

Game, dataset and fuzzy match: making the most of multiple datasets

Contributed by ORC International
Wednesday, 05 March 2008
Last Updated Monday, 14 July 2008

3Â March 2008

Datasets held by public sector organisations are a hot topic. This article describes how "fuzzy matching"â€™ data within and across databases can help the public sector to produce essential evaluative, audit and management information.

Since 1998 a large team at ORC International has been working with the Department for Work and Pensions on individual level data to help reduce the number of long-term unemployed.

At the start, there was just a single database for recording details of all 20,000 participants in the Governmentâ€™s New Deal for Young People programme. Now, there is an over-arching database for all clients recorded on New Deal, Jobcentre and Jobcentre Plus computer systems. The team at ORC International now manage and apply regular updates to 40 databases storing millions of records accumulated since 1998 and deliver 35-40 cumulative data extracts a month using the statistical package SAS.

ORC International has the technology to cross-reference records of individuals with data from, for example, job seeking activity, benefit claim records and other programme-specific information. The technique is called "fuzzy matching".

Fuzzy matching

Fuzzy matching techniques enable the linkage of multiple records for an individual across and within data sets. This eliminates errors by finding duplications and inconsistencies such as variations in the spellings of names, more than one National Insurance number due to the use of temporary numbers, or minor data entry mistakes.

The use of geographical codes then helps take the fuzzy matching process to a more detailed level of analysis. Jobseekersâ€™ postcodes are regularly mapped against a range of geographical variables including local authority district, electoral ward, parliamentary constituency, "travel to work" area and government regional office.

Collecting data from bespoke web sites

Some data on service provision is collected using bespoke websites " a key feature of the project since 2002. Initially, ORC International developed a website to capture the details for those attending European Social Fund Provision. Over the years, other bespoke web tools have been added. Examples include Employment Retention and Advancement, Employment Zones, Progress to Work and Incapacity Benefit. Adaptability is an essential characteristic of ORC Internationalâ€™s data systems, for example, random allocation techniques have been applied to enable the assignment of new clients to programme and control groups.

Future developments

ORC International looks forward to helping with what are growing data challenges facing the public sector. We are well-placed to provide technological and research solutions for data security issues and do not predict that these important concerns will slow down the trend of ever more organisations seeking to improve their effectiveness through joined-up data systems.