NOVARICA | Executive Brief

INTELLIGENT TEXT INGESTION: OVERVIEW AND PROMINENT PROVIDERS

DECEMBER 2020

Summary

Tools for identifying, extracting, and processing semi-structured and unstructured text have proliferated in recent years. Several such tools now exist in both insurance-specific and horizontal forms. Many insurers use them to tackle heavily manual processes in new business applications, loss runs, exposure data, and claims.

As these solutions evolve, they are also providing benefits beyond ingestion, including handwriting recognition, data enrichment, and uses to automate and modernize entire workflows.

Providers profiled include AWS, ABBYY, ACORD, Automation Hero, Chisel AI, Cinnamon AI, Coforge, CogniSure, Convr, Friendly, Google, Groundspeed Analytics, Hyperscience, Intellect SEEC, Microsoft, omni:us, and Vidado.

Contents

Primary Report Contact

Deb Zawisza
VP, Research and Consulting
dzawisza@novarica.com

Page Count

11

CONTACT US TO LEARN MORE

833-668-2742 | inquiry@novarica.com | novarica.com

This copy of this Novarica report is provided by ABBYY. ABBYY did not sponsor or participate in the creation of this report, but has obtained license from Novarica after publication to distribute for industry education purposes.

INTRODUCTION

Insurers in the US collectively process millions of documents annually across most of their business processes, from distribution and underwriting to finance and claims. Manual document-based processes are still in place, and for some insurance market segments, there is still a lot of manual data entry into core business systems.

There are now technology solutions that insurers can leverage to scan electronic documents and gather text for automating the process of updating their core business systems. These solutions can help solve two types of problems: extracting text from digital documents that are cleanly indexed (structured data) and extracting unstructured text from digital documents. The latter has traditionally been a bigger challenge for insurers, but there has recently been an increase in the number of software vendors offering technology solutions that address the challenge of transforming unstructured text from documents into structured data sets for automated processing.

Unstructured text-based documents are currently a focus for insurers who receive applications, medical notes, invoices, and other documents that can be processed by intelligent text ingestion (ITI) software solutions supported with artificial intelligence (AI) and machine learning (ML) technology. These solutions are broader than optical character recognition (OCR) and intelligent character recognition (ICR) technologies, which focus on identifying characters in unstructured text. ITI solution providers have created technologies to help insurers and other companies digitally ingest these types of documents and allow the insurers to take action on the incoming data.

This report provides profiles of several vertical and horizontal industry solution providers that offer technology solutions for unstructured text extraction, classification, and related tools.

© 2020, Novarica, Inc. All rights reserved. Reproduction or redistribution prohibited without permission.

INSURANCE USE CASES

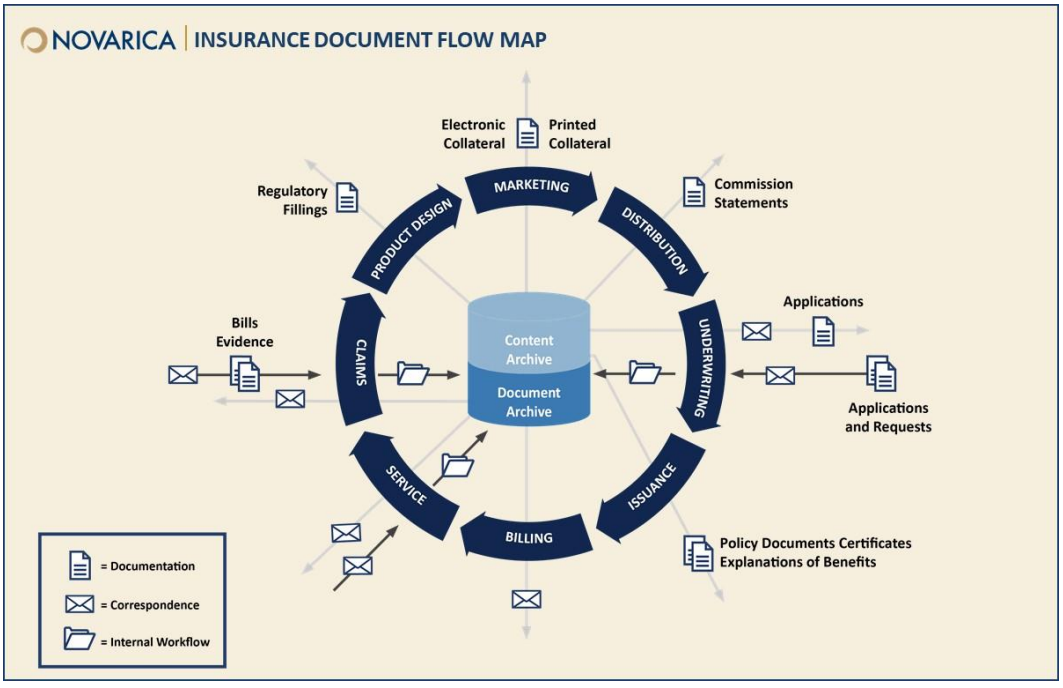
Use cases for ITI solutions typically involve business processes that are dependent on paper images or non-standard electronic files. Specialty lines have a wide set of use cases in new business and underwriting given the variety of document types submitted by brokers. However, other property/casualty lines of business as well as life lines rely on forms or other text-based artifacts for underwriting and claims processing.

