GRC Database Information
Nieuwe Prinsengracht 80-hs
1018 VV AMSTERDAM
The Netherlands
020-528 61 20 (telephone)
020-422 93 11 (‘fax)
graham@grcdi.nl
http://www.grcdi.nl

Settlements

NB: Tables are provided as is. The publisher is not responsible or liable for errors or damage resulting from the use of these tables.

Highlighted text indicates additions/alterations since the last document version.

A great deal of information about this file is also to be found at http://www.grcdi.nl/settlements.htm . Please refer also to that webpage if you have not already seen it.

The attributes of the table are as follows:

- Records contain the settlement name and postal code as they appeared in the incoming record.
- Many of the records contain settlement names which have been typed incorrectly, are in the wrong language and so on. This table contains for a proportion of the settlement names correct upper and lower case versions, making this table an ideal source for standardising and correcting settlement name data, allowing improvements in data quality, in de-duplication rates and in postal validation rates.
- This data comes from real World sources, and can contain real World data errors. Though we mark data which we know is not correct (e.g. settlement fields containing data which is not the settlement, clearly mis-matched settlements/postal codes and so on), there can be no guarantee that a combination is correct on the ground.
- Coverage is not 100%. Data will come mainly from areas with the greatest population and greatest economic activity (i.e. the most businesses). Some settlements may never appear in these files.
- Latitude/longitude information (for a limited number of countries, and sometimes with limited coverage) are for postal code centroids rather than place name centroids. As this is very difficult information to obtain, we have often used geocoding from several populated places within a postal code area and found a point that lies equidistant between them. This will

not always identify the centre of a postal code centroid, but is more accurate than accepting the geocoding of a single settlement within a postal code area as being the same as for the postal code area itself. Postal codes referring to non-geographical addresses (large users, post office boxes etc.) will not be geocoded. No guarantees can be given as to the accuracy of the geocoding provided. Longitudes west of the Greenwich meridian (0°) and latitudes to the south of the equator are shown as negative numbers. Longitudes to the east of the Greenwich meridian and latitudes to the north of the equator are shown as positive numbers. For information on which countries are geocoded, and for coverage, please refer to http://www.grcdi.nl/settlements.htm and http://www.grcdi.nl/cities_count.pdf

The data is provided in Windows code page 1252, complying to ISO-8859-1, MS-DOS code page 850. Data is in Latin script only. A Microsoft Access file contains also the data for those place names containing characters which cannot be reproduced with the Windows code page 1252.

Where an incoming settlement name contains country name or region information which is not required for postal purposes, this has been removed in the correction fields.

NOTE: On the basis of real-World data records, we have created a number of new records which can be identified within the file because they have a count value of 0. For example, if we found this combination:

ASMTERDAM Amsterdam	1000 AA	count 1, corrected: AMSTERDAM

We have added (if not already existing)

asmterdam	1000 AA	count 0
Asmterdam	1000 AA	count 0
AMSTERDAM	1000 AA	count 0
Amsterdam	1000 AA	count 0

This makes the table more complete for settlement name correction purposes.

To aid manipulation and limit the size of each file, the full file is split into EIGHTEEN sections, with names ending:

BR_A – Brazil records (Postal codes 0 and non-numeric) (URNs commence with BA)
BR_B – Brazil records (Postal codes 1) (URNs commence with BB)
BR_C – Brazil records (Postal codes 2,3) (URNs commence with BC)
BR_D – Brazil records (Postal codes 4,5,6) (URNs commence with BD)

BR_E – Brazil records (Postal codes 7,8,9) (URNs commence with BE)
CA_A – Canada records (Postal codes J, K, L, M, N, P (Western Québec, Ontario). (URNs commence with CA)
CA_B – Canada records (All other postal codes). (URNs commence with CB)
ES – Spain records (URNs commence with E)
FR – France records (URNs commence with FR)
GB – United Kingdom records. (URNs commence with G)
IN – India records. (URNs commence with I)
JP – Japan records. (URNs commence with J)
MX – Mexico (URNs commence with MX)
NL – The Netherlands records. (URNs commence with N)
PT – Portugal records. (URNs commence with P)
US_A – United States records (Postal codes 0-4). (URNs commence with UA)
US_B – United States records (Postal codes 5-9). (URNs commence with UB)
XX – all other country records. (URNs commence with R)

Unicode_xx - A Microsoft Excel file containing also the data for those place names containing characters which cannot be reproduced with the Windows code page 1252.

