Article
Henrique Dias · Jan 13, 2022 4m read

Open Exchange

How to find the dataset you need?

Hey community! How are you doing?

I hope to find everyone well, and a happy 2022 to all of you!

Over the years, I've been working on a lot of different projects, and I've been able to find a lot of interesting data.

But, most of the time, the dataset that I used to work with was the customer data. When I started to join the contest in the past couple of years, I began to look for specific web datasets.

I've curated a few data by myself, but I was thinking, "This dataset is enough to help others?"

So, discussing the ideas for this contest with @José Pereira , we decided to approach this contest using a *different perspective*.

We thought of offering a variety of datasets of any kind from two famous data sources. This way, we can be empowering the users to find and install the desired dataset in a quick and easy way.

Socrata

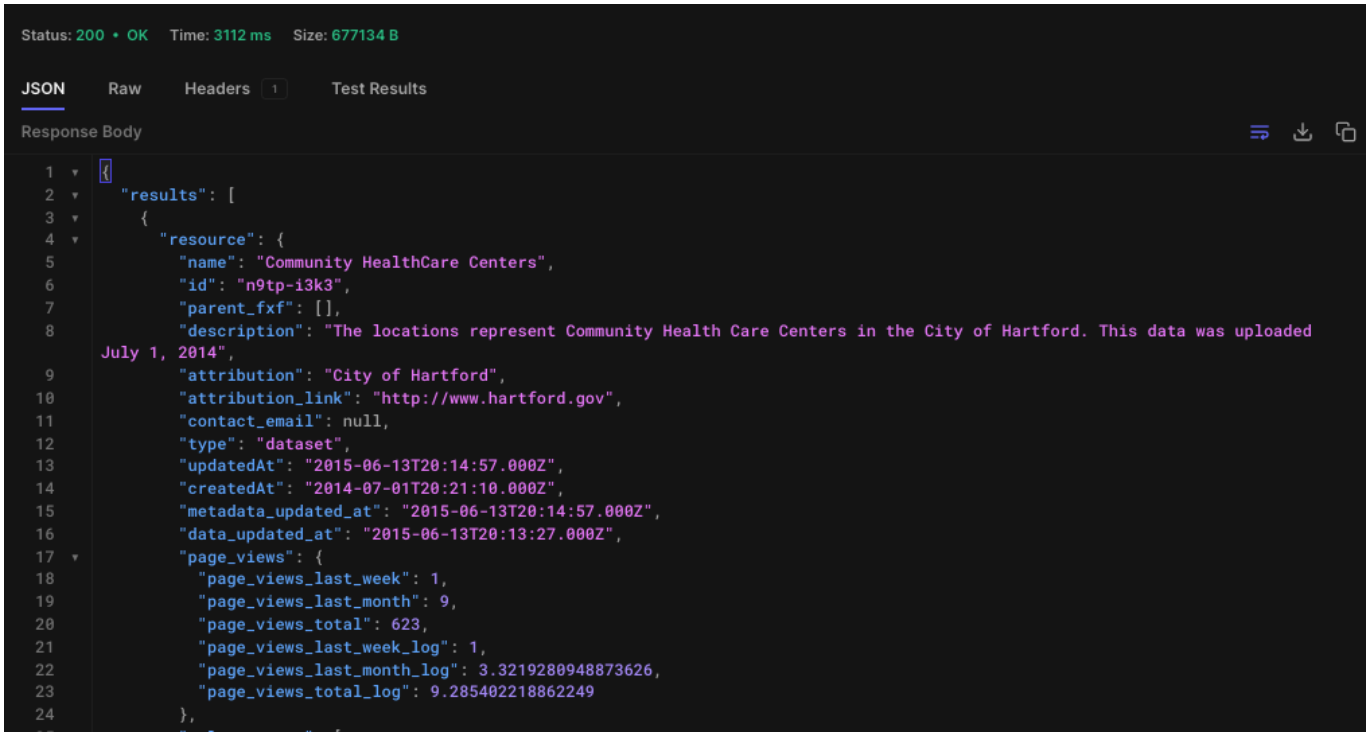
The Socrata Open Data API allows you to programmatically access a wealth of open data resources from governments, non-profits, and NGOs around the world.

