



Information Extraction in Banking and Insurance

An Overview of AI Capabilities

A collaboration between Iron Mountain and Emerj



IRON MOUNTAIN®



EMERJ

AN OVERVIEW OF AI CAPABILITIES

Artificial intelligence can be used to increase efficiency in a variety of business process spanning many different sectors. This white paper reflects on (a) how AI can be used to digitize paper documents in order to create a searchable digital database accessible across an enterprise and (b) how AI can be used to search through these databases and extract information from them.

Specifically, this white paper focuses on how AI-based information extraction applications can drive revenue and reduce risk in banking and insurance.

Emerj Artificial Intelligence spoke with Iron Mountain's Director of Product Management, Anke Conzelmann, about these applications and their use in a variety of sectors, and quotes from this interview are used to add color to the various points discussed in this paper. Statistics from Emerj's [AI in Banking Vendor Scorecard and Capability Map](#) are used to the same effect.

“

Even the largest banks store old physical documents in file cabinets and boxes off the bank's premises.

INFORMATION EXTRACTION IN BANKING Compliance, Contracts, and More

Large banks deal with millions of documents every day across their corporate offices and numerous branches. Although one might assume that these documents are digital, in many cases, even the largest banks store old physical documents in file cabinets and boxes off the bank's premises. Even those that are kept on-site might be relegated to storage units amongst hundreds of thousands of other documents.

”

The state of organization in the digital space isn't much better. One bank's multiple departments might all store their digital documents differently. They

might each use a different system or save the documents in the same system in a totally different way. In addition, document layouts often change over time. A bank's employment contracts might look completely different from those it had employees sign twenty years ago, for example, and these contracts might differ in both format and content.

Banks pay employees and retained legal teams to spend a majority of their time finding and reading through these documents to find information relevant to them in the moment. Depending on where the documents are stored, this could take weeks or months, and they might need to do the same process over again to find different information at a later date.

As a result, if a customer were to ask the bank to purge all of the information it has on that customer, the bank would have a very difficult time fulfilling this request. That puts them at risk of noncompliance with regulations such as GDPR and the California Consumer Privacy Act. In the case of the former, noncompliance could result in a fine of €10 million (\$11.2 million) or 2% of the company's yearly revenue, whichever is higher.

In addition, banks are spending on search and discovery: not analysis. They would obviously rather have their employees and lawyers spend as much of their time making decisions based on the information they find during their searches as possible.

Artificial Intelligence, namely natural language processing (NLP) and machine vision, could be one way to remedy this situation.

This chapter of the white paper begins with what is for many banks one of the most challenging problems at their companies: what to do with their backlogs of paper documents.

Digitizing Paper Documents and Microfiches

Common in the financial sector, including investment banking, is the storage of documents on microfiche. Microfiches are small index card-size films that contain microscopic images of documents. A microfiche reader zooms in on these images, allowing bank employees to read the documents.

“

Machine vision software could help with digitizing the documents found within a microfiche.... These documents are often still relevant to established banks that have been in business for decades.

”

Microfiches often contain past account statements and customer information. Employees that need to collect this information need to slide microfiches under a reader to figure out what each document on the microfiche contains. They might go through several microfiches to find the right information or get all the information required to respond to a request. Digitizing these microfiches could save employees time, allowing them to focus on higher value activities and enhance the customer experience.

Machine vision software, particularly optical character recognition (OCR), could help with digitizing the documents found within a microfiche. OCR is a type of artificial intelligence that works specifically to transcribe printed and hand-written text into digital text. Document digitization may be a helpful AI use-case in banking right now in this current phase of the AI Zeitgeist (which Emerj calls “Emergence”), mainly for the reasons described at the top of this paper.

Microfiches contain documents from before digital became the primary storage method in banking, but these documents are often still relevant to established banks that have been in business for decades.

Conzelmann exemplifies their relevance with an estate owner who asserts the bank holds millions for a particular estate. An employee at the bank may need to verify or disprove this by producing multiple months or even years of statements on microfiche. Conzelmann details the laborious process of finding these statements on microfiche:

[The employee needs to] go find the right microfiche in the right box, put it on [their] microfiche reader, find the right square on [the] microfiche, get that digitized, and that was month one. Hopefully [the account] only ha[s] one page statements because, if [it] ha[s] two pages, [they’re] doing it twice for that particular month. [Searching across] 24 months takes a really long time. [The employee is] not adding any value through this; all [they’re] doing is responding to a request from a customer.

