**IBM Analytics**
White Paper

October 2017

# Advanced global name recognition technology

*Frankie Patman Maguire, Linguist/Architect,*
*IBM Global Name Recognition*

## Contents

## Introduction

Despite many remarkable advances 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 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 that originate in one writing system and are then transferred into another. For example, the same Chinese person (whose name is normally written in Han characters in China) may have one set of information recorded under the Romanized surname “Xue” and another under “Hsueh” in the United States.

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 US Census records. The original Soundex method of generating “keys” was later implemented as a software-based algorithm, and today it is the most widely used alternative to exact matching when names are involved in automated search and retrieval systems.

There have been many attempts to improve on Soundex over the years. However, they are all still key-based systems—and therefore they all suffer from the same fundamental deficiencies. While the key-based approach is certainly compact and efficient, it 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. In the first study, only 33 percent of the matches returned by the Soundex algorithm were correct. Even more significant, the algorithm failed to discover 25 percent of correct matches.[1] In the second study, just 36 percent of Soundex returns were correct, while more than 60 percent of correct names were never returned.[2]

Obviously, for mission-critical government applications such as terrorist watchlists, immigration tracking, visa applications and fraud detection, failing to identify 25 to 60 percent of target names within a database is unacceptable. IBM has worked with government agencies around the world over the past two decades to develop advanced technology for improving name matching 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 solve the problem of name matching. This approach is now also available for commercial organizations.

## Delivering high-precision name matching

The IBM® InfoSphere® Global Name Management platform is designed to address the challenges posed by large, multicultural databases in which both predictable and random name-spelling variations are present in a significant number of records. The solution offers:

**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. This knowledge allows the correct set of matching techniques to be applied to the name. InfoSphere Global Name Management software automates cultural identification of names—enabling levels of speed and consistency that humans cannot provide.

**Automatic application of linguistic rules for the cultural and language context.** A full name must be parsed, and then possible word-order variations and shortened forms must be identified. Spelling variants for each part of the name must be 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 more than 750 million names from around the world. These names are used to provide the automated statistical and linguistic methods required for accurate name matching.

**Noise tolerance (such as tolerance for 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 that does not correspond to any orthographic or phonological principles.

**Recognition of equivalent but dissimilar name variants.** Most cultures have names that are understood and accepted as interchangeable equivalents, perhaps used in different social circumstances. Nicknames and pet names ("Betty" for "Elizabeth" or "Paco" for "Francisco," for example) 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.

**Ranked returns with the best matches presented first.** 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. Matching names that are most similar to the query name are returned before those that are less similar.

**Syntactic flexibility.** Because names are particularly susceptible to misinterpretation when they are captured in electronic form from oral or written origins, an advanced name-searching system must reasonably overcome any differences in whitespace placements or even field placements (within a database record). For example, InfoSphere Global Name Management detects Asian names in which the surname order is accidentally reversed or Middle Eastern names with prefixes mistakenly classified as what Americans call “middle names” and matches them with their standard counterpart versions.

**Special treatment for names of organizations.** Organization names differ significantly from personal names. For example, they do not share the given name and family name distinctions typical of personal names and they may contain special characters not used in personal names, such as digits. They may also contain sequences of words that together create a semantically meaningful term, which should be searched as a unit, and they often contain legal business registration terms that provide little distinctive information. IBM global name recognition capabilities apply search and match strategies created specifically to handle the special challenges posed by organizational names.

**Capacity for adjustment and tuning.** Name searching is a nondeterministic 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 among related names, so “good” search results may depend primarily on the linguistic and cultural knowledge of the user. InfoSphere Global Name Management provides numerous mechanisms for fitting search results to business rules by adjusting the quality and quantity of the matches it produces.

**Support for end-user education and assistance.** Because determining the degree of similarity between names can often be complex, and because the conventions and rules governing names may differ from one language and culture to the next, results for name matching may be confusing to someone unfamiliar with names from a particular culture. IBM provides an automated name reference tool to help the end user of a system understand the output of name matching operations that may return names from parts of the world that are unfamiliar to the user.

