

PREDICTS 2 ETL Process

1 CONTENTS

1. Summary.....	1
2. Useful links.....	1
3. Extract names	2
3.1 Load unresolved names from database.....	2
3.2 Load manually resolved names.....	2
3.3 Merge streams.....	2
4. Clean + transform	2
4.1 String cleaning.....	2
4.2 Build GBIF query.....	3
5. Query GBIF.....	3
5.1 GBIF species match query	3
5.2 Transforming results	3
6. Evaluate GBIF matches	4
6.1 Handling unmatched records.....	4
6.2 Handling matched records.....	4
6.3 Check match confidence	5
7. Write out results.....	5
7.1 Resolved names	5
7.2 Unresolved names	7

1. SUMMARY

This document records the ETL process used by NHM Informatics to resolve taxonomic names gathered from biodiversity papers as part of the PREDICTS2 project. The GBIF species endpoint is used to resolve straightforward names (including synonyms) and to identify names in need of manual resolution by an appropriately qualified human. The steps laid out in the rest of this doc should be read in conjunction with ETL diagram.jpg.

2. USEFUL LINKS

ETL platform: [Pentaho Data Integration \(Community edition v8.3\)](#)

GBIF Species API (as of 01/04/2019 – 30/09/2020): <https://www.gbif.org/developer/species>

3. EXTRACT NAMES

3.1 LOAD UNRESOLVED NAMES FROM DATABASE

Unique (original species name, source name) pairs without a resolved name are extracted from `dbo.predicts_2.species_record`.

```
SELECT DISTINCT so.source_name, sr.species_name AS 'original_species_name',
sr.high_level_taxa AS 'high_level_taxa', sr.species_name AS 'original_species_name'
FROM species_record sr, diversity_report dr, sample_event se, site s,
study st, source so
WHERE sr.name_resolution_id IS NULL
AND sr.diversity_report_id = dr.diversity_report_id
AND dr.sample_event_id = se.sample_event_id
AND se.site_id = s.site_id
AND s.study_id = st.study_id
AND st.source_id = so.source_id
ORDER BY sr.species_name ASC
;
```

Query 3-1 - unresolved name extraction

3.2 LOAD MANUALLY RESOLVED NAMES

A locally-synced copy of `needs_manual_name_resolution.xlsx` is read in. It contains unique (original species name, source name) value pairs and a `manually_corrected_species_name` field where the PREDICTS2 team should enter any manually corrected taxonomic names we've been unable to match against GBIF.

If `manually_corrected_species_name` is NULL, the record is filtered out shortly after load.

3.3 MERGE STREAMS

As each record from 3.1. is processed, its (source name, original species name) fields are checked against the records from 3.2. On match, `manually_corrected_species_name` is returned and attached to the record.

field	description
<code>source_name</code>	Name of source/study the name was found in
<code>original_species_name</code>	Verbatim, unresolved species name
<code>manually_corrected_species_name</code>	Corrected form of name

Table 1 - fields present in stream at end of extract phrase

4. CLEAN + TRANSFORM

4.1 STRING CLEANING

`name_string_cleaned` field is created. It contains a copy of `original_species_name` where “. | _ | ,” characters have been replaced by whitespace. Non-standard delimiters can interfere with GBIF's fuzzy-match search. E.g.,

- [Boykinia.major](#): match confidence 94 against (correct) **genus**-level result
- [Boykinia major](#): match confidence 98 against (correct) **species**-level result

Occurrences of the strings ["species", "spp", "sp"] are also removed in `name_string_cleaned`. Use of these strings to e.g., indicate an uncertain species name impacts on the fuzzy-match search:

- [Antennaria](#): tied match confidence 99 against genus-level result
- [Antennaria spp](#): match confidence 74 against genus-level result

4.2 BUILD GBIF QUERY

If `manually_corrected_species_name` is NULL, new field `name_to_use` is set to the value of `name_string_cleaned`. Else, `name_to_use` = `manually_corrected_species_name`.