What the bank wants to do is be able to search for an account number across a date range and find all of the statements relevant for the particular estate, but this is challenging for large banks that store their documents across disparate sources, including microfiche.

An OCR software could in theory transcribe the text in the documents that were digitized from microfiche so that employees could search the documents and find relevant information within them faster and so that another employee doesn't have to search for that same document on microfiche in the future.

Information Extraction for Mortgages and Contracts

There are numerous use-cases for natural language processing in banking, and those for document search and related functions are by far the most common. In fact, according to Emerj's research, roughly 23% of AI vendors selling into banking offer NLP software for information retrieval: in other words, search. Seven top 100 banks also claim they're using NLP for information retrieval as well, including JP Morgan Chase.

“

Digital documents are certainly better than physical ones for the purposes of standard search functions, but AI search promises to extract information from thousands of documents...

Banks could digitize all of their paper documents and still have difficulty searching for relevant information within them. Digital documents are certainly better than physical ones for the purposes of standard search functions, but AI search promises to extract information from thousands of documents when that information isn't identical in every document it exists. This kind of functionality is right now only possible with artificial intelligence.

Suppose a bank digitizes mortgage agreements from paper. This likely allows employees to read through them faster, but they may still need to read most of the contract to find the information relevant for them. An AI-based document search or information extraction application, what Iron Mountain calls “document understanding,” could allow the employee to find information such as:

- Mortgages of a certain amount issued within a certain date range
- Mortgages issued in a certain geolocation within a certain date range
- Mortgage agreements containing specific clauses or iterations of those clauses

Document search and information extraction applications could present these mortgages to a bank employee even if the information within them is not in the same format or isn't said in the same way.

”

This applies to the Mortgage QA process as well: a bank must ensure the information within a customer's loan file is complete (all the forms are present) and that all of their forms are filled out completely. An information extraction software may be able to pull out the customer's name, social security number, the APR, and other pertinent information as it appears in various places across all of the customer's documents, even if that information is written differently in different places of the documents for different loans.

A customer named Robert might sign their name "Bob," for example, or they might have missed a digit in their social security number. In theory, the NLP software would still extract this information as the customer's name and social security number, but would flag the inconsistency as an exception. This would allow a bank employee to verify and correct the information or ask the customer to update their information as needed.

Information Extraction for Human Resources and Compliance

In areas where data privacy laws such as GDPR are in effect or will be soon, banks will likely need to figure out how to find all of the information they have on a customer or employee, be able to produce this information if requested, and be able to prove that they've purged it if the customer or employee ever asks them to do so. This can be a challenge in and of itself, but gets even more difficult for customers and employees whose information is stored in part within physical documents. Conzelmann explains with a personal anecdote:

I've been at Iron Mountain a long time. When I first came, there were physical pieces of paper that were filled out as part of my employee file. Those are still sitting somewhere. But there was also digital information that was collected when I had my review last month. So how does [a bank] go across those different repositories of information...and be able to answer questions like...'Give me all of this employee's personal information across all of these channels.'

“

A search application that allows a bank's HR or customer service department to quickly find all of an employee's or customer's information may be necessary in the future as data privacy laws become more ubiquitous.

”

Finding where that information is located is paramount for banks that want to remain within compliance, and although the integration time will differ depending on the way the bank is already organizing their digital documents, a bank might benefit from implementing an AI-based document search system for compliance purposes.

Under these data privacy laws, banks need to be able to present a customer or employee's personal information to them when they ask for it. A search application that allows a bank's HR or customer service department to quickly find all of an employee's or customer's information may be necessary in the future as data privacy laws become more ubiquitous in many parts of the world.

Emerj's research corroborates that there's a need for AI-based solutions for compliance. They found that 12 AI vendors offer compliance solutions to banks, about 15% of the number of AI vendors selling into banking. They also found that, on average, AI vendors offering compliance solutions were relatively credible, scoring a 3.1 out of 4.0 on their Average Expertise and Funding score, which scores a vendor on the AI experience of their team and how much funding they've raised.

This indicates that banks looking to adopt AI for compliance at their companies are likely to work with an AI vendor that has the technical staff to back up the claims they make about their software.

In other words, compliance vendors are likely to know what they're talking about when it comes to artificial intelligence and machine learning.

