# The Linked Data for Libraries Project: A Progress Report

Dean B. Krafft and Tom Cramer

OLF Forum

October 27, 2015







## Linked Data for Libraries (LD4L)

- Nearing the end of a two-year \$999K grant to Cornell, Harvard, and Stanford
- Partners have worked together to assemble ontologies and data sources that provide relationships, metadata, and broad context for Scholarly Information Resources
- Leverages existing work by both the VIVO project and the Hydra Partnership

# Vision: Create a LOD standard to exchange all that libraries know about their resources



# Overview

## Specific LD4L goals

- Free information from existing library system silos to provide context and enhance discovery of scholarly information resources
- Leverage usage information about resources
- Link bibliographic data about resources with academic profile systems and other external linked data sources
- Assemble (and where needed create) a flexible, extensible LD ontology to capture all this information about our library resources
- Demonstrate combining and reconciling the assembled LD across our three institutions

## LD4L working assumptions

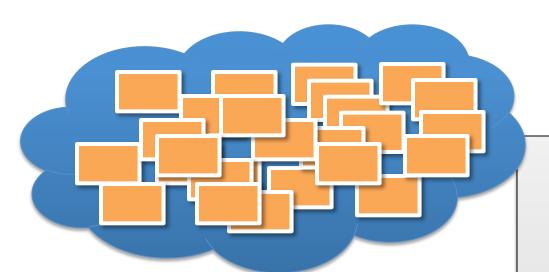
- Trying to do conversion and relation work at scale--with full sets of enterprise data
  - Harvard: 13.6 million bibliographic records
  - Stanford and Cornell: roughly 8 million bib records in each collection
- Trying to understand the pipeline / workflows that will be needed for this
- Looking to build useful, value-added services on top of the assembled triples

### LD4L data sources

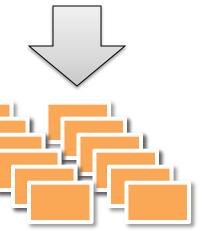
Person Data Bibliographic Data • CAP, FF, MARC VIVO • MODS ORCID • EAD ISNI VIAF, LC Usage Data Circulation Citation Curation • Exhibits Research Guides Syllabi

Tags

# Selected Use Cases



42 raw use cases



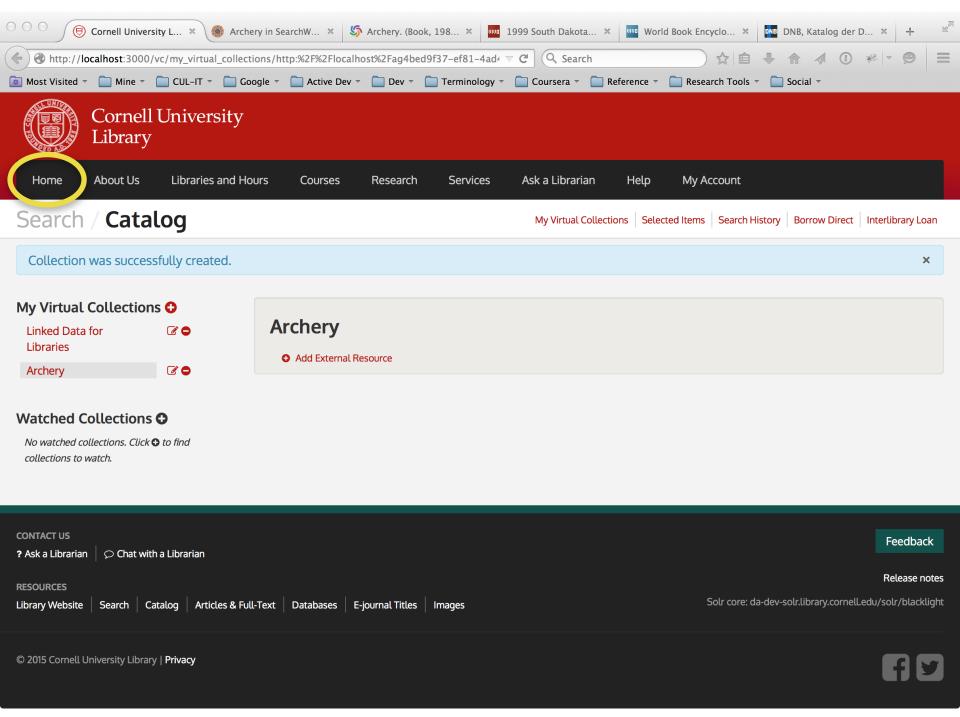
12 refined use cases in 6 clusters...

