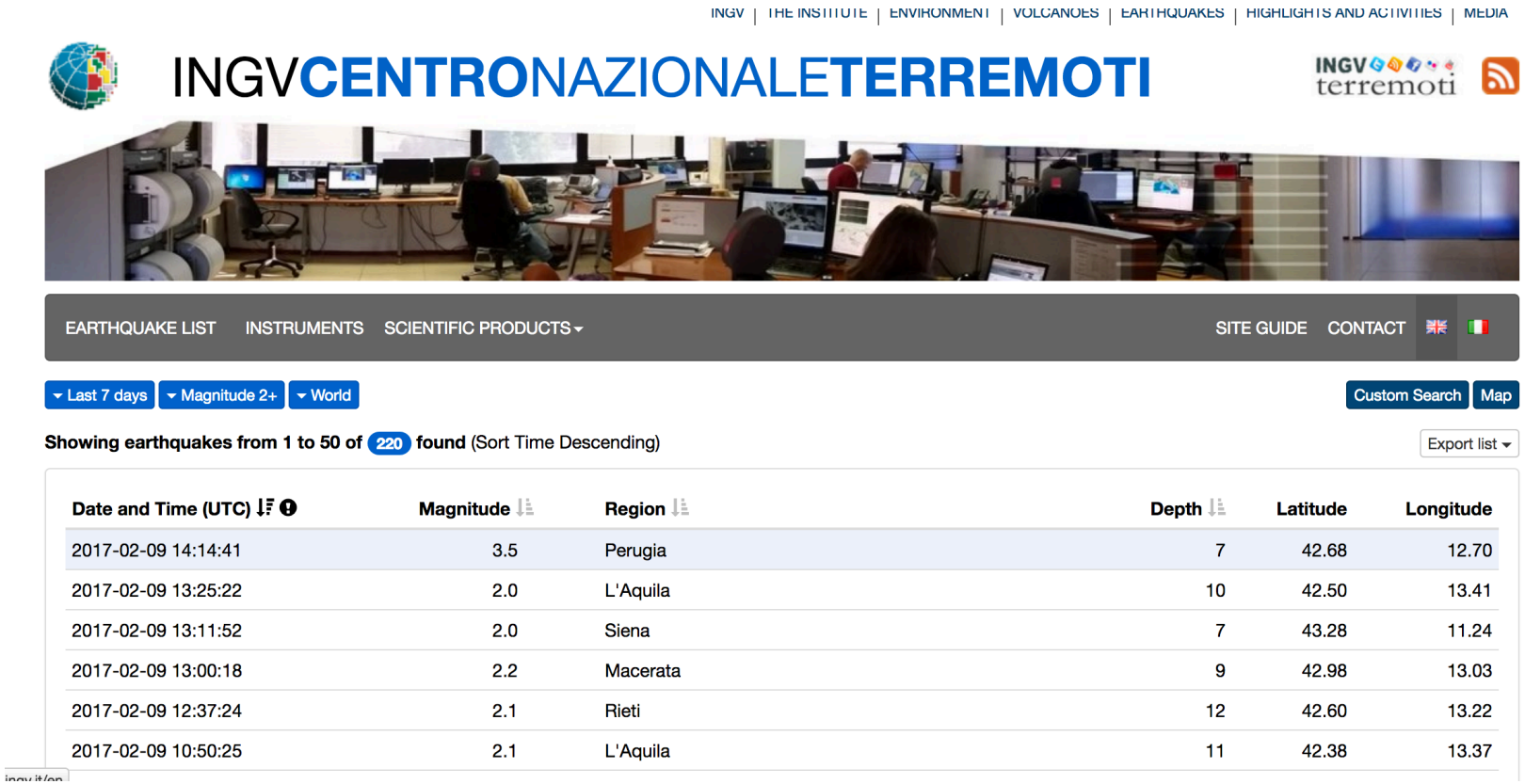
Examples of Data **Wrangling**

Rosa Filgueira

Seismology

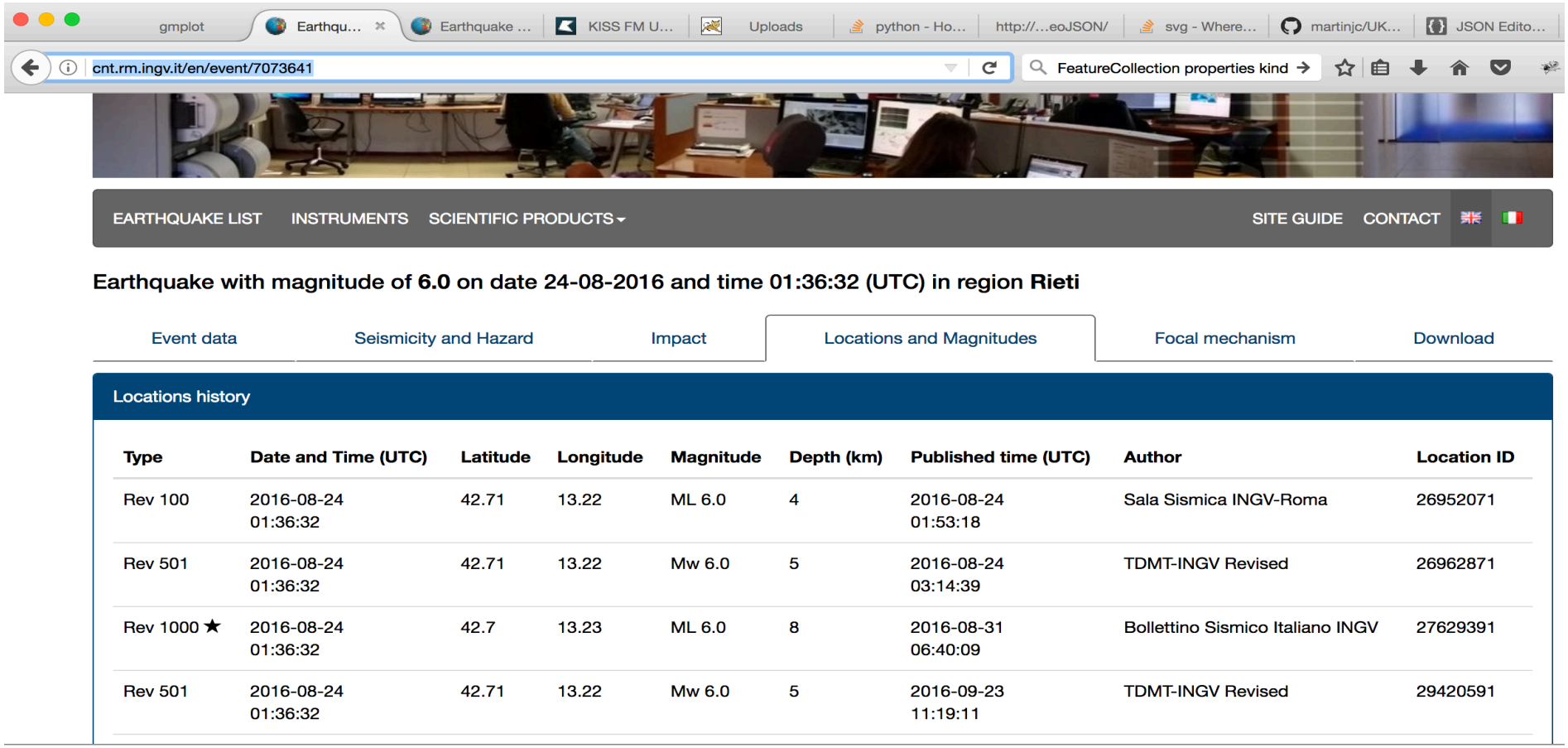
- INGV- CENTRONAZIONALE - TERREMOTI
  http://cnt.rm.ingv.it/en



Date and Time (UTC)	Magnitude	Region	Depth	Latitude	Longitude
2017-02-09 14:14:41	3.5	Perugia	7	42.68	12.70