THE BOTTOM LINE

What **BANKS** Need to Know

Banks have options when it comes to natural language processing solutions for extracting information from digital documents. They have much less options for digitizing their paper documents, although Emerj suspects that the pool of solutions for this use-case will widen over time.

The bottom line is that banks perhaps more than any other financial institutions are dealing with an inordinate amount of documents in a variety of formats both physical and digital, and they struggle with searching through these documents to generate customer analytics, solve customer support inquiries, and perhaps most importantly, remain compliant with local and regional laws.

This will likely become more difficult with the continued introduction of data privacy laws, and banks could lose hundreds of millions of dollars in fines if they aren't efficient and organized enough to provide customers with their personal information and purge it on request.

Banks that have the resources to commit to building an AI-based information extraction product in-house or working with a credible AI vendor may come out ahead of even the largest banks that struggle to digitize their millions of legacy documents, let alone implement an AI search function for them.

Conzelmann puts the potential value of AI-based search applications succinctly:

The power of machine learning and AI is that you can do this at scale across millions of documents even when you're dealing with disparate, different-looking content.

Banks could save millions on time-consuming processes that involve manually searching through paper documents, microfiches, PDF scans, and digital forms in a variety of file types. Although absolutely no bank should implement AI without a thorough understanding of the data, talent, time, and resources it requires, banks that are truly ready for AI might want to consider a document digitization or search application. Emerj suspects that applications like these are likely to become universal in the coming decade, especially in response to GDPR and similar regulations.

“

Banks could save millions on time-consuming processes that involve manually searching through paper documents....

”

INFORMATION EXTRACTION IN INSURANCE Claims and Underwriting

“

People are generally very good at finding patterns within datasets, but this ability dulls as we're presented with more and more data. A team of chief claims officers... might still spend months sifting through millions of claim forms to garner any reasonably accurate insights from them.

”

Customer data is essential for insurance firms to stay competitive in the coming decade. Insurance companies at present have backlogs of data on past and existing customers in the form of policy agreements, applications, and claims forms. They've also collected millions of images showing car damage, property damage, and personal injuries.

Patterns exist within this data that could inform the decisions of various insurance departments. Discovering these patterns, however, is a challenge. People are generally very good at finding patterns within datasets, but this ability dulls as we're presented with more and more data. A team of chief claims officers, for all intents and purposes experts when it comes to dealing with claims data, might still spend months sifting through millions of claims forms to garner any reasonably accurate insights from them.

This challenge is compounded because large insurance enterprises are still not entirely digital. In other words, this backlog of claims forms and policy agreements is still partly a collection of paper documents. Older documents are likely stored off-site in various locations across the region the insurance firm is operating in. Global firms may even store these documents in other countries.

What this means is that there are entire time periods of insurance data that are difficult to access at any given moment. Most insurance firms also still accept paper claims forms and applications, and they take payment and send claims payouts via check.

Not only that, but even digital information can be stored in systems that don't communicate with one another. The claims department at a large division of a global insurance enterprise might use a completely separate system for dealing with claims forms than the underwriting department at another division of the same enterprise. As a result, insurance firms struggle to keep all of their customer data in the same location.

“

Some of the largest insurance enterprises in the US are already implementing AI solutions for functions such as customer service.

”

For example, if an employee at a nation-wide auto insurance enterprise wanted to figure out the optimal premium that a customer should pay, they would need to find patterns across similar customers. Perhaps the customer is in their 40s, puts 300 miles on their car every week, and lives in a high-crime area. How much is this customer worth to the insurer?

That isn't something one can accurately determine without aggregating the lifetime value of every customer of a similar demographic. This would require underwriters to sift through thousands of past customer records, including the claims that customers of this demographic tend to file, the length of which they stay on the policy, and how much their premiums have been historically (which could vary wildly for a number of reasons).

These documents may or may not be digitized, and so underwriters may in some cases need to look through boxes of paper documents in order to find policy agreements, claims forms, and other documents belonging to customers that fit the demographic. This is a rigorous and time-consuming task, and so underwriters tend to settle for historical precedent that's easily accessible to them when determining premiums.

Artificial Intelligence, on the other hand, is quite good at dealing with large volumes of data. Whether or not AI upends the insurance industry remains to be seen, but some of the largest insurance enterprises in the US are already implementing AI solutions for functions such as customer service.

