PRESENTING LEADING fellows

Amanda Whitmire  Stanford University
Chris Wiley  UIUC
Chenyue Jiao  UIUC
Ateanna Urii  U of Texas Rio Grande Valley
Chris Rauch  Drexel University

Principal Investigators

Jane Greenberg, Drexel University
Kenning Arlitsch, Montana State University
Erik Mitchell, University of California, San Diego
Rachel Frick, OCLC
Jake Williams, Drexel University
**DATA**

- 4 Hubs, 18 nodes
- 31 mentors
- 8 Drexel LIS/data sci. faculty
- DEI Task Force + Advisory Board
- 24 Fellows

**LEADING sites**

- Academy of Natural Sciences (ANS)
- AI-Collaboratory, University of Maryland iSchool (AIC-MD)
- California Digital Library, University of California Office of the President (CDL)
- Loretta C. Duckworth Scholars Studio, Temple University Libraries (LCDSS)
- Digital Scholarship, Tish Library, Tufts University (DTTu) (2021 only)
- Kislak Center for Special Collections, Univ. of Pennsylvania Libraries (UPenn)
- LYRASIS (LYR)
- Montana State University Library (MSU)
- Movement Alliance Project (MAP)
- OCLC - R&D (OCLC-R)
- Smithsonian Libraries (SL)
- UC San Diego Library (UCSD)
- University of New Mexico (UNM)
- University of North Texas (UNT)
- University of Rochester (ROC)

**Additional LEADING 2022 sites**

- Dryad
- National Science Foundation, Public Access Repository
- Research Organization Registry (ROR)
- Ronin Institute/U.S. Research Data Alliance
2021 Fellows

https://cci.drexel.edu/mrc/leading/fellows/
LEADING Fellows/Sites

Amanda Whitmire, Head Librarian & Bibliographer, Harold A. Miller, Library, Stanford Libraries
LEADING Site: Academy of Natural Sciences
thalassa@stanford.edu

Chris Wiley, PhD Student, University of Illinois—Urbana Champaign
LEADING Site: Montana State University Library
cawiley@illinois.edu

Chenyue Jiao, PhD Student, University of Illinois—Urbana Champaign
LEADING Site: University of New Mexico Libraries with Montana State University Library
cjiao4@illinois.edu

Ateanna Uriri, Metadata Librarian, University of Texas at Rio Grande Valley Libraries
LEADING Site: UC San Diego
ateanna.uriri@utrgv.edu

Chris Rauch, PhD Student, College of Computing & Informatics, Drexel University
LEADING Site: California Digital Library
cr625@drexel.edu
PROJECT GOALS AND OBJECTIVES

• Automate identification of taxonomic names in ANSP corpus
• Use NLP to identify relationships b/w species names + locations + dates to capture species occurrences in the corpus

PROGRESS / ACCOMPLISHMENTS

See: https://amandawhitmire.github.io/blog/

R code/Rmd
• Used Global Names Finder, a tool to ID taxonomic names in text to create lists of taxonomic names from corpus, formatted for use in Python NLP tool, spaCy

Python Notebooks
• Dialed in how named entity recognition works w/ custom entity sets in spaCy

Streamlit App to visualize NLP

APPRAOCH

Workflow documentation

WHY LEADING?

• Dedicated time to learning skills that make me more effective in my job

TAKE AWAY

Time extremely well spent for a mid-career librarian: my organization, researchers and I all benefit from new skills.
PROJECT GOALS AND OBJECTIVES

1. Assess MSU Faculty perception on the usefulness and effectiveness of the library

2. Examine the citation data and determine the degree to which the collection supports faculty

PROGRESS / ACCOMPLISHMENTS

1. MSU Faculty rates their skills at finding information as effective and that the library has been effective in meeting their needs.

2. Faculty are concerned about the loss of journal loss

CITATION ANALYSIS RESULTS

- Sixty-six faculty and their articles were analyzed in this study using SCOPUS.
- This resulted in 1987 cited references.
- 80% of Journals
- 9.7% Dissertations, book chapters
- Less than 1% are web accessible resources/reports
PROJECT GOALS AND OBJECTIVES

The Repository Analytics and Metrics Portal (RAMP) is “a web service that aggregates use and performance use data of institutional repositories” 1.

**Question**: Does metadata affect the click rates of IR contents and how they affect?

**Objective**: To understand the relationship between some metadata fields and click rates of IR contents.

**Implication**: IR search engine optimization

**APPROACH**

- Data: Thesis and dissertation from 35 IRs during January 2019 and May 2019; Title, abstract, keyword, and subject were extracted
- Text mining & Natural language processing
- Statistical analysis

**FINDINGS**

**Metadata can affect the click rates of items in IR.**

- The shorter and clear titles are likely be clicked;
- Titles containing a colon or hyphen, more noun words, and some positive or negative words are associated with higher clicks;
- Context information in titles can affect the clicks;
- The shorter abstract, more keywords and subject can attract more attention.