2017-02-09 13:25:22	2.0	L'Aquila	10	42.50	13.41
2017-02-09 13:11:52	2.0	Siena	7	43.28	11.24
2017-02-09 13:00:18	2.2	Macerata	9	42.98	13.03
2017-02-09 12:37:24	2.1	Rieti	12	42.60	13.22
2017-02-09 10:50:25	2.1	L'Aquila	11	42.38	13.37

Seismology

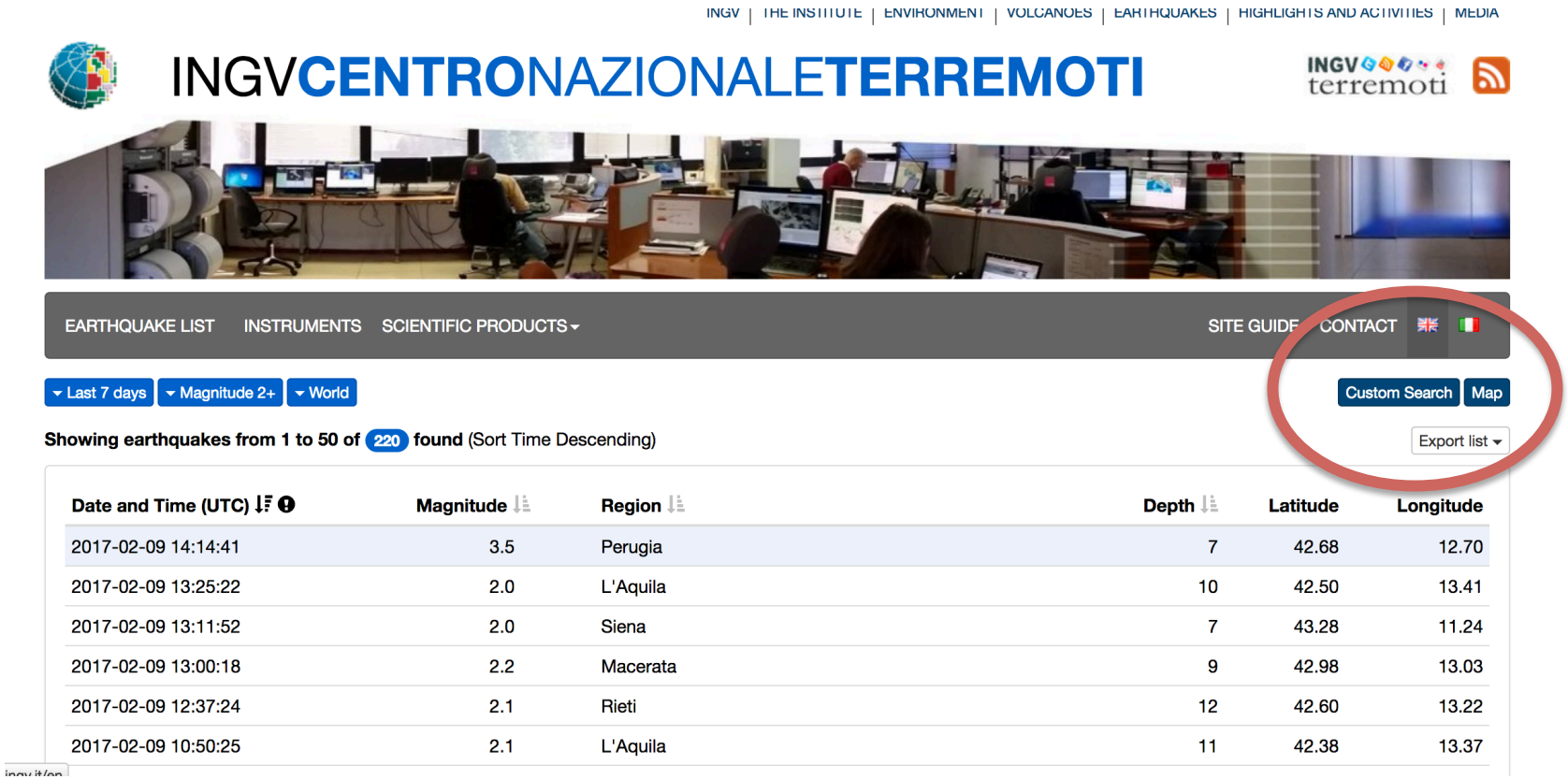
- INGV- CENTRONAZIONALE - TERREMOTI
  http://cnt.rm.ingv.it/en



