

TrustYou Import Data Format ^{*}

version 2.2

Alin Sanislav
TrustYou GmbH[†]

Costin Cozan
TrustYou GmbH

Adrian Achihăei
TrustYou GmbH

October 7, 2010

Contents

1	Introduction	2
1.1	Purpose	2
2	Data format	2
2.1	Entity elements	2
2.1.1	Hotel entity elements	3
2.1.2	Restaurant entity elements	5
2.2	Review elements	7
2.2.1	Hotel review elements	8
2.2.2	Restaurant review elements	9
3	ID Tables	11
4	Matching	11
5	Quality Check	11
5.1	Duplicates Filtering	12
6	Document changes	13

^{*}©TrustYou 2010 - All rights reserved, proprietary information

[†]Agnes-Pockels-Bogen 1 80992 München Tel: 089-54802925 Fax: 089-381644539

1 Introduction

1.1 Purpose

The purpose of this document is to define the XML format of documents which can be interpreted by TrustYou Processing. Please make sure you received a copy of XML schema along with this document. The actual schema will contain more constraints than those listed here.

2 Data format

XML documents will be properly UTF-8 encoded. The XML root element should be called <root>.

Date format. All dates defined in this document are an epoch time-stamp representation (normally an integer between 599612400 and 5996124000000). Epoch is defined as milliseconds since January 1, 1970, 00:00:00 GMT. A negative number is the number of milliseconds before January 1, 1970, 00:00:00 GMT.

All uid tags will contain a valid UUID see [RFC 4122](#)

2.1 Entity elements

Mandatory fields:

type would denote the entity type. i.e. type=1 → hotel, type=2 → restaurant, a.s.o. (see „3.IDs Tables“ section)

source the source id which is an internal TrustYou source id (please ask for such an id if it's not known)

name name of entity (hotel name, restaurant name ...)

uid uniquely identifies an entity (valid UUID see [RFC 4122](#))

Optional fields that might improve matching and processing quality if available or applicable:

id source's internal entity id

mark/total the note/mark/score given by a specific source to the entity. It must be a numeric value, exactly how it is being provided by the source (no normalization must be applied to it)

stars standard rating of the entity, for example number of stars for hotels or restaurants

ratenow the absolute URL where the user can be guided to write a review for the entity

price a general price found on the page. if available it must be number, either integer or decimal (like 56 or 78.50 or 109,49)

price_currency string representation of the currency (symbol or name)

picture absolute URL to a picture which should represent the thumbnail

description entity's description

createdate when this entity element was generated as time-stamp in epoch (milliseconds since January 1, 1970, 00:00:00 GMT. A negative number is the number of milliseconds before January 1, 1970, 00:00:00 GMT)

reviews_cnt the number of reviews available for this entity if available

2.1.1 Hotel entity elements

In addition to entity fields the hotel entities must and could contain the following fields:

Mandatory fields:

address/city the city name

address/country the country name

address/url absolute URL to the hotel's webpage (source deep link)

Optional fields that might improve matching and processing quality if available or applicable:

reviews_cnt Number of reviews count as appear publicly on the site at the time of the snapshot. For the providers of daily feeds this is actually a mandatory field.

city_rank If the source provides an absolute rank of the hotel in the city list will be crawled in the following format: `<city_rank>11</city_rank>` or `<city_rank>1/13</city_rank>` So if the source provides the upper bound we shall extract it if not there will be only the rank. The regular expression to validate will be then `<city_rank>[[:digit:]]+(/[[:digit:]]+)*</city_rank>`

address/street the street address (as clean as possible; it may contain other address elements if they can't be separated in their own fields)

address/zip the zip code of the hotel

address/rawAddress the entire address string, which is not split into separate address fields. could contain any combination of address/city, address/country, address/street, address/zip fields.

address/geo/lat the geolocation latitude in decimal degrees

address/geo/long the geolocation longitude in decimal degrees

amenities amenities applicable for the hotel

giataId id assigned by GIATA (<http://www.giata.de>)

contactInfo/phone hotel's telephone numbers (if there are more than one, should be included in the same field)

contactInfo/fax hotel's fax numbers (if there are more than one, should be included in the same field)

