CASE STUDY

Education Sector Client

Client Requirement:

The client received company information in a daily feed from over 100 different suppliers, to a total of over 350,000 business records within a 3-month period, which included company name and postcode (no other address attributes), and a rough, free-text description of what the company did. The feeds contained multiple non-consistent records per company; around 90% of the businesses were British, the rest being international.

The client required a 3rd party to tidy and enhance this data on a daily basis, so that at the end of the 3 month period they would have one unified database containing consistent business information.

Major requirements consisted of:

- Standard Industry Classification (SIC) codes appending
- Company names standardized to remove typing errors and allow for analysis by company
- Employee figures adding
- Flagging of private and public sector companies
- Flagging of charities
- Each day's data to be returned within 4 working days, extended to 6 days during peak flow
- Postcode verification and correction
- The addition of other elements such as website address and Company Registration Number
- Consistency across records for the same company

Budget was limited, time was tight and there was a need for a high degree of accuracy and for as many of the data fields as possible to be populated. The number of records received each day varied, with an expected peak of around 20,000 records a day, so flexibility to deal with a sudden influx was required.

Oblong Solution:

In order to send and receive xml data on a daily basis we created a secure web interface, with the client systems notifying ours when a file was ready to collect and vice versa.

We developed a bespoke automated system to process as much of the data as possible. As this project was to be performed on an annual basis (350,000 plus business records a year with a potential large overlap in companies reported) we created our system to learn from previous data, to recognise previously coded companies and descriptors, and where possible to re-use previous attributes for the same companies.

While the project was live, we processed the incoming client data on a daily basis as follows:

Using our Unity matching software we matched each incoming company name and postcode to our internal business universe and to previously cleaned data from this project, and appended SIC codes and other attributes.

To further populate SIC codes and fill any gaps we used our AutoSIC software which recognises keywords in company names, or the free-text descriptions of what the company did.

Finally, we used the list of previously processed descriptions of what companies do and their given SIC to see if we could find a similar description and append SIC codes accordingly.

Records which could not be automatically SIC coded using any of the above processes were manually reviewed, which included much of the international data. We would review the descriptor and if necessary search for the company online in order to SIC code it.

In order to ensure accuracy, every record was manually checked at least once before it left the building. The vast majority of results were returned to the client within 24 hours of being received.

The client had no control over the flow of the data coming in from the different suppliers, and we found that we reached peaks of over 50,000 records per day, at which time we were still able to return the data within the deadline agreed.

Once all of the data had been received, we then ran it through various processes in order to ensure consistency across records. There were often multiple records for the same company, but as each record had been received on different days, from different suppliers, with different free-text descriptions of what they did and, frequently, incorrect or missing postcodes, there was scope for inconsistency. Therefore we not only subjected the final data to further processing to improve on this, but we also scrutinised the data manually to ensure the correct information was returned.

We automatically linked together records for the same company and then checked these links manually to ensure they were correct. These links allowed us to keep data consistent for the same company. We then manually checked the standard names, SIC codes, public/private flags, charity flags and employee figures, ensuring accuracy, consistency and especially that any large household-name company had all relevant information included.

Additional databases were brought in to create the flags required for public companies and charities, although many companies registered as charities are not necessarily doing charitable work as their main function, therefore these also needed to be manually vetted.

We looked at each and every record multiple times to ensure that the data was accurate and consistent.

Other data was appended at no additional cost, including website address and Companies House Registration Number.

Because each year we use the previous year's results to improve the systems, over time we reduce the need for manual work, therefore making the project more efficient and keeping costs down.