Solving the Address Search & Matching Problem for Tax and Revenue Applications

An overview of the identity search & matching problem space, applications in Tax and Revenue applications and an introduction to Identity Systems technology

WP_TAX_070228

Introduction

Tax and Revenue Departments have various needs to find, match, group and investigate records based on identity data. Applications range from taxpayer inquiry through to compliance and investigation systems. Data in such systems suffers from the same types of natural and unavoidable error and variation found in other large identity data stores, but in addition there are other factors at work that affect data quality and completeness.

- Most individuals and organizations will only do the absolute minimum that it takes to "be in compliance"
- There are many avenues for tax avoidance and fraud within today's complex tax systems
- Tax systems must track people over long periods of time

Anyone who has ever searched a telephone book for a name like Ricky Smith knows that the endless variations and errors in this class of data make for an almost impossible task. (Possible matches include Rick Smythe, Richard Smithe, Smith Richard A., Ricardo A Smith, R. A. Smithe, etc). Tax and Revenue systems that use names, company names, and addresses to find or match information records face the same challenges. The error and variation in such identity data is compounded by the volume of information records being searched against, and the need to perform searches in real-time. In addition, it is typical in such systems – *especially compliance and tax discovery systems*, that apart from the accidental error and variation – the identity data is subject to **deliberate** abnormal or extreme variation.

Traditional phonetic search techniques like Soundex and NYSIIS don't address all aspects of the problem space, do not provide matches in ranked order, cannot address words that are in the wrong order (i.e. Lee Kwok Ki and Kwok Ki Lee), and do not handle data from multiple countries/character sets/languages efficiently.

The current fiscal realities for most Government departments increase the need for better identity searching and matching to help raise revenue that may otherwise be uncollected. Identity Systems technology can help Tax and Revenue agencies reduce costs and discover non-compliant taxpayers.

Reliable, accurate and fast identity searches are fundamental requirements of Tax Administration, Discovery and Compliance systems.

The identity search problem is complex – and requires sophisticated tools and techniques to address and solve the problem.

Use of Identity Systems Technology in Tax and Revenue Systems

Tax Administration

Many of Identity Systems customers use our products in both batch and real-time applications to match tax returns to taxpayer master files or for taxpayer enquiry searches.

In batch applications – it is common that tax returns are initially searched using SSN/TIN and taxpayer name and/or address is used to verify the match (to address the possibility of erroneous SSN/TIN on either the return or taxpayer record). If no match is found, Identity Systems technology is used to perform fuzzy, high-performance search and match using name, address and other attributes to confirm that this return is about a new taxpayer.

Online or real-time searches are also used to search and locate taxpayer records for customer service, call-in change requests, or dispute resolution.

Data Warehouse

Identity Systems' technology is used to create and maintain authoritative data stores for tax account and tax return information – to support a variety of uses including administration and tax discovery. Such data warehouse creation requires the ability to reliably match and link data about the same person or organization, often across disparate databases and systems. Identity Systems ability to work with data in any format is a significant benefit to such projects. Identity Systems has software that can perform this work off-line in batch using data extracts and data feeds, and also product for supporting such a system from operational data stores, via data synchronization and background relationship discovery utilities.

Compliance and Tax Discovery

Government revenue departments must work to enforce compliance within the tax regime to maximize allowable revenue and to demonstrate both to the politicians and to the public that efforts are made to expose tax avoiders and recoup losses. To make the best use of resources, an important component of compliance investigation is in identifying high-probability and high-worth cases to investigate. Such identification is greatly assisted by the process of matching identity data between different tax systems, and sometimes by matching tax data to third part data (e.g. the "Yellow Pages"). Apart from the *inevitable* error and variation that exists in identity data, data sourced from different systems suffers from an additional component of variation introduced by the different systems' input methods, different users, controls, forms, database designs and other constraints and practices.

