TEXAS STATE
UNIVERSITY LIBRARIES

Bringing Linked Data into Libraries via Wikidata

PCC Wikidata Pilot Project @
Texas State University

Mary Aycock, Database and Metadata Management Librarian
Nicole Critchley, Assistant Archivist, University Archives
Amanda Scott, Wittliff Collections Cataloging Assistant

https://digital.library.txstate.edu/handle/10877/14815

Presentation Outline

- The problem of hidden metadata
- Introduction of projects
- Methodologies
- Impacts
- Challenges & Summary

The Problem: Hidden Metadata

Currently carefully curated metadata in GLAMs often isolated in various silos

- MARC-based library catalogs
- PDF finding aids
- Authority records in traditional catalogs and digital repositories

001			n 93022619
003			DLC
005			20210205070216.0
008			930311 n | a c a n n a a b n | n a a a
010			n 93022619
024	7		Q21520276|2wikidata|1https://www.wikidata.org/entity/Q21520276
035			(OCoLC)oca03332939
040			DLC|beng|cDLC|dTxSmTSU
100	1		Mclean, Robert J. C.
373			Texas State University|2naf
670			Immobilized biosystems, c1993:|bCIP t.p. (Robert J.C. Mclean, Dept. of Microbiology and Immunology, Queen's University, Kingston, Canada)
670			Texas State University Department of Biology faculty directory, viewed February 4, 2021|b (Robert McLean, Ph.D., Regents' Professor; Research Interests: In my lab we study the biology of bacteria as they naturally grow in surface-adherent communities (biofilms))

Linked data has been proposed as a solution to de-silo this rich metadata

Wikidata as a Possible Solution?

- Linked open data platform
- Easy to use graphical interface
- Stability: Created in 2012 as part of the Wikimedia Foundation
- Ready-made infrastructure (Wikibase) & ontology
- Shared values of knowledge creation
- SPARQL querying
- Used by Google and other agents (Siri, etc.)

Program for Cooperative Cataloging Wikidata Pilot

- Program for Cooperative Cataloging (PCC): International coalition focused on metadata in GLAM institutions
- Goals of PCC pilot
  - Comparing ease of use and benefits of Wikidata to other registries (LCNAF, ISNI)
  - Assessing the productivity and quality assurance tools that exist (or should exist)
  - Learning about the culture of the Wikidata community
- Benefits to our library
  - Gaining familiarity of working
  - Applying linked data knowledge gained over the past years
  - Bringing more visibility to affiliated persons, organizations, collections, and projects within our university

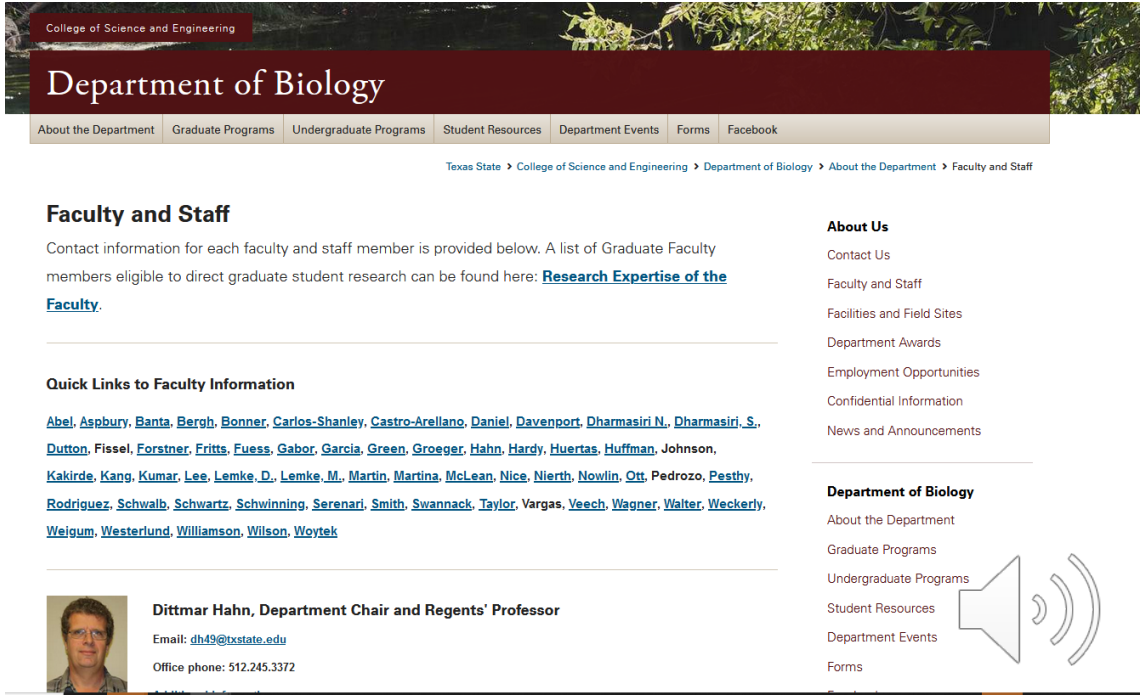
Pilot Project Introductions

Increasing visibility of faculty and collections

- Faculty
  - Participate in creating/enhancing Wikidata items for faculty outside the traditional authority silo
- University Archives oral history collections
  - Increase access and visibility to oral histories by creating entries with rich biographical information not easily discovered otherwise
- Wittliff Collection finding aids
  - Increase access and visibility to collection by linking to PDF finding aids

