MAKING SMITHSONIAN OPEN ACCESS ACCESSIBLE WITH PYTHON AND DASK

Smithsonian

DASK

Mike Trizna

Smithsonian OCIO Data Science Lab

May 4, 2021 | *csv,conf,v6*

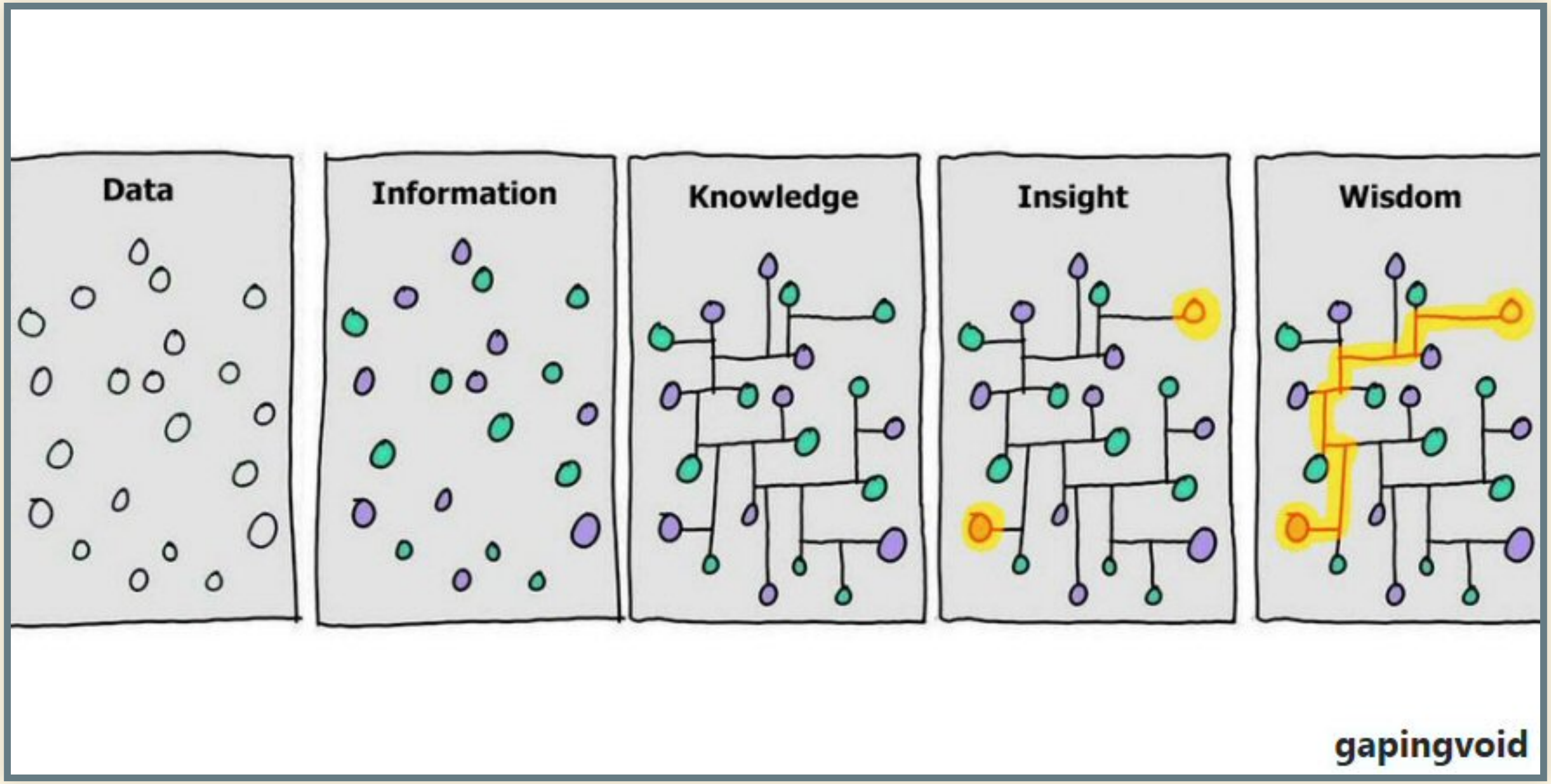
WHAT IS THE SMITHSONIAN INSTITUTION?

Yes, there are the museums (19 of them, mostly in Washington, DC), but we also have 21 libraries and archives, 9 research centers ... and a zoo.

