

# Semantic Identity Resolution Platform

Advanced Master Data Management  
with web-sourced reference data

## Contents

Introduction .....	3
Customer Master Data Management .....	3
The Shortfall of Existing MDM Solutions .....	3
Digital Trowel's Semantic IR Platform .....	5
Information Extraction .....	6
Data Cleansing .....	6
Match and Merge (deduplication) .....	7
Summary .....	8

## Introduction

Digital Trowel (DT) stands at the forefront of information extraction technology. DT's Semantic Identity Resolution (IR) platform encompasses the core modules required to enhance existing customer master data management (MDM) solutions. DT's proprietary technology yields near-human intelligence and is unique in its ability to process unstructured text and provide both rapid and accurate semantic fuzzy matching using web-sourced reference data.

## Customer Master Data Management

### Accurate and comprehensive representation of customer data

MDM is a rapidly developing discipline in which IT and business enterprises collaborate to achieve accuracy and accountability of shared master data assets. Enterprises are increasingly turning towards MDM as a means to optimize business processes by rationalizing and sharing master data sets.

Customer master data management – also referred to as customer data integration (CDI) – enables the accurate and comprehensive representation of customer information across channels and business lines, ultimately delivering an actionable form to touchpoints.

## The Shortfall of Existing MDM Solutions

### Inaccurate and duplicate customer records

Customer data is replete with information carrying the same meaning but with completely different textual representations. Examples include foreign, abbreviated or incorrect spellings of names of people and companies. Conversely, customer information also includes information with similar textual representation and a completely different meaning. Current MDM solutions utilizing a mathematical matching engine process this data incorrectly, leading to inaccurate or duplicate customer records that require costly and error-prone manual intervention.

### Information is subject to time decay

Customer data information ages quickly. Every day companies hire new employees, move locations and change their contact information. According to the US Department of Labor, employed professionals change jobs every four years. If only 2% of any CDI data set becomes outdated every month, within a year 25% of the database is out of date. Current sources, such as company websites, press releases and other web-sourced content provide up-to-date information; however existing IR solutions lack the tools to read unstructured text on the Internet and to accurately parse and link it to the correct person or company.

### Lengthy processing times as size of database increases

As the amount and type of data that users gather about their customers increases, MDM solutions are having an increasingly difficult time handling all the content. Traditional IR systems that run on record-by-record matching cannot handle the hundreds of millions of data points that a large modern database contains. For example, Digital Trowel regularly matches a database of 400 million records before de-duplicating them.

The key challenges to handling large databases are:

- **Performance.** IR algorithms must perform a very large number of comparisons. The comparisons need to be extremely fast and resource-efficient.
- **Distribution.** As IR is a computational-intensive process, the processes need to be built to easily share the IR workload across multiple computing cores.

## Digital Trowel's Semantic IR Platform

### Converting unstructured text into actionable intelligence

Digital Trowel's Semantic Identity Resolution Platform fills the void left by existing MDM solutions. Based on DT's information extraction technology, the platform constitutes a breakthrough in the areas of information extraction and fuzzy semantic matching, based on web-sourced reference data. DT's proprietary technology yields near-human intelligence, effectively converting unstructured text into actionable intelligence.

DT's fully customizable Semantic Identity Resolution Platform is comprised of three core modules:

- 1. Information Extraction** – Semantic algorithms extract accurate structured data from unstructured free text sources.
- 2. Data Cleansing** – Data is refined through standardization processes and then validated, facilitating high quality match results.
- 3. Match and Merge** – Natural language processing algorithms provide rapid and accurate semantic fuzzy matching of terms with similar meanings but different textual representations and the disassociation of terms with similar textual representations but different meanings. Digital Trowel web-sourced data provides a master reference database larger and more accurate than other MDM providers.

Additional information and examples from each module are provided below.

## Digital Trowel's Information Extraction Technology

*The first implementation of the Hybrid Computational Linguistics Model*

*Digital Trowel's (DT) text mining technology provides the infrastructure for its Semantic MDM Platform. It incorporates the cutting-edge information extraction and semantic matching technology required to deliver near human intelligence.*

*DT has developed the industry's most advanced and accurate text mining technology – CaRE (CRF Assisted Relation Extraction). This technology is the first implementation of the Hybrid Computational Linguistics Model for Information Extraction. CaRE is based on CRF (conditional random fields), a mathematical probabilistic model. CaRE incorporates state-of-the-art supervised machine learning algorithms which learn the semantics of specific words and phrases according to their statistical distribution in text, otherwise known as NER (Named Entity Recognition). The flexible interface between NER (supervised machine learning) and rulebooks (knowledge-based semantic rules) provides synergetic balancing with accuracy results exceeding 90%.*

*CaRE is based on weighted discriminative context-free grammars (CFG), interpretable and trainable as CRFs, featuring a flexible interface between the parsing component and the token classification components such as part-of-speech (PoS) tagger and NER. This interface allows the grammar to selectively modify the classification results and adapt generically trained sequence classifiers to specific domains of relation extraction.*

*In addition, DT's Semantic MDM Platform includes a matching engine based on proprietary algorithms, optimized for extremely fast and accurate semantic fuzzy matching of customer data. This unique approach allows matching of entities with the same meaning but with very different textual representations. It is based on a hybrid of linguistic and machine learning algorithms for the location and extraction of weighted synonyms, rather than the common mathematical algorithms used in other platforms.*

## Information Extraction

### Extracting detailed structured data from free text

The information extraction module incorporates highly sophisticated linguistic and machine learning algorithms that scan unstructured information sources (such as business and professional profiles, announcements, event transcripts, blogs, share price data and a variety of government databases) and extract accurate and detailed structured data. DT's sophisticated engine takes complicated free text and extracts structured entities and relationships (text / HTML in and XML out).

