Matching Records Without a Common Identifier - The UK Experience

Steven VALE and Mike VILLARS
Office for National Statistics, Cardiff Road, Newport, NP9 1XG, UK
E-mail: steven.vale@cec.eu.int ; mike.villars@ons.gov.uk

Abstract: The UK does not yet have the benefit of a common identification number for businesses, therefore the matching of records from the different administrative sources used for the ONS business register requires the use of other fields, such as name, address and post code. This paper describes how the ONS has tackled this problem, the issues faced, the software used, and some ideas concerning future enhancements.

Keywords: matching, matching software, namekey

1. Introduction

The Inter-Departmental Business Register (IDBR) is the basis for a wide range of statistical business surveys in the UK. It is maintained by the Office for National Statistics (ONS). The IDBR uses data from a variety of administrative and statistical sources, the most important of which are Value Added Tax (VAT) records and Pay As You Earn (PAYE) income tax records.

All UK businesses with an annual turnover of £51,000 or more are required to register for VAT purposes, similarly all businesses with one or more employees are required to register for PAYE purposes. There is, therefore, a considerable overlap between the records held by these two sources, and thus considerable potential for duplication of records on the IDBR.

The IDBR is updated from the VAT register on a weekly basis and from the PAYE register on a quarterly basis. To minimise duplication it is essential to check that new units, or "births", from these sources are genuine, and that they haven't already been added from another source. The VAT and PAYE records each have their own system of unit identification numbers, and the definitions of units vary, reflecting the different needs of the two systems. This makes the matching of units from these sources a vital but rather complex task.

---

1 Steven Vale is currently on a 3 year secondment from the ONS to Eurostat
2 There are a few exceptions to this rule, mainly concerning businesses operating in VAT exempt industries
2. The matching software

The software tools used for record matching are SSA-NAME3 and SSA-EXTENSIONS produced by Search Software America (SSA). SSA-NAME3 is the foundation and includes fundamental key and search algorithms. SSA-EXTENSIONS includes “Scoring” and “Compound Name” routines.

SSA-NAME3 contains a key building algorithm designed to resolve various problems encountered in name based matching including:

- Differences in names due to spelling and phonetic variation, handwriting and transcription errors, reading and keying errors.
- Problems created by the high volume of common words in names and especially the large groups of people with a few very common family names.
- The mixed use of words, initials and abbreviations in names.
- Use of “nick-names”, anglicisation of names, etc.

Input files are processed in four phases:

CLEANING - This routine edits the name string, removing special characters and replacing lower-case with upper-case.

FORMATTING - This routine edits the name string into separate words, removing “stop words”; replacing selected words and concatenating prefix words.

STANDARDISATION - This routine “standardises” the name, for example removing double characters.

KEY GENERATION - This builds storage key and a table of key ranges to be used in database searches.

SSA-EXTENSIONS is a collection of routines that enhance the SSA-NAME3 algorithms, including:

- Generation routines for ‘scoring’ or ‘matching’ algorithms to compare records.
- Compound name routines where there is a need to recognise multiple key generation because the name contains more than one entry, i.e., “John Smith and James Dean”.


SSA-NAME3 uses various definition files to control its behaviour. Many features of SSA-NAME3 are driven from tables that are user-defined. The definition of these
Matching Records Without a Common Identifier - The UK Experience

tables is in the form of high level descriptions contained in a definition file, following a syntax defined by SSA. This definition file is passed through a process known as generation to produce a module in a language suitable for the target system, i.e., ‘C’ for the IDBR. These modules include:

- **Edit List** - Using knowledge of the data held to build the ‘edit rules’. It should be a representative file of the raw input data used in the application and contain unedited names including titles, suffixes and company names such as ‘PLC’, ‘INC’, ‘SPA’, etc.

- **Frequency Tables** - Once the edit-list exists a frequency analysis can be run on the formatted names from a representative file of names. This produces a frequency table for the population. It consists largely of tables used to control the key building and the search logic.

- **Scoring Schemes** - The scoring functions of SSA-NAME3 are implemented by the application programmers/analysts defining scoring schemes. They describe the parameters to be used when scoring two records.

- **Service group** - Several SSA-NAME3 algorithms may be grouped together with a scoring control table to form a single module or program known as a Service Group.

- **Authorisation Module** - The authorisation module holds the signature of each module in the current generation. At run-time these are tested against the actual modules being used. If a mismatch occurs an error is produced to prevent the accidental use of a formatting routine different to that used when loading SSA namekeys into the database.

### 4. Generation of the SSANAME3 table on the IDBR.

The name and trading style for each address record held on the IDBR was put through the SSA-NAME3 algorithm to bring back namekeys for each word in the name/trading style. This can bring back up to 20 different keys for each record. The resulting table is used as the target for all possible matching from any input source.

At present there are around 20.7 million rows held on SSANAME3 table created from 6.8 million rows on the address table.