Insurers ingest external documents in three key functional areas:

- **Underwriting,** where distributors or prospective insureds may provide supplemental information, schedules, copies of current policies, ACORD forms, and loss runs. Ingestion of this information can feed analytic or robotic process automation (RPA) processes to create a more automated underwriting process.
- **Claims,** where claimants and service providers send insurers voluminous documentation ranging from medical bills to repair estimates to legal notices to medical records.
- **Service,** where policyholders send information related to good student discounts, premium audit requests, mailing address changes, and payment information.

The figure below looks at both inbound and outbound document flow within the insurance life cycle and highlights the areas where insurers are ingesting external documents.

Figure 1: Insurance Document Flow Map



© 2020, Novarica, Inc. All rights reserved. Reproduction or redistribution prohibited without permission.

THE MECHANICS OF TEXT INGESTION

Text ingestion vendors employ similar processes across their solutions to address their basic purpose—scan a document, identify data, name the data, and store the data in structured form. The "identify" and "name" portions of the process typically utilize AI and ML technologies to compare what the solution ingests to previously scanned data and determine a definition match with a certain confidence level. For simple illustration, the solution can recognize that a policy number is a policy number and not an address or a last name.

Various vendor solutions utilize their versions of AI, ML, and deep learning technologies within their products. They often market their solutions by stating the average percentage of a given document's data that is successfully converted into structured data and the number of days it takes to train a new document. It is strongly advisable for insurers to conduct proofs of concept with multiple vendor solutions to properly evaluate the efficacy of these solutions against an insurer's key document types.

Document Types

Ingestion solutions typically have the capability to handle multiple types of documents, enabling carriers to leverage a common approach for structured and unstructured text. Documents that are less structured have a greater reliance on AI and machine learning capabilities to improve accuracy.

- **Structured Text:** Structured text-based documents are digital documents where each textual data element is indexed with a known position on the document, field name, field definition, and field length. An example of a structured text-based document is ACORD 130, an insurance application for a workers' compensation policy.
- **Semi Structured Text:** Semi-structured text-based documents contain characteristics of both unstructured and structured documents. An example of a semi-structured document is an Excel spreadsheet of a property schedule where the column headers may not be structured; however, once the column header is identified, the values in the column can be identified (example: total insured value).
- **Unstructured Text:** An unstructured text-based document is a digital document that contains text without any structure to each textual data element. Each such element on the unstructured document exists without definition, indexing, or size. An example of an unstructured text-based document is an insurance application for a D&O policy.

Document Processing

At their core, all unstructured text solutions work through four stages: document scan, document identification, text field identification and extract, and text storage. The scanning process involves OCR conversion of optical characters into text. Once the scan is complete, the system must identify the type of document and then identify fields of text within it. Once this is complete, the data can be extracted and stored.

