**BIEN database entities related to data provenance and ownership**

Prepared by: Brad

Date: 14 February 2013

**1. Introduction**

The following are BIEN database entities are intended to track where data come from and who owns them. The purpose of tracking these entities is to ensure proper attribution and support data access conditions.

**2. Changes requested by GBIF to support acknowledgement of GBIF, its data providers and data resources.**

Among other changes to the BIEN website, GBIF has requested that we acknowledge GBIF as a "global data indexer" on a separate page that would list each GBIF data provider ("data publisher" in GBIF terminology), each data set ("data resource" in GBIF terminology) and the number of records in the data set. In addition, we need to provide links to the description of each data publisher and each data resource within the GBIF website. Table 1 provides an example of a few records from the resulting table of data publishers and resources.

Table 1. Table acknowledging GBIF data publishers and data resources, requested for BIEN website.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| GBIF data publisher | Links to data publisher | Contributing datasets | Links to data resources | Nos. of data records |
| Herbarium of the University of Arizona | <http://data.gbif.org/datasets/provider/269/> | The University of Arizona Vascular Plant Herbarium | <http://data.gbif.org/datasets/resource/7900/> | #### |
| Herbarium of the University of Arizona | <http://data.gbif.org/datasets/provider/269/> | The Robert L. Gilbertson Mycological Herbarium | [http://data.gbif.org/datasets/resource/7901/](http://data.gbif.org/datasets/resource/7900/) | #### |
| Duke University Herbarium | <http://data.gbif.org/datasets/provider/292> | DUKE Vascular Plant Collection | <http://data.gbif.org/datasets/resource/81043> | #### |
| Duke University Herbarium | <http://data.gbif.org/datasets/provider/292> | DUKE Bryophyte Collection | <http://data.gbif.org/datasets/resource/8104> | #### |

In fact, there are currently no resources within GBIF for "The Robert L. Gilbertson Mycological Herbarium" nor for the " DUKE Vascular Plant Collection" (the other two do exist). But I wanted to provide a hypothetical example of what different resources from the same publisher would look like. Incidentally, "The University of Arizona Vascular Plant Herbarium" and "The Robert L. Gilbertson Mycological Herbarium" both reside in the same building and have the same Index Herbariorum acronym (ARIZ). However, because they are maintained as separate databases (the former accessible and the latter not), GBIF would treat them as separate resources from the same data publisher.

**3. Additions to BIEN schema to allow tracking of data provenance and ownership, and to comply with GBIF requests**

To support displaying on our website the above information request by BIEN, as well as to ensure proper attribution of the creators of the data we manage, the BIEN database must maintain a distinction between the entity that first published a dataset and entity that provided the data to BIEN. We also should maintain a distinction between the primary data publisher and the dataset itself; currently, we conflate both(for example, in the current BIEN schema, "ARIZ" to refers to (1) specimen records from The University of Arizona Herbarium and the institution called The University of Arizona Herbarium). Finally, we must also author of the data, *when that author also has authorship rights,* such as the obligations of acknowledgement or requirement to offer co-authorship.

I believe that all the above concepts are covered by the following entities. Terms in square brackets are provided for the purpose of defining other terms, but are not explicitly tracked within the BIEN database:

**dataset** (GBIF syn: data resource). One or more data records managed together as a unit, and limited in scope in one or more ways, for example, to a particular region ("Madidi National Park forest plots"), a particular author ("Alwyn Gentry Transect Dataset"), a particular taxon ("University of Arizona Mycological Collection"), a particular research project ("La Selva secondary forest plots") or a particular data publisher (typically, a herbarium, such as New York Botanical Garden). In practice, BIEN inherits dataset definitions from its dataproviders. However, there may be cases where we choose to subdivide a dataset, particular if the subdivision have different dataowners and/or different conditions or use or acknowledgement rules.

