The Basics of

International Address Hygiene

a plain language introduction to the terminology and concepts of address correction in international prospect and customer databases

Created by

WorldVu™
International Information for Businesses

Distributed by

GLOBAL ADDRESS DATA ASSOCIATION

Introduction

It has always been a puzzle as to why the process of accomplishing something so intrinsically tied to the success of companies who do business internationally should be given so little attention by those companies. Address hygiene is, almost always, an afterthought. It is often the last thing to be budgeted and the first thing to be cut. And, as when a new senior executive decides to cut the direct marketing budget, this is always a bad idea.

Part of the problem is that the apparent complexity and "secret lingo" of the process makes it difficult to fully appreciate its intrinsic impact on business performance and marketing effectiveness. Yes, it does make a difference if the title in the address to a prospective male customer in Germany is correct. Yes, it does matter if the customer's street address has changed due to construction. Yes, it does matter if you have three different spellings of the street in Nairobi where your most important client in Kenya apparently lives.

And, no, the file you are renting from company X consisting of prospects for your product in France will not be free of name misspellings, non-current addresses, duplicates, and completely unintelligible data.

Without question, international address hygiene can be a complex and complicated process, but learning the process need not be complex and complicated. One must simply recall that address hygiene is a series of processes, and these processes are more easily understood if learned separately. Each of these processes has as its goal the correction, improvement, or provision of one or more elements or aspect of the address. And the ultimate goal is to maintain contact with customers and prospects, be it to fulfill existing obligations, or to increase ROI with new sales or relationships.

An understanding of the basics of international addressing and the component processes of international address hygiene will allow managers, marketers and others to more fully understand the nuances when discussing the subject with programming staff or with service bureaus, and of course with the Treasurer and senior management in general.

It is with this goal that WorldVu and the Global Address Data Association provide this work. To allow for better understanding, the most examples we use in this work will be familiar to those in Europe or North America. The same principles apply to the processing of names and addresses from other regions, although specifics differ because of the differences in naming and addressing practices.

International Addressing Basics

A few simple rules to start with:

1. The format of an address – any address – will depend on the type of address and the country in which it is located.

   The basic address types include a number of ***elements*****:** street or building name, post office box or bag, general delivery or *poste restante*, postal code, city or town. There are also specially-formatted military or diplomatic addresses. In addition to the basic address information, an address may optionally include a building designation, such as a building name or number; suite, apartment or floor information; a corporate name, department or internal mail delivery destination; or other similar details.

   Finally, while each address is made up of address elements, such as a house or building number, a street name, a postal code, city or town, in the case of international mail, there is also the all-important country name.

2. Each country specifies its own address requirements, including what elements are used and their placement within the address formats used by that country.

   An ***address format*** may also be called an ***address template*** and is the address "block" as written on an envelope. Most countries have multiple address formats for the different address types that are used. For example, the address format for post office box addresses will differ from street or building addresses.

3. There are about 30 different very ***basic formats*** used worldwide as defined by the elements used and their placement in the address.

   In the U.S., these elements include a state abbreviation and ZIP (or postal) code, both placed to the right of the town or city name. In Germany, no state or province name or abbreviation is used and the PLZ (or postal code) is placed to the left of the town or city name. Taking placement of address elements, punctuation and spacing into account, there are hundreds of variations on the basic formats. At this time, no one has an exact count because all the addresses worldwide have not been defined.*

4. Not all countries – in fact, only about 50 or 60 – have postal code or address databases which are kept reasonably current and made available to address hygiene service providers. The service bureaus that do international address hygiene use a variety of data as reference sources, ranging from the official governmental or postal data available on the market and other commercially-available data, which in some cases is constructed by vendors or by the service bureau itself on a proprietary basis.

   In short, international address hygiene is part science, part politics, part software selection, and part business. Results can vary significantly from bureau to bureau.

---

* The Universal Postal Union is developing a standard, designated S42 that describes and defines all address formats in each country. This is a lengthy exercise being carried out by volunteers in cooperation with each country's address or postal authority. To date, XX countries have participated. This project is the first systematic attempt to define all the addresses used worldwide and will create a single set of formats, or templates, accepted by the major standards authorities, including the International Standards Organization (ISO).