Some examples of the types of compliance projects run by Identity Systems customers include:

- matching federal tax information to state taxpayer databases to identify individuals and companies that may have paid federal taxes but are not in the state system;
- cross matching employment tax reports from businesses (from withheld employee taxes) to the corporate Franchise Tax databases to verify if the business has or should be paying corporate franchise taxes; cross matching using 1099's (USA contractor) from individuals that reported this type of income against Franchise Tax data to find the companies that should pay corporate franchise tax;

- matching US Federal Aviation Administration (FAA) airplane registration records with reported "sales and use taxes" forms to find businesses that own airplanes and did not pay the required taxes;
- matching information from US Coast Guard and Parks and Wildlife Commission against "sales and use taxes" to find boat owners that fail to report/pay use taxes;
- matching US Customs data against "sales and use taxes" to find people that bought or imported works of art or any other high-value taxable items and failed to report/pay use taxes.

The Identity Search Problem

Names (and other identity data) suffer from unavoidable error and variation.

The variations that occur include spelling, typing and phonetic error; synonyms & nicknames; Anglicization, ethnic, and foreign versions of names; initials, truncation and abbreviation; prefix and suffix variations; compound names; account names; missing words, extra words and word sequence variations, as well as format, character and convention variations.

Variation examples

Person Names:

Nicknames: William, Bill, Billy, Will

Name Variation: Chris, Kris, Christie, Krissy, Christy, Christine, Tina

Abbreviation/Spelling: Mohammed, Mohd, Mohamad, Mhd, Muhammad

Foreign Versions: Peter, Pete, Pietro, Piere, Pierre

Spelling Variation: Johnson, Johnsen, Johnsson, Johnston, Johnstone, Jonson

Suffix Variation: Smith II, Smith 11, Smith Jr, Smith jnr

Anglicization: De La Grande, Delagrande, D L Grande

Out of Order: Henry Tun Lye Aun; Mr Aun Tun Lye (Henry)

Initials/Order: Frank Lee Adam; A. Frank Lee; Lee Frank

Titles: Dr. Henry Lee, Henry Lee, M.D., Mr. Henry Lee

Company Names

Town Park Plaza Hotel, Park Plaza, Hotel Plaza, Town Park

John Deer Engg Labs, John Deare Laboratories, Engineering Research Labs c/o John Deere Inc

IBM, International Business Machines, Intl. Bus. Mach., I.B.M., IBM Inc.

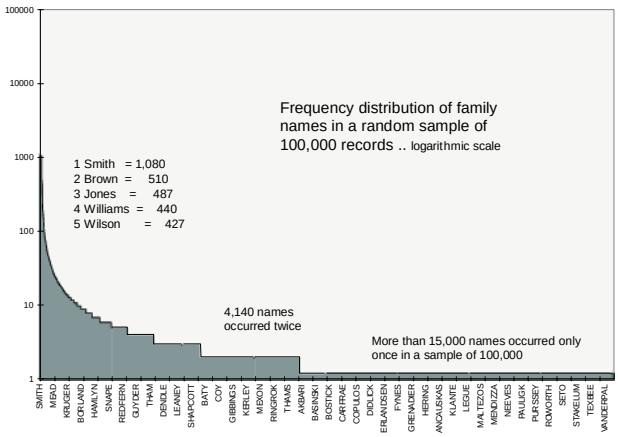
Addresses:

Jackson Rd. East Hartford
117-2a Jackson Rd East, Hartfrd
2a East Jackson, Hartford
117a Jackson Rd, E. Hartford

Grd. Fl. 192 Aberdeen St Southhead
192/1 Aberdeen Sreet Sth. Hd.
Ground Floor 192 Aberdeen St South Head

Common and uncommon words – the performance problem

The words we use to label things are chosen from a very different vocabulary than meaningful language. There are no dictionaries, spell checkers or rules for the names of people, places, things or even addresses. The vocabulary in use for people's first names includes in excess of 2,500,000 words in the USA alone, yet as much as 80% of the population may have names from as few as 500 words.



Accurate and high-performance name searching must perform for the uncommon names as well as for very common words. This is an extremely difficult challenge when a database of 100,000,000 people may contain 100,000 John Smiths, or Juan Rodriguezs or 1 Main Streets.

