Reconciliation Data for Providers - Notes from Nature

File Information

So you have completed an expedition in Notes from Nature and gotten the reconciled transcripts back and are wondering what to do with all of these docs! Here, we explain what each of the files are and how to use them. We also explain the logic behind the reconciliations.

Each file will have the same four digit workflow number and the same expedition name. For example, the new world swallowtail expedition is workflow number 2054, therefore all of the files will have '2054_New_World_Swallowtail' in the name.

Some of you will have provided data about the subjects at the time of expedition creation. This data can be found at the end of the files with each of your field names combined with "subject_". So if you included "collection code" information when you provided your manifest files, that field will be included in the output files as "subject_collection code".

There are three files you get:

####_expedition.summary.html
> -- overview of the reconciliation success as well as highlighting records that might want to be double checked and why. This is an HTML file and should open in any browser (with the possible exception of Internet Explorer).

####_expedition.raw_transcripts.csv
> -- contains all the transcriptions for each record ~3 rows per record

####_expedition.reconciled.csv
> -- contains the reconciled transcriptions == 1 row per record

Summary File:

This file has everything you need to understand what your citizen scientists did and how we reconciled each record. The file gives the overview of the reconciliation process, including the summary statistics for the number of records, number of transcripts and how many were reconciled with no conflicts. It also provides all the records, which you can examine fully. Lets take a look at the first part of the file:

Date:	2017-01-04 12:23
Number of Subjects:	315
Number of Transcripts:	1,003
Transcripts per Subject:	3.18

Reconciliation Summary

Field	Type	Reconciled Exact Matches	Fuzzy Matches	All Blank	One Transcript	Total	No Matches
Country	select	315		0	0	315	0
County	select	304		7	2	313	2
StateProvince	select	315		0	0	315	0
ScientificName	text	268	47	0	0	315	0
Location	text	87	223	2	2	314	1
HabitatDescription	text	102	152	19	30	303	12
CollectedBy	text	282	33	0	0	315	0
CollectorNumber	text	291	4	4	13	312	3
Month	select	315		0	0	315	0
Day	select	315		0	0	315	0
Year	select	315		0	0	315	0

You can see, for example, that for this expedition, the field "Location" had 87 exact matches, 223 fuzzy matches, 2 records where the three transcribers all left the field blank, and 2 records where only one of three transcribers provided content.

The second part of this file gives you access to all of the reconciled transcripts for further examination. First note that you can filter your outputs to only examine problem records, or problem records for a particular field. You might want to skip records that didn't have any issue, and just focus on the problematic ones. Below you can see the default view with the "Filter Rows" drop down set to "Show All"

Reconciliation Detail

Filter Rows: [Show All]

	subject_id	classification_id	Country	County	StateProvince	ScientificName	Location	HabitatDescription	CollectedBy	CollectorNumber	Month	Day	Year	subject_catalogNumber
+	2060359		United States of America	Martin	Florida	Cassytha filiformis	Jonathan Dickinson State Park (S. Fla. Water Management District wetland site 26). T40S, R42E, S1/2 of NE1/4 Sec 9.	Frequent parasite on shrubs of open flatwoods bordering marsh.	Loran C. Anderson	17494	6 - June	10	1997	000094322

Note the "+" sign next each record. Let's look at a problematic record, with the plus sign clicked so you can "open" that record and look at it.

subject_id	classification_id	Country	County	StateProvince	ScientificName	Location	HabitatDescription	CollectedBy	CollectorNumber	Month	Day	Year
2060386		United States of America	Jackson	Florida	Lindera benzoin		Frequent, slightly elevated places, floodplain woodland, near the river, Apalachicola River, E of Sneads.	R. K. Godfrey	78388	2 - February	18	1981
		Exact match, 3 of 3 records with 0 blanks	Exact match, 3 of 3 records with 0 blanks	Exact match, 3 of 3 records with 0 blanks	Normalized exact match, 3 of 3 records with 0 blanks	No text match on 3 records with 1 blank	Normalized exact match, 3 of 3 records with 0 blanks	Normalized exact match, 2 of 3 records with 0 blanks	Normalized exact match, 3 of 3 records with 0 blanks	Exact match, 3 of 3 records with 0 blanks	Exact match, 3 of 3 records with 0 blanks	Exact match, 3 of 3 records with 0 blanks
	13347387	United States of America	Jackson	Florida	Lindera benzoin		Frequent, slightly elevated places, floodplain woodland, near the river, Apalachicola River, E of Sneads.	R. K. Godfrey	78388	2 - February	18	1981
	13415852	United States of America	Jackson	Florida	Lindera benzoin	Not specified.	Frequent, slightly elevated places, floodplain woodland, near the river, Apalachicola River, E of Sneads.	R.K. Godfrey	78388	2 - February	18	1981
	13424306	United States of America	Jackson	Florida	Lindera benzoin	Apalachicola River, E of Sneads.	Frequent, slightly elevated places, floodplain woodland, near the river, Apalachicola River, E of Sneads.	R. K. Godfrey	78388	2 - February	18	1981

First, to orient you to the outputs. The first row shows the reconciled outputs, if there are any. You'll note that the "Location" field is blank. The second row, in grey, is a description of what happened during reconciliation. There are categories we have developed and that are hopefully understandable, but more details are available online. The rows following in green are the actual transcripts that were generated by the citizen scientists.

The problem field for this record is "Location", and we noted this by highlighting it in pink. In all other fields, there were exact matches. Normalized versus non-normalized exact matches is explained in more detail in the full description of the reconciliation tools we provide online.

The location field does not have final reconciled transcript, because there was no text match for the 2 records that were not blank. You can see the entries from the citizen scientists. In this case, a check of the locality information on the label is almost certainly warranted.

Raw_classifications File Info:

The goal is for each record to be transcribed three times before it is classified as finished. Therefore, each record will have 3 (or more) transcriptions. This file contains all of those raw transcripts. Difficult records might have more transcripts or if a record is skipped a lot it might have fewer. However, the majority of the records will have three classifications. This file can be used to check the reconciliations for more difficult records or to do your own reconciliations.

Reconciled File Info:

This file contains the reconciled data. The three transcriptions for each record have been examined and a single reconciled transcript determined. The reconciliation logic for each field is

spelled out below. This is the file that generally can be used to find the final data for each record.

Summary File Info:

This file contains the decision rules that were used for each reconciliation - the details are spelled out below but essentially if all of the classifications matched then it will say 'all agree' or if two of the three matched it will say 'majority rules' meaning the most common transcript was selected. If there is a question about how the reconciliations were done this is the file to check.

More information:

For more information about how we do reconciliation, we provide a more detailed description of the logic and categories etc. here. We also are more than happy to answer any questions you have. Please contact Michael Denslow: *notesfromnature.pm@gmail.com*