social 👻 | |
| | Cornell University Search Cornell | |
| | Index Log in Across Cornell Search | |
| | Home People Organizations Research Events | |
| | 1999 South Dakota archery deer hunter survey report Report I Overview Additional Document Info View All Mathematical actions view profile in RDF format Gigliotti, Larry M Boulanger, Jay Jenks, Jonathan A Hubbard, Daniel E publication date 2000 | e |
| | | |

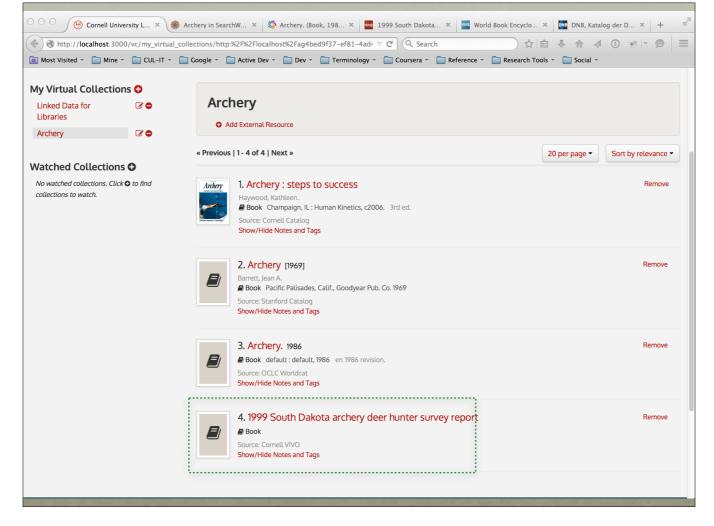
For VIVO Cornell, the URI is a permalink. Click the permalink symbol beside the title of the resource and copy the URI.



Click + Add External Resource under the virtual collection title Archery in the header of the main content area of the page.



Paste in the URI and Save changes.



The report is added to the Archery virtual collection.

The host 'vivo.cornell.edu' is registered, so follow the Cornell VIVO process.

Process to add:

* use content negotiation to get text/turtle

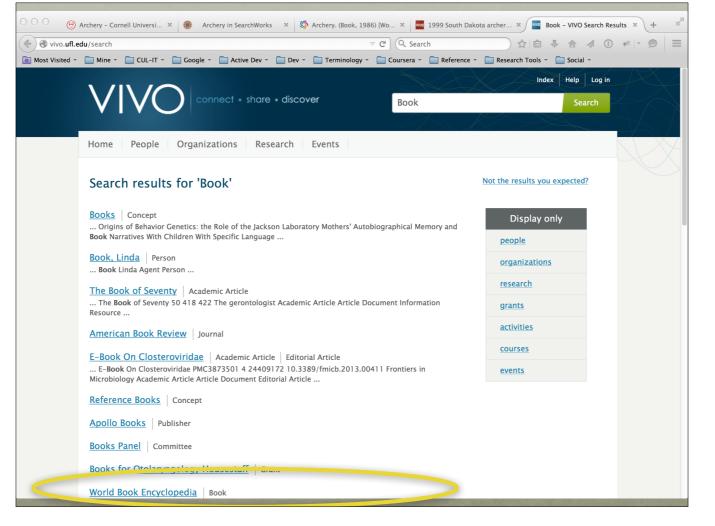
* extract display metadata from BIBO/VIVO triples

* author and publisher are URIs and the triples with details about these are not provided by the first content negotiation. For this demo, only one content negotiation is performed. Additional content negotiation requests could be made to get the author and publisher details.

* add triples to add the URI to the aggregates list for the Archery aggregation and add the triples for a proxy resource according to the ORE ontology

| | ⊽ ୯) (୦ | Search → Construction Construc | 0 */- ⊘ ≡ |
|---|--|---|-----------|
| Home People Organizations | nare • discover Research Events | Index Help Log ir Search | |
| Welcome to VIVO VIVO is a scholarly networking and understanding and collaboration at Browse or search for information o courses, and organizations. Search VIVO | mong all disciplines. n people, publications, grants, | Manage My Page Log in now to manage your profile page. Gatorlink Login | |
| Book Browse by People (186,439) Organizations (13,563) | Search Clinical Faculty Consultant | | |
| Research (77,529) | Courtesy Faculty | | |

