

Data Standardization & Normalization



Optimize your data

Unlock the power of data and you'll drive better business performance.

Data. It's one of your most valuable commercial assets. But poor data quality causes too many business problems. Partner with us and discover a suite of services to support better data management, maintenance and deployment - covering higher-quality inputs, data cleansing and standardization.

Have access to 7 main countries in Latin America with a complete range of services: Mexico, Argentina, Brazil, Chile, Colombia, Peru and Uruguay.



Working with MSLA Data Management Services you have access to the following features:



Data profiling

- The process generates simple to advanced profiling information including basic data statistics (mean, median, frequency, variation, etc.) and details (structure, content, classifications, etc.) to identify data errors and weak points in your data collection processes.
- Creates a data quality assessment table showing possible issues for each individual column in a customer database, including the number of profanities found, inconsistently ordered names, invalid ZIP codes, etc.



Data Standardization & Formatting

- Customize and create rules (triggering) to standardize data.
- Transform and standardize data such as names, addresses, dates, product numbers, and phone numbers.
- Extract structured information from freeform text.
- You will have the address format for each member country, viz.:
 - ✓ the type and position of the postcode;
 - ✓ the description of the coding system;
 - ✓ the format of domestic addresses
- Once the addresses are standardized and deduplicated, the address verification solution compares your addresses against each postal service database to validate the deliverability of the addresses. If information is incorrect or missing, address verification updates the records.



Duplicate identification

After addresses are standardized, it is quite possible that there are duplicate records, especially if this is a merger of several lists or different departments or offices contribute to the list. Using deduplication software and protocols, it is possible to identify possible duplicate records.



Data Profiling

Data profiling is the process of examining, analyzing, and creating useful summaries of data. The process yields a high-level overview which aids in the discovery of data quality issues, risks, and overall trends. Data profiling produces critical insights into data that companies can then leverage to their advantage.

MSLA Data Profiler analyzes data before it's merged into your warehouse, then helps ensure consistent data quality once it's there.

Our Data Profiling main features are:

String Analyzer

The string analyzer provides general purpose profiling metrics for string column types.

List MSLA21022019-Puerto Rico.xlsx | Analysis results | DataCleaner

Analysis results | List MSLA21022019-Puerto Rico.xlsx

String analyzer

	Nombre	Dirección	Provincia	Código	Teléfono
Row count	192	192	192	192	192
Null count	0	0	0	0	0
Blank count	0	0	0	0	0
Entirely uppercase count	1	0	0	0	0
Entirely lowercase count	0	0	0	0	0
Total char count	5292	4103	1679	576	1916
Max chars	56	46	15	3	10
Min chars	7	5	5	3	9
Avg chars	27,563	21,37	8,745	3	9,979
Max white spaces	6	9	3	0	2
Min white spaces	1	0	0	0	1
Avg white spaces	3,089	3,349	1	0	1,979
Uppercase chars	714	633	244	0	0
Uppercase chars (excl. first letters)	522	414	52	0	0
Lowercase chars	3977	2258	1243	0	0
Digit chars	6	505	0	576	1344
Diocritic chars	73	25	0	0	0
Non-letter chars	601	1212	192	576	1916
Word count	784	834	244	192	192
Max words	7	10	2	1	1
Min words	2	1	1	1	1

