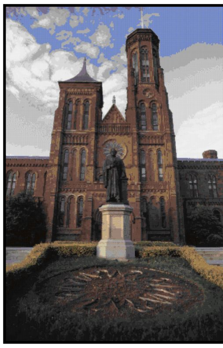


Data-intensive approaches to digitized museum collections

Dr. Rebecca B. Dikow
Data Science Lab
Smithsonian Institution

@rdikow
@SIDataScience

DATA SCIENCE LAB
Smithsonian

A diversity of locations

Heterogeneous digital data

Lack of purpose-built software tools

Smithsonian Open Access Initiative

CREATE. IMAGINE. DISCOVER.

Smithsonian

Open Access Launch Figures

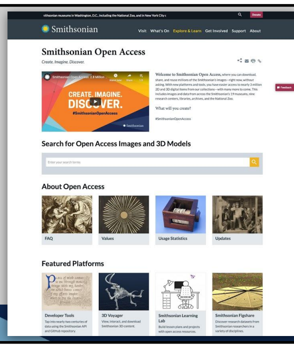
2.8 million 2D and 3D objects
14 million metadata records
173 years of staff-created data

2,809 3D models
40,500 design objects
2.67 million scientific specimen images
20,000 library volumes

Open Access Vision

Make the nation's collection available to people around the world for any purpose: to make discoveries, build new knowledge, and to develop new art and creative projects to help us see the world a little differently.

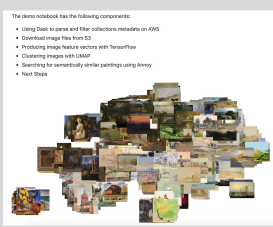
https://si.edu/openaccess



One way to engage directly with the OA collections data:

https://github.com/sidatasciencelab/siopenaccess

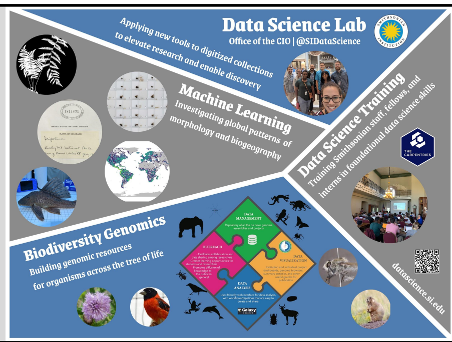
https://sidatasciencelab.github.io/siopenaccess/saam_clustering_tutorial.html



A diversity of locations

Heterogeneous digital data

Lack of purpose-built software tools

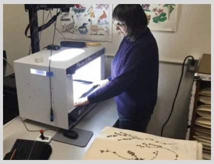


Data Science is a team effort!

Geographic patterns of morphological diversity in ferns and fern allies

Digitizing the US National Herbarium

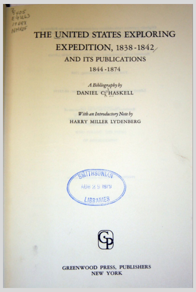
Scaling from thousands to millions:

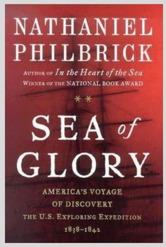


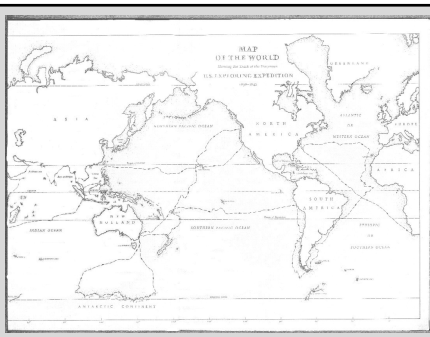
Smithsonian Institution
NATIONAL MUSEUM of NATURAL HISTORY
Smithsonian DIGITIZATION

First pilot projects: detecting mercury staining and family ID

Wilkes Expedition: 1838-1842

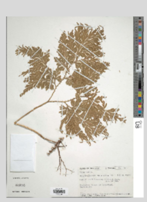






Building a training dataset: mercury staining

- Visually inspected thousands of herbarium sheets for the presence of mercuric chloride crystallization
- Final dataset had ~7K "stained" and 7K "clean" sheets, partitioned 80% for training, 20% for testing/validation