- **Document Scan:** Scanning digital documents is the ability to "read" a document using OCR technologies, convert an unknown text into a recognizable text, and store it in a computer file. For example, OCR technology can scan the label "policy number" on a document along with its content, e.g., "abc123," and then convert them into the literal "policy number" and "abc123." This first step is essential to the overall process.

© 2020, Novarica, Inc. All rights reserved. Reproduction or redistribution prohibited without permission.

- **Document Identification:** Documents submitted, trained, and processed by an unstructured text ingestion product in the past are compared with existing documents scanned from previous runs that aided the solution's internal learning curve. Identification of new documents is added to the existing base of processed documents.
- **Text Field Identification and Extract:** As part of the ingestion of data, each text field is evaluated, and trained algorithms are used to identify a policy number, for instance, label that field, and add it to the field identification ML rules.
- **Text Storage:** Extracted text data are then stored internally in data sets and customer-defined data formats with appropriate labels and other entity characteristics that can be leveraged by vendor services and client processing.

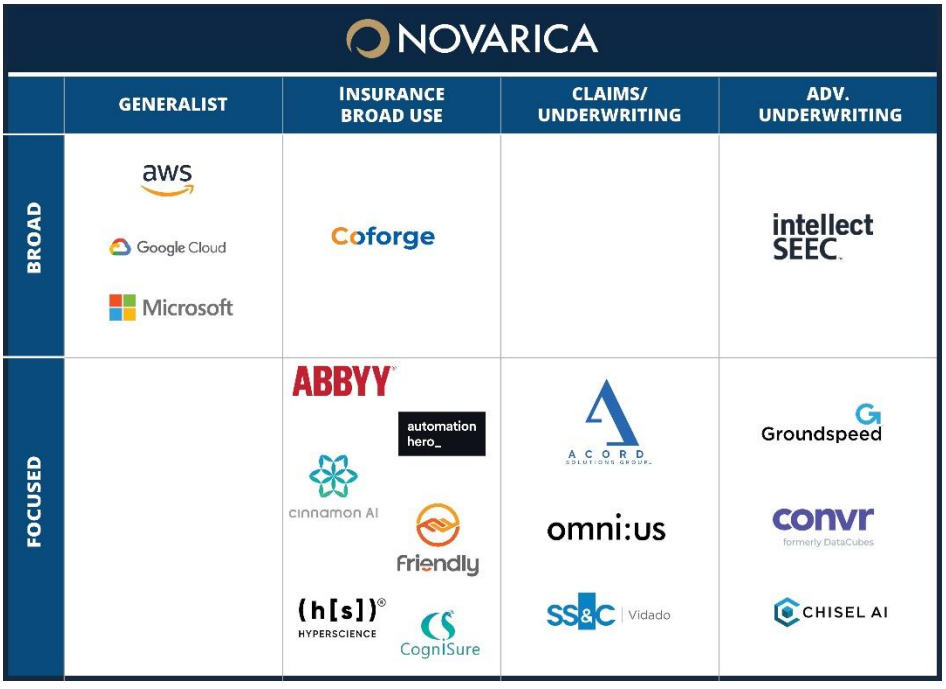
SOLUTION TYPES AND PROMINENT PROVIDERS

Although unstructured text ingestion is a relatively new technology for the insurance industry, benefits are already being realized in process efficiency, routing timeliness, and improvement in errors and omissions. Use cases are primarily focused on underwriting and claims due to the large amount of unstructured information flowing from agents, insureds, and third parties.

Vendors offer different types of solutions that support a wide variety of use cases and can be segmented into four categories: generalist tools, broad use insurance-specific tools, claims/underwriting-specific tools, and ITI advanced underwriting tools. Insurers can leverage tools from all the categories depending on their product mix and processing challenges. Most tools can integrate with RPA tools to fully automate processes.

Insurers who process digital documents without the benefit of an ingestion tool rely on manual processes to determine the necessary data to capture, where to route the document, and what the next steps in processing are. A successful unstructured text ingestion tool implementation can replace this manual capture, review, and routing process.

