Band 15 / 2018

Wissensentwicklung mit IBM Watson in der Zentraldokumentation (ZentDok) der Landesverteidigungsakademie

Entwicklungen und Anwendungen in der Open-Source Informationsbereitstellung des ÖBH

Klaus Mak, Hans Christian Pilles, Markus Bertl und Joachim Klerx

Schriftenreihe der
Landesverteidigungsakademie

UNSER HEER

Schriftenreihe der
Landesverteidigungsakademie

Klaus Mak, Hans Christian Pilles, Markus Bertl, Joachim Klerx

Wissensentwicklung mit IBM Watson in der Zentraldokumentation (ZentDok) der Landesverteidigungsakademie

Entwicklungen und Anwendungen in der Open-Source Informationsbereitstellung des ÖBH

15/2018
Wien, Oktober 2018

Impressum:

Medieninhaber, Herausgeber, Hersteller:
Republik Österreich / Bundesministerium für Landesverteidigung
Rossauer Lände 1
1090 Wien

Redaktion:
Landesverteidigungsakademie
Zentraldokumentation und Information
Stiftgasse 2a
1070 Wien

Schriftenreihe der Landesverteidigungsakademie

Oktober 2018
ISBN 978-3-903121-55-3

Druck:
ReproZ W 18-
Stiftgasse 2a
1070 Wien

Inhaltsverzeichnis

1 Zusammenfassung 7
2 Abstract 9
3 Einleitung 11
3.1 Ausgangsituation in der Wissensentwicklung an der ZentDok 11
3.2 Daten für Anwendungsbeispiele 14
3.3 Metadaten 15
3.3.1 Interne Metadaten 16
3.3.2 Externe Metadaten 18
3.4 Annotation 18
3.5 Suche und Analyse 20
4 Pre-Prozesse 23
4.1 Datenmanagement 24
4.2 Pre-Processing 26
4.3 Semantische Vorarbeiten und Werkzeuge, ProTerm 26
5 IBM Watson Explorer (WEX) 29
5.1 IBM Watson 29
5.2 IBM Watson Explorer Deep Analytics Edition 29
5.3 IBM Watson Explorer - Analytical Components (WEX/AC) 29
5.4 IBM WEX/AC Komponenten 30
5.4.1 Crawling 30
5.4.2 Parsing 31
5.4.3 Indexing 31
5.4.4 Inhaltsanalyseminer 31
5.5 IBM WEX/AC Analysemodule 32
5.5.1 Suche 33
5.5.2 Erweiterte Suche 35

5.5.3 Facetten 36
5.5.4 Zeitreihen 38
5.5.5 Abweichungen 40
5.5.6 Trends 40
5.5.7 Facettenpaare 42
5.5.8 Verbindungen 43
5.5.9 Dashboard 45
5.5.10 Meinungen (Sentiment) 46
5.5.11 Berichte 47
5.6 Content Analytics Studio 47
5.6.1 UIMA Pipeline 49
6 IBM i2 51
6.1 IBM i2 Analyst's Notebook Premium 53
6.1.1 Netzwerkanalysen 54
6.1.2 Zeitreihenanalysen 56
6.1.3 Histogramme und Heatmaps 57
6.1.4 Georeferenzierung 58
6.2 IBM i2 Analyze 59
6.3 IBM i2 Enterprise Insight Analysis (EIA) 60
6.4 Intelligence Portal 60
6.5 Szenarien für die Datenhaltung 61
6.5.1 IBM i2 ANBP mit Local Analysis Repository (LAR) 62
6.5.2 IBM i2 ANBP mit Group Analysis Repository (GAR) 63
6.5.3 IBM ANBP mit Group Analysis Repository (GAR) mit Information Store und ESRI 64
7 IBM Cloud 65
7.1 IBM Cloud - AI Services Überblick 66
7.2 IBM Watson Knowledge Studio (WKS) 67

7.2.1 WKS-Prozess 69
8 Zukünftige Entwicklungsschritte 71
9 Abbildungsverzeichnis 79
10 Glossar 81
11 Stichwortverzeichnis 82
12 Literaturverzeichniss 83
13 Autoren 87

1 Zusammenfassung

An der Zentraldokumentation der Landesverteidigungsakademie (ZentDok/LVAk) werden seit mittlerweile 50 Jahren Anstrengungen unternommen, um der Gesamtorganisation des ÖBH qualitativ hochwertige offene Fachinformationen zur Verfügung zu stellen. Aber auch technische und semantische Möglichkeiten zur Unterstützung der Wissensentscheidung und Entscheidungsunterstützung aller Bedarfsträger waren stets ein Grundauftrag dieser Organisation. Diese ständige operative und anwendungsorientierte Entwicklungsarbeit findet in den letzten Jahren ihren vorläufigen Höhepunkt in der Auseinandersetzung mit dem anspruchsvollen Programmpaket IBM Watson. Es wurden und werden alle Anstrengungen unternommen und in dieser Publikation beschrieben, um die Wissensentwicklung und bereits realisierte und zukünftige Anwendungsmöglichkeiten dieser Zukunftstechnologie für die Bereitstellung von offenen Fachinformationen im ÖBH nutzbar zu machen. Dabei sollen die Potentiale von AI/KI (Artificial Intelligence, Künstliche Intelligenz), ML (Machine Learning) sowie Visualisierungstechnologien für weitere Möglichkeiten ihres Einsatzes im ÖBH erkenn- und nutzbar werden.

Diese Publikation zeigt in Grundzügen die Funktionalität der Softwarepakete IBM Watson Explorer Deep Analytics Edition und IBM i2 Enterprise Insight Analysis. Der Watson Explorer stellt ein mächtiges Tool zur Auswertung von unstrukturierten Daten (Text) dar. Er bietet die Möglichkeit, riesige Textmengen intelligent durchsuchbar zu machen, Entitäten auf Basis von Ontologien, Regeln und künstlicher Intelligenz zu extrahieren und über Korrelationen in Echtzeit miteinander in Verbindung zu setzen. i2 ist ein Tool für Netzwerkanalysen das hilft, aus Daten Strukturen und Verbindungen zu erkennen. Es werden unter anderem Möglichkeiten der Visualisierung, Social Network Analysis und Geoanalysen geboten, um versteckte Verbindungen und Muster zu erkennen. In den Kapiteln wird neben einer generellen Beschreibung der Softwareprodukte auch immer wieder auf die Anwendung zur Wissensentwicklung in der ZentDok eingegangen und die dort umgesetzte Projektkonfiguration (mit Unterstützung der Fa. BIConcepts, Herr Fuchslueger und Herr Bertl), vorgestellt.

Weiters wird ein Einblick in die Watson Services der IBM Cloud und die zukünftigen Entwicklungsschritte in diesen Bereichen gegeben.

2 Abstract

For 50 years now, every effort has been made at the Central Documentation Department of the National Defence Academy to provide the Austrian Army with high-quality open-source information products. But also, all technical and semantic possibilities to support knowledge development and decision making processes of all users have always been a basic task of this organization. In recent years, this constant operative and application-oriented development work has reached its preliminary culmination in the confrontation with the sophisticated IBM Watson software package. All efforts have been taken, to develop and realize the use of these future technologies. The potentials of AI (Artificial Intelligence), ML (Machine Learning) and visualization technologies for further possibilities of their use in the Austrian Army shall be recognized and tested.

The aim of this publication is to describe the main functionality of the software bundles IBM Watson Explorer Deep Analytics Edition, IBM i2 Enterprise Insight Analysis and the Watson Services in the IBM Cloud. Watson Explorer is a powerful tool for analyzing unstructured data (text mining). It enables its user to perform intelligent queries on a vast amount of data, extract relevant entities using ontologies, rules and artificial intelligence. It offers functionality to perform content analytics tasks and real time correlation analysis to link entities together. i2 is a network analysis tool for turning data into intelligence using for instance visualization of networks, social network analysis and geospatial views so that hidden connections and patterns inside a network can be uncovered. Additionally, to the general functionality of the software, the use of this tools in the Central Documentation Department is shown.

