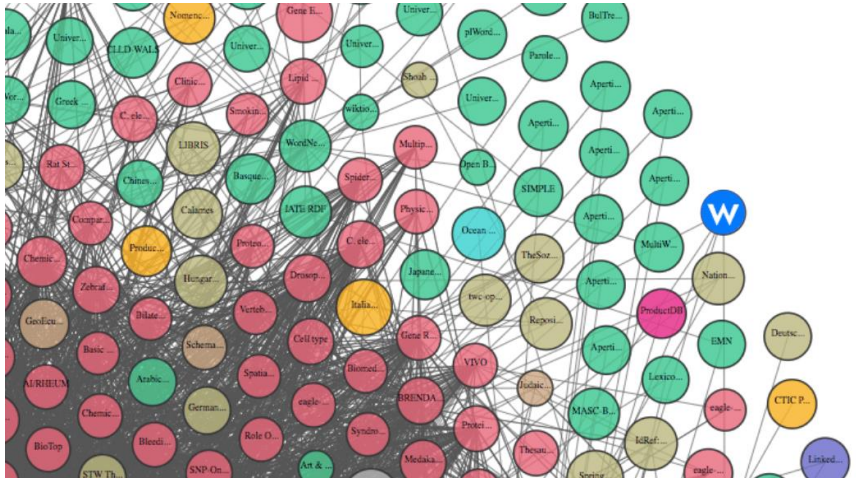


Getting Start with Wikibase : a Linked Open Data Approach to Name Authority Control in Institutional Repositories

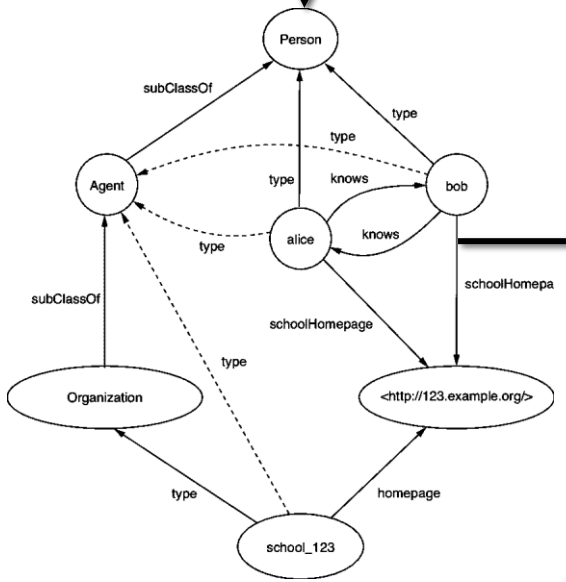
Greta Heng
Email: jieheng2@illinois.edu
GitHub: https://github.com/gretaheng/Wikibase_IR
4/30/2021

Baras, John S.

String

http://viaf.org/viaf/35102971

Permanent Identifiers



Identity Management (Thing)

- differentiating entities
- creating identifiers
- **establishing relationships among entities**

WHERE ARE WE NOW?

Outline

Introduction

Project Parameters

- Project Goals
- Project Value
- Project Design
- Stakeholders
- Project Activities
- Performance Measurement

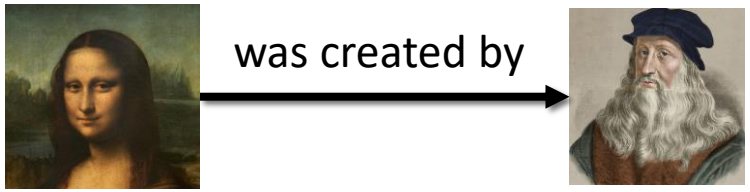
Potential Challenges

Discussion

Conclusions

1. Introduction:
Concepts & Background

- **Name authority control**: the process of choosing one authorized form of the name to provide consistency in its use
- **Institutional repositories (IRs):** publish and archive the research outputs
  - DSpace -> Samvera (in development)
  - Metadata: XML -> RDF



RDF Data Model

Our Library's Institutional Repository

1. Introduction

Concepts

- **Linked open data (LOD)** solutions for authority control
  - Transform names in the format of text strings to web-resolvable entities
  - Link to and virtually merge different authority files
- **Name authority reconciliation:**
  - Reconcile the text strings of names against existing identifiers
  - The ability to communicate among identifier systems is crucial
  - Name authority files / identifiers:
    - Library of Congress Name Authority File (LCNAF)
    - Virtual International Authority File (VIAF)
    - International Standard Name Identifier (ISNI)
    - Wikidata
    - Open Researcher and Contributor ID (ORCID)

Traditional and well supported

Potential authority source and have more granular information

Preferred by funders

1. Introduction:

What is Wikibase? Why Wikibase?

