Using APIs to Acquire Data

Concepts and demo using R & Python to access COVID-19 datasets using Harvard Dataverse API

Philip Durbin
Will Beasley
Stefan Kasberger

Agenda

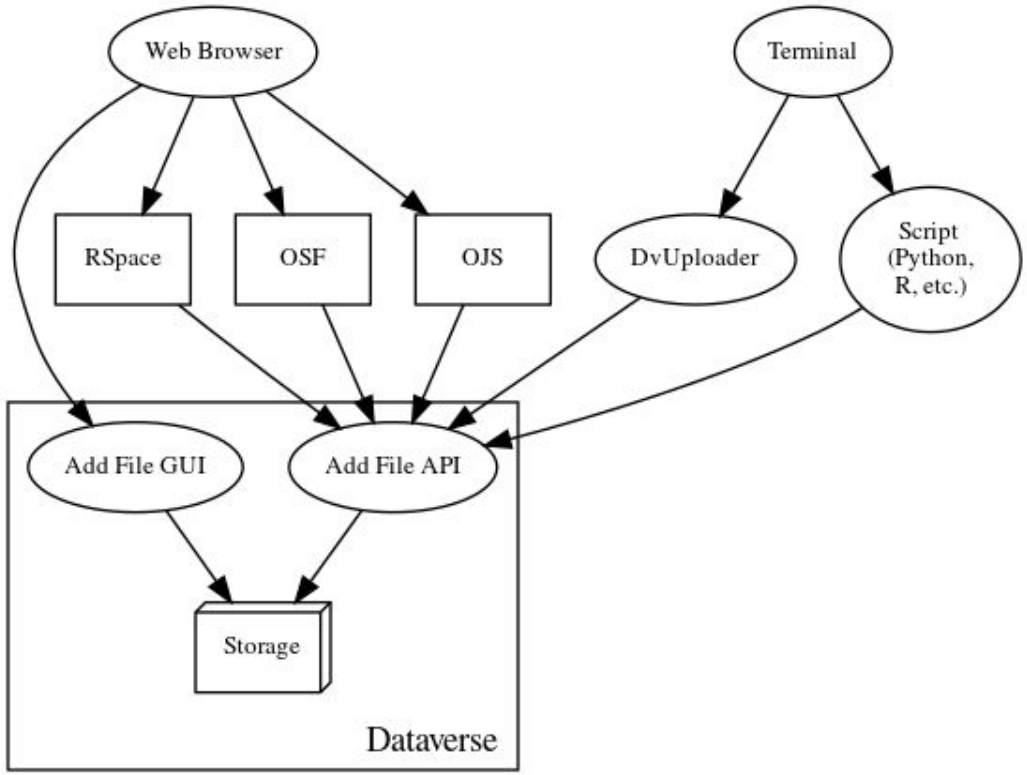
- API Concepts
- Downloading with R
- Downloading with Python
- Questions

API Concepts

What is an API?

- API = Application Programming Interface
- Traditional APIs: Tidyverse, Pandas, etc.
- Web APIs (RESTful APIs): Twitter, GitHub, etc.

Dataverse API example: Add File API



Dataverse API use cases

- Integration
- Automation
- External Tools (e.g. Data Explorer)
- Reproducibility
  - Continuously updating data
  - Published research
- ... where your imagination takes you

Downloading a COVID-19 dataset

HARVARD Dataverse

Add Data | Search | About | User Guide | Support | Sign Up | Log In

Data (China Data Lab)

Harvard Dataverse > China Data Lab Dataverse > Resources for COVID-19 > Data >

US COVID-19 Daily Cases with Basemap

Version 46.0

China Data Lab, 2020, "US COVID-19 Daily Cases with Basemap", https://doi.org/10.7910/DVN/HIDLTK, Harvard Dataverse, V46, UNF:6:I+QTrceV0xEn3GGLAskwEQ== [fileUNF]

Cite Dataset Learn about Data Citation Standards.

Access Dataset

Contact Owner | Share

Dataset Metrics

14,936 Downloads

Description Updated to Nov. 29, 2020. It contains COVID-19 Daily Cases with US basemap, including state, county-level, and metropolitan data.

