Preparing for a Name Disambiguation Application for Institutional Repositories at Texas A&M University: the Planning and Test Preparation Phases

Charity Stokes, Tatyana Chubaryan, James Creel, Jeannette Ho

Texas A&M University Libraries

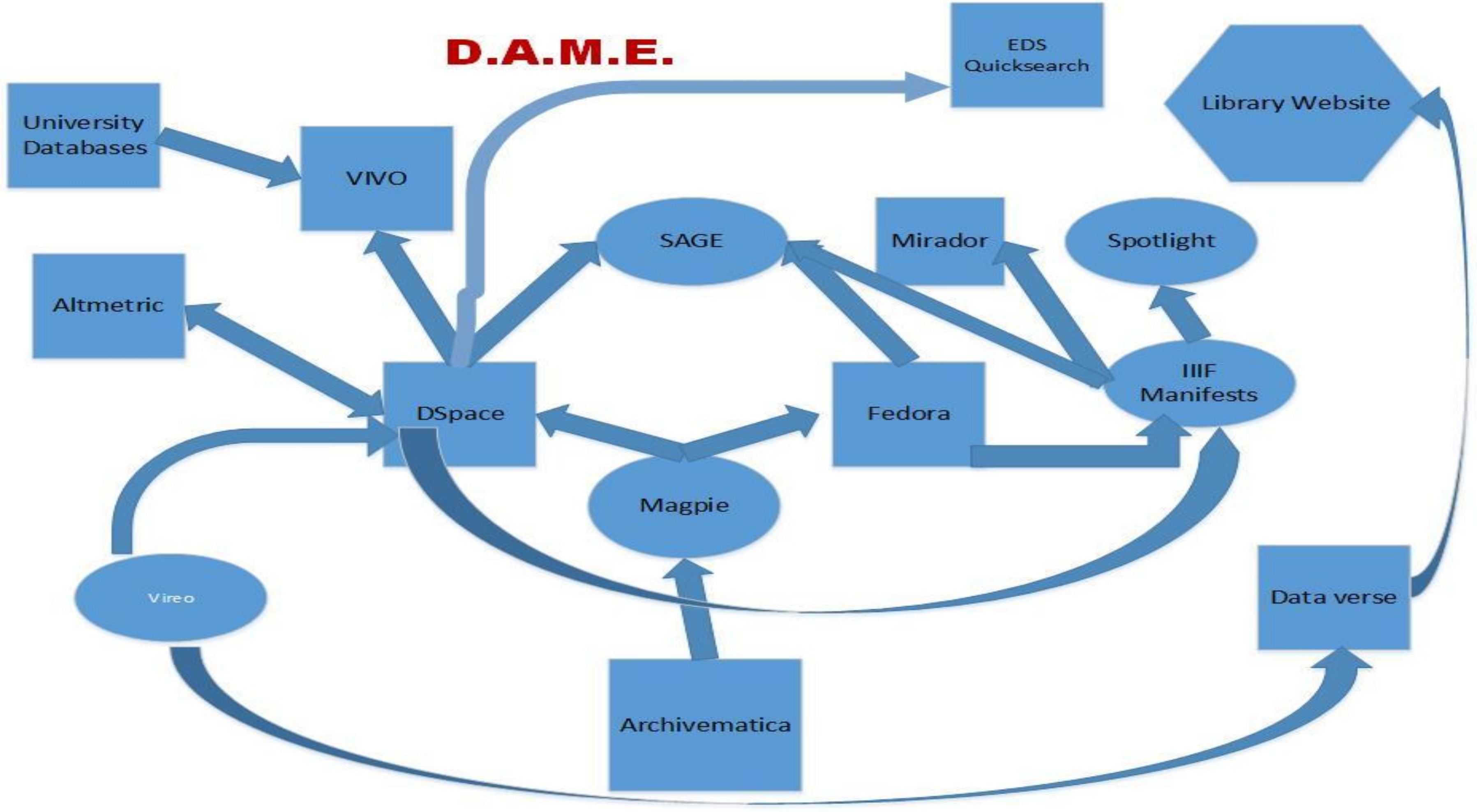
May 26, 2021
Texas Conference on Digital Libraries

LIBRARIES
TEXAS A&M UNIVERSITY

Lack of disambiguation: a problem that has grown and has become more urgent as our systems and collections has become more complex

- Problem for users who can't find all of a person's works under the same name
- Problem for Digital Initiatives staff who want to harvest metadata from VIVO instance.
  - Expressed desire for consistent forms of names
- Problem across collections that were entered over time in these platforms
- Problem for coordinating searches across various platforms
  - DSpace, Fedora, Avalon, catalog, VIVO
- Problem of multiple sources of metadata
  - Individuals adding items, Vireo, bulk ingest, MARC records, etc.

