



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>

Addresses

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

GRC Database Information do not guarantee that each address component string or other string listed is found on the ground in the real world. Strings found in a country in one language region have in some cases been applied to other countries within the same language region to improve accuracy and coverage.

Highlighted text indicates changes since the release of the previous version.

The table *Addresses* contains strings which are found in addresses, and indicates which part or parts of an address these strings might represent.

The data is provided in Windows code page 1252, complying to ISO-8859-1, MS-DOS code page 850. Diacritical characters (accents) for most Western European languages are reproduced in the table. Those for Eastern European and non-European languages have been replaced with equivalent characters. Though our database systems can store Unicode data, we cannot enter it in normal use. We therefore provide fields with the same address strings with Unicode place holders so that you can translate the data to Unicode should you wish to do so.

Strings and wildcards

The wildcard \$ is used in the field ALTERN to show that the string is normally found concatenated to another string. Thus **\$straße** might be found as **Bahnhofstraße**, **Cwm\$** as **Cwnbran** and so on.

In all cases, the strings must be applied in context to be effective. A string that indicates a thoroughfare type when assessing a street address may indicate a company type in another context. For example, the abbreviation **\$stra** in Germany is indicative of the thoroughfare type **Straße** when found in a street address string, and has been shown as such in the table, but the same string can also indicate a company name or a city name and can therefore not be applied willy-nilly on any strings.

The strings stored in the fields UP_CORRECT and DN_CORRECT are those that ensure the greatest consistency, without unnecessarily sacrificing data fullness. For **companies**, the default **legal forms** given in the fields UP_CORRECT and DN_CORRECT are the contracted forms of the string, as these are what are generally accepted in normal use. For other company strings, normally the expanded form is given in these fields. For **thoroughfares**, the default is that the correct forms given in the field UP_CORRECT and DN_CORRECT are the expanded forms of the string. However, in some cases expansion can lead to doubt. In Dutch, for example, the abbreviation *PR* can indicate the word *PRINS* or the word *PRINSES*. Because the accuracy of any expansion cannot be guaranteed, for these strings the correct forms given are the abbreviations, so that upon application of this table to real-world data, accuracy and consistency can be maintained. Furthermore, where a string may occur naturally in the language concerned, it has not been given an expanded form. For example, the thoroughfare type *chaussee* is found in Dutch. A word ending with *ch.* (note the full-stop) in a street address field is assumed to be an abbreviation of *chaussee*, whereas a word ending in *ch* (note no full-stop) the abbreviation is retained because this can be a normal word ending.

Whilst the fields UP_CORRECT and DN_CORRECT contain strings to try to ensure consistency, the fields FULL_UP and FULL_DN contain the full version of a string where the full version can be assumed from what is written in the ALTERN field. Thus, for the Dutch strings PR, PRINS and PRINSES, the table will contain:

<i>Altern</i>	<i>Up_Correct</i>	<i>Dn_Correct</i>	<i>Full_Up</i>	<i>Full_Dn</i>
PRINSES	PR.	Pr.	PRINSES	PrinSES
PRINS	PR.	Pr.	PRINS	Prins
PR	PR.	Pr.	PR.	Pr.
PR.	PR.	Pr.	PR.	Pr.
PrinSES	PR.	Pr.	PRINSES	PrinSES
Prins	PR.	Pr.	PRINS	Prins
Pr	PR.	Pr.	PR.	Pr.
Pr.	PR.	Pr.	PR.	Pr.
etc.

The UP_CORRECT and DN_CORRECT fields contain the most consistent acceptable abbreviation for these alternative forms, whilst the fields FULL_UP and FULL_DN contain the long versions, provided a long version can be extrapolated from an abbreviated version. Using all of this information, a company should be able to link all abbreviations and long forms which are linked to one another via their long or short forms, and apply rules as necessary depending on the needs of their database or their business rules.

NB: As new data is found and analysed, and new strings are added to the tables, the relationships between the data in the fields ALTERN, UP_CORRECT, DN_CORRECT, FULL_UP and FULL_DN will change – these fields are being regularly re-assessed and their contents can change.

Note also that consistency is based upon context. FL. in The United Kingdom could be an abbreviation for FLOOR or FLAT, both sub-building types, so UP_CORRECT and DN_CORRECT contain FL. If the context were to be different, such as one string being a sub-building type, the other a settlement indicator, the standardisation in these fields would not be considered necessary.

Note that apparent language anomalies are not necessarily written in full. AVE is found in Dutch data, and may be assumed to be an abbreviation of the English or French AVENUE, but as this is not certain, it is retained in its original form in the tables.