For example, the platform will extract relevant information about an individual from a profile on a corporate website. Yielded entities may include names, businesses, contact details, business events, employment history, education records, board membership, associations and more.

## Data Cleansing

### Standardizing and adding value to raw data

In the data cleansing module, the raw data extracted from numerous sources is first broken down into entities such as company, name, address, phone numbers, web links, employment history, education records and more. We then refine and normalize these entities.

This module adds value to the raw data by providing more than 20 different types of data cleansing processes. By correcting misspellings and standardizing abbreviations, the data cleansing module facilitates accurate matching and merging. In addition, it simplifies data storage and improves overall processing. Several examples of data cleansing are listed below:

- **Names of people and companies** – Parses and normalizes names of businesses and people; handles spelling mistakes, slang and abbreviations.
- **Employment history** – Divides data into core position name, specialty and region. For example: “Senior Vice President, Sales and Operations, Asia-Pacific”, where “Senior Vice President” is the core position name, “Sales and Operations” is the specialty and “Asia-Pacific” is the region.
- **Education history** – Standardizes the names of educational institutions and extracts education history with exact dates from biographies and curriculum vitae.
- **Addresses** – Validates and converts data into a standard format including nine-digit ZIP code, longitude and latitude. Partial addresses such as “Greater NY Area” are supported.

- **Dates** – Takes special account of partial dates (e.g. “summer 2008” or simply “2008”) and relative dates (e.g. “seven years ago”). Partial dates are kept along with an accuracy level (full, month and year, season and year, year only). Relative dates are converted to actual dates at the moment of article publication.
- **Telephone numbers** – Divides number into country code, area code and phone number. The combination of the area code and the exchange is validated.

### Automatic quality control and data validation

The platform includes an automatic quality control module that periodically scans all content that was processed and changed. Valid records are approved for further processing and invalid records are marked for further testing. The platform includes a large collection of validation patterns comprising expressions that validate various data entries including names, addresses, phone numbers, web links and so forth. An example of a simple validation pattern would be “all phone number beginning with 555.” Valid data entries known in advance – white lists – are included to prevent filtering accurate data. Invalid data entries known in advance – black lists – are included to exclude irrelevant data and profanities.

## Match and Merge (de-duplication)

### Data corroboration, matching and merging

The elimination of duplicate records is performed using the platform’s unique understanding of natural language. The process, based on DT’s proprietary algorithms, identifies duplications and recognizes spelling variations, synonyms and antonyms. The matching module is also capable of conducting fuzzy matching on detailed profile information in employment and education history. For example, the system will recognize that “IBM,” “International Business Machines,” and “Big Blue” all refer to the same company. It also standardizes names of businesses and educational institutions and detects and eliminates duplicate foreign spellings of names (e.g. “Filip” and “Philippe”).

The process identifies variations and abbreviations of names (“Bob” versus “Robert”), alternative title displays (“Dr. Robert Smith” versus “Robert Smith MD”) and nicknames (“Peggy” versus “Margaret”). In addition, the platform’s semantic fuzzy matching automatically extracts domain specific synonyms such as “Tom’s Diner” and “Tom’s Restaurant” in order to increase the probability of a true match. The platform further improves the quality of data records by disassociating terms with different meanings and similar textual representations. For example, different terms such as “Audio” and “Auto” in “Southern Audio Services” and “Southern Auto Services” are detected in order to reduce the number of false matches. Matching is also performed on partial addresses such as “Greater NY area” and “New York City”.

All matched records are then processed in order to corroborate their data and link data fields. Several parameters are used for this stage including source priority, credibility, data currency and more. Every matched set of companies is transformed into a single company record. All of the auxiliary records and original data such as addresses, phone numbers, web links, and employment records are linked to the new master record. Each data segment is then examined to ensure that there are no missing sections. Subsequently, the merged data records are stored in a database.

### Data Append

Once a data segment is fully standardized, matched and merged, the system can append data to any missing fields. For example, a company record that misses a phone number is sent back to a phone enrichment process that looks up the company on the web and searches for matching phone numbers.

### Configurable Rules

The matching engine works against a configurable set of rules that can be calibrated and optimized per data source. Source-specific rules can be used if there exists some source specific knowledge such as the Secretary of State registration number or DUNS number.

### Scalability & Performance

The matching engine is performance-optimized, able to process hundreds of millions of records in a few hours. Due to a sophisticated clustering technology, the matching algorithms are linearly scalable so that doubling the number of computing cores nearly doubles the processing throughput.

## Summary

DT's unique hybrid approach of semantic Identity Resolution generates a singularly accurate version of the truth based on its unmatched ability to read unstructured texts. DT's platform achieves actionable customer intelligence by incorporating cutting-edge information extraction, data cleansing and de-duplication technologies. DT's web sourced reference data provides the ultimate master database for high quality and up-to-date matching. DT's unique solution achieves near-human intelligence, increasing the efficiency and quality of any MDM system.

- 1 Park 80 Plz ▪ W 2nd floor Saddle Brook, New Jersey 07663-5242 ▪
  - Tel: +1 (201) 931-9200 ▪ Fax: +1 (646) 349-3532 ▪
  - [www.digitaltrowel.com](http://www.digitaltrowel.com) ▪ [info@digitaltrowel.com](mailto:info@digitaltrowel.com) ▪