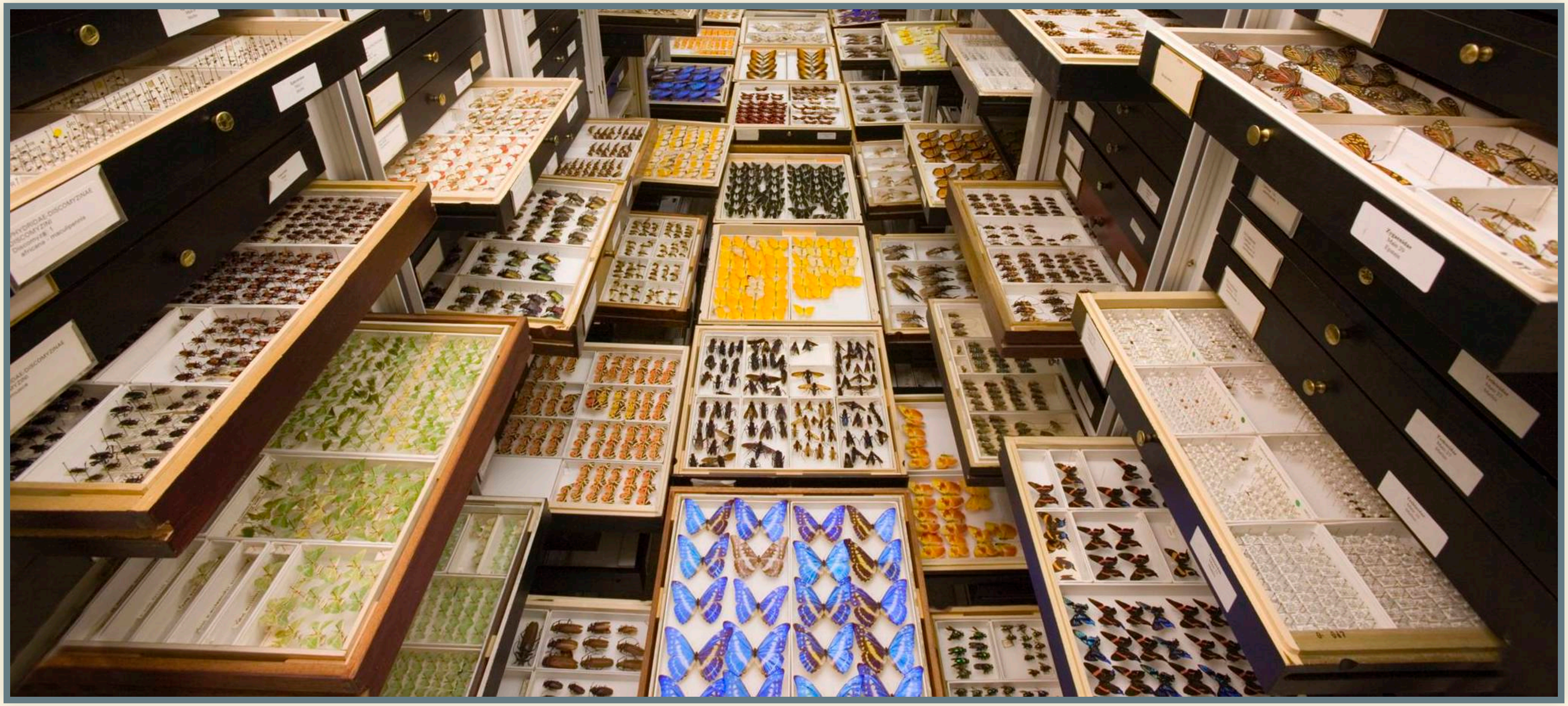
SMITHSONIAN MISSION

Founded in 1846 from the bequest of Englishman James Smithson with the condition:

> *"under the name of the Smithsonian Institution, an establishment for the* ***increase and diffusion of knowledge.***"



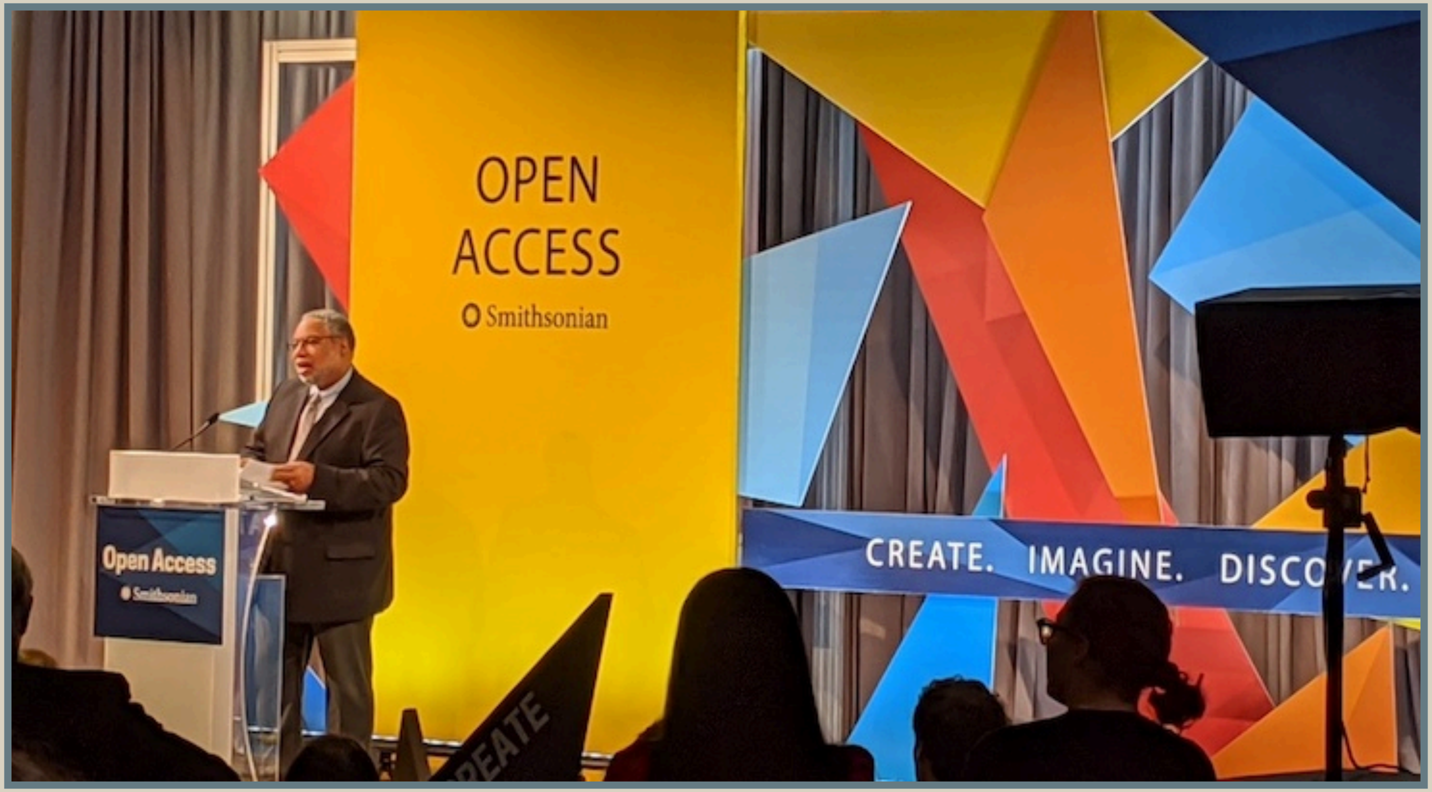
The Smithsonian has been increasing and diffusing *Knowledge* since 1846, but what about all of that *Data*?



All of that *data* and *info* that fed into *knowledge*, *insight*, *and wisdom* were dutifully cataloged and stored.

LAUNCH EVENT

February 25, 2020



SI OPEN ACCESS RELEASE

Of the Smithsonian's 155 million objects, 2.1 million library volumes and 156,000 cubic feet of archival collections:

- 2.8 million 2-D and 3-D images
- Over 17 million collection metadata objects

TERMS OF USE

Before February 2020, all Smithsonian museums and units made their data searchable and sometimes able to download, but through each individual unit. Many different use agreements.

SI Open Access put all media and metadata in one place, and all Open Access media is CC0.

SO HOW CAN ALL OF THIS DATA BE ACCESSED?

I will cover 3 different ways.

All 3 share metadata records in the same deeply-nested JSON structure.

