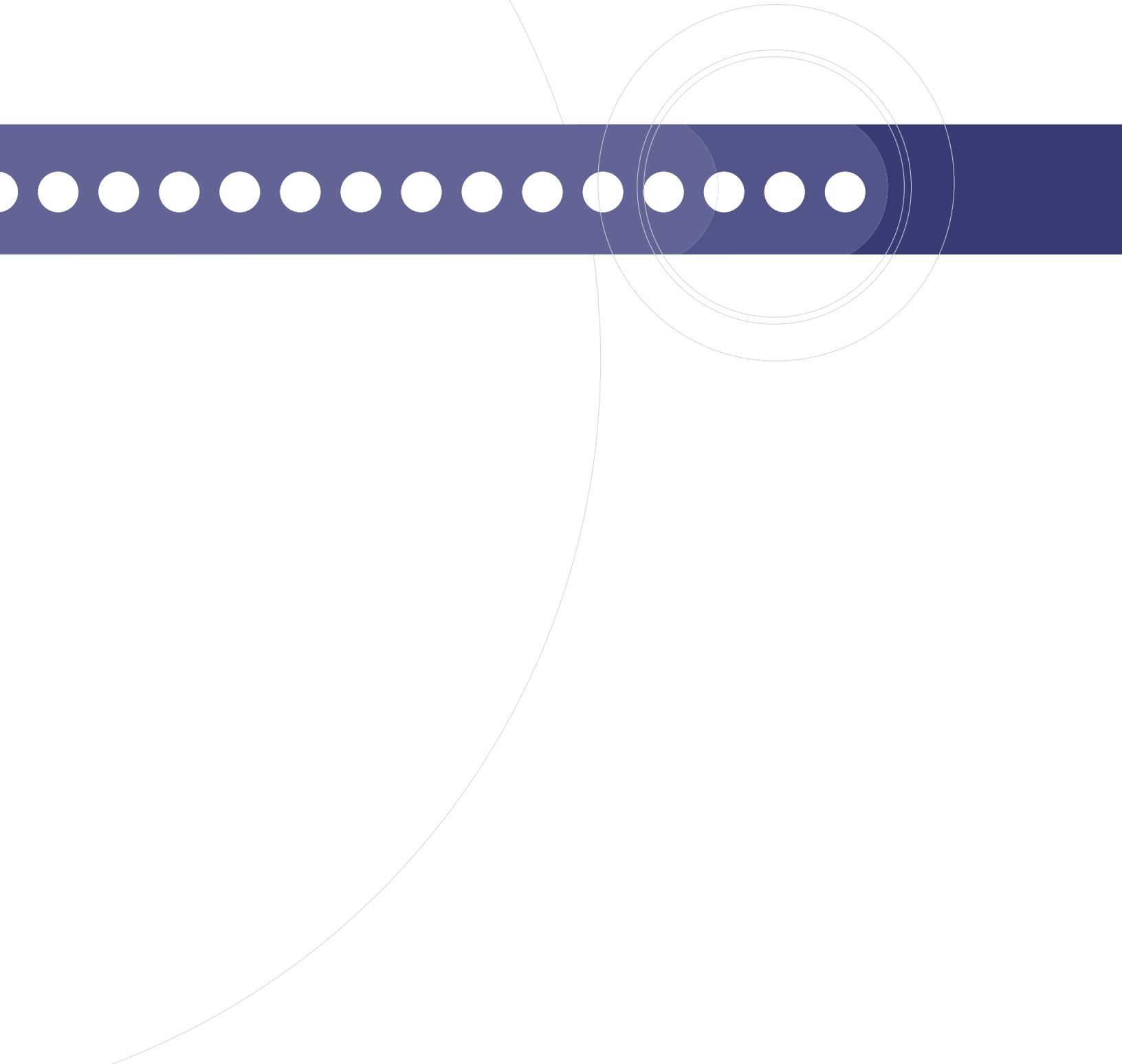


Data Quality and Identity Resolution



This document contains Confidential, Proprietary, and Trade Secret Information (“Confidential Information”) of Informatica Corporation and may not be copied, distributed, duplicated, or otherwise reproduced in any manner without the prior written consent of Informatica.

