Understanding international data

Practical advice for retailers looking to go global

Independent White Paper

Commissioned by Postcode Anywhere

Author: Graham Rhind

PostcodeAnywhere

Foreword

Are you ready to go global?

The internet has created an international marketplace. In the UK, online sales generated from international markets are expected to soar sevenfold to £28 billion by 2020.

With over 130 different address formats in the world, 6,000 languages and innumerable accents, it's easy to see why poor international data is the guilty secret that continues to blight big business.

The problem is two-fold: alienated foreign customers are abandoning their shopping carts in droves, while the only quick fix for incorrectly captured address details is to pay third-party delivery firms an expensive re-routing fee.

CEOs and CIOs who want to operate successfully and maintain healthy margins need to find more efficient ways of collecting and holding accurate global data.

If we were starting from scratch, it would be a daunting task. Address formats are not only defined by geography; indigenous cultural and language differences are the barriers that restrict access to accurate and effective international address capture

> **Every lost parcel costs retailers money**

Fortunately international addressing experts are approaching the problem pragmatically, so what was once a headache can now be simple to put right.

The challenges facing organizations who wish to embrace the opportunities of globalisation are explored and identified in this white paper.

Guy Mucklow CEO,
Postcode Anywhere

> We never have time to do it properly the first time round. We always seem to find time to do it twice.

Poor Quality Data: The Pandemic Problem that Needs Addressing

Graham Rhind

It is rarely the case that a database contains addresses from only one country – globalisation of markets and data gathering via the internet means any company can expect to be collecting data from almost any corner of the world.

We tend to regard moving our business interests over borders as the process of going global. If done properly, it is, in fact, a process of going local. "Web globalisation is personalisation on a global scale" (MarketingSherpa, 2007). If done properly, globalisation takes account of language, culture, demographics and geography to become personalisation outside of our national boundaries. To this list I would add law, which is an important definer in some cases concerning language and personal name use.

The impacts of poor international data quality are the same as for those of poor national data quality: difficulty in using the data to generate campaign lists, with resulting weak response rates; inability to match and integrate data from other data sources; inability to report and draw accurate conclusions from the data; problematic creation of single views for each customer; fragmented processes and inability to draw comparisons across national boundaries; poor use of marketing resources; the creation of barriers to best practice and, very importantly, the huge costs in time and money that poor quality data brings with it. The Data Warehousing Institute (TDWI) reported that "poor quality data costs United States businesses a staggering $611 billion a year in postage, printing and staff overhead."

> Poor quality data costs US business $611 billion a year.

Studies consistently find that well over 95% of businesses agree that inaccurate data costs them money.

I contend that most of those who do not agree with that assertion are in denial.

Yet a company which is properly prepared for international data is a very rare beast. Reports show that a majority of respondents agree that companies overlook the complexities of managing international data, and about the same proportion fail to understand the impact of poor data. My own experience would suggest that most of the minority of companies who did not agree with these statements are themselves living in a state of ignorance – companies who think they manage international data well rarely do so. Consistently, around 90% of respondents rate data accuracy and address quality as challenges in international data management.

It is very quick and easy to set up an internet data collection page, for example, with very little thought about its structure and the requirements of the customers using it, and only be faced with resolving the data quality issues thrown up by it at a much later stage in the process, when not all quality issues can be resolved anymore.

Going global requires preparation. Thinking of the world as your market is easy. Acting to make it happen is another matter altogether, especially when it comes to your data. This white paper plots the philosophy to adopt to ensure that your move into global data is a success.

The most important preparation is to find out about global data. The world is a complex place. There are about 240 countries and territories, thousands of languages written in many different scripts, over 130 postal address formats and at least 40 personal name formats. Postal code systems don't just look different, they also work differently. Dates and numbers are written differently. The list goes on. Without the sturdy foundation of a good knowledge of variations in data globally, you will find yourself forever fighting data fires, and never achieving any level of data quality.

We are all affected by our cultural background, and considering global diversity is not at the top of our minds when we are planning international marketing operations! It is, however, one of the steps to ensuring high quality international data capture.