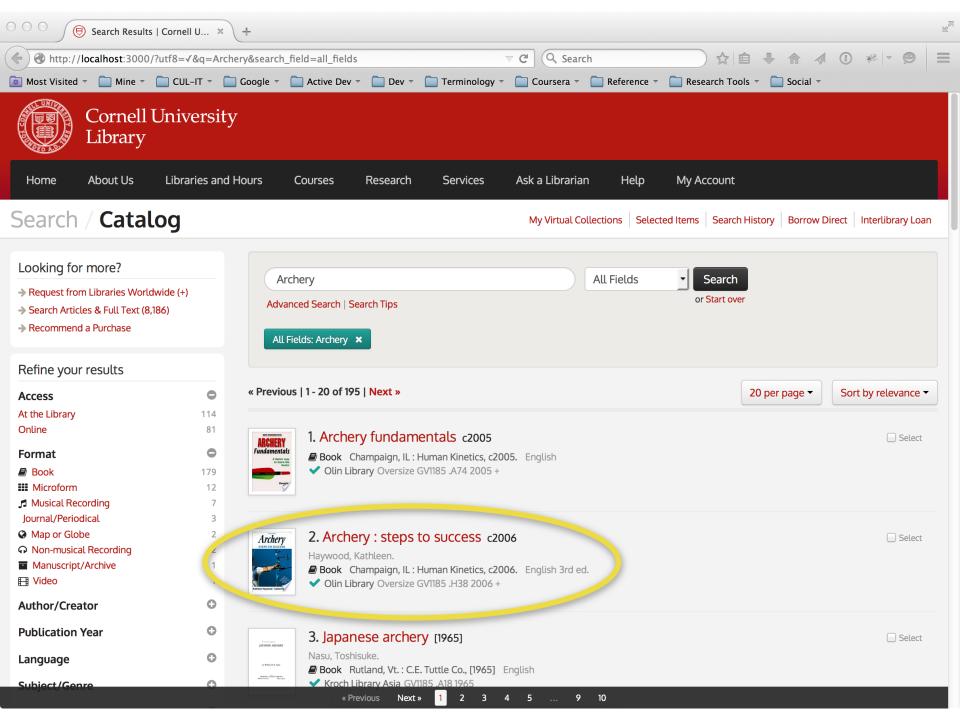
#### LD4L Use Case Clusters

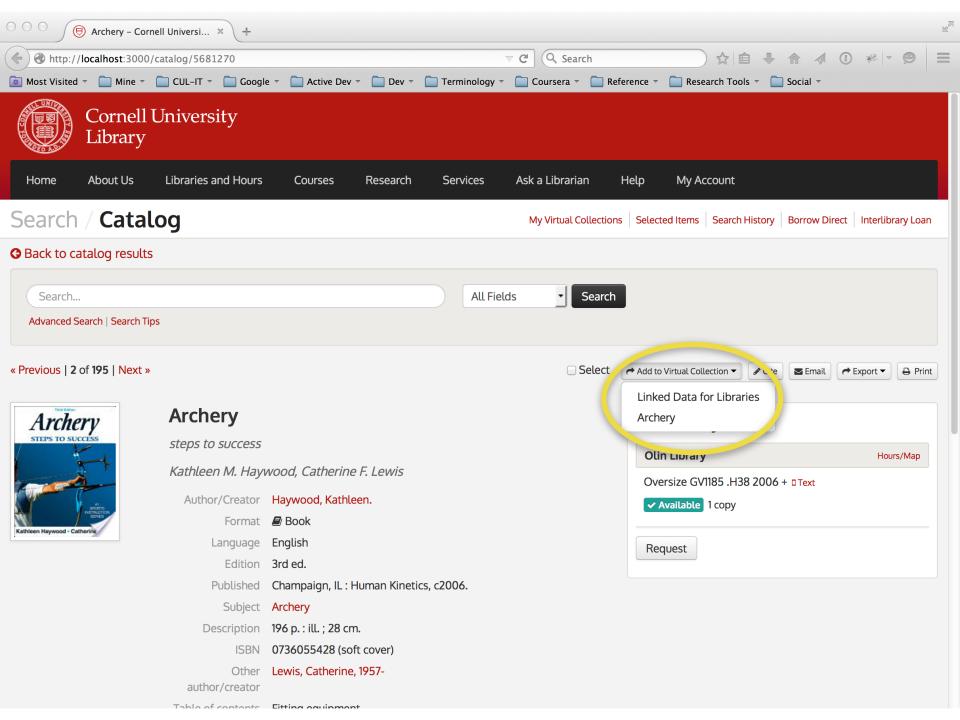
- 1. Bibliographic + curation data
- 2. Bibliographic + person data
- 3. Leveraging external data including authorities
- 4. Leveraging the deeper graph (via queries or patterns)
- 5. Leveraging usage data
- 6. Three-site services, e.g. cross-site search