While every attempt has been made to ensure that the information in this document is accurate and complete, some typographical errors or technical inaccuracies may exist. Informatica does not accept responsibility for any kind of loss resulting from the use of information contained in this document. The information contained in this document is subject to change without notice.

The incorporation of the product attributes discussed in these materials into any release or upgrade of any Informatica software product—as well as the timing of any such release or upgrade—is at the sole discretion of Informatica.

Protected by one or more of the following U.S. Patents: 6,032,158; 5,794,246; 6,014,670; 6,339,775; 6,044,374; 6,208,990; 6,208,990; 6,850,947; 6,895,471; or by the following pending U.S. Patents: 09/644,280; 10/966,046; 10/727,700.

This edition published May 2008

Table of Contents

Introduction	2
History and Trends	3
Data Quality and Identity Resolution Projects	4
Common Challenges	6
Data Quality and Identity Resolution	7
Conclusion	8
FAQs	9



A wide range of customer requirements is driving the need for an integrated platform that provides broader Data Quality and Identity Resolution capabilities to address evolving use cases.

Introduction

Over the past decade, Data Quality processes have evolved beyond the traditional cleansing of names and addresses to increase efficiency rates and postal discounts for marketing campaigns. The current accepted definition for Data Quality is an end-to-end data management process that includes profiling, parsing, cleansing, standardizing, matching, merging, and monitoring phases. This broader Data Quality process is further seen as part of an integrated and scalable data integration platform designed to support reusability across an organization as part of an Integration Competency Center (ICC), providing shared data services across the enterprise. In parallel, the market for Identity Resolution processes has evolved from simple search and match functionality for customer service, fraud detection, and security screening processes to today's highly accurate, high-performance, real-time, cross-language search and match services across multiple applications. This paper describes how Identity Resolution complements and extends the application of Data Quality and data integration processes into business applications.

History and Trends

Traditionally, Data Quality processes were introduced to organizations in ways such as cleansing customer data for marketing campaigns, standardizing data prior to a data migration project, or providing matching functionality as part of a business application customer-screening process. As business requirements have changed, organizations today recognize that a proactive approach to Data Quality can enable a wide range of business imperatives, including compliance, improved decision-making, increased operational efficiency, and cost-reduction initiatives. Business intelligence and data governance applications require Data Quality scorecards and Data Quality monitoring processes. Supply chain management processes require Data Quality business rules for all master data types, including customer, supplier, product, asset, and financial data. Fraud, security, and screening applications require high-precision matching to avoid high-risk mistakes. Data entry applications require high-performance, real-time search and match capabilities to achieve the response times required by the business processes. Master Data Management (MDM) and Customer Data Integration (CDI) applications depend on an end-to-end Data Quality process.

Similarly, Identity Resolution processes have evolved to assist organizations in dealing with the challenges of handling identity data: the information that specifically and accurately identifies a client, a prospect, a supplier, a taxpayer, a criminal suspect, a product. By its very nature, identity data is subject to unavoidable error and variation—spelling and typographical errors, transliteration differences, nicknames, abbreviations—that can compromise the performance of basic search and match processes, generating false positives or missing matches entirely.

As reliable, fast identity searches—in addition to accurate matching, duplicate discovery, and relationship linking—have become fundamental requirements of many applications, from Customer Relationship Management (CRM) to Anti-Money Laundering (AML), more sophisticated Identity Resolution technology and techniques have become critical.

