VIEW POINT

SOV CLEANSING FOR INSURANCE

Abstract

Data quality is the key to achieving data consistency, completeness, timeliness and conformity in the data journey. It enables the ability to take control of data, aids empowered business decisions, and; if not properly curated, can be a stumbling block for data-driven organizations that use data to derive intelligent and informed business decisions. Unfortunately, as organizations grow organically, data entities within them also grow like hydra-headed monsters. Most data sources are riddled with various inaccuracies that make them unreliable, and it becomes worse with potential risks or perils.

IDSML is an Infosys proprietary end-to-end Python-based solution that addresses the challenges and automates the SOV cleansing process for underwriters. This PoV will talk more about it and its in-built analytical MDM capabilities, ML-based techniques, and the comprehensive set of tools for data profiling and data standardization using Google and Bing's APIs customized for insurance underwriting processes.

Infosys®
Navigate your next

Overview

Data quality is a concerning issue and a key stumbling block for insurers and reinsurers alike, who use data for deriving intelligent and informed business decisions. Unfortunately, across the insurance industry, most data sources are riddled with various inaccuracies that make them unreliable with potential risks.

More than 75% of Insurance organization leaders concur that accurate location-based data is 'important' to business operations, with more than 50% agreeing it is 'critical'.

This process of data rectification and curating validated information is essential to make business processes intelligent for calculating insurance premiums for catastrophic insurance premiums.

IDSML, an ML-based data quality solution, enables multiple large insurance companies to automate the SOV cleansing process. A Schedule of Values (SOV) in insurance is used for describing properties and their associated attributes covered by any policy. The toolset performs automatic mapping of fields, cleansing, and transformations and generates the Risk Management Services (RMS) based Account and Location file, which is to be used for CATASTROPHIC (CAT) Modeling analysis for the below line of businesses (LOBs):

- Property Open Market
- Inland Marine
- Cargo
- Marine & Energy
- Terrorism
- Binders

Key process interventions and opportunities

During the quote and rating process, there are multiple opportunities to free up the underwriter's capacity by eliminating non-value add tasks. Mostly there are non-value add tasks involving data formatting of spreadsheets to prepare them for calculating coverages which enable actuarial valuation metrics.

Infosys tools automate the process flow using the customised features given below:

- Automated text field mapping
- Location based address cleansing and standardization
- Primary and secondary modifiers update as per business rules and CAT modules

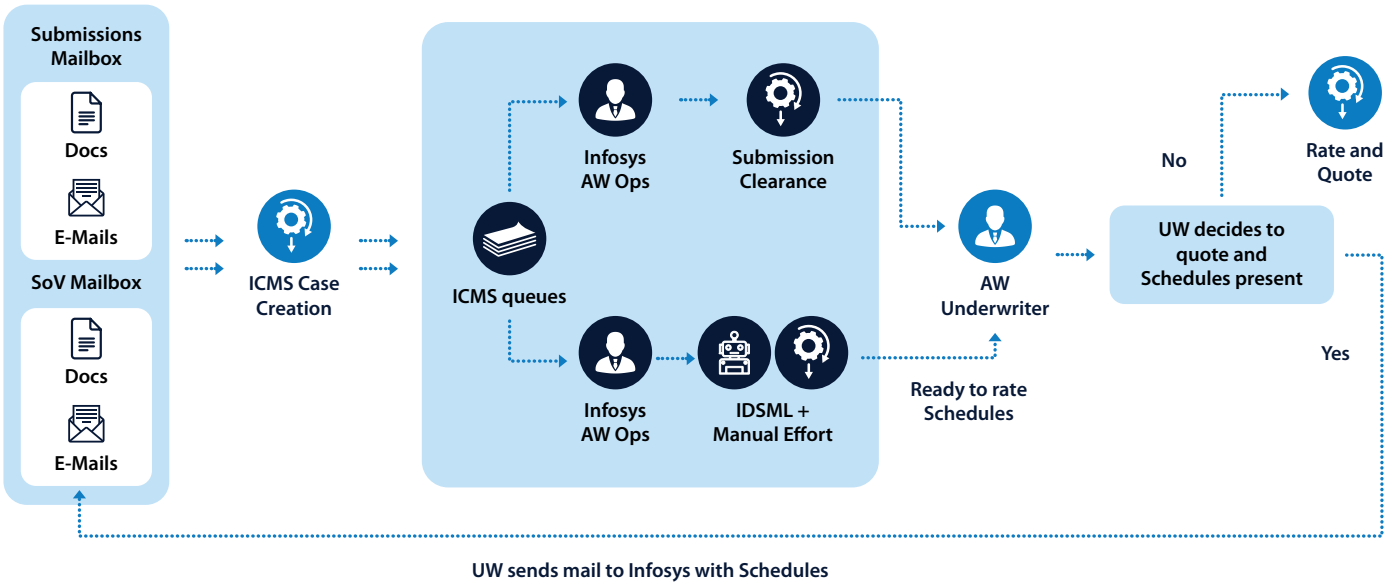


***Figure 1**: Business view representation of RMS SOV process & automation opportunities*

Key activities details performed are as below

Data profiling →ML based data cleansing → data enrichment (Google API)

IDSML python-based module is used to accept the input excel file and map the columns to client-provided dictionary-based values enriched with business rules. Natural language processing using Stemming, Tokenization, and Lemmatization techniques is used for cleansing the descriptive fields and mapping them to get accurate codes.

IDSML also automates the mapping of raw SOV files to CAT modeling standard format using string similarity algorithms (Cosine and Jaro-Wrinkler).

A workflow-based approach is used to enable checks and balances by different roles during the lifecycle.

