



IBM **Information Management** software

Advanced Global Name Recognition Technology

*Dr. John C. Hermansen
IBM Distinguished Engineer
Chief Technology Officer
IBM Global Name Recognition*



Contents

2 Introduction

4 Elements of the IBM High-Precision Name Matching System

8 IBM Global Name Recognition – Leading the Industry in Advanced Name Recognition

9 IBM Global Name Recognition Technologies

10 IBM Global Name Analytics

10 IBM Global Name Scoring

11 IBM Global Name Reference Encyclopedia

12 Platforms Supported

12 For Additional Information

Introduction

Despite many remarkable advances made in other areas of business automation, automated processing and matching of personal names in databases has languished for decades without significant theoretical or practical advances. The purpose of this paper is to highlight the issues, requirements, and technologies available for automated advanced name recognition.

The problem to be solved is a familiar one for many people: a name is entered in one database with the surname “Rodgers,” and in a different database as “Rogers.” A person’s name is recorded as “Dayton,” but should actually be spelled “Deighton.” The problem is greatly compounded with names originating outside North America. For example, the same Chinese person may have one set of information recorded under the surname “Xue,” and another under the surname “Hsueh.”

The earliest attempt at coping with name variation was the Russell Soundex matching algorithm, developed around 1910 as an aid in the manual analysis of U.S. Census records. The original Soundex method of generating ‘keys’ was later implemented as a software-based algorithm, and is today the most widely used alternative to exact-matching when names are involved in automated search and retrieval systems. Over the years, there have been many attempts to improve on Soundex, but they are all still key-based systems and, therefore, suffer from the same fundamental deficiencies that plague Soundex.

While it is certainly compact and efficient, the key-based approach falls well short of solving many of the problems associated with searching for names. Two extensive studies examined the results of the

basic Soundex algorithm, using statistical measures to gauge accuracy.

- Study #1 Results: Only 33% of the matches that would be returned by Soundex would be correct. Even more significant was the finding that fully **25% of correct matches would fail to be discovered by Soundex**. (Alan Stanier, September 1990, *Computers in Genealogy*, Vol. 3, No. 7)
- Study #2: Only 36.37% of Soundex returns were correct, while **more than 60% of correct names were never returned by Soundex**. (A.J. Lait and B. Randell, 1996)

Obviously, for mission-critical federal applications such as terrorist watch-lists, INS tracking, visa applications, and fraud detection, failing to identify 25-60% of target names within a database is unacceptable. The Federal Government recognized this deficiency, and worked with IBM Global Name Recognition over the past two decades to develop advanced technology for improving performance across multiple cultures. This approach hinges on the latest advances in computational linguistics – the application of statistics, mathematics, linguistics research, and computational expertise to the problem of name matching. This approach is now also available for commercial organizations.

IBM Global Name Recognition technology is the ONLY name searching patented software since Soundex!

Elements of the IBM High-Precision Name Matching System

In order to meet the challenges posed by large, multi-cultural databases in which both predictable and random name-spelling variations are present in a significant number of records, an IBM Global Name Recognition solution provides:

1. ***Culture-specific matching criteria.*** Naming systems differ significantly from one culture to the next—in the relative order in which parts of a name appear, in the consistency with which they are written in romanized form, in the way they are abbreviated, and in which parts are considered mandatory for identification. To identify all potential matches accurately, IBM technologies must first determine a name’s culture of origin. Such knowledge allows the correct set of matching techniques to be applied to the name. IBM accomplishes cultural identification automatically, adding speed and consistency that humans cannot be expected to provide.
2. ***Automatic application of linguistic rules for the culture/ language context.*** A full name must be parsed, and possible word order variations and shortened forms must be identified. Spelling variants for each part of the name are calculated. There are many possible approaches to this step—rule-based, algorithmic, statistical/probabilistic, or combinations of these. Furthermore, variants may be based on either phonetic (pronunciation) or alphabetic similarity. IBM has accumulated over 750,000,000 names from every country in the world. These names are used to provide the automated statistical and linguistic methods required for accurate name matching.

