iDigBio: What's in a name – dataset, collection, institution, publisher, recordset?

We are often asked questions such as, "How many collections are there?" "Who is participating in digitization activity?" "How much progress has been made toward digitizing collections?" Although answers to these questions appear to be straight-forward, they often require paragraphs of qualifiers to answer accurately.

This document defines:

- terms used to refer to various entities related to counts in iDigBio
- where to go to find counts, which are periodically updated

Definitions are established within two competing contexts: an institutional hierarchy and aggregated data. These competing contexts are a result of the funding model of the Advancing Digitization of Biodiversity Collections (ADBC) Program as well as implementation decisions made by digitization staff and collection managers.

Definitions based on institutional hierarchy

The definitions below are used to provide information on institutions and collections. Curation and maintenance of this data resource are ongoing at iDigBio. Definitions of terms in the List of Collections can be found at: https://github.com/iDigBio/idb-us-collections.

Institution: A physical entity from which iDigBio receives digital data and media of vouchered specimens. Examples include:

- *Museums* - e.g., Florida Museum of Natural History (FLMNH/UF/FLAS), Bishop Museum (BISH)
- *Universities/colleges* - e.g., University of California, Berkeley (UCMP)
- *Research institutions* - e.g., Morton Arboretum (MOR), Paleontological Research institute (PRI)

Institutions exist for iDigBio in two contexts:

1) counts of institutions receiving ADBC funding.
2) counts of institutions not receiving ADBC funding

Institutions usually have symbolic codes. In some cases, these are maintained in lists developed by organizations or societies. For example, the Index Herbariorum and the American Society of Ichthyologists and Herpetologists maintain and curate lists of institution codes for herbaria and fish collections, respectively.

Index Herbariorum
American Society of Ichthyologists and Herpetologists

These codes are included in metadata of digitized specimen records, as well as used in publications to identify specimen use and in invoices to track loans between institutions. Because institutional codes are not always unique, it can be a challenge to use them to identify a particular institution. They sometimes are used in combination with a collection code to facilitate recognition of data.
Source of the count: The current number is 520 as of August 2017.

Collection: A taxonomic division within an institution; e.g., an insect collection, an herbarium.
A challenge in tracking and counting collections is that often taxonomic groups within a taxonomic division of an institution are subdivided and also referred to as collections. Collections can be nested within collections; e.g., lichens, bryophytes, fungi, algae, vascular plants might be considered "collections" within a plant collection. Also a single dataset can contain the entire institution's digitized holdings (e.g., those of Harvard Herbaria and Bishop Museum). Efforts have been made by iDigBio to flatten the data into easily recognizable, aggregated taxon-based collections.
Source of the count: ask Kevin Love

Collections in the US are represented in iDigBio's *List of US collections* available at:
https://www.idigbio.org/portal/collections.

> This list attempts to include all US collections, whether or not they have published data to an aggregator. Data providers and collections staff are asked to provide and update information in this list.

Definitions from the view of aggregated data

The following are in reference to aggregated data:

Publisher: An individual entity from which data have been ingested. Publishers make data available in a Darwin Core archive through:

1) IPTs, which may be a GBIF country node; e.g., Norway, Sweden, Denmark, or an institution; e.g., Yale, MCZ, NYBG, SI – Smithsonian.
2) Other sources of Darwin Core archives; e.g., Symbiota nodes, PNW node – Consortium of Pacific Northwest Herbaria. These are IPT-like, and are custom built for the data providers.

Source of the count: The current number is 75 as of March 2018.
http://api.idigbio.org/v2/view/publishers

Recordset: An individual set of data that has been mobilized, published, and ingested by the iDigBio portal. A recordset can be various combinations of institutional datasets:

1) An institution's digitized specimen records for a single taxonomic group (e.g., HUH – Harvard Herbaria) – homogenous taxon group (typical institutional dataset)
2) An institution's digitized specimen records for two or more taxonomic groups (e.g., BISH – Bishop Museum) – heterogeneous taxon groups
3) Only a portion of an institution's digitized specimen records for a single taxonomic group (e.g., University of Kansas, Biodiversity Institute and Natural

History Museum – Herpetology, FMNH Ferns) – a subset of a homogenous taxon group

A challenge in relating recordsets to collections and institutions is that recordsets may involve combinations of these three types of recordsets. A recordset published to iDigBio may consist of specimens from a particular geographic area and include specimens from several taxonomic collections. Others may consist of records mobilized and published collectively by an aggregator (e.g., Morphbank) or a TCN and involve specimens from several institutions, or datasets containing only parts of a collection (e.g., an herbarium's invasive vascular plants from the Great Lakes region). Because of the various ways in which datasets are mobilized, they often do not map neatly to collections or even to institutions.

Source of the total count: iDigBio homepage. The current number is 1552 as of March 2018.
http://search.idigbio.org/v2/summary/count/recordsets?rsq={%22data.ingest%22:%22true%22}&limit=5000

Source of the ADBC counts:
https://www.idigbio.org/wiki/index.php/TCN_Resources - TCNs.2FPENs_At_a_Glance

Counts for publishers and datasets can be accurately counted as they relate to iDigBio. But for reasons described above, counts cannot be provided for collections and institutions.

Authors: McCaffrey, Page, Love, August 2017

Appendix

iDigBio Counts – as of March 2018

Name of Count	Count	Reference
Institutions		
* ADBC	396	https://www.idigbio.org/wiki/index.php/TCN_Resources#TCNs.2FPENs_At_a_Glance
* Total	708	By special request
Collections - USA		https://www.idigbio.org/portal/collections
Datasets ingested	1552	http://search.idigbio.org/v2/summary/count/recordsets?rsq={%22data.ingest%22:%22true%22}&limit=5000
Publishers	75	http://api.idigbio.org/v2/view/publishers