- **Wikibase: name authority database**
  - The software platform underlying Wikidata
  - Open source
  - Provide a robust infrastructure for knowledge graphs, triple stores, and SPARQL queries
  - Design local name authority RDF data model
  - Benefit from Wikidata

2. Project Parameters

1. Project Goals
2. Project Value
3. Project Design
4. Stakeholders
5. Project Activities
6. Project Performance Measurement

2.1 Project Goals

- Ensure that an author/department search on the IR will retrieve all and only those works associated with that author/department
- Build local authority files using Wikibase
- Improve local authority data management and scholarly communication using LOD

2.2 Project Value:
Visibility, Dissemination of Knowledge & Discovery Experience

Name	Num of works
Program on International Policy Attitudes (PIPA)	265
Baras, John S.	233
Shneiderman, Ben	187
Makowski, Armand M.	88
Krishnaprasad, Perinkulam S.	85
Plaisant, Catherine	72
Stewart, G. W.	69
Hanna, William John	68
Goeringer, Paul	65
Abed, Eyad H.	63

Table 1 Ten Authors with the Highest Numbers of Works Published on the Library's IR

Wikidata Property
ethnic group
given name
doctoral advisor
occupation
award received
sex or gender
ISNI
Library of Congress authority ID
Mathematics Genealogy Project ID
Google Scholar author ID
Google Knowledge Graph ID
MR Author ID
educated at
doctoral student
Academic Tree ID
instance of
WorldCat Identities ID
VIAF ID
Nationale Thesaurus voor Auteurs ID
DBLP author ID

Table 2 All Properties of Baras, John S. on Wikidata

2.3 Project Design

Phase I

July 2021-Oct. 2021

- Data Collection
- Information Gathering
- Authority Data Model Discussion

Phase II

Nov. 2021-June 2022

- Wikibase Installation
- Data Cleaning
- Reconciliation
- Authority Data Model Design

Phase III

July 2022-Feb. 2022

- Build Authority Files on Wikibase
- IR Authority Control
- Documentation, Training and Workflow

Phase IV

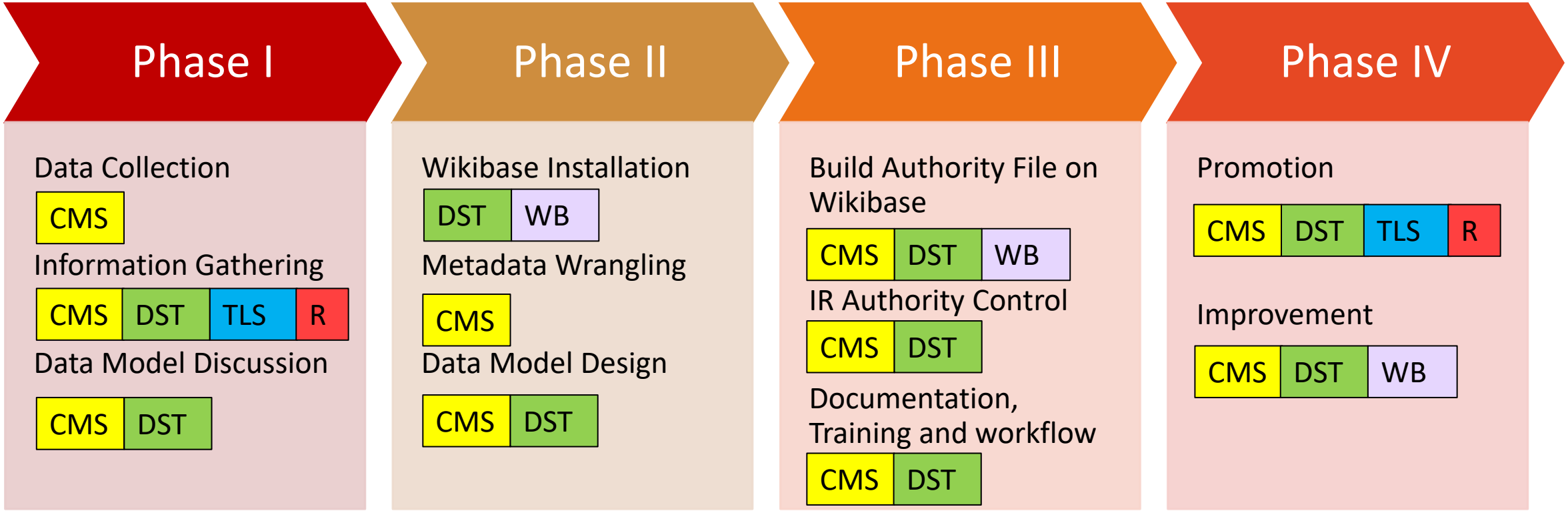
Mar. 2023-June 2023

- Improvement
- Promotion

2.4 Stakeholders

- **Internal Stakeholders**
  - Cataloging and Metadata Services CMS
    - Metadata librarian
  - Digital Services and Technologies DST
    - IT specialists
    - Digital services librarian and data librarian
  - Researchers R
  - Teaching and Learning Services TLS
    - Liaison librarian
- **External Stakeholders**
  - Wikibase WB
    - Developer

2.4 Stakeholders



2.5.1 Project Activities:

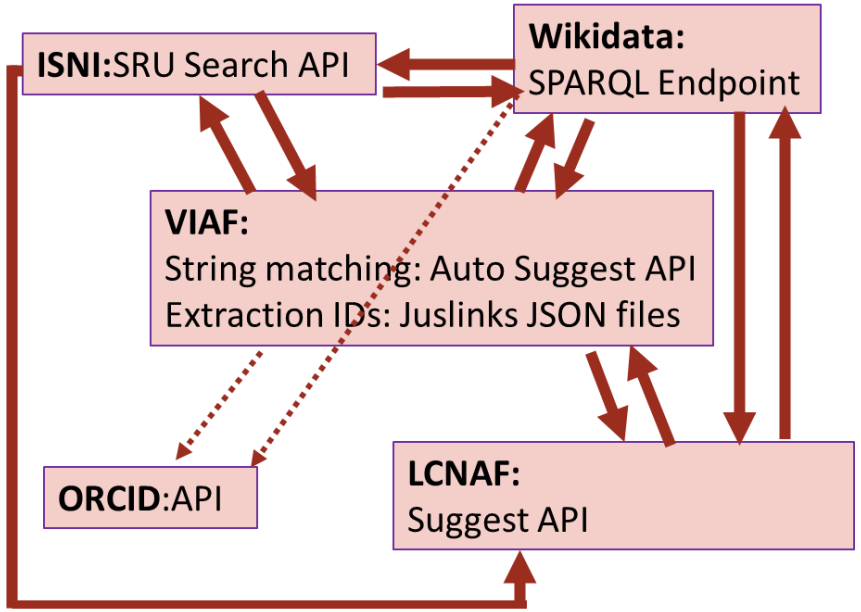
Ensure that an author / department search on the IR will retrieve all and only those works associated with that author / department