**LEADING WHY? and/or TAKE AWAY/s**

- Received much help and support
- Learned data science techniques and skills
- Feel confidence in solving problems involving data analytics
- Realized that data sciences can inform right decisions and provide better services

---
1. https://rampanalytics.org
PROJECT GOALS AND OBJECTIVES
• Make sense of what was in the collection
• Understand how it is used in its current state and how it could be used/transformed by applying different tools
• Make organizing/processing/transforming items easier for UCSD Library Staff

PROGRESS/ACCOMPLISHMENTS
• Created a detailed spreadsheet of Oral History audio files available on the collection site
• Created code to scrape metadata from collection image gallery
• Created PDF Mining introductory demo guide in Jupyter Notebook
• Created Streamlit app site that converts MP3 to WAV files, transcribes audio to text, and performs Named Entity Recognition with related Wikidata QIDs on text

APPROACH
• PDF mining, web scraping, Named Entity Recognition, and transcription automation using a variety of Python tools/packages, including spaCy, Streamlit, IBM Watson’s Speech to Text

LEADING WHY? and/or TAKE AWAY/s
• General professional development
• An opportunity to work with other people/institutions
• Related to above bullet point, see what collections and processes other institutions were working on/using
PROJECT GOALS AND OBJECTIVES

- Promote YAMZ as a community driven metadata dictionary.
- User testing, software review and community outreach to enhance YAMZ’s ability to facilitate and incentivize high-quality feedback and refinement of vocabularies and ontologies
- Implement a sustainable development cycle.

APPROACH

- Conduct codebase analysis and review related publications.
- Standardize codebase
- Publish plan for enhancements and receive feedback.
- Develop and implement user engagement plan.

PROGRESS /ACCOMPLISHMENTS

- Codebase modernization
- Presentation to 14th International Conference on Metadata and Semantics
- Research and publication in Spring Communications in Computer and Information Science (CCIS no. 1537)

LEADING WHY? and/or TAKE AWAY/s

Maintaining continuity over time when project participants change can be difficult. LEADING provided a way to ensure the long-term continuity of a promising project across funding cycles and invites insights and contributions from the community it seeks to serve.
Alternate Definitions for Ablation

Term: Ablation
Definition: (1) Combined processes (such as sublimation, melting, evaporation) which remove snow or ice from the surface of a glacier or from a snow-field. Also used to express the quantity lost by these processes. (2) Reduction of the water equivalent of a snow cover by melting, evaporation, wind and avalanches.

Term: Ablation
Definition: (1) combined processes (such as sublimation, fusion or melting, evaporation) which remove snow or ice from the surface of a glacier or from a snow-field; also used to express the quantity lost by these processes (2) reduction of the water equivalent of a snow cover by melting, evaporation, wind and avalanches.

Term: Ablation
Definition: (1) All processes that reduce the mass of the glacier. (2) The mass lost by the operation of any of the processes of sense 1, expressed as a negative number. The main processes of ablation are melting and calving (or, when the glacier nourishes an ice shelf, ice discharge across the grounding line). On some glaciers sublimation, loss of windborne snow and avalanching are significant processes of ablation. ‘Ablation’, unqualified, is sometimes used as if it were a synonym of surface ablation, although internal ablation, basal ablation, and frontal ablation, especially calving, can all be significant in some contexts.
LEADING Data Challenge

The Event:
Three teams competing for a cash prize over a period of 3 days. Team members were assigned roles Team Lead, Coder, Googler, Visualizer.

The Challenge:
Create a Challenge Statement
Analyze the data
Create a visualization
Present Findings

The Data:
5 Million WorldCat records, coded as children’s literature published between 1800-2021 in English-language cataloging included: Dewey Decimal Classification (DDC); Library of Congress Classification (LCC); Faceted Access to Subject Terminology (FAST); Virtual International Authority File (VIAF) IDs; Work IDs; Library Holdings.

The Winning Challenge Statement:
Explore gaps between languages represented in library collections and languages represented in local communities.
### Impacts

- **Scholarly and scientific outputs:** papers, posters, presentations, code, collaboration across projects (+25 outputs)
- **Doctoral dissertations:** 2 coming out of LEADS; **jobs:** 2 from LEADING
- **Bridging:** LIS Ph.D. students + early-to-mid career professionals
- **MSU:** RAMP project
- **UCSD:** Farmworker Movement Project; LEADING Community, C. Goldman
- **OCLC data challenge:** LEADS and LEADING Fellows
- **Drexel/Data science education resources** on LEADING’s Github site

### Acknowledgements

- IMLS LB21 Program
- **31 LEADING Mentors**
- LEADING Diversity, Inclusivity, & Equity Task Force (DIET)
- LEADING Advisory Board
- Fellows LEADING and LEADS
- Diversity, Equity & Inclusion Council (CCI/Drexel)
- Drexel University, UCSD, MSU + OCLC admin. support
- **Sam Grabus, LEADING Project Manager**
Questions/Discussion