Save results

Escribe aquí para buscar

19:51 08/03/2019

Date/Time Analyzer

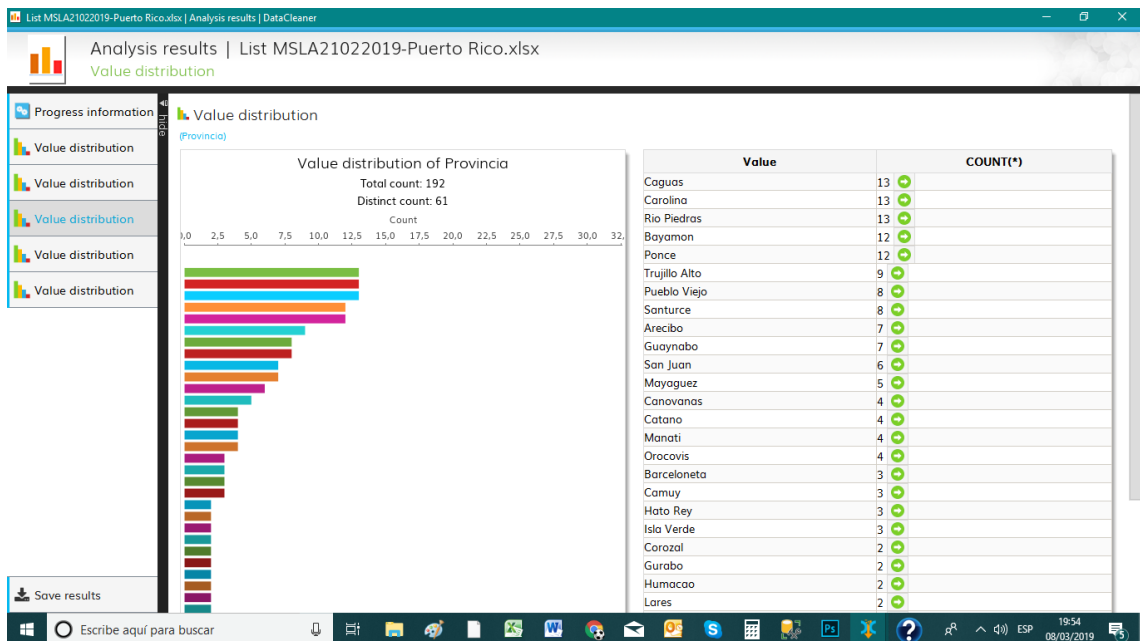
The Date/time analyzer provides general purpose profiling metrics for temporal column types such as DATE, TIME and TIMESTAMP columns.



DATE OF BIRTH	DATE OF BIRTH (as date)
19761019	Tue Oct 19 00:00:00 COT 1976
19470202	Sun Feb 02 00:00:00 COT 1947
19440828	Mon Aug 28 00:00:00 COT 1944
19720521	Sun May 21 00:00:00 COT 1972
19720913	Wed Sep 13 00:00:00 COT 1972
19570627	Thu Jun 27 00:00:00 COT 1957
19661203	Sat Dec 03 00:00:00 COT 1966
19660308	Tue Mar 08 00:00:00 COT 1966
19750507	Wed May 07 00:00:00 COT 1975
19580914	Sun Sep 14 00:00:00 COT 1958
19650514	Fri May 14 00:00:00 COT 1965
19700831	Mon Aug 31 00:00:00 COT 1970
19650221	Sun Feb 21 00:00:00 COT 1965
19600924	Sat Sep 24 00:00:00 COT 1960
19580109	Thu Jan 09 00:00:00 COT 1958
19680116	Tue Jan 16 00:00:00 COT 1968
19760602	Wed Jun 02 00:00:00 COT 1976
19741005	Sat Oct 05 00:00:00 COT 1974
19760329	Mon Mar 29 00:00:00 COT 1976
19650619	Sat Jun 19 00:00:00 COT 1965
19550309	Wed Mar 09 00:00:00 COT 1955
19571013	Sun Oct 13 00:00:00 COT 1957
19560823	Thu Aug 23 00:00:00 COT 1956
19690720	Sun Jul 20 00:00:00 COT 1969
19571206	Fri Dec 06 00:00:00 COT 1957
19630412	Fri Apr 12 00:00:00 COT 1963
19701007	Wed Oct 07 00:00:00 COT 1970
19620302	Fri Mar 02 00:00:00 COT 1962
19670317	Fri Mar 17 00:00:00 COT 1967
19570511	Sat May 11 00:00:00 COT 1957
19431212	Sun Dec 12 00:00:00 COT 1943

Value Distribution

The value distribution (often also referred to as 'Frequency analysis') allows you to identify all the values of a particular column.



Data Standardization & Formatting

This service allows you to process a set of addresses that have a basic format requested by the customer, allowing to structure, standardize, validate (depending on the region) and georeference addresses with master files of streets and addresses.

