ZyLAB®
eDiscovery & Information Risk Management

Breaking the Fraud Triangle

ZyLAB Fraud Prevention and Detection Program

A ZyLAB Solution for
Corporate Internal Investigation
& Security Departments

By Johannes C. Scholtes, Ph.D.
Professor at Text-Mining

December 2013

Contents

Introduction 3

Managing the Dark Side of Big Data 4

Content analytics and text mining 5

Finding without knowing exactly what to look for 5

Faceted search and information visualization 7

The Fraud Triangle 8

ZyLAB Fraud Prevention and Detection Program 9

Dealing with privileged, data protection and privacy concerns 13

Intelligent redaction 14

Defensibility and chain of custody 15

Validating the automated processes 16

Exploratory search 17

De-Duplication 19

Dealing with documents in other languages 21

Multi-Media Search 24

Customer Case: The FBI Enron Case 31

Customer Case: OLAF 32

Business Drivers 32

Special Benefits of ZyLAB to OLAF 33

Customer Case: UN War Crimes Tribunals 34

About the Author 35

About ZyLAB 36

Introduction

Economic crimes such as corruption and fraud are difficult to detect and prevent, but the financial and reputational consequences and the growing public and political demand for harsh action on corporates whose employees break the law are forcing companies to review their security and compliance policies to limit the extent to which fraud can take place.

In many companies, fraud is detected more often by anonymous tips or by accident, than through pro-active internal audits. One of the challenges is the complexity of “Big Data” (see next chapter) and the fact that almost 80% of enterprise content today is unstructured and therefor seems hard to examine.

In his Fraud Triangle, criminologist Dr. Donald R. Cressey identifies three components that are present where fraud exists: (1) incentive or pressure, (2) opportunity, and (3) rationalization form the three angles of the so-called Fraud Triangle. Breaking this Fraud Triangle is the key to fraud deterrence and implies that if an organization removes one of the elements in the Fraud Triangle, the likelihood of fraudulent activities is highly reduced.

Most anti-fraud activities only focus on the structured data, mostly financial administrations and ERP systems. However, structured data is only 10-20% of all data. Often, the components leading to the identification of the angles of the triangle are hidden in the vast volume of unstructured data, formed by e-mails, user documents, presentations, and web content.

In this whitepaper we illustrate how Fraud Triangle Analytics (FTA) can support corporate security officers and internal auditors with internal investigations on their Big Data collections to prevent and detect fraud as early as possible. We show how Big Data can be researched, extracted and presented in a transparent structure so the results of the investigation can be used to automatically detect potential fraudulent activities and prevent these in the future. In addition, we’ll show how to deal with confidential, privacy or privileged information and data protection concerns.

Managing the Dark Side of Big Data

The ongoing information explosion is reaching epic proportions and has earned its own name: Big Data. Gartner, and now much of the industry, use the so-called "3Vs" model to classify Big Data: "Big data is high **V**olume, high **V**elocity, and/or high **V**ariety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization." This clearly describes the type of data auditors and internal investigators have to deal with today.

Big Data encompasses both challenges and opportunities. The opportunity, as focused on by many auditors, is to use the collective Big Data to identify and recognize patterns of behavior and collect evidence of fraudulent behavior. But there is also a dark side to Big Data: investigating and analyzing Big Data collections is an enormous challenge and puts lots of pressure on internal investigative teams and resources. New data formats (**multimedia**, in particular), **different languages**, **cloud** and other **off-site locations** and the continual increase in regulations and legislation—which may contradict previous protocols—add even more complexity to this puzzle.

Defensibility, auditing, quality control and the chain of custody are paramount for classification processes during internal and law enforcement investigations. If you cannot explain exactly how the classification and correlation process are implemented and executed, you will have a hard time in court.

Information protected under **privacy and data protection regulations** or **attorney-client privileged information** for example requires constant attention during audits and investigations. Violations of such regulations and rights are very counterproductive and result in many problems later in the investigation or during trials and even lead to an unfavorable judgment or dismissals.

At the same time, every investigative team has **limited resources and will have to answer to very strict deadlines**. This requires strict resource planning and constant monitoring of progress.

Several of the above mentioned requirements are contradictory in nature. In essence, we need computers and advanced algorithms to deal with these contradicting requirements and at the same time battle the data explosion. This is where data content analytics come into play.

Content analytics and text mining

Content analytics such as text mining and machine learning technology from the field of artificial intelligence can be used very effectively to manage Big Data. Think of tasks such as, identifying exact and near-duplicates, structuring and enriching the content of text and multimedia data, identifying relevant (semantic) information, facts, events, and ultimately, predicting or anticipating the most logical next thing to happen or to classify unstructured information automatically according to the requirements of a particular investigation. As a result of these content analytics efforts, users can explore and understand repositories of Big Data better and also apply combinations of advanced search and data visualization techniques easier.

Finding without knowing exactly what to look for

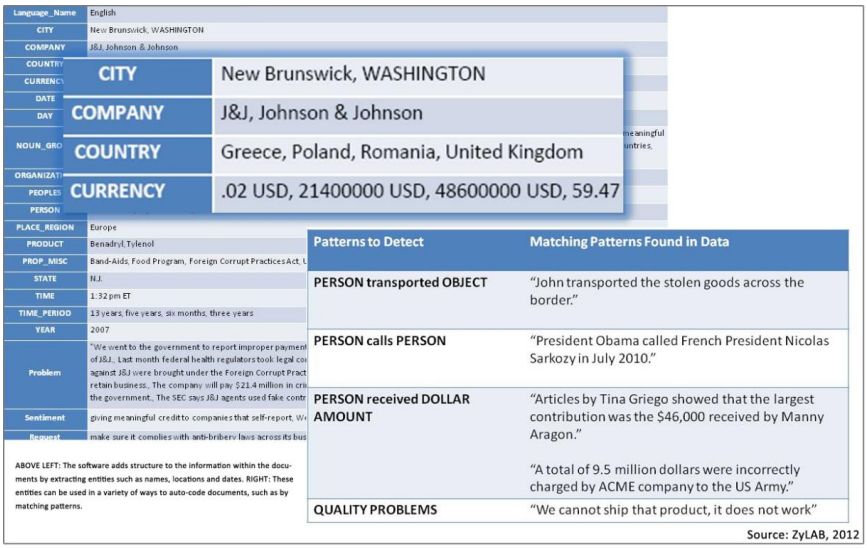
Traditional search requires a user to know what they are looking for. Text mining attempts to discover information in a pattern that is not known beforehand through the use of advanced information extraction techniques as well as machine learning. By focusing on patterns and characteristics, text analytics can produce better search results and deeper data analysis, thereby providing quick retrieval of information that otherwise would remain hidden.

One of the most compelling differences with regular (web) search is that typical search engines are optimized to find only the most relevant documents; they are not optimized to find all potentially relevant documents. The majority of commonly used search tools are built to retrieve only the most popular hits—which simply doesn't meet the demands of law enforcement or legal investigations that require a more exploratory type of search.

In addition, regular search does not provide any mechanism to identify, predict or classify specific (hidden) patterns, which could provide valuable insights into Big Data. This is where text analysis can make a big difference as it is particularly interesting in areas where users must discover new and unknown insights from Big Data.

This is achieved by enriching the original data with additional meta information that allows for not only more sophisticated search capabilities, but also for different context specific functions such as sorting, organizing, categorizing, classifying, grouping and structuring data based on additional meta-information. In addition, utilizing this additional meta-information will open a whole spectrum of additional search techniques, such as clustering, visualization, advanced (semantic) relevance ranking, automatic document grouping, predicting patterns and categorization.

- Concept extraction: extraction of predefined (full-text) queries that identify document and meta-information content with specific combinations of keywords or (fuzzy and wildcard) word patterns in.
- Entity extraction: extraction of basic entities that can be found in a text such as: people, companies, locations, products, countries, and cities.
- Fact extraction: these are relationships between entities, for example, a contractual relationship between a company and a person.
- Attributes extraction: extraction of properties of the found entities, like title, a person's age and social security number, addresses of locations, quantity of products, car registration numbers, and type of organization.
- Events extraction: these are interesting events or activities that involve entities, such as: "one person speaks to another person", "a person travels to a location", and "a company wires money to another company".
- Sentiment detection: finding documents that express a sentiment and determine the polarization and importance of the sentiment expressed.
- Concept, Context and Discourse Detection: find the *red-line* and the story: every crime or fraud has its own story line.
- Extended natural language processing: Part-of-Speech (POS) tagging for pronoun, co-reference and anaphora resolution, semantic normalization, grouping, and entity boundary detection.

