iDigBio v3 API Preview

Format Changes, Integrated Search, and Automated Indexing

API Context & Primer

The iDigBio Application Programming Interfaces (API) provide a method for users and programs to access and interact with data stored in the iDigBio system.

- In v1 of iDigBio, there were 2 main APIs: Data and Search.
  - Data: create, read, update and delete
  - Search: discovery
- In v3 these have been combined into a single API.

API Context & Primer

iDigBio's API's follow the RESTful paradigm wherever possible, the core principle of this paradigm is that the API itself does not maintain any state, instead it provides the API client with all the necessary state information, and links to further actions.

API Context & Primer

iDigBio's API's use the HTTP protocol for interaction, the same protocol that your web browser uses. Data returned by the APIs is formatted using a format known as JSON, the JavaScript Object Notation, which is reasonably compact and has wide support both within web browsers, and in specialized tools for interacting with APIs.

V1 API Recap

- Restful
- All core properties namespaced to idigbio (idigbio:etag)
- No support for generic binary endpoints, only images
- Manual indexing with search format controlled by external scripts.

Format Changes

New Format:

```
{
        version: "0",
        uuid: "0000...",
        dateModified: "2013….",
        etag: "95e9….",
        scope: "00d9….",
        type: "records",
        links: {
                _self: "....",
                _scope: "....",
                “mediarecord: [
                        “abcd…”,
                        “bcde…”
                ]
        }
}
```

Old Format:

```
{
        "idigbio:uuid":"0000...",
        "idigbio:version":0,
        "idigbio:etag":"95e9….",
        "idigbio:dateModified":"2013….",
        "idigbio:links":{
                "record":"http://api.."
        }
}
```

Format Changes (cont.)

- Drop the idigbio namespace (more JSON compliant)
- All records now have an inherent type
- System generated links now have an underscore in front of them to distinguish them from other links
  - Examples of system links are _self (self reference) and _scope (scope reference)

List endpoint functionality changes

- Sorted based on descending modification date by default (newest results first)
- Date range supporting (unix timestamp or ISO date)
- Count/limit pagination as in the v1 APIs
- Ability to retrieve limited counts of data along with the item list (currently 1000 max pending testing)
- New /object list endpoint that can query all object types at the same time
- Integrated querying.
  - Simple term searching (ex. ?scientificname=puma%20concolr)
  - full query support ?query={....}
- All methods are combinable (date range + limit/count, query + date range)

Automated Action support

- API now publishes data events to an internal queue that supports multiple consumer types
- This functionality supports
  - automated indexing of newly updated records
  - auto-thumbnailing for new images
  - automatic decomposition of newly inserted datasets
  - automated fetching of newly registered data endpoints
- Put together, this allows the iDigBio system to be completely automated for most of the data ingestion process.

Things that haven't changed

- Still a RESTful JSON API
- Backwards compatible with v1
- Still focused on delivering nearly-raw data
  - Consistency of data formats and content across collections and across disciplines remains an issue. We're working on solutions, but need to start a broader conversation on how to deliver those solutions and their results to the databases of record

Data Ingestion Process

Ingesting data into iDigBio follows the following process.

1. A new collection is registered on portal.idigbio.org with a dataset endpoint.
2. iDigBio's automated systems download the dataset and process it.
3. iDigBio Staff review the dataset.
4. A summary report is sent to the curator for approval. (see next slide)
5. The records and mediarecords are ingested into the public data APIs
6. At this point the records are searchable via the public portal
7. In the background, image links in mediarecords are visited and derivative images (thumbnail and webview) are generated.
8. Another background process checks RSS datasets once a week for updates and non-RSS datasets once a month.
9. The process repeats for the updated records from step 4.
10. A post-ingestion summary report is sent to the curator.

Data Reporting

- New automated decomposition scripts generate summary metrics in the step immediately prior to data ingestion.
- Our thought is that, at least initially, we will email these out to curators and have some sort of manual confirmation before data is actually ingested
- Once a data set is published, these will go out periodically as new data is fetched and ingested.

Data Reporting (cont.)

25286
Specimen Records

Specimen Records Updated:
25286
Specimen Records Created:
0

Field Fill Percentages

Minimum Recommended Terms

dwc:occurrenceID (Occurrence ID)	100.0%
dwc:scientificName (Scientific Name)	100.0%
dwc:recordedBy (Collected By)	99.74%
dwc:locality (Locality)	95.82%
dwc:eventDate (Date Collected)	98.41%

Other Terms

dwc:kingdom (Kingdom)	100.0%
dcterms:language (Language)	100.0%
dwc:collectionCode (Collection Code)	100.0%
dwc:basisOfRecord (Basis of Record)	100.0%
dwc:locationID (Location ID)	100.0%

33209
Media Records

Media Records Updated:
33209
Media Records Created:
0

Field Fill Percentages

Minimum Recommended Terms

ac:bestQualityAccessURI (Best Quality Access URI)	100.0%
dc:format (Format)	0%
dcterms:title (Title)	100.0%
xmpRights:UsageTerms (License Terms)	0%
xmpRights:WebStatement (License URL)	61.3%
ac:licenseLogoURL (License Logo URL)	61.3%

Other Terms

ac:mediumQualityAccessURI (Medium Quality Access URI)	100.0%
ac:providerID (Provider ID)	100.0%
dwc:occurrenceID (Occurrence ID)	100.0%
ac:thumbnailFormat	100.0%

New Login & AAA

iDigBio has also been working on a new Authentication and Authorization system. Some highlights of this work are:

- Dropping social authentication from the system
- Transparent OpenID based single sign-on to all iDigBio websites
  - Once the system has been battle tested, we may be able to offer this login service to other interested projects (users could log in to your systems using a single common iDigBio account)

The authentication component of this should be live sometime before the main v3 launch, possibly before the summit. The authorization component will take a little longer, although the back ends to support it will be deployed with the v3 APIs.

Lastly: A Reminder

- Beta versions of new iDigBio software are published to the beta endpoints periodically
  - http://beta-portal.idigbio.org
  - http://beta-api.idigbio.org (new API isn't here yet)
- When major feature groups are completed, I'll be sending emails out to CYWG members, TCN PIs, and other interested parties for feedback. Feel free to forward these email along to others who may be interested, or provide me with more emails to add to the list.
- Even outside of feedback pushes, iDigBio greatly appreciates comments on what parts of our software you like, what you don't like, and improvements and features you would like to see.
- Our #1 goal is to make sure that our software is meeting community needs. The less feedback we get, the farther we risk straying from that goal.
- Our launch goal is to have the new API up sometime in the next few months, and to have several Feedback pushes on our v3 Portal interfaces over the next couple of months before its launch in December.
- Feel free to start conversations on IDIGBIO-CYWG-L@lists.ufl.edu .