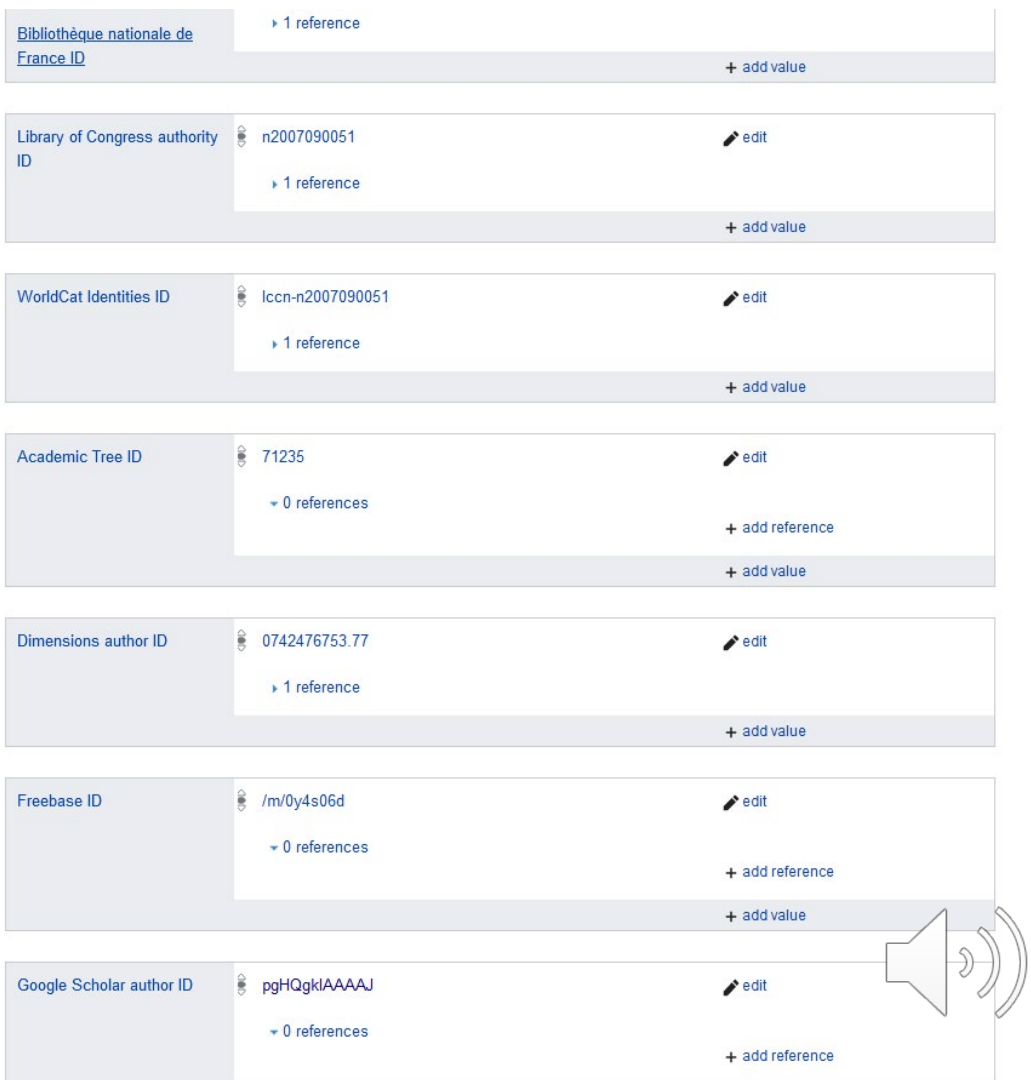
Texas State University Faculty Project: Identity Management Orientation

- Using ***identifiers*** instead of unique strings (names)
  - Strings → Things
- Using data external to library silos
- Empowering communities to create own items



Hub of Identifiers

Snippet of identifiers for Dr. Jill Pruetz: https://www.wikidata.org/wiki/Q11727176



University Archives oral history collection

- Bring attention to oral histories and the rich demographical data that exists for the participants
  - Bits and pieces of demographic data in several places, spreadsheets (not publicly available), website, the collection or PDF finding aid (not easily searched)
- Go to where the people already are, Wikipedia and Wikidata
- Help us connect these oral histories to other collections, increasing exposure