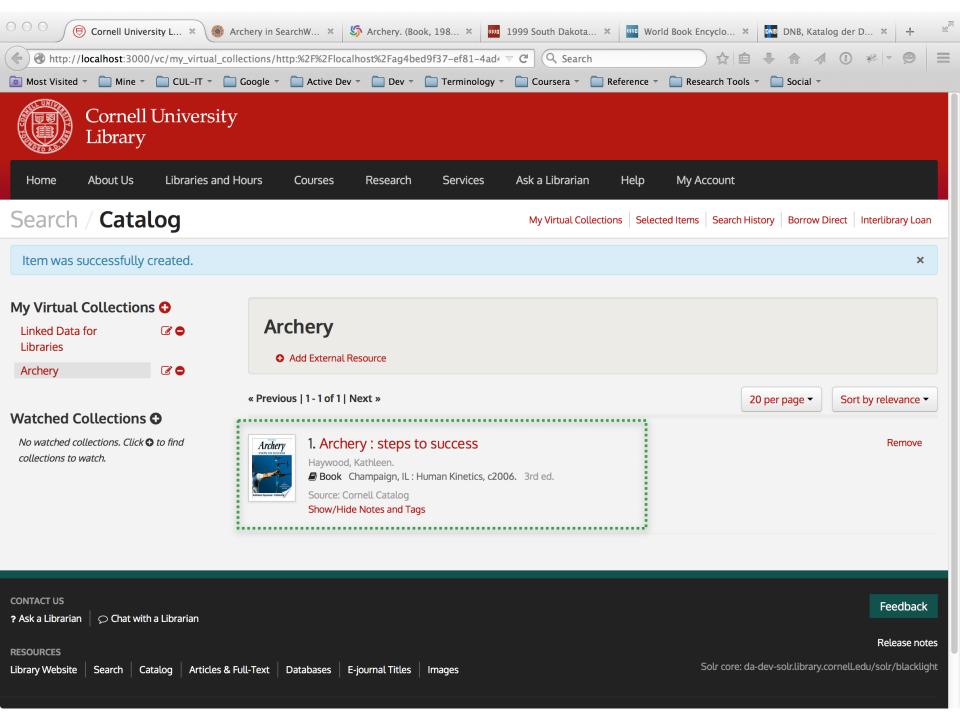
### 1.1 Build a virtual collection

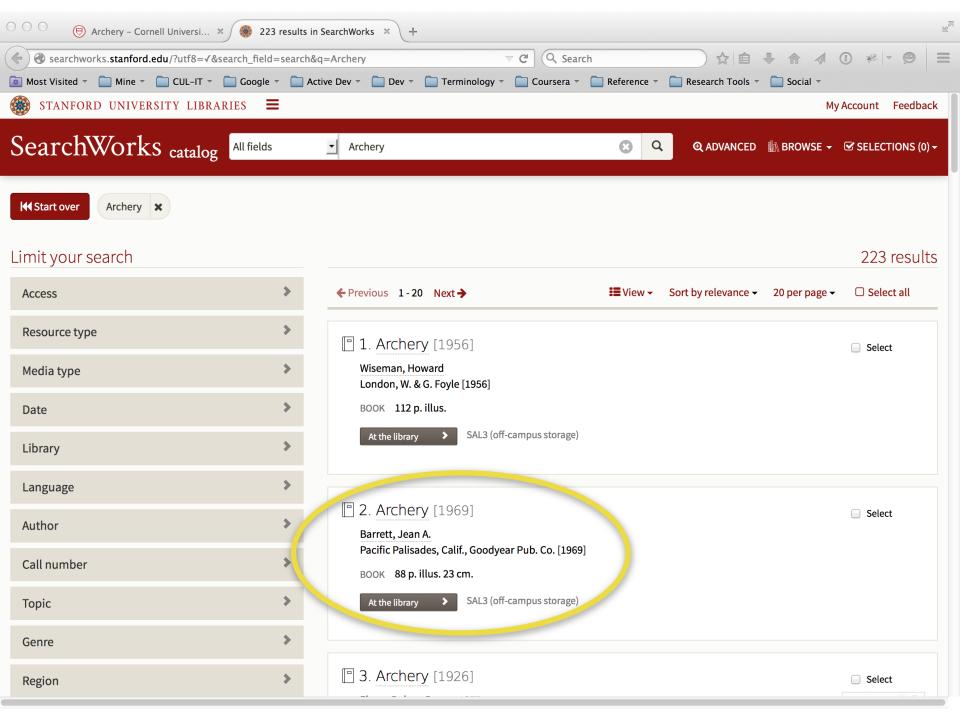
- Goal: allowing librarians and patrons to create and share virtual collections by tagging and optionally annotating resources
- Implementations
  - Cornell
  - Stanford