- **Data Collection**
  - 32,976 author names on IR (accessed on 4/14/2021):
    - Name, work title, affiliation, deposit time, and subjects
  - University's department
    - Department names, structure, the names of faculty and staff in each department
- **Data cleaning: inconsistency and disambiguation**
  - Pseudonyms and variant names
    - Abbazio, Jessica M. | Abbazio, Jessica Marie
  - Discrepancies in the format
    - Fineman, Yale [transcriber, performer] | Fineman, Yale
  - Misspellings
  - Same name

2.5.1 Project Activities:

Ensure that an author / department search on the IR will retrieve all and only those works associated with that author / department

- **Reconciliation**
  - Selected authority files: LCNAF, VIAF, Wikidata, ORCID, ISNI, & …
  - Methods



Identifier Systems Communication Flowchart

Query against VIAF through its Auto suggest API using Python

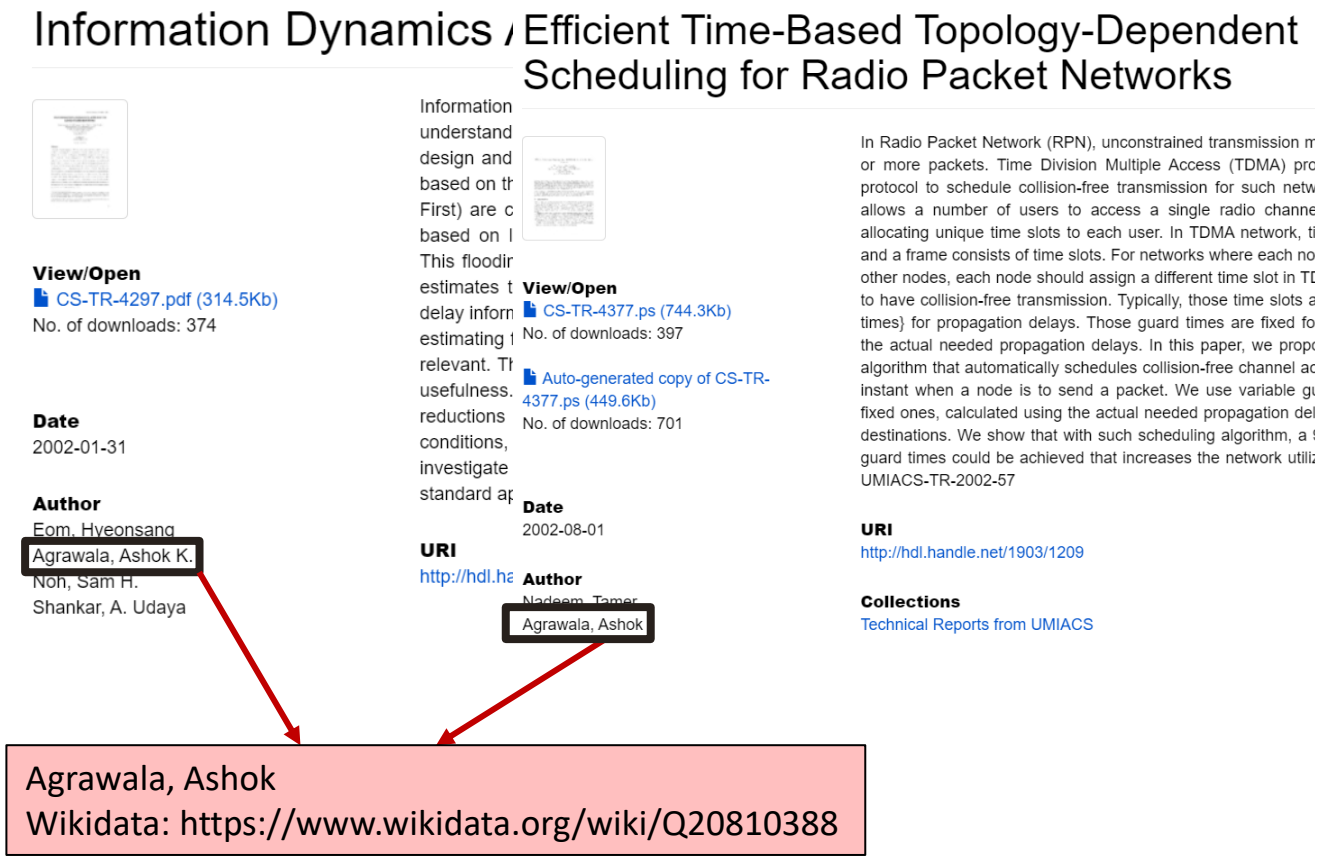
- *IF* return one VIAF ID: parse VIAF Justlinks JSON, get the name's ORCID, ISNI, Wikidata, and LCNAF IDs, if any
- *IF* return many VIAF IDs: use additional data to aid disambiguation