contactInfo/email hotel's contact email address

contactInfo/homepage hotel's homepage URL

Sample:

```
<?xml version="1.0"?>
<root>
  <entity>
    <id>114374</id>
    <uid>e211b190-31d0-11df-8c61-001fe218b3e3</uid>
    <giataId>9123843</giataId>
    <type>1</type>
    <source>25</source>
    <name>Hotel Crowne Plaza</name>
    <stars>5</stars>
    <mark>
      <total>5.3</total>
    </mark>
    <reviews_cnt>260</reviews_cnt>
    <createdate>1268836557489</createdate>
    <ratenow>http://src.de/hotelbewertung.php?lang=en&hid=123</ratenow>
    <address>
      <url>http://src.de/hotel-Reiseinformationen+123.html</url>
      <city>Andorra la Vella</city>
      <country>Andorra</country>
      <street>Prat de la Creu 88</street>
      <zip/>
      <geo>
```

```

    <lat>42.507</lat>
    <long>1.525</long>
  </geo>
  <rawAddress>Schoene Str. 12 85598 Munich</rawAddress>
</address>
<contactInfo>
  <phone>+376 (0)89 123131</phone>
  <fax>+376 (0)89 123000</fax>
  <email>info@sampleplaza.com</email>
  <homepage>http://www.simpleplaza.com</homepage>
</contactInfo>
<price>73</price>
<price_currency>EUR</price_currency>
<picture>http://www.hotelimage.net/images/hotel/586/586078.jpg</picture>
<description>Mitten im Herzen von Baden-Baden, jedoch ruhig gelegen im
  hoteleigenen Park, erstrahlt der Glanz des Hotels Belle Epoque
</description>
<amenities>Parkplatz, Restaurant,Zimmerservice, Konferenz- und
  Veranstaltungenr&#xE4;ume, Bar, 24-Stunden-Rezeption</amenities>
</entity>
</root>

```

2.1.2 Restaurant entity elements

In addition to entity fields the restaurant entities must and could contain the following fields:

Mandatory fields:

address/city the city name

address/country the country name

address/url absolute URL to the hotel's webpage (source deplane)

Optional fields that might improve matching and processing quality if available or applicable:

description any information related to restaurant's type and offerings

specific any information related to restaurant's specific (could be more than one specific)

price any information related to the price range/class of the restaurant (will be later post-processed)

address/street the street address (as clean as possible; it may contain other address elements if they can't be separated in their own fields)

address/zip the zip code of the restaurant

address/rawAddress the entire address string, which is not split into separate address fields. could contain any combination of address/city, address/country, address/street, address/zip fields

address/geo/lat the geolocation latitude in decimal degrees

address/geo/long the geolocation longitude in decimal degrees

contactInfo/phone restaurant's telephone numbers (if there are more than one, should be included in the same field)

contactInfo/fax restaurant's fax numbers (if there are more than one, should be included in the same field)

contactInfo/email restaurant's contact email address

contactInfo/homepage restaurant's homepage absolute URL

openingHours opening hours

picture absolute URL to a picture which should represent the thumbnail

Sample:

```
<?xml version="1.0"?>
<root>
  <entity>
    <id>6038</id>
    <uid>e211b190-31d0-11df-8c61-001fe218b3e3</uid>
    <type>2</type>
    <source>356</source>
    <name>Ristorante L'Angelo</name>
    <mark>
      <total>4.25</total>
    </mark>
    <reviews_cnt>43</reviews_cnt>
    <createdate>1268836557489</createdate>
    <ratenow>http://src.de/rest/bewertung/6038</ratenow>
    <address>
      <url>http://src.de/rest/6038</url>
      <city>M&#xFC;nchen</city>
      <country>Deutschland</country>
```

```

    <street>Musterstra&#xDF;e 98</street>
    <zip>85312</zip>
    <rawAddress>Musterstra&#xDF;e 98 M&#xFC;nchen</rawAddress>
    <geo>
      <lat>54.794</lat>
      <long>5.44</long>
    </geo>
  </address>
  <price>17219</price>
  <contactInfo>
    <phone>+49 (0)89 961131</phone>
    <fax>+49 (0)89 960000</fax>
    <email>info@ristorante.de</email>
    <homepage>http://www.ristorante.de</homepage>
  </contactInfo>
  <openingHours>Mo - Sa: 11:00 &#xA0; 1:00; So: 9:00 - 1:00</openingHours>
  <picture>http://src.de/cdn/rest/6038.jpg</picture>
  <description>mit Mittagstisch, Vegetarische Gerichte, fr&#xFC;hst&#xFC;ck, pasta,
nudeln, cocktails, longdrinks, wein, weinkarte</description>
  <specific>Italienisch</specific>
</entity>
</root>

```

2.2 Review elements

Mandatory fields:

uid uniquely identifies a review (valid UUID see [RFC 4122](#))

source source id which is an internal TrustYou source id (please ask for such an id if it's not known)

type would denote the review's entity type. i.e. type=100 → hotel review, type=200 → restaurant review, a.s.o. (see „3.IDs Tables“ section)