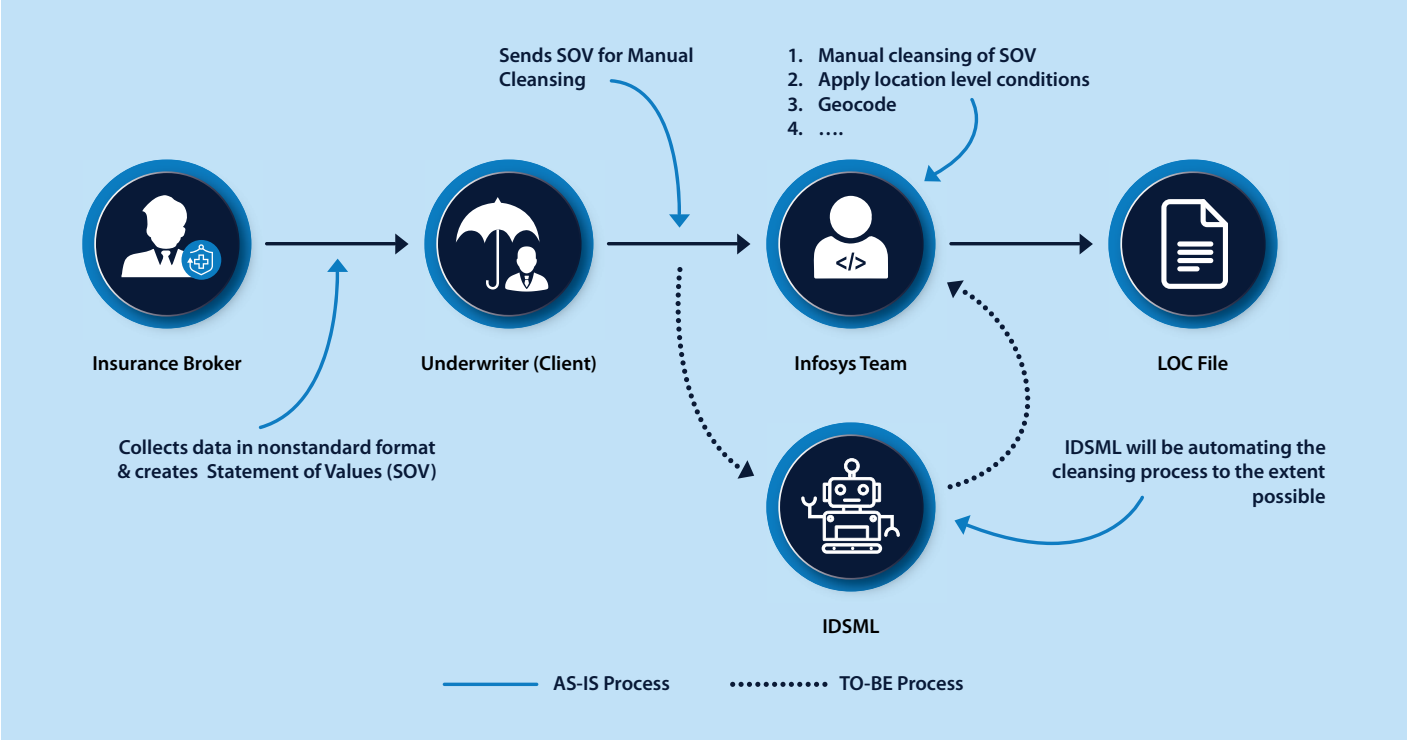


***Figure 2**: Representative view of IDSML automation*

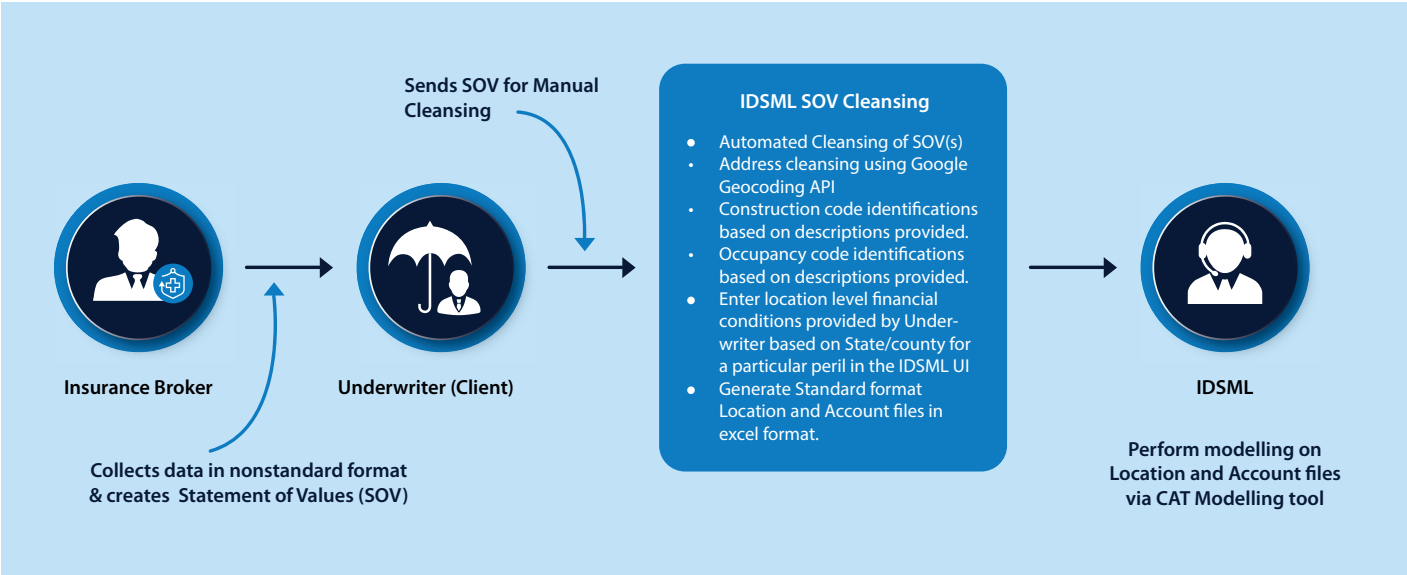


***Figure 3**: IDSML solution enablers*

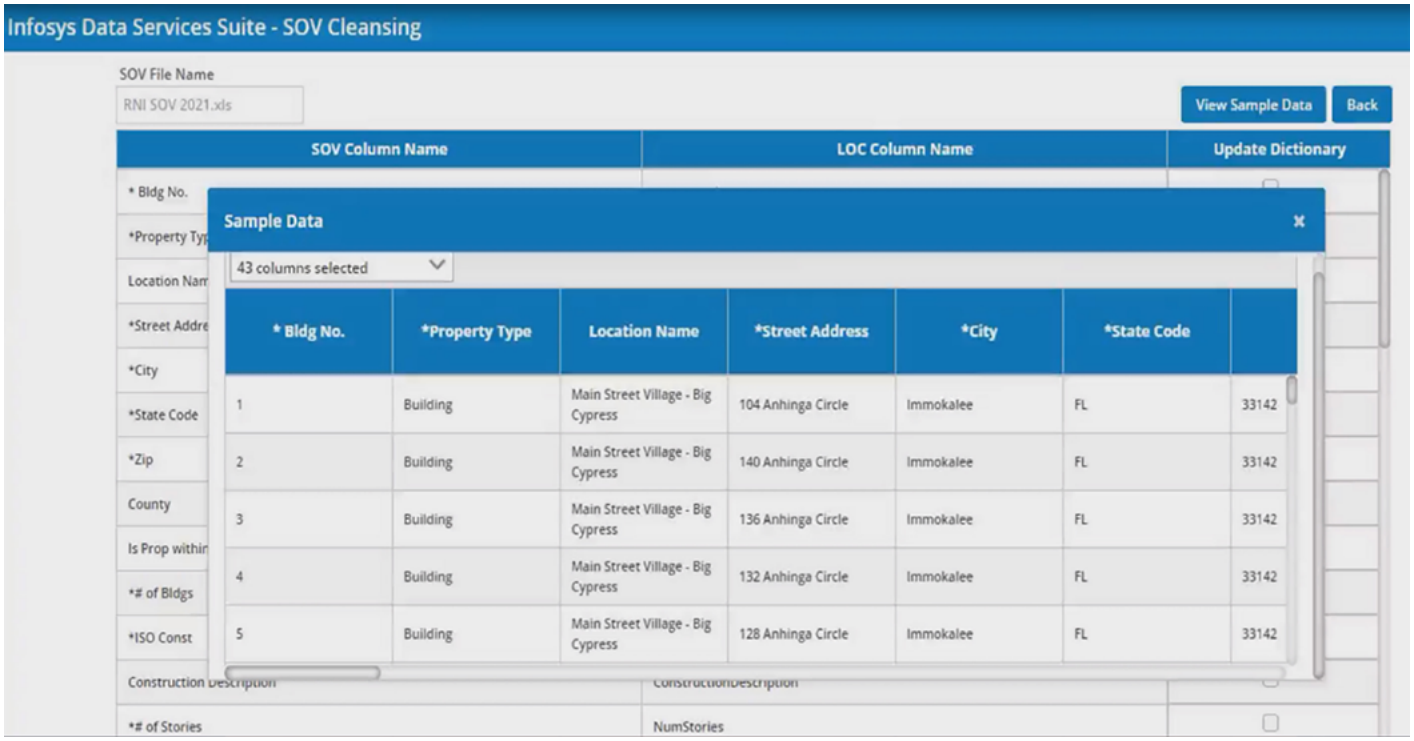


***Figure 4**: IDSML field mapping (Cosine/Jaro Wrinkler-based)*

String similarity algorithms automate mapping from raw input files to CAT modelling standard format using algorithms such as Cosine and Jaro Wrinkler. The data is matched using a continuously evolving and custom-defined threshold value for each field against a list of values using the business rules.

