

# Accessing Digital Collections Data Sources for Research: A Tour of iDigBio Data Services

Matthew Collins - iDigBio

[mcollins@acis.ufl.edu](mailto:mcollins@acis.ufl.edu)

Deborah Paul - iDigBio

[dpaul@fsu.edu](mailto:dpaul@fsu.edu)



Island Biology 2016, University of the Azores at Angra do Heroísmo

# What's in iDigBio's Repository?

## Vouchered specimen records!

- Metadata in Darwin Core, Audubon Core, and user-defined fields
- Images of specimens
- All groups including paleo
- US and international



List	Labels	Media	Recordsets	Total: 28,164		
Family	Scientific Name	Date Collected	Country	Institution Code	Basis of Record	Columns
ACERACEAE	Acer abchasicum Rupr.	no data	Brasil	Rio de Janeiro Botani...	PreservedSpecimen	<a href="#">view</a>
ACERACEAE	Acer abchasicum Rupr.	no data	Brasil	Rio de Janeiro Botani...	PreservedSpecimen	<a href="#">view</a>
Sapindaceae	Acer acuminate	no data	United States	UCMP	PreservedSpecimen	<a href="#">view</a>
Aceraceae	Negundo aceroides	1850-10-5	no data	MSC	PreservedSpecimen	<a href="#">view</a>
Aceraceae	Negundo aceroides	1887-5-4	United States	MSC	PreservedSpecimen	<a href="#">view</a>
Aceraceae	Acer negundo L. ssp. ...	1891-06-20	USA	UConn	Preserved Specimen	<a href="#">view</a>
Aceraceae	Negundo aceroides	1895-4-27	United States	MSC	PreservedSpecimen	<a href="#">view</a>
Aceraceae	Negundo aceroides	1895-8-25	United States	MSC	PreservedSpecimen	<a href="#">view</a>
Aceraceae	Negundo aceroides	no data	no data	MSC	PreservedSpecimen	<a href="#">view</a>
ACERACEAE	Acer acinatum Hort.	no data	Brasil	Rio de Janeiro Botani...	PreservedSpecimen	<a href="#">view</a>
Sapindaceae	Acer acuminatum Wal...	1893-07-20	United States	MO	PreservedSpecimen	<a href="#">view</a>
Sapindaceae	Acer acuminatum	no data	India	UCMP	PreservedSpecimen	<a href="#">view</a>
ACERACEAE	Acer acuminatum Wal...	no data	Brasil	Rio de Janeiro Botani...	PreservedSpecimen	<a href="#">view</a>

