AUSTRIAN LAW JOURNAL

ISSN: 2409-6911

DOI: 10.25364/1.3:2016.1.5
www.austrian-law-journal.at

# Semantische Analyse unstrukturierter Daten

## Review und Analyse: Big-Data Ansatz bei internen Untersuchungen anhand eines Beispiels

**Jörg Fuchslueger***, Wien

***Kurztext**: In der juristischen Fallbearbeitung steht man immer häufiger vor der Herausforderung, unüberschaubare Massen an Unterlagen berücksichtigen zu müssen. Relevante Information ist verborgen in hunderttausenden E-Mails, Office-Dokumenten und PDF-Dateien. All diese Unterlagen zu lesen, zu verstehen und zu beurteilen, ist in der Praxis nicht mehr möglich. Der vorliegende Artikel zeigt ein Vorgehensmodell, wie mit semantischer Inhaltsanalyse die notwendige Informationsgewinnung in umfangreichen, elektronisch gespeicherten Textbeständen durchgeführt werden kann.*

***Schlagworte:** E-Discovery; interne Untersuchungen (Internal Investigations); Big Data; Compliance; semantische Analyse; juristische Fallbearbeitung.*

## I. Einleitung

Im Zuge unternehmensinterner Sachverhaltsaufklärungen (*Internal Investigations*) sind immer häufiger unüberschaubare, große Datenbestände zu überprüfen. Damit verbunden sind nicht nur kritische Zeit- und Kostenentwicklungen. Untersuchungsteams stoßen an die Grenze des Machbaren.

Selbst bei Einsatz elektronischer Review-Plattformen[1] und computerunterstützten Methoden wie Predictive Coding[2] sind bisherige Vorgehensmodelle bei den heutigen Datenmengen nicht mehr zielführend. Der Trend zeigt auf immer schneller und noch weiter wachsende Datenvolumen. Wissensbasierte Ansätze, Cognitive Computing[3] und heute verwertbare Teilbereiche in der Entwicklung natürlich-sprachlicher Informationssysteme[4] gewinnen an Bedeutung, um sichergestellte, elektronisch vorhandene Informationsquellen für die Aufklärungsarbeit weiterhin sinnvoll nutzen

* *Jörg Fuchslueger* ist Head of Content Analytics der BIConcepts IT Consulting GmbH. Der Beitrag gibt den Inhalt des Vortrags vom 11. 12. 2015 am 7. Thementag „Recht und IT" zum Thema „Compliance-Management. Standards – Tools – Haftung" am ReSoWi-Zentrum der Universität Graz wieder.

1 Durch Scan und Schrifterkennung (OCR) werden auch „Papierdokumente" zu ESI (Electronically Stored Information) und können damit computerunterstützt durchsucht und abgearbeitet werden.

2 Maschinelle Klassifikation von Dokumenten auf Basis manuell erstellter Lernmengen.

3 Ein von IBM geprägter Begriff; dahinter steht ein Konzept, das eine wachsende Anzahl konkreter Services im Umfeld „künstlicher Intelligenz", wie zB die Umwandlung von Sprachinformation in Text (Voice to Text), Erkennung von Objekten in Bildmaterial uvam, als Webservices oder bereits integriert in einer fertigen Entwicklungsumgebung zur Verfügung stellt. http://www.ibm.com/cloud-computing/bluemix/watson/ (abgefragt am 31. 3 2016).

4 Natural Language Processing (NLP) gilt als Teilbereich der künstlichen Intelligenzforschung, um den Dialog Mensch–Maschine auf Sprachebene zu ermöglichen.

**Fundstelle:** *Fuchslueger*, Semantische Analyse unstrukturierter Daten, ALJ 1/2016, 68–77 (http://alj.uni-graz.at/index.php/alj/article/view/61).

zu können. Die zunehmenden Datenmengen erfordern zusätzlich einen Umdenkprozess – klassisch lesen, selbst nach aufwendigen Verfahren der Datenreduktion, funktioniert nicht mehr.

## II. Grundlagen

Ehe die semantische Analyse unstrukturierter Daten am konkreten Beispiel dargestellt wird, sollen die einschlägigen Grundbegriffe geklärt werden, um so das Verständnis sowohl der Herausforderung als auch der Lösung durch BIConcepts zu ermöglichen.

### A. Strukturierte versus unstrukturierte Daten

Strukturierte Daten sind zumeist in Datenbanken gespeichert und werden mit genormten Abfragesprachen recherchiert und analysiert. Um auch Nicht-Datenbank-Spezialisten und -Anwendern einen benutzerfreundlichen Zugriff zu ermöglichen, wird mit semantischen Modellen gearbeitet. Semantische Modelle ermöglichen – ohne Detailkenntnis der Datenstruktur – auch Datenbank-Laien, selbstständig Abfragen durchzuführen und Berichte und Auswertungen zu erstellen. Data-mining-Modelle berechnen bei Bedarf Muster und daraus wiederum Abhängigkeiten oder Auffälligkeiten. In der strukturierten Datenwelt haben wir gelernt, mit entsprechenden Abfragen und Auswertungen Antworten auf unsere Fragen zu bekommen. Selbst bei umfangreichen Datenbanken oder sehr komplexen Verknüpfungen zwischen Tabellen ist bei vielen Unternehmen Technologie und Know-how vorhanden, um selbstständig Recherchen durchzuführen. *Ad-hoc*-Abfragen und Analysen in strukturierten Daten sind etablierte Methoden der Informationsgewinnung. Neue Erkenntnisse entstehen nicht durch das Lesen einzelner Datensätze, sondern durch Interpretation von Berichten, Auswertungen, Statistiken, Diagrammen und Visualisierungen.

