

Examples of Data Wrangling

Rosa Filgueira

Seismology

- INGV- CENTRONEAZIONALE - TERREMOTI
<http://cnt.rm.ingv.it/en>

INGV | THE INSTITUTE | ENVIRONMENT | VOLCANOES | EARTHQUAKES | HIGHLIGHTS AND ACTIVITIES | MEDIA

 INGVCENTRONEAZIONALETERREMOTI 



EARTHQUAKE LIST INSTRUMENTS SCIENTIFIC PRODUCTS ▾ SITE GUIDE CONTACT  

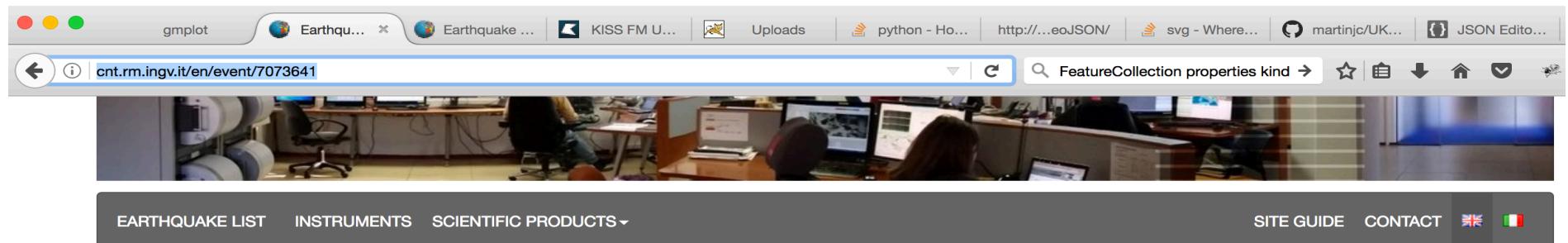
▼ Last 7 days ▼ Magnitude 2+ ▼ World Custom Search Map Export list ▾

Showing earthquakes from 1 to 50 of 220 found (Sort Time Descending)

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>



The screenshot shows a web browser window with the URL <http://cnt.rm.ingv.it/en/event/7073641>. The page displays information about a magnitude 6.0 earthquake that occurred on August 24, 2016, at 01:36:32 UTC in the Rieti region. The interface includes a navigation bar with links for EARTHQUAKE LIST, INSTRUMENTS, SCIENTIFIC PRODUCTS, SITE GUIDE, CONTACT, and flags for English and Italian. Below the navigation bar is a large image of a control room with multiple computer monitors displaying seismic data. The main content area features tabs for Event data, Seismicity and Hazard, Impact, Locations and Magnitudes (which is currently selected), Focal mechanism, and Download. The Locations and Magnitudes tab displays a table titled "Locations history" with four rows of data:

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- CENTRONEAZIONALE - TERREMOTI
<http://cnt.rm.ingv.it/en>

INGV | THE INSTITUTE | ENVIRONMENT | VOLCANOES | EARTHQUAKES | HIGHLIGHTS AND ACTIVITIES | MEDIA

 INGVCENTRONEAZIONALETERREMOTI 



EARTHQUAKE LIST INSTRUMENTS SCIENTIFIC PRODUCTS ▾ SITE GUIDE CONTACT  

▼ Last 7 days ▼ Magnitude 2+ ▼ World

Custom Search Map Export list ▾

Showing earthquakes from 1 to 50 of 220 found (Sort Time Descending)

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

webservices

Seismology

Starttime
Endtime
Magnitude
Area



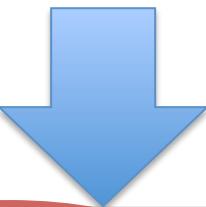
webservices.ingv.it/fdsnws/event/1/query?starttime=2017-02-02T00%3A00%3A00&endtime=2017-02-09T23%3A59