MSLA works with all street layers information from each Postal Office and private sources to guarantee the most accurate results on your service request.

It should be mentioned that the sale of this product may be performed as a standard product (only until the validation process depending on the region) or as expanded product (including geocoding), depending on customer needs.








CUSTOMER BENEFITS

- Certification of the existence of an address.
- Guaranteed database update address.
- Contactability increases real contact with customers.
- Decreasing return correspondence up to 40%.
- Data cleansing.
- Addresses under standard format.
- There is no duplication of addresses.
- Assigning the exact location of information on a digital map.

TARGET MARKET

All companies or institutions whose input information flows have an address, which is used in your business processes for customer contactability and / or availability of services offered.

SERVICES PROVIDED BY COUNTRY

Country	Data Cleansing	Data Standardization	Deduplication	Geocoding	Level
Argentina 	Yes	Yes	Yes	Yes	STREET
Brazil 	Yes	Yes	Yes	No	N/A
Chile 	Yes	Yes	Yes	No	N/A
Colombia 	Yes	Yes	Yes	No	N/A
Mexico 	Yes	Yes	Yes	Yes	STREET
Peru 	Yes	Yes	Yes	Yes	STREET
Uruguay 	Yes	Yes	Yes	Yes	STREET

LEGEND

Data Cleansing/Standardization: Yes = service available address standardization; NO = service standards address not available

Geocoding: Yes = geocoding service available; NO = not available geocoding service

Level: Street = street detail; LOCATION = location detail

Deduplication: Yes = Deduplication service available; NO = unavailable service Deduplication

ADDRESS FORMAT BY COUNTRY



For purposes of addressing mail from within the USA, the name of the country is MEXICO. In Spanish, the 'e' has an acute accent: México. In Spain and parts of Latin America, some people prefer the more phonetic spelling, "Mejico" (just as in the USA, some prefer to write "Tejas").

Mexico has states like Jalisco, Sonora, etc, which are included in the address. The state for Mexico City is DF (Federal District), similar to Washington DC in the USA or in Australia Canberra ACT (DF is divided into delegations Mexico City Including, St. Jerome, etc.)

BUT WILL CHANGE THE NOMECLATURE from "Mexico City" to "Cuidad de Mexico DF" due to a constitutional amendment adopted in 2016.

Postal codes are 5 digits. Examples:

The states of Mexico and their official abbreviations are:

AGS	Aguascalientes	MOR	Morelos
BCN	Baja California Norte	NAY	Nayarit
BCS	Baja California Sur	NL	New Lion
CAM	Campeche	OAX	Oaxaca
CHIS	Chiapas	PUE	Puebla
CHIH	Chihuahua	QRO	Querétaro
COAH	Coahuila	QROO	Quintana Roo
CABBAGE	Colima	SLP	San Luis Potosi
DF	Federal District	SIN	Sinaloa
DGO	Durango	SON	Sonora
GTO	Guanajuato	TAB	Tabasco
GRO	Guerrero	TAMPS	Tamaulipas
HGO	Hidalgo	TLAX	Tlaxcala
JAL	Jalisco	WATCH	Veracruz
MEX	Mexico (State)	YUC	Yucatán
MICH	Michoacán	ZAC	Zacatecas

It is Important to put Colonia for District (when known) in Mexican addresses, for example:

Latin American Faculty of Social Sciences
Km.13 Carretera al Ajusco, Colonia Héroes de Padierna
Section 20-021, Delegacion Alvaro Obregon
01000 Mexico, DF
MEXICO

The 5-digit zip code goes on the left, then a city or the town, a comma, and the state abbreviation.