International Data

Most large identity databases contain data from multiple languages, countries and cultures which often have different structures, follow different parsing rules, and have different variation characteristics. Also, if transliteration, Romanization, character set conversions and other such transformations are employed, a new class of error and variation are introduced.

Names are truly many-to-many

It is obvious that two people or companies, or products, can have exactly the same name. It is also obvious that, even ignoring error and variation, people places and things have more than one name:

- People have maiden names and married names;
- People have aliases and professional names;
- Companies have registered names, trading names and division names;
- Places have several addresses, on two separate streets, old addresses, billing addresses, postal addresses etc.
- People and places can have names in more than one language.
- The relationship between a name and that which it names is quite naturally a true "many to many relationship."

Indexing these "many to many" relations requires careful design in the majority of today's search applications. Search techniques that depend on two fields, one for "name" and one for "maiden name", or "alias" are not good. When we are searching for a person's name, or address we do not know which "role" it plays. We do not know if it is a birth name, married name or maiden name, we do not know if it is a current or prior address. In order to address this problem effectively, it is necessary to have several keys or index entries pointing to the same identity.

The Standardization, Cleansing, and Parsing Dilemma

There are multiple Marketing objectives that have conflicting needs in the way that name and address data is captured, stored, and used. In many marketing systems, this conflict has not been recognized, leading to a bias in one area, and a less than satisfactory solution in another.

Examples of such conflicts are:

- The address most useful for reaching the prospect and fostering a good relationship is the one the prospect provides; the address most useful for achieving a cheap mailing rate is the one the Post Office provides.
- The name most useful for matching is an unformatted original; the name most useful for personalization will be formatted and enhanced;

- The address most useful for matching or creating household groupings is the unformatted or partially formatted original; the address most useful for statistical analysis will be formatted and enhanced.

Therefore, to achieve high-quality searching and matching of identity data, and to support good prospect/customer relationships, it is necessary to store the original, unenhanced data. In addition, to support marketing needs such as personalization or postal enhancement, it is necessary to store the formatted, enhanced data (or a derivative, e.g., postal barcode), or view these needs as 'output' processes to be applied to the data as it is taken out of the system.

Furthermore, when focusing on searching and matching addresses, it is important to understand exactly how the address is to be used. Traditional address matching methods rely on the cleaning and formatting of an address to build a stable "match-code." However, many address types, such as the location for a particular utility service or a record for land titles, cannot easily be "formatted."

Another aspect to consider in relation to mailing addresses is the source of the address. Increasing commercial globalization and Internet commerce has led to a much greater diversity of addresses reaching our systems. In current times, for example, it is not uncommon for a bookstore in the USA to deliver to a client in Japan.

The format of mailing addresses in some countries is relatively simple and well understood. However, many other countries have much more complex addresses and less rigorous standards for their writing. In addition, localized web sites often request information formatted in the home country's address format rules, which further complicates the issue and results in even messier data.

When designing name and address systems where the system has to operate in a variety of countries, or upon databases that are partitioned for different countries, it is necessary to do all data acquisition and database design in such a fashion as to minimize the differences between countries. This is the model on which many web-based systems must be based. Any organization dealing with addresses from multiple countries cannot rely on searching and matching processes that require that the addresses be cleaned and formatted.

Popular but less successful techniques

Text Search

The idea that text retrieval packages can successfully be used for name search applications is only believed by users who are unaware of the "good" data that they miss. Even when "Full Text" indexes have phonetic algorithms, or "expert" rule bases for Name searches the indexing mechanism will result in an inefficient process. Imagine the cost of finding all the index values for the records containing John and then joining them with those that contain Smith to discover the subset John Smith. In addition, most text search packages fail to recognize that Bill Smith and William Smith could be the same person.

Typical free text indexing techniques do not allow high performance or high quality retrieval of data containing names. Text searches are not for the serious searcher dealing with identity data.

WildCard Searches:

Wild-card searches do overcome SOME error and variation in the name. That is why users believe in them. In reality they only work if the searcher correctly guesses the right characters to include or exclude, and the database had no error in the characters actually used. They cannot find all the relevant records, do not address nicknames and abbreviations, nor the fact that different records have different errors.