For this initial release, we are using Socrata APIs to search and download and speficic dataset.

Open the API tool of your preference like Postman, Hoppscotch

```
GET> https://api.us.socrata.com/api/catalog/v1?only=dataset&q=healthcare
```

This endpoint will return all healthcare related datasets, like the image below:



Now, get the ID. In this case the id is: "n9tp-i3k3"

Go the the terminal

```
IRISAPP>set api = ##class(dc.dataset.importer.service.socrata.SocrataApi).%New()

IRISAPP>do api.InstallDataset({"datasetId": "n9tp-i3k3", "verbose":true})

Compilation started on 01/07/2022 01:01:28 with qualifiers 'cuk'
Compiling class dc.dataset.imported.DsCommunityHealthcareCenters
Compiling table dc_dataset_imported.DsCommunityHealthcareCenters
Compiling routine dc.dataset.imported.DsCommunityHealthcareCenters.1
Compilation finished successfully in 0.108s.

Class name: dc.dataset.imported.DsCommunityHealthcareCenters
Header: Name VARCHAR(250),Description VARCHAR(250),Location VARCHAR(250),Phone_Number
 VARCHAR(250),geom VARCHAR(250)
Records imported: 26
```

After the command above, your dataset it's ready to use!

Kaggle

Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

In June 2017, Kaggle announced that it passed 1 million registered users, or Kagglers, and as of 2021 has over **8 million registered users**. The community spans **194 countries**. It is a diverse community, ranging from those just starting out to many of the world's best known researchers.

This what I call a huge community, right?!

To use the datasets from Kaggle, you need to register on the website. After that, you need to create an API token to use Kaggle's API.

Now, just like with Socrata, you can use the API to search and download the dataset.

```
GET> https://www.kaggle.com/api/v1/datasets/list?search=appointments
```

Now, get the ref value. In this case the ref is: "joniarroba/noshowappointments"

The parameters below "*your-username*", and "*your-password*" are the parameters provided by Kaggle when you create the API token.

```
IRISAPP>Set crendtials = ##class(dc.dataset.importer.service.CredentialsService).%New
()

IRISAPP>Do crendtials.SaveCredentials("kaggle", "<your-username>", "<your-password>")

IRISAPP>Set api = ##class(dc.dataset.importer.service.kaggle.KaggleApi).%New()

IRISAPP>Do api.InstallDataset({"datasetId":"joniarroba/noshowappointments", "credenti
als":"kaggle", "verbose":true})

Class name: dc.dataset.imported.DsNoshowappointments
Header: PatientId INTEGER,AppointmentID INTEGER,Gender VARCHAR(250),ScheduledDay DATE
,AppointmentDay DATE,Age INTEGER,Neighbourhood VARCHAR(250),Scholarship INTEGER,Hiper
tension INTEGER,Diabetes INTEGER,Alcoholism INTEGER,Handcap INTEGER,SMS_received INTE
GER,No-show VARCHAR(250)
Records imported: 259
```

After the command above, your dataset it's ready to use!

Graphic User Interface

We're offering a GUI to install the dataset to make things easier. But this is something that we like to discuss in our next article. In the meanwhile, you can check a sneak peek below while we are polishing a few things before the official release:

Video Demo

How is the behavior of downloading a bigger dataset? +400.000 records aren't enough?! How about 1 MILLION RECORDS?! Let's see it!

https://www.youtube.com/embed/0T8wXRsaJso
[This is an embedded link, but you cannot view embedded content directly on the site because you have declined the cookies necessary to access it. To view embedded content, you would need to accept all cookies in your ***Cookies Settings****]*

Voting

If you liked the app and think we deserve your vote, please vote for **iris-kaggle-socrata-generator**!

https://openexchange.intersystems.com/contest/current

#Contest #Data Import and Export #Data Model #Unstructured Data #InterSystems IRIS #InterSystems IRIS for Health
Check the related application on InterSystems Open Exchange