ePANDDA

Seth Kaufman (seth@whirl-i-gig.com)

The Project

- Lots of tremendously useful data and APIs out there: PBDB, iDigBio, iDigPaleo, VertNet, Morphosource, Etc.
- Possibilities seen for an API layer to foster interoperation between these resources
- Accelerate discovery by:
  - Removing data integration hurdles
  - Providing a “one stop shop” for a broad range of useful data
  - Bridge conceptual divides between data sets in new ways
- ePandda =
  Enhancing **PA**leontological and **N**eontological **D**ata **D**iscovery **A**PI

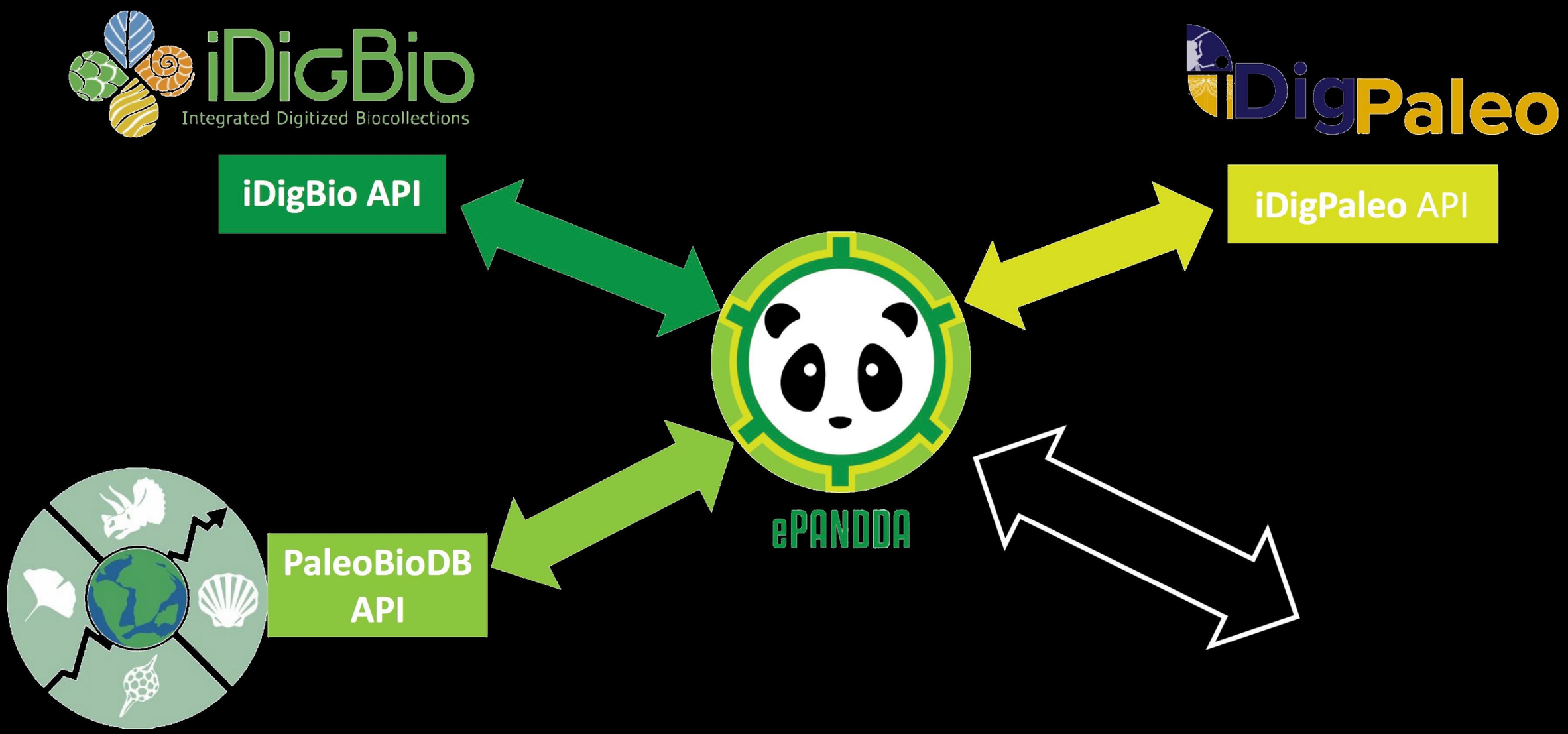
The Project

- What ePandda is *not*:
  - A web site (or any user interface)
  - A new database