Type	Date and Time (UTC)	Latitude	Longitude	Magnitude	Depth (km)	Published time (UTC)	Author	Location ID
Rev 100	2016-08-24 01:36:32	42.71	13.22	ML 6.0	4	2016-08-24 01:53:18	Sala Sismica INGV-Roma	26952071
Rev 501	2016-08-24 01:36:32	42.71	13.22	Mw 6.0	5	2016-08-24 03:14:39	TDMT-INGV Revised	26962871
Rev 1000 ★	2016-08-24 01:36:32	42.7	13.23	ML 6.0	8	2016-08-31 06:40:09	Bollettino Sismico Italiano INGV	27629391
Rev 501	2016-08-24 01:36:32	42.71	13.22	Mw 6.0	5	2016-09-23 11:19:11	TDMT-INGV Revised	29420591

Seismology

- INGV- CENTRONAZIONALE - TERREMOTI
  http://cnt.rm.ingv.it/en

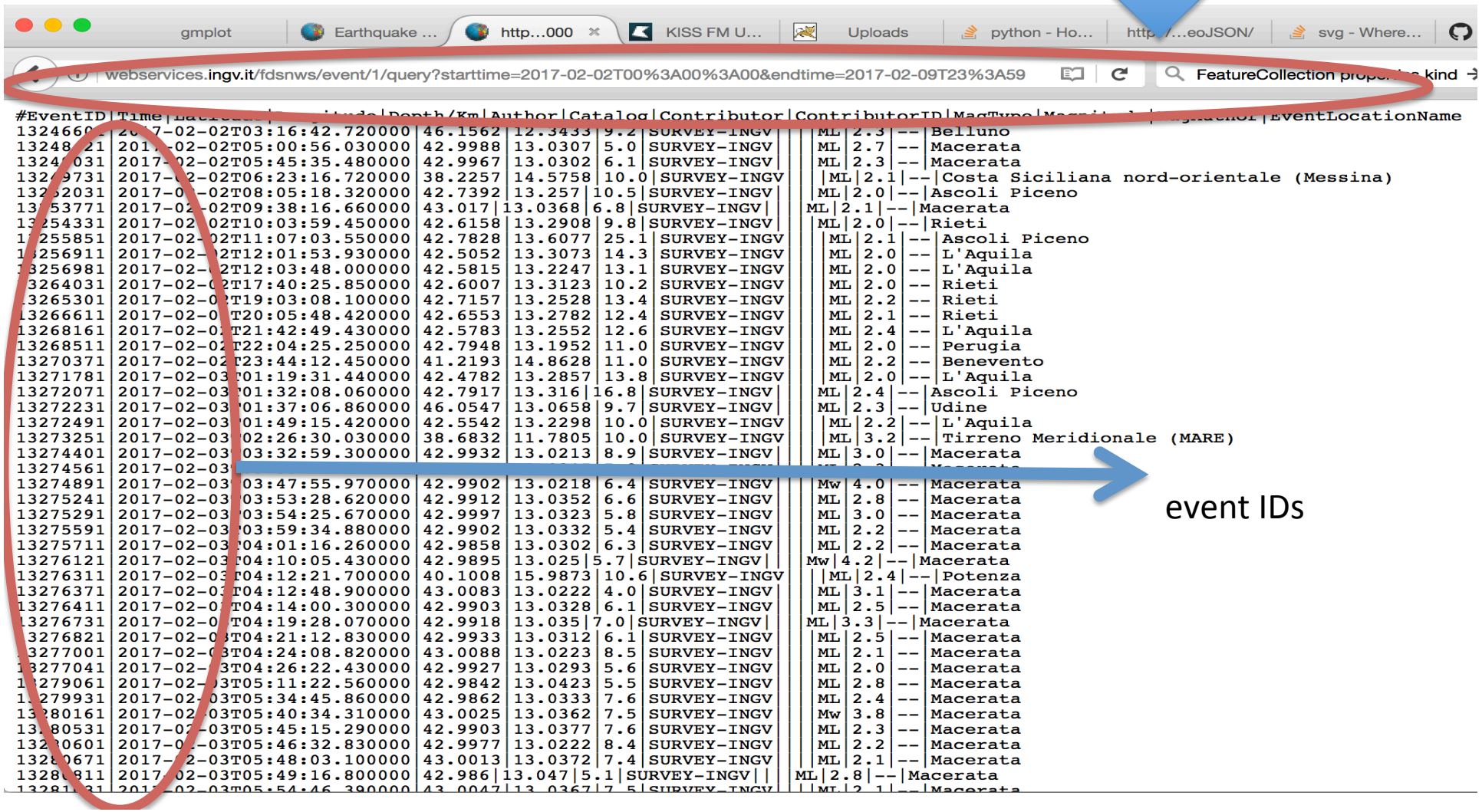


Date and Time (UTC)	Magnitude	Region	Depth	Latitude	Longitude
2017-02-09 14:14:41	3.5	Perugia	7	42.68	12.70
2017-02-09 13:25:22	2.0	L'Aquila	10	42.50	13.41
