Reducing the need for scrap and rework with web data collection

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.

Though the Internet was once heralded as a solution for enabling cheap and effective data collection, experience has shown that this data is often too polluted to be useful in any business intelligence sense. 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 scrap and rework.

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

When a decision is made to collect customer information on the web, the short-term view of how to achieve this is usually chosen. A web data collection form can be up and running within a few hours. It does not usually require any special budgetary measures and answers the pressures from other company departments to get the data as quickly and cheaply as possible. Normal company structures militate against budget being made 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 with what they are familiar. They use the same fields, the same field labels and the same screen layout that they know from their own country. It is overlooked 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, it is overlooked that these visitors have personal details that do not coincide to 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" may be filled in by one visitor with a form of address, by another with a job title and by yet a third with an academic title. In other countries, not only do people's name and address details consist of different components, they are also written in different ways. Their information may be too long to fit into the given fields; and required fields, for state or postal code, for example, will require 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.

Data is collected, but it arrives in the database confused, concatenated, abbreviated, mis-fielded and completely useless. As the quantity of the data increases, it becomes clear that a major cleansing program is necessary to effectively use the data. Expensive software is acquired. This costs much more than it would have to create a good data entry system at the beginning. A better data structure must be identified, though 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.

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

Companies do not often reach the end point of this path. 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, this cycle has been followed by almost all companies. Not only does this result in bad data and its consequences (like poor customer image), but also an image and morale problem within the company. The data is not regarded as accurate, and is therefore underutilized. 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. This is the only way that data can be collected on the web accurately enough to be fully used for business intelligence, without expensive scrap and rework.

Graham Rhind is an acknowledged expert in the field of data management. He runs his own consultancy company, GRC Database Information, based in The Netherlands, where he researches postal code and addressing systems, collates international data, runs a busy postal link website and writes data management software. Graham speaks regularly on the subject and is the author three books on the topic of international data management.
http://www.grcdi.nl
graham@grcdi.nl