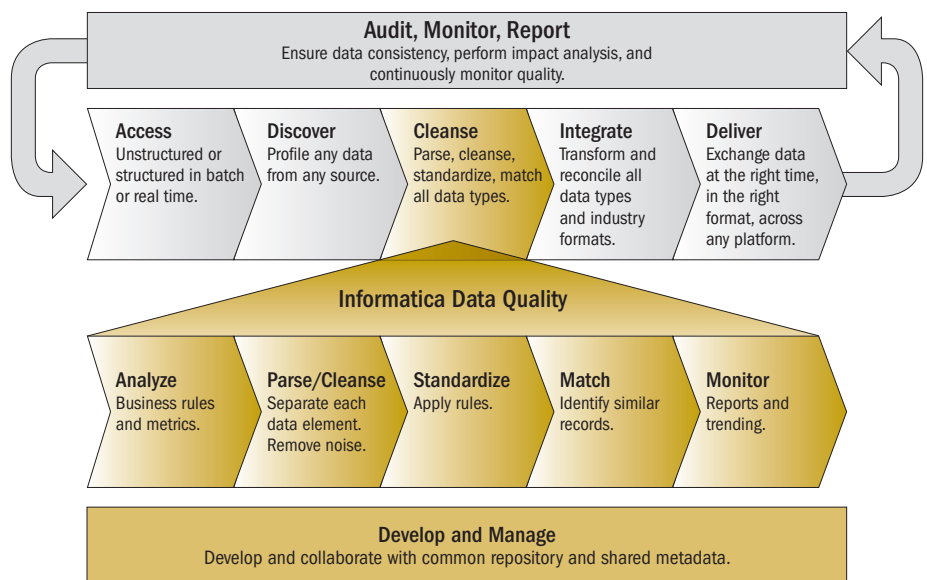
Accurate, high-performance Identity Resolution must perform for uncommon names as well as for very common words. This is an extremely difficult challenge when a database of 100,000,000 people may contain 100,000 John Smiths, Juan Rodriguezes, or 1 Main Streets

Data Quality and Identity Resolution Projects

Most experts agree that the ideal Data Quality process, deployed in real time or batch, follows the standard steps of profiling, parsing, cleansing, standardizing, matching, merging, and monitoring. There are many use cases that rely on some or all of these process steps to fulfill the primary requirement. For example, compliance-related projects require Data Quality reporting and monitoring to build confidence in the data. Data migration projects require profiling to ensure the data movement process does not fail. Data warehouse projects require cleansing to ensure the data in the warehouse is high quality for business decision-making.

While the concept of cleaning, enhancing, or otherwise standardizing names and addresses to make searching and matching easier is very enticing, there can be a conflict between the idea of “cleaning” and the “unpredictability” of identity data. Incorrect assumptions made by formatting and cleaning routines can corrupt good data, and records rejected by the cleaning process can mean lost opportunities to match.

The Informatica Product Platform Automating the Entire Data Integration Lifecycle



What if the primary requirement is to search and match (e.g., for high-risk applications or where there is a high cost or impact of missing a match), and any attempt to cleanse the data before matching may introduce additional errors to the matching process? What if a customer does not want to be matched or linked and intentionally provides false data? What if the data is sparsely populated or partially incomplete? What if there are spelling, typing, or transcription errors; nicknames, synonyms, or abbreviations; foreign or anglicized words; prefix or suffix variations; the concatenation or splitting of words; noise words or punctuation; casing or character set variations? What if the number of sources, the time to market and/or the available project resources dictate that cleansing is not currently appropriate or a realistic project deliverable?

Customer Relationship Management and Master Data Management systems require real-time search and match functionality to support the core processes of data management. A range of fraud (e.g., Anti-Money Laundering) and security screening processes (e.g., in government agencies such as border protection) require high-performance, real-time searching and matching built into their core applications. Global organizations require cross-language matching to gain a single, global view of customers. In many cases, the organization will not own—and therefore cannot modify—the source data, which may have missing fields, errors, or a different structure.