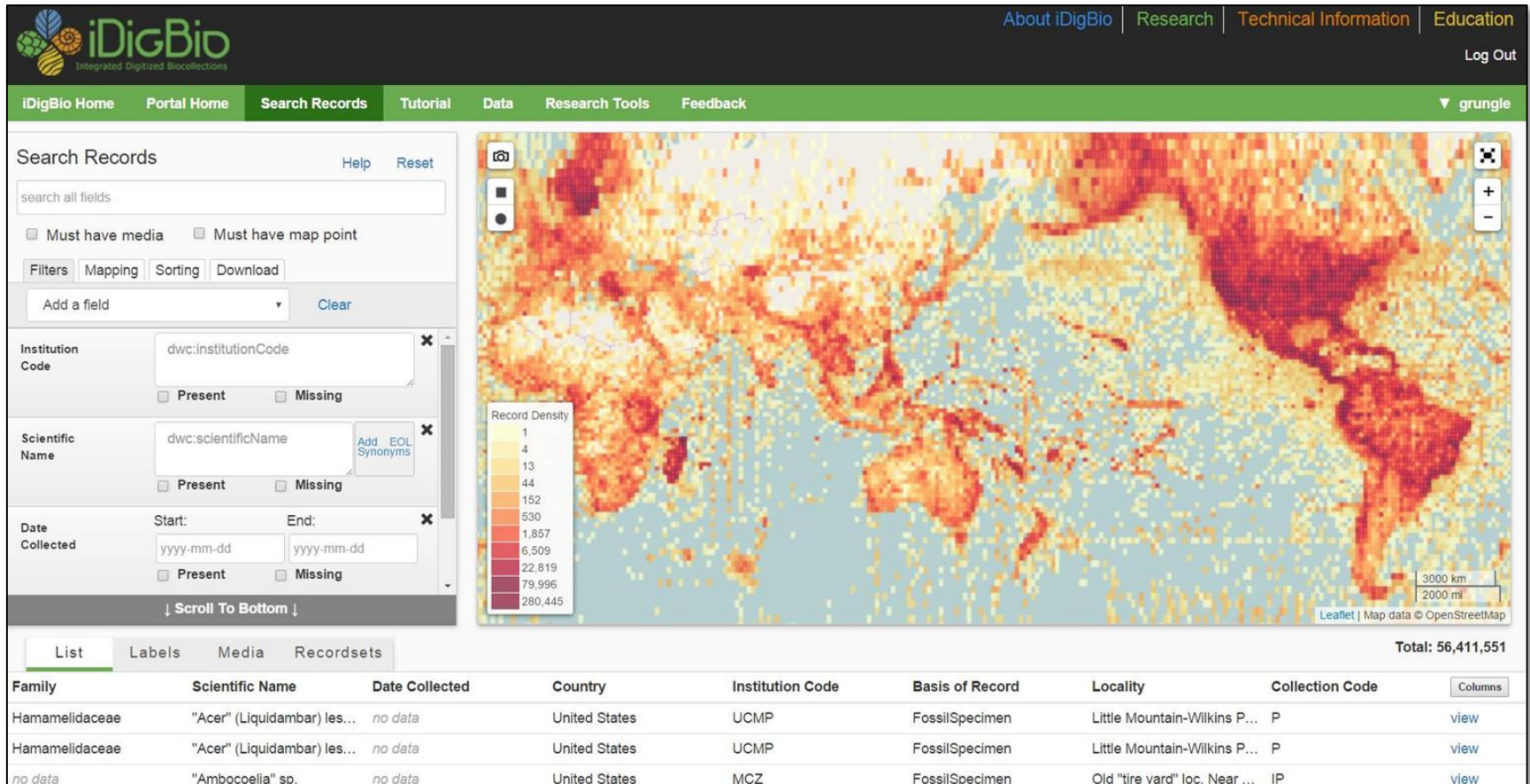


# Accessing iDigBio Data

Advantages of the different ways

1. **Website/portal** - good to see what's there
2. **Downloads** - flexible analysis, permanence of your custom data set
3. **API** - exploration of data, building applications
4. **Packages** - easier than direct use of API, literate programming, notebooks, exploration
5. **Computation** - big data, cross-dataset analysis

# Seeing What's There - The Portal



The screenshot displays the iDigBio portal interface. At the top, there is a navigation bar with links for 'About iDigBio', 'Research', 'Technical Information', and 'Education', along with a 'Log Out' button. Below this is a secondary navigation bar with 'iDigBio Home', 'Portal Home', 'Search Records', 'Tutorial', 'Data', 'Research Tools', and 'Feedback'. A 'grungle' dropdown menu is also visible.

The main content area is divided into two sections. On the left is the 'Search Records' panel, which includes a search input field, filter options for 'Must have media' and 'Must have map point', and buttons for 'Filters', 'Mapping', 'Sorting', and 'Download'. It also features a 'Add a field' dropdown and a 'Clear' button. Below these are three filter sections: 'Institution Code' (with a text input 'dwc:institutionCode' and 'Present'/'Missing' checkboxes), 'Scientific Name' (with a text input 'dwc:scientificName', an 'Add EOL Synonyms' button, and 'Present'/'Missing' checkboxes), and 'Date Collected' (with 'Start' and 'End' date inputs and 'Present'/'Missing' checkboxes). A 'Scroll To Bottom' button is at the bottom of this panel.

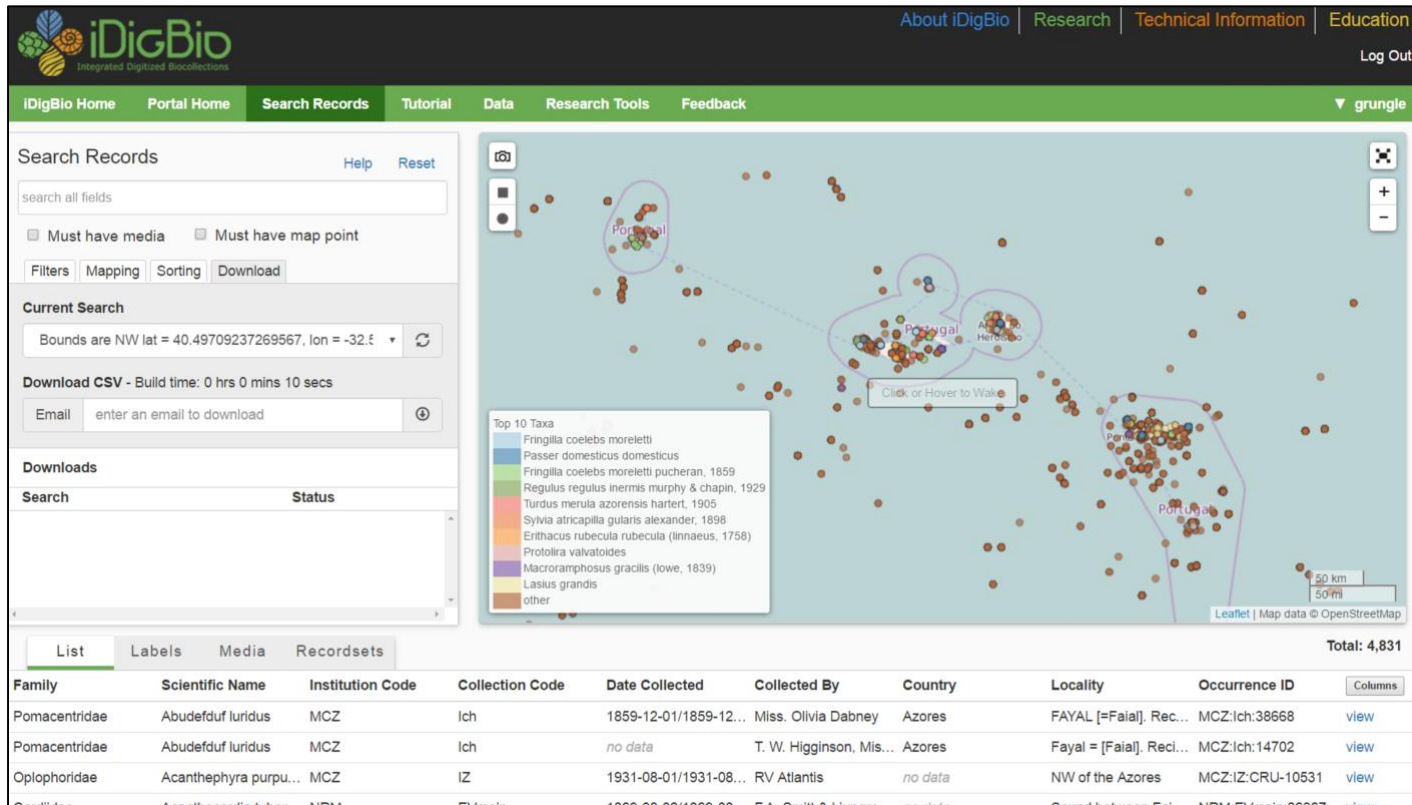
On the right is a world map showing record density. A legend titled 'Record Density' provides a color scale from light yellow (1) to dark red (280,445). The map includes zoom controls and a scale bar (3000 km / 2000 mi). Attribution text at the bottom right of the map reads 'Leaflet | Map data © OpenStreetMap'.