Wild-card searches normally return too many irrelevant candidates and will always miss many of the relevant candidates. Wild-card searches are not for the serious searcher.

Match-codes

A Match-code is an entity key built from a combination of the entity's attributes. The Match-code for: John Smith dob: 22 Feb 1979 2/234 33RD St., New York, NY12345, might be: SMITHJ79NY. Some match-coding aims to build a unique key for entities, and could add a sequence number to the end. Match-codes require that the entity is first strictly formatted into its pieces; that all pieces used are in the "stable" order; and that there are no errors in the pieces used.

If the above was always true then Match-Codes would be successful. Typically Match-Codes do find "correct records", but they frequently miss all the other good candidates. If there is no ability to overcome error and variation in the name parts, all records with such error and variation will be missed. Missing word, extra word and word order variations are often missed unless the searcher permutes the search criteria.

Match-Codes fail if any one piece of the data used to build the code is not identical. Match-Codes are not for the serious search or matching application.

N-Gram Indexing

An n-gram is a set of "n" consecutive characters extracted from a word or code. Typical values for "n" are 2 or 3. These extracted n-grams are subsequently indexed for all names or addresses in the database. At search time, the idea is that words or codes that are similar between the search and file data will have a high proportion of n-grams in common. N-grams are particularly well suited to string and text searching; however, unless supported by extensive rule bases for phonetic and synonym variation, as well as for noise words, they do not readily overcome the typical error and variation found in identity data, nor do they easily scale to very large data volumes.

Yesterday's Soundex's, NYSIIS, and other simple phonetic algorithms

In the early 1900's the Russell Soundex, was developed to provide a stable manual filing code for the USA Census documents. The development of this algorithm for encoding the last name of a person was based upon phonetics and certain classes of typical spelling and filing errors. It was a simple set of rules to convert a last name word into a four, five or six digit number that had a high probability of being the same for two words that were variations of each other.

Since then, many Algorithms with a similar objective have been developed and modified. In 1969 the New York State Identification Intelligence System project evaluated the then popular algorithms. This included evaluation of last names on New York State criminal records to see how effective algorithms were at overcoming the error and variation. To this they used records that were previously paired using fingerprints. They evaluated such algorithms as: Soundex and many of its variants; LA County Sheriff; Consonant coding; Phonic standard and extended; Michigan Lien; and several Extract list based systems. The end result of their project was to define a popular algorithm known as NYSIIS which they proved was better optimized for their data at that point in time.

At Identity Systems we have conducted many such exercises on data from many customers and countries and have found that, even in one country, no single algorithm is optimum for all naming data.

Fundamentally the choice of the optimum word stabilization algorithm depends upon the source of the data and the frequency distribution of certain classes of error in the data. For example carefully written and captured Criminal Records have a very different error distribution than urgent verbal requests for information over a radio. While such stabilization algorithms can be a critical piece of a name search engine, the algorithms of the past are not suitable for use as database keys with today's data or volumes. Typically they either cause too many records to be found, or they miss too many relevant records.

Where stabilization algorithms are used, the choice of the right algorithm can be critical and must be based upon the true distribution of the error and variation in both the population of file data, and the population of the search data. The objective of word stabilization algorithms should be limited to retain as much of the original words form as is commensurate with the objective of "not missing valuable records."

However, the problem of phonetics and spelling is only one part of the general name search problem – and as such a complete name search solution cannot be solely dependent on just a phonetic algorithm. The design of suitable database keys to maximize quality is a separate and much more complex exercise.

Identity Systems' Approach:

The ideal identity search solution:

- Overcomes the error and variation in the data
- Maintains excellent performance even in extreme volume situations
- Finds all valid candidates without generating false matches.