Kann dieser Ansatz auch bei unstrukturierten Daten funktionieren? Können Dokumenttypen aller Art, wie Millionen von E-Mails samt deren Anhängen, zumeist PDF, Office-Dokumente oder, fachlich und inhaltlich gesprochen, Verträge, Rechnungen, Angebote, Präsentationen oder Berichte, ähnlich einer Datenbank inhaltlich ausgewertet werden? Wir sind davon überzeugt und unsere Antwort lautet: Ja. Über semantische Analysemodelle können unstrukturierte Daten entsprechend strukturiert werden. Damit lässt sich auch diese Art von unstrukturierter Datenvielfalt selbst bei enormen Datenmengen in angemessener Zeit abfragen und inhaltlich auswerten. Ohne jedes einzelne Dokument lesen zu müssen, werden Auffälligkeiten automatisiert erkannt, Zusammenhänge visualisiert und umgehend verwertbare Ergebnisse gewonnen. Mit diesem praktikablen Ansatz kann der Datenexplosion bei internen Untersuchungen begegnet werden. Die technische Grundlage natürlich-sprachlicher Informationsverarbeitung liefert dabei die IBM Watson Explorer Analytical Components.

### B. IBM Watson Explorer Analytical Components

BIConcepts ist ein Software Entwicklungs- und IT-Dienstleistungsunternehmen in Wien. BI steht für Business Intelligence, das Unternehmen beschäftigt sich mit der Informationsaufbereitung aus großen Datenmengen. In diesem Technologieumfeld ist BIConcepts langjähriger IBM Premium Business Partner. IBM investiert laufend in den Bereich Analytik und seit einigen Jahren auch massiv in den Bereich Cognitive Computing. Basierend auf spezieller Software, die ua auch „Sprachverständnis“ (Natural Language Processing, NLP) einsetzt, werden diverse innovative

Lösungen entwickelt. IBM bündelt diese Entwicklungen in der „IBM Watson Group". Der Namensgeber für das Computersystem, das 2011 gegen die besten Spieler in der Quizshow „Jeopardy!" (ähnlich der „Millionenshow") antrat, war *Thomas J. Watson*, einer der Mitbegründer von IBM. IBM Watson hat in dieser Quizshow gegen die beiden Rekordsieger gewonnen und damit eindrucksvoll bewiesen, dass ein Computersystem natürlich-sprachliche Fragen automatisiert erfassen und richtig beantworten kann.

IBM Watson Explorer Analytical Components steht heute als Komponente einer Standardsoftware zur semantischen Suche und Analyse von umfangreichen Textdatenbeständen zur Verfügung. Mit diesem Verfahren werden aus natürlich-sprachlicher Information Erkenntnisse gewonnen, ohne Texte Satz für Satz lesen zu müssen. Auf Basis dieser Technologie hat BIConcepts speziell für den Anwendungsfall interner Untersuchungen gemeinsam mit Fachexperten wie Forensikern, Juristen, Gutachtern und Gerichts- und Wirtschaftssachverständigen ein System entwickelt, das umfangreiche Untersuchungsdaten vorab semantisch analysiert und in einem „intelligenten" Datenraum dem Untersuchungsteam für die Recherche performant und sicher zur Verfügung stellt: Intelligent Content Investigation (ICI).

## III. Intelligent Content Investigation (ICI)

Es gibt einige Kriterien, die ein Big Data-Recherchetool erfüllen muss, um direkt von Juristen bzw entsprechenden Experten eines Untersuchungsteams als gewinnbringende Unterstützung ihrer Arbeit akzeptiert zu werden. Unter „Big Data" verstehen wir nicht nur die sehr großen Datenmengen (Volume), sondern auch die Herausforderung durch die Vielfalt (Variety) der Informationsobjekte wie E-Mails, Präsentationen, Verträge, Angebote, Chatprotokolle, allgemeiner Schriftverkehr uvm sowie deren unterschiedlichste Quellen wie zB Server, Laptops, Smartphones, Mobile- oder Cloudspeicher.[5] Big Data-Analyse bedeutet, all diese Daten für den Endanwender in einer einheitlichen Oberfläche und in einem einheitlichen Format in einem zentralen System für die Recherche bereitzustellen.