  - A data collection project
  - A digitization project
- ePandda is an API providing access to existing data sets
- Any web sites and user interfaces produced by the project exist to document the structure and support the use of the ePandda API

The Project

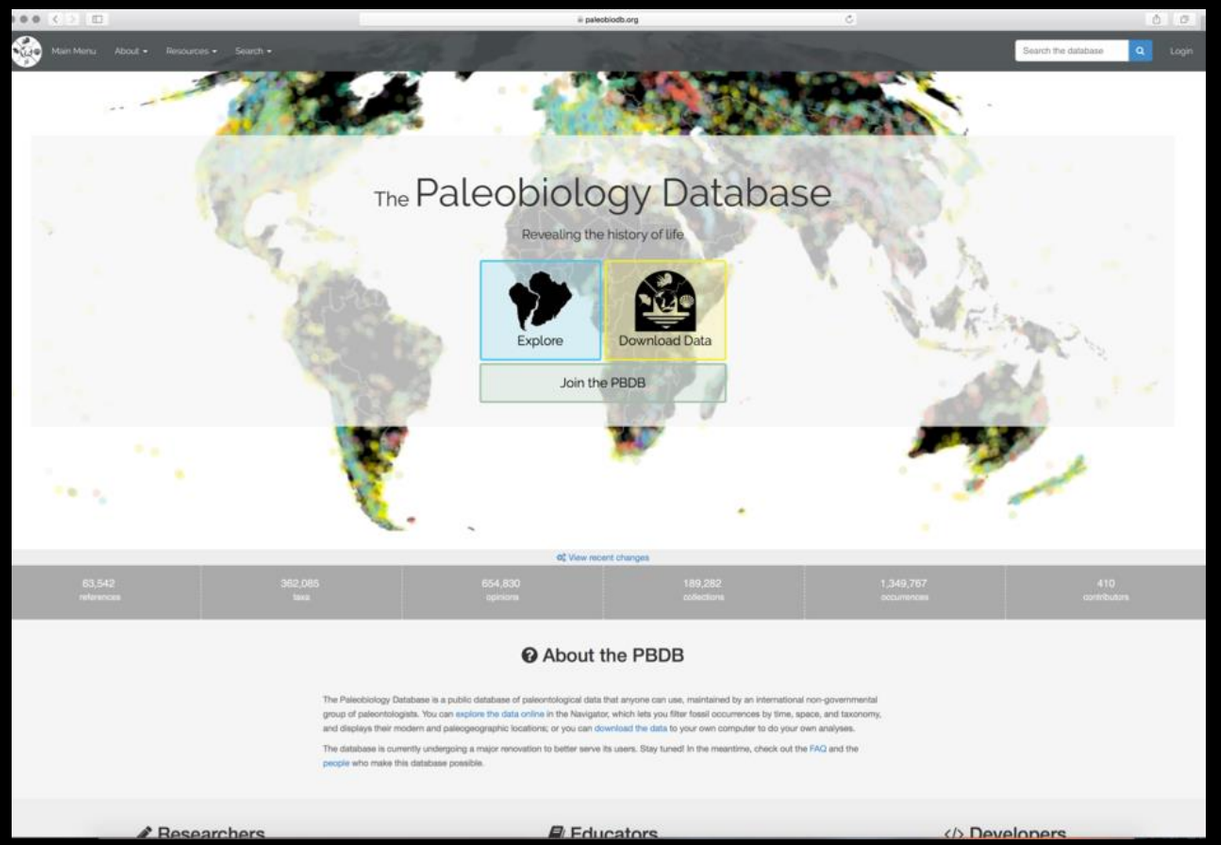
- Grant submitted by Chris Norris, Susan Butts (Yale), Talia Karim (University of Colorado), Jocelyn Sessa (Drexel), Gil Nelson (FSU/iDigBio), Mark Uhen (GMU/PBDB) and others
- Initial development focussed upon providing accessibility for the PBDB and iDigBio
  - Two of the largest and most used data sets for paleo research
  - PBDB is largely centered on occurrences
  - iDigBio is centered on specimens