## IBM InfoSphere Global Name Management: A leader in advanced name recognition

IBM is a leader in providing high-precision software for mission-critical name matching and searching. Since its acquisition of Language Analysis Systems, Inc., IBM has advanced the use of computational linguistics expertise and technology to address the complex problem of multicultural name matching and searching. It is no longer necessary to task programmers with adapting older, key-based approaches to try to solve this persistent problem. IBM offers off-the-shelf software and linguistics expertise for truly advanced solutions. InfoSphere Global Name Management can deliver outstanding results in the following areas:

- Fully automated multicultural name matching
- Database deduplication and record merging
- Terrorist watchlist checks
- Fraud detection
- Predictive data mining
- Improved precision for culturally sensitive customer relationship management (CRM) applications

InfoSphere Global Name Management software continues to demonstrate its value every day within government, aviation and financial institutions, allowing those organizations to respond effectively to new federally mandated watchlist applications. In addition, the value of this technology is becoming evident for mission-critical applications in retail, tourism, healthcare and other industries with culturally broad customer bases.

Deployed at customer locations worldwide, InfoSphere Global Name Management provides out-of-the-box support for enterprises dealing with hundreds of cultures and languages. It helps organizations minimize fraud, increase sales, expedite collections and improve handling of critical customer relations.

## IBM InfoSphere Global Name Management technologies

IBM's patented name recognition technology was designed to address several common name-recognition issues. The software can:

- Identify names by culture
- Search for multicultural names in a database
- Parse a name into surname and given name
- Generate frequency statistics for name tokens
- Generate potential 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 analysis of more than 750 million names from around the world

IBM global name recognition technologies are available on Microsoft Win32, UNIX and Linux platforms. Interfaces are available in C++, Java, SOAP and XML-over-IP for most products. Check with IBM for specific availability.

## Name analysis techniques

InfoSphere Global Name Management name analysis components are designed to address the specific 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 name analysis software is designed to meet the unique demands of organizations and governments that rely upon data sets from cultures around the globe.

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

## Name search, match and scoring techniques

InfoSphere Global Name Management enables users to search for multicultural names in a database and provides the most likely variations, which enhances the accuracy of name searching and the quality of identity verification initiatives. This capability helps address the problem of inexactness due to transliteration, as well as the profusion of naming and syntactical schemes that can make it difficult to distinguish a "Saddam Hussein" from a "Prince Hussein." Users can search and recognize foreign names using InfoSphere Global Name Management capabilities in systems such as those that screen potential threats and perform background checks across multiple geographies and cultures.

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

## InfoSphere Global Name Management end-user support tool

InfoSphere Global Name Management provides a comprehensive, interactive encyclopedia of names, name use and name variations for use by analysts, investigators, researchers and developers within global public and private organizations. It contains culture-specific information about names, their use, their meanings and patterns of spelling variations. Each name entered by a user is automatically analyzed to show:

- Cultural and linguistic classification
- Most prominent spelling variants
- Gender associations and probabilities
- Titles
- Affixes
- Qualifiers
- Countries where the name occurs most frequently

The encyclopedic section of the tool includes information for names in many languages, including Arabic, Chinese, Hispanic, French, German, Indian, Korean, Pakistani, Russian/Slavic, Thai, Japanese and Western African languages.

## For more information

To learn more about IBM global name recognition products and services, including those discussed in this white paper, contact your IBM representative or visit: **ibm.com**/us-en/marketplace/ibm-infosphere-global-name-management

Additionally, IBM Global Financing provides numerous payment options to help you acquire the technology you need to grow your business. We provide full lifecycle management of IT products and services, from acquisition to disposition. For more information, visit: **ibm.com**/financing

## About the author

Frankie Patman Maguire has 30 years of experience in linguistic research and applied linguistics. She has focused on automated name analysis and search since 1999 and is a named inventor on several patents related to processing person and organization names. She holds a Master of Science in Linguistics from Georgetown University.

Software Group
Route100
Somers, NY 10589

Produced in the United States of America
October 2017

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

The client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions. THE INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS" WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

The client is responsible for ensuring compliance with laws and regulations applicable to it. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the client is in compliance with any law or regulation. Statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

[1] Alan Stanier, September 1990, Computers in Genealogy, Vol. 3, No. 7.

[2] A.J. Lait and B. Randell, Sept. 1995, Dept. of Computing Science, University of Newcastle upon Tyne, "An Assessment of Name Matching Algorithms," unpublished.

Please Recycle