Caltech Archives

The Caltech People Project

Tommy Keswick & Mariella Soprano
ArchivesSpace Online Forum 2020

Caltech Library

Introduction

Tommy Keswick

Digital Technologies Development Librarian

Mariella Soprano

Senior Archivist for Collection Management

Agenda

- Conceptualizing the Caltech People Project
- Initial Data Processing
- Formulating a Strategy
- Implementation
- Future Workflows and Questions

Conceptualizing the Project

- The Problem/Goal
- Why ArchivesSpace as a Solution?
- Integrating Data

Conceptualizing the Project

- The Problem/Goal
    - Library would like to provide services across all the repositories
        - Linking individuals associated with Caltech to their bodies of work
        - Linking to external information resources about Caltech people
    - Many repositories across the Library (and Archives)
        - No authority files exist for names of people or groups
        - IDs for people are different across repositories
        - Name formats are different across publications
        - No way to separate Caltech people from their co-authors on publications (in metadata)

Conceptualizing the Project

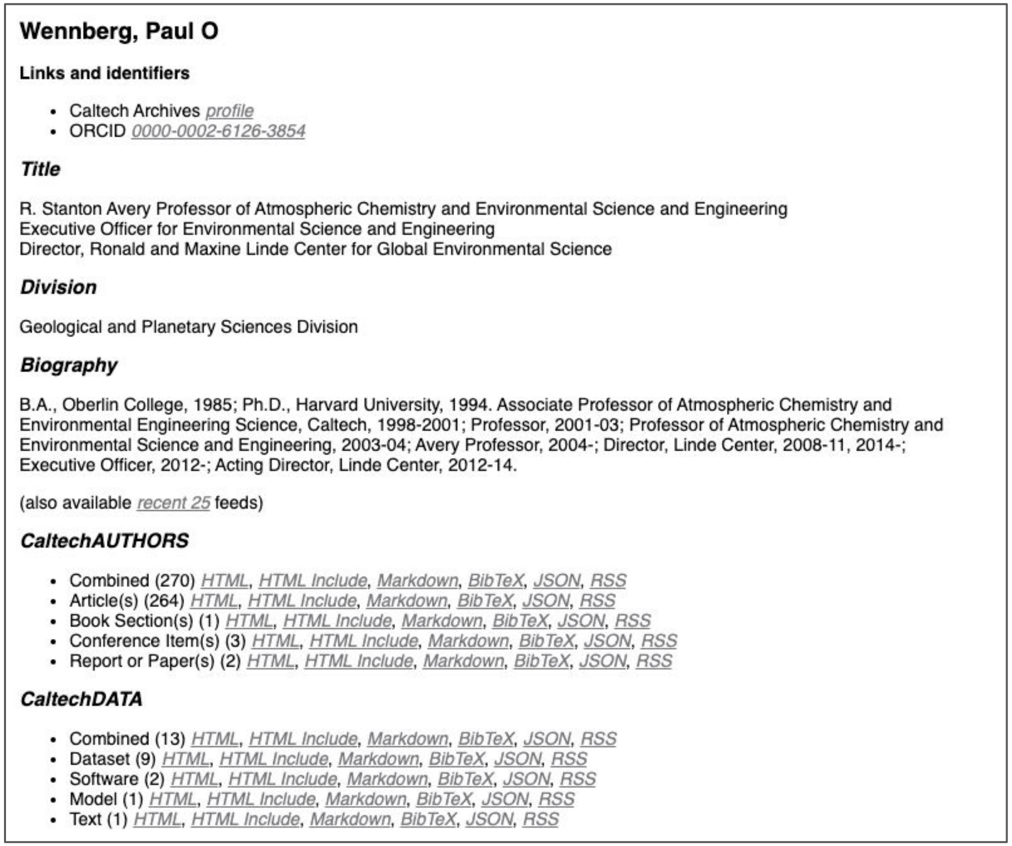
- Why ArchivesSpace as a Solution?
  - Institutional History
    - Today’s research and publications are tomorrow’s archival collections
  - Agent Records module provides functionality we need
    - No other system that we use can store as much authoritative data
  - Community adoption gives us confidence
    - Even if ArchivesSpace evolves into other projects in the future, the concept of Agents will likely persist

Conceptualizing the Project

- Integrating ArchivesSpace data with other sources
  - Services built on top of the authority file
    - Built with the ability to utilize the links between data sources that ArchivesSpace holds
  - Aggregate information from internal and external sources
    - ArchivesSpace data can be harvested and used to link to outside sources
  - Index it all externally for a new fast discovery interface
    - Searching for a person will bring up everything we know about them from the sources we care about

