

Identity Resolution & Data Quality Solutions

Matching, Address Validation, Profiling, Screening, Compliance

MerlinMerge® SpeedPro

Product Data Sheet

I. INTRODUCTION

Founded in 1993, Intelligent Search Technology, Ltd. (IST) has devoted its resources to the development of the fastest and most accurate search and matching software. In fact, an independent benchmark test conducted by the New York State Division of Criminal Justice demonstrated that IST's software significantly outperformed the competition. IST is a privately-held corporation and has shown continued growth since its inception. We are located in the northern suburbs of New York City in White Plains, NY.

Our *MerlinMerge® SpeedPro* product is an easy-to-use data-cleansing tool that intelligently searches for and identifies duplicate records. Searches can be based upon any type of data, such as names and dates, companies and phone numbers, and other criteria. It uses IST's proprietary fuzzy search and matching technology to identify and/or eliminate duplicate records within one or more data sources.

MerlinMerge SpeedPro offers:

- ▶ Merge/purge operations.
- ▶ Batch Match processing.
- ▶ Deduplication processing.
- ▶ Household determination.

MerlinMerge SpeedPro works with regular text files, however, can connect to SQL Server, Oracle, MSAccess, DB2, Sybase, and Teradata.

Our proprietary searching and matching technology is also the underlying intelligence behind all of IST's other products, including:

NameSearch® - This powerful fuzzy searching and matching product overcomes data variations due to misspellings, transpositions, acronyms, abbreviations, nicknames, etc. The result is more accurate search results while virtually eliminating false positives. NameSearch comes with an extensive nickname rulebase which includes not only English, but also multi-cultural names.

ISTwatch® - ISTwatch is an OFAC compliance software solution that helps companies perform identity screening, while complying with government and industry regulations. It screens against several government lists, including OFAC, FinCEN, FBI's Most Wanted Terrorist and Hijack Suspect lists, Denied Persons List, and international lists from the European Union, United Kingdom, United Nations, France, and more. In seconds, ISTwatch performs terrorist searches interactively or in batch, delivering throughput in excess of 25,000 records per second. And its API allows easy integration into other enterprise applications.

CorrectAddress® - Utilizes the same powerful searching engine as our NameSearch, but applied against the US Postal Service database. It is a CASS™-certified, address standardization and validation tool that will help to better manage address lists and ensure that mail sent is deliverable. It enables users to cleanse, verify and standardize their addresses in real-time and batch mode, and also comes with a powerful API enabling easy integration into most commercially available databases, or custom web-based or other enterprise applications. In addition to validating the addresses, CorrectAddress also adds the Carrier Route, LOT, ZIP+4 and Delivery Point Barcodes (DPBC) to every deliverable address. Add-ons available for CorrectAddress include Delivery Point Validation™, Geocoding, and LACSLINK.

CorrectAddress® Plus -Includes all of the functionality of CorrectAddress and MerlinMerge SpeedPro, packaged together in one powerful application.

Identity Resolution & Data Quality Solutions

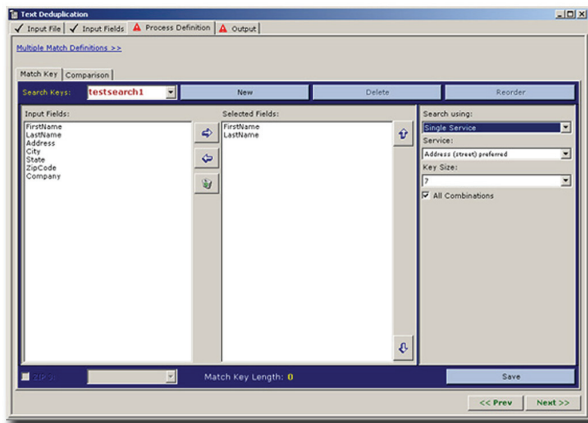
Matching, Address Validation, Profiling, Screening, Compliance

II. MERLINMERGE SPEEDPRO CAPABILITIES

A. MATCHING, SEARCHING, COMPARISON

MerlinMerge SpeedPro enables systems to find and match records using personal names, corporate names, addresses, social security, phone and account numbers and other identifying information.