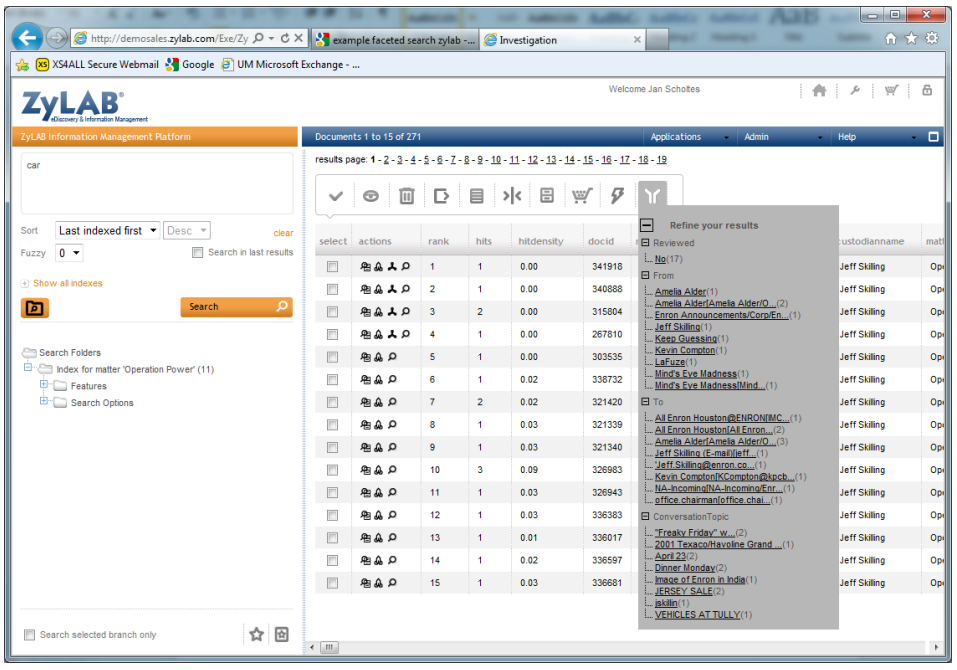


Examples of Extracted Entities, Attributes, Facts, Events and Sentiments

Faceted search and information visualization

Following content analytics processing, the original data is enriched with additional meta-information. Now a new, broader range of analysis and search techniques can be applied. Using the enriched data, it is now possible to organize and visualize data, make complex statistical analysis, call-up similar documents when searching, sort & search on specific features, cluster on attributes, navigate using the complete text of a document and using the available document attributes, etc.

Faceted search is one of these advanced search techniques, also called faceted navigation or faceted browsing. Faceted search uses a collection of information that is represented using a faceted classification,[1] allowing users to explore by filtering available information. Facets are often derived by analysis of the text using entity extraction techniques or from pre-existing fields in the database such as author, description, language, and format. This approach permits existing data to have this extra metadata extracted and presented as a navigation facet.



The gray box provides an example of automatically derived faceted search to refine results from the ENRON data set based on certain values in the facets.

[1] Faceted Classification is an analytic-synthetic classification scheme. It takes information and divides it into categories that express its different facets or attributes rather than assigning it a single, rigid value (Wikipedia)

The Fraud Triangle

A popular model to identify fraud is the *Fraud Triangle*[2] that explains the three factors that cause someone to commit occupational fraud:

1. Motive (or pressure) – the need for committing fraud (need for money, etc.);
2. Rationalization – the mindset of the fraudster that justifies them to commit fraud;
3. Opportunity – the situation that enables fraud to occur (often when internal controls are weak or nonexistent).

The fraud triangle originated from Donald Cressey's hypothesis[3] that "Trusted persons become trust violators when they conceive of themselves as having a financial problem which is non-shareable, are aware this problem can be secretly resolved by violation of the position of financial trust, and are able to apply to their own conduct in that situation verbalizations which enable them to adjust their conceptions of themselves as trusted persons with their conceptions of themselves as users of the entrusted funds or property."

Breaking the Fraud Triangle is the key to fraud deterrence and implies that an organization must remove one of the elements in the fraud triangle in order to reduce the likelihood of fraudulent activities. "Of the three elements, removal of Opportunity is most directly affected by the system of internal controls and generally provides the most actionable route to deterrence of fraud"[4].

This is why more and more organizations increase the number of internal audits and fraud investigations. As ever more evidence of non-compliance and fraud is hidden in unstructured data, analyzing this data to find the three key elements of the Fraud Triangle is a good start to trigger more in-depth investigations.

---

[2] Source: http://www.acfe.com/fraud-triangle.aspx.
[3] Donald R. Cressey, Other People's Money (Montclair: Patterson Smith, 1973) p. 30.
[4] Cendrowski, Martin, Petro, The Handbook of Fraud Deterrence.

ZyLAB Fraud Prevention and Detection Program

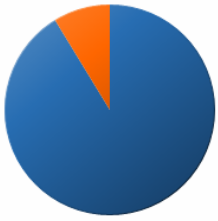
ZyLAB Fraud Prevention and Detection Program is an extensive program to detect and prevent fraud. The program combines components of the ZyLAB eDiscovery and Information Risk Management Platform with best practice templates, implementation guidelines and customizable, multilingual libraries to detect and prevent fraud. The basis for the new program is Fraud Triangle Analytics (FTA) that can be used with the data in an existing data repository, or as part of a data monitoring scheme.

Fraud Triangle Analytics is a powerful technique that through libraries of specific keywords that relate to incentive or pressure, opportunity, and rationalization, connects electronic communications to the angles of the Fraud Triangle. With the use of FTA lexical, syntactic and semantic patterns that indicate possible fraudulent activities are recognized in almost any kind of data, regardless of location or language.

ZyLAB has developed a libraries of keywords for every angle of the Fraud Triangle, syntactic and semantic patterns which identifying fraud related information with the highest possible precision and recall.

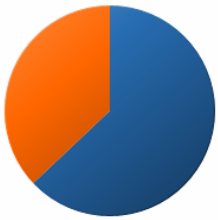
After the raw data has been processed and has gone through the Fraud Triangle Analytics scenario, all documents responding to one of the three Fraud Triangle angles are automatically identified and labeled as Potential Responsive to Pressure, Rationalization or Opportunity; the three different angles of the Fraud Triangle. Because these documents are automatically labeled during the analytics scenario, they will automatically be visualized in the facetted view of ZyLAB's review interface.

Fraud Triangle Pressure



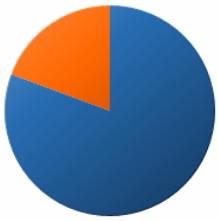
No	Yes
32938 Documents	3148 Documents

Fraud Triangle Rationalization



No	Yes
22717 Documents	13369 Documents

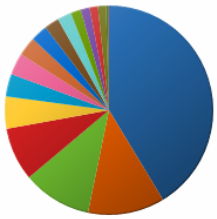
Fraud Triangle Opportunity



No	Yes
29091 Documents	6995 Documents

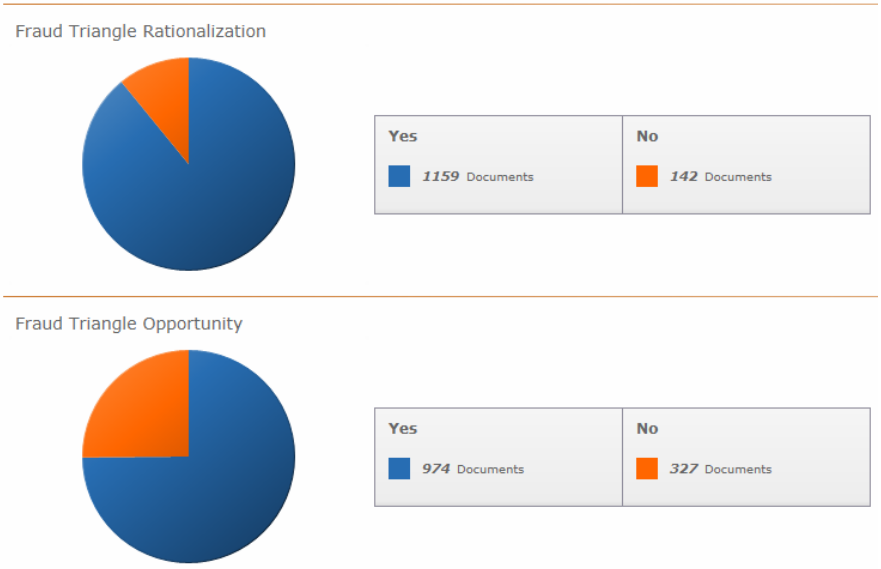
In order to get an overview of custodians of these potential responsive documents, one could drill down in the data by clicking one of the three facets of the Fraud Triangle, like “Pressure”.