Results: mercury staining and family ID models

ACTUAL \ PREDICTED	unstained	stained
unstained	882	46
stained	77	682

ACTUAL \ PREDICTED	clubmoss	spikemoss
clubmoss	858	59
spikemoss	23	901

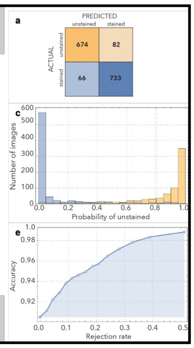
Biodiversity Data Journal 5: e21139
doi: 10.3897/BDJ.5.e21139

Research Article

Applications of deep convolutional neural networks to digitized natural history collections

Eric Schuettpelz[‡], Paul B. Frandsen[§], Rebecca B. Dikow[§], Abel Brown[|], Sylvia Orli[‡], Melinda Peters[‡], Adam Metallo[§], Vicki A. Funk[‡], Laurence J. Dorr[‡]

‡ National Museum of Natural History, Smithsonian Institution, Washington, DC, United States of America
§ Office of the Chief Information Officer, Smithsonian Institution, Washington, DC, United States of America
| NVIDIA, Santa Clara, CA, United States of America



How can we scale this work across collections?

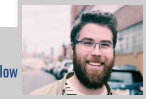
- Need to be sure collection-specific features are "masked"
- Potential sources of bias include lighting, labels, color bar, stamps, barcodes

White et al., in review, *Applications in Plant Sciences*

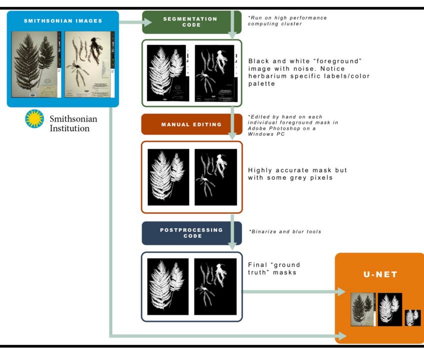
Applications in Plant Sciences
APPLICATION ARTICLE
INVITED SPECIAL ARTICLE
For the Special Issue: Machine Learning in Plant Biology: Advances Using Herbarium Specimen Images

Generating segmentation masks of herbarium specimens and a data set for training segmentation models using deep learning

Alexander E. White[1,2], Rebecca B. Dikow[1,4], Makinnon Baugh[3], Abigail Jenkins[3], and Paul B. Frandsen[1,3]

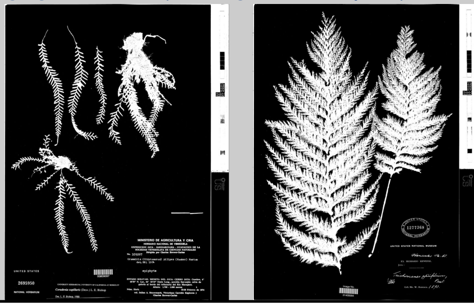


Alex White, postdoctoral fellow

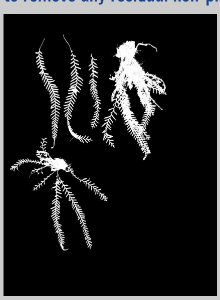


Original Smithsonian images:

After running segmentation code (built using PlantCV and OpenCV):



After manual processing to remove any residual non-plant material:



These processed images are called masks: images of identical resolution that define the identity of each pixel in the original image.

400 ground-truth masks were used to train a U-Net:

U-Net: Convolutional Networks for Biomedical Image Segmentation

Olaf Ronneberger, Philipp Fischer, and Thomas Brox

Computer Science Department and BIOSS Centre for Biological Signalling Studies,
University of Freiburg, Germany
ronneber@informatik.uni-freiburg.de,
WWW home page: http://lmb.informatik.uni-freiburg.de/

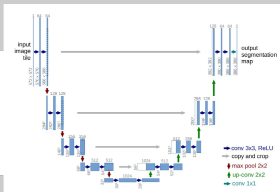
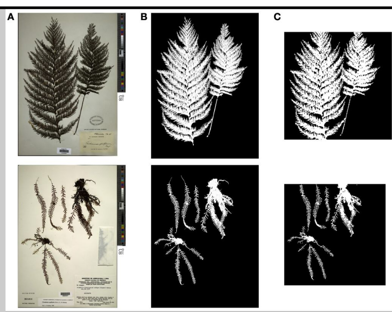


Fig. 1. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.



