A Dataset Is Worth >1000 Words

Whole Hog Series, Beyond Text Session Lightning Talk: January 14, 2015

Eleni Castro, Research Coordinator
Harvard Institute For Quantitative Social Science (IQSS)
Harvard Dataverse - dataverse.harvard.edu
Contact: ecastro@g.harvard.edu

Diane Sredl, Data Reference Librarian
Harvard Library, Lamont Library
Contact: govdocs@fas.harvard.edu
What is research data?
What stories can data tell us?
Help Inform Policy

New Bike Crash Map Offers Comprehensive Look at Where and How Accidents Happen

The 'cutting edge' project breaks down the time, day, and even weather conditions of each incident.

By Steve Annear | Boston Daily | June 18, 2014, 10:30 a.m.


Source: http://worldmap.harvard.edu/maps/boston-bikes
Harvard University economists Carmen Reinhart and Kenneth Rogoff have acknowledged making a spreadsheet calculation mistake in a 2010 research paper, “Growth in a Time of Debt” (PDF), which has been widely cited to justify budget-cutting. But the authors stand by their conclusion that higher government debt is associated with slower economic growth. Here’s what you need to know:

Data was not publicly available!

Understand our past

For more Digital Humanities projects see:
http://www.darthcrimson.org/
And the present

A LOOK AT 14,939,154 #PARIS #BATACLAN #PARISATTACKS #PORTEOUVERTE TWEETS

Submitted by Nick Ruest on Sat, 10/24/2015 - 14:49

IMAGES
We are able to create a list of images tweeted in our dataset by using image_urls.py.

$ python ~/twarc/utils/image_urls.py paris-valid-deduplicated.json > paris-tweets-images.txt
$ cut paris-tweets-images.txt | sort | uniq -c | sort -n > paris-tweets-images-uniq.txt $ cut paris-tweets-images-uniq.txt | wc -l
$ tail paris-tweets-images-uniq.txt

From the above, we can see that there were 6,879,441 total images tweets, representing 46.06% of total tweets, and 660,470 unique images. The top 10 images tweeted were as follows:

1. 49,041 Occurrences

APPEL À TÉMOINS

ABDESLAM SALAH
né le 16 septembre 1989
à Bruxelles (Belgique)

Individu faisant l’objet d’un mandat de recherche.

SIGNALEMENT :
1 m 75, yeux marron

CONTACT :
Si vous disposez d’informations permettant de le localiser, contactez immédiatement
le 197 Alerte attentat.

Individu dangereux, surtout n’intervenez pas vous-même.

Source: http://ruebot.net/post/look-14939154-paris-bataclan-parisattacks-porteouverte-tweets

Source: http://www.theguardian.com/uk/2011/aug/24/twitter-study-post-riot-plans#img-1
Understand outcomes of deadly viruses

Clinical Illness and Outcomes in Patients with Ebola in Sierra Leone

Schaffner, John; Shaffer, Jeffrey; Goba, Augustine; Gbakie, Michael; Giro, Stephen; Colubri, Andrea; Seafor, Rachel; Kanneh, Lansana; Moigboi, Alex; Momoh, Mambu; Fullah, Mohammad; Moses, Lina; Brown, Bethany; Andersen, Kristian; Winnicki, Sarah; Schaffner, Stephen; Park, Daniel; Yozwiak, Nathan; Jiang, Pan-Pan; Kargbo, David; Jalloh, Simbirie; Fonnie, Mbalu; Sinnah, Vandi; French, Issa; Kovoma, Alice; Kamara, Fatima; Tucker, Veronica; Koniwa, Edwin; Sellu, Josephine; Mustapha, Ibrahim; Foday, Momoh; Yillah, Mohamed; Kanneh, Franklyn; Saffa, Sidiki; Massally, James; Boiser, Matt; Branco, Luis; Vandi, Mohamed; Grant, Donald; Happi, Christian; Gevao, Sahr; Fletcher, Thomas; Fowler, Robert; Bausch, Daniel; Sabeti, Pardis; Khan, Humar; Garry, Robert, 2015, "Clinical Illness and Outcomes in Patients with Ebola in Sierra Leone", http://dx.doi.org/10.7910/DVN/29296, Harvard Dataverse, V1

If you use these data, please add this citation to your scholarly resources. Learn about Data Citation Standards.

Description

This data comprises a total of 213 cases evaluated for Ebola virus infection at the Kenema Government Hospital in Sierra Leone between May 25 and June 18, 2014. Outcome data was available for 87 of 106 EBOV positive cases. Metabolic panels were performed on 98 Ebola virus disease and non-Ebola virus disease illness patients with adequate samples volumes. Ebola virus load was determined in 83 cases with adequate samples volumes by quantitative polymerase chain reaction (qPCR) at Harvard University. Sign and symptom data was obtained on 44 patients with a clinical chart that were admitted to Kenema Hospital. The metabolic panels were obtained from serum samples analyzed with a Piccolo Blood Chemistry Analyzer and Comprehensive Metabolic Reagent Discs (Abaxis).

Source: Harvard Dataverse http://dx.doi.org/10.7910/DVN/29296
The mysteries held in the stars above

Measure and Map Our Galaxy

We need your help looking through tens of thousands of images from the Spitzer Space Telescope. By telling us what you see in this infrared data, we can better understand how stars form.

See also: theastrodata.org for Harvard-Smithsonian datasets

Source: http://www.milkywayproject.org/
And most important: Modern Romance

Where can data be found?

Start at → re3data.org

Only a sample of what is out there!

HUBBLESITE
ADS
zenodo
PDB
biosharing
DATA.GOV
American Mineralogist Crystal Structure Database
What tools can I use to analyze data?

Some examples of tools:

- Refine
- STATA
- R
- R Studio
- SPSS
- Excel
- GitHub
- Python
- Leaflet
- WorldMap
- Gephi
- plotly

Best Practices to Remember:

- use open or common file formats
- describe your data for discovery and reuse (metadata)
- have a data management plan ([DMPTool at Harvard](https://dataverse.harvard.edu/))
- archive & share your data for reuse (e.g., [Harvard Dataverse](https://dataverse.harvard.edu/))
Further Reading & Resources at Harvard

- Finding Data Resources at Harvard Library: http://guides.library.harvard.edu; or contact: govdocs@fas.harvard.edu

- Harvard Open Data Assistance Program Weekly Office Hours (Weds 11am-1pm) http://projects.iq.harvard.edu/odap/open-hours

- Data Science Services: Free research and training services for Harvard affiliates http://rtc.iq.harvard.edu/

Books
- Dataclysm: Who We Are When We Think No One Is Looking by Christian Rudder http://id.lib.harvard.edu/aleph/014155996/catalog
- Big data, little data, no data : scholarship in the networked world by Christine L. Borgman http://id.lib.harvard.edu/aleph/014297454/catalog