`searchTerm` is created. It contains a URL-encoded copy of `name_to_use`, which is appended to the end of the GBIF fuzzy match URL: <http://api.gbif.org/v1/species/match?verbose=true&name=>

field	description
<code>source_name</code>	Name of source/study the name was found in
<code>original_species_name</code>	Verbatim, unresolved species name
<code>manually_corrected_species_name</code>	Corrected form of name
<code>name_string_cleaned</code>	Cleaned version of species name
<code>name_to_use</code>	Version of the name that will be used to query the GBIF API
<code>searchTerm</code>	GBIF URL

Table 2 - fields in stream at end of clean/transform

5. QUERY GBIF

5.1 GBIF SPECIES MATCH QUERY

Multi-threaded calls to the GBIF API are made using the `searchTerm` URL created in the previous step (round robin distribution). Results are received in JSON and stored in a new field: `result`.

5.2 TRANSFORMING RESULTS

JSON record stored in the `result` field is flattened. At this stage, only values relating to the quality and nature of the GBIF match are extracted and appended to the record (the `result` field is unchanged). The current system timestamp is also added to the record

New field name	JSON path	Type	Description
<code>usageKey</code>	<code>\$.usageKey</code>	Integer	GBIF ID of usage
<code>matchType</code>	<code>\$.matchType</code>	String	Match type (controlled vocabulary)
<code>status</code>	<code>\$.status</code>	String	Taxonomic status (controlled vocabulary)
<code>acceptedUsageKey</code>	<code>\$.acceptedUsageKey</code>	Integer	GBIF ID of accepted usage if status = SYNONYM
<code>confidence</code>	<code>\$.confidence</code>	Integer	Measure of confidence in top match (0-100)
<code>note</code>	<code>\$.note</code>	String	Field containing more detailed match information.

alt	<code>\$.alternatives</code>	String	List of alternative matches in decreasing confidence order. JSON.
------------	------------------------------	--------	---

Table 3- fields added after query stage

6. EVALUATE GBIF MATCHES

6.1 HANDLING UNMATCHED RECORDS

The value of `matchType` is used to identify records without a clear ‘best’ match against the GBIF backbone. These fall into one of three categories:

- no matches found:** `alt` is NULL, `matchType` = ‘NONE’
- > 1 equally good match found:** `note` starts with ‘Multiple equal matches’, `matchType` = ‘None’
- no ‘good’ matches found:** `alt` is not NULL, `matchType` = ‘NONE’

Records of type a. or b. are written to `needs_manual_name_resolution.xlsx` when the pipe terminates. See section 7.2 for detail.

To handle records of type c. (i.e., matches exist, but the confidence level is below the GBIF quality-control confidence cut-off), the value of `alt` is flattened as described in section 5.2. This may comprise more than one record.

The `alt` result with the highest `confidence` is identified, extracted and fed back into the matched record process (see section 6.2). Most of these matches have low `confidence` and will require manual resolution, but this path was worth handling separately to allow finer control over the cut-off value.

6.2 HANDLING MATCHED RECORDS

Where a single, good match is found in the GBIF backbone, the `matchtype` field is examined and used to filter the results into two streams: synonyms and accepted names.

- Synonyms:** `acceptedUsageKey` is used to construct a second GBIF query to retrieve the accepted name. The record is then parsed in the same step as non-synonyms, see next.
- Non-synonyms:** These are flattened to extract taxon information, hierarchy and GBIF indicators:

Name	Path	Type
scientificName	<code>\$.scientificName</code>	String
canonicalName	<code>\$.canonicalName</code>	String
rank	<code>\$.rank</code>	String
t_status	<code>\$.taxonomicStatus</code>	String
kingdom	<code>\$.kingdom</code>	String
kingdomKey	<code>\$.kingdomKey</code>	Integer
phylum	<code>\$.phylum</code>	String
phylumKey	<code>\$.phylumKey</code>	Integer
order	<code>\$.order</code>	String
orderKey	<code>\$.orderKey</code>	Integer

family	\$.family	String
familyKey	\$.familyKey	Integer
genus	\$.genus	String
genusKey	\$.genusKey	Integer
species	\$.species	String
speciesKey	\$.speciesKey	Integer
class	\$.class	String

Table 4 – taxonomic fields extracted

NB: descriptions for most of these fields are given in the GBIF API docs: <https://gbif.github.io/gbif-api/apidocs/org/gbif/api/vocabulary/package-summary.html>, and <https://www.gbif.org/developer/species>

6.3 CHECK MATCH CONFIDENCE

Records are evaluated according to their `confidence` measure: only records with a value ≥ 95 are written to the database.