One of the ways to do this is to make no assumptions except that all assumptions are incorrect. All assumptions need to be checked. The assumption, for example, that exclusively Italian is spoken in Italy may seem quite correct, but there are 24 spoken languages in Italy, and at least two of these, German and French, have important legal and cultural weight in two provinces. Collecting data in Italy without knowing this or taking it into account will cause problems, both with your customers and with your data.

Make no assumptions except that all assumptions are incorrect.

1. Do not fall into the “repent at leisure” trap. Find out about the cultural and data norms of the countries for which you will be collection and managing data.
2. Be prepared to spend money at the collection stage – it saves more money in the longer term.
3. Make your data collection systems mirror the data formats of the country and language as closely as possible – this makes data entry comfortable and increases accuracy.
4. Use auto-complete or other address validation software at the data entry stage. There is no better way of collecting correct data than interacting with the source – your customer. Reducing the distance between your customer and their data increases data quality.
5. Never compromise on data quality. Make auditing your data a regular process – anomalies need to be identified and resolved at source before affecting overall data quality.
6. Don’t reinvent the wheel – make use of the resources available to you.
7. Be guided by common sense.

Companies who gather data via the internet often have only themselves to blame for the poor-quality data they obtain due to the lack of thought put into their input forms.

From “Practical International Data Management”

> **American databases often allow only 35 characters for a postal town name.**

You may decide that your database's field lengths are more than sufficient to hold any international data because it will hold any data you've ever needed it to nationally. Could it hold the 225 characters of the name of the Sultan of Brunei? Or the 163 characters of the full name of the Thai capital, Bangkok? American databases often allow only 35 characters for a postal town name, because this allows storage of all American place names, but used internationally it would mean 20% of Brazilian place names having to be truncated or abbreviated.

Knowing about the world does not entail that you must act on all of the knowledge, but it leaves you prepared for the issues which arise from going global. You may choose, for example, to have your web form support the address formats of only the most economically important countries, or, for countries with a number of minority languages, only to present your form in the main national languages. Provided you make these decisions on the basis of broad knowledge, you will be prepared for, and able to work around, any issues that it might produce in your data.

Related to the knowledge issue is the rule that it is better to collect good data than correct bad data. 90% of data management can be achieved using 10% of the total effort, and this effort is best concentrated at the data entry stage. Gender coding is a good illustration of this. Collecting names such as Joan, Jan, Nicola and José, all of which can be male of female according to culture, and then trying to gender code them after capture will always result in errors, poor data quality, and irritated customers.

> **Failure to understand that this world we live in is a mosaic of languages and cultures is costing data managers enormous amounts of frustration and money.**

It is better to collect good data than correct bad data.

The immediacy of, for example, the internet allows people to forget common sense and to start collecting data without giving any thought to the quality of the data being collected. Look at virtually any data entry screen or form, and you will notice that it is almost without exception based on the national data properties of the company concerned. If the company is in the United States, you will be asked for a state and a zip code (though some will think they are being fully international by asking for a "postal code" instead). Dutch forms will ask for a preposition (van de, de la) for your name.

French forms will ask you for the building number before the street name, whilst German forms will ask for it after, and also ask for your academic title, which is essential there as it forms part of the form of address. Many forms will make the postal code a required field, even though over 70 states and territories of the world still have no postal code system.

Given that we rarely use anything other than computers to collect data these days, and computers are a dynamic medium, it is a pity that data entry forms usually have a single structure, which the user will need to wrestle with to get their data into.

Field lengths often cause problems...

- The world's longest postal code is 10 digits long
- The world's longest settlement name is 85 characters long
- The world's longest personal name is over 100 characters long

The wrestling match with your form that the user has is as nothing to the wrestling the companies behind these forms will have to do to make sense of the data they have collected.

You should alter your data processes to fit cultural and linguistic norms, not vice versa.

Data entry forms should ideally match the patterns and norms of the place concerned. The fields should be presented in the order in which the user would expect to see them (especially important for personal names and postal addresses). Only relevant fields should be presented, and fields only made required when they can always be completed. This can be achieved by asking users for their country and preferred language and then presenting them with the data entry form which fits their requirements. This costs more time and money in the design stage, but saves much more in data cleansing after capture and, importantly, allows increased data quality.