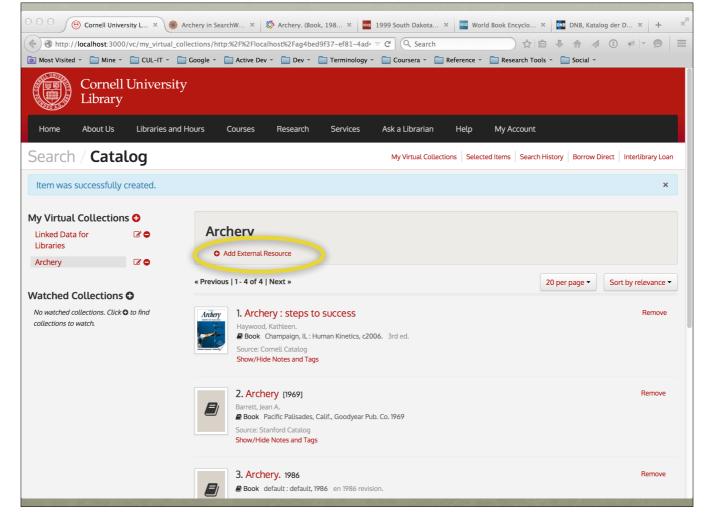
Search for 'Book' at another VIVO site that is not registered in the Virtual Collection system. NOTE: UFL doesn't have any books on Archery in its VIVO.



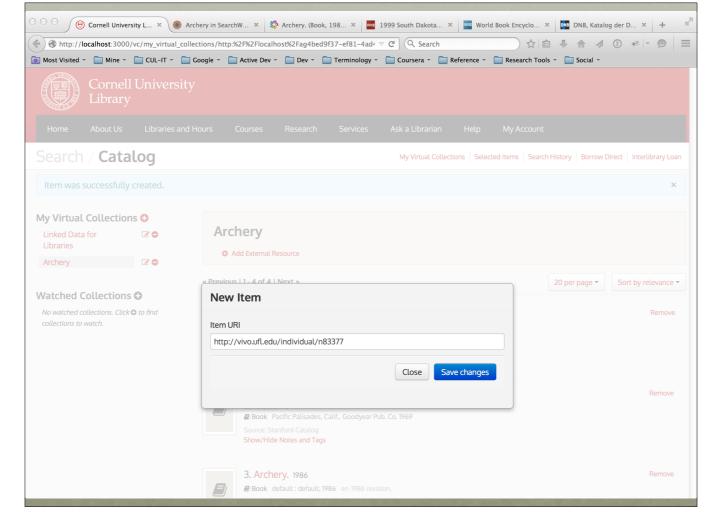
Select one of the resources that is a Book.

| | rchWorks × SArchery. (Book, 1986) [Wo × 🚾 1999 South Dakota archer. © C Q. Search e Dev ~ Dev ~ Terminology ~ Coursera ~ Reference ~ Res shore • discover | |
|--|--|--|
| Home People Organizations World Book Encyclopedia Background Image: Contributor Balmer, Todd Scientific Advis Scientific Advis | Research Events | |

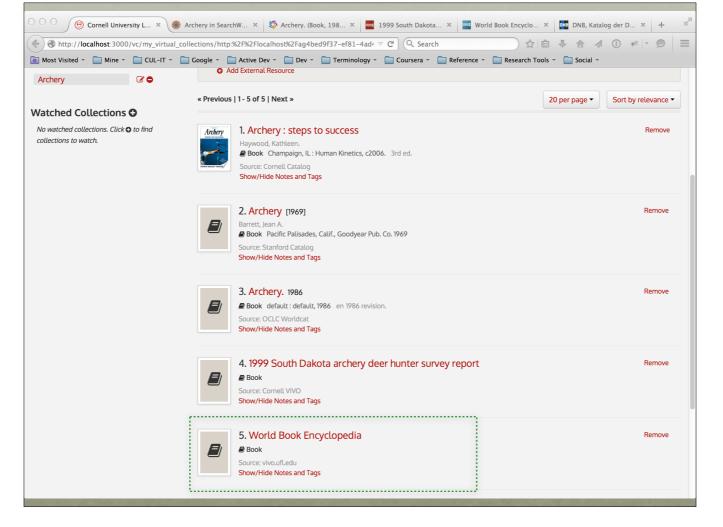
For VIVO at UFL, the URI is a permalink. Click the permalink symbol beside the title of the resource and copy the URI.