Our digital ecosystem (aka Digital Asset Management Ecosystem or “The DAME”)



The DAMEname working group (DAME=Digital Asset Management Ecoystem)

- **A working group was formed in 2018 to make recommendations for a "robust name authority system for Texas A&M affiliates and entities," including:**
  - An approach that would use existing standards (e.g., ISNI, ORCIDs), etc. alongside a newly minted URI-based identifier
  - A name application, if applicable
  - Basic technical needs for implementation
  - Estimated time/effort to implement the solution
  - Evaluate VIVO a a potential authority file system
  - Priorities among the need for authorities for A&M faculty members, staff, students, colleges, departments, etc.
- **While we did not rule out names of organizations, colleges and departments, etc. or subjects, our focus was on PERSONAL NAMES as the FIRST step**

Activities of the working group (Phase I: June to December 2018, Phase 2: January 2019 to the present)

- Conducted a literature review
- Reviewed existing standards for authority control and identifiers (focusing on persons)
- Examined existing “name apps”
  - NAMES (University of North Texas)
  - CEDAR (University of Houston)
  - Others
- Developed use cases (i.e., what would we like our app to do?)
- Recommended building our OWN app
- Explored and developed proposal for infrastructure and data sources for the app

Purposes of proposed app:

- **To serve as a tool for library personnel to manage identities in our DAME**
  - Every new name that gets input into IR should get minted with unique URI in the app (UUID)
  - Identify names for clean-up and reconciliation in the IR
- **To serve as a tool for metadata providers in OAKTrust and other repositories in our DAME to consistently select names that accurately identify and disambiguate authors**
- **To allow users of TAMU repositories to identify authors (and eventually organizational entities and subjects)**

Things we would like our “authority control” app to do:

- **Mint unique URI (using UUIDs) for each person in our repository**
  - That can be re-used by other components within our DAME (e.g., Fedora, Avalon, Spotlight, possibly FOLIO?, etc.).
  - Eventually, would like to link these URIs with ones associated with each digital “item” a person created or contributed to answer the question: “Show me all of this person’s works, despite form of name.”
- **Allow searching of both canonical name and name variants of each person or entity, in addition to retrieval via UUID**
  - App should cluster variants to answer the question: “Do I have all of this person’s works despite form of name”?
- **Disambiguate names for metadata providers and repository users so they can identify which name matches person being searched for**
  - To answer the question: “Is this the person that I am searching for?”
  - Use of contextual information

Things we would like our "authority control" app to do (continued):

- **Identify names that need to be disambiguated (i.e., reconciled) after they are input into the IR for retrospective clean-up (e.g., self-submitted and legacy collections)**
- **Enable metadata providers to select appropriate names at the point of entering metadata to ensure consistency**
  - Via "type aheads" or drop-down lists that would provide choices that display the "canonical" form of name (plus additional contextual data?)
  - To do this, names in the app would need to be accessible to:
    - Authors who self-submit their works and metadata into OAKTrust (via Manakin self-submission form or Vireo ETD management software)
    - People who do NOT utilize a user interface when supplying metadata for the IR (e.g., spreadsheets for batch ingests)

Things we would like our “authority control” app to do (continued):

- **Our ultimate goal: Enable users (e.g.., “the public”) to identify names of authors, etc. when searching our IR**
  - Via similar “type ahead” or drop-down list used for metadata providers
  - Need user interface that will allow them to access names in the app with enough disambiguating information for them to tell “This is the author I’m looking for” and retrieve all of their works.

What our app will NOT do:

- Disambiguate faculty members who are no longer affiliated with TAMU (left, retired, etc.)
- Disambiguate researchers whose work is included in our repositories but have not ever been directly affiliated with TAMU
  - Researchers that have co-authored with TAMU faculty but are not affiliated with TAMU
  - Individuals (e.g., artists, donors, etc.) featured in special collections that have no other connection with TAMU