Financial institutions use identity data for many critical processes, such as customer or account setup; new customer credit and loan applications; fraud reports; Anti-Money Laundering; and bankruptcy, to name a few. All of these processes involve pulling data from a variety of systems, many of which will be legacy and will have non-standardized data that needs to be searched and matched (see Figure 1 below).

Government departments and agencies utilize identity data (concerning citizens, voters, residents, employers or employees, overseas visitors, registered organizations, and multinationals) for a range of processes, such as identification; welfare eligibility and tracking; healthcare and community services; police, court, and intelligence systems; tax systems; customs and immigration; company and business registers; patent and trademark registers; license and vehicle registers; address searches; and directory inquiries. Often the government agencies are not “allowed” to change the data for some sources and therefore a cleansing or standardizing strategy may not completely answer their requirements.

Insurance companies use identity data for many critical purposes, such as existing customers or policies; new customers or policies; rejected policy application files; fraud reports, bankruptcy, and other lists; government-sanctioned alert lists; marketing databases; call centers; or any other area where the names and addresses of people, organizations, products, or other entities need to be searched, matched, grouped, screened, or linked. In an ideal world, all the data sources will be cleansed and standardized as an early project goal to enable the most accurate match results; however, it is sometimes not a realistic short-term goal, as the data is owned by many different organizations.

Search and match software is used in police operations, criminal history and intelligence, missing persons, court administration, roads and traffic infringement, penal and probation systems, stolen property, criminal statistics, child welfare and maintenance, and entry alert. In many of these applications, “smart” fuzzy indexing is needed to get around the challenges of missing or false data.

Healthcare companies and organizations use patient and provider identity data for many critical purposes, including patient and provider inquiry systems, pharmaceutical and poison databases, immunization databases, and fraud discovery and investigation.

Data in tax and revenue 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. Tax systems must track people over long periods of time, and there are many avenues for tax avoidance and fraud within today’s complex tax systems. To discover tax and noncompliance, systems must deal with multiple sources of taxpayer information, with identity data (e.g., name, address, TIN, SSN) of varying quality.

Product Category	Account	Last Name	First Name	MI	Street	City	Zip
Online Checking	7A-301-22451	Jones	Jonathan	J	124 Oxford	Redwood City	94061
Personal Savings	11-301-13421	Jones	Jackie	J	32 Selby Lane Atherton, CA 94027		
Personal Savings	1A-492-8874112	Jones	Jon		124 Oxford	Redwood City	94061
Commercial Checking	SB3-001-9865KL	Jones Concrete & Masonry			PO Box 62	San Francisco	94111

Figure 1: Variant data in the banking industry.

Common Challenges

Reviewing these use cases, it becomes clear that they share a number of common challenges, particularly as they are extended to interact with broad, global systems. In these instances, the characteristics of all such application environments typically include:

- Data in the records to be matched is sparse or limited
- Original data entry is not focused on accuracy
- The business application does not control the data and may not alter it (cleansing is not an option)
- Screening is needed
- Data volumes are massive
- User needs fast response times
- Data must be matched across languages, character sets, and country scripts (e.g., Arabic versus Latin script)

That last challenge is critical. Most large identity databases contain data from multiple languages, countries, and cultures that often have different structures, follow different parsing rules, and have different variation characteristics (see Figure 2 below). Also, if transliteration, Romanization, character set conversion, and other such transformations are employed, a new class of error and variation is introduced.

Regardless of what formatting or standardizing techniques may be later used on foreign data, for example, to help with its recognition or matching, the integrity of the data should be maintained throughout the process (regardless of whether it is converted to Unicode along the way). The data should be captured in as free-form a format as possible, such as a single long field for the full name or an array of lines for the address. In addition, insofar as the original character set, country, or language is known, this information should be captured and stored with the data. This will later help search and match systems apply country-specific rules to the data from different countries or populations. (This need for population-specific rules arises because of the conflicts that exist between different countries and languages, both at the character level as well as the token level.)