Subject Earth and Environmental Sciences; Medicine, Health and Life Sciences; Social Sciences

Keyword COVID-19, US, map

Related Publication Hu, T., Guan, W., Zhu, X.,..., & Bao, S. (2020). Building an Open Resources Repository for COVID-19 Research, Data and Information Management, 4(3), 130-147. doi: https://doi.org/10.2478/dim-2020-0012

Files | Metadata | Terms | Versions

us_state_con Find

Filter by

File Type: All Access: All

Sort

1 File Download

us_state_confirmed_case.tab

Tabular Data - 85.8 KB - Nov 30, 2020 - 255 Downloads

326 Variables, 51 Observations - UNF:6:HjQAJHds25/KtR+Uet/kRg==

Preview | Metadata | Versions

Export Metadata

File Metadata

Preview

Download URL

Use the Download URL in a Wget command or a download manager to avoid interrupted downloads, time outs or other failures. User Guide - Downloading via URL

`https://dataverse.harvard.edu/api/access/datafile/4201597`

File UNF

UNF:6:HjQAJHds25/KtR+Uet/kRg==

Original File MD5

f3f907820426bb9c13279aa7ee40ba44

Publication Date

2020-11-30

Size

85.8 KB

Type

Tab-Delimited

Variables

326

Observations

51

Deposit Date

2020-11-30

Dataverse API Guide

HARVARD Dataverse

Add Data | Search | About | User Guide | Support | Sign Up | Log In

Deposit and share your data. Get academic credit.

Harvard Dataverse is a repository for research data. Deposit data and code here.

Add a dataset +

Organize datasets and gather metrics in your own repository.

A dataverse is a container for all your datasets, files, and metadata.

Add a dataverse +

Publishing your data is easy on Harvard Dataverse!

Learn about getting started creating your own dataverse repository here.

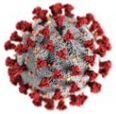
Getting started

Find data across research fields, preview metadata, and download files

Search over 103,300 datasets... | Find

VIEW ALL DATA >

Featured



COVID-19 Data Collection

A curated collection of COVID-19 data deposited in the Harvard Dataverse repository.

API Guide

Contents:

Data Access API

The Data Access API provides programmatic download access to the files stored under Dataverse. More advanced features of the Access API include format-specific transformations (thumbnail generation/resizing for images; converting tabular data into alternative file formats) and access to the data-level metadata that describes the contents of the tabular files.

Contents:

Basic File Access

Basic access URI:

`/api/access/datafile/$id`

Note

Files can be accessed using persistent identifiers. This is done by passing the constant `:persistentId` where the numeric id of the file is expected, and then passing the actual persistent id as a query parameter with the name `persistentId`.

Example: Getting the file whose DOI is *10.5072/FK2/J8SJZB*

```
GET http://$SERVER/api/access/datafile/:persistentId/?persistentId=doi:10.5072/FK2/J8SJZB
```

Parameters:

`format`

the following parameter values are supported (for tabular data files only):

Value	Description
original	“Saved Original”, the proprietary (SPSS, Stata, R, etc.) file from which the tabular data was ingested;
RData	Tabular data as an R Data frame (generated; unless the “original” file was in R);
prep	“Pre-processed data”, in JSON.
subset	Column-wise subsetting. You must also supply a comma separated list of variables in the “variables” query parameter. In this example, 123 and 127 are the database ids of data variables that belong to the data file with the id 6: `curl 'http://localhost:8080/api/access/datafile/6?format=subset&variables=123,127'`.

Basic Download By Dataset

The basic form downloads files from the latest accessible version of the dataset. If you are not using an API token, this means the most recently published version. If you are using an API token with full access to the dataset, this means the draft version or the most recently published version if no draft exists.

A curl example using a DOI (no version):

```
export API_TOKEN=xxxxxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxxx
export SERVER_URL=https://demo.dataverse.org
export PERSISTENT_ID=doi:10.70122/FK2/N2XGBJ

curl -O -J -H "X-Dataverse-key:$API_TOKEN" $SERVER_URL/api/access/d
ataset/:persistentId/?persistentId=$PERSISTENT_ID
```

The fully expanded example above (without environment variables) looks like this:

```
curl -O -J -H X-Dataverse-key:xxxxxxxx-xxxx-xxxx-xxxx-xxxxxxxxxxxx
https://demo.dataverse.org/api/access/dataset/:persistentId/?persis
tentId=doi:10.70122/FK2/N2XGBJ
```

Client Libraries (pyDataverse, etc.)