2017-02-09 13:11:52	2.0	Siena	7	43.28	11.24
2017-02-09 13:00:18	2.2	Macerata	9	42.98	13.03
2017-02-09 12:37:24	2.1	Rieti	12	42.60	13.22
2017-02-09 10:50:25	2.1	L'Aquila	11	42.38	13.37

Seismology

webservices

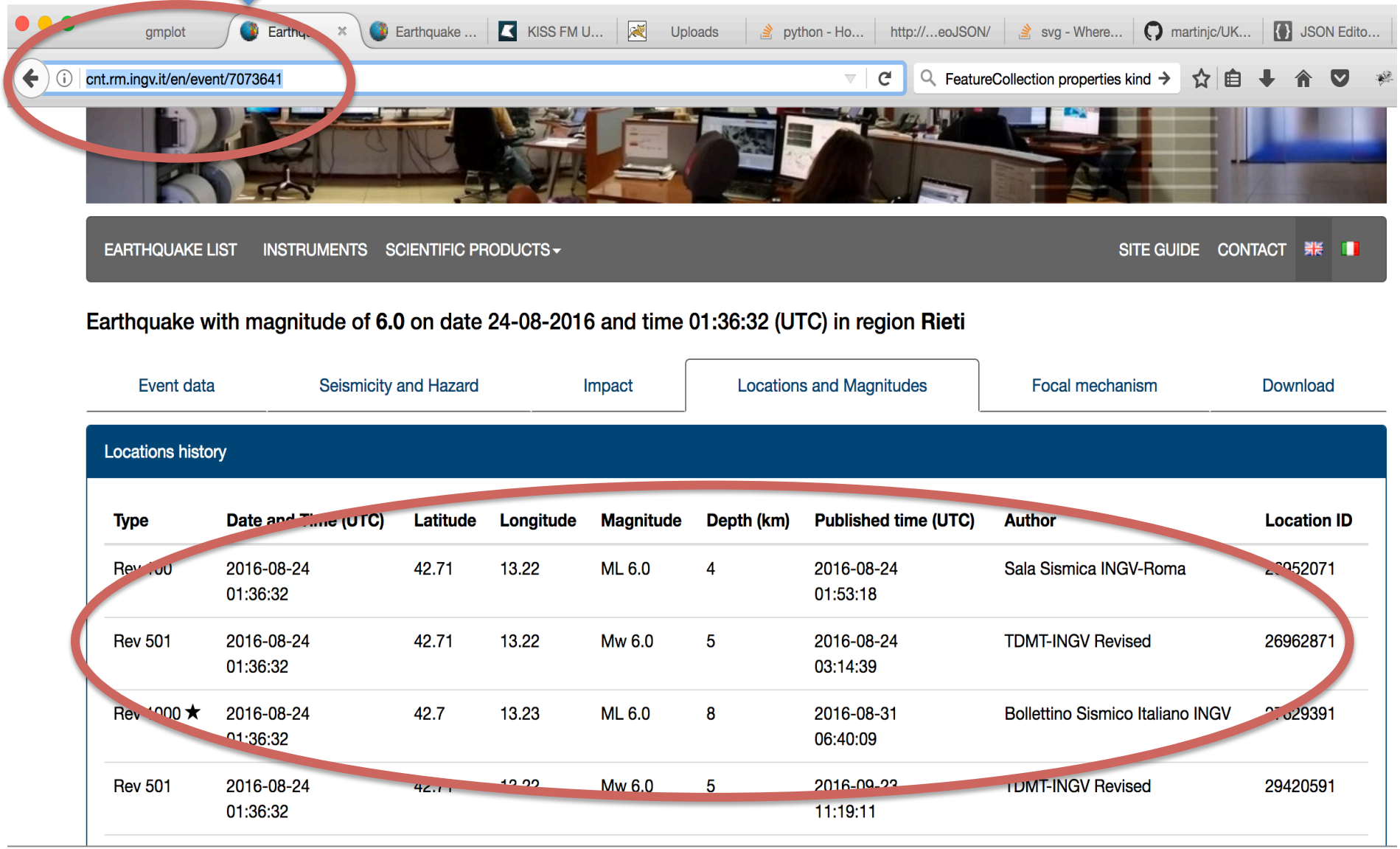
Starttime
Endtime
Magnitude
Area



event IDs

Seismology

HTML



Type	Date and Time (UTC)	Latitude	Longitude	Magnitude	Depth (km)	Published time (UTC)	Author	Location ID
Rev 500	2016-08-24 01:36:32	42.71	13.22	ML 6.0	4	2016-08-24 01:53:18	Sala Sismica INGV-Roma	26952071
Rev 501	2016-08-24 01:36:32	42.71	13.22	Mw 6.0	5	2016-08-24 03:14:39	TDMT-INGV Revised	26962871
Rev 1000 ★	2016-08-24 01:36:32	42.7	13.23	ML 6.0	8	2016-08-31 06:40:09	Bollettino Sismico Italiano INGV	27529391
Rev 501	2016-08-24 01:36:32	42.71	13.22	Mw 6.0	5	2016-09-23 11:19:11	TDMT-INGV Revised	29420591