What is International Address Hygiene?

International address hygiene is a set of computer processes applied to address data to correctly format the addresses and correct errors in them. While some companies that mail large quantities have software for international hygiene in-house, the process is most often done using an outside service bureau. Whether the hygiene is done in-house or by a service bureau, the principles are the same with the department running the hygiene process taking the place of a service bureau. After processing by a service bureau, the corrected file of addresses may be returned to the client and used to update their in-house database or used to produce envelopes or mailing labels.

Because each country has its own address requirements and formats, all international address standardization and hygiene is done country-by-country. In effect, the data for each country is processed separately. The sophistication or accessibility of the necessary tools from each country will determine the availability and level of some services with the complete hygiene protocol.

The basic processes in international address hygiene are

- address standardization
- address formatting
- address correction
- address verification.

Other services often provided by service bureaus that specialize in address hygiene may include

- duplicate identification
- merge/purge
- contact name parsing and standardization
- gender identification
- latitude and longitude appending.

We discuss each of these processes below.

Address Standardization

Address standardization is a basic requirement for most other processing because the address must be in a standard, recognized form to compare to other addresses, whether to identify duplicate addresses or to verify that the address is correct. The address standardization procedure first parses the address into its elements -- house number, apartment, suite, street name, city name, province, postal code, etc. The elements are then arranged into a standard set of fields.

For example, in one address hygiene processing run of a large file the address for the Brazilian Direct Marketing Association was found to be written in the three different ways, shown below:

ABEMD Edificio Italia 50 Av. Sao Luis, 13º a. Sao Paulo 01046-926 Brazil	Associação Brasileira de Marketing Directo 50 Av. Sao Luis Sao Paulo – SP 01046-926 Brazil	Direct Marketing Assoc. Edificio Italia - 13 andar 50 Avenida Sao Luis 01046-926 Sao Paulo Brazil

These three different versions would be parsed into the following elements as a step in standardizing and correctly formatting the address. Parsing is frequently used as part of the

standardization process, although some variation exists among service bureaus on how the standardization is accomplished.

organization	ABEMD	Associação Brasileira de Marketing Directo	Direct Marketing Assoc.
building name	Edificio Italia		Edificio Italia
building number	50	50	50
street	Av. Sao Luis	Av. Sao Luis	Avenida Sao Luis
floor designation	13º a.		13 andar
city	Sao Paulo	Sao Paulo	Sao Paulo
province		SP	
postal code	01046-926	01046-926	01046-926
country	Brazil	Brazil	Brazil

Address Formatting

Once the address is separated into its component parts, as shown above, the address can then be formatted in accordance with the postal requirements of the country where it is located. Of course, in the case of multiple addresses, as in the above example, it may be desirable to eliminate the duplicate records. (Duplicate identification and merge/purge are discussed briefly below.)

The final single address entry with complete address information, using the same fields as above, would be

organization	Associação Brasileira de Marketing Directo
building name	Edificio Italia
building number	50
street	Avenida Sao Luis
floor designation	13 andar
city	Sao Paulo
province	SP
postal code	01046-926
country	Brazil

The fields can be recombined into a correctly formatted address as it would appear on an envelope or label for mailing, and thus we get:

Associação Brasileira de Marketing Directo
Edificio Italia
Avenida Sao Luis 50, 13º andar
Sao Paulo - SP
01046-926
BRAZIL

The service bureaus which provide international address hygiene will often produce labels or envelopes or recommend lettershops that can do so, or assist their clients in the correct placement of fields.

Address Correction

The address correction process consists of comparing the addresses in the file being processed against a list of deliverable addresses, accurate street names, and extant cities and towns. The

service bureaus obtain this data from the relevant government agency in the country in question, usually the post, or from a commercial source. The nature of the information available from government, postal and private sources will obviously make a difference in whether this service is available for a particular country.