Table structure (for all files except Unicode_xx)

NOTE: the field structure may differ in each file section to prevent the files sizes becoming bloated. This information is provided in the *field length* column.

Field name	*Field type*	*Field length*	*Contents*
URN	Character	10	A **unique code** for the combination of data in this field. This number should be used as a reference if there are any queries about the data. The file is split into multiple parts to aid manipulation. Please see above for information about these codes.
COUNTRY	Character	45	The **country name** in full.
GRCID	Character	3	A unique **country code** used by GRC Database Information
ISO2	Character	3	The **ISO 3166* 2-digit code** for this country.
ISO3	Character	3	The **ISO 3166* 3-digit code** for this country.
ISONUMERIC	Character	3	The **ISO 3166* 3-digit numeric code** for this country.
CITY	Character	70	The **settlement name** as it appeared in the incoming file. This is raw data and can contain any character of any type – ASCII control characters, Unicode characters, punctuation etc.
PC	Character	20	The **postal code** applying to this settlement as found in the incoming file. **Note**: if this field contains additional, repeated or extra address information such as a country code, a place name or a sorting code then that information may be retained in

			this field. If a postal code is incorrectly formatted but would be correct if correctly formatted, then it is retained in the data table. If this field contains clearly incorrect data, it is not supplied.
CORR_PC	Character	15	The **postal code** which is **corrected** in terms of length, format and allowable characters. For example, for the code in the PC field: *GB-W 1A4WW* this field will contain: *W1A 4WW* Where the PC field is known to contain a legacy postal code (i.e. a postal code from an older postal code system), this field contains '*****'
UP_CORRECT	Character	70 Length 80 in file names ending br_a br_b br_c br_d br_e	This is the **full upper-case version** of the alternative settlement name given in the field CITY, including those diacritics which can be stored in Windows code page 1252. To understand what we regard as "correct", please see the section below.
DN_CORRECT	Character	70 Length 80 in file names ending br_a br_b br_c br_d br_e	This is the **full mixed-case version** of the alternative settlement name given in the field CITY, including those diacritics which can be stored in Windows code page 1252. To understand what we regard as "correct", please see the section below.
UNI_UP	Character	70 Length: 95 in files with the name ending in xx Length: 1 and always empty in files with names ending: br_a br_b br_c br_d br_e ca_a ca_b gb in nl us_a	This is the **full upper-case version** of the corrected settlement name given in the field UP_CORRECT with Unicode place holders. This field is only filled if the diacritical marks in the settlement name cannot be reproduced in Windows ANSI 1252. The Unicode value for the missing diacritical mark(s) are show as the Unicode number between pointed brackets. Thus the city of *GDAŃSK* will be shown in this field as *GDA<0143>SK.*

		us_b	
UNI_DN	Character	70 Length: 95 in files with the name ending in xx Length: 1 and always empty in files with names ending: br_a br_b br_c br_d br_e ca_a ca_b gb in nl us_a us_b	This is the **full mixed-case version** of the corrected settlement name given in the field DN_CORRECT with Unicode place holders. This field is only filled if the diacritical marks in the settlement name cannot be reproduced in Windows ANSI 1252. The Unicode value for the missing diacritical mark(s) are show as the Unicode number between pointed brackets. Thus the city of *Gdańsk* will be shown in this field as *Gda<0144>sk.*
INCL_EXTRA	Logical	1	Is TRUE if the UP_CORRECT and DN_CORRECT fields contain data other than one or more place names, such as a state or province name or code. NOTE: This field is filled automatically for these countries: American Samoa, Australia, Canada, Guam, Italy, Marshall Islands, Micronesia, Northern Mariana Islands, Palau, Puerto Rico, Spain, USA, US Virgin Islands. For other countries the field is used incidentally.
COUNT	Numeric	10	The number of times that the **CITY/PC combination** has been found in real world data analyzed by GRC Database Information.
LEVEL1	Character	50	Province or state information for this postal code area. Please see information below in **Country specific information** section.
POSTAL	Character	50 Length: 3 br_a br_b br_c br_d br_e Length: 1 and always empty in files with names ending: gb	Province or state information as used in postal addresses for this postal code area. Please see information below in **Country specific information** section.

