The Taxonomy building process

Here are some observations and suggested methods based on the experience gained from the Botanical taxonomy building project.

# Initially

There should be a prioritized list of taxonomy spines for ingestion into the end spine. (IE. The taxon product we will end up matching specimen names against).

1. Local taxonomies that are in use. Their names are likely to be present on the collection specimen labels. The Specify taxonomy being an obvious component.
2. Authoritative taxonomies that are widely accepted, such as Catalog of Life, CoL , or low hanging fruits like the GBIF taxonomy, which contains: National checklist of all species occurring in Denmark, ITIS, Dyntaxa, etc. Many checklists not mentioned here will already be present in the GBIF backbone.
3. Ask curators and local taxonomy experts for checklists and recommendations.
4. Make sure to always add a column “taxonomic\_source” that states which checklist the added records hail from. Like ‘Specify’ or ITIS or ‘Dyntaxa’.

# Process

Initially one checklist is uploaded to the developer database. The GBIF taxonomy would be a good place to start. Personally, I am using Postgres as it has a great palette of SQL tools with which to clean and manipulate data.

The addition of new spines will result in a specific number of DB records. This number should steadily grow as more spines are added. This is a simple check to perform.

**Important: Clean each spine before making any joins or mergers with the development taxonomic spine! Having to step back in the dev spine is a pain in the ace.**

Import each spine (after cleaning) into its own table in the dev DB. After each new spine is joined to the dev spine: A “unit test” should be run to see if what is expected is there!

1. Unit test could be a number-of-records check
2. Or a list of names check
3. Duplicates check
4. Did we lose any names since last iteration?
5. All of the above

This should be scripted in Python for easy initiation and completeness.

# Ongoing testing

There should be a versioning system in place so that taxonomy versions could be compared to each other. This would satisfy /Process-imprimis “number-of-records” check.

# 