Optional fields that might improve matching and processing quality if available or applicable:

id id of the review from the source side

puid parent's id that uniquely identifies an entity (valid UUID see [RFC 4122](#))

date review's date in epoch (milliseconds since January 1, 1970, 00:00:00 GMT. A negative number is the number of milliseconds before January 1, 1970, 00:00:00 GMT)

text the text of the review

mark/total the note/mark/score that accompanies the review, normalized to scale 0 → 100

mark/score a XML tag containing the pair `<note> </note> <detail> </detail>` (e.g. `<mark> <total> 5.8 </total> <score> <note> 5.3 </note> <detail>Service</detail> </score> <score>...`)

response_to preferable the parent review uid (valid UUID see [RFC 4122](#)). A response from the hoteliers will be crawled as a normal review and this field will identify that represents a response to a parent review.

address/url the absolute URL to the review

title review title

author text node which identifies the author (e.g. `<author> Hanna L. </author>`)

author_location separate (not embedded) text node that contains the location of author if available (e.g. `<author>Hanna L.</author> <author_location> France </author_location>`)

author_age separate (not embedded) text node that contains an integer or range representing the age of the author (e.g. `<author_age>25-30</author_age>`)

author_type a coma separated list of author classifications on the source (Family, Young Couple, Business ...)

createdate when this review element was generated as time-stamp in epoch (milliseconds since January 1, 1970, 00:00:00 GMT. A negative number is the number of milliseconds before January 1, 1970, 00:00:00 GMT)

2.2.1 Hotel review elements

In addition to review fields the hotel reviews must and could contain the following fields:

Mandatory fields:

name hotel's name to which the review belongs to (must be same value as the name element from a hotel entity)

address/city the city where the hotel is (must have the same value as the address/city element from a hotel entity)

address/country the hotel country (must have the same value as the address/country element from a hotel entity)

address/rawAddress the entire address string, which is not split into separate address fields. could contain any combination of address/city, address/country, address/street, address/zip fields

Sample:

```
<?xml version="1.0"?>
<root>
  <review>
    <uid>e211b190-31d0-11df-8c61-001fe218b3e3</uid>
    <puid>bfffe190-31d0-11df-8c61-221fe218b3fe</puid>
    <id>1001680</id>
    <type>100</type>
    <source>70</source>
    <name>Hotel Crowne Plaza</name>
    <address>
      <url>http://src.de/bericht-Hotelbewertungen_Sample+Plaza.html</url>
      <city>Andorra la Vella</city>
      <country>Andorra</country>
      <rawAddress>Prat de la Creu 88 Andorra</rawAddress>
    </address>
    <date>1268836557489</date>
    <title/>
    <author>Katja</author>
    <text>Das Hotel liegt in der N&#xE4;he (ca. 10 Minuten zu Fuss) zum
Stadtzentrum und verf&#xFC;gt &#xFC;ber eine Tiefgarage</text>
    <mark>
      <total>53</total>
    </mark>
    <createdate>1268836557489</createdate>
  </review>
</root>
```

2.2.2 Restaurant review elements

In addition to review fields the restaurant reviews must and could contain the following fields:

Mandatory fields:

name restaurant's name to which the review belongs to (must be same value as the name element from a restaurant entity)

address/city the city where the restaurant is (must have the same value as the address/city element from a restaurant entity)

address/country the restaurant's country (must have the same value as the address/country element from a restaurant entity)

address/rawAddress the entire address string, which is not split into separate address fields. could contain any combination of address/city, address/country, address/street, address/zip fields

Sample:

```
<?xml version="1.0"?>
<root>
  <review>
    <uid>e211b190-31d0-11df-8c61-001fe218b3e3</uid>
    <puid>bffff190-31d0-11df-8c61-e345218be3e3</puid>
    <id>6038</id>
    <type>200</type>
    <source>356</source>
    <name>Ristorante L'Angelo</name>
    <address>
      <url>http://www.sourcURL.de/rest/6038</url>
      <city>München</city>
      <country>Deutschland</country>
      <rawAddress><span class="street-
address">Musterstraße 98</span>, <span
class="postal-code">85312</span> <span
class="locality">München</span></rawAddress>
    </address>
    <date>1268836557489</date>
    <title/>
    <author>Gaby</author>
    <text>Sehr freundliche und familiäre Atmosphäre, ausgefallene und
ausgesprochen schmackhafte Gerichte. Auch einfache Gerichte wie z.B.
Pizza werden angeboten und sind ausgesprochen gut.</text>
    <mark>
      <total>98</total>
    </mark>
    <createdate>1268836557489</createdate>
  </review>
</root>
```

3 ID Tables

Table 1: Document type IDs

No.	TrustYou type ID	TrustYou type description
1.	1	Hotel entity element
2.	2	Restaurant entity element
3.	100	Hotel review element
5.	200	Restaurant review element

4 Matching

This section provides an outline of the algorithm used by TrustYou to match reviews with entities and the implications for the data provider. This is a technical description of the quality implications and not a contractual view.