3. **Noise tolerance. (e.g., typographical errors)** Once culture-specific knowledge has been used to isolate and align those portions of the name to be compared, the character-level comparisons take into consideration the possibility of random keying, which correspond to no orthographic or phonological principle.
4. **Recognition of equivalent but dissimilar name variants.** (e.g., Elizabeth and Betty, or Paco and Francisco) In most cultures, names are found which are understood and accepted as interchangeable equivalents, perhaps used in different social circumstances. Nicknames and pet names (ELIZABETH ~ BETTY) are prominent examples of given name (first name) variants in wide use among English-speaking and Western European societies. IBM Global Name Recognition technology automatically recognizes name variants from multiple cultures.
5. **Ranked returns, with the best matches presented first.** Matching names that are most similar to the query name are returned before those that are less similar. IBM technology includes a means to measure the degree of similarity between two names and ranks them accordingly in search results, using sophisticated intelligence about the sound, spelling, and patterns of variation known to occur in each culture.

6. ***Statistical and probabilistic search aids.*** Knowing the relative frequency of a specific name within a particular population allows a correspondingly greater emphasis to be placed on the discriminatory value of the information supplied by other name parts in a search transaction. IBM closely integrates statistical and probabilistic information with the matching and ranking logic of the search algorithm, based on its extensive archives of name data from every country in the world. This type of statistical information becomes crucial, for example, when dealing with Korean names, since approximately 75% of the population share the top half-dozen surnames. Such statistical information, derived from our unique corpus of names, is critical in order to determine “typical” or “distinctive” name patterns for a particular culture or society.

7. ***Syntactic flexibility.*** Because names are particularly susceptible to misinterpretation when they are captured in electronic form from oral or written origins, differences in white-space placements or even field placements (within a database record) must be overcome to a reasonable degree in an advanced name searching system. For example, Oriental names in which the surname order is accidentally reversed, or Middle Eastern names with prefixes mistakenly classified as what Americans call “middle names” should be reliably and efficiently matched with their more standard counterpart versions.

8. ***Capacity for adjustment and tuning.*** Name searching is a non-deterministic problem, meaning that it is not always possible to obtain definitive results. Exact-matches in name search results are easy to identify, but there are many shades of similarity and equivalence possible to discern among related names, so “good” search results may depend more than anything on the linguistic and cultural knowledge of the user. IBM Global Name Recognition technologies provide numerous mechanisms for fitting search results to business rules by adjusting the quality and quantity of the matches it produces.

9. ***Support for end-user education and assistance.*** Because names can often be complex or even ambiguous in nature, results for name matching may be confusing to someone uneducated in a particular culture. IBM provides automated name reference tools that greatly assist the end-user of a system to understand the advanced output of matching names from around the world.

IBM Global Name Recognition – Leading the Industry in Advanced Name Recognition

IBM is the leader in providing high precision software for mission-critical name matching and searching. Since 1984, IBM (through Language Analysis Systems, Inc. which was acquired by IBM in March, 2006) has pioneered the use of computational linguistics expertise and technology to solve the complex problem of multi-cultural name matching and searching. It is no longer necessary to task programmers with “tweaking” older, key-based approaches to try to solve this persistent problem. IBM offers off-the-shelf software and linguistics expertise for truly advanced solutions.

IBM Global Name Recognition delivers previously unavailable results for client requirements in the following areas:

- Fully-automated, high-performance, multi-cultural name matching
- Ethnic-based marketing campaigns
- Database de-duping and cleansing
- Terrorist watch-list checking
- Fraud detection
- Predictive data mining
- Improved precision for culturally-sensitive CRM applications

It is now a matter of public record that, immediately after the September 11 terrorist attacks, federal authorities used IBM Global Name Recognition technology to quickly expand their investigation, tracking the hijackers to their Florida connections. The software continues to prove its value everyday within federal, aviation and financial

institutions, allowing those organizations to respond effectively to new federally mandated Watch List applications. The value of this new technology is also becoming evident to companies that now recognize the importance that this high-performance software has for mission-critical commercial applications.

IBM Global Name Recognition software currently runs in over 250 locations worldwide, supporting hundreds of cultures and languages out-of-the-box.

IBM Global Name Recognition software enables any organization to minimize fraud, increase sales, expedite collections, and improve its handling of critical customer relations. Over 15 years in its development, this specialized technology has proven its effectiveness across a wide variety of platforms, and is surprisingly easy to integrate. This software is now available as a commercial off-the-shelf solution for government agencies and commercial enterprises.

