

SEAMLESS

**Address Matching and De-duplication
software.**



White Paper

“ Why do we need data matching software ? ”

At any time 15% of the data in a typical customer database is inaccurate.

CRM implementations and data warehouses rely on quality data to give accurate results. Without this quality the benefits of expensive software cannot be realised. However only a minority of companies actively pursues data quality initiatives. The opportunity exists therefore to gain a competitive edge by taking advantage of tools designed to improve data quality.



Some of the most common reasons for requiring a data matching solution would be:

- In order to de-duplicate addresses, which can save a company involved with mail shots large amounts of time, money and embarrassment.
- To combine information from various sources, to allow a company to better understand its own business and enable detailed analysis through the use of data marts and warehouses.
- To combine information about customers and to quickly identify them, a critical success factor in the implementation of any CRM strategy.

SEAMLESS Success Story

In a recent implementation of *SEAMLESS* a customer database of over 1.6 million records was compared against a *PAF database resulting in a hit rate of 96%.

* PAF = Post Office Address Format, the UK definitive guide to names and addresses.

“ Why is it called *SEAMLESS* ? ”

It is an acronym, which stands for **Server Embedded Address Matching and Location Enhancement Software Suite**.

Server Embedded ...

Typically it is implemented as a set of database procedures which means it can be built into your application, potentially reducing the need for costly and problematic interfaces between systems.

Address Matching and

The main function of the software. Obviously it will match more than just addresses, in fact any text at all.

Location Enhancement ...

One of the options with **SEAMLESS** is to enrich customer details including addresses, phone numbers, email addresses etc. SEAMLESS matches against standard address data sources from most countries in the world.

Software Suite.

A set of database procedures, functions and techniques which combine to provide industry best results.

As the name suggests *SEAMLESS* has been designed to fit right in alongside your existing applications with a minimum of fuss. It can be set to work either in batch mode with little or no intervention, or built into an interactive form.



SEAMLESS Bureau Services

C&C Group also operate a data matching service bureau, which makes use of the full *SEAMLESS* technology. We can offer a full service where we extract your data and put it back again. Fast turn around guaranteed! Try us.

“ What problems can **SEAMLESS** overcome ? ”

SEAMLESS allows data from different sources to be easily merged into a single data set, without duplication.

Matching information can be difficult because of :

- Different storage formats
- Misspellings and transpositions
- Use of abbreviations and nicknames
- Global address changes
- Phonetic spellings
- Different levels of detail
- Different standards



SEAMLESS employs a series of sanitisation methods, which greatly improve our chances of finding matches.

- Format independence
- Name frequency histograms
- Nickname translations
- Abbreviation expansion
- Noise reduction
- PAF updates

A number of scoring techniques are then applied, from which a result matrix is derived. The results can be analysed and rules applied to determine the matches.

- Character Counting
- Consecutive Character Matching
- Phonetic Consecutive Character Matching
- Word Matching
- Phonetic Word Matching
- Weighted character counting (Scrabble method)

Together the scoring techniques give consistently better results than those achieved by employing a single method only.

SEAMLESS Implementations cover many sectors of EMEA Business

Transport * Utilities * Pharmaceuticals * Healthcare * Hi Tech

“ How does **SEAMLESS** handle nicknames ? ”

SEAMLESS is supplied with a supporting database containing an extensive set of equivalent names, abbreviations and shortforms. The supporting database has recently been extended to cover Irish names and more countries will be added shortly.

Before any matching techniques are applied - names and addresses are scanned for shortforms and nicknames. For instance Bob will be set to its equivalent Robert for the purpose of matching and Av. will be extended to Avenue.

Additionally a database of known names and frequencies has been built up which allows us to put a higher value on matches where the names involved are more rare.

The supporting database also contains a set of noise words, these are things which are of little value when matching and usually serve to make the matching process more difficult. Examples would be ‘Jr’ or ‘Esq’ at the end of someone’s name.

Standard sanitisation procedures are also incorporated into the **SEAMLESS** engine. These remove multiple spaces, excessive punctuation and converts the strings to be matched to the same case prior to comparison. All of these basic tasks are performed before any matching techniques are applied in order to increase the chances of success.



SEAMLESS - the matching tool of choice for data professionals

A major data supplier to the UK Pharmaceutical Industry has recently selected C&C Group and **SEAMLESS** as their external data matching partners.