### A. ICI als Schnittstelle

ICI versteht sich als Schnittstelle zwischen den Nutzern und dem auszuwertenden Datenmaterial. In der Bedienung darf es für den Anwender keine IT-bedingten Hürden geben, wie etwa unbekannte Fileformate, unterschiedlichste E-Mail-, Betriebs- und Datenträgersysteme, inkompatible Zeichencodes oder komplexe Metadaten. Suchabfragen und Auswertungen müssen trotz enormer Datenmenge rasch und unmittelbar erfolgen. Nur dadurch wird die Methode der *Ad-hoc*-Recherche auch akzeptiert. Diese erfolgt über eine einfach zu bedienende Weboberfläche mit einer sicheren Verbindung zur Big Data-Rechercheplattform. Ein ICI-Server kann bspw „Inhouse" betrieben oder in einem externen Hochsicherheitsrechenzentrum für den Zeitraum der Untersuchung angemietet werden.

### B. Benutzerfreundlichkeit

Einfachheit in der Bedienung ist für die Akzeptanz jeder Applikation wesentlich. Bei einer Anwendungssoftware für interne Untersuchungen, mit der in der Regel nicht tagtäglich gearbeitet wird,

5 Vgl *Brücher*, Rethink Big Data (2013) 41 ff.

ist das umso wichtiger. Benutzerfreundlichkeit wird dadurch erzielt, dass der Aufbau und das Verhalten der Oberfläche in den Grundprinzipien jenen Anwendungen entsprechen, die uns vertraut sind. ICI funktioniert wie eine Webseite mit übersichtlichen Navigationselementen und einem klaren Aufbau in der Ergebnisdarstellung. Alle Möglichkeiten der Facettensuche bzw der linguistischen Suche, der damit verbundenen Suchsyntax, der Interpretation der unterschiedlichen Auswertungen und dem Review einzelner Dokumente können innerhalb eines halben Tages geschult werden.

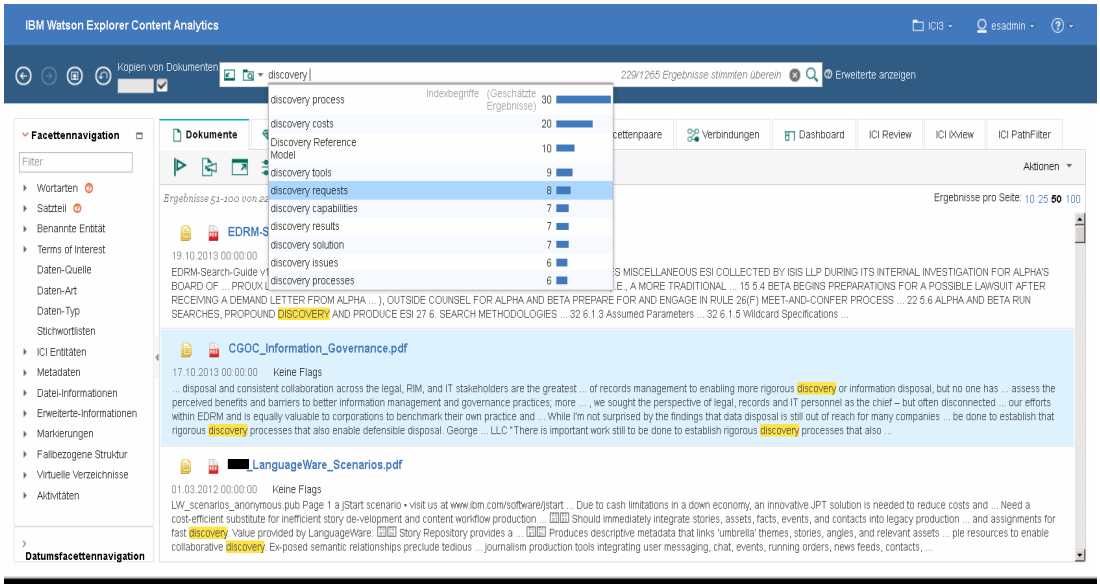


**Abb 1:** ICI-Weboberfläche mit einfacher Navigation und übersichtlicher Ergebnisdarstellung.

## C. Struktur und Übersicht anhand eines Beispiels

Als Erklärungsmodell der Funktionsweise von ICI soll ein frei erfundener und sehr vereinfachter „klassischer" Compliance-Fall dienen: Im Zuge von IT-Ausschreibungen soll es bei der Angebotseinholung von Subleistungen für bestimmte Technologiekomponenten durch einen oder mehrere Mitarbeiter zur Bevorzugung eines speziellen Anbieters gekommen sein. Daher soll untersucht werden, ob es bei der Lieferantenauswahl für Subdienstleistungen in einer bestimmten Abteilung des Unternehmens „Unregelmäßigkeiten" gab.

Angenommen, dem Untersuchungsteam stehen die E-Mail-Postfächer des Hauptverdächtigen und von fünf weiteren involvierten Personen, deren Laptops sowie weitere projektbezogene Verzeichnisablagen von drei Fileservern als Informationsquellen zur Verfügung. Nach dem Prozessieren dieser Daten ist es nicht unrealistisch, dass in Summe über 500.000 Textdokumente mit weit über zwei Millionen Seiten vorliegen.

