Poor quality data
the pandemic problem
that needs addressing

Independent White Paper
Commissioned by Postcode Anywhere
Author: Graham Rhind

Foreword

BACK TO BASICS

The Internet has had a major impact on global business as organisations, regardless of size, potentially have access to a universal customer-base. As technologies develop at a rapid pace, the most important tool deployed by any established marketing department is still the data base.

However creative, strategically planned and implemented your marketing campaign is, its success will be reliant on the accuracy of the data used. Postcode technology has now become the tried and tested instrument of the marketer in the UK. The problem still remains that this prescriptive formula has not been so successfully deployed in other countries. With over 10,000 languages in the world, 125 different address formats and countless accents it is little wonder that access to accurate International data continues to blight big business.

Companies who wish to operate successfully and secure a slice of the world-wide pie need to find more efficient ways of collecting and holding accurate global data. There are a number of challenges, explored and identified in this White Paper, facing those organisations wishing to embrace the opportunities that the trend towards globalisation holds.

Whilst address formats are defined by geography; indigenous cultural and language differences are the barriers that can restrict access to accurate and effective international address capture. However prevention is better than cure and this document presents a textbook case that a "back to basics" approach i.e. the collection of good quality data at source, is still the best way forward.

Guy Mucklow
Managing Director, Postcode Anywhere

About Postcode Anywhere

Postcode Anywhere is the brand name for Fernhill Solutions Limited, a UK based software company which was set up by Jamie Turner and Guy Mucklow in 1999.

One of the key drivers for setting up Postcode Anywhere was the recognition by the founders of the potential of the internet in providing a much more efficient medium for delivering data. Compared with traditional addressing software applications which were expensive to produce, complicated to install and obsolete by the time they were delivered, the web services or data on demand approach, is both cost effective and easy to deploy whilst also having the added benefit of being centrally managed, thus removing the headache of updates for the database administrator.

Starting out with UK address management solutions, Postcode Anywhere has now branched out to provide one of the most comprehensive selections of web services available in the market; ranging from international address capture, geocoding solutions, bank and lifestyle data as well as maps and route planning.

The company, is one of the leading web service providers, with over 5,000 customers worldwide ranging from Fortune 500 and FTSE businesses, Government bodies through to "one man bands". Its technology is universally recognised as leading edge, a fact acknowledged by the growing list of national and international awards that it has won in recent years.

Poor quality data the pandemic problem that needs addressing

For any company with a sizeable database, the requirement to manage and process non-UK data will be there. Even if that database contains only UK addresses, it will surely contain data such as names which originate in another culture. Furthermore, it is rarely the case that a database contains only UK addresses – globalisation of markets and, importantly, of data gathering means, such as via the Internet, means that any company can expect to be collecting data from almost any corner of the world.

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

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 / Byte Level, 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.

Companies collecting and holding global data need to find ways to manage it which meets many challenges, which is easy to deploy and is cost effective.

The impacts of poor international data quality are the same as for those of poor quality national data: 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 in 2002 that "poor quality data costs United States' businesses a staggering USD 611 billion a year in postage, printing and staff overhead".

Studies of UK businesses find consistently that well over 95% of respondents agree that inaccurate data costs their company 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 any more.

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, some 6000 languages written in tens of different scripts, over 120 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.

Going Global

Language

Culture & Law

Demographics

Geography

Going Local

"**Find out about global 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

Equally. 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 able to be prepared for, and 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.

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 that 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 60 states and territories of the world still have no postal code system.

> "It is better to collect good data than correct bad data"

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. 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.

> "Alter your data processes to fit cultural and linguistic norms, not vice versa"

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.

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 targetting.

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

Works Cited

MarketingSherpa / Byte Level. (2007). Web Globalization Report 2007. USA: MarketingSherpa Inc. / Byte Level Research LLC.

About the author:

Graham Rhind is an acknowledged expert in the field of data management He runs his own data consultancy company, GRC Database Information, in The Netherlands, http://www.grcdi.nl where he researches postal code and addressing systems, collates international data, runs a busy postal link website and writes data management software. Graham also regularly speaks on the subject and is the author of Building and Maintaining a European Direct Marketing Database, The Global Sourcebook of Address Data Management and Practical International Data Management - a guide to working with global names and addresses.

PostcodeAnywhere

Enigma House, Elgar Business Centre
Moseley Road, Hallow, Worcester, WR2 6NJ
Tel: 0870 241 3310

www.postcodeanywhere.co.uk