3 Einleitung

Die Entwicklungen sowie Anwendungen, die hier dargestellt werden, dokumentieren den Einsatz der Softwarepakete IBM Watson (Watson Explorer Deep Analytics Edition) und IBM i2 (i2 Enterprise Insight Analysis und Analyst's Notebook Premium) in der ZentDok der Landesverteidigungsakademie. Dies stellt einen Meilenstein in der kontinuierlichen Entwicklung der Open-Source-Informationsvermittlung für das ÖBH durch die ZentDok dar. Neben dem Nutzen für die unmittelbare Arbeit der Experten in der ZentDok, sollen in absehbarer Zeit sämtliche Bedarfsträger des ÖBH diese neuen Anwendungen eines der modernsten Softwarepakete in der Fachinformationsversorgung und weiteren Anwendungsgebieten des ÖBH für ihre operative Arbeit nutzen können.
Im ersten Teil dieser Publikation werden die Projektkonfiguration und ihre Implementierung in die operative Arbeit der ZentDok beschrieben. Dies geschieht an Hand von konkret umgesetzten Aufgabenstellungen. Im zweiten Teil werden nach Analyse der Erstergebnisse und weiterer Entwicklungsaufgaben hochwertigste Verfahren der künstlichen Intelligenz (AI/KI) wie auch des Machine Learning (ML) beschrieben und deren Nutzen und eventuelle Einsatzmöglichkeiten im ÖBH vorgestellt.

3.1 Ausgangsituation in der Wissensentwicklung an der ZentDok

In Abbildung 1 wird die Ausgangssituation für alle Projektschritte skizziert. Die drei Hauptphasen des Projektes werden unterteilt in die Pre-Prozesse, die Watson-Prozesse und die i2-Prozesse. Grundsätzlich sind alle Phasen inhaltlich zusammenhängend, da im Pre-Prozess bereits alle Zielvorgaben für weitere Verarbeitungsschritte im Detail zu berücksichtigen sind. Durch die notwendige Vorbereitung und die modulartige Konfiguration der Softwarepakete ergab sich diese Projektkonfiguration fast zwangsläufig.

Alle Projektplanungsschritte, deren Dokumentation und deren Operationalisierung werden im Prozess- und Wissensmanagement-Werkzeug ADONIS-PROMOTE modelliert und abgebildet sowie in weiterer Folge als Mustermodelle und Musterprozesse im internen Netz oder als HTML-Export den Nutzern zur Ausbildung und zur kollaborativen Weiterentwicklung zur Verfügung gestellt. Dieser Prozess

wird bei Göllner, Mak und Woitsch, (2010a) und Mak und Woitsch, (2005) beschrieben.

Diese Vorgangsweise ermöglicht in weiterer Folge auch eine Übernahme in das Strategie- und Performancemanagement-Werkzeug ADOSCORE um ein sogenanntes Performance Management System aufzubauen. Damit wird eine Möglichkeit zur Steuerung über Kennzahlen umgesetzt (Göllner, Mak und Woitsch, 2010b).

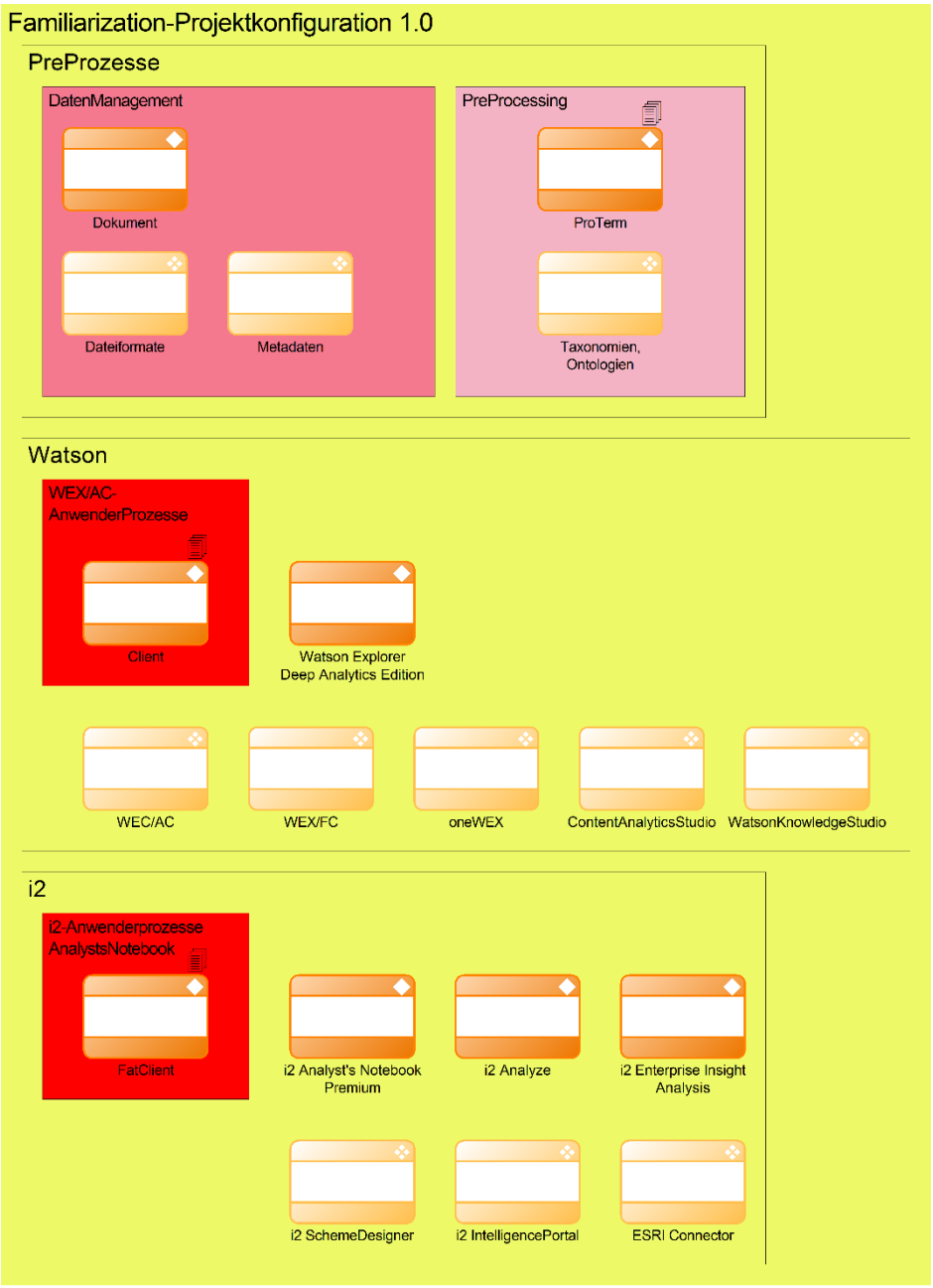


Abbildung 1: Projektkonfiguration 1

3.2 Daten für Anwendungsbeispiele

Alle Daten, die für die ersten Anwendungsbeispiele herangezogen werden, sind Open-Source-Dokumente aus verschiedensten Repositorien der Zentraldokumentation. In Abbildung 2 wird beispielsweise die Meldungsplattform für Cyber-Informationen des Cyber-Dokumentations- und Forschungszentrums der ZentDok (CDFZ) dargestellt (Mak et al., 2015). Mehr als 100.000 Dokumente, geordnet in einem Kategoriesystem und versehen mit verschiedensten Meta-Tags, können so weiterverarbeitet werden. Die Gesamtzahl der Dokumente der ZentDok für die beschriebenen Anwendungsbeispiele erreichen derzeit ca. drei Millionen und sind mit Masse in deutscher und englischer Sprache sowie mehr als 30 weiteren Sprachen verfügbar.

Abbildung 2: News Erfassung und Visualisierung in Ushahidi

Die CDFZ Daten bestehen aus Meldungen, die im Internet mittels Open Source Informationsarbeit (OSInfo) als relevant identifiziert und erfasst

werden. Bei der Datenerfassung werden die Daten mit einer ganzen Reihe von Kategorien versehen, die sich aus den analytischen Anforderungen heraus ergeben haben. Einige Grundkategorien, wie z.B. Land, Ort, soziale und politische Faktoren ergeben sich aus elementaren Kategoriesystemen, wie z.B. dem Doppelvektoren Modell (Göllner et al., 2014). Andere spezifischere Kategorien, wie z.B., Cyber War, kritische Infrastruktur, Konflikttyp ergeben sich aus den Motiven für die unterschiedlichen analytischen Ansätze zur Auswertung. Weitere Informationen dazu können bei Mak et al., (2015) gefunden werden.

Prinzipiell lassen sich alle verschiedenen strukturierten und unstrukturierten Datensätze in die Wissensentwicklung integrieren. Für die beschriebenen Fallbeispiele wurde eine Auswahl unter Bezug auf Verfügbarkeit, Kosten und Nutzen getroffen. Alle Datensätze bestehen jeweils aus Metadaten und Content Daten. Im Folgenden wird beschrieben, welche der möglichen Metadaten genauer untersucht wurden.