Custodian

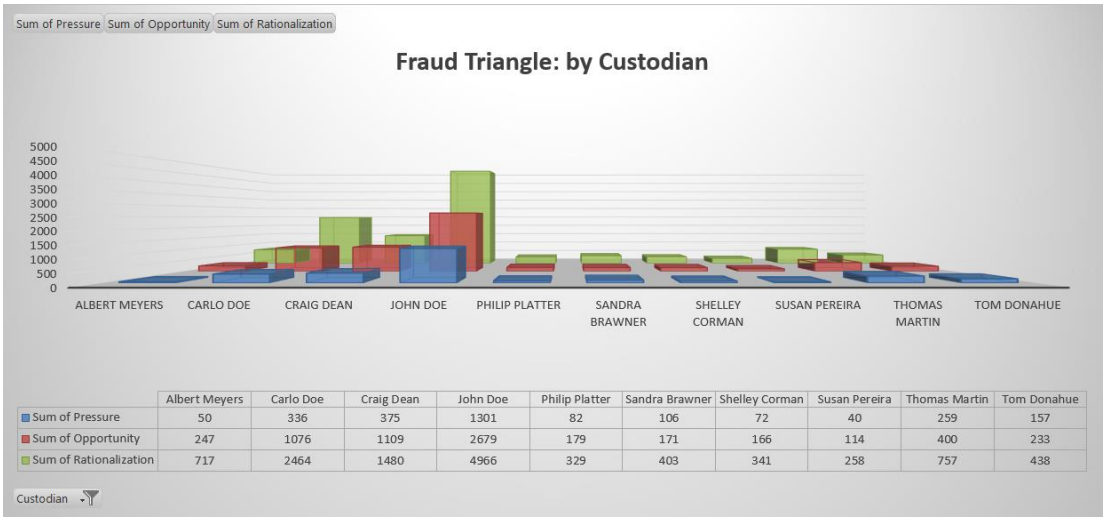


John Doe 1301 Documents	Craig Dean 375 Documents	Carlo Doe 336 Documents
Matt Motley 110 Documents	Sandra Brawner 106 Documents	Judy Townsend 93 Documents
Jeff King 69 Documents	Albert Meyers 50 Documents	Monique Sanchez 44 Documents

Drilling down further, one could select a specific custodian, like John Doe and see how many of his documents are potentially responsive to the other Fraud Triangle angles like “Rationalization” and “Opportunity”.

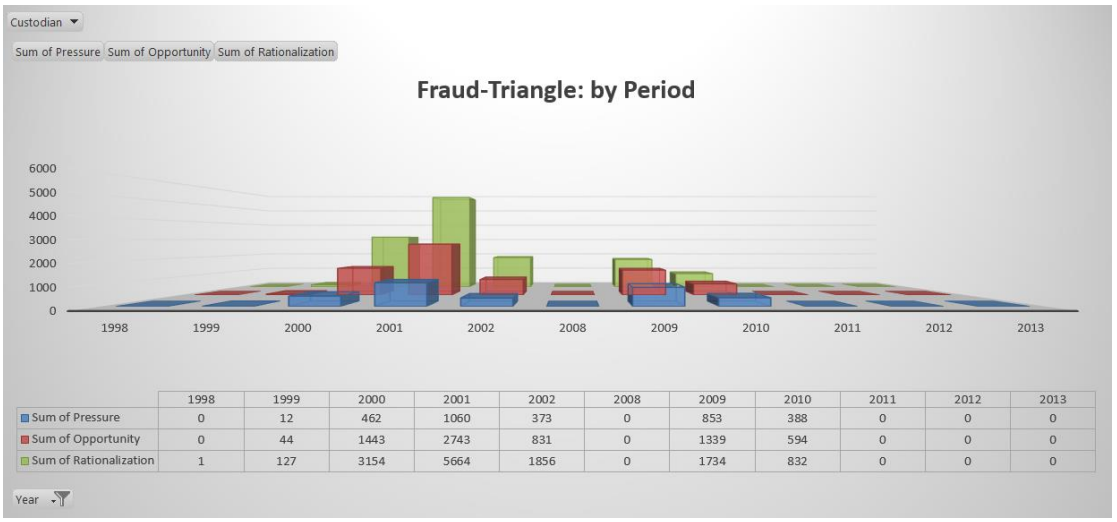


For further Fraud analytics ZyLAB provides for an analytics tool that puts all angles from the Fraud Triangle in perspective of Custodian, Time and Frequency.

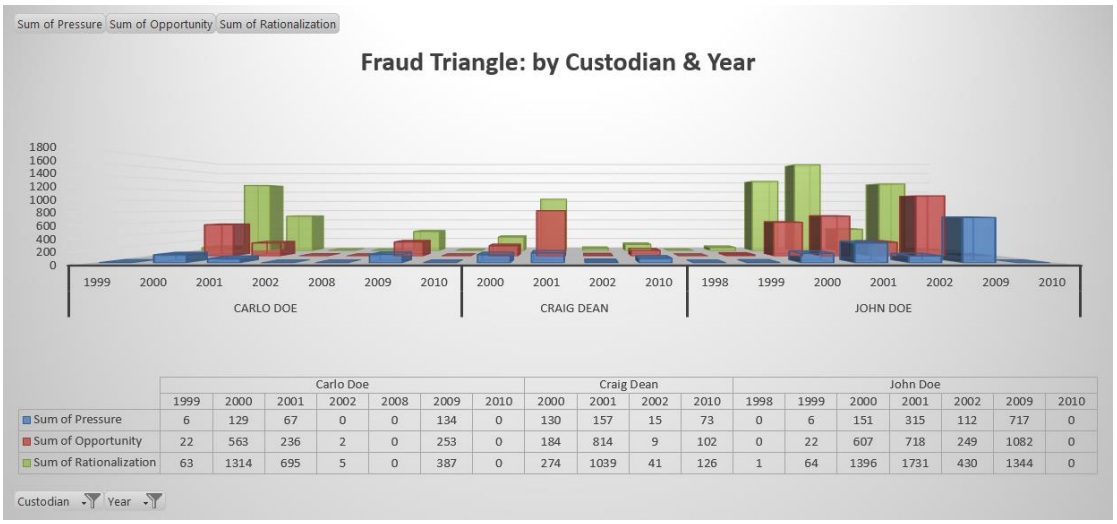


	Albert Meyers	Carlo Doe	Craig Dean	John Doe	Philip Platter	Sandra Brawner	Shelley Corman	Susan Pereira	Thomas Martin	Tom Donahue
Sum of Pressure	50	336	375	1301	82	106	72	40	259	157
Sum of Opportunity	247	1076	1109	2679	179	171	166	114	400	233
Sum of Rationalization	717	2464	1480	4966	329	403	341	258	757	438

Custodian



	1998	1999	2000	2001	2002	2008	2009	2010	2011	2012	2013
Sum of Pressure	0	12	462	1060	373	0	853	388	0	0	0
Sum of Opportunity	0	44	1443	2743	831	0	1339	594	0	0	0
Sum of Rationalization	1	127	3154	5664	1856	0	1734	832	0	0	0

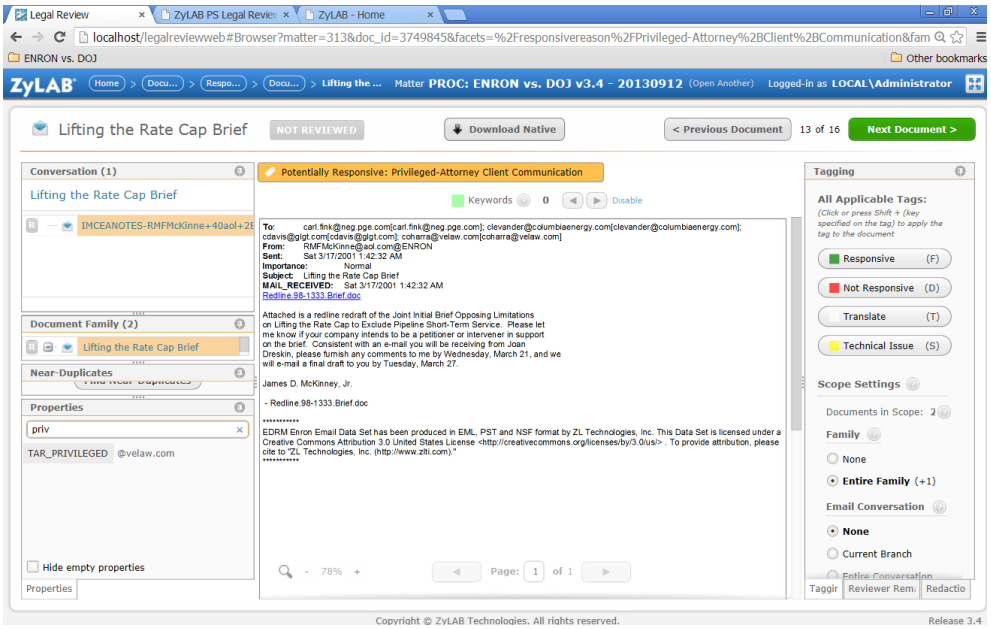


	Carlo Doe							Craig Dean				John Doe						
	1999	2000	2001	2002	2008	2009	2010	2000	2001	2002	2010	1998	1999	2000	2001	2002	2009	2010
Sum of Pressure	6	129	67	0	0	134	0	130	157	15	73	0	6	151	315	112	717	0
Sum of Opportunity	22	563	236	2	0	253	0	184	814	9	102	0	22	607	718	249	1082	0
Sum of Rationalization	63	1314	695	5	0	387	0	274	1039	41	126	1	64	1396	1731	430	1344	0

Dealing with privileged, data protection and privacy concerns

Even during an internal investigation, it is important to be compliant with Privileged, Data Protection and Privacy regulations. Without technology, it is almost impossible to conduct an investigation without (unknowingly) violating such regulations.

