

Matching API

Adrian Achihăei
TrustYou GmbH*

December 6, 2011

Contents

1 Abstract	2
2 Input Format	2
3 Algorithm and Accuracy considerations	3
4 Output Format	3

*Agnes-Pockels-Bogen 1 80992 München Tel: 089--54802925 Fax: 089--381644539

1 Abstract

The purpose of the API is to match various types of entity ids (the ones used by TrustYou systems and your own ones) based on other informations like name, country ISO code, address, geocoordinates. If your records are collected from more sources the algorithm will be able to cluster them based on the provided attributes and associate them the same unique id.

2 Input Format

Data will be accepted in tsv (tab separated values) format without header chunked into files no larger than 20MB. No quotation signs will be used, the columns will be separated by tab `\t` character and the records will be separated by new line `\n` character. Each chunk of data will contain complete records and upon upload you will receive an universal identification number of your processing task. After processing finishes you will receive a notification per email and the optional callback URL will be called with the processing identification number concatenated to it.

Please make sure the individual records and or columns does not contain new line character (record separator) since no quotation format is assumed. Also the supported text encoding is UTF-8.

The following columns in the precise order will be provided. In case the data in a mandatory column is not provided the following character sequences `\N` `N/A` or `NULL` will be used for `` ` not available"`

id The first column will contain your own identification number for the particular entity. This id will be supplied in the output mapping file with the corresponding TrustYou id if matched. The API allows for specifying external IDs defined by other sources. Additional external IDs can be included as NAME:VALUE pairs after the actual ID. Additional NAME:VALUE pairs have to be separated from the actual and from each other using a semi-colon. The third line in the input example shows how to add ID information from GIATA and HolidayCheck. Supported external ID names are: GIATA, HOLIDAYCHECK, BOOKING, TRIPADVISOR, TRUST, MICROS.

name The second column will contain entity name (e.g. hotel name) as stored on your system.

country Entity country ISO code (e.g. the hotel is located in DE, US, GB, ...)

longitude Longitude (e.g. 11.4253890) part of geographic coordinates of the entity represented by the record

latitude Latitude (e.g. 48.7621450) part of geographic coordinates

address Last column will contain further address information like street, city name in any language, zip code.

Here is an example of data uploaded to be processed with matching API

```
df69ceb3f58b    Hotel Pius Hof  DE      11.4186
                48.7829 Gundekarstrasse 4, 85057 Ingolstadt
0bfe1f83da41    Hotel ARA                DE
                11.4594 48.7777 Schollstrasse 10a, 85055 Ingolstadt
b0da6dc89ac1;GIATA:225947;HOLIDAYCHECK:162862  GB
                -0.4848 53.8206 18A East End, Walkington
...
```

3 Algorithm and Accuracy considerations

The tools used to cluster the data are based on a set of specialized algorithms designed during specific data analysis conducted at TrustYou Labs. The algorithms provides good accuracy on various test data sampled and we are quite confident in the quality of the results however the more complete is the input set of data the better are the results. A computed accuracy factor will be provided in the result file represented as a floating number between 0 and 1. Lower numbers mean a reduced confidence that the records were correctly matched. One shall use the result data only if a significantly large percentage from the accuracy values are close to 1. Our statisticians will be happy to assist you choosing an adequate threshold based on you data usage requirements.

4 Output Format

The output will contain exactly the information contained in the input data and additionally contains information about the matching result in each line.

id The first column will contain your own identification number for the particular entity as provided in the input data.

name same as in input.

country same as in input.

longitude same as in input.

latitude same as in input.

address same as in input.

TrustYou id The second column will contain an universal identification number. In case the accuracy is above 0.5 it can be used to request data about entity over our Widget API. "n/a" will be displayed if no match could be found.

accuracy A computed accuracy factor in the range 0..1

comment An optional text containing comments.

Here is an example of data downloaded after processing was complete.

```
df69ceb3f58b    Hotel Pius Hof    DE    11.4186
                48.7829 Gundekarstrasse 4, 85057 Ingolstadt    2
                d1cfa38-2f1d-11e0-b72e-afc8628f099b    0.8997
0bfe1f83da41    Hotel ARA    DE
                11.4594 48.7777 Schollstrasse 10a, 85055 Ingolstadt
                n/a    0
b0da6dc89ac1 ;GIATA:225947;HOLIDAYCHECK:162862    GB
                -0.4848 53.8206 18A East End, Walkington
                4ae1afaa-2f1d-11e0-b50c-231f0d83dc59
                0.9238
```

...