Query against other authority sources through their API services + SPARQL Endpoint using Python

- *IF* return one ID: that's it!
- *IF* return many IDs: use additional data to aid disambiguation

2.5.1 Project Activities:

Ensure that an author / department search on the IR will retrieve all and only those works associated with that author / department



ORCID + REPOSITORIES

DSpaceCRIS

- Open-source extension of DSpace that includes CRIS features
- Developed by CINECA (Italy) and Hong Kong University
- Includes ORCID full-featured ORCID integration

469.0 KB

Date

2020-06-12

Authors

Zhu, He

Wu, Li-Tzy

DukeSpace, Institutional Repository at Duke Univesity

2.5.2 Project Activities:

Build local authority files using Wikibase

➢ **Investigation**

- Data model
- Descriptive metadata workflow
- Technical difficulties

➢ **Data Model**

- Entities:
  - Person
  - Corporate
- Relationships
  - Person and person
  - Corporate and corporate
  - Person and corporate
- Required attributes for entities:
  - Person: preferred name, gender…
  - Corporate: preferred name …
- Required relationships:
  - Person and corporate: employer, has employee
- Optional attributes
  - Person: field of work, pseudonyms….
  - Corporate: variant names…
- Optional relationships: has student, has-sub-unit…

2.5.2 Project Activities:
Build local authority files using Wikibase

➢ **Objects**

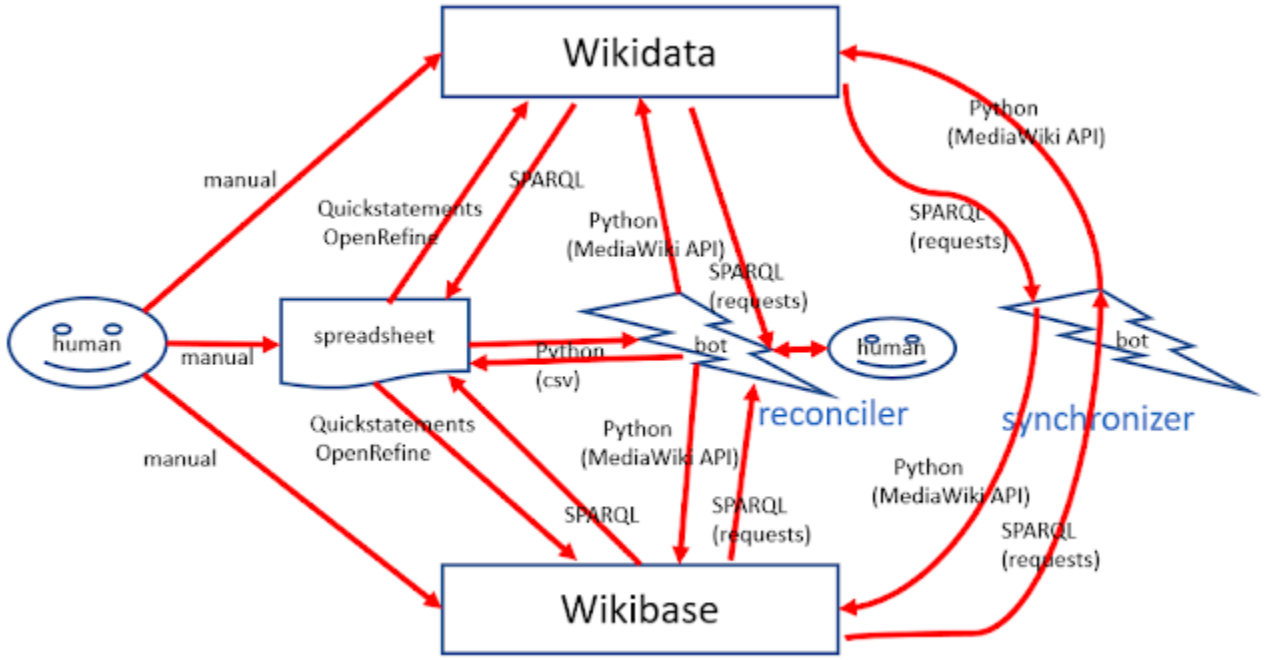
- All current departments
- Scholars who have identifiers in at least one selected name authority system
  - Raw data: 28,388 of 32,976 names only has one work deposited on IR