3.3 Metadaten

Metadaten sind beschreibende Informationen über Inhalt, Beziehung, Struktur und Bedeutung von Daten, prozessbezogene Informationen über deren Verarbeitung sowie die Beschreibung von Anwendungen und Zugriffsrechten. Als solche enthalten sie eine erstaunliche Anzahl an Informationen über die Struktur der zu analysierenden Daten und Wissensobjekte. Im Zuge der Analyse werden soweit wie möglich zusätzliche Metadaten (sowie Entitäten) extrahiert. Deswegen sollte in der Gliederung der Metadaten nach internen (siehe Kapitel 3.3.1) und externen (siehe Kapitel 3.3.2) Metadaten unterschieden werden.

Arten von Metadaten basierend auf Bossert, (2014):

- *Administrative Metadaten* sind Angaben zum Sachbearbeiter, Dateinamen, Speicherort oder zur Provenienz.
- *Technische Metadaten* beinhalten unter anderem Angaben über das Dateiformat, über die Datenmenge oder über die Erstellungs-Software.
- *Deskriptive Metadaten* beschreiben den Inhalt (z.B. Autor, Titel, …) eines digitalen Objektes.

- *Strukturelle Metadaten* informieren über die Beziehungen und Verknüpfungen zu anderen Objekten.

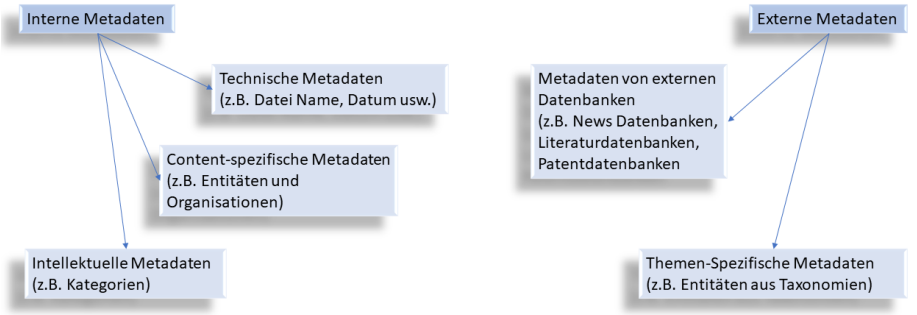


Abbildung 3: Übersicht über die verwendeten Kategorien von Metadaten in IBM Watson Explorer

Metadaten werden im Watson Explorer über Indexfelder bereitgestellt, da diese strukturgebenden Daten üblicherweise zur weiteren Analyse, zur Gruppierung, Filterung und Suche verwendet werden. Die Indexfelder können über Facetten im Inhaltsanalyseminer zur Auswertung bereitgestellt werden.

3.3.1 Interne Metadaten

Aus Sicht der ZentDok lassen sich bei den internen Metadaten technische, Content-spezifische und intellektuelle Metadaten unterscheiden. Im Folgenden werden jeweils kurze Beispiele zu allen verwendeten internen und externen Metadaten angeführt.

Interne Metadaten sind Daten, die aus dem Datensatz selbst abgeleitet werden und zur Strukturierung des Datensatzes beitragen, wobei die gewünschte Strukturierung jeweils mit der Motivation für die Analyse korrespondiert. Deswegen sind manche Metadaten für eine Analyse relevant und andere sind es nicht.

Technische Metadaten sind die erste Kategorie der internen Metadaten. Diese werden aus den Dateieigenschaften oder aus dem Workflow abgeleitet:

- Dateiname
- Dateigröße
- Datum Erstellung
- Datum Änderung
- …

Content-spezifische Metadaten sind unter Nutzung div. Methoden und Werkzeuge aus dem Content für die jeweilige Objektgruppe im Inhaltsanalyseminer extrahierte und über Facetten bereitgestellte relevante Entitäten:

- Personen
- Organisationen
- Politische Parteien
- Krisenherde
- Terrorgruppen
- …

Intellektuelle Metadaten sind händisch erfasste Entitäten bzw. Kategorien, die über Facetten bereitgestellt werden. Im Bereich CDFZ sind dies alle Kategorien aus der News-Erfassung für die Cyber-Meldeplattform. Alle Meldungen werden über Kategorie-Systeme getaggt (angereichert):

- Cyber Thema
- Site-Info
- Regionen
- DocType
- Horizon Expert Tagging
- Horizon Expert Tagging – Mil
- CROWD-Recherche
- 6W-Bericht
- CROWD-Rechercheberich
- CROWD-MetaRecherche-KritInfra
- Konferenzen
- Semantisches Feld

3.3.2 Externe Metadaten

Externe Metadaten sind relevante Entitäten aus externen Datenquellen (also nicht aus dem Content extrahierte).

Themen-spezifische Metadaten sind relevante Entitäten, die aus externen Datenquellen gewonnen und über Facetten bereitgestellt werden:

- Listen - Hacker-Namen, Hacker-Gruppen, Hacker-Operationen, …
- Taxonomie – Begriffe der kritischen Infrastruktur
- …

3.4 Annotation

Als Annotationen werden hier intelligente „Regeln" verstanden die es erlauben aus einem Text domänenspezifische Inhalte wie Namen, Nummern oder Objekte zu extrahieren und auswertbar zu machen. Zum Erstellen einer Annotation können zum Beispiel regelbasierte Verfahren wie „Regular Expressions" (RegEx), Wörterbücher oder Ontologien zum Einsatz kommen wie auch das Machine Learning. Eine Kombination der Verfahren ist ebenfalls möglich. Im weitesten Sinne kann gesagt werden, dass durch Annotationen aus unstrukturierten Texten strukturierte Daten erzeugt werden. Die Annotationen erlauben es auch, Textinhalte, die normalerweise nur von Menschen verstanden werden, der Software „verständlich" zu machen. Sie erzeugen also Wissen aus dem Text und bringen Intelligenz in den Text. Annotationen tragen als wesentlicher Bestandteil zu einem guten Analyseergebnis bei. Entwickelt werden Annotationen im Watson Explorer Umfeld entweder mit dem Content Analytics Studio (siehe Kapitel 5.6 Content Analytics Studio) auf Basis der Unstructured Information Management Architecture (UIMA) oder dem Watson Knowledge Studio (siehe Kapitel 7.2 Watson Knowledge Studio).

Die Abbildung 4 stellt die verschiedenen Annotationsarten gegenüber und zeigt mit welchen Tools diese entwickelt werden können.

Rule Based	**Machine Learning**	**Hybrid**
• Auf Regelbasis • Nachvollziehbar • Lernaufwand • Ab einer gewissen Komplexität schwer wartbar	• Auf Basis von Training und Statistik • Wenig Lernaufwand für den Anwender. Die Maschine lernt • Übersichtlich • Schwer nachvollziehbar • Benötigt eine ground truth	• Kombination von rule based und ML • Kombinierte Vorteile der anderen Methoden • Regeln verringern Trainingsaufwand und erhöhen die Präzision von ML
Content Analytics Studio	Watson Knowledge Studio	Watson Knowledge Studio

Abbildung 4: Annotationsarten

Gemeinsam mit den Metadaten werden Annotationen über Facetten im Watson Explorer zur Analyse bereitgestellt.

Für die ersten Anwendungsbeispiele wurden folgende Facetten implementiert:

- Cyber Thema
- Risiko
- Hacker
- Hacker Gruppen
- Hacker Namen
- Hacker Spaces
- Substantivfolgen
- Rubrik
- Personennamen (mit und ohne Titel)

3.5 Suche und Analyse

Ein wichtiger Teil im Wissensmanagement (WM) ist die Suche. Diese hat sich in den letzten Jahren drastisch weiterentwickelt, von der einfachen Stichwortsuche über den Unified Access verschiedener Datenquellen bis hin zu semantischen Analysen, Entity Extraction und kognitiven Fähigkeiten. Auch die Suchoberflächen haben sich verändert. Für ein modernes WM wird eine auf jede Benutzergruppe zugeschnittene Oberfläche erwartet und benötigt. All diese Funktionen steigern den Business Value einer Such- und Analyselösung. Diese Entwicklung wird in Abbildung 5 gezeigt. Das Softwarepaket Watson Explorer Deep Analytics *Edition* deckt dabei alle oben genannten Funktionen ab (ibm.com, 2018c).