Information extraction/document understanding is a more nascent use-case for AI in insurance. That said, Emerj suspects that in the coming few years, this use-case will become more ubiquitous in the insurance industry. This is because information extraction software promises to reduce the time that underwriters and other insurance employees spend searching through documents.

The ability to search through digitized documents is made possible with natural language processing; the ability to digitize paper documents so that they're searchable with an NLP software is made possible with machine vision. More specifically, optical character recognition serves to read printed and handwritten letters and transcribe them into digital text.

Digitizing Paper Claims Forms and Images

Insurance enterprises struggle to answer simple questions about how to price their policies for maximum profit and how to accurately adjust claims for minimal claims leakage. This in part is due to the inability to access historical customer data that in many cases is stored in physical documents.

“

Two different adjusters might look through the same claims... and come up with different payout amounts. Both of these amounts might be more than what the damage actually costs to repair.... Artificial intelligence could help claims adjusters reduce claims leakage.

”

Digitizing these documents is the first step in extracting information from them, and it's a necessary step for feeding the data in these documents into an artificial intelligence algorithm.

At present, a claims adjuster that wants to determine the optimal payout to a customer whose home is partially flooded may need to search through past paper claims forms to get a sense of what customers were paid historically for similar damages.

The key is that “similar damages” is subjective and requires discretion on the part of the claims adjuster. Adjusters often need to look at the images customers provide and make an assessment about how much repairs might cost based on a variety of factors.

Two different adjusters might look through the same claims form and the same images and come up with different payout amounts. Both of these amounts might be more than what the damage actually costs to repair, and the insurance company won't find this out until later.

Artificial intelligence could help claims adjusters reduce claims leakage, but only if the claims forms and images attached to them are digitized. Employees at the insurance firm could scan physical documents and photographs, turning them into PDFs or image files.

Conzelmann pointed to another feature robust platforms may offer: the ability to find similar images. According to Conzelmann, “Adjusters can simply ask for similar images to the one showing the damage for the claim they are working on and quickly find relevant claims that had similar damage.”

Then, an OCR software could transcribe the letters on the documents into digital text, thus making the text “machine readable,” or ready for feeding into a machine learning algorithm. After training the algorithm to suit the insurance firm’s purposes, an employee would in theory be able to search for specific information within these documents.

For example, they might be able to pull up historical claims forms for property damage of a certain amount. This would reduce the time adjusters spend searching through paper documents for the same information.

Machine vision software for image recognition could also classify images of damage by damage severity and by the amount that was paid out to the customer for that damage. This classification could be used as a factor for determining the optimal payout on a claim.

“

Even digital documents can be unorganized. Many exist in a variety of different systems across an insurance enterprise’s divisions and branches.... AI could help search through these disparate data sources.

”

This would entail a prescriptive analytics capability that would use a customer’s demographics, the text information on their claims form, and the images attached to their claims to suggest the optimal payout for that customer’s claim. This is also why claims processing and adjustment are underdeveloped use-cases for AI in insurance. They require a robust network of machine learning capabilities involving natural language processing, optical character recognition, machine vision for image recognition, and prescriptive analytics.

Information Extraction for Underwriting

Although prescriptive analytics capabilities are rare in insurance due to the varied types of data (text, image, numeric), claims adjusters and underwriters can still use natural language processing software to search through their stores of documents once they’re digitized. This could prove beneficial because even digital documents can be unorganized. Many exist in a variety of different systems across an insurance enterprise’s divisions and branches. They may even exist in different folders and organizational structures within the same department at the same

branch. Conzelmann spoke to us about how AI could help search through these disparate data sources, emphasizing the value of AI for this scenario:

In addition to enriching the metadata by extracting information from the documents, there could be metadata that you have in a repository already, it could be metadata that's available out in the market for purchase, it could be publicly available information...the key is to be able to create the relationship between all of these different bits and pieces and making it all part of the metadata that's attached to an asset.

The "asset" in this case could be a particular insurance customer or an insured property.

An information extraction and document search application could prove useful for searching through digital documents across the insurance firm's numerous branches if those documents are stored in the cloud or some file-share program.

For example, an underwriter might be able to answer the question "Should I onboard this customer?" much faster than they would if they had to manually search through digital documents one by one for information that might help them answer that question. Instead, the underwriter could pull up records from past customers similar to the customer they're looking to onboard. The underwriter could then search through these records for information about claims the customer has made and customer lifetime value, and this could give them a better idea of whether or not to onboard the potential new customer. It might also inform the premiums they offer that customer.