Num. of Works on IR	Num. of Scholars
1	28,388
2-10	4,331
11-50	239 ★
51-265	18 ★

Statistics about Number of Works on IR and Number of Scholars

2.5.2 Project Activities:
Build local authority files using Wikibase

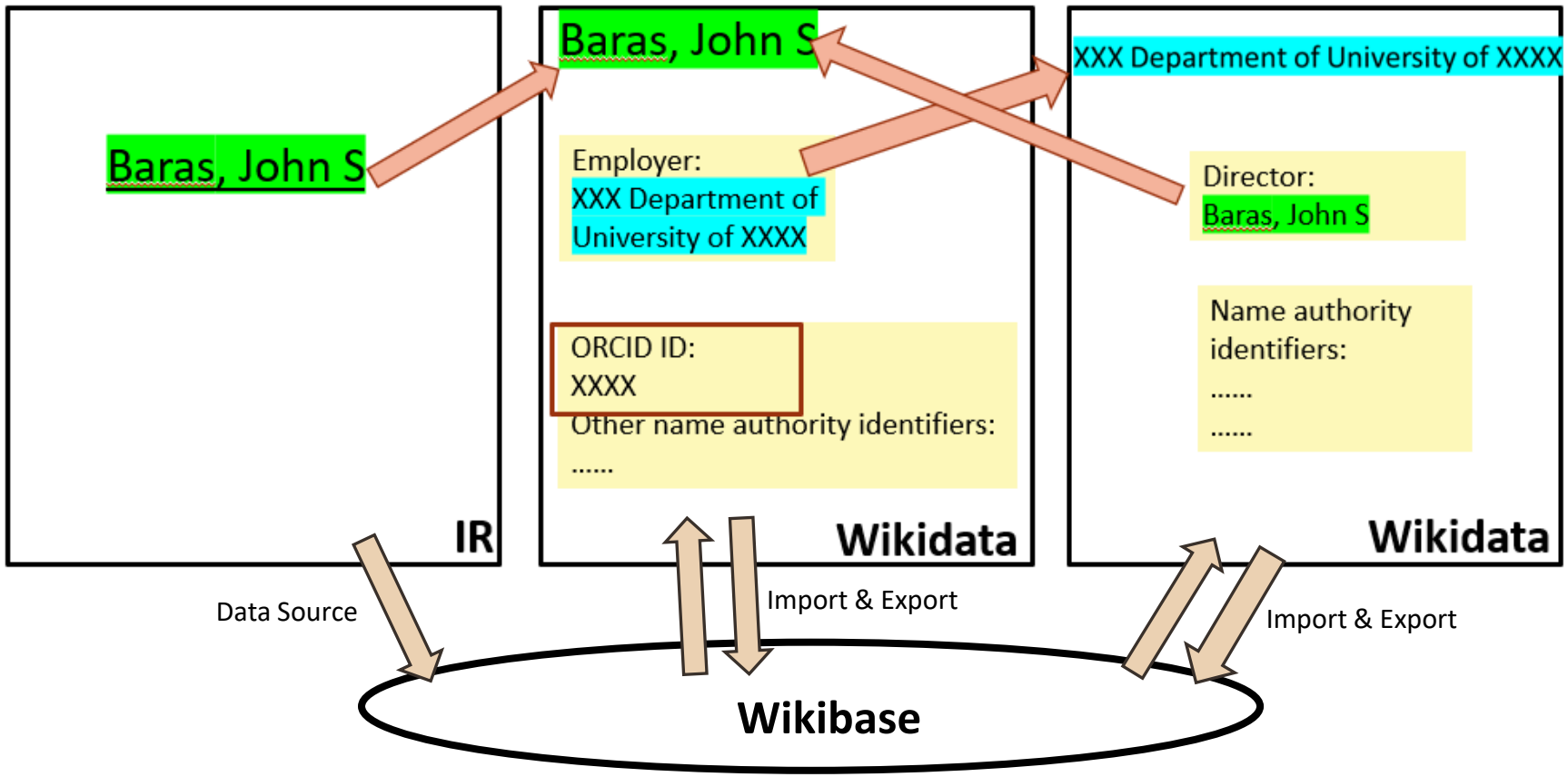
- **Draw data from Wikidata & push data to Wikidata**
  - Name has Wikidata page
  - Name does not have wikidata page but exist in other selected authority files
- **Approaches**
  - Manual work
  - OpenRefine
  - SPARQL
  - WikibaseImport / WikibaseIntegrator
  - Bot: make edits without the necessity of human decision-making



[VanderBot, built by Vanderbilt University]

2.5.2 Project Activities:

Build local authority files using Wikibase



2.5.3 Project Activities:

Improve local authority data management and scholarly communication

- **Documentation, training, and workflows**
  - Stakeholder considerations
  - Update IR deposit instruction
  - Provide trainings for staff and librarians in the DST
  - Create a WikiProject for this project, update progress and documentation on WikiProject, explain the data model there
- **Promotion**
  - Present progress of the project at potential conferences and webinars
  - Encourage the university community to register their name in authority systems