It is common to see the postal code written on the right, but this is an old form (say, pre-2000):

(Person's Name)

Avenida Castillo Chapultepec No.47
Colonia Cd.Chapultepec
Cuernavaca, MOR 62380
MEXICO

The composition of the standardized output is:

Field	Description
Tipo_via	Street type
nom_via	Street name
tipo_asen	Type of Asentamiento (Colonia/Fraccionamiento)
nom_asen	Name of Asentamiento
nom_loc	Locality name
nom_mun	Municipio name
nom_ent	State name
d_codigo	Postal code
Latitud	Latitude
Longitud	Length

Number of addresses in the database: 1,013,557



All Brazilian states and a 5 + 3-digit postal code (CEP Endereçamento Postal Code) goes on the left. The state goes on the right, separated by a dash. There should be no other punctuation. Example:

20071-003 Rio de Janeiro-RJ

The state for Brasilia is DF (Federal District), like Washington DC, eg:

70084-970 Brasilia-DF

Always use the exact spacing and punctuation shown above - no periods, commas, etc. Never include CEP in the address; it just means postal code. For example, if you have an address like:

Rio de Janeiro, RJ CEP 20071-003

It should be written like:

20071-003 Rio de Janeiro-RJ

The states of Brazil and Their official abbreviations are:

AC	Acre	MA	Maranhão	RN	Rio Grande do Norte
AL	Alagoas	MT	Mato Grosso	RS	Rio Grande do Sul
AP	Amapá	MS	Mato Grosso do Sul	RJ	Rio de Janeiro
A.M	Amazon	MG	Minas Gerais	RO	Rondônia
BA	Bahía	PR	Paraná	RR	Roraima
EC	Ceará	PB	Paraíba	SC	Santa Catarina
DF	Distrito Federal	PA	Pará	SE	Sergipe
ES	Espirito Santo	PE	Pernambuco	SP	São Paulo
GO	Goiás	PI	Piauí	TO	Tocantins

The composition of the standardized output is:

Field	Description
Calle	Street name
Números	Street number
Rango km	Distance in km
Complementary address	Additional information beside the street for better location
Municipio	Municipio name
City	City name
State	State name
CEP	Postal code

Number of addresses in the database: 606,007

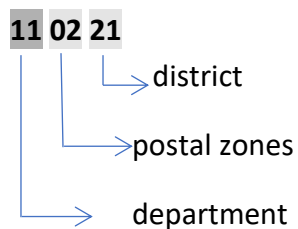


Postcode

Postcode type and position

6 digits postcode after the locality

Coding method



Ministerio de Tecnologías

Edificio Murillo Toro

Cra 8ª entre calles 12 y 13

BOGOTÁ, D.C. 111711

COLOMBIA

Example

Urban

Ministerio de Technologies
Edificio Murillo Toro
Cra. 8a entre calles 12 y 13
BOGOTA 111711
COLOMBIA

addressee
building
street + premises
locality + postcode
country

Urban with sub-locality

Calle 32a Sur Transversal 68b
#5
Alqueria La Fragua
LOCALIDAD KENNEDY 110841
BOGOTA D.C.

street + premises
sub-locality
locality + postcode
province

Rural

Adriana Gómez
Arroyón
PLANETA RICA 233057
CORDOBA

sub-locality
locality + postcode
province

The composition of the standardized output is:

Field	Description
Calle	Street name
Código_Centro_Poblado	City code number
Centro_Poblado	City name
Código_Mun	Municipio Id number
Municipio	Municipio name
Código_Dpto	Departamento Id number
Departamento	Departamento name
Código_Postal	Postal code

Number of addresses in the database: 8,658,985



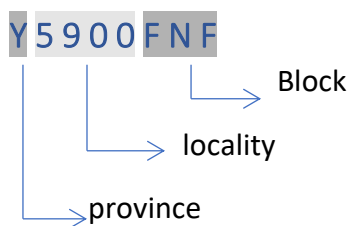
ARGENTINA

Post code

Post code type and position

8 alphanumeric characters (1 letter, 4 digits and 3 letters) to the left of the locality name.

Coding method



SIGNUM SRL
Juan López
San Martín 230
Piso 4 Dpto. A
Y5900FNF Villa María
Buenos Aires

Address format: Argentine Post recommendations:

- each line should contain a maximum of 40 characters;
- the following symbols should be avoided: full stop (.), dash (-), accents (´, `), (o) and parentheses ((),);
- in general, it is not necessary to indicate the type of thoroughfare if it is a street (CALLE); other types of thoroughfare must be specified;
- the thoroughfare type (CALLE) is obligatory where the thoroughfare name is a number or letter (e.g. CALLE 7, CALLE A);

- where the thoroughfare name ends in a number, the abbreviation “Nº” should be used to separate the number in the name of the street from the street number (e.g. 17 DE OCTUBRE DE 1945 Nº 1340);
- fonts with letters between 3 and 7 mm high and no more than 7 mm wide are recommended; the preferred fonts are Courier 12 or 15 and Helvetica 12 or 15.