Original image Ground truth Unet prediction

Paper, code, model, and data available:

White et al., 2020: https://doi.org/10.1002/aps3.11352

https://github.com/sidatasciencelab/fern_segmentation

Original images (https://doi.org/10.25573/data.9922148)
Curated masks (https://doi.org/10.25573/data.9922232)
Metadata (https://doi.org/10.25573/data.11771004)

Uncovering the Scientific Impact of Women at the Smithsonian Using Machine Learning

BECAUSE OF HER STORY

Smithsonian

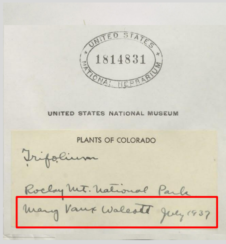
EST. 1846

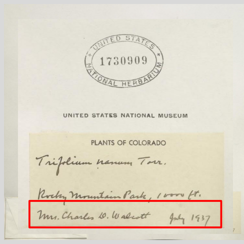
Mirian Tsuchiya, postdoctoral fellow

Mary Vaux Walcott

Mary Vaux Walcott, July 1937

Mrs. Charles B. Walcott, July 1937





More details about this project: https://datascience.si.edu/news/whatsinaname

The Funk List:

includes more than 400 current and past Smithsonian women in science

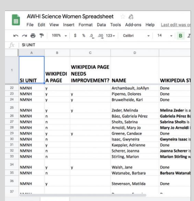


Photo of Vicki Funk by Mauricio Diazgranados

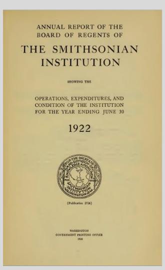
How do we measure scientific impact?

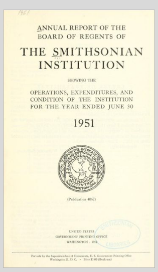
- Publications
- Service
- Collections
- Public outreach

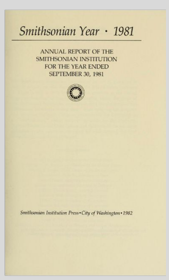
From left to right: Vicki Funk, Sophie Lutterlough, and Jessie Cohen

Machine learning tools can help us connect women on the Funk List to Smithsonian archives and collections data to help us better understand their scientific impact.

Smithsonian Annual Reports

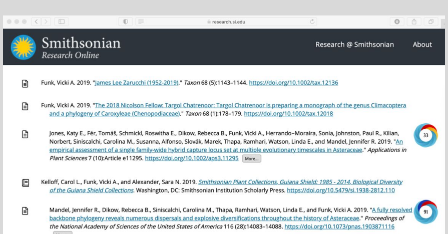






Biodiversity Heritage Library

SRO – Smithsonian Research Online



Sample search for publications by Vicki Funk

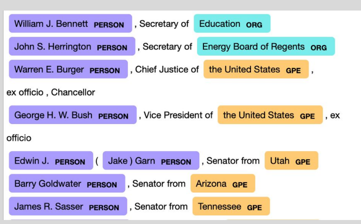
Smithsonian Libraries

Methods

Used a combination of Natural Language Processing in spaCy and shell scripting to:

- Extract and count mentions of women on the Funk List in Annual Reports
- Count publications for women on the Funk List from Smithsonian Research Online
- Extract and count occurrences first names and words related to science in the Annual Reports

NER – Named Entity Recognition

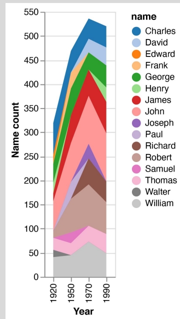


A portion of the 1985 Annual Report - this section lists the members of the Board of Regents

Methods

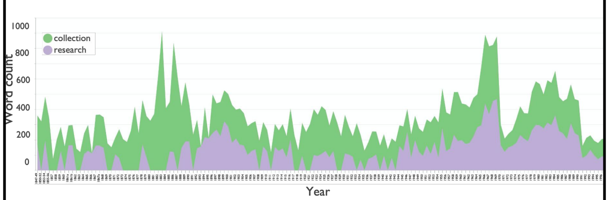
- Included all women from the Funk List no longer at the Smithsonian – 127 total
- Analyzed Annual Reports from 1846-1999
- Downloaded all citations from SRO