### 5. Applications in use on the IDBR

Applications have been set up to provide a name matching service for:

- VAT births.
- PAYE births.
- On-line name matching

335
A standard name-matching environment for all ad-hoc requests for name matching.

The VAT births, PAYE births and the standard system all use a similar matching routine in a batch environment described below, whereas the on-line matching is screen based.

6. Batch Processing

To illustrate the batch processing system we will follow the example of a typical quarterly update of PAYE data. A table of births is obtained as the input from the quarterly PAYE update file. This holds the unit reference, name, trading style, address and postcode. The name field is used to generate a range of SSA namekeys. From this range the key produced from the right most name will be used as the major key in name matching.

By inputting Steven Vale to obtain SSAname keys for all the names it will produce:

```
  NAME:  STEPHEN     STAFAN
         VALE        VAL

  KEYS:
    1  XJXMS$$-
    2  YLVO$$$
```

YLVO$$

relates to the Surname (major name) and will be used in the initial SSA matching.

The major name key (or stab key) is used to check against a table of namekeys generated from the names of each record held on the address table on the IDBR, to provide a count of expected matches. It will start by using the 1st range. If the count is zero it will move down to the 2nd, 3rd range etc. Each level of range will increase the possibility of finding matches. This is done by changing the depth at which the original range of keys is obtained to get a wider spread of keys.

```
  STAB:  YLVO$$$
         1  YLVO$$ YLVRZZZZ
         2  YLTO$$ YLVZZZZ
         3  YLS$$ YLZZZZZ
         4  YK$$ YPZZZZZ
```

If the count is > 0 and < 200 then these matches are treated as good matches. If the count is > 199 and the employment is < 10, no further action is taken. If the employment is > 9 then all the matches found are treated as good matches.

For the good matches the input name, address and postcode are compared with the name, address and postcode of each of the good matches to produce 3 scores (out of 100) which are added together and divided by 3 to create a total score. If the total score
is >79 then this is determined to be a definite match. If the score is between 60 and 79 then this is determined to be a possible match.\footnote{Lower values are used for corporate units, as they are required by law to have unique names} Any other score is deemed to be too small and is regarded as a non-match.

Duplicates on the definite match list are then removed, as well as records on the possible match list which also appear on the definite match list.

The records on the definite match list are then linked automatically to their corresponding units on the IDBR. The records on the possible match list, and larger non-match records are reported for clerical checking. Those that are not matched clerically can, if they meet certain criteria, be sent a Business Register Inquiry form to confirm their details.

For a typical PAYE update (40,000 to 50,000 new records), around 37% of records are definite matches and 35% are possible matches (of which approx. 80% can be matched clerically).

7. On-line name matching

The on-line name matching follows a similar process;

The operator enters a name on the screen and it is passed to SSA-N\textsc{ame3}.

SSA-N\textsc{ame3} returns a search table containing the start and end value for a key sequential access for each possible depth of search.

The application program chooses the first depth and reads the set of records between the Start and End key for that depth.

A record or records are returned to the application.

The application program displays all records and if necessary it progresses through the depths until the user is satisfied.

There are a couple of limits imposed by the IDBR on-line matching. The first is that the depth cannot go beyond 80. The range of namekeys is too wide to be of use at this depth. There is also a maximum limit to the number of rows that can be returned. The maximum limit can be raised from its default of 200 rows to 999 rows if required.

8. Computer system details

The computer is a NUMA machine running 8 Pentium processors. SSA software is written mostly in C. We have developed this to run in a UNIX (Dynix) environment where the SSA programs are called from an Ingres application.
9. Some problems encountered

Major problems were mostly sorted out during the original setting up of the system where we were guided by Search Software America. An example of the sort of problem found since is the use of “Trading as” or “T/A” in the name e.g. Mike Villars T/A Mike’s Coffee Bar. This caused problems because although keys were created for all the words in the name, Bar would be used as the major key. This would bring back too many choices as there are thousands of bars in the UK. Therefore we split the name so that the last word prior to “T/A” i.e. Villars is the major key. This increases the quality of matches.

A different type of problem that is not so easy to solve is the gradual growth over time of the number of small unmatched units, leading to increasing duplication on the IDBR. Checking such units clerically is very labour intensive. Various possible solutions have been tried, including fine tuning of the matching parameters, re-running the matching process at a different point in time to compensate for time lags, and using extra information, such as the company number, where available. These have all met with some success, particularly the latter, which greatly reduced the number of corporate non-matches, but new ways are constantly being sought to reduce this problem. A register quality project supported by funding from Eurostat is currently looking at this problem in detail.

10. Future developments

We are constantly looking at ways to obtain better matches. One possibility currently being tested is the use of a new software package QuickAddress, based on the Post Office’s Address File (PAF). This is used to improve the quality of addresses in the input files, reducing the non-match rates by 2-3%. Other possible links to geographical referencing systems are also being explored.

A project to develop a single business register in the UK is currently being considered. If approved, this will provide a single point of registration, and will hold basic information about all UK businesses. This basic information, including a standard identifier, will then be passed to the various administrative and statistical bodies within government. This will therefore have a considerable impact on the process of matching records, solving some problems, but also introducing new ones.