This implies:

- Intelligent and scalable **algorithms**, which, through the use of rich keys and search strategies, return all of the candidates an expert user would consider as being the same as the search criteria.
- These algorithms must be able to cope with data from the real world. This includes data that is not formatted or cleaned or which contains missing, extra, truncated, out of order, non standard, or noise characters/words, initials, abbreviations, nicknames, numbers, codes, and concatenations.
- The algorithms need a customizable rule base to incorporate the knowledge of the expert user, and a default population rule base in the case where the user is not that experienced.
- The algorithms require phonetic and orthographic correction functionality, to address spelling and typing errors.
- Intelligent matching routines must be available and able to be tuned to mimic the expert user making a choice as to which candidates are the correct matches. Such matching routines need to take into account all of the error and variation in the identities' attributes, as well as weighting the attributes as the user would.
- The Algorithms must work well regardless of the country of origin and language of the data and must insulate the application developer from the differences between country and language.

Identity Systems' software is designed and built on these premises and is being used by over 500 commercial and governmental organizations around the world.

A Representative List of Identity Systems Government and Revenue Customers

USA Users include:

Federal Bureau of Investigation
Internal Revenue Service (IRS)
District of Columbia, Office of Tax & Revenue
Maine, Revenue Services
City of Detroit – Tax
City of Pittsburgh - Tax
Georgia, Dept of Revenue
Massachusetts, Dept of Revenue
New York City – Dept of Finance
Ohio – Dept of Taxation
Pennsylvania – Dept of Revenue
Tennessee – Dept of Finance
California – Board of Equalization
California – Franchise Tax Board
Kansas City – Dept of Finance
Texas – Comptroller of Public Accounts
Wisconsin – Dept of Revenue
Washington – Dept of Revenue

International Users include:

British Columbia – Dept of Finance (Canada)
Australian Tax Office
Inland Revenue Authority of Singapore
UK Inland Revenue
Victorian State Revenue Dept (Australia)
Western Australia State Revenue
Bureau of Internal Revenue Philippines
NSW Office of State Revenue
VIC State Revenue Office
WA State Revenue Office
QLD Office of State Revenue
NZ Revenue
Canada Customs & Revenue Agency

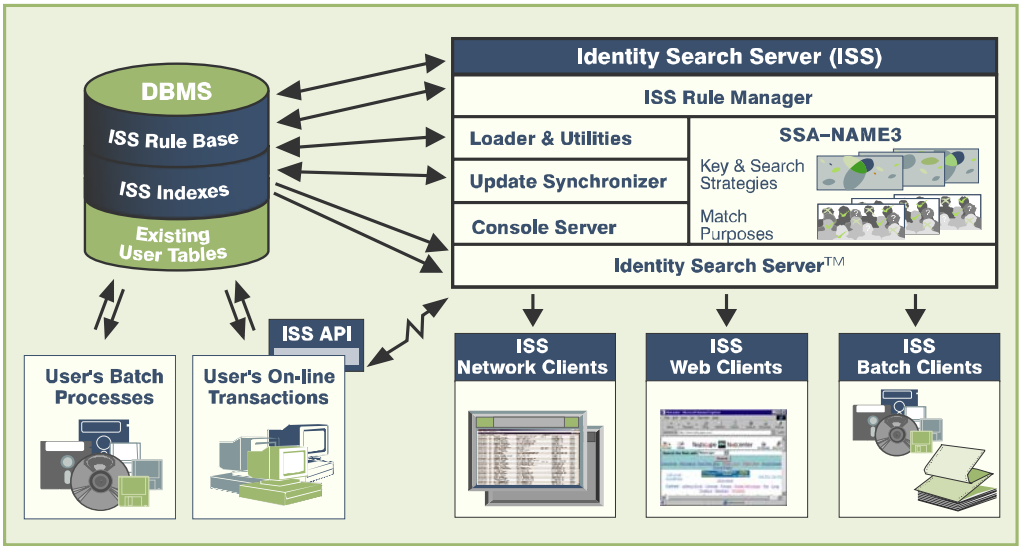
Identity Systems Products

There are three products:

- **Identity Search Server (ISS):** Online / Batch
- **SSA-NAME3:** Core SDK/API (included in other products)
- **Data Clustering Engine (DCE):** Batch