Code - GitHub

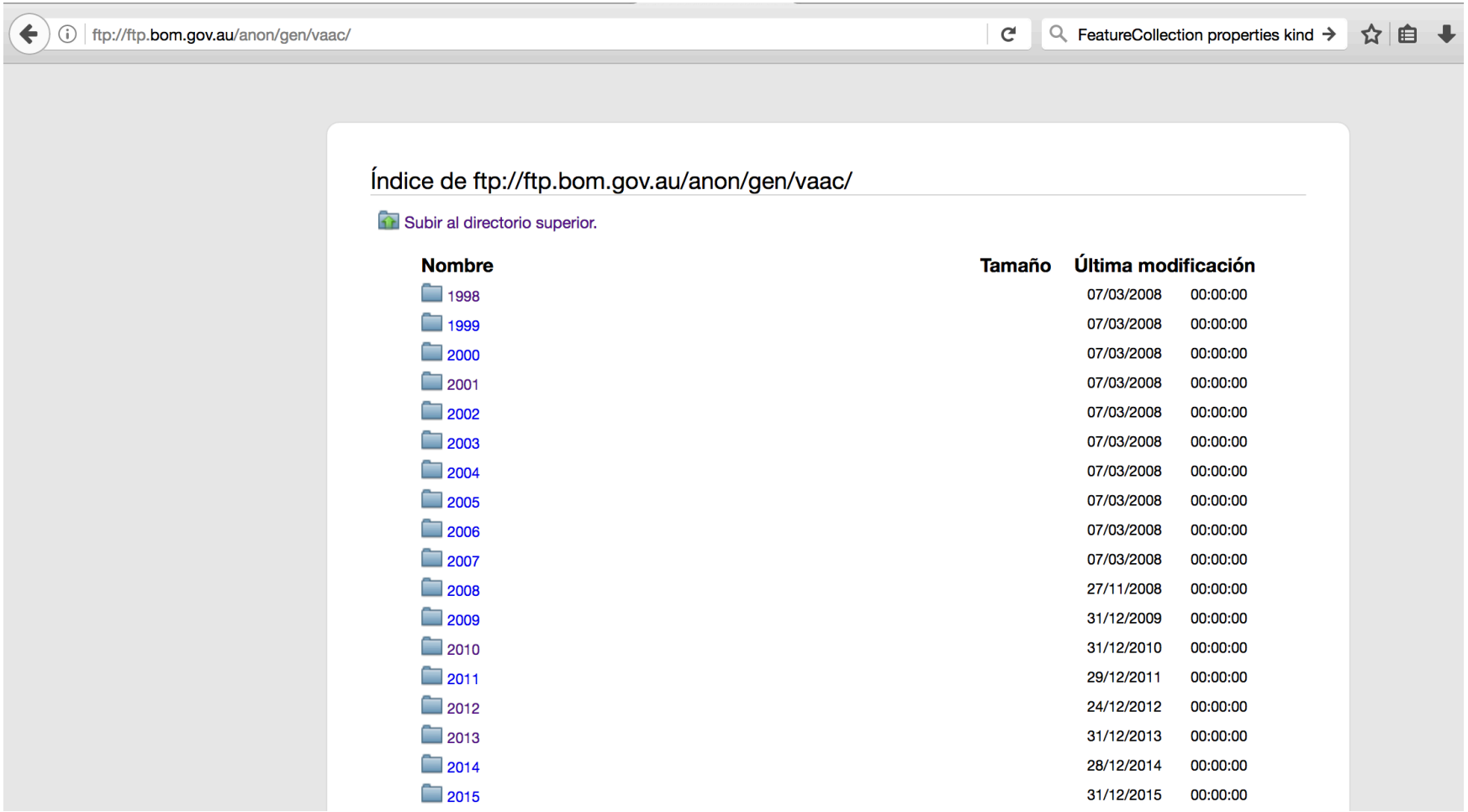
- Simple python code for downloading data in memory – I used some html libraries for parsing HTML – Everything is streamed – Not intermediate files

- It could be easily be transformed into a data-pipeline workflow to download/parser several events in parallel – not need for this now –

- https://github.com/rosafilgueira/Data_Wrangling/blob/master/Seismology_Example/download_cnt.py