“ Without getting too technical, how does it all work ? ”

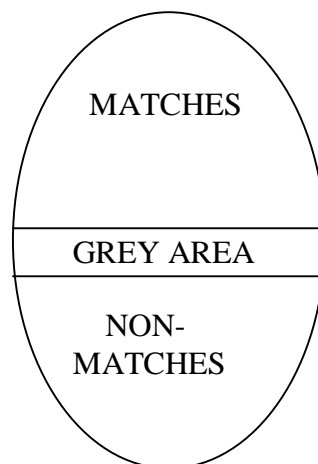
A function is called which has two input parameters. These input parameters are the two strings to be compared. The function returns a set of output parameters each of which represents the result of applying one of the matching techniques to the two input strings.

A result of 0 would indicate that there is no match at all between the two strings. A score of 100 would represent a perfect match between the two strings. In practice scores will be distributed over the entire range from 0 to 100.



This black box approach means that the *SEAMLESS* matching engine simply needs to be fed potential matches and returns a result array. All that remains to be done is to analyze the result array and apply rules to determine whether or not the potential match is indeed a match.

The rules can be determined by running a sample of potential matches through the *SEAMLESS* engine and analysing the results. There are two rules that need to be set, one identifies the point above which all comparisons are considered to be matches, and the other a point below which any comparisons are not matches. What is left will be a zone containing potential matches – the grey area, the validity of which can only really be confirmed one way or the other by human inspection.



SEAMLESS harnesses the power of the latest DBMS systems

SEAMLESS is optimised for use with Oracle 8i & Oracle 9i.

“ How do we sort through the grey area? ”

A simple helper application can help sift through the grey area. This is designed to look for records which cannot be automatically electronically matched and require the use of the human eyeball to make the final decision.

The candidate records are identified by analysing the scores and thresholds determined earlier and due to the unique scoring methods employed by SEAMLESS the most likely matches can be displayed first. This is the final part of the matching process and helps to squeeze out the last few extra percent of matches. The example below is a helper application that has been customised extensively.



Address Key	Address Type	Address Score	Paresa Score	Cares Score	Matched Method	Organisation Name
29425888	S	100	100	100	AUTO	MASON BROWN ASSOCIATES
23726784	S	71	71	40	79	SEAMLESS
28423288	S	62	63	22	13	SEAMLESS
21060488	S	62	62	22	13	SEAMLESS

SEAMLESS Plans and development

SEAMLESS is constantly being improved upon. The supporting databases are being enlarged to incorporate new countries and research is underway into new matching techniques.

“ How do the matching techniques differ ? ”

SEAMLESS incorporates many matching techniques, each of which contributes a score to the final result matrix.

The types of techniques currently applied include :

Character Counting

This gives us a basic count of the number of characters in the address. A recent enhancement was to extend this to a weighted character count method, based on tile scoring analogous to the Scrabble method.



Consecutive Character Matching

Gives us an indication of the occurrence of groups of characters in each address. This is a configurable number but tests have shown that comparing each group of four characters gives the best results for normal data sizes, and groups of three characters work better when the data size is small (e.g. names).

Phonetic Consecutive Character Matching

This works the same way as the consecutive character matching described above, but applies a phonetic algorithm to each resulting character group. The phonetic matching therefore scores highly on similar sounding character groups, this allows for common misspellings.

Word Matching

Word matching gives a score based on the number of words that match across the two strings.

Phonetic Word Matching

Phonetic word matching gives a score based on the number of words that sound the same across the two strings.

SEAMLESS Customisation Options

SEAMLESS can be easily optimised for specific applications by extending the supporting database through a simple form interface.

“ What makes *SEAMLESS* different from other products? ”

We believe that the matching techniques which have been developed by C&C Group represent an original and unique approach to the problem of data matching, and that the results speak for themselves.

The product has been developed as a result of our experience in data matching, always with the aim of replicating the intuitive way that the human brain recognizes that two pieces of information are related whilst harnessing the processing power available with modern day computers. The result is a highly powerful, accurate and cost effective matching tool that can be embedded *seamlessly* into your applications.



SEAMLESS - Want further information?

Contact C&C Group at the address below or visit our website www.candc-uk.com.

We will be happy to arrange a full demonstration and discuss any specific requirements you may have.

C&C Group
Bourne House
475 Godstone Road
Whyteleafe
Surrey
CR3 0BL

Telephone
01883 621006
Facsimile
01883 621007

Email
Seamless@candc-uk.com
Internet
www.candc-uk.com