WEB API

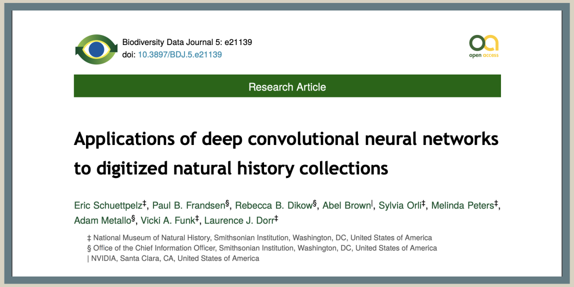
🔗: http://edan.si.edu/openaccess/apidocs/

- API Key needed (but free and painless to register)
- Great for getting a feel for record structure

HOWEVER YOU WILL QUICKLY RUN INTO LIMITATIONS

- Records are extensively indexed, but can only search indexed fields.
- Row limit of 1000 per API call

EXAMPLE: HERBARIUM SHEET MERCURY DETECTOR



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

open access

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

2017 paper that described building a machine learning model to detect herbarium sheets that had been stained with mercury.

https://doi.org/10.3897/bdj.5.e21139

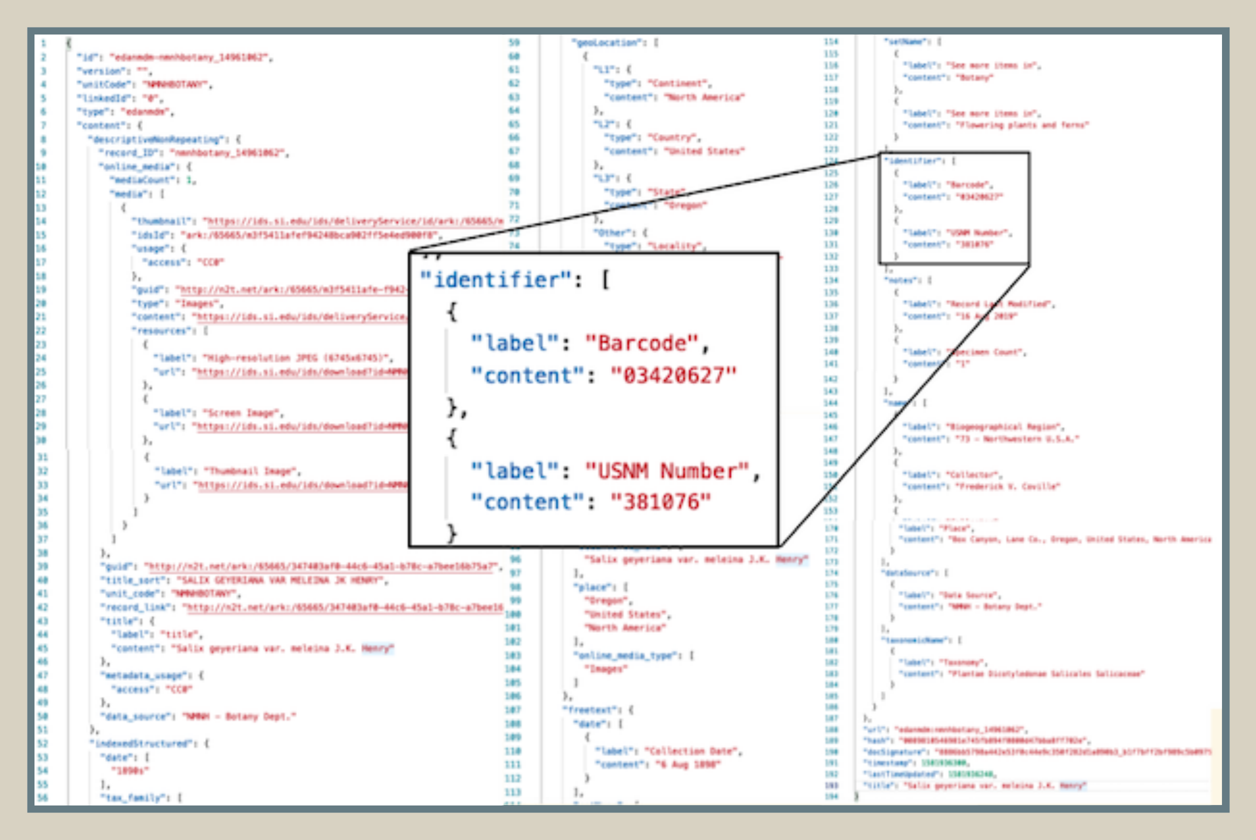
EXAMPLE: HERBARIUM SHEET MERCURY DETECTOR

I wanted to create a new model on same dataset (2017 is ancient history in Machine Learning)

All training images are shared on Figshare, but photos are resized and I wanted original metadata