		nl	
LEVEL2	Character	50 Length: 1 and always empty in files with names ending: br_a br_b br_c br_d br_e ca_a ca_b nl	Region information for this postal code area. Please see information below in **Country specific information** section.
DATE_ADDED	Date	8	Date that the record was added. For all records added on or before 19th January 2004, this field contains the date 19th January 2004.
DATE_CHANG	Date	8	Date that the city/postal code combination was last encountered in a file. Note: this does NOT refer to any changes which are made in any other fields in the file, which are constantly being updated.
LATITUDE	Numeric	20.9	The latitude of the centroid of the postal code, to two places after the decimal point. See explanation about this information above, and for each country concerned below. NB: Latitude information is available for a limited number of countries only, and coverage is often not complete. Please refer to http://www.grcdi.nl/settlements.htm and http://www.grcdi.nl/cities_count.pdf
LONGITUDE	Numeric	20.9	The longitude of the centroid of the postal code, to two places after the decimal point. Negative numbers are west of the Greenwich meridian. See explanation about this information above, and for each country concerned below. NB: Longitude information is available for a limited number of countries only, and coverage is often not complete. Please refer to http://www.grcdi.nl/settlements.htm and http://www.grcdi.nl/cities_count.pdf

* Please note that the table contains separate entries for the entity Somaliland (GRCID: *SOA*). This entity does not have an ISO 3166 code so only the GRCID country code has been assigned.

Table structure (for file Unicode_xx)

Unicode_xx is a Microsoft Excel file which has been added to contain places names which contain characters which cannot be reproduced in the Windows code page 1252. The way in which these characters can be transferred to, and read by, other software may be limited.

Note that this file does not contain place names with overstrike Unicode characters, nor those with the character “Latin small letter turned e” (Unicode: <01DD>)

Field name	*Field type*	*Field length*	*Contents*
URN	Character	10	A **unique code** linking this file back to the data in the file **settlements_xx**
UNI_UP	Character	95	This is the **full upper-case version** of the corrected settlement name given in the field UP_CORRECT in the file SETTLEMENTS_XX. This field is only filled if the diacritical marks in the settlement name cannot be reproduced in Windows ANSI 1252.
UNI_DN	Character	95	This is the **full mixed-case version** of the corrected settlement name given in the field DN_CORRECT in the file SETTLEMENTS_XX. This field is only filled if the diacritical marks in the settlement name cannot be reproduced in Windows ANSI 1252.

What is “correct”?

The place name used in the “corrected” fields is the name which is correct according to a number of principles, including correct in (one of) the local languages, culturally correct (used by the inhabitants) or postally correct. It will contain diacritical marks (if appropriate) and punctuation (where valid). It will always be a local version of the name.

When raw data (that stored in the CITY field) contains a smaller area than a place name, such as a street name, it is excluded from the file when processed. If it contains a larger area, such as a province name, this data is retained in the UP_CORRECT and DN_CORRECT fields where postally valid. Please refer to the country specific information, below, for additional information.

Where a postal code covers a number of separate places (e.g. villages), then if a village name was used originally, it is retained, even though the post office may prefer a different place name to be used. On the other hand, names of sections (districts) of cities (e.g. *Manhattan* instead of *New York*) are corrected to the name of the city.

When there are no overriding cultural considerations, the postal version is used. Thus *NEW YORK (NY)* and *NEW YORK NY* are both culturally accurate, so the latter is used as it is preferred postally.

Choices which have been made when choosing the correct version are presented below, where they differ in specifics from these principles.

The postal codes are not “corrected” when they are found “incorrect” in the incoming data. However, for them the rule applies that if the postal code field contains additional, repeated or extra address information such as a country code, a place name or a sorting code then that information may be retained in this field and supplied in that form. If a postal code is incorrectly formatted but would be correct if correctly formatted, then it is retained in the data table and supplied. If this field contains clearly only non-postal code or incorrect data, it is not supplied.

Country specific information

Transliteration: where the original place names are in a script other than the Latin script used in English (e.g. Greek, Arabic, Chinese, Russian etc.), no single transliterated version exists – many versions can be regarded as correct. In this table, as far as possible we have used a single transliteration version for the whole country, depending upon sources available to us.

Andorra: LEVEL1 contains *parish*.

Algeria: LEVEL1 contains *wilayas*. LEVEL2 is empty.

Australia: LEVEL1 and POSTAL contain *states and territories*. LEVEL2 is empty except for overseas territories (Christmas Island, Cocos (Keeling) Islands and Norfolk Island). UP_CORRECT and DN_CORRECT fields will contain state/territory abbreviations if these were included in the raw data in the CITY field.