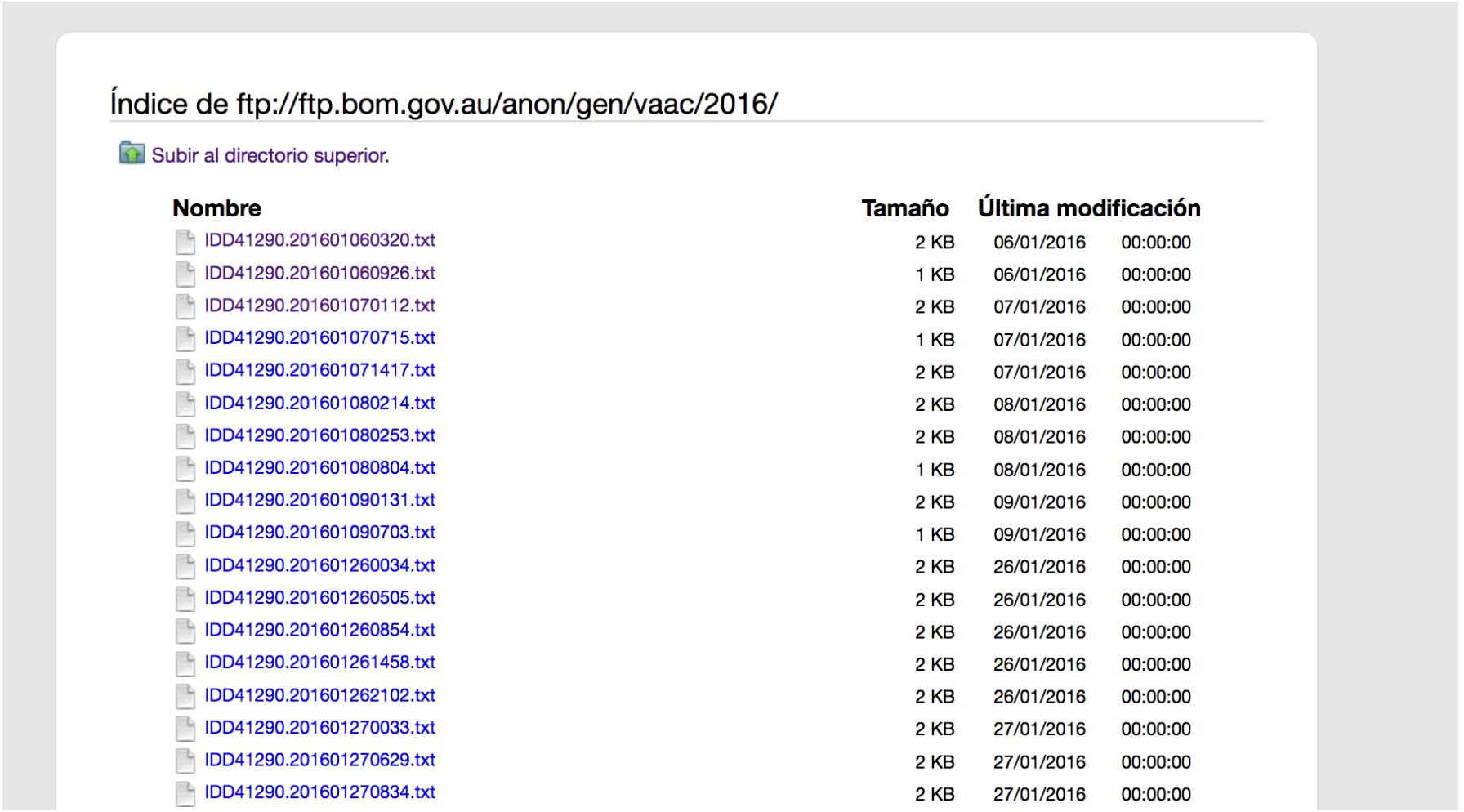
Volcanology

- **ftp**://ftp.bom.gov.au/anon/gen/vaac/



Volcanology

- ftp://ftp.bom.gov.au/anon/gen/vaac/2016



Índice de ftp://ftp.bom.gov.au/anon/gen/vaac/2016/

Subir al directorio superior.

Nombre	Tamaño	Última modificación
IDD41290.201601060320.txt	2 KB	06/01/2016 00:00:00
IDD41290.201601060926.txt	1 KB	06/01/2016 00:00:00
IDD41290.201601070112.txt	2 KB	07/01/2016 00:00:00
IDD41290.201601070715.txt	1 KB	07/01/2016 00:00:00
IDD41290.201601071417.txt	2 KB	07/01/2016 00:00:00
IDD41290.201601080214.txt	2 KB	08/01/2016 00:00:00
IDD41290.201601080253.txt	2 KB	08/01/2016 00:00:00
IDD41290.201601080804.txt	1 KB	08/01/2016 00:00:00
IDD41290.201601090131.txt	2 KB	09/01/2016 00:00:00
IDD41290.201601090703.txt	1 KB	09/01/2016 00:00:00
IDD41290.201601260034.txt	2 KB	26/01/2016 00:00:00
IDD41290.201601260505.txt	2 KB	26/01/2016 00:00:00
IDD41290.201601260854.txt	2 KB	26/01/2016 00:00:00
IDD41290.201601261458.txt	2 KB	26/01/2016 00:00:00
IDD41290.201601262102.txt	2 KB	26/01/2016 00:00:00
IDD41290.201601270033.txt	2 KB	27/01/2016 00:00:00
IDD41290.201601270629.txt	2 KB	27/01/2016 00:00:00
IDD41290.201601270834.txt	2 KB	27/01/2016 00:00:00