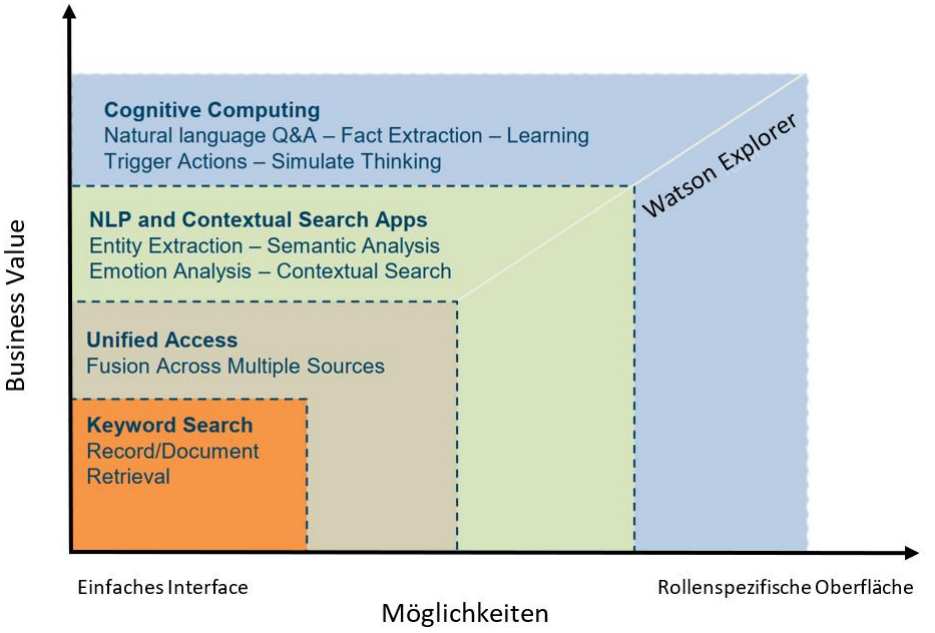


Abbildung 5: Entwicklungsschritte der Suche

Ein weiterer wichtiger Teil im Knowledge Discovery und Management Bereich ist, neben der Suche, die Analyse von Daten. Auch hierzu stellt der Watson Explorer einige Funktionen wie Facettenanalyse, Häufigkeiten, Korrelationsanalyse, Zeit-, Abweichungs- und Trendanalysen sowie viele weitere Funktionen zur Verfügung (ibm.com, 2018c). Dies wird genauer im Kapitel 5.5 IBM WEX/AC Analysemodule beschrieben.

4 Pre-Prozesse

Die oben angeführten Dokumentenbestände, die in weiterer Folge durchsucht und analysiert oder deren Zusammenhänge visualisiert werden sollen, werden durch das Qualitätsmanagement System (QMS) der ZentDok im sogenannten Pre-Processing-Verfahren aus den unterschiedlichen Repositorien für die Weiterverarbeitung mit Watson Explorer (WEX) oder i2 Enterprise Insight Analysis (EIA) vorbereitet. Neben der Erfassung von verschiedensten Metadaten, werden auch im Bereich der Textanalyse notwendige inhaltliche Vorarbeiten geleistet. So können mit ProTerm (einem intern entwickelten Textanalysesystem) und weiteren eigenentwickelten Named Entity Recognition Tools (NER) wichtige Entitäten mit hoher Genauigkeit, aber auch mit hohem Aufwand, für eine qualitativ hochwertige Weiterverarbeitung zur Verfügung gestellt werden. Im nächsten Entwicklungsschritt sollen auch Methoden aus dem Bereich künstliche Intelligenz, Machine Learning und Natural Language Processing zum Einsatz kommen.

Die Pre-Prozesse umfassen, immer abgestimmt auf die Zielsetzung, die Schritte

- Formulierung und Dokumentation von Fragestellungen an ein Analysesystem für die Zielerreichung
- Festlegung von erforderlichen Repositorien
- Analyse der Datenqualität und Struktur der Repositorien
- Analyse und Identifizierung sowie Dokumentation von zu erwartenden Sichten (Datenfelder, Facetten)
- Identifizierung von vorhandenen und zu extrahierenden Metadaten sowie Entitäten
- Strukturierte Extraktion von Metadaten und Entitäten
- Verarbeitung und Vorbereitung von Metadaten und Entitäten für ein Mapping in Indexfelder und deren Anzeige in der Facettenstruktur
- Programmierung eines Custom Crawlers (bei Bedarf) für den Import von Daten sowie Metadaten und Entitäten in WEX/AC

- Anpassung und Abbildung des Pre-Prozesses bei erweiterten, geänderten oder neuen Fragestellungen, Problemstellungen und Anforderungen
- Erstellung von Musterprozessen und dadurch eventuelle Modellerweiterungen

4.1 Datenmanagement

Wie in Abbildung 6 dargestellt, können die Pre-Prozesse in zwei Hauptschritte unterteilt werden:

- dem Datenmanagement und
- dem Pre-Processing selbst

Beim Datenmanagement müssen die Voraussetzungen für die Weiterverarbeitung in WEX und i2 geschaffen werden. Dies bedarf einer genauesten Analyse der Funktionalitäten und den zu erfüllenden Anforderungen aus der zu erreichenden Aufgabenstellung. Die Importmodalitäten des Datenbestandes betreffen die Inhalte, deren Struktur und deren Metadaten. Faktoren dabei ist etwa die Sprache, die jeweiligen Formate, Qualität und Quantität sowohl des Contents wie auch der Metadaten.

In den dargestellten Anwendungsbeispielen beschränken sich die Sprachen auf Deutsch und Englisch.

In Abbildung 6 wird der Prozess von der jeweiligen Anfrage bis zu den Watson Explorer Prozessen abgebildet. Die Fragenanalyse und die Vergleichsprozesse mit allen Möglichkeiten, die das Softwarepaket bietet, stellen dabei eine große Herausforderung dar. Zusätzlich müssen alle Datenbestände, deren Qualität und Quantität, mitberücksichtigt werden. Dies stellt auch für ein bereits eingespieltes Team eine neue Herausforderung dar. Die Erstellung von Mustermodellen und Musterprozessen für die jeweilige Weiterverarbeitung und jeweilige Domain sind unabdingbare Forderungen, um eine unternehmensweite Verwendung in weiterer Folge realisieren zu können.

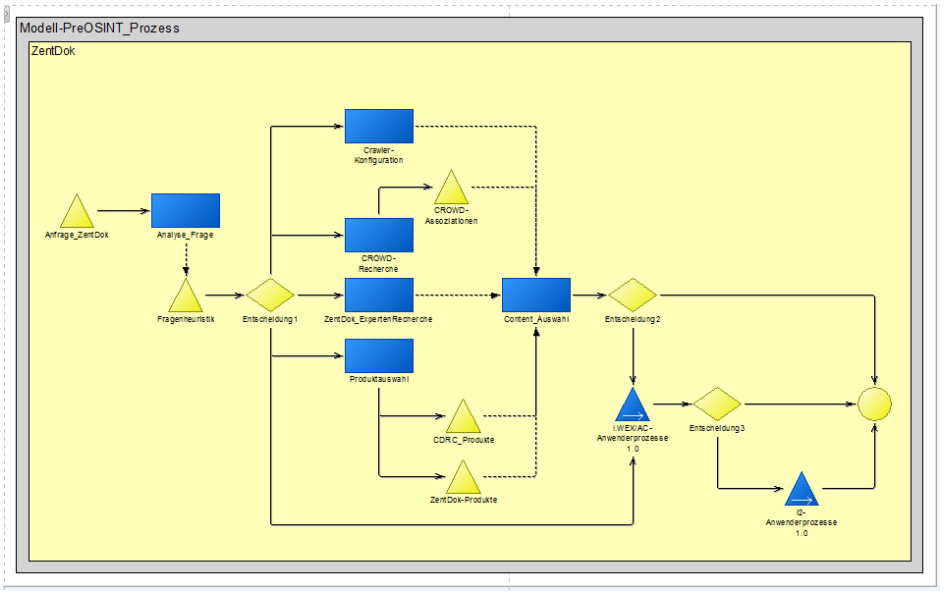


Abbildung 6: ProOSINT Prozess

Der jeweilig eingeteilte Prozessverantwortliche und Entscheidungsträger hat auch die qualitätsgesicherte Modellierung zu übernehmen, um so eine kontinuierliche und innovative Weiterentwicklung aller neuen Anforderungen und deren Dokumentation im Netzwerk sicherzustellen.
Nach erfolgter Auswahl der Dokumente, die für die Weiterverarbeitung zur Entscheidung 2 und 3 bestimmt wurden, müssen die Strukturen und die Schnittstellen der Inputmechanismen angepasst werden. Ein Modell für eine Anfrage entsteht. Dies muss von internen Entwicklern geleistet werden – auch dies bringt neue Herausforderungen und oft neue Überraschungen mit sich.
Alle einmal durchgeführten Anpassungen sind unternehmensweit im Wissensmanagement-System abrufbereit.