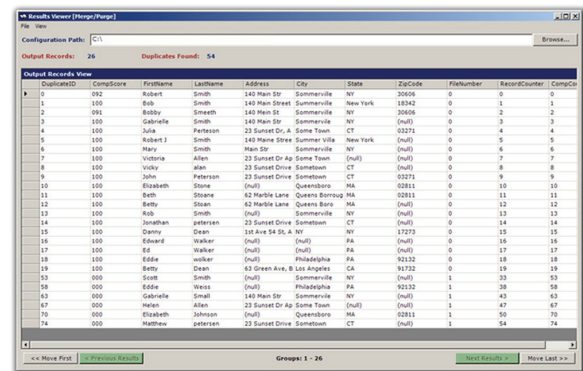
The first aspect of the matching process involves the creation of intelligent keys. This facility is used for the retrieval of records regardless of variation.



Matching is achieved through advanced comparison functions that utilize neural net technology, rule-based intelligence and advanced heuristical pattern recognition. The matching functionality will deliver comparison scores that approximate values generated by an individual with significant linguistic expertise. These comparison routines enable systems to make decisions without human intervention.

B. DEDUPLICATION (DEDUPING)

The process of identifying and removing duplicate records is called "deduplication" or "deduping". Deduplication is a key operation when dealing with large volumes of data and especially when integrating data from multiple sources.



ID	Confidence	FirstName	LastName	Address	City	State	ZipCode	PhoneNumber	RecordCounter
0	100	Robert	Smith	140 Main St	Summerville	NV	20006	0	0
1	100	Bob	Smith	140 Main Street	Summerville	New York	10342	0	1
2	100	Bobby	Smith	140 Main St	Summerville	NV	20006	0	2
3	100	Gabrielle	Smith	140 Main St	Summerville	NV	(null)	0	3
4	100	Jane	Peterson	23 Sunset Dr A	Some Town	CT	02213	0	4
5	100	Robert J	Smith	140 Main Street	Summerville	New York	(null)	0	5
6	100	Mary	Smith	Main St	Summerville	NV	(null)	0	6
7	100	Victoria	Alan	23 Sunset Dr Ap	Some Town	(null)	(null)	0	7
8	100	Vicky	Alan	23 Sunset Drive	SomeTown	CT	(null)	0	8
9	100	John	Peterson	23 Sunset Drive	SomeTown	CT	02213	0	9
10	100	Elizabeth	Stone	(null)	Queensboro	MA	02811	0	10
11	100	Beth	Stoane	62 Marble Lane	Queens Boroug	MA	02811	0	11
12	100	Betty	Stoan	62 Marble Lane	Queens Boro	MA	(null)	0	12
13	100	Rob	Smith	(null)	Summerville	NV	(null)	0	13
14	100	Jonathan	petersen	23 Sunset Drive	SomeTown	CT	(null)	0	14
15	100	Danny	Dean	1st Ave 54 St A	NV	NV	17273	0	15
16	100	Edward	Walker	(null)	(null)	PA	(null)	0	16
17	100	Ed	Walker	(null)	(null)	PA	(null)	0	17
18	100	Eddie	walker	(null)	Philadelphia	PA	19132	0	18
19	100	Betty	Dean	63 Green Ave	B Los Angeles	CA	91732	0	19
20	100	Sam	Smith	(null)	Summerville	NV	(null)	0	20
21	100	Eddie	Weiss	(null)	Philadelphia	PA	91732	1	21
22	100	Gabrielle	Small	140 Main St	Summerville	NV	(null)	1	22
23	100	Hean	Alan	23 Sunset Dr Ap	Some Town	(null)	(null)	1	23
24	100	Elizabeth	Johnson	(null)	Queensboro	MA	02811	1	24
25	100	Matthew	petersen	23 Sunset Drive	SomeTown	CT	(null)	1	25

The truth is that all databases contain duplicate records, for a number of reasons - spelling and transposition errors, misheard names and addresses, different people entering data often from multiple locations, merging external data, etc. This can present a huge problem for an organization but often it does not emerge until a larger project is undertaken, or the organization expands or an unhappy customer complains. As data is collected from a wide variety of sources, the number of duplicate entries will keep growing. Most databases contain at least 3% duplicate records and in many cases, significantly more.