**[dataauthor]:** The persons or institutions who created the data on which a dataset is based. Note that digitization of data (for example, digitization of a specimen by a herbarium) does not by itself constitute authorship. A herbarium that digitizes a specimen is the datapublisher, not a dataauthor, of that specimen. The dataauthor of a specimen is the collector. For purposes of attribution, BIEN only tracks who are also dataowners (see below).

**dataowner**: A dataauthor who retains at least minimum ownership rights. A test of minimum ownership rights is acknowledgement: if you are obliged to cite a dataauthor, they they are also a dataowner. Collectors of individuals specimens are not considered data owners as they are not typically acknowledged. Dataowner may have more extensive ownership rights, including the right to restrict access to their data and the right to be offer coauthorship in publications which use their data.

**[dataprovider]:** A persons or institution that provides electronic access to a dataset. For purposes of attribution, BIEN only tracks dataproviders that are either primarydataproviders or biendataproviders (see below), therefore this

**primarydataprovider** (GBIF syn: datapublisher): The dataprovider that first provides electronic access to a dataset. The dataprovider must not have obtained the data from another dataprovider; they either digitized the dataset themselves, or are making available a digital copy of the dataset prepared by someone else but not previously made available. If a dataset was not published elsewhere prior to inclusion in the BIEN database, then BIEN is the datapublisher for that dataset. BIEN's policy is that datapublishers must always be acknowledged.

**biendataprovider**: The dataprovider that provides a dataset to BIEN. The biendataprovider may or may not be a datapublisher. Most data aggregators such as GBIF will be dataproviders but not datapublishers. If a dataset was digitized by BIEN for the first time, then BIEN is the biendataprovider. BIEN's policy is that biendataproviders must always be acknowledged.

In terms of Table 1, above, " GBIF data publisher"= **primarydataprovider, "**Contributing datasets"=**dataset**. Table 2 is a derived view showing examples of each of these entities in the BIEN database.

**4. Implementation in the BIEN database**

Coming up…

Table 2. Examples of the above entities in the BIEN database

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **datatype** | **dataset** | **dataowner** | **primarydataprovider** | **biendataprovider** |
| specimen | University of Arizona Herbarium Vascular Plant Collection | University of Arizona Herbarium | University of Arizona Herbarium | University of Arizona Herbarium |
| specimen | Herbarium of The New York Botanical Garden | The New York Botanical Garden | The New York Botanical Garden | The New York Botanical Garden |
| specimen | University of Belen Herbarium | University of Belen Herbarium | University of Belen Herbarium | Species Link |
| specimen | Herbario Nacional do Brasil | Herbario Nacional do Brasil | Herbario Nacional do Brasil | Species Link |
| specimen | University of Minnesota Herbarium | University of Minnesota Herbarium | University of Minnesota Herbarium | GBIF |
| specimen | Smithsonian National Herbarium | Smithsonian National Herbarium | Smithsonian National Herbarium | GBIF |
| specimen | Paris Herbarium | Paris Herbarium | Paris Herbarium | GBIF |
| plot | CTFS Forest Plot Dataset | The Center for Tropical Forest Science | The Center for Tropical Forest Science | The Center for Tropical Forest Science |
| plot | FIA Forest Plot Database | Forest Inventory and Analysis National Program | Forest Inventory and Analysis National Program | Forest Inventory and Analysis National Program |
| plot | The Alwyn GentryForest Transect Dataset | Missouri Botanical Garden | Missouri Botanical Garden | SALVIAS |
| plot | The Enquist lab transects | Brian Enquist, Brad Boyle | SALVIAS | SALVIAS |
| plot | RAINFOR 0.1 ha transects | Multiple owners (RAINFOR) | RAINFOR | SALVIAS |
| plot | The Brad Boyle transects | Brad Boyle | SALVIAS | SALVIAS |
| plot | Madidi Forest Plots | Peter Jorgensen | BIEN | Peter Jorgensen |