FeatureCollection property kind

#EventID	Time	Latitude	Longitude	Depth/km	Author	Catalog	Contributor	ContributorID	Magnitude	Type	EventLocationName
1324660	2017-02-02T03:16:42.720000	46.1562	12.3433	9.2	SURVEY-INGV		ML	2.3	--	Belluno	
1324821	2017-02-02T05:00:56.030000	42.9988	13.0307	5.0	SURVEY-INGV		ML	2.7	--	Macerata	
1324931	2017-02-02T05:45:35.480000	42.9967	13.0302	6.1	SURVEY-INGV		ML	2.3	--	Macerata	
13249731	2017-02-02T06:23:16.720000	38.2257	14.5758	10.0	SURVEY-INGV		ML	2.1	--	Costa Siciliana nord-orientale (Messina)	
13252031	2017-02-02T08:05:18.320000	42.7392	13.257	10.5	SURVEY-INGV		ML	2.0	--	Ascoli Piceno	
13253771	2017-02-02T09:38:16.660000	43.017	13.0368	6.8	SURVEY-INGV		ML	2.1	--	Macerata	
13254331	2017-02-02T10:03:59.450000	42.6158	13.2908	9.8	SURVEY-INGV		ML	2.0	--	Rieti	
13255851	2017-02-02T11:07:03.550000	42.7828	13.6077	25.1	SURVEY-INGV		ML	2.1	--	Ascoli Piceno	
13256911	2017-02-02T12:01:53.930000	42.5052	13.3073	14.3	SURVEY-INGV		ML	2.0	--	L'Aquila	
13256981	2017-02-02T12:03:48.000000	42.5815	13.2247	13.1	SURVEY-INGV		ML	2.0	--	L'Aquila	
13264031	2017-02-02T17:40:25.850000	42.6007	13.3123	10.2	SURVEY-INGV		ML	2.0	--	Rieti	
13265301	2017-02-02T19:03:08.100000	42.7157	13.2528	13.4	SURVEY-INGV		ML	2.2	--	Rieti	
13266611	2017-02-02T20:05:48.420000	42.6553	13.2782	12.4	SURVEY-INGV		ML	2.1	--	Rieti	
13268161	2017-02-02T21:42:49.430000	42.5783	13.2552	12.6	SURVEY-INGV		ML	2.4	--	L'Aquila	
13268511	2017-02-02T22:04:25.250000	42.7948	13.1952	11.0	SURVEY-INGV		ML	2.0	--	Perugia	
13270371	2017-02-02T23:44:12.450000	41.2193	14.8628	11.0	SURVEY-INGV		ML	2.2	--	Benevento	
13271781	2017-02-03T01:19:31.440000	42.4782	13.2857	13.8	SURVEY-INGV		ML	2.0	--	L'Aquila	
13272071	2017-02-03T01:32:08.060000	42.7917	13.316	16.8	SURVEY-INGV		ML	2.4	--	Ascoli Piceno	
13272231	2017-02-03T01:37:06.860000	46.0547	13.0658	9.7	SURVEY-INGV		ML	2.3	--	Udine	
13272491	2017-02-03T01:49:15.420000	42.5542	13.2298	10.0	SURVEY-INGV		ML	2.2	--	L'Aquila	
13273251	2017-02-03T02:26:30.030000	38.6832	11.7805	10.0	SURVEY-INGV		ML	3.2	--	Tirreno Meridionale (MARE)	
13274401	2017-02-03T03:32:59.300000	42.9932	13.0213	8.9	SURVEY-INGV		ML	3.0	--	Macerata	
13274561	2017-02-03T03:47:55.970000	42.9902	13.0218	6.4	SURVEY-INGV		Mw	4.0	--	Macerata	
13275241	2017-02-03T03:53:28.620000	42.9912	13.0352	6.6	SURVEY-INGV		ML	2.8	--	Macerata	
13275291	2017-02-03T03:54:25.670000	42.9997	13.0323	5.8	SURVEY-INGV		ML	3.0	--	Macerata	
13275591	2017-02-03T03:59:34.880000	42.9902	13.0332	5.4	SURVEY-INGV		ML	2.2	--	Macerata	
13275711	2017-02-03T04:01:16.260000	42.9858	13.0302	6.3	SURVEY-INGV		ML	2.2	--	Macerata	
13276121	2017-02-03T04:10:05.430000	42.9895	13.025	5.7	SURVEY-INGV		Mw	4.2	--	Macerata	
13276311	2017-02-03T04:12:21.700000	40.1008	15.9873	10.6	SURVEY-INGV		ML	2.4	--	Potenza	
13276371	2017-02-03T04:12:48.900000	43.0083	13.0222	4.0	SURVEY-INGV		ML	3.1	--	Macerata	
13276411	2017-02-03T04:14:00.300000	42.9903	13.0328	6.1	SURVEY-INGV		ML	2.5	--	Macerata	
13276731	2017-02-03T04:19:28.070000	42.9918	13.035	7.0	SURVEY-INGV		ML	3.3	--	Macerata	
13276821	2017-02-03T04:21:12.830000	42.9933	13.0312	6.1	SURVEY-INGV		ML	2.5	--	Macerata	
13277001	2017-02-03T04:24:08.820000	43.0088	13.0223	8.5	SURVEY-INGV		ML	2.1	--	Macerata	
13277041	2017-02-03T04:26:22.430000	42.9927	13.0293	5.6	SURVEY-INGV		ML	2.0	--	Macerata	
13279061	2017-02-03T05:11:22.560000	42.9842	13.0423	5.5	SURVEY-INGV		ML	2.8	--	Macerata	
13279931	2017-02-03T05:34:45.860000	42.9862	13.0333	7.6	SURVEY-INGV		ML	2.4	--	Macerata	
13280161	2017-02-03T05:40:34.310000	43.0025	13.0362	7.5	SURVEY-INGV		Mw	3.8	--	Macerata	
13280531	2017-02-03T05:45:15.290000	42.9903	13.0377	7.6	SURVEY-INGV		ML	2.3	--	Macerata	
13280601	2017-02-03T05:46:32.830000	42.9977	13.0222	8.4	SURVEY-INGV		ML	2.2	--	Macerata	
13280671	2017-02-03T05:48:03.100000	43.0013	13.0372	7.4	SURVEY-INGV		ML	2.1	--	Macerata	
13280811	2017-02-03T05:49:16.800000	42.986	13.047	5.1	SURVEY-INGV		ML	2.8	--	Macerata	
1328121	2017-02-03T05:54.16390000	43.0047	13.0367	7.5	SURVEY-INGV		Mt	2.1	--	Macerata	

