# Airport Information Scraper

Alex Van Kooy
Amit Nijsure
Fernando Ramirez
Abdulaziz Alquzi

# Introduction and Purpose of Dataset

Airports around the world are mainly identified by two main identifiers or codes

1. International Civil Aviation Organization (ICAO) code – A four-letter code designating aerodromes around the world.
2. International Air Transport Association (IATA) code – A three-letter geocode designating many airports and metropolitan areas around the world.

Latitude and longitude information is a required attribute for any airport so that Map APIs can pinpoint the exact location of the airport.

The purpose of this dataset is using web scraping to collect information on airports around the world such as their names, city and country, IATA code, ICAO code and location coordinates

An umbrella objective would be to cross-reference the latitude and longitude information of these airports with the UFO Sightings dataset.

# Potential Users and Applications

The intent of this dataset is for it to be used as a lightweight cross reference source. The initial inspiration was to cross reference reported ‘UFO’ sightings to with airport locations.

Beyond debunking ‘UFO’ claims the dataset is still useful for a simple augmentation to any other dataset that contains similar geolocation information.

# Source of Data

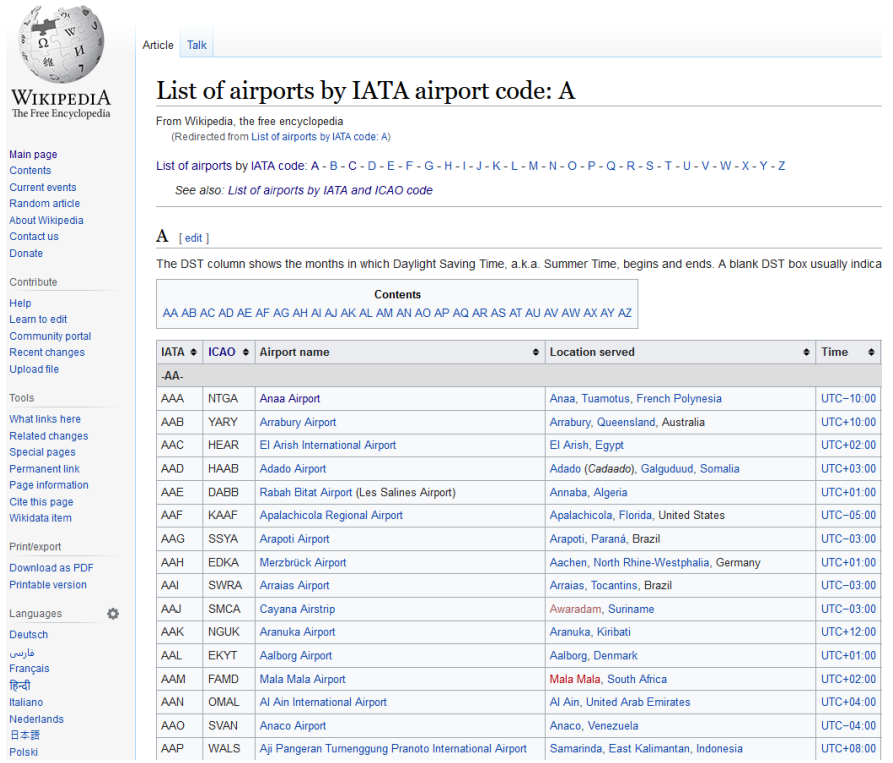
We are currently using Wikipedia as the source of all the data that needs to be scraped

The main purpose of using Wikipedia is that it is most widely updated information source there is on the internet.

Also we are using a publicly available OpenFlights Dataset in order to look for missing information.

# Approach to Acquiring the Data

- Public, standardized, and non-controversial information?
  - Wikipedia!
- Other sources
  - Provide similar information
  - More difficult to gain access (FAA)
- Our plan
  - Scrape(bs4) the IATA pages for basic information
  - Use the airport URLs to collect location data.
  - Check for the 429 rate limit alert in responses



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export
Download as PDF
Printable version

Languages
Deutsch
فارسی
Français
हिन्दी
Italiano
Nederlands
日本語
Polski

Article Talk

## List of airports by IATA airport code: A

From Wikipedia, the free encyclopedia
(Redirected from List of airports by IATA code: A)

List of airports by IATA code: A - B - C - D - E - F - G - H - I - J - K - L - M - N - O - P - Q - R - S - T - U - V - W - X - Y - Z

*See also: List of airports by IATA and ICAO code*

### A [ edit ]

The DST column shows the months in which Daylight Saving Time, a.k.a. Summer Time, begins and ends. A blank DST box usually indica