Casing

The default casing for strings within the fields DN_CORRECT and FULL_DN is the casing that would be found if this string were to be found in the middle of a line within an address block. Thus, the English preposition *on* and the French thoroughfare type *rue* are written with a lower case first letter. However, the language rules for the country concerned should be followed when applying the data to real World situations. If the English preposition *on* is written at the start of a line, it will take a capital letter, whilst *rue* will retain its lower case first letter.

Where the strings are in a country where words are concatenated with each other, such as The Netherlands and Germany, the casing of the full string is usually in lower case, for example *straat*, the exception being where the word is normally found written on its own. When the word is found on its own, such as *Kerk Str.* then the first letter of the full string should be upper cased like this: *Kerk Straat*. Equally, should a word which is in the table with an upper case first letter be found concatenated to another word, its first letter should be lowered.

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. These have alternative case forms. For example, if we found this element:

USA AVE.

We have added (if not already existing)

USA Ave.

USA ave.

Tagging and statistics

The files contain (for some countries) two sets of indicators to allow you to assess the strings. The first is a set of logical values (*TRUE* or *FALSE*) to indicate that the string refers linguistically to a specific address element. A “true” value in the tables may be applied for each string to more than one category. This is because a string may apply to more than one address component. For example, *COURT* may apply to a building or to a street. Users should bear this in mind when applying the data from these tables to the real world situation. The logical flags indicate that the word is indicative of this type of address element, but does not preclude this string being used in other parts of addresses. Thus *STREET* is tagged as being indicative of a thoroughfare, but there is a settlement name called *STREET* in the United Kingdom. The string *WATER* is tagged as *OTHER*, but it can be used as part of a company name, building name, street address string,

settlement name, personal name and so on. Where available, the statistics fields will indicate to what extent the string is used within differing address sections.

For countries where strings are often found concatenated with each other, such as Germany, each string has been added to the table in a number of different formats, for example:

Straße
 Straße-
 Straße\$
 Straße-\$
 \$Straße
 \$-Straße

etc. Not each of these formats are necessarily found on the ground in the country concerned, but the strings have been retained in the tables for two reasons: firstly, to show which combinations do not appear; and secondly to allow for incorrect data-entry by users in the format concerned. The statistics for the string in each format are not necessarily exclusive of each other. For example, **Straße\$** will also count occurrences counted by **Straße-\$**, **Zwei\$** will also count occurrences of **Zweite**, and so on.

Utilisation

When using the tables for standardisation, validation and so on, it is advisable to sort the table on the basis of length of the string (in characters) in the field *ALTERN*, so that the longest is at the top of the file and the shortest at the bottom. This is because a number of shorter strings exists in the file which are also components of longer strings, such as *BOX* and *P.O. BOX*. It is better to locate and standardise (in this example) *P.O. BOX* before locating and standardising *BOX*.

Table structure

<i>Field name</i>	<i>Field type</i>	<i>Field length</i>	<i>Contents</i>
URN	Numeric	10	A unique number which the combination of data in this field. This number should be used as a reference if there are any queries about the data.
COUNTRY	Character	45	The country name in full.
GRCID	Character	3	A unique country code used by GRC Database Information
CONT	Character	3	A code indicating the continent upon which the country is. In some cases, where interpretations of continental location differ, or where a country may be interpreted as being on more than one

			continent, a continent has nevertheless been assigned. The codes used are: <ul style="list-style-type: none"> • AFR (Africa) • ANT (Antarctica) • ASI (Asia) • AUS (Australasia/Oceania) • EUR (Europe) • NAM (North America) • SAM (South America)
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.
STANDAR_UP	Character	60	This is the standardised upper-case version of the alternative address element string given in the field ALTERN. This version does not necessarily contain locally recognised standardisations; this data can therefore be used for data processing but not for data which will be output. This field is empty if INVALID=.T.
STANDAR_DN	Character	60	This is the standardised mixed-case version of the alternative address element string given in the field ALTERN. This version does not necessarily contain locally recognised standardisations; this data can therefore be used for data processing but not for data which will be output. This field is empty if INVALID=.T. Note: A string may commence with a lower case letter in this field for one of two reasons: <ul style="list-style-type: none"> • in this country, the address element type is normally written with a lower-case first letter when this data is found written within an address block line (e.g. thoroughfares in France) • the string may be written concatenated to the rest of the name, such as in street names in Germany. In this case, the