4.2 Pre-Processing

In diesem Arbeitsschritt können drei Verfahren, auch in Kombination, zur Anwendung kommen. Dies ist abhängig neben der voran zu stellenden Fragenanalyse, von der Datenmenge, der Datenqualität und der jeweiligen Sprache der Dokumente. Da WEX sowohl über Content-Analyse als auch über Metadaten-Analyse zu den Ergebnissen kommt, ist eine semantische Vorarbeit für hohe Qualitätsansprüche oft notwendig. Auch bei Sprachen, die nicht im IBM-Spektrum liegen, sind Eigenentwicklungen und Eigenanwendungen erforderlich.

Ziel des Pre-Processing ist die Konvertierung und Datenaufbereitung der Ausgangsdaten in ein Format, welches sinnvoll weiterverarbeitet werden kann.

4.3 Semantische Vorarbeiten und Werkzeuge, ProTerm

Die Analyse von Fachtexten stellt eine semantische und linguistische Herausforderung dar. Neben grammatikalischen Verfahren kommen auch statistische Verfahren zur Anwendung, um relevante Aussagen oder Fachtermini aus Texten extrahieren zu können. Neben der Analyse selbst, liegt auch ein Schwergewicht auf der Entwicklung oder der Auswahl von vorhandenen Vergleichskorpora zur genauen Bestimmung von Inhalten in neuen Dokumentenbeständen durch Vergleiche. Danach können eventuell neue relevante Begriffe maschinell oder intellektuell erfasst werden.
Mit dem Analysetool ProTerm, einer Eigenentwicklung der ZentDok, können statistische Verfahren zur Textanalyse herangezogen werden.

In Abbildung 7 werden türkischsprachige Dokumente analysiert und die relevanten Begriffe für die notwendige Weiterverarbeitung vorbereitet. Diese anspruchsvolle Aufgabe erfordert neben sehr guten Sprachkenntnissen auch die jeweils erforderlichen Linguistik- und Semantikkenntnisse der zu analysierenden Sprache.

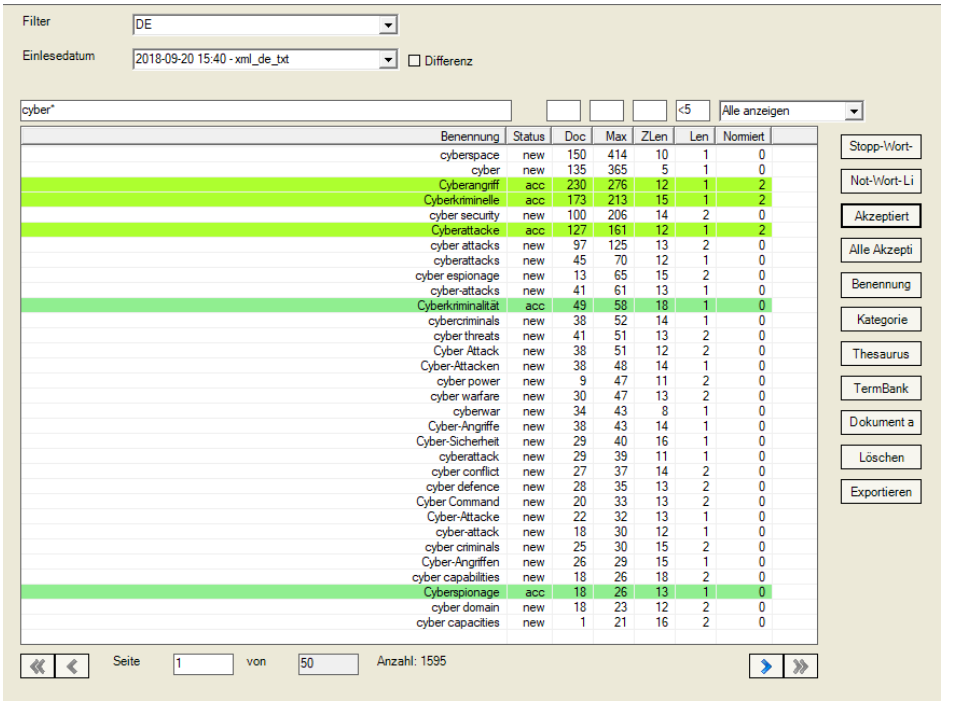


Abbildung 7: ProTerm Textanalyse - Modul NewTerm – Beispiel „cyber“

5 IBM Watson Explorer (WEX)

5.1 IBM Watson

IBM Watson ist ein übergeordneter Begriff für Software und Services von IBM die im Umfeld von „Cognitive Computing“ angesiedelt sind. Unter Cognitive Computing werden Lösungen und Methoden verstanden, die stark in Richtung maschinelles Lernen (ML) oder Deep Learning, meist über neuronale Netze und/oder mit dem „Verstehen“ der natürlichen Sprache (Natural Language Processing, NLP) zu tun haben. Der Begriff Cognitive Computing im Bereich Watson wird bei High, (2012) ausführlich beschrieben.

5.2 IBM Watson Explorer Deep Analytics Edition

IBM Watson Explorer Analytical Components ist eine Server-basierte Standardsoftware von IBM auf Basis von UIMA für die „intelligente“ Suche und Analyse (Textmining) innerhalb umfangreicher Textdatenbestände.

IBM Watson Explorer Deep Analytics Edition besteht aus den Programmteilen:

- Analytical Components (WEX/AC) = Text Mining
- Foundational Components (WEX/FC) = Suche, 360° View
- oneWEX = die Zusammenführung von WEX/AC und WEX/FC, für Suche, Analyse und Text Mining

Die oneWEX Komponente wird ab Version 12 der Watson Explorer Deep Analytics Edition mit ausgeliefert (ibm.com, 2018c).

5.3 IBM Watson Explorer - Analytical Components (WEX/AC)

IBM Watson Explorer - Analytical Components bietet Funktionen, die eine breite Anwendung für Suchmöglichkeiten sowie für Content Analyse, kombiniert mit Metadatenanalyse, verschiedensten Visualisierungsmöglichkeiten sowie statistische Auswertungen, ermöglichen.

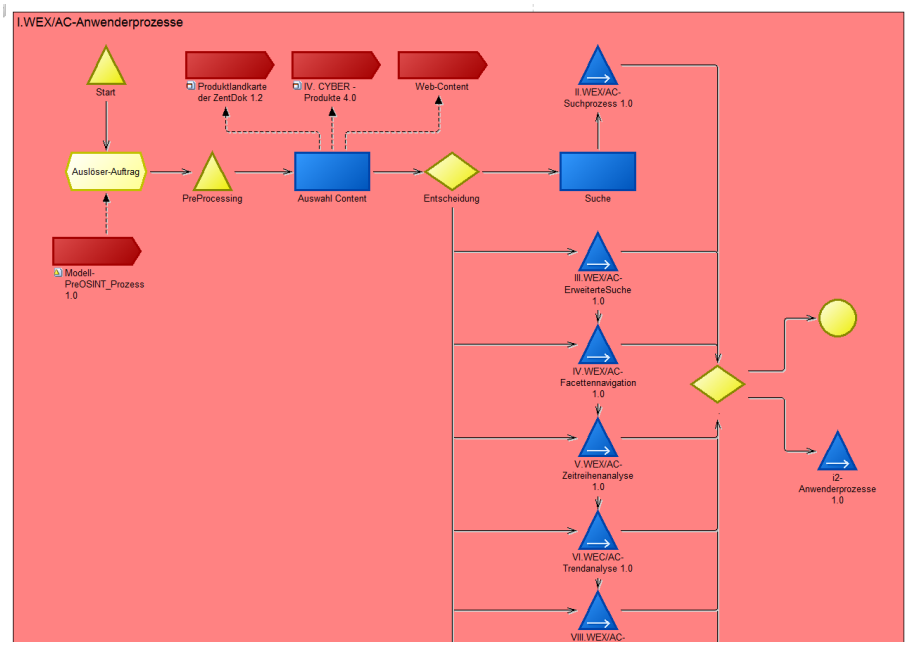


Abbildung 8: WEX/AC - Anwenderprozess

5.4 IBM WEX/AC Komponenten

IBM WEX/AC kann in 4 wesentliche Komponenten bzw. Funktionsbereiche unterteilt werden.

