Using Web Data as a source of Open Source Intelligence

CorkSec, 2017-10-10

Johannes Ahlmann

The internet contains many open and openly-available datasets that can be used to gather intelligence on people and organizations.

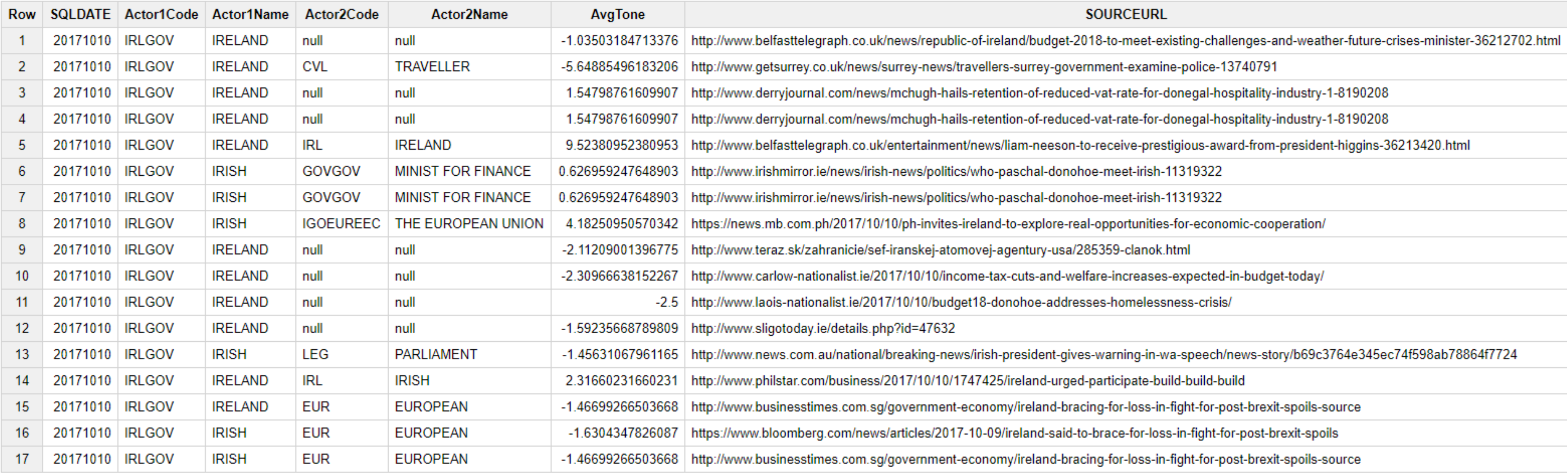
This talk will outline possible approaches to gathering such intelligence.

GDELT

Monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages and identifies the people, locations, organizations, counts, themes, sources, emotions, counts, quotes, images and events.

```
SELECT SQLDATE, Actor1Code, Actor1Name, Actor2Code, Actor2Name, AvgTone, SOURCEURL FROM [gdelt-bq:gdeltv2.events] WHERE MonthYear == 20
1710 AND Actor1Code == 'IRLGOV' LIMIT 1000;
```