As a simple example of the challenges of dealing with international data, consider the abbreviation St. In an English or American address, it stands for the word “street”; in a French address, it means Saint.

Record 1 	Record 2 	Record 3 
Peg Mc Cary	Margaret MacClary	Grietje McClary
Abdulaziz A Rahman Al Sugair	Abd A Rhman Hammed Al-Shugair	ريق صلل ن م ح ر ل ادب ع زي ز ع ل ادب ع
George Papadopoulos	Georgios Papadopoulos	Γεώργιος Παπαδόπουλος
Saito Kyoko	齊藤 京子	Kyouko Saitou
William Kwok	W. Kwok Ki Hoh	Mr. Billy H Kwok

Figure 2: Different types of character sets must be evaluated.

Data Quality and Identity Resolution

Reviewing these projects we've previously discussed makes it possible to outline the requirements of a broader Data Quality search and match solution. It should feature the following:

Requirement	Description
Access to identity data	• Broad data access from a range of sources using a data integration platform
Profiling	• Profiling the sources to identify the best attributes to choose for indexing and matching
Data cleansing and standardizing	• Cleansing and standardizing of sources where applicable to increase match accuracy
Transformation	• Transformation of sources to an optimum data structure for searching and matching
Real-time, high-performance, highly accurate, searching and matching	<ul style="list-style-type: none"> • "Smart" fuzzy indexing, to overcome spelling, phonetic, transliteration, and multi-country data, missing/out-of-order words, and other errors and variation • Flexible search strategies, to balance performance and comprehensiveness of search • Cross-language searches • Matching algorithms that emulate a human expert's ability to determine a match based on numerous attributes • Speed and scale, in order to perform high-volume searches quickly against very large databases
Monitoring	Measure Data Quality over time using a set of Data Quality dimensions (e.g., completeness, conformity, consistency, duplicates, integrity, and accuracy)

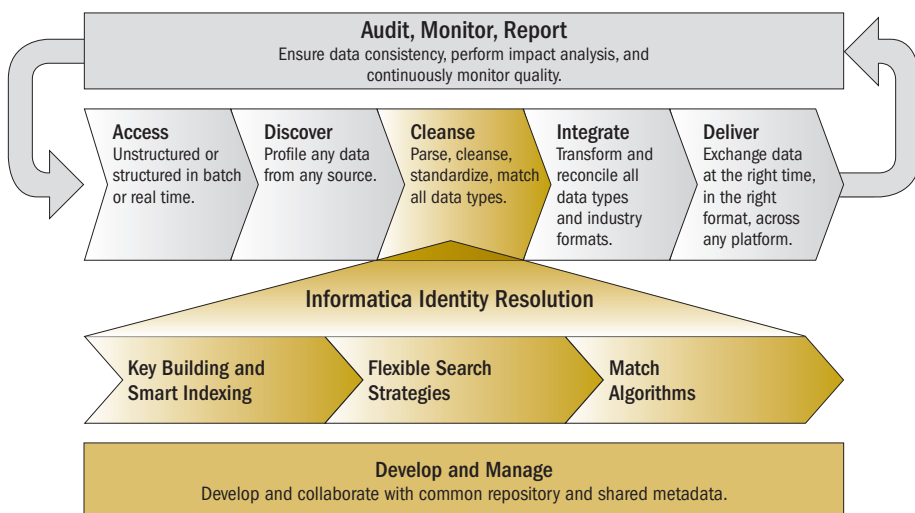
The key to effective Identity Resolution technology is emulating an intelligent business user's ability to determine a match based on a variety of factors, overcoming spelling, phonetic, and other errors and omissions in the data while offering the speed and scale to perform high-volume searches quickly against very large databases.

At the heart of such an approach will be 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.

