CIVIC DIGITAL FELLOWSHIP

Improving Access and Interoperability between the ADREC Metadata Shares and DMS using Natural Language Processing.

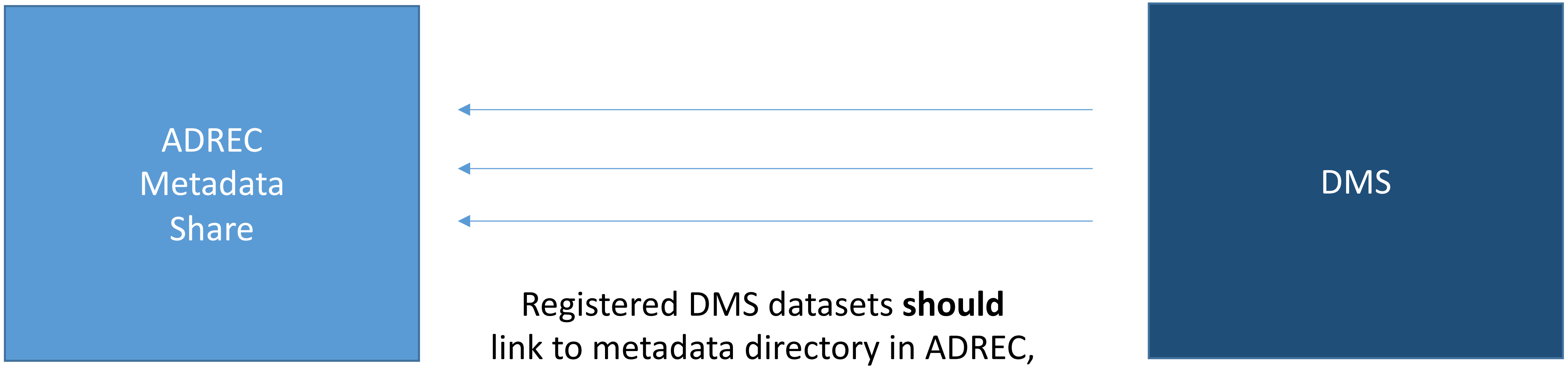
Nikasha Patel

Supervised by Harold Saintelien, Crissman Nichols, and Michael Castro

Policy and Data Stewardship Branch

Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau.

Issue: ADREC resources inaccessible from DMS



ADREC Metadata Share

DMS

Registered DMS datasets **should** link to metadata directory in ADREC, **but** there are missing links

Additionally, datasets registered in DMS have different names than in ADREC

Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau.

