CIVIC DIGITAL FELLOWSHIP

Improving Access and Interoperability between the ADREC Metadata Shares and DMS

Sophie Kaplan

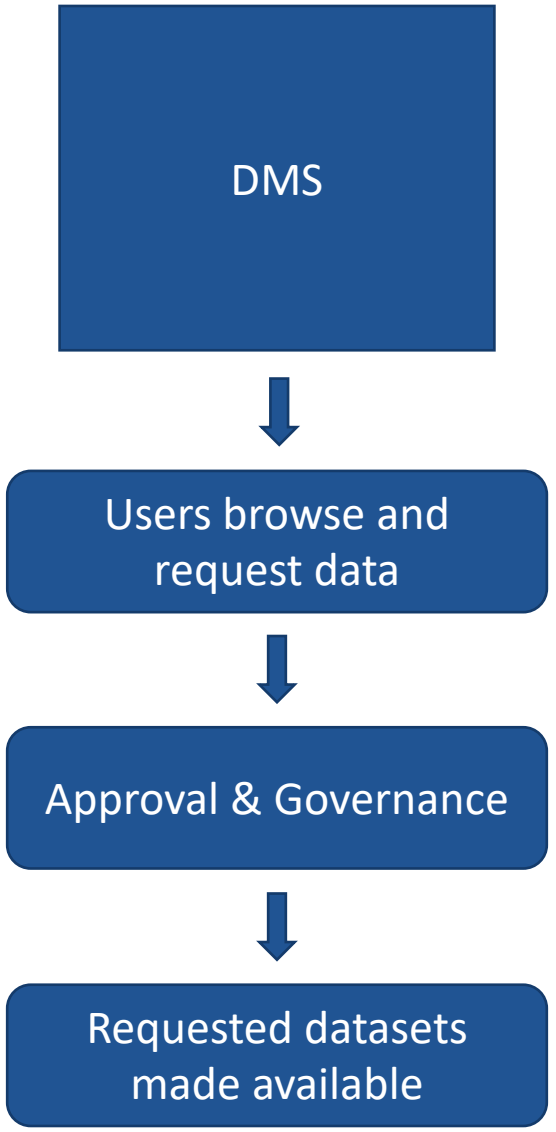
Supervised by Harold Saintelien, Crissman Nichols, and Michael Castro

Policy and Data Stewardship Branch

Issue: ADREC resources inaccessible

ADREC

- 6000 directories and subdirectories
- Rich metadata resource that is currently inaccessible to users of the DMS



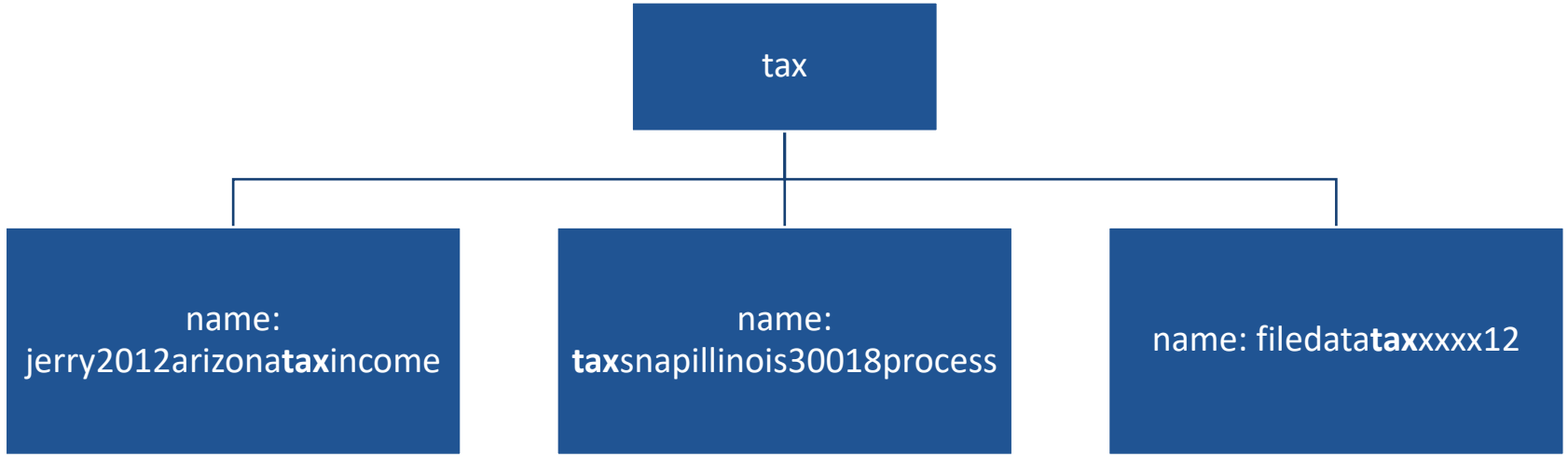
Project Objectives

Mature and optimize delivery of Census data by implementing a:

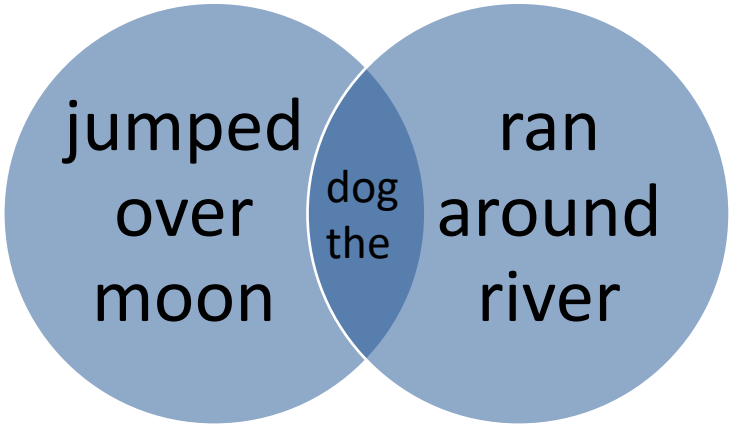
1. **Method for traversing the ADREC metadata share**
2. **Crosswalk between the ADREC and DMS nomenclature**
3. **API to access ADREC metadata associated with a given dataset/series in the DMS**

Finding Language Associations

- **Fuzzy matching algorithms had quadratic runtime and low hit rate**
- **Needed to make the problem smaller**
  - **Group files with blocking predicates**
  - **Use string similarity algorithms to find further associations**

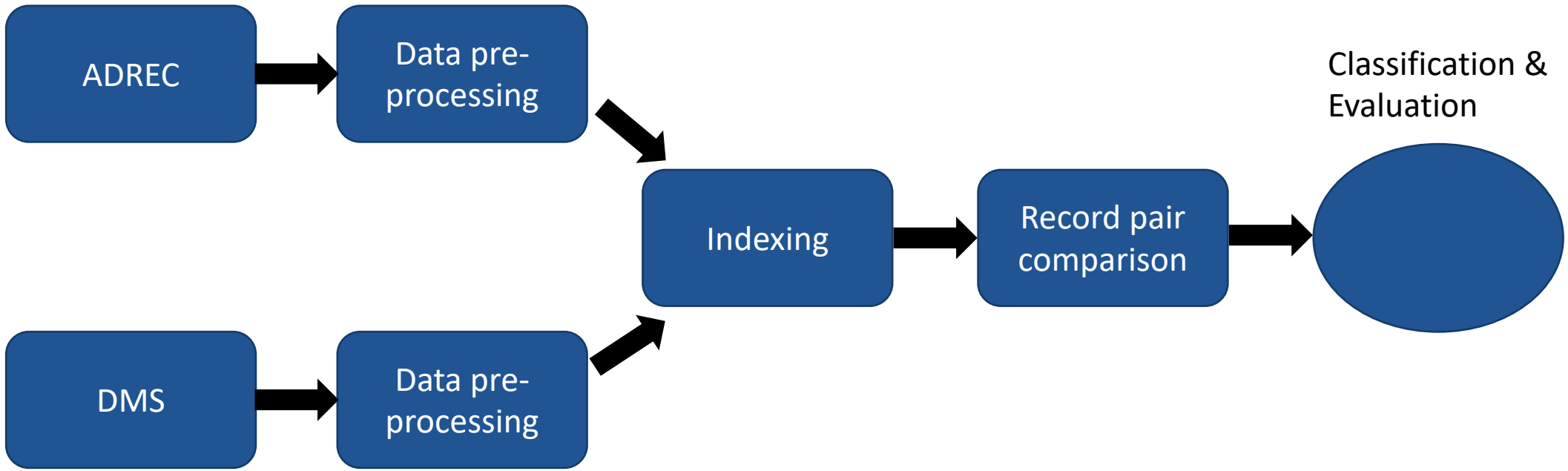


Jaccard String Similarity



$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Record Linkage Process



Deliverables

- ✓ **Web Search Engine**

  Tools and Skills: Java, Spring Boot, Thymeleaf, Maven, Subversion, Junit

- ✓ **Database of Associations**

  Tools and Skills: Python, Record Linkage, Subversion

- ✓ **In Progress: API**

  Tools and Skills: Java, REST API Principles, Spring Boot, SQL, Maven, Subversion, Junit

Future Steps: Expand classification methods, develop a method for keeping the DMS current on changes to the ADREC metadata share