Wittliff Collections finding aids and archival resources

- Focusing primarily on adding links to finding aids
  - Where no Wikidata item exists a new one is created
- Generating more exposure for finding aids, associated collections, and the Wittliff Collections as a whole
  - Raising awareness for Wittliff archival holdings and associated entities
- Ultimately, creating a single base for all publicly available biographical data

Methodologies

Methodologies

- Each project worked independently according to expertise
- Researched and compiled data from multiple sources and created spreadsheets, i.e. faculty, special collections, oral history subjects
- Created data models via WikiProjects template: https://www.wikidata.org/wiki/Wikidata:WikiProject_PCC_Wikidata_Pilot/Texas_State_University_Libraries

Data Model for Oral History Project

Basic statements [edit]

Property	Value	Usage note
Label	Person's name as given on university website	Do not include titles or degrees, if LCNF exists, use that instead
Alias	Other form of name in use or previously used	
Description	based on field of work	See description guidelines (concise, lower case)
instance of (P31)	human (Q5)	no reference needed
occupation (P106)	Based on field of work	
date of birth (P569)		Careful with PII (Personal Identifiable Information) for living persons
date of death (P570)		
educated at (P69)	If available	
has works in the collection (P6379)		Texas 150: Sesquicentennial Oral History Project (Q98832214)
archives at (P485)	Include if other collections within University Archives or another institution exist	Include permalink if available
SNAC ARK ID (P3430)	Include if already exists	Include permalink
Library of Congress authority ID (P244)	LCCN from national authority record, if exists	Include permalink
Find A Grave memorial ID (P535)	Include if already exists	Include permalink

archives at (P485)

the institution holding the subject's archives

papers at | correspondence at | archive location

has works in the collection (P6379)

collection that has works of this person or organisation (use archive location P485 for the archives)

works in collection | has works in the collection(s) | work in collection | has work in the collection | museum housing this person's work

oral history at (P9600)

No description defined

oral histories at | oral history in collection at | oral history in collection of | oral history held by

Wittliff collections (Q8028599)

Special collections at Texas State University

University Archives, Texas State University (Q98802224)

Administrative and historical archives of Texas State University

University Archives

Texas 150: Sesquicentennial Oral History Project (Q98832214)

oral history project celebrated the State's 150th anniversary in 1986, interviewees shared memories about Texas State University and life in (edit
Texas, offering a glimpse into important people, places, and events

Texas 150

Methods and Goals

- Editing items
  - Manually edited existing records OR
  - Used tools such as OpenRefine or Quick Statements to bulk create Wikidata items
- Goals:
  - Faculty project: Create/edit records for at least 100 faculty
  - Oral History project: Create or enhance records for the subjects of the Texas150 Oral History collection (70)
  - Wittliff project: Add “Archives at” property for finding aids on A-Z index (152)