Below the map is a tabbed interface with 'List', 'Labels', 'Media', and 'Recordsets' tabs. The 'List' tab is active, showing a table with the following columns: Family, Scientific Name, Date Collected, Country, Institution Code, Basis of Record, Locality, Collection Code, and Columns. The total number of records is displayed as 'Total: 56,411,551'.

Family	Scientific Name	Date Collected	Country	Institution Code	Basis of Record	Locality	Collection Code	Columns
Hamamelidaceae	"Acer" (Liquidambar) les...	no data	United States	UCMP	FossilSpecimen	Little Mountain-Wilkins P...	P	view
Hamamelidaceae	"Acer" (Liquidambar) les...	no data	United States	UCMP	FossilSpecimen	Little Mountain-Wilkins P...	P	view
no data	"Ambocoelia" sp.	no data	United States	MCZ	FossilSpecimen	Old "tire yard" loc. Near ...	IP	view

<http://portal.idigbio.org/>

# Getting What's There - Downloads



The screenshot shows the iDigBio search interface. The top navigation bar includes links for About iDigBio, Research, Technical Information, and Education. The main navigation bar has options for iDigBio Home, Portal Home, Search Records, Tutorial, Data, Research Tools, and Feedback. The user is logged in as 'grungle'.

The 'Search Records' section includes a search box, filters for 'Must have media' and 'Must have map point', and buttons for 'Filters', 'Mapping', 'Sorting', and 'Download'. The 'Current Search' section shows the search bounds: NW lat = 40.49709237269567, lon = -32.5. The 'Download CSV' section shows a build time of 0 hrs 0 mins 10 secs and an email input field.

The 'Downloads' section has a table with columns for 'Search' and 'Status'. Below this is a tabbed interface with 'List', 'Labels', 'Media', and 'Recordsets' tabs. The 'List' tab is active, showing a table of search results.

The map on the right shows a distribution of records in the Azores. A 'Top 10 Taxa' legend is displayed, listing the following taxa with their corresponding colors:

- Fringilla coelebs moreletti (blue)
- Passer domesticus domesticus (green)
- Fringilla coelebs moreletti pucheran, 1859 (light blue)
- Regulus regulus inermis murphy & chapin, 1929 (yellow)
- Turdus merula azorensis hartert, 1905 (orange)
- Sylvia atricapilla gularis alexander, 1898 (red)
- Erithacus rubecula rubecula (linnaeus, 1758) (purple)
- Protollira valvatoides (pink)
- Macroramphosus gracilis (lowe, 1839) (brown)
- Lasius grandis (dark brown)
- other (grey)

The table below the map shows the following data:







Family	Scientific Name	Institution Code	Collection Code	Date Collected	Collected By	Country	Locality	Occurrence ID	Columns
Pomacentridae	Abudefduf luridus	MCZ	Ich	1859-12-01/1859-12-...	Miss. Olivia Dabney	Azores	FAYAL [=Faial]. Rec...	MCZ:Ich:38668	<a href="#">view</a>
Pomacentridae	Abudefduf luridus	MCZ	Ich	no data	T. W. Higginson, Mis...	Azores	Fayal = [Faial]. Rec...	MCZ:Ich:14702	<a href="#">view</a>
Ophiophoridae	Acanthephyra purpu...	MCZ	IZ	1931-08-01/1931-08-...	RV Atlantis	no data	NW of the Azores	MCZ:IZ:CRU-10531	<a href="#">view</a>
Certhiidae	Acanthopneuste tubu...	NBM	EV	1860-08-02/1860-08-...	E.A. Smith & Livings...	no data	Sand between Fai...	NBM:EV:1860-08-02-...	<a href="#">view</a>

The total number of records is 4,831.