			<p>element should be written with a lower-case first letter when concatenated but an upper-case first letter when found standalone.</p>
ALTERN	Character	60	<p>An alternative form of the address element string. This includes correct forms, abbreviated forms and common mis-typings. The compiler of this list has been very careful to exclude strings which may be confused with other string types in applying the data in this table to address databases in practice. For example, the string <i>CH</i> is an abbreviation of the thoroughfare <i>CHAUSSEE</i> in German, but it is also a normal word ending (e.g. <i>KIRSCH</i>). To prevent the application of this table which would cause all occurrences of <i>KIRSCH</i> to become <i>KIRSCHAUSSEE</i>, the alternative <i>CH</i> has not been added to the table, but the alternative <i>CH.</i> (i.e., followed by a full-stop) is included.</p> <p>The use of punctuation and casing has also been carefully applied to reduce application errors.</p>
FULL_UP	Character	60	<p>This is the full (where possible) upper-case version of the alternative address element string given in the field ALTERN. This field is empty if INVALID=.T.</p>
FULL_DN	Character	60	<p>This is the full (where possible) mixed-case version of the alternative address element string given in the field ALTERN. This field is empty if INVALID=.T.</p> <p>Note: A string may commence with a lower case letter in this field for one of two reasons:</p> <ul style="list-style-type: none"> • in this country, the address element type is normally written with a lower-case first letter when this data is found written within an address block line (e.g. thoroughfares in France) • the string may be written concatenated to the rest of the

			name, such as in street names in Germany. In this case, the element should be written with a lower-case first letter when concatenated but an upper-case first letter when found standalone.
UNI_FULLLUP	Character	60	Where the upper-case string reproduced in the field FULL_UP should contain characters which cannot be reproduced under Windows code page 1252, these strings are shown in this field with a Unicode value replacing the character which cannot be reproduced. The Unicode value is given between <> brackets. Thus, the Hungarian <i>LÉPCSO</i> should have a double acute accent on the <i>O</i> . This is therefore reproduced in this field as <i>LÉPSC<0150></i> .
UNI_FULLLDN	Character	60	Where the mixed-case string reproduced in the field FULL_DN should contain characters which cannot be reproduced under Windows code page 1252, these strings are shown in this field with a Unicode value replacing the character which cannot be reproduced. The Unicode value is given between <> brackets. Thus, the Hungarian <i>lépsc</i> should have a double acute accent on the <i>o</i> . This is therefore reproduced in this field as <i>lépsc<0151></i> .
INVALID	Logical	1	A 'T' indicates that the string is an obscenity or otherwise indicates a non-name or address entry, such a punctuation marks, "n/a", "Private" etc.. These strings indicate the <u>full</u> contents of a field (i.e. "Private" on its own is an invalid entry, whilst in combination with other strings it can be part of a valid address such as "Private Road"), and no corrected forms are provided for these strings.
DIRECTIONA	Logical	1	A 'T' in this field indicates that this is a directional , that is, a string indicating a compass bearing, such as North, South etc.
COMPANY	Logical	1	A 'T' in this field indicates that this

			string is indicative of a company , either a legal form (e.g. LIMITED) or another indicator or a company, such as <i>INHABER</i> or <i>AIRLINE</i> .
DEPARTMENT	Logical	1	A 'T' in this field indicates that this string is indicative of a department within a company.
SUBBUILDIN	Logical	1	A 'T' in this field indicates that this string is indicative of a sub-building , that is, a part of a building such as a <i>ROOM</i> or an <i>ENTRANCE</i> .
BUILDING	Logical	1	A 'T' in this field indicates that this string is indicative of a building .
UNIVERSITY	Logical	1	A 'T' in this field indicates that this string is indicative of a university .
STREET	Logical	1	A 'T' in this field indicates that this string is indicative of a thoroughfare .
ZONE	Logical	1	A 'T' in this field indicates that this string is indicative of a district or zone , such as <i>INDUSTRIAL ESTATE</i> .
SETTLEMENT	Logical	1	A 'T' in this field indicates that this string is indicative of a settlement such as <i>VILLAGE</i> .
POSTBOX	Logical	1	A 'T' in this field indicates that this string is indicative of a postbox string, such as <i>PO Box</i> , <i>POSTFACH</i> .
SAL_PREFIX	Logical	1	A 'T' in this field indicates that this string is indicative of routing to an individual, for example <i>Care of</i> or <i>Attention</i> .
SALUT	Logical	1	A 'T' in this field indicates that this string is indicative of a form of address or a job title , such as <i>REVEREND</i> , <i>MAGISTRATE</i> .
PERS_NAME	Logical	1	A 'T' in this field indicates that this string is indicative of a personal name , such as <i>JOSEPH</i> , <i>PEPE</i> .
MAN_MADE	Logical	1	A 'T' in this field indicates that this string is indicative of a man-made feature other than a building or thoroughfare, such as <i>CANAL</i> , <i>MEADOW</i> .
GEOGRAPHIC	Logical	1	A 'T' in this field indicates that this string is indicative of a geographical feature , such as <i>HILL</i> , <i>RIVER</i> .
BIOLOGICAL	Logical	1	A 'T' in this field indicates that this string is indicative of a plant or animal