Apart from the fact that it is never possible to produce as good quality data from batch cleaning after data entry than from data which is validated on entry, it costs many times more to clean data after the fact than to collect good data at source. Budgetary structures in many companies prefer to spend more when they can see the poor quality data which needs cleaning than to spend money on prevention, but when it comes to data quality, it is the most short-sighted approach. Many respondents to surveys cite lack of budget as a reason for data quality problems.

> **Many companies continue to underestimate the extent of the differences which exist between and within countries and cultures. Do not assume that everybody living in France speaks French.**

Den Haag, or 's-Gravenhage, are the local Dutch versions of the city name The Hague. Research on almost 1 million "cleaned" Dutch address records shows this city 's name written in 57 ways, 53 of them wrong (over 50% of all cases). Some of the ways in which the name was written would have defied the best batch cleansing tools to correct.

Knowing the complexity of international data collection and management, an important consideration is to not reinvent the wheel. Companies specialising in international data collection, cleansing, validation and management will always be better placed to achieve higher quality international data than any company starting from scratch. Making use of the tools and expertise available will always dramatically reduce learning curves, reduce errors and save money.

It is important to keep in mind that, though address formats are defined by geography, cultural and language variations are a function of person. You must never assume that a person in a country has that country's nationality or speaks that country's languages. Data collection forms that change language according to IP address or other pointer are a constant irritant to some customers and will significantly reduce response and data quality. Where possible, the user should be given a choice on the language in which any form is presented to them.

Naturally, the choice of languages for the forms will be based on commercial decisions, but also on the knowledge you have gleaned about the countries you are targeting.

Though going global is an inevitable step for most companies, and data can be collected quickly and easily, good international data collection is a challenge. To achieve the same quality of international data as you would expect from your national data, ensure that you are prepared for all the diversity you will encounter, and utilise available knowledge and tools to achieve top quality data collection. Only in this way will you be able to go local as you go global.

In Sweden, address order may be defined by whether the correspondence is private or may be opened by a secretary:

Private:
Sven Janssen
Widgets AB

May be opened by others:
Widgets AB
Sven Janssen

Appendix A: The Postal Code

Most database managers recognise the importance of the postal code. Postal codes are formed and formatted differently in each country. In some countries, they indicate an area as large as a municipality (for example, Belgium), in others they indicate areas as small as a group of houses or a single company (for instance, in The Netherlands, United Kingdom, the United States). Some countries have postal codes which cover a large range of possibilities, from several municipalities to a single company (Germany). Many countries do not have postal codes.

> **Some countries allow companies and individuals to have more than one postal code.**

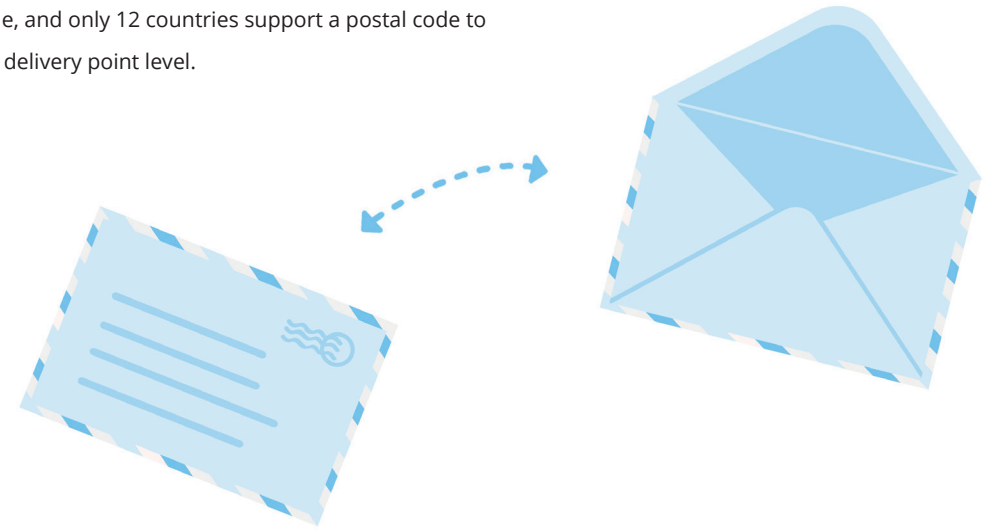
Some countries allow companies and individuals to have more than one code. The Netherlands, for example, allows two codes, one for the street address, the other for the mailing address. German companies have a third.