Support for “download-and-code” workflow



# What's in a Download from iDigBio

Name ^	Date modified	Type	Size
 meta.xml	2/21/2016 11:25 PM	XML Docu...	30 KB
 multimedia.csv	2/21/2016 11:25 PM	Microsoft E...	1 KB
 multimedia_raw.csv	2/21/2016 11:25 PM	Microsoft E...	2 KB
 occurrence.csv	2/21/2016 11:25 PM	Microsoft E...	920 KB
 occurrence_raw.csv	2/21/2016 11:25 PM	Microsoft E...	737 KB
 records.citation.txt	2/21/2016 11:25 PM	Text Docu...	1 KB

```

records.citation.txt
1 http://www.idigbio.org/portal (2016),
2 Query: {"filtered": {"filter": {"and": [{"term": {"recordset":
  "ef30a918-b583-41f1-9ac4-4a37591b515a"}}]}},
3 1376 records, accessed on 2016-02-21T23:23:08.677528,
4 contributed by 1 Recordsets, Recordset identifiers:
5 http://www.idigbio.org/portal/recordsets/ef30a918-b583-41f1-9ac4-4a37591b515a (1376 records)
6

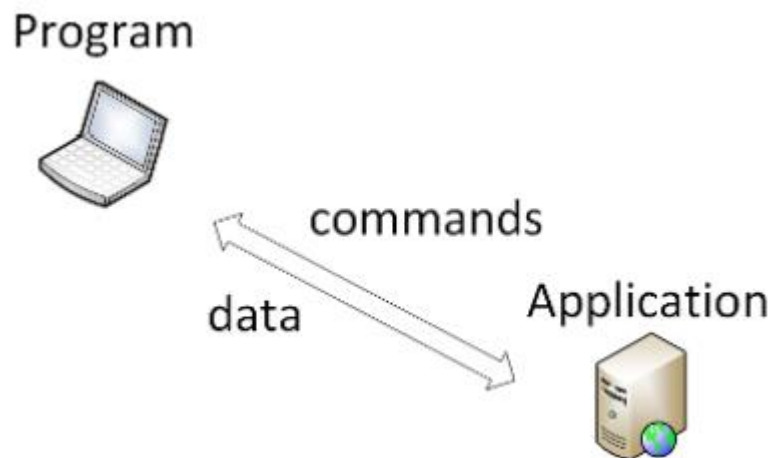
```

Processed records, raw records, media metadata, record metadata, citation information - everything you need to do reproducible analyses

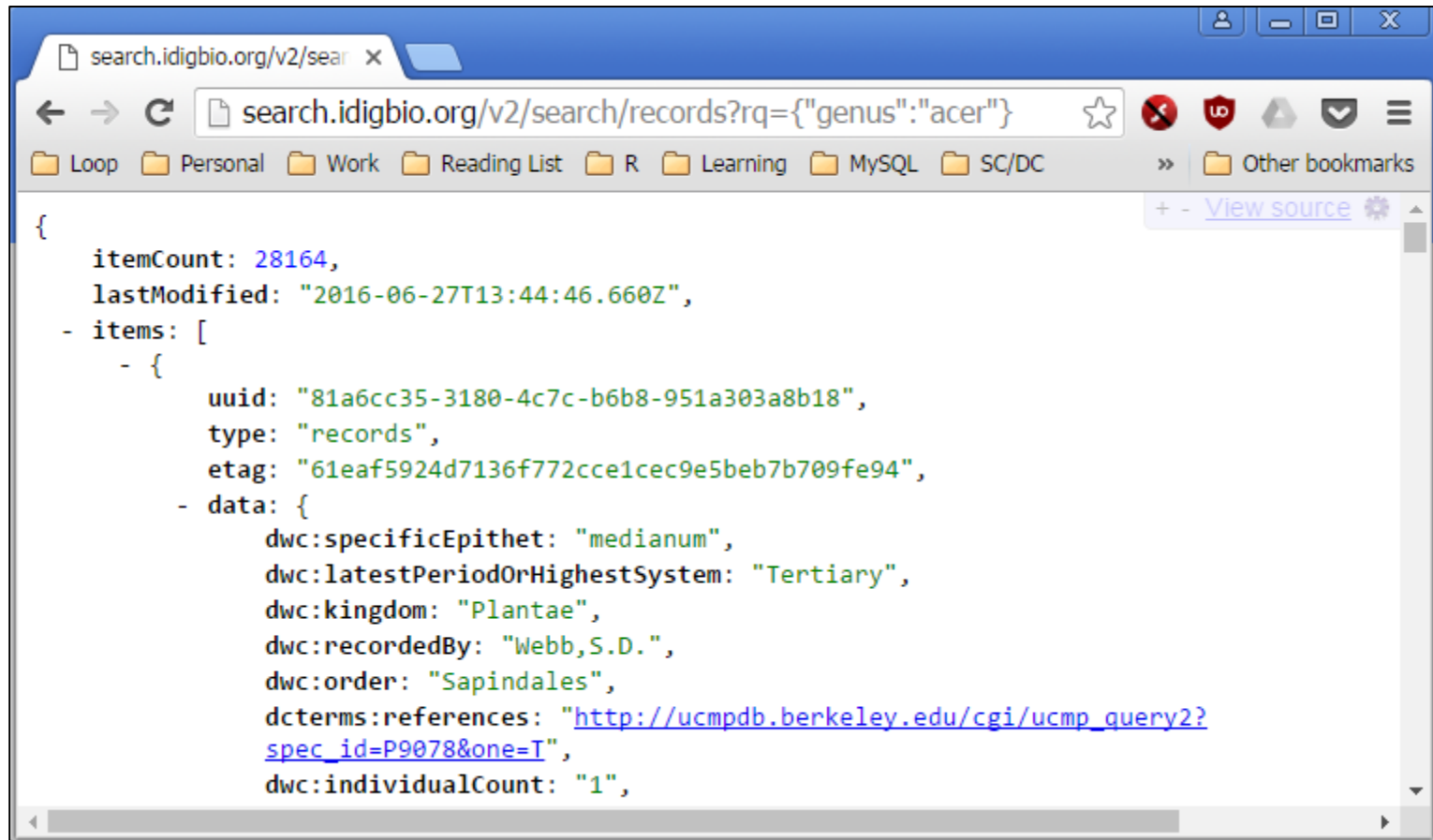
# Getting What's There - API

API: Application Programming Interface

- I want to do some **Programming**
- using parts of someone else's **Application**
- and I need an **Interface** that describes the commands and data to do it.