5.4.1 Crawling

Anbindung von Datenquellen (Datenbank, Filesysteme, HTML, XML, Mailserver, Content Management Systeme, Web, …) in denen Textinformationen gespeichert sind. Der Crawler belässt die Daten dort wo sie gespeichert sind, holt sich relevante Metadaten sowie extrahiert den reinen Textinhalt, behält aber einen Verweis auf die Originaldatei, sodass diese später jederzeit wieder abgerufen werden kann.
WEX stellt fertige Crawler für den Zugriff auf diverse Systeme zur Verfügung. Diese Crawler können direkt über die Admin Konsole in WEX/AC (z.B. für Webinhalte, Datenbanken oder Files) konfiguriert aber auch, für spezielle Anwendungen (z.B. PDF + Metadaten-XML) oder zur

Anbindung von nicht out oft the Box unterstützten Systemen, über in Java programmierte „Custom Crawler“ eingebunden werden.

5.4.2 Parsing

Der reine Textinhalt wird von der „Parsing Engine“ im ersten Schritt strukturiert (Tokenization, Sätze, Absätze, Seiten). Die „Tokens“ werden mit Annotationen „angereichert“. Nach der Spracherkennung werden über „Part of Speech“ (POS-Tagger: Wortarten, Wortartfolgen, Wortstamm) erkannte Entitäten (Personen, Firmen, Orte) optional auch „Sentiment“ (negative/positive Phrasen) standardmäßig annotiert. Dies passiert auf UIMA Basis (siehe Kapitel 3.4 und Kapitel 5.6.1). All diese Funktionalitäten werden fertig mit dem Standardprodukt ausgeliefert, können aber bei Bedarf in der Oberfläche des Watson Explorers an die eigenen Bedürfnisse angepasst werden.
Zusätzlich können beliebige, individuelle Annotationen implementiert werden, die zuvor im Content Analytics Studio (Windows Client Software auf Eclipse Basis) entwickelt und getestet wurden. Optional können Annotatoren mit dem Watson Knowledge Studio (WKS) entwickelt werden. WKS ist eine kostenpflichtige Cloud Service von IBM und erlaubt zusätzlich zu RegEx, Wörterbüchern und Ontologien auch Annotationen auf Basis von „Machine Learning“ zu entwickeln.

5.4.3 Indexing

Der reine Text sowie die dazugehörigen Metadaten und Annotationen werden vom Indexserver in einem Lucene ähnlichen Index gespeichert. Dabei werden die Metadaten und Annotationen in definierten Feldern abgelegt. Die Inhalte dieser Felder können als „Facette“ im Frontend zur Auswahl (Filter) und zur Analyse bereitgestellt werden. Neben dem reinen Suchindex kann mit WEX/AC ein Miningindex erstellt werden, dieser liefert zusätzlich zur Häufigkeit eine in Echtzeit berechnete Abweichungsanalyse bzw. Auffälligkeits- und Korrelationsanalyse für alle Facetteneinträge.

5.4.4 Inhaltsanalyseminer

Am Ende stellt der Suchserver sowohl ein webbasiertes Frontend als auch eine REST API zur Verfügung, über die der gesamte Index durchsucht und ausgewertet werden kann. Um den Miningindex zu analysieren bietet der

Inhaltsanalyseminer eine sehr intuitive und im Verhältnis zur Komplexität der umfangreichen Analysemöglichkeiten einfach zu bedienende Oberfläche.

Abbildung 9: WEX/AC - Analysemodule im Inhaltsanalyseminer

Genauere Beschreibungen findet man bei Chen et al., (2014) und ibm.com, (2018c).

5.5 IBM WEX/AC Analysemodule

WEX stellt im „Inhaltsanalyseminer" Analysemodule über ein webbasiertes HTML5 Interface zur Verfügung. Es können bei speziellen Anforderungen oder Darstellungen, die nicht standardmäßig von WEX unterstützt werden auch eigens programmierte Widgets in die Weboberfläche eingefügt werden. Der Inhaltsanalyseminer ist multiuserfähig und unterstützt eine Userverwaltung mit ausgefeilten Security Modellen sowohl auf Collection als auch auf Dokumentenebene.

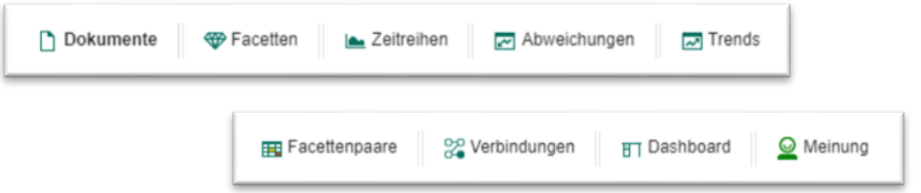


Abbildung 10: WEX/AC – Analysemodule

- *Dokumente* listet Dokumente mit Titel und eine intelligente Zusammenfassung basierend auf der aktuellen Suche auf
- *Facetten* listet die Inhalte einzelner Facetten mit Häufigkeit und Korrelation zur aktuellen Suchabfrage auf
- *Zeitreihen* zeigt Frequenzänderungen über eine Zeitperiode
- *Abweichungen* zeigt die Abweichung von Facetteninhalten über eine Zeitreihe
- *Trends* ermöglichen auffällige Zu- oder Abnahmen über eine Zeitperiode zu erkennen
- *Facettenpaare* zeigt eine zweidimensionale Facettenauflistung mit Häufigkeit und Korrelation
- *Verbindungen* visualisiert die Zusammenhänge zwischen Facettenwerten
- *Dashboard* zeigt Analysen in verschiedenen Charts und Tabellen und ganzen Analysemodulen, die nach Belieben konfigurierbar sind
- *Meinungen* zeigt den Sentiment der sich hinter Facetten und Textcontent befindet

5.5.1 Suche

Der über Webbrowser aufrufbare Inhaltsanalyseminer erlaubt den Zugriff für berechtigte Anwender auf freigegebene Objektgruppen. Die direkte Suche erfolgt über die Eingabe im Suchfeld. Der Suchsyntax ist in ein bis zwei Stunden geschult und lehnt sich an den Lucene Standard an, bietet aber noch diverse Erweiterungen. Die Ansicht „Dokumente“ zeigt eine Text-Zusammenfassung der Treffer mit Highlighting und gewählter Sortierreihenfolge (hier nach Relevanz). Die Relevanz wird von WEX aufgrund eines internen intelligenten Algorithmus auf Basis von der Anzahl

und der Position der Suchergebnisse im Text ermittelt und kann vom Anwender bei Bedarf beeinflusst werden. Auch andere Algorithmen zur Relevanzberechnung können herangezogen werden.

Die dahinter liegenden Dokumente sind verlinkt und können mit einem Klick heruntergeladen werden. Im Userinterface wird standardmäßig eine intelligente Zusammenfassung zu jedem Dokument angezeigt um ein effektives Querlesen zu ermöglichen. Diese Zusammenfassung ist dynamisch und wird auf Basis eines Algorithmus abhängig von der aktuellen Suche erstellt, um dem Anwender möglichst die Informationen auf einen Blick zu bieten, die bei der aktuellen Suche relevant sind. Die Zusammenfassung ist durch verschiedene Parameter an die Bedürfnisse des Ansenders anpassbar.

Der Watson Explorer unterstützt auch Natural Language Queries (NLQ). Dies erlaubt es dem Anwender, eine natürlich sprachliche Frage einzugeben und Dokumente, in denen diese Frage behandelt wird beziehungsweise Antworten vorkommen, retourniert zu bekommen. Dabei wird die Frage durch eine trainierbare NLQ Engine automatisch in eine WEX Query Syntax umgewandelt.