Data Sources



Paleobiology Database

- Primarily literature-based
- Occurrence data (specimen module in progress)
- Analytic and Mapping tools for researchers
- New educational resources



https://paleobiodb.org

iDigBio

- Specimen-based with images
- Neontological and paleontological data
- Continuous development of community research tools

iDigBio
Integrated Digitized Biocollections

https://idigbio.org

Technical Overview

- API is implemented in Python using the Flask framework
- API is organized as a set of endpoints
  See https://api.epandda.org
- Endpoints query prepared indices
- Indices are generated by a data harvesting and assembly framework
- Indices return key values for related records in iDigBio and PBDB
  - We don’t fully replicate iDigBio and PBDB…

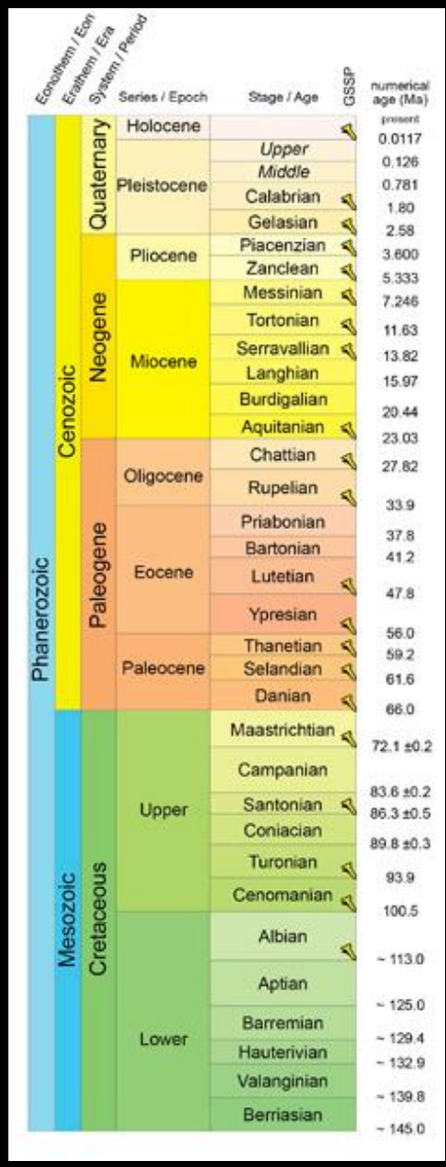
Creating Indices

- Indices allow endpoints to plugin in query parameters and return keys to specific records in the underlying data resources (iDigBio, PBDB, Etc.)
- Indices are constructed by querying iDigBio and PBDB APIs, with local storage provided by ElasticSearch
    - Snapshots of iDigBio/PBDB datasets are generated by API calls
    - Snapshots are used to locally resolve PBDB and iDigBio keys into limited sets of data that can be included in the ePandda response for convenience
- Current indices are implemented using ElasticSearch and MongoDB, although API framework is agnostic

Building an Index: Normalization

- Source data is cleaned up in several ways:
  - removing unwanted artifacts (Eg. stripping characters and tags)
  - parsing JSON objects
  - correcting or normalizing values against other sources (APIs, authorities, Etc.)
  - in all instances original values are retained in searchable fields