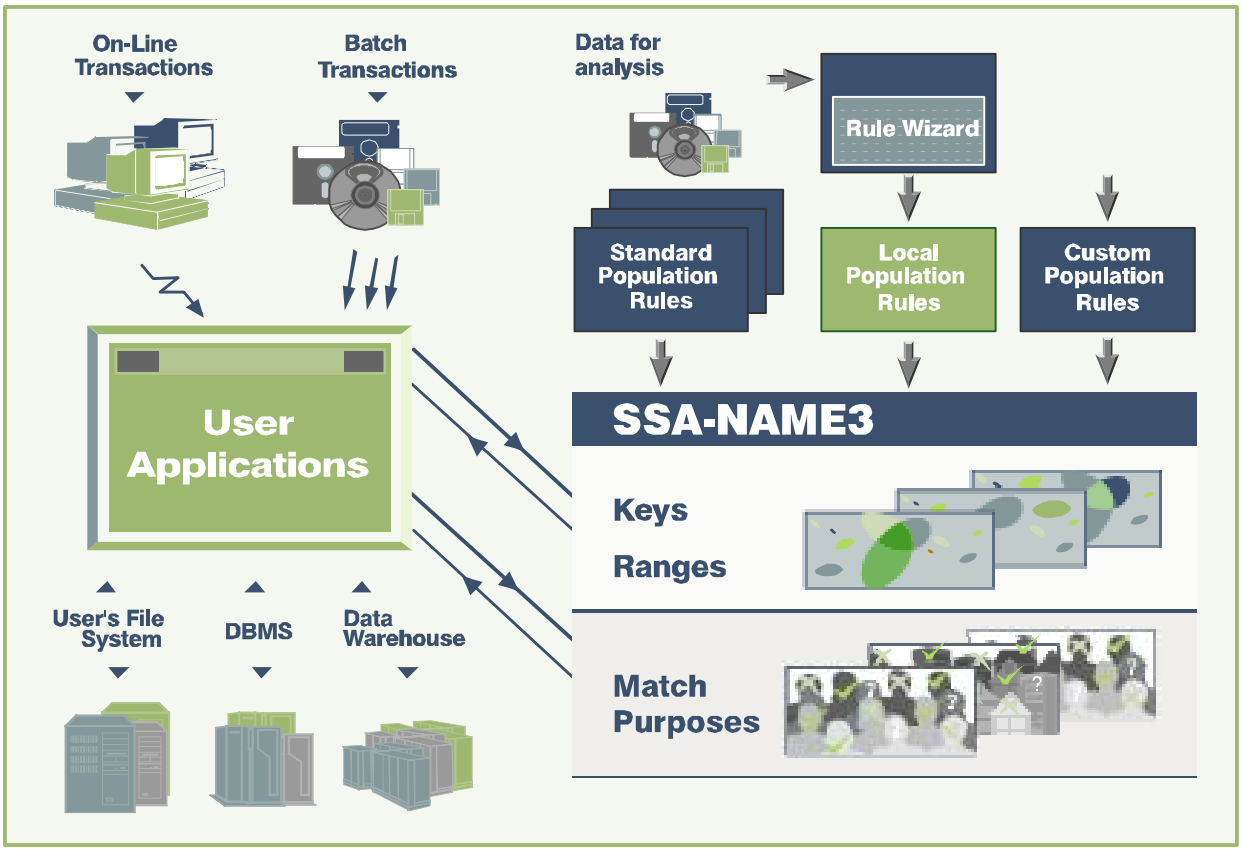
All these products can search, match, and rank identity **data of any quality or format** and do not require any "cleaning" or "scrubbing" of the source data nor is the source modified in any way by Identity Systems products. **All countries/languages** are supported, both singularly and in combination. All products can be used to centrally index and simultaneously search, match, and group identity data from disparate sources such as from Current Sales, Legacy, and External databases. An overview of each product follows.

The Identity Search Server (ISS) is the product that provides **on-line and batch searching, matching and duplicate discovery** for all types of identification data stored in relational databases (**DB2/UDB, Oracle, and SQL/Server)** on OS/390, NT, AIX, HP/UX, Linux, & Solaris platforms. High-performance indexes are automatically maintained **without changes to existing application programs**. Custom search clients can be developed using an easy-to-use API. ISS uses an **n-tier** architecture, making it the ideal solution for organizations implementing the latest technologies, including web-based applications. With ISS, the capability to **centrally** index identity information from **disparate** source tables, databases, and computers is particularly advantageous since the central index can then be used to search, match, rank, group, and maintain all the data **simultaneously** and in **real time**.