query



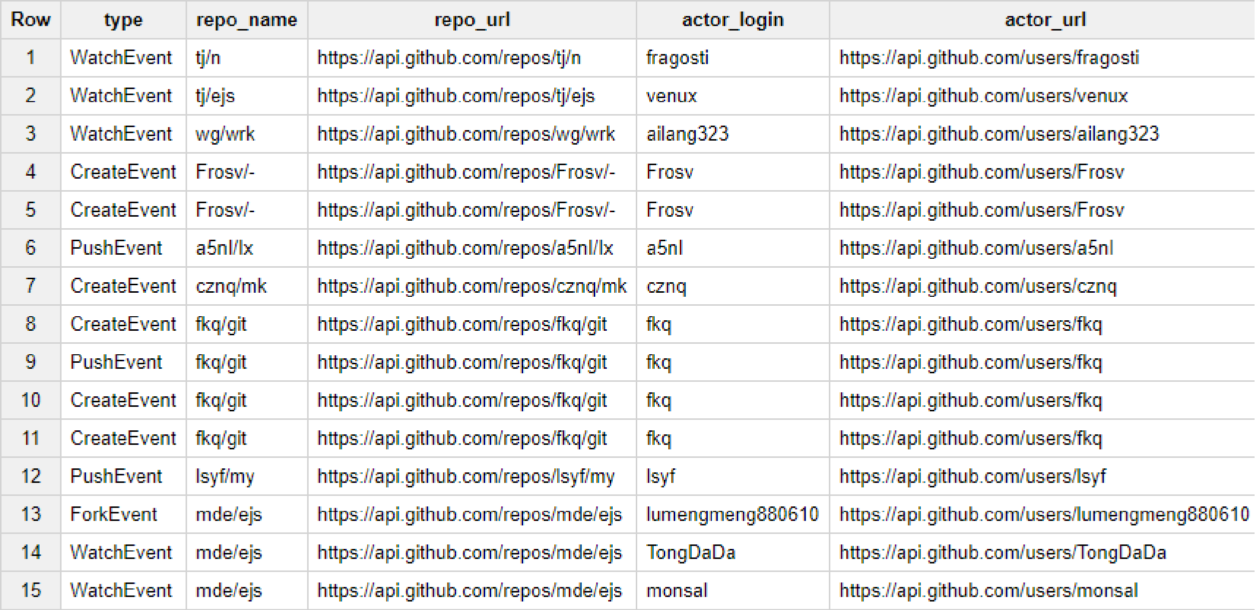
Row	SQLDATE	Actor1Code	Actor1Name	Actor2Code	Actor2Name	AvgTone	SOURCEURL
1	20171010	IRLGOV	IRELAND	null	null	-1.03503184713376	http://www.belfasttelegraph.co.uk/news/republic-of-ireland/budget-2018-to-meet-existing-challenges-and-weather-future-crises-minister-36212702.html
2	20171010	IRLGOV	IRELAND	CVL	TRAVELLER	-5.64885496183206	http://www.getsurrey.co.uk/news/surrey-news/travellers-surrey-government-examine-police-13740791
3	20171010	IRLGOV	IRELAND	null	null	1.54798761609907	http://www.derryjournal.com/news/mchugh-hails-retention-of-reduced-vat-rate-for-donegal-hospitality-industry-1-8190208
4	20171010	IRLGOV	IRELAND	null	null	1.54798761609907	http://www.derryjournal.com/news/mchugh-hails-retention-of-reduced-vat-rate-for-donegal-hospitality-industry-1-8190208
5	20171010	IRLGOV	IRELAND	IRL	IRELAND	9.52380952380953	http://www.belfasttelegraph.co.uk/entertainment/news/liam-neeson-to-receive-prestigious-award-from-president-higgins-36213420.html
6	20171010	IRLGOV	IRISH	GOVGOV	MINIST FOR FINANCE	0.626959247648903	http://www.irishmirror.ie/news/irish-news/politics/who-paschal-donohoe-meet-irish-11319322
7	20171010	IRLGOV	IRISH	GOVGOV	MINIST FOR FINANCE	0.626959247648903	http://www.irishmirror.ie/news/irish-news/politics/who-paschal-donohoe-meet-irish-11319322
8	20171010	IRLGOV	IRISH	IGOEUREEC	THE EUROPEAN UNION	4.18250950570342	https://news.mb.com.ph/2017/10/10/ph-invites-ireland-to-explore-real-opportunities-for-economic-cooperation/
9	20171010	IRLGOV	IRELAND	null	null	-2.11209001396775	http://www.teraz.sk/zahranicie/sef-iranskej-atomovej-agentury-usa/285359-clanok.html
10	20171010	IRLGOV	IRELAND	null	null	-2.30966638152267	http://www.carlow-nationalist.ie/2017/10/10/income-tax-cuts-and-welfare-increases-expected-in-budget-today/
11	20171010	IRLGOV	IRELAND	null	null	-2.5	http://www.laois-nationalist.ie/2017/10/10/budget18-donohoe-addresses-homelessness-crisis/
12	20171010	IRLGOV	IRELAND	null	null	-1.59235668789809	http://www.sligotoday.ie/details.php?id=47632
13	20171010	IRLGOV	IRISH	LEG	PARLIAMENT	-1.45631067961165	http://www.news.com.au/national/breaking-news/irish-president-gives-warning-in-wa-speech/news-story/b69c3764e345ec74f598ab78864f7724
14	20171010	IRLGOV	IRELAND	IRL	IRISH	2.31660231660231	http://www.philstar.com/business/2017/10/10/1747425/ireland-urged-participate-build-build-build
15	20171010	IRLGOV	IRELAND	EUR	EUROPEAN	-1.46699266503668	http://www.businesstimes.com.sg/government-economy/ireland-bracing-for-loss-in-fight-for-post-brexit-spoils-source
16	20171010	IRLGOV	IRISH	EUR	EUROPEAN	-1.6304347826087	https://www.bloomberg.com/news/articles/2017-10-09/ireland-said-to-brace-for-loss-in-fight-for-post-brexit-spoils
17	20171010	IRLGOV	IRISH	EUR	EUROPEAN	-1.46699266503668	http://www.businesstimes.com.sg/government-economy/ireland-bracing-for-loss-in-fight-for-post-brexit-spoils-source