Count of mentions of 10 most common first names in four Annual Reports



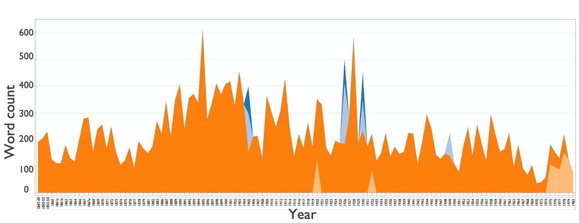
Annual Report word counts through time

Mentions of the words: research and collection



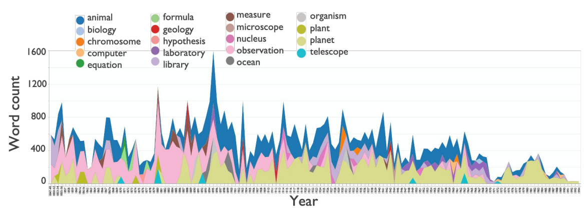
Annual Report word counts through time

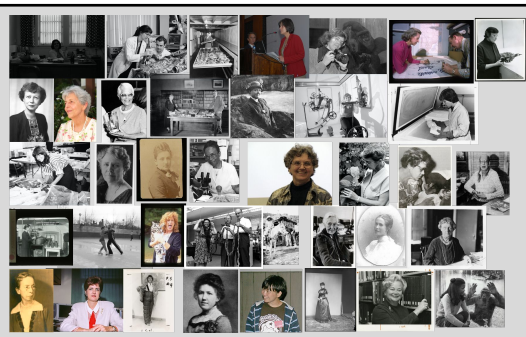
Mentions of the words: man, male, women, and female



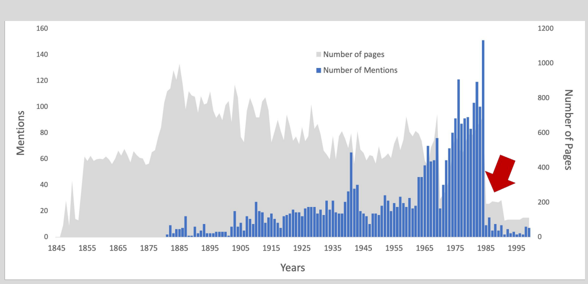
Annual Report word counts through time

Mentions of some common science words

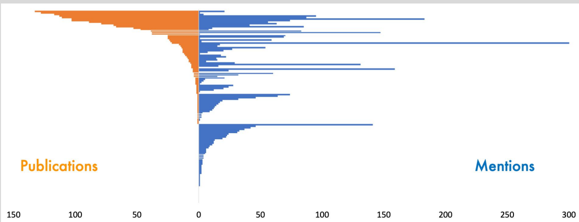




Total number of mentions per Annual Report



Number of mentions in the report do not correspond to scientific publications



Number of mentions in the report do not correspond to scientific publications

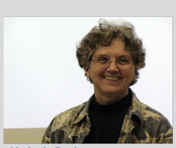
Betty Jane Meggers
NMNH
Tenure: 1954-2012
Mentions: 183
Publications: 103

Frances Theresa Densmore
NMNH
Tenure: 1907-1957
Mentions: 330
Publications: 22

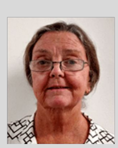
Matilda C. Stevenso
NMNH
Tenure: 1889-1915
Mentions: 159
Publications: 5

Number of mentions in the report do not correspond to scientific publications

JoGayle Howard
NZP/SCBI
Tenure: 1994-2011
Mentions: 4
Publications (until 1999): 117

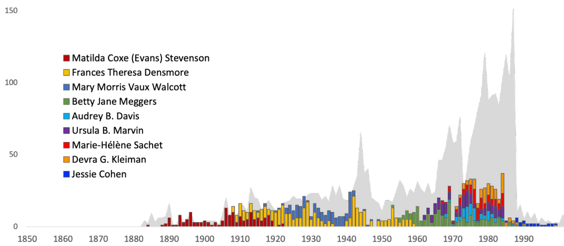


Vicki A. Funk
NMNH
Tenure: 1981-2019
Mentions: 21
Publications (until 1999): 133



Pamela B. Vandiver
MCI
Tenure: 1985-2003
Mentions: 1
Publications: 128

Who are the most-mentioned women in each decade?



find more online: https://datascience.si.edu/AWHISymposium

Many contributors, many thanks

Partners:
NMNH Botany
OCIO DPO
OCIO DAMS
Smithsonian Institution Archives
American Women's History Initiative
United States Holocaust Memorial Museum
Tiana Curry
Megan Glenn
Liz Harmon
Effie Kapsalis
Ryan King
Katrina Lohan
Grace May
Richard Naples
Jenn Schneider
Keri Thompson
Mike Trizna

Funding:
Smithsonian Women's Committee
Smithsonian Office of the Provost
Smithsonian Office of the CIO