In some countries, there are certain letters or numbers which are known not to occur in any postal code. A common one is that 0 does not appear at the beginning of a numeric postal code. In Canada, the first digit of the postal code cannot be D, F, I, O, Q, U, W or Z, and so on.

Over seventy countries in the world have no postal code, and only 12 countries support a postal code to the delivery point level.

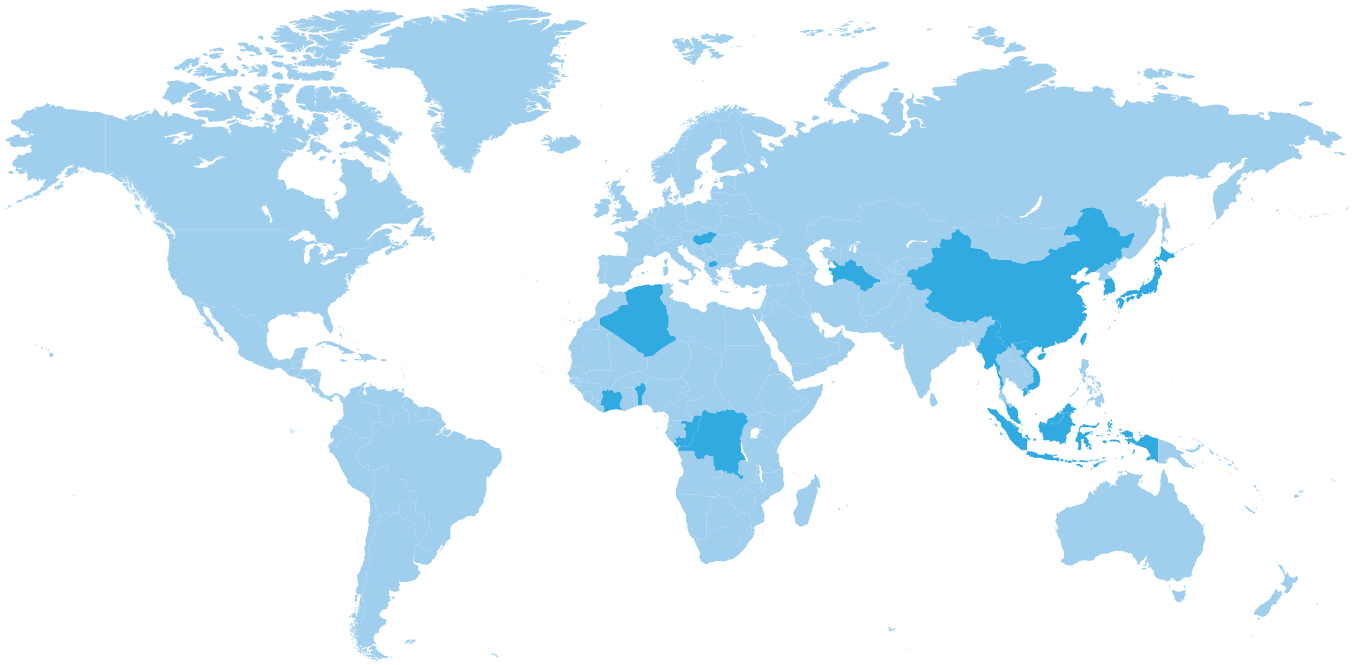
Worldwide address figures

- 40 different personal name formats
- 240 countries
- 130+ different address formats
- 6,000 languages

