CLAIMS

1. An automated computer implemented method of deduplicating point of interest records of computer implemented database, comprising:

standardizing data of a pair of records (100);

classifying (200) the pair of records into a category of pre-defined categories;

carrying out a category specific determination of non-duplicate (300);

if the category specific determination of non-duplicate has failed to determine that the pair of records is not duplicated, tokenizing (400) the data of the pair of records into a pair of tokenized records;

determining whether core words of the pair of records are linked;

if it is determined that the core words are linked, computing a score of cross information and distance of the pair of tokenized records (600);

determining (700) whether the score of cross information and distance passes a pre-determined threshold;

if the score of cross information and distance passes the pre-determined threshold, determining that the pair of records is a duplicate; and

if the score of cross information and distance does not pass the pre-determined threshold, determining that the pair of records is a non-duplicate.

2. The method of claim 1, wherein the category is a point address category, and
the category specific determination of non-duplicate comprises determining that the pair of records have identical street labels (316), and
the category specific determination of non-duplicate comprises determining that the pair of records have identical house numbers (310).

3. The method of claim 1, wherein the category is a non-point address category and the category specific determination of non-duplicate comprises determining that the pair of records refer to a same entity (322).

4. The method of any one of the previous claims, further comprising linking tokens of the pair of tokenized records, wherein linking comprises:

linking common tokens (510) of the pair of tokenized records; and

linking similar tokens (530) of the pair of tokenized records.

5. The method of claim 4, wherein linking similar tokens (530) comprises:

checking for concatenation; and/or

checking for compound words.

6. The method of claim 4 or claim 5, wherein linking similar tokens (530) comprises computing edit distances.

7. The method of claim 6, wherein determining whether core words are linked (538) is based on the edit distances.

8. The method of any one of the previous claims, wherein standardizing data of the pair of records is based on one or both of: a country specific dictionary, and a language specific dictionary.

9. An electronic data processing system for deduplicating point of interest of computer implemented database, comprising:

a standardizer to standardize data of a pair of records (100), wherein equivalent words with different spelling are normalized to a single spelling;

a classifier (200) to classify the pair of records into a category of pre-defined categories;

a category specific determiner to determine if the pair of records is a category specific non-duplicate,

a tokenizer to, if the category specific determiner failed to determine that the pair of records is category non-duplicate, tokenize (400) the data of the pair of records into a pair of tokenized records; and

a scorer to determine whether core tokens of the pair of records are linked or not,

wherein if it is determined that the core tokens are linked, the scorer computes a score of cross information and distance of the pair of tokenized records and determines (700) whether the score of cross information and distance passes a pre-determined threshold;

if the score of cross information and distance passes the pre-determined threshold, the scorer determines that the pair of records is a duplicate; and

if the score of cross information and distance does not pass the pre-determined threshold, the scorer determines that the pair of records is a non-duplicate.

10. The electronic data processing system of claim 9,
wherein the category is a point address category, and to determine if the pair of records is a category specific non-duplicate comprises:

to determine if the pair of records have different street labels, and/or

to determine if the pair of records have different house numbers.

11. The electronic data processing system of claim 9, wherein the category is a non-point address category and to determine if the pair of records is a category specific non-duplicate comprises determining if the pair of records refer to different entities.

12. The electronic data processing system of any one claims 9 to 11, further comprising a linker to link tokens of the pair of tokenized records, the linker comprising:

a common linker to link common tokens of the pair of tokenized records; and

a similarity linker to link similar tokens of the pair of tokenized records.

13. The electronic data processing system of claim 12, wherein to link common tokens comprises one or both of:

a concatenation checker configured to check for concatenated words;

a compound checker configured to check for compound words.

14. The electronic data processing system of claim 12 or claim 13, wherein to link similar tokens comprises computing edit distances.

15. The electronic data processing system of claim 14, wherein to link similar tokens comprises determining that core words are linked based on the edit distances.

16. The electronic data processing system of any one of claims 9 to 15, wherein to standardize data of the pair of records is based on one or both of: a country specific dictionary, and a language specific dictionary.

17. A data processing apparatus configured to carry out the method of any of the claims 1 to 8.

18. A non-transitory computer-readable medium storing computer executable code comprising instructions for deduplicating point of interest of computer implemented database according to the method of any one of claims 1 to 8.