event IDs

event IDs



Seismology

HTML

The screenshot shows a web browser window with the URL cnt.rm.ingv.it/en/event/7073641 highlighted by a red circle. Below the browser is the Seismology HTML page. At the top, there's a navigation bar with links for EARTHQUAKE LIST, INSTRUMENTS, SCIENTIFIC PRODUCTS, SITE GUIDE, CONTACT, and language options (English, Italian). The main content area displays information about an earthquake with a magnitude of 6.0 on August 24, 2016, at 01:36:32 UTC in the Rieti region. Below this, a table titled 'Locations history' is shown, which is also circled in red.

Type	Date and Time (UTC)	Latitude	Longitude	Magnitude	Depth (km)	Published time (UTC)	Author	Location ID
Rev 700	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	27329391
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

Earthquake with magnitude of 6.0 on date 24-08-2016 and time 01:36:32 (UTC) in region Rieti

Event data

Seismicity and Hazard

Impact

Locations and Magnitudes

Focal mechanism

Download

Locations history

Type	Date and Time (UTC)	Latitude	Longitude	Magnitude	Depth (km)	Published time (UTC)	Author	Location ID
Rev 700	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	27329391
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](https://github.com/rosafilgueira/Data_Wrangling/blob/master/Seismology_Example/download_cnt.py)

Volcanology

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

The screenshot shows a web browser window displaying the contents of the specified FTP directory. The address bar at the top shows the URL: `ftp://ftp.bom.gov.au/anon/gen/vaac/`. The main content area is titled "Índice de `ftp://ftp.bom.gov.au/anon/gen/vaac/`". Below the title, there is a link labeled "Subir al directorio superior." (Up one level). The content is presented as a table with three columns: "Nombre" (Name), "Tamaño" (Size), and "Última modificación" (Last modified). The "Nombre" column lists the years from 1998 to 2015, each preceded by a blue folder icon. The "Tamaño" and "Última modificación" columns show the size as 00:00:00 and the last modification date as either 07/03/2008 or 27/11/2008 for the earlier years, and dates ranging from 31/12/2009 to 31/12/2015 for the later years.

Nombre	Tamaño	Última modificación
1998	07/03/2008	00:00:00
1999	07/03/2008	00:00:00
2000	07/03/2008	00:00:00
2001	07/03/2008	00:00:00
2002	07/03/2008	00:00:00
2003	07/03/2008	00:00:00
2004	07/03/2008	00:00:00
2005	07/03/2008	00:00:00
2006	07/03/2008	00:00:00
2007	07/03/2008	00:00:00
2008	27/11/2008	00:00:00
2009	31/12/2009	00:00:00
2010	31/12/2010	00:00:00
2011	29/12/2011	00:00:00
2012	24/12/2012	00:00:00
2013	31/12/2013	00:00:00
2014	28/12/2014	00:00:00
2015	31/12/2015	00:00:00