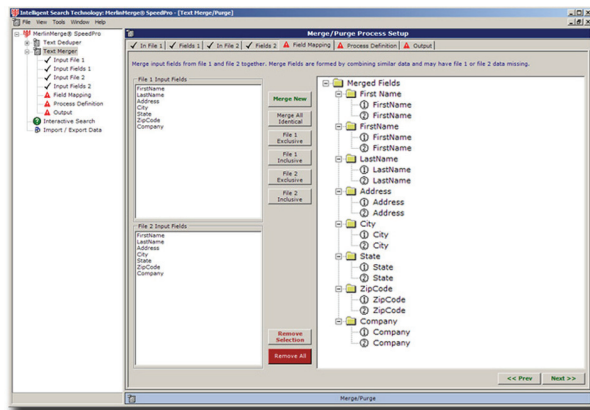
The main challenge in this task is identifying when a pair of records refers to the same entity in spite of various data inconsistencies. *MerlinMerge SpeedPro* is specifically designed to provide a solution to this issue.

Identity Resolution & Data Quality Solutions

Matching, Address Validation, Profiling, Screening, Compliance

C. MERGE/PURGE

Merging data refers to the process of integrating data from multiple sources. When combining the data from two or more separate lists into a single one, it is often the case that certain records from the lists repeat. The process of removing the repeated records is referred to as “purging” the records.



Whether we refer to redundant data, wrong data, missing data or miscoded data, every company has some of each, probably residing in several different departments.

MerlinMerge SpeedPro provides an intuitive interface designed specifically for merging two separate lists into a single one.

D. HOUSEHOLD DETERMINATION

A variation of deduplication is called household determination. Household determination involves combining records for different individuals living at the same address into a single entry. For example, records for “John Smith”, “Jane Smith” and “Junior Smith” containing the same address information become a single record for “The Smith Household”. The same can be done with people working in the same company.

A householding merge/purge considers last name and address, including variations in the name and address. Records with unique last names and the same address are not considered a match.

E. DATA QUALITY

The problem of poor data quality processed by an information system is widespread in the industrial, government and academic environments. Poor data quality has a negative impact on the competitiveness of an organization and can cause many other problems especially when working on bigger projects. The *MerlinMerge SpeedPro* software reduces the cost of doing business by improving the accuracy and usability of your data.

F. PERFORMANCE

Computer technology always faces the issue of speed versus accuracy. This is especially relevant when dealing with large data sets. Tasks such as comparing two strings take a lot of processing time. In addition reading through millions of records is very I/O intensive. The more processing time spent, the better the results, but sometimes fast performance is more important.

MerlinMerge SpeedPro does an excellent job at finding the right balance between speed and accuracy. Through many years of experience, testing and optimization our technology experts have fine-tuned *MerlinMerge SpeedPro* to achieve a very high degree of accuracy while preserving fast performance. Additionally you are able to change the matching criteria to achieve a balance of speed and accuracy that is custom fit to your organization.

Identity Resolution & Data Quality Solutions

Matching, Address Validation, Profiling, Screening, Compliance

III. INTELLIGENCE

MerlinMerge SpeedPro uses the searching and matching intelligence of our extremely powerful NameSearch® technology to achieve unparalleled accuracy and speed while overcoming variations due to misspellings, transcriptions, transpositions, acronyms, phonetics, sequence differences, nicknames and many other common errors found in data.