The confidence measure and cut-off point were selected following testing and analysis of available options. Some 600 unmatched names were fed into the pipe while variables such as cut-off level and similarity measure (including alternative fuzzy-matching algorithms e.g., Jaro, Levenshtein, Wunsch) were modified between runs.

The F1 score of confidence $\geq 95 = 0.95$, with precision of 1.

7. WRITE OUT RESULTS

7.1 RESOLVED NAMES

Accepted matches are de-duplicated and written to two tables in `dbo.predicts_2_live`:

- `tbl.gbif_species_reference` holds taxonomic data retrieved from GBIF: one record per species/genus record. If the GBIF `usage_key` is already present in the table, the record is updated.

predicts_2. gbif_species_reference	Stream field	Update on duplicate key
usage_key (PK)	usage_key	Y
scientificName	scientificName	Y
canonicalName	canonicalName	Y
rank	rank	Y
taxon_status	taxon_status	Y
kingdom	kingdom	Y
phylum	phylum	Y
order	order	Y
family	family	Y
genus	genus	Y
species	species	Y

class	class	Y
kingdomKey	kingdomKey	Y
phylumKey	phylumKey	Y
orderKey	orderKey	Y
familyKey	familyKey	Y
genusKey	genusKey	Y
speciesKey	speciesKey	Y
updated	name_parser_timestamp	Y

Table 5 - contents of `dbo.predicts_2.gbif_species_reference`

- b. `tbl.name_resolution` holds the audit trail/results of the name resolution process:

predicts_2.name_resolution	Stream field
name_resolution_id (PK)	[auto-increment on insert]
gbif_usage_key (FK)	usage_key
original_species_name	original_species_name
manually_corrected_species_name	manually_corrected_species_name
name_string_cleaned	name_string_cleaned
gbif_query_species_name	gbif_query_species_name
name_parser_timestamp	name_parser_timestamp
gbif_url	gbif_url
synonym_usage_key	synonym_usage_key
match_type	match_type
resolved_via_synonym	resolved_via_synonym
alternate_match_used	alternate_match_used
confidence	confidence
sim_name_confidence	sim_name
sim_authorship_confidence	sim_authorship
sim_classification_confidence	sim_classification
sim_rank_confidence	sim_rank
sim_status_confidence	sim_status

Table 6 - contents of `dbo.predicts_2.name_resolution`

The newly-minted `name_resolution_id` is returned when a row is written to `name_resolution`.

This key is used to update a third, occurrence-level table: `species_record`. Within this table, `original_species_name` is used to identify the appropriate line(s) to be updated:

Insert / update

Step name: set species_record.name_resolution_id

Connection: mattw

Target schema:

Target table: species_record

Commit size: 100

Don't perform any updates: ☐

The key(s) to look up the value(s):

#	Table field	Comparator	Stream field1	Stream field2
1	species_name	=	original_species_name	
2	name_resolution_id	IS NULL		

Get fields

Update fields:

#	Table field	Stream field	Update	
1	name_resolution_id	name_resolution_id	Y	

Get update fields

Edit mapping

Help OK Cancel SQL

Figure 1 - Insert/update criteria for name_resolution

7.2 UNRESOLVED NAMES

Unresolved names fall into one of the following categories:

- No results:** GBIF was unable to find a single match for this name
- Tied results:** > 1 record has the highest confidence match value
- No reliable results:** GBIF matches with a confidence score ≥ 94 (max. 100)

These are written to a new copy of needs_manual_name_resolution.xlsx for investigation + resolution by the P2 team. The fields included are:

field	description
source_name	Name of research study/source from dbo.predicts_2.source
non_standard_name_note	Indicator of type of name from dbo.predicts_2.species_record. E.g., 'Common name' or 'Uncertain taxa'
high_level_taxa	Higher taxonomy indicator (can vary in level). Value from dbo.predicts_2.species_record
original_species_name	Original species name from dbo.predicts_2.species_record
manually_corrected_species_name	Null (to be populated by PREDICTS2 team)
name_parser_timestamp	Date the pipeline was last run
gbif_best_match	Match(es) with highest confidence suggested by GBIF. Ties results are '/' delimited.
gbif_best_match_confidence	Confidence score of highest match suggested by GBIF

Table 7 - composition of manual resolution spreadsheet