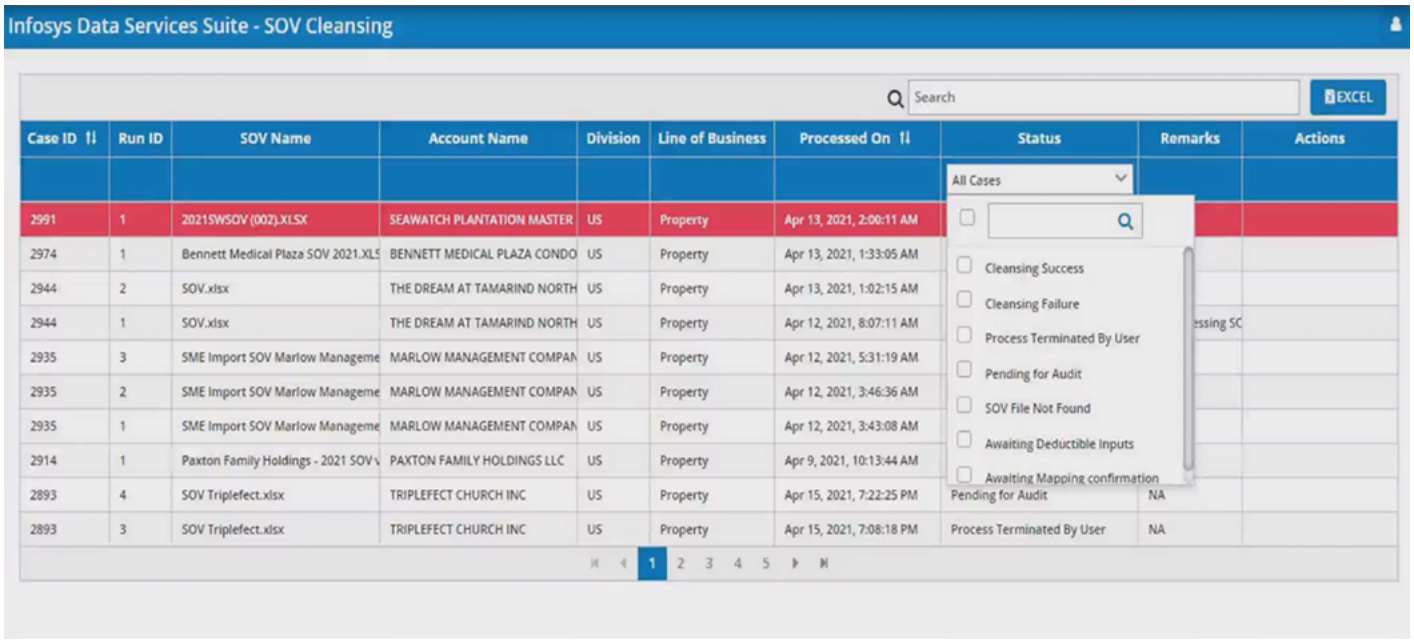


***Figure 5**: IDSML workflow for lifecycle management*

Subsequently, address cleansing to standard format is enabled using a Google API connector, which parses and updates the free text to standard addresses as per API-provided values. This process involves the following tool-based activities:

- Addresses are captured from different sources and in a non-standard format
- Shuffled address values
- Spelling mistakes are corrected
- Pin code corrections

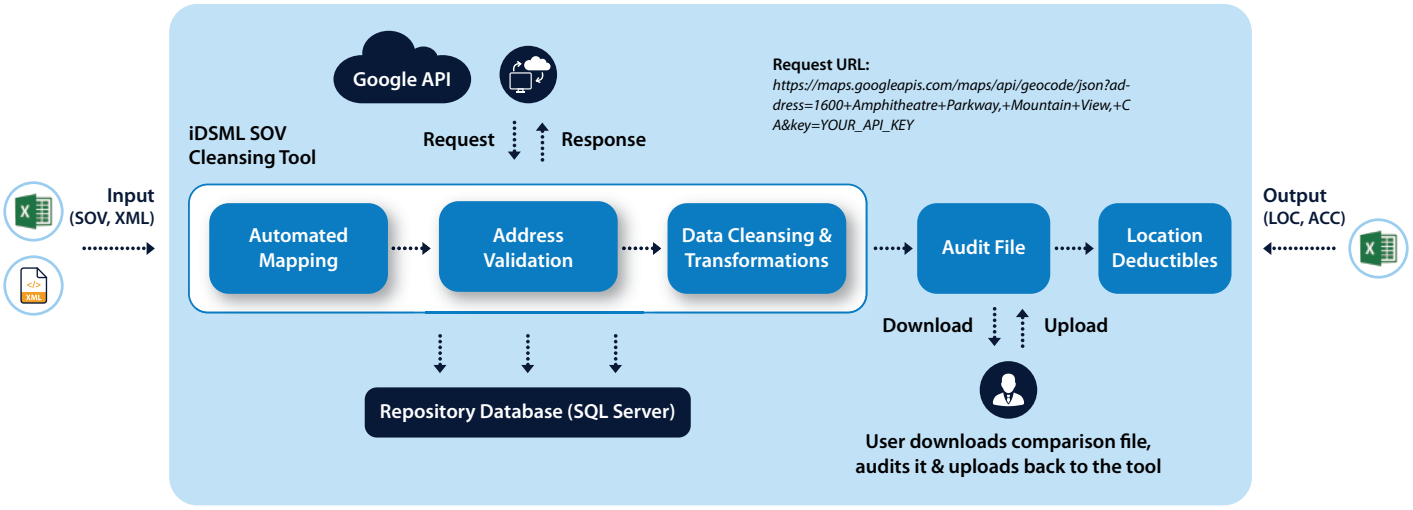
IDSML subsequently leverages wrapper logic to select a similar address and standardise it.

H
Address
640 West Broadway Long Beach NY 11561
48890 Milmont Dr Ste 107D Freemont CA 94538
800, 804, 812 A&BAnacapa St.Santa BarbaraCA93101
800 B-1, 800 B-2, 812C, 814Presidio Ave.Santa BarbaraCA93101
125-127, Suites 101-102, 201- 204E. De La Guerra St.Santa BarbaraCA93101
402 - 436E. Gutierrez St.Santa BarbaraCA93101
411E. Montecito St.Santa BarbaraCA93101
4185Carpinteria Ave.CarpinteriaCA93013
4187Carpinteria Ave.CarpinteriaCA93013
4189Carpinteria Ave.CarpinteriaCA93013

DD	DE
Latitude	Longitude
40.5841786	-73.6825923
37.4623737	-121.9217997
34.42130410000001	-119.6984869
34.4213369	-119.6981451
34.4218323	-119.6970342
34.4207801	-119.6884234
34.4202269	-119.6874552
34.4041766	-119.5352914
34.4040225	-119.5347414
34.4040653	-119.5346346

I	J	K	L	M	N	O
Address	City	StateCode	County	Deductible Region	Zip	Country
640 West Broadway	Long Beach	NY	Nassau County	Other	11561	US
48890 Milmont Drive	Fremont	CA	Alameda County	California	94538	US
804 Anacapa Street	Santa Barbara	CA	Santa Barbara County	California	93101	US
814 Presidio Avenue	Santa Barbara	CA	Santa Barbara County	California	93101	US
204 East De La Guerra Street	Santa Barbara	CA	Santa Barbara County	California	93101	US
436 East Gutierrez Street	Santa Barbara	CA	Santa Barbara County	California	93101	US
411 East Montecito Street	Santa Barbara	CA	Santa Barbara County	California	93101	US
4185 Carpinteria Avenue	Carpinteria	CA	Santa Barbara County	California	93013	US
4187 Carpinteria Avenue	Carpinteria	CA	Santa Barbara County	California	93013	US
4189 Carpinteria Avenue	Carpinteria	CA	Santa Barbara County	California	93013	US