Click + Add External Resource under the virtual collection title Archery in the header of the main content area of the page.



Paste in the URI and Save changes.

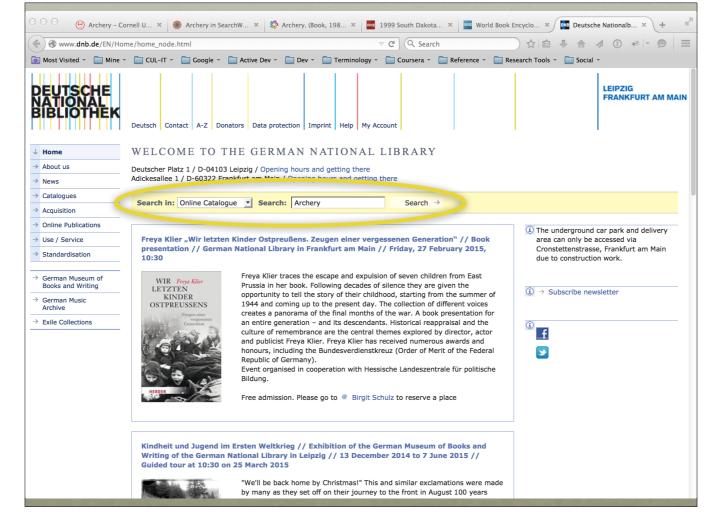


The report is added to the Archery virtual collection.

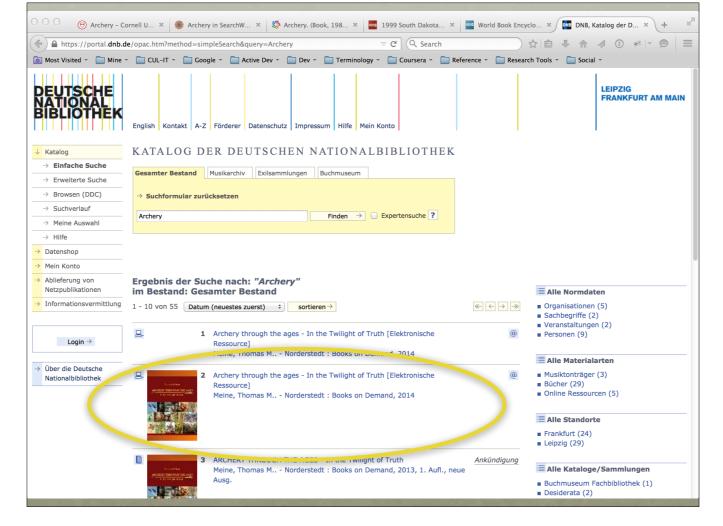
The host 'vivo.ufl.edu' is NOT registered, so use generic processing. Process to add:

- * use content negotiation to attempt to get application/rdf+xml
- * recognize that the rdf type is bibo:Book and use Cornell VIVO process for metadata extraction
- * extract display metadata from BIBO/VIVO triples

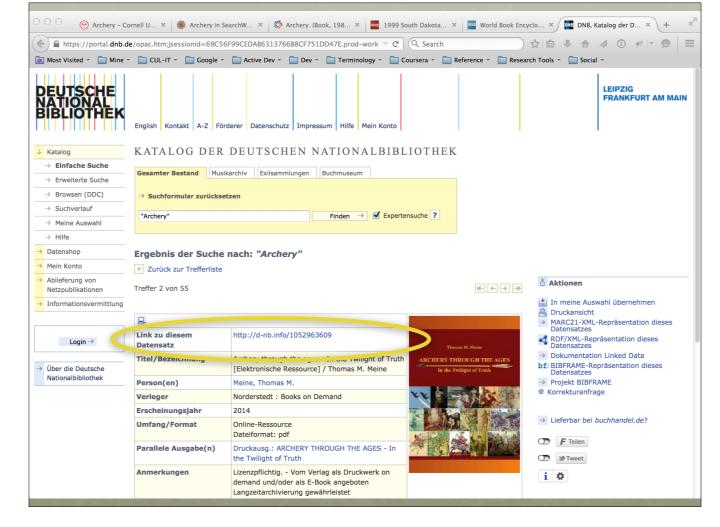
* add triples to add the URI to the aggregates list for the Archery aggregation and add the triples for a proxy resource according to the ORE ontology