### 1. Aufbereitung der Datenmenge

Zunächst wird das Datenmaterial in drei Schritten aufbereitet:

#### a. Indexierung und Strukturierung (Crawling)

ICI indexiert und strukturiert in einem ersten Schritt automatisch alle Dokumente auf Basis eines Standardanalysemodells (Big Data-Ansatz). Das Programm öffnet dabei jede einzelne Datei und

extrahiert daraus die Metadaten und den Text. Dieser Vorgang wird als „Crawling“ bezeichnet. Der sehr rechenintensive Vorgang kann in Abhängigkeit der Hardware und der Datenmenge einige Stunden in Anspruch nehmen.

**b. Tokenization und syntaktische Aufbereitung**

Die sprachliche Analyse des Textes wird im nächsten Schritt mit einem sog „Parser“ durchgeführt. Da die weitere semantische Analyse eines Textes bereits von der zugewiesenen Sprache abhängig ist, wird vom Parser auch die Sprache festgestellt. Bei der „Part of Speech“-Erkennung werden zunächst Seiten, Absätze und Sätze in sog „Token“ zerlegt (Tokenization) und dann in weiterer Folge morphologisch (Personalformen, Fallmarkierungen) und syntaktisch (Subjekt, Objekt, Modifikator, Artikel etc) analysiert. Dieser Vorgang führt ua auch jedes Wort auf seine lexikalische Grundform zurück. Der ICI-Standard beinhaltet die Sprachen Deutsch und Englisch. Derzeit können bei Bedarf noch 25 weitere Sprachen eingesetzt werden.

**c. Semantische Auswertung (Parsing Rules)**

Der dritte Schritt ist die semantische Analyse. Dabei kann auf alle vorangegangenen Analysen zurückgegriffen werden. Semantische Analysen werden durch Regeln, die sog „Parsing Rules“, implementiert. Dabei wird einzelnen Wörtern oder ganzen Satzteilen eine Bedeutung zugeordnet. Zum Verständnis sollen einige semantische Analysen aus dem ICI-Standardmodell vorgestellt werden.

Personen, Firmen, Orte, Geldbeträge, Kontoverbindungen, Betrags- und Datumsangaben innerhalb des Textes zu annotieren, ist Teil des ICI-Standardanalysemodells. Parsing Rules werden von Fachexperten mit spezieller Software, wie bspw dem IBM Watson Studio, entwickelt und getestet. IBM Watson Studio ermöglicht anhand von relevanten Beispieltexten neue Modelle direkt zu entwickeln und die Ergebnisse der einzelnen Regeln sofort zu evaluieren. Unterschiedlichste Wörterbücher oder Wissensmodelle können eingebunden werden und werden in der Regeldefinition verwendet. Eine Parsing Rule erkennt zB Firmenbezeichnungen und annotiert diesen Satzteil mit „Firma“. Die Ergebnisse der Parsing Rule können normalisiert im Index als Facette „Firmen“ abgelegt werden, es entsteht also vereinfacht eine Liste aller gefundenen Firmen aus den gesamten Daten. Der Index beinhaltet natürlich auch die exakten Orte des Vorkommens jedes einzelnen Eintrages. Ein weiteres Beispiel sind Datumsangaben: Normalisiert bedeutet in diesem Fall, dass zB Textpassagen wie „5. Februar 2016“, „05/02/16“ oder „2016-02-05“ mit einer entsprechenden Parsing Rule in der Facette „Datum“ mit dem Wert „05.02.2016“ abgelegt werden. Zusätzlich wird diese Facette als Datumsfacette definiert und ermöglicht damit, einen bestimmten Zeitraum abzufragen. Auch diese Facette (Liste) liefert sofort, ob und welche Datumsinformationen in allen (kein Filter bzw keine Suchabfrage aktiv) bzw in den Inhalten der aktuell gefilterten Dokumente vorkommen. Die Möglichkeiten zur Definition von Parsing Rules sind vielfältig.

## 2. Facettenbildung

Die normalisierten Parsing-Ergebnisse werden als „Facetten“ im Index abgelegt. Wichtig für die Anwender ist dabei die Repräsentation dieser Ergebnisse in einer übersichtlichen, gut gegliederten Facettenstruktur und damit deren einfache Anwendung und Auswertbarkeit.