Conceptualizing the Project

- Integrating ArchivesSpace data with other sources



Wennberg, Paul O

Links and identifiers

- Caltech Archives profile
- ORCID 0000-0002-6126-3854

Title

R. Stanton Avery Professor of Atmospheric Chemistry and Environmental Science and Engineering
Executive Officer for Environmental Science and Engineering
Director, Ronald and Maxine Linde Center for Global Environmental Science

Division

Geological and Planetary Sciences Division

Biography

B.A., Oberlin College, 1985; Ph.D., Harvard University, 1994. Associate Professor of Atmospheric Chemistry and Environmental Engineering Science, Caltech, 1998-2001; Professor, 2001-03; Professor of Atmospheric Chemistry and Environmental Science and Engineering, 2003-04; Avery Professor, 2004-; Director, Linde Center, 2008-11, 2014-; Executive Officer, 2012-; Acting Director, Linde Center, 2012-14.

(also available *recent 25* feeds)

CaltechAUTHORS

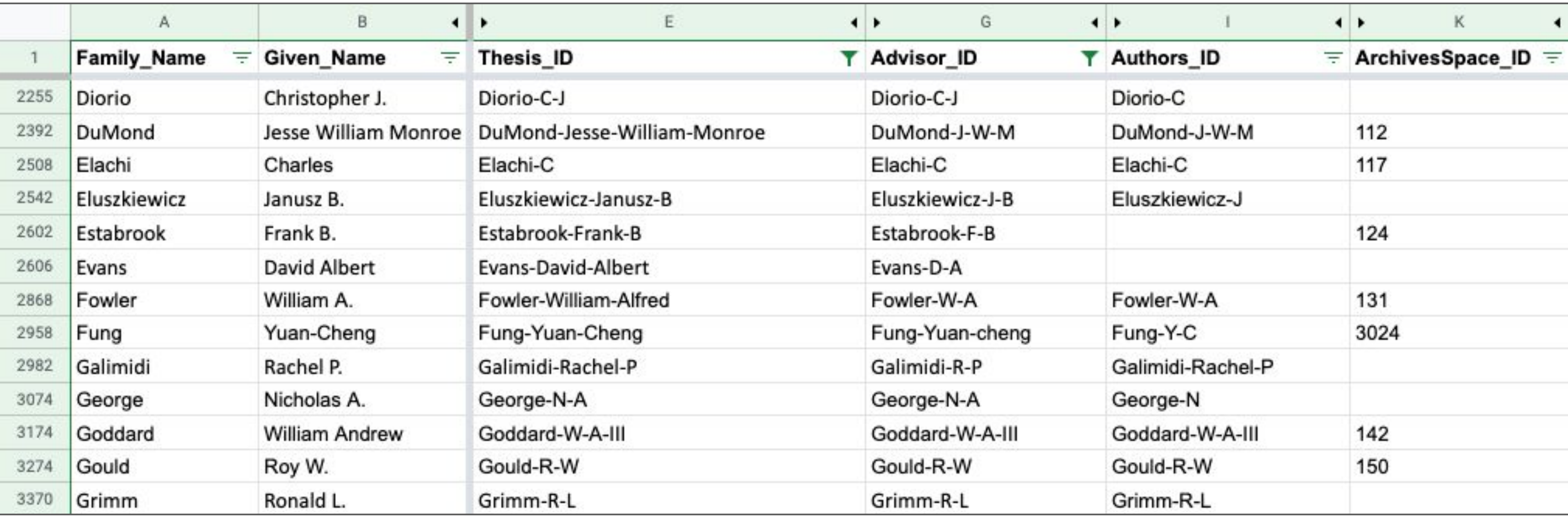
- Combined (270) HTML, HTML Include, Markdown, BibTeX, JSON, RSS
- Article(s) (264) HTML, HTML Include, Markdown, BibTeX, JSON, RSS
- Book Section(s) (1) HTML, HTML Include, Markdown, BibTeX, JSON, RSS
- Conference Item(s) (3) HTML, HTML Include, Markdown, BibTeX, JSON, RSS
- Report or Paper(s) (2) HTML, HTML Include, Markdown, BibTeX, JSON, RSS