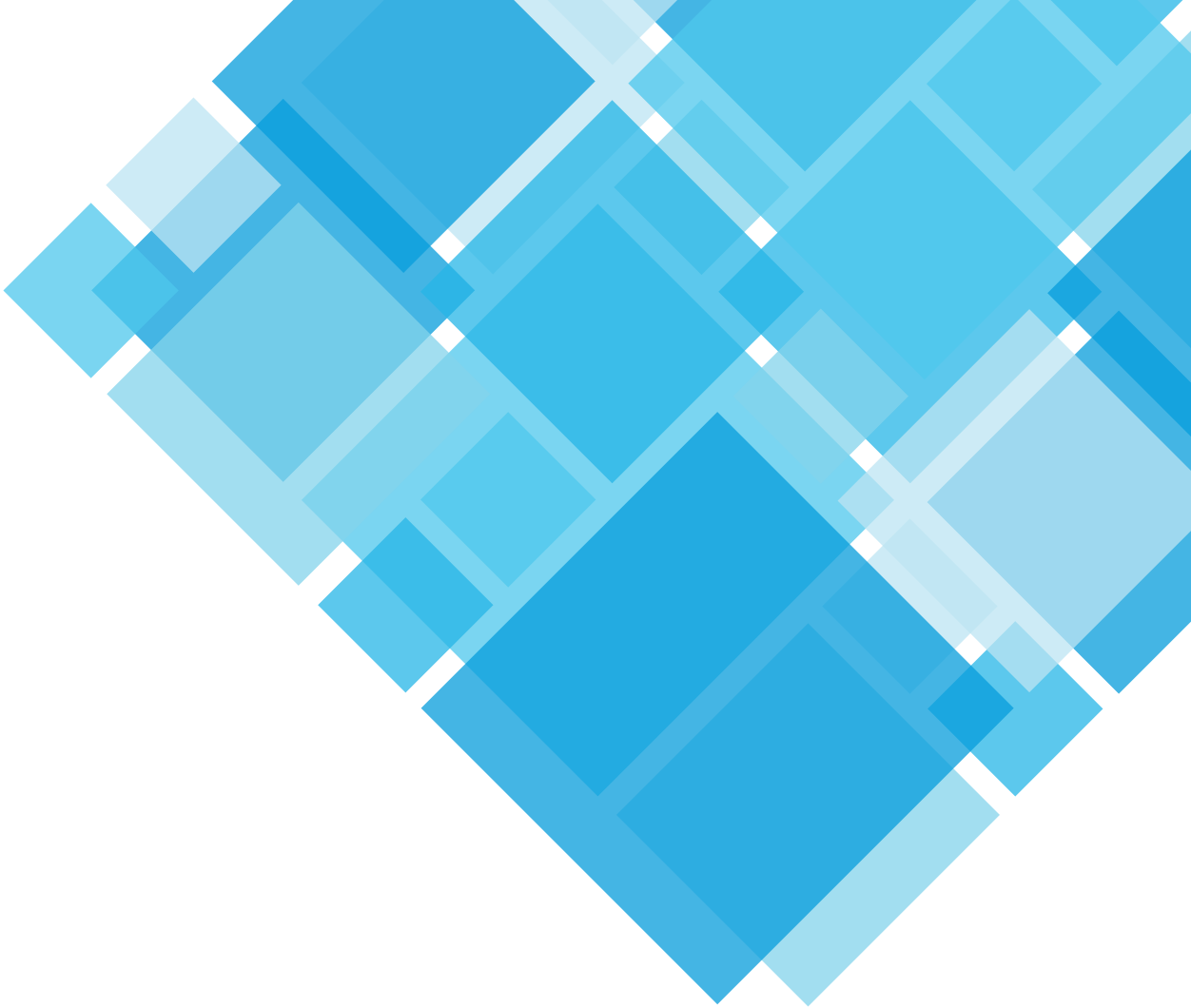


Appendix B: Countries where people usually write their names with the family name first

- Algeria
- Benin
- China (not Tibetan)
- Congo (Kinshasa)
- Hong Kong
- Hungary
- Indonesia (Chinese and Malay)
- Ivory Coast
- Japan (Japanese style)
- Macao
- Macedonia
- Malaysia (Chinese and Malay)
- Myanmar
- Singapore (Chinese and non-Muslim Malays)
- South Korea
- Taiwan
- Turkmenistan
- Vietnam



There are pitfalls which you need to consider when collecting international data... these pitfalls need to be addressed before the event, rather than afterwards when correction can be horrendously expensive.