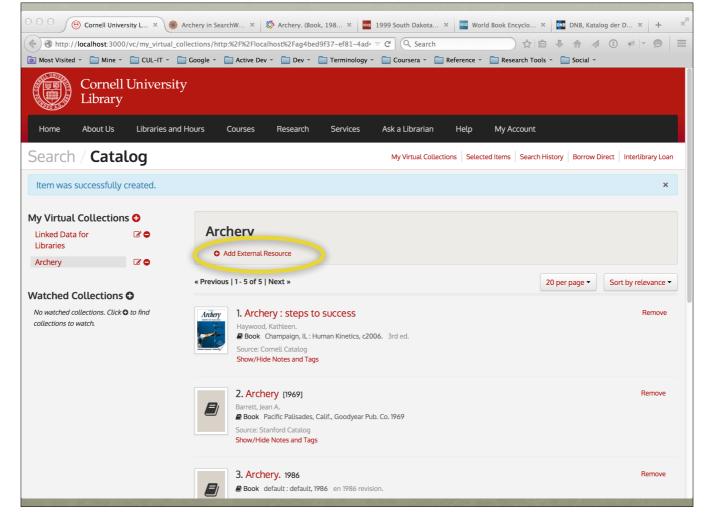
Search for 'Archery' at another site that is not registered in the Virtual Collection system.



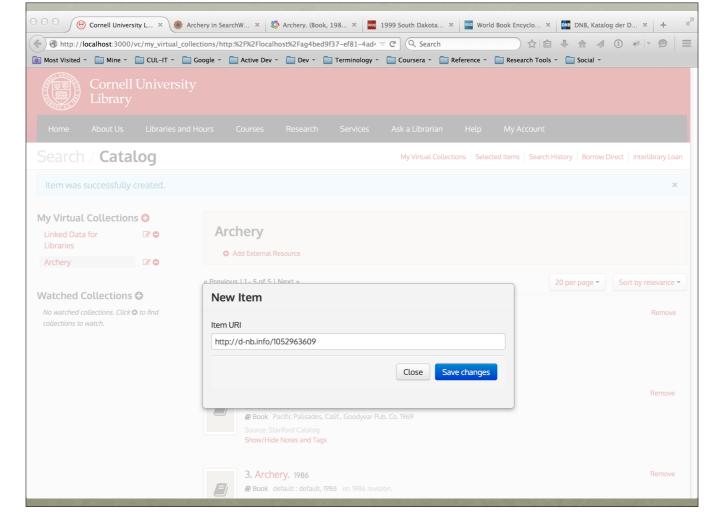
Select one of the resources related to archery.



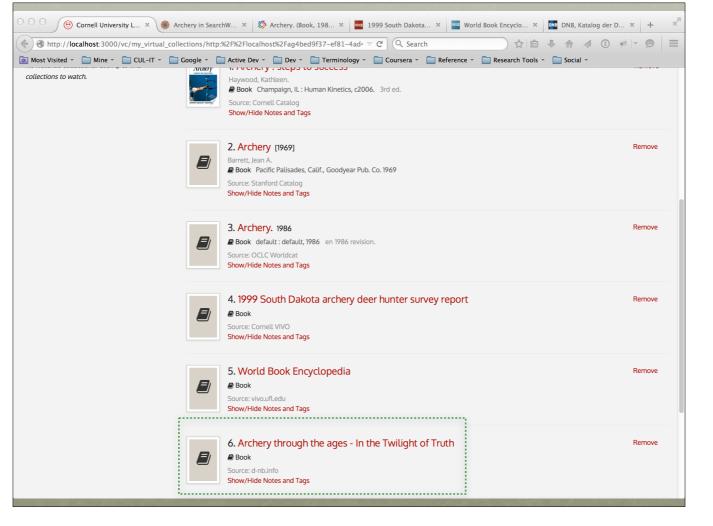
For Deutsche National Bibliothek, the URI is a permalink. Copy the permalink listed as the first item in the table of information about the resource on the resource detail page.