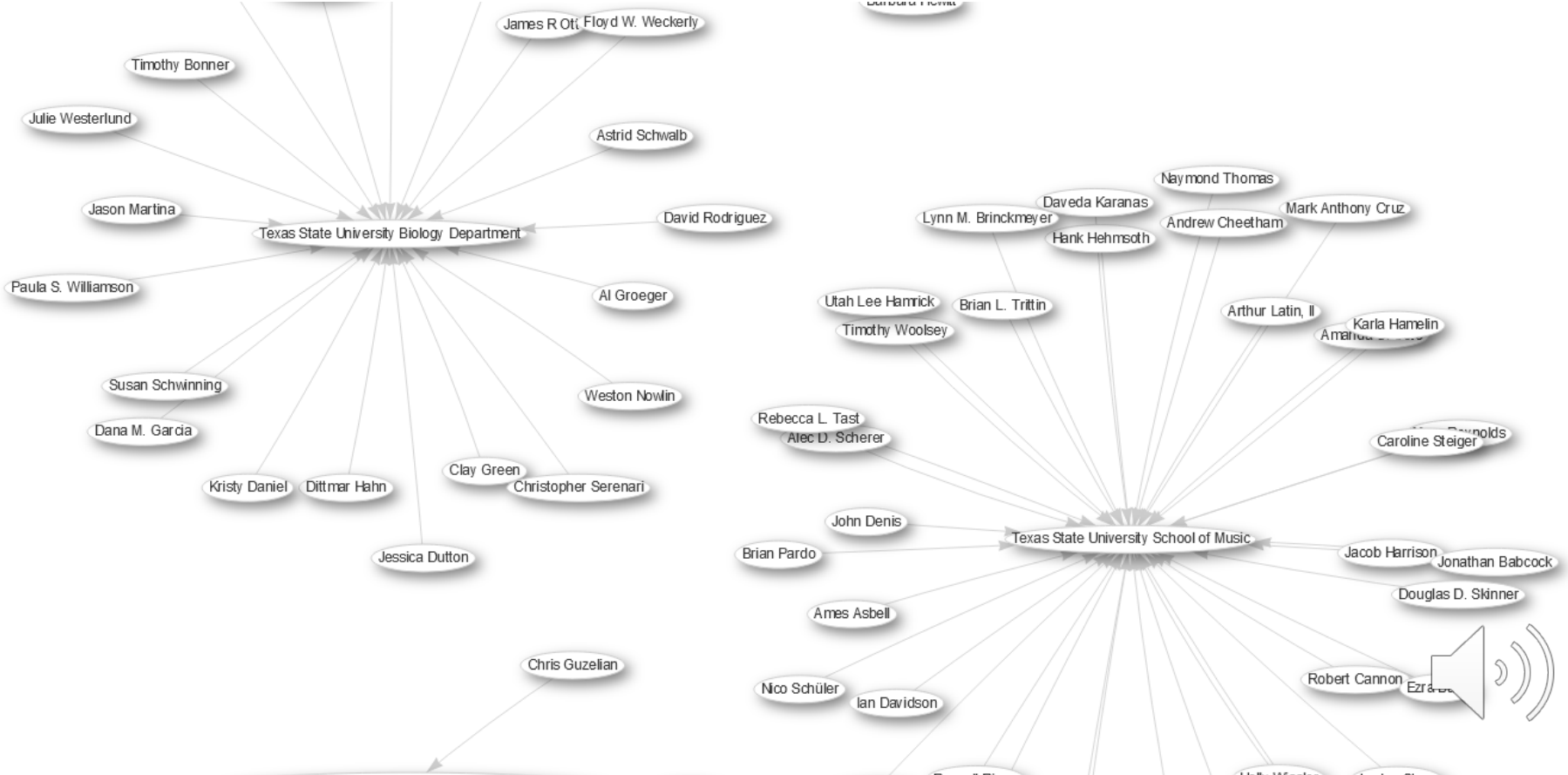
Wittliff Archives Collections

- Of the 152 archives in our Southwestern Writers Collection 7 remain to be reviewed
- Analyzing each archive individually to determine need for entry
  - Once current list is complete, Wittliff Gallery's photographer collection will be assessed
- Including reference links to Wittliff Collections, the University Libraries repository and other archival institutions as needed
  - Both Wittliff Collections and the repository sites are being monitored for added collections and available biographical data during the current finding aid transitional period
    - Updates result in earlier Wittliff Writers project list to be outdated
      - Ex. Sergio Troncoso Papers added after the project list was pulled

Abstract:
The Sergio Troncoso Papers span 1975-2020 and are divided into six series: Personal, Published Works, Magazine and Journal Contributions, Publicity, Student Letters, and Digital Objects. The bulk of the collection is drafts of his early work, including *The Last Tortilla and Other Stories*, *The Nature of Truth*, *Crossing Borders: Personal Essays*, and *From This Wicked Patch of Dust*. Also of note are his academic essays from his graduate school studies at Yale.

Impacts

Faculty metadata now on linked data graph



Faculty also linked to their digital scholarship

List of publications RSS

Search:

Date	Work	Type	Pages	Venue	Authors
2020-06-25	Phase II spatial patterning of vulture scavenged human remains	scholarly article		Forensic Science International	Kate Spradley
2017-02-20	Morphological variation among late holocene Mexicans: Implications for discussions about the human occupation of the Americas.	scholarly article		American Journal of Physical Anthropology	Kate Spradley, Brianne Herrera, Mark Hubbe
2016-09-01	Metric Methods for the Biological Profile in Forensic Anthropology: Sex, Ancestry, and Stature	review article, scholarly article			Kate Spradley
2016-09-01	The Role of the Anthropologist in the Identification of Migrant Remains in the American Southwest	review article, scholarly article			Kate Spradley
2014-02-06	Ancestry assessment using random forest modeling.	scholarly article		Journal of Forensic Sciences	Kate Spradley, Joseph T Hefner

Wikidata Query Service

author: list-of-publications.sparql

Service to university

SpaceX launch carries Texas State's experiment on biofilm formation to International Space Station

RESEARCH & INNOVATION Jayme Blaschke | December 4, 2020

Robert J. McLean (Q21520276)