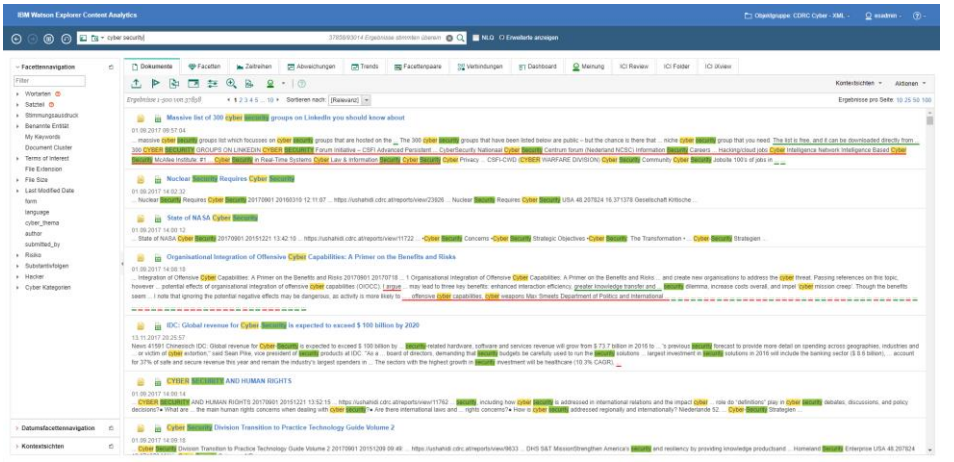


Abbildung 11: WEX/AC – Volltextsuche mit Treffer Highlighting und Sentiment

5.5.2 Erweiterte Suche

Die erweiterte Suche bietet eine grafische Oberfläche um Suchanfragen absetzten zu können ohne die WEX Suchsyntax zu kennen. Sie bietet also einen idealen Einstiegspunkt für unerfahrene Anwender um das System und die Suchfunktionen kennenzulernen. Alle Funktionen, die über die Oberfläche der erweiterten Suche abgedeckt werden, sind auch direkt über die Suchsyntax verfügbar.

In Abbildung 12 sind alle Prozessschritte aufgenommen und beschrieben. Alle Modelle werden den Anwendern zur Verfügung gestellt.

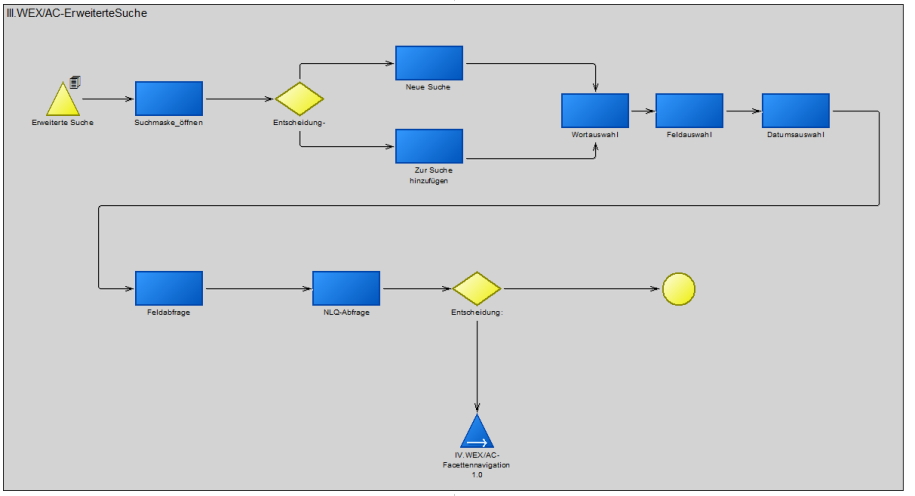


Abbildung 12: WEX/AC - Erweiterte Suche - Modell Suchprozess

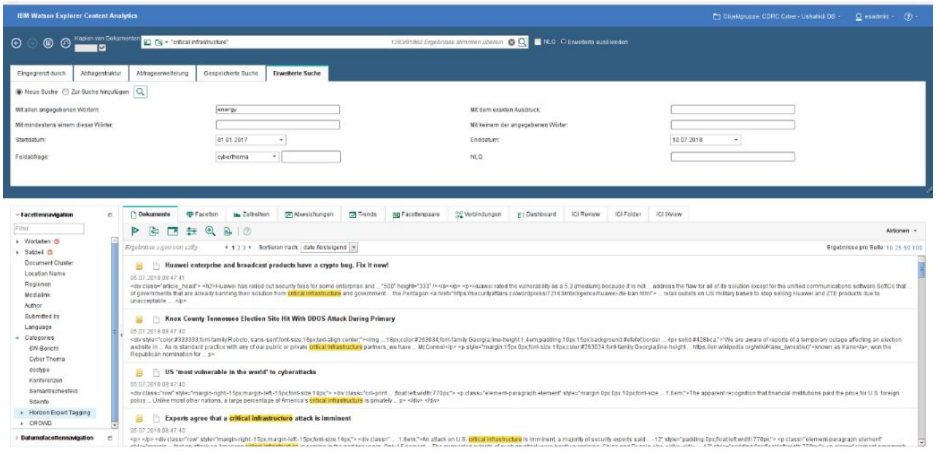


Abbildung 13: WEX/AC - Erweiterte Suche - Beispiel

5.5.3 Facetten

In Facetten werden sowohl die bei der Textanalyse annotierten (Character Rules, Dictionaries, Parsing Rules und Machine Learning) „Treffer“ aus Volltexten angezeigt, als auch die Treffer von später durch den Anwender hinzugefügten benutzerdefinierten Kategorien. Die Annotation erfolgt sowohl über IBM Standardregeln als auch über jene im Content Analytics Studio (CAS) oder Watson Knowledge Studio benutzerdefinierte Zeichen-, Wörterbücher- Parsing-, oder Machine Learning Regeln. Für die Annotation können sowohl die im CAS also auch über Watson Knowledge Studio (WKS) formulierten Regeln/Modelle herangezogen werden. Neben im Inhaltsanalyseminer standardmäßig bereit gestellter Facetten können auch benutzerdefiniert Facetten in Content Analytics Studio und Watson Knowledge Studio kombiniert und formuliert werden – die eigentliche Stärke von WEX/AC.
Benutzerdefinierte Kategorien werden auch als Facetten dargestellt. Dabei handelt es sich um Suchen, die der Anwender jederzeit hinzufügen kann und die dann analytisch zur Auswertung zur Verfügung stehen.

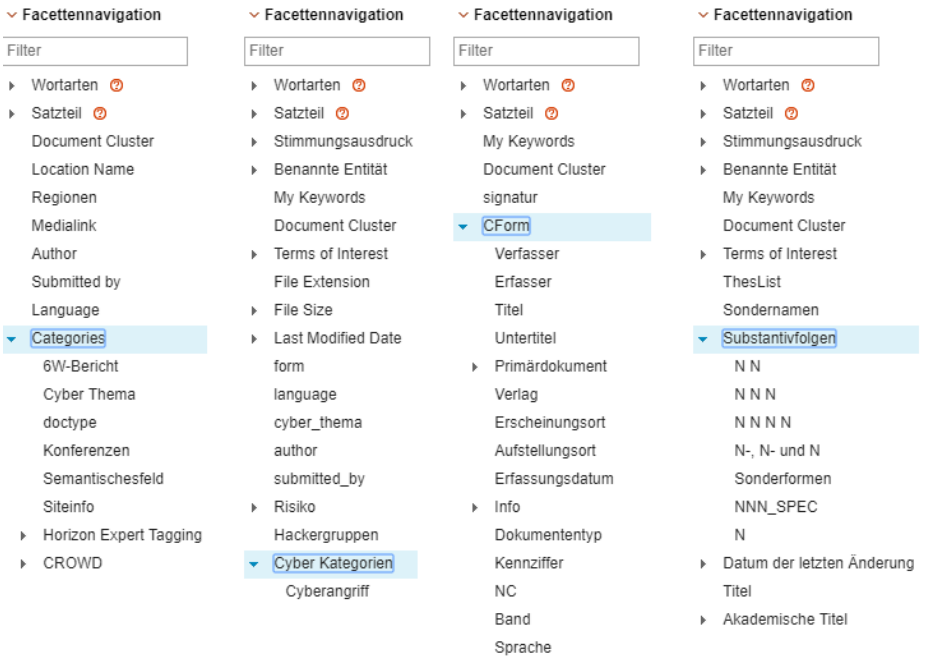


Abbildung 14: WEX/AC – Facettennavigation,
Beispiele unterschiedlicher Inhaltsanalyseminer-Objektgruppen