Examples Home delivery:

SIGNUM SRL
 JUAN LOPEZ
 SAN MARTIN 230 street name and number
 PISO 4 DPTO. A floor + department (or office, district, etc.)
 Y5900FNF VILLA MARIA postcode + locality
 ARGENTINA

P.O. Box delivery:

SR. JUAN LOPEZ
 CASILLA DE CORREOS 432 P.O. Box
 CORREO CENTRAL post office name
 C1000WAE CAPITAL FEDERAL postcode + locality
 ARGENTINA

Delivery to a farm, rural school, etc.:

PROF. JUAN LOPEZ
 ESCUELA RURAL 45 name of farm, rural school, etc.
 X5187YAB SAN CLEMENTE postcode + locality
 ARGENTINA

The composition of the standardized output is:

Field	Description
id_tipo_camino	Street type id number
Camino	Street type
Nombre_calle	Street name
Barrio	Neighborhood
id_localidad	Locality id number
Localidad	Locality name
Código Postal	Postal code
Provincia	Province name

Number of addresses in the database: 325,353



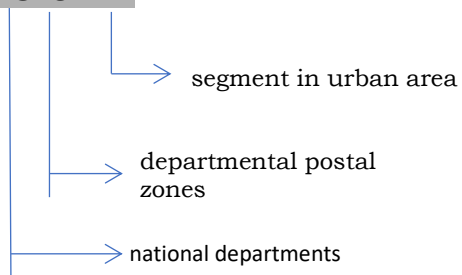
Post code

Post code type and position

5 digits after the street and sub-locality

Coding method

15 0 74



Sr. José Galvez Lora
 Av. Larco 1301, Miraflores
 → 15074
 LIMA
 PERU

Address format:

- The addresses should be written in “block letters” by preference.
- It’s necessary to indicate the type of street for example:

Type of road in Spanish	Abbreviation of type of road in Spanish	Type of road in English
Calle	Calle	Street
Jirón	Jr.	Urban road composed by several streets or street sections between corners.
Avenida	Av.	Avenue
Pasaje	Psje.	Passage

Or any other type of street should be specified.

- When the name of a street ends with a number (Calle 10), the number sign “Nº” should be used to separate the numeric end of the street name from the number of the house in this street: (for example: Calle 10 Nº. 985)

Examples

Sr. José Galvez Lora	addressee
Av. Larco 1301, Miraflores	street, premise, sub-locality
15074	postcode
LIMA	province
PERU	country

List of provinces

01 Amazonas	14 Lambayeque
02 Ancash	15 Lima
03 Apurimac	16 Loreto
04 Arequipa	17 Madre de Dios
05 Ayacucho	18 Moquegua
06 Cajamarca	19 Pasco
07 Callao	20 Piura
08 Cusco	21 Puno
09 Huancavelica	22 San Martin
10 Huánuco	23 Tacna
11 Ica	24 Tumbes
12 Junín	25 Ucayali
13 La Libertad	

The composition of the standardized output is:

Field	Description
CATEG_VIA	Street id number
CATEG_TXT	Street type
NOMBRE_VIA	Street name
NOMBRE_ALT	Complementary address
CUADRA	Block number
URBANIZACION	Neighborhood
CODDIST	District code
DISTRITO	District name
CODPROV	Province code
PROVINCIA	Province name
CODDPTO	State code
DEPARTAMENTO	State name
UBIGEO	Postal code
LONGITUD	Length
LATITUD	Latitude

Number of addresses in the database: 1,101,777



Post code

Post code type and position

7 digits to the left of the name of the "Comuna"

Coding method

872 0019

Sequential number
that identifies a block face

3 digits that identify a postal
Distribution area ("Sector"), usually a "Comuna"

Señorita
María Teresa Torres
El Juncal 050, Edificio B, Piso 2
8720019 Quilicura
Región Metropolitana
Chile

Address format: Recommendations concerning fonts:

- height of characters: between 2 and 7 mm;
- space between address lines: at least 1 mm;
- pitch: between 6 and 12 characters per inch;
- space between characters: at least 0.4 mm (for whole height);
- matrix printer characters are permitted provided they consist of a large number of dots arranged very close together;
- proportionally spaced print is machine-processable.