microbiologist

R.J.McLean | Robert McLean | Robert J.C. McLean | Robert Mclean | Robert McLe

Recoin: Most relevant properties which are absent

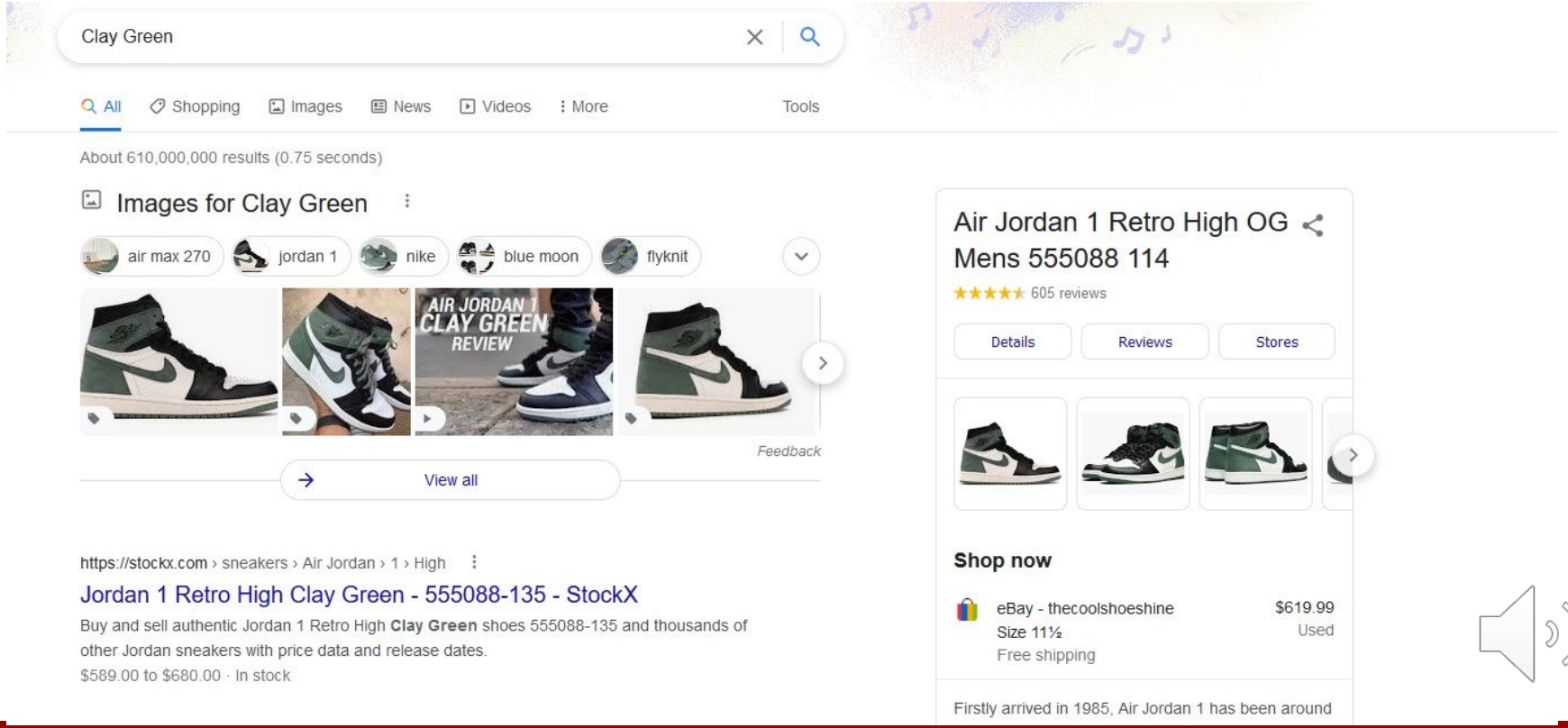
In more languages

Statements

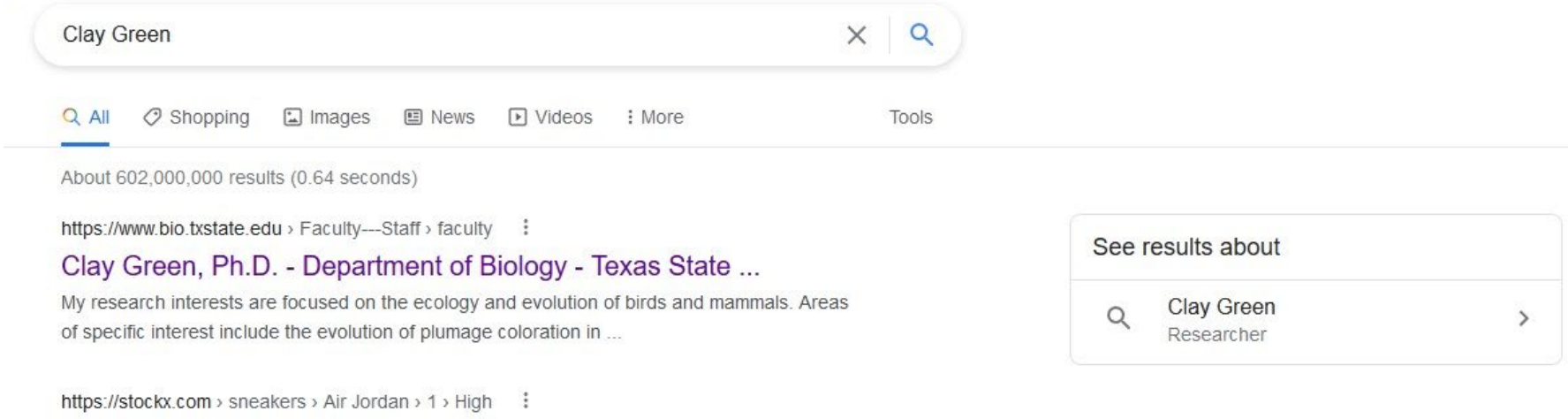
instance of	human
	0 references

Tangible results in search results

We have seen an increase in Google rankings, particularly for those faculty with names easily confused with other entities: For instance, a search of “Clay Green” yielded shoes:



Dr. Clay Green two weeks after Wikidata work



University Archives oral histories impact:

In the beginning

Wikipedia page?	Wikidata page?	LCNAF?	UA webpage	Dspace link	SNAC entry	Related collections?
3 (plus one mentioned in related article)	3	16	70	69	14	6

Now, all 70 participants have rich entries that link out to related collections

Ruth Bain example

Bain was interviewed for one of our oral history projects, but we don't have her papers.