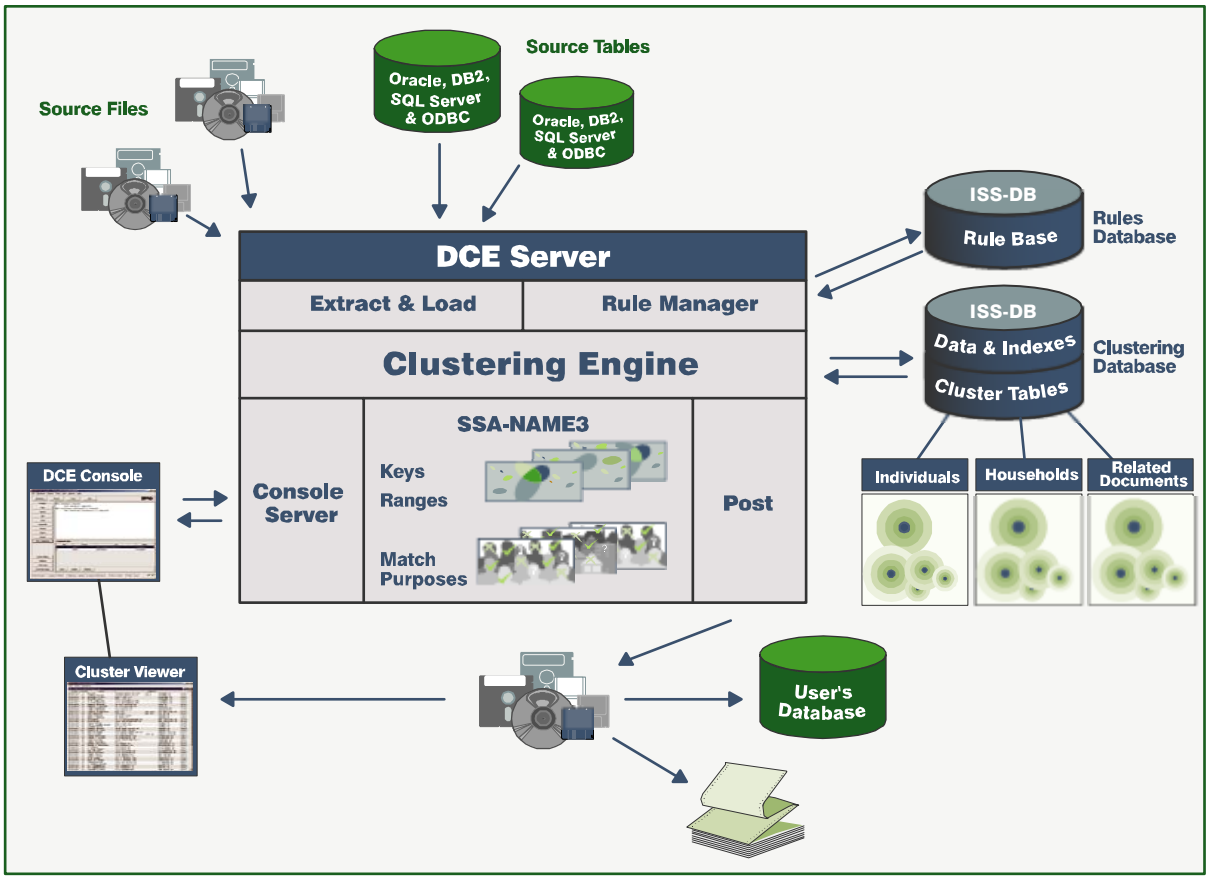


Customizable Identity Search Server™

SSA-NAME3 is Identity Systems **core technology** and included in all the other Identity Systems products. It is a system development kit that enables organizations to build business application programs to **search, match, rank, and analyze** records about people, companies, products, addresses and various other identity data. It can be executed on most any platform and operate on data stored in any database. The software contains algorithms for computing 'fuzzy' search keys and strategies, as well as algorithms for complex matching of identity data.