An underwriter could make their decision about the customer in a matter of minutes as opposed to the hours or days it may take them to do so manually. This has clear savings benefits for the insurance company, as well as customer experience benefits. It could allow an insurance firm to move closer to offering "on demand insurance," the ability for an insurance company to onboard a customer when the customer needs insurance (such as the day they're diagnosed with an illness).

“

With AI, an underwriter could make a decision about the customer in a matter of minutes as opposed to the hours or days it may take to do so manually.

”

Insurance firms are scrambling to cater to millennial customers, who more than any other generation expect a level of speed congruent with their experience growing up with the internet. They don't find it necessary to show up at a physical location and discuss their insurance policies. They want to be able to apply via chatbot or email, and they want to start their policies very shortly thereafter. AI-based information extraction software could help with this, potentially giving insurance firms that integrate it an edge over their competitors.

THE BOTTOM LINE

What **INSURANCE FIRMS** Need to Know

Claims processing and underwriting are two areas of insurance that could benefit from AI-based information extraction/document search software. That said, neither are developed use-cases for AI in insurance right now. This will likely change over time as AI becomes more accessible to businesses, perhaps with autoML or a shift in the culture of innovation at older enterprises. At that point, AI use-cases in insurance will likely move from the cost-saving benefits of document search applications to more complex machine learning systems that involve document search, machine vision, and prescriptive analytics, allowing for capabilities that drive growth, such as tailor-made insurance policies.

For now, information extraction and document digitization software could reduce the time underwriters and claims adjusters spend searching for information through paper and digital documents that they regularly use to make decisions about premiums and claims payouts. A less laborious and more organized search process could

“

Purchasing a platform such as that offered by Iron Mountain can remove the need for a company to train its own ML engine to recognize the documents being ingested, and requires less resources dedicated to the project.

”

result in more profitable premiums and less claims leakage, although without a prescriptive analytics function, the premium and payout amounts are still left up to underwriters and adjusters (in other words, human error).

That said, artificial intelligence is difficult to integrate into existing businesses. At the same time, there are ways to mitigate spend and achieve a quicker time to market. Currently there aren't many AI vendors that offer products clients can use “out of the box” or that are “plug and play,” so to speak.

Purchasing a platform such as that offered by Iron Mountain can remove the need for a company to train its own ML engine to recognize the documents being ingested, and requires less resources dedicated to the project, as well as the ability to gain a competitive advantage by bringing a business' AI solution online faster than it would be able to if it had to build it in-house.

Those that offer something close to this are often in customer service or similar horizontals that don't differ much from company to company, although it's very likely that these products still require training on the part of the client. In most cases, however, the AI vendor will work with the client to train the software, and the client may not require a team of in-house data scientists.

As such, working with an AI vendor like Iron Mountain will often require less from the client than building an AI application in house. Iron Mountain specifically claims their information extraction software comes built-in with AI capabilities.

In summation, insurance firms might want to consider AI-based document search and document digitization solutions, especially older firms that have legacy systems and stores of physical documents in a variety of disparate locations. But in doing so they should consider their business needs and the time and resource-intensive nature of an AI project before rushing to work build a solution from scratch with an AI vendor.

About
Emerj

Emerj helps organizations develop AI strategies and make AI investments to reduce risk and improve their competitiveness. From the World Bank to global pharmaceutical companies, we work with leading organizations that need to make critical decisions about AI and technology strategy.

Our market research and advisory services are focused exclusively on competitive artificial intelligence strategy and AI market insights. Through our proprietary research methods and network of global advisors, we keep a pulse on what's possible and what's working, mapping emerging AI capabilities and helping companies leverage them to win.

About
Iron
Mountain

Iron Mountain Incorporated (NYSE: IRM) is the global leader for storage and information management services. Trusted by more than 225,000 organizations around the world, the company helps customers lower cost and risk, comply with regulations, recover from disaster, and enable a more digital way of working.

Iron Mountain InSight is a content services platform that provides actionable business insights and predictive analytics through Machine Learning (ML)-based classification of a company's physical and digital information, which adds structure, context, and meta-data to information to make it more usable. The resulting enriched content can then enable enhanced automated governance and workflow throughout an oil and gas organization.