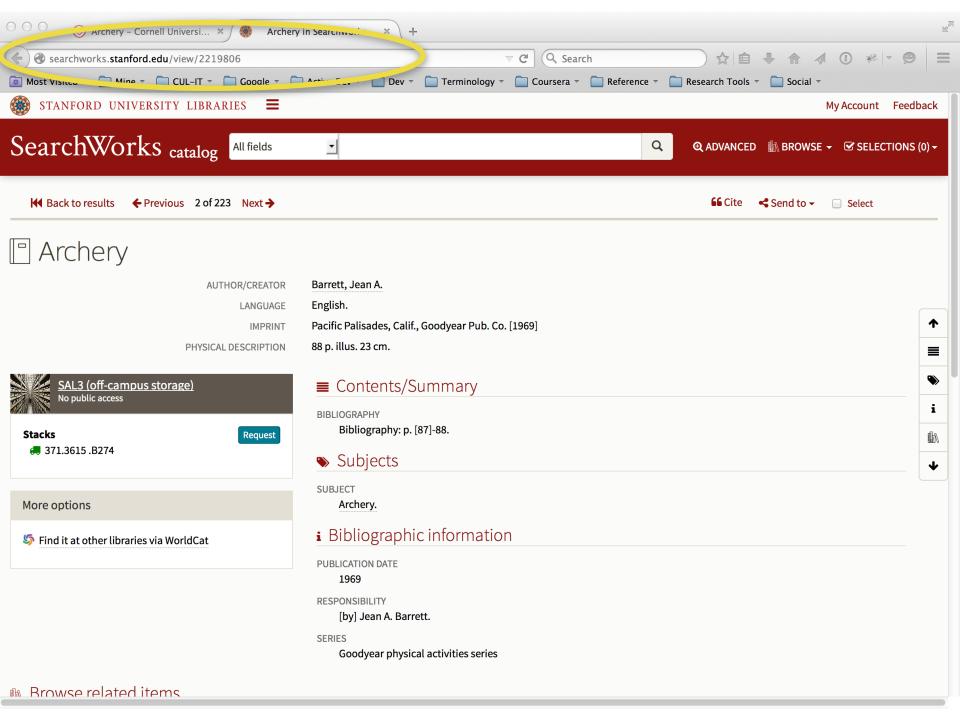


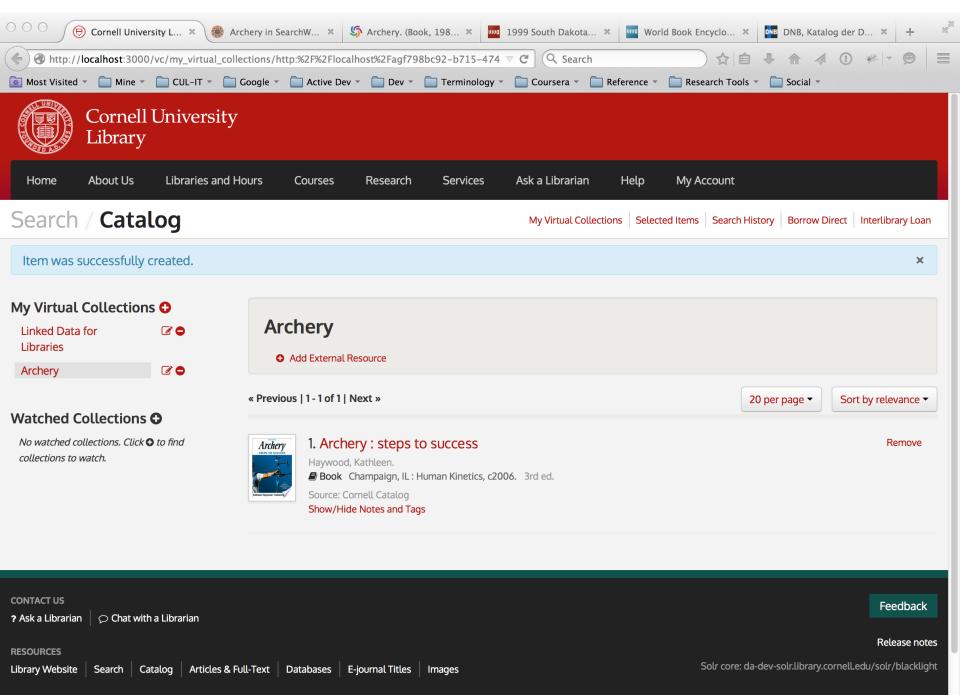


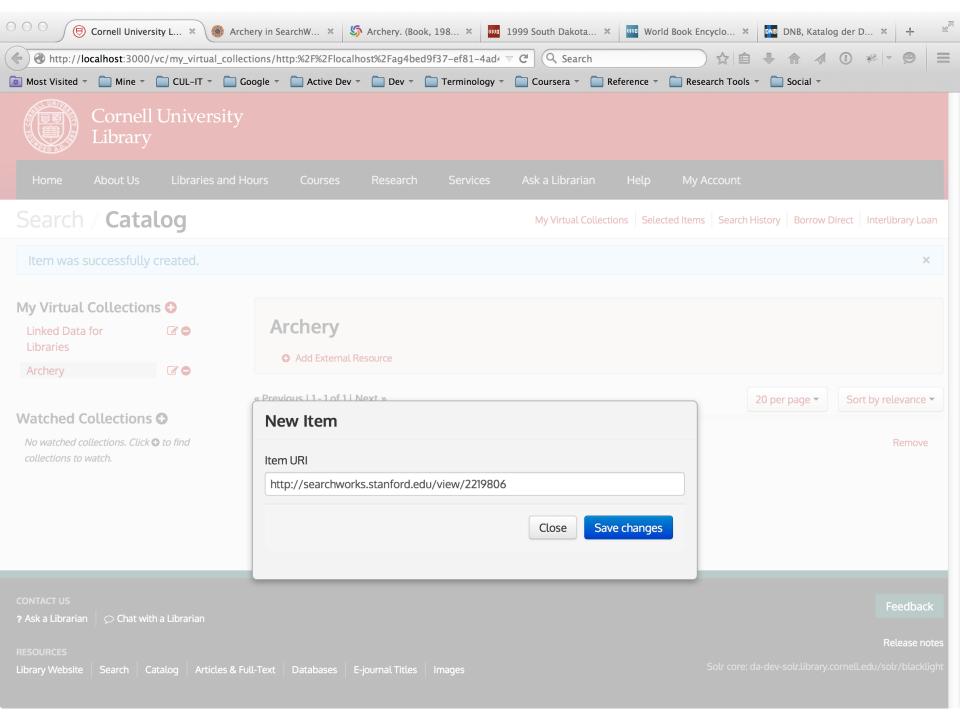


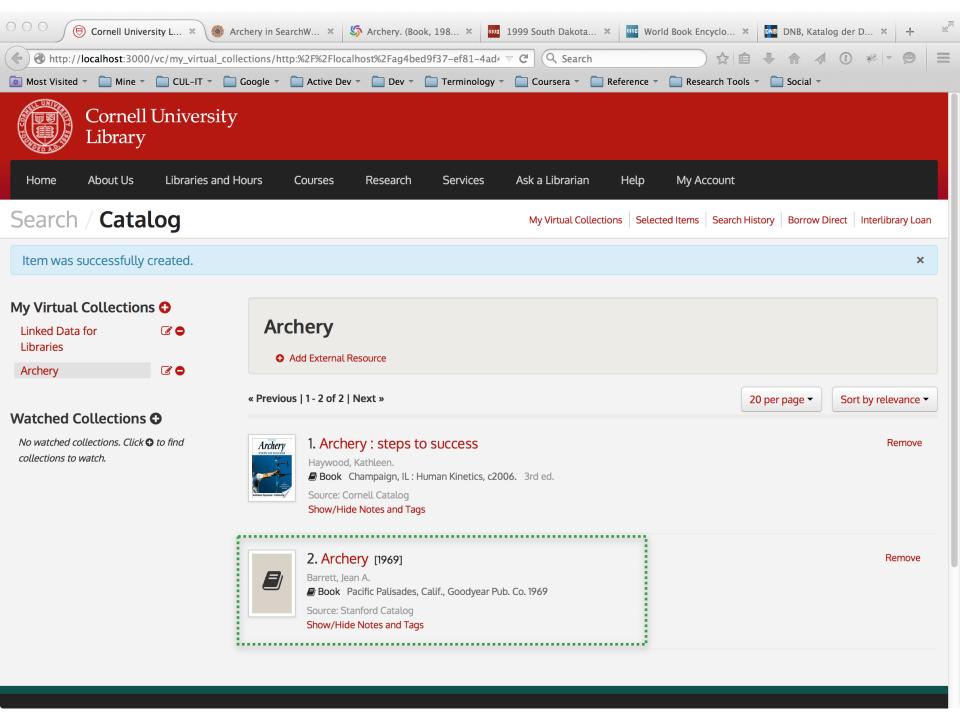


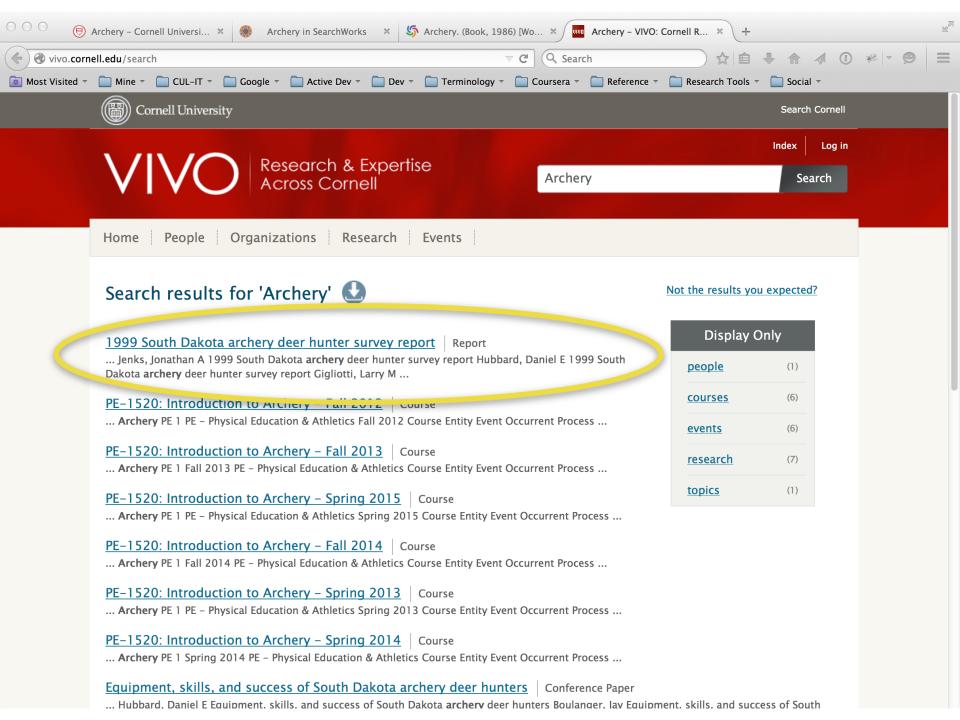


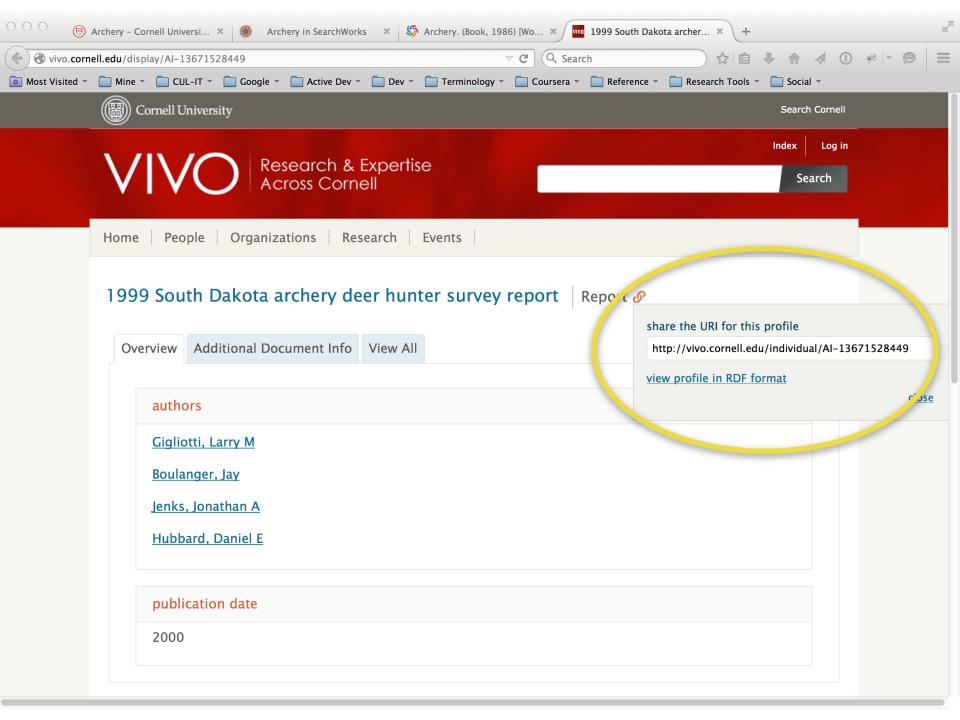


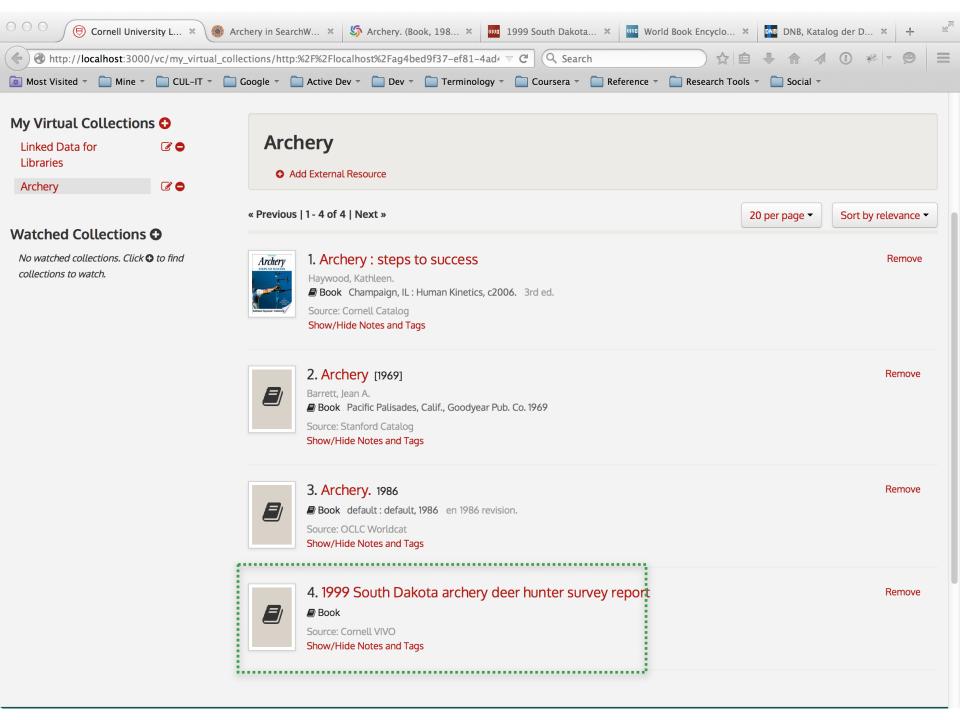






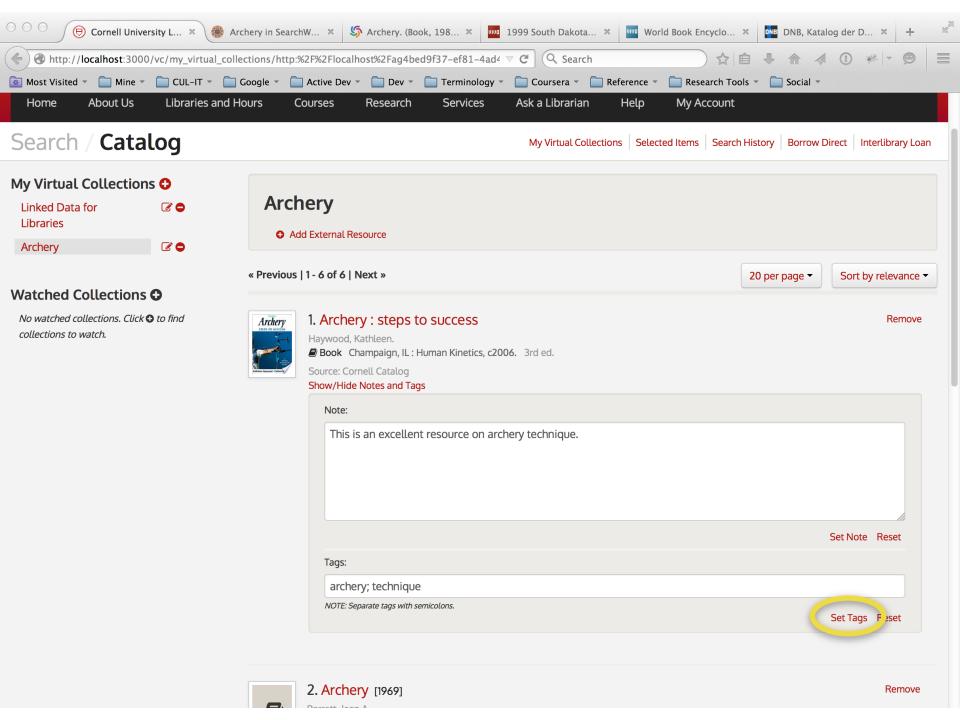






# 1.2 Tag scholarly information resources to support reuse

- Goal: provide librarians tools to create and manage larger online collections of catalog resources