CaltechDATA

- Combined (13) HTML, HTML Include, Markdown, BibTeX, JSON, RSS
- Dataset (9) HTML, HTML Include, Markdown, BibTeX, JSON, RSS
- Software (2) HTML, HTML Include, Markdown, BibTeX, JSON, RSS
- Model (1) HTML, HTML Include, Markdown, BibTeX, JSON, RSS
- Text (1) HTML, HTML Include, Markdown, BibTeX, JSON, RSS

Initial Data Processing

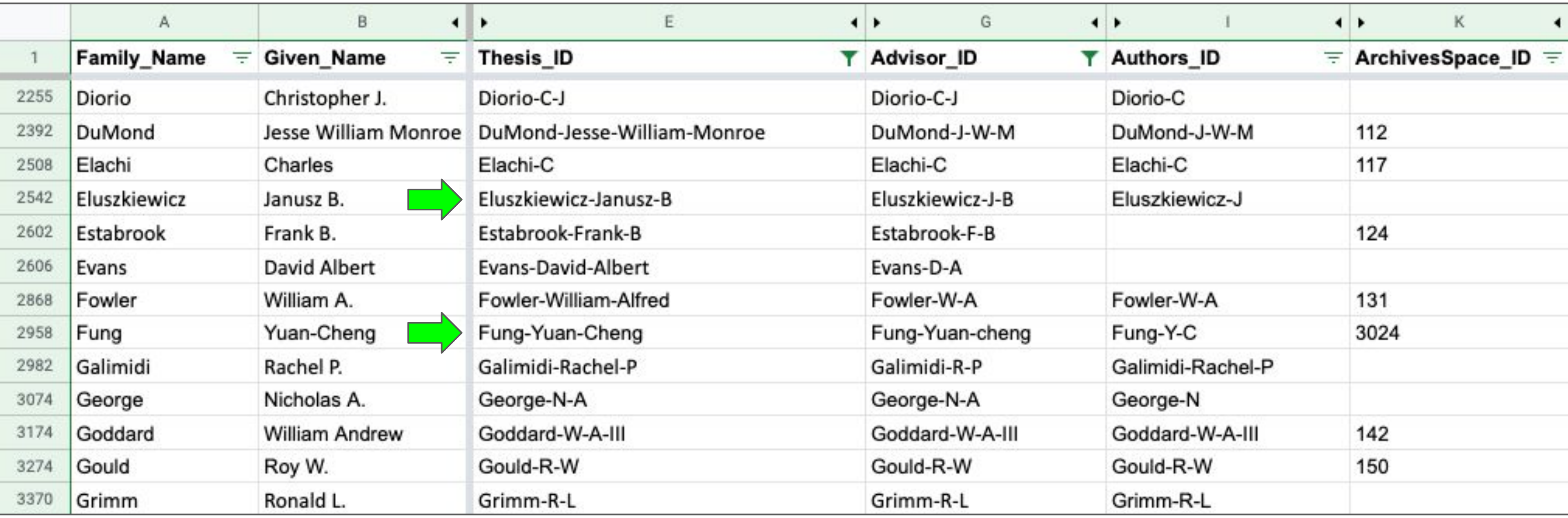
- It always starts with a spreadsheet…



	A	B	E	G	I	K
1	Family_Name	Given_Name	Thesis_ID	Advisor_ID	Authors_ID	ArchivesSpace_ID
2255	Diorio	Christopher J.	Diorio-C-J	Diorio-C-J	Diorio-C	
2392	DuMond	Jesse William Monroe	DuMond-Jesse-William-Monroe	DuMond-J-W-M	DuMond-J-W-M	112
2508	Elachi	Charles	Elachi-C	Elachi-C	Elachi-C	117
2542	Eluszkiewicz	Janusz B.	Eluszkiewicz-Janusz-B	Eluszkiewicz-J-B	Eluszkiewicz-J	
2602	Estabrook	Frank B.	Estabrook-Frank-B	Estabrook-F-B		124
2606	Evans	David Albert	Evans-David-Albert	Evans-D-A		
2868	Fowler	William A.	Fowler-William-Alfred	Fowler-W-A	Fowler-W-A	131
2958	Fung	Yuan-Cheng	Fung-Yuan-Cheng	Fung-Yuan-cheng	Fung-Y-C	3024
2982	Galimidi	Rachel P.	Galimidi-Rachel-P	Galimidi-R-P	Galimidi-Rachel-P	
3074	George	Nicholas A.	George-N-A	George-N-A	George-N	
3174	Goddard	William Andrew	Goddard-W-A-III	Goddard-W-A-III	Goddard-W-A-III	142
3274	Gould	Roy W.	Gould-R-W	Gould-R-W	Gould-R-W	150
3370	Grimm	Ronald L.	Grimm-R-L	Grimm-R-L	Grimm-R-L	

