**Informatics Tools for Biodiversity Data Conversion**

Kittinan Ponkaew

University of California, San Diego

**Abstract**

Maintaining and accessing biodiversity data is an extremely important issue for many institutions, but such data is usually stored in multiple differing formats and have not-uniform entry tags, thus making access difficult. Creating a centralized database that contains all data in a unified format addresses these problems. This project was a collaboration between the San Diego Natural History Museum (SDNHM), Balboa Park Online Collaborative (BPOC), and the UCSD PRIME program. SDNHM has multiple collections of taxonomy, each stored in its own database, thus creating a central database containing all the information that can be updated periodically in an efficient manner is critical for the future use of SDNHM collections.

**Introduction**

Each collection of taxonomy in SDNHM is stored in its own database, making accessing these data uniformly difficult. This project aims to address that problem by aggregating each separate database into a central database. This central database will then allow SDNHM to present its data to various audiences: the public, Global Biodiversity Information Facility (GBIF), etc. The goal is to aggregate multiple taxonomy databases from SDNHM into a centralized database. Tools for such conversion process will be needed, and made in such a way that they are reusable by others who will be working on the project in the future, and will allow the museum to update the database as they collect more information, without having forcing all departments to use the centralized database to store their information. This will allow each department to keep the way they are

collecting data the same and minimize the amount of changes they need to make to the way data are collected.

**Method**

Specify 6.5 (University of Kansas, Lawrence, KS), a biodiversity collection application, was used as the centralized data aggregator. Specify offers various tools and features that made it the application of choice. Once of the biggest benefit of specify is its ability to export data into Darwin Core Archive (DCA) format. This will allow the data to be exported and uploaded to GBIF and possibly other institutions in a uniform format. Specify toolsets interface nicely with Microsoft Excel 2003, making MS Excel a format to convert data into in the case that the current taxonomy database cannot directly interface with Specify. Another feature of Specify that falls in line with achieving our goals is its ability to export data into Apache Solr (Apache Software Foundation), a search platform. This allows SDNHM to create a web portal, letting users, both scientific and non-scientific, to search through the centralized database for information. This also gives SDNHM a finer control of how to present data to the users.

To convert data into specify, Python based tools were created (Python 2.7). Python offer several plugins that made it suitable for the work. Such plugins include the ability to write and read data from MS Excel 2003 spreadsheet. Since Specify accepts data from MS Excel, Python scripts were written to work with it, before being imported into Specify.

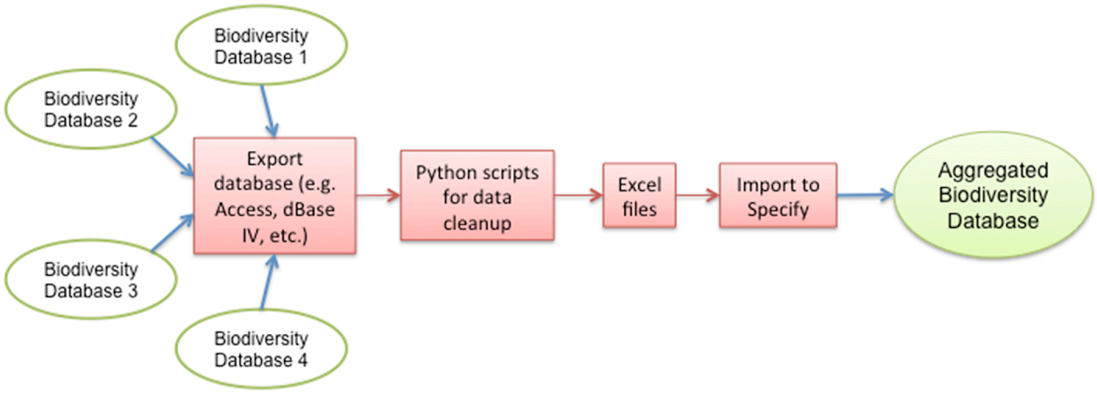


*Figure 1: Data conversion flow chart, courtesy of Jason Hagg.*

The first step of processing these data was to export them into excel files. Data in Microsoft Access did not pose much problem since the two were fairly compatible. Since some database were in Excel format already, this process was made easier. The reason for exporting into excel was because the data in the database may have several information in excess of what was needed. This process allows the data to be parsed from excel and be cleaned up before being imported into Specify. The next step was to create python scripts for each of the database in order to clean up the data. A lot of the data were not needed yet, for example, information on preparation type, and accession data. The data type of taxonomy data were then extracted from the original excel file and formatted into another excel file each with the label and format that Specify will quickly accept. When the data are imported into Specify, the interface recognizes the individual labels, and knows which category to parse the data into.

One of the biggest challenges for this project was extracting the data uniformly. What this means is that paleontology database has different significant information compared entomology database, such as stratigraphy and geochronology. Specify has a

feature to separate divisions, or collections, allowing individual collections to be stored separately even though they are still in the same database. Another problem was to prevent data from conflicting with each other. These taxonomy data are tagged with catalog numbers from their departments, so numbers will repeat themselves. By extracting the data from the original database, and parse it accordingly, custom tags were added in such a way that each data entry can be traced back to which department it came from and what the original number was, without causing conflict in the centralized database.

**Conclusion**

Even though not all databases were converted, the project was very successful. At the beginning of the project, there was a lot of pessimism as to how much of the database will be compatible, how long it will take, and how much of the data can be used. However as the project progress, due to a lot of similarities between multiple databases, when one database's conversion process finished, the same process can be used again with another database, thus saving time by not having to start everything from scratch. A lot of testing went into the project to ensure that everything works out the way it supposed to. The tools used are reusable, making future conversion easier to approach in order to update new data. Furthermore, this new centralized database forces SDNHM to look at how they format data, how these data can be use, what is necessary and not, and will allow further progress to be made.

**Acknowledgements**

I would like to thank the following people for the opportunity, and this experience a success:

**UCSD**

- Dr. Gabriele Wienhausen
- Dr. Peter Arzberger
- Teri Simas
- Jim Galvin
- Tricia Taylor
- Dr. Jason Haga

**NICT**

- Dr. Shinji Shimojo
- Takata Tomoaki

**BPOC**

**National Science Foundation**

**University of Kansas** (For their great software)