# Getting What's There - API Through a Web Browser

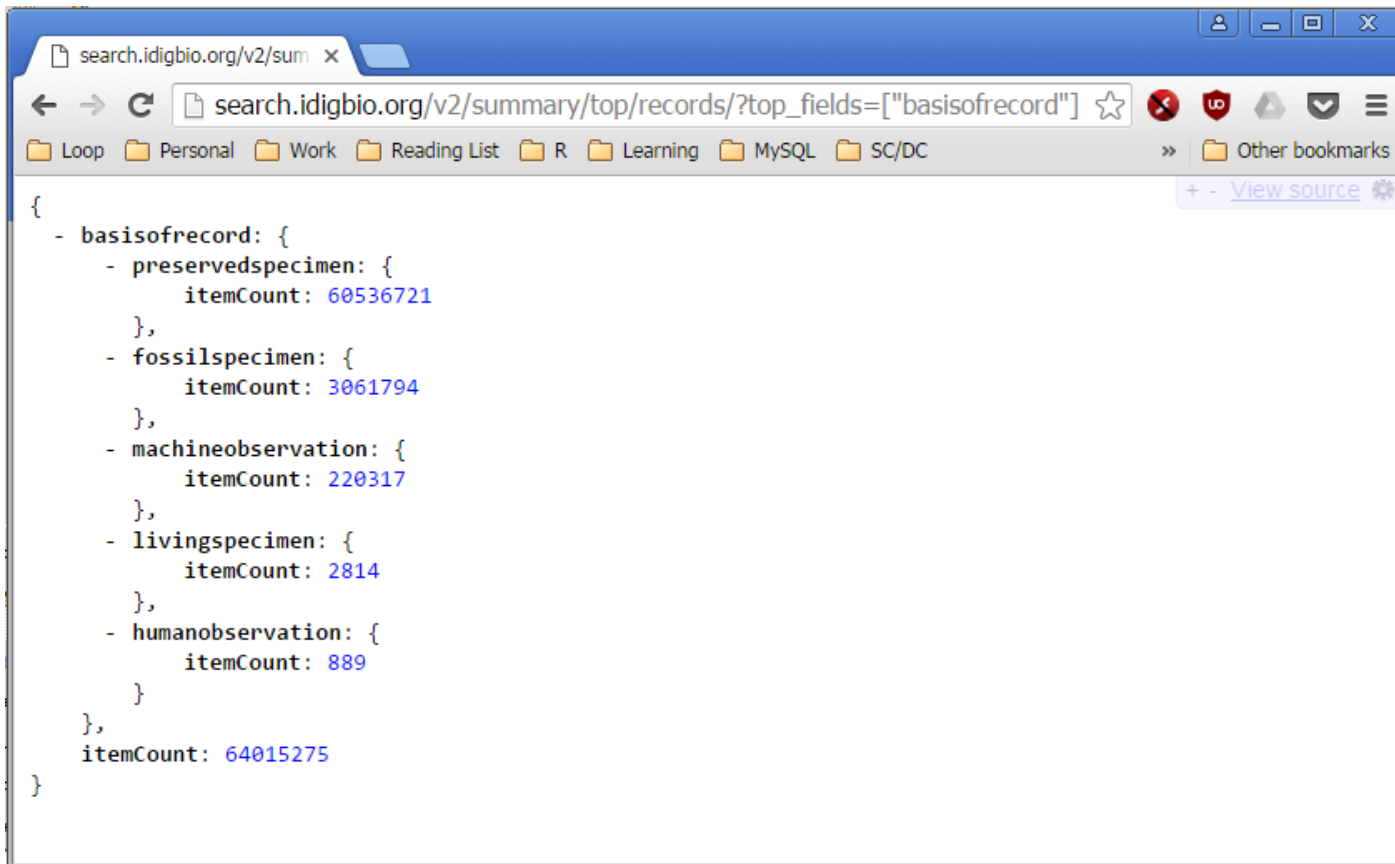


The screenshot shows a web browser window with the address bar containing the URL `search.idigbio.org/v2/search/records?rq={"genus":"acer"}`. The browser's address bar also shows a search bar with the same query. The main content area displays a JSON response from the API. The response includes the total number of items (28164), the last modified timestamp, and a list of items. The first item in the list is a record for the genus *Acer* with the following details:

```
{
  itemCount: 28164,
  lastModified: "2016-06-27T13:44:46.660Z",
  - items: [
    - {
      uuid: "81a6cc35-3180-4c7c-b6b8-951a303a8b18",
      type: "records",
      etag: "61eaf5924d7136f772cce1cec9e5beb7b709fe94",
      - data: {
        dwc:specificEpithet: "medianum",
        dwc:latestPeriodOrHighestSystem: "Tertiary",
        dwc:kingdom: "Plantae",
        dwc:recordedBy: "Webb,S.D.",
        dwc:order: "Sapindales",
        dcterms:references: "http://ucmpdb.berkeley.edu/cgi/ucmp\_query2?spec\_id=P9078&one=T",
        dwc:individualCount: "1",
```