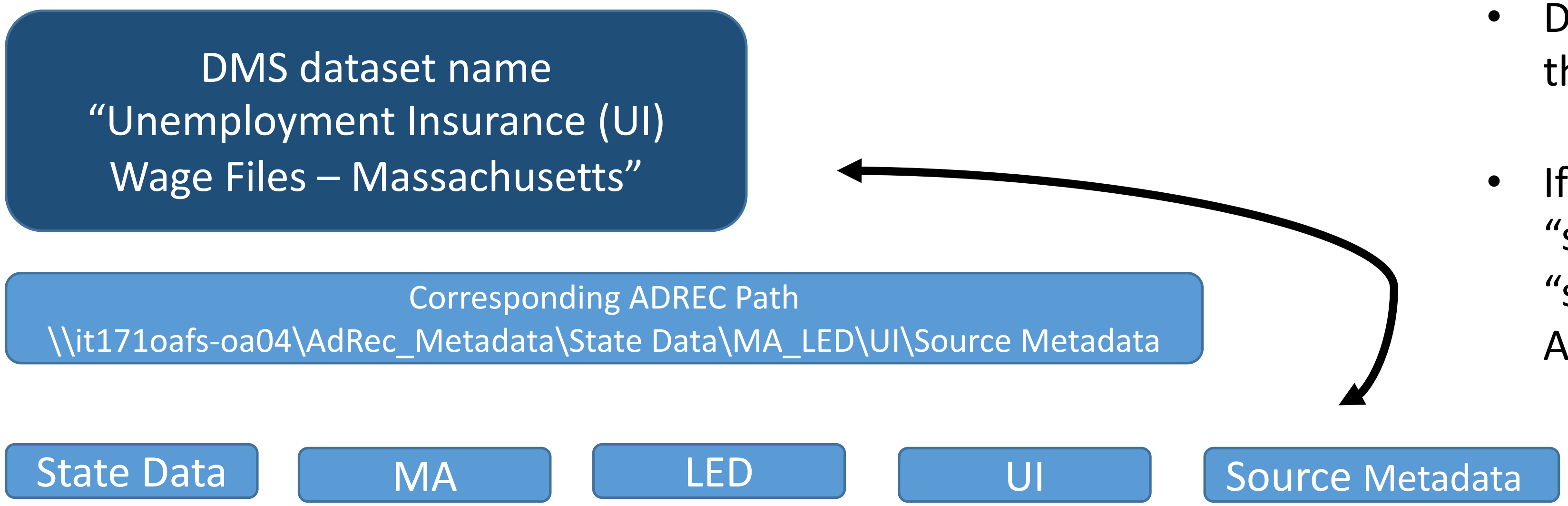
Project Objectives

Continue upon the previous CDF's work to

1) Find the missing links between ADREC and DMS efficiently and accurately
2) API to access ADREC metadata associated with a given dataset/series in the DMS

Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau.

Solution and Breakthrough



DMS dataset name
“Unemployment Insurance (UI)
Wage Files – Massachusetts”

Corresponding ADREC Path
\\it171oafs-oa04\AdRec_Metadata\State Data\MA_LED\UI\Source Metadata

- DMS dataset name corresponds to the file path to an ADREC directory
- If the DMS dataset name is like a “sentence,” then it’s related to the “sentence” that spells out the ADREC directory

Employ natural language processing “bag of words” technique to compute a similarity score between pairs of dataset names and relevant ADREC paths

Bag of Words: [“cats”, “dogs”, “I”, “like”]

Sentence: “I like dogs”

Corresponding Vector: [0, 1, 1, 1]

Solution and Breakthrough

Step 1: Pre-process data by larger categories

State Data

Federal Data

Commercial Data

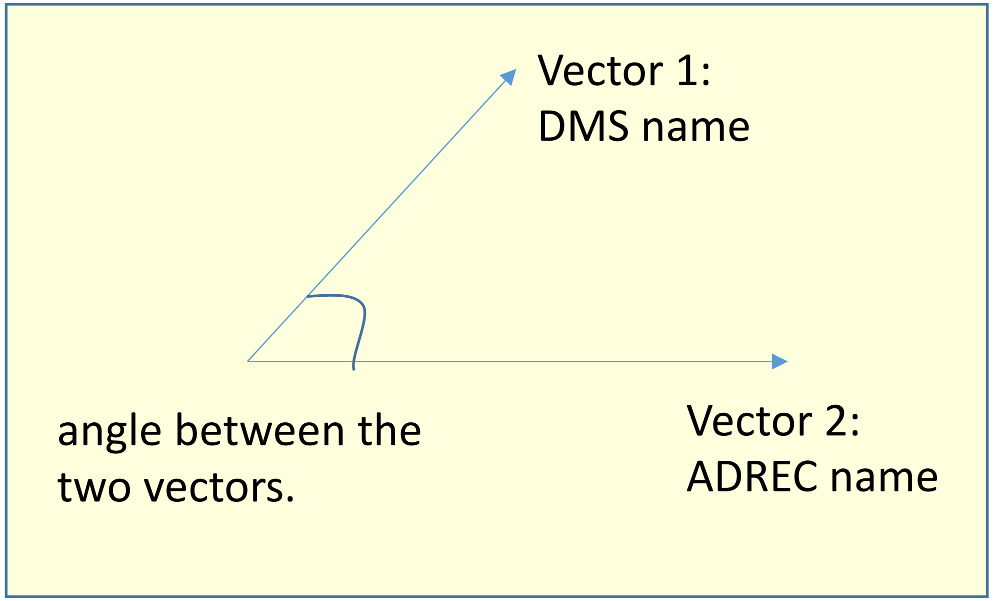
Step 2: Create dictionaries that map common acronyms to specific terms found in DMS

Step 3: Replace terms in DMS dataset names with corresponding acronyms found in ADREC

Step 4: Vectorize DMS dataset names and ADREC file paths using "bag of words"

Example vector: [0 1 0 0 2 0 3 1]

Step 5: Calculate the cosine similarity between pairs of vectors.



Conclusions

✓ Accurate database of matched records

Tools and Skills: Python, basic NLP toolkits (Sklearn), Subversion

✓ In Progress: API

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

For future consideration: Standardizing DMS naming schemes

Need for Census to adopt consistent naming schemes for a secure, private search engine that doesn't compromise Title 13/26/FTI data

Next step: Incorporating linkage results into an API

Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau.