Looking for Big Data (BD) in the "Wild"

Tidewater Big Data Enthusiasts
Chuck Cartledge
Developer

April 23, 2020 at 12:18 Noon

Contents

List of Tables **ii**

List of Figures **ii**

1 Introduction **1**

2 Ways to get data **2**
2.1 Create your own . . . 2
2.2 Download a file . . . 5
2.3 Download using an Application Program Interface (API) . . . 8

3 Selected Big Data Sources **25**
3.1 Aggregator . . . 26
3.2 Aviation . . . 41
3.3 Developers . . . 51
3.4 Education . . . 56
3.5 General . . . 58
3.6 Geographic information . . . 69
3.7 Geographic location . . . 73
3.8 Government . . . 74
3.9 Social . . . 105
3.10 Weather . . . 110
3.11 Zip code . . . 123

4 System performance **124**

5 References **125**

List of Tables

1 Contents of the file Medicare_Provider_Util_Payment_PUF_CY2013.zip. . . . 8

List of Figures

1 Nicholas Felton's annual report. . . . 3
2 Aaron Parecki records his GPS position every 5 seconds. . . . 4
3 Centers for Medicare and Medicaid Services download site. . . . 6
4 Medicare data file viewed normally. . . . 7
5 Medicare data file shown with tabs. . . . 7
6 Project Gutenberg home page. . . . 9
7 Project Gutenberg's copy of Romeo and Juliet. . . . 10
8 Flickr's API page. . . . 12
9 Flickr's flickr.photos.search API link. . . . 13
10 Flickr's flickr.photos.getWithGeoData API page. . . . 14
11 Flickr's API key page. . . . 15
12 Flickr's API key type selection page. . . . 16
13 Flickr's application description page. . . . 17
14 Flickr's API exploration link. . . . 18
15 Flickr's API exploration page. . . . 20
16 Flickr's API exploration page with bounding box data. . . . 21
17 Flickr's API exploration calling method. . . . 22
18 Flickr's API exploration with data and XML format. . . . 23
19 Flickr's API exploration with data and JSON format. . . . 24
20 70+ websites to get large data repositories for free home page. . . . 26
21 Anywhere that I could download a sample database? : learnprogramming home page. . . . 27
22 Data Portals home page. . . . 28
23 Datasets Archive home page. . . . 29
24 Datasets for Data Mining and Data Science home page. . . . 30
25 enigma home page. . . . 31
26 Finding Data on the Internet home page. . . . 32
27 Linked Data - Connect Distributed Data across the Web home page. . . . 33
28 Oxford Internet Institute University of Oxford home page. . . . 34
29 Real Time Data Resources home page. . . . 35
30 Thirty Thousand Feet home page. . . . 36

31 Top 10 Weather APIs home page. 37
32 University Corporation for Atmospheric Research home page. 38
33 Where can I find large datasets open to the public? - Quora home page. . . . 39
34 Wikipedia:Database download home page. 40
35 Flight Aware home page. 41
36 Flight Stats home page. 42
37 Flight View home page. 43
38 OAG (Official Airline Guide) home page. 44
39 QPX Express API home page. 45
40 Travel Boutique home page. 47
41 Travel Fusion home page. 49
42 Wego Flights home page. 50
43 GeoCommons home page. 51
44 Global Data. Local Context home page. 52
45 OpenStreetMap home page. 53
46 Quandl Financial and Economic Data home page. 54
47 TIGER Products home page. 55
48 Delve Datasets home page. 56
49 Statistics Online Computational Resource (SOCR) home page. 57
50 Bing home page. 58
51 Blog - AriLamstein.com home page. 59
52 Dogpile home page. 60
53 DuckDuckGo home page. 62
54 Google home page. 63
55 Our World in Data home page. 64
56 Stanford Large Network Dataset Collection home page. 65
57 UCI Machine Learning Repository home page. 66
58 WolframAlpha home page. 67
59 Yahoo home page. 68
60 GeoNames home page. 69
61 Open Source Data and APIs — MaxMind home page. 70
62 Places API - CityGrid V2 - CityGrid Media Documentation home page. . . . 72
63 MaxMind Open Source Data and APIs home page. 73
64 Austin Texas Data Portal home page. 74
65 Bureau of Economic Analysis home page. 75
66 Bureau of Labor Statistics home page. 76
67 Census Bureau Economic Statistics home page. 77
68 City Maps - Robert Wood Johnson Foundation home page. 78
69 Datasets - data.sa.gov.au home page. 79
70 Earthquake Hazards Program home page. 80

71 FAA Web Services, Airport Service home page. 81
72 Federal Reserve Bank of St. Louis, Economic Research home page. 82
73 Gapminder Data home page. 83
74 General Social Survey home page. 85
75 Massachusetts Bay Transportation Authority home page. 86
76 National Aeronautics and Space Administration, Goddard Institute for Space Studies home page. 87
77 National Flight Data Center (NFDC) home page. 88
78 National Longitudinal Surveys — A Program of the U.S. Bureau of Labor Statistics home page. 89
79 NYC Open Data home page. 90
80 Parking in San Francisco, CA home page. 91
81 SNAP Retailer Locator home page. 92
82 Social Security Administration home page. 93
83 The National Archives home page. 94
84 THe National Map home page. 95
85 The National UFO Reporting Center home page. 96
86 The World Bank home page. 97
87 UK Government Open Data Portal home page. 98
88 United States Department of Agriculture, Economic Research Service home page. 99
89 USDA Food Composition Database home page. 100
90 US Department of Transportation home page. 101
91 US Government Open Data Portal home page. 102
92 USGovXML.com: US Government Web Services and XML Data Sources. home page. 103
93 Washington Metropolitan Area Transit Authority API home page. 104
94 American Customer Satisfaction Index home page. 105
95 American Customer Satisfaction Index home page. 106
96 Linkedin API home page. 107
97 Meetup API home page. 108
98 YELP API home page. 109
99 AccuWeather home page. 110
100 Aeris Weather home page. 111
101 Dark Sky Forecast home page. 112
102 National Data Buoy Center home page. 113
103 National Weather Service home page. 114
104 OpenWeatherMap home page. 115
105 Real-time weather forecasts home page. 116
106 WeatherBug home page. 117

107 Weather Source home page. 118
108 Weather Underground home page. 119
109 World Weather Online home page. 120
110 World Weather OnLine home page. 121
111 Zip Codes home page. 123
112 How long it took to create the home page images. 124

1 Introduction

One way to view things in the Big Data (BD) world is as a collaboration by three individuals or groups [2]. These are:

- *Data visionary*: the one with the vision of how different data could be combined,
- *Data scientist*: the one with the technical and analytical skills to unify disparate sources of data, and
- *Data broker*: the one with raw data

A simple example will clarify the things that each of these groups brings to the table.

- *Statement by data visionary*: What is the relationship between: (1) US annual change in inflation, (2) US annual adjusted unemployment rate, (3) political party in US House of Representatives and Congress, and (4) US major wars and conflicts for the 20^{th} and 21^{st} centuries? The goal to be able to make a statement like: When party X is in control and we are/are not involved in a major conflict, then we can expect unemployment to do something and inflation to do something.
- *Data broker*:
  1. Inflation data is available from the US Bureau of Labor Statistics[1]
  2. Unemployment level data is available from the US Bureau of Labor Statistics[2]
  3. A "good enough" listing of how many members of each part for each house of congress[3]
  4. A "good enough" listing of US major wars and conflicts[4]
- *Data scientist*: how to display all the data on a single X-Y plot, where time is on the X axis? The intent being to be able to look up from the X axis and see if the US was in a conflict, which party was controlling the House and the Senate, and what the change in unemployment and inflation was from the previous year.

In the following sections, we'll talk about becoming your own *data broker*.

[1] `http://data.bls.gov/pdq/SurveyOutputServlet`
[2] `http://data.bls.gov/cgi-bin/surveymost?r4`
[3] `http://en.wikipedia.org/wiki/Party_divisions_of_United_States_Congresses`
[4] `http://en.wikipedia.org/wiki/List_of_wars_involving_the_United_States`

2 Ways to get data

2.1 Create your own

Perhaps the most overlooked source of Big Data is you.

Everyday you create data, some of which is captured, most is lost. The data that is captured can tell a lot about you, and can be displayed in interesting ways. Nicholas Felton has been collecting and publishing personal data since 2005 [1] (see Figure 1). Your smartphone can provide all sorts of data about where you are when and how fast you're going. Aaron Parecki is the co-founder of IndieWebCamp, and maintains oauth.net. He is known for having tracked his location at 5 second intervals since 2008 [3] (see Figure 2). If you are wearing a Fitbit (or other wearable sensor), them you are creating data all the time. Now it is a matter of collecting and accessing it. Locating, understanding, and manipulating the data can be a challenge.

Even if you aren't wearing a wearable device, most likely you are carrying one. Your smartphone is carrying an array of sensors, including:

1. Motion sensors that can tell the difference between waling and driving,
2. A barometer for measuring atmospheric pressure,
3. A gesture sensor that detects hand movements through infrared rays,
4. Gyroscope to measure acceleration,
5. Magnetometer to measure magnetic lines of flux,
6. GPS to tell where you are around the world,
7. WIFI to connect to the world, and also to tell how close you are to a broadcast station or router,
8. Camera(s) to see,
9. Microphone(s) to listen,
10. Speaker(s) to speak,
11. Temperature and pressure (on the screen).

Undoubtedly, there will be more as technology gets smaller, as they become more energy efficient, and people begin to want more and more. One immediate advantage (depending on where you live), will be that you can have your insurance rate reflect how you really drive.

Figure 1: Nicholas Felton's annual report.

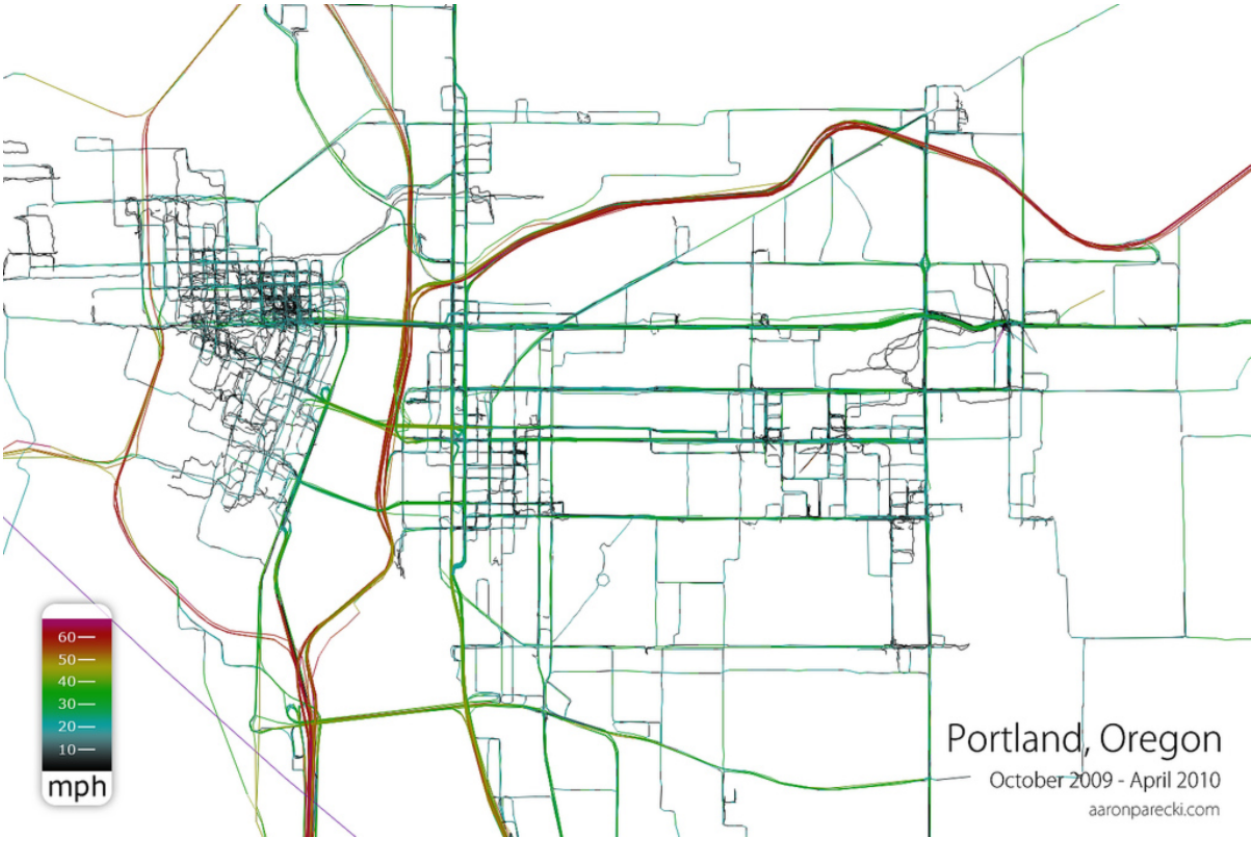


Figure 2: Aaron Parecki records his GPS position every 5 seconds.

An app running on your smartphone could communicate with the insurance company in real-time to report on your current driving activities[5]. Currently this is an "opt-in" application, but what could happen when the "opt-inners" out number the "opt-outters"? What will that say about your own perception of your driving habits? It could almost be Orwellian.

2.2 Download a file

Another way to acquire enough BD to have "fun" with is to download a file from somewhere that contains things of interest. Sometimes it is restively easy to find somewhere to download the file, other times it may take a while. Once you have located the file, you'll have to process it in some manner. The exact processing details will depend on how the file is laid out, and what tools you have available. All of this sounds very esoteric, and unfullfilling so we'll work through a couple of examples. The first will be to download a file from Medicare with tab separated values (TSV)[6]. The second will be a unformatted text file.

The Centers for Medicare and Medicaid Services[7] makes available a vast array of data relating to all their programs. As a specific example of a TSV, we will look at Medicare Payments during calendar year 2013 (see Figure 3)[8]. We are interested in the Medicare Physician and Other Supplier PUF, CY2013, tab delimited format file for download. Some particulars about the file Medicare_Provider_Util_Payment_PUF_CY2013.zip.

1. It is 497,014,400 bytes of compressed data.
2. It contains three other files (see Table 1).
3. The data file has a header record.
4. The data fields are delimited by tab characters that are normally invisible (see Figure 4).
5. The a single tab is between each data field. The editor can make the tabs visible (see Figure 5).

In the final analysis: the file is large, it is well structured, most modern languages will be able to process the file without trouble.

[5] `https://www.washingtonpost.com/news/the-switch/wp/2016/01/04/the-big-data-of-bad-driving-and-how-insurers-plan-to-track-your-every-turn/`

[6] TSV files are a subset of a general class of files called "delimited text" files. All delimited files strive to identify and use a particular character (either a TASB, or a SPACE, or a COMMA, or DOUBLE QUOTES) that the originiator believes will not occur normally in the data.

[7] `https://www.cms.gov/`

[8] `https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier2013.html`

Figure 3: Centers for Medicare and Medicaid Services download site.

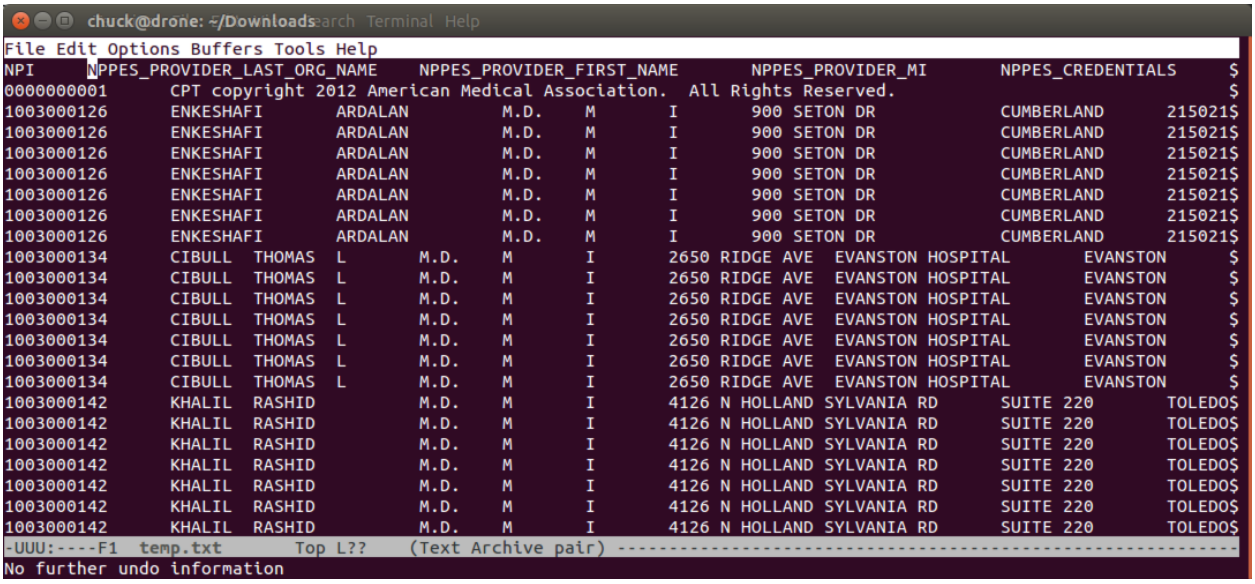


Figure 4: Medicare data file viewed normally.

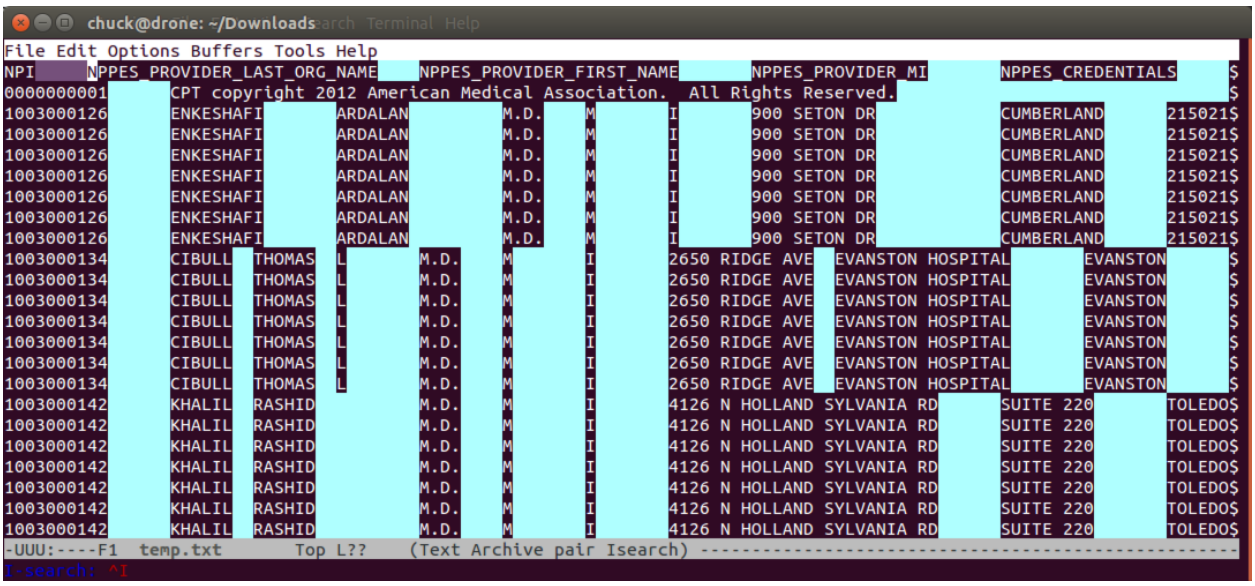


Figure 5: Medicare data file shown with tabs. Some editors go to great lengths to expand each tab character into some number of spaces (often 8) to try and make the text "pretty" on the display.

Length	Date	Time	File name
24,011	27-May-2015	14:47:38	CMS_AMA_CPT_license_agreement.pdf
3,650	14-Apr-2015	13:42:36	Medicare-Physician-and-Other-Supplier-PUF-SAS-Infile.sas
2,209,344,403	9-Apr-2015	16:14:28	Medicare_Provider_Util_Payment_PUF_CY2013.txt

Table 1: Contents of the file Medicare_Provider_Util_Payment_PUF_CY2013.zip. The data is almost 18 months old, and is the most recent available. The payment file is over 2.2Gigabytes in size and has 9,287,878 lines of data.

Project Gutenberg (PG)[9] (see Figure 6) is a volunteer effort to digitize and archive cultural works, to "encourage the creation and distribution of eBooks."[10] It was founded in 1971 by Michael S. Hart and is the oldest digital library. Most of the items in its collection are the full texts of public domain books. The project tries to make these as free as possible, in long-lasting, open formats that can be used on almost any computer. As of 3 October 2015, Project Gutenberg reached 50,000 items in its collection. The PG serves as a good source of text for many samples of free text. In this example we will look at the play: Romeo and Juliet (see Figure 7).

Some particulars about the PG version of Romeo and Juliet:

1. It has 5,557 lines.
2. It has 27,424 words.
3. It has 153,666 characters.
4. It has a PG specific header that is 289 lines long.

In the final analysis: the file is small, it is free form text, most modern languages will be able to load the file without trouble. However making sense of the file is a natural language processing (NLP) problem, and some computer languages are better than others when dealing with NLP tasks.

2.3 Download using an Application Program Interface (API)

Conceptually, application program interface (API) is a contract between one piece of software and another. One of these pieces of software will provide a service (commonly called the "server"), to the other (commonly called the "client"). The client provides data using the correct protocols and formats to the server, and the server will "fulfill" the contract by performing the appropriate action. Some actions that a server could perform include:

[9] `https://www.gutenberg.org/`

[10] `http://www.gutenberg.org/wiki/Gutenberg:Project_Gutenberg_Mission_Statement_by_Michael_Hart`

Figure 6: Project Gutenberg home page.

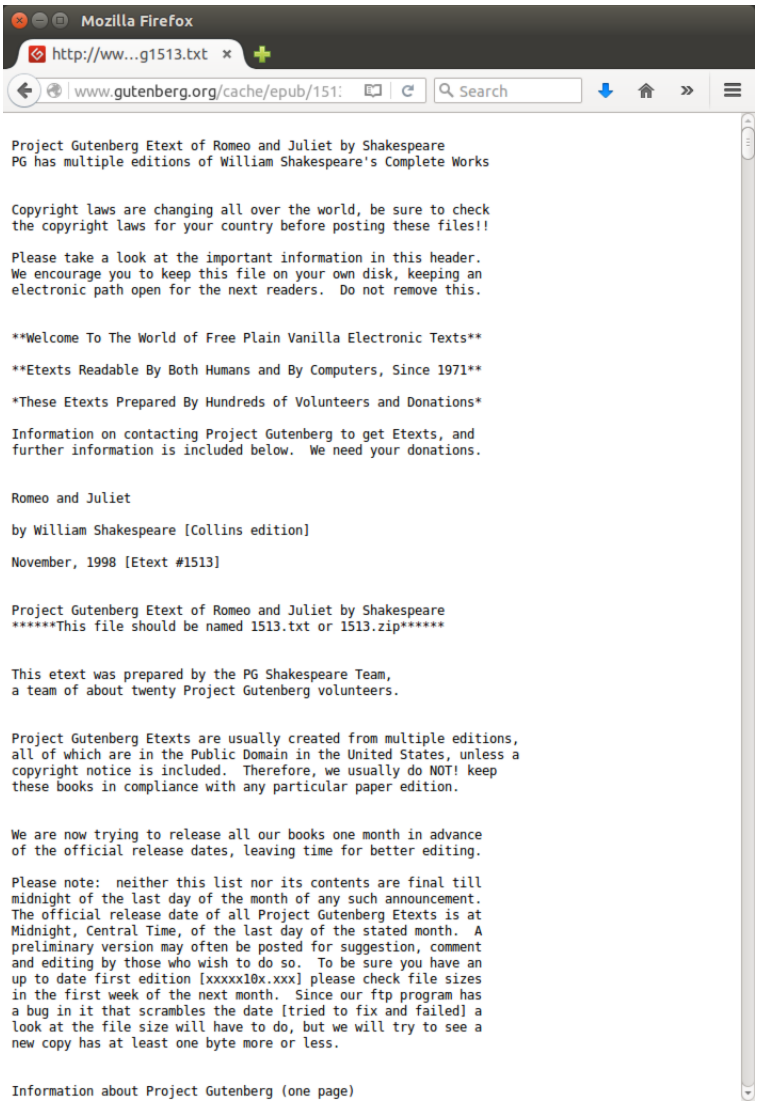


Figure 7: Project Gutenberg's copy of Romeo and Juliet.

- Return a web page when the client (a Web browser, using the hypertext transport protocol (HTTP)), properly requests one, or
- Compute a trigonometric function when the sin button on a calculator is pressed, or
- Encapsulate raw bits from a network interface into data frames as part of the Open System Interconnection (OSI) protocol.

It is the responsibility of the client to understand the details of the contract when requesting something from the server. The only standard that exists across all APIs is a functional one. To wit; if the client meets its side of the contract, then the server must meet its side. The server can unilaterally change the terms of the contract.

As an example, we will look at a moderately complex API to programmatically download pictures from Flickr.

1. Flickr has an extensive API[11] to allow applications to interact with Flickr and do all the things that a human would do, plus more (see Figure 8)
2. If you scroll down the Flickr API page, we are interested in using "flickr.photos.search" API (see Figure 9).
3. Flickr's flickr.photos.search API[12] "contract" has one required value, and wide range of optional values (see Figure 10).
4. The flickr.photos.search API requires a Flickr API key[13]. Applying for the key is completely on-line and is quick and easy.
5. If your interest is commercial, then apply for a commercial key. If your interest is browsing or investigative, then apply for a non-commercial key (see Figure 12).
6. You fill in all the necessary information, and press the SUBMIT key (see Figure 13).
7. Flickr will respond with a page showing the public and private API keys associated with your application (screen not shown). The public key is the one that the flickr.photos.search requires. The secret key is one that you should keep private to yourself and only use when necessary to authenticate yourself to Flickr.
8. The Flickr API page lists all the optional and required arguments with a brief explanation of the type of data associated with the argument. At the bottom of the page is a link to explore how the API operates (see Figure 14).

[11] `https://www.flickr.com/services/api/`
[12] `https://www.flickr.com/services/api/flickr.photos.search.html`
[13] `https://www.flickr.com/services/api/misc.api_keys.html`

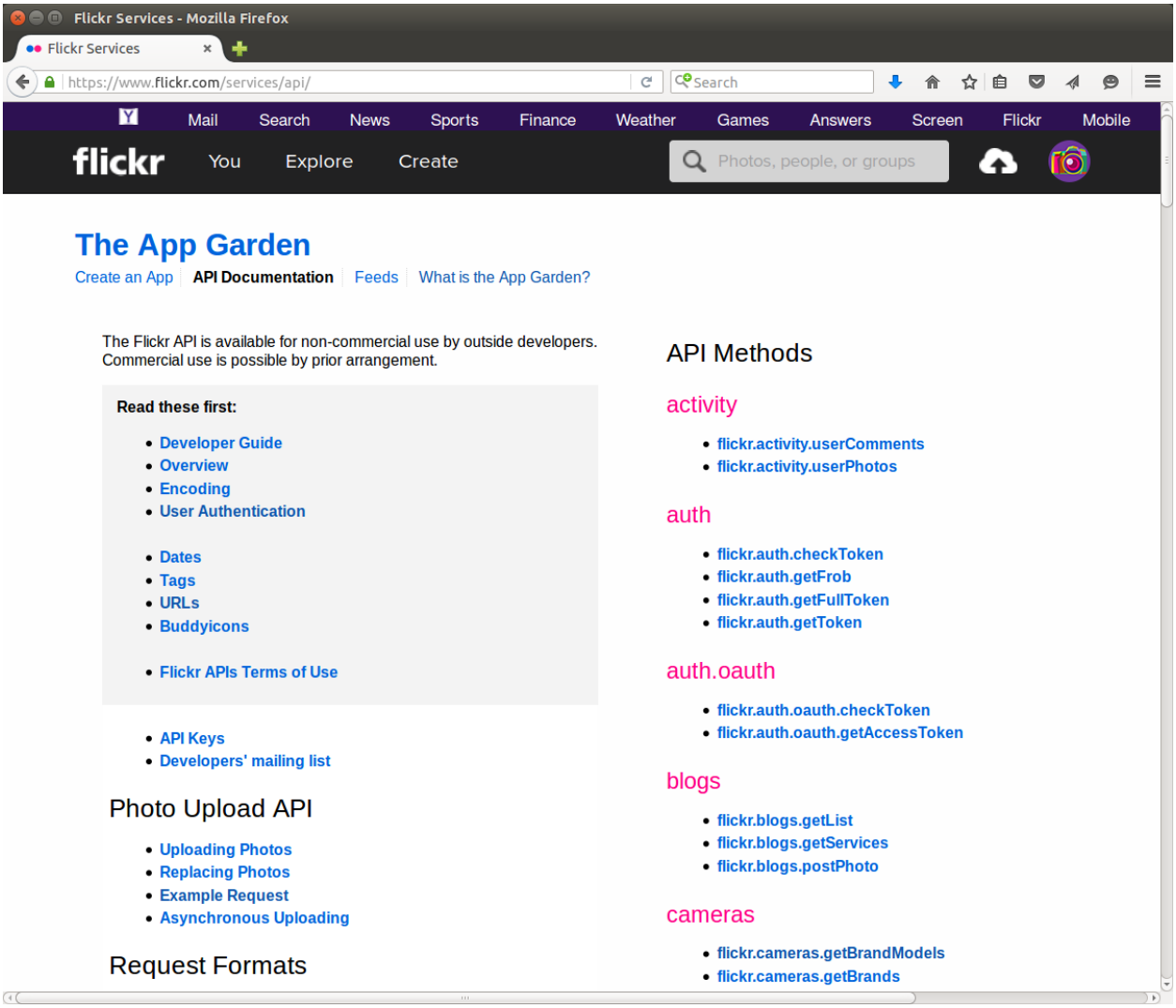


Figure 8: Flickr's API page.

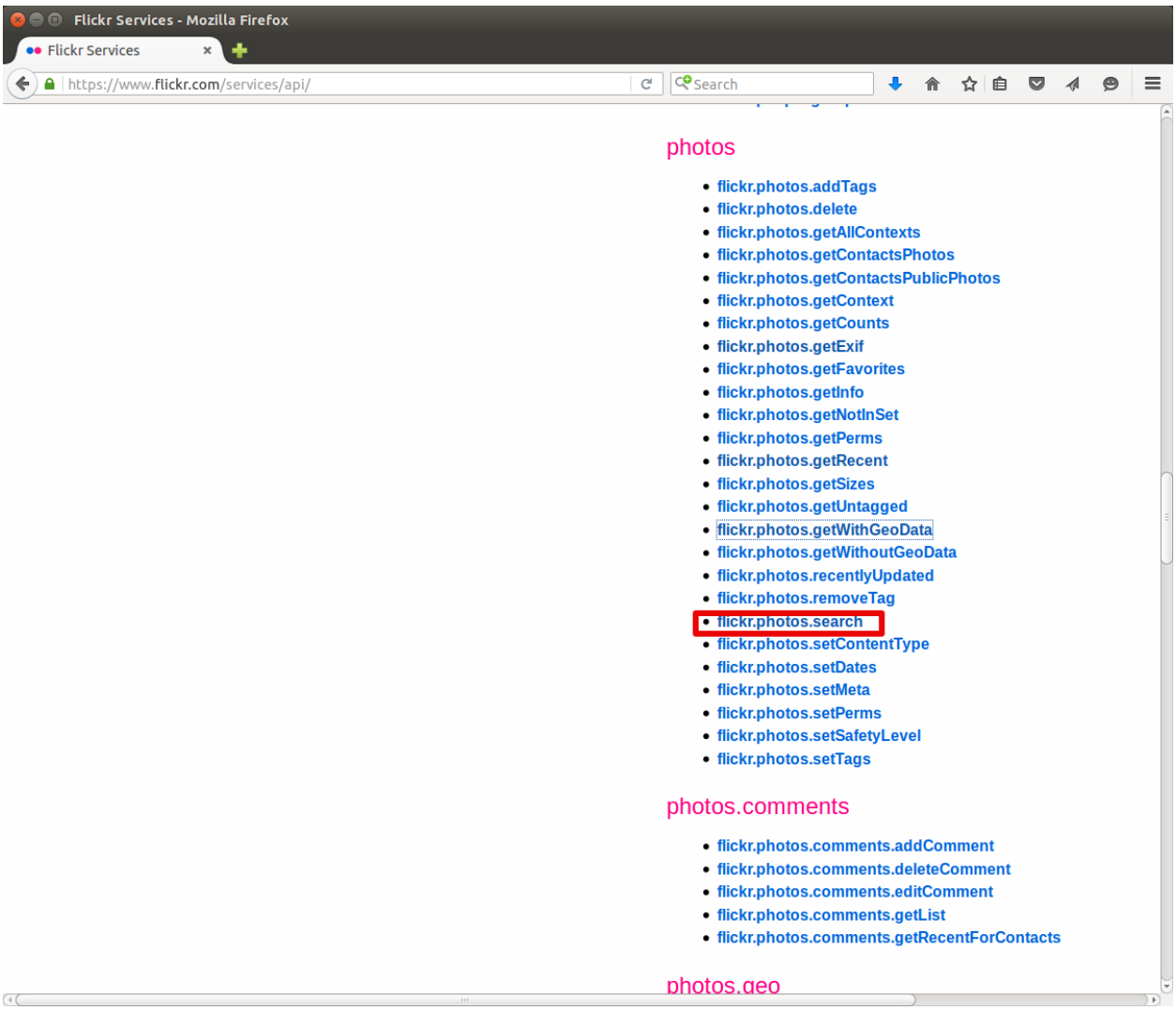


Figure 9: Flickr's flickr.photos.search API link. The link of interest is highlighted in red.

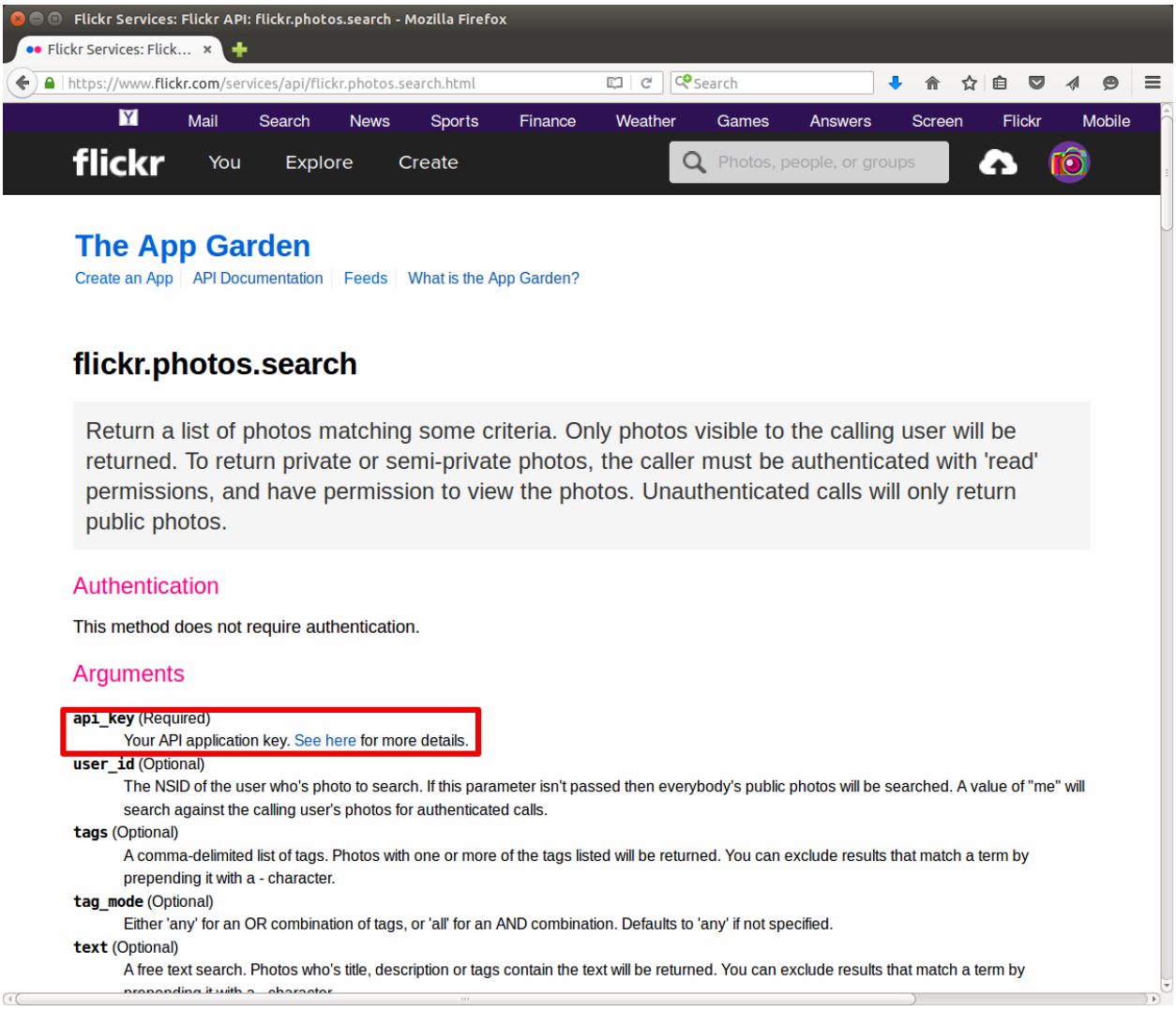


Figure 10: Flickr's flickr.photos.getWithGeoData API page. The single required data value is highlighted in red.

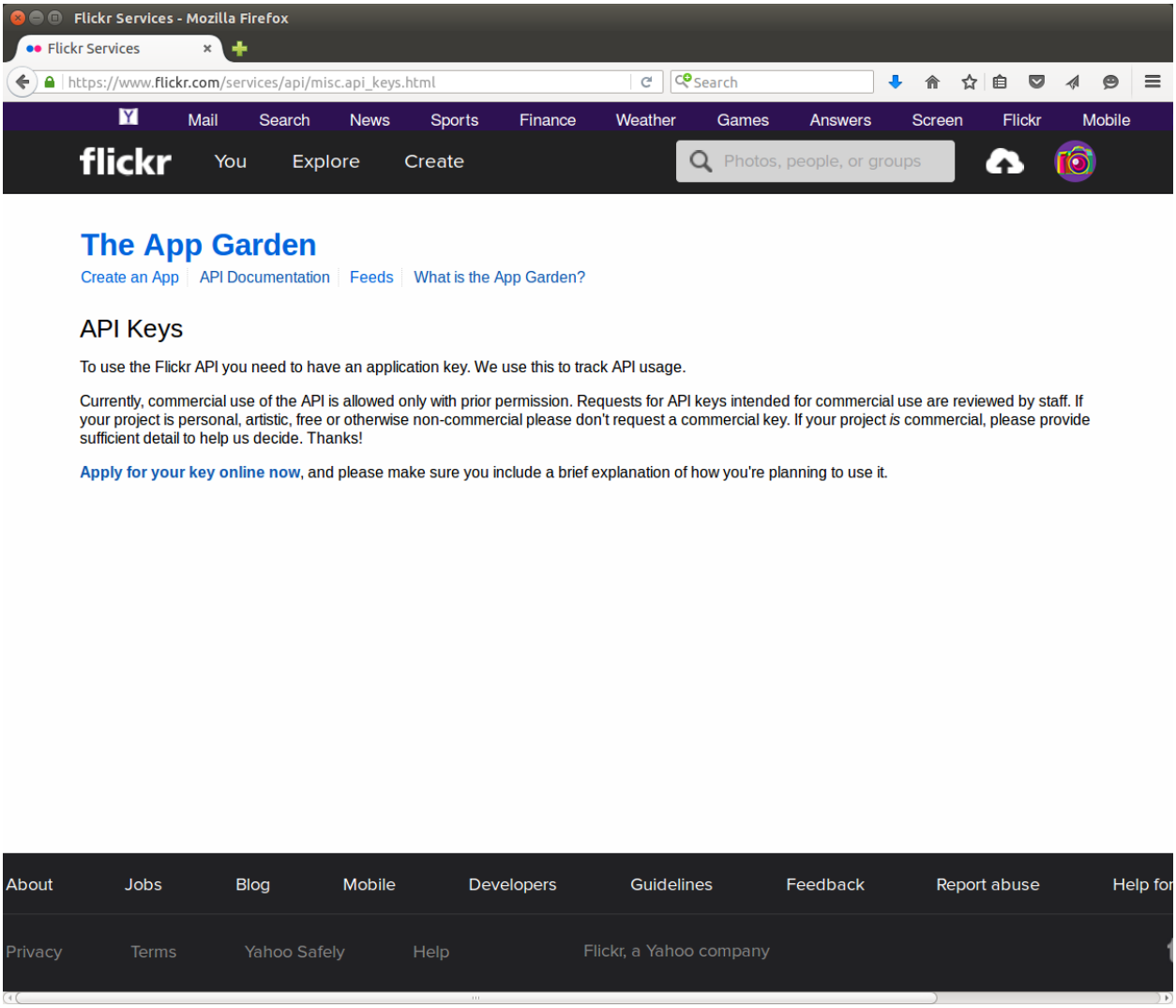


Figure 11: Flickr's API key page.

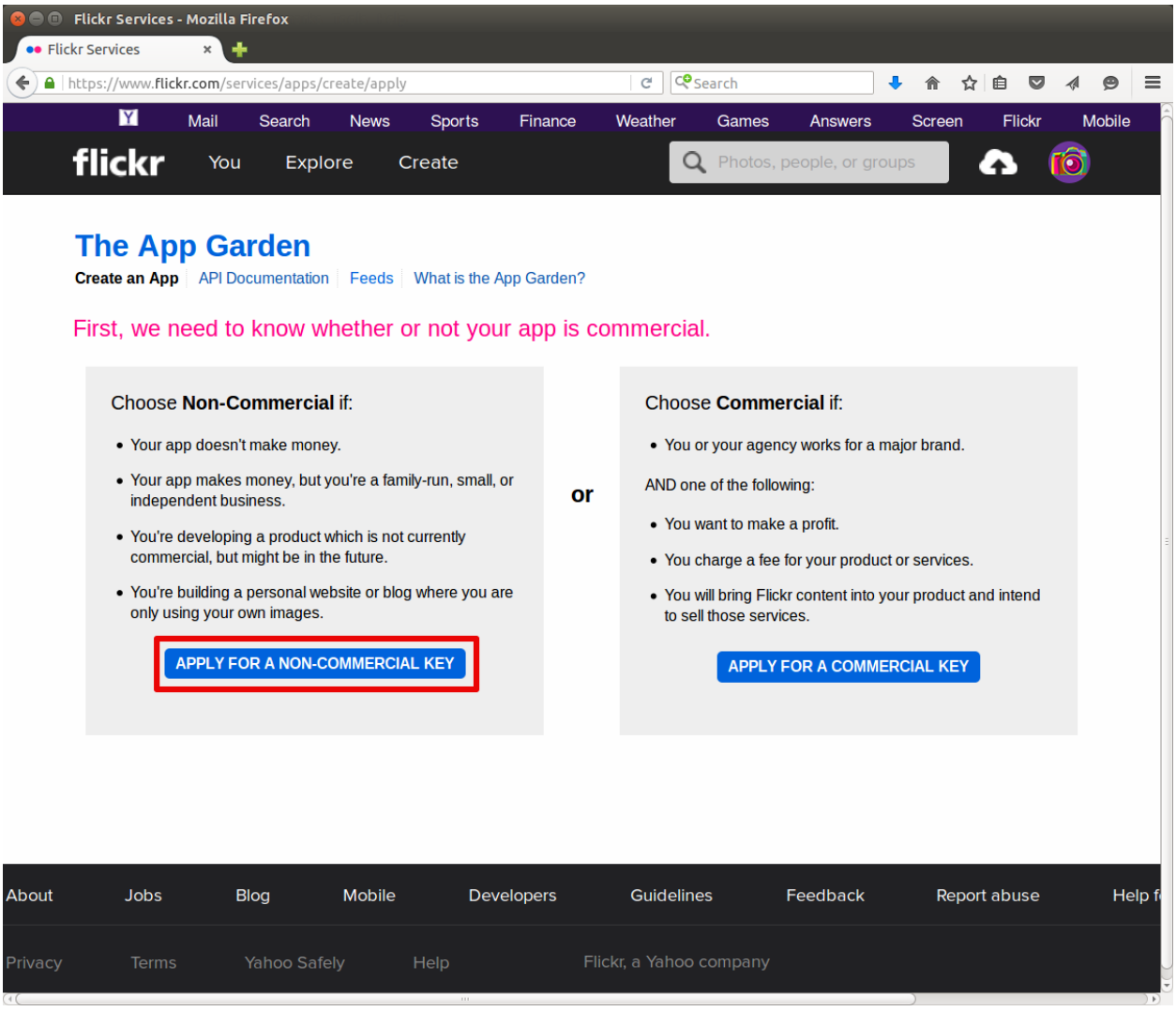


Figure 12: Flickr's API key type selection page. The non-commercial key selection is highlighted.

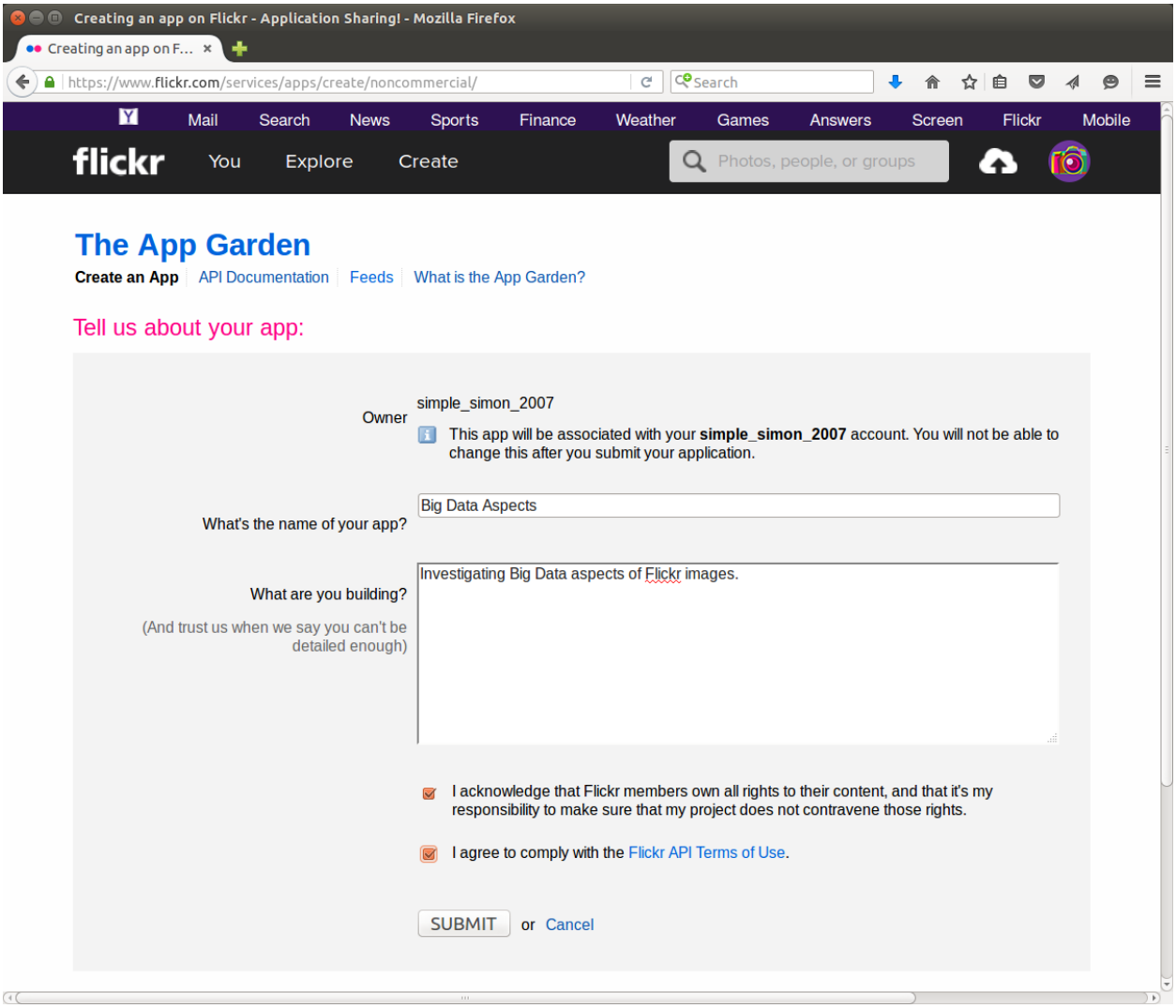


Figure 13: Flickr's application description page.

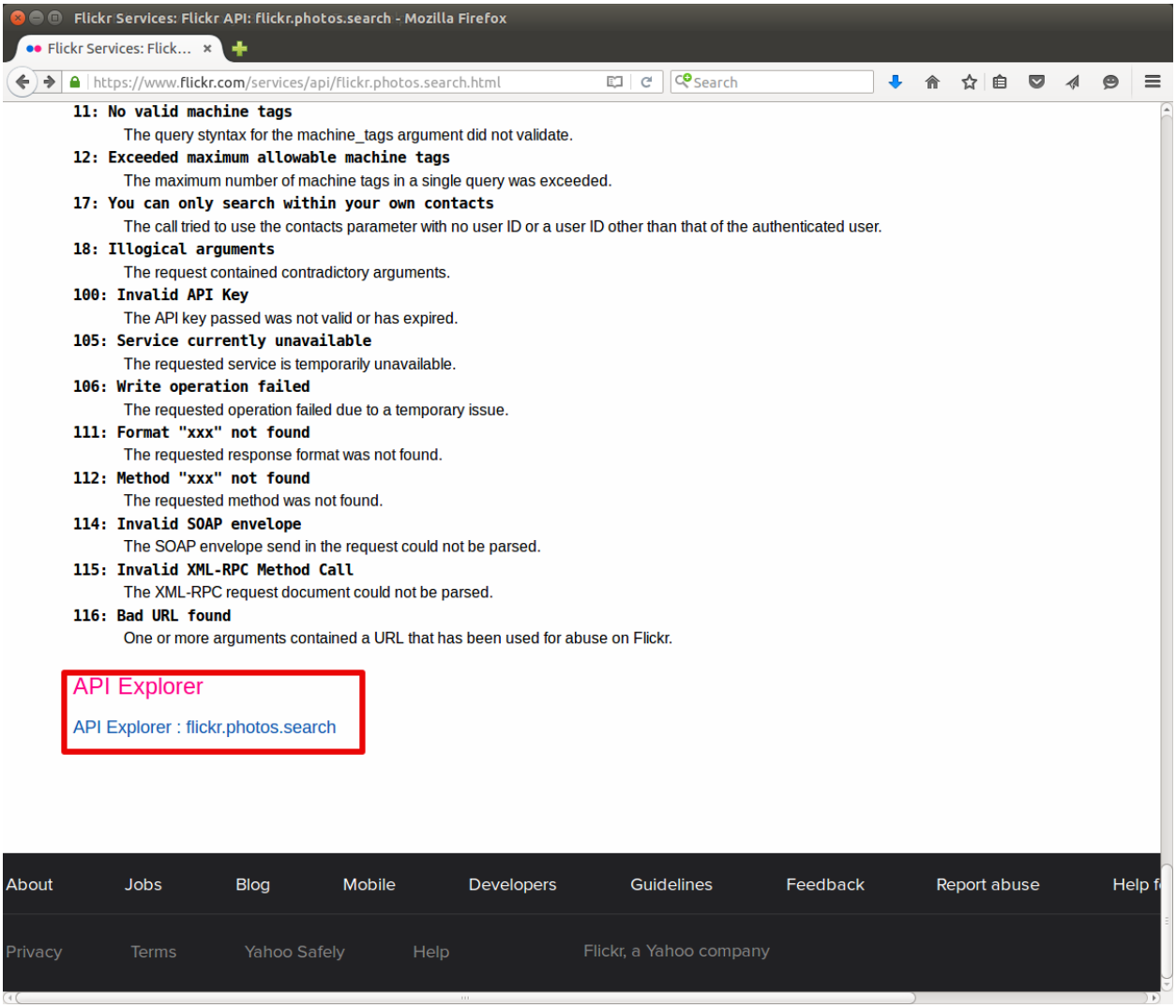


Figure 14: Flickr's API exploration link. The link is highlighted in red.

9. The Flickr API explorer page (see Figure 15)[14] allows you to fill out data for almost all the fields associated with the flickr.photos.search API. The only field that is not available is the API key field. The exploration field uses a default value for the key, so you don't have to enter one.

10. For this exploration, we are looking for pictures tagged with a position between 36 and 37 North, and 75 and 76 West

11. We are now ready to interact with the API and get data back in an XML format (see Figure 17).

12. After pressing the Call Method button, the page is refreshed. At the bottom of the page is the data returned from the call, and the URL used get the data from Flickr (see Figure 18). There are several items of interest in the response window. They include:

    (a) The number of pages of data that match the request (580 in this example).

    (b) The data shown comes from a particular page (1 in this example).

    (c) Each page has a number of picture references (100 in this example).

    (d) The total number of pictures (57,993 in this example).

13. The same API method call made a few minutes later only selecting JSON (JavaScript Object Notation) as the output format (see Figure 19). The difference between XML (eXtensible Markup Language) is more a matter of how you think about the data and the tools that you have available to manipulate the data rather than the data itself. The values are slightly different than the XML image because time has passed and the picture owners may have removed pictures from Flickr. Remember that Flickr is a semi-real time application because people can change things without notifying you.

Sample flickr JSON and XML response files are attached[15].

[14] `https://www.flickr.com/services/api/explore/flickr.photos.search`

[15] The JSON file is here. The XML file is here.

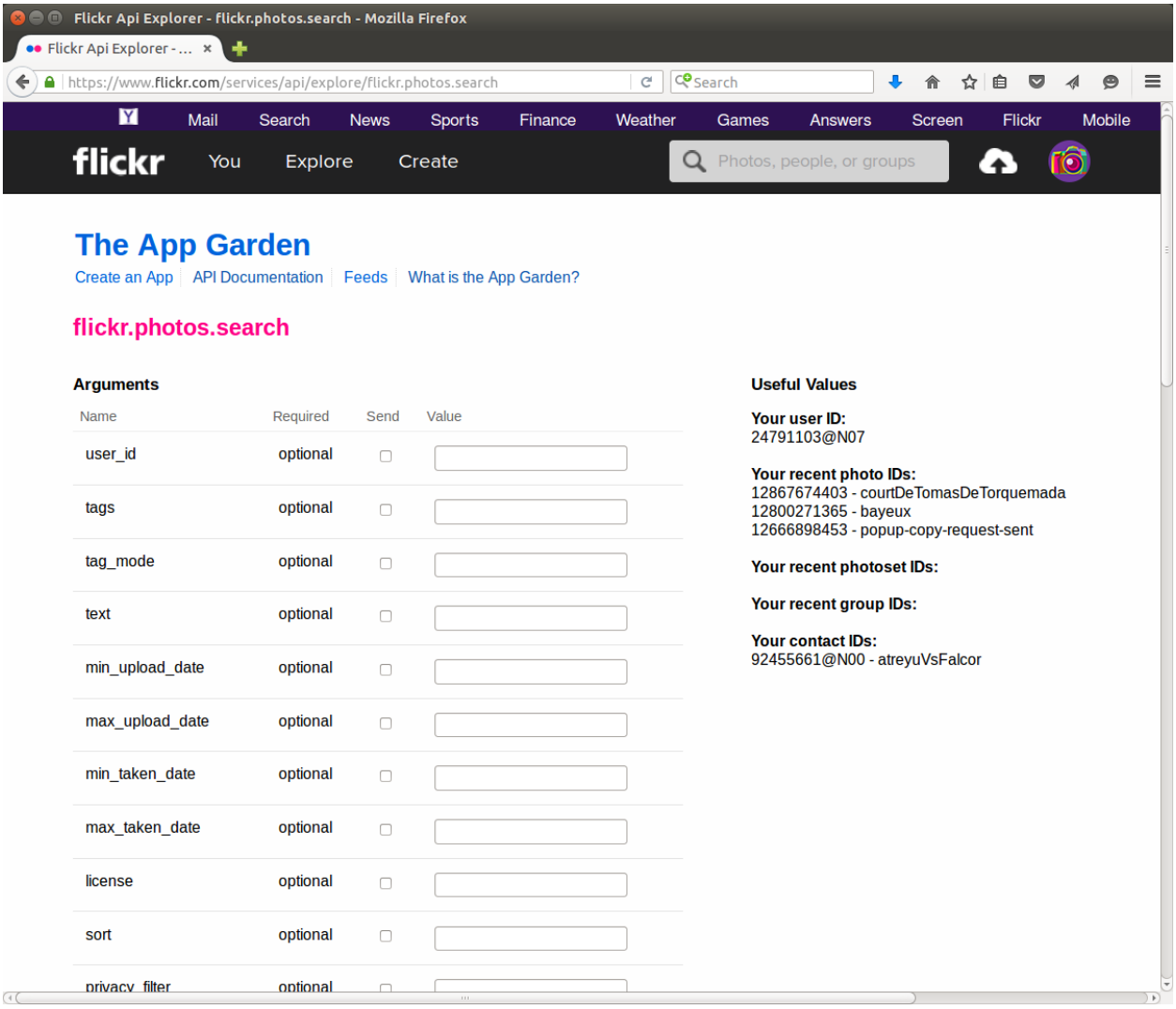


Figure 15: Flickr's API exploration page.

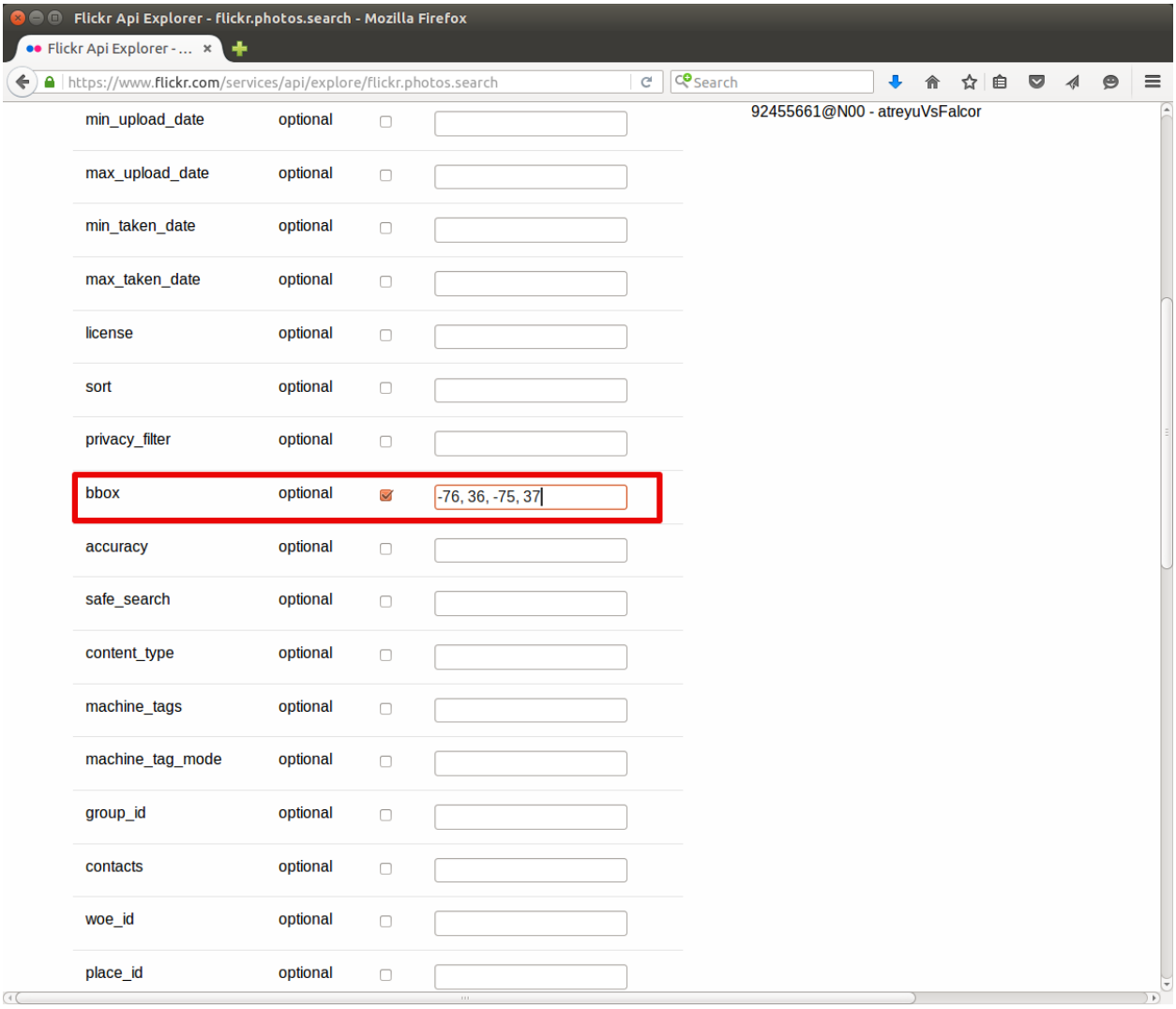


Figure 16: Flickr's API exploration page with bounding box data.

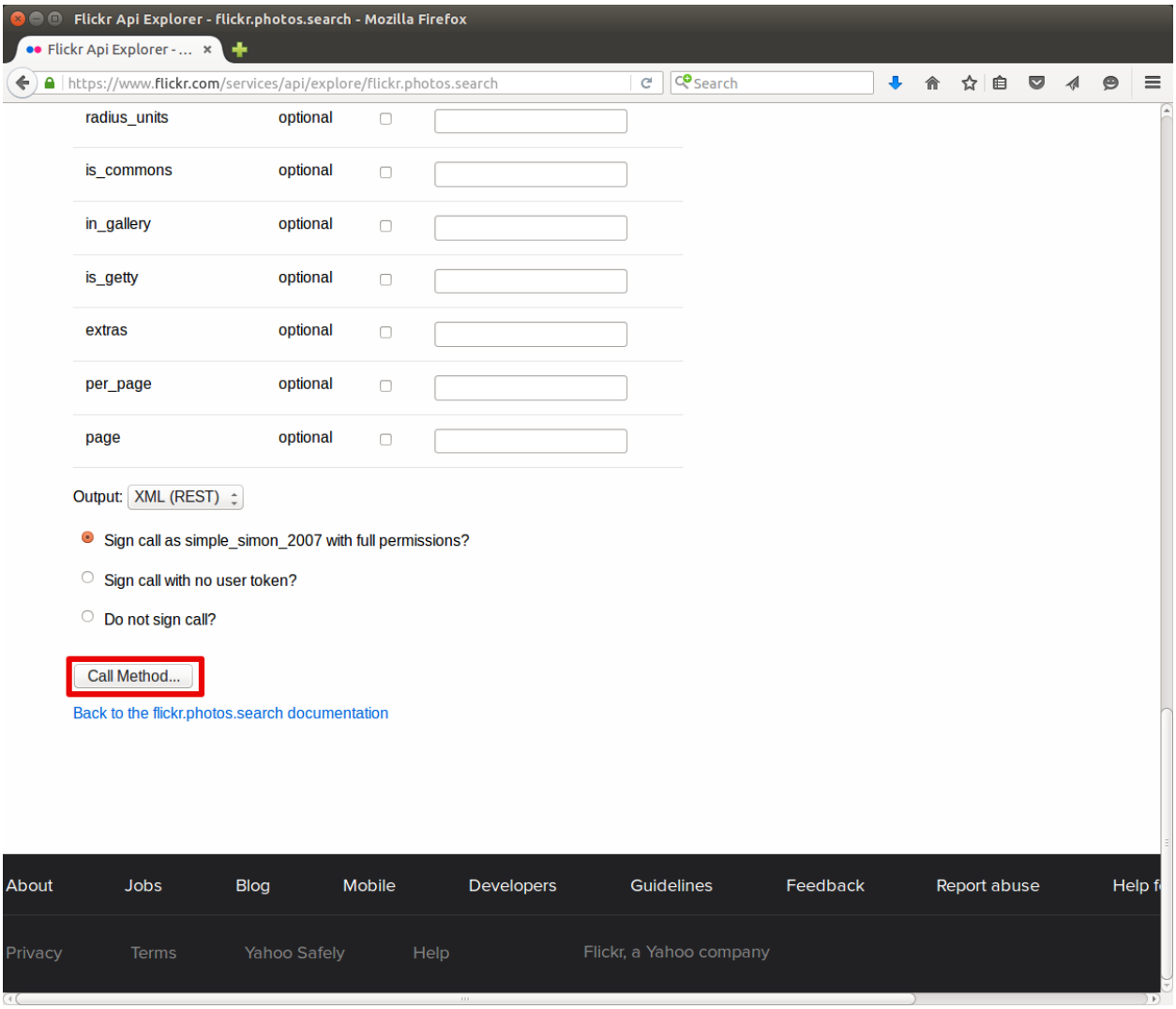


Figure 17: Flickr's API exploration calling method. The Call Method button is highlighted in red.

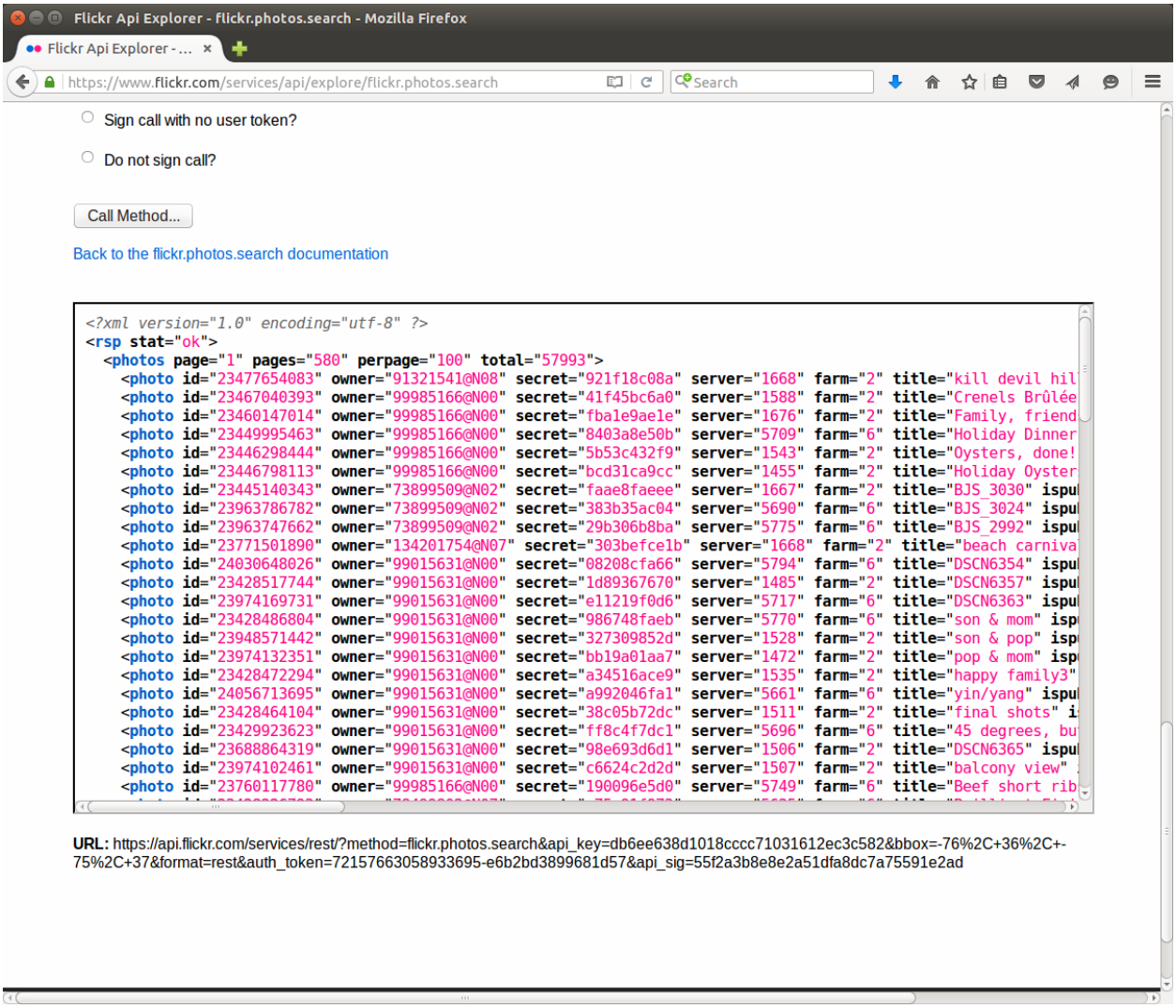


Figure 18: Flickr's API exploration with data and XML format.

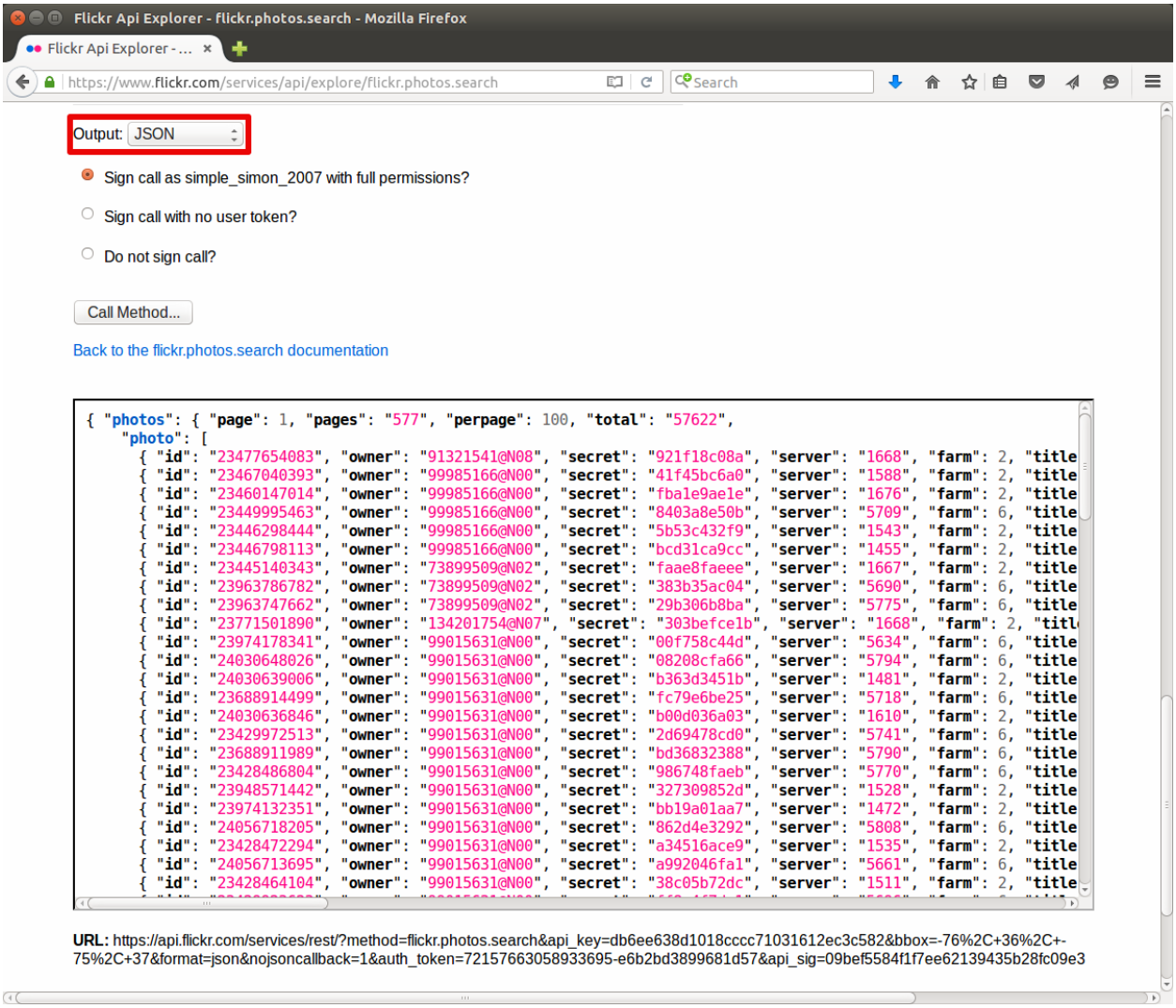


Figure 19: Flickr's API exploration with data and JSON format.

3 Selected Big Data Sources

The sites listed in this section are attached to this PDF. The R driver program used to collect the home pages for the sites in this report is attached as well. The R worker program used to do the "heavy lifting" is also attached.

3.1 Aggregator

1. 70+ websites to get large data repositories for free: A blog dealing with all things Big Data. (see Figure 20)
   `http://bigdata-madesimple.com/70-websites-to-get-large-data-repositories-for-free/`

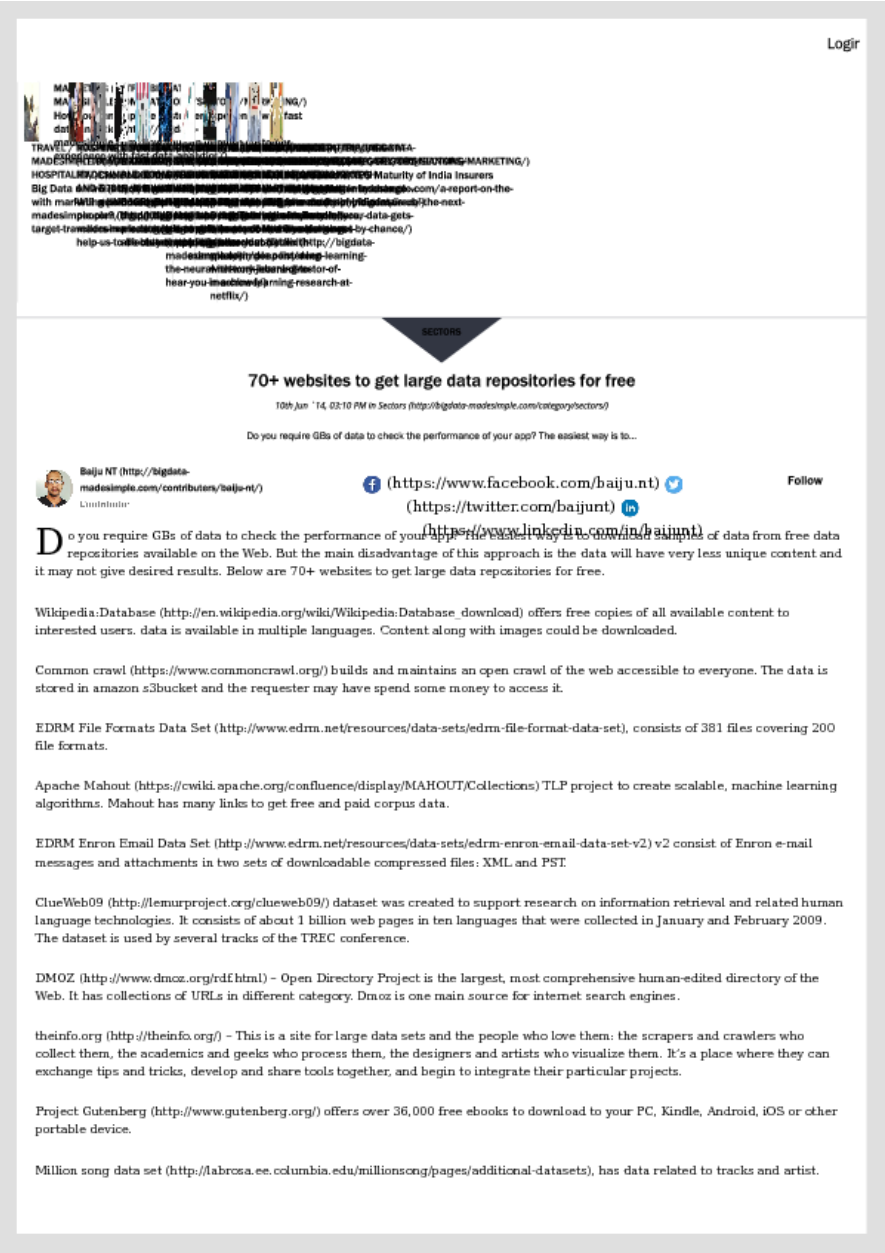


Figure 20: 70+ websites to get large data repositories for free home page.

2. Anywhere that I could download a sample database? : learnprogramming: A general question posed by many, and containing links to other sources. (see Figure 21) `https://www.reddit.com/r/learnprogramming/comments/2wr4mm/anywhere_that_i_could_download_a_sample_database/`

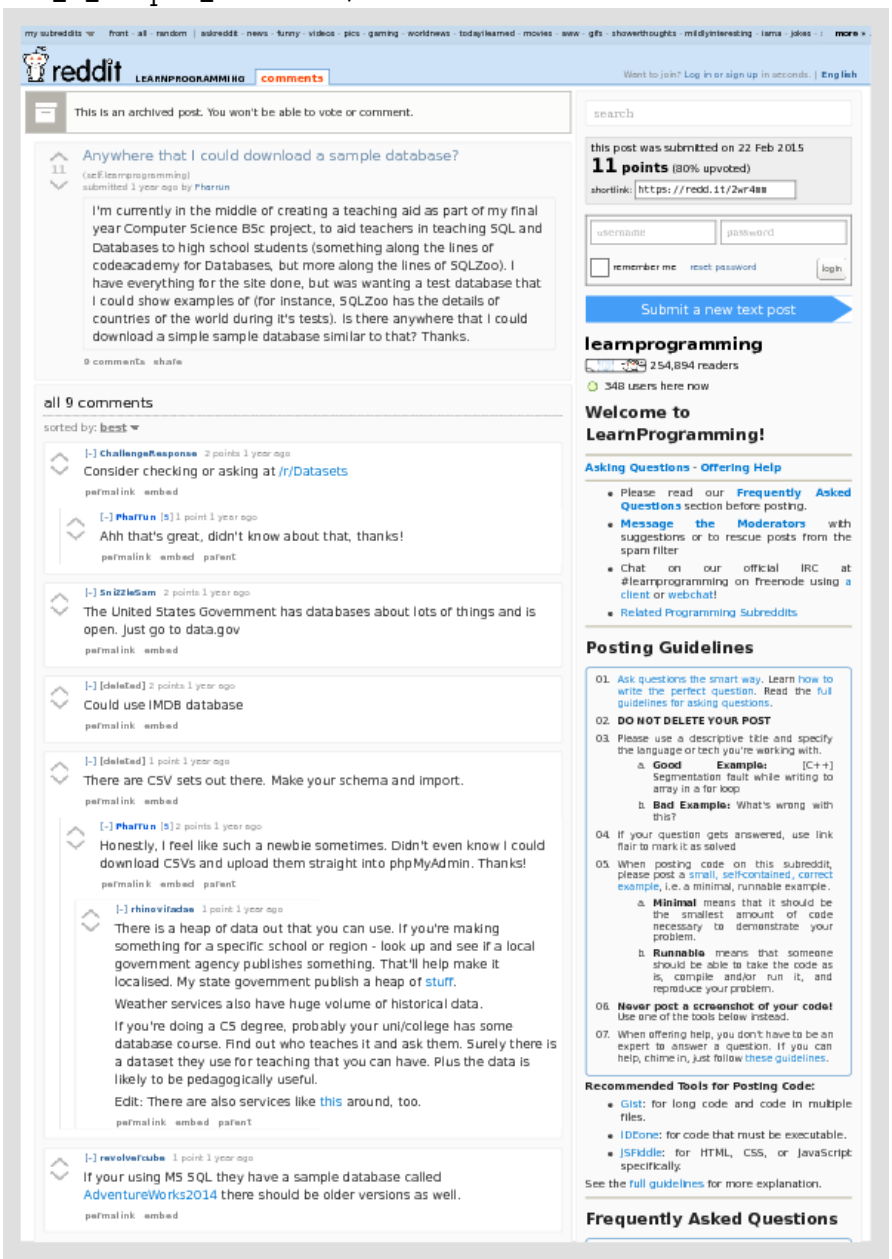


Figure 21: Anywhere that I could download a sample database? : learnprogramming home page.

3. Data Portals: A comprehensive list of open data portals from around the world. (see Figure 22)
`http://dataportals.org/`

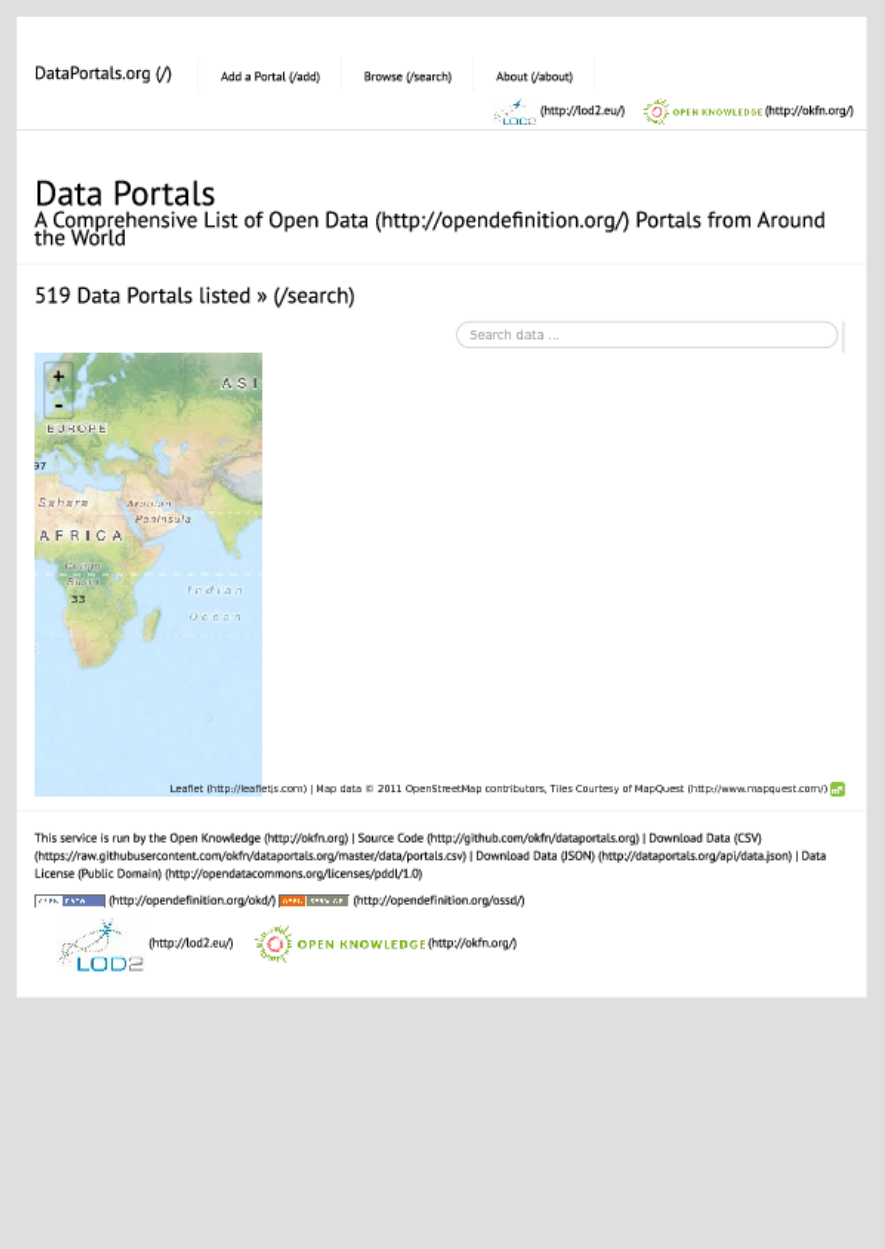


Figure 22: Data Portals home page.

4. Datasets Archive: A place to ask about the location of different datasets. (see Figure 23) `https://www.reddit.com/r/Datasets`

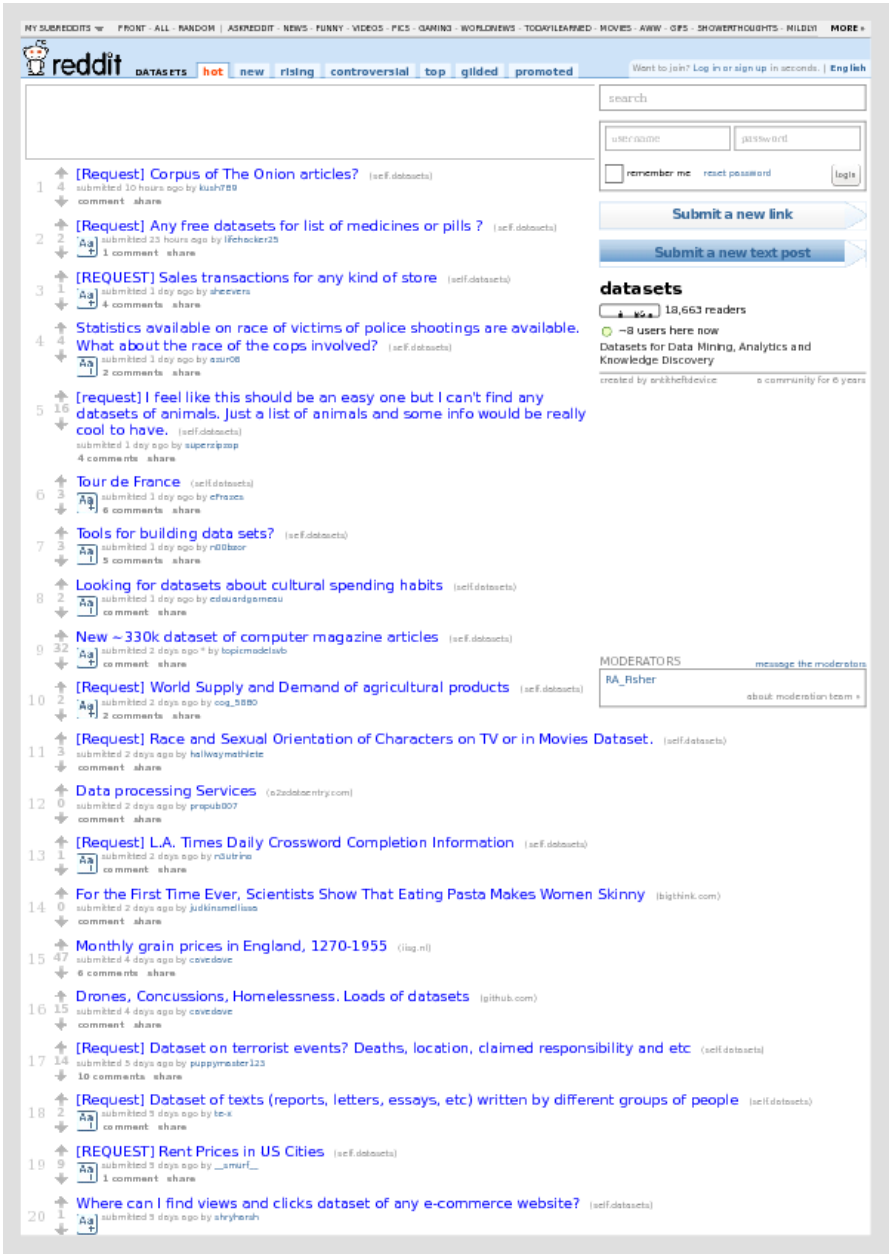


Figure 23: Datasets Archive home page.

5. Datasets for Data Mining and Data Science: A good source about data mining, analytics, big data, and data science. (see Figure 24)
`http://www.kdnuggets.com/datasets/index.html`

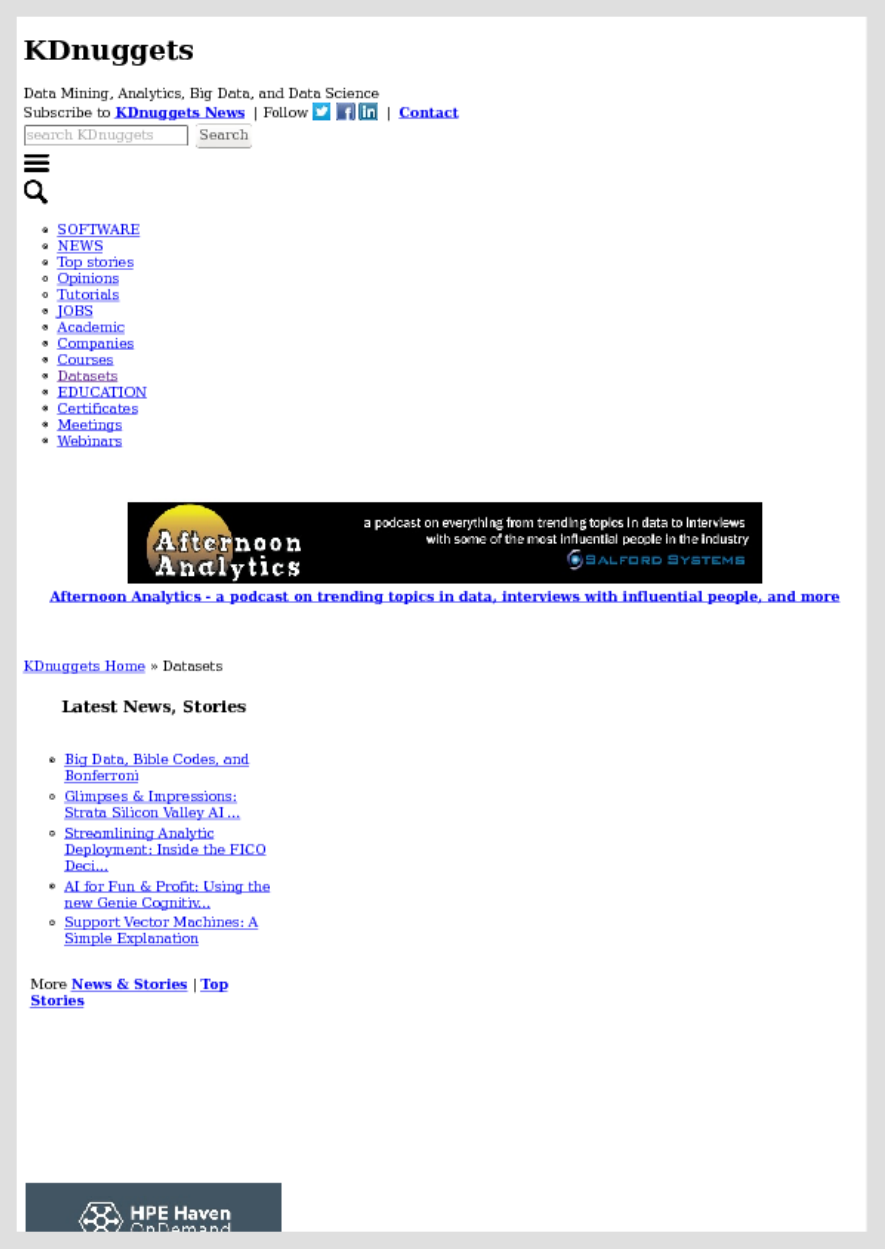


Figure 24: Datasets for Data Mining and Data Science home page.

6. enigma: A provider of data and data services. (see Figure 25)
`http://enigma.io/`

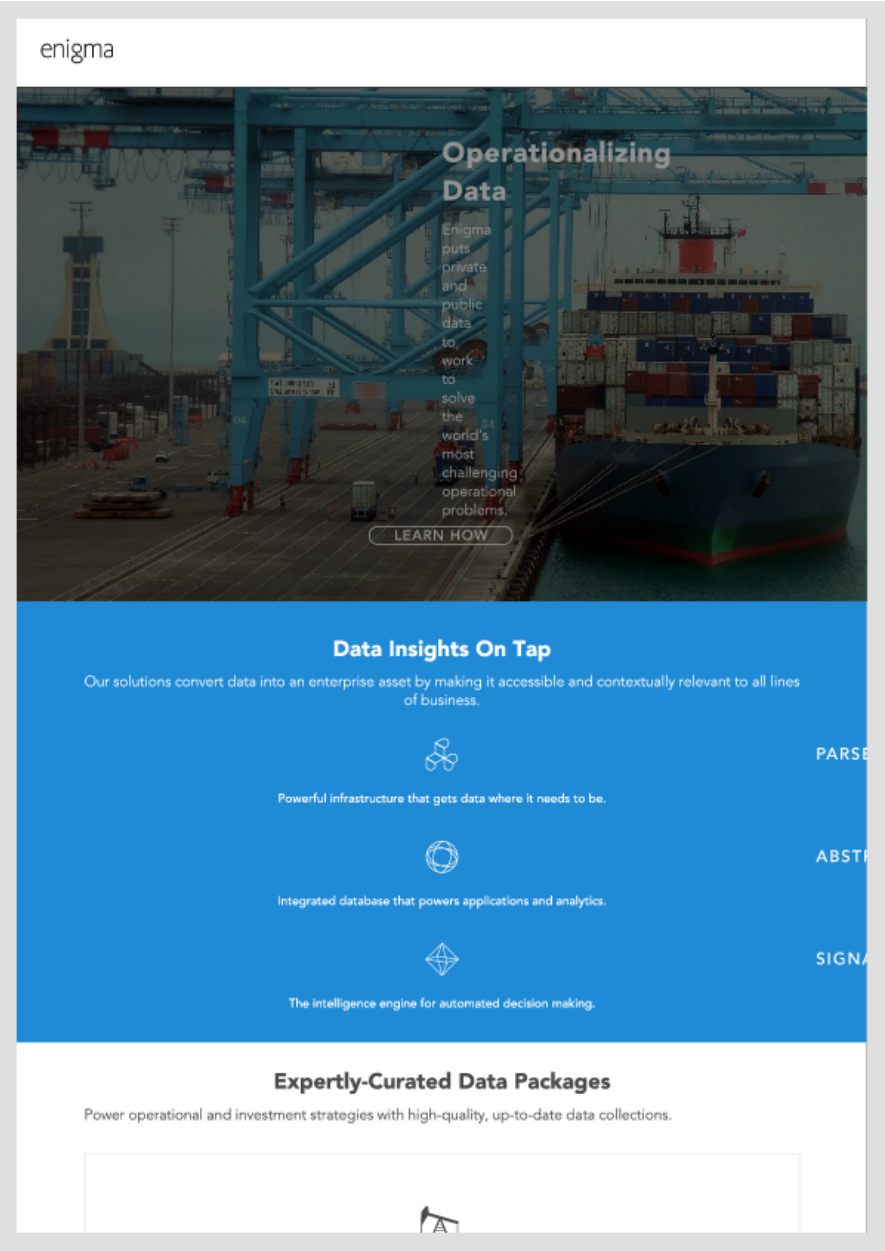


Figure 25: enigma home page.

7. Finding Data on the Internet: The following list of data sources has been modified as of 3/18/14. Most of the data sets listed below are free, however, some are not. (see Figure 26) `http://www.inside-r.org/howto/finding-data-internet`

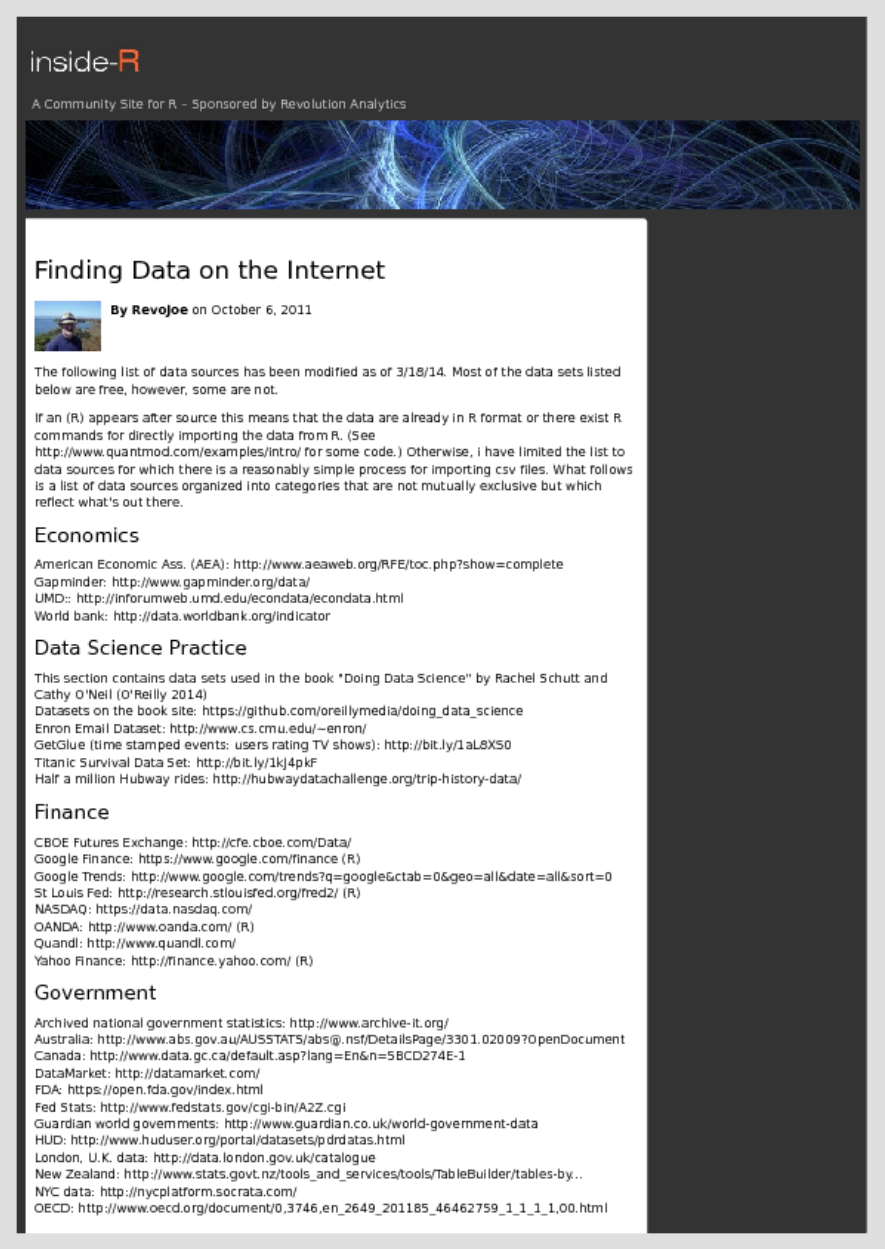


Figure 26: Finding Data on the Internet home page.

8. Linked Data - Connect Distributed Data across the Web: Linked Data is about using the Web to connect related data that was not previously linked, or using the Web to lower the barriers to linking data currently linked using other methods. More specifically, Wikipedia defines Linked Data as "a term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF." (see Figure 27)
`http://linkeddata.org/`

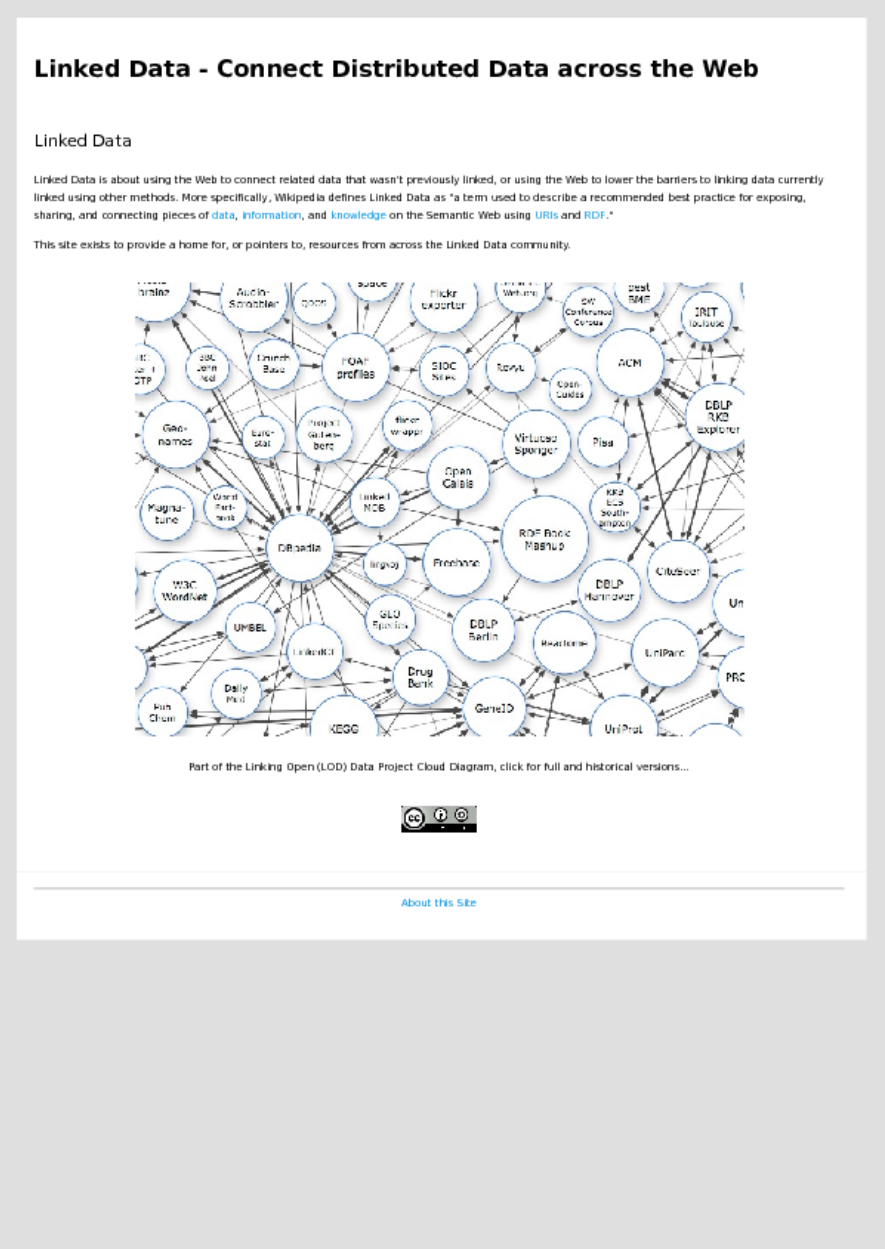


Figure 27: Linked Data - Connect Distributed Data across the Web home page.

9. Oxford Internet Institute University of Oxford: The Oxford Internet Institute was founded in 2001 at the University of Oxford, as an academic centre for the study of the societal implications of the Internet. (see Figure 28)
`http://geography.oii.ox.ac.uk/`

Figure 28: Oxford Internet Institute University of Oxford home page.

10. Real Time Data Resources: Many of our classroom projects use real time data sites that are difficult to find. This page contains links to some of the most compelling Internet web sites for educational use. These include a host of sites that have real time weather, air, water, and satellite information that will enhance traditional classroom lessons. (see Figure 29)
`http://www.k12science.org/materials/resources/realtimedata/`

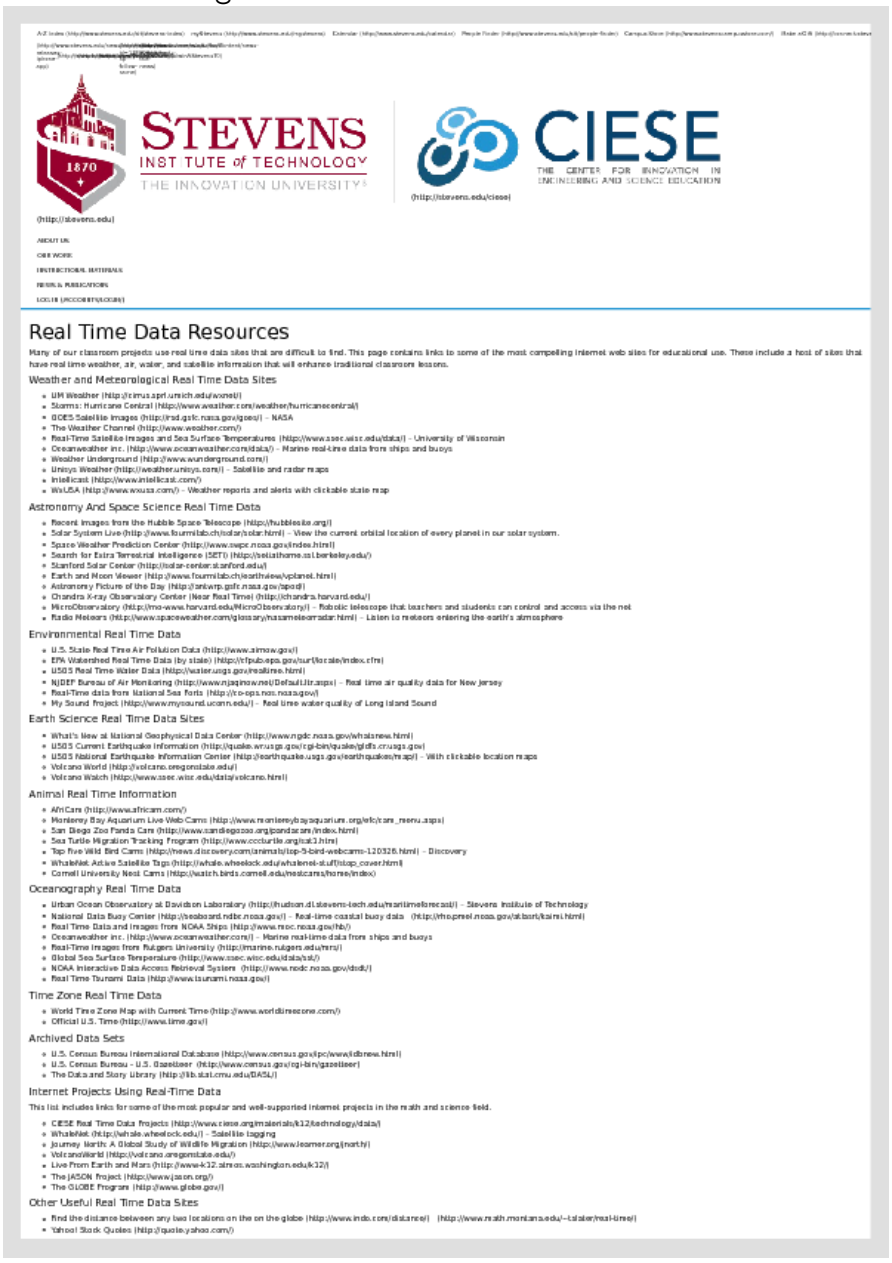


Figure 29: Real Time Data Resources home page.

11. Thirty Thousand Feet: Thirty Thousand Feet is an aviation directory with thousands of links to aviation web pages, aviation news, and other sources of commercial, military, and general aviation information. (see Figure 30)
`http://www.thirtythousandfeet.com/track.htm`

Figure 30: Thirty Thousand Feet home page.

12. Top 10 Weather APIs: Weather has become a pretty hot topic, especially in technology circles. Weather has not only become a standard app found on nearly every smartphone and mobile device, but it's being used by many technology companies in new and innovative ways. (see Figure 31)
`http://www.programmableweb.com/news/top-10-weather-apis/analysis/2014/11/13`

Figure 31: Top 10 Weather APIs home page.

13. University Corporation for Atmospheric Research: Search and access 178 data sets covering the Atmosphere, Ocean, Land and more. Explore climate indices, reanalyses and satellite data and understand their application to climate model metrics. This is the only data portal that combines data discovery, metadata, figures and world-class expertise on the strengths, limitations and applications of climate data. (see Figure 32)
`https://climatedataguide.ucar.edu/`

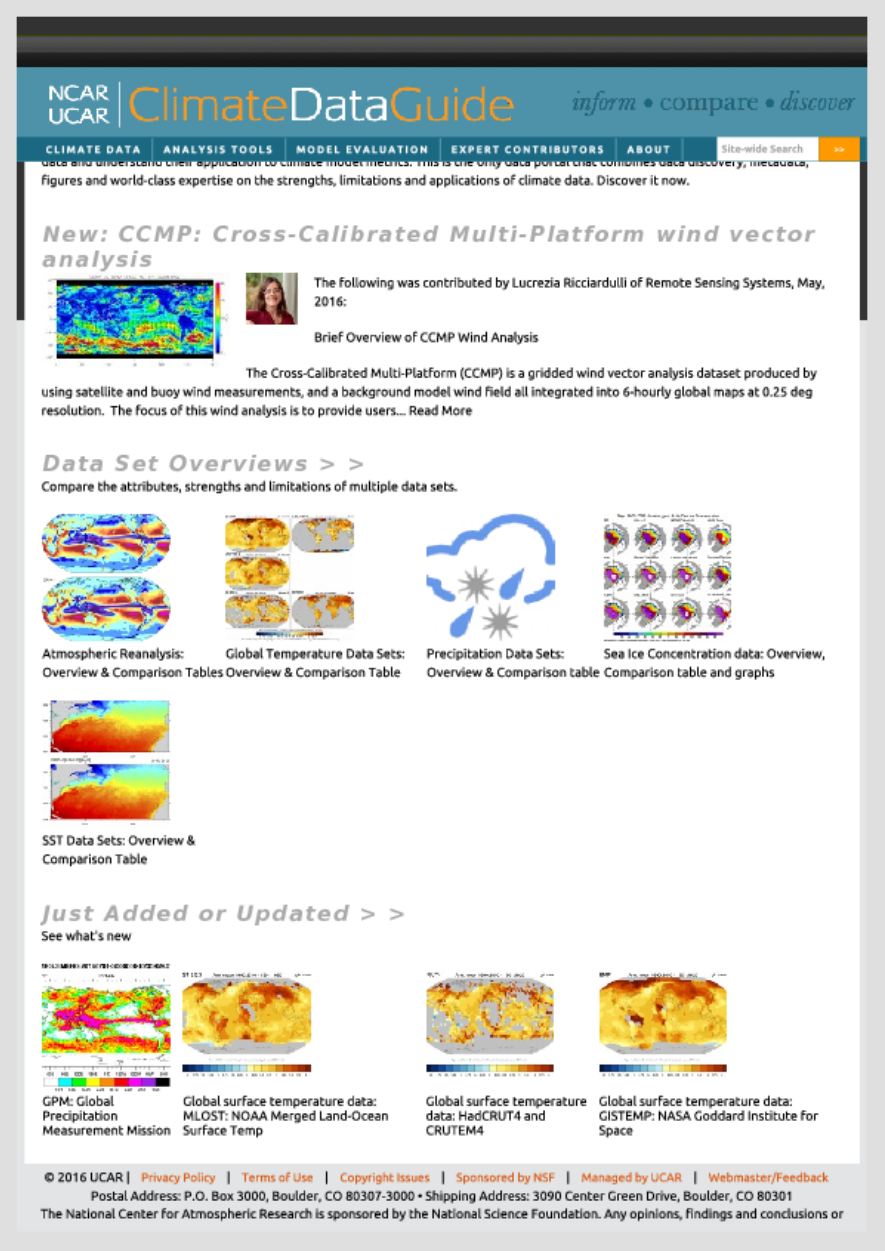


Figure 32: University Corporation for Atmospheric Research home page.

14. Where can I find large datasets open to the public? - Quora: The best answer to any question, especially about where to find Big Data. (see Figure 33)
`http://www.quora.com/Where-can-I-find-large-datasets-open-to-the-public`

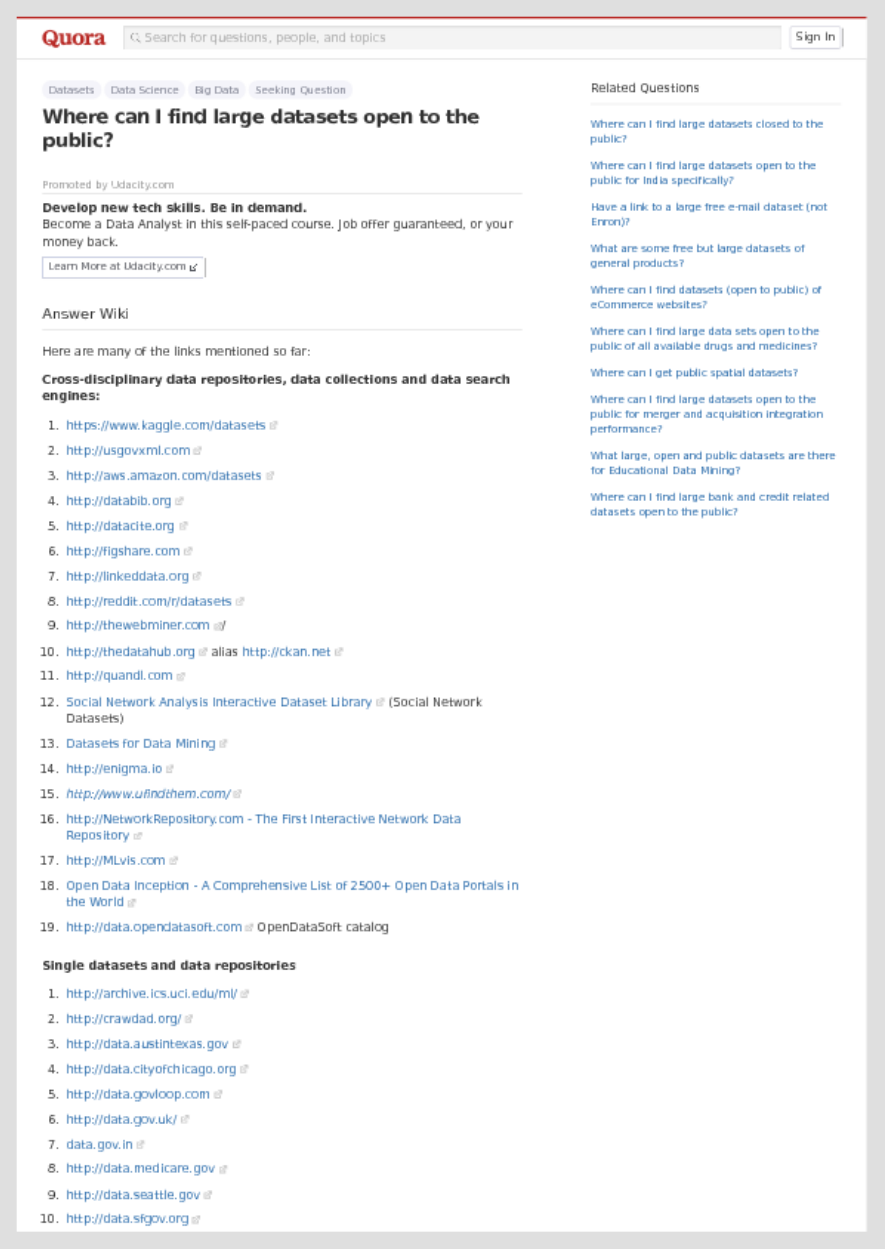
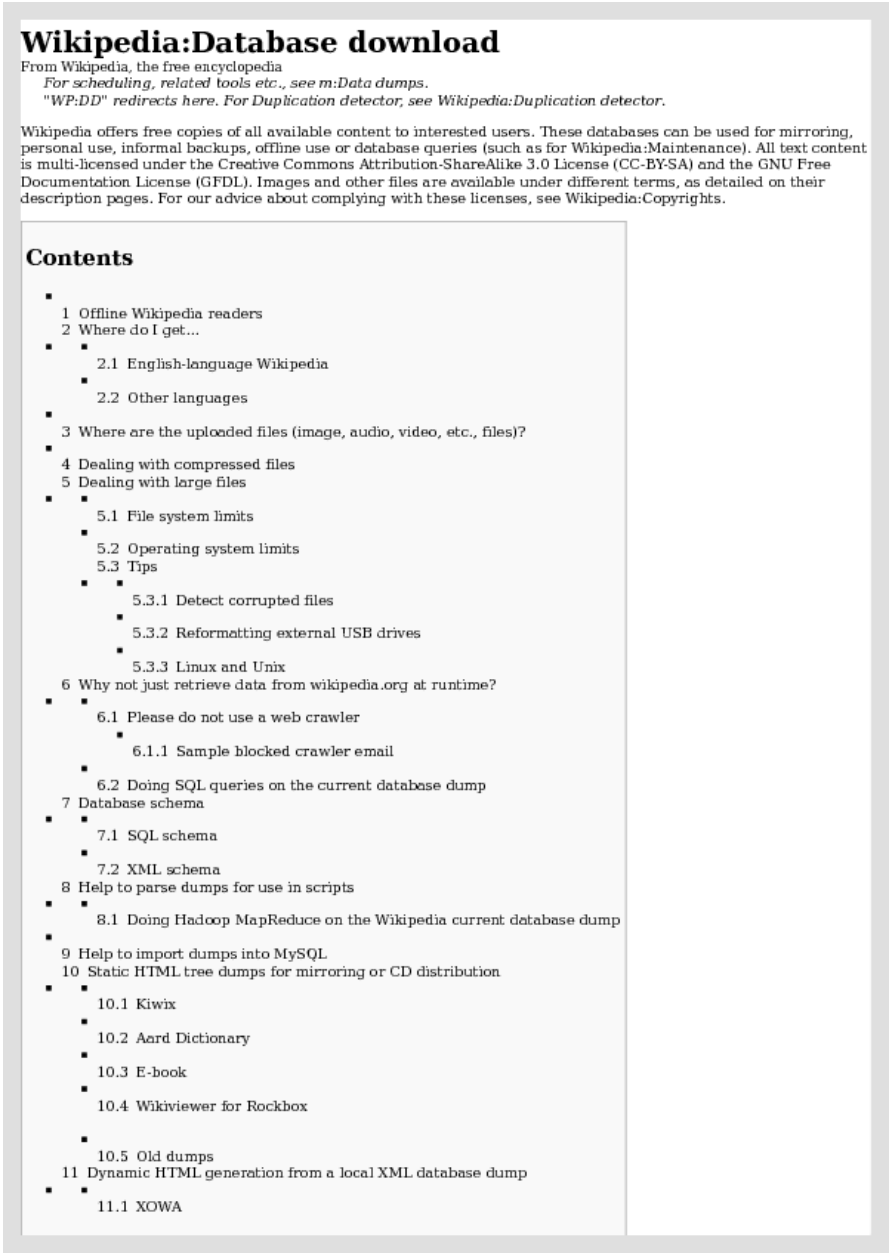


Figure 33: Where can I find large datasets open to the public? - Quora home page.

15. Wikipedia:Database download: Wikipedia offers free copies of all available content to interested users. These databases can be used for mirroring, personal use, informal backups, offline use or database queries (such as for Wikipedia:Maintenance). All text content is multi-licensed under the Creative Commons Attribution-ShareAlike 3.0 License (CC-BY-SA) and the GNU Free Documentation License (GFDL). Images and other files are available under different terms, as detailed on their description pages. (see Figure 34)
`https://en.wikipedia.org/wiki/Wikipedia:Database_download`



Wikipedia:Database download

From Wikipedia, the free encyclopedia

For scheduling, related tools etc., see m:Data dumps.

"WP:DD" redirects here. For Duplication detector, see Wikipedia:Duplication detector.

Wikipedia offers free copies of all available content to interested users. These databases can be used for mirroring, personal use, informal backups, offline use or database queries (such as for Wikipedia:Maintenance). All text content is multi-licensed under the Creative Commons Attribution-ShareAlike 3.0 License (CC-BY-SA) and the GNU Free Documentation License (GFDL). Images and other files are available under different terms, as detailed on their description pages. For our advice about complying with these licenses, see Wikipedia:Copyrights.

Contents

1 Offline Wikipedia readers
2 Where do I get...
- 2.1 English-language Wikipedia
- 2.2 Other languages

3 Where are the uploaded files (image, audio, video, etc., files)?
4 Dealing with compressed files
5 Dealing with large files
- 5.1 File system limits
- 5.2 Operating system limits
- 5.3 Tips
  - 5.3.1 Detect corrupted files
  - 5.3.2 Reformatting external USB drives
  - 5.3.3 Linux and Unix

6 Why not just retrieve data from wikipedia.org at runtime?
- 6.1 Please do not use a web crawler
  - 6.1.1 Sample blocked crawler email
- 6.2 Doing SQL queries on the current database dump

7 Database schema
- 7.1 SQL schema
- 7.2 XML schema

8 Help to parse dumps for use in scripts
- 8.1 Doing Hadoop MapReduce on the Wikipedia current database dump

9 Help to import dumps into MySQL
10 Static HTML tree dumps for mirroring or CD distribution
- 10.1 Kiwix
- 10.2 Aard Dictionary
- 10.3 E-book
- 10.4 Wikiviewer for Rockbox
- 10.5 Old dumps

11 Dynamic HTML generation from a local XML database dump
- 11.1 XOWA

Figure 34: Wikipedia:Database download home page.

3.2 Aviation

1. Flight Aware: Harness FlightAware's infrastructure to build an awesome aviation app using FlightXML. (see Figure 35)
   `http://flightaware.com/commercial/flightxml/`

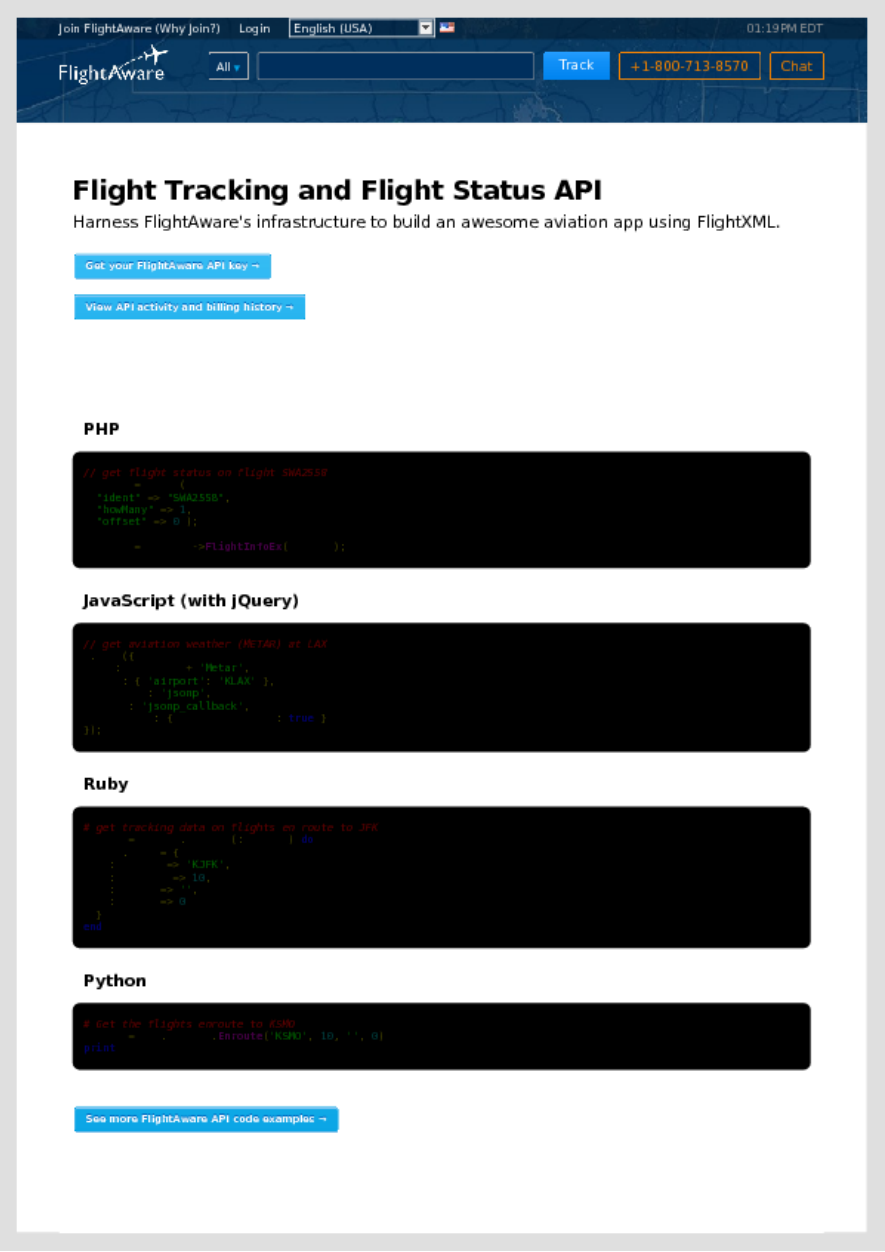


Figure 35: Flight Aware home page.

2. Flight Stats: A collection of flight related information including arrival, departures, and current status (see Figure 36)
`https://developer.flightstats.com/`

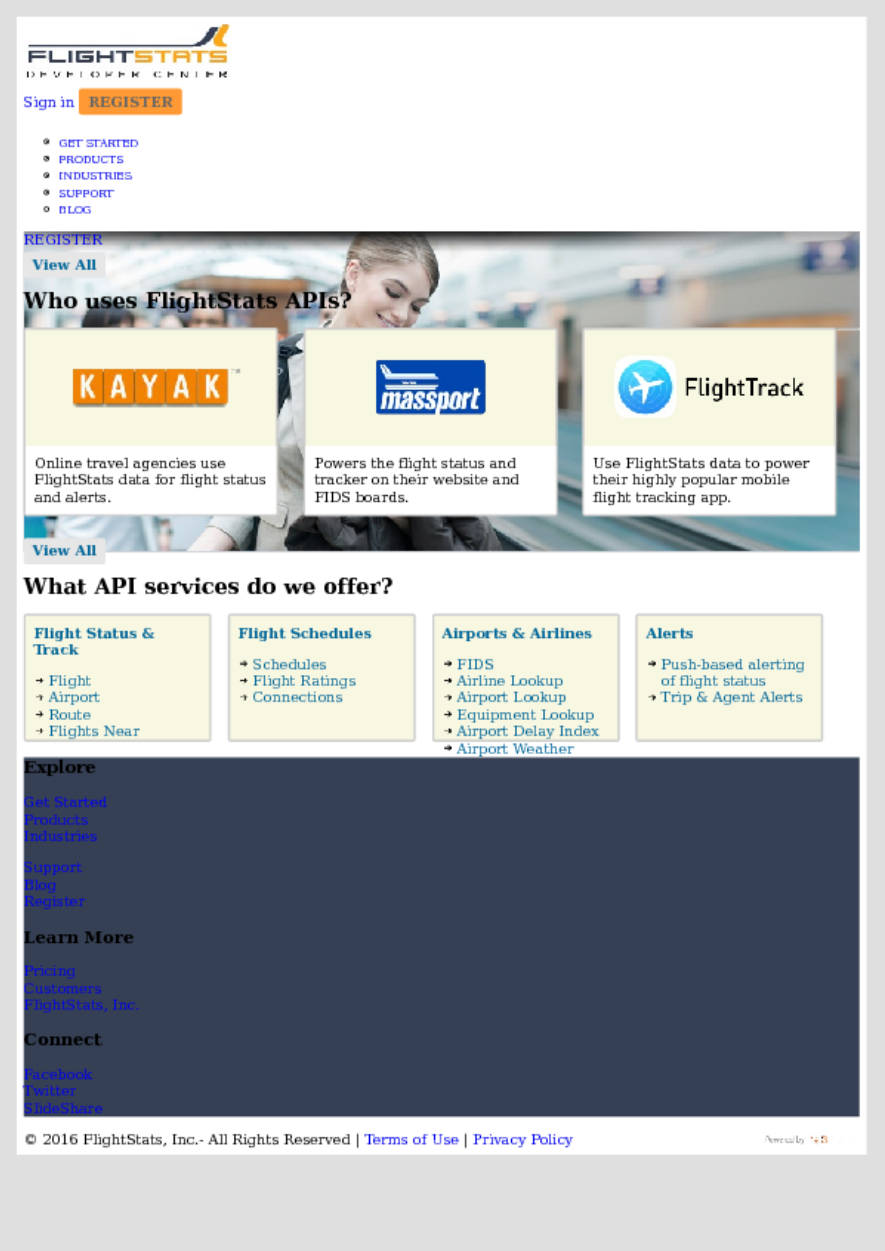


Figure 36: Flight Stats home page.

3. Flight View: FlightViews comprehensive, fast, accurate global real-time flight information aggregates data from hundreds of sources on more than 130,000 daily flights into a single truth data stream. Data customers and developers access FlightViews flight information through flexible and easily consumed APIs to power displays and applications that inform travelers and the businesses that provide travel services. (see Figure 37)
`http://corporate.flightview.com/products/data-feeds-apis`

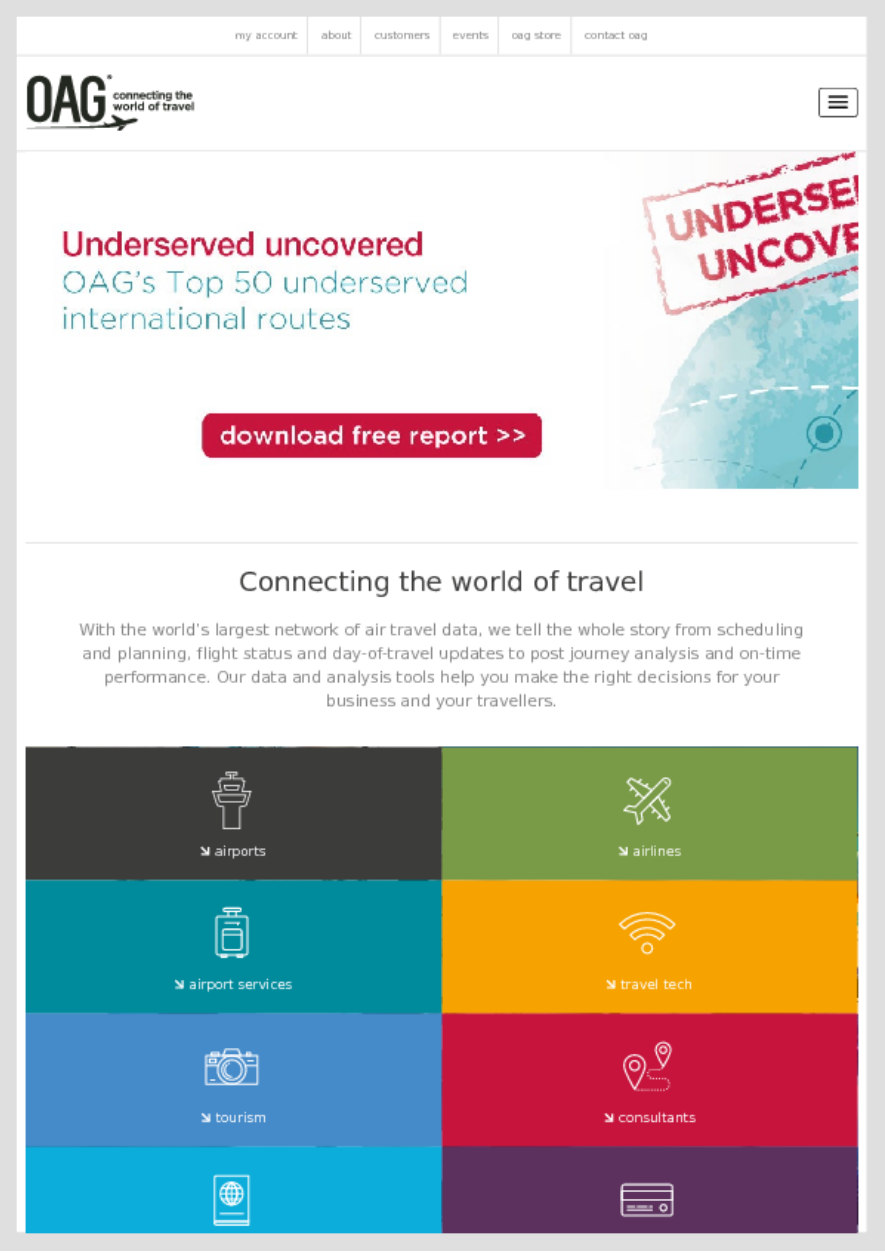


Figure 37: Flight View home page.

4. OAG (Official Airline Guide): OAG has the knowledge, experience and capability to provide reliable, high quality, up-to-the-minute aviation data, offering essential information and innovative travel planning solutions. Our products are vital to strategic and commercial planning, driving key business decisions and delivering Absolute Aviation Advantage across the worldwide air transport industry. (see Figure 38)
`http://www.oag.com/Flight-Schedules/Schedules-OnDemand`

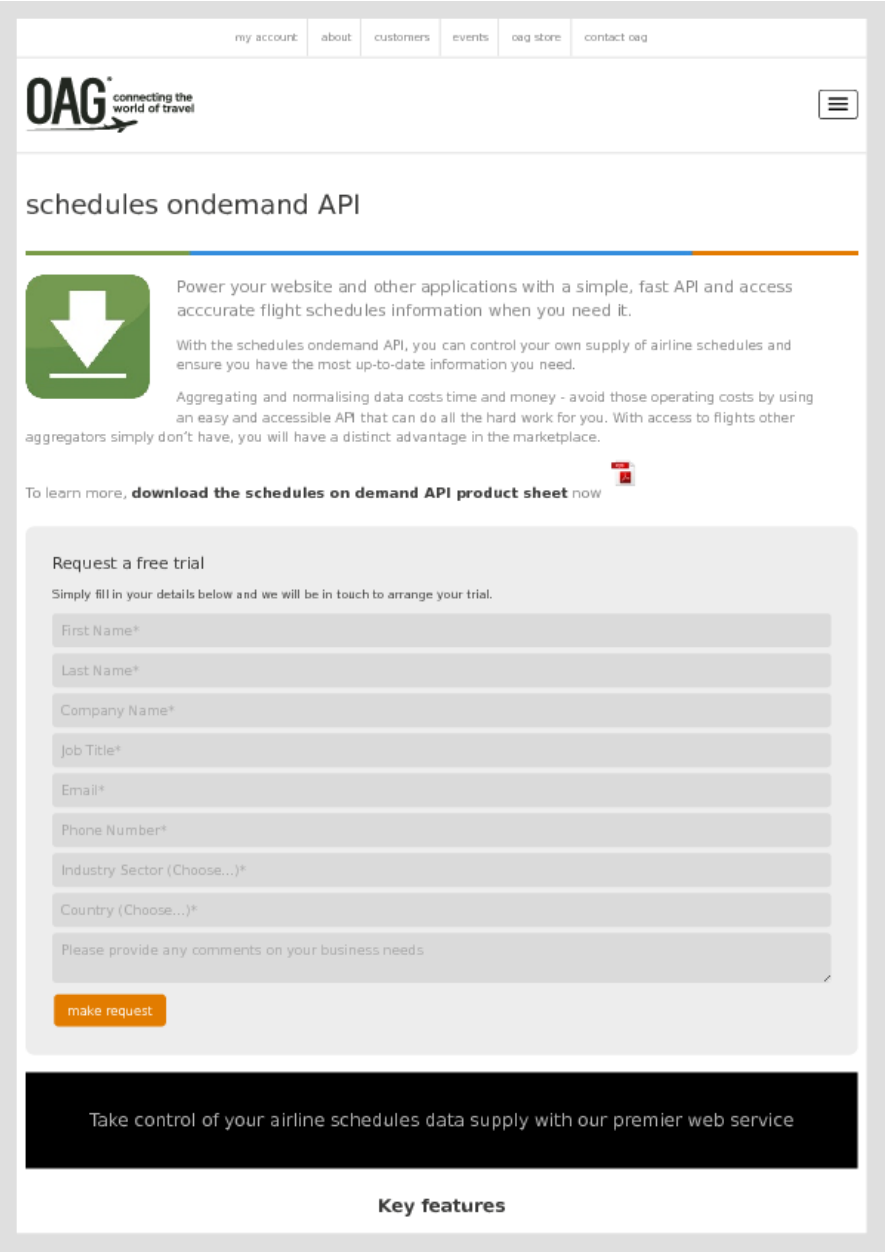


Figure 38: OAG (Official Airline Guide) home page.

5. QPX Express API: Global airline pricing and shopping in a single, standard API. (see Figure 39)
   `https://developers.google.com/qpx-express/`

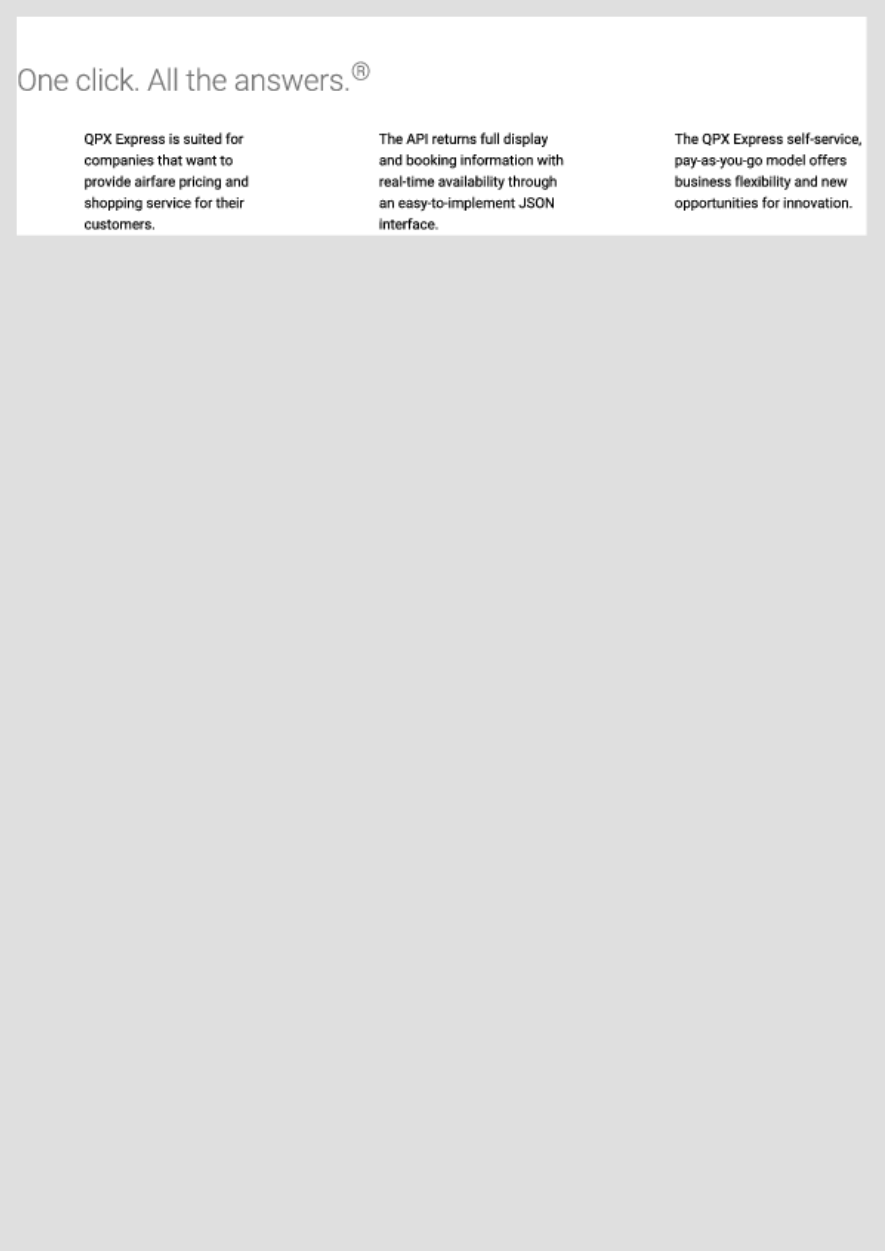


Figure 39: QPX Express API home page.

6. Travel Boutique: In recent years, the flight booking segment of the travel sector has become increasingly competitive. With the largest tour operators and websites offering customers a wide range of options all in one place, it's difficult for the smaller travel agents to compete. As ever, at Travel Boutique Online, we've developed a solution which allows small- and medium-sized travel agents in India to overcome this challenge. This solution will enable you to sell a vast inventory of flights from a large pool of airlines direct to your customers. (see Figure 40)
`http://www.travelboutiqueonline.com/Flight_api.aspx`

Figure 40: Travel Boutique home page.

7. Travel Fusion: TF.Flight API is the largest direct connect LCC and (a growing number of) FSC airline carriers platform available from a single API. TF.Flight API aggregates information from hundreds of airlines, providing agents, e-commerce platforms, search and mobile services with a single system to manage, search, book and support this vast content. TF.Flight API is available via an API, any Agency Desktop, and web-based (agency/user login) portals. (see Figure 41)
`https://www.travelfusion.com/corporate/page/tf-flight-api`

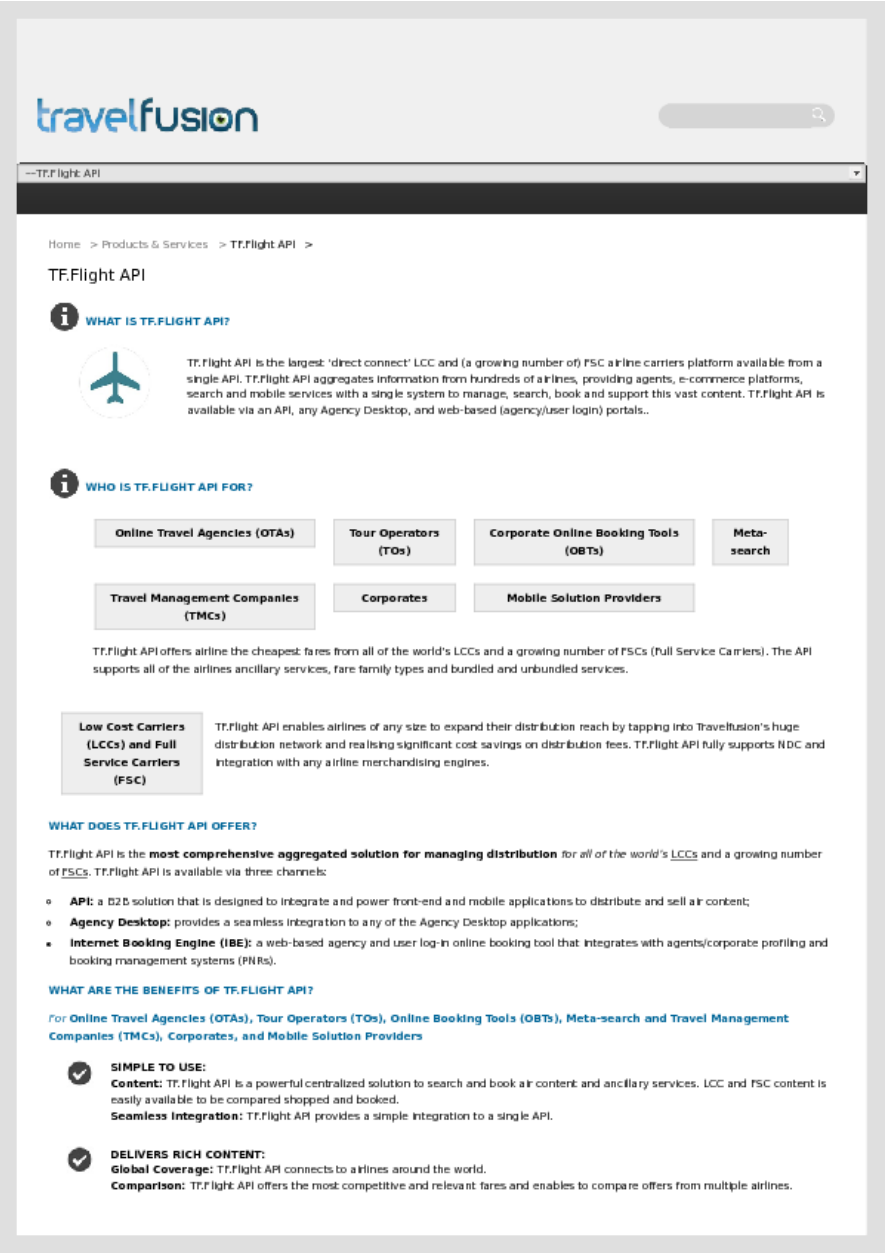


Figure 41: Travel Fusion home page.

8. Wego Flights: Wego Flights API allows clients to search for flight fares and availability in real-time across Wego's inventory of partners. The service is available via standard HTTP POST request and responses are made available in JSON format. (see Figure 42)
`http://support.wan.travel/hc/en-us/articles/200191669-Wego-Flights-API`

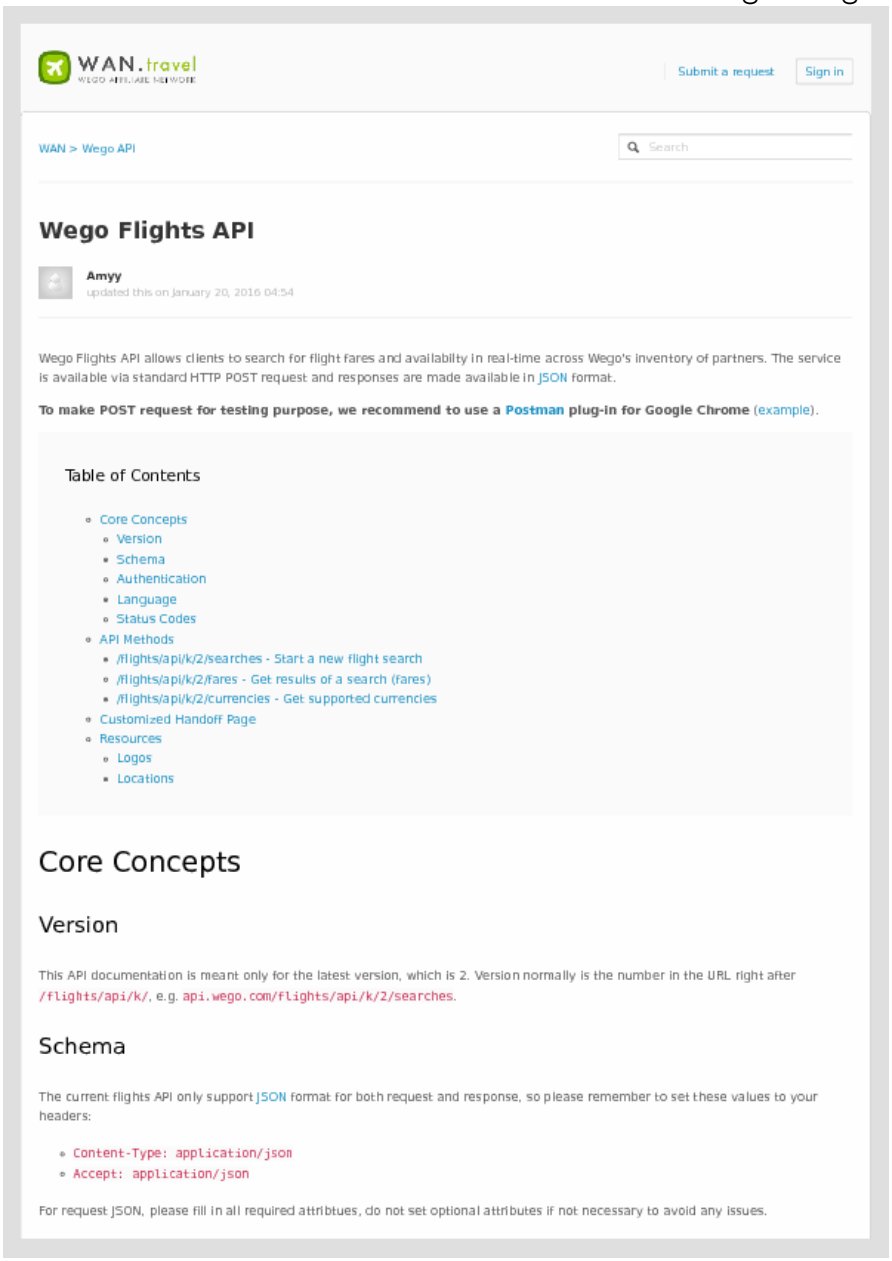


Figure 42: Wego Flights home page.

3.3 Developers

1. GeoCommons: GeoCommons is a community contributed collection of open data from around the world. Uploaded by the public, data are often from public and open government website and sources. (see Figure 43)
   `http://geocommons.com/`

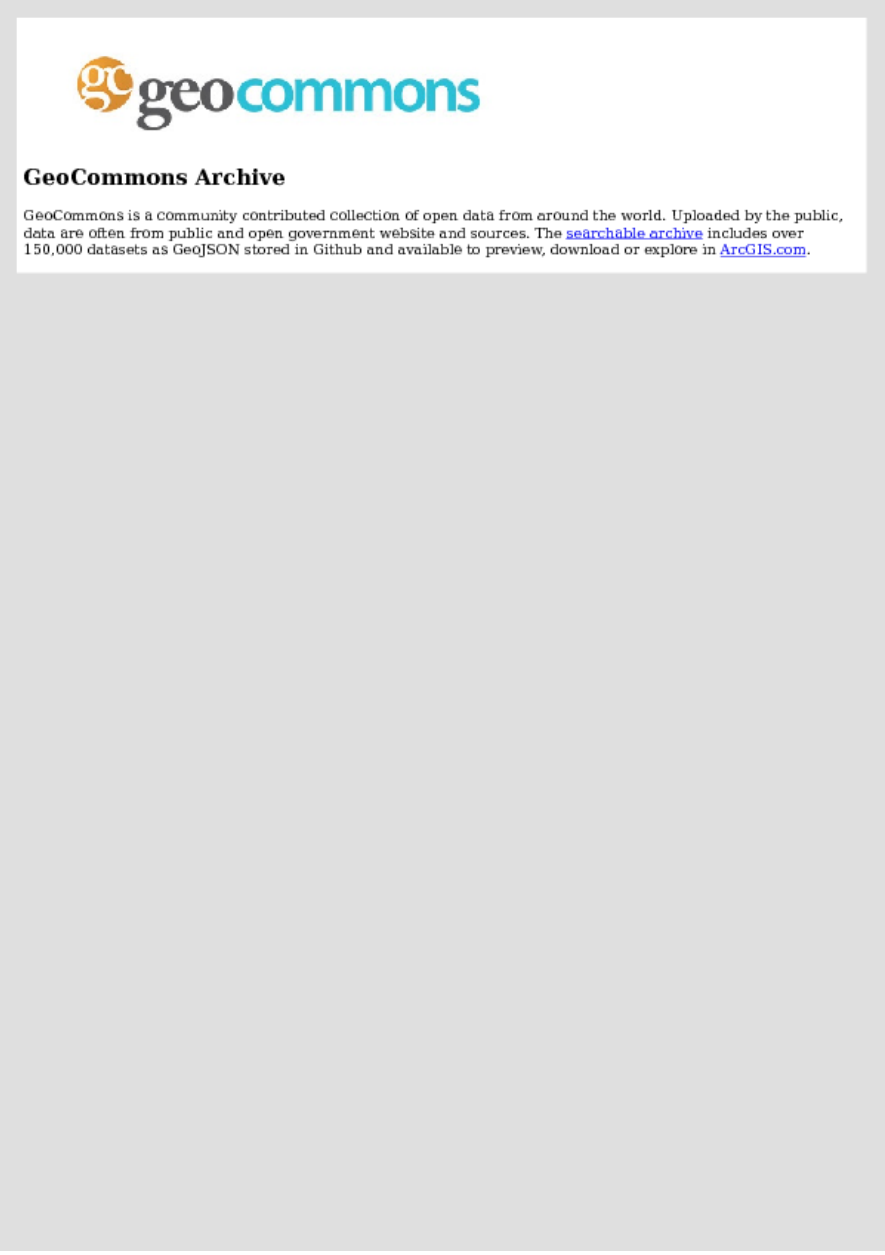


Figure 43: GeoCommons home page.

2. Global Data. Local Context: The best location data for Mobile Advertising, Developers, and Enterprise solutions. (see Figure 44)
`https://www.factual.com/`

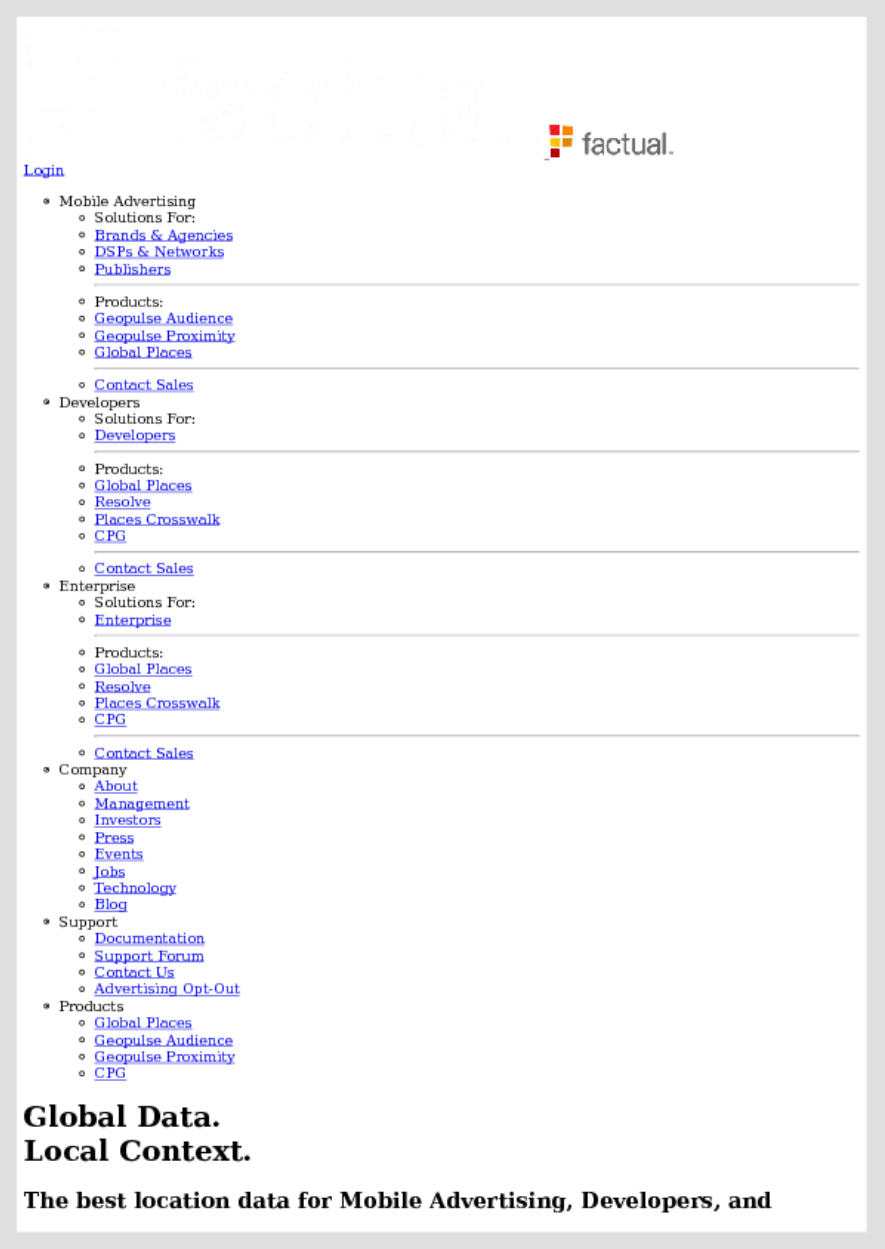


Figure 44: Global Data. Local Context home page.

3. OpenStreetMap: OpenStreetMap is a map of the world, created by people and free to use under an open license. (see Figure 45)
   `http://www.openstreetmap.org`

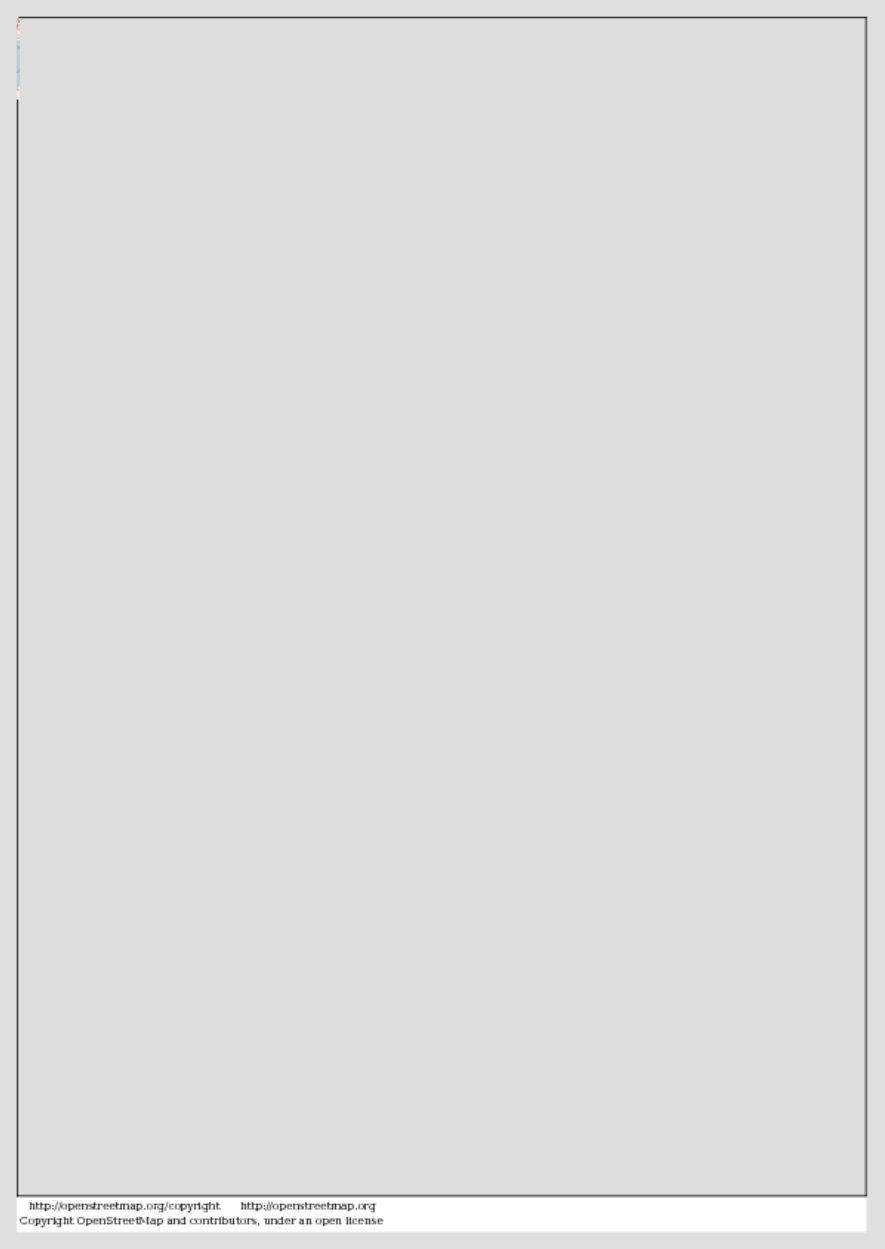


Figure 45: OpenStreetMap home page.

4. Quandl Financial and Economic Data: Quandl helps data analysts save time, effort and money by delivering high-quality financial and economic data in the precise format they need. (see Figure 46)
   `https://www.quandl.com/`

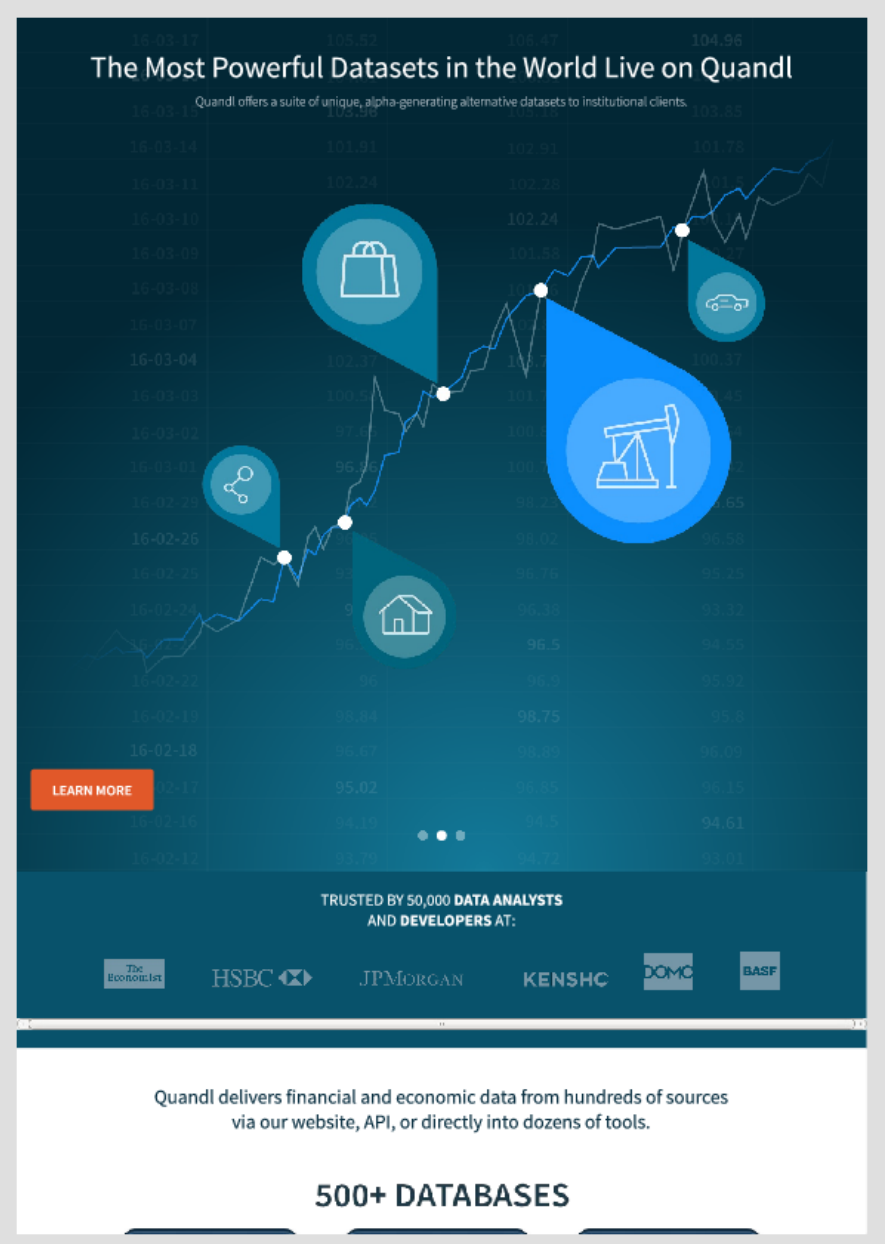
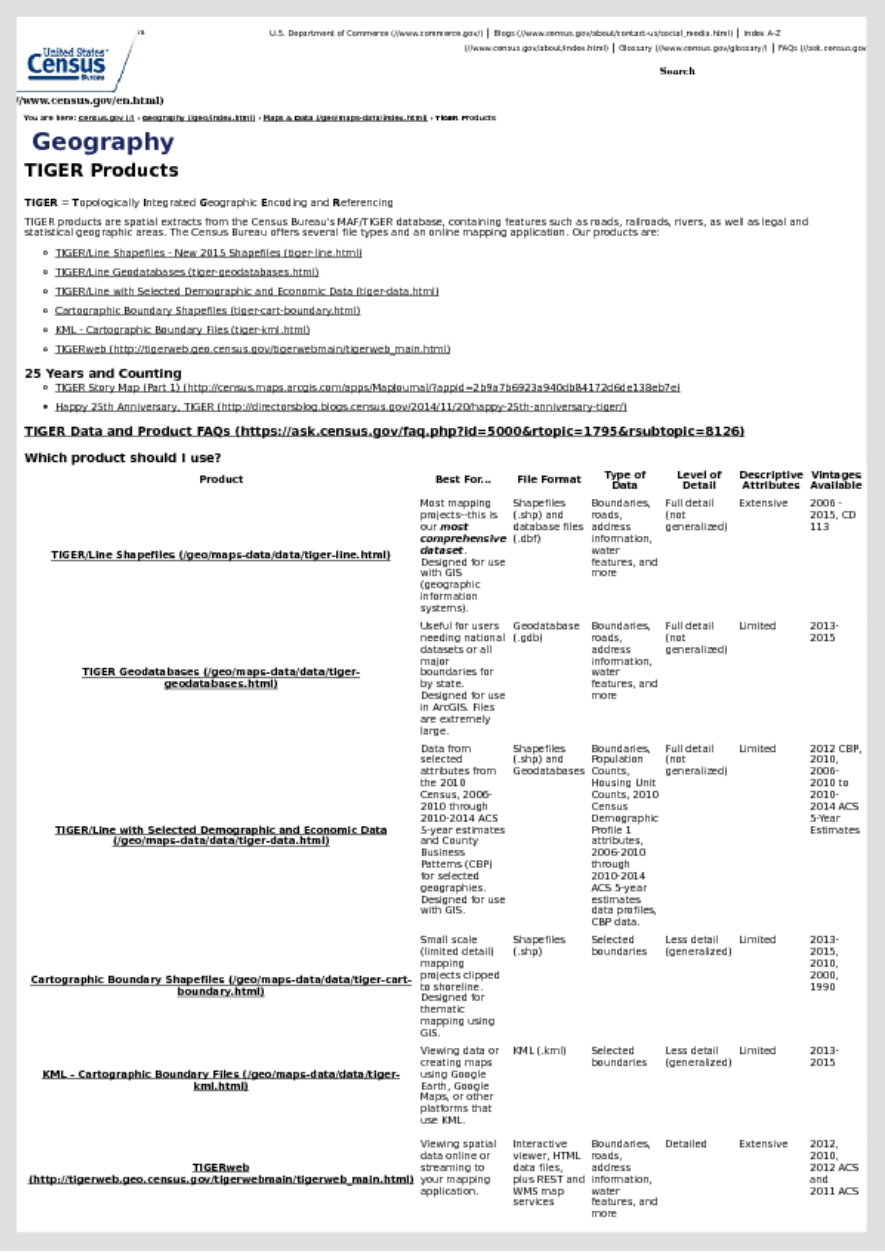


Figure 46: Quandl Financial and Economic Data home page.

5. TIGER Products: TIGER = Topologically Integrated Geographic Encoding and Referencing. TIGER products are spatial extracts from the Census Bureau MAF/TIGER database, containing features such as roads, railroads, rivers, as well as legal and statistical geographic areas. The Census Bureau offers several file types and an online mapping application. (see Figure 47)
`http://www.census.gov/geo/maps-data/data/tiger.html`



U.S. Department of Commerce (//www.commerce.gov/) | Blogs (//www.census.gov/about/contact-us/social_media.html) | Index A-Z (//www.census.gov/about/index.html) | Glossary (//www.census.gov/glossary/) | FAQs (//ask.census.gov

Search

(/www.census.gov/en.html)

You are here: census.gov (.) › geography (/geo/index.html) › Maps & data (/geo/maps-data/index.html) › TIGER Products

Geography

TIGER Products

TIGER = **T**opologically **I**ntegrated **G**eographic **E**ncoding and **R**eferencing

TIGER products are spatial extracts from the Census Bureau's MAF/TIGER database, containing features such as roads, railroads, rivers, as well as legal and statistical geographic areas. The Census Bureau offers several file types and an online mapping application. Our products are:

- TIGER/Line Shapefiles - New 2015 Shapefiles (tiger-line.html)
- TIGER/Line Geodatabases (tiger-geodatabases.html)
- TIGER/Line with Selected Demographic and Economic Data (tiger-data.html)
- Cartographic Boundary Shapefiles (tiger-cart-boundary.html)
- KML - Cartographic Boundary Files (tiger-kml.html)
- TIGERweb (http://tigerweb.geo.census.gov/tigerwebmain/tigerweb_main.html)

25 Years and Counting

- TIGER Story Map (Part 1) (http://census.maps.arcgis.com/apps/MapJournal/?appid=2b9a7b6923a940db84172d6de138eb7e)
- Happy 25th Anniversary, TIGER (http://directorsblog.blogs.census.gov/2014/11/20/happy-25th-anniversary-tiger/)

TIGER Data and Product FAQs (https://ask.census.gov/faq.php?id=5000&rtopic=1795&rsubtopic=8126)

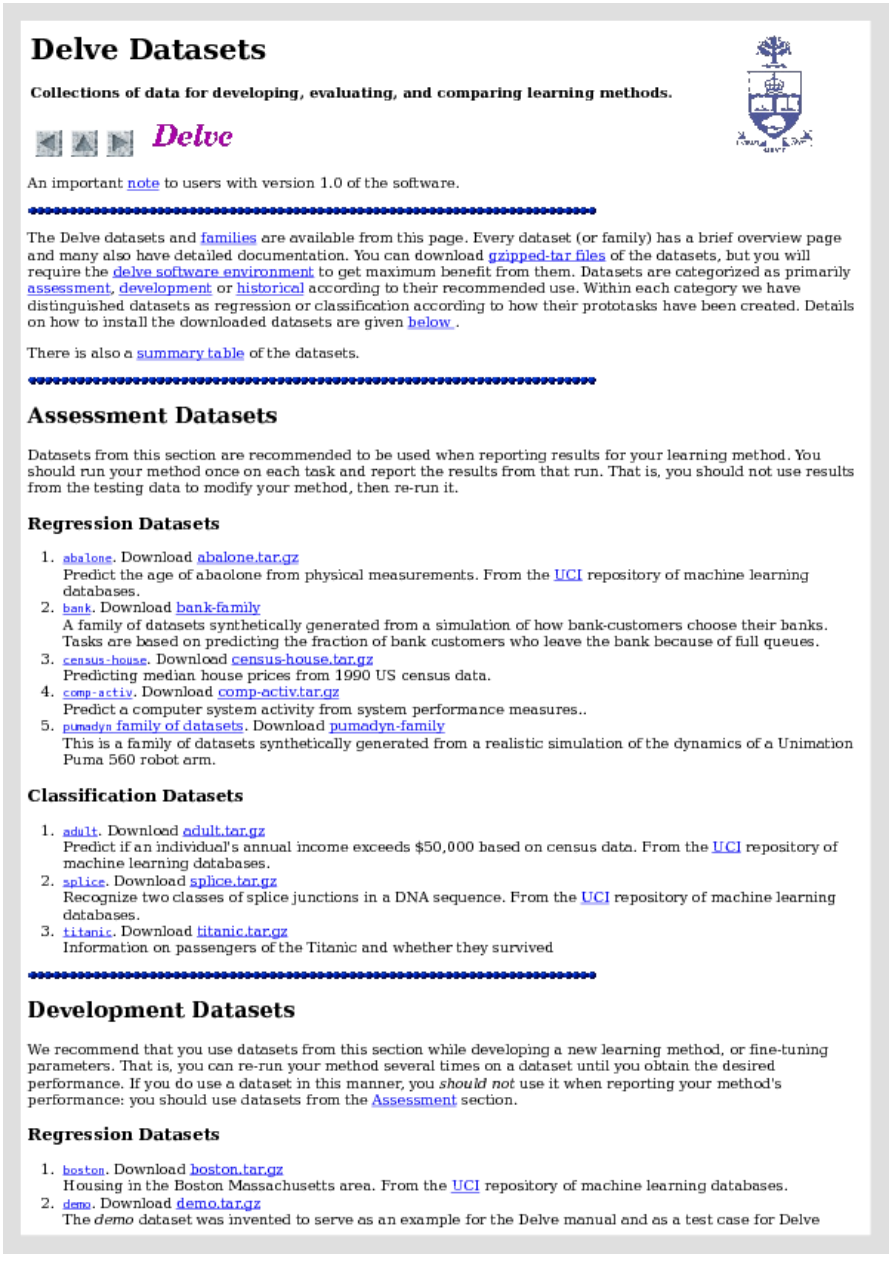
Which product should I use?

Product	Best For...	File Format	Type of Data	Level of Detail	Descriptive Attributes	Vintages Available
TIGER/Line Shapefiles (/geo/maps-data/data/tiger-line.html)	Most mapping projects--this is our ***most comprehensive dataset***. Designed for use with GIS (geographic information systems).	Shapefiles (.shp) and database files (.dbf)	Boundaries, roads, address information, water features, and more	Full detail (not generalized)	Extensive	2006 - 2015, CD 113
TIGER Geodatabases (/geo/maps-data/data/tiger-geodatabases.html)	Useful for users needing national datasets or all major boundaries for by state. Designed for use in ArcGIS. Files are extremely large.	Geodatabase (.gdb)	Boundaries, roads, address information, water features, and more	Full detail (not generalized)	Limited	2013-2015
TIGER/Line with Selected Demographic and Economic Data (/geo/maps-data/data/tiger-data.html)	Data from selected attributes from the 2010 Census, 2006-2010 through 2010-2014 ACS 5-year estimates and County Business Patterns (CBP) for selected geographies. Designed for use with GIS.	Shapefiles (.shp) and Geodatabases	Boundaries, Population Counts, Housing Unit Counts, 2010 Census Demographic Profile 1 attributes, 2006-2010 through 2010-2014 ACS 5-year estimates data profiles, CBP data.	Full detail (not generalized)	Limited	2012 CBP, 2010, 2006-2010 to 2010-2014 ACS 5-Year Estimates
Cartographic Boundary Shapefiles (/geo/maps-data/data/tiger-cart-boundary.html)	Small scale (limited detail) mapping projects clipped to shoreline. Designed for thematic mapping using GIS.	Shapefiles (.shp)	Selected boundaries	Less detail (generalized)	Limited	2013-2015, 2010, 2000, 1990
KML - Cartographic Boundary Files (/geo/maps-data/data/tiger-kml.html)	Viewing data or creating maps using Google Earth, Google Maps, or other platforms that use KML.	KML (.kml)	Selected boundaries	Less detail (generalized)	Limited	2013-2015
TIGERweb (http://tigerweb.geo.census.gov/tigerwebmain/tigerweb_main.html)	Viewing spatial data online or streaming to your mapping application.	Interactive viewer, HTML data files, plus REST and WMS map services	Boundaries, roads, address information, water features, and more	Detailed	Extensive	2012, 2010, 2012 ACS and 2011 ACS

Figure 47: TIGER Products home page.

3.4 Education

1. Delve Datasets: Collections of data for developing, evaluating, and comparing learning methods. (see Figure 48)
   `http://www.cs.toronto.edu/~delve/data/datasets.html`



Delve Datasets

Collections of data for developing, evaluating, and comparing learning methods.

Delve

An important note to users with version 1.0 of the software.

The Delve datasets and families are available from this page. Every dataset (or family) has a brief overview page and many also have detailed documentation. You can download gzipped-tar files of the datasets, but you will require the delve software environment to get maximum benefit from them. Datasets are categorized as primarily assessment, development or historical according to their recommended use. Within each category we have distinguished datasets as regression or classification according to how their prototasks have been created. Details on how to install the downloaded datasets are given below .

There is also a summary table of the datasets.

Assessment Datasets

Datasets from this section are recommended to be used when reporting results for your learning method. You should run your method once on each task and report the results from that run. That is, you should not use results from the testing data to modify your method, then re-run it.

Regression Datasets

1. abalone. Download abalone.tar.gz
   Predict the age of abalone from physical measurements. From the UCI repository of machine learning databases.
2. bank. Download bank-family
   A family of datasets synthetically generated from a simulation of how bank-customers choose their banks. Tasks are based on predicting the fraction of bank customers who leave the bank because of full queues.
3. census-house. Download census-house.tar.gz
   Predicting median house prices from 1990 US census data.
4. comp-activ. Download comp-activ.tar.gz
   Predict a computer system activity from system performance measures..
5. pumadyn family of datasets. Download pumadyn-family
   This is a family of datasets synthetically generated from a realistic simulation of the dynamics of a Unimation Puma 560 robot arm.

Classification Datasets

1. adult. Download adult.tar.gz
   Predict if an individual's annual income exceeds $50,000 based on census data. From the UCI repository of machine learning databases.
2. splice. Download splice.tar.gz
   Recognize two classes of splice junctions in a DNA sequence. From the UCI repository of machine learning databases.
3. titanic. Download titanic.tar.gz
   Information on passengers of the Titanic and whether they survived

Development Datasets

We recommend that you use datasets from this section while developing a new learning method, or fine-tuning parameters. That is, you can re-run your method several times on a dataset until you obtain the desired performance. If you do use a dataset in this manner, you *should not* use it when reporting your method's performance: you should use datasets from the Assessment section.

Regression Datasets

1. boston. Download boston.tar.gz
   Housing in the Boston Massachusetts area. From the UCI repository of machine learning databases.
2. demo. Download demo.tar.gz
   The *demo* dataset was invented to serve as an example for the Delve manual and as a test case for Delve

Figure 48: Delve Datasets home page.

2. Statistics Online Computational Resource (SOCR): The Statistics Online Computational Resource (SOCR)[1] is an online multi-institutional research and education organization. SOCR designs, validates and broadly shares a suite of online tools for statistical computing, and interactive materials for hands-on learning and teaching concepts in data science, statistical analysis and probability theory. (see Figure 49)
`http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data`

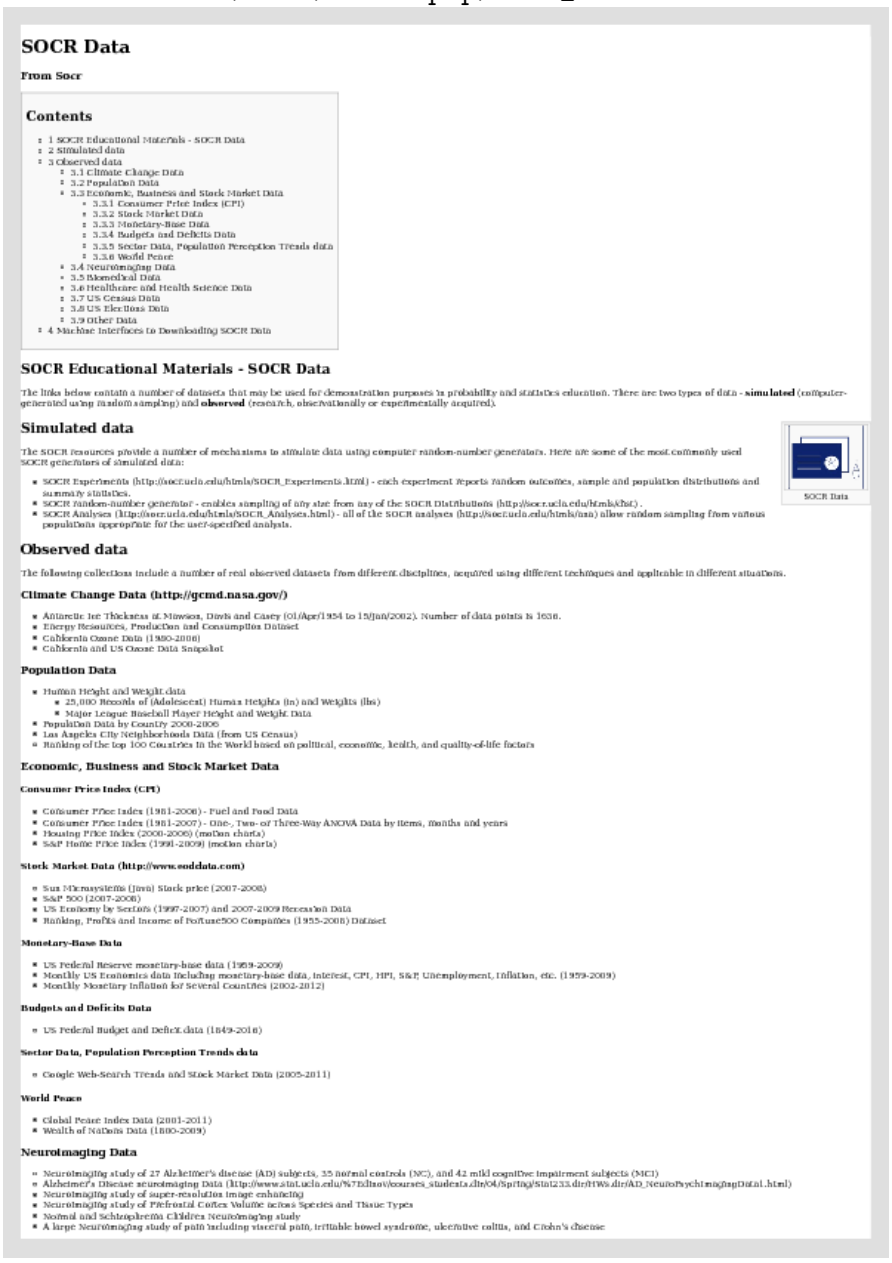


Figure 49: Statistics Online Computational Resource (SOCR) home page.

3.5 General

1. Bing: Bing (known previously as Live Search, Windows Live Search, and MSN Search) is a web search engine (advertised as a "decision engine") from Microsoft. (see Figure 50)
   `http://www.bing.com/`

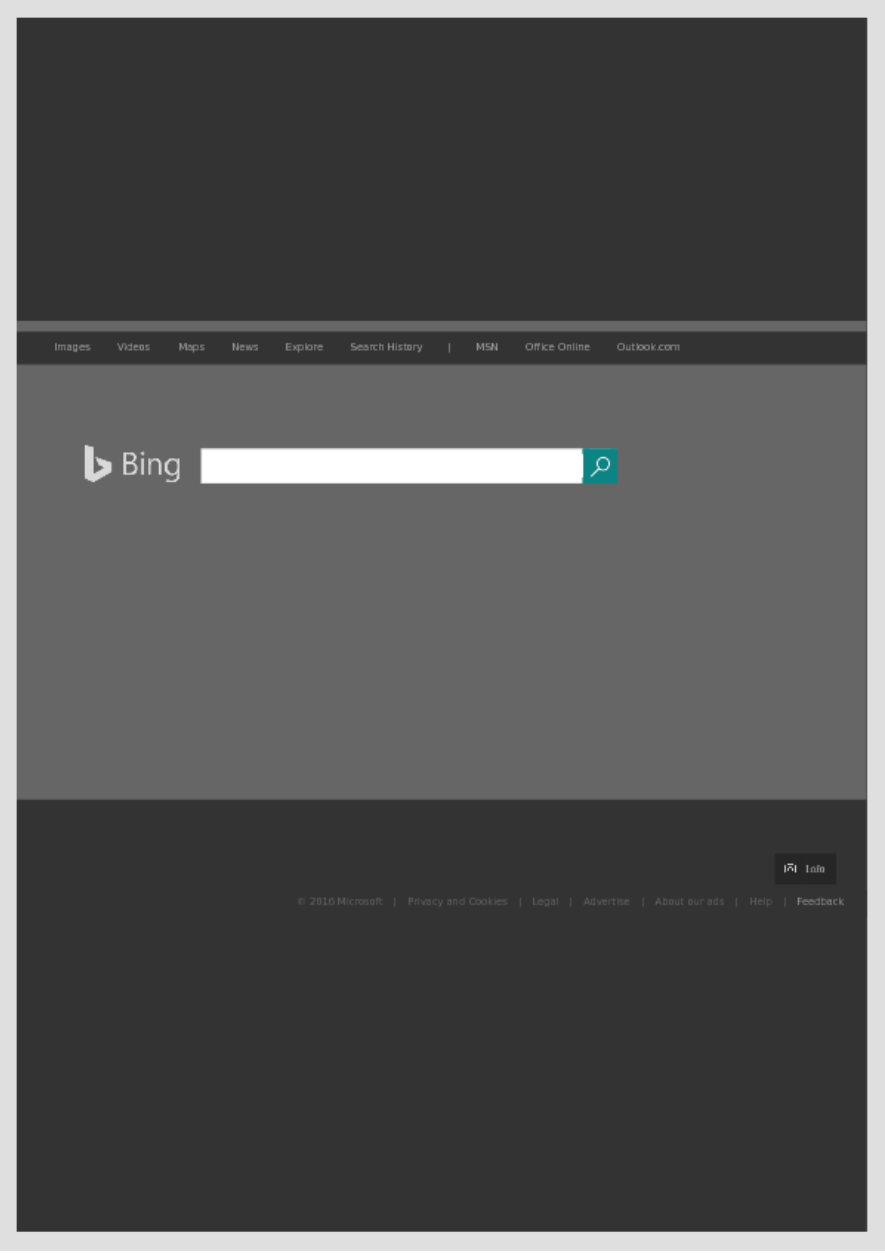


Figure 50: Bing home page.

2. Blog - AriLamstein.com: Combining Big Data and R (see Figure 51)
`http://www.arilamstein.com/blog/`

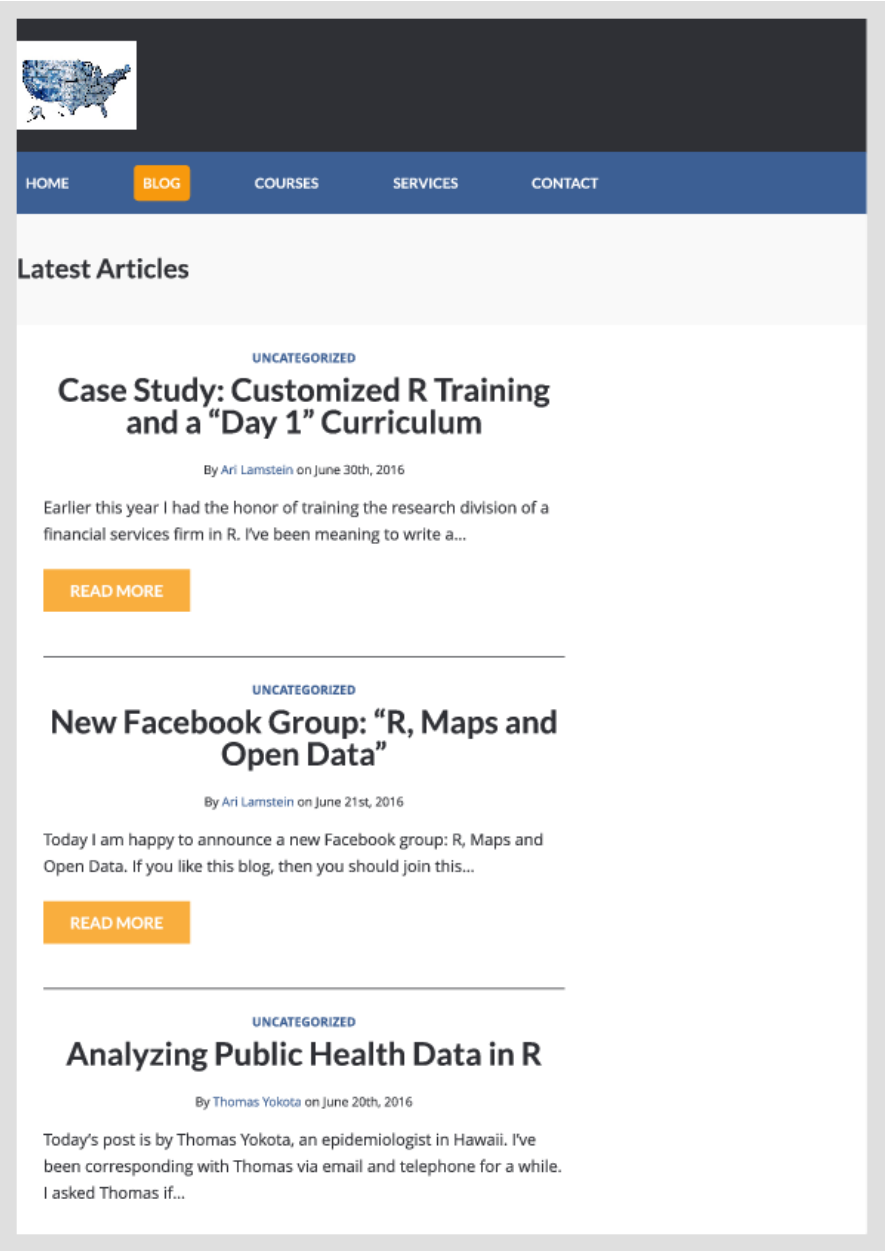


Figure 51: Blog - AriLamstein.com home page.

3. Dogpile: Dogpile is a search engine that fetches results from Google, Yahoo! and Yandex, and includes results from several other popular search engines, including those from audio and video content providers. It is a registered trademark of Blucora, Inc.. (see Figure 52)
`http://www.dogpile.com/`

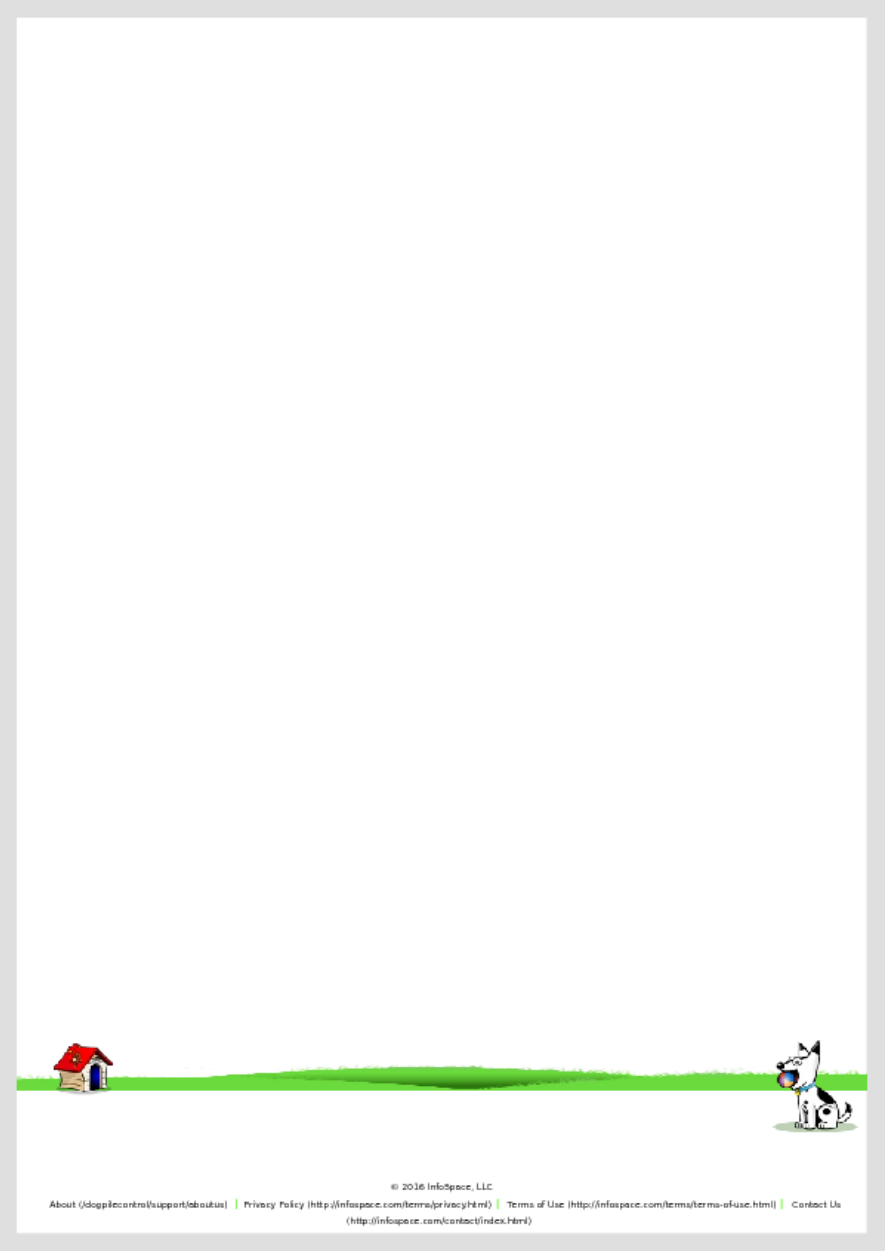


Figure 52: Dogpile home page.

4. DuckDuckGo: DuckDuckGo is an Internet search engine that emphasizes protecting searchers privacy and avoiding the filter bubble of personalized search results.[1] DuckDuckGo distinguishes itself from other search engines by not profiling its users and by deliberately showing all users the same search results for a given search term. DuckDuckGo emphasizes getting information from the best sources rather than the most sources, generating its search results from key crowd-sourced sites such as Wikipedia and from partnerships with other search engines like Yandex, Yahoo!, Bing, and Yummly. (see Figure 53)
`https://duckduckgo.com/`

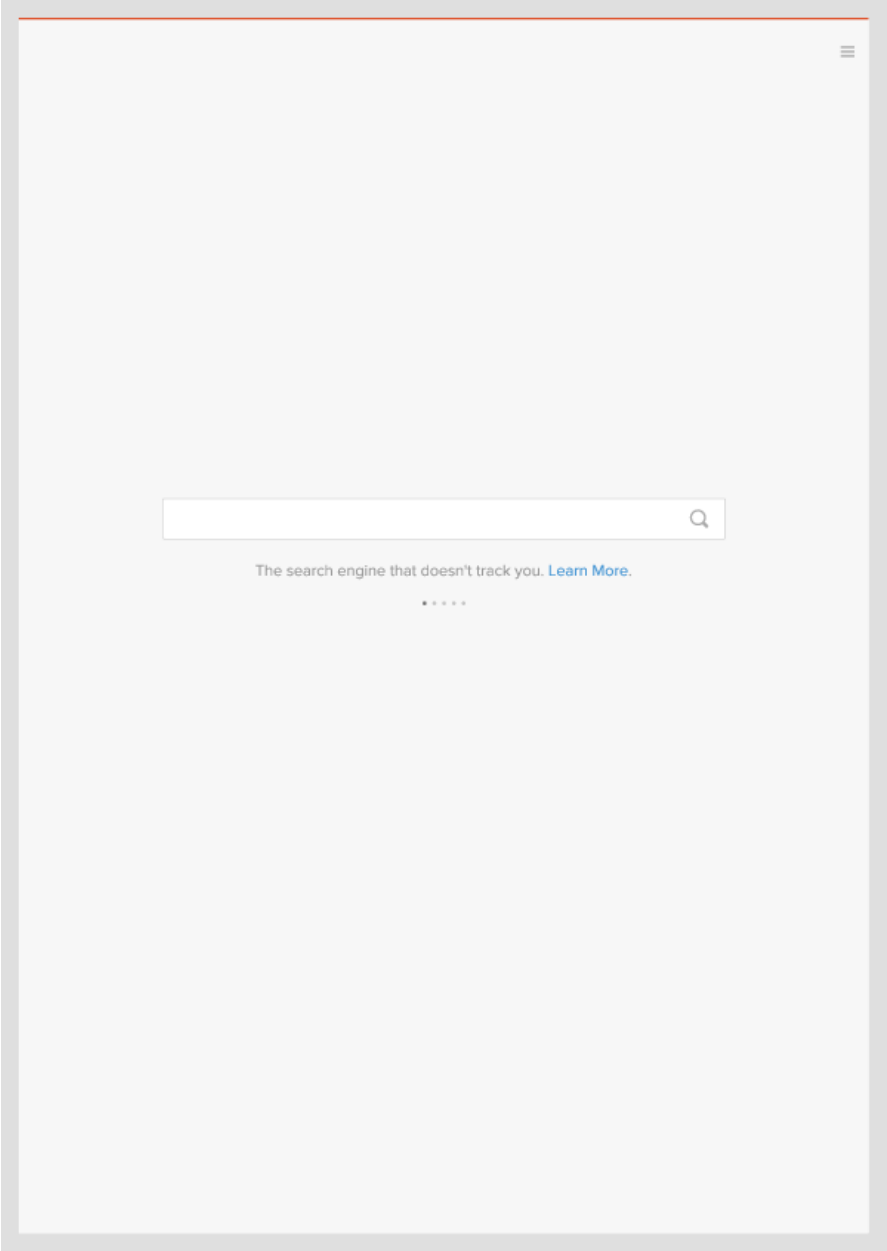


Figure 53: DuckDuckGo home page.

5. Google: Google Inc. is an American multinational technology company specializing in Internet-related services and products. These include online advertising technologies, search, cloud computing, and software. Most of its profits are derived from AdWords, an online advertising service that places advertising near the list of search results. (see Figure 54)
`http://www.google.com`

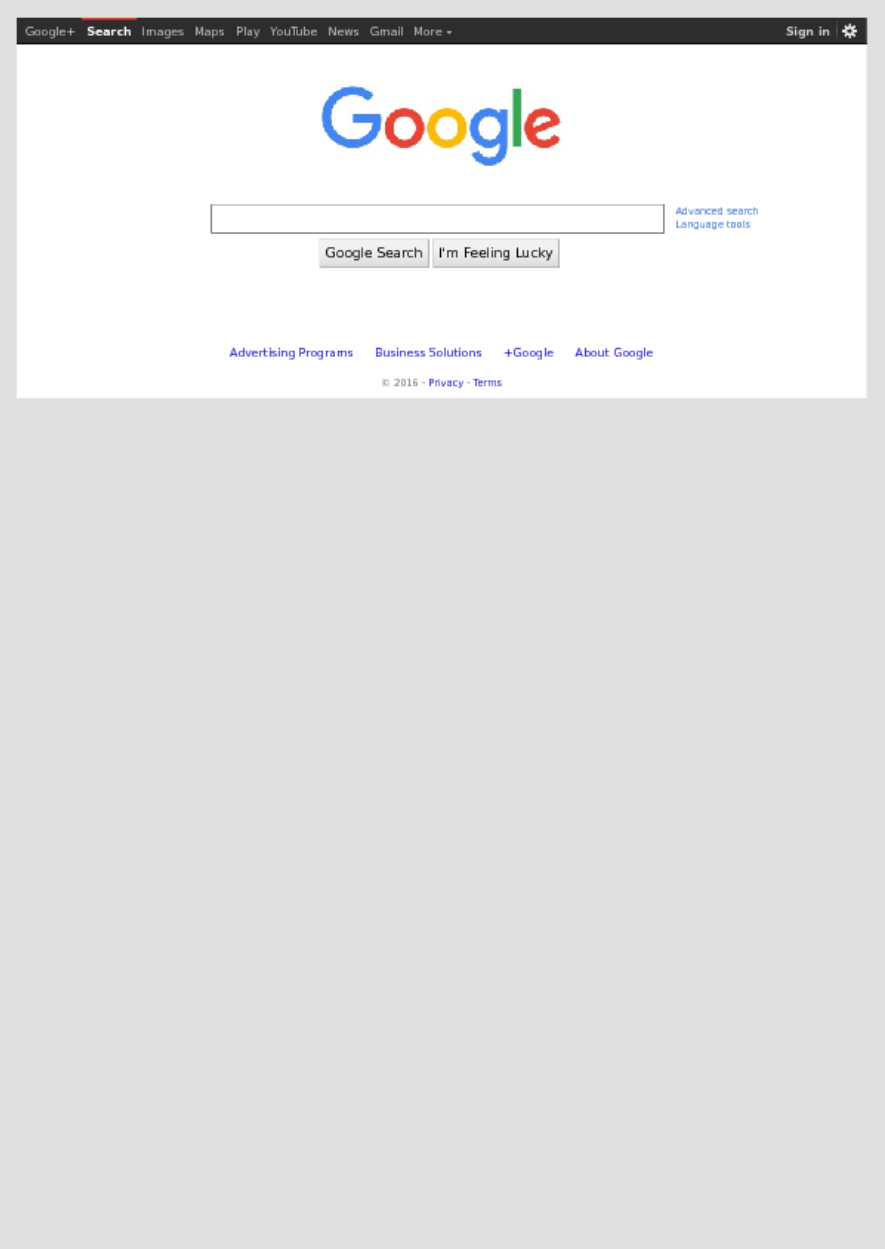


Figure 54: Google home page.

6. Our World in Data: Explore the ongoing history of human civilization at the broadest level, through research and data visualization. (see Figure 55)
`https://ourworldindata.org/`

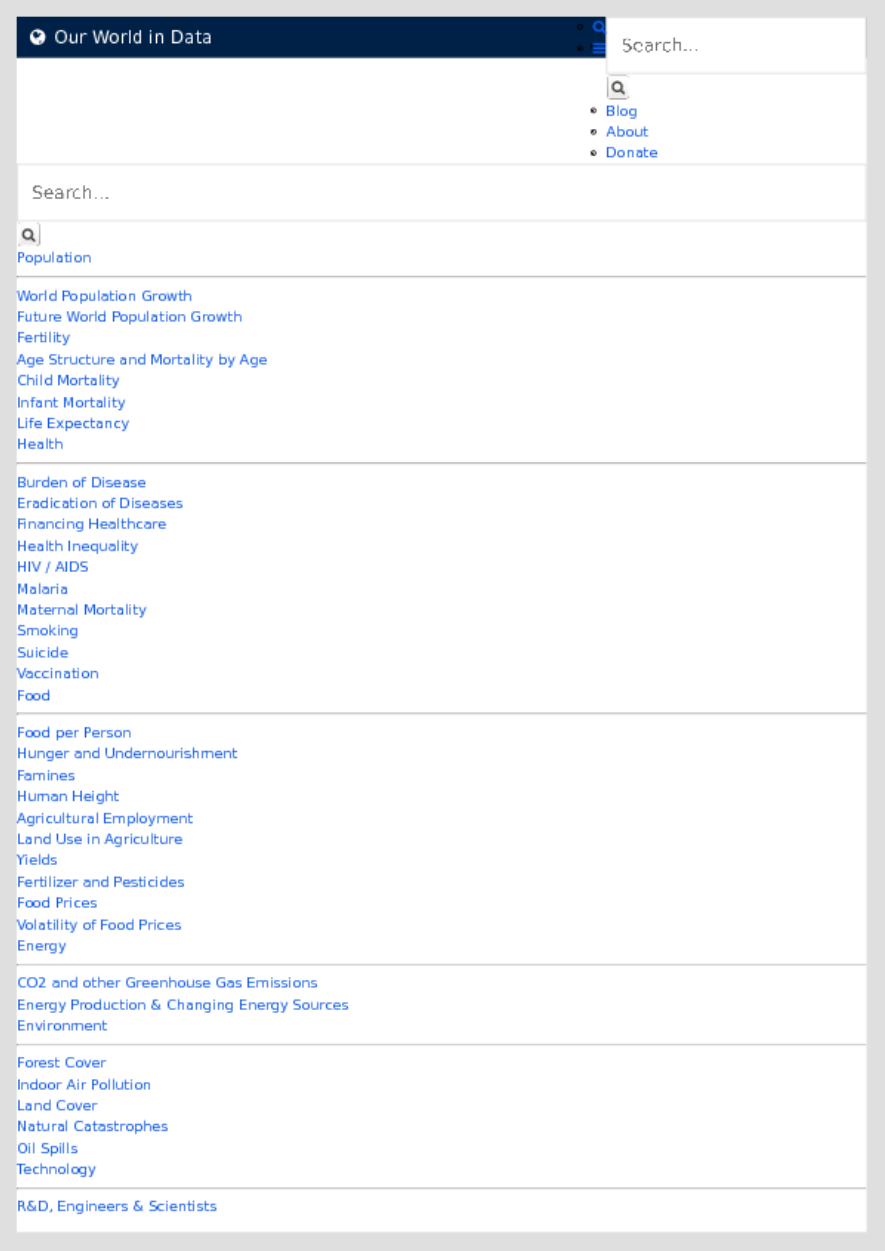


Figure 55: Our World in Data home page.

7. Stanford Large Network Dataset Collection: The SNAP library is being actively developed since 2004 and is organically growing as a result of our research pursuits in analysis of large social and information networks. Largest network we analyzed so far using the library was the Microsoft Instant Messenger network from 2006 with 240 million nodes and 1.3 billion edges. (see Figure 56)
`http://snap.stanford.edu/data/index.html`

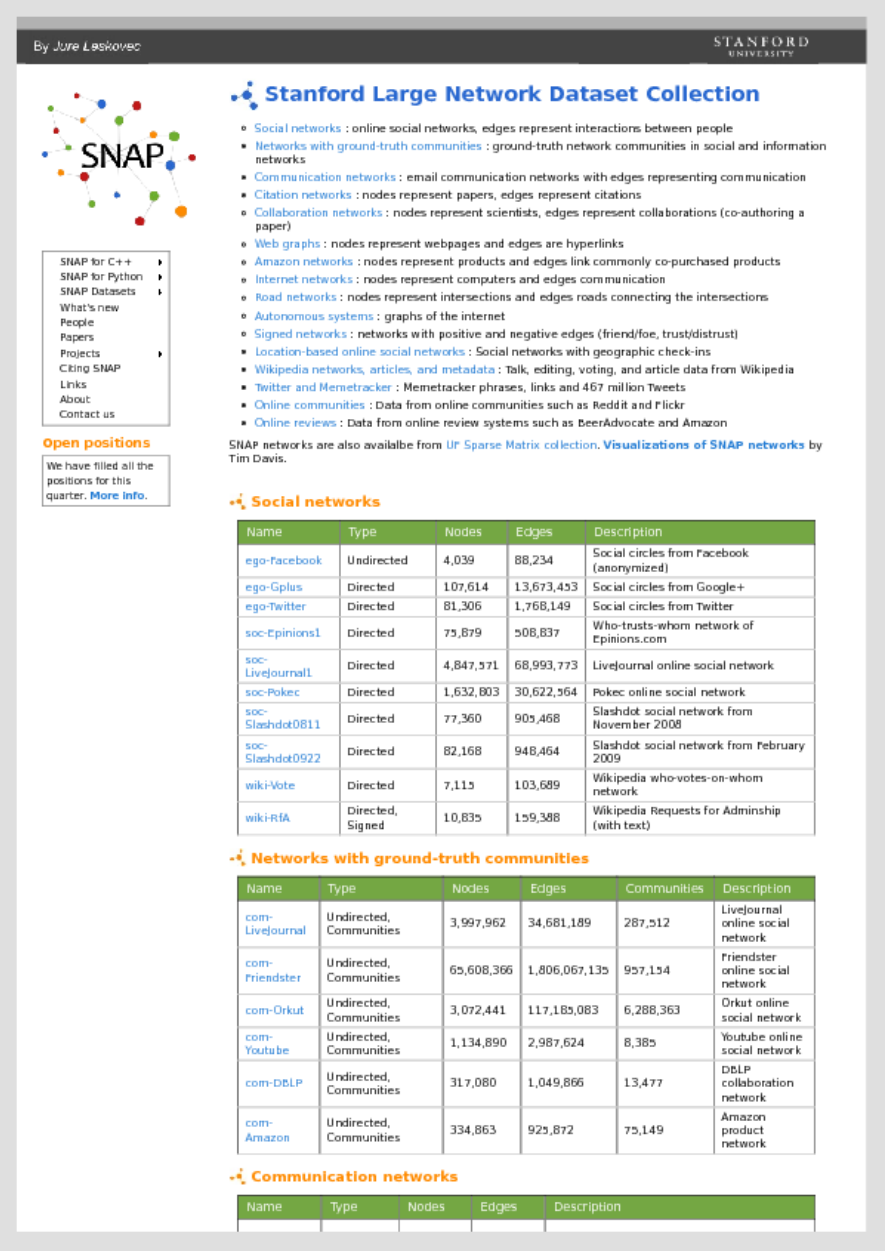


Figure 56: Stanford Large Network Dataset Collection home page.

8. UCI Machine Learning Repository: All sorts of datasets. (see Figure 57)
`https://archive.ics.uci.edu/ml/datasets.html`

Figure 57: UCI Machine Learning Repository home page.

9. WolframAlpha: A specialized data and mathematical search engine. (see Figure 58) `http://www.wolframalpha.com/`

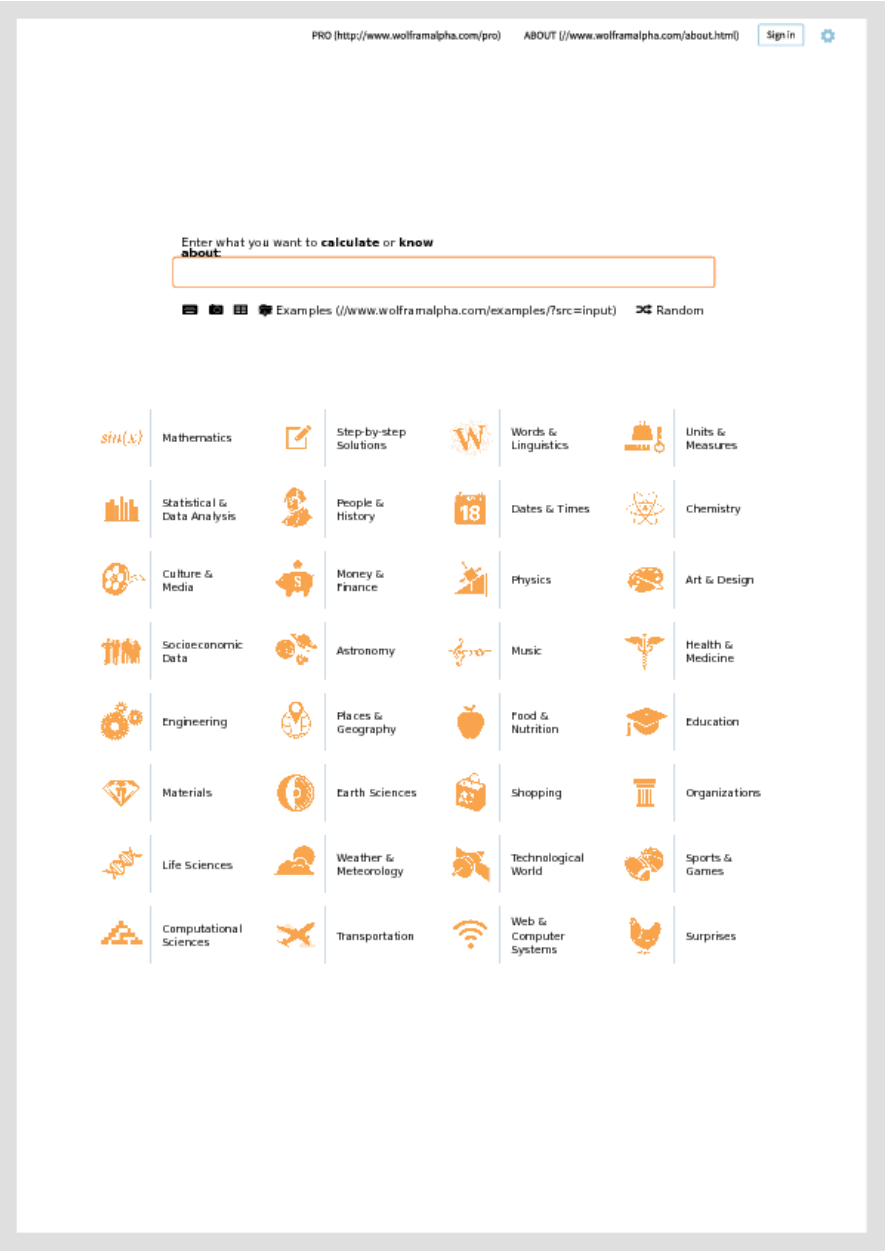


Figure 58: WolframAlpha home page.

10. Yahoo: Yahoo Inc. (styled as Yahoo!) is an American multinational technology company headquartered in Sunnyvale, California. It is globally known for its Web portal, search engine Yahoo! Search, and related services, including Yahoo! Directory, Yahoo! Mail, Yahoo! News, Yahoo! Finance, Yahoo! Groups, Yahoo! Answers, advertising, online mapping, video sharing, fantasy sports and its social media website. (see Figure 59)
`https://www.yahoo.com/`

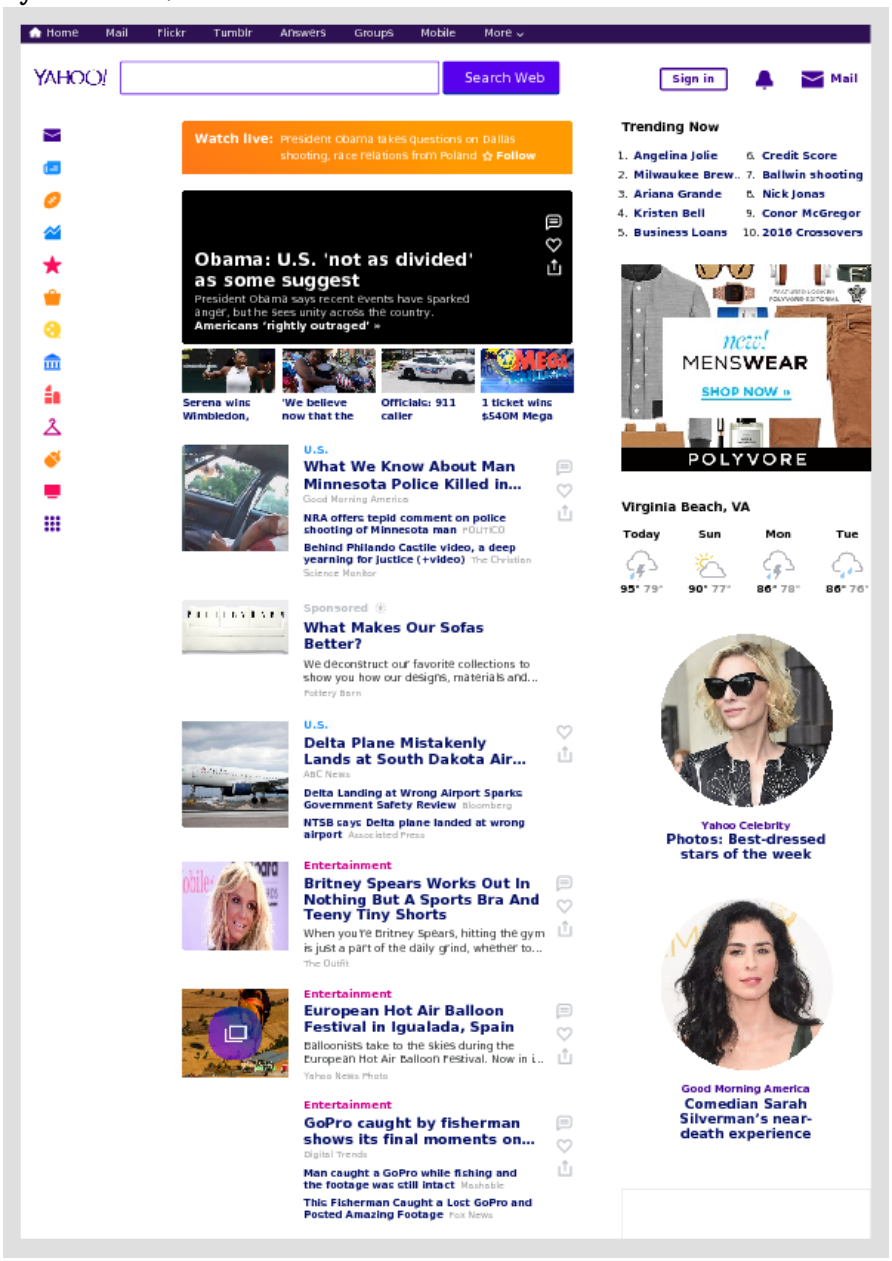


Figure 59: Yahoo home page.

3.6 Geographic information

1. GeoNames: The GeoNames geographical database covers all countries and contains over eight million place names that are available for download free of charge. (see Figure 60)
   `http://download.geonames.org/export/dump/`

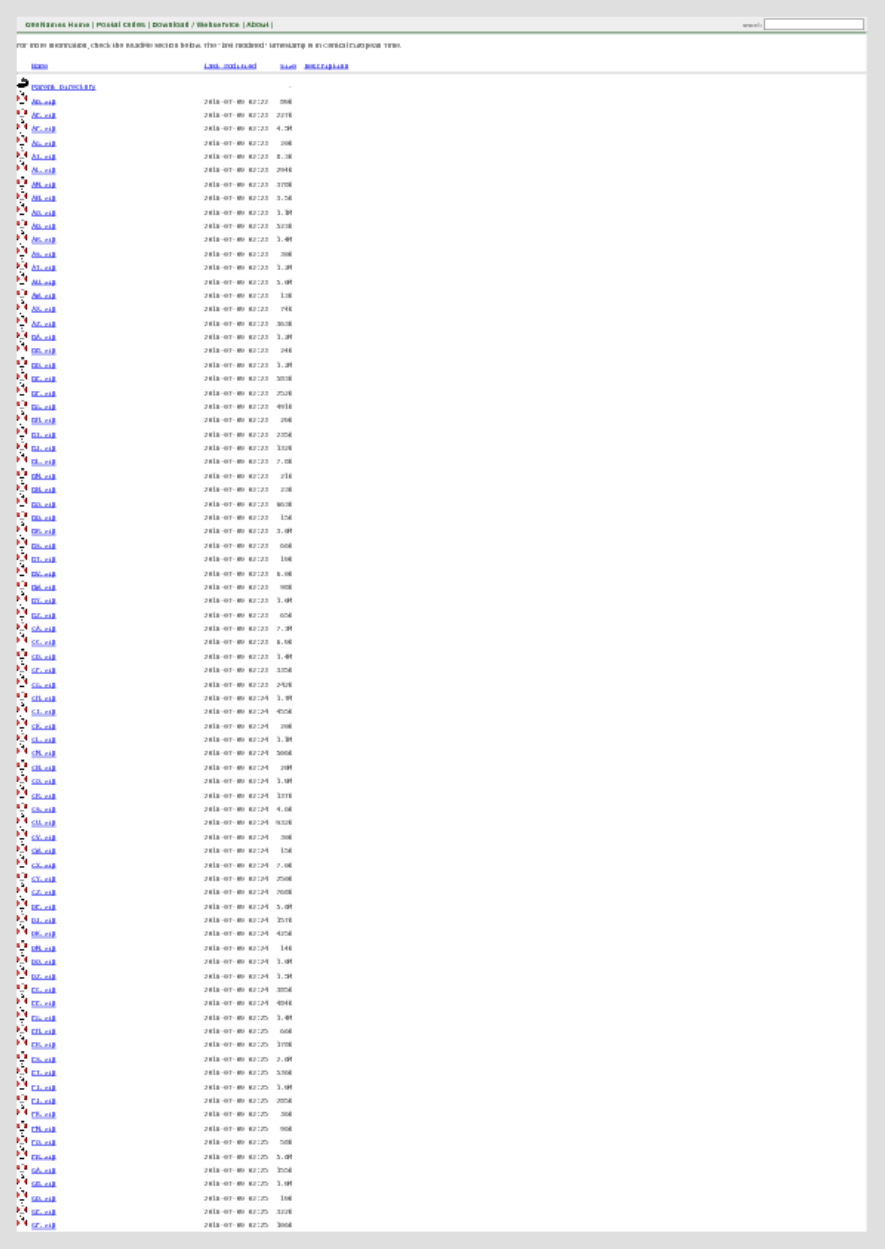


Figure 60: GeoNames home page.

2. Open Source Data and APIs — MaxMind: Add geographic information to an application using their datasets, and APIs. (see Figure 61)
`https://www.maxmind.com/en/open-source-data-and-api-for-ip-geolocation`

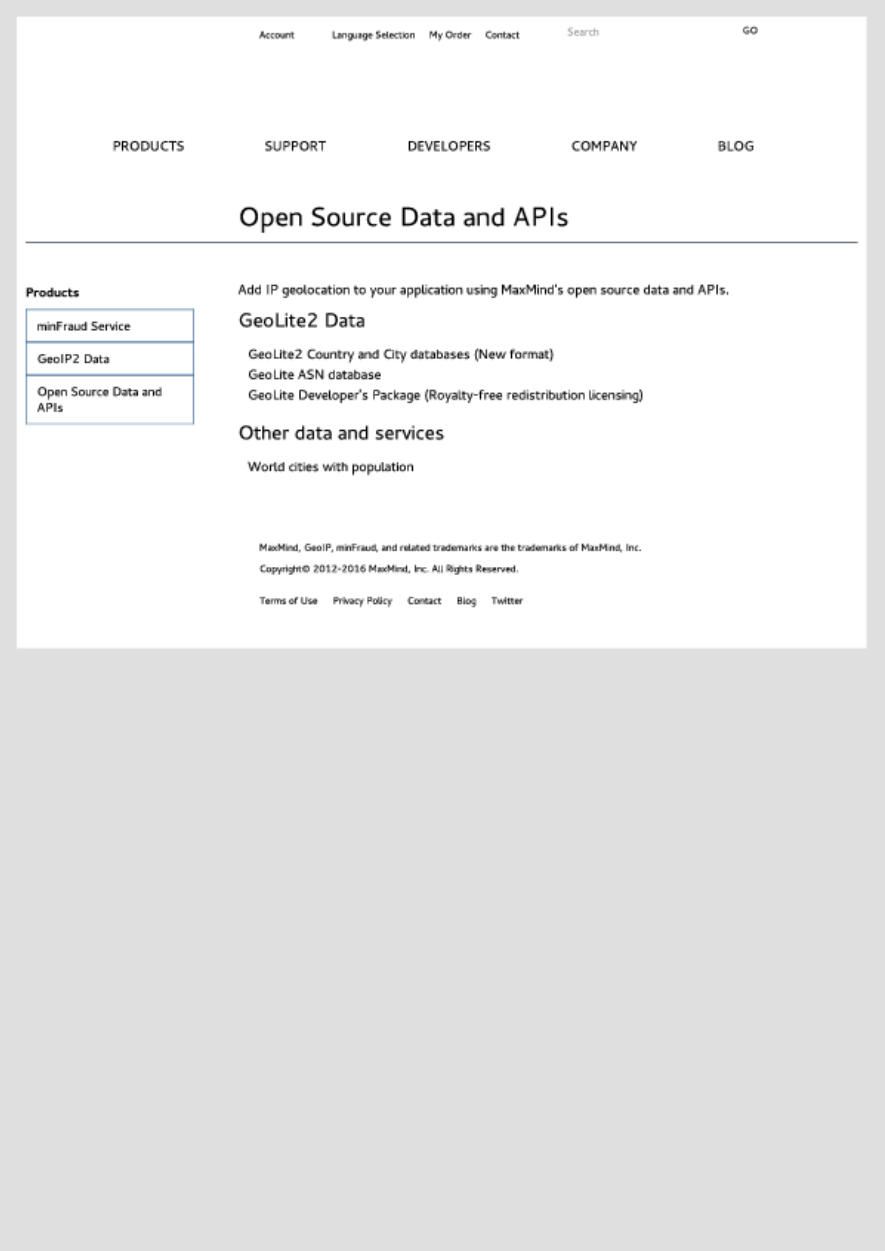
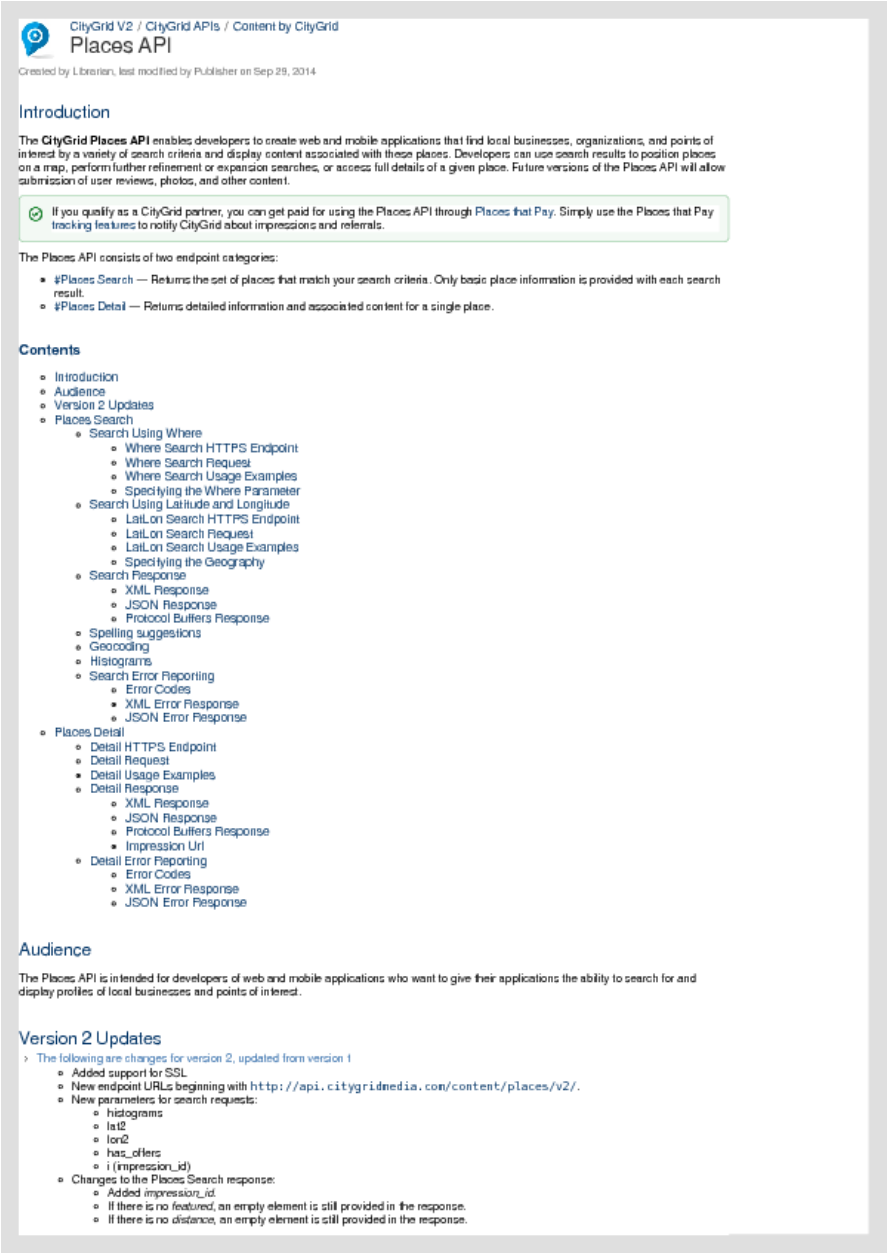


Figure 61: Open Source Data and APIs — MaxMind home page.

3. Places API - CityGrid V2 - CityGrid Media Documentation: The CityGrid Places API enables developers to create web and mobile applications that find local businesses, organizations, and points of interest by a variety of search criteria and display content associated with these places. Developers can use search results to position places on a map, perform further refinement or expansion searches, or access full details of a given place. Future versions of the Places API will allow submission of user reviews, photos, and other content. (see Figure 62)
`http://docs.citygridmedia.com/display/citygridv2/Places+API#PlacesAPI-SearchWhereNotes`



CityGrid V2 / CityGrid APIs / Content by CityGrid

Places API

Created by Librarian, last modified by Publisher on Sep 29, 2014

Introduction

The **CityGrid Places API** enables developers to create web and mobile applications that find local businesses, organizations, and points of interest by a variety of search criteria and display content associated with these places. Developers can use search results to position places on a map, perform further refinement or expansion searches, or access full details of a given place. Future versions of the Places API will allow submission of user reviews, photos, and other content.

If you qualify as a CityGrid partner, you can get paid for using the Places API through Places that Pay. Simply use the Places that Pay tracking features to notify CityGrid about impressions and referrals.

The Places API consists of two endpoint categories:

- #Places Search — Returns the set of places that match your search criteria. Only basic place information is provided with each search result.
- #Places Detail — Returns detailed information and associated content for a single place.

Contents

- Introduction
- Audience
- Version 2 Updates
- Places Search
  - Search Using Where
    - Where Search HTTPS Endpoint
    - Where Search Request
    - Where Search Usage Examples
    - Specifying the Where Parameter
  - Search Using Latitude and Longitude
    - LatLon Search HTTPS Endpoint
    - LatLon Search Request
    - LatLon Search Usage Examples
    - Specifying the Geography
  - Search Response
    - XML Response
    - JSON Response
    - Protocol Buffers Response
  - Spelling suggestions
  - Geocoding
  - Histograms
  - Search Error Reporting
    - Error Codes
    - XML Error Response
    - JSON Error Response
- Places Detail
  - Detail HTTPS Endpoint
  - Detail Request
  - Detail Usage Examples
  - Detail Response
    - XML Response
    - JSON Response
    - Protocol Buffers Response
    - Impression Url
  - Detail Error Reporting
    - Error Codes
    - XML Error Response
    - JSON Error Response

Audience

The Places API is intended for developers of web and mobile applications who want to give their applications the ability to search for and display profiles of local businesses and points of interest.

Version 2 Updates

› The following are changes for version 2, updated from version 1

- Added support for SSL
- New endpoint URLs beginning with `http://api.citygridmedia.com/content/places/v2/`.
- New parameters for search requests:
  - histograms
  - lat2
  - lon2
  - has_offers
  - i (impression_id)
- Changes to the Places Search response:
  - Added *impression_id*.
  - If there is no *featured*, an empty element is still provided in the response.
  - If there is no *distance*, an empty element is still provided in the response.

Figure 62: Places API - CityGrid V2 - CityGrid Media Documentation home page.

3.7 Geographic location

1. MaxMind Open Source Data and APIs: Add IP geolocation to your application using MaxMind's open source data and APIs. (see Figure 63)
   `https://www.maxmind.com/en/open-source-data-and-api-for-ip-geolocation`

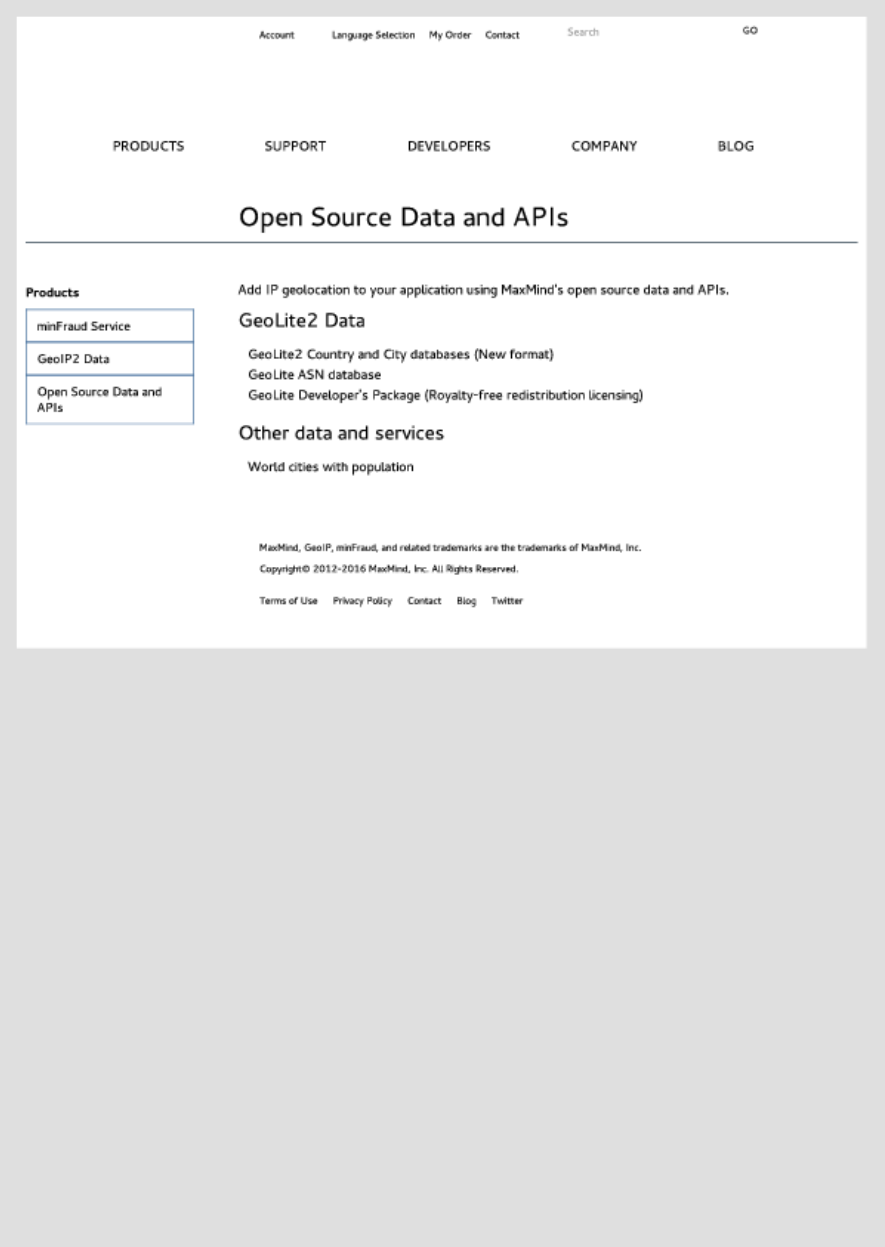


Figure 63: MaxMind Open Source Data and APIs home page.

3.8 Government

1. Austin Texas Data Portal: The Data Portal is your open government resource. Data portal benefits include: transparency, user-friendly data presentation and opportunities for community app development. (see Figure 64)
   `https://data.austintexas.gov/`

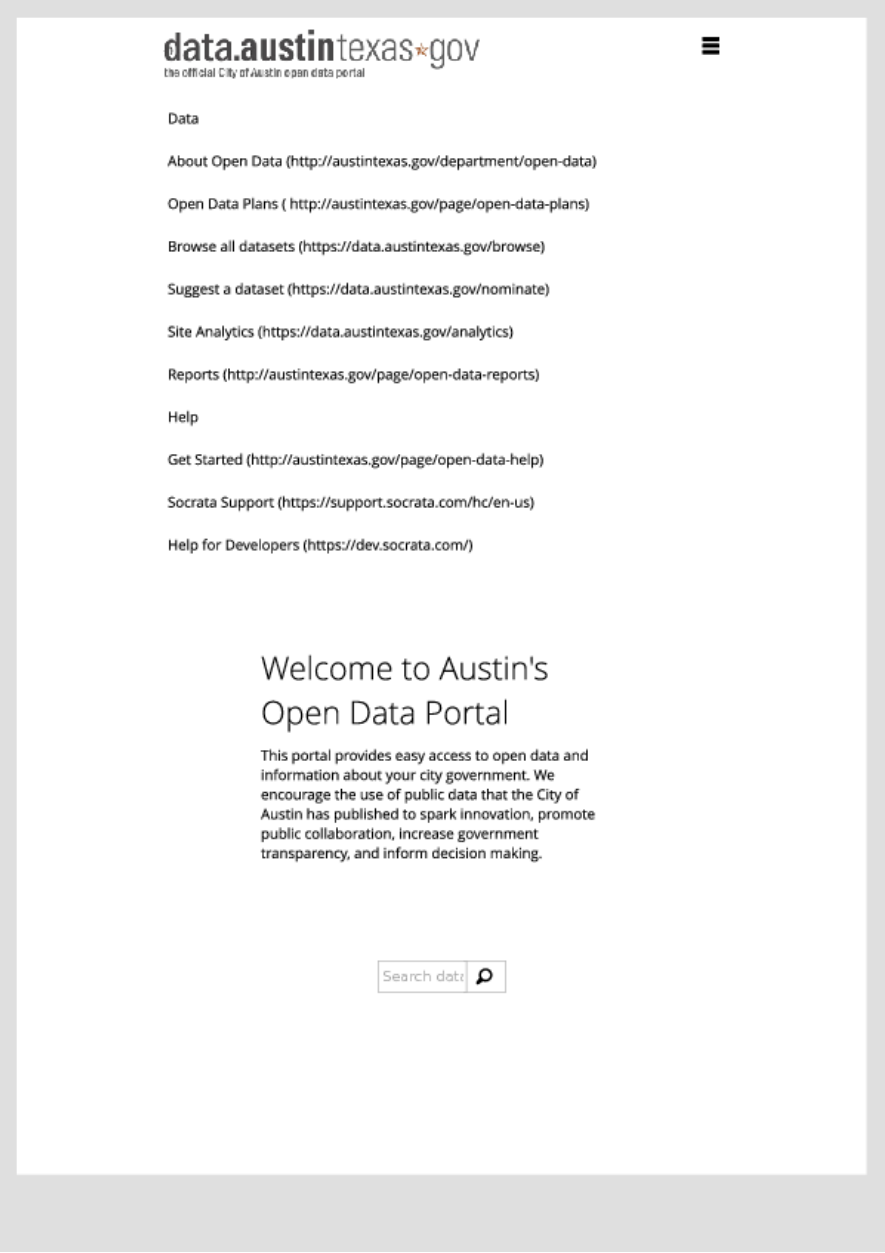


Figure 64: Austin Texas Data Portal home page.

2. Bureau of Economic Analysis: The Bureau of Economic Analysis (BEA) promotes a better understanding of the U.S. economy by providing the most timely, relevant, and accurate economic accounts data in an objective and cost-effective manner. (see Figure 65)
`http://www.bea.gov/index.htm`

Figure 65: Bureau of Economic Analysis home page.

3. Bureau of Labor Statistics: The Bureau of Labor Statistics of the U.S. Department of Labor is the principal Federal agency responsible for measuring labor market activity, working conditions, and price changes in the economy. Its mission is to collect, analyze, and disseminate essential economic information to support public and private decision-making. As an independent statistical agency, BLS serves its diverse user communities by providing products and services that are objective, timely, accurate, and relevant. (see Figure 66)
`http://www.bls.gov/`

Figure 66: Bureau of Labor Statistics home page.

4. Census Bureau Economic Statistics: Historical and current information about various industries. (see Figure 67)
   `https://www.census.gov/econ/`

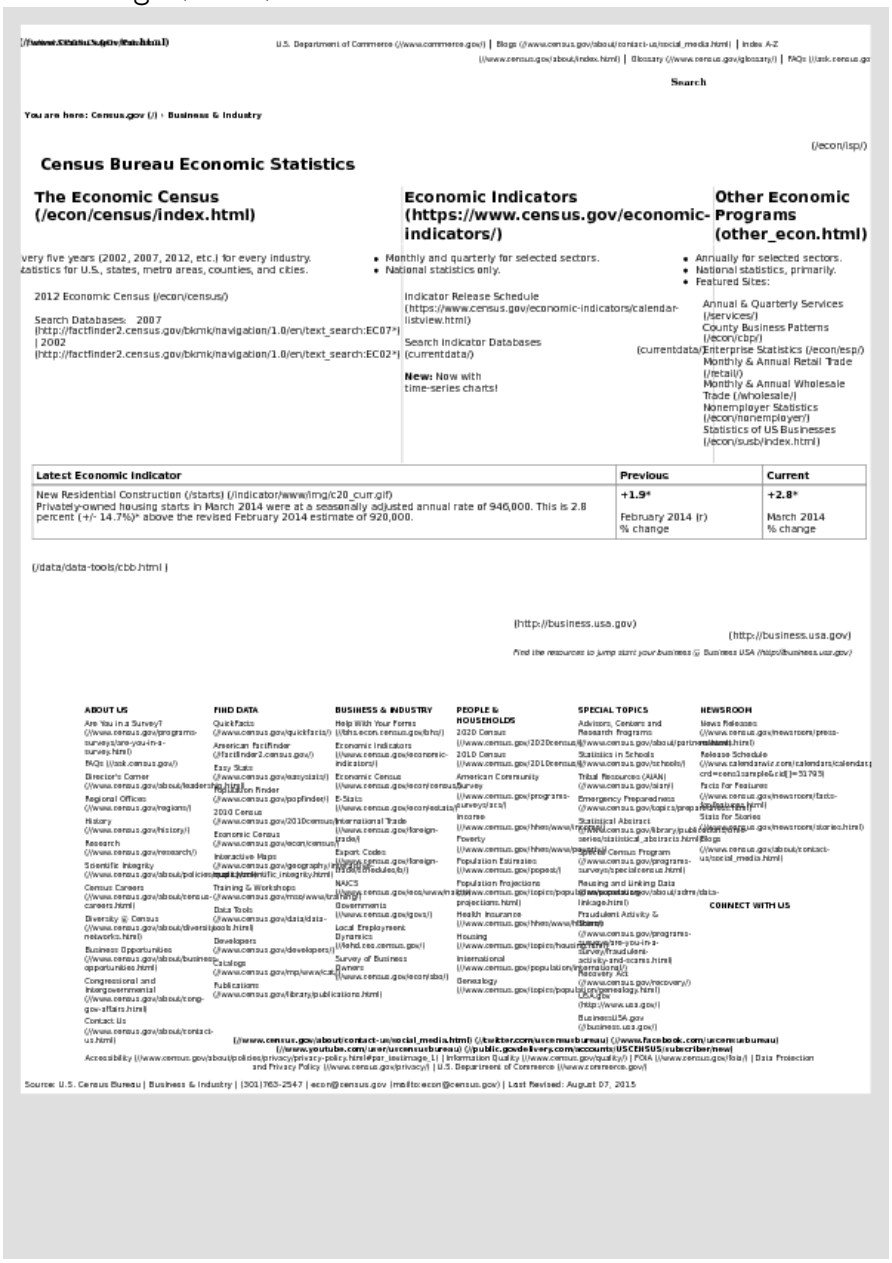


Figure 67: Census Bureau Economic Statistics home page.

5. City Maps - Robert Wood Johnson Foundation: A discussion about how your life expectancy can be estimated by your subway stop. (see Figure 68)
`http://www.rwjf.org/en/library/articles-and-news/2015/09/city-maps.html`

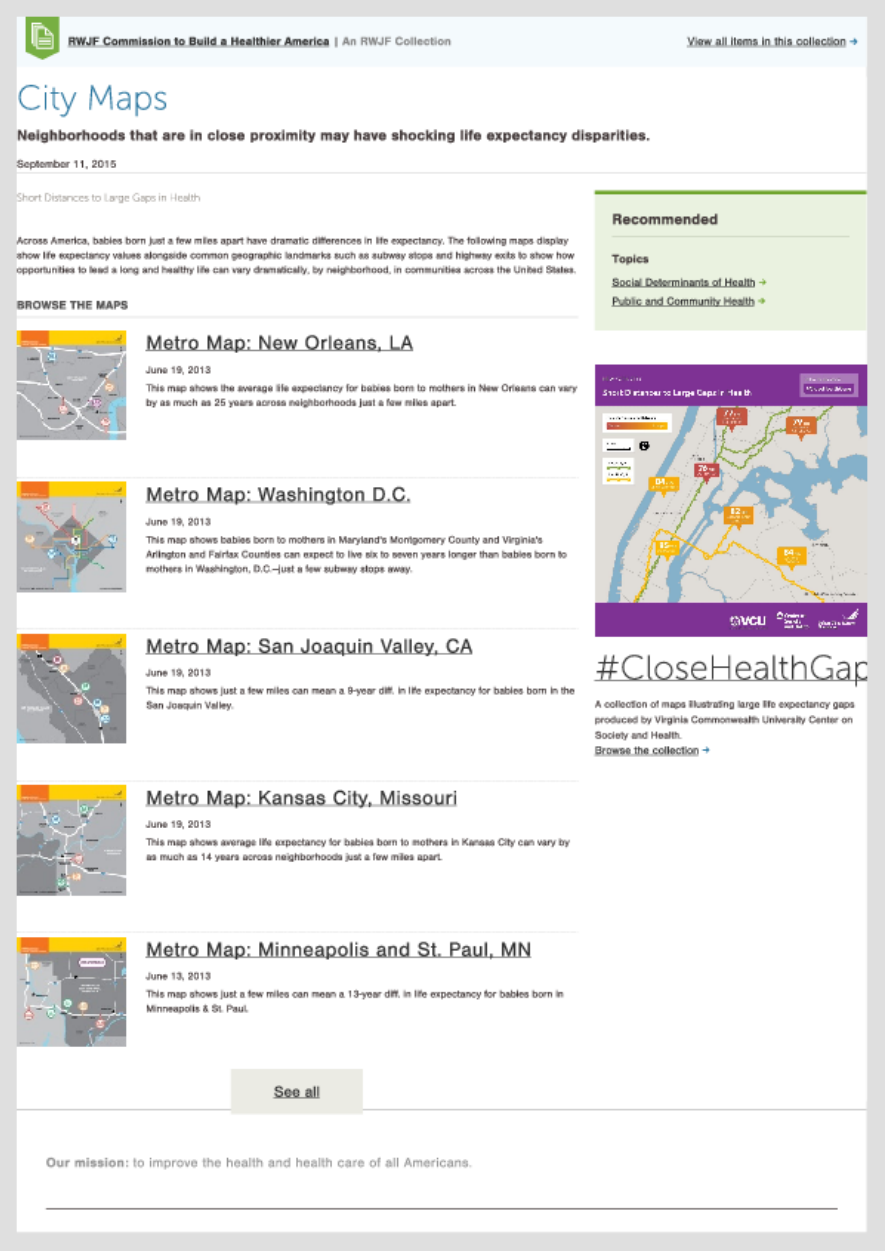


Figure 68: City Maps - Robert Wood Johnson Foundation home page.

6. Datasets - data.sa.gov.au: A collection of datasets made available by the South Australia government. (see Figure 69)
`https://data.sa.gov.au/data/dataset`

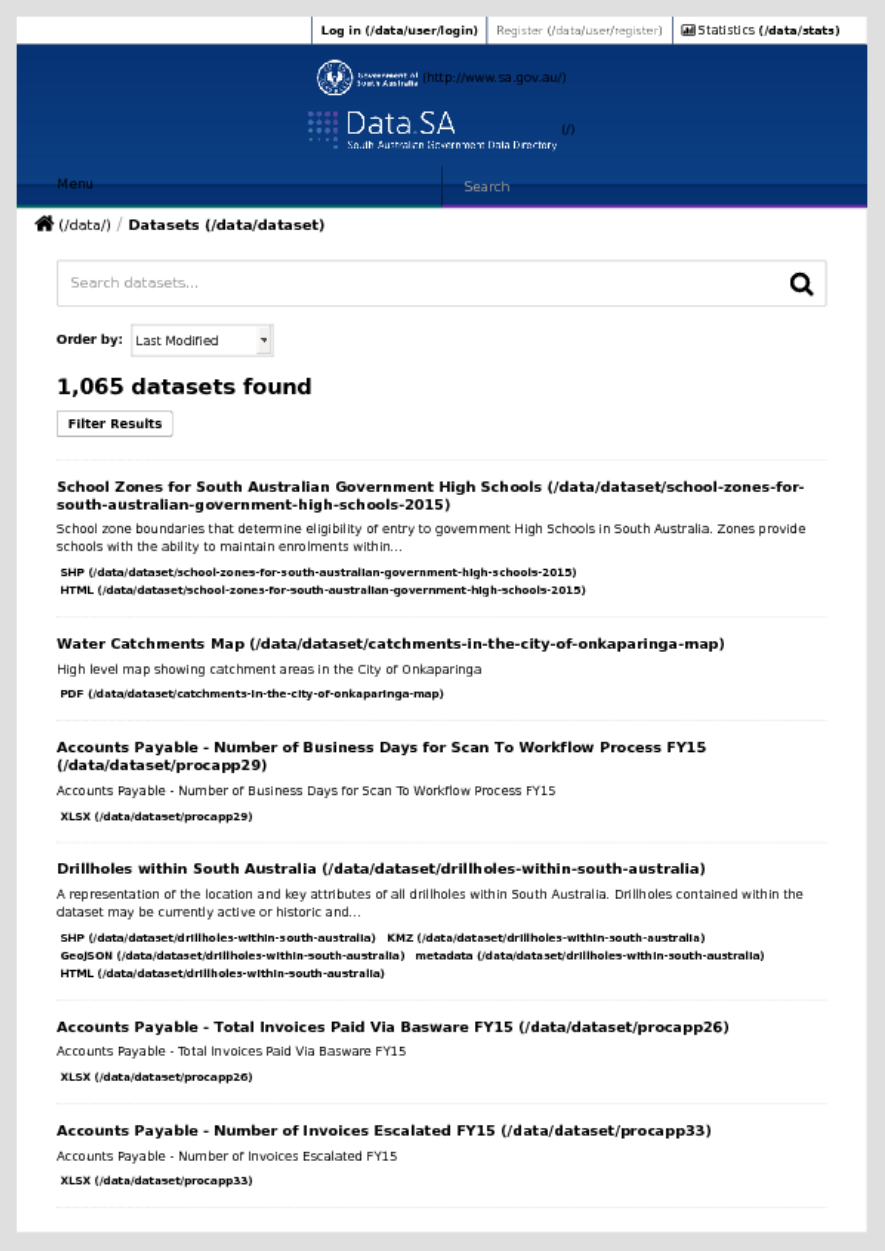


Figure 69: Datasets - data.sa.gov.au home page.

7. Earthquake Hazards Program: The Earthquake Notification Service (ENS) is a free service that sends you automated notifications to your email or cell phone when earthquakes happen. (see Figure 70)
`http://earthquake.usgs.gov/earthquakes/feed/v1.0/`

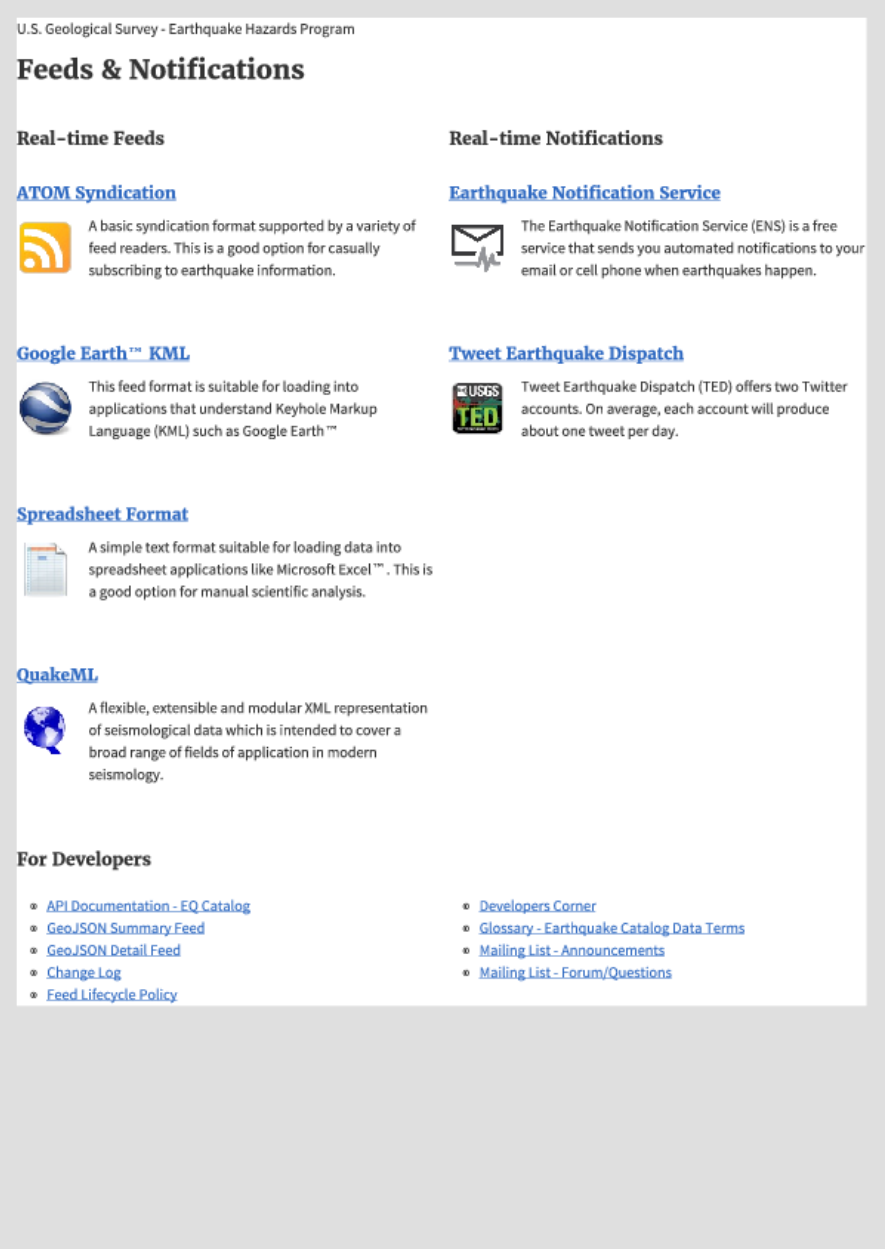
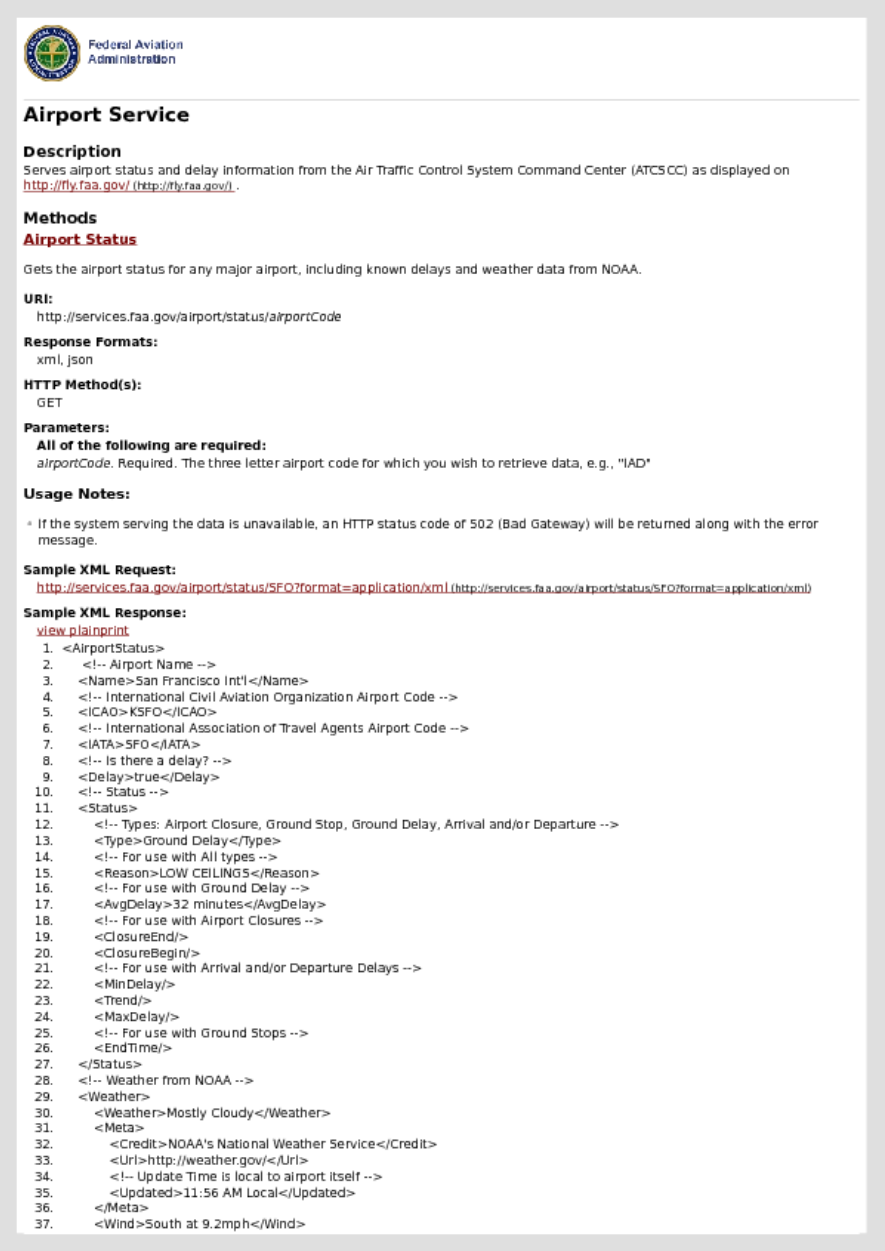


Figure 70: Earthquake Hazards Program home page.

8. FAA Web Services, Airport Service: Serves airport status and delay information from the Air Traffic Control System Command Center (ATCSCC) as displayed on http://fly.faa.gov/ (see Figure 71)
   `http://services.faa.gov/docs/services/airport/`



Federal Aviation Administration

Airport Service

Description

Serves airport status and delay information from the Air Traffic Control System Command Center (ATCSCC) as displayed on http://fly.faa.gov/ (http://fly.faa.gov/) .

Methods

Airport Status

Gets the airport status for any major airport, including known delays and weather data from NOAA.

URI:
http://services.faa.gov/airport/status/*airportCode*

Response Formats:
xml, json

HTTP Method(s):
GET

Parameters:
All of the following are required:
airportCode. Required. The three letter airport code for which you wish to retrieve data, e.g., "IAD"

Usage Notes:

- If the system serving the data is unavailable, an HTTP status code of 502 (Bad Gateway) will be returned along with the error message.

Sample XML Request:
http://services.faa.gov/airport/status/SFO?format=application/xml (http://services.faa.gov/airport/status/SFO?format=application/xml)

Sample XML Response:
view plainprint

```
<AirportStatus>
  <!-- Airport Name -->
  <Name>San Francisco Int'l</Name>
  <!-- International Civil Aviation Organization Airport Code -->
  <ICAO>KSFO</ICAO>
  <!-- International Association of Travel Agents Airport Code -->
  <IATA>SFO</IATA>
  <!-- Is there a delay? -->
  <Delay>true</Delay>
  <!-- Status -->
  <Status>
    <!-- Types: Airport Closure, Ground Stop, Ground Delay, Arrival and/or Departure -->
    <Type>Ground Delay</Type>
    <!-- For use with All types -->
    <Reason>LOW CEILINGS</Reason>
    <!-- For use with Ground Delay -->
    <AvgDelay>32 minutes</AvgDelay>
    <!-- For use with Airport Closures -->
    <ClosureEnd/>
    <ClosureBegin/>
    <!-- For use with Arrival and/or Departure Delays -->
    <MinDelay/>
    <Trend/>
    <MaxDelay/>
    <!-- For use with Ground Stops -->
    <EndTime/>
  </Status>
  <!-- Weather from NOAA -->
  <Weather>
    <Weather>Mostly Cloudy</Weather>
    <Meta>
      <Credit>NOAA's National Weather Service</Credit>
      <Url>http://weather.gov/</Url>
      <!-- Update Time is local to airport itself -->
      <Updated>11:56 AM Local</Updated>
    </Meta>
    <Wind>South at 9.2mph</Wind>
```

Figure 71: FAA Web Services, Airport Service home page.

9. Federal Reserve Bank of St. Louis, Economic Research: Download, graph, and track 294,000 US and international time series from 81 sources. (see Figure 72)
`https://research.stlouisfed.org/fred2/`

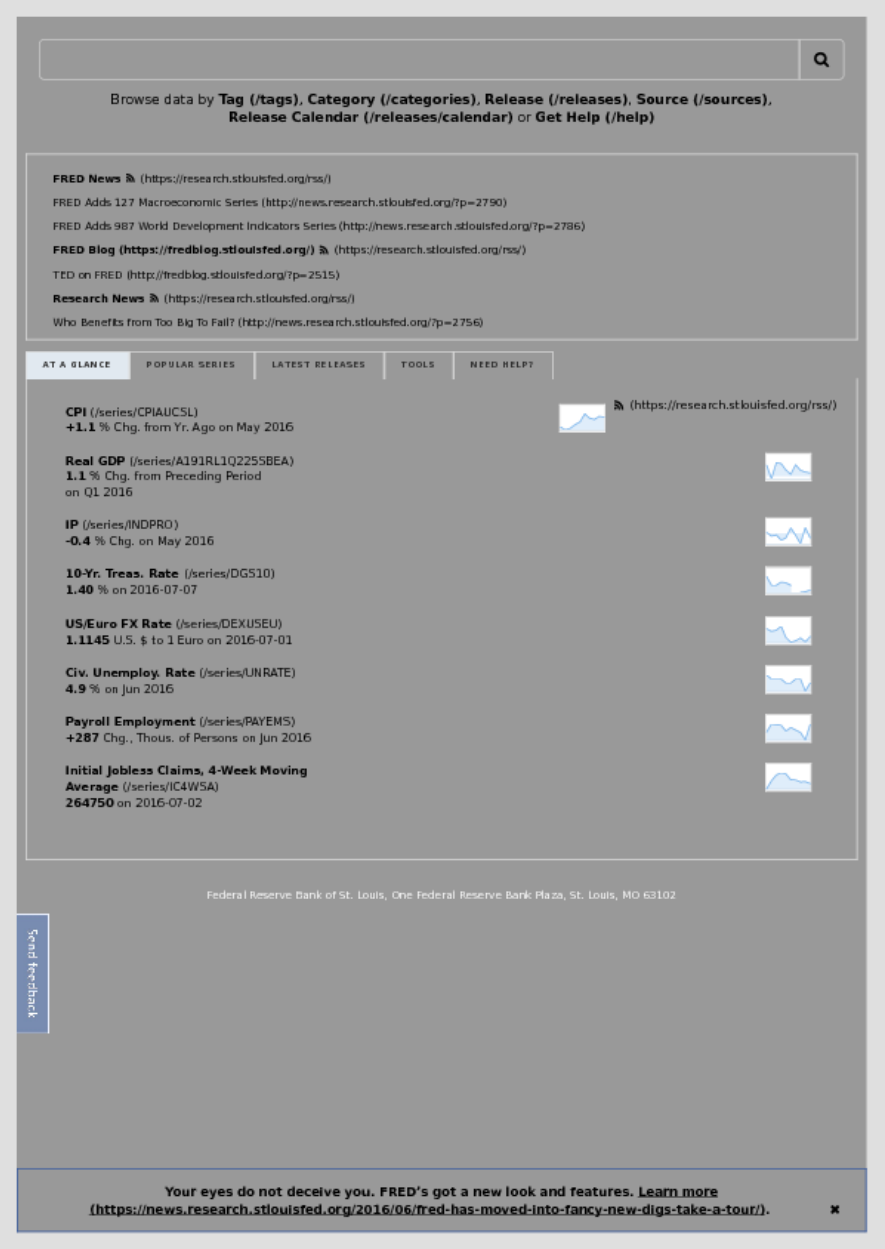


Figure 72: Federal Reserve Bank of St. Louis, Economic Research home page.

10. Gapminder Data: Gapminder is a non-profit venture promoting sustainable global development and achievement of the United Nations Millennium Development Goals by increased use and understanding of statistics and other information about social, economic and environmental development at local, national and global levels. (see Figure 73)
`http://www.gapminder.org/data/`

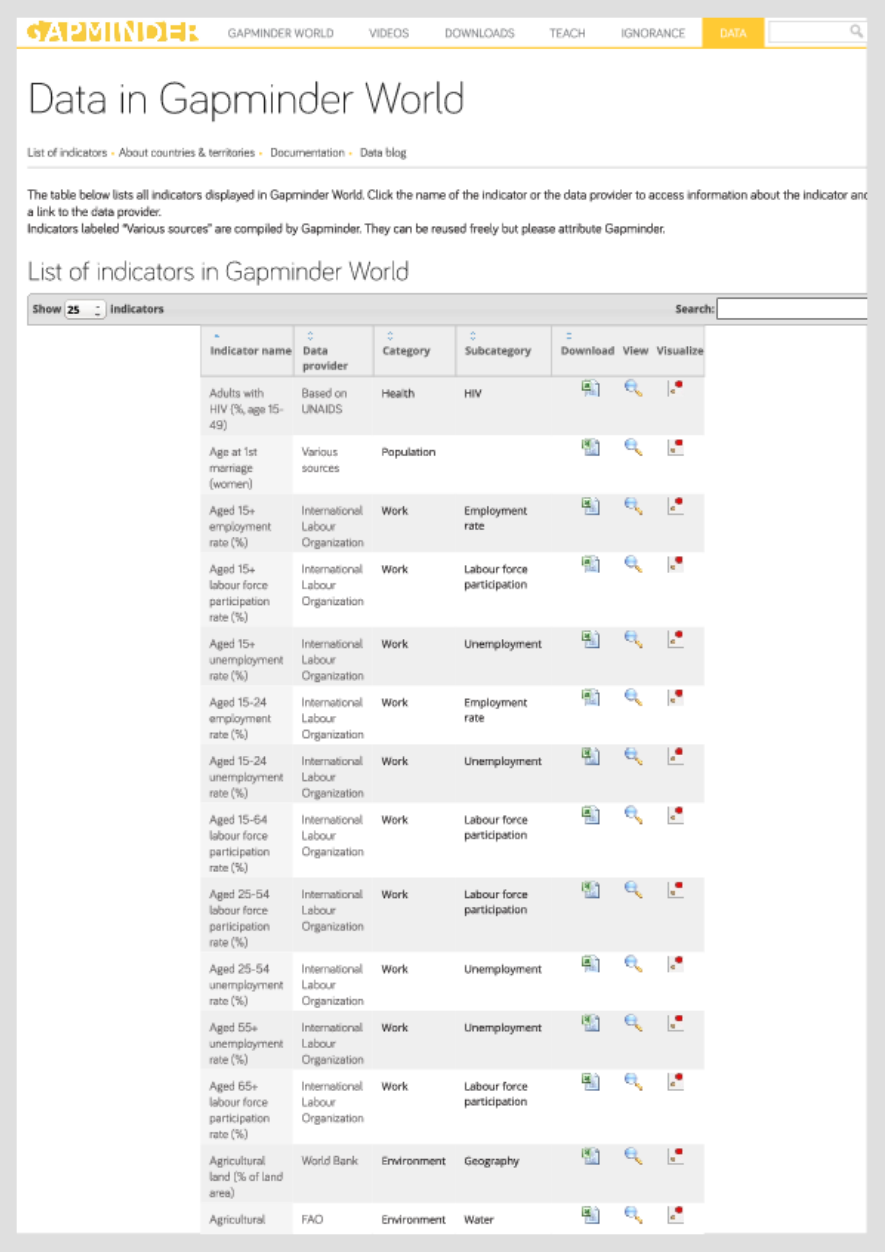


Figure 73: Gapminder Data home page.

11. General Social Survey: The GSS gathers data on contemporary American society in order to monitor and explain trends and constants in attitudes, behaviors, and attributes. Hundreds of trends have been tracked since 1972. In addition, since the GSS adopted questions from earlier surveys, trends can be followed for up to 70 years. The GSS contains a standard core of demographic, behavioral, and attitudinal questions, plus topics of special interest. Among the topics covered are civil liberties, crime and violence, intergroup tolerance, morality, national spending priorities, psychological well-being, social mobility, and stress and traumatic events. Altogether the GSS is the single best source for sociological and attitudinal trend data covering the United States. It allows researchers to examine the structure and functioning of society in general as well as the role played by relevant subgroups and to compare the United States to other nations. The GSS aims to make high-quality data easily accessible to scholars, students, policy makers, and others, with minimal cost and waiting. (see Figure 74)
`http://www3.norc.org/GSS+Website`

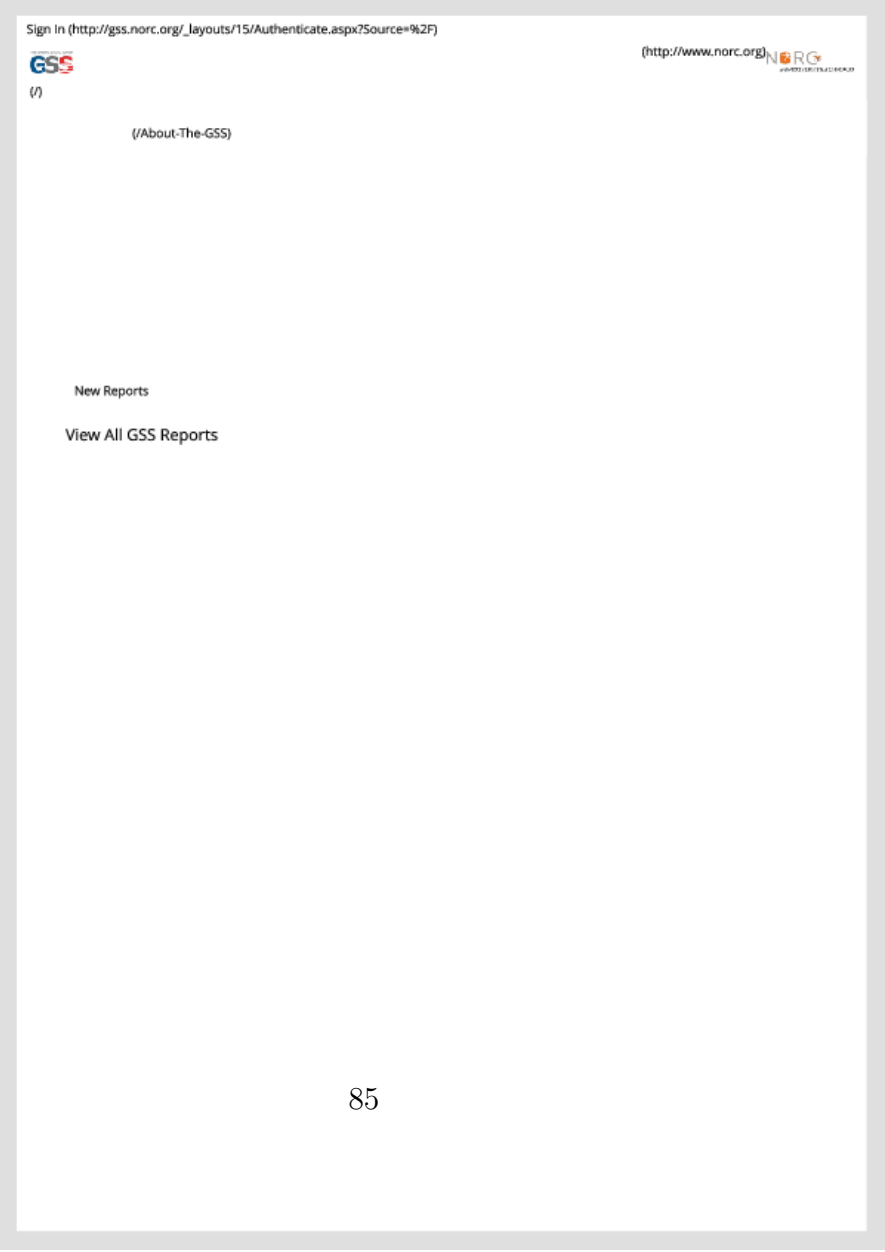


Figure 74: General Social Survey home page.

12. Massachusetts Bay Transportation Authority: Contains schedule, alert, vehicle position, and arrival prediction data. The same data is available in GTFS-realtime. (see Figure 75)
`http://realtime.mbta.com/portal`

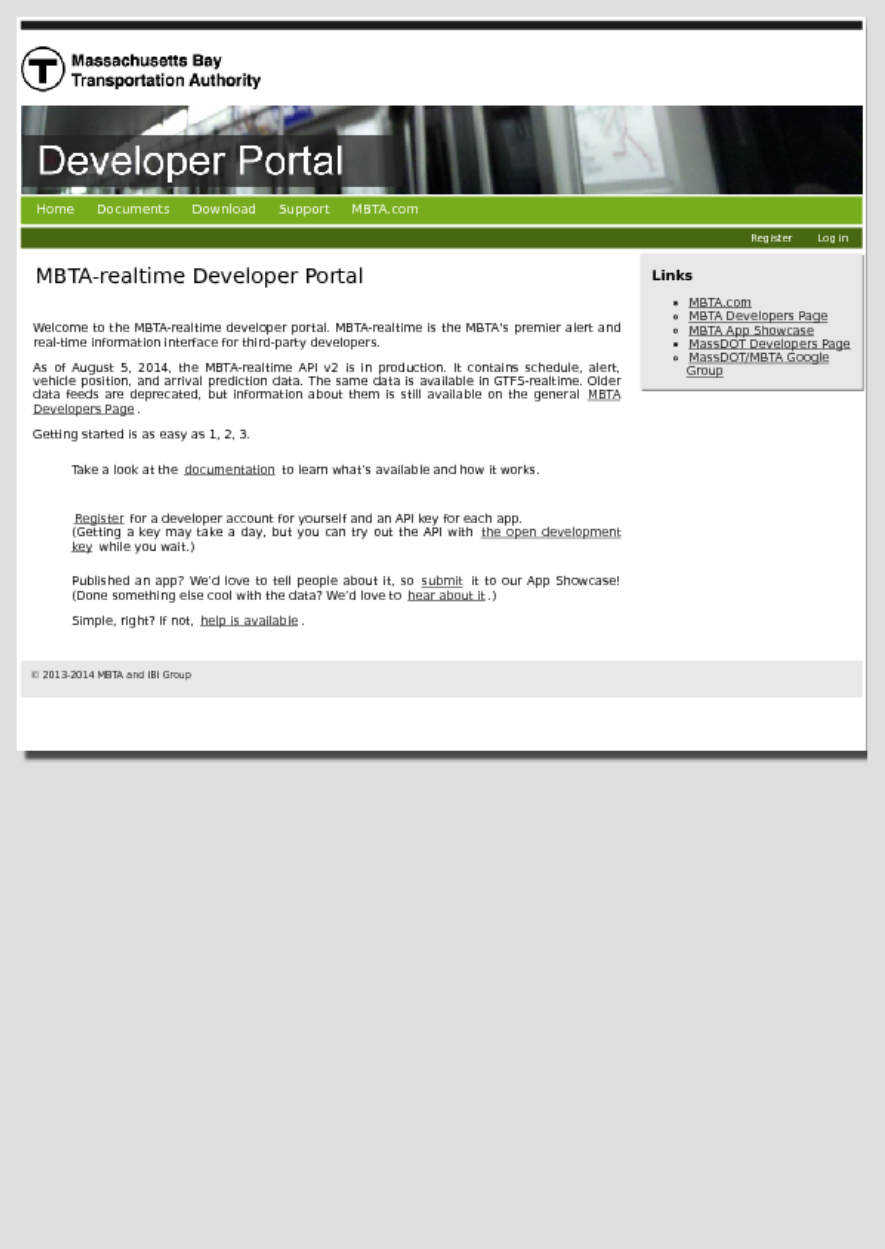


Figure 75: Massachusetts Bay Transportation Authority home page.

13. National Aeronautics and Space Administration, Goddard Institute for Space Studies: Datasets and derived material are available from the NASA GISS websites. (see Figure 76) `http://data.giss.nasa.gov/`

Figure 76: National Aeronautics and Space Administration, Goddard Institute for Space Studies home page.

14. National Flight Data Center (NFDC): The NFDC is part of the FAA's Aeronautical Information Services group (AJV-5). This web portal is intended to service the aviation community providing the ability to browse for aeronautical data, submit data to the FAA for airport or chart updates, or communicate with FAA specialists. (see Figure 77)
`https://nfdc.faa.gov/xwiki/bin/view/NFDC/WebHome`

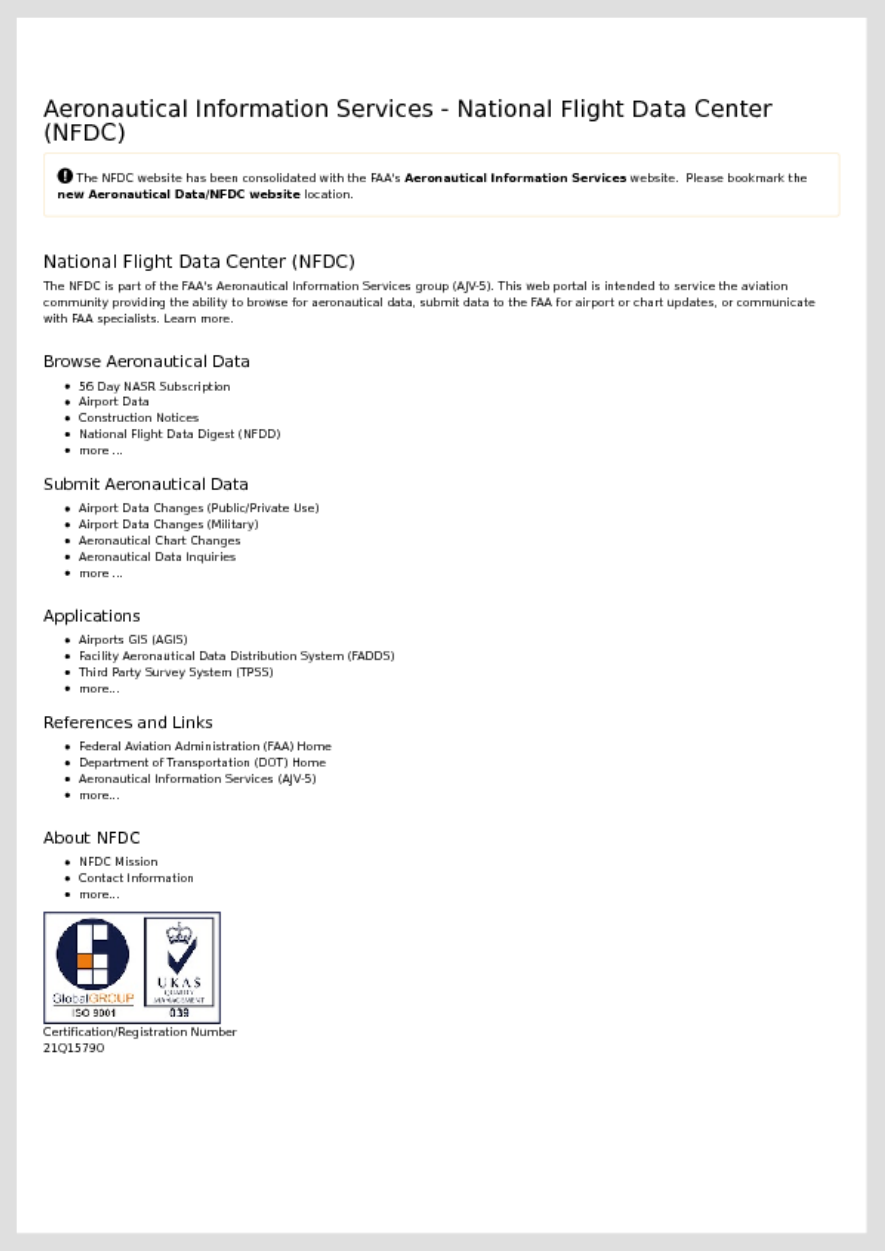


Figure 77: National Flight Data Center (NFDC) home page.

15. National Longitudinal Surveys — A Program of the U.S. Bureau of Labor Statistics: The NLS, sponsored by the U.S. Bureau of Labor Statistics, are nationally representative surveys that follow the same sample of individuals from specific birth cohorts over time. The surveys collect data on labor market activity, schooling, fertility, program participation, health, and much, much more. (see Figure 78)
`https://www.nlsinfo.org/#NLSY79`

Figure 78: National Longitudinal Surveys — A Program of the U.S. Bureau of Labor Statistics home page.

16. NYC Open Data: Open data government data about New York state. (see Figure 79) `https://data.cityofnewyork.us/`

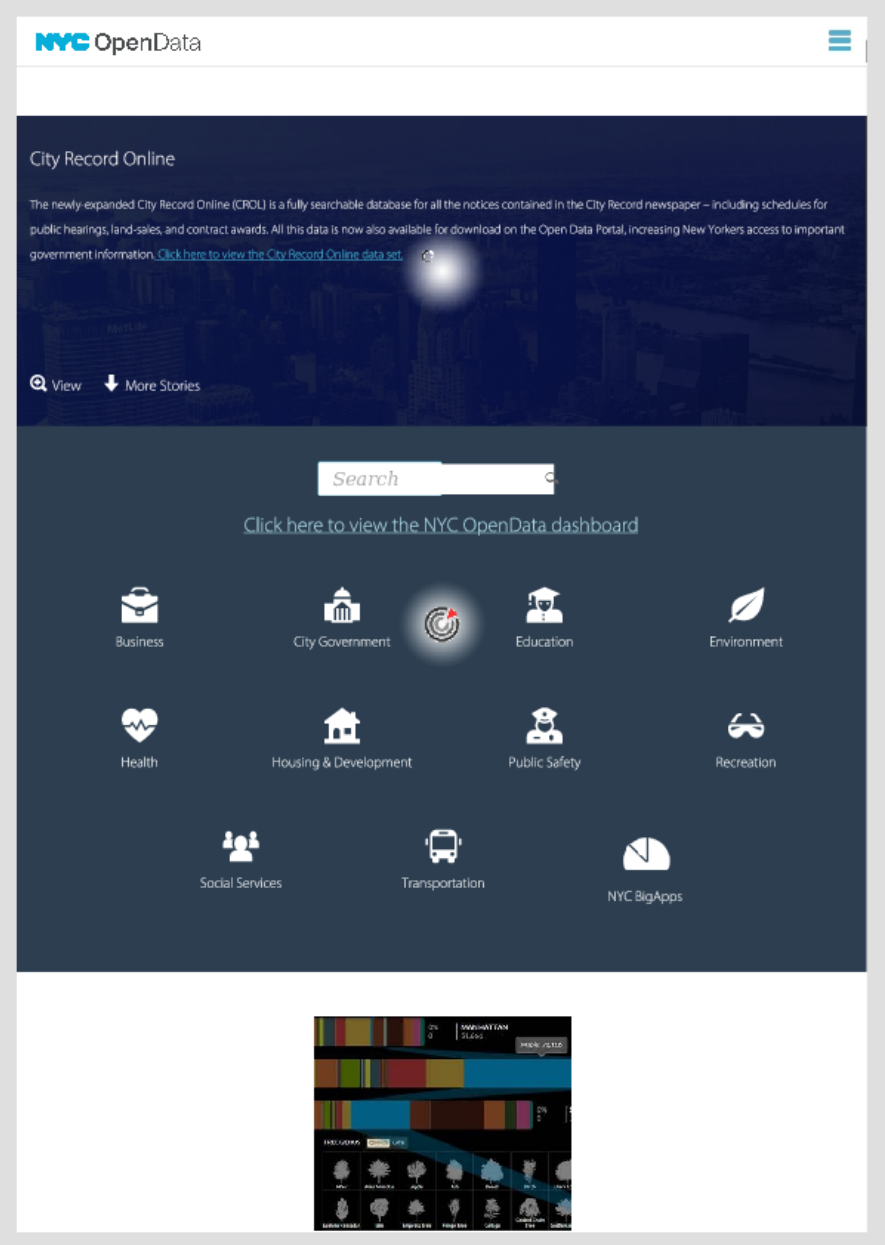


Figure 79: NYC Open Data home page.

17. Parking in San Francisco, CA: The PDF contains the URL to the API that provides real-time parking information for San Francisco, CA. (see Figure 80)
`http://sfpark.org/resources/sfpark-service-api-reference-updated/`

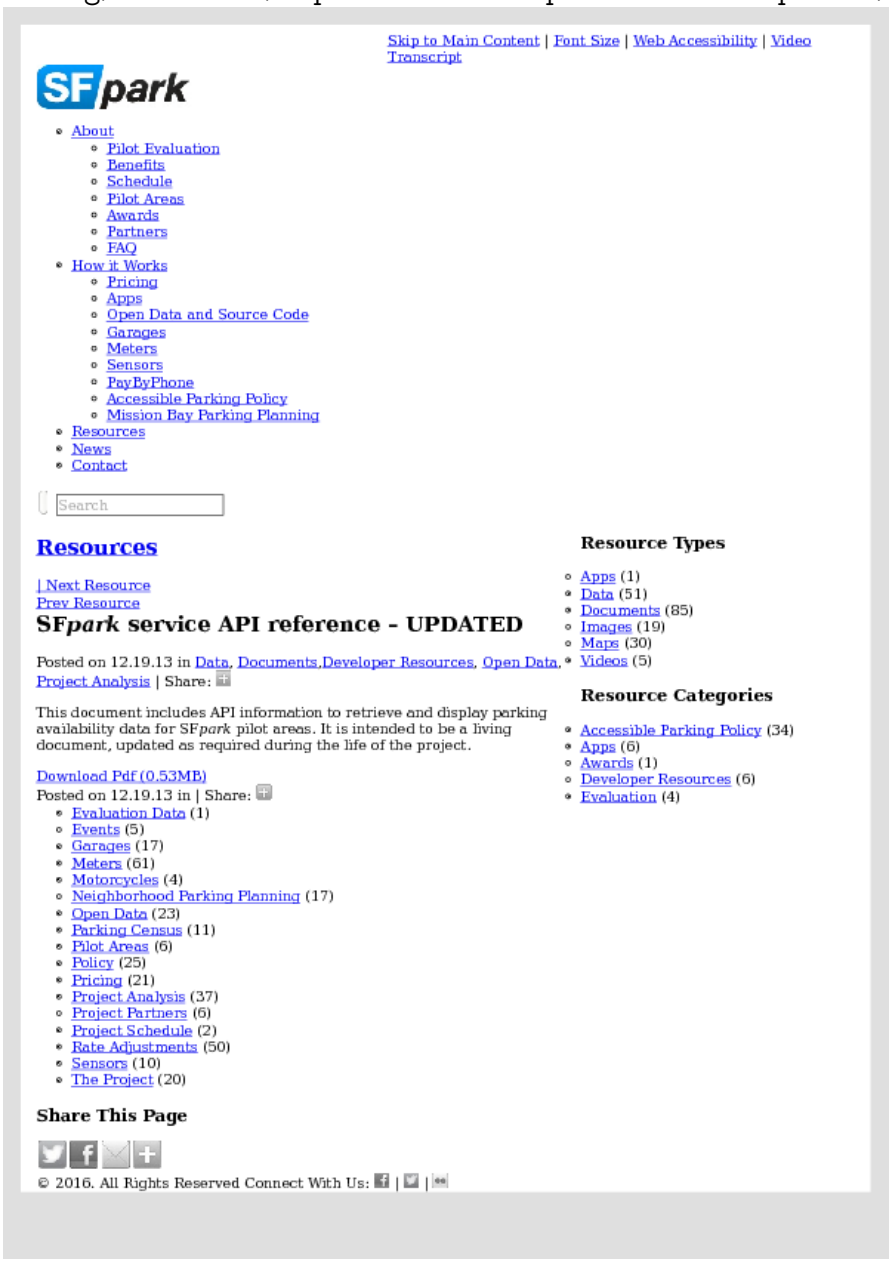


Figure 80: Parking in San Francisco, CA home page.

18. SNAP Retailer Locator: The location of retailers that take/honor Supplemental Nutrition Assistance Program (SNAP) cards. (see Figure 81)
`http://www.fns.usda.gov/snap/retailerlocator`

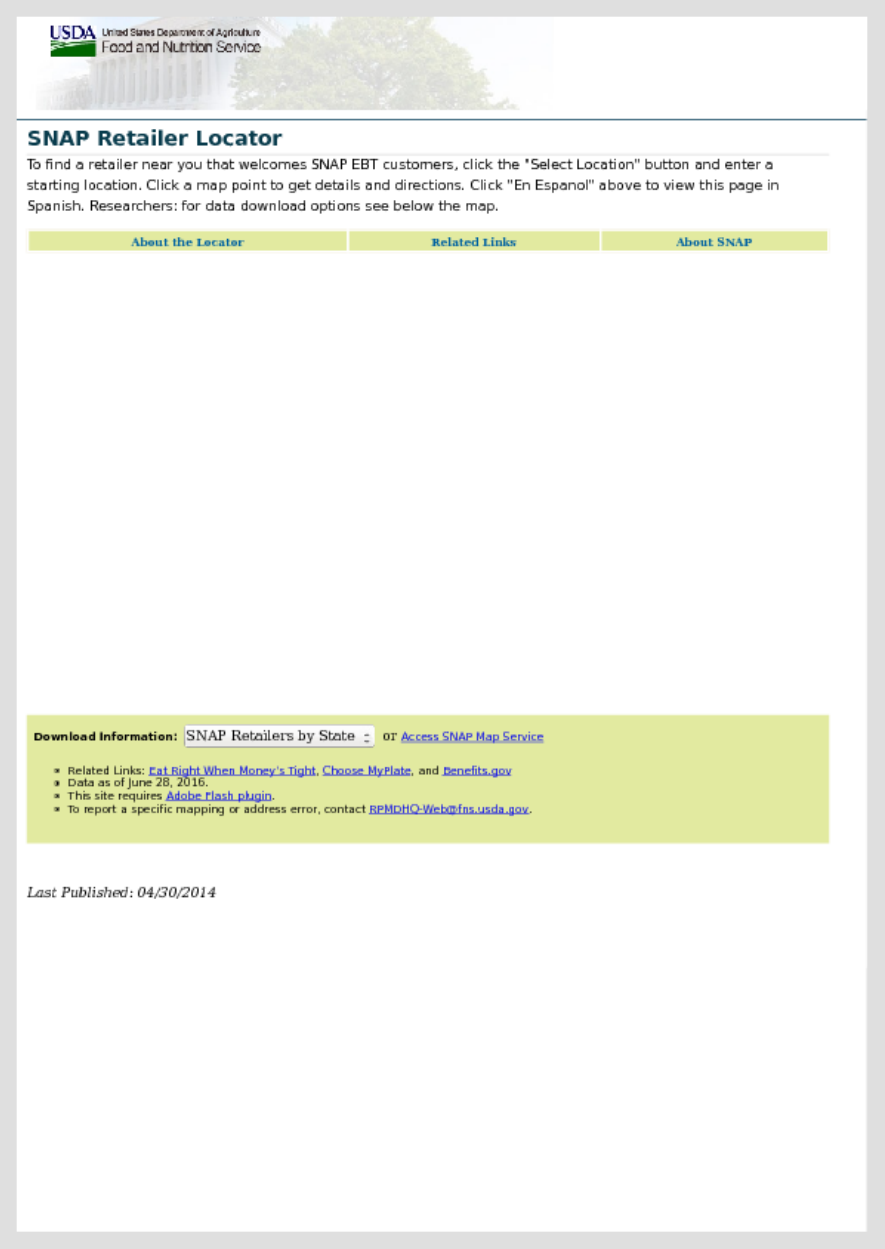
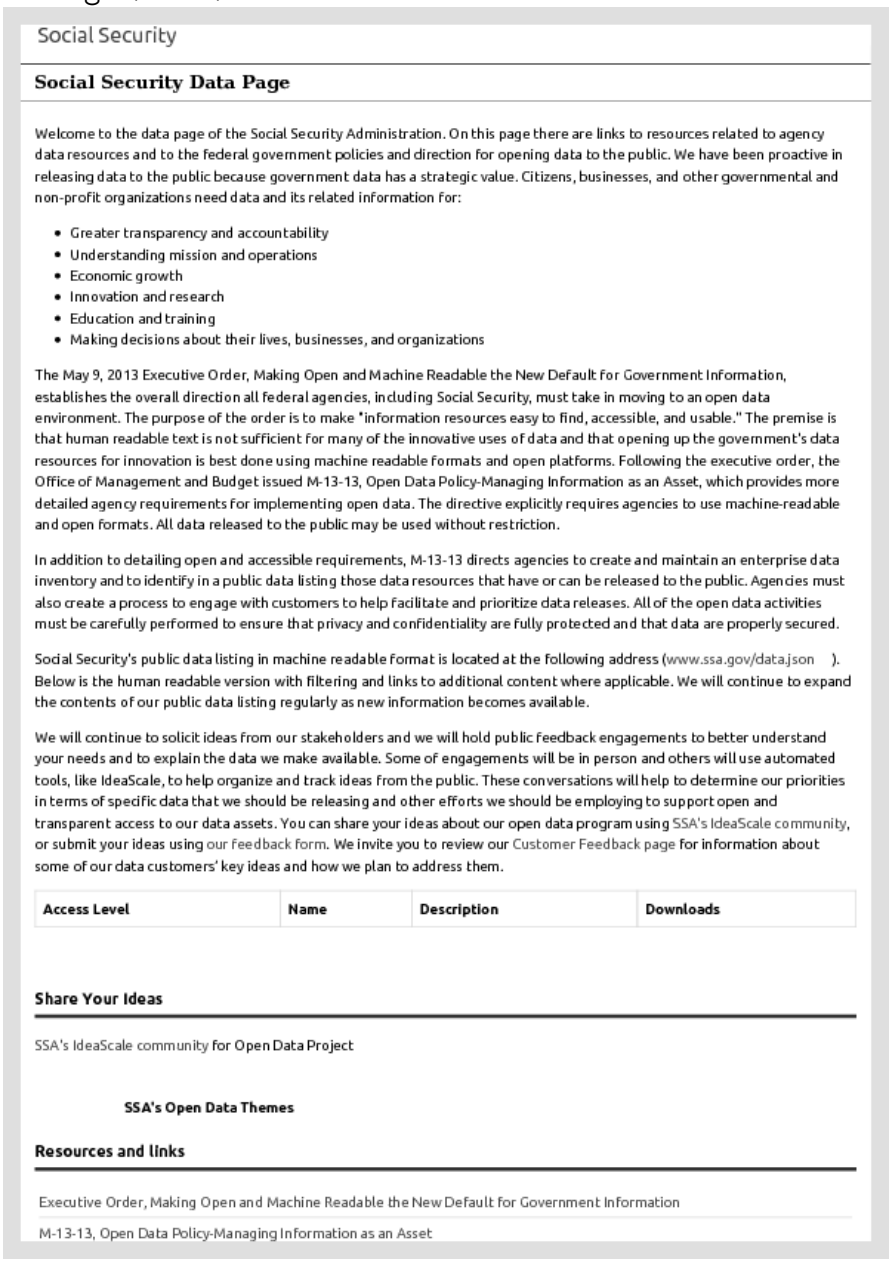


Figure 81: SNAP Retailer Locator home page.

19. Social Security Administration: Welcome to the data page of the Social Security Administration. On this page there are links to resources related to agency data resources and to the federal government policies and direction for opening data to the public. We have been proactive in releasing data to the public because government data has a strategic value. (see Figure 82)
`https://www.ssa.gov/data/`


Social Security

Social Security Data Page

Welcome to the data page of the Social Security Administration. On this page there are links to resources related to agency data resources and to the federal government policies and direction for opening data to the public. We have been proactive in releasing data to the public because government data has a strategic value. Citizens, businesses, and other governmental and non-profit organizations need data and its related information for:

- Greater transparency and accountability
- Understanding mission and operations
- Economic growth
- Innovation and research
- Education and training
- Making decisions about their lives, businesses, and organizations

The May 9, 2013 Executive Order, Making Open and Machine Readable the New Default for Government Information, establishes the overall direction all federal agencies, including Social Security, must take in moving to an open data environment. The purpose of the order is to make "information resources easy to find, accessible, and usable." The premise is that human readable text is not sufficient for many of the innovative uses of data and that opening up the government's data resources for innovation is best done using machine readable formats and open platforms. Following the executive order, the Office of Management and Budget issued M-13-13, Open Data Policy-Managing Information as an Asset, which provides more detailed agency requirements for implementing open data. The directive explicitly requires agencies to use machine-readable and open formats. All data released to the public may be used without restriction.

In addition to detailing open and accessible requirements, M-13-13 directs agencies to create and maintain an enterprise data inventory and to identify in a public data listing those data resources that have or can be released to the public. Agencies must also create a process to engage with customers to help facilitate and prioritize data releases. All of the open data activities must be carefully performed to ensure that privacy and confidentiality are fully protected and that data are properly secured.

Social Security's public data listing in machine readable format is located at the following address (www.ssa.gov/data.json). Below is the human readable version with filtering and links to additional content where applicable. We will continue to expand the contents of our public data listing regularly as new information becomes available.

We will continue to solicit ideas from our stakeholders and we will hold public feedback engagements to better understand your needs and to explain the data we make available. Some of engagements will be in person and others will use automated tools, like IdeaScale, to help organize and track ideas from the public. These conversations will help to determine our priorities in terms of specific data that we should be releasing and other efforts we should be employing to support open and transparent access to our data assets. You can share your ideas about our open data program using SSA's IdeaScale community, or submit your ideas using our feedback form. We invite you to review our Customer Feedback page for information about some of our data customers' key ideas and how we plan to address them.

Access Level	Name	Description	Downloads

Share Your Ideas

SSA's IdeaScale community for Open Data Project

SSA's Open Data Themes

Resources and links

Executive Order, Making Open and Machine Readable the New Default for Government Information

M-13-13, Open Data Policy-Managing Information as an Asset

Figure 82: Social Security Administration home page.

20. The National Archives: The United Kingdom National Archives. (see Figure 83)
`http://www.nationalarchives.gov.uk/`

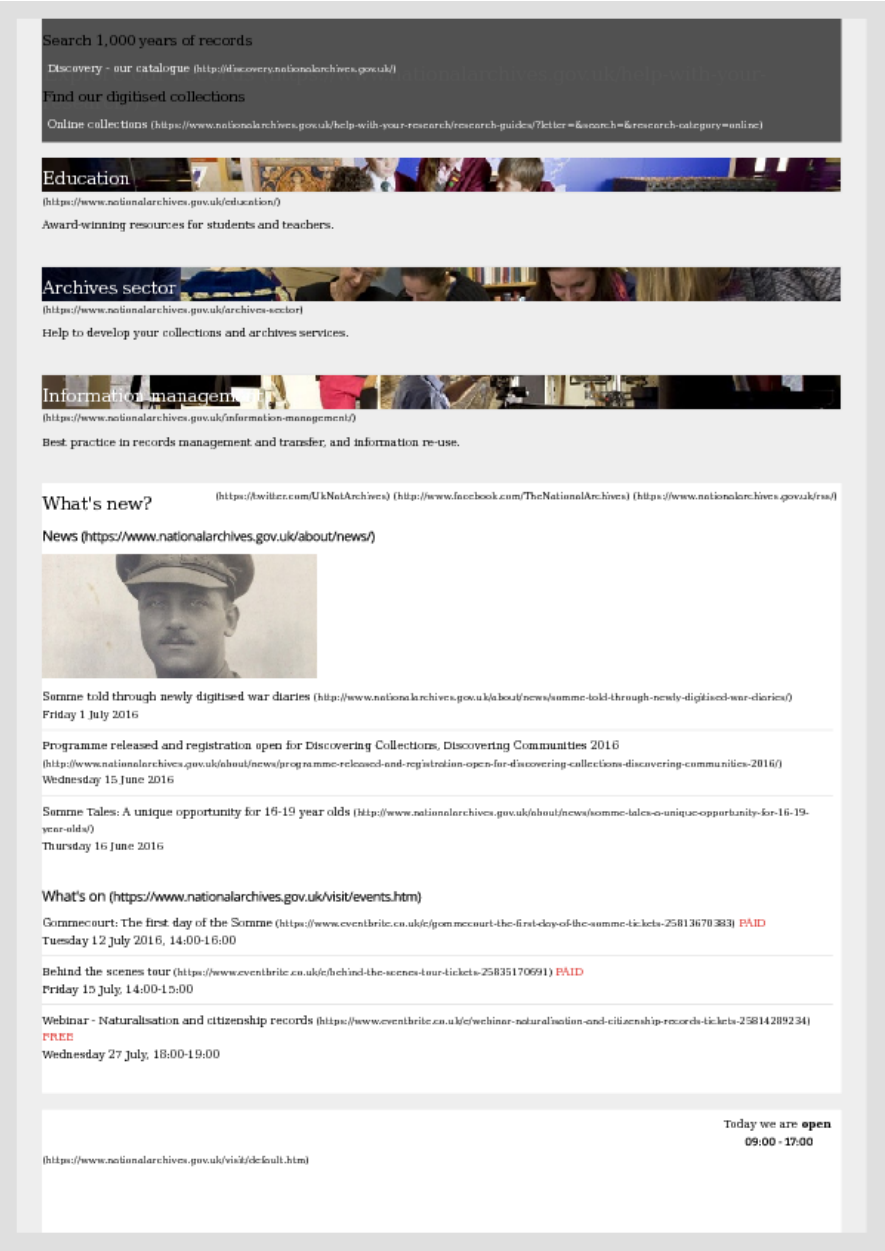


Figure 83: The National Archives home page.

21. THe National Map: As one of the cornerstones of the U.S. Geological Survey's (USGS) National Geospatial Program, The National Map is a collaborative effort among the USGS and other Federal, State, and local partners to improve and deliver topographic information for the Nation. (see Figure 84)
`http://nationalmap.gov/`

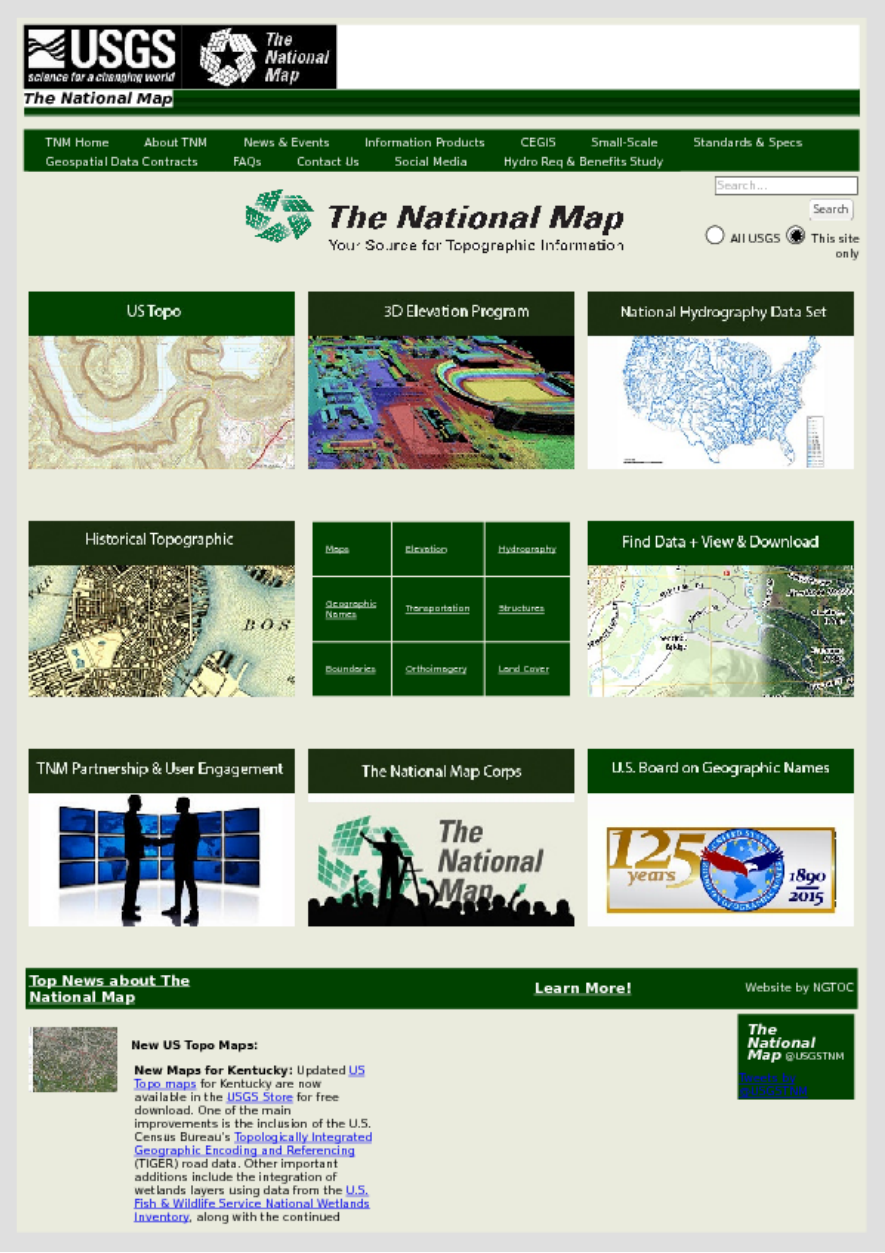


Figure 84: THe National Map home page.

22. The National UFO Reporting Center: Dedicated to the Collection and Dissemination of Objective UFO Data (see Figure 85)
`http://www.nuforc.org/webreports.html`

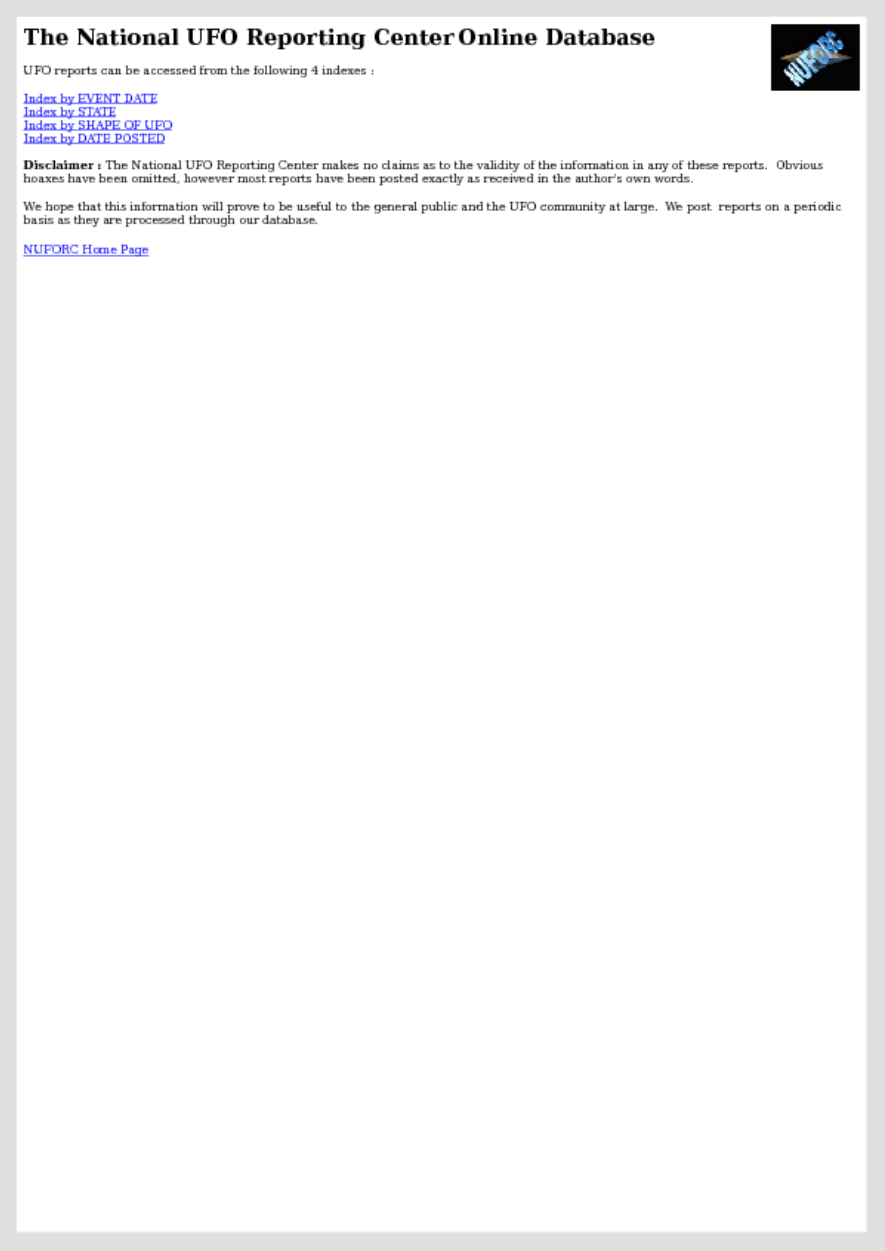


Figure 85: The National UFO Reporting Center home page.

23. The World Bank: These datasets were compiled for World Bank research and are provided free of cost to foster the creation of new knowledge. (see Figure 86)
`http://www.worldbank.org/`

Figure 86: The World Bank home page.

24. UK Government Open Data Portal: U.S. Government open data portal. (see Figure 87) `https://data.gov.uk/`

Figure 87: UK Government Open Data Portal home page.

25. United States Department of Agriculture, Economic Research Service: The ERS mission is to inform and enhance public and private decision making on economic and policy issues related to agriculture, food, the environment, and rural development. With over 300 employees, The Economic Research Service is a primary source of economic information and research in the U.S. Department of Agriculture. (see Figure 88)
`http://www.ers.usda.gov/data-products/.aspx`

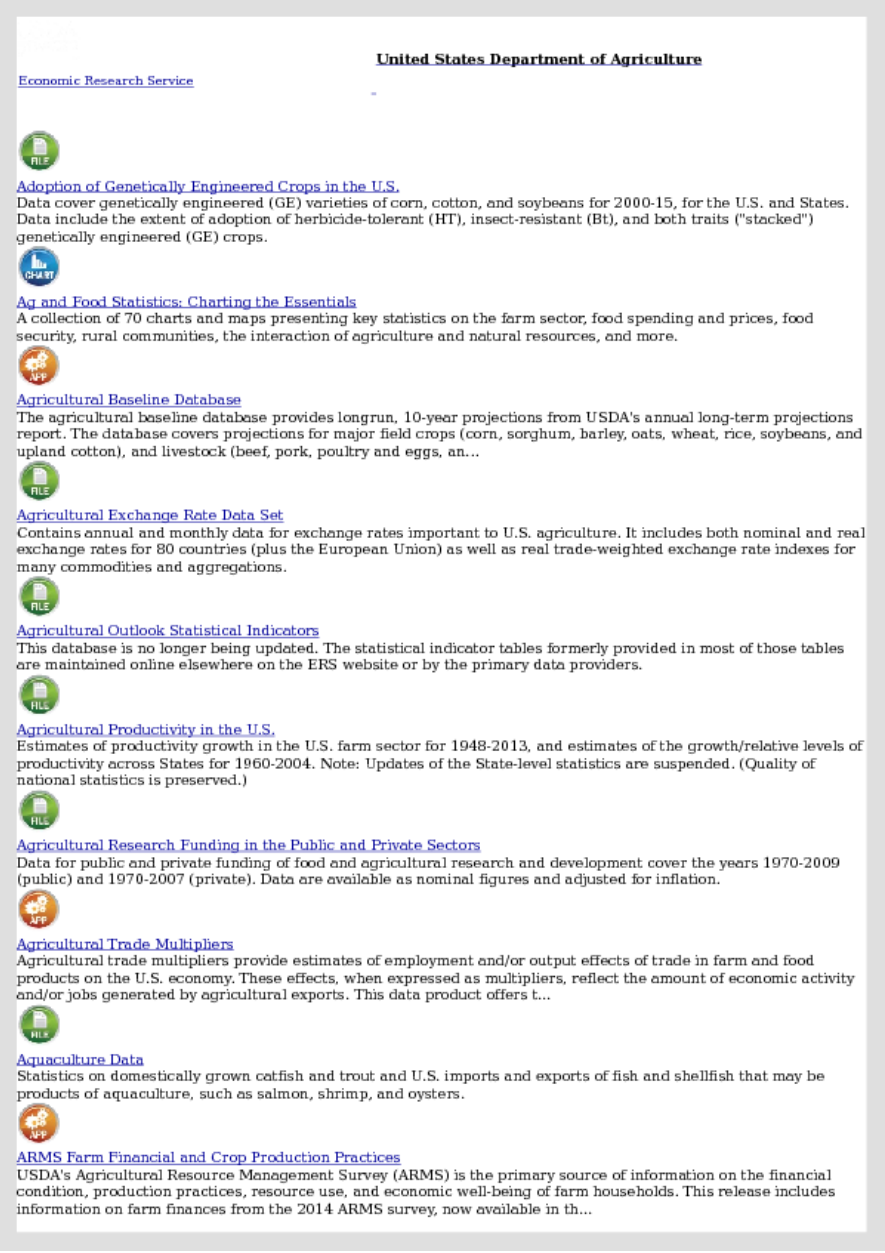


Figure 88: United States Department of Agriculture, Economic Research Service home page.

26. USDA Food Composition Database: Access to Release 28 of the USDA National Nutrient Database for Standard Reference. You can either view the data here or download the data files and documentation in several different formats for use on your computer. An online search is also provided so you can look up the nutrient content of 8,789 different foods directly from this home page. (see Figure 89)
`https://ndb.nal.usda.gov/`

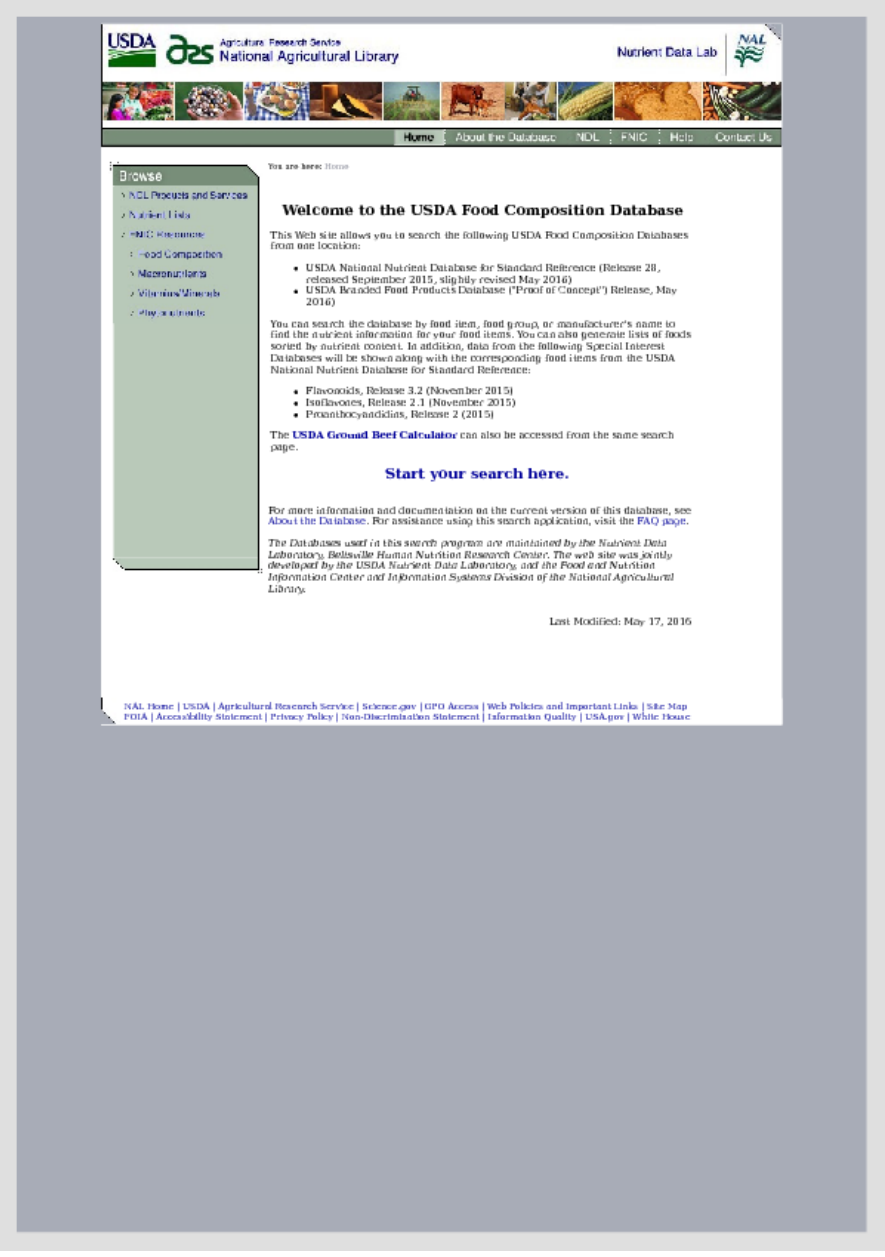


Figure 89: USDA Food Composition Database home page.

27. US Department of Transportation: Catalog of available datasets. (see Figure 90) `https://catalog.data.gov/organization/dot-gov`

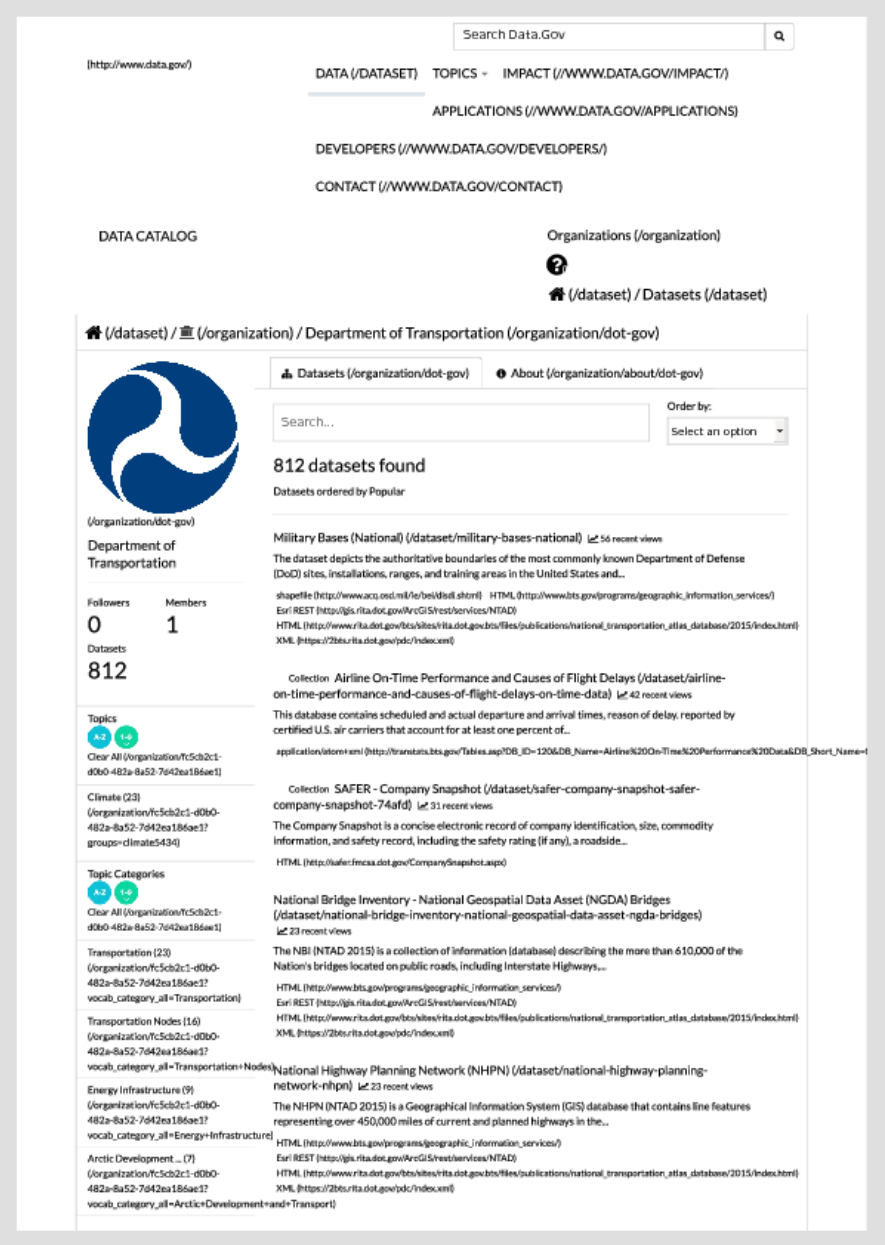


Figure 90: US Department of Transportation home page.

28. US Government Open Data Portal: U.S. Government open data portal to over 190,000 datasets. (see Figure 91)
`http://www.data.gov/`

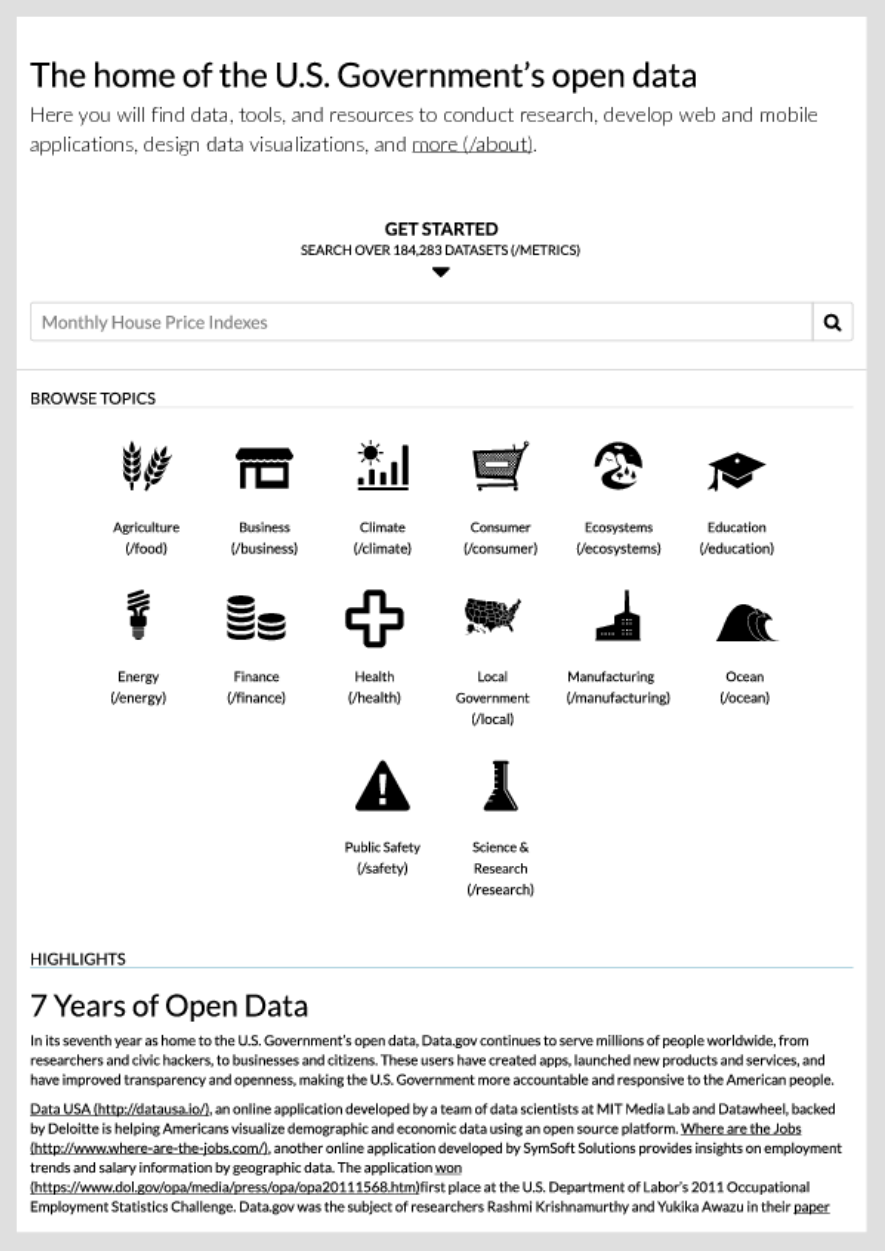


Figure 91: US Government Open Data Portal home page.

29. USGovXML.com: US Government Web Services and XML Data Sources.: USGovXML.com is an index to publicly available web services and XML data sources that are provided by the US government. USGovXML.com indexes data sources from all 3 branches of government as well as its boards, commissions, corporations and independent agencies. (see Figure 92) `http://usgovxml.com/`

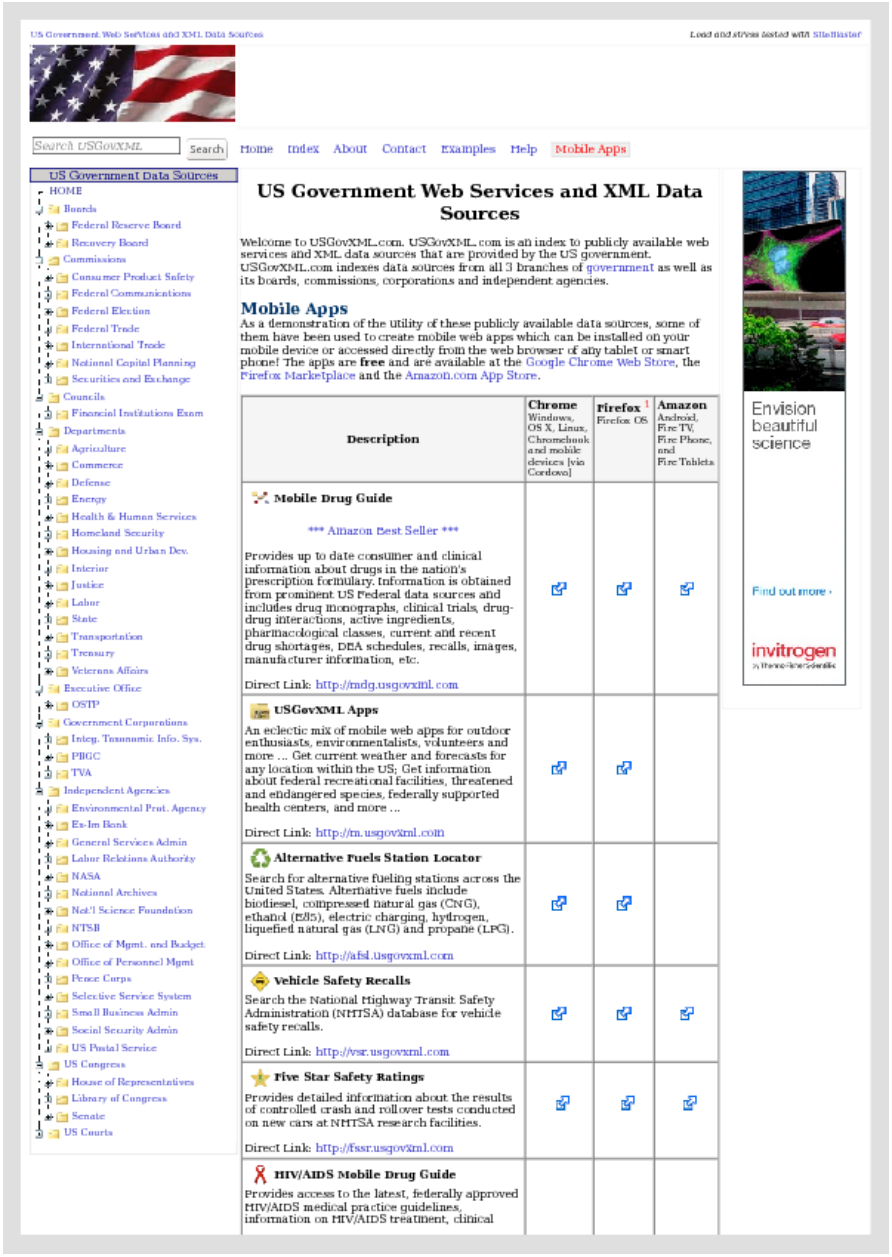


Figure 92: USGovXML.com: US Government Web Services and XML Data Sources. home page.

30. Washington Metropolitan Area Transit Authority API: API with real time information about bus, train, and stations in the Washington, DC Metro system. (see Figure 93)
`https://developer.wmata.com/`

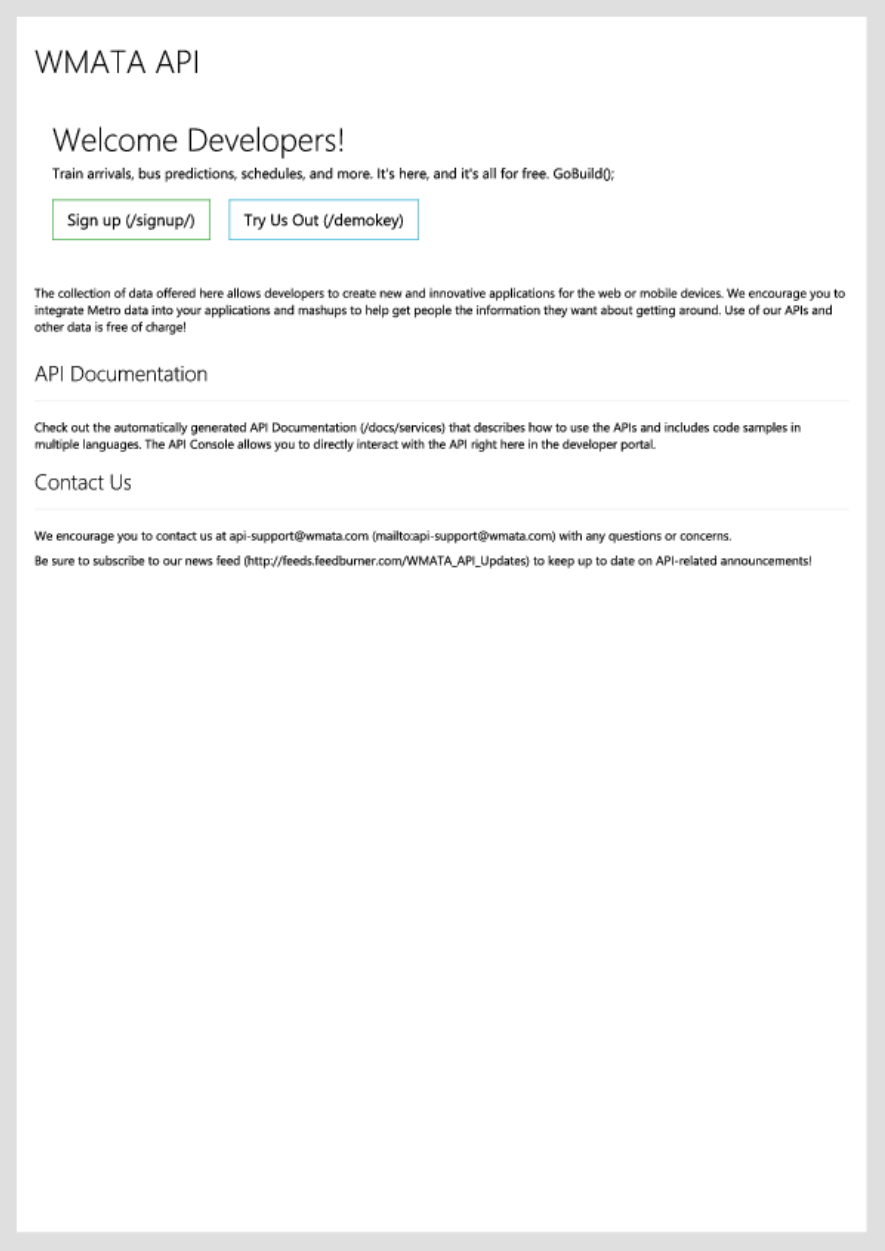


Figure 93: Washington Metropolitan Area Transit Authority API home page.

3.9 Social

1. American Customer Satisfaction Index: The American Customer Satisfaction Index provides unique customer experience benchmarking capabilities that come from the Indexs one-of-a-kind, cross-industry structure. (see Figure 95)
   `http://theacsi.org/`

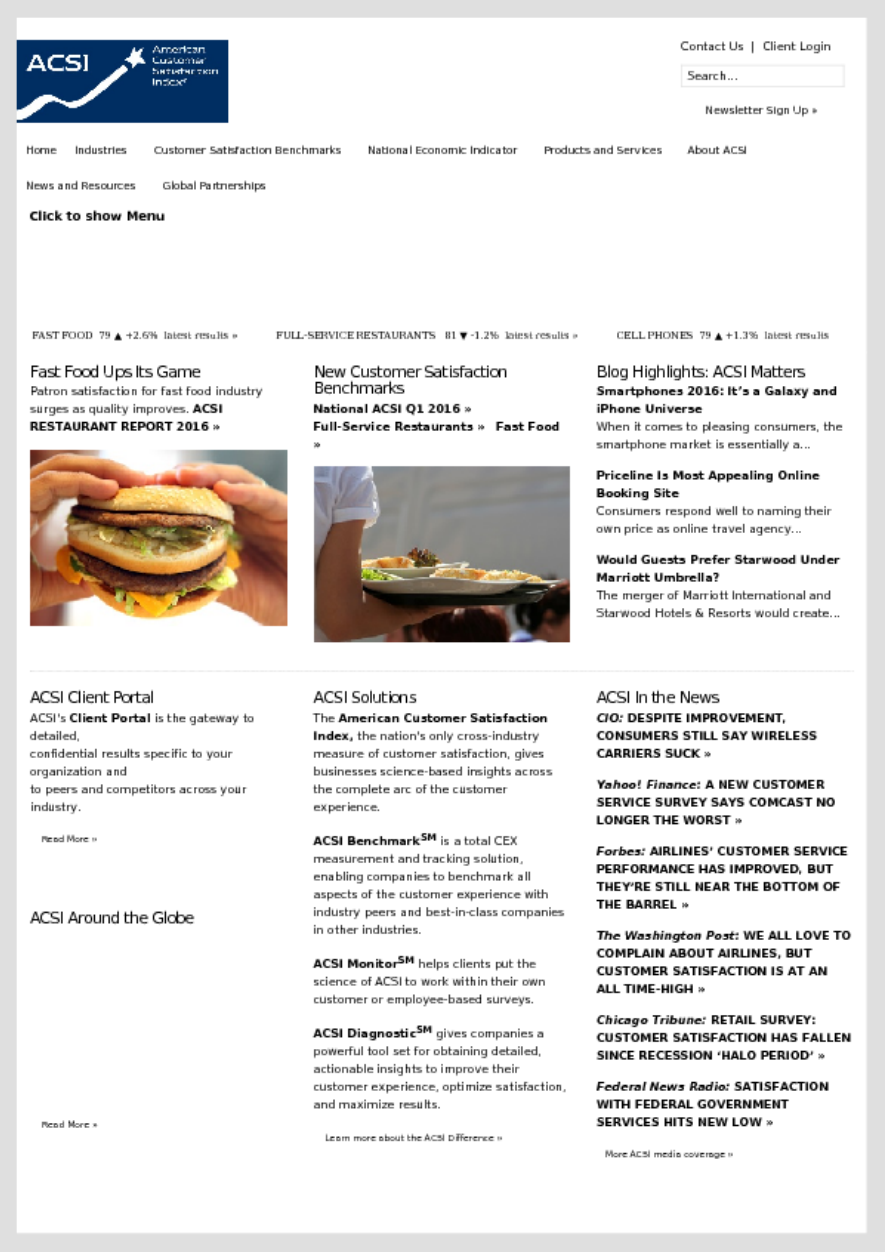


Figure 94: American Customer Satisfaction Index home page.

2. American Customer Satisfaction Index: The American Customer Satisfaction Index provides unique customer experience benchmarking capabilities that come from the Indexs one-of-a-kind, cross-industry structure. (see Figure 95)
`http://theacsi.org/`

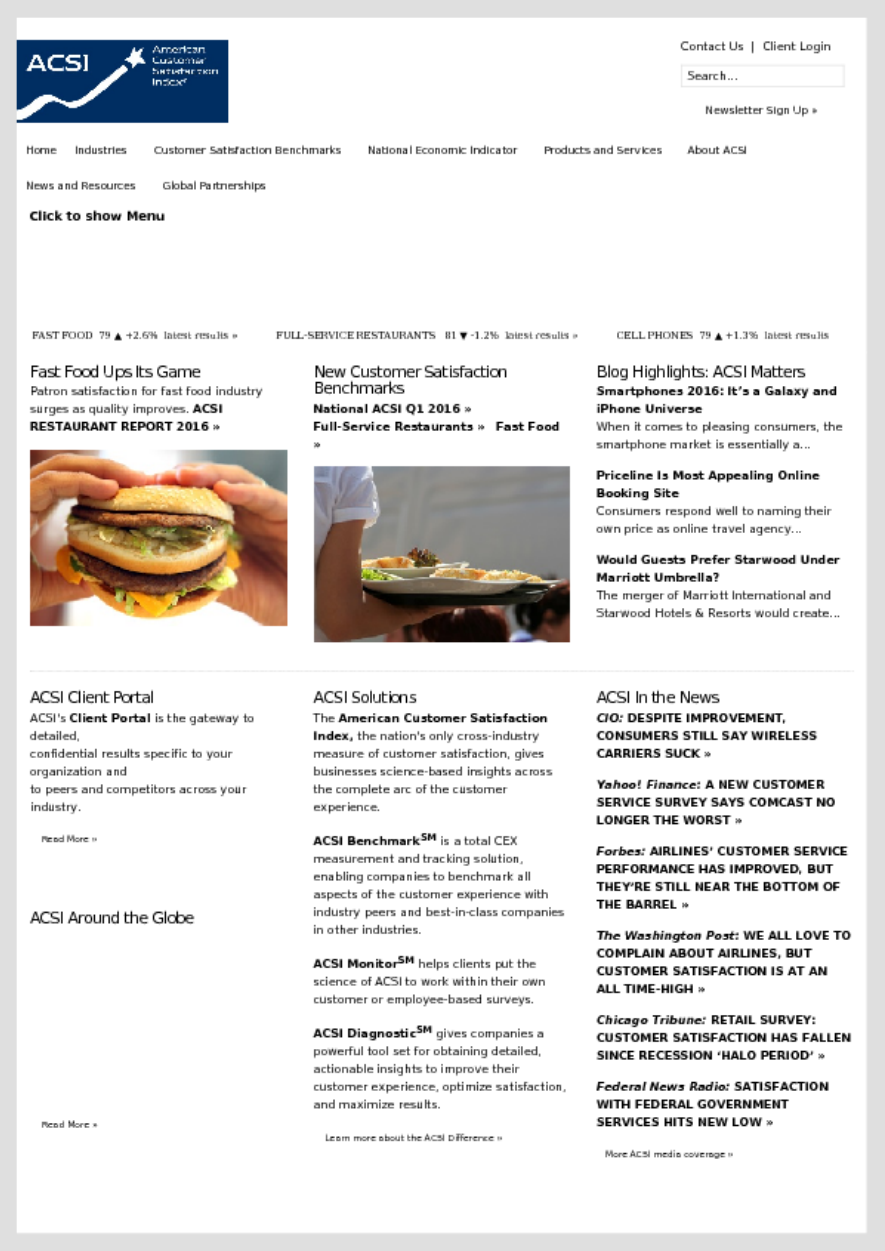


Figure 95: American Customer Satisfaction Index home page.

3. Linkedin API: The foundation of all digital integrations with LinkedIn. (see Figure 96) `https://developer.linkedin.com/docs/rest-api`

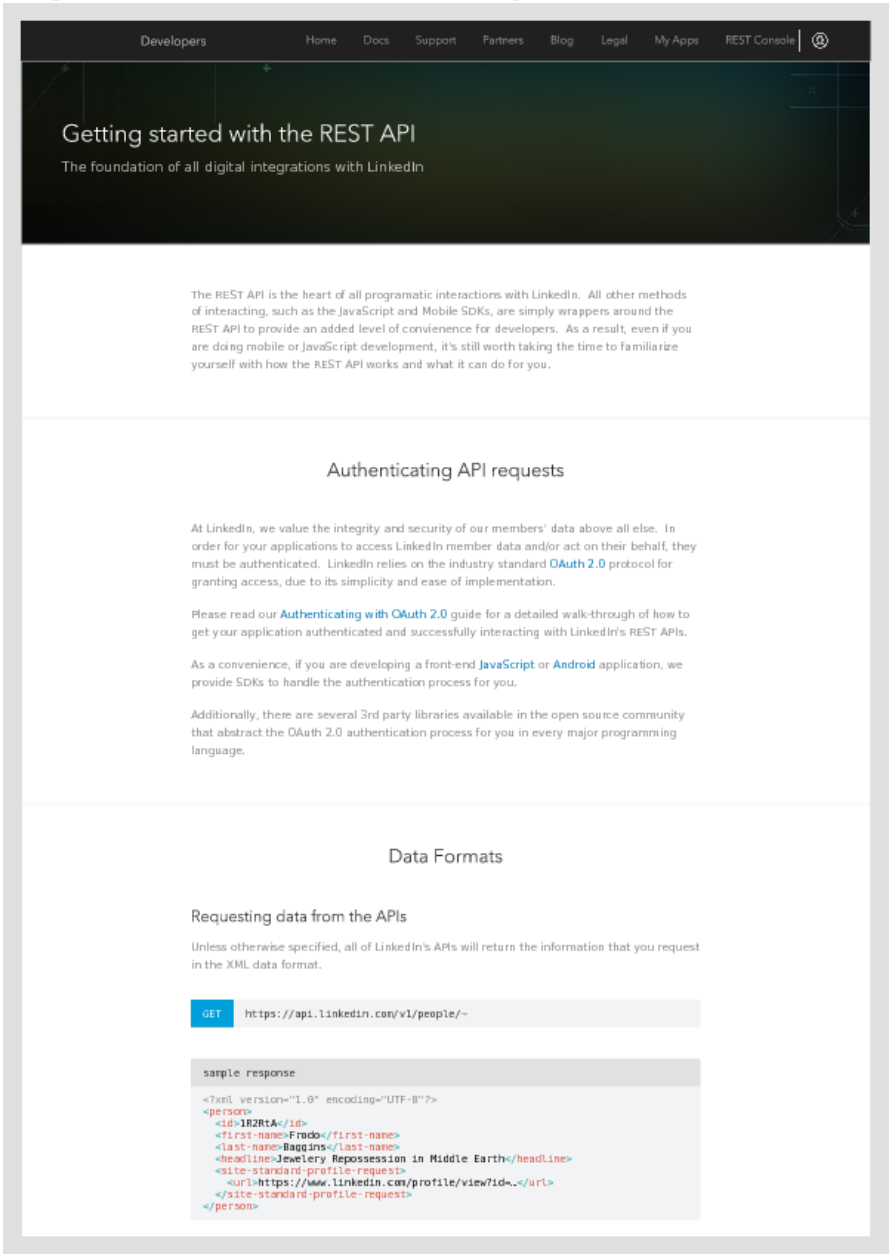


Figure 96: Linkedin API home page.

4. Meetup API: The Meetup API provides simple RESTful HTTP and streaming interfaces for extending your community using the Meetup platform from your own apps. (see Figure 97) `http://www.meetup.com/meetup_api/`

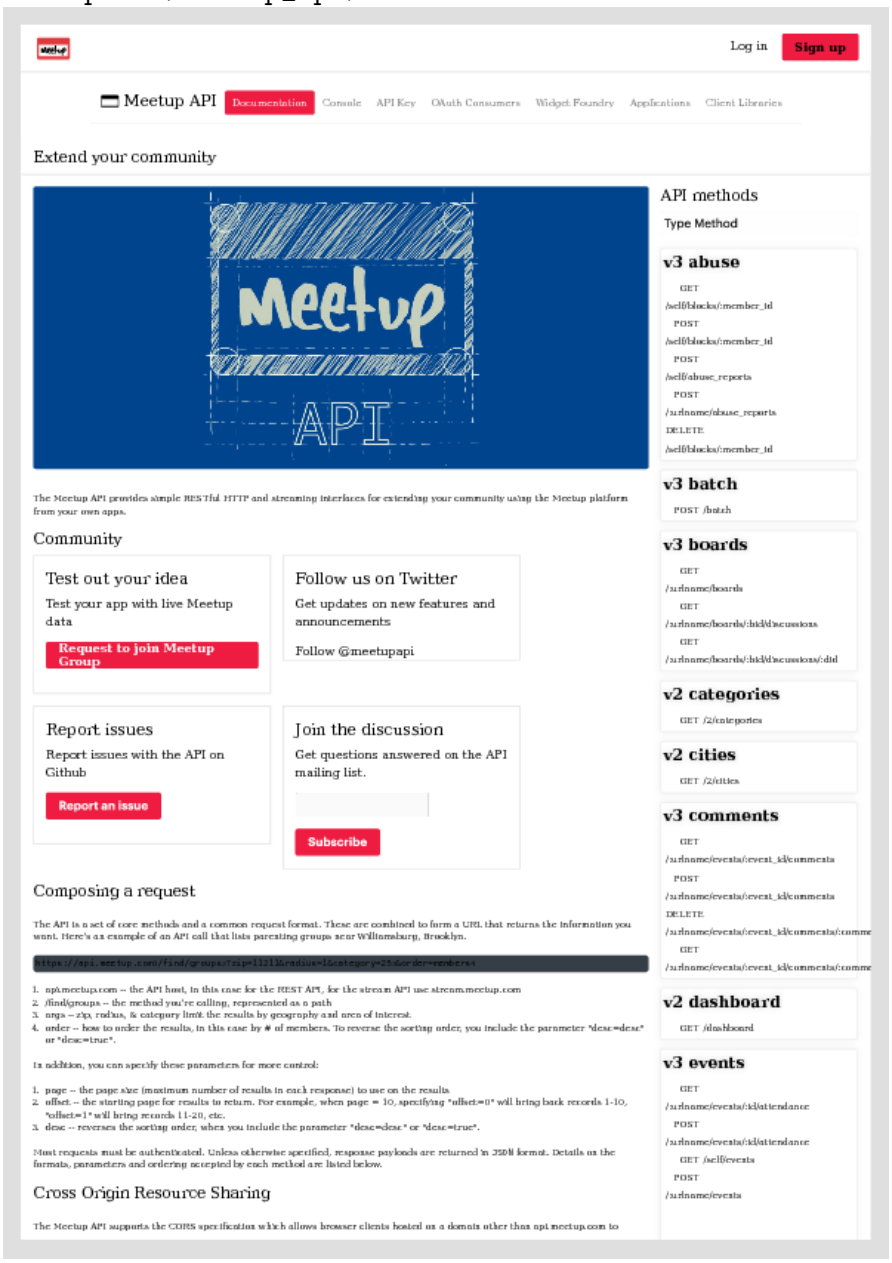


Figure 97: Meetup API home page.

5. YELP API: Search over 50 million local businesses from 32 countries. Enhance your app with Yelp ratings, reviews, photos and much more. Simple and fast API with powerful category and geo search filters. (see Figure 98)
`https://www.yelp.com/developers`

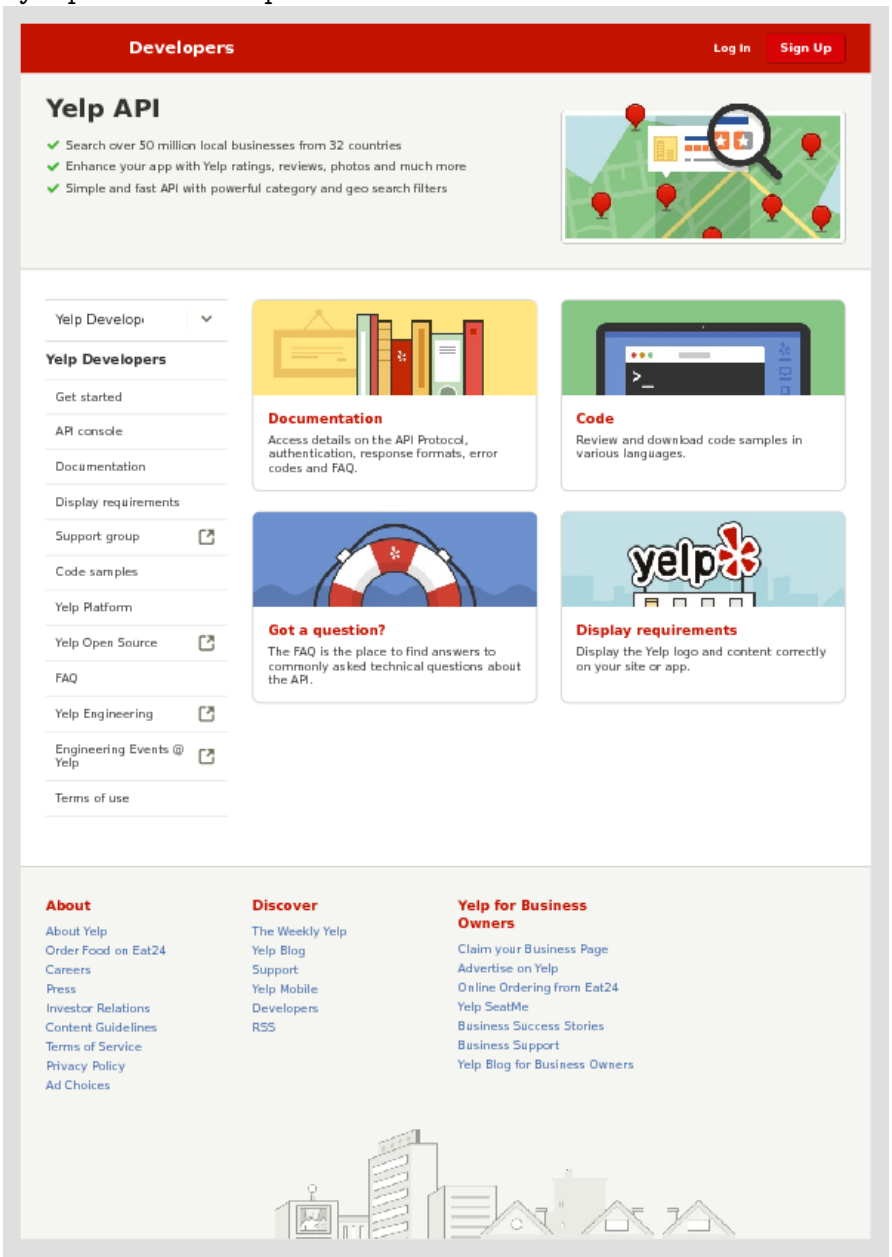


Figure 98: YELP API home page.

3.10 Weather

1. AccuWeather: The AccuWeather API provides subscribers access to location based weather data via a simple RESTful web interface (see Figure 99)
   `http://apidev.accuweather.com/developers/`

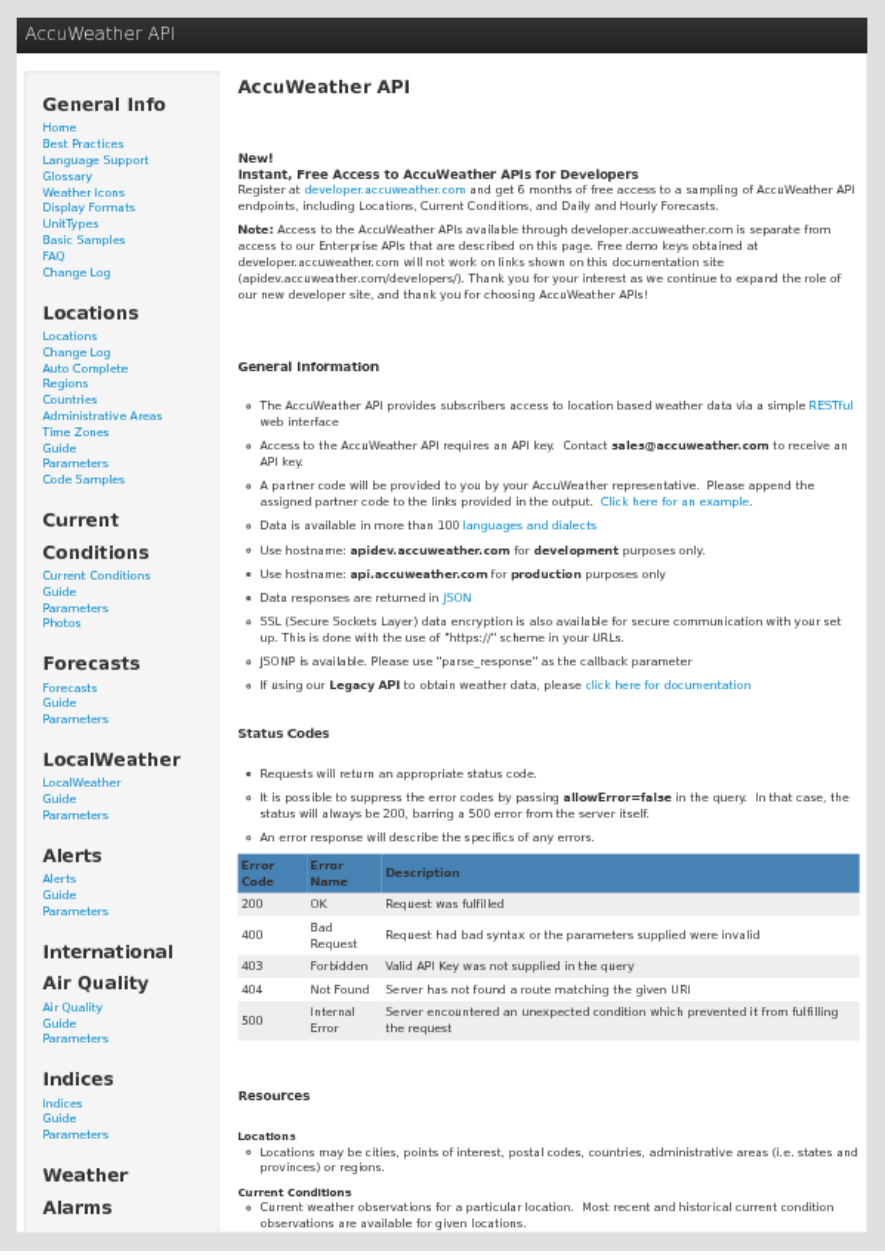


Figure 99: AccuWeather home page.

2. Aeris Weather: An advanced weather API to power all of your custom applications, offering a breath of fresh air from the basic to the most complex solutions. (see Figure 100)
`http://www.aerisweather.com/develop/`

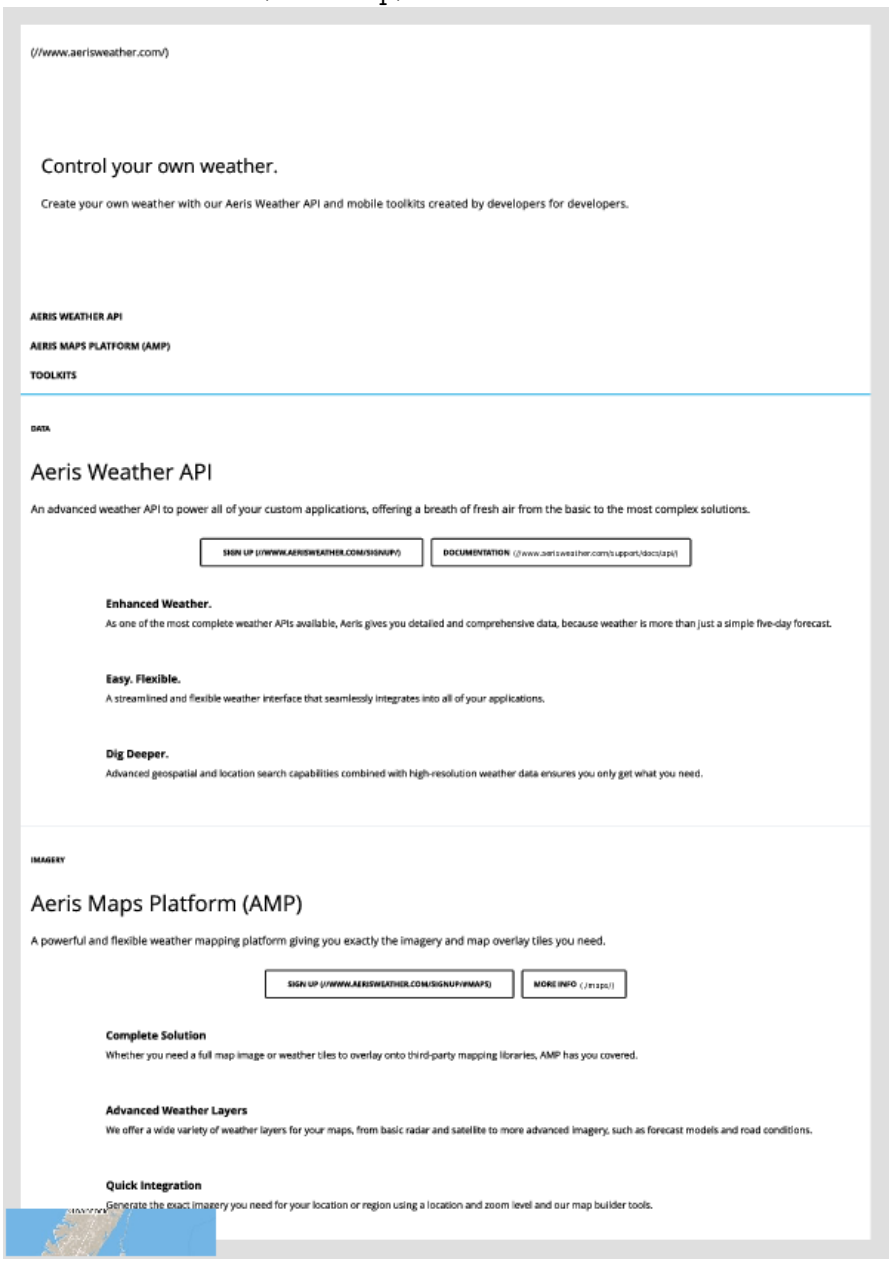


Figure 100: Aeris Weather home page.

3. Dark Sky Forecast: The same API that powers Forecast.io and Dark Sky for iOS can provide accurate shortterm and longterm weather predictions to your business, application, or crazy idea. (see Figure 101)
   `https://developer.forecast.io/`

Figure 101: Dark Sky Forecast home page.

4. National Data Buoy Center: The National Data Buoy Center (NDBC) is a part of the National Oceanic and Atmospheric Administration's (NOAA) National Weather Service (NWS). NDBC designs, develops, operates, and maintains a network of data collecting buoys and coastal stations. (see Figure 102)
`http://www.ndbc.noaa.gov/`

Figure 102: National Data Buoy Center home page.

5. National Weather Service: The National Weather Service is a component of the National Oceanic and Atmospheric Administration (NOAA). NOAA is an Operating Unit of the U.S. Department of Commerce. Our Mission Provide weather, water, and climate data, forecasts and warnings for the protection of life and property and enhancement of the national economy. (see Figure 103)
`http://www.weather.gov/`

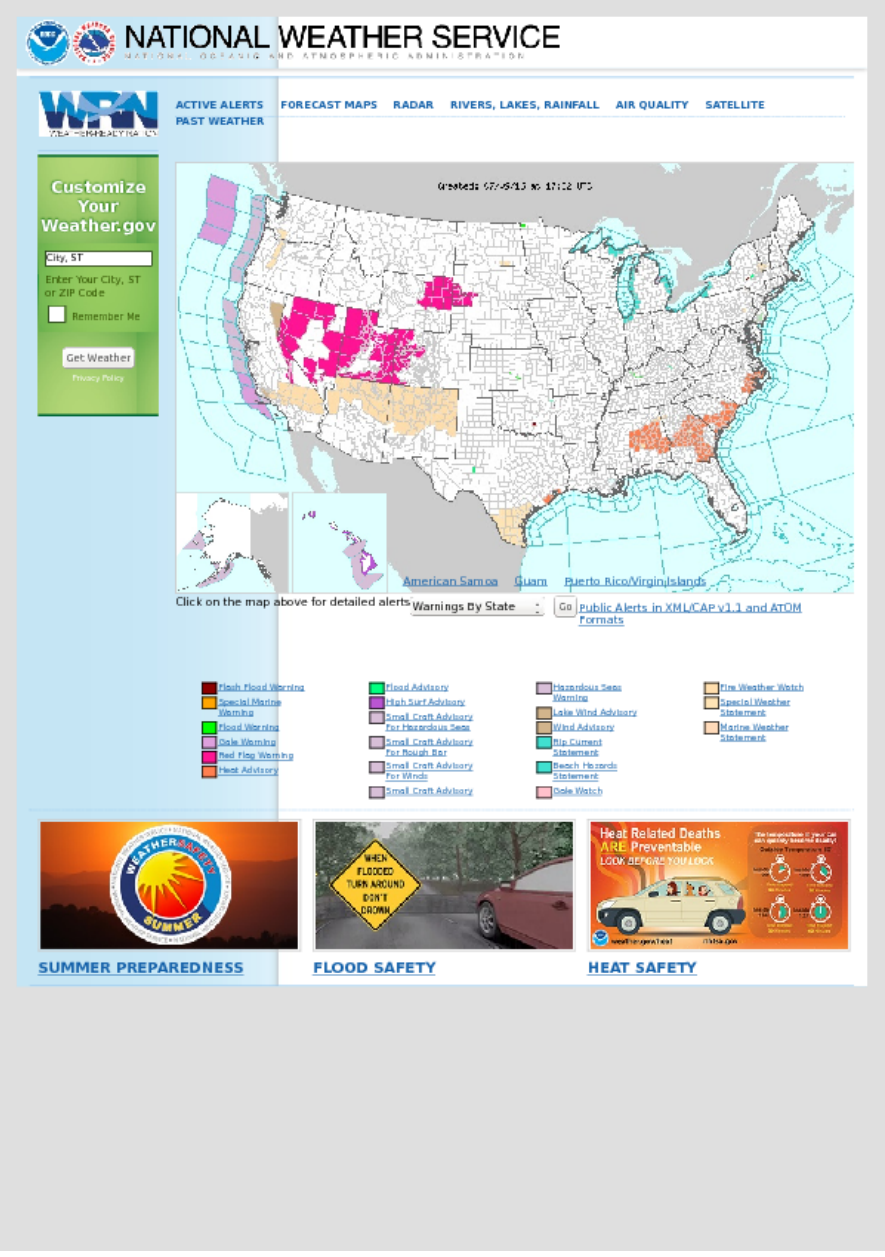


Figure 103: National Weather Service home page.

6. OpenWeatherMap: Our weather API is simple, clear and free. We also offer higher levels of support, please see our paid plan options. To access the API you need to sign up for an API key if you are on a free or paid plan. (see Figure 104)
`http://openweathermap.org/api`

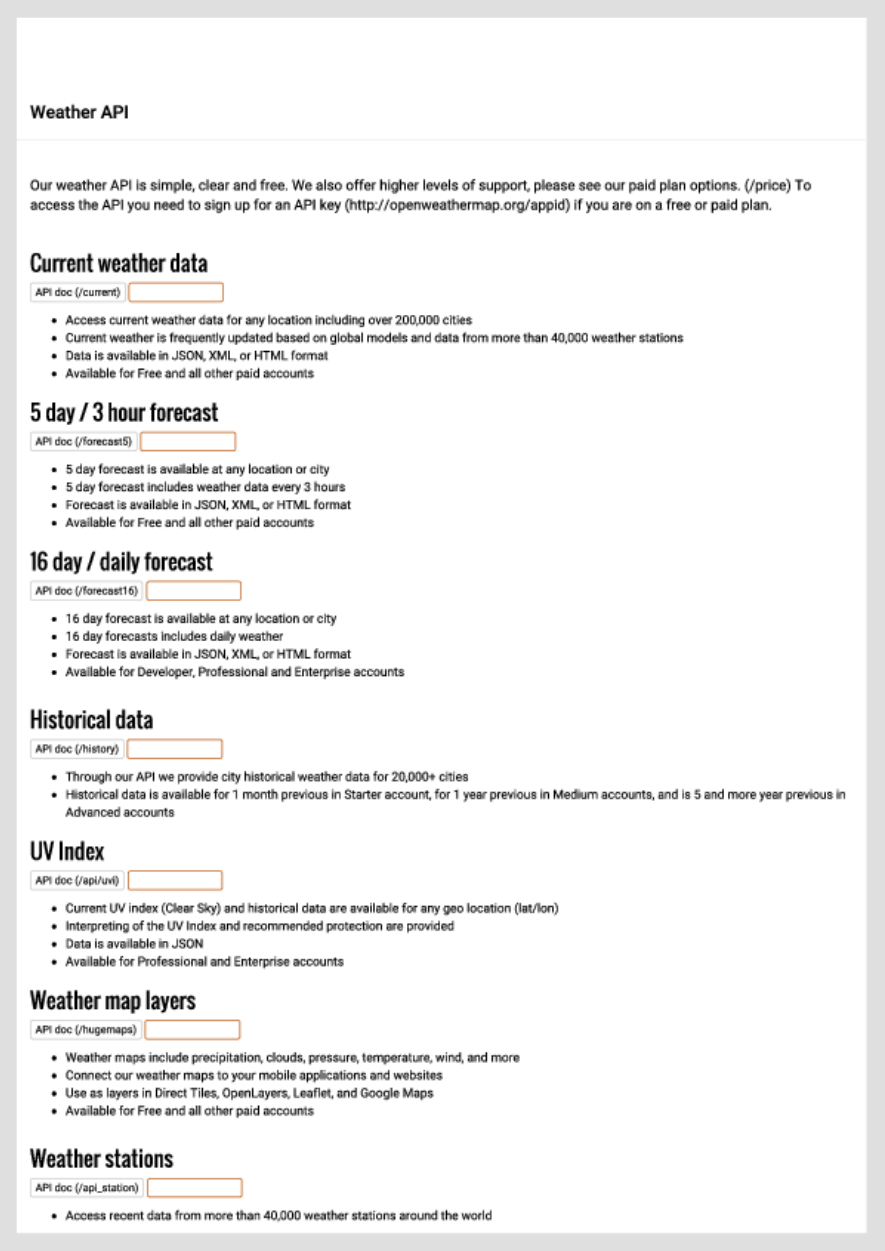


Figure 104: OpenWeatherMap home page.

7. Real-time weather forecasts: The easiest, most advanced, weather API on the web. (see Figure 105)
`https://developer.forecast.io/`

Figure 105: Real-time weather forecasts home page.

8. WeatherBug: With our APIs you can integrate the power of hyper localized data and the most comprehensive weather information into your products, services and apps. PulseAPI goes beyond typical weather data by providing the most reliable weather intelligence trusted by more than 40 million customers around the globe. We offer a wide range of options based on your volume and specific needs. Extend your product capabilities into a new dimension of intelligent weather offering with PulseAPIs. (see Figure 106)
`http://business.weatherbug.com/products/api-data-feeds/`

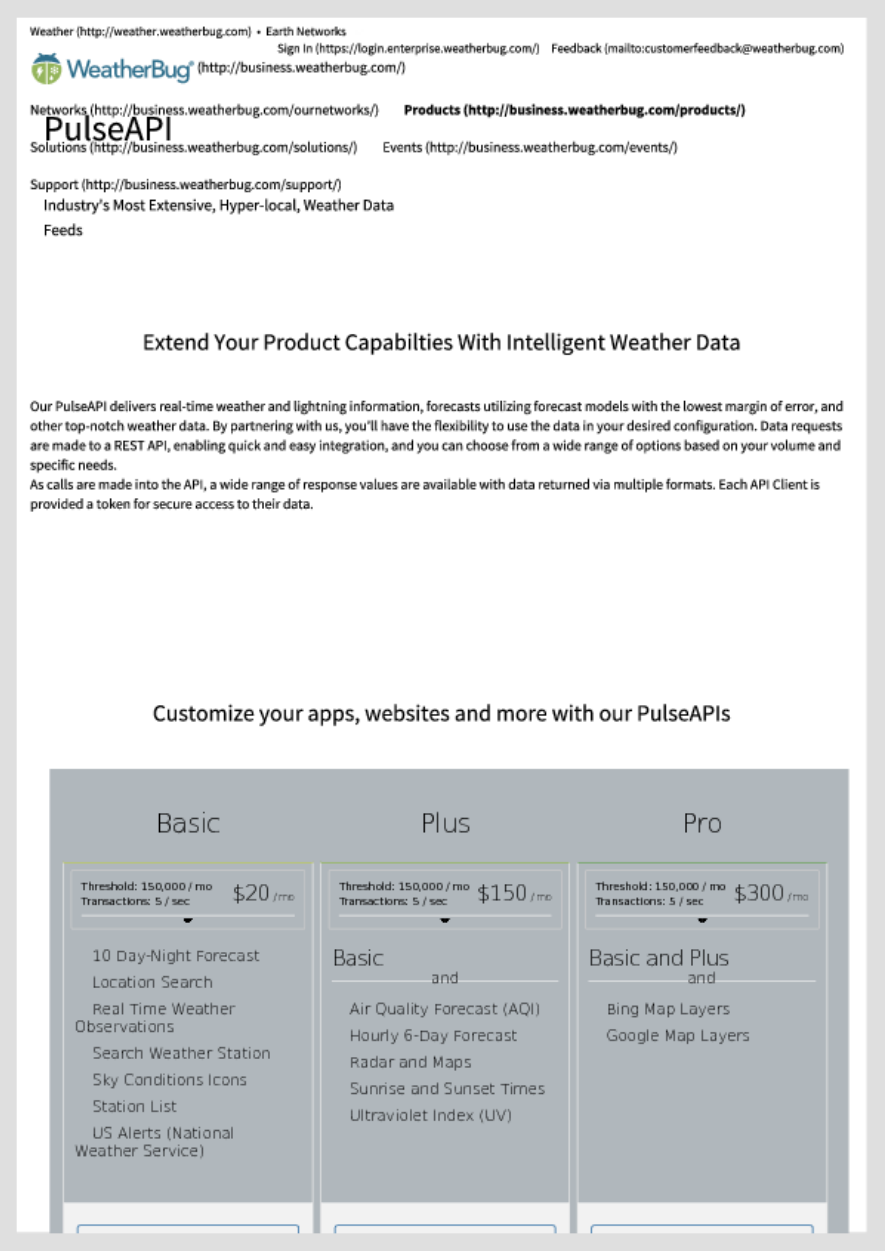


Figure 106: WeatherBug home page.

9. Weather Source: Powerful Weather API built systems that demand speed, reliability, and heavy load. (see Figure 107)
`http://weathersource.com/weather-api`

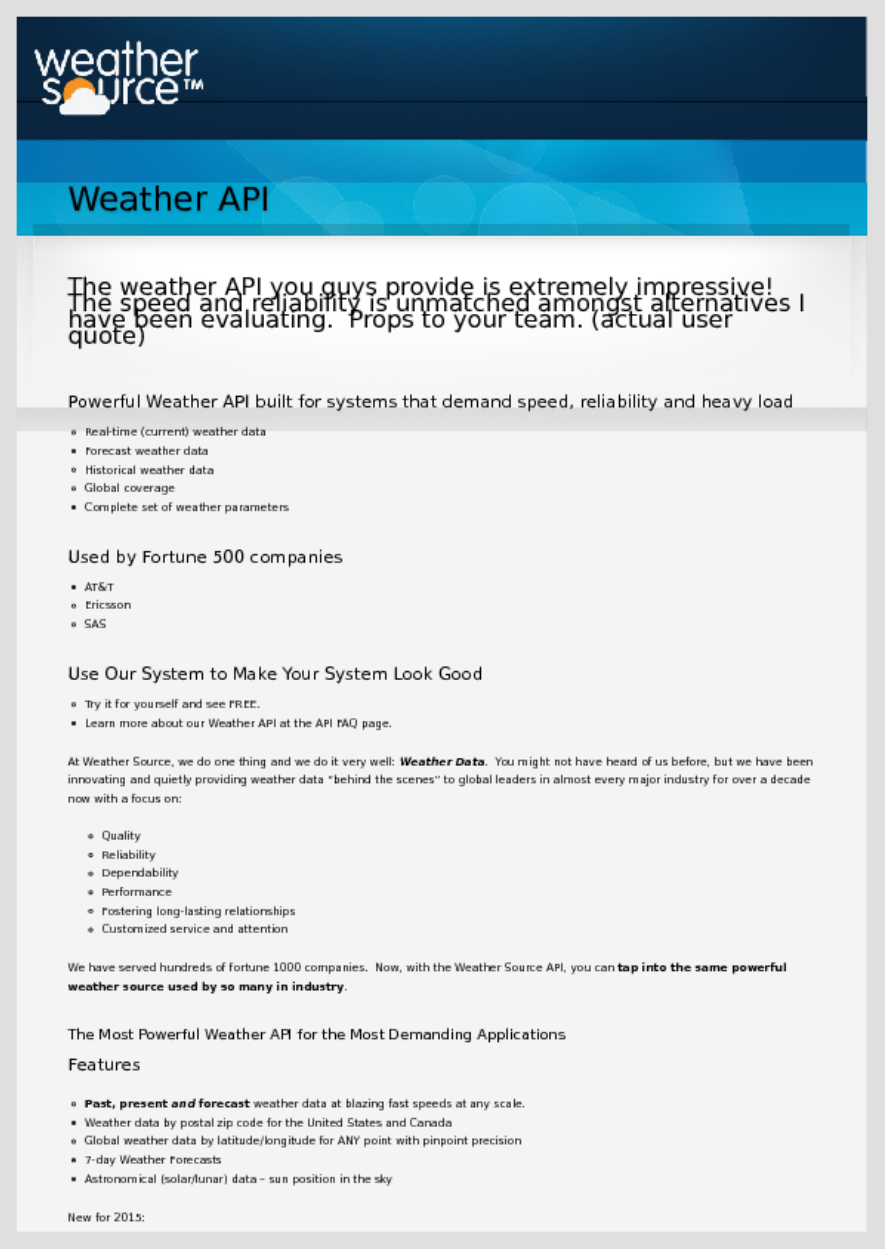


Figure 107: Weather Source home page.

10. Weather Underground: Reliable data, accurate forecast, & global coverage in 80 languages. (see Figure 108)
`http://www.wunderground.com/weather/api`

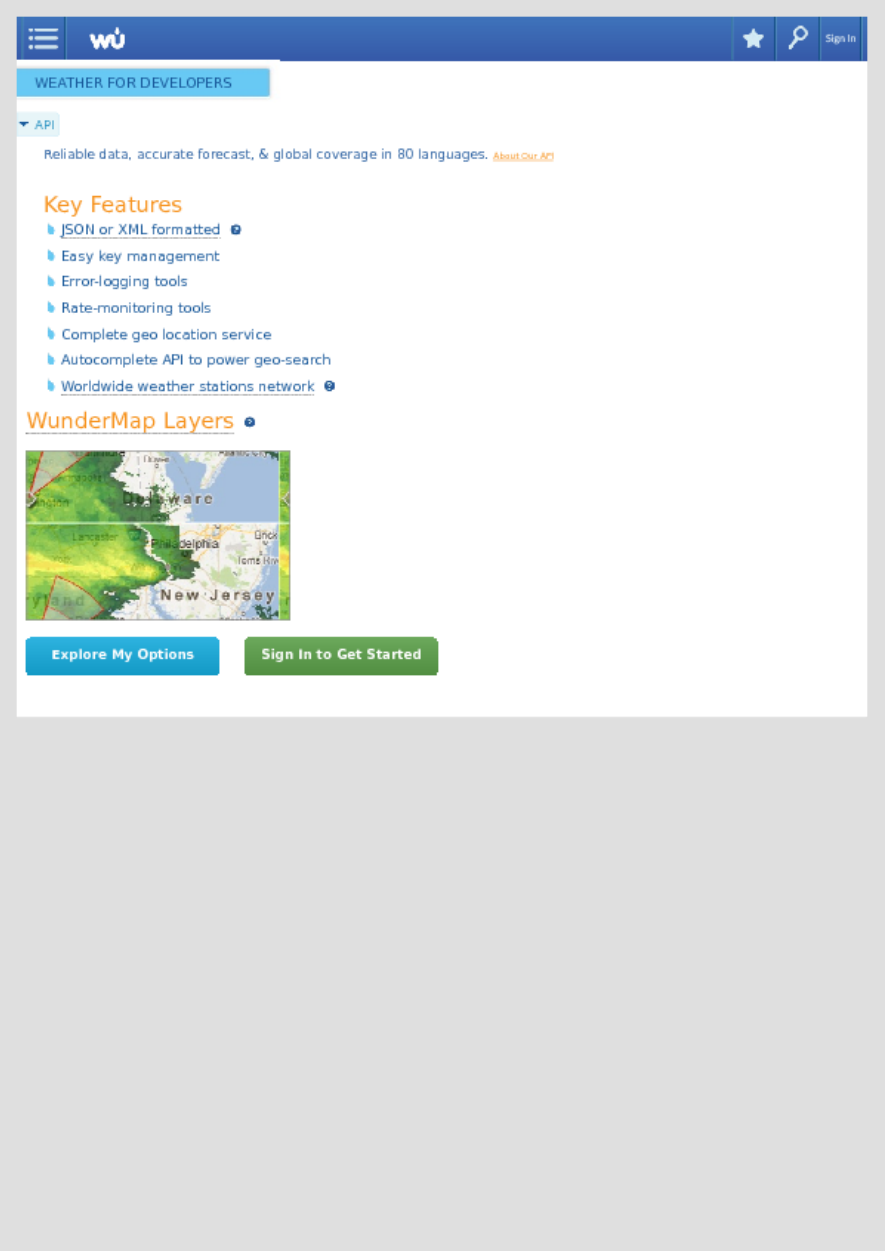


Figure 108: Weather Underground home page.

11. World Weather Online: World Weather Onlines weather API (application programming interface) allows developers and programmers to access current, past and future weather data for use in apps and on websites. (see Figure 109)
`https://developer.worldweatheronline.com/`

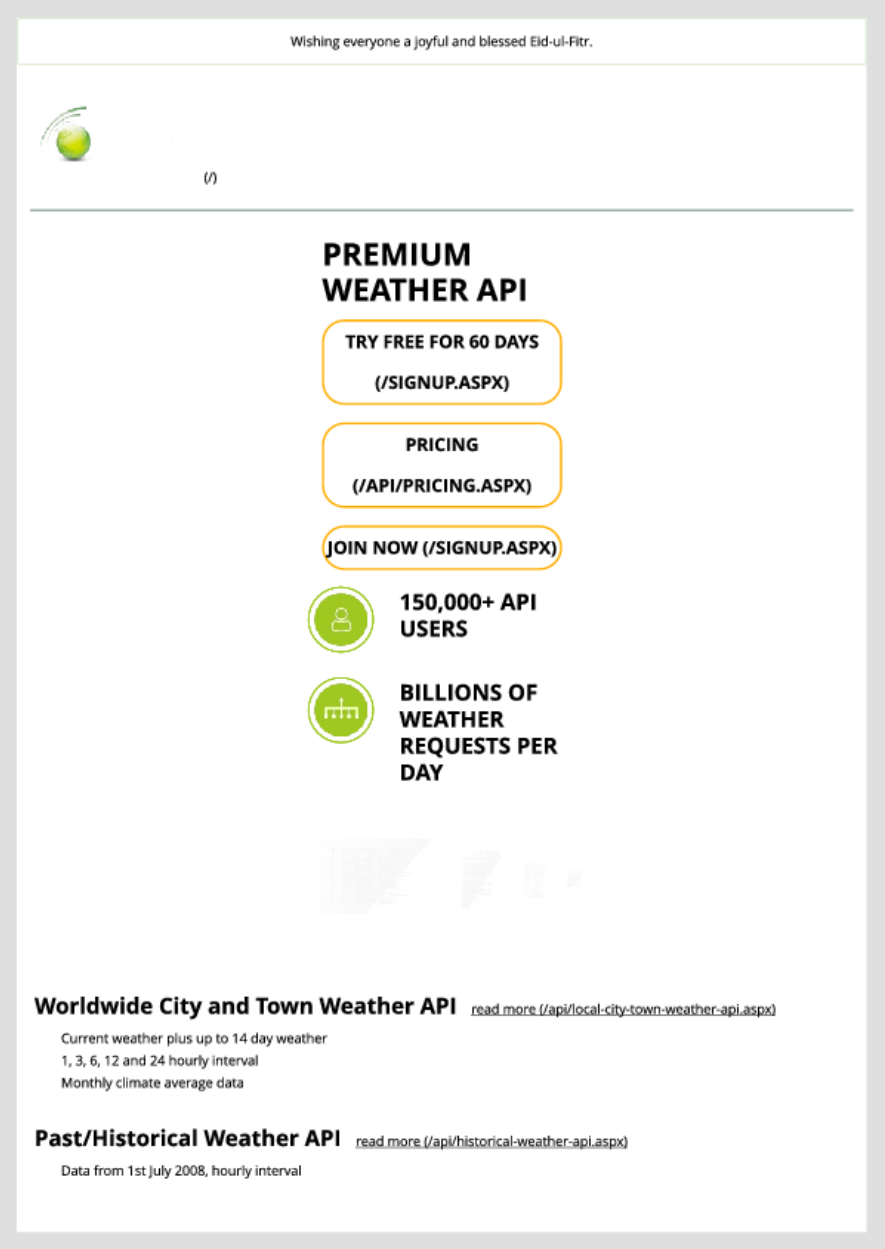


Figure 109: World Weather Online home page.

12. World Weather OnLine: Our weather API allows developers and programmers to embed weather data into their applications and websites. The local weather API provides access to present weather conditions and up to 15 days weather forecast across worldwide locations. (see Figure 110)
`http://us.worldweatheronline.com/api/local-city-town-weather-api.aspx`

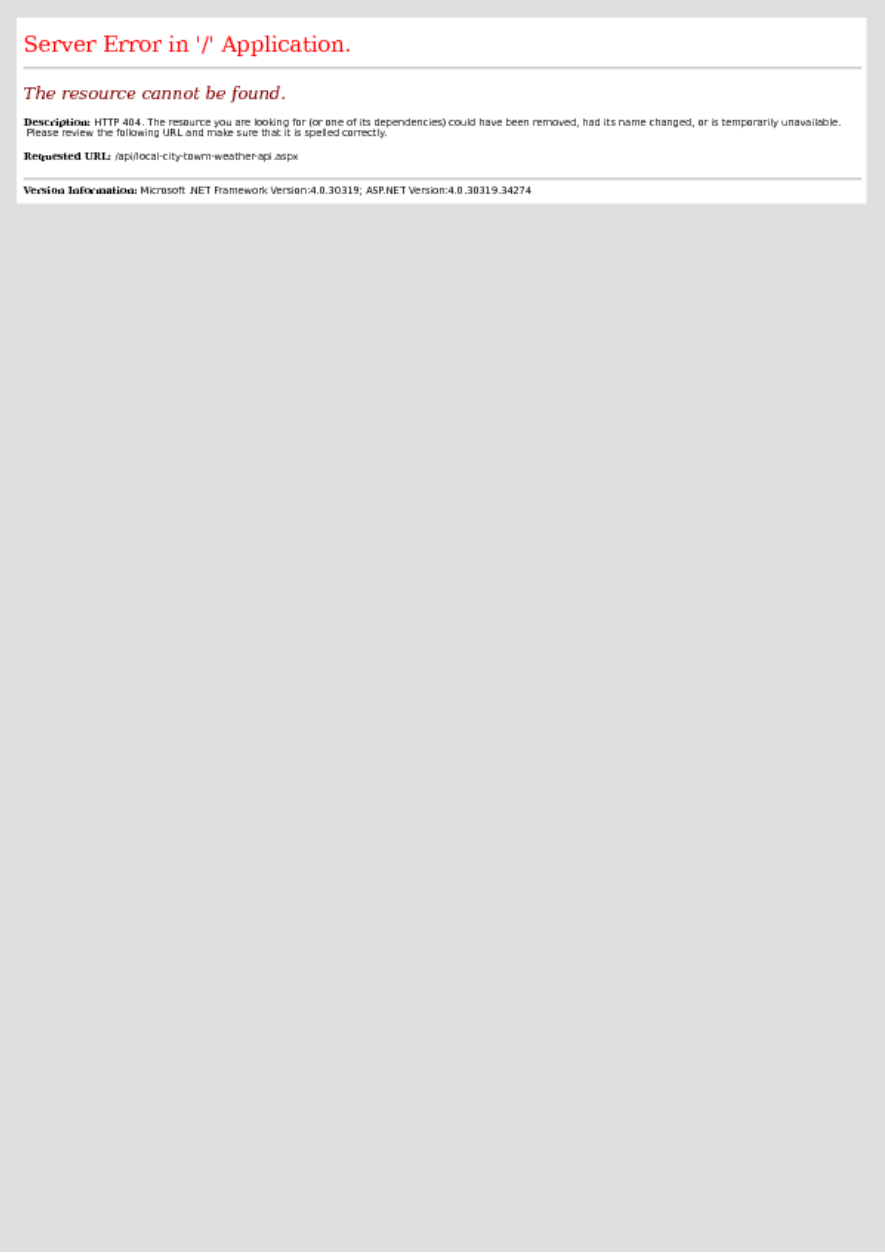


Figure 110: World Weather OnLine home page.

3.11 Zip code

1. Zip Codes: Our ZIP Code Database can be downloaded in both Microsoft Excel and CSV formats and easily opened in most spreadsheet applications or imported into the database software of your choosing. The data comes from authoritative sources such as the United States Postal Service (2011), US Census Bureau (2010), the Internal Revenue Service (2008), and Yahoo. (see Figure 111)
   `http://www.unitedstateszipcodes.org/`

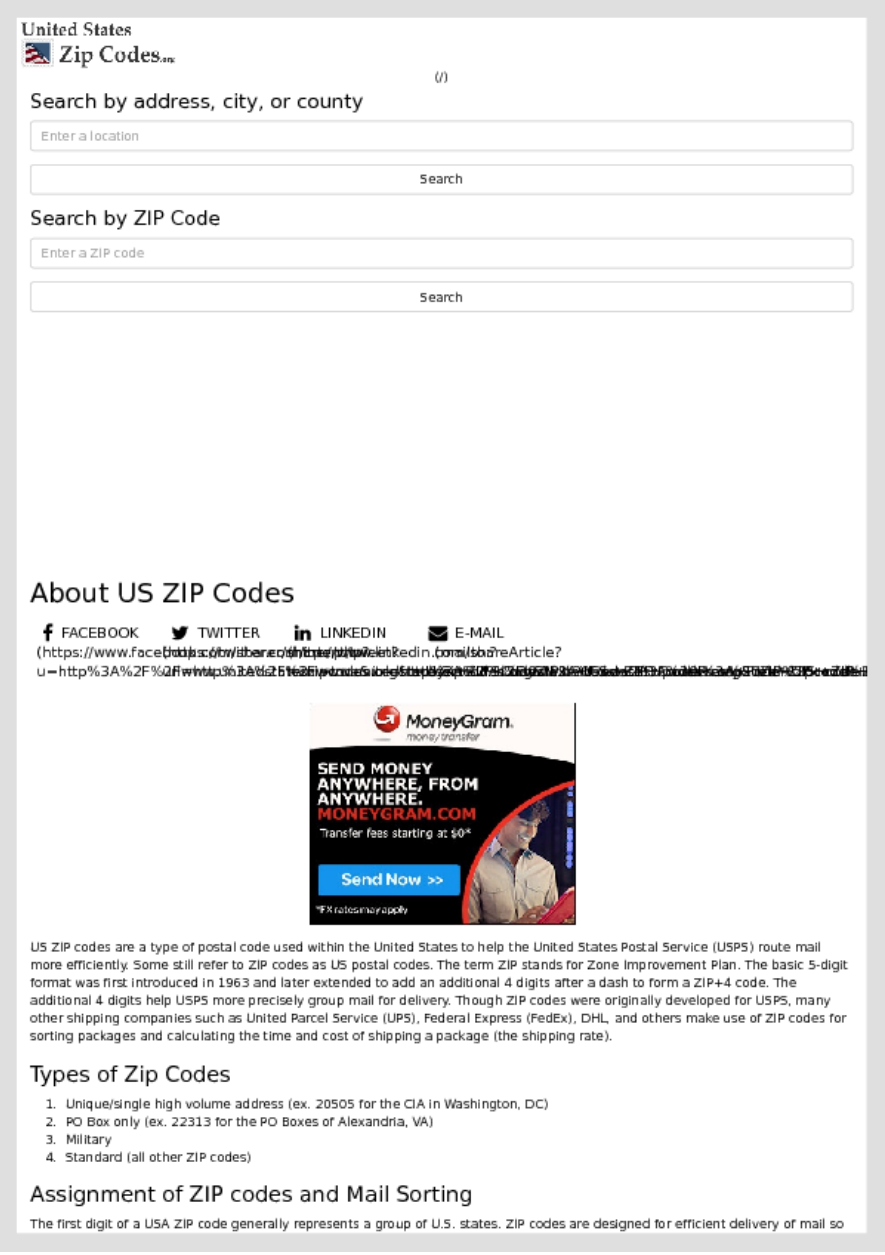


Figure 111: Zip Codes home page.

4 System performance

A graph showing system performance creating the home page images for all BD sources in this report has been created (see Figure 112).

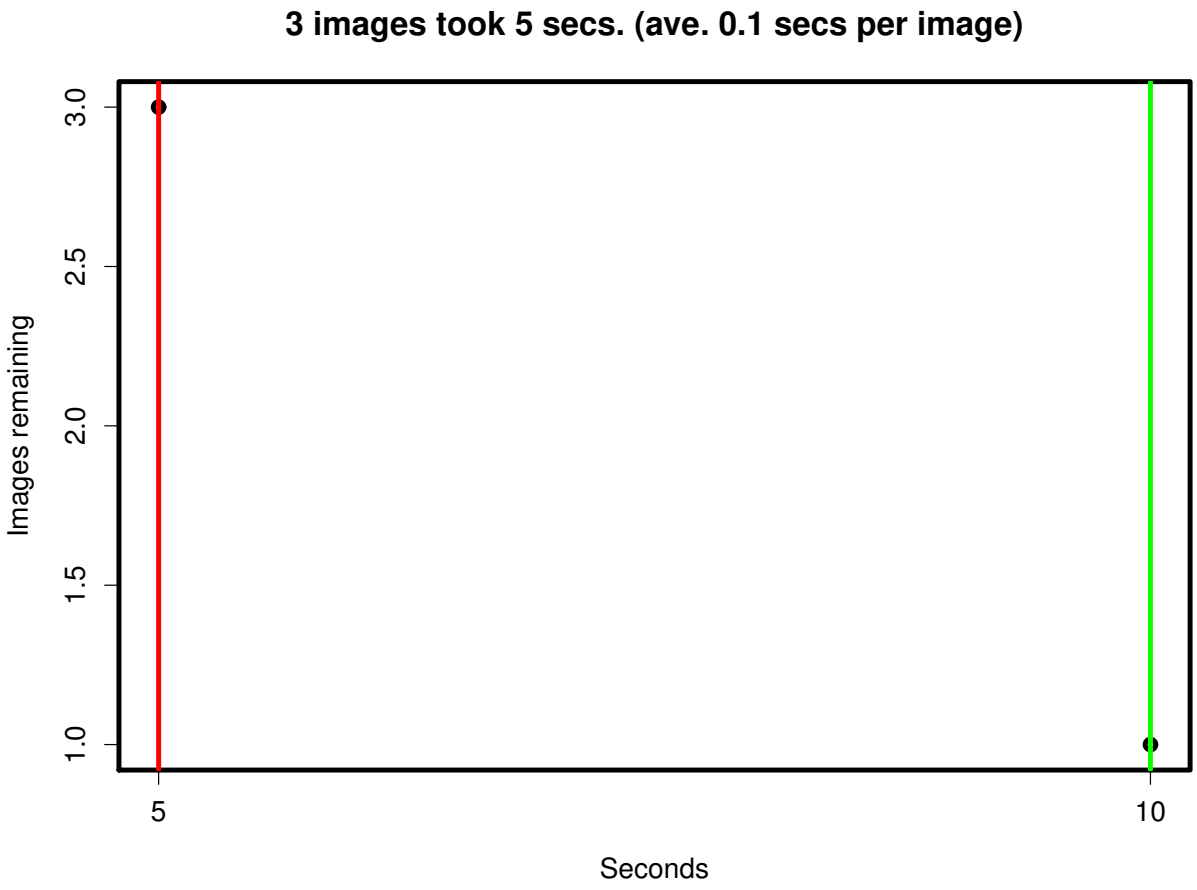


Figure 112: How long it took to create the home page images. The red lines are when 50 percent of the images were created. The green vertical line is when the first image was created.

5 References

[1] Nicholas Felton, `http://feltron.com/`, 2014.

[2] Viktor Mayer-Schönberger and Kenneth Cukier, *Big data: A revolution that will transform how we live, work, and think*, Houghton Mifflin Harcourt, 2013.

[3] Aaron Parecki, `http://aaronparecki.com/`, 2015.