The Data Clustering Engine (DCE) is a **stand-alone, batch** data **grouping and investigation engine** for all forms of identification data and can execute on AIX, HP/UX, Linux, Windows NT, Solaris, & Compaq Tru-64 Unix platforms. The DCE performs a thorough analysis of the relationships between people, companies, addresses, products and other entities. This powerful software is used by numerous leading corporations and institutions for **de-duplication, file matching, householding, fraud investigation**, and various other analyses. It **does not require programming** and is highly scalable for the processing of large and extreme data volumes.



Identity Systems

Identity Systems is the pioneer in enabling organizations to build and maintain high-quality identity data search and matching software solutions. Identity Systems have been in this area of specialization for more than 20 years. Identity Systems now has over 600 clients worldwide who rely on its robust enterprise-wide software. Identity Systems became a wholly owned subsidiary of Nokia in 2006.

Company History

Identity Systems was originally established in Australia in 1987 as Search Software America (SSA) with the purpose of formally developing and marketing its founders experience in building identity search and matching systems. This experience had shown that, while the significance of identity matching in different systems varies considerably, there has always been the same basic set of concerns:

- **Performance problems arise because the most frequently occurring names are also usually those upon which searches are most often performed.**
- **Traditional phonetic algorithms create poor response time and prevent the user from locating a match among many candidates. Traditional match-codes fail to find matches when data is invalid or incorrectly parsed.**
- **True phonetics is only a subset of the errors in names, addresses and other identity data.**

Some of our projects emphasized the need to quickly achieve a match, if there was one. Others placed their emphasis on proving that there was no match at all. One project presented the unusual opportunity to empirically develop and modify an algorithm designed to solve phonetic, orthographic, and Anglicization problems in more than 2,000,000 hand-written credit records. During the project development activity, some 300,000 computerized matches were compared to manual matches done by expert searchers. Whenever the searcher found data not found by the system, the algorithm was refined.

Another project involved the re-processing of 25 million records of international data from virtually every country in the world, where it was known that at least 99% of the records were in fact pairs about the same person. The project also demanded a performance breakthrough since the design objective was to develop an identity search system to handle over 30,000 inserts per day and to support a 50,000,000 record on-line database. Based on the fact that the project's purpose was to identify the records that were not pairs and the population was smaller than the error rate in the data, the successful solution required considerable research.

The experience gained from such projects and Identity Systems work with customers around the world for the last 20 years, has been incorporated into ongoing development and maintenance of Identity Systems product suite. Although the identity search and matching problem may never be completely solved, Identity Systems enables organizations to reach unparalleled levels of quality and performance.

Contact Us

For more information about Identity Systems and our products, please visit our website, or request information from one of the contacts below. For other locations and distributors, visit www.identitysystems.com/contact.htm.

Americas

Identity Systems
1445 East Putnam Avenue
Old Greenwich CT 06870
USA
tel. +203-698-2399
USASales@identitysystems.com

EMEA

Identity Systems
IDS House
9 Headley Road
Woodley, Reading RG5 4JB
UNITED KINGDOM
tel. +44-118-944 9688
UKSales@identitysystems.com

France

Identity Systems
(Nokia)
164 Bd Victor Hugo
St Ouen 93400
FRANCE
tel: +33-1-4915 1515
frsales@identitysystems.com

Germany

Identity Systems
Lyoner Straße 26
D-60528 Frankfurt/Main
GERMANY
tel. +49-69-677 33 462
desales@identitysystems.com

Australia & New Zealand

Identity Systems
Nokia House
Level 6, 19 Harris Street
Pyrmont, NSW 2009
AUSTRALIA
tel. +61-2-9571 1300
AUSSales@identitysystems.com

South East & East Asia

Identity Systems
(Nokia)
438B Alexandra Road
#07-00 Alexandra Technopark
SINGAPORE 119968
tel. +65-6723 1620
SEASales@identitysystems.com