Click + Add External Resource under the virtual collection title Archery in the header of the main content area of the page.



Paste in the URI and Save changes.



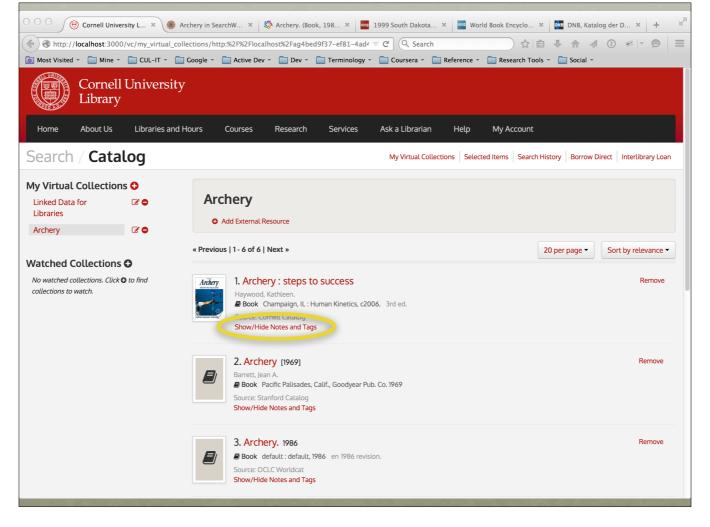
The book is added to the Archery virtual collection.

The host 'd-nb.info' is NOT registered, so use generic processing. Process to add:

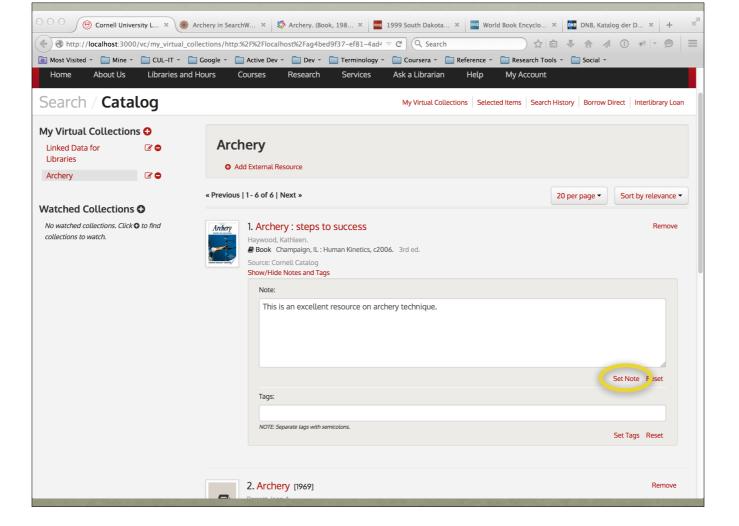
Process to add:

- * use content negotiation to attempt to get application/rdf+xml
- * recognize that the rdf type is bibo:Document and use Cornell VIVO process for metadata extraction
- * attempt to extract display metadata from BIBO/VIVO triples

* add triples to add the URI to the aggregates list for the Archery aggregation and add the triples for a proxy resource according to the ORE ontology

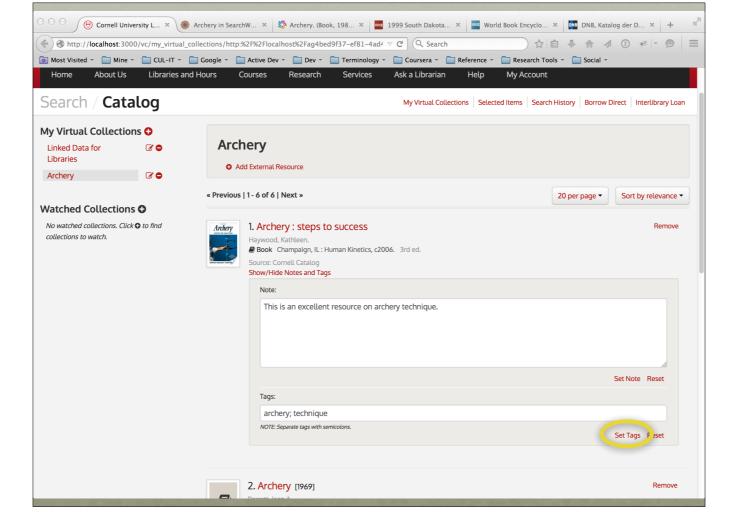


Click Show/Hide Notes and Tags to annotate a resource.