***Figure 6**: Address standardization using APIs*

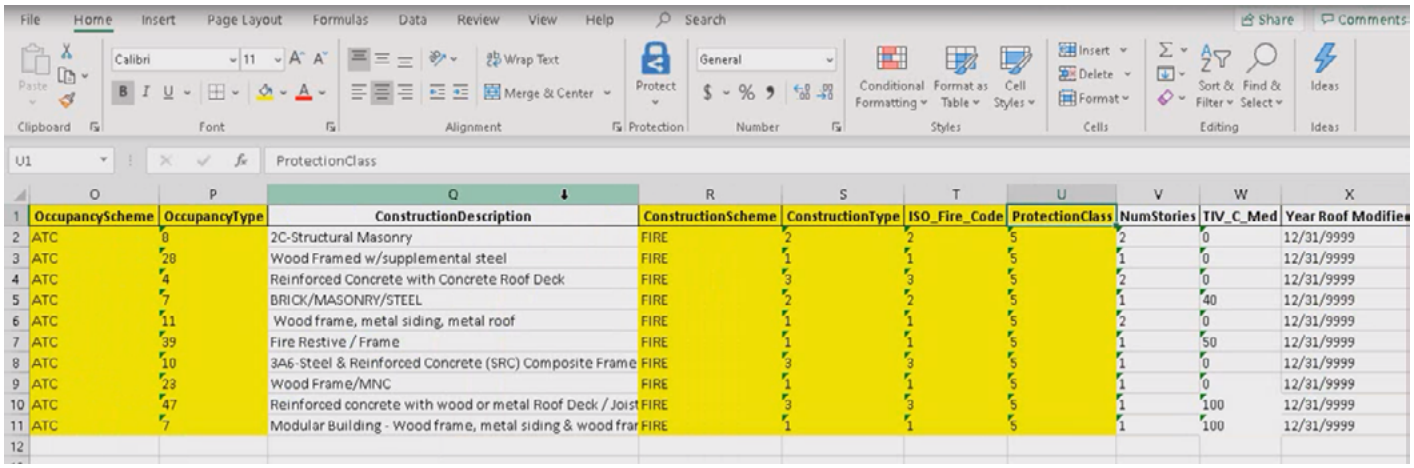


***SOV** – Schedule of Values, **LOC** – Location File, **ACC** – Account File*

***Figure 7**: Plugin integration of Google & Bing APIs*

Subsequently, primary and secondary modifiers undergo data profiling and cleansing as below. Based on the text field input, the natural language string is processed by techniques such as Stemming, Tokenization, Lemmatization, stop word identification and special characters removal for cleansing and enrichment using standardized lookup values.

Missing coverages are calculated and updated on a pro-rated basis.



	O	P	Q	R	S	T	U	V	W	X
1	OccupancyScheme	OccupancyType	ConstructionDescription	ConstructionScheme	ConstructionType	ISO_Fire_Code	ProtectionClass	NumStories	TIV_C_Med	Year Roof Modifie
2	ATC	0	2C-Structural Masonry	FIRE	2	2	5	2	0	12/31/9999
3	ATC	28	Wood Framed w/supplemental steel	FIRE	1	1	5	1	0	12/31/9999
4	ATC	4	Reinforced Concrete with Concrete Roof Deck	FIRE	3	3	5	2	0	12/31/9999
5	ATC	7	BRICK/MASONRY/STEEL	FIRE	2	2	5	1	40	12/31/9999
6	ATC	11	Wood frame, metal siding, metal roof	FIRE	1	1	5	2	0	12/31/9999
7	ATC	39	Fire Restive / Frame	FIRE	1	1	5	1	50	12/31/9999
8	ATC	10	3A6-Steel & Reinforced Concrete (SRC) Composite Frame	FIRE	3	3	5	1	0	12/31/9999
9	ATC	23	Wood Frame/MNC	FIRE	1	1	5	1	0	12/31/9999
10	ATC	47	Reinforced concrete with wood or metal Roof Deck / Joist	FIRE	3	3	5	1	100	12/31/9999
11	ATC	7	Modular Building - Wood frame, metal siding & wood frar	FIRE	1	1	5	1	100	12/31/9999
12										

***Figure 8**: IDSML primary & secondary SOV cleansing outcome*

V
CONSTTYPE
Mixed Masonry & Wood Frame
Mixed Masonry & Wood Frame
50% Frame50% Reinforced Concret
50% Frame50% Reinforced Concret
50 Steel, 35% Wood Frame, 15% Re
Frame
Frame
Concrete Tilt-Up
Concrete Tilt-Up
Concrete Tilt-Up

W	X	Y	Z	AA
onstructionDescriptic	ConstructionScheme	ConstructionType	ISO_Fire_Code	ProtectionClass
Mixed Masonry & W	FIRE	1	1	5
Mixed Masonry & W	FIRE	1	1	5
50% Frame50% Rein	FIRE	1	1	5
50% Frame50% Rein	FIRE	1	1	5
50 Steel, 35% Wood	FIRE	1	1	5
Frame	FIRE	1	1	5
Frame	FIRE	1	1	5
Concrete Tilt-Up	FIRE	4	4	5
Concrete Tilt-Up	FIRE	4	4	5
Concrete Tilt-Up	FIRE	4	4	5