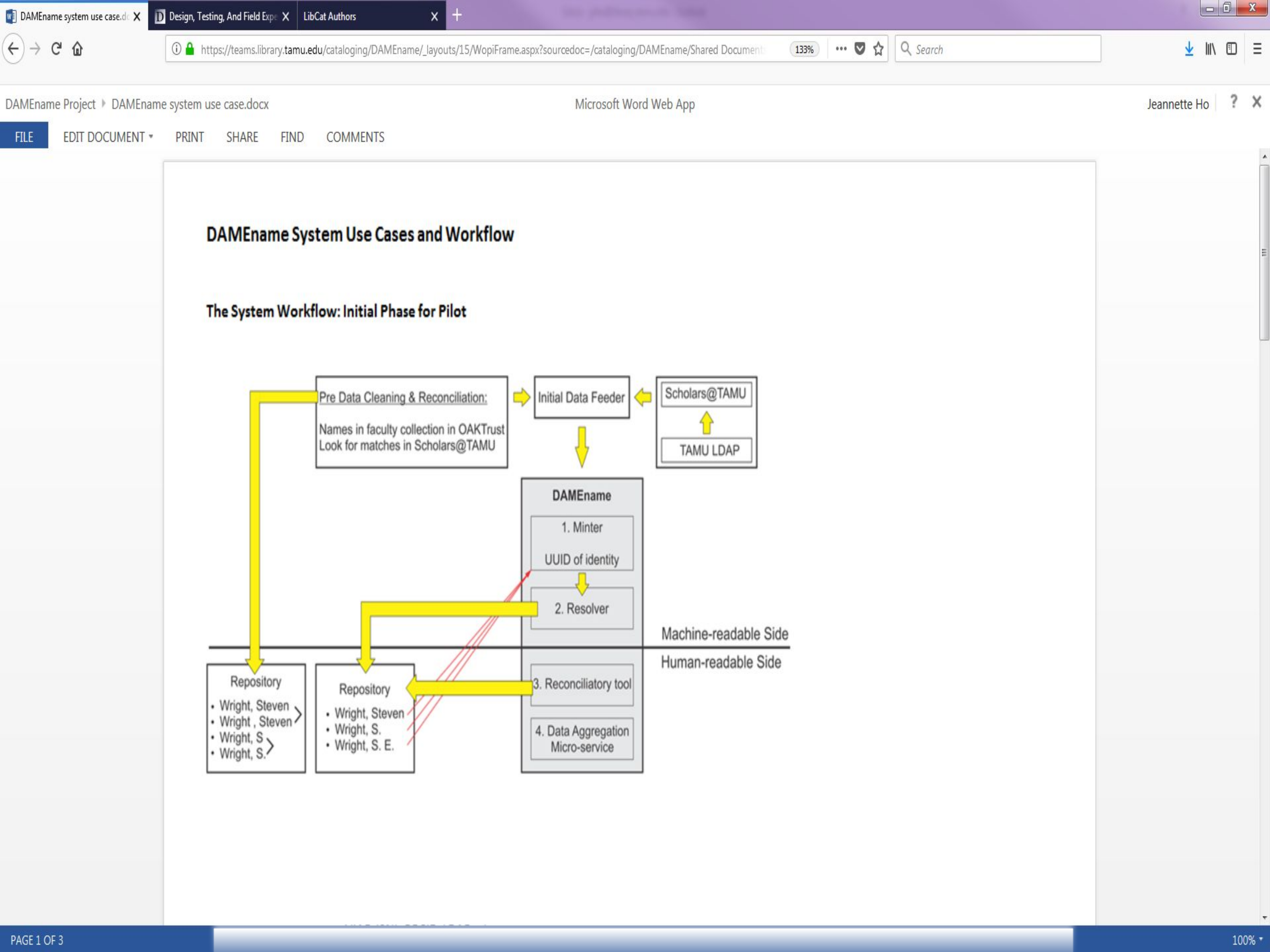
Why develop our own app?

- Although we liked their app, our situation at TAMU differs from UNT:
  - A lot of metadata and collections are SELF-SUBMITTED by faculty/students outside of Libraries
  - Lack of enough staffing to manually input information about names into an app and need to automate the population of the app
- Complexity of our digital ecosystem that consists of various platforms and databases
- Need for an app that can work with linked data in the future

What should be stored in the app? At a MINIMUM, it should include:

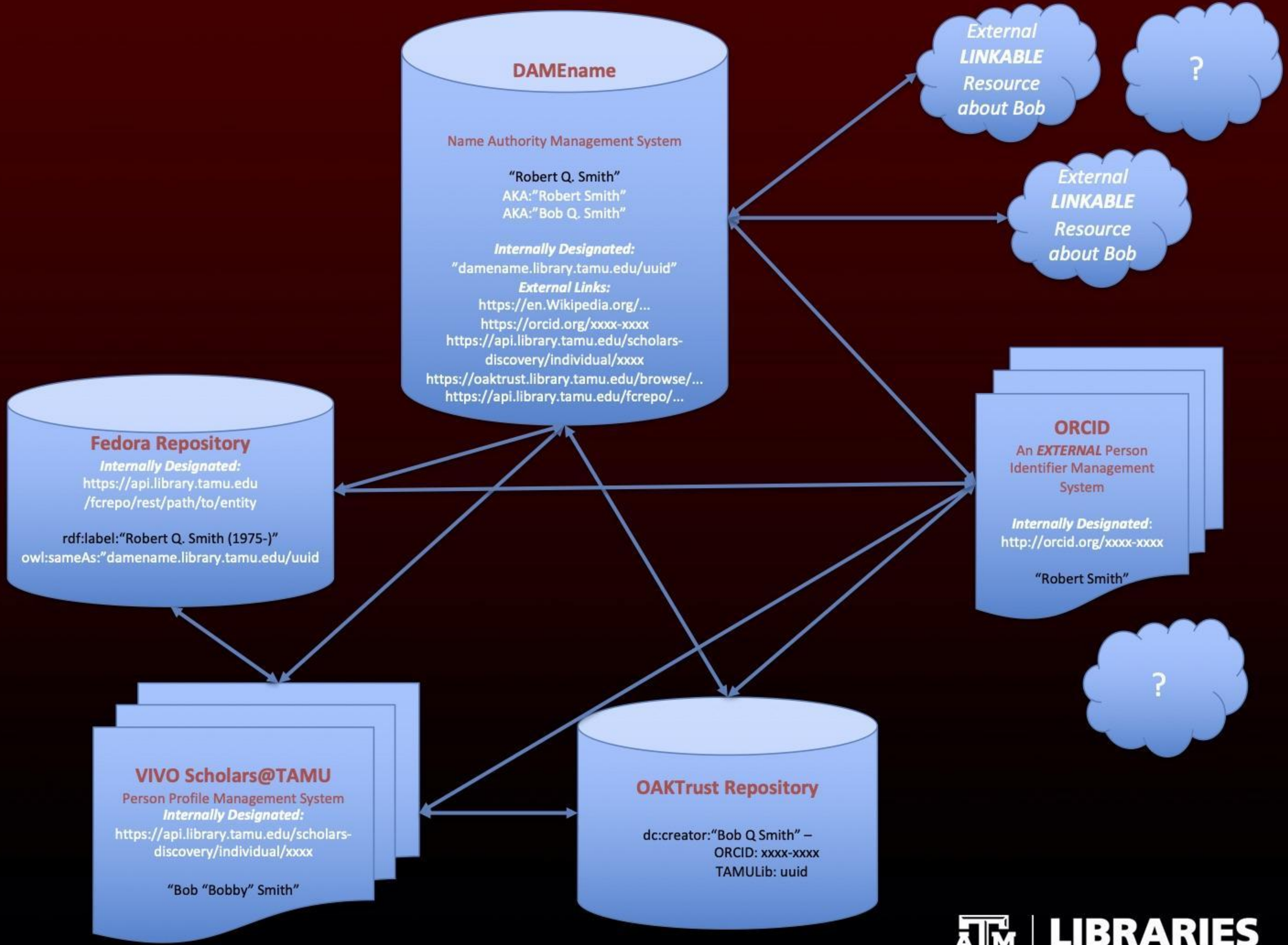
- **Unique identifier (UUID)**
- **“Canonical” name (“Murry, Robert D.) from LDAP, our TAMU directory**
- **Name variants (“Murray, Bob”, “Murray, Bobby”)**
- **Links to external sources**

  – We are actually managing IDENTITIES rather than NAMES!
    - The unique URI (the UUID) is the central component that ties all name variants and links to external identifiers together


DAMEname system use case.d
Design, Testing, And Field Expe
LibCat Authors
https://teams.library.tamu.edu/cataloging/DAMEname/_layouts/15/WopiFrame.aspx?sourcedoc=/cataloging/DAMEname/Shared Document
133%
Search
DAMEname Project
DAMEname system use case.docx
Microsoft Word Web App
Jeannette Ho
FILE
EDIT DOCUMENT
PRINT
SHARE
FIND
COMMENTS

DAMEname System Use Cases and Workflow

The System Workflow: Initial Phase for Pilot



DAMEname

Name Authority Management System

"Robert Q. Smith"
AKA:"Robert Smith"
AKA:"Bob Q. Smith"

Internally Designated:
"damename.library.tamu.edu/uuid"
External Links:
https://en.Wikipedia.org/...
https://orcid.org/xxxx-xxxx
https://api.library.tamu.edu/scholars-discovery/individual/xxxx
https://oaktrust.library.tamu.edu/browse/...
https://api.library.tamu.edu/fcrepo/...

External ***LINKABLE*** *Resource about Bob*

?