Email communication or transcripts and recordings of conversations between suspects and their attorneys can easily end up in official case files, resulting in reduced sentences, possible mistrial and disciplined action against the officers responsible for the disclosure and violations of the rights of the suspects.

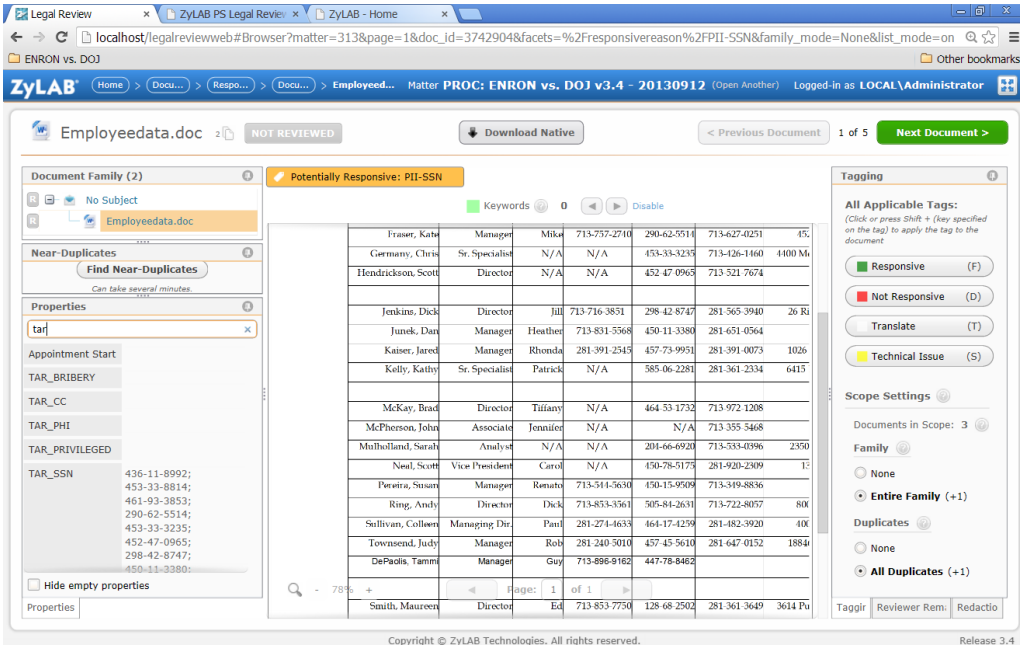


Example of an Attorney-Client Privileged E-mail from the ENRON Investigation

Not only can evidence containing protected personal identifiable information or protected health information accidently be disclosed; it can also become public information like the Enron Data Set is an industry-standard collection of email data that was previously hosted by EDRM and in 2012 became an Amazon Web Services Public Data Set. The Enron Data Set has served for many years as an industry-standard collection of email data for electronic discovery training and is a valuable public resource for all sorts of researchers from all disciplines. It has never been a secret that the data set that was originally made available by the Federal Energy Regulatory Commission (FERC)

contained a high level of personally identifiable information (PII) about the company's former employees.

ZyLAB assisted to protect the privacy of hundreds of individuals by locating private data in the data set. By using ZyLAB Visual Classification technology in combination with the existing deep processing, content analytics and search capabilities, several hard to find items like documents containing social security and credit card numbers, protected health information, 1040 tax forms, and even pornographic images have been identified.

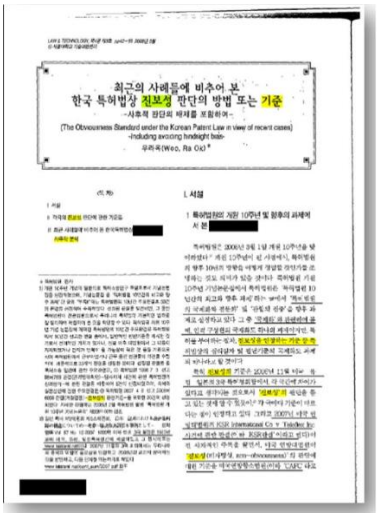


Example of a Document from the Public ENRON set still Containing SSN's.

Intelligent redaction

Once PII or other confidential information is identified, the necessary next steps can be taken. ZyLAB's intelligent redaction software tool supports automatic bulk redactions on keywords or patterns (i.e., any numbers appearing in the format of a Social Security Number) as well as on-the-fly electronic redactions of content.

Pre-emptive, role-based redactions of sensitive keywords and role-based folder access can reduce the risk of certain groups of people having an opportunity to leak trade secrets and un-redacted content.

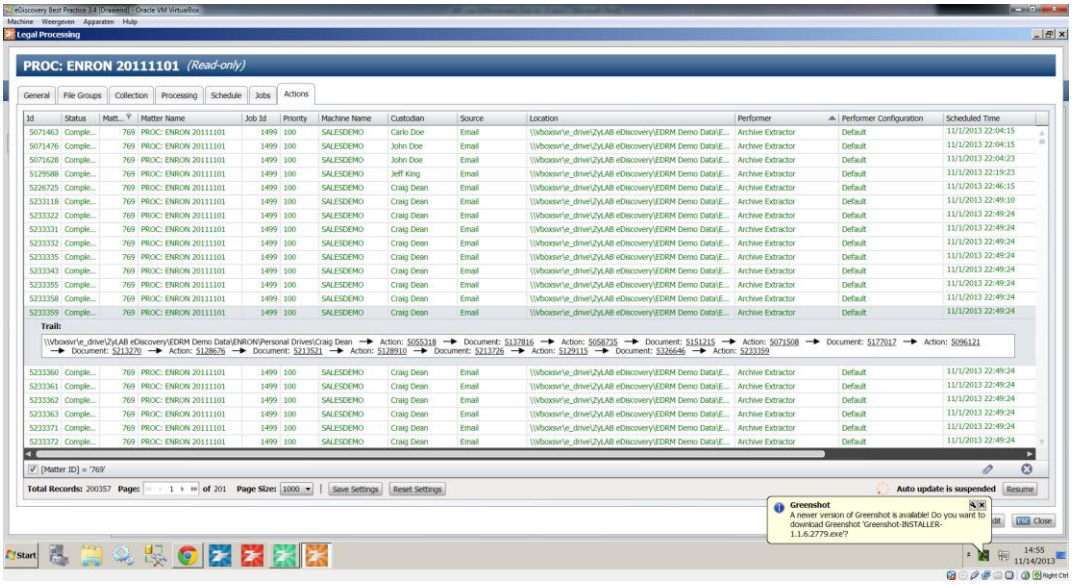


ZyLAB's software redacts in multiple languages, and multiple directions.

Defensibility and chain of custody

Using any type of automated processes also requires an unbiased evaluation of the results and defensible processes. In other words, the quality and reliability of the automatic structuring, enrichment, classification and prediction techniques needs to be measured by using existing best-practices. Only then will end-users accept the usage of such technology for mission-critical processes.

All automated processes have to be logged and documented, and they must always be applied with defensibility, auditing, quality control and the chain of custody in mind to avoid trouble down the road. This is paramount when automating legal processes like eDiscovery. Without being able to explain in court exactly how the automated process was implemented and executed, your counsel will have a hard job defending your case against opposing counsel. Existing case law referring to the usage of the technology is also essential in this process.



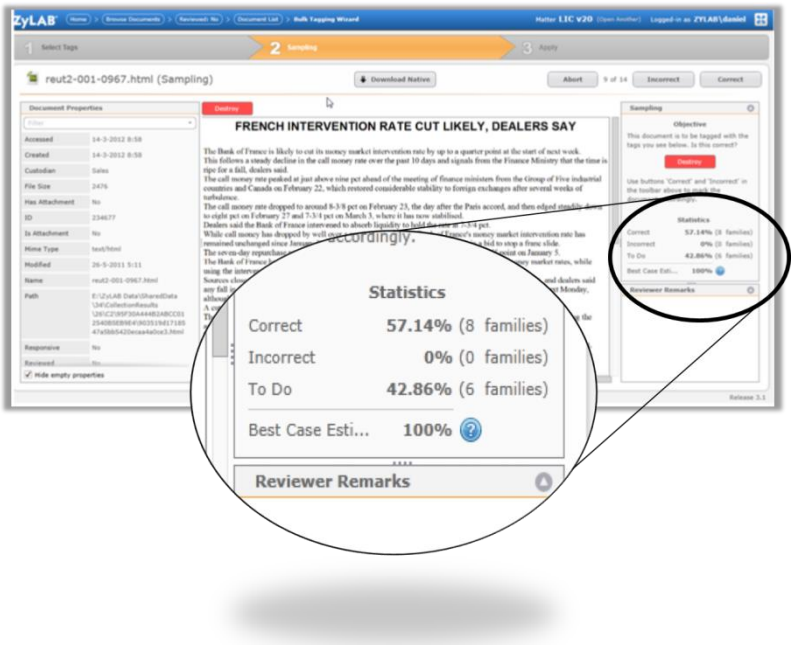
Full Chain of Custody Logging of all Automated Actions

Validating the automated processes

Good data sampling empowers practitioners to make informed decisions based on real-time insight to accelerate various iterative eDiscovery and records management processes whilst keeping the automated processes defensible. Data sampling is relevant for the defensibility and quality control of eDiscovery in relation to any automated or machine assisted process, where data is identified, collected, preserved, processed, analyzed, reviewed or produced.