Figure 2: Prominent Providers of Intelligent Text Ingestion for Insurers



© 2020, Novarica, Inc. All rights reserved. Reproduction or redistribution prohibited without permission.

Generalist Tools

Generalist tools have the capability to ingest text from any type of artifact and industry, but they typically need to be trained on specific uses cases. Tools in this segment can be either leveraged or trained to extract text from invoices, medical records, legal documents, policy documents, and other artifacts. Insurance carriers can also tap into the broader AI platforms available from these vendors to build a more general text ingestion capability.

Some of the prominent providers in this category include:

- **AWS,** which offers its Amazon Comprehend product to ingest unstructured text from documents and create insights and relationships within the text. Amazon offers a distinct version of the tool called Comprehend Medical that is designed to address medical terminology and natural language, including but not limited to medical conditions, doctor notes, medication, dosage, etc.
- **Microsoft,** which offers Azure Cognitive Services to ingest raw text for ingesting unstructured text from documents and create structured data. Insurers can link to its Form Recognizer tool with an API to extract text, key/value pairs, and tables from documents. With just a few samples, Form Recognizer can tailor its understanding to an insurer's documents, both on-premises and in the cloud.
- **Google,** which offers its Cloud Document AI product that combines OCR technology and deep learning technology to read and understand documents. Google Cloud's Vision OCR and form parser technology uses deep learning neural network algorithms to perform text, character, and image recognition in over 200 languages.

Broad Use Insurance-Specific Tools

Broad use insurer-specific tools can automate manual processes associated with reviewing documents, identifying key data (e.g., insured name, address, deductibles, limits, etc.), and entering this data into core insurance systems. The use of a business rules engine in the ingestion solution enables carriers to apply rules of engagement to each incoming document.

Business rules typically include routing documents to the appropriate personnel (e.g., by line of business and document type), prioritizing submissions (e.g., by executing real-time adjusted appetite rules), formulating criticality of data consumed (e.g., highlighting the high-risk losses from loss runs), and identifying missing information or cross-field edits that help the carrier ensure that data is complete and accurate.

Some of the prominent providers in this category include the following:

- **ABBYY** was founded in 1989 and is headquartered in Milpitas, CA, with offices in Europe and Asia-Pacific. ABBYY enables organizations to gain a better understanding of their business processes and content with its Digital Intelligence platform. ABBYY's Content Intelligence technologies and solutions capture and transform content locked within documents, forms, and correspondence into the structured data needed for automation in the cloud or on-premise. This includes invoices, claims, new account opening and customer onboarding documents, purchase orders, bills of lading, and contracts. ABBYY's clients include MelitaUnipol Insurance Agency, Ltd. and AOK, Germany's largest public health insurer.
- **Automation Hero** is a startup founded in 2017 and headquartered in San Francisco, CA. Its intelligent process automation platform combines unstructured data ingestion; process mining; ETL; RPA; and native AI such as intelligent OCR, sentiment analysis, and dark data extraction. Its back end focuses on three key areas: augmenting employee decision-making, cutting repetitive and time-consuming tasks, and automating common customer asks. The platform offers both unattended and attended automations. Its attended version is

© 2020, Novarica, Inc. All rights reserved. Reproduction or redistribution prohibited without permission.

personified by Robin, an adaptive and interactive AI assistant. Automation Hero's live insurer clients include Signal Iduna. Automation Hero reports that five customers are currently live on its solution in North America.