Wikidata allows us to connect both of these.

archives at	Austin History Center	edit
	2 references	
	reference URL	https://legacy.lib.utexas.edu/taro/aushc/00136/00136-P.html
	retrieved	1 September 2020
		+ add reference
		+ add value

has works in the collection	Texas 150: Sesquicentennial Oral History Project	edit
	2 references	
	reference URL	https://www.univarchives.txstate.edu/research-collections/core-collections/070-media-artifacts/media-oral-history/oral-history-name-index/bain-ruth.html
	retrieved	1 September 2020
		+ add reference
		+ add value

Texas 150 Oral History collection monthly Digital Collections usage statistics

Website options:

PDF Transcript- October 2, 1986

HTML Transcript- October 2, 1986

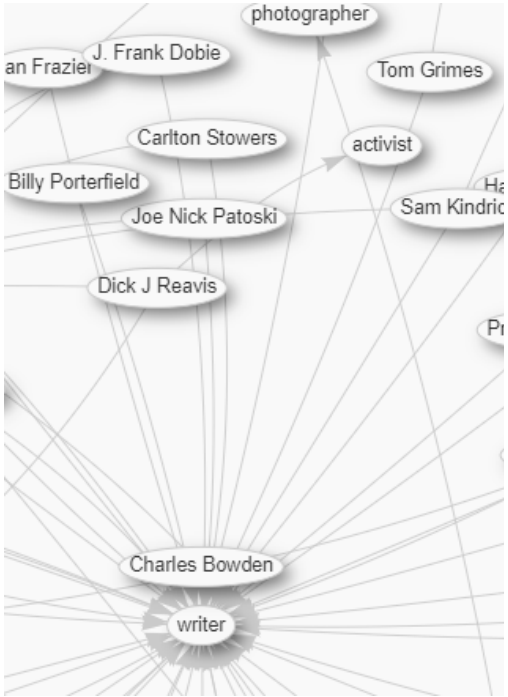
Links and data collection started

Creating entries and adding links to Wikidata

Texas 150 Oral History collection monthly Digital Collections usage statistics			
month	collection views	item views	downloads
2019-09	10	84	271
2019-10	1	98	282
2019-11	6	97	199
2019-12	6	54	72
2020-01	13	1600	149
2020-02	4	94	116
2020-03	2	93	82
2020-04	7	124	110
2020-05	16	193	146
2020-06	13	232	0
2020-07	16	223	58
2020-08	8	276	121
2020-09	3	1220	124
2020-10	6	1773	180
2020-11	7	68	180
2020-12	5	1683	66
2021-01	6	1583	48
2021-02	80	1024	10
2021-03	21	1879	27
2021-04	7	1712	21
2021-05	4	1554	56
2021-06	18	1701	37
2021-07	9	1935	38
2021-08	3	1982	30
2021-09	5	1758	47
2021-10	10	2036	32

Wittliff Archives Collections

- Wikidata impacting technical user
  - Revisiting entities to assess need for additional data
  - Adding values to existing popular entities
    - Ex. Lonesome Dove (mini-series) ‘cast member’ property
  - Recognizing the need for the creation of more linked data to yield high returns and inspire wide data creation that benefits all users

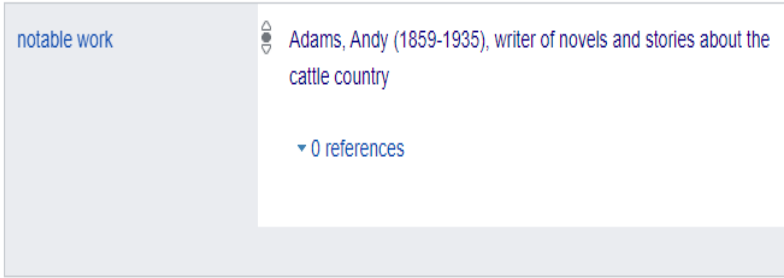


Additional work Wikidata is inspiring

- Incorporating the addition of InfoBoxes in Wikipedia pages for individuals
- Editing Wikipedia pages after discovering errors in Wittliff Collections references

Simmons has, since 2008, donated his papers to an archive in the Witliff Collections at Texas State University.[2][4]