Google BigQuery

Bigquery hosts a variety of public datasets that can be analyzed using familiar SQL. Users can query this data directly in the Bigquery web UI or programmatically using the Bigquery REST API. These data sets are freely hosted and accessible to everyone. You can query this data up to 1TB per month for free.

- github archive preview

```
SELECT type, repo.name, repo.url, actor.login, actor.url FROM [githubarchive:day.20171010] LIMIT 1000;
```



Row	type	repo_name	repo_url	actor_login	actor_url
1	WatchEvent	tj/n	https://api.github.com/repos/tj/n	fragosti	https://api.github.com/users/fragosti
2	WatchEvent	tj/ejs	https://api.github.com/repos/tj/ejs	venux	https://api.github.com/users/venux
3	WatchEvent	wg/wrk	https://api.github.com/repos/wg/wrk	ailang323	https://api.github.com/users/ailang323
4	CreateEvent	Frosv/-	https://api.github.com/repos/Frosv/-	Frosv	https://api.github.com/users/Frosv
5	CreateEvent	Frosv/-	https://api.github.com/repos/Frosv/-	Frosv	https://api.github.com/users/Frosv
6	PushEvent	a5nl/lx	https://api.github.com/repos/a5nl/lx	a5nl	https://api.github.com/users/a5nl
7	CreateEvent	cznq/mk	https://api.github.com/repos/cznq/mk	cznq	https://api.github.com/users/cznq
8	CreateEvent	fkq/git	https://api.github.com/repos/fkq/git	fkq	https://api.github.com/users/fkq
9	PushEvent	fkq/git	https://api.github.com/repos/fkq/git	fkq	https://api.github.com/users/fkq
10	CreateEvent	fkq/git	https://api.github.com/repos/fkq/git	fkq	https://api.github.com/users/fkq
11	CreateEvent	fkq/git	https://api.github.com/repos/fkq/git	fkq	https://api.github.com/users/fkq
12	PushEvent	lsyf/my	https://api.github.com/repos/lsyf/my	lsyf	https://api.github.com/users/lsyf
13	ForkEvent	mde/ejs	https://api.github.com/repos/mde/ejs	lumengmeng880610	https://api.github.com/users/lumengmeng880610
14	WatchEvent	mde/ejs	https://api.github.com/repos/mde/ejs	TongDaDa	https://api.github.com/users/TongDaDa
15	WatchEvent	mde/ejs	https://api.github.com/repos/mde/ejs	monsal	https://api.github.com/users/monsal

  - find company employees; what is company up to; what kind of people is it hiring
  - github project health
  - find similar github projects
- stackoverflow
- hacker news
- reddit, reddit_posts
- etc.