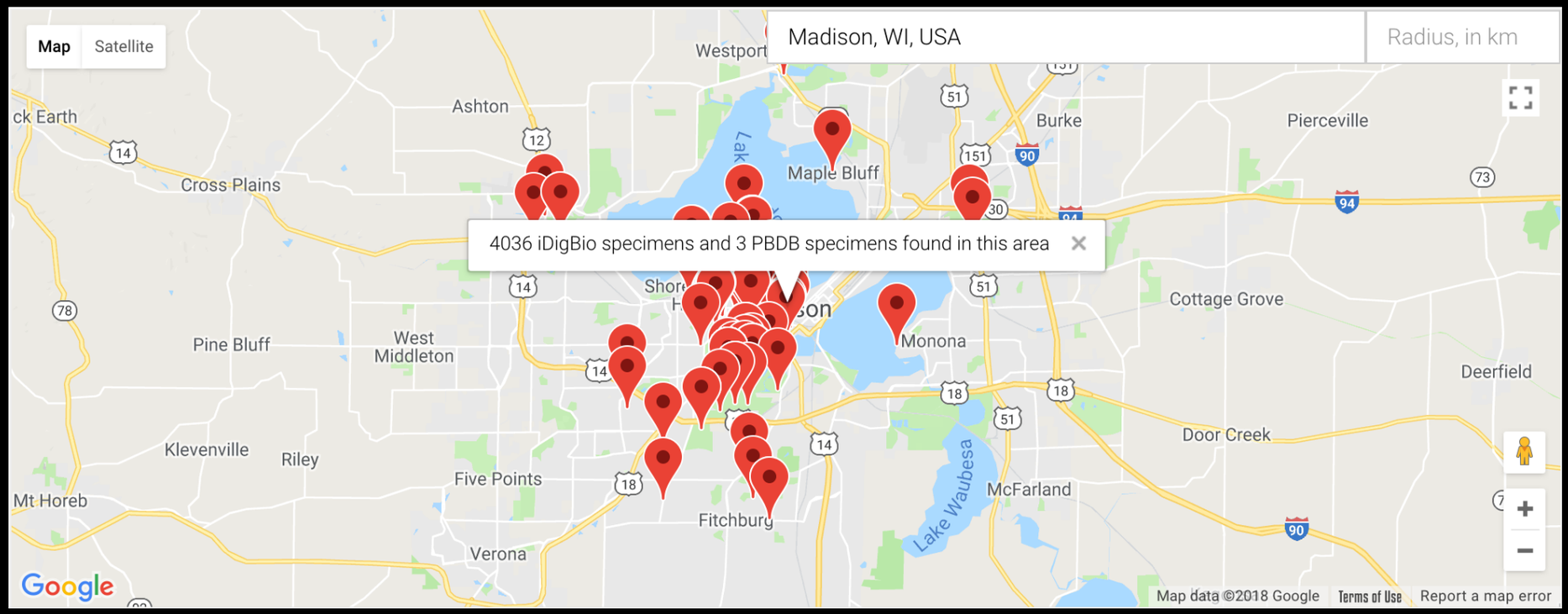
- **Chronostratigraphy**
  - We rely on the ICS timescale to resolve strata to their bounds in Ma
  - We also match local and variant strata names against this to enable wider searches



Building an Index: Normalization

- **Locality**
  - All country, state/province and county names can be checked against the GeoNames API
    - Initial implementation did this, but currently disabled for performance reasons (with the exception of country names)
  - This functions to correct language, accent, and common typo errors in the source data
  - Geo-coordinates are also normalized into an easily searchable format

that enables distance and bounding-shape queries



Building an Index: Hierarchies

- To support broad-to-narrow and narrow-to-broad searching hierarchical data structures are used
- **Taxonomy:**
  - Hierarchies can be built using the PaleoBio API, falling back to the GlobalName parser API (or others) as necessary
    - Initial implementation did this, but is currently disabled for performance reasons
    - Uncertainty around benefits of applying this to PBDB/iDigBio data
- The resolvers will match the lowest level in the hierarchy as possible
- These will generate fully searchable hierarchies for all terms
- If a match cannot be found, or is invalidated, the verbatim hierarchy can be made available for search

