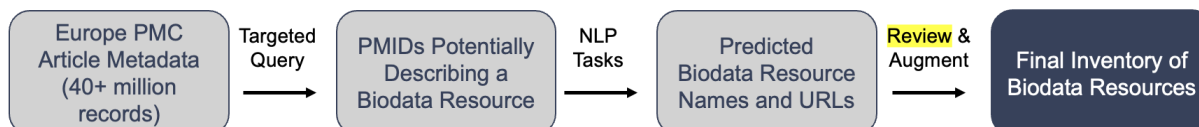


Manual Review Process for the Biodata Resource Inventory

Purpose: This document accompanies the Biodata Resource Inventory project and provides guidance for reviewing the preliminary inventory of predicted resources that results from the Natural Language Processing tasks. This is the “review” step highlighted in yellow:



In the 2022 inventory (Imker et al. 2023 <https://doi.org/10.1371/journal.pone.0294812>) and for any subsequent updates to the inventory, a curator reviews flagged records prior to augmenting the inventory with additional metadata and finalizing the inventory. The flags indicate that human judgment should be used to determine if low probability records should be removed from the inventory and if potential duplicate records should be merged within the inventory. The file for manual review comes from the Biodata Resource Inventory pipeline as a CSV. In this file, there are three columns that indicate a record has been flagged for manual review: `duplicate_urls`, `duplicate_names`, and `low_prob`.

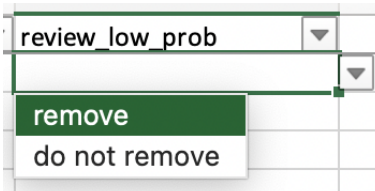
This guide was updated to V2 following the first update of the inventory in July 2024 to add new resources with associated papers published in 2022 and 2023.

STEP 1: Set up manual review spreadsheet

1. The file to review is named **predictions.csv** and is found in `out/new_query/for_manual_review` after running `! make update_inventory`
2. Open the CSV file in MS Excel and immediately save as with “_review_V1” appending the file name. If work is done over multiple days, recommend saving each new day as a new version iteratively to be able to go back in case of mistakes.
3. Data -> Filter to be able to sort all columns (including new) without scrambling. **THIS IS ESSENTIAL.**
4. Create drop-downs for consistent, structured values for review columns:
 - a. Create a new sheet/tab and name as “Review Values”
 - b. On this sheet, create a table of options **exactly as below** (changing any values will break the script that processes the manual review results):

review_low_prob	review_dup_urls	review_dup_names
remove	merge on record with best name prob	merge all "dup name" IDs
do not remove	do not merge	do not merge
	conflicting record(s) to be removed	merge only:

		conflicting record(s) to be removed
--	--	-------------------------------------

- c. Lock the Review Values tab to prevent accidental changes (right click on tab name at bottom -> Protect Sheet... -> allow users to select locked and unlocked cells -> OK)
 - d. Back in the manual review sheet/tab, highlight entire column for "review_low_prob"
 - i. Data -> Validation -> Allow; select list; highlight the "remove and do not remove" cells for "review_low_prob" on the Review Values tab
 - ii. Check that only those values are available now for that column
 
 - iii. Repeat for "review_dup_urls" and "review_dup_names"
4. Optional: Conditional format best_name_prob to be green = high, red = low

STEP 2: Review flags for low_prob using "review_low_prob" column

1. Sort by low_prob
2. Review each, referring to text column as needed
 - a. If extracted name is erroneous, select "remove" from dropdown
 - i. Guidelines:
 - Allow either full name or short name/abbreviation - as long as correct, retain in the inventory
 - Do not remove records for names missing a general type, e.g. "database" or "DB" or "catalog" or "data hub" or "library"
 - e.g. do not remove: "tautomeric" when actual name is just "tautomeric database"
 - e.g. remove: "HGD" when actual name is "HGD mutation database"
 - Flag for removal when any key part of the name was missing, e.g. "GlycositeAtlas" (predicted) vs "N-GlycositeAtlas" (actual)
 - ii. note reason for removal in "review_notes_low_prob"
 - FALSE POS: CLASS
 - FALSE POS: INCORRECT NAME
 - FALSE POS: PARTIAL NAME
 - FALSE POS: URL scramble
 - iii. When complete, double check numbers between "low_prob" and "review_low_prob" to make sure all have been reviewed.

STEP 3: Review flags for duplicate URLs using “review_dup_urls” column

1. Sort by extracted_url so duplicate urls line up
2. Review each, referring to text column as needed
 - i. Guidelines:
 - If URLs refer to the same resource and merging the records on the record with best name probability will not result in an erroneous name, select “merge on record with best name prob”
 - If URLs DO NOT refer to the same resource OR merging the records will result in an erroneous name overriding a correct name, select “do not merge”
 - If the record is flagged and removed because of the evaluation in Step 2 above, select “conflicting record(s) to be removed”
 - ii. Note any odd cases in “review_notes_dup_url”
 - iii. When complete, double check numbers between “duplicate_urls” and “review_dup_urls” to make sure all have been reviewed.

STEP 4: Review flags for duplicate names using the “review_dup_names” column

1. Sort by best_name so duplicate names line up
2. Review each, referring to text column as needed
 - i. Guidelines:
 - If predicted names are accurate and applies to all of the records (just variation between URLs), select “merge all “dup name” IDs”
 - The “main” URL will be either the one associated with the most recent publication or, in the case of a tie for newest publication date, the one associated with the highest name probability
 - If predicted names are accurate for **only some of the records**, select “merge only:” and put the IDs to be merged in the “review_notes_dup_names” column (comma separated) with no other text, e.g. “26481361, 31647100”
 - For any records that should not be merged, (e.g. different resources or merging would cause some other issue), select “do not merge”
 - If the record is flagged and removed because of the evaluation in Step 2 above, select “conflicting record(s) to be removed”
 - ii. When complete, double check numbers between “duplicate_names” and “review_dup_names” to make sure all have been reviewed.

STEP 5: (new in 2024) Remove special characters, if needed

1. The prediction.csv file may contain a few (e.g., < 5) names or URLs with long strings of special characters (e.g., Ç√É¬Ç√Ç¬É ... on MacOS or ÃfÂfÃ,ÂfÃf... on PC) appended to them due to encoding errors. Steps may be taken to mitigate this issue in future updates to the models and/or pipeline, but for the current

instantiation, remove these strings or they may throw errors for the remainder of the pipeline. *This is the only direct remediation to the predicted values made.* Note that because this remediation happens after strict depublication and flagging, there could be a few missed deduplications in the final inventory.

STEP 6: (new in 2024) Check date format and save csv file

1. The publication date format must be D/M/YYYY for the remainder of the pipeline, change if necessary using format cells -> date -> type 3/12/2004
2. Save tab as predictions_v[#].csv

STEP 7: (new in 2024) Perform a pre-check to ensure all flags are accounted for

1. Prior to initiating the remainder of the pipeline, run R script, STEP_1_precheck_manual_reviewed.R, to checked that all flags are accounted for: https://github.com/1heidi/inventory_update
2. Correct any issues as needed, recheck, and save file via R as **predictions.csv** (must be this name exactly)

STEP 8: Add the reviewed inventory into directory for further processing

1. Add final manually reviewed file **predictions.csv** to out/new_query/manually_reviewed
2. Run the post-review pipeline. Scripts within the pipeline will first check that all flagged records have been reviewed and contain valid review values; if not, errors will indicate where there are issues.