Wikipedia Infoboxes and category information is a huge treasure trove of information.

Whether information about entities like companies or universities, or using redirects and multi-lingual entries to compile lists of aliases.

- yago demo
  - YAGO is a huge semantic knowledge base, derived from Wikipedia WordNet and GeoNames. Currently, YAGO has knowledge of more than 10 million entities (like persons, organizations, cities, etc.) and contains more than 120 million facts about these entities.
  - YAGO is an ontology anachored in time and space
- dbpedia - bubble navigator, spotlight
  - DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web.
  - as tables

Toggle line numbers

```
"Karlsruhe_Institute_of_Technology": {
    "foundingDate": "2009-10-01",
    "label": "Karlsruhe Institute of Technology",
    "president_label": "Holger_Hanselka",
    "type_label": "Public university",
    "country": "http://dbpedia.org/resource/Germany",
    "numberOfDoctoralStudents": "831",
    "city": "http://dbpedia.org/resource/Karlsruhe",
    "country_label": "Germany",
    "facultySize": "7177",
    "state_label": "Baden-Württemberg",
    "numberOfStudents": "24528",
    "point": "49.00944444444445 8.411666666666667",
    "city_label": "Karlsruhe",
}
```

Twitter allows to stream any tweets, or filter for particular keywords in realtime.

The volume/throuhput is restricted to I believe 1/6th of all available tweets, but for all/most practical purposes a filtered stream represents the totality of twitter messages for a given filter in realtime.

Twitter allows to track 400 keywords, follow 5,000 userids and define 25 location boxes.

```
https://stream.twitter.com/1.1/statuses/filter.json?track=#jobs,#hiring,#job,#career
```

```
{
  "entities": {
    "urls": [
      {
        "url": "https://t.co/gI3p5KT1Pu",
        "expanded_url": "http://snapjobsearch.com/jobs/view/4014840/",
        "display_url": "snapjobsearch.com/jobs/view/4014…",
      }
    ],
    "user_mentions": [],
    "hashtags": [
      {
        "text": "Columbus",
      },
      {
        "text": "OH",
      },
      {
        "text": "ComputerITServices",
      },
      {
        "text": "job",
      },
      {
        "text": "hiring",
      }
    ],
  },
  "text": "Medical Practice Rep Mount Carmel Medical Group East, #Columbus, #OH, #ComputerITServices https://t.co/gI3p5KT1Pu #job #hiring",
  "source": "<a href=\"http://snapjobsearch.com\" rel=\"nofollow\">SJS_US</a>",
  "lang": "en",
  "created_at": "Mon Oct 09 22:00:12 +0000 2017",
}
```

Common
Crawl

Common Crawl is a nonprofit 501(c)(3) organization that crawls the web and freely provides its archives and datasets to the public.

The latest crawl as of September 2017 now contains 3.01 billion web pages and over 250 TiB of uncompressed content.