2.6 Project Performance Measurement

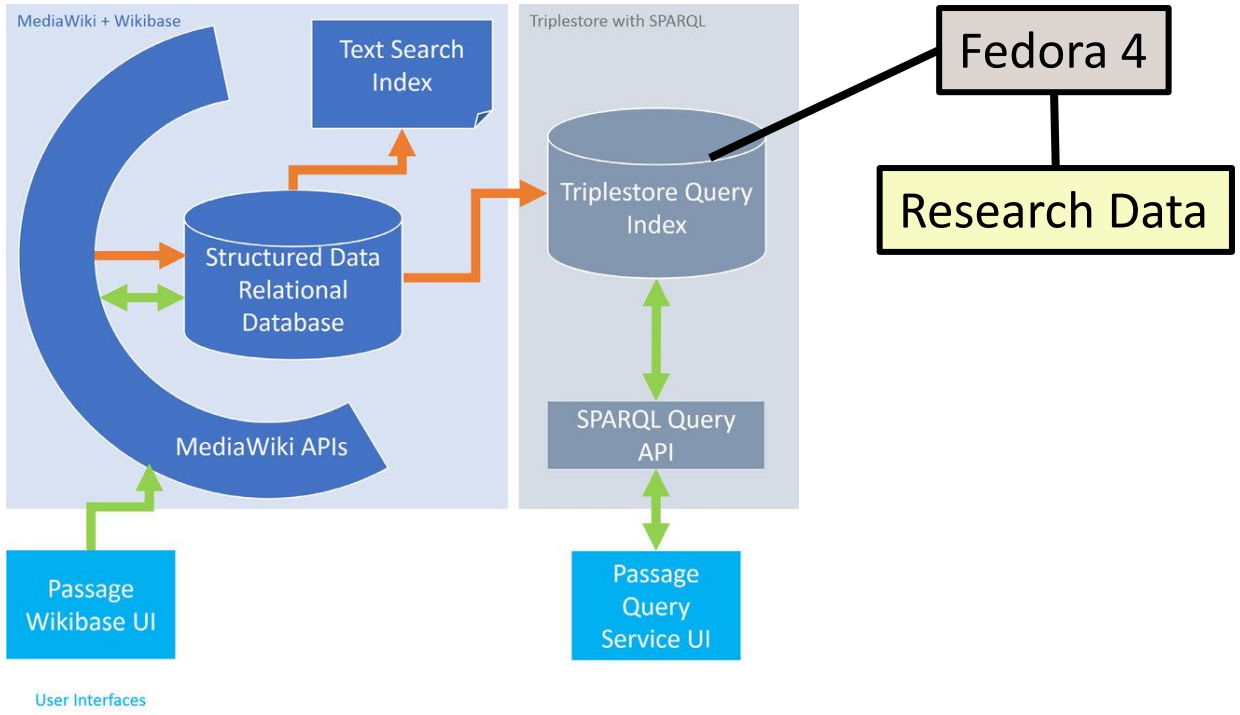
- On-time: complete activities on-time
- Performance: achieve project goals
- Stakeholder support & engagement:
  - Gain support
  - Translate needs
  - Encourage contribution

3. Potential Challenges:

- **Metadata challenges:**
  - Inconsistency and ambiguation
    - Abbazio, Jessica M. | Abbazio, Jessica Marie
  - Identifier deduplication
    - Name: 'Shneiderman, Ben’
      - 6 VIAF IDs: ['108743798', '870156991005861180004', '310043169', '638616109897692964005', '309041480', '121472706750357000 7’]
        - Name identifiers: '108743798', '870156991005861180004’
        - Expression identifiers for the author’s book:'310043169', '638616109897692964005', '309041480', '121472706750357000 7’

3. Potential Challenges:

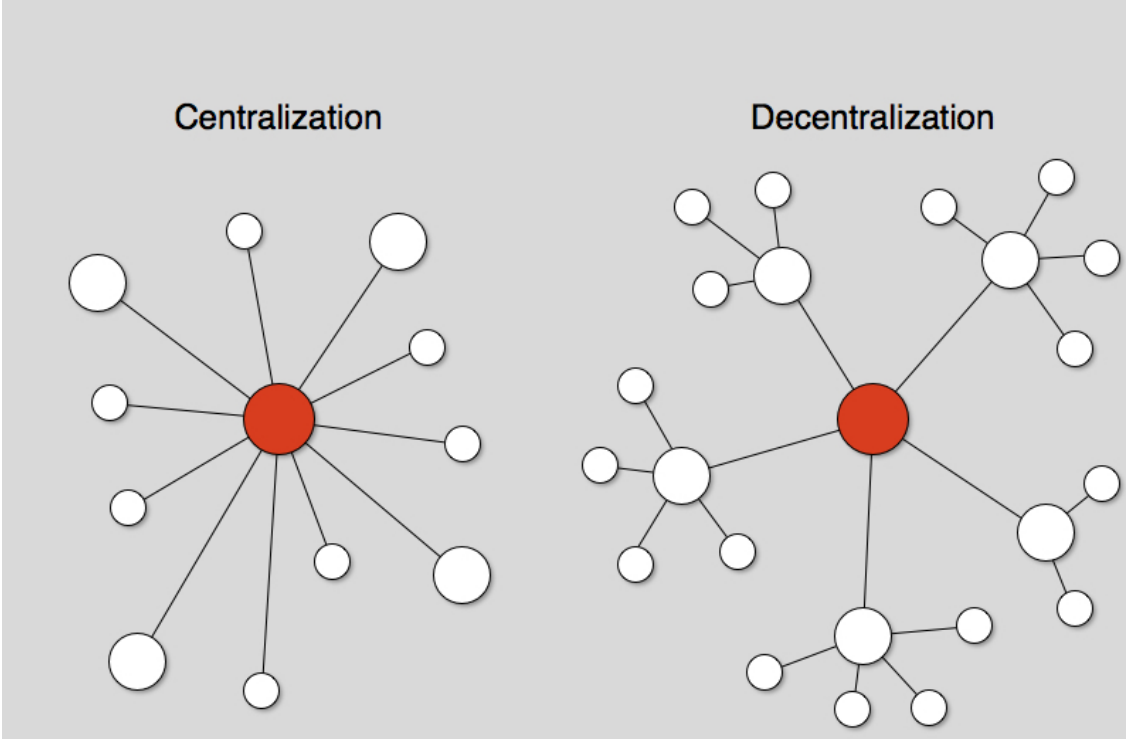
- Technical challenges:
  - Wikidata installation
  - Data exchange between Wikibase and Wikidata
  - Consolidated management

