**An Informal Report on the Feasibility of Extracting Various TNU Fields from Digital Born PDFs**

If a field is omitted and you think I should have commented on it please let me know. I’ve made plenty of assumptions about how taxonomic conventions work, if any of them are dangerously wrong please let me know that as well.

**“Required” fields:**

**TaxonomicName:**

Most fields “belong” to a scientific name, so if the name is not detected initially many other fields will fail to be detected as well. It is very important that the system does not fail to identify scientific names. Due to the nature of scientific names, it is impossible to correctly identify scientific names *100%* of the time. To resolve this conflict, it is necessary that some amount of manual/review is done to ensure all names are correctly extracted.

For names which contain identifiers like sp. it is definitely possible to identify each name’s general location, but can be difficult to choose the bounds around the name and would require some input from the user to ensure the name is correctly identified. Eg for the string:

“Where they discovered Goodenia asteriscus Peter Lang, sp. Nov. was…”

there can be issues where “discovered” might be interpreted as part of the name, or conversely where “lang” is not interpreted as part of the name. There are methods of mitigating this issue like adding “discovered” to a list of words to ignore and using things like capitalisation and italics as hints, but they are not foolproof and can fail if authors format their names inconsistently. The main benefit of this identification approach is that *names with registered identifiers will never be overlooked entirely.*

*For names which do not use identifiers:* After doing some research I have discovered some attempts to automatically grab scientific names including:

“TaxonGrab ([Koning et al. 2005](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6351503/#B4507597)), TaxonFinder ([Leary et al. 2007](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6351503/#B4507607)), OrganismTagger ([Naderi et al. 2011](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6351503/#B4507628)), NetiNeti ([Akella et al. 2012](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6351503/#B4503796)), SPECIES tagger ([Pafilis et al. 2013](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6351503/#B4503846)), BiOnym ([Berghe et al. 2015](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6351503/#B4507638)) and Global Names recognition and discovery tool (GNRD) ([Pyle 2016](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6351503/#B4507706)).”

An interactive example of one of the better ones can be found at <http://gnrd.globalnames.org/>. Shortcomings of this system include that it highlights names that are just mentioned in passing and is not completely accurate, meaning it could possibly fail to detect a valid scientific name. Furthermore these tools only select the name itself, so the problem of selecting the bounds of the name (to include things like authority, year, etc.) is still the same as with the identifier approach.

Sub-fields:

If we can correctly identify and parse scientific names then we should automatically have many of the other fields, including:

taxonomicName, taxonomicNameStringWithAuthor, uninomial, Genus, infragenericEpithet , specificEpithet, infraspecificEpithet, basionymAuthorship, taxonRank, verbatimNameString, Microreference, namePublishedInYear, scientificNameAuthorship, basedOn, basionymAuthorship, combinationAuthorship, combinationExAuthorship, basionymExAuthorship and kindOfName.

This assumption relies on:

1. The necessary information being present in the name (eg. sometimes names may not specify authorship/year)
2. The name being formatted according to standard conventions (variations can be dealt with, as long as they are logical)

Currently GNParser is the tool used by us to parse name strings, you can try it here: <https://parser.globalnames.org/?q=Glabu+tomentosum+Nutt.+ex+Seem> . The parser has some issues, (eg. it cannot deal with identifiers like sp. or sec. as part of the name) but is very fast and has a >99% accuracy when tested on a variety of complex scientific names.

**TaxonomicNameUsage/Concept:**

My understanding of this field is far from perfect, but I believe that (usually) this field will either state that something is a protonym, a subsequent usage/variant or a synonym. Protonyms will be the easiest to deal with due to the sp. Nov. identifier, subsequent usages could be detected by searching existing databases.

Synonyms and variations will probably be very difficult to detect in plain text, however if each usage is attached to a name, then as long as the name is detected the user can be made aware that the program was unable to extract the name’s usage. The user can then be given an opportunity to manually determine the usage without risking a false negative. Because I don’t fully understand what a TaxonomicNameUsage can entail, I don’t want to definitively say non-basic usages can’t be detected in plain text, but it seems very likely that this is the case.

TaxonomicNameUsage is tied to fields including kindOfNameUsage, protonym/basionym and other fields which are essentially just the different names/resources/references that the usage concerns and thus can’t be detected if the name usage itself is not correctly identified. If all “relationships” belong to named taxa then TaxonomicNameUsage would also give the relationshipType and hasRelationship fields.

**NomenclaturalCode:**

From my understanding there is a finite number of codes and the codes would be in the form of “Code Identifier” - “Number” or something similar (eg. ICZN-123456) If this is the case then identifying the code should be trivial provided that it is present somewhere in the PDF text (although for multiple of the case studies this does not seem to be the case). Obviously if the data is not in the PDF there is no way of extracting it, although it could be possible to make an educated guess based on the contents of the paper (eg. papers talking about leaf measurements are probably either ICN or ICNCP).

**“Non-Required” fields:**

**CultivarGroup:**

If I understand correctly, this is denoted by group/Gp and there is a finite list of groups available. This should be easy to both detect and classify.

**namePublishedInYear and scientificNameAuthorship:**

Sometimes this information isn’t present in the name explicitly. The program can try and find the date and authorship of the paper itself and the user could choose to default to this value instead if no name is specified.

**TypeOfType:**

If it’s explicitly mentioned (which seems to be the case mostly) then this is easy.

**Other Type Details:**

Assuming that TypeOfType is identified, finding the gender should be easy. Finding the name/ID of the type may or may not be difficult based on the naming conventions used, which I would have to talk to you about because I’m struggling to find anything consistent online.

**NameAccordingTo:**

This field seems to be usually tied to the “sec.” identifier within scientific names. Unfortunately GNParser does not handle this identifier, but it should be possible for us to create our own solution for this specific case.