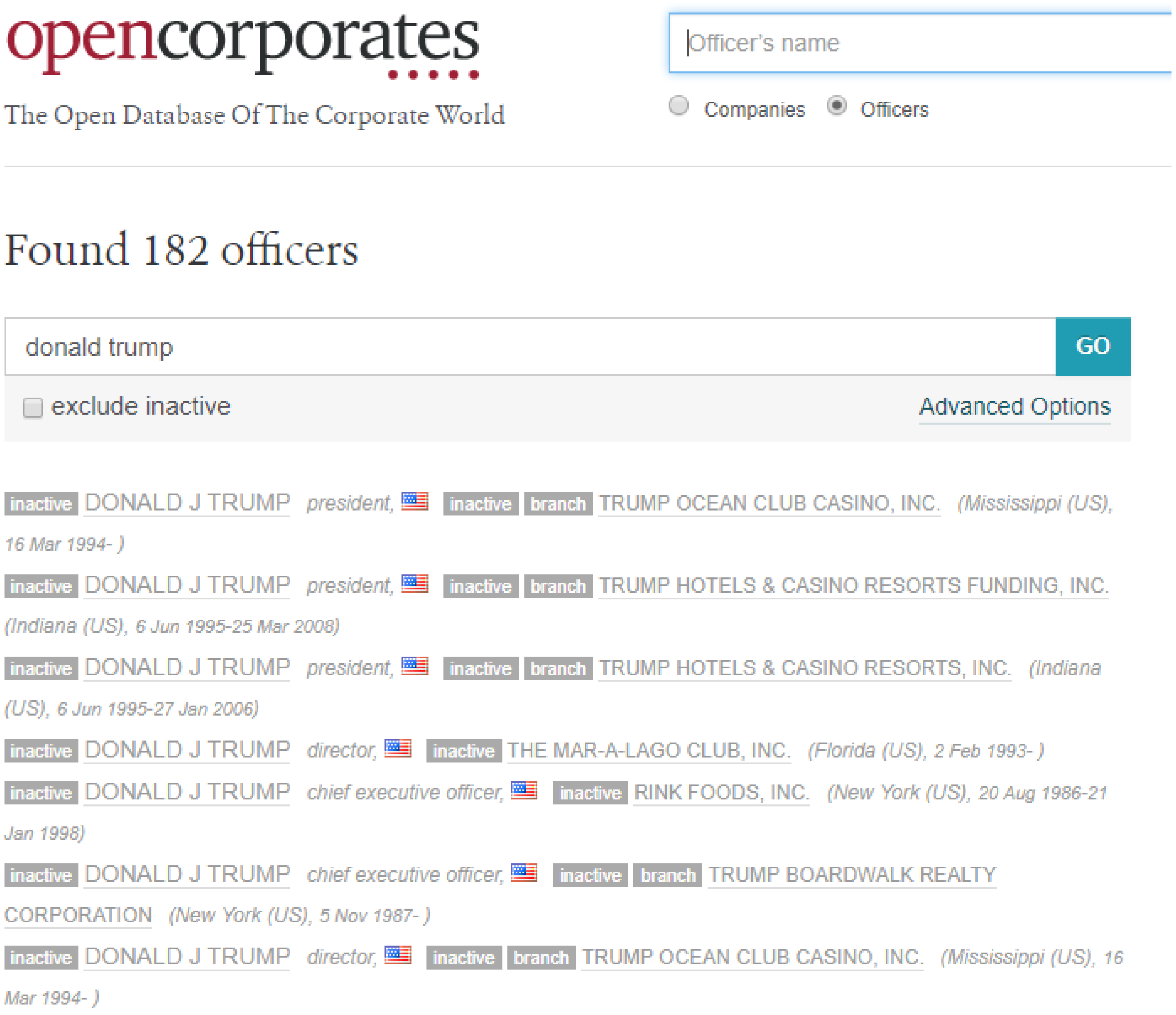
The data is available on Amazon S3 and can be processed relatively cheaply using Amazon EC2.

- commoncrawl
- commonsearch datasets

```
○  facebook.com 244660.58
○  twitter.com 164232.66
○  blogger.com 77521.93
○  youtube.com 62967.95
○  plus.google.com 61344.234
○  instagram.com 39883.676
○  linkedin.com 34856.848
○  wordpress.org 33809.844
○  google.com 27425.883
○  pinterest.com 25640.172
○  ... [112M hostnames]
```

Open Corporates

- open corporates is the largest open database of companies and company data in the world, with in excess of 100 million companies in a similarly large number of jurisdictions.
- 136M companies, 178M officers



Found 182 officers

inactive DONALD J TRUMP president, inactive branch TRUMP OCEAN CLUB CASINO, INC. (Mississippi (US), 16 Mar 1994-)

inactive DONALD J TRUMP president, inactive branch TRUMP HOTELS & CASINO RESORTS FUNDING, INC. (Indiana (US), 6 Jun 1995-25 Mar 2008)

inactive DONALD J TRUMP president, inactive branch TRUMP HOTELS & CASINO RESORTS, INC. (Indiana (US), 6 Jun 1995-27 Jan 2006)

inactive DONALD J TRUMP director, inactive THE MAR-A-LAGO CLUB, INC. (Florida (US), 2 Feb 1993-)