With data sampling, a random subset of documents is selected automatically (based on certain predefined statistical properties or distribution) and the quality of, for instance, the processing, bulk tagging, or automatic redaction of these documents is verified manually by human reviewers.

Depending on the outcome of the manual verification, the results from the automated process is accepted or rejected. When it is rejected, the parameters of the automatic process can be adjusted and then re-sampled to verify a satisfactory improvement. Alternatively, the team can opt to perform the work manually if automation is deemed inappropriate for that process.



Data sampling helps to explain and defend the application of automatic technology in court. This is not a trivial task given the fact that most of these techniques are not commonly understood by those who may be involved in the law suit. Data sampling on the other hand, is a well understood method in court.

Exploratory search

To most people, there is just one type of search: type in a couple of keywords and hope that you will find what you are looking for. When looking at search, we can very clearly differentiate search for: web, e-commerce, enterprise, desktop, mobile, social, real time, discovery and information governance purposes.

In each of these different applications, search, relevance ranking, relevance feedback, user interaction, result navigation and document viewing have different purposes and also work differently under the hood. Basically, we can differentiate two main types of search: lookup search and exploratory search.

Lookup search is a search where a user typically knows exactly what he or she is looking for, where a user typically searches data where the content is well known and where the user is primarily interested in just the best document or web site and not in the return of all the potentially relevant documents and web sites. Lookup search is also referred to as web-search or portal search.

Typical lookup search engines are almost all web-search engines, but also the open source Java engine from Lucene. These engines work well for web portals, personal search, mobile or social search, but they are less suited for discovery, compliance, investigative, and legal search because they only return the best and not all potentially relevant results. They lack advanced navigation and often only have relevance feedback based on popularity or on other non-exhaustive techniques.

Also, these engines cannot handle wildcards, fuzzy searches and other “find more” or “find similar” search techniques very well (it may work, but often it becomes very slow on larger collections or it is limited to generative techniques that are not exhaustive).

Exploratory search is a specialization of information exploration which represents the activities carried out by searchers who are either:

1. Unfamiliar with the domain of their goal (e.g. the need to learn about the topic in order to understand how to achieve their goal).
2. Unsure about the ways to achieve their goals (technology or the process).
3. Unsure about their goals in the first place.

This is exactly the case in discovery, compliance, investigative, intelligence and information governance search applications.

Consequently, exploratory search covers a broader class of activities than typical information retrieval, such as investigating, evaluating, comparing, and synthesizing, where new information is sought in a defined conceptual area; exploratory data analysis is another example of an information exploration activity. Typically, therefore, such users generally combine querying and browsing strategies to foster learning and investigation.

Precision and Recall

Precision is the ability to retrieve the most precise results. Higher precision means better relevance and more precise results but may imply fewer results returned. For a query, recall means the ability to retrieve as many documents as possible that match or are related to a query. Recall may be improved by fuzzy searches, wildcards and synonym expansion. In information retrieval, there's a classic tension between recall and precision. By specifying more recall (trying to find all the relevant items), you often get a lot of junk. If you limit your search by trying to find only precisely relevant items, you can miss important items because they don't use quite the same vocabulary. ZyLAB offers an extensive set of tools to improve your recall and increase the precision.

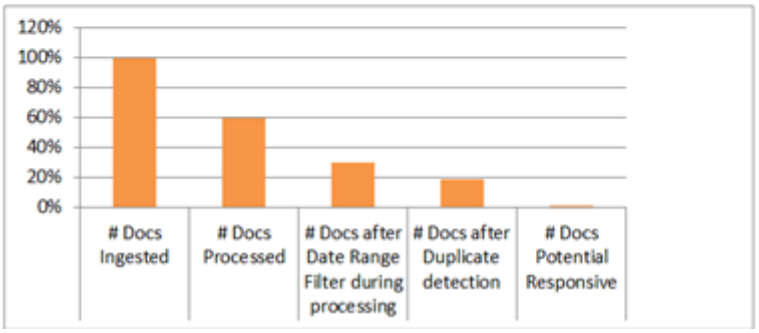
ZyLAB's precision and recall tools	
Precision	Recall
Boolean queries, Proximity queries, Quorum queries, relevance ranking, sorting, keyword in context, refine results, key field searches, progressive search, document preview, visualization	Fuzzy Search, Wildcard options, synonyms, vocabulary, text mining, translations, automatic notification, transliteration assistant

Exploratory search techniques are breaking through in search applications, but the required additional information to use them effectively depends completely on content analytics, text-mining technology and advanced result navigation and visualization. Also, document based relevant feedback, taxonomy support and extensive metadata management are essential tools. These will be discussed further on in this white paper.

De-Duplication

Email collections, electronically stored documents and the like all contain duplicate and near-duplicate messages and documents. These duplicate and near duplicates form large problems in applications where large volumes of electronic data are searched and reviewed by humans.

# GB Ingested	1.243	Gb	
# GB Filtered by Date Range Filtering on collection	746	Gb	
# Docs Ingested	12.283.931	Docs	100%
# Docs Processed	7.372.335	Docs	60%
# Docs after Date Range Filter during processing	3.733.463	Docs	30%
# Docs after Duplicate detection	2.345.987	Docs	19%
# Docs Potential Responsive	221.170	Docs	2%
# Pages OCR-ed	2.622.371	Pages	



An example of Automatically Reducing the Size of the Investigative Data Set

Detection of duplicate and near-duplicate emails (properties, email body and attachments), electronic documents or other electronic content (all referred to as objects), can help to reduce the data set with 90% or more, leaving only 10% or less of the original data.

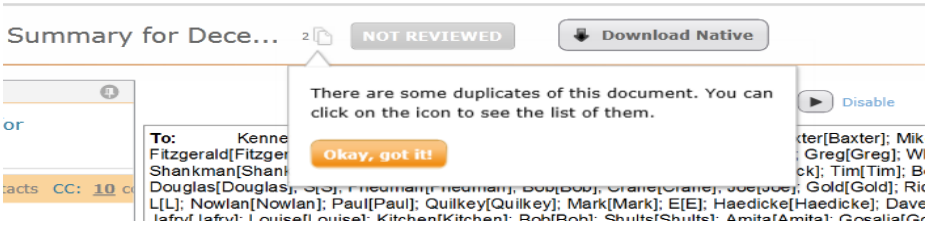
De-duplication methods

Several different de-duplication methods exist:

- Exact de-duplication: Detecting exact duplicates can be done reliably by using hashing techniques. In such cases (a combination of) the document textual content, properties or binary content is hashed with a MD-5, SHA-1 or other hashing algorithm. If two documents are exactly the same or if they have exactly the same document properties, then the hash must be exactly the same.
- Filtering with NIST on known files: The National Software Reference Library (NSRL-NIST) Reference Data Set (RDS) is a list of known electronic files from commercial software and data providers that do not contain any relevant data for investigations or eDiscovery. Known hashes from these files can be used to filter such files from large electronic data collections.
- Near de-duplication: these files are created for many reasons like the creation of slightly different versions of a document, different formats of a

document (like creation of a PDF from a Word file) or forwarded or copied and blind-copied emails.

For each type of duplicate, techniques exist that can identify, deal with or create reports of identified and remove (near)-duplicates.



Dealing with documents in other languages

High-stake investigations are not limited by national boundaries and no auditor can afford to miss relevant information because it is in a foreign language and the cost of translation is too high.

Multi-lingual text collection

Multi-lingual text collection hide more complexities than it initially look like, because, in addition to differences in character sets and words, text analysis makes intensive use of statistics as well as the linguistic properties (such as conjugation, grammar, tenses or meanings) of a language. These language dependencies need to be addressed when dealing with non-English content.

First, basic low-level character encoding differences can have a huge impact on the general searchability of data. Whereas English is often represented in basic ASCII, ANSI or UTF-8, foreign languages can use a variety of different code-pages and UNICODE (UTF-16), all of which map characters differently. Before an archive with foreign language content can be full-text indexed and processed, a 100% matching character mapping process must be performed. Because this process may change from file to file, and may also be different for various electronic file formats, this exercise can be significant and labor intensive. In fact, words that contain such language-specific special characters as ñ, Æ, ç, or ß (and there are hundreds more like them) will not be recognized at all if the wrong language model is chosen or if the language is not known.