***Figure 9**: IDSML primary & secondary SOV cleansing*

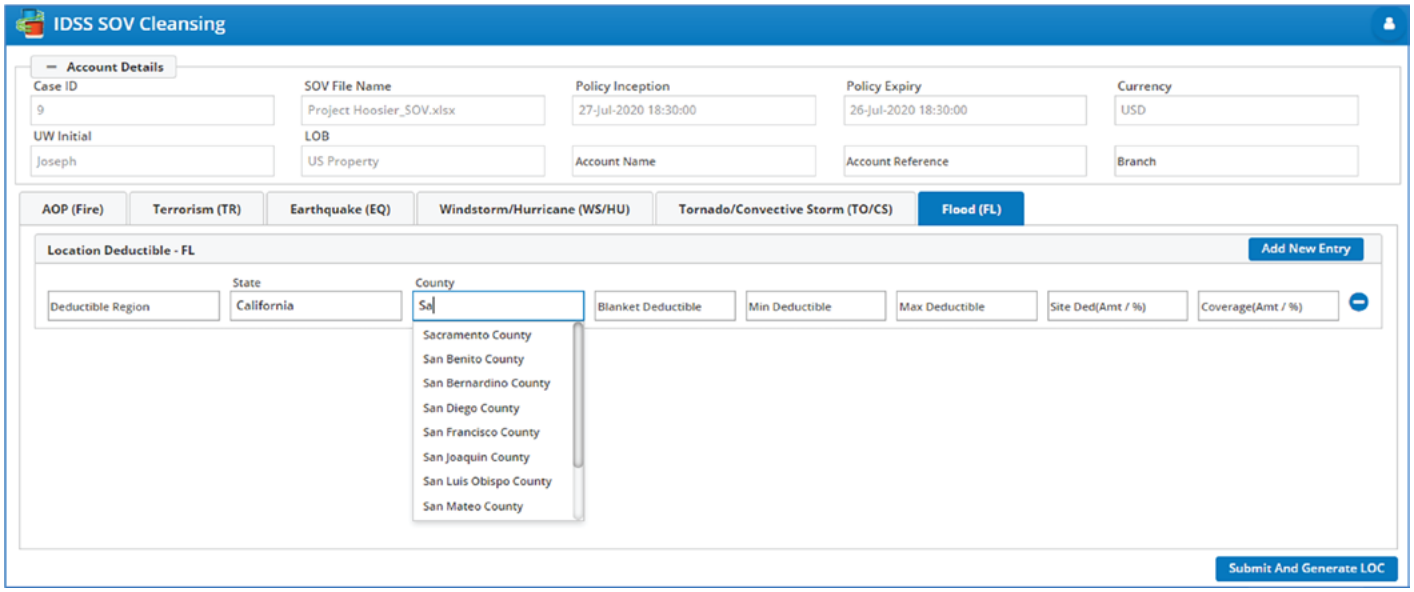


***Figure 10**: IDSML allocation for location level deductibles*

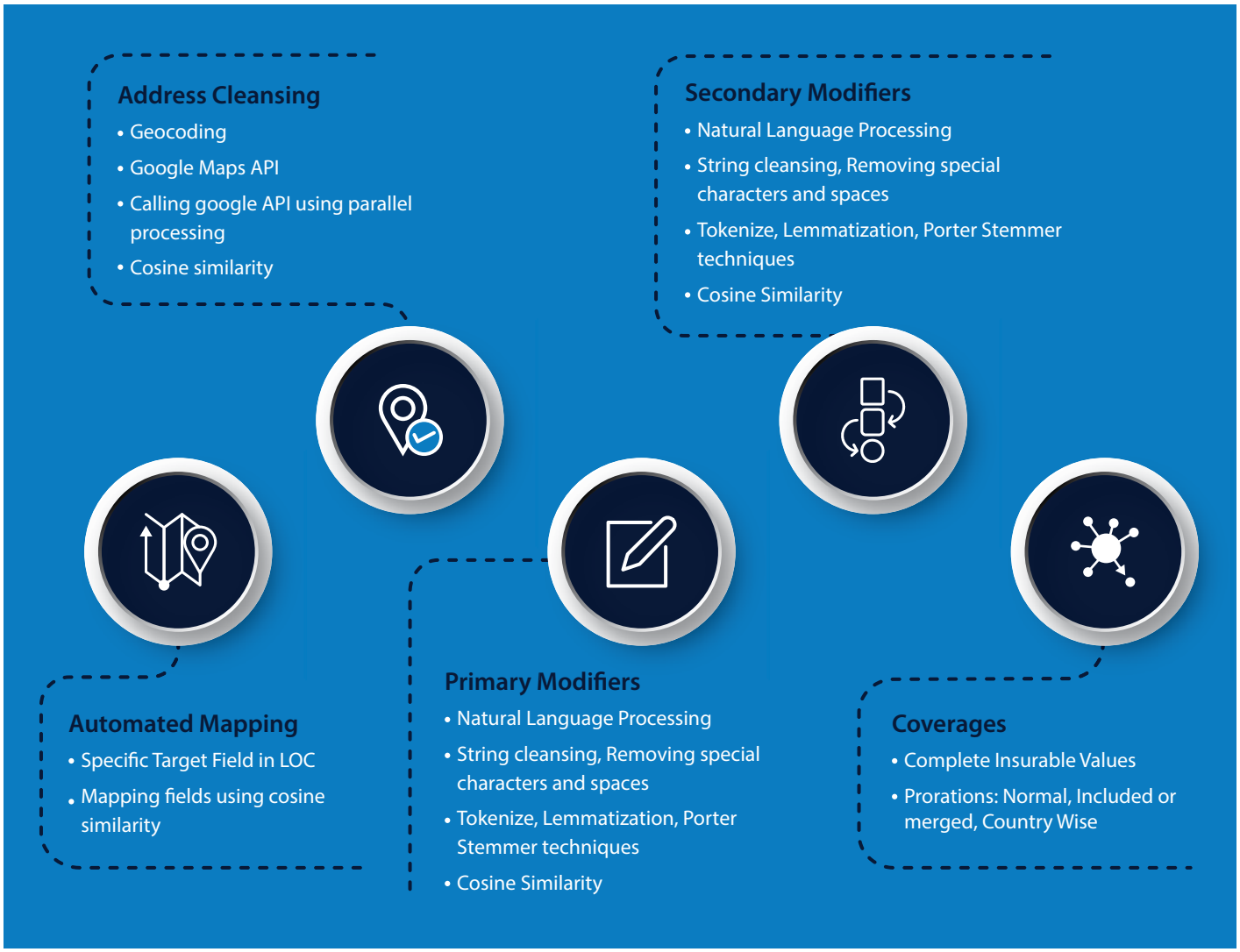


***Figure 11**: Key procedures in end-to-end lifecycle*

Product Description

IDSML is an Infosys proprietary end-to-end solution that addresses the challenges faced in the insurance domain and is currently customized for SOV cleansing for underwriters.

IDSML has been developed to build Analytical MDM capabilities using traditional and ML-based techniques. It has a comprehensive set of tools for data profiling, data standardization, and address standardization using Google and Bing APIs customized for insurance underwriting processes. It leverages supervised and unsupervised learning models to detect data anomalies and help in missing value correction. The MDM module has different matching techniques using deterministic, fuzzy, phonetic or hybrid approaches and ML-based approaches to identify duplicates and generate a golden record using survivorship rules.

Address Cleansing

- Address is standardized and cleansed using Google Maps API
- Option for geocoding of locations available

Coverages

Individual Coverages are identified and Pro-rated based on the rules and Total Insurable Values are calculated.

Automated Mapping

- SOV fields mapping to specific target fields in the LOC template
- Matching performed using dictionary, fuzzy tech, semantic models