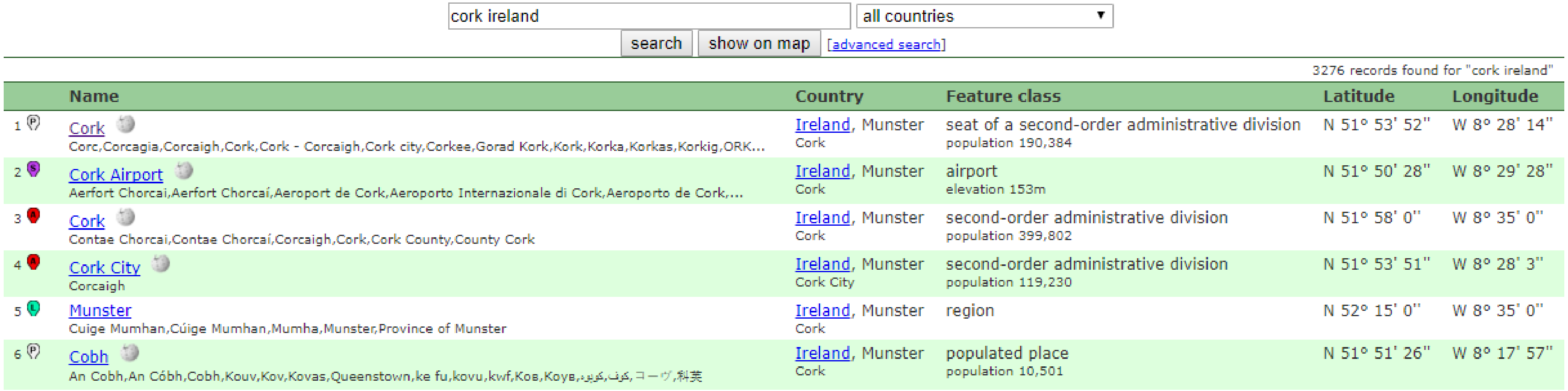
inactive DONALD J TRUMP chief executive officer, inactive RINK FOODS, INC. (New York (US), 20 Aug 1986-21 Jan 1998)

inactive DONALD J TRUMP chief executive officer, inactive branch TRUMP BOARDWALK REALTY CORPORATION (New York (US), 5 Nov 1987-)

inactive DONALD J TRUMP director, inactive branch TRUMP OCEAN CLUB CASINO, INC. (Mississippi (US), 16 Mar 1994-)

Geonames

- The geonames geographical database is available for download free of charge under a creative commons attribution license.
- It contains over 10 million geographical names and consists of over 9 million unique features,
    - whereof 2.8 million populated places and 5.5 million alternate names.



cork ireland | all countries ▼

search | show on map | [advanced search]

3276 records found for "cork ireland"

	Name	Country	Feature class	Latitude	Longitude
1	Cork Corc,Corcagia,Corcaigh,Cork,Cork - Corcaigh,Cork city,Corkee,Gorad Kork,Kork,Korka,Korkas,Korkig,ORK...	Ireland, Munster Cork	seat of a second-order administrative division population 190,384	N 51° 53' 52''	W 8° 28' 14''
2	Cork Airport Aerfort Chorcai,Aerfort Chorcaí,Aeroport de Cork,Aeroporto Internazionale di Cork,Aeroporto de Cork,...	Ireland, Munster Cork	airport elevation 153m	N 51° 50' 28''	W 8° 29' 28''
3	Cork Contae Chorcai,Contae Chorcaí,Corcaigh,Cork,Cork County,County Cork	Ireland, Munster Cork	second-order administrative division population 399,802	N 51° 58' 0''	W 8° 35' 0''
4	Cork City Corcaigh	Ireland, Munster Cork City	second-order administrative division population 119,230	N 51° 53' 51''	W 8° 28' 3''
5	Munster Cuige Mumhan,Cúige Mumhan,Mumha,Munster,Province of Munster	Ireland, Munster Cork	region	N 52° 15' 0''	W 8° 35' 0''
6	Cobh An Cobh,An Cóbh,Cobh,Kouv,Kov,Kovas,Queenstown,ke fu,kovu,kwf,Ков,Коув,كوف,كوبه,コーヴ,科芙	Ireland, Munster Cork	populated place population 10,501	N 51° 51' 26''	W 8° 17' 57''