EXAMPLE: HERBARIUM SHEET MERCURY DETECTOR

Unfortunately the "barcode" term from the supplementary materials is not an indexed field.



```
{
  "id": "edanmdm-nmnhbotany_14961062",
  "version": "",
  "unitCode": "NMNHBOTANY",
  "linkedId": "0",
  "type": "edanmdm",
  "content": {
    "descriptiveNonRepeating": {
      "record_ID": "nmnhbotany_14961062",
      "online_media": {
        "mediaCount": 1,
        "media": [
          {
            "thumbnail": "https://ids.si.edu/ids/deliveryService/id/ark:/65665/m
            "idsId": "ark:/65665/m3f5411afef94248bca902ff5e4ed900f8",
            "usage": {
              "access": "CC0"
            },
            "guid": "http://n2t.net/ark:/65665/m3f5411afe-f942-
            "type": "Images",
            "content": "https://ids.si.edu/ids/deliveryService,
            "resources": [
              {
                "label": "High-resolution JPEG (6745x6745)",
                "url": "https://ids.si.edu/ids/download?id=NMN
              },
              {
                "label": "Screen Image",
                "url": "https://ids.si.edu/ids/download?id=NMN
              },
              {
                "label": "Thumbnail Image",
                "url": "https://ids.si.edu/ids/download?id=NMN
              }
            ]
          }
        ]
      },
      "guid": "http://n2t.net/ark:/65665/347403af0-44c6-45a1-b78c-a7bee16b75a7",
      "title_sort": "SALIX GEYERIANA VAR MELEINA JK HENRY",
      "unit_code": "NMNHBOTANY",
      "record_link": "http://n2t.net/ark:/65665/347403af0-44c6-45a1-b78c-a7bee16
      "title": {
        "label": "title",
        "content": "Salix geyeriana var. meleina J.K. Henry"
      },
      "metadata_usage": {
        "access": "CC0"
      },
      "data_source": "NMNH - Botany Dept."
    },
    "indexedStructured": {
      "date": [
        "1890s"
      ],
      "tax_family": [
```

```
      "geoLocation": [
        {
          "L1": {
            "type": "Continent",
            "content": "North America"
          },
          "L2": {
            "type": "Country",
            "content": "United States"
          },
          "L3": {
            "type": "State",
            "content": "Oregon"
          },
          "Other": {
            "type": "Locality",
```

```
        "Salix geyeriana var. meleina J.K. Henry"
      ],
      "place": [
        "Oregon",
        "United States",
        "North America"
      ],
      "online_media_type": [
        "Images"
      ]
    },
    "freetext": {
      "date": [
        {
          "label": "Collection Date",
          "content": "6 Aug 1890"
        }
      ],
```

```
      "setName": [
        {
          "label": "See more items in",
          "content": "Botany"
        },
        {
          "label": "See more items in",
          "content": "Flowering plants and ferns"
        }
      ],
      "identifier": [
        {
          "label": "Barcode",
          "content": "03420627"
        },
        {
          "label": "USNM Number",
          "content": "381076"
        }
      ],
      "notes": [
        {
          "label": "Record Last Modified",
          "content": "16 Aug 2019"
        },
        {
          "label": "Specimen Count",
          "content": "1"
        }
      ],
      "name": [
        {
          "label": "Biogeographical Region",
          "content": "73 - Northwestern U.S.A."
        },
        {
          "label": "Collector",
          "content": "Frederick V. Coville"
        },
        {
          "label": "Place",
          "content": "Box Canyon, Lane Co., Oregon, United States, North America
        }
      ],
      "dataSource": [
        {
          "label": "Data Source",
          "content": "NMNH - Botany Dept."
        }
      ],
      "taxonomicName": [
        {
          "label": "Taxonomy",
          "content": "Plantae Dicotyledonae Salicales Salicaceae"
        }
      ]
    }
  },
  "url": "edanmdm:nmnhbotany_14961062",
  "hash": "[illegible]",
  "docSignature": "[illegible]
  "timestamp": [illegible],
  "lastTimeUpdated": [illegible],
  "title": "Salix geyeriana var. meleina J.K. Henry"
}
```

```
"identifier": [
  {
    "label": "Barcode",
    "content": "03420627"
  },
  {
    "label": "USNM Number",
    "content": "381076"
  }
```

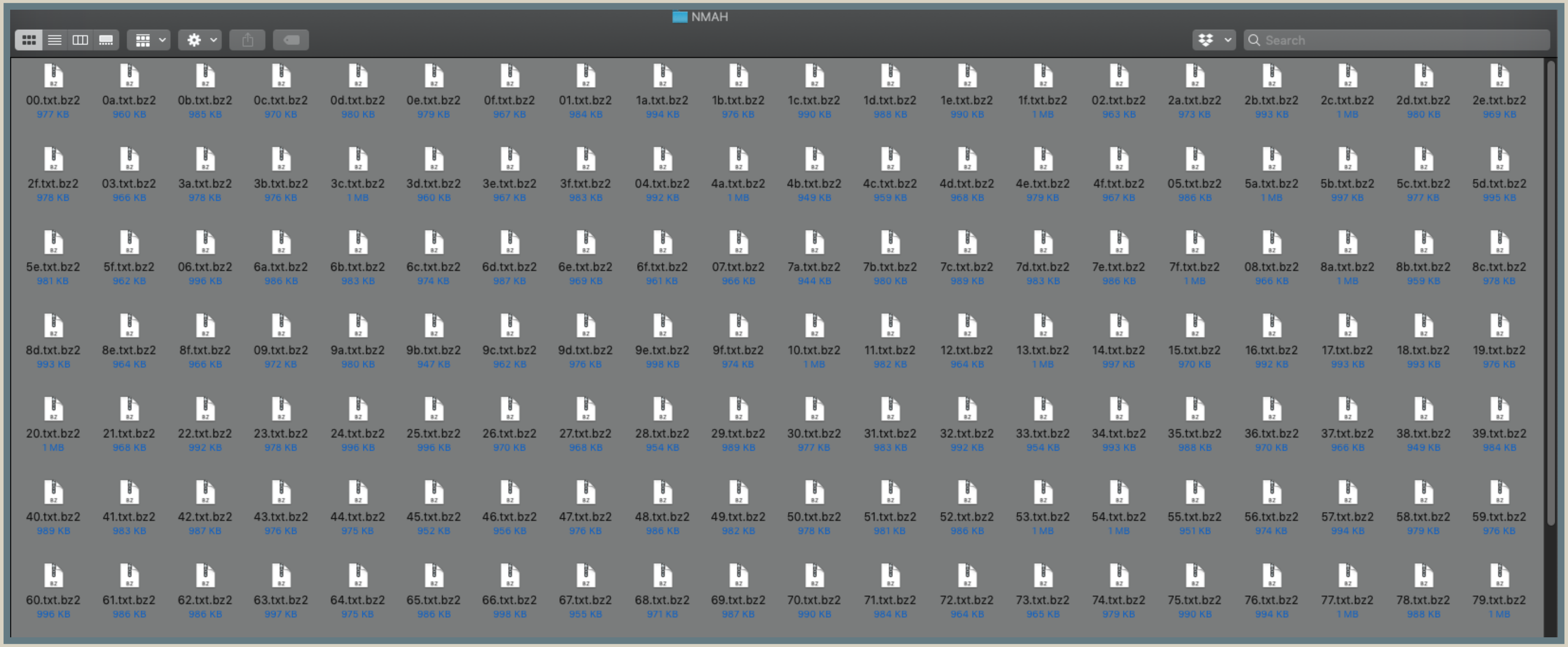
FULL DATASET SOURCES: GITHUB AND AWS

- AWS S3 for all metadata and images
- GitHub for versioned metadata

HOW IT'S PACKAGED

- Files are serialized as line-delimited JSON and compressed with bzip2.
- Directories are organized by owning unit and files are distributed by first two characters of content serialization hash.