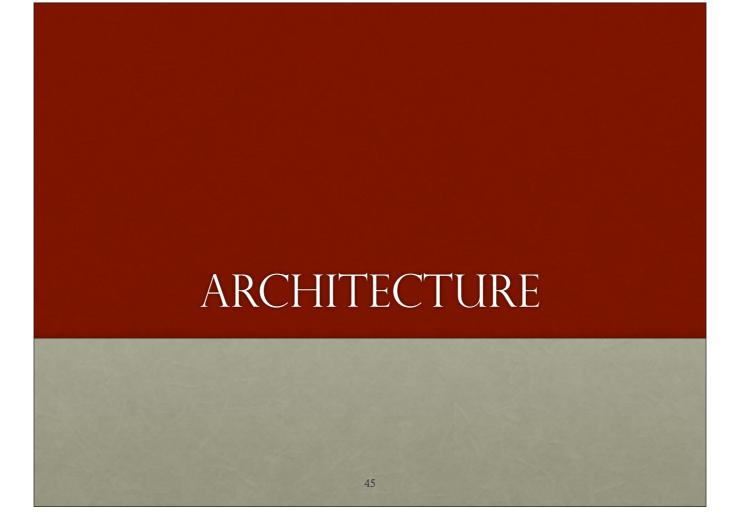
Type in text for the note and click Set Note.

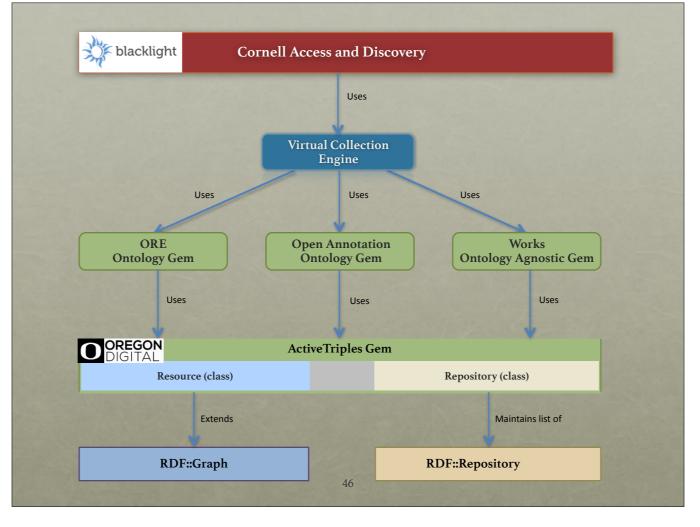
The note is saved as a comment annotation in the triplestore using triples as defined in the Open Annotation ontology's.



Type in text for the note and click Set Tags.

Each tag is saved as a tag annotation in the triplestore using triples as defined in the Open Annotation ontology's.





Two major pieces of the infrastructure existed before work began on the system. The **Cornell Access and Discovery** system is the live search system for the catalog that was implemented with blacklight. The goal was to integrate the virtual collection management code into the Access and Discovery system.

The other existing work was ActiveTriples which is a gem that was extracted from the work done on the Oregon Digital project. By extending the ActiveTriples::Resource class, models can be defined to have predicates as properties that can hold the object values of a set of triples for a common subject. This allows for triples to be manipulated in logical clusters instead of each triple being manipulated individually, thus reducing the likelihood of errors being introduced and increasing programming productivity. ActiveTriples also provides a means for persisting triples into a triplestore and later resuming all the related triples back into a new instance of a resource model. ActiveTriples makes use of several of the classes defined in the ruby rdf project on github. I highlight two here, the RDF::Graph and RDF::Repository, but there are many other classes that can be used for querying and transforming triples, and connecting to various triplestore implementations.