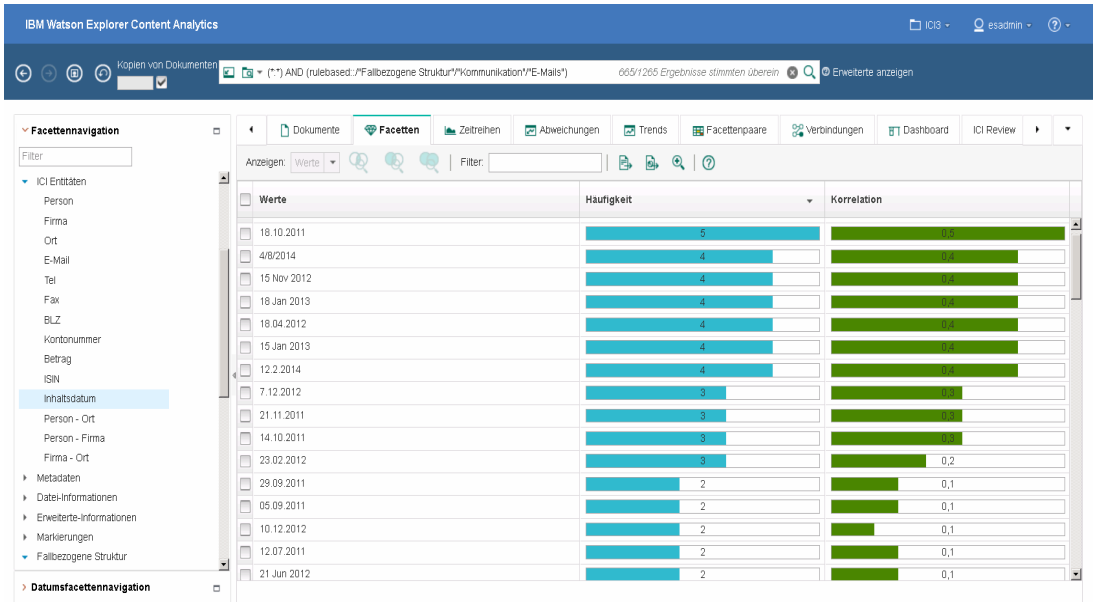


**Abb 2:** ICI-Facettenstruktur und -Anzeige (Beispiel: gefundene Datumsangaben innerhalb von E-Mails und deren Anhänge in einer für den Untersuchungsfall relevanten Sender-Empfänger-Gruppe). In gleicher Weise können für alle Facettenpunkte Listen und Auswertungen in Bezug auf jede beliebige *ad hoc* erstellte Teildatenmenge unmittelbar angezeigt, interpretiert und als weiterer Filter angewandt werden.

Zusätzlich zu den Facetteneinträgen aus der semantischen Analyse werden auch weitere Strukturpunkte durch vorhandene Metadaten der einzelnen Dokumente im Standardmodell automatisch erzeugt. Gute Beispiele dafür sind Sender, Empfänger und das Sendedatum eines E-Mails oder bei Dokumenten der Bearbeiter und das letzte Speicherdatum. Die Datumsfelder ermöglichen Suchabfragen mit exakten Datumseinschränkungen auf einer Zeitachse und unmittelbare Auswertungen mit Zeitreihen, Abweichungen und Trends. Das Analysemodell optimiert Datumsinformationen hinsichtlich verschiedener Datumsformate und Zeitzonen.

Auch die einzelnen Datenquellen (zB Laptop von Herrn A, Fileserver XY, Projektlaufwerk Angebote) oder auch Dokumentarten (zB E-Mail, Schriftdokumente, Tabellen, Präsentationen) werden in den Metadaten-Facetten sehr übersichtlich zur Analyse und Auswahl angeboten. Die Facettenstruktur hat den Vorteil, dass Metadaten nicht zu sehr „IT-spezifisch“, sondern von der Sprachbegrifflichkeit für den Endanwender möglichst verständlich und in deren spezifischem Sprachgebrauch zur Auswahl angeboten werden.

Die Standard-Facettenstruktur von ICI ist damit sehr praxisbezogen und anwenderorientiert. Sie liefert sofort ein erstes Bild über alle Inhalte. Das Untersuchungsteam ist damit in einer frühen Phase der Überprüfung bereits in der Lage, die „Verwertbarkeit“ der sonst unüberschaubaren Datenmenge zu beurteilen und das weitere Vorgehen zu planen.

## 3. Fallbezogenes Analysemodell

Zu Beginn der Untersuchung, eventuell auch nach einer ersten Sichtung der Daten, wird vom Untersuchungsteam zumeist ein Fragenkatalog erstellt. Bei umfangreichen Daten oder komplexen Sachverhalten ist eine Aufteilung in einzelne Untersuchungspunkte (also der Fragen/Themen, nicht

der Daten!) oft zweckmäßig und bildet eine gute Basis für die Arbeitsaufteilung innerhalb eines größeren Untersuchungsteams.

Auf Basis des Fragenkatalogs wird überprüft, ob die im Standardmodell vorhandene Facettenstruktur ausreicht oder ob es weitere fallspezifische Anforderungen für semantische Analysen gibt. Falls erforderlich werden diese als Teil des Servicepaketes von ICI zusätzlich fallbezogen implementiert. Ein praktisches Beispiel war die zusätzliche Annotation aller ISIN (Identifikation für ein Wertpapier) im Fall eines Kapitalanlagebetruges.