A. SPELLING AND KEYBOARD ERRORS

Spelling and keyboard errors account for many of the duplicates that may be found in a database. Using intelligent key building and advanced comparison routines *MerlinMerge SpeedPro* successfully overcomes spelling errors including: multiple typographical errors, letter transpositions, incomplete words, etc.

B. RULEBASE EXPERTISE

A rulebase expert system is used to identify nicknames. Entities such as "Bill," "William," "Bob" and "Robert" are often used interchangeably to identify individuals. The rulebase is also used to identify noise words. Noise words are elements in a name that do not help in the identification of a candidate. Examples of noise words are "Incorporated," "Corporation," "Limited," "Junior," "Senior," "Avenue," and "Street." Often there are times where elements in a name contribute to the identity but should be treated as less important. In these cases, the rulebase does not treat them as noise words but recognizes that they are less significant. Some examples are: associate, board, international and services. Other variations are caused by the use of common prefixes. Names such as "McDonnell" are confused with "MacDonnell." Prefix recognition provides the facility for handling these classes of problems. The rulebase can also recognize diminutives. Frequently there are names that end in a diminutive such as "i.e." or "y." In these cases, it is useful to identify the root and apply the rule. For example, you would want "Bill," "Billie" and "Billy" to find "William" or "Willie."

BILL YARA	WILLIAM YARA
BOBBY KENNEDY	ROBERT KENNEDY
JIM P PHILLIPS SR	JAMES P PHILLIPS
SMITH AND ASSOCIATES	SMITH
MCDONELL CORPORATION	MCDONELL
MR MATT J THOMAS	MATTHEW J THOMAS
MARINA DELSOLE	MARINA DEL SOLE
DR LEONARD MACCOY MD	LEONARD MCCOY

C. PHONETIC ERRORS

Discrepancies caused by phonetic errors account for 20-25% of all name variations.

Traditional solutions such as Soundex and NYSIIS used for solving name variations only deal with phonetic errors. These solutions involve the standardization of easily confused sounds. For example, "PH"'s would be treated as "F"'s. Linguistic rules are generated to phonetically tokenize a name. These phonetically tokenized words serve as the basis for name retrieval. In some instances these rules help to find names that are difficult to spell. Unfortunately, the distribution pattern of common names becomes even more skewed. For example, inquiries on "John" also return "Joan," "Jim," "Jane," "Jimmy," "Jenn" and other names which fall in the "JAN" phonetic pattern. By aggravating the skew in distribution of names, both quality and performance are sacrificed.

MerlinMerge SpeedPro addresses problems due to phonetics by employing analysis routines to determine the extent of phonetic tokenization. This enables *MerlinMerge® SpeedPro* to overcome problems due to phonetics without the negative consequences incurred with all other methods of name search.

Identity Resolution & Data Quality Solutions

Matching, Address Validation, Profiling, Screening, Compliance

D. SANITIZATION

While processing the data, *MerlinMerge SpeedPro* expedites a process called *sanitization* that removes noise characters, extra spaces, control characters and converts lower case letters to uppercase. Examples of noise characters are: "@," "#," "\$," "%," "^," "&," "*", "(", ")", "{", "[", "]". The following characters are handled separately and have special meanings: commas, hyphens, and quotes. Commas usually indicate the insertion of a last name. Sanitization places words followed by commas at the end of the string. Quotes are deleted and the space between them is removed. A space replaces the hyphens.

Before Sanitization	After Sanitization
Scott Lions	SCOTT LIONS
Smith, John F.	JOHN F SMITH
Rose Stone-Shield	ROSE STONE SHIELD
James O'Tool	JAMES OTOOL
James O. Tool	JAMES OTOOL
Owen, Tool, James	JAMES OWEN TOOL
# Williams , \$Richard	RICHARD WILLIAMS

The sanitization process also uses a small rulebase. The rulebase is applied after all the alpha characters have been converted to upper case letters and extra blanks are removed. This rulebase is used to recognize words that contain noise characters or prefixes that could be affected by the sanitization process.