HOW IT'S PACKAGED



00.txt.bz2 977 KB	0a.txt.bz2 960 KB	0b.txt.bz2 985 KB	0c.txt.bz2 970 KB	0d.txt.bz2 980 KB	0e.txt.bz2 979 KB	0f.txt.bz2 967 KB	01.txt.bz2 984 KB	1a.txt.bz2 994 KB	1b.txt.bz2 976 KB	1c.txt.bz2 990 KB	1d.txt.bz2 988 KB	1e.txt.bz2 990 KB	1f.txt.bz2 1 MB	02.txt.bz2 963 KB	2a.txt.bz2 973 KB	2b.txt.bz2 993 KB	2c.txt.bz2 1 MB	2d.txt.bz2 980 KB	2e.txt.bz2 969 KB
2f.txt.bz2 978 KB	03.txt.bz2 966 KB	3a.txt.bz2 978 KB	3b.txt.bz2 976 KB	3c.txt.bz2 1 MB	3d.txt.bz2 960 KB	3e.txt.bz2 967 KB	3f.txt.bz2 983 KB	04.txt.bz2 992 KB	4a.txt.bz2 1 MB	4b.txt.bz2 949 KB	4c.txt.bz2 959 KB	4d.txt.bz2 968 KB	4e.txt.bz2 979 KB	4f.txt.bz2 967 KB	05.txt.bz2 986 KB	5a.txt.bz2 1 MB	5b.txt.bz2 997 KB	5c.txt.bz2 977 KB	5d.txt.bz2 995 KB
5e.txt.bz2 981 KB	5f.txt.bz2 962 KB	06.txt.bz2 996 KB	6a.txt.bz2 986 KB	6b.txt.bz2 983 KB	6c.txt.bz2 974 KB	6d.txt.bz2 987 KB	6e.txt.bz2 969 KB	6f.txt.bz2 961 KB	07.txt.bz2 966 KB	7a.txt.bz2 944 KB	7b.txt.bz2 980 KB	7c.txt.bz2 989 KB	7d.txt.bz2 983 KB	7e.txt.bz2 986 KB	7f.txt.bz2 1 MB	08.txt.bz2 966 KB	8a.txt.bz2 1 MB	8b.txt.bz2 959 KB	8c.txt.bz2 978 KB
8d.txt.bz2 993 KB	8e.txt.bz2 964 KB	8f.txt.bz2 966 KB	09.txt.bz2 972 KB	9a.txt.bz2 980 KB	9b.txt.bz2 947 KB	9c.txt.bz2 962 KB	9d.txt.bz2 976 KB	9e.txt.bz2 998 KB	9f.txt.bz2 974 KB	10.txt.bz2 1 MB	11.txt.bz2 982 KB	12.txt.bz2 964 KB	13.txt.bz2 1 MB	14.txt.bz2 997 KB	15.txt.bz2 970 KB	16.txt.bz2 992 KB	17.txt.bz2 993 KB	18.txt.bz2 993 KB	19.txt.bz2 976 KB
20.txt.bz2 1 MB	21.txt.bz2 968 KB	22.txt.bz2 992 KB	23.txt.bz2 978 KB	24.txt.bz2 996 KB	25.txt.bz2 996 KB	26.txt.bz2 970 KB	27.txt.bz2 968 KB	28.txt.bz2 954 KB	29.txt.bz2 989 KB	30.txt.bz2 977 KB	31.txt.bz2 983 KB	32.txt.bz2 992 KB	33.txt.bz2 954 KB	34.txt.bz2 993 KB	35.txt.bz2 988 KB	36.txt.bz2 970 KB	37.txt.bz2 966 KB	38.txt.bz2 949 KB	39.txt.bz2 984 KB
40.txt.bz2 989 KB	41.txt.bz2 983 KB	42.txt.bz2 987 KB	43.txt.bz2 976 KB	44.txt.bz2 975 KB	45.txt.bz2 952 KB	46.txt.bz2 956 KB	47.txt.bz2 976 KB	48.txt.bz2 986 KB	49.txt.bz2 982 KB	50.txt.bz2 978 KB	51.txt.bz2 981 KB	52.txt.bz2 986 KB	53.txt.bz2 1 MB	54.txt.bz2 1 MB	55.txt.bz2 951 KB	56.txt.bz2 974 KB	57.txt.bz2 994 KB	58.txt.bz2 979 KB	59.txt.bz2 976 KB
60.txt.bz2 996 KB	61.txt.bz2 986 KB	62.txt.bz2 986 KB	63.txt.bz2 997 KB	64.txt.bz2 975 KB	65.txt.bz2 986 KB	66.txt.bz2 998 KB	67.txt.bz2 955 KB	68.txt.bz2 971 KB	69.txt.bz2 987 KB	70.txt.bz2 990 KB	71.txt.bz2 984 KB	72.txt.bz2 964 KB	73.txt.bz2 965 KB	74.txt.bz2 979 KB	75.txt.bz2 990 KB	76.txt.bz2 994 KB	77.txt.bz2 1 MB	78.txt.bz2 988 KB	79.txt.bz2 1 MB

BUMMER, WE NEED TO GO THROUGH EVERY SINGLE FILE ONE AT A TIME?

If I'm looking across all units, that's 9,728 files to process!

BUT THERE'S ACTUALLY A BENEFIT TO SO MANY FILES TO PROCESS

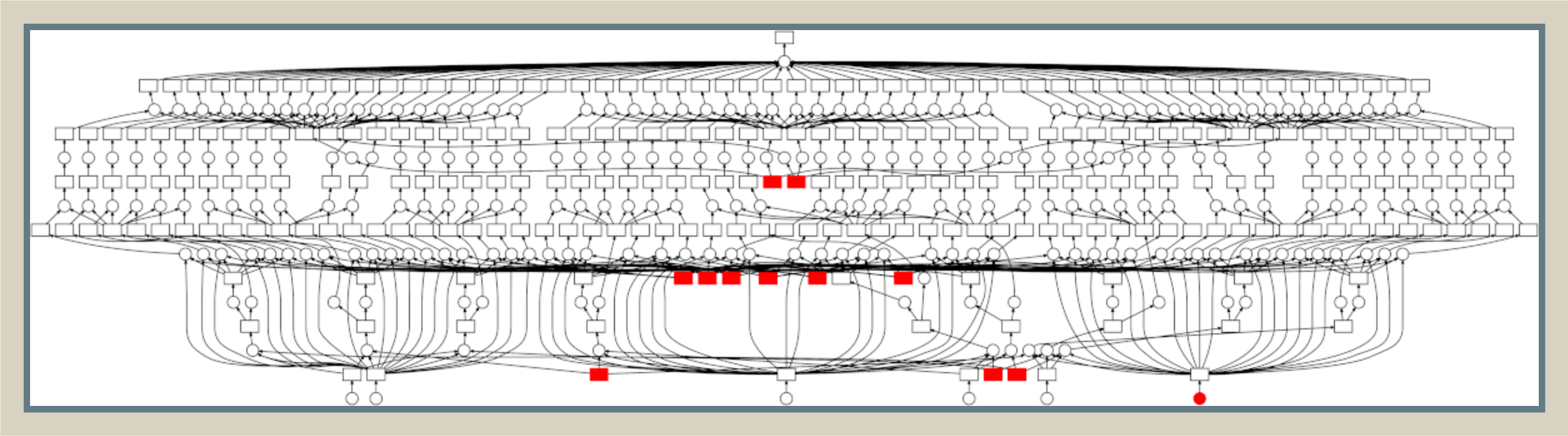
WE CAN MULTITASK!