The work underway for the LD4L use case includes the definition of ontology gems which **inherit** all the **triple manipulation functionality** provided by ActiveTriples and the virtual collection engine which uses the ontology gems to build the bigger concept of a Virtual Collection and provide the UI integration into the Access and Discovery system. By separating the code into ruby supported components, we achieve code isolation for improved testing and a **higher degree of reusability which lowers the programming effort**. We saw this earlier in Stanford's **Triannon** system which was able to make use of the Open Annotation Gem for working with annotation triples in ruby. Each gem defines models for classes in the ontology by **extending ActiveTriples::Resource** and by **adding additional methods** that provide commonly required **services for a specific ontology**. For example, when you add a new item to an aggregation, the ORE ontology gem adds a new proxy resource for the item and updates the aggregation which holds a list of URIs for all items in the list.

WORKING WITH AN ONTOLOGY GEM

Setup required for all examples.

require 'ld4l/open_annotation_rdf'

```
ActiveTriples::Repositories.add_repository :default, RDF::Repository.new
p = LD4L::FoafRDF::Person.new('p4')
```

Example creating a comment annotation.

ca = LD4L::OpenAnnotationRDF::CommentAnnotation.new('c10')
ca.hasTarget = RDF::URI("http://example.org/bibref/br3")
cb = ca.setComment("This book is a good resource on archery technique.")
ca.annotatedBy = p
ca.setAnnotatedAtNow
ca.persist!

Triples created...

<http://localhost/9c8c8126-2d31-48be-81d8-3cd4748a3351> a

<http://www.w3.org/2011/content#ContentAsText>,

<http://purl.org/dc/dcmitype/Text>;

<http://purl.org/dc/terms/format> "text/plain";

<http://www.w3.org/2011/content#chars> "This book is a good resource on archery technique." .



PROCESSING SPEED FOR COLLECTING & DISPLAYING

- Real time metadata extraction is slow
 - simple triple content negotiation is slow
 - multiple content negotiation calls is required by some sites
 - worst case occurs for sites providing only marcxml leading to a conversion step before extracting metadata
- Working at scale
 - solr indexing & caching
 - periodic re-indexing of resource metadata
 - length of time to re-index

The first challenge is processing speed for collecting and displaying metadata about each resource in the collection. Right now, this impacts every view of a virtual collection as the metadata is re-extracted with every refresh of the display. This is simply a limitation of the system not being far enough along in the development process.

49

In the near future, this problem will be relegated to adding a new item.

The plan is to use a **solr index to cache** the extracted metadata and use **periodic re-indexing** or a lazy re-index based on last date extracted to bring the metadata up to a current state. A continuing **challenge** once indexing is in place is the length of **time to re-index** as the number of items across all collections grows. We anticipate the potential for re-indexing to happen on the scale of a day.

BATCH PROCESSING FOR LARGE SCALE COLLECTIONS

- Rules for gathering records
 - range of call numbers
 - keywords
 - owned by a particular library
- Batch metadata manipulation
- Processing ranges of items

As we move from the current small scale support to large, hundreds of thousands, scale, we will be adding in batch processing. At Cornell, we currently have CuLLR which is a system that allows us to gather resources into a collection and provides a website that can be navigated and searched as though it is a separate library. The technology for this system is no longer current, but we learned a lot about batch processing and want to carry that knowledge forward into the new system.

50

For example, CuLLR supports rules like identifying a range of call numbers, matching keywords, and ownership by a particular library, as well as combinations of rules, like all call numbers starting with K that are owned by the Law library.

The second batch processing approach will allow common metadata to be set for all or large selections of items in a collection to be set in one step.

As a collection grows, it will need to operate on a range of items instead of the entire collection. This is particularly challenging for ordered lists.

INTEGRATION WITH OTHER INSTITUTIONS

- Not much linked data exists.
- Many ontologies.
- Many interpretation of ontology usage.
- Varying support for content negotiation.
- How does the user locate the URI to use?

RESOURCES

- RDF Ruby Development:
 - <u>ruby-rdf</u> rdf, linkeddata, rdf-do
 - <u>ActiveTriples</u> manipulate triples as objects
- LD4L
 - <u>ORE Gem</u> aggregations representing lists with items
 - <u>OA Gem</u> comment, tag, and semantic tag annotations
 - <u>Works Gem</u> ontology agnostic providing multiple models for works, metadata extraction into a simplified set of display metadata
 - <u>Virtual Collection Engine</u> UI for Virtual Collections

52

SIMILAR WORK

• projecthydra-labs/rdf-proxy list

• <u>CuLLR</u> (Virtual Libraries at Cornell - non linked data)

53