Bei der Erstellung eines fallbezogenen Analysemodells können, wie auch oft bei manuellen Reviews, Stichwortsuchen Verwendung finden. Die semantische Umsetzung mit ICI liefert dabei im Vergleich zu Volltextsuchen messbar[6] höherwertige Ergebnisse. Ein weiterer wesentlicher Vorteil bei der Umsetzung von Stichwortsuchen mit ICI gegenüber Systemen mit Volltextsuche ist die Möglichkeit der zusätzlichen Differenzierung der Stichwortbegriffe. Bei der Modellierung von Stichwortlisten mit einem semantischen Analysemodell können Begriffe speziell annotiert werden, wenn diese in einem relevanten Zusammenhang zu einer bestimmten Sache, einer bestimmen Person oder einer definierten Handlung stehen.

In unserem Beispiel mit den vermuteten Unregelmäßigkeiten bei der Auftragserteilung nehmen wir, frei erfunden, die Firma bzw den Begriff „HAL" in die Stichwortliste auf. HAL ist eine internationale Firma mit Niederlassungen in fast allen Ländern der Welt. Es besteht die Anforderung, zu differenzieren und dabei Dokumente und Konversationen möglichst mit Bezug auf Niederlassungen von HAL in Österreich speziell zu betrachten. Diese Inhalte sollen über einen im Analysemodell hinterlegten Strukturpunkt unter „HAL Österreich" gefiltert und analytisch ausgewertet werden können.

„HAL" in den Textdokumenten automatisch zu finden und in einem Strukturpunkt des Facettenbaumes unter „HAL" abzubilden, ist ein einfacher Schritt. Selbst wenn die Firma mit Leerzeichen (H A L) oder mit Trennzeichen (H-A-L), aber auch wenn der ausgeschriebene Firmenwortlaut im Text steht, wird dieser Textteil als „HAL" erkannt und annotiert. Nach der Umsetzung der neuen Anforderung und erweiterten Parsing Rule findet der Anwender im Facettenbaum unter Firmen nach wie vor den Eintrag HAL, aber darunter einen weiteren fallbezogenen Strukturpunkt „HAL Österreich". Bei jeder beliebigen weiteren Recherche über den Strukturpunkt „HAL Österreich" werden alle Ergebnisse gezeigt und alle Begriffe markiert und ausgewertet, die „HAL Österreich" betreffen. Beispielsweise auch „Wienerstraße 12" wenn diese als Adresse einer österreichischen HAL-Niederlassung mit der Parsing Rule verknüpft ist. Wir sprechen hier auch von einem Wissensmodell, da die Parsing Rule „HAL Österreich" über sinnvolle Informationsverknüpfungen (Excelsheets, Datenbanken, Wikis, Ontologien etc) um die „Intelligenz" angereichert wird, welche Textinhalte auf einen praktikablen Bezug zu dieser Firma hinweisen.

Synonyme (zB für „Diskont", wie Preisminderung, Preisnachlass, Prozente, Verbilligung, Vergünstigung uam) werden über Standardwörterbücher zur Verfügung gestellt oder, bei ganz speziellem Wortgebrauch, durch zusätzliche Synonymlisten ergänzt.

6 Mit Stichproben manuell überprüfte Trefferquoten (engl *„recall"*), Genauigkeit (engl *„precision"*) und Ausfallquote (engl *„fallout"*).

Mehrdeutige Wörter, sog „Homonyme“, werden oft als Herausforderung semantischer Analysemodelle diskutiert. Häufiges Beispiel ist „Golf“ mit den Bedeutungen „Auto“, „Sport“ oder „Meerbusen“. Unsere Projekterfahrungen zeigen, dass es in einem fallspezifischen Analysemodell ausreichend ist, nur jene Homonyme abzubilden, die für den konkreten Untersuchungsfall auch eine entsprechende Relevanz haben. Wenn zB eine Person mit dem Namen „Golf“ eine Rolle spielt, dann wird die Regel zur Annotation dieser Person die Problematik der Mehrdeutigkeiten entsprechend berücksichtigen. Falls es zusätzlich auch relevant wäre, zwischen „Golf: Auto“ und „Golf: Sport“ zu unterscheiden, werden auch dazu geeignete Parsing Rules eingestellt. Die Anzahl relevanter Homonyme bei internen Untersuchungen und die Erstellung von Mehrdeutigkeitsregeln bleiben aber meist in einem sehr überschaubaren Rahmen.