ENTER DASK

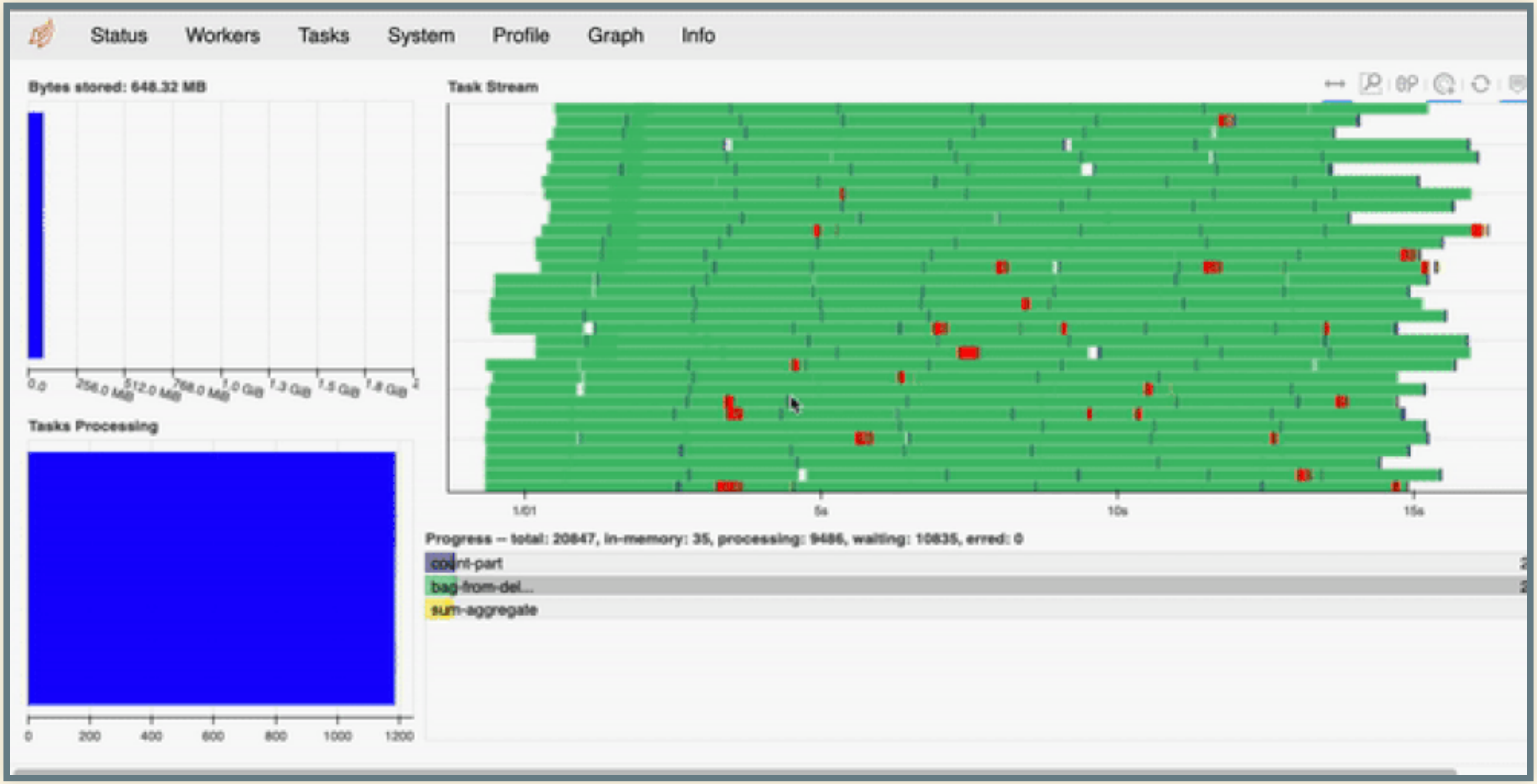
https://dask.org/

WHAT IS DASK?

Dask lets you set up a mini cluster on your machine ... or on an actual compute cluster



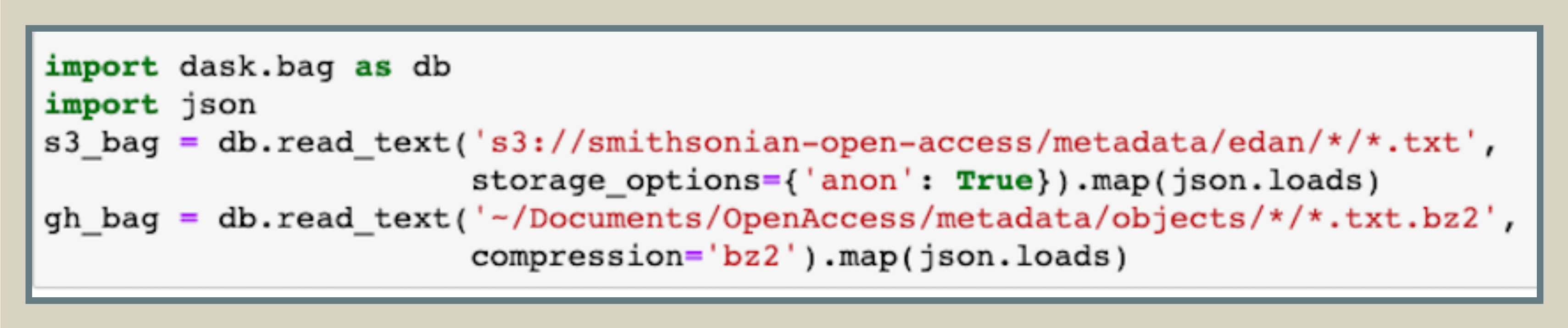
DASK DASHBOARD



Progress -- total: 20847, in-memory: 35, processing: 9486, waiting: 10835, erred: 0