The degree to which complete Data Quality and Identity Resolution processes are implemented within an organization is affected by several interrelated dependencies, including business drivers, levels of control over the data, and the degree of data fragmentation. A broader data cleansing and standardizing process will always improve the accuracy of matching processes. A platform to access all identity data will enable faster results.

The Informatica Product Platform

Identity Resolution Complements and Extends Data Quality





Conclusion

Data Quality is evolving rapidly. Increasingly, it is about providing complete, accurate, up-to-the-minute information to broad, global systems that support strategic goals. This change is creating a new series of challenges to building effective Data Quality processes, from the need to process more data in less time to the necessity to accurately handle multiple languages and character sets. As a result, an enterprise requires Data Quality solutions that deliver highly accurate, high-performance, cross-language searching and matching with the most accurate results possible. A data integration platform, a broader Data Quality process to cleanse and monitor Data Quality, and a highly accurate and flexible matching engine, which can operate reliably on the original data regardless of effort, variation, format, or language differences, are necessary to deliver Identity Resolution business applications.

FAQs

Q: What is Identity Resolution?

Identity data—the information that specifically and accurately identifies a client, prospect, customer, supplier, taxpayer, criminal suspect, or product—is a special case, subject to different rules and unavoidable errors and omissions. Identity Resolution helps meet the challenges of dealing with identity data by emulating an intelligent user’s ability to recognize matches but also to discover connections—between people, accounts, products—that might be hidden in the data.

Q: Is Identity Resolution the same as Data Quality?

Identity Resolution complements and extends Data Quality. Data Quality suites deliver profiling, parsing, cleansing, standardizing, matching, and monitoring solutions. Identity data will always contain a certain amount of unavoidable and conflicting error and variation, as well as a percentage that cannot be corrected by other Data Quality processes. In addition, Identity Resolution technology can perform cross-language searching and matching and highly scalable, real-time searching and matching, which typically are not a focus of the Data Quality match step. These capabilities are very complementary to Informatica Data Quality, and customers will now benefit from very specialized capabilities for customer Identity Resolution built up over 20 years.

Q: Why is Identity Resolution important?

Identity Resolution is a critical component of many different types of applications—for example, Customer Relationship Management, Master Data Management, and Customer Data Integration. Informatica’s software allows the application to not only mimic the expert business user in order to find similar identity records, it can also perform this function across more than 60 languages. Despite data having errors, variation, and duplication, Identity Resolution delivers the highest possible reliability when searching, matching, or grouping data based on names, addresses, descriptions, and other identification data. This technology has many applications, including the following (not exhaustive):

- Fraud detection (credit, Telco, and insurance industries)
- Anti-Money Laundering (banking and finance industries)
- Immigration control, border security, or other people issues associated with law enforcement
- Patient record matching (healthcare and insurance industries)
- Identity matching (data bureaus and credit agencies)
- Tax payments (federal, state, and municipal agencies)

Q: Why is cross-language match so important?

In today’s economy, every company that wants to succeed does business with customers, partners, and suppliers across the globe. Informatica is the only company to offer the ability to search and match across multiple languages simultaneously. A company like HP, with customers in 178 countries, needs the ability to match names represented in multiple languages, not just across a single language (e.g., “Hanjin Shipping” in Korean script can match against that same name spelled out in Latin characters).



INFORMATICA[®]

Worldwide Headquarters, 100 Cardinal Way, Redwood City, CA 94063, USA
phone: 650.385.5000 fax: 650.385.5500 toll-free in the US: 1.800.653.3871 www.informatica.com

Informatica Offices Around The Globe: Australia · Belgium · Canada · China · France · Germany · Japan · Korea · the Netherlands · Singapore · Switzerland · United Kingdom · USA

Copyright © 2008 Informatica Corporation. All rights reserved. Printed in the U.S.A. Informatica and the Informatica logo are trademarks or registered trademarks of Informatica Corporation in the United States and in jurisdictions throughout the world. All other company and product names may be trade names or trademarks of their respective owners.

6859 (05/30/2008)