External ***LINKABLE*** *Resource about Bob*

Fedora Repository
Internally Designated:
https://api.library.tamu.edu/fcrepo/rest/path/to/entity

rdf:label:"Robert Q. Smith (1975-)"
owl:sameAs:"damename.library.tamu.edu/uuid

ORCID
An ***EXTERNAL*** Person Identifier Management System

Internally Designated:
http://orcid.org/xxxx-xxxx

"Robert Smith"

?

VIVO Scholars@TAMU
Person Profile Management System
Internally Designated:
https://api.library.tamu.edu/scholars-discovery/individual/xxxx

"Bob "Bobby" Smith"

OAKTrust Repository

dc:creator:"Bob Q Smith" –
ORCID: xxxx-xxxx
TAMULib: uuid

“Data Aggregation Micro-service”: Where will disambiguating data come from?

- **It will link to multiple external sources of linked data, including Scholars (our local VIVO instance)**
- **Some possibilities that ranked the highest by the DAMEName working group:**
  - Library of Congress National Authority File (LCNAF)
  - ISNI (International Standard Identifier)
  - ORCID
  - VIAF (Virtual International Authority File)
  - Scopus
- **Other ones we looked at:**
  - Researcher ID — MS Academic
  - Wikidata — Dimensions
  - Google Scholar

Preparing for testing the future app prototype:

- In fall of 2019, a team consisting of members from the cataloging unit and one librarian from the Office of Scholarly Communications
  - Cleaned up names of advisors of theses and dissertations from OAKTrust repository
  - Identified variants of names that would be used to test the app prototype

Preparing for testing the future app prototype:

- Ran spreadsheet of names through OpenRefine:
  - Inserted periods and spaces between initials (JM to “ J. M.”) and put names in inverted order where needed using GREL expressions in OpenRefine (“J.M. White” to “White, J. M.”)
  - Clustered names if predominant form and one variant
  - For rest of names, retained variants to test how well the name app prototype will “catch” them and flag them for manual review.
  - Flagged “interesting” cases that we wanted to retain to test the app
  - Cleaned up obvious misspellings, stray punctuation marks and instances of “Jr.” (Charles, Jr. Mchael became “Charles, Michael, Jr.”) and numbering

Preparing for testing the future app prototype:

- **Examples of name variants that we did NOT correct or cluster:**

  — Kim, Sung Hyun vs. Kim, Hyun Sung
  — Smith, James A. vs. Smith James Allen
  — Walker, Duncan M. (Hank) vs. Walter, Duncan M.
  — James, Tony vs. James, Tony R.
  — Chen, Li vs. Li, Chen
  — Choice of diacritics differed for same name

Issues to be resolved in the future:

- **What contextual data can be used for disambiguation?**
  - Some possibilities: college or department, title, subject areas, publication titles, etc.)
  - Will we control for colleges and departments? (establish as separate entities in the app?
- **Where will contextual data used to disambiguate names come from?**
  - Some possibilities: VIVO Scholars database at TAMU, LDAP TAMU directory, external linked data sources (ORCID, ISNI, etc.).
  - Wikidata?

Issues to be resolved (continued):

- What will user interface look like for librarians/staff performing reconciliation? For metadata providers who will utilize the app?

- How will app interact with the public interface of repositories within the TAMU digital ecosystem?
  - Can users of the IR (faculty, students, etc.) benefit from contextual data it can "link out" to?
  - Maybe a link from name in IR to a knowledge card that utilizes links inside the app to pull in data from external sources?
  - Future integration with FOLIO?

Issues to be resolved (continued):

- **How to deal with name changes?**
  - Treat as name variants or mint new identifier within the app for them?
  - Associate each name with dates when each name was used?
  - Persons vs. organizations

- **Will data be cached?**
  - "Linking out" to external systems may slow performance of the app
  - Keeping local "copies" of such data can help speed things up

Next steps:

- Develop a prototype (projected date: not before TAMU's FOLIO implementation in September 2021)
- Test the prototype on the electronic thesis and dissertation collection
  - Run the names through the app that we cleaned up
- Analyze results of the test
  - Did it catch all of the variant names that we flagged for human review and reconciliation by staff? What is the success rate?
  - What are problems with the app that need to be tweaked?
  - What are the frequencies of different types of problems?
  - How much work does cleaning them up involve? (Implications for staffing & workflows)

In the future, we hope to :

- Plan how to deal with legacy data
- Eventually connect personal identities with their works in the IR
- Eventually include other types of entities in the app besides people (e.g., organizations, subjects)

Questions?

Our contact info:

jaho@library.tamu.edu

jcreel@library.tamu.edu

charity.martin@library.tamu.edu

tchubar@library.tamu.edu