Volcanology

- ftp://ftp.bom.gov.au/anon/gen/vaac/2016/IDD41290.201601060320.txt

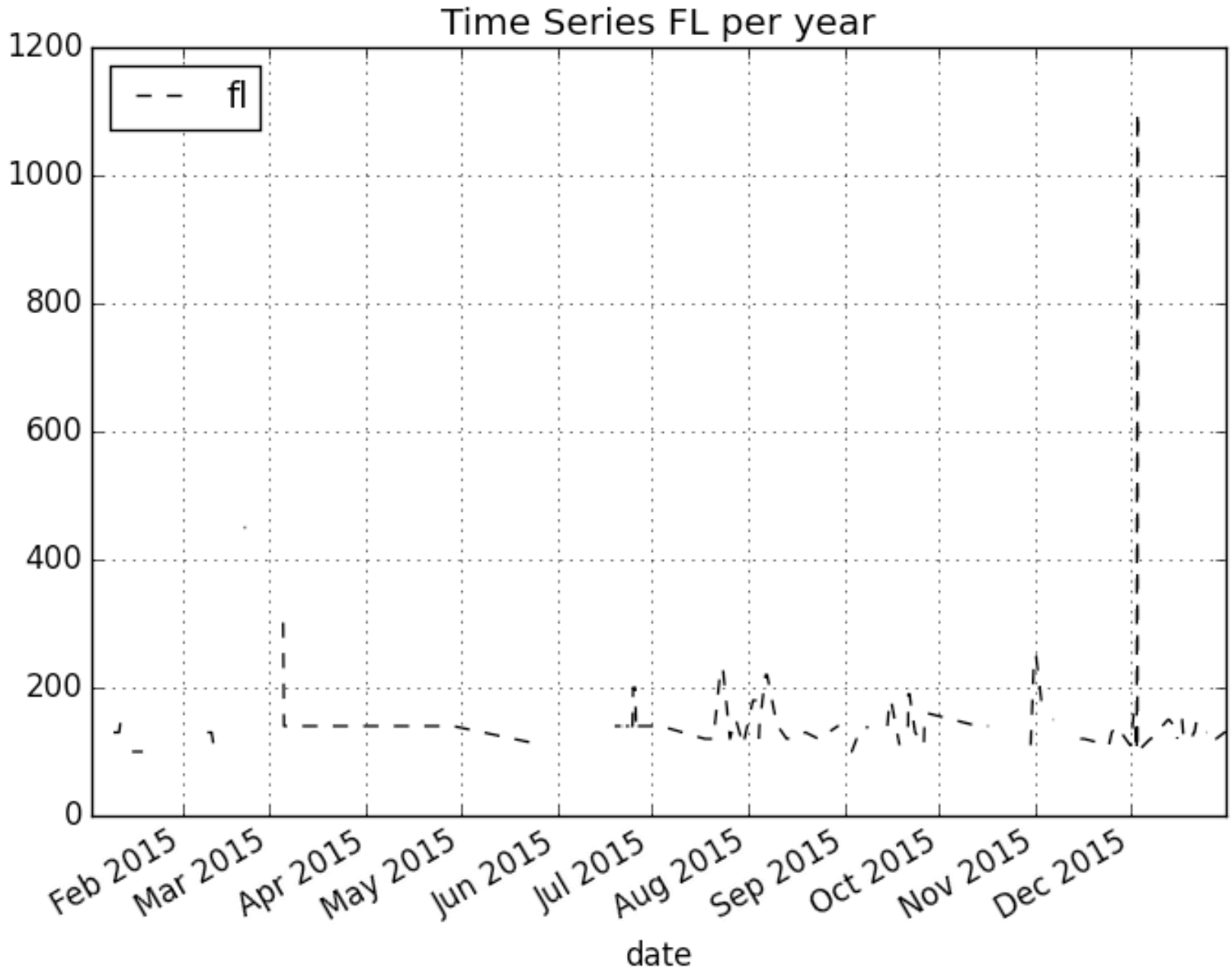
```
IDD41290

VA ADVISORY
DTG: 20160106/0330Z
VAAC: DARWIN
VOLCANO: SINABUNG 261080
PSN: N0310 E09824
AREA: INDONESIA
SUMMIT ELEV: 2460M
ADVISORY NR: 2016/1
INFO SOURCE: GROUND REPORTS, HIMAWARI-8
AVIATION COLOUR CODE: RED
ERUPTION DETAILS: MINOR VA/STEAM EMMISSION TO FL140
        EXTENDING TO THE WEST
OBS VA DTG: 06/0330Z
OBS VA CLD: SFC/FL140 N0314 E09824 - N0310 E09810 - N0305
        E09811 - N0304 E09817 - N0310 E09826 MOV SW 5KT
FCST VA CLD +6 HR: 06/0930Z SFC/FL140 N0314 E09824 - N0313
        E09811 - N0304 E09754 - N0248 E09738 - N0235 E09737 - N0222
        E09748 - N0237 E09809 - N0309 E09827
FCST VA CLD +12 HR: 06/1530Z NO VA EXP
FCST VA CLD +18 HR: 06/2130Z NO VA EXP
RMK: VA EXPECTED TO DISSIPATE WITHING 12 HRS.
NXT ADVISORY: NO LATER THAN 20160106/0930Z
Copyright Commonwealth of Australia 2011, Bureau of Meteorology (ABN 92 637 533
532).  Users of these web pages are deemed to have read and accepted the
conditions described in the Copyright, Disclaimer, and Privacy statements
```

