How the GND4C-Toolbox works

Prepared by

Michael Markert & Erdal Ayan (Software Developer)

Contact:

Email: erdal_ayan@yahoo.com

Table of Contents

The toolbox generates scores for results from the GND on imported datasets. It is a five-step process for each entity type – Importing, Querying, Matching, Scoring and Recommending.

Step-1: Data Import for Person Objects (and Buildings Objects)

Datasets

In the toolbox at the moment there are basically two options in order to import datasets from both DigiCult's .json Api (NDS) and private datasets. Here "private datasets" refer to the datasets which are not available on DigiCult's .json Api but come from a partner institution's own sources.

Options

- NDS Data Import
    - in this option the user can import the (online) datasets formatted in the available .json Api via toolbox's import module
    - in this option the user can also prefer to use a .csv file to import the NDS datasets if the dataset was prepared as .csv
    - at the moment this option works only for Person objects.
- Private Data Import
    - In this option the user can prepare the datasets (both Person and Building objects) in accordance with the .json standards that were created and optimized for this purpose. The instructions can be found in the import module option.

Before the import process, the user should login to the Toolbox environment with correct credentials and give a particular "name" for the dataset and then either check the dataset before saving it via listed dataframe or download the dataset as .cvs or directly save the dataset. When the import process is successfully completed and saved in the Toolbox's database, then the user can continue with querying, matching and scoring parts defined below.

Step - 2: Querying for Person Objects

- For the querying of person objects, all combinations of name elements (name, surname, name addition, personal name, prefix, counting, birth and death dates) in the source data and all alternative names are queried via lobid.org as far as there are at least two name elements or only a single element for the objects with single elements.
  Example combinations for the object name: "Hugo Abensperg Traun"

1	Hugo~2+Abensperg~2
2	Hugo~2+Traun~2
3	Abensperg~2+Traun~2
4	Hugo~2+Abensperg~2+Traun~2

With the “~2” in the query every string element internally on the lobid side is matched even to strings that are different in up to 2 positions, including deletions, additions, and replacement. So “Traun~2” would also deliver names with the elements “Daun” or “Traum” in the response (for further information on the underlying concept see https://en.wikipedia.org/wiki/Damerau%E2%80%93Levenshtein_distance).

- Name additions (von und zu, de la) and punctuations are cleaned from the candidate query content/parameters before creating the query search terms..
- Only the first elements of each GND response are processed, you can choose the number (as default 20, 40 to 60 at the moment) on the matching page.
- If dateOfBirth and/or dateOfDeath is provided, all queries are repeated including these values. This makes sure that for common names like Stefan Müller the ones with fitting life dates are ranked higher and thus can be found in the processed part of the response.
- When the search terms are constructed with the chosen query parameters, the query sets are iteratively sent to the Lobid platform and if the returning response is positive and includes .json objects at all, the matching process starts.

Step - 3: Matching

- Initially, the returning .json objects are converted into pandas dataframes and cleaned (punctuations, strange characters, extra spaces, etc) and standardized (lower cases).
- Attributes used for matching process are listed as following:

Source data (NDS or Private Datasets)	*Target data (GND via Lobid)*
Preferred names	Preferred names
Preferred names	Variant names
Non-preferred names	Preferred names
Non-preferred names	Variant names
Birthdate	Birthdate
Birth Place	Birth Place
Deathdate	Deathdate
Death Place	Death Place
Profession(s) (via codes)	Profession(s) (via codes)
Period of Activity (Start and End)	Birthdate and Deathdate

Step-4: Scoring

The source data object is evaluated against all responses in several steps:

Names

max_name_score

- The number of identical name elements is counted, e.g. source *Stefan Müller* vs. *Stefan Klaus Müller-Schmidt* in GND(lobid) response => max_name_score is 2. This goes for all name variants that are provided in the source and in the Lobid response.
- For the scoring process, always the highest value is used, independently of it being a ‘preferred’ or ‘alternative/variant’ name.

IstMaxNameScore

- This variable counts the number of highest values of max_name_score for one comparison. So if there are the name variants *Stefan Müller* and *Stefan Richard Müller* in the source and *Stefan Klaus Müller-Schmidt* as well as *Stefan Klaus Müller* in one GND object, IstMaxNameScore would be 4 (because 2 elements of both name variants in the the source match 2 name elements (max_name_score is 2) of both name variants in the GND). The underlying assumption is that if there is a high similarity between several source and GND variant names it is more likely that this is a match.