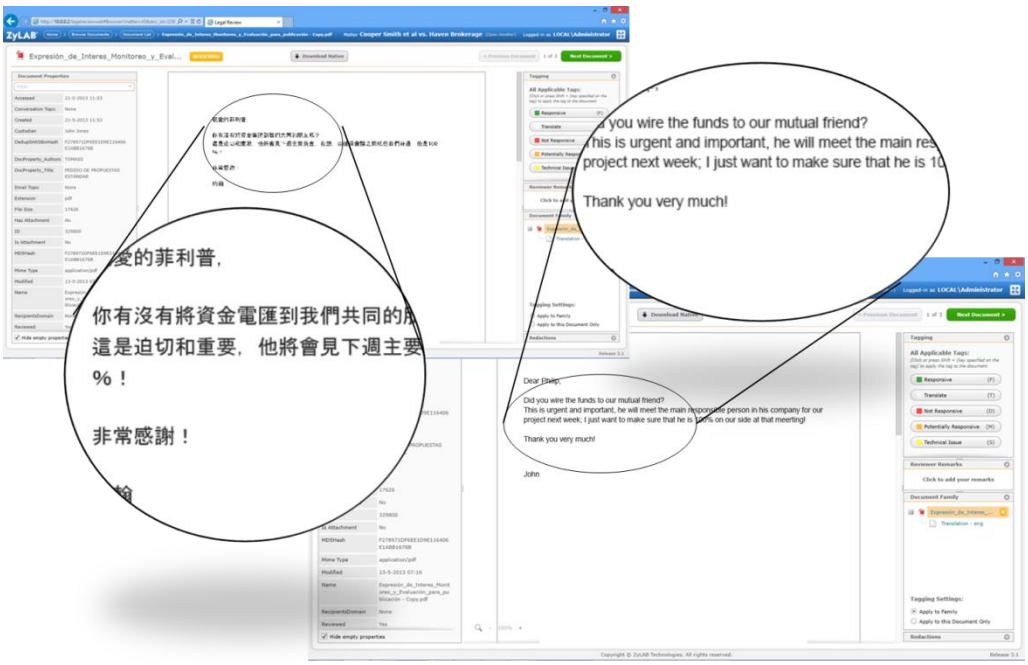
Next, the language needs to be recognized and the files need to be tagged with the proper language identifications. For electronic files that contain text that is derived from

an Optical Character Recognition (OCR) process or for data that needs to be OCRed, this process can be complex.

Text-analysis applications use more advanced linguistic analysis and often depend heavily on specific language characteristics and statistics to deal with complex information extraction, handling of co-references or anaphora and negations. Without proper knowledge of the underlying language, these techniques will not work properly, so it is important that the technology you use can deal with multi-lingual documents and multi-lingual document collections.

Machine Translation

Google's new algorithms are based on Statistical Machine Translation (SMT) methods in combination with translation memory. This method differs very much from the more traditional grammar-based methods. For many years, the field of computational linguistics consisted on the one hand of research based on Chomsky's theories of generative grammars and on the other hand more statistical approached. Because of the complexity, non-robustness and slow processing of the grammatical approach, statistical approaches are favored more and more by the research community.



Over the years, the statistical and grammatical methods have more or less merged, where the better working approach is now based on statistical algorithms in combination with large corpora of natural language which is tagged with (simple) linguistic properties

and linguistic structures. Linguistic probabilities are automatically calculated from large collections of data.

Statistical Machine Translation works on the same principle: from a large collection sentence pairs in the source and target language, a SMT algorithm can derive the most probable translation for a particular sentence, phrase or word in a specific context. This approach really leads the evolution of this effort with innovative technology that overcomes many of the problems of traditional automated translation. While the translations may not legally admissible in court, they do provide great insights in the content of large document and e-mail collections.

Now, why are SMT suddenly so good? There are two major reasons for this:

(i) After 9-11, the US intelligence forces were in great need of translations for languages such as Urdu, Pashtu, and Farsi. There were not enough screened translators and it was impossible to teach enough existing and newly (screened) employees to translate all the available data. Machine Translation was the only option. The problem with existing (grammar-based) translation tools was that training the system for a specific domain required the vendor to be involved. This was of course a problem because of the highly confidential nature of the data. And last but not least, understanding the training process required deep knowledge of computational linguistics, another hard to find talent. Statistical Machine Translation can automatically learn from sets of examples, a process that can be done in-house. Also, the SMT was able to process the often corrupted data much more robust than the traditional approaches. So, basically, SMT did a better job on all requirements and US Intelligence agencies invested heavily in this new technology, making it even better.

(ii) Due to the availability of large translation databases from for instance the United Nations and the European Union, training the SMT algorithms is much easier than it was in the past. Finally, Moore's law, resulting in twice the amount of data every 18 months, is in our advantage!

There is one golden rule in all statistical linguistic algorithms: THE ONLY GOOD DATA IS MORE DATA. And for that reason, I expect these algorithms only to become better and better, because the one thing that we can sure about, and that is that we will end of with even MUCH more data in a few years from now.

Here is what the advantages of Statistical Machine Translation are in more detail:

- **Volume:** SMT technology has the unique ability to handle a high volume of translations – quickly. It is an ideal solution for companies that have continuous publishing/translation cycles, large volumes of digital content, and growing demands to distribute more multilingual information.
- **Speed:** SMT delivers the highest throughput commercially available for statistically-based, automated translation solutions and unprecedented speeds for translating digital content. Additionally, the speed at which a company can get up-and-running with a SMT solution is significantly faster than other options – from evaluation, to integration, to deployment.
- **Accuracy & Training:** SMT offers you the ability to train translation systems to a specific domain or subject area to radically increase translation accuracy in-house. This process utilizes existing translated content to teach the software the terminology and style of the requested domain. This is especially interesting for intelligence and security organizations dealing with very confidential data. There is no need for you to disclose your data to a 3rd party.
- **Robustness:** There is no other approach that can process incomplete, misspelled, inconsistent and otherwise wrong data better than Statistical Machine Translations. Where other approaches fail dramatically, SMT can easily work with such data and still produce meaningful results.
- The applications of Machine Translation are endless: next to the obvious ones in intelligence, law enforcement and law enforcement, there are many other applications in the fields of eDiscovery, compliance, information governance, auditing, and of course knowledge management.

These translation solutions accelerate the way the world communicates by “unlocking” large volumes of digital content that would not be translated without automation.

Multi-Media Search

Today’s Enterprise Big Data contains large volumes of electronically stored information (ESI) that is non-textual such as images, video and audio. Processing, searching and classifying these types of information without textual information add significant cost and risks. ZyLAB’s technology allows internal investigators to search natively in audio files

and in the audio component of for instance a video, in addition, ZyLAB introduced Visual Classification technology to automatically categorize images based on their content.

State-of-the-Art in Audio and Speech Search

For the Enron case, nearly a dozen FBI analysts spent 3 months transcribing 2,800 hours of audio so they could search for key phrases in the transcript. With the ZyLAB Audio Search Bundle available today, they could perform those same searches directly on the audio files – not a speech-to-text transcript – within about 5 minutes and instantly replay the segments to verify their relevance.

Written text, such as transcripts from audio recordings, cannot fully convey intent, nuance or emotion which is only discernible by human listeners. Additionally, speech-to-text technology is generally limited to dictionary entries. In contrast, the ZyLAB Audio Search Bundle transforms audio recordings into a phonetic representation of the way in which words are pronounced so that internal investigators can search for dictionary terms, but also proper names, company names, or brands without the need to "re-ingest" the data.

There are three approaches to audio discovery capability, each with unique features:

1. The Human Listening Approach

Until recently the only practical solution to locating information within audio data was human listening. Skilled people do have some significant advantages in terms of interpretation and judgment, especially for picking up on subtle nuances and inflections. However, there are obvious and major weaknesses:

- A person can only listen to one call at a time.
- People have a limited attention span and fallible memory, affecting the volume of data that can be recalled and the number of terms that can be searched for.
- Finally, even the most skilled analysts are only able to operate at just faster than real-time, they can miss critical items and can't work 24/7.

2. Speech to Text Technology

"Speech to Text" technology, also known as Large Vocabulary Continuous Speech Recognition (LVCSR), converts the speech content of audio into text, processing it using a large vocabulary dictionary. This process requires a sophisticated language model for

good recognition, resulting in heavy processing needs. If a certain word or name is not included in the dictionary it will never be found.

Speech to Text technology can search an audio transcript many times faster than real time; typically 2-4GB of text may be searched in 0.1 seconds. However, converting the audio to that searchable text is processor hungry and is frequently achieved at only 2–3x faster than real time. When ad hoc searching needs to be applied, a large dictionary is required to carry out the recognition. This further limits the volume (or speed) of data that can be practically processed. On the positive side some Speech to Text systems can produce a transcript that might be simply read as text.

For all these reasons, we have all been disappointed by the results of speech-to-text technology in the last decades. Because of the two-step conversion process with two phases of quality loss (first sound to phoneme and second phoneme to text), the problems will probably not be solved easily. So, what is the solution?