			name , such as <i>SWAN</i> , <i>SHEEP</i> .
COLOUR	Logical	1	A ‘T’ in this field indicates that this string is indicative of a colour , such as <i>WHITE</i> .
NUMBER	Logical	1	A ‘T’ in this field indicates that this string is indicative of a number (numeric, cardinal or ordinal) such as <i>ONE</i> , <i>FIRST</i> .
ARTICLE	Logical	1	A ‘T’ in this field indicates that this string is indicative of a definite or indefinite article , such as <i>THE</i> , <i>A</i> .
PREPOSITIO	Logical	1	A ‘T’ in this field indicates that this string is indicative of a preposition or positional strings, such as <i>IN</i> , <i>BEHIND</i> , <i>BY</i> .
ADJECTIVE	Logical	1	A ‘T’ in this field indicates that this string is indicative of an adjective , such as <i>LARGE</i> , <i>WIDE</i> .
OTHER	Logical	1	A ‘T’ in this field indicates that this is a string which commonly occurs in addresses and can be recognized and standardized, but does not belong to any of the above-mentioned groups.
ENGLISH	Character	50	The English translation (for some strings) of the non-English language string. Contains “Invalid” for all strings where INVALID=.T.
ISO_639_2	Character	3	The ISO-639-2 (alpha-3, terminological) language code indicating the language region in which this string is found (for multi-lingual countries or countries where another language is used for some address elements). Codes currently applied: <ul style="list-style-type: none"> • afr (Afrikaans) • ara (Arabic) • aze (Azerbaijani) • bul (Bulgarian) • cat (Catalan) • ces (Czech) • cym (Welsh) • dan (Danish) • deu (German) • ell (Greek) • eng (English)

			<ul style="list-style-type: none"> • esp (Spanish) • est (Estonian) • fin (Finnish) • fra (French) • fry (Frisian) • gla (Gaelic, Scots) • gle (Irish) • heb (Hebrew) • hin (Hindi) • hrv (Croatian) • hun (Hungarian) • isl (Icelandic) • ita (Italian) • ind (Indonesian) • jpn (Japanese) • kal (Greenlandic) • kaz (Kazakh) • kor (Korean) • lav (Latvian) • lit (Lithuanian) • ltz (Lëtzebuergesch) • mkd (Macedonian) • mlt (Maltese) • mol (Moldavian) • msa (Malay) • nld (Dutch) • nor (Norwegian) • pap (Papiamentu) • pol (Polish) • por (Portuguese) • ron (Romanian) • rus (Russian) • slk (Slovakian) • slv (Slovenian) • srp (Serbian) • swe (Swedish) • tgl (Tagalog) • tha (Thai) • tur (Turkish) • ukr (Ukrainian) • vie (Vietnamese) • zho (Chinese)
ISO_639_2_2	Character	3	As a single string may be used in more than one language (for example,

			<i>AVENUE</i> in English and French), this is the ISO-639-2 (alpha-3, terminological) language code indicating the second language region in which this string is found (for multi-lingual countries). See above for further explanation.
REAL_WORLD	Numeric	10	The number of occurrences of this string found in real-world address data. NB: During analysis, data is sorted on the basis of frequency already found in data, and for each field, only one string is counted. Thus, for a field containing <i>1st floor Station House</i> , the string <i>House</i> (being the string from this field found most often) will be counted, but the strings <i>floor</i> and <i>Station</i> will not. This count is therefore useful as an indicator of commonality of a string, but should not be regarded as exact.
ALL_WORLD	Numeric	10	The number of real world address records analysed for this string for this country.
IS_COUNTED	Logical	1	TRUE if this string has been analysed against the postal tables (i.e., for the 12 previous fields); FALSE if it has not.
DATE_ADDED	Date	8	Date of addition of this record to the table (dates before 1 st October 2000 are dated 1 st October 2000)
DATE_CHANG	Date	8	Date of last change made to this record

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

End

Graham Rhind
GRC Database Information
Nieuwe Prinsengracht 80-hs
1018 VV AMSTERDAM
THE NETHERLANDS
020-5286120 (telephone)
020-4229311 ('fax)
graham@grcdi.nl
<http://www.grcdi.nl/addresses.htm>

