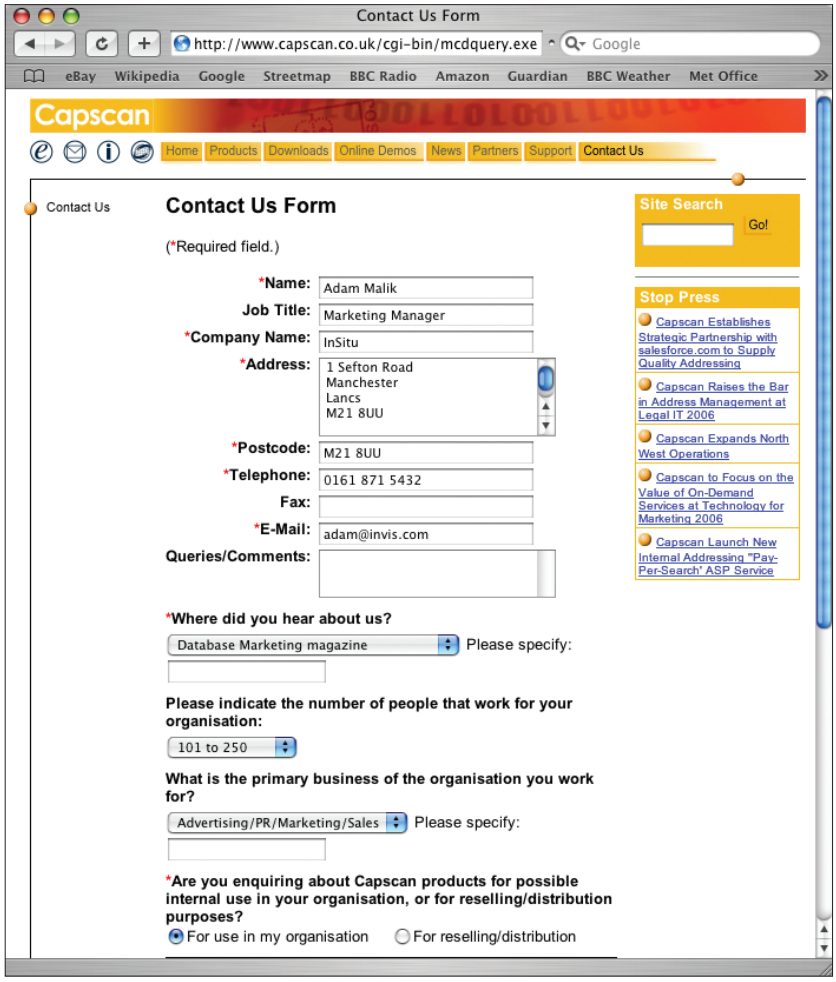
NOTEBOOK

Garbage in, garbage out

Graham Rhind discusses how to reduce the amount of laborious post-processing when collecting international customer data online.

Though the web was once heralded as a great way to collect customer data cheaply and effectively, experience has shown that this data is often too polluted to be useful in any business application, let alone analysis. This is not due to the medium, but to the poor understanding that most companies have of how to achieve quality data collection on the web without expensive data re-engineering.



When collecting data on the web, companies must allow diverse visitors to record their information in a way that is familiar and comfortable to them.

Quick and dirty

The path normally followed by companies when choosing to collect customer data on the web is defined by the company decision-making structure and general ignorance of global diversity. Inevitably this dictates that scrap and rework will be needed on the data collected. The path normally looks like this.

When someone decides to collect customer information on the web, they usually take a short-term view of how to achieve this. A web data collection form can be up and running within a few hours. It does not usually require any special budget and is usually a response to pressure from other company departments to get the data as quickly and cheaply as possible. Most company structures militate against making money available to research and implement good data collection practices at the start of the process.

Little or no thought is given to this data collection page. The employees concerned stick to what they are familiar with. They use the same fields, the same field labels and the same screen layout that they know from their own country. Somehow, they forget that a web page can be viewed from any place in the world, and that people from outside the company's home country are likely to want to enter their details too. Similarly, they forget that these visitors have personal details that do not fit with the local norms.

Ambiguous or country or language biased field labels will mean different things to different site visitors, causing them to provide different information based upon their interpretation. For example, a field entitled "Title", even within the UK, might be interpreted to require either a form of address (Mr, Mrs etc.) or a job title. A response from Germany, on the other hand, is likely to regard this as being the place to enter an academic title.

Equally, using the labels "First name" and "Last name", intended to collect given name and family name, will collect this information in reverse from most people, as the majority of the world's population write their names in a different order to us in the UK. The large numbers of people worldwide without a family name will be at a loss as to what to fill in.

Using labels such as "prefix" and "suffix" will, again, collect different pieces of information as visitors to the site write their personal information in a different order to UK visitors. Their information may be too long to fit into the given fields; and required fields, for state or postal code, for example, will force them to enter nonsense information if their addresses do not contain such details. They may have more information than they can fit comfortably in the company's web form. Because of this, they are required to shoehorn their data into the available space.

Continuous flow

Data is collected, but it arrives in the database confused, concatenated, abbreviated, mis-fielded and completely useless. As the quantity of the data increases and its potential value is appreciated, it then becomes clear that it needs a major cleansing

program to use it effectively. The next step? Buy some expensive software.

This costs much more than creating a good data entry system at the start would have done. Choosing a better data structure will be one of the first tasks, though again this would have been better tackled before any data gathering began. In fact, the best way of getting top quality information from a customer is interacting with him or her at the time of data collection. This is true regardless of how expensive the software is, how many hours of labour are put into the process and how many processes are run.

Next, the data is processed and a certain percentage is improved, but the stream of poor data from the initial collection point continues. Thus, the data is assessed, scrapped and reworked as a continual process.

Companies do not often reach the end of this road. Data remains of poor quality, with the resultant business process failures when the data (or information from that data) is used. Although these results are clearly not effective, almost all companies have followed this cycle. Not only does this result in bad data and its consequences (like poor brand image), but can also mean an image and morale problem within the company.

The data is not regarded as accurate, and is therefore under-used. Budget is difficult to pin down to correct the problem because people are not confident about the outcome, and expensive processes do not show enough improvement to increase confidence. As people consider what has been spent already, they are reluctant to spend more.

The budgeting for web data collection should be moved to the beginning of the process, which is the design and execution of data collection processes and applications. With a small amount of investment and research, higher quality data can be collected from website visitors. Data collection pages, which dynamically alter form structure, order and language to the country and language of the visitor, allow visitors to record their information in a way that is familiar and comfortable to them.

Field labels and lengths can be adjusted; and validation, both full postal and individual component validation, such as postal code length, can be implemented to reduce data pollution as much as possible. Getting it right in the first place is the only way that data can be collected on the web accurately enough to be properly useful for business intelligence, without expensive – and often pointless – post-processing. ■

Getting it right in the first place is the only way that international data can be collected on the web accurately enough to be properly useful

Graham Rhind is an acknowledged expert in the field of data management and runs his own consultancy company, GRC Database Information, based in The Netherlands. (http://www.grcdi.nl, graham@grcdi.nl)