Exploits

- Much personal data available, but not legally accessible
- https://haveibeenpwned.com/
  - exploit.In (711M)
  - antipublic (593M)
  - River City (457M)
  - etc.

Web Crawling

- Example:
    - Given a list of existing clients, crawl their websites, extract fields of interest, and identify what they are talking about.

Toggle line numbers

```
{
  "domain": "1to1media.com",
  "num_pages": 24,
  "social": [
    "https://www.linkedin.com/company/2556633",
    "https://twitter.com/Wendys?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor",
    "http://www.slideshare.net/1to1Media",
    "http://twitter.com/1to1media",
    "@1to1media",
    "http://twitter.com/judithaquino",
    "http://www.facebook.com/1to1media",
    "http://www.youtube.com/user/1to1Videos",
    "http://www.linkedin.com/in/judithaquino",
    "https://www.pinterest.com/1to1media/"
  ],
  "most_common": [
    "customer experience",
    "customer relationship",
    "customer loyalty",
    "customer satisfaction",
    "customer journey"
  ]
}
```

- website technologies webdata.org
  - Toggle line numbers

```
{
  "Programming Languages": [
    "Python"
  ],
  "JavaScript Frameworks": [
    "jQuery"
  ],
  "Web Servers": [
    "Apache"
  ],
  "Wikis": [
    "MoinMoin"
  ],
  "Font Scripts": [
    "Font Awesome"
  ],
  "Web Frameworks": [
    "Twitter Bootstrap"
  ]
}
```

- metadata
  - Toggle line numbers

```
{
  "datePublished": "2017-10-10T15:43:02+0100",
  "@context": "http://schema.org",
  "associatedMedia": {},
  "liveBlogUpdate": [
    [
      {
        "datePublished": "2017-10-10T14:02:51+0100",
        "@type": "BlogPosting",
        "author": {
          "@id": "https://www.theguardian.com#publisher"
        },
        "url": "https://www.theguardian.com/business/live/2017/oct/10/markets-uk-trade-manufacturing-growth-productivity-imf-global-economy-business-live?page=with:block-59dcc4ebe4b076f91939d34a#block-59dcc4ebe4b076f91939d34a",
        "articleBody": "And here is Larry Elliott’s analysis of the IMF report:",
        "publisher": {
          "@id": "https://www.theguardian.com#publisher"
        },
        "headline": "UK suffers productivity blow, as goods trade deficit hits record high - business live"
      }
    ]
  ],
  "description": "Britain imported more from the rest of the world than ever before in August, but managed to export more to Europe",
  "publisher": {
    "logo": {},
    "name": "The Guardian",
    "@context": "http://schema.org",
    "@type": "Organization",
    "@id": "https://www.theguardian.com#publisher"
  },
  "dateModified": "2017-10-10T15:43:02+0100",
  "coverageStartTime": "2017-10-10T15:43:02+0100",
  "coverageEndTime": "2017-10-10T15:43:02+0100",
  "url": "https://www.theguardian.com/business/live/2017/oct/10/markets-uk-trade-manufacturing-growth-productivity-imf-global-economy-business-live",
  "headline": "UK suffers productivity blow, as goods trade deficit hits record high - business live",
  "@type": "LiveBlogPosting"
}
```

- record extraction pydepta of site
- reddit cryptocurrency, sentiment analysis