- Implementation
  - More automation
  - Batch processes as well as individual editing



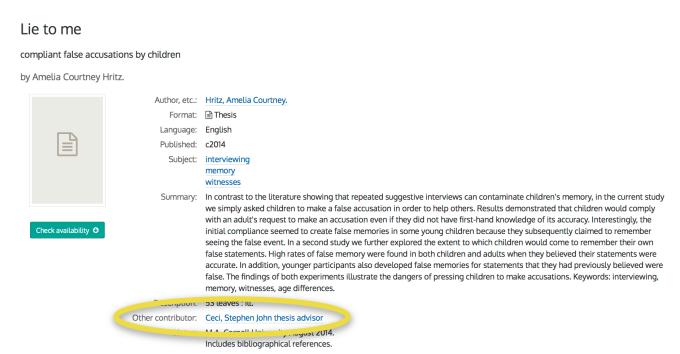
Discussion: How is this approach relevant to my needs and library?

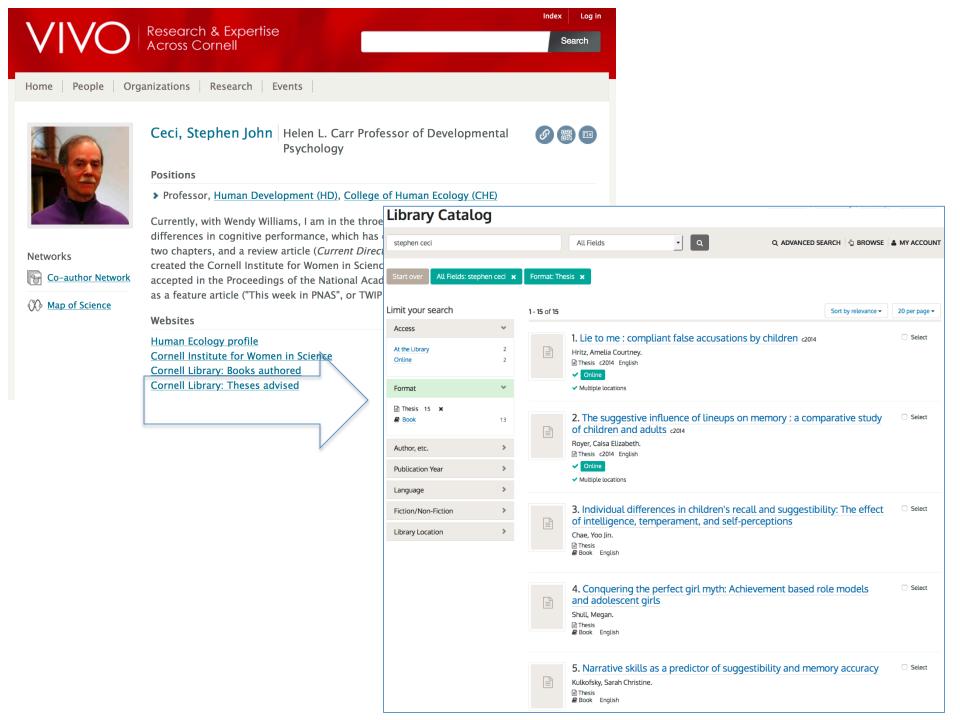
# 2.1 See and search on works by people to discover more works and better understand people

- Goal: link catalog search results to researcher networking systems to provide current articles, courses
- Implementation
  - Adding VIVO URIs to MARC records for thesis advisors
  - Adding links to VIVO records linking back to faculty works and their students' theses
  - Raises important issues about URI stability

#### Thesis Advisors and VIVO

- Cornell Technical Services is including thesis advisors in MARC records using NetIDs from the Graduate school database
   e.g., 700 1 ‡a Ceci, Stephen John ‡e thesis advisor ‡0
- Advisors are looked up against VIVO to get URIs for the faculty members