If address data available from those sources does not contain exact addressing information, it will not be possible to correct addresses. The basic principle is that sufficient information must be available to make a unique match between the address data and an address being corrected. For examples, where building numbers, street names, and postal codes are used, the address data would need to include details about what numbers on what streets are in specific postal codes..

Difficulties with correcting addresses can even occur where data is quite robust. For example, the address "1960 East Pratt Street in U.S. ZIP code 21231" does not exist. If the data correction software is properly written, the program will examine the possible number transpositions within that ZIP code. If it disclosed that there is only in one possible combination: 1906, it would then correct the address accordingly. However, if multiple possibilities existed, as for example, 6190 East Pratt Street and 9160 East Pratt Street, the address would be identified as incorrect but not automatically corrected.

Some spelling errors may be corrected in this process as well. In the same way that number transpositions occur, common typographical errors or very close spellings can be automatically corrected to the most likely correct spelling. Using the example of Pratt Street, Pratt might have been spelled erroneously as Oratt, Prstt or some other close variation. Since no such street exists and Pratt is a likely alternative, the spelling could be automatically corrected. Again, this is dependent on the availability of a list of all valid street names.

Address Verification

"Address verification" is one of those ambiguous concepts. It is related to address correction but concerns deliverability. It also can mean different things in different countries. It may mean that the address in question actually exists; or it may mean that it is a number existing within a range of valid addresses; or it might only mean that the postal code is correct for that city or town name.

In the least developed countries, it may only mean that the city or town exists in that country. As noted in the introduction, for some countries service bureaus may have access to information to supplement what is provided by the government or postal authorities. This is usually offered to customers as an additional service.

About 50 countries provide "street" or "premises" data, that is, information of what numbers recognized for postal delivery exist on particular streets, available from government or postal sources. For these countries, particular street names within a locality or postal code – say University Way in postal code 123456 – can be verified. Some of these countries provide a range of numbers that are valid for particular street names – say 12 – 58 University Way in postal code 123456. This, of course, does not always guarantee your prospect will get your mailing or that the list you rented was well put together. Although a street number may be verified because it is within an existing range, the number may be an empty lot or the premises may be vacant.

Verification that the postal code and the city or town addressed go together is available in approximately another 75 countries in addition to the approximately 50 countries that verify street addresses. In some, the verification may be to a district within a city. In many of the remaining international destinations, verification of city or town names is available. Obviously, these latter

levels of verification provide considerably less assurance of a deliverable address because the street name or the building number has not been verified.

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.

While there are numerous deduplication programs on the market, they generally all undertake the same process. Each element in the addresses in the file is examined to determine if another address in the file exactly or closely matches it.

If matches are identified, an importance (or weight) is given to each exact or near match. These weights are then processed using a formula to determine how likely it is that the records are duplicates. If a sufficient number of elements match closely, the score is high and the records are considered potential duplicates. (The formula that takes into account the weight of each element is often referred to as an algorithm.)

For example, an exact match of both the given (personal) names ***and*** the family names (James Cooper) would have a higher score than just a match of given names (James Joyce and James Cooper). If the postal code and street name also match in the two addresses (Uncas Street, 12845), the matching score would be higher still. However, a match of only the postal code and given (personal) names would result in a lower match score (James, 12845).

The criteria for duplicate identification can usually be adjusted to stricter or looser standards. A test run of a portion of the complete file or files can indicate if suspected duplicates are slipping past or if too many unique records are being flagged as duplicates. How to handle this should be discussed in advance with the potential service bureaus.

Merge/Purge

Merge/purge is the common shortened form for the process of merging information from multiple records for the same individual or company from a database or from multiple lists and purging the unwanted duplicates. The process is similar to the usual merging and purging of domestic customer files. In the example at the beginning of this discussion, three records for the same company were standardized, the address information was combined (merged), and one record would be retained while the other two would be removed (purged).

Where a record such as this one also includes an individual's name or other information, such as sales or payment history, the decision on whether information should be combined or removed becomes more complicated. The purpose of the processing will dictate how these potential duplicates should be handled.