The postal address must contain a minimum of information for correct sorting and delivery (see example 1).

Other information related to the building, floor, village, apartment, should be placed to the right side of the street name and number. This part of the address is called “rest of address” (see example 2).

Addresses of items addressed to a Post Office Box should have the postal code of the Post office or agency of Correos de Chile (see example 3).

If the comuna / locality doesn't have a postal code for a block face, the comuna postal code should be used (for the whole commune / locality) (see example 4).

Most rural addresses do not have street numbers; these addresses are identified by the company name, the street name but no number (see examples 5 and 6).

Examples

Example 3

Señora
Fernanda Genoud
Casilla 13-D, Sucursal Plaza de Armas
8329001 SANTIAGO
REGION METROPOLITANA
CHILE

name of addressee
P.O. Box + name of post office
postcode (of the post office) + name of comuna regio
country

Example 4

Señor
Pedro López
Empresa Automotriz
Avenida Arturo Prat 567
3930000 BULNES
VIII REGION DEL BIOBIO
CHILE

name of addressee
company name
street name + number
postcode (of the commune) + name of comuna
region
country

Example 5

Señor
Manuel Vera
Empresa Metalúrgica Panamericana
Norte S/N Km. 15 8700000
QUILICURA
REGION METROPOLITANA
CHILE

name of addressee
company name
street name without number + rest of address
postcode (of commune) + name of comuna
region
country

Example 6 Señor

Enrique González
Camino Público S/N, Tunca Arriba
2970000 SAN VICENTE DE T.T.
VI REGION DE O'HIGGINS CHILE

name of addressee
street name without number + rest of address
postcode (of commune) + name of comuna
region
country

The composition of the standardized output is:

Field	Description
Código_vía	Street id number
Nombre_vía	Street name
Código_Comuna	Comuna id number
Nombre_Comuna	Comuna name
Código_Provincia	Province id number
Nombre_Provincia	Province name
Código_Región	Region id number
Nombre_Región	Region name
Código_Postal	Postal Code

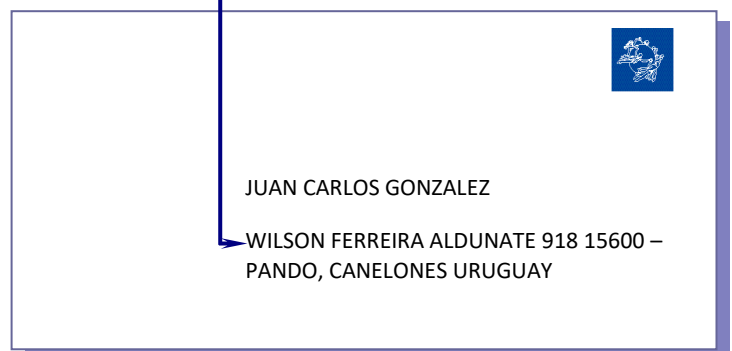
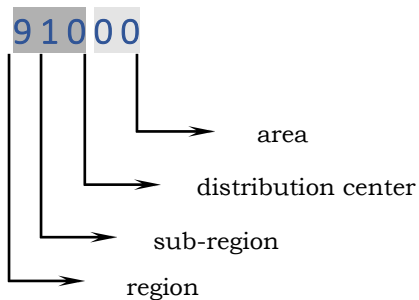
Number of addresses in the database: 755,119



Postcode Postcode type and position

5 digits to the left of the locality

Coding method



**Position of
the address
on the
envelope**

Bottom right

**Alignment
of address
lines** Left-aligned

**Address
format**

Uruguayan Post recommendations:

Ordering of lines in address from specific to general (from top to bottom).