# 4.1 Identifying related works

- Goal: finding additional resources beyond those directly related to any single work using queries or patterns, as for example changes in illustrations over a series of editions of a work
- Implementation
  - Explored by modeling non-MARC metadata from Cornell Hip Hop Flyer collection using LinkedBrainz
  - Availability of data will influence richness of discoverable context

### Hip Hop flyers



# Pilot: linking Hip Hop flyer metadata to MusicBrainz/LinkedBrainz data

- Model non-MARC metadata from Cornell Hip Hop Flyer Collection in RDF
  - Test LD4L BIBFRAME for describing flyers originally catalogued using ARTstor's Shared Shelf
  - Use Getty Art & Architecture Thesaurus to create bf:Work sub-classes
  - Test the use of other ontologies for describing other entities including Event ontology and Schema.org
- Use of URIs for performers to recursively discover relationships to other entities via dates, events, venues, graphic designers, work types and categories

#### MusicBrainz



Overview Releases Recordings Works Events Relationships Aliases Tags Details Edit

Legal name: Kevin Donovan

#### Wikipedia

**Kevin Donovan** (born April 19, 1957), better known by the stage name **Afrika Bambaataa**, is an American DJ from the South Bronx, New York. He is notable for releasing a series of genre-defining electro tracks in the 1980s that influenced the development of hip hop culture. Afrika Bambaataa is one of the originators of breakbeat DJing and is respectfully known as "The Godfather" and "Amen Ra of Hip Hop Kulture," as well as the father of electro funk. Through his co-opting of the street gang the Black Spades into the music and culture-oriented Universal Zulu Nation, he has helped spread Hip Hop Culture throughout the world.

Continue reading at Wikipedia... Wikipedia content provided under the terms of the Creative Commons BY-SA license

#### Discography

#### Album

Year	Title	Artist	Rating	Releases
1984	Unity	Afrika Bambaataa & James Brown	www	1
1986	Beware (The Funk Is Everywhere)	Afrika Bambaataa	****	2
1988	The Light	Afrika Bambaataa	statatat	1
2004	Dark Matter Moving at the Speed of Light	Afrika Bambaataa	****	2
_	Zulu Nation War Chant	Afrika Bambaataa	statatate	1

#### Album + Compilation

Year	Title	Artist	Rating	Releases
1997	Zulu Groove	Afrika Bambaataa	***	1
1999	Electro Funk Breakdown	Afrika Bambaataa	***	2
1999	The 12" Mixes	Afrika Bambaataa	****	1
2001	Looking for the Perfect Beat 1980-1985	Afrika Bambaataa	****	3
2003	Afrika Bambaataa Presents Eastside	Afrika Bambaataa	****	1

#### Reconciling mo:Release with bf:Audio



Releases

Legal name: Kevin Donovan

#### Wikipedia

Overview

**Kevin Donovan** (born April 19, 1957), better known DJ from the South Bronx, New York. He is notable for 1980s that influenced the development of hip hop cul breakbeat DJing and is respectfully known as "The Gc the father of electro funk. Through his co-opting of th culture-oriented Universal Zulu Nation, he has helped

Recordings

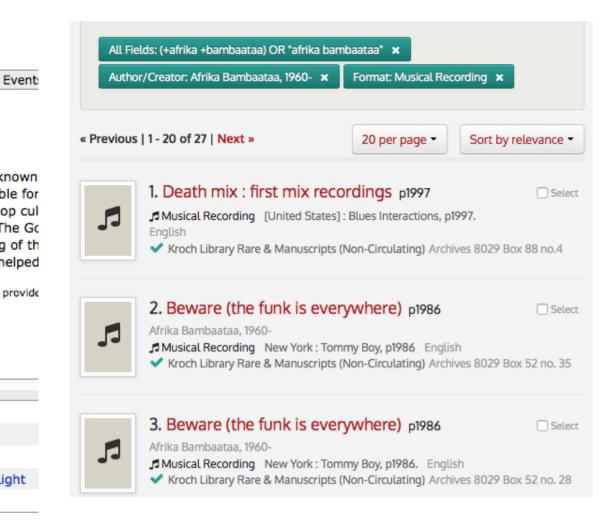
Works

Continue reading at Wikipedia... Wikipedia content provide

#### Discography

#### Album

Year	Title
1984	Unity
1986	Beware (The Funk Is Everywhere)
1988	The Light
2004	Dark Matter Moving at the Speed of Light
_	Zulu Nation War Chant



## **Takeaways**

- Able to map large parts of our metadata to RDF using multiple ontologies to discover more relationships to more entities (still some mapping and reconciliation work to do)
- Largely predicated on manual workflows for preprocessing, URI lookups, and unstable software for RDF creation
- Need more URIs for both linking to and linking from in order to take advantage of queries and patterns

Discussion: What other kinds of context do libraries have that we could add as linked data to aid in discovery and understanding?

# Assembling the Ontology

LD4L data sources

#### Bibliographic Data

- MARC
- MODS
- EAD

#### Person Data

- CAP, FF, VIVO
- ORCID
- ISNI
- VIAF, LC

#### Usage Data

- Circulation
- Citation
- Curation

# BIBFRAME basic entities and relationships

• Creative Work - a resource reflecting a conceptual essence of the cataloging resource.

• Instance - a resource reflecting an individual, material embodiment of the Work.

Authority - a resource reflecting key authority concepts that have defined relationships
reflected in the Work and Instance. Example of Authority Resources include People, Places,
Topics, Organizations, etc. One important concept in Authority is domain, which is the entity
taking responsibility for the recognition, organization and maintenance (to ensure integrity) of
the authoritative resources.