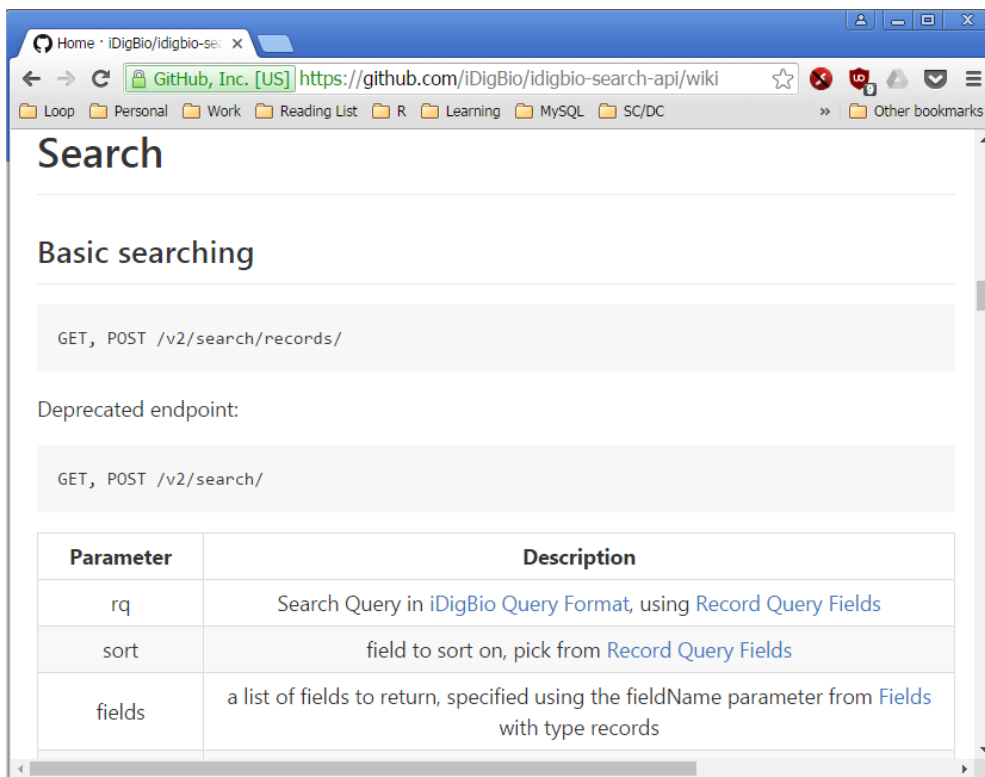
# Analyzing What's There - Summary API Through a Web Browser



```
{
  - basisofrecord: {
    - preservedspecimen: {
      itemCount: 60536721
    },
    - fossilspecimen: {
      itemCount: 3061794
    },
    - machineobservation: {
      itemCount: 220317
    },
    - livingspecimen: {
      itemCount: 2814
    },
    - humanobservation: {
      itemCount: 889
    }
  },
  itemCount: 64015275
}
```

# API Documentation

[https://www.idigbio.org/wiki/index.php/IDigBio\\_API](https://www.idigbio.org/wiki/index.php/IDigBio_API)