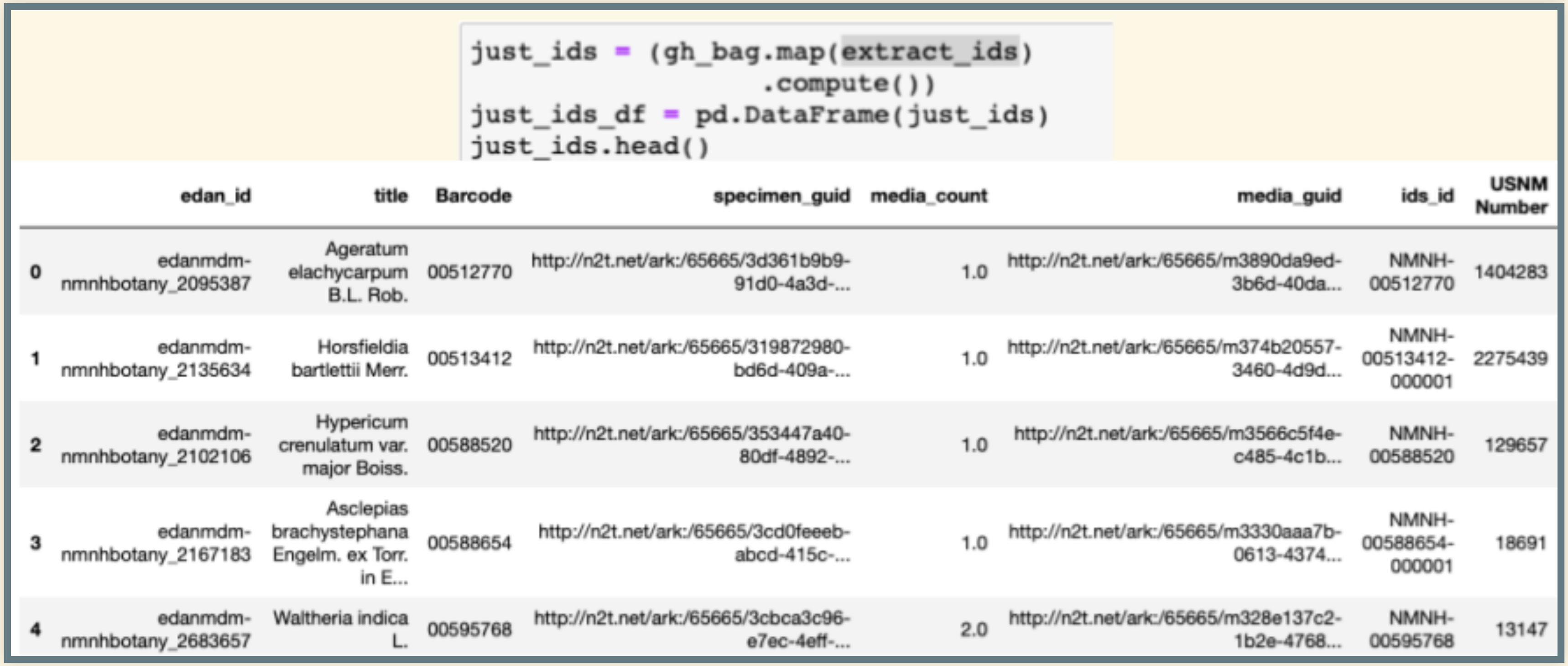
WHAT IS DASK?

Dask is more well-known for parallel processing of DataFrames, but it also contains a really useful catch-all "Bag" type.



```python
import dask.bag as db
import json
s3_bag = db.read_text('s3://smithsonian-open-access/metadata/edan/*/*.txt',
                      storage_options={'anon': True}).map(json.loads)
gh_bag = db.read_text('~/Documents/OpenAccess/metadata/objects/*/*.txt.bz2',
                      compression='bz2').map(json.loads)
```

EXAMPLE: HERBARIUM SHEET MERCURY DETECTOR

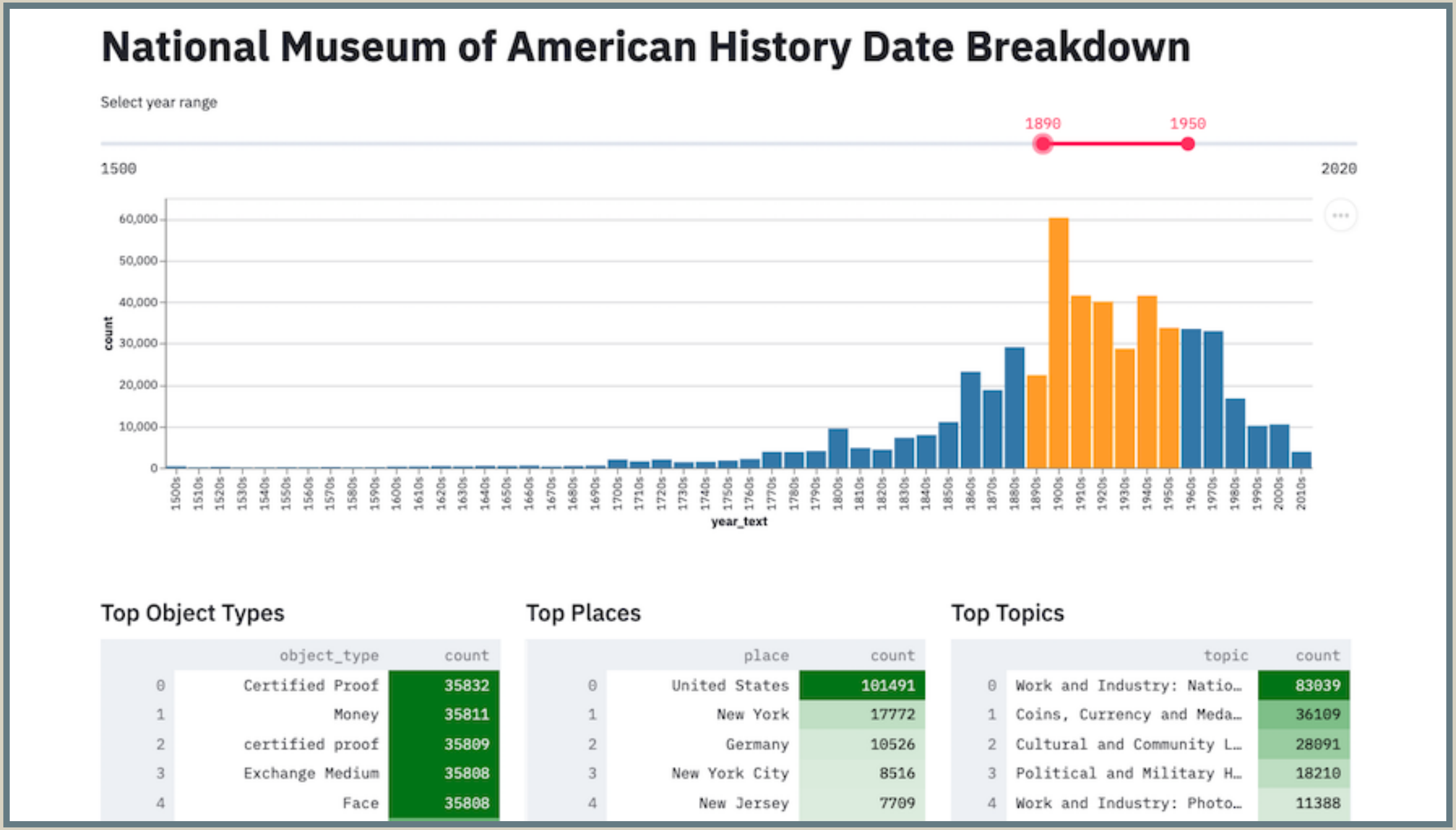


```
just_ids = (gh_bag.map(extract_ids)
                  .compute())
just_ids_df = pd.DataFrame(just_ids)
just_ids.head()
```