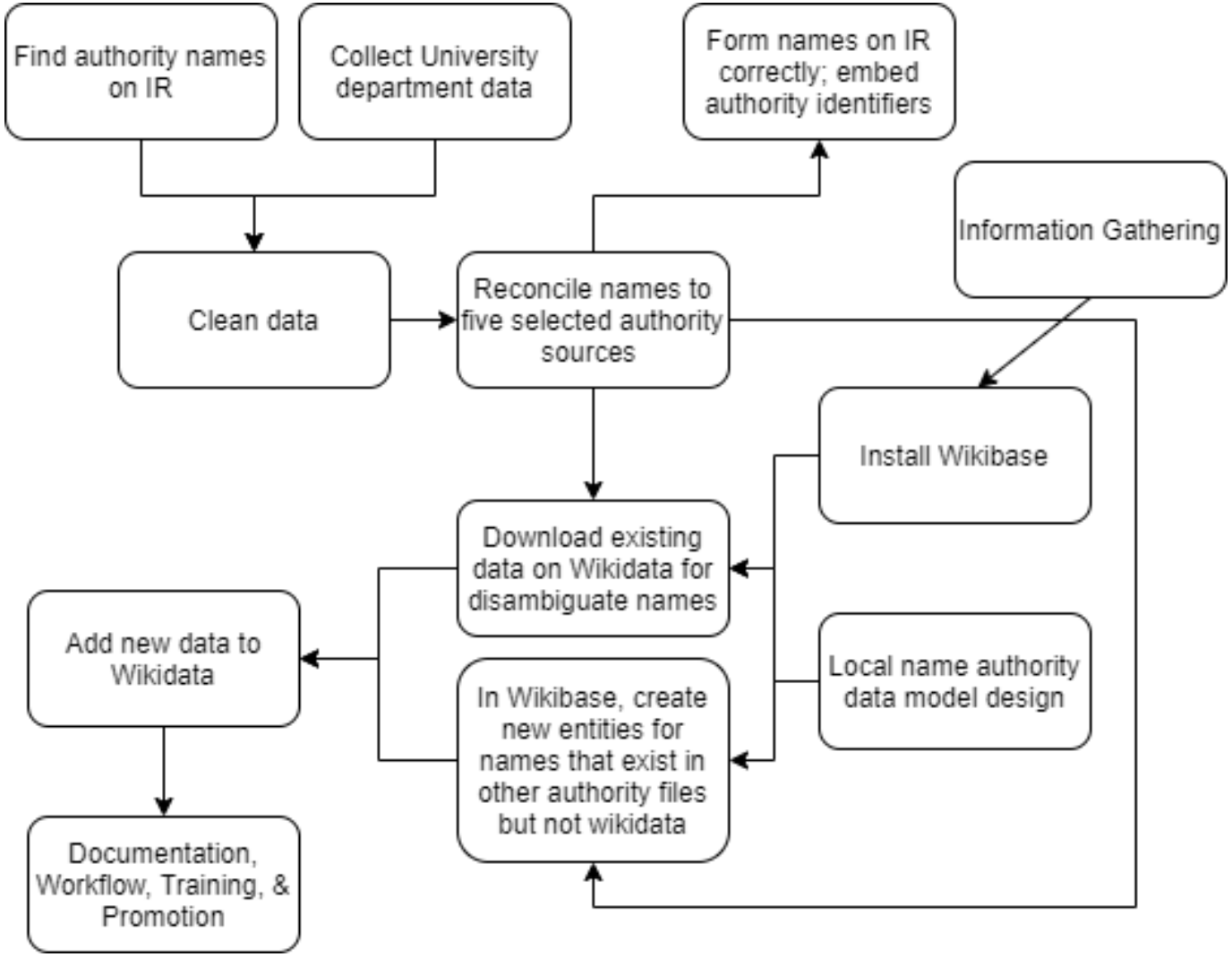


MediaWiki and Wikibase Environment, [OCLC Passage Project]

4. Discussion

- Tension between centralized authority and distributed authority




Find authority names on IR
Collect University department data
Form names on IR correctly; embed authority identifiers
Information Gathering
Clean data
Reconcile names to five selected authority sources
Install Wikibase
Download existing data on Wikidata for disambiguate names
Add new data to Wikidata
Local name authority data model design
In Wikibase, create new entities for names that exist in other authority files but not wikidata
Documentation, Workflow, Training, & Promotion

Conclusion:

Reference

Bakauf, Steve. "VanderBot part 2: The Wikibase data model and Wikidata identifiers". Steve Bakauf's Blog (blog). February 7, 2020. http://baskauf.blogspot.com/2020/02/vanderbot-part-2-wikibase-data-model.html.

Downey, Moira. "Assessing author identifiers: Preparing for a linked data approach to name authority control in an institutional repository context." *Journal of Library Metadata* 19, no. 1-2 (2019): 117-136.

Godby, Carol Jean, and Karen Smith-Yoshimura. "From records to things: managing the transition from legacy library metadata to linked data." *Bulletin of the Association for Information Science and Technology* 43, no. 2 (2017): 18-23.

Godby, Jean, Karen Smith-Yoshimura, Bruce Washburn, Kalan Davis, Karen Detling, Christine Fernsebner Eslao, Steven Folsom, Xiaoli Li, Marc McGee, Karen Miller, Honor Moody, Holly Tomren, and Craig Thomas (2019). *Creating Library Linked Data with Wikibase: Lessons Learned from Project Passage.* Dublin, OH: OCLC Research. https://doi.org/10.25333/faq3-ax08.