## 4. Benutzerdefinierte Kategorien

Zusätzliche fallbezogene Strukturpunkte, mit ICI als „Benutzerdefinierte Kategorien“ bezeichnet, werden durch das Wissen über die Relevanz von bestimmtem Personen, Produkten, Projekten, Themengebieten oder Konversationspfaden (wie bestimmte Sender-Empfänger-Gruppen) definiert. Auch spezifische Zeiträume, wie zB der Zeitraum, in der Person A Geschäftsführer der Firma B war, oder Zeiträume für spezielle Projekte oder Angebotsfristen werden als benutzerdefinierte Kategorien hinterlegt. Diese Art von Strukturpunkten kann vom Untersuchungsteam beliebig definiert und eigenständig hierarchisch in der Facettennavigation einfach abgebildet werden. Die Zuordnung von Dokumenten zu solchen Strukturpunkten erfolgt vom Anwender über passende Abfragen, wobei neben freier Sucheingabe die gesamte vorhandene Facettenstruktur dafür genutzt werden kann. Auch für diese Strukturpunkte werden sämtliche verfügbaren Analysen, wie zB Häufigkeiten, Auffälligkeiten, Zeitreihen, Abweichungen, Korrelationen oder Zusammenhänge in Echtzeit berechnet und dargestellt. In der Literatur finden wir diese Funktionalität oft unter „virtuelle Verzeichnisse“. Das heißt, das System erlaubt eine Art Folder-Struktur anzulegen, in der Dokumente nicht dadurch vorkommen, indem sie dort physisch gespeichert oder verknüpft sind. Dokumente scheinen in diesen „Ordnern“ auf, sobald sie einer Abfrage bzw Regeln des semantischen Analysemodells entsprechen. ICI kann dabei zB einen Strukturpunkt (Folder) definieren, in dem alle Rechnungen für ein bestimmtes Projekt aufgelistet werden, sofern sich diese über deren Inhalte semantisch klassifizieren lassen. Die Möglichkeiten mit dieser Methode Ordnung in die zu untersuchende Datenmenge zu bringen, sind dabei umfangreich und beeindruckend.

## 5. Auswertung der Ergebnisse

Zurück zum Beispielsfall: Es gilt immer noch herauszufinden, ob zu bestimmten Angeboten oder Projekten im Kontext zu bestimmten Technologiekomponenten spezielle Informationen zwischen zwei bestimmten Personen ausgetauscht wurden. Dabei spielt es keine Rolle, in welche Richtung die Informationen geflossen sind, ob die Personen direkt adressiert oder auf Kopie gesetzt wurden oder ob sich die Informationen im E-Mail selbst oder in einem Anhang befinden. Für diese Fälle wird bspw in ICI mit dem verfügbaren Abfrage-Wizard eine dafür passende Abfrage erstellt und diese einem neuen Strukturpunkt im Facettenbaum, zB unter Kommunikationspfade – „Person A mit B“, zugeordnet.

Das fallbezogene Analysemodell erlaubt es, mit wenigen Schritten bestimmte Projekte und Technologiekomponenten zu filtern. Diese Ergebnismengen können in jedem Strukturpunkt analytisch ausgewertet und durch zusätzliche Abfragen oder Filter weiter eingeschränkt werden. Zum Beispiel führt bei unserem angenommenen Fall, eine tiefergehende Betrachtung auf Beträge im Text zwischen 6,500.000,00 und 8,000.000,00 Euro, zu einer weiteren brauchbaren Reduktion der Dokumentenmenge. Auch ein Dokument mit dem Textinhalt „… gut wäre unter 7 Mio zu bleiben …“ wird berücksichtigt, da der Text „7 Mio“ im Facettenindex durch die semantische Analyse auch mit dem Betrag 7,000.000,00 annotiert ist.

Angenommen, es verbleiben von der Ursprungsmenge nun rund 2.500 Dokumente, die unseren bisherigen Abfragekriterien entsprechen. Wir wissen aufgrund der bisherigen Analyse, dass diese Dokumente mit bestimmten Projekten und Technologien zu tun haben und in irgendeiner Form Betragsangaben zwischen 6,5 und 8 Millionen beinhalten. Wir gehen jetzt der Frage nach, ob zu den bereits als kritisch definierten Kommunikationspfaden E-Mails vorliegen bzw ganz allgemein, ob es dazu auffälligen weiteren E-Mail-Verkehr gegeben hat. Das Analysemodell liefert auf beide Fragen die sofortige Antwort.

Zum Verständnis: Die jetzt gefilterte Datenmenge wird statistisch in Echtzeit mit der gesamten restlichen Datenmenge (Big Data-Analytik) verglichen. Daraus ergeben sich „auffällige“ weitere Kommunikationspfade, Personen, Firmen, Kontonummern, was auch immer das Analysemodell ermöglicht.