- Wrap APIs and make them more comfortable from your language.
- Ideally, provide a layer of protection against API changes.
  - (That said, if APIs change, you should complain!)

Downloading with R

https://github.com/IQSS/dataverse-client-r

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

download.R US_cases

Filter Cols: « < 1 – 50 > »

	fips	NAME	POP70	HHD70	POP80	HHD80	POP90	HHD90
1	01	Alabama	3434507	1031615	3886040	1340563	4040576	15
2	02	Alaska	225211	61547	393155	130004	549808	1
3	04	Arizona	1768275	538809	2705322	953006	3663266	13
4	05	Arkansas	1901082	608500	2253450	805730	2350107	8
5	06	California	19838084	6570174	23575384	8624938	29724503	103
6	08	Colorado	2181196	684818	2828761	1042987	3278284	12
7	09	Connecticut	2830466	872390	3103666	1093929	3285685	12
8	10	Delaware	547983	164756	594063	206973	666161	2
9	11	District of Columbia	755915	262166	638332	253124	606886	2
10	12	Florida	6609892	2225578	9536558	3667906	12936271	51
11	13	Georgia	4583982	1367090	5457519	1869746	6477997	23
12	15	Hawaii	743601	199357	931792	287576	1108041	3

Showing 1 to 13 of 51 entries, 326 total columns

https://github.com/IQSS/dataverse-client-r

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

download.R US_cases

Source on Save Run Source

```r
library("dataverse")
library("tidyverse")
# specify the Digital Object Identifier (DOI) for the dataset
DOI <- "doi:10.7910/DVN/HIDLTK"
# specify version to ensure the file doesn't change
dataset_version = 46
# retrieve the contents of the dataset
covid <- get_dataset(DOI, version = dataset_version)
# view contents
glimpse(covid, max.level = 1)
# view available files
covid$files$filename
# get data file for COVID-19 cases
US_cases_file <- get_file("us_state_confirmed_case.tab", dataset = DOI)
# read the data into a data frame
US_cases <- read_csv(US_cases_file)
# inspect the data
head(US_cases) # 50 states plus DC by 314 days
```

47 is out

28:1 (Top Level) R Script

https://github.com/IQSS/dataverse-client-r

Downloading with Python

What is pyDataverse?

- Requires: Python >=3.6
- Use-Cases: data access, data migrations
- Features: API wrapper + (meta)data manipulation and validation
- Target audience: DevOps, Data Scientists, Researcher
- Open Source → get involved!
- Release of v0.3.0 is coming

github.com/gdcc/pyDataverse

gdcc / **pyDataverse**

Code | Issues 24 | Pull requests 1 | Actions

master | 3 branches | 4 tags

skasberger Merge branch 'develop'

src/pyDataverse	change release number to v0.2.1
tests	add test_models_datafiles.py
tools	fix travis issue py27 and dist

github.com/gdcc/pyDataverse

AUSSDA

```python
# cell #1
import io
import pandas as pd
from pyDataverse.api import Api

doi = "doi:10.7910/DVN/HIDLTK"
base_url = "https://dataverse.harvard.edu"
api_token = ""

# cell #2
api = Api(base_url, api_token)
resp = api.get_dataset(doi)
datafiles = resp.json()["data"]["latestVersion"]["files"]

for df in datafiles:
    filename = df["dataFile"]["filename"]
    datafile_id = df["dataFile"]["id"]
    print(f'Filename is "{filename}", datafile ID is "{datafile_id}"')

# cell #3
datafile_id = "4274786"
resp = api.get_datafile(datafile_id)
print(resp.content)

# cell #4
data = io.StringIO(str(resp.content,'utf-8'))
us_states_cases = pd.read_csv(data, sep="\t")
print(us_states_cases.head(10))
```

Questions?