Austria: LEVEL1 contains *Bundesländer*. LEVEL2 is empty. Geocoding is for the full length of the postal code (4 digits). In corrected place name fields, SANKT written as SANKT rather than ST.

Azerbaijan: CITY field contains many post office names.

Belarus: LEVEL1 contains *voblasts’*. Minsk city is not distinguished from Minskaya.

Belgium: LEVEL1 contains *regions and provinces*. LEVEL2 contains *federal area* (Vlaanderen, Wallonie or Bruxelles / Brussel). Geocoding is for the full length of the postal code (4 digits). For *Brussels*, the corrected version will

contain one of the local language version (*Brussel*, *Bruxelles* or *Brüssel*) unless the original version is in another language, when the English form *Brussels* is used.

Brazil: UP_CORRECT and DN_CORRECT fields will contain state abbreviations if these were included in the raw data in the CITY field. LEVEL1 contains state name, POSTAL the two-letter state abbreviation.

Canada: LEVEL1 and POSTAL contain *provinces and territories***.** LEVEL2 is empty. *Québec* is given in French, *New Brunswick* in English and French, the other names in English. UP_CORRECT and DN_CORRECT fields will contain province/terrority abbreviations if these were included in the raw data in the CITY field.

China: LEVEL1 contains provinces.

Costa Rica: LEVEL1 contains provinces.

Croatia: LEVEL1 contains *zupanije (counties)*.

Denmark: Records from Greenland and Faeroe Islands has been allowed when found in Danish databases. Geocoding is for the full length of the postal code (4 digits). LEVEL1 contains region name. The boundaries between postal code regions 6 and 7 do not coincide with regional boundaries so accuracy for these areas cannot be guaranteed.

Dominican Republic: LEVEL1 contains provinces.

Egypt: LEVEL1 contains governorates.

Federated States of Micronesia: LEVEL1 and POSTAL contain the United States Postal Service state indicator. LEVEL2 is empty.

Finland: LEVEL1 contains *sub-region* for postal code areas which are not shared between provinces. The boundaries between postal code do not always coincide with regional boundaries so accuracy for these areas cannot be guaranteed. LEVEL2 contains *province*.

France: LEVEL1 contains *departments and overseas territories*. LEVEL2 contains *regions* (for metropolitan France only). Geocoding is for the full length of the postal code (5 digits). Upper case corrected versions do not contain diacritical marks, the mixed case versions do (including those on upper-case letters within the mixed-case place name version). Numbers following *Cédex* have preceding zeroes removed: *Cédex 5* instead of *Cédex 05.*

Germany: City districts (Ortsteil, Stadtteil), (e.g. *Charlottenburg*) have (as much as possible) been assigned corrected names as the name of the city (e.g. *Berlin)*, not the name of the district. Geocoding is for the full length of the postal code (5 digits); SANKT is written in full in place names. LEVEL1 contains *Bundesländer*.

Guatemala: LEVEL1 contains *departamentos*.

India: The CITY, UP_CORRECT and DN_CORRECT fields contain, in many cases, a post office name or address, street ranges or other information indicating postal code coverage. LEVEL1 contains *states and union territories*. LEVEL2 is empty. **Note**: there may be an overlap in postal areas between states and union territories not available in this table.

Indonesia: LEVEL1 contains *province* (where postal code districts are not shared between provinces).

Italy: LEVEL1 and POSTAL contain *provinces* where postal codes are not shared between provinces. LEVEL2 contains *regions*. Geocoding is for the full length of the postal code (5 digits). UP_CORRECT and DN_CORRECT fields will contain state abbreviations if these were included in the raw data in the CITY field.

Jamaica: LEVEL1 contains parishes.

Kenya: The CITY, UP_CORRECT and DN_CORRECT fields contain, in many cases, a post office name or address.

Kosovo: Places in Kosovo have two postal codes: those used by the Serbian postal services, and those used by the Kosovan postal services. Given Serbia's non-recognition of Kosovo's secession, and the likelihood that Serb minorities within Kosovo will continue to use Serbian postal codes, Kosovan place names are listed for both Serbia and Kosovo. In the former case Serb postal codes are given, in the latter case Kosovan.

Laos: LEVEL1 contains provinces.

Luxembourg: UP_CORRECT and DN_CORRECT fields contain the local (Luxembourgeois) version of each place name, rather than the French or other version. Geocoding is for the full length of the postal code (4 digits)

Madagascar: LEVEL1 contains parishes.