Contents
AA AB AC AD AE AF AG AH AI AJ AK AL AM AN AO AP AQ AR AS AT AU AV AW AX AY AZ

| IATA | ICAO | Airport name | Location served | Time |
|---|---|---|---|---|
| -AA- | | | | |
| AAA | NTGA | Anaa Airport | Anaa, Tuamotus, French Polynesia | UTC−10:00 |
| AAB | YARY | Arrabury Airport | Arrabury, Queensland, Australia | UTC+10:00 |
| AAC | HEAR | El Arish International Airport | El Arish, Egypt | UTC+02:00 |
| AAD | HAAB | Adado Airport | Adado (*Cadaado*), Galguduud, Somalia | UTC+03:00 |
| AAE | DABB | Rabah Bitat Airport (Les Salines Airport) | Annaba, Algeria | UTC+01:00 |
| AAF | KAAF | Apalachicola Regional Airport | Apalachicola, Florida, United States | UTC−05:00 |
| AAG | SSYA | Arapoti Airport | Arapoti, Paraná, Brazil | UTC−03:00 |
| AAH | EDKA | Merzbrück Airport | Aachen, North Rhine-Westphalia, Germany | UTC+01:00 |
| AAI | SWRA | Arraias Airport | Arraias, Tocantins, Brazil | UTC−03:00 |
| AAJ | SMCA | Cayana Airstrip | Awaradam, Suriname | UTC−03:00 |
| AAK | NGUK | Aranuka Airport | Aranuka, Kiribati | UTC+12:00 |
| AAL | EKYT | Aalborg Airport | Aalborg, Denmark | UTC+01:00 |
| AAM | FAMD | Mala Mala Airport | Mala Mala, South Africa | UTC+02:00 |
| AAN | OMAL | Al Ain International Airport | Al Ain, United Arab Emirates | UTC+04:00 |
| AAO | SVAN | Anaco Airport | Anaco, Venezuela | UTC−04:00 |
| AAP | WALS | Aji Pangeran Tumenggung Pranoto International Airport | Samarinda, East Kalimantan, Indonesia | UTC+08:00 |

# Approach to Preprocessing Data

- Information is standardized by international agreements, so very little preprocessing is required.
- Our focus was on the Latitude and Longitude information which had it's own set of problems:
  - No accuracy standard. Some are defined to the decimal seconds and others only to the minute.
  - Minute and second marks (′ and ″ ) are not quotes but prime ticks in Unicode.
  - Direction is a N/S E/W convention
- The above combination raised the concern of useability and clarity. We decided to convert the string notation to a signed integer.
- Any airport that did not have a page in Wikipedia was searched for using another public, precompiled airport dataset for their Lat/Long. The number of airports missing information in both were noted in the README file to provide the most complete and useable dataset possible.

# Distribution Approach


- After completion and being compiled (Wiki & Github -open flight data) the dataset can be accessible as open-source.
  - Csv. File can be obtainable from GitHub or Kaggle for further data analysis from pre-processed data collection.
- Rights for individuals
  - Data is open-source based on user profile and commits (unique ID)
  - Easily publicly accessible for "shared tasks" for groups interested in cross-examining flight patterns.
- README & system requirements

# Discussion of Access Rights

- Wiki is great because of its openness!
  - Data is free and users are encouraged to obtain/explore datasets
- Broad language in robots.txt stating that web-crawlers or spiders may be blocked if "that go way too fast"
  - Limit of 429 rate limit responses was predetermined
- Individual airport wiki pages were able to be successfully scrapped.
- Dataset has a high reproducibility level and can be generated with minimal criteria (no -access keys or tokens)

# Issues and Limitations

One of the main issue faced while scraping data from the Wikipedia pages was the dead or problematic links for some airports.

These airports have been collected in a separate data structure by our code and set aside for fetching the latitude and longitude information from some other data sources.

Information of some airports which could not be collected from Wikipedia was fetched from the OpenFlights Dataset. More work needs to be done to fetch the same for the rest of the airports.

# Team and Contributions

- Alex
  - Built the functionality to parse and collect the information on the IATA page then navigate to the individual airport pages to convert and collect the Lat/Long information.
- Amit
  - Iterating the skeleton data collection code over all 26 alphabets to scrape information for airports
  - Collecting miscellaneous data for each alphabet such as size of collected data, number of missing values filled using other sources.
  - Preparing metadata document (README.md)
- Fernando
  - Collection of alphabetical airports, post-bs4 processing, using csv.dictwriter of scrapped wiki airports
  - Collection served to contribute as main dataset file
- Abdulaziz
  - Iterating through the main dataset to find the missing data.
  - Collecting the missing data from the other dataset, refill it to the main dataset.