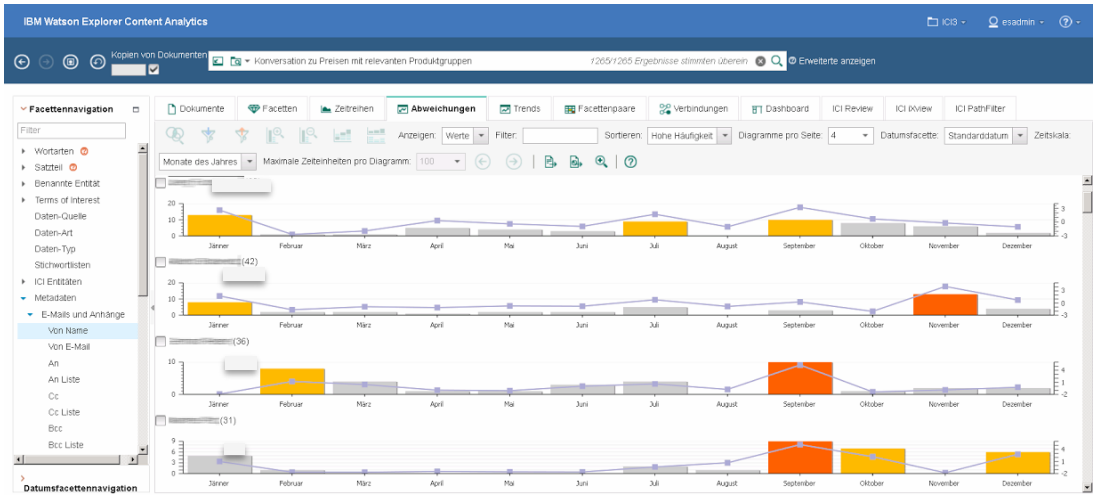


**Abb 3** zeigt eine von vielen möglichen Auswertungen in einer durch eine semantische Suche gefilterten Dokumentenmenge. Die Grafik zeigt „auffällige“ (rot) Absender von E-Mails auf der Zeitachse und ermöglicht unmittelbar eine weitere Selektion (Klick auf einzelne Balken), um bei Bedarf einen Review oder eine tiefergehende Analyse der „dahinterliegenden“ Dokumente durchzuführen. Weitere Darstellungen erlauben Trend- und Netzwerkanalysen bzw eine Verwendung der Daten in frei wählbaren Visualisierungs- oder speziellen Analysetools.

Der zuvor erstellte Strukturpunkt in den Kommunikationspfaden „Person A mit B“ zeigt in den jetzt gefilterten 2.500 Dokumenten rund 300 relevante Dokumente an. Diese können quergelesen, relevante Inhalte exportiert oder für die weitere Bearbeitung geflaggt werden.

Unser Beispielsfall sei nach rund 10 Stunden Recherchearbeit abgeschossen. Dabei wurde die Bearbeitung aller Fragestellungen (Such-, Auswertungs- und Ergebnisprotokolle) nachvollziehbar

dokumentiert und es wurden aus mehr als 500.000 Dokumenten 20 beweisrelevante Dokumente für den Untersuchungsbericht exportiert. Die zumeist umfangreicheren echten Fälle benötigen auch mit semantischer Analyse etwas mehr Zeit.

## IV. Fazit: Semantische Analyse ist ein Lernprozess

Die Menge an Möglichkeiten eines analytischen Systems bedingen auch einen Lernprozess, um diese in ihrer vollen Funktionalität nutzen zu können. Zusammenhänge und Pfade können zB nicht nur – wie oben erwähnt – zwischen Sender und Empfänger von E-Mails dargestellt werden, sondern auch zwischen allen möglichen Strukturpunkten. Durch die Kombination von Facetten (Facettenpaare) können zB Buchungstexte und Geldbeträge – zusammen mit Kontoinformationen – analysiert oder Firmennamen und Projekte in Bezug auf Personen dargestellt werden.

Der Lernprozess für den Anwender liegt vor allem in dieser neuen Arbeitsweise. *Ad-hoc*-Analysen in einem riesigen Datenbestand durchzuführen, interaktiv zu arbeiten, tiefer in die Daten hineinzugehen, Auswertungen zu interpretieren, bei Bedarf noch tiefer zu gehen oder wieder zurück nach oben, das alles erfordert ein Umdenken. Schon mit ersten Erfahrungen wird die Definition von Analysemodellen viel einfacher. Aktuelle Entwicklungen von IBM, besonders im Bereich des kognitiven Computing, werden es ermöglichen, nicht nur natürlich-sprachlich auszuwerten, sondern auch die Eingaben an das System mit freier Texteingabe durchzuführen.

Suche setzt voraus zu wissen, wonach gesucht wird. Mit semantischer Suche kann intelligent gesucht werden. Auffälligkeiten können gerade auch da auftreten, wo diese gar nicht vermutet wurden. In einem konkreten Anwendungsfall mit ICI wurde der wichtige Hinweis in einem E-Mail mit einer Schimpfsuada gefunden. Bestimmte Personen, die auf keiner Untersuchungsliste standen, wurden vom Analysemodell mit hoher Auffälligkeit in diesen Nachrichtentexten entdeckt. Die weitere Suche und Analyse mit Bezug auf diese Personen hat dann zu wichtigen neuen Netzwerken und Erkenntnissen in jenem Fall geführt. Mit semantischer Analyse wurden auch wichtige Inhalte entdeckt, die im Vorfeld nicht als relevant erkannt wurden und nach denen nicht direkt gesucht worden wäre.

Die Arbeit mit ICI wird von den Anwendern als sehr praxistauglich und effizient beurteilt. IBM hat die Lösung von BIConcepts mit der Nominierung zum Beacon Award 2016 weltweit in der Kategorie „Watson Outstanding Cognitive Computing Solutions“ ausgezeichnet.