Before Sanitization	After Sanitization	Sanitization (without rulebase expertise)
c\o	CARE OF	C O
Mc Donald, Old	OLD MCDONALD	MC OLD DONALD
%	CARE OF	

E. WORD SEQUENCE VARIATIONS

Many searching problems are caused by sequence variations. The inability to determine the order of words for a particular entity occurs at both data entry and inquiry time. The name "Frank Lee" for example, could have been "Lee Frank." This problem is particularly pervasive in company names. Names such as "International Business Machines," "Anderson Consulting" and "Kemper Insurance Company" are examples where the left-most word is most significant. Conversely, "Edward S. Gordan Real Estate Company" and "Paul Mitchell Hair Products" are examples where the left-most word is less significant. The inability to predict the significant name with respect to word position causes many searches to fail.

Merging foreign database files causes other sequence variations. This frequently occurs when external lists are purchased or companies consolidate information. Inconsistent methodologies for data capture make the standardization of name fields impossible. Aggravating the sequence problem are those instances in which company names are intermixed with personal names. All of these factors, in addition to human error, contribute to identification problems caused by sequence variations. *MerlinMerge® SpeedPro* provides a facility for handling these problems.

To understand this better an analogy can be drawn between a telephone book and a database system. When looking for "Frank Lee," the "L" section is searched. If the name is not there, the search is continued by looking in the "F" section. In order to find "Frank Lee" we had to search two separate sections of the phone book. Suppose we were looking for "Frank Lee Ray." To ensure success we must search all the permutations. This is an extremely arduous and time consuming process for both people and computers. By listing "Frank Lee" in both the "L" and "F" sections, regardless of order, only one section would need to be searched.

Using this approach, *MerlinMerge SpeedPro* is able to overcome word sequence variations without sacrificing performance.

Identity Resolution & Data Quality Solutions

Matching, Address Validation, Profiling, Screening, Compliance

F. ACRONYM RECOGNITION

Corporate name searching concretely illustrates the pragmatic difficulties in developing solutions that find correct information without missing likely candidates. People readily understand the similarities between "Triple A towing" and "AAA towing" yet computerized systems would need to employ a knowledge-based algorithm to recognize the relationship between "Triple A" and "AAA."

The deployment of intelligence through knowledge-based systems greatly benefits search and matching algorithms by identifying nicknames, shortened forms, noise words and other circumstances that require experience to return a more comprehensive result set. However, knowledge-based systems are limited by the breadth and depth of their lexicon. Contrary to names such as "IBM" and "AT&T," the vast majority of acronyms lie outside the scope of knowledge-base processing. For example, our clients often used the "IST" acronym interchangeably with "Intelligent Search Technology" yet it would be unreasonable to expect the inclusion of "IST" in a knowledge-based system.

MerlinMerge SpeedPro with its corporate search algorithms and acronym recognition functionality significantly advances the ability to seek and match corporate name data.

IV. CURRENT LIST OF CUSTOMERS AND PARTNERS

IST has over fifteen hundred satisfied customers, including many Fortune 500 companies, Federal, State, and Local government agencies, and companies representing Banking/Financial Services, Consulting, Law Enforcement, Defense, Education, Healthcare, Retail, Manufacturing, Insurance, Entertainment, and many other industry sectors. Some examples are:

AARP	JPMorgan Chase
ACNielsen	Lexis Nexis
American Express	Mass Mutual
American Financial Group	Microsoft Corporation
American Red Cross	Monsanto
AT&T	National Education Association
BAE Systems	Neiman Marcus
Business Objects	State & Local agencies from nearly all the 50 States
Choicepoint	Transunion
Citigroup	Turner Broadcasting
Epson America	U.S. Dept. of Defense
General Electric	U.S. Social Security Administration
Hilton Hotels	Wells Fargo
Intuit	Zurich Financial Services