- **Cinnamon AI** was founded in 2016 and is headquartered in Tokyo, Japan. Cinnamon AI develops AI-based business platforms to free up human resources by eliminating repetitive tasks. It offers a computer system that has human intelligence features like recognition and judgment that help select appropriate and flexible correspondence according to the partner and circumstances. It does this based on features derived from accumulated data, allowing users to focus on creative work. Cinnamon AI's clients include Dai-ichi Life Insurance Group, Nissay, and Tokio Marine. Cinnamon reports that ten customers are currently live on its solution in North America.
- **Coforge** (formerly NIIT Technologies) was founded in 2004 and is headquartered in New Delhi, India. It offers its clients its Self-Learning Intelligent Content Extraction ("SLICE") tool, which uses open-source technology for unstructured text extraction. Coforge intends for the tool to be used as part of a broader workflow, as the system can determine not only what sort of document it is examining and what is in it, but also what to do with it. It is designed to pull in external data (both structured and unstructured) from multiple sources (e.g., PDF, image, handwritten text, content from email body, Excel, Word, etc.). During the process, it cleans the unstructured data, then structures the data in a meaningful format and ingests it into the underlying systems like agent/broker portal, underwriting workbench, policy administration system, and claims processing systems. Coforge reports that four customers are currently live on its solution in North America.
- **CogniSure** was founded in Warrenville, IL in 2019 and offers an AI platform that the company notes is purpose-built for the insurance industry. The company uses AI and ML to extract unstructured data from insurance documents (PDF, Excel, scanned files) and convert it to JSON, XML, CSV, and customer-specific output formats through API. The AI algorithms can extract data from multiple unstructured sources, with a primary focus on loss runs. CogniSure believes its ability to handle complex, multi-formatted loss run data is a competitive differentiator in the space. Extracted data is enriched with third-party data sources. CogniSure AI notes that it is now ISO 27001:2013 certified. The company notes that it has engaged in multiple pilots, including pilots with insurance carriers for submission insights, brokers for loss insights (e.g., BrokerTech Ventures or BTV), and lenders for policy insights. No funding information is currently available for CogniSure AI. Cognisure reports that it currently has 12 pilots running with customers in North America.
- **Friendly** was founded in 2018 and is headquartered in San Francisco, CA. Friendly's text ingestion tool was designed specifically for insurance use cases. Its first cases were in the life/health/annuity sector, but it now has carrier and broker clients in the property/casualty space. Its ingestion solution can be used for policy and claim document ingestion, such as allowing brokers to upload a declaration page and request a quote based on the same coverages. Friendly's clients include Principal, American Enterprise, and PPS. Friendly reports that six customers are currently live on its solution in North America.
- **Hyperscience** was founded in 2014 and is headquartered in New York, NY, with offices in Sofia, Bulgaria and London, UK. Hyperscience automates processes using proprietary ML techniques to improve throughput and reduce manual errors. The Hyperscience Intelligent Document Processing solution has been implemented at leading financial services, insurance, health-care, and government organizations around the world, including QBE, TD Ameritrade, Voya Financial, and Mutual of Omaha. Insurance makes up approximately 25% of its revenue today.

© 2020, Novarica, Inc. All rights reserved. Reproduction or redistribution prohibited without permission.

Claims/Underwriting-Specific Tools

These tools are tailored for forms and artifacts associated with claims and underwriting, including ACORD forms for property/casualty and life/annuity. The technology is pretrained to recognize the form and extract the data. ACORD form ingestion improves accuracy and efficiency in both functions. The ability to analyze incoming documents containing unstructured blocks of information can allow insurers to identify key claims trends like medical conditions, medication intake, workplace injuries, and location of accidents. Another example is in workers' compensation and medical liability, where important medical details or injury information that could impact the outcome of the claim are often "hidden" in unstructured text.

Some of the prominent providers in this category include the following:

- **ACORD** Transcriber is a document digitization solution that automatically extracts data from and populates data onto ACORD forms and other documents using custom mapping tools and AI frameworks. It is not only pretrained on all ACORD forms and versions, but can also be used for unstructured text, images, and spreadsheets. It also enables users to access, complete, and generate HTML forms. The use of microservice APIs facilitates integration across platforms. ACORD reports that 15 customers are currently live on Transcriber in North America.
- **omni:us** was founded in 2015 and is headquartered in Berlin, Germany. Its AI solution is designed exclusively for the insurance vertical and currently focuses on high-volume property/casualty and health claims. omni:us's AI solution is able to create a broad and deep database from all kinds of incoming document types (classification and extraction), reducing manual data processing. This approach forms the foundation for additional omni:us cognitive insurance AI models, such as Completeness Check or Coverage Check. It features APIs for integration with other systems, such as fraud detection vendors (e.g., FRISS), and omni:us also partners with RPA providers for data transfer to and from legacy systems. omni:us currently has 19 active clients, including Allianz, AmTrust, and Vienna Insurance Group. It has five clients live in North America.
- **Vidado,** acquired in March 2020 by SS&C Technologies, was founded in 2011 in the San Francisco Bay area. Thanks to its advanced ML capabilities combined with the industry's largest human-verified data set, Vidado's AI is a highly accurate document automation solution for enterprise forms, including handwriting, low-DPI scans, and blurry faxes. It validates and enriches the content before sending to downstream systems and workflows—facilitating improvements in straight-through processing. Vidado supports a variety of workflows in production across the insurance industry, including supporting 20+ life insurers, 15+ health insurers, as well as enterprises in health care and financial services.

© 2020, Novarica, Inc. All rights reserved. Reproduction or redistribution prohibited without permission.

ITI Advanced Underwriting Tools

These tools have features and capabilities designed primarily for commercial insurance business submission and underwriting. Commercial insurance including excess and specialty business submission is a complex and paper-intensive process. Limited standardization is in place, and business submission may require extensive loss runs and non-standard applications. These technologies expand the use of text ingestion to more fully automate business processes, including risk identification and selection.

ML and AI play a larger role in this suite of tools to expedite the underwriting process as well as to classify risks appropriately. Some solutions incorporate the use of third-party data to provide a more accurate view of the risk and exposures.

Some of the prominent providers in this category include the following:

- **Chisel AI** was founded in 2016 and is headquartered in Toronto, Canada. It applies natural language processing and ML to unstructured and structured data sources such as insurance documents. The company's platform is built for commercial lines insurers, reinsurers, and brokers; it is compatible with a variety of file transfer formats and analytics solutions. Insurance use cases include data extraction, quote comparison, policy checking, submission triage, and submission prioritization. Chisel AI has completed multiple funding rounds led by Venrock, with the most recent round completed in 2020. It is currently in use at Zurich North America and two large global brokers, and it is conducting pilots with several other carriers and brokers.
- **Convr** (formerly DataCubes) was founded in 2016 and is based in Schaumburg, Illinois. Convr is a decision science platform that automates the underwriting process for commercial property/casualty carriers. The company uses ML to digitize data from policy submission documents such as applications and loss runs. It has collected third-party data from thousands of different sources, which are used to derive insights that automatically answer insurer underwriting questions. The company also offers a score for policy submission evaluation and prioritization. Convr has raised $17.7M in total funding. It raised $15.2M in the latest Series B funding in November 2019, led by Palm Drive Capital.
- **Groundspeed Analytics** is a data science company that helps insurance companies, brokers, and reinsurers increase value from data in the submission intake and underwriting processes. Its Fusion Data Pipeline-as-a-Service transforms unstructured data from insurance files and combines it with third-party data to create insights and predictions. Groundspeed works with five of the top 12 commercial insurers and seven of the top 15 brokers, including Travelers, Aon, and Liberty Mutual. The company has raised a total of $32M across two funding rounds. Its most recent round of $30M in Series B funding occurred in July 2018, led by Oak HC/FT. Other investors include ManchesterStory Group, Michigan Angel Fund, Service Provider Capital, and Tappan Hill Ventures. Groundpseed reports that 15 customers are currently live on its solution in North America.
- **Intellect SEEC,** the insurance brand for Intellect Design Arena, is headquartered in Piscataway, NJ. Its newly introduced product Intelligent Data Extract (IDX) is an intake tool for document unbundling, extraction, validation, and enrichment of data from broker-submitted bundles, including Excel documents such as statements of values, ACORD and custom forms, and others. It can be used with other tools within Intellect's platform, such as Risk Analyst, which triangulates thousands of external data elements to offer intake validation and enrichment, proactive marketing, and predictive analytics for underwriting and Xponent, which is an underwriting workflow offering rate, quote, issue, endorse, and renew capabilities. Intellect SEEC reports that five commercial lines customers are currently live on its solution in North America.