IBM Global Name Recognition Technologies

IBM's patented name recognition technology is exactly suited for each type of problem. Some of the highlights include the ability to:

- Identify name by culture and relative frequency
- Search for multi-cultural names in a database
- Parse a name into Surname and Given Name
- Generate frequency statistics for name tokens
- Generate all variants of a name
- Generate additional attributes such as gender
- Quickly train field personnel in advanced multicultural name searching techniques
- Utilize rich name data gained from the comprehensive study of over 1 billion names from around the world

IBM Global Name Analytics

IBM Global Name Analytics is designed to address the specific needs and demands of managing multicultural data sets. Unlike traditional data cleansing capabilities that have been designed primarily to manage data assets in westernized, romanized cultures, IBM Global Name Management is designed to meet the unique demands of organizations and governments that rely upon data sets from cultures around the globe.

IBM Global Name Analytics identifies and classifies what cultural background a given name comes from and recognizes whether a name is predominantly male or female. It automatically parses culturally diverse personal name information into surname and given name components to ensure name data is consistent and accurate across your systems. These capabilities enable organizations to improve data quality, retain customer information, and treat multiple cultures sensitively by accurately parsing and storing customers' names within their automated systems.

IBM Global Name Scoring

IBM Global Name Scoring allows a user to identify identical and fuzzy clear text and phonetic name matches more effectively, overcoming the vagueness and inexactness of transliteration, pronunciation, and the wild profusion of naming and syntactical schemes.

IBM Global Name Scoring enables users to search for multi-cultural names in a database and provides the most likely variations more effectively, improving the accuracy of name searching and the quality of identity verification initiatives. This capability overcomes

the vagueness and inexactness of transliteration, as well as the wild profusion of naming and syntactical schemes that make it difficult to distinguish the Saddam Hussein's from the Prince Hussein's. Users can search and recognize foreign names, screen potential threats, and perform background checks across multiple geographies and cultures.

IBM Global Name Scoring provides a phonology-oriented search capability that provides ranked search results based on similarity of pronunciation. Phonetic matching applies language-specific letter-to-sound rules in order to identify potential pronunciations for names, so that two superficially dissimilar names can be matched by a shared spoken form.

IBM Global Name Reference Encyclopedia

The IBM Global Name Reference Encyclopedia is a comprehensive, interactive database of names, name use, and variations for use by analysts, investigators, and researchers within global public and private organizations. It contains culture-specific information about names, their use, meanings, and patterns of spelling variations. Each name is automatically analyzed to show: cultural/ethnic classification, most prominent spelling variants, gender associations/probabilities, titles, affixes, qualifiers, and countries where name occurs most frequently.

The IBM Global Name Reference Encyclopedia includes information and analyses for names from around the world: Anglo/European, Arabic, Chinese, Hispanic, French, German, Indian, Korean, Pakistani, Russian/Slavic, Thai, Japanese, Western African cultures, and more.

Platforms Supported

IBM Global Name Recognition technologies are available across Win32, Unix, and Linux platforms. Interfaces are available in C++, JNI, SOAP, and XML-over-IP for most products. (Please check specific availability)

Contact Information

Dr. John (Jack) Hermansen
Chief Technology Officer, IBM Global Name Recognition
(703)834-6200 x222 • jhermansen@us.ibm.com

Thomas Woodcheke
Worldwide Sales Manager, IBM Global Name Recognition
(703)834-6200 x253 • twoodche@us.ibm.com

Leonard Shaefer
Directory of Development, IBM Global Name Recognition
(703)834-6200 x228 • lshaefer@us.ibm.com

Timothy Paydos
Director of Marketing, Threat & Fraud Intelligence
(860)408-1639 • tpaydos@us.ibm.com

Additional Information

For the latest information about our products and services, see the following website: www.ibm.com/software/data/globalname/



© Copyright IBM Corporation 2006, 2005

IBM (United States of America)
Entity Analytic Solutions
6600 Bermuda Rd, Suite A
Las Vegas, Nevada
United States of America, 89119

Printed in the United States of America
06-06
All Rights Reserved.

DB2, IBM, the IBM logo, and the On Demand logo are trademarks of International Business Machines Corporation in the United States, other countries or both.

Other company, product and service names may be trademarks or service marks of others.

References in this publication to IBM products or services do not imply that IBM intends to make them available in all countries in which IBM operates.

 Printed in the United States of America on recycled paper containing 10% recovered post-consumer fiber.

IBM's customers are responsible for ensuring their own compliance with relevant laws and regulations. It is a customer's sole responsibility to obtain advice of competent legal counsel as to the identification and interpretation of laws and regulations that may affect a customer's business and any actions required to comply with such laws. IBM does not provide legal, accounting or audit advice or represent or warrant that its services or products will ensure that a customer is in compliance with any law.