1. `<puid>` will contain a valid UUID the same value as its parent entity `<uid>`. The tag `<puid>` is no longer mandatory (this information is expected to be provided for most of the sources where this is naturally available for the provider)
2. In case `<puid>` is not provided the matching will be done based on the pair of elements `<source_id>` (mandatory for all data) and `<id>` – source’s internal entity id. This combination should be unique within the entities provided by a source. The internal entity id should be available for the most entities and reviews.
3. The remaining orphan reviews are acceptable only for a small percentage of data, where `puid` could not be generated (relation not known) and no `<id>` could be provided. For this small set a lower quality matching based on all attributes will be attempted or the orphans will not be included in the data at all.

5 Quality Check

This section gives a few hints about a few minimal quality checks to be done before submitting a data snapshot because they will be enforced during the data import.

XML will be verified against the schema or similar means (e.g. no double address tag is allowed)

If two or more <entity> XMLs will contain the same uid than the data snapshot is invalid

If there is a <puid> with no counterpart entity/uid, that particular review is considered orphan and the bug will be reported.

The combination <source_id> <id> shall be unique within the whole entity set. Duplicate <source_id> <id> within entity set will be reported as bug.

After counting all orphans (reviews that do not have puid nor id) if the percentage from the total is too high the data snapshot is invalid (limit to be defined based on experience we will gain with the specific data provider/partner)

5.1 Duplicates Filtering

Dimension definition A dimension represents one attribute or XML tag of the entity or review respectively (e.g. name, source, stars, city, country ... are all dimensions of the entity)

The case where the data provider assigns by mistake the same value on one dimension (e.g. name or description) will be regarded as false duplicate and be eliminated before delivery.

We strongly suggest that a further check for the real duplicates is included on data provider side as well.

Real duplicate definition For entities: Mark as duplicate all that have identical md5sum(name, city, country, stars, mark_total, street.filtered_alpha(), zip, description.filtered_alpha(), amenities.filtered_alpha()) For reviews: Mark as duplicate all that have identical md5sum(city, country, name, date, title.filtered_alpha(), author, text.filtered_alpha(), mark_total) Iterate the sorted marked duplicates (should be a tiny percentage of data), compare and remove the real duplicates.

filtered_alpha() means stripping all characters outside [[:alphanum:]]+ only for the purpose of duplicate removal not from the actual data.

6 Document changes

Table 2: Changes list

Version	Author	Changes	Date
0.1	Costin Cozan	First draft	13.07.08
0.2	Costin Cozan, Alin Sanislav	Reviewed and finalized first version	10.08.08
0.3	Costin Cozan, Alin Sanislav	Updated all sections	14.09.08
0.4	Alin Sanislav	Added expert reviews section	14.11.08
0.5	Alin Sanislav	Removed amenities field from entity	19.11.08
0.6	Alin Sanislav	Added new fields for reviews: <ul style="list-style-type: none">• createdate	24.11.08
0.7	Alin Sanislav	Added new fields for entities: <ul style="list-style-type: none">• createdate• uid• mark/total• price_currency	
0.7	Alin Sanislav	Added tables with TrustYou IDs	27.11.08
0.8	Alin Sanislav	Fixed the URL tag position in review samples	01.12.08
0.9	Alin Sanislav	Corrected the review type samples to contain the IDs specified in „3.IDs Tables“ section	02.12.08

Version	Author	Changes	Date
0.10	Alin Sanislav	Moved review's text, date and mark fields from mandatory to nice to have fields	05.12.08
1.0	Alin Sanislav	<ul style="list-style-type: none"> • Added geo/lat, geo/lang,contactInfo and giataId elements for hotel entities • Added restaurant elements 	20.02.09
1.1	Alin Sanislav	<ul style="list-style-type: none"> • Added address/rawAddress element for hotel/restaurant entities and reviews 	23.02.09
1.2	Alin Sanislav	Added paid element for reviews	13.11.09
1.3	Alin Sanislav	Added reviews_cnt entities element for	22.01.10
2.0	Adrian Achihăei	Clean-up expert chapters. Added matching description	28.07.2010
2.1	Adrian Achihăei	Extra Review fields	06.09.2010
2.2	Adrian Achihăei	Fixes and further descriptions of fields as suggested by Justin Gilbreath revision	07.10.2010