An address has a maximum of 8 lines and a minimum number of 3 lines. No blank lines to be included.

The order of the lines is as follows:

1. organisation
2. function and unit
3. addressee name
4. street + premises + building + floor/door
5. supplementary data
6. sub-locality
7. postcode + locality + province
8. country (only for international mail)

An address line has a maximum of 40 characters including spaces. There should be one space between words.

Capital letters are recommended for the last 3 lines of the address.

Uruguayan Post prefers fixed-width fonts like Verdana, Lucinda Console, Courier New (10 to 12 points).

The uses of punctuation marks are acceptable in the line containing the building & thoroughfare address, and in the line for the district information.

The locality is separated from the postcode by a dash (-), and from the country by a comma (,). The locality is required unless it is identical to the name of the department.

Examples

Hospital de Clínicas
Subdirector de Desarrollo
Dr. Sergio Carrasco Avenida Italia s/n Piso 1B Esquina Américo Ricaldoni
PARQUE BATTLE
11600 – MONTEVIDEO
URUGUAY

Dr. Augusto Rodríguez
Chaná 1215, apto. 152
70200 – ROSARIO, COLONIA
URUGUAY

List of provinces (departamentos)

NAME	Code ISO (3166-2)
ARTIGAS	UYAR
CANELONES	UYCA
CERRO LARGO	UYCL
COLONIA	UYCO
DURAZNO	UYDU
FLORES	UYFS
FLORIDA	UYFD
LAVALLEJA	UYLA
MALDONADO	UYMA
MONTEVIDEO	UYM
PAYSANDU RIO	UYPA
NEGRO RIVERA	UYRN
ROCHA	UYRO
SALTO SAN JOSÉ	UYSA
SORIANO	UYSJ
TACUAREMBÓ	UYTA
TREINTA Y TRES	UYTT

The composition of the standardized output is:

Field	Description
codigo_via	Street id number
nombre_via	Street name
num_puerta	Street number
letra_puerta	Street letter
Km	Km number
Manzana	Block
Solar	building site
codigo_localidad	Locality id number
Localidad	Locality name
departamento	Department name
codigo_postal	Postal code
Latitud	Latitude
Longitud	Length

Number of addresses in the database: 39,836

PROCESS:

This service is provided in three stages:

a) Automatic process

Intelligent Standardization Tool (IST), which allow to differentiate the correct data, standardize addresses and assigning the postal code. This thread will perform the following tasks: debugging, validation, standardization and allocation of the postal code. The software performs the following functions:

Identifies the syntax elements of a sentence by using a grammar that defines valid language.

Read the record and seeks similar on the Master Streets Base, solving spelling problems and typing errors, and complete records for uniquely written. Identifies the following errors: inconsistent or incomplete addresses.

Fix, in a fast and effective way, a large percentage of database problems due to the effectiveness of this software.

b) Semi-assisted process

It is done to the records that cannot be resolved automatically. It is done with the assistance of operators who perform the search task to arrive at a solution.

With support of IST, records are not standardized in the previous step is divided into sub lots. Operators analyze the causes that prevented automatic normalization so as to find common parameters error, resolve and make a new automatic process. It includes the following tasks: analysis, debugging, validation, standardization and allocation of the postal code.

c) Manual process

All records that have not been resolved by the above processes are analyzed by Manual process. This thread will perform the following tasks: collecting, analyzing, debugging, validation, standardization and allocation of the postal code.

This process is carried out with the assistance of specialized operators with expert knowledge of the problems of local addresses, using maps and consultation of the Postal Service from each country for resolving cases.