Home · iDigBio/idigbio-se: X

GitHub, Inc. [US] | <https://github.com/iDigBio/idigbio-search-api/wiki>

Loop Personal Work Reading List R Learning MySQL SC/DC >> Other bookmarks

## Search

### Basic searching

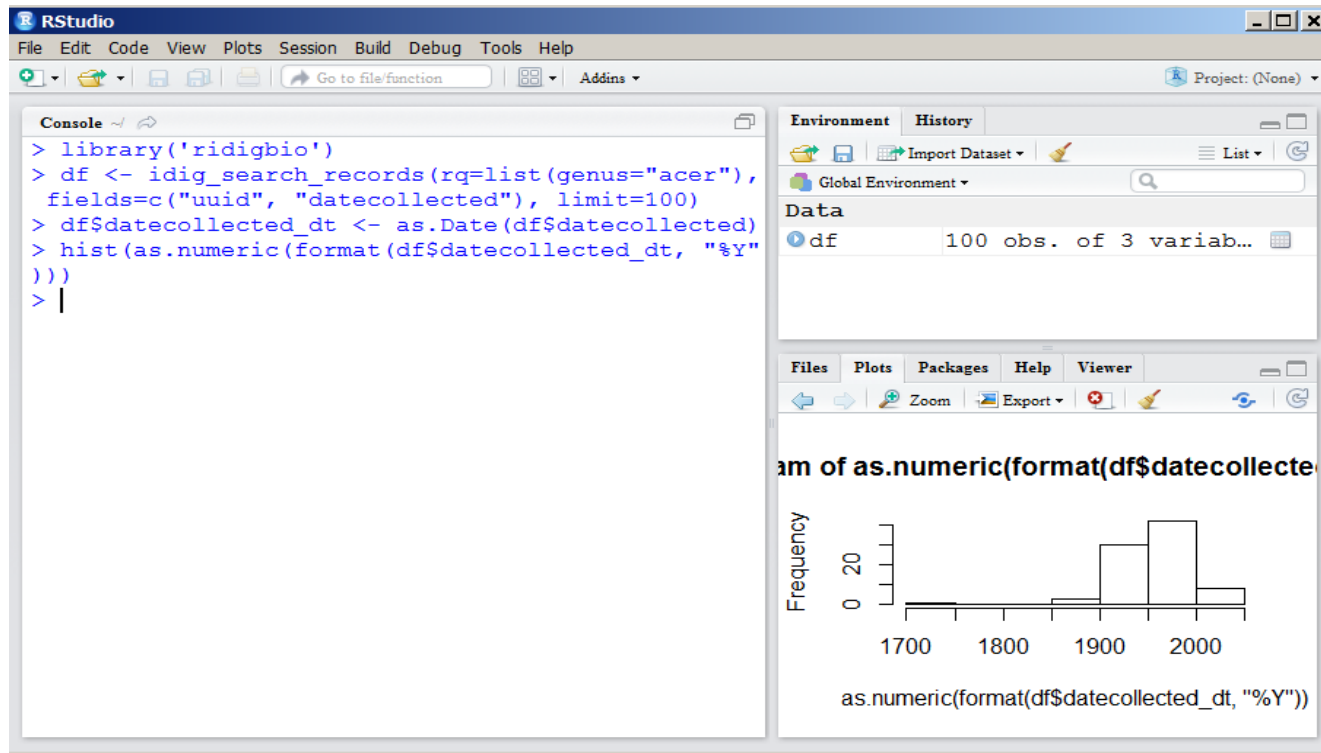
GET, POST `/v2/search/records/`

Deprecated endpoint:

GET, POST `/v2/search/`

Parameter	Description
rq	Search Query in iDigBio Query Format, using <a href="#">Record Query Fields</a>
sort	field to sort on, pick from <a href="#">Record Query Fields</a>
fields	a list of fields to return, specified using the <code>fieldName</code> parameter from <a href="#">Fields</a> with type records