volcano_name
colour_code
advice_date,
advice_time
time, date, fl,

Volcanology- Code - GitHub

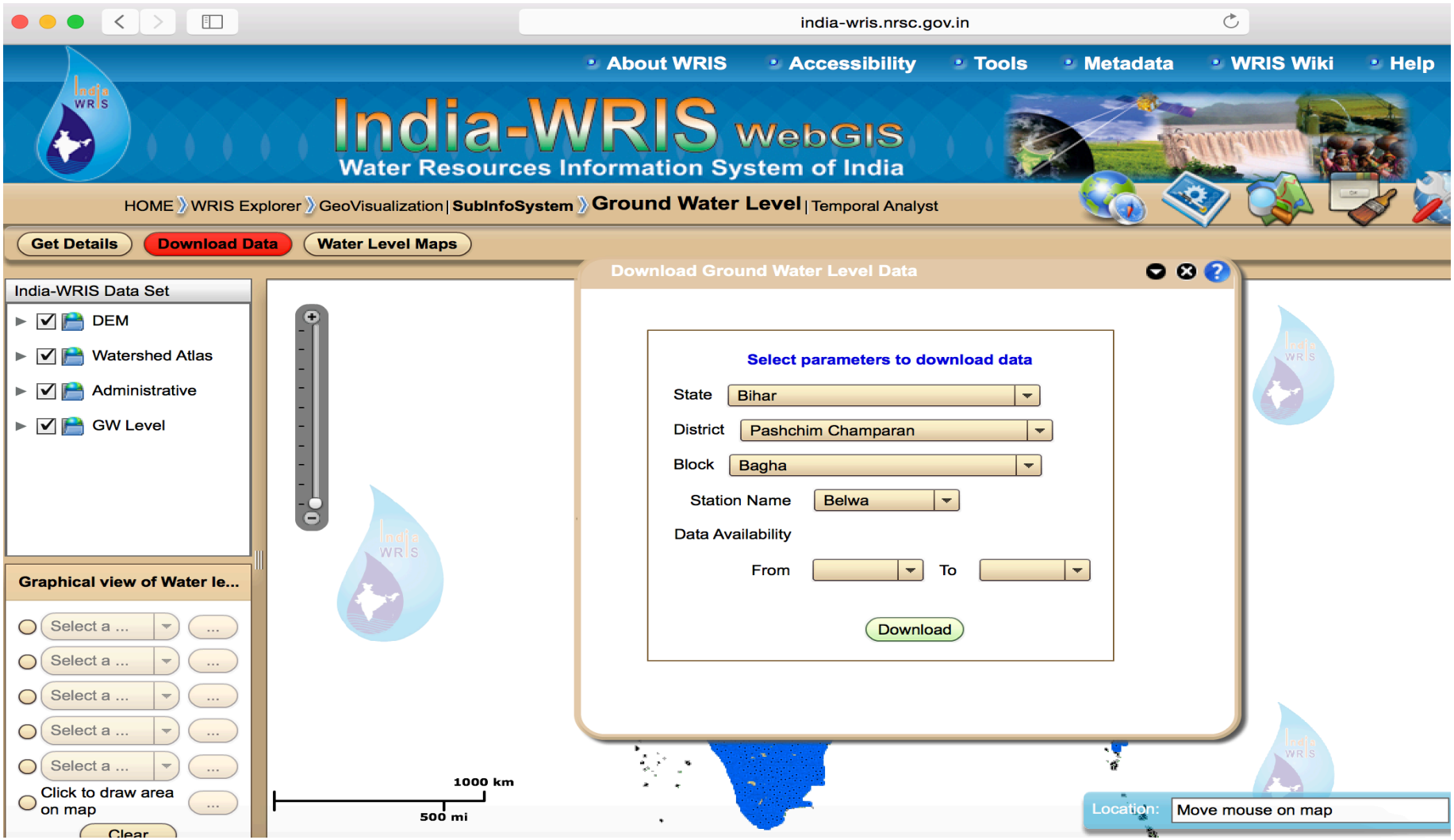
- - **Data-pipeline workflow** to download and to extract the selected data from different URLs in parallel - (new version soon -total streaming) - **dispel4py + python**
  - https://github.com/rosafilgueira/Data_Wrangling/blob/master/Volcanology_Example/vaac_preprocess.py

- **Python program** to interpolate FLs values per volcano/year – Statistics - Pandas & Python
  - https://github.com/rosafilgueira/Data_Wrangling/blob/master/Volcanology_Example/vaac_process.py


Time Series FL per year
fl
1200
1000
800
600
400
200
0
Feb 2015
Mar 2015
Apr 2015
May 2015
Jun 2015
Jul 2015
Aug 2015
Sep 2015
Oct 2015
Nov 2015
Dec 2015
date

Hydrology

http://www.india-wris.nrsc.gov.in/GWLevelApp.html?UType=R2VuZXJhbA==?UName=



Hydrology

- The data is ordered into State, District, Block, Station name, and the user interface lets you select the times for which they should be downloaded. However, it only lets you download one file at a time.

- Would you be able to write something clever so all the GW level data in the sates Bihar and Uttar Pradesh are downloaded for all districts, blocks, and stations? I won't need all of them, but it might be easier to download them all and then see which ones are within our catchment.

- Example of parameters:
  - State: Bihar
  - District: Pashchim Champaran
  - Block: Bagha
  - Station Name: Belwa
  - The range of time we need is whatever is available,

Social-Computing- Code - GitHub

- **Tweets classification – Supervised Method – Logistic Regression**
  - Labeling tweets - **Python**
  - Create and Training a classifier– **ScikitLearn ML python**
  - Applying the classifier **– Data-pipeline workflow –**
    - For labeling automatically more tweets in parallel
    - For creating statistics with the tweets labeled

- https://github.com/rosafilgueira/Data_Wrangling/tree/master/SocialComputing_Example