© 2020, Novarica, Inc. All rights reserved. Reproduction or redistribution prohibited without permission.

CONCLUDING THOUGHTS

ITI can transform manual processes through automation, provide quicker cycle times, improve accuracy, and generate new insights. Defining the right set of use cases for a pilot is critical to enable the technology to achieve accuracy and throughput expectations. Building on a base of success is more effective than taking a broad approach that will have too many variables and downsides.

As much as insurers would like to eliminate paper or paper images, some aspects of the business are still a long way from being fully digitized. ITI provides an opportunity to take advantage of fully digital processes without fully digital input. The technology landscape will continue to mature as carriers deploy more uses and as vendors continue to invest in their solutions.

Related Research

- *Novarica Research Digest: Underwriting*
- *Novarica Research Digest: Property/Casualty Claims*
- *Document ECM/CCM Systems*
- *Artificial Intelligence Use Cases in Insurance*

© 2020, Novarica, Inc. All rights reserved. Reproduction or redistribution prohibited without permission.

ABOUT NOVARICA

Novarica helps more than 100 insurers make better decisions about technology projects and strategy. Our research covers trends, best practices, and vendors, leveraging relationships with more than 300 insurer CIO members of our Research Council. Our advisory services provide enterprise phone and email consultations on any topic for a fixed annual fee. Our consulting services range from strategic blueprints and roadmaps to vendor evaluations. Other special programs include our Silicon Valley Tour, InsureTech Summits, online learning courses, and more. https://novarica.com

Authors

Deb Zawisza is a Vice President of Research and Consulting at Novarica. She has expertise in insurance technology leadership and transformation with over 25 years of experience. Prior to joining the firm, she was SVP/CIO for Claims/Loss Control at Travelers Insurance. Deb has also served as CIO and CTO at The Phoenix Companies (now Nassau Re) and as a Senior Principal Consultant at PwC. In addition, she has held various IT leadership roles at Aetna across multiple lines of business. She attended Rensselaer Polytechnic's MBA program and a has a BBA from Adelphi University. She can be reached directly at dzawisza@novarica.com.

Charlie Kirchofer is a lead associate at Novarica. Prior to joining Novarica, he worked in market research and commercial due diligence; was a freelance editor, translator, and English and German language instructor; and was adjunct professor of Security Studies at the School of Criminology and Justice Studies at UMass Lowell. He has a PhD in War Studies from King's College London, an MA in International Relations from Webster University, and a BA in Linguistics and German from Binghamton University. He can be reached directly at ckirchofer@novarica.com.

DISCLAIMER

THIS REPORT CONTAINS NOVARICA ANALYST OPINION BASED ON PERSONAL EXPERIENCE, INFORMATION PROVIDED BY THIRD-PARTY RESEARCH SUBJECTS, AND SECONDARY RESEARCH. NOVARICA MAKES NO WARRANTIES, EXPRESS OR IMPLIED, CONCERNING THE QUALITY, FITNESS FOR A PARTICULAR PURPOSE, TITLE OR NON-INFRINGEMENT OF THIS REPORT, OR THE RESULTS TO BE OBTAINED THEREFROM OR ANY SYSTEM OR PROCESS THAT MAY RESULT FROM CUSTOMER'S IMPLEMENTATION OF ANY RECOMMENDATIONS NOVARICA MAY PROVIDE. NOVARICA EXPRESSLY DISCLAIMS ANY WARRANTY AS TO THE ADEQUACY, COMPLETENESS OR ACCURACY OF THE INFORMATION CONTAINED IN THIS REPORT. THE CUSTOMER IS SOLELY RESPONSIBLE FOR ANY BUSINESS DECISIONS IT MAKES TO ACHIEVE ITS INTENDED RESULTS.

LAST UPDATED: March 1, 2021

© 2020, Novarica, Inc. All rights reserved. Reproduction or redistribution prohibited without permission.