ADDRESS CORRECTION OUTPUT

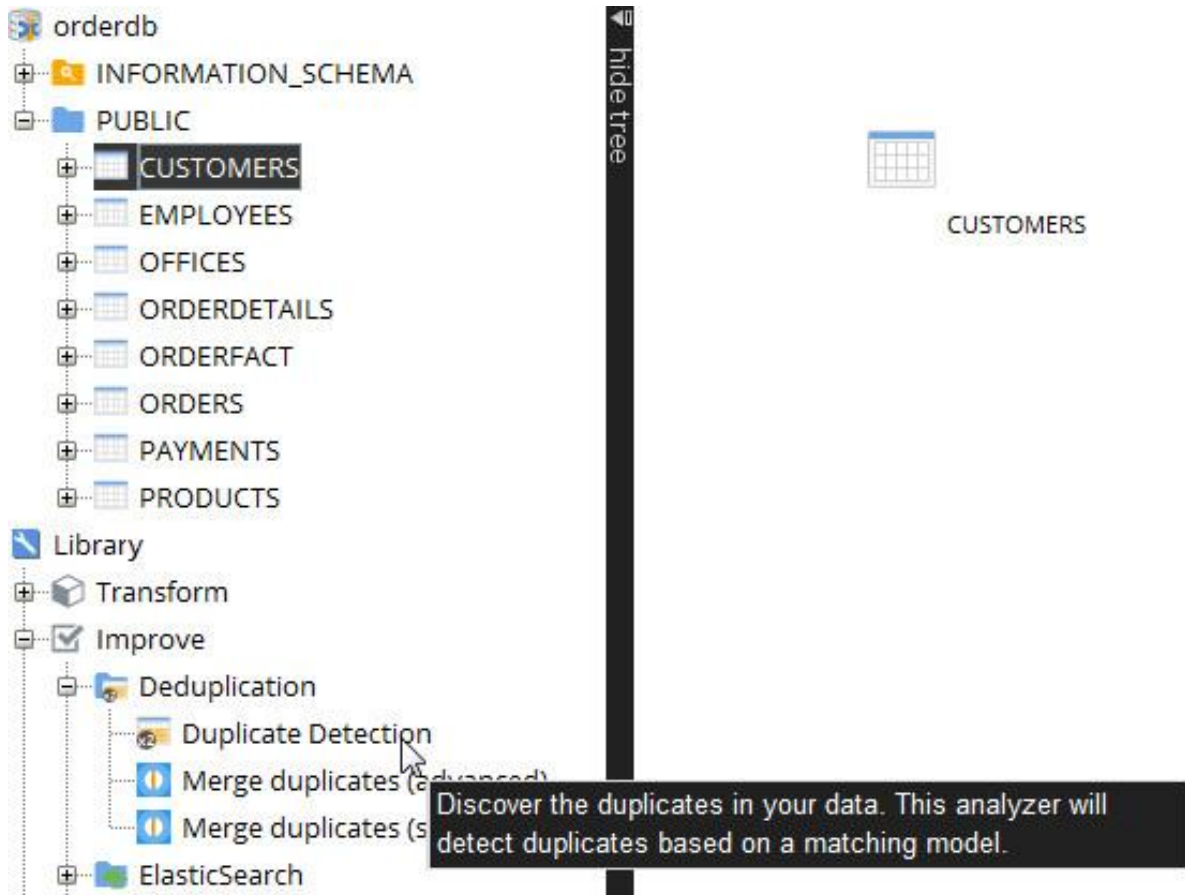
The normalization processing delivers the following output:

Code	Output label	Description
NO	Normalized	MSLA returns all the normalized data. It also informs the possible contingencies that have occurred to obtain said resolution
NF	No found	Address not found
AM	Ambiguous	MSLA returns more than one direction with different solutions
ND	No Data	No data is provided
NA	No Available	Addresses that cannot be interpreted by the address grammar.
ERR	Error	Error in the entered data. The reason for it is also, ERR.

Duplicate Identification

After addresses are standardized, it is quite possible that there are duplicate records, especially if this is a merger of several lists or different departments or offices contribute to the list. Using deduplication software and protocols, it is possible to identify possible duplicate records.

The 'Duplicate detection' function allows you to do fuzzy matching of duplicate records - records that represent the same person, organization, product or other entity.



For more information:

t: (511)705-5506

e: lrrios@mslainternational.com

w: <https://mslainternational.com>

Registered office:

Av. Circunvalación del Golf Los Inkas 208

Santiago de Surco, Lima, Peru.

©MSLA International. All rights reserved