Krznarich, Liz. "What's New at ORCID?" June 26, 2017. https://slideplayer.com/slide/12963651/

Neatrour, Anna, and Jeremy Myntti. "Western Name Authority File: A pilot regional name authority project." *Journal of Library Metadata* 19, no. 1-2 (2019): 19-38.

Parker, Bria, and Adam Gray. "Rethinking the University of Maryland Authority File for the Linked Data Environment." *Journal of Library Metadata* 19, no. 1-2 (2019): 69-81.

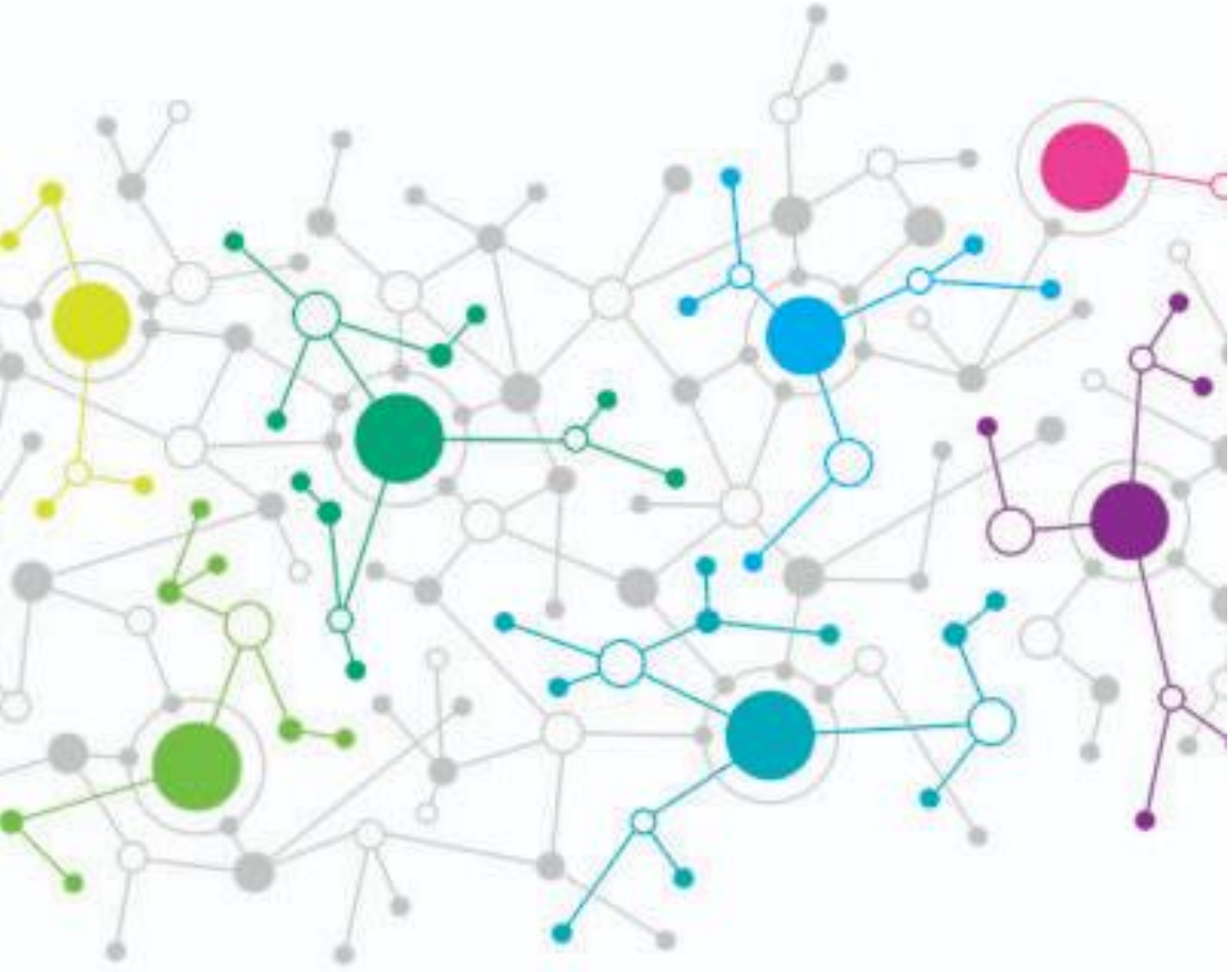
Reference

Smith-Yoshimura, Karen. *Transitioning to the next generation of metadata*. OCLC, 2020. https://library.oclc.org/digital/collection/p267701coll27/id/17015/ .

Smith-Yoshimura, Karen, and Eric Celeste. "Researcher Identifiers—What's in a Name (or URI)?." (2014).

Stanford Wikidata Working Group, "Data models and vocabularies." Accessed April 20, 2021. https://www.wikidata.org/wiki/Wikidata:WikiProject_Stanford_Libraries

Tian, Cindy, and Greta Heng. "Links All the Way Down: Emblematica as a LOD Case Study." Findings presented at 2020 LD4 Conference on Linked Data in Libraries. July 20, 2020. https://ld42020.sched.com/event/cjIP/linked-data-for-chronological-period-definitions-emblematica-as-linked-data-case-study-modeling-cultural-heritage-information-on-the-medieval-middle-east.



Thank you! Questions? Suggestions?

Greta Heng
Email: jieheng2@illinois.edu
GitHub: https://github.com/gretaheng/Wikibase_IR
4/30/2021