Initial Data Processing

- It always starts with a spreadsheet…



	A	B	E	G	I	K
1	Family_Name	Given_Name	Thesis_ID	Advisor_ID	Authors_ID	ArchivesSpace_ID
2255	Diorio	Christopher J.	Diorio-C-J	Diorio-C-J	Diorio-C	
2392	DuMond	Jesse William Monroe	DuMond-Jesse-William-Monroe	DuMond-J-W-M	DuMond-J-W-M	112
2508	Elachi	Charles	Elachi-C	Elachi-C	Elachi-C	117
2542	Eluszkiewicz	Janusz B.	Eluszkiewicz-Janusz-B	Eluszkiewicz-J-B	Eluszkiewicz-J	
2602	Estabrook	Frank B.	Estabrook-Frank-B	Estabrook-F-B		124
2606	Evans	David Albert	Evans-David-Albert	Evans-D-A		
2868	Fowler	William A.	Fowler-William-Alfred	Fowler-W-A	Fowler-W-A	131
2958	Fung	Yuan-Cheng	Fung-Yuan-Cheng	Fung-Yuan-cheng	Fung-Y-C	3024
2982	Galimidi	Rachel P.	Galimidi-Rachel-P	Galimidi-R-P	Galimidi-Rachel-P	
3074	George	Nicholas A.	George-N-A	George-N-A	George-N	
3174	Goddard	William Andrew	Goddard-W-A-III	Goddard-W-A-III	Goddard-W-A-III	142
3274	Gould	Roy W.	Gould-R-W	Gould-R-W	Gould-R-W	150
3370	Grimm	Ronald L.	Grimm-R-L	Grimm-R-L	Grimm-R-L	

Initial Data Processing

- It always starts with a spreadsheet…
  - Data collection grew beyond the Library-managed repositories
    - Caltech Directory (HR)
      - Contains titles and positions
    - ORCID
      - Author-managed form of name
    - VIAF
    - LCNAF
    - ISNI
    - Wikidata
    - SNAC

Formulating a Strategy

- Stakeholder Discussions
- Data Mapping
- Scripting

Formulating a Strategy

- Stakeholder Discussions
  - Library Developers, Repository Librarians, Systems Librarian, Archivists, Researchers

Formulating a Strategy

- Data Mapping
  - Phase 1
    - Name Authority will be local
      - Harvesting from other sources is a future possibility
    - Notes will contain current and historical appointments from the Caltech Directory
      - Parsing of this data is a future possibility
    - Related Agents will link to Corporate Entity records of Caltech units
      - Caltech, Caltech Faculty, Caltech Students, divisions/departments, etc.
      - Potential enhancement with Dates parsed from Directory data
    - External Documents used for links to Caltech repositories
      - Linking the disperate IDs from the spreadsheet
  - … and looking forward to enhancements

Formulating a Strategy

- Scripting
    - Starting with a CSV file containing all our data
    - We have an ad hoc ArchivesSnake/API study group
    - Began learning how to retrieve records, parse records, and update or create records

Implementation

- Local Development Environment for testing creating & updating records
  - ArchivesSpace Docker + database dump from Lyrasis
  - Challenge: could not successfully back up and restore indexes to avoid lengthy reindexing
- Leveraging ArchivesSnake and the ArchivesSpace API
  - Basically, creating a Python script using ArchivesSnake to get and post records
    - If an Agent already exists (meaning, we have an ID in the spreadsheet), then we update that existing record, otherwise we create new Agent records
- Some challenges:
  - The order of Notes matters for display in the Public UI
    - We had to prepend new notes to existing notes for our desired display order
  - **The Doozy** 😕: Related Agents are not paginated upon display (staff view & edit, public view)
    - Records with many Related Agents load that much HTML to display in a browser

Future Workflows and Questions

- Who initiates a new person record?
  - If data comes from the Directory…
  - If data comes from our Institutional Repositories…
- How much automation or human curation?
  - What is the role of Archivists in approving/rejecting new records?
  - What is the technical process?
- When do we start Phase 2?
  - Parsing dates and appointments from the Directory
  - Automatically importing from external name authorities
- How do we overcome the Related Agents issue?
  - Develop a plugin?
  - Get involved in ArchivesSpace development?

Thank You!

Tommy Keswick
tkeswick@caltech.edu

Mariella Soprano
mariella@caltech.edu

Caltech Library

Caltech Archives