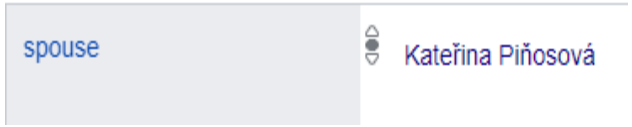
  - Corrected spelling and added link to Wittliff Collections

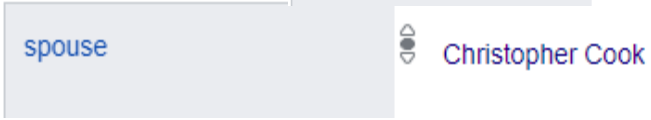


- Linking existing Wikidata pages to relevant individual's page (ex. Wilson M. Hudson's encyclopedia entry)
  - Hudson lacked an entry in Wikidata but was present in article as "author name string'
  - Creation of individual page resulted in ability to link Hudson as 'author'

Wikidata inspires even more

- Discovering and correcting bot data errors
- Reconciling missing properties in linked pages
  - Connecting related entities to one another

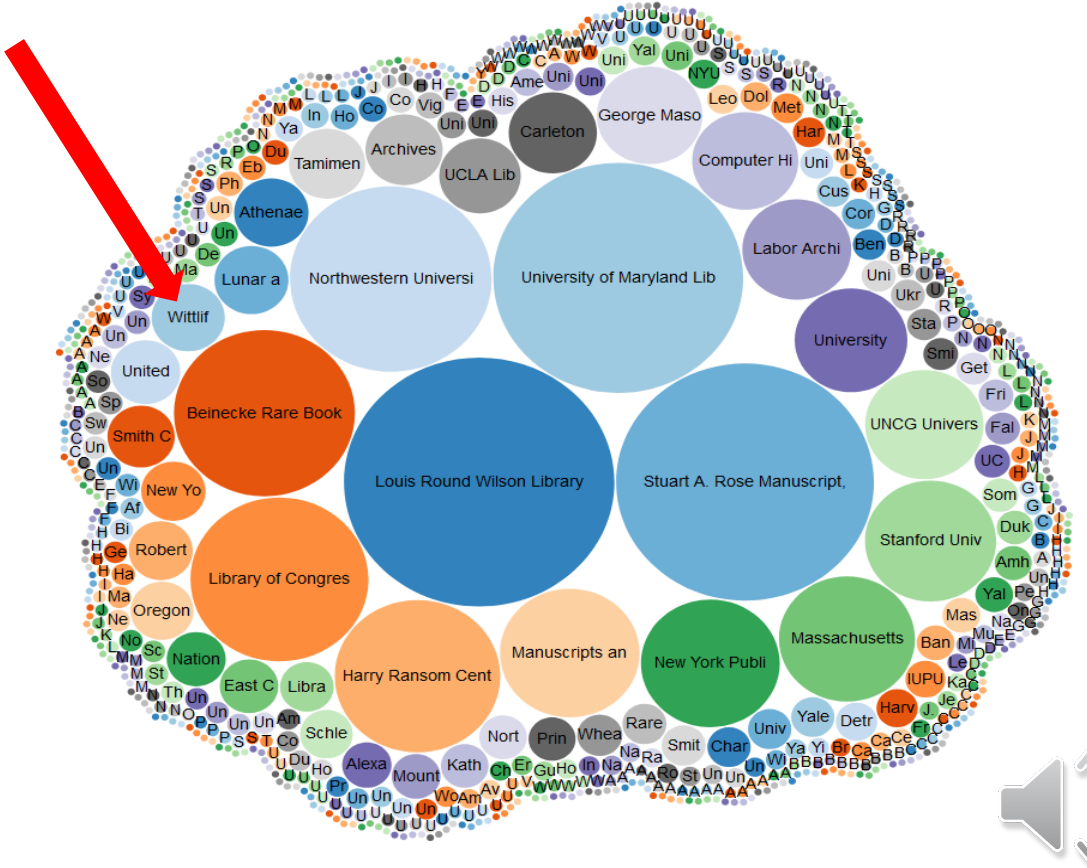




SPARQL query

Wittliff Collections' visibility in comparison with all US participants.

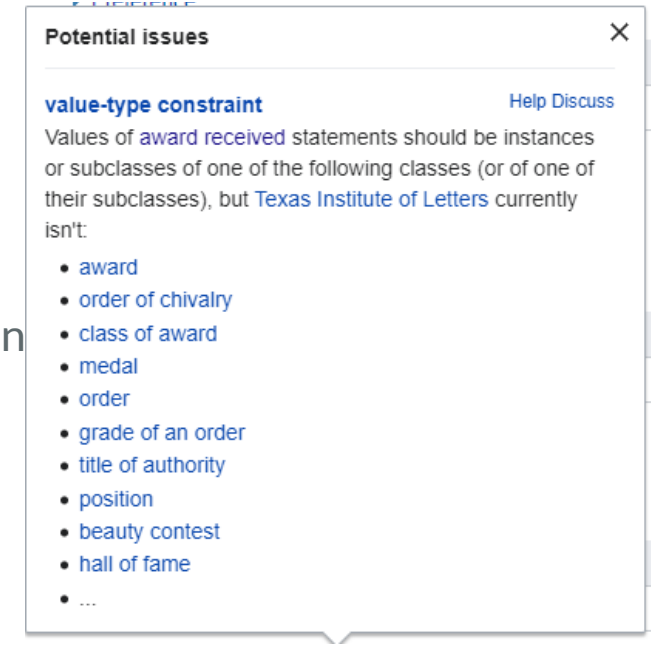
Only includes pages with an 'Archives at' property.