IstMx_x_Anz

- In the GND especially with common names several potential matches have a high IstMaxNameScore (as you can see with the example above). Therefore IstMx_x_Anz was introduced to count the number of respective GND person objects. In the total score it serves as a negative correction factor: If several objects in the GND response have the same (highest) IstMaxNameScore, IstMx_x_Anz negatively (see section Total Score) contributes to the total score to avoid false auto-matches and misleading rankings of ‘wrong’ GND person objects.

Dates

- The years of birth and death are compared. If the date is identical in source and response the score is 1, if it differs about up to 5 years, the score is reduced by 0.1 for each year, e.g. 1902 vs. 1898 => score is 0.6.
- If there is no date to compare on either side the score is 0.
- If there are dates in original data AND the GND but the difference is larger than 5 years, the score is -1.

Period of Activity

- If a periodOfActivity is provided in the source data, the “start” and “end” values are compared to birthdates and deathdates in the GND. The assumption is that people need to be alive to be active.
- If the periodOfActivity starts within 100 years after birth, the value is 1, else -1. If the periodOfActivity starts within 100 years before death, the value is 1, else -1.
- Same goes for the ends of activity periods.
- If a value is missing, the score is set to 0.
- All four values are summed up, thus the maximum value here is 4 (start and end of a period of activity is within the phase between birth and death)

Birth and Death Places

- If a place of birth or death is provided and it is identical in the source and the GND the value is 1, if it differs the value is -1.
- If a value is missing, the score is set to 0.

Professions

- The evaluation of professions or ‘jobs’ is the most complex part of the toolbox scoring as simple value to value comparisons often are not useful here. Johannes Bracht (digiCULT-Verbund e. G.) developed a solution to match professions and provided a list containing hierarchically listed and grouped names of professions based on the professions in the GND. Every profession is provided with an identification number with up to six subgroups, e.g. Alphornbläser has the ID 004.002.003.016.002.001 with 004 being Musik, 004.002 Musiker, 004.002.003 Instrumentalist, 004.002.003.016 Blasmusiker and so on.
- The evaluation is positive when the profession from source and GND matches at least on a higher level, e.g. when one side says Alphornbläser (004.002.003.016.002.001) and the other says Straßenmusiker (004.003.001). The scoring value is lower or higher depending on the number of matching xxx-Elements in the profession ID.
  If only the first three digits (first group) are identical the score is 0.25, if the first two groups are identical (six digits) the score is 0.75, and if three or more groups are identical (at least nine digits) the score is 1. If there is no value or not even the first group matches, the score is 0.
- If more than one profession is provided in the source and/or the GND, all values are compared and the highest one is used.
- The list is fixed at the moment so if a profession is missing in it but exists in the dataset the score will always be 0.

Total Score

- All parameter values are summed up to generate a total score.
- As data elements of entities besides names (like life dates or professions) are not evenly distributed neither in the source data nor the GND, corrective multipliers were introduced by Johannes Bracht (digiCULT e.G.). They are meant to redefine the ‘role’ of a parameter during scoring, i.e. profession information is much less important for GND matching than the date of birth so the multiplier for this parameter is much lower.
- Additionally, the results are shifted towards negative values to have a clear limit: A total_score below 0 most likely is a false match, a total_score above 0 and below 2.5 needs to be checked carefully and a total_score above 2.5 is most likely a positive match.
- All values for these multipliers were defined during the toolbox development based on test datasets from different institutions using linear regression. Thus, they do not fit to every use case, but it is possible to change all parameters on the scoring page of the toolbox to adapt them to project needs.
- The basic calculation is
  total_score =
  -8.41255 +
  max_name_score * 1.298 +
  lstMx_x_An * -0.283 +
  lstMaxNameScore * 5.07 +
  max_job_score * 2.778 +
  G_birthdate_score * 1.139 +
  H_deathdate_score * 1.2015 +
  O_birthplace_score * 0.857 +
  P_deathplace_score * 2.024 +
  activity_period_ijklmn_score * 0.178
- Hint: If you don’t have total_scores > 2.5 but there are positive matches, try changing lstMx_x_An to 0.

Step-5: Recommending

- in the works