Contact Name Parsing and Standardization

The names of individuals can be separated into component parts – honorific (Mr., Ms., etc.), given or personal name, "middle" name, family name, suffix – in the same way that address elements can be separated. This permits addressing a letter with the individual's given name or in a more formal manner.

Because the order of name components is not the same in all countries, expertise is required to properly identify the components. For example, the family name is traditionally written before the

given name in some Asian cultures, particularly Chinese, Japanese and Korean. Further, in some countries honorifics do not always precede the name, as they do in the Americas and Western Europe.

Some individuals providing their names on forms may "fix" their name to make it correspond to (usually) the European convention. This compounds the difficulty of identifying the components correctly, since some names will be in one order and others will be transposed. The skill of the service bureau is important to correct results.

This also suggests the wisdom of using different name and address intake systems which first ask for the country name, and which then change format to one which elicits information in the "local" form for both name and address.

Gender Identification

Gender identification, much like spelling correction, will depend on the expertise of the particular firm providing the service. To determine gender, one must usually identify the individual's personal or given name as separate from the family name. If the name is not already separated into these divisions, the name needs to be parsed and standardized.

Each service bureau will have its own unique list of honorifics and given names by gender. Patronymics may aid in this process because they are frequently gender-specific. Some names and some honorifics are used be both men and women in some languages and it would not be possible to determine the gender from the name or the honorific appearing in the file.

In the example below, Robert and Rob would likely be men. R. A. and Bobbie could be either men or women, since Bobbie is used as a nickname for both Robert and Roberta.

assigned gender	man	?	?	man
full name	Robert A. Reeve	Bobbie Reeve	R. A. Reeve	Rob Reeve, Jr.
given name	Robert	Bobbie	R.	Rob
middle name	A.		A.	
family name	Reeve	Reeve	Reeve	Reeve
suffix				Jr.

Latitude and Longitude Appending

Knowing the exact geographic location of addresses permits mapping them. This has a number of business and public service applications. However, in most postal systems it is not possible to derive a geolocation from a postal address. One would need a file that matches the two of them. Once again, the ability to do this varies by both country and service bureau. Most postal authorities and government agencies do not now collect this information, although the Royal Mail in the U.K. has begun a project to do so and other developed countries are likely to follow. Because of this, data from private companies is used for this process.

Note on Country Variations

As noted, there are vast differences among countries in what postal address data is available. The most developed countries often have a complete database of all delivery addresses, with regular updates. Other countries may have a database of all postal codes and the associated cities, a list of cities and towns, or no information. Where they exist, address databases or listing are most often maintained by a country's postal authority or by a government agency.

That such a database exists does not necessarily mean it is available. Some countries restrict access to their address files for security reasons, limit their use to domestic or authorized agents, charge very high usage fees, or otherwise limit access. Restrictions due to privacy concerns also affect some files, particularly change of address files.

Where a complete database of all postal addresses is available, it is possible to compare a mailer's file to the country's address file and determine if the mailer's addresses match in the same way so that deduplication can be undertaken. This also establishes that the address is correct and deliverable, although it does not guarantee that the company or individual is at that address. For those countries with less information, deduplication and address verification efforts will be limited to verification that postal codes and towns or cities agree or are distinct.

The address-related services dependent on services and files provided by the postal authorities or a government agency within a country include

- address verification
- national change-of-address
- suppression of deceased, gone away (moved) and "Do Not Mail" addresses
- postal presorting.

Service Bureau Variation

Many service bureaus supplement the information available from postal authority or government sources. This creates variations in what each bureau provides. Each service bureau will also compile tables of common spelling errors, names, honorifics, etc. over time with new information added to improve and update the lists. These tables are proprietary and vary from one firm to another.

Those vendors of international address hygiene services who provide gender identification or spelling correction will have created their own lists or tables of honorifics, gender-based names, common misspellings, logical letter or character replacements, and so on for each language and writing system that they can process.

There will be greater differences among service providers when working with different writing systems (Chinese, Cyrillic, Greek, etc.) or with foreign-language spelling correction or gender identification based on their knowledge of the unique aspects of names and addresses created by linguistic and cultural differences. Determining your requirements is the best first step to finding the best vendor for your needs.