Challenges and Conclusion

Addressing the learning curve

- Defining properties
  - Ex. Difference between start/end time and point in time qualifiers
- Unavailable or undefined headings
  - Leads to a need to create more pages
    - Cannot link to an item that does not exist
  - Experienced primarily in the ‘Awards received’ property
- Revisiting pages created early in the project to apply newly learned skills



Data maintenance

- Keeping up with changes
  - Now we need to go back and update with new ArchivesSpace links
  - TARO finding aids - new platform and the redirect will only work for a few months
  - Migration: Future plans for new Digital Collection repository platform
- General link rot
- Bots adding PII (Personal Identifiable Information)

Addressing questions of sustainability

- Can be difficult to find time in existing workflow
    - Side project status
- However, library silo systems aren't sustainable and also relatively invisible. What makes the most sense to put our efforts toward?
- Wikidata exposes our data to the public and others can build upon it, breaking down silos. It's what makes Wikidata a good resource.

Further Considerations & Avenues

- Investigating adding citations to Wikidata
  - Titles from faculty authors and the Wittliff Writers archive
  - Articles from digital journals to Wikidata
- Assessment of Wikidata work
- Continue developing expertise in finding aids
- Expand editing in Wikipedia
- How to incorporate Wikidata into our platforms?
- Project's future
  - Adding images to Wikimedia Commons to assist in further differentiating between entities
  - Periodic quality control to ensure links remain active
  - Edit or create new pages as more archives become available
  - Explore the possibility of using Wikidata as base for storing local authority records

Conclusion

- Time consuming - rabbit hole of possibilities
- Wikidata can increase visibility
- Forward thinking data initiatives
- Allows us to serve our institution better

Sites specific to project

Bulk data entry guides for OpenRefine and QuickStatements: https://digital.library.txstate.edu/handle/10877/13529

PCC Wikidata Pilot meetings: Meeting Notes, Slides & Recordings

LD4 Affinity Group page: Wikidata:WikiProject LD4 Wikidata Affinity Group/Affinity Group Calls

Texas State Wikidata PCC pilot page with data models: Wikidata:WikiProject PCC Wikidata Pilot/Texas State University Libraries

Further reading

Allison-Cassin, S., & Scott, D. (2018). Wikidata: A platform for your library's linked open data. Code4Lib Journal, 40. https://journal.code4lib.org/articles/13424

Bianchini, C., Bargioni, S., & Girolamo, C. C. P. di S. (2021). Beyond VIAF: Information Technology and Libraries, 40(2), Article 2. https://doi.org/10.6017/ital.v40i2.12959

Carlson, S., Lampert, C., Melvin, D., & Washington, A. (2020). Linked data for the perplexed librarian.

Cooey, N. (2019). Leveraging Wikidata to Enhance Authority Records in the EHRI Portal. Journal of Library Metadata, 19(1–2), 83–98. https://doi.org/10.1080/19386389.2019.1589700

Lemus-Rojas, M., & Odell, J. D. (2018). Creating Structured Linked Data to Generate Scholarly Profiles: A Pilot Project using Wikidata and Scholia. Journal of Librarianship and Scholarly Communication, 6(1). https://doi.org/10.7710/2162-3309.2272

Further reading

Ruttenberg, J. (2019). ARL White Paper on Wikidata: Opportunities and Recommendations. 60. https://www.arl.org/resources/arl-whitepaper-on-wikidata/

Shenoy, K., Ilievski, F., Garijo, D., Schwabe, D., & Szekely, P. (2021). A Study of the Quality of Wikidata. ArXiv:2107.00156 [Cs]. http://arxiv.org/abs/2107.00156

Tharani, K. (2021). Much more than a mere technology: A systematic review of Wikidata in libraries. The Journal of Academic Librarianship, 47(2), 102326. https://doi.org/10.1016/j.acalib.2021.102326

Vrandecic, D. (2013). The Rise of Wikidata. IEEE Intelligent Systems, 28(4), 90–95. https://doi.org/10.1109/MIS.2013.119

TEXAS STATE UNIVERSITY LIBRARIES

Questions?

Mary Aycock, mba18@txstate.edu

Nicole Critchley, ncritchley@txstate.edu

Amanda Scott, als3@txstate.edu

https://digital.library.txstate.edu/handle/10877/14815

ALKEK LIBRARY