Works Cited

MarketingSherpa / Byte Level. (2007). Web Globalization Report 2007. USA.
MarketingSherpa Inc. / Byte Level Research LLC.
Graham Rhind / Practical International Data Management (2001).

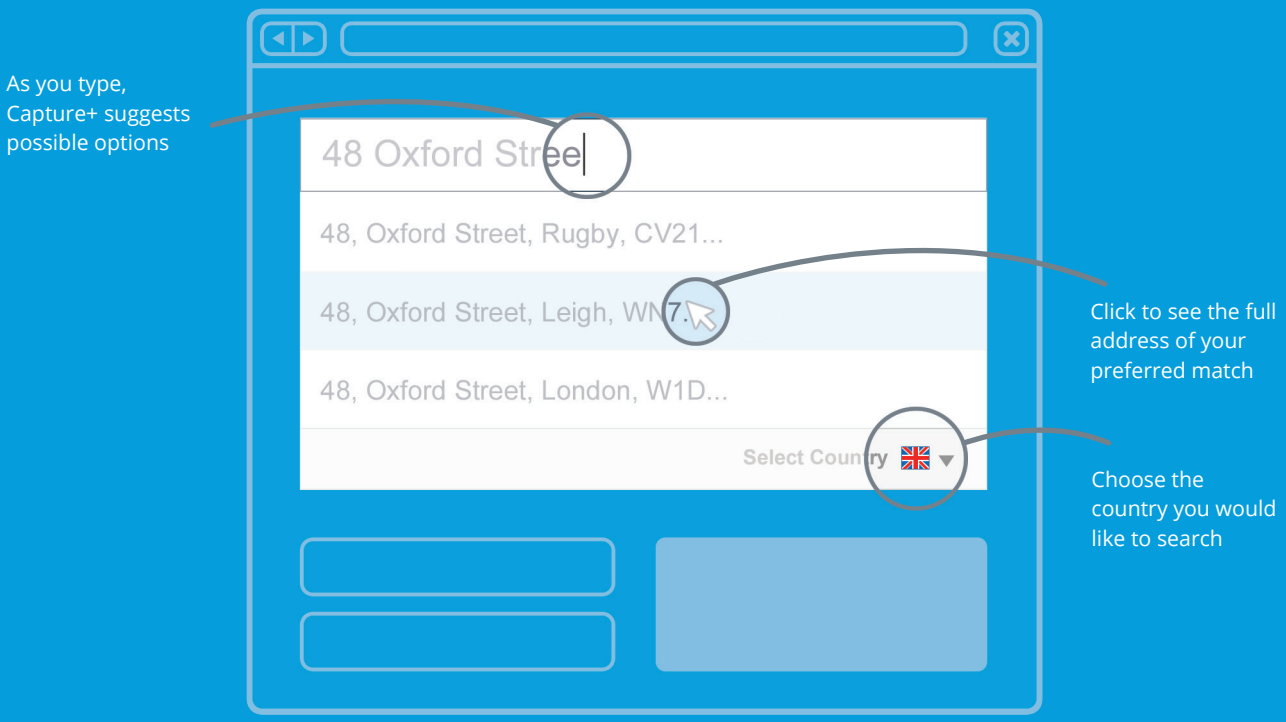
About the Author:

Graham Rhind is an acknowledged expert in the field of data quality. He runs his own consultancy company, GRC Database Information, based in Germany, where he researches postal code and addressing systems, collates international data, runs a busy postal link website and writes data management software. You can find him on Twitter via @grahamrhind.

Delivering internationally?

Make it easy for people to enter UK and international addresses with Capture+

How it works



✓ on web forms ✓ at checkouts ✓ in CRM systems

To find out more visit www.postcodeanywhere.com

UK | 0800 047 0495 U.S. | 1-866-838-9075

Rest of the World | +44 1905 888 550

theteam@postcodeanywhere.com

PostcodeAnywhere