Building an Index: Hierarchies

- **Chronostratigraphy & Lithostratigraphy**
  - The PaleoBio and ICS authorities are used to generate reference hierarchies that can be used to search across levels
  - Names are converted to MYA as as normalized value
    - This enables children and potentially siblings to be searched and matched on
- **Locality:**
  - Similarly the locality index is searchable as an hierarchy and will return child records

ePandda API

- Organized around endpoints
  - **occurrences**: Primary query interface. Allows queries on any supported field in PBDB or iDigBio and returns a correlated result set
  - **publications:** Query unified PBDB/iDigBio index for publication data and correlated specimens.
  - **annotations:** Attach and retrieve annotations to any PBDB or iDigBio data entity in ePandda. [In progress]
  - Everything else:
    - **stats, bug_report, full_match_results**

ePandda API Matching Parameters

- The occurrence and publications endpoints match records against three criteria
  - Locality
  - Taxonomy
  - Chronostratigraphy
- These parameters can be matched against different levels of each hierarchy to control the specificity of matches
- Generally all three are required to query ePandda for matching results, but this can be adjusted

https://api.epandda.org


```
{
    "description": "ePANDDA REST API guide",
    "routes": [
        {
            "description": "Human created Annotations endpoint (ORCID verified )",
            "methods": "HEAD,POST,OPTIONS,GET",
            "name": "Create Annotations",
            "url": "/annotations/create"
        },
        {
            "description": "Returns full data sets for matching criteria returned from the main Occurrence endpoint",
            "methods": "HEAD,POST,OPTIONS,GET",
            "name": "Full Result Response",
            "url": "/full_match_results"
        },
        {
            "description": "Searches PBDB and iDigBio for publication <=> specimen matches and returns match groups
based on taxonomy, bibliograhy and locality",
            "methods": "HEAD,POST,OPTIONS,GET",
            "name": "Publications Index",
            "url": "/publications"
        },
        {
            "description": "Returns openAnnotations for linked data in ePANDDA.",
            "methods": "HEAD,POST,OPTIONS,GET",
            "name": "Annotations",
            "url": "/annotations"
        },
        {
            "description": "Searches PBDB and iDigBio for occurrences and returns match groups based on taxonomy,
chronostratigraphy and locality",
            "methods": "HEAD,POST,OPTIONS,GET",
            "name": "Occurrence Index",
            "url": "/occurrences"
        },
        {
            "description": "File Bug Reports Here",
            "methods": "HEAD,POST,OPTIONS,GET",
            "name": "ePandda Bug Reports",
            "url": "/bug_report"
        },
        {
            "description": "Interesting API statistics",
            "methods": "HEAD,POST,OPTIONS,GET",
            "name": "API statistics",
            "url": "/stats"
        },
        {
            "description": "Summary of available endpoints",
            "methods": "HEAD,POST,OPTIONS,GET",
            "name": "API Info",
            "url": "/"
        }
    ],
    "timeReturned": "2018-05-11 09:05:35.860819",
    "v": 1.0
}
```

https://epandda.org

109,925,680 records
API status: OK
Updated: 5/11/2018 @ 01:05 pm

ePANDDA BETA

enhancing Paleontological and Neontological Data Discovery API

ABOUT | DOCUMENTATION | SANDBOX | EXAMPLES

ePANDDA is a new application programming interface (API), created as a collaboration by members of the Academy of Natural Sciences of Drexel University, Florida State University, George Mason University, University of Colorado Museum of Natural History, and the Yale Peabody Museum of Natural History. Its purpose is to connect the Paleobiology Database (PBDB), iDigPaleo, and iDigBio and will allow seamless searching and data discovery among the three databases.

NOW SEARCHING:

108,583,031
iDigBio specimens

109,925,680
specimen & occurrence records

1,342,649
PaleoBio DB occurrences

VIEW SAMPLE

Questions?

Seth Kaufman (seth@whirl-i-gig.com)