3. Pure Phonetic Search Technology

Phonetic search seems to provide the best solution to search large collections of audio and video files. Especially when you have a lot of multi-speaker telephony data. A system incorporating Phonetic search technology transforms audio recordings into a phonetic representation, rather than written words. Next, user queries are also converted into phoneme sequences and are then matched by using fuzzy technology to the recognized sound recordings.

It includes a model for the way in which words are pronounced and is therefore not limited to only searching for words in a dictionary. This means that searches for personal or company names or brands can be successfully conducted – without the need to “re-ingest” the data.

Phonetic search uses fuzzy matching of the recognized phonemes and the phonemes of the words one is looking for. A user definable threshold specifies where to cut off the search results.

Fuzzy matching: have we not seen this before?

Many years ago, all vendors tried to develop the perfect Optical Character Recognition (OCR) engine. However, the random character of the errors and the often low scan quality resulted in a more pragmatic approach where instead of aiming for the impossible 100% correct OCR quality, fuzzy searches solved the mismatch between misrecognized

OCR text and search queries. This, in combination with display of the retrieved keywords in the original images and hit highlighting resulted in the most powerful search solutions for scanned material (paper, faxes and other bitmaps).

In phonetic search, the exact same pragmatic approach is taken, instead of aiming for the impossible 100% correct speech-to-text, speech is converted into phonemes and end-user queries are also converted into phonemes, which are matched against the sound collections by implementing a fuzzy search between the phonemes in the query and the recognized phonemes. The result: a completely tunable search tool that can be used in eDiscovery, law enforcement and compliance applications where 100% recall is paramount.

Phonetic Search for your organization

Phonetic speech search engines are creating new opportunities for a wide range of organizations by delivering the ability to quickly, accurately and economically search large volumes of data to return relevant results. Phonetic search engines use a fraction of the hardware required by traditional solutions to deliver greater depth of audio search. It has the flexibility to use multiple search items, leading to greater accuracy and relevancy of results. And exact numbers of audio files relating to a specific topic can be easily determined, even across extremely large data-sets.

With the ZyLAB Audio Search Bundle, auditors, internal investigators and attorneys can identify and collect audio recordings from various sources with far greater efficiency and effectiveness than was ever possible with manual processing. The software supports multiple search techniques simultaneously, such as Boolean and wildcard, leading to greater accuracy and relevance of results. The fast, iterative search helps to reduce the size of the data set and the costs for review.

A search using phonetic recognition technology will run up to eighty thousand times fasteraster than real time. Using a single core of a typical Intel processor, eight hours of audio data can be searched in under a second. Preparation of the searchable content is conducted at rates up to 80x faster than real time.

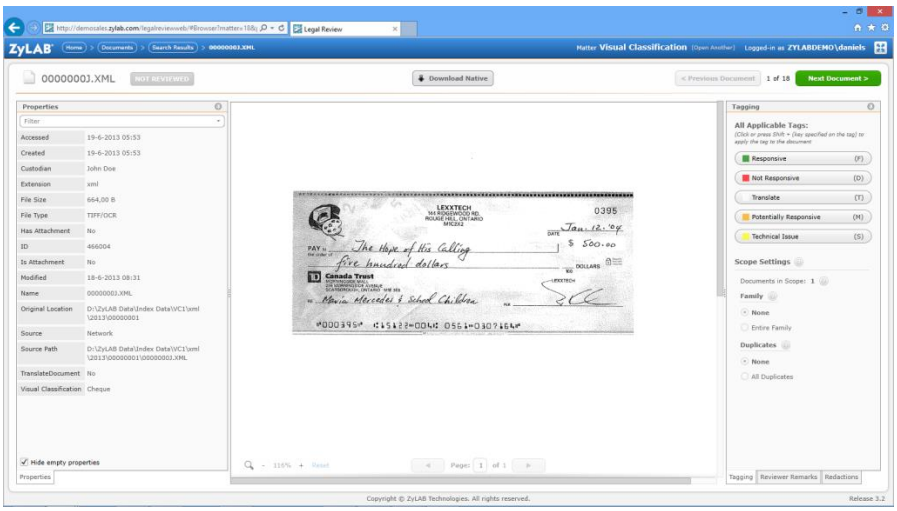
Visual Classification

ZyLAB's native visual search and categorization functionality automatically recognizes pictures and identifies among others people, babies, elderly people, flowers, cars, planes, in and outdoor scenes, and many other concepts. The new functionality is perfectly

usable for the identification of images of personal identifiable information (PII), potential intellectual property, handwritten notes, check's, ID's, and other information that otherwise cannot be recognized automatically.

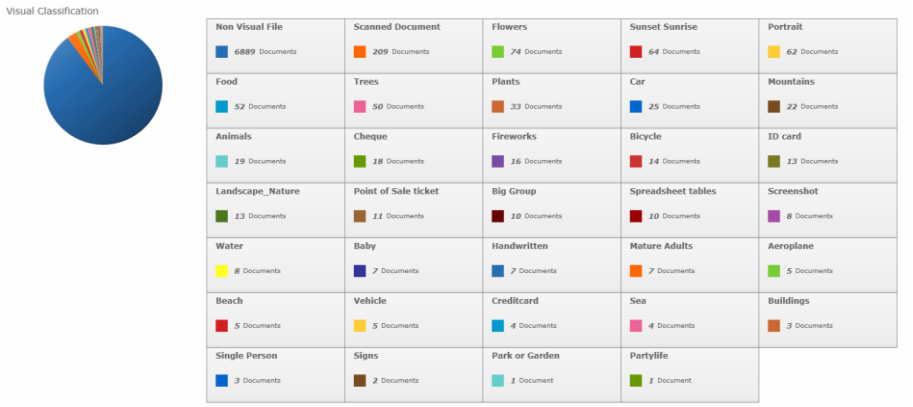
Identifying ID's

If such an image did not contain textual information, then automatic Optical Character Recognition (OCR) or any other text-based search and technology assisted review tools were useless and an expensive and long manual review process was the only alternative. The native visual search and categorization technology that we now offer as a new add-on to the ZyLAB eDiscovery and Production Platform, is traditionally used in law enforcement and intelligence applications such as video surveillance and picture classification. The cost and time savings for reviewers are paramount. This technology enables reviewers to exclude all non-relevant images beforehand, find all images that contain handwritten notes instantly for further investigation, and quickly and effectively locate potentially confidential information such as copies of credit cards and ID's.



Identifying Receipts and Checks

ZyLAB's Visual Classification can detect a standard set of "concepts". These concepts are matched to the visual content of an image and for each concept a confidentiality score is calculated. Based on these scores the images can be tagged. Available concepts are among others animals, beach, big group, buildings, cars, city life, crowd, daytime outdoor, drawing, female, fireworks, flower, pornographic / child abuse, guns, indoor, Manga Comic, mountains, nighttime, one or more persons, outdoor, overlaid text, park garden, party life, persons, plants, scanned checks, credit cards, handwritten documents, ID cards, letters, magazine or newspapers, point of sales tickets, spreadsheets, screenshots, sea, small groups and vehicles.



Various Categories of Visual Classification

Customer Case: Security Breach

Our customer is a leading bank from Europe and one of the world's largest banking/financial services and insurance multinational by revenue. Its primary businesses are retail banking, direct banking, commercial banking, investment banking, asset management and insurance services. The bank is serving over 85 million individual and institutional clients in more than 45 countries.

The Business Case

The consequences of a security breach can be very serious and not only include the cost of resolving the breach and potential lost revenues but also expose the bank to potential lawsuits. So when the customer suspected an ex-employee to have leaked confidential information to future business relations for his own personal benefit, a full scale investigation was started.

The scale of this investigation that included multiple email boxes of multiple custodians and regions, exceeded the capacities of the internal investigation team, which normally handled cases involving maximal 10 GB of data. ZyLAB was called in to support the investigators with rapidly analyzing the more than 300GB of collected information for this case.

ZyLAB's Solution

In close cooperation with the internal investigation team of the bank, ZyLAB used advanced filter techniques to establish the details and magnitude of the data leakage.

After a first filtering based on the external features of the collected files, drastically diminishing the amount of data, more advanced filtering based on keywords, risky data combinations and sentiments, reduced the volume to an easy and quick to review 3GB. This way ZyLAB provided the investigators full control over the data and the speeded the completion of the breach investigation.

Customer Case: The FBI Enron Case

Fast, Accurate Processing of Evidence for the FBI, Houston Texas

- eDiscovery processing for the most dramatic corporate collapses in history.
- Required a system to archive, search, find, organize, share and produce any Enron document or electronic file within seconds.
- Involved more than 300 GB of data including 2.5 million pages (7,000 documents), 100 GB of PST email, and 100 GB of electronic files.
- Decentralized, full-text exploratory, legal searching.
- The first trial in FBI history in which both the Prosecution and the Defense worked with ZyLAB technology and the same electronic data.

ZyLAB delivered immediate results and value

ZyLAB provides world-class eDiscovery software and services that are directly aligned with the Electronic Discovery Reference Model. This case required only a small portion of our end-to-end solution.