# Analyzing What's There - Packages



R ridigbio: <https://cran.r-project.org/web/packages/ridigbio/index.html>

Python idigbio: <https://pypi.python.org/pypi/idigbio>

# Analyzing What's There - Applications



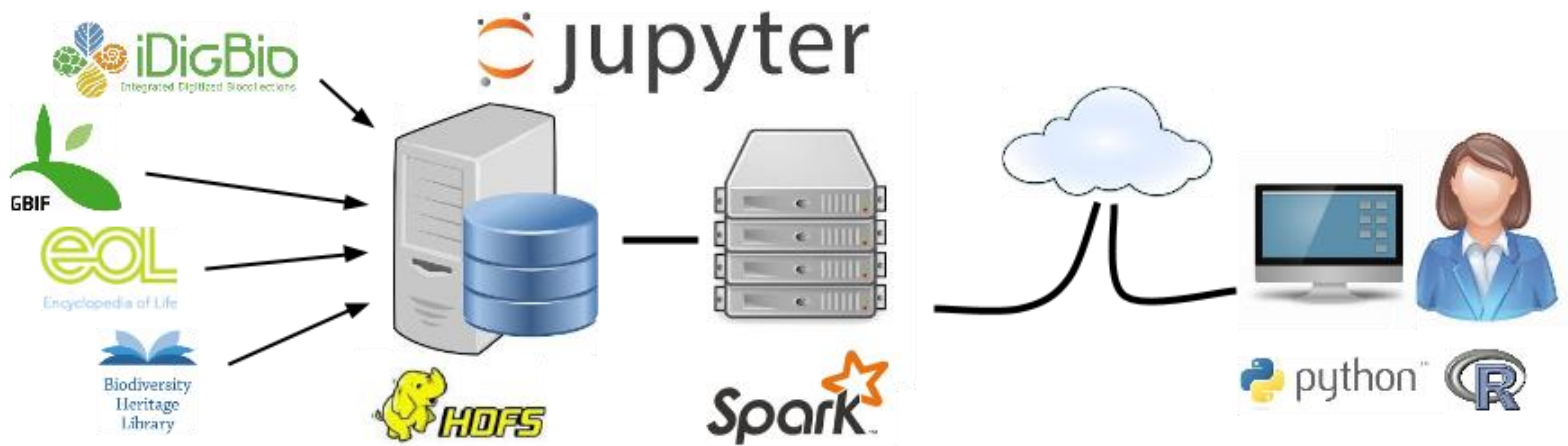
<http://lifemapper.org/>

# Analyzing What's There - GUODA

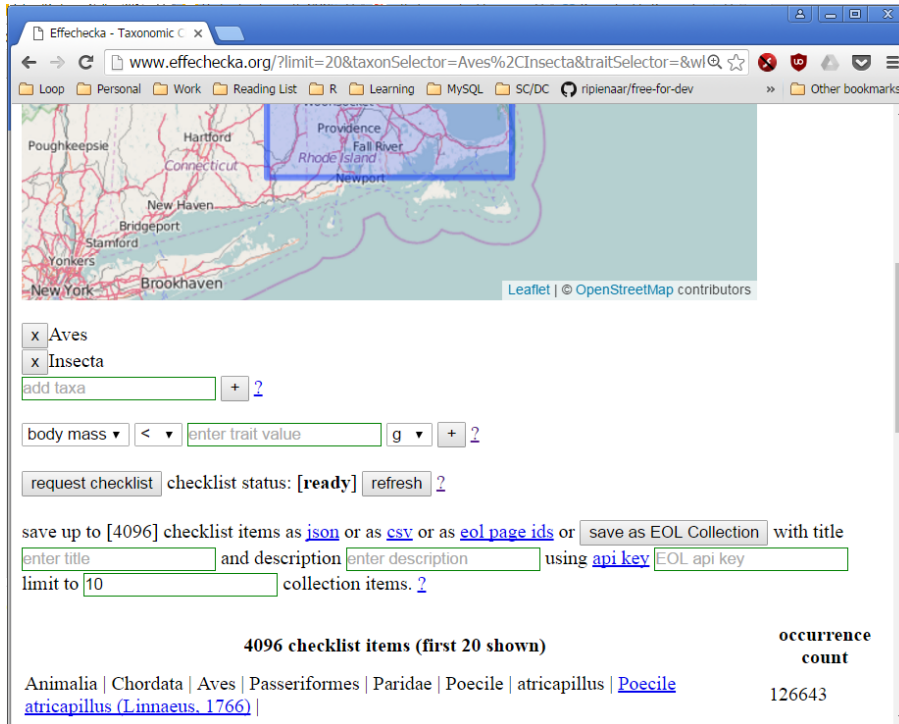
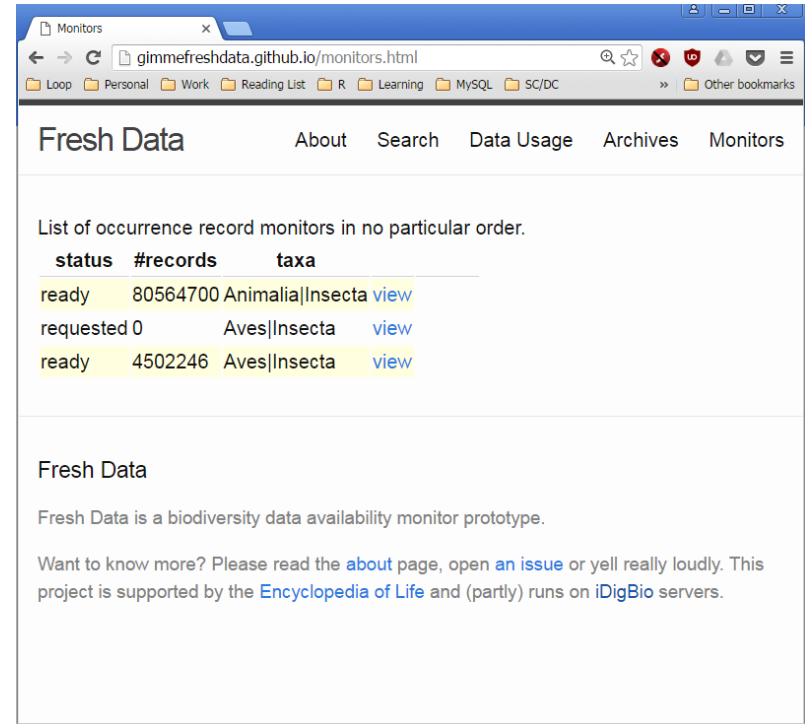
## Global Unified Open Data Access

An informal collaboration between technologists from organizations like EOL , ePANDDA, and iDigBio as well as independent biodiversity informaticists.

<http://guoda.bio>



# Analyzing What's There - Apps on GUODA

<http://www.effechecka.org/>

<http://gimmefreshdata.github.io/>



# It's Not Just Us!

Everyone has portals, downloads, APIs, and packages



Where to find some more of them:

<https://ropensci.org/packages/>

<https://www.biodiversitycatalogue.org/services>

# Challenges - Citation of Data

Easy access can complicate citation of data

- How do you cite 1000's of records in limited space?
- Aggregated datasets, who to cite?
- Permanently recording which records were used

<http://www.gbif.org/resource/80761>

<https://www.idigbio.org/content/idigbio-terms-use-policy>

# Challenges - Data Quality

We keep the raw data as provided to us but we interpret the data and adjust it to make it searchable such as:

- Flip and sign-correct lat & lon
- Format dates
- Fill higher taxonomy from GBIF backbone

Data Quality flags are provided as fields

GBIF/TDWG working group for standardizing data changes (with GBIF & ALA)

# Opportunities - Reproducible Research Workflows

What is all this for? Why have anything beyond just data dumps? Building efficient, collaborative, and reproducible scientific research!

- Collaborate with peers
- Iterate quickly on ideas
- Re-use and build on the data and code of others
- Publish the complete resources (data and code) to reproduce research

# Thank you!



[www.idigbio.org](http://www.idigbio.org)



[facebook.com/iDigBio](https://facebook.com/iDigBio)



[twitter.com/iDigBio](https://twitter.com/iDigBio)



[vimeo.com/idigbio](https://vimeo.com/idigbio)



[idigbio.org/rss-feed.xml](http://idigbio.org/rss-feed.xml)



[webcal://www.idigbio.org/events-calendar/export.ics](http://webcal://www.idigbio.org/events-calendar/export.ics)