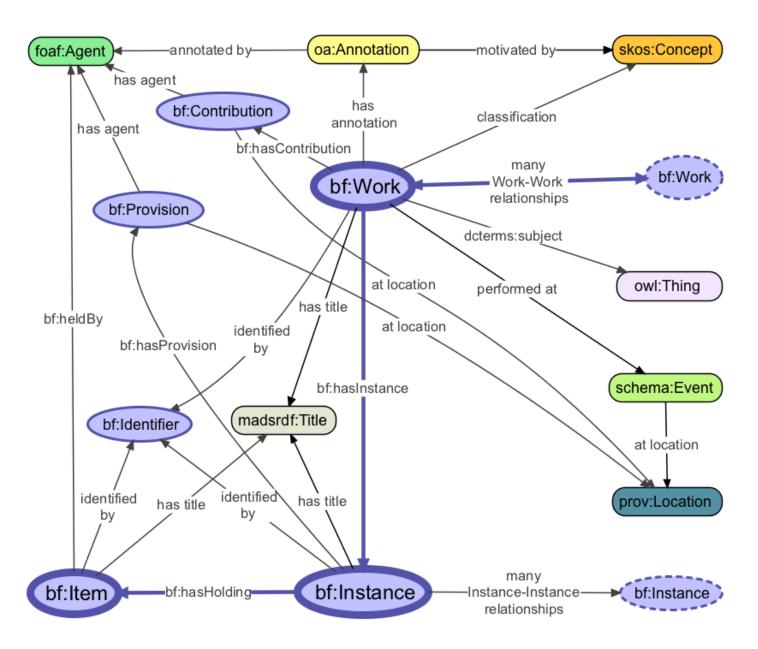
 Annotation - a resource that enhances our knowledge about another resource when knowing, minimally, 'who' is doing the annotating is important. Library Holdings, Cover Art and Reviews are examples types.

subject creator Work hasInstance Instance publisher format publishedAt

http://bibframe.org/vocab-model/

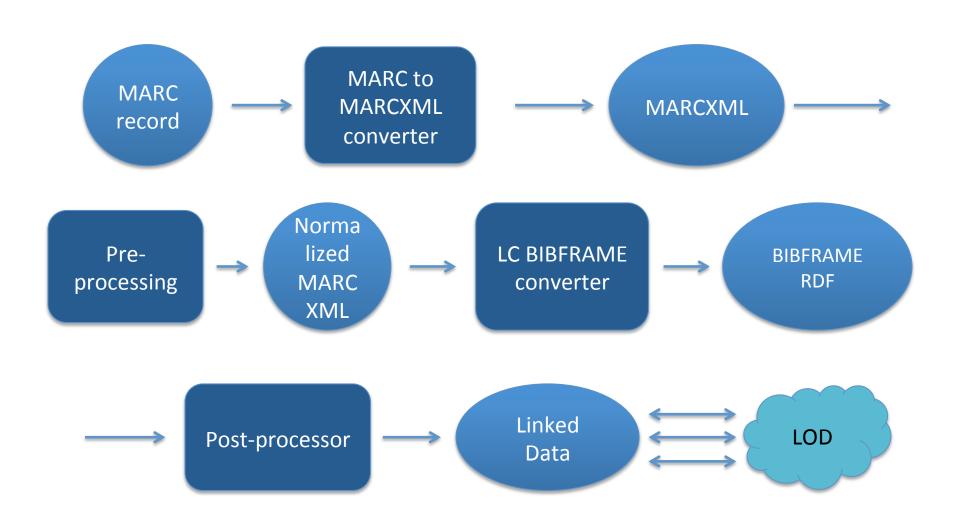
# Linked data best practices (partial) from Sanderson report

- Clarify and limit scope
- Use URIs in place of strings (identification of the resource itself vs. resource description)
- Reuse existing vocabularies and relate new terms to existing ones
- Only define what matters (and inverse relationships do)
- Remove authorities as entities in favor of real world URIs
- Reuse the Open Annotation ontology vs. reinventing the wheel



LD4L Ontology (simplified for display)

#### Stages of LD4L data transformation



#### Future processing challenges

- Join with VIVO/CAP/Profiles data as a coherent, richer local authority picture
- Extend to full variety of different types of catalog records
- Address issues of entity resolution and linking in the real world for works, people, organizations, events, places, and subjects
- Integrate with other linked data sources via common global identifiers and shared ontologies

# LD4L Workshop: Discussion and Outcomes

#### LD4L Workshop

- February, 2015 at Stanford
- 50 attendees doing leading work in linked data related to libraries, from around the world
- Review & vet the LD4L work done to date
  - Use cases
  - Ontology
  - Technology
  - Prototypes
- Plot development moving forward

https://wiki.duraspace.org/display/ld4l/LD4L+Workshop+Agenda

#### **Topics**

- Curation of Linked Data
- Techniques & Technology
  - Entity resolution (strings to things)
  - Reconciliation (things to things)
  - Converters & validators
- New Uses, Use Cases & Services (WHY?)
- Community (WHO)

#### Workshop Recommendations

- Our goal should be that others outside the library community use the linked data that we produce
- We must create applications that let people do things they couldn't do before – don't talk about linked data, talk about what we will be able to do
- Local original assertions (new vs. copy cataloging) should use local URIs even when global URIs exist
- Look to LD to bring together physically/organizationally dispersed but related collections
- Libraries must create a critical mass of shared linked data to ensure efficiency and benefit all of us

#### Opportunity



LD4L -> LD4Everyone

Library data on the Web, **fed** and **used** by the entire Web

#### **Current Capacity**



- Clear business case / value proposition
- Ontology
- Community (who is "we"?)
- Tools & Infrastructure

#### Challenges



- Ontology & vocabulary mapping
- Demonstrate uses & services
- Build / polish tools for everything! conversion, reconciliation, management, publishing, editing, and more!
- Develop Community
  - Knowledge experts with technologists
  - Among libraries
  - Among LAMs
  - With Industry
  - With Internet

Discussion: What are the challenges for my library in moving forward toward a Linked Data future?

## Summary



#### **Project Outcomes**

- Open source extensible LD4L ontology compatible with VIVO ontology, BIBFRAME, and other existing library LOD efforts
- Open source LD4L semantic editing, display, and discovery system
- Project Hydra compatible interface to LD4L, using ActiveTriples to support Blacklight search across multiple LD4L instances

#### LD4L Team



### Questions?

More Info: <a href="http://ld4l.org">http://ld4l.org</a>