Primary Modifiers

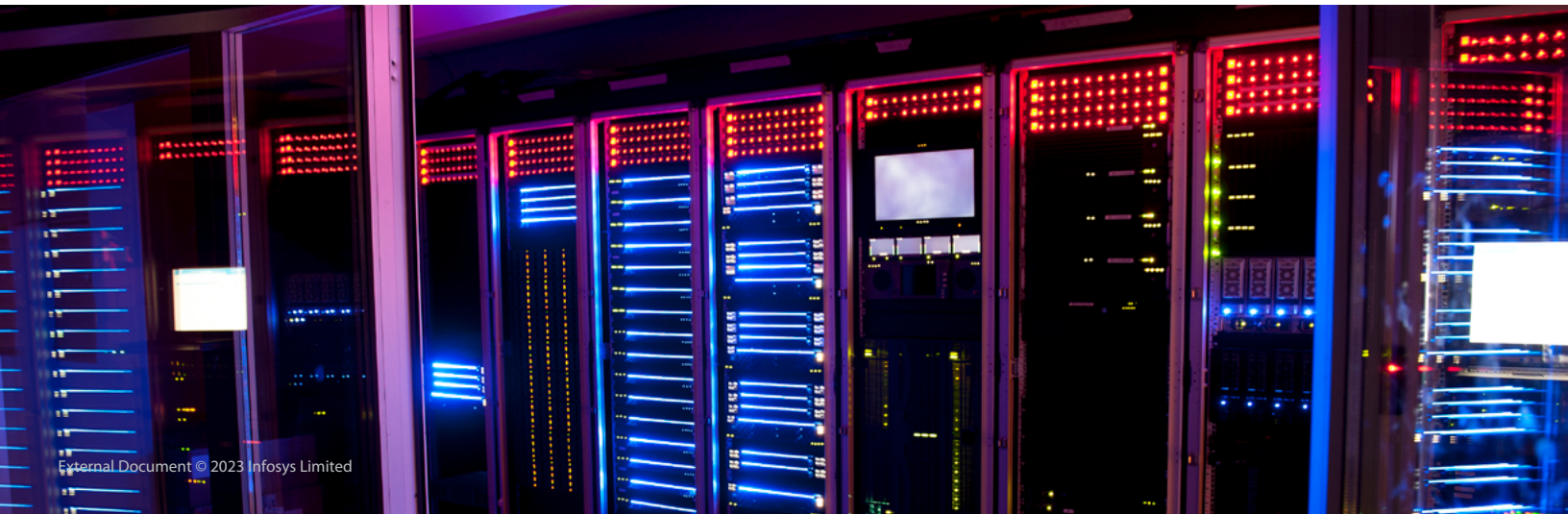
- Construction code & scheme identifications based on similarity checks & Natural Language Processing
- Occupancy types & scheme identifications based on similarity checks & Natural Language Processing

Secondary Modifiers

Secondary Modifiers columns are identified based on lookups, similarity and Natural Language Processing techniques.

User Interface available to enter location level financial conditions or deductible information provided by Underwriter based on State/county for a particular peril.

***Figure 12**: IDSML SOV cleansing capabilities*



1.1. Data Quality and Master Data Management

Profiling examines parameters as data values, ranges, frequency distributions, metadata mismatches, and other non-standard record formats, etc. The outcome of data profiling and data analysis will help decide the cleansing rules to not only standardize the data objects but also enrich them to meet business regulatory requirements.

1.2. Data Profiling

Data profiling is the process of examining the data in an existing data source and collecting information about it. Profiling is carried out on the legacy data extracts to gather relevant quality statistics from them. This activity feeds other data migration processes like data analysis, cleansing, data mapping, etc.

It is a technically led activity and requires extensive collaborations with functional and business SMEs/groups from clients and domains for support and input.