Malaysia: LEVEL1 contains states and federal territories.

Marshall Islands: LEVEL1 and POSTAL contain the United States Postal Service state indicator. LEVEL2 is empty.

Mexico: LEVEL1 contains *states*. LEVEL2 is empty.

Micronesia: LEVEL1 and POSTAL contain the United States Postal Service state indicator. LEVEL2 is empty.

Mozambique: LEVEL1 contains provinces.

Netherlands: The postal service sometimes uses a province indicator where more than one settlement of a given name exists in the country. This is not a requirement if the postal code has been given. In this file the province indicators have been retained if given in the CITY field, but have not been added if not given in the incoming data.
LEVEL1 contains *provinces*. LEVEL2 is empty.
Geocoding is for the first four digits of the postal code.

New Zealand: The CITY, UP_CORRECT and DN_CORRECT fields contain, in some cases, a post office name or address.

Norway: LEVEL1 contains *county* for postal code areas which are not shared between counties. The boundaries between postal code do not always coincide with regional boundaries so accuracy for these areas cannot be guaranteed.

Oman: LEVEL1 contains *regions*.

Pakistan: LEVEL1 contains *territories* and *provinces*.

Palau: LEVEL1 and POSTAL contain the United States Postal Service state indicator. LEVEL2 is empty.

Papua New Guinea: LEVEL1 contains the province name. POSTAL and LEVEL2 are empty.

Philippines: Place names may be replaced with large user names for some postal codes. LEVEL1 contains provinces.

Poland: LEVEL1 contains *province* names.

Portugal: LEVEL1 contains *district* names.

Puerto Rico: LEVEL1 and POSTAL contain the United States Postal Service state indicator. LEVEL2 is empty.

Romania: LEVEL1 contains *judete (county)* names

Russia: LEVEL1 contains *oblasts*, autonomous republics, *okrugs* and *krays*.

San Marino: LEVEL1 and POSTAL contain the Italian Postal Service province indicator. LEVEL2 is empty.

Spain: LEVEL1 and POSTAL contains *provinces*. LEVEL2 contains *regions*. Geocoding is for the full length of the postal code (5 digits). UP_CORRECT and DN_CORRECT fields will contain state abbreviations if these were included in the raw data in the CITY field.

South Korea: LEVEL1 contains provinces and metropolitan cities.

Sweden: Geocoding is for the first three digits of the postal code. LEVEL1 contains *county* for postal code regions which do not include ground in multiple counties. Note that the postal code area equivalents are not exact and place names along the county borders may be misassigned.

Switzerland: Canton abbreviations have been removed from settlement names in corrected data fields as they are not required for addressing. Geocoding is for the full length of the postal code (4 digits). LEVEL1 contains the full canton name in the local language(s), POSTAL the canton abbreviation.

Taiwan: Settlement names are listed in both Tongyong Pin Yin (Romanized phonetic system) transliteration and Han Yu Pin Yin (The United National Mandarin Phonetic System) transliterations; and other, non-formalised, transliterations. We have attempted to use the same transliteration for each place name to aid consistency.

Thailand: LEVEL1 contains *changwat* (provinces).

Tunisia: LEVEL1 contains *governorate* names.

Turkey: LEVEL1 contains *province* names; LEVEL2 contains *bölge* (*census-defined regions*).

Ukraine: LEVEL1 contains *oblasti* names.

United Kingdom: Geocoding is for the whole of the outward section of the postal code, and the first digit of the inward section of the code. UP_CORRECT and DN_CORRECT fields will contain county names if these were included in the raw data in the CITY field. LEVEL1 contains county/unitary authority. NOTE: administrative areas and postal code areas are NOT contiguous in the UK. We have only added this information at the request of customers and cannot guarantee it being 100% accurate. Administrative structures are complex in the UK and we have maintained the most accurate possible solution but exclude all postal code areas which straddle administrative district boundaries.

Furthermore, a place may topographically be in one administrative area whilst a ceremonial, even non-existent, county is used in addresses.
About 90% of entries have an administrative area assigned to this field.

United States of America: LEVEL1 and POSTAL contain *state*. LEVEL2 contains *county*. UP_CORRECT and DN_CORRECT fields will contain state abbreviations if these were included in the raw data in the CITY field.

Uruguay: LEVEL1 contains *departamentos*.

Vietnam: LEVEL1 contains provinces.

End