	edan_id	title	Barcode	specimen_guid	media_count	media_guid	ids_id	USNM Number
0	edanmdm-nmnhbotany_2095387	Ageratum elachycarpum B.L. Rob.	00512770	http://n2t.net/ark:/65665/3d361b9b9-91d0-4a3d-...	1.0	http://n2t.net/ark:/65665/m3890da9ed-3b6d-40da...	NMNH-00512770	1404283
1	edanmdm-nmnhbotany_2135634	Horsfieldia bartlettii Merr.	00513412	http://n2t.net/ark:/65665/319872980-bd6d-409a-...	1.0	http://n2t.net/ark:/65665/m374b20557-3460-4d9d...	NMNH-00513412-000001	2275439
2	edanmdm-nmnhbotany_2102106	Hypericum crenulatum var. major Boiss.	00588520	http://n2t.net/ark:/65665/353447a40-80df-4892-...	1.0	http://n2t.net/ark:/65665/m3566c5f4e-c485-4c1b...	NMNH-00588520	129657
3	edanmdm-nmnhbotany_2167183	Asclepias brachystephana Engelm. ex Torr. in E...	00588654	http://n2t.net/ark:/65665/3cd0feeeb-abcd-415c-...	1.0	http://n2t.net/ark:/65665/m3330aaa7b-0613-4374...	NMNH-00588654-000001	18691
4	edanmdm-nmnhbotany_2683657	Waltheria indica L.	00595768	http://n2t.net/ark:/65665/3cbca3c96-e7ec-4eff-...	2.0	http://n2t.net/ark:/65665/m328e137c2-1b2e-4768...	NMNH-00595768	13147

DIGITAL HUMANITIES INTERNSHIP PROJECT

PATRICK MCMANUS FROM GEORGE MASON UNIVERSITY



National Museum of American History Date Breakdown

Select year range

1890 – 1950

1500 – 2020

Top Object Types

	object_type	count
0	Certified Proof	35832
1	Money	35811
2	certified proof	35809
3	Exchange Medium	35808
4	Face	35808

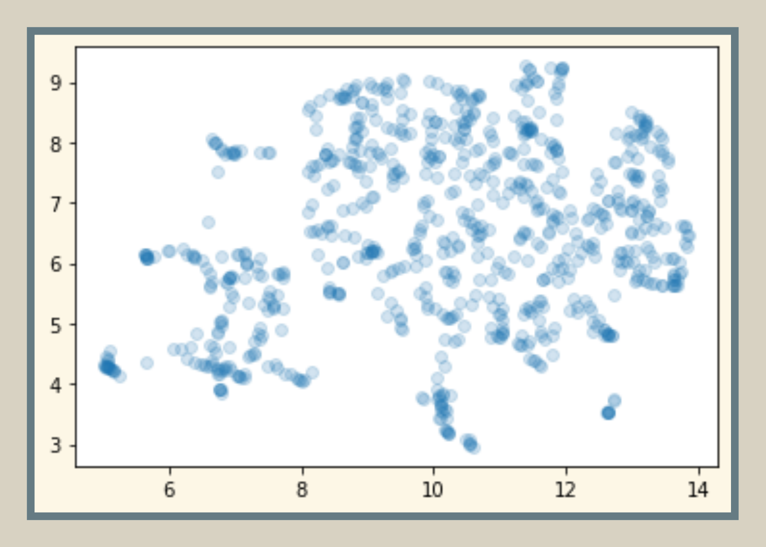
Top Places

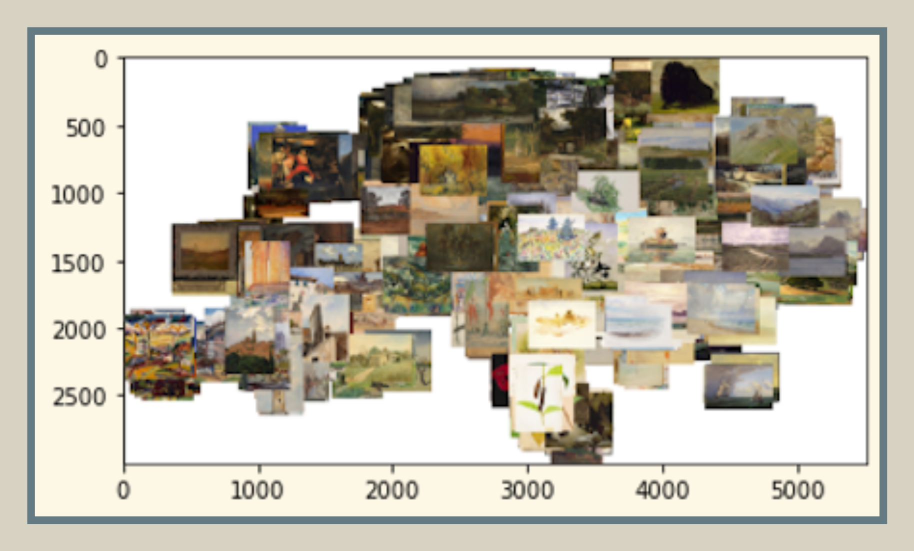
	place	count
0	United States	101491
1	New York	17772
2	Germany	10526
3	New York City	8516
4	New Jersey	7709

Top Topics

	topic	count
0	Work and Industry: Natio…	83039
1	Coins, Currency and Meda…	36109
2	Cultural and Community L…	28091
3	Political and Military H…	18210
4	Work and Industry: Photo…	11388

EXAMPLE: SEMANTIC CLUSTERING AMERICAN ART PAINTINGS





Full interactive notebook (through Binder) available at https://github.com/sidatasciencelab/siopenaccess.

QUESTIONS?

https://www.si.edu/openaccess

https://github.com/MikeTrizna/CSVConf2021_siopenacc