- Business Subject Matter Experts – to identify known business process and data quality issues associated with the use of the source system and data to support the activities of the business area
- Technical Subject Matter Experts – to identify technical and data quality issues associated with the source system
- Data Governance Board – to guide the Data Cleansing/Migration team in the correct resolution of identified data quality issues

The benefit of data profiling is to improve data quality, shorten the implementation cycle of major projects, and improve the understanding of data for the users. Discovering business knowledge embedded in the data itself is one of the significant benefits derived from data profiling. Data profiling is one of the most effective technologies for improving data accuracy in corporate databases.

1.3. Sample Data Quality Issues in Insurance

#	Category	Examples	Fix for One-Time Cleansing	Long Term Fix
1	Attributes with Blank or Invalid Values **(Completeness)**	• Mobile Number, email address, country code, gender details have blank or invalid values • International prefix for mobile numbers has not been configured	Cleansing rules can be defined to correct the values/derive them from other attributes if possible	Data capture mechanisms should be improved to have drop downs, checkboxes, and radio buttons wherever possible instead of text boxes and auto population of few attributes
2	Lack of Accurate Information **(Accuracy)**	• Customer address • Different email address provided in different instances for the same customer	• Business rules can be defined to get address details from reference databases • Survivorship rules can be defined to identify the most appropriate value for attribute	Data capture mechanism to be improved by having important attributes as mandatory fields
3	Lack of Data **(Completeness)**	Employment information, PCI Information, company size related data, third party involved in the accident, etc., is not documented	Business rules can be defined to get these details from Third party reference databases based on other attributes if possible	Data capture mechanism to be improved by having important attributes as mandatory fields
4	Inconsistent formats in applications or LoBs **(Inconsistency)**	Gender	Data standardizations rules	Data standardization at source
5	Capturing More information than needed **(Accuracy)**		Source fields assessment exercise has to be carried out to identify irrelevant fields and decommission them	Source fields assessment exercise to be carried out to identify irrelevant fields and decommission them
6	Duplicate Records **(Duplication)**	• Records with almost same name, mobile no., and email address • If a customer owns a company and is the CEO too, then he will have two separate customer IDs	De-duplication rules can be defined to check for duplicate records based on a set of attributes	Before creating a new customer, validation mechanism to check if a customer record with same name/email /mobile no. exists in the database needs to be implemented
7	Lack of validation or controls to identify if the customer already exists **(Duplication)**	• Policies being issued with variants of the client names	De-duplication rules can be defined to check for duplicate records based on a set of attributes	Business process change to implement duplicate customer checks at the source itself
8	Orphan Records **(Duplication)**	Orphan customers without policy	Business rules need to identify orphan customers and delete them after merging any important attribute to the master record	Business process change to delete orphan customers once it turns out to be a policy or after a certain time period
9	Data Integration Issue **(Integrity)**	• Disconnect between member details and customer details across applications • Independent applications for each LoB	Linking and matching rules can be defined to a certain extent across the data sources	MDM helps in identifying the golden customer record
10	No reports to identify **Data Quality** Issues	• Periodic reports to Identify how many duplicate records are created in the last month	DQ reports can be configured	DQ reports can be configured

***Table 1**: Common data quality issues*

1.4. Data Cleansing

An iterative cleansing process starts as soon as the data quality has been assessed. Data cleansing includes the following iterative steps:

- Elimination of obsolete records
- Removal of duplicate records
- Correcting inaccurate records
- Correcting incomplete records

Data cleansing activities are performed to enable the availability of clean business data.

- Data needs to be standardized and formatted correctly to make sense (E.g., addresses should be derived from standard addresses as per Google/Bing, while legacy applications can allow a free text format).
- Mandatory data fields need to be populated with Not Null values to ensure all fields with Not Null attribute values hold correct data, or else blanks will be considered failures.
- Data de-duplication needs to be performed to master data to populate across defined business applications. A strategy to identify, remove and rectify duplicate records can be agreed upon either at the source or during the extraction and conversion process.
- Data cleansing should pick up inaccurate data fields and apply business rules on data exclusion or transformation and standardization to relevant entities.

1.4.1. Identify and resolving missing data

The recommended approaches to resolving such issues include:

- Using standard database tables or excel worksheets for data manipulation
- Governance guidelines around master data during data population at source
- Populating missing data with data load programs either by calculation or by mapping tables

They help in achieving the following graduated Data Maturity for organizations.

The primary focus of IDSML has been to create a single platform that can help enable rules-based alleviation of all data issues for insurance underwriters for the SOV cleansing process and help derive maximum value.

It incorporates natural language processing and similarity mapping algorithms to not only standardize the data objects but also enrich the data objects to meet insurance business requirements.

About the Authors

Eggonu Vengal Reddy

Eggonu Vengal Reddy is a Principal Product Architect with over 20 years of experience in Data Management, specifically Data Warehousing, Data Modeling, Big Data, and Data Science. He has provided architecture and design to develop tools and solutions to handle enterprise-wide database migrations, master data management, data quality and wrangling, explorative analysis, and feature engineering in the Machine Learning life cycle.

Tushar Subhra Das

Tushar Subhra Das is a Senior Business Data Analyst with over 10 years of experience in Data Migration and Governance. He has worked with Europe and Australia-based insurance and logistics clients for Data migration, MDM and Data Quality, and process governance.

In his current role, Tushar is responsible for APAC and EMEA data migration deployments and enhancements, including product developments for iDSS as the next-generation industry-standard data management platform.

Gopinadh Bapatla

Gopinadh Bapatla is a Technology Architect with over 20 years of experience in Application Development, Data Analysis, Data Science, Machine Learning, and Big Data. He has provided architecture and design to develop tools and solutions to handle data quality by data wrangling, explorative analysis, feature engineering, and building Machine Learning models.

For more information, contact askus@infosys.com

Infosys®
Navigate your next