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>

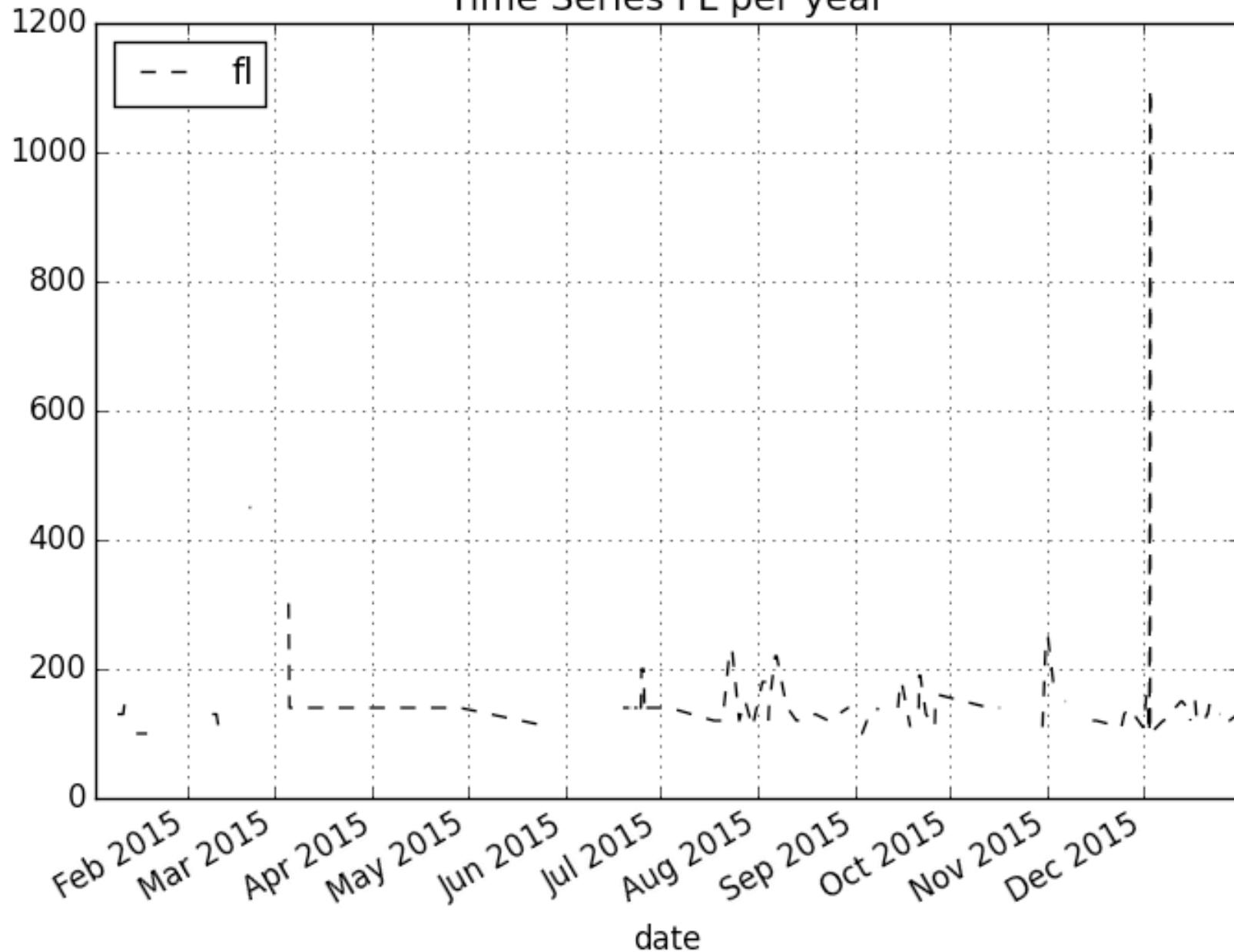
VA ADVISORY
DTG: 20160106/0330Z
VAAC: DARWIN
VOLCANO: SINABUNG 261080
PSN: N0310 E09824
AREA: INDONESIA
SUMMIT ELEV: 2460M
ADVISORY NR: 2016/1
INFO SOURCE: GROUND REPORTS, HIMAWARI-8
AVIATION COLOUR CODE: RED
ERUPTION DETAILS: MINOR VA/STEAM EMISSION TO FL140
EXTENDING TO THE WEST
OBS VA DTG: 06/0330Z
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 WITHIN 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

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



Hydrology

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

The screenshot shows the India-WRIS WebGIS interface. The top navigation bar includes links for About WRIS, Accessibility, Tools, Metadata, WRIS Wiki, and Help. Below the header is a banner featuring the India-WRIS logo and a satellite image. The main menu path is HOME > WRIS Explorer > GeoVisualization | SubInfoSystem > Ground Water Level | Temporal Analyst. A toolbar below the menu contains icons for Get Details, Download Data, and Water Level Maps. On the left, a sidebar titled "India-WRIS Data Set" lists options like DEM, Watershed Atlas, Administrative, and GW Level, with GW Level selected. A "Graphical view of Water le..." section contains dropdown menus for selecting areas. In the center, a modal window titled "Download Ground Water Level Data" allows users to select parameters: State (Bihar), District (Pashchim Champaran), Block (Bagha), Station Name (Belwa), and a date range from and to. A "Download" button is at the bottom of the modal. At the bottom of the page, there's a map of India with a scale bar showing 1000 km and 500 mi, and a location bar that says "Location: Move mouse on map".

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 states 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](https://github.com/rosafilgueira/Data_Wrangling/tree/master/SocialComputing_Example)