PROCESSING

- ZyLAB scanned 2 million pages of documents, converted them to searchable OCR TIFF files, and added key fields. To meet the quick turnaround required, ZyLAB balanced the load across 10 OCR workstations working in parallel. ZyLAB then indexed all of the electronic evidence to make it full text searchable.

LEGAL REVIEW

- ZyLAB provided remote access to the central data storage hub at the FBI field office in Houston. The teams from Washington D.C., Houston, and San Francisco could login to search, review and analyze the evidence simultaneously.

PRODUCTION AND DISCLOSURE

- In addition to facilitating collaboration among the members of the Prosecution, the ZyLAB archive and eDiscovery software was accessible via web server to the Defense. This was central to their production and disclosure strategy.

Customer Case: OLAF

The European Anti-Fraud Office (OLAF) was created by Commission Decision 1999/352/CE. OLAF is charged with protecting the financial and other interests of the Community against fraud and irregular conduct in respect to both the income and expenditure of the Communities, wherever this may have occurred.

The mission of OLAF is to protect the financial interests of the EU, to fight fraud, corruption and any other irregular activity affecting the EU budget, including misconduct within the European Institutions.

OLAF achieves its mission by conducting, in full independence, internal and external investigations. It also organizes close and regular co-operation between the competent authorities of the Member States in order to co-ordinate their activities. OLAF supplies Member States with the necessary support and technical know-how to help them in their anti-fraud activities. It contributes to the design of the anti-fraud strategy of the European Union and takes the necessary initiatives to strengthen the relevant legislation.

Business Drivers

OLAF acts on allegations of fraud, corruption and any other irregular activity affecting the EU budget. The initial facts are carefully verified and assessed. If there is a serious suspicion that the alleged wrongdoing took place after this initial process, an investigation is opened. OLAF uses ZyLAB for large-scale investigations that involve enormous information collections. With ZyLAB OLAF can find entities and facts without knowing they existed before the start of an investigation.

An important issue for law enforcement agencies to consider is the internal workflow of all data and, of course, data protection regulations. OLAF addresses these issues and applies ZyLAB' search and text-mining technology to solve cases that would otherwise not have been solved. As a result, internal and external fraud is prosecuted and valuable EC resources are protected.

Special Benefits of ZyLAB to OLAF

- Entity extraction allows OLAF to extract company and individual names from large data collections within 24 hours, allowing the investigators and analysts to kick-off additional seizures and investigations immediately. This is very important in the beginning of an investigation to prevent additional suspects from fleeing or destroying evidence.
- OLAF can easily find hidden patterns and code words by using ZyLAB's Content Analytics and Text Mining: for instance "people/company pays People/Company", "people meet people", etc. No longer needed to maintain large queries of names and companies.
- Find hidden transactions and bank accounts: no need to search for (known or unknown account number), just find all wire transfer patterns and export account numbers to XLS for verification.
- Criminals use several ways to hide fraudulent information: encrypted ZIP, non-searchable PDF. TIFF or other bitmaps, deeply embedded objects in email attachments, etc. In several cases, the fact that ZyLAB could identify this information, resulted in finding essential evidence.
- Data protection and privacy regulations compliance
- MOREQ-2 Records Management compliant.
- Easy to integrate with other OLAF systems and tools (ENCASE, FTK, Analyst Notebook, Oracle Case System, Documentum, etc.).
- Works on highly secure encrypted network.
- Various tax authorities in the EU also use ZyLAB. Easy to exchange data and results from investigations.
- OLAF can create extracts from the CASE FILES and share them on (encrypted) CD and DVD to prosecutors.

Quote:

"OLAF, the antifraud unit of the EC, confiscated 1 Tb of email data in 2009. Without ZyLAB's professional text mining tools, it would have been impossible to analyze this data in-depth and on-time. ZyLAB uncovered digital information and code names that would otherwise have been hidden and the suspects would have walked away. Because of ZyLAB, suspects could have been identified and prosecuted." Eric Yperman, Team Leader Operational Intelligence

Customer Case: UN War Crimes Tribunals

The UN War Crime Tribunals and International Court of Justice required ZyLAB to use XML as a file format in order to make sure they would have access to all data files in the future for long term prosecutions, litigations and historical record. The UN archived e-mail, but also evidence, court records and almost all administrative materials.

In the independent update "Enabling Prosecution of the Unspeakable" industry analyst IDC highlights the following best practices that global operations should consider when looking for archiving for eDiscovery:

- Prepare for multimedia operations.
- Understand the demands of recall in search.
- Embrace the challenge of scale.
- Expand to true Multilanguage processing.
- Build for repeatability.

In the update IDC further concludes that "The UN Information Management team supporting the war crimes tribunals was able to bring into place a complex system with many innovative features in search processing in a short time frame through establishing a strong partnership with its selected vendor, ZyLAB, and through taking advantage of highly capable processing capabilities within the software to reduce or eliminate the bottlenecks and delays inevitably associated with manual discovery operations. As a result, the prosecutor within the ICTY tribunal was able to carry out trials with evidence that had been transformed into digital formats and analyzed and produced with full accessibility and consistency for prosecutors, defense attorneys, and judges. This is a significant accomplishment for a steady-state commercial enterprise. The UN work is all the more impressive, growing as it did out of the horror of the prosecution of war crimes and the need of the international community to respond quickly and fairly in executing justice."[5]

[5] IDC UPDATE "ZyLAB: Enabling Prosecution of the Unspeakable" by Hadley Reynolds, June 2011. The full update is available on request.

About the Author

Dr. Johannes C. Scholtes is Chairman and Chief Strategy Officer of ZyLAB. From 1987 to 2009 he acted as President / CEO of ZyLAB. Scholtes has been involved in deploying in-house eDiscovery software with organization such as the UN War Crimes Tribunals, the FBI-ENRON investigations, the EOP (White House), and thousands of other users worldwide.

Scholtes holds the extraordinary Chair in Text Mining from the Department of Knowledge Engineering at the University of Maastricht and he is a senior research fellow of the Dutch School for Information and Knowledge Systems (SIKS).

Before joining ZyLAB in 1989, Scholtes was lieutenant in the intelligence department of the Royal Dutch Navy. Scholtes holds a M.Sc. degree in Computer Science from Delft University of Technology and a Ph.D. in Computational Linguistics from the University of Amsterdam.

About ZyLAB

ZyLAB's industry-leading, modular eDiscovery and enterprise Information Management (IM) solutions enable organizations to manage boundless amounts of enterprise data in any format and language, to mitigate risk, reduce costs, investigate matters and elicit business productivity and intelligence.

For 30 years ZyLAB has been a dominant player in compliance and eDiscovery related solutions, due in part to its' advanced capabilities for multi-language support, searching, content analytics, document reviewing, and e-mail and records management (for both scanned and electronic documents).

While the ZyLAB eDiscovery & Production system is generally implemented to investigate a specific legal matter, it is a solid and robust foundation from which to pursue proactive, enterprise-wide objectives for information management. Those broader goals are achieved through the use of the ZyLAB Compliance & Litigation Readiness system.

The ZyLAB e-Discovery system is directly aligned with the Electronic Discovery Reference Model (EDRM) and features modules for forensic sound collection, culling, advanced e-mail conversion (Exchange and Lotus Notes) and legal review.

The company's products and services are used on an enterprise level by corporations, government agencies, courts, and law firms, as well as on specific projects for legal services, auditing, and accounting providers. ZyLAB systems are also available in a Software-as-a-Services (SaaS) model.

ZyLAB's products are open and scalable, with installations managing some of the largest collections of mission-critical data in the world. The award-winning ZyLAB Information Management Platform brings our core capabilities into a single solution that provides an optimal framework for six, specialized, all-in-one system deployments.

Currently the company has sold 1.7 million user licenses through more than 9,000 installations. All of our solutions include full installation, project management and integration services. Current customers include The White House, Amtrak and US Army OIGs, US Department of Treasury, The EPA, National Agriculture Library, Royal Library of the Netherlands, FBI, Arkansas and Ohio state police forces, German customs police, Danish national police, War Crimes Tribunals for Rwanda, Cambodia, and the former

Yugoslavia, KPMG, PricewaterhouseCoopers, Deloitte, Akzo Nobel, Sara Lee, Pacific Life, Siemens, Dow Automotive and Lloyds of London.

ZyLAB is positioned by Gartner, Inc. as one of the strongest “Visionaries” in the 2013 Magic Quadrant for eDiscovery Software and has received numerous other industry accolades over the last 3 decades.

ZyLAB is certified and registered as compliant with the International Standards Organization (ISO) 9001:2000. ZyLAB also lets Microsoft, Oracle and other infrastructure providers regularly certify critical components that work closely with their infrastructure. ZyLAB was certified under the US-DoD 5015.2 records management standard and ZyLAB is compliant with the European MoReq2 standard and various other regulations

Headquartered in Amsterdam, the Netherlands and McLean, Virginia, ZyLAB also serves local markets from regional offices in New York, Barcelona, Frankfurt, London, Paris, and Singapore. To learn more about ZyLAB visit www.zylab.com

ZyLAB®

eDiscovery & Information Risk Management

www.zylab.com