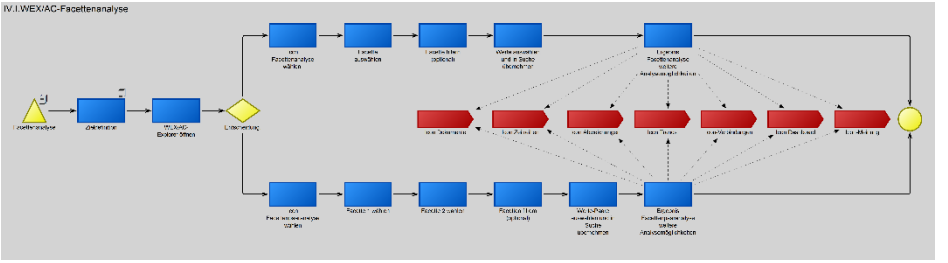


Abbildung 15: WEX/AC - Facettenanalyse - Modell

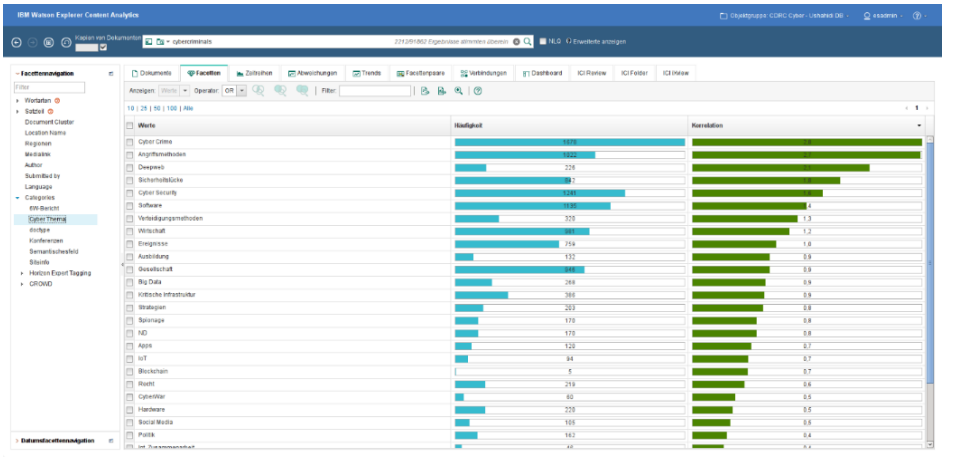


Abbildung 16: WEX/AC - Facettenanalyse - Beispiel

Die Facettenansicht zeigt die Einträge der in der Facettennavigation gewählten Facette. Neben der Häufigkeit ist auch die „Korrelation“ ein von IBM Watson Explorer aus dem Miningindex berechneter Wert, der die „Auffälligkeit“ des Facetteneintrages in der Abfragemenge durch Vergleich mit der Gesamtdokumentenmenge (Big Data Analytik) angibt (1 = Normwert).

5.5.4 Zeitreihen

Zeitreihen (Jahre, Monate, Tage, Tage der Woche, …) zeigen die Häufigkeit der Treffer in den Dokumenten bezogen auf die aktuelle Abfrage und der gewählten Datumsfacette an. Abweichungen zeigen die Häufigkeit und zusätzlich die „Auffälligkeit“ bezogen auf eine gewählte Facette. Die Reihenfolge der Diagramme kann über die Häufigkeit oder den Index (Auffälligkeitswert für die Zeitperiode) gesteuert werden.

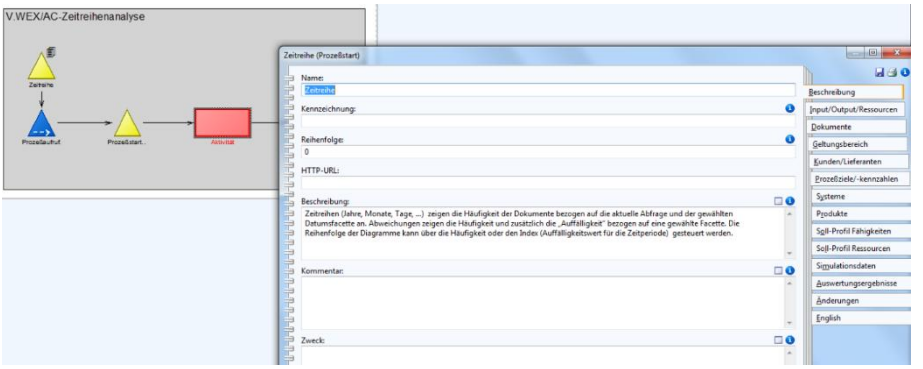


Abbildung 17: WEX/AC - Zeitreihenanalyse - Prozessbeschreibung

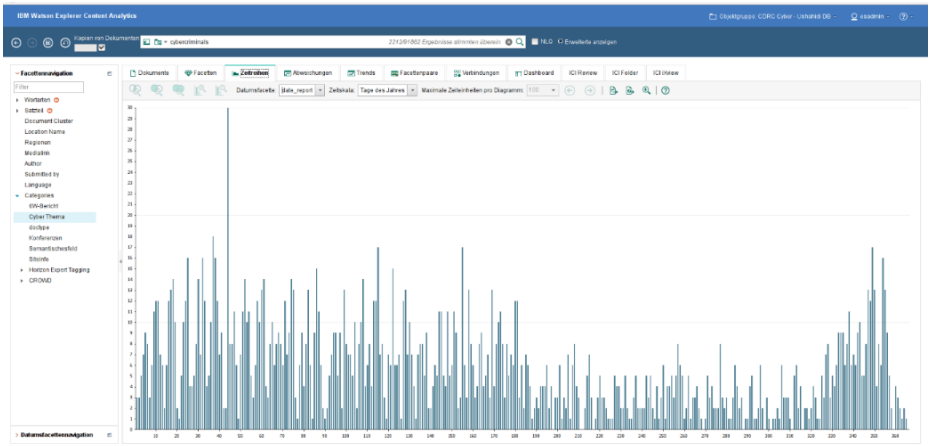


Abbildung 18: WEX/AC - Zeitreihenanalyse - Beispiel

5.5.5 Abweichungen

Diese Ansicht zeigt wie Facettenwerte über eine Zeitperiode abweichen (eine „Was ist passiert?“ Analyse). Dabei wird für jeden Facettenwert ein Diagramm erstellt, bei dem die Farbe der einzelnen Balken die Auffälligkeit anzeigt. Grau ist nicht auffällig, gelb, orange und rot zeigen Auffälligkeiten an.

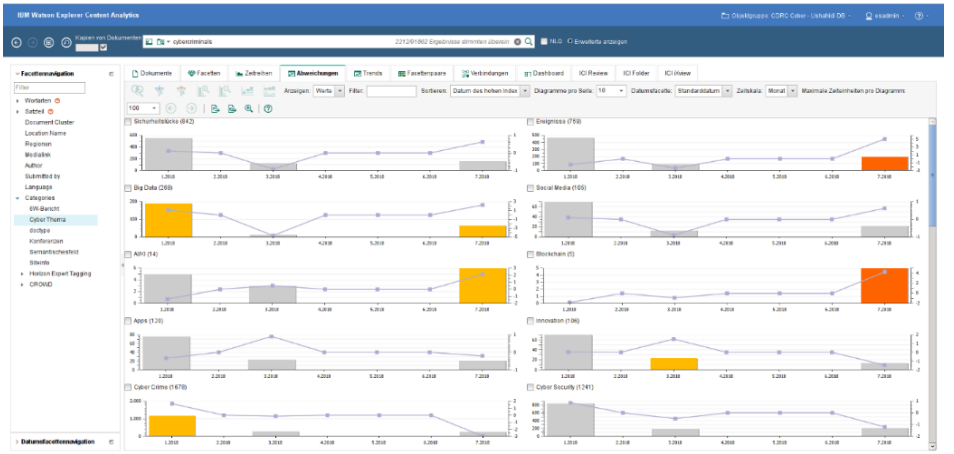


Abbildung 19: WEX/AC - Abweichungen - Beispiel

5.5.6 Trends

Trends zeigen die Häufigkeit und zusätzlich die „Auffälligkeit“ bezogen auf eine gewählte Facette. Die Reihenfolge der Diagramme kann über die Häufigkeit oder den Index (Auffälligkeitswert für die Zeitperiode im Vergleich zu den Vorperioden) gesteuert werden. Über die Sortierung „Letzter Index“ liefert die Trendanalyse hochwertige Ergebnisse im Bereich „Early Warning“ (FCC = Federal Communications Commission, ICO = Initial Coin Offering). Im Unterschied zum Abweichungen-Tab zeigt Trends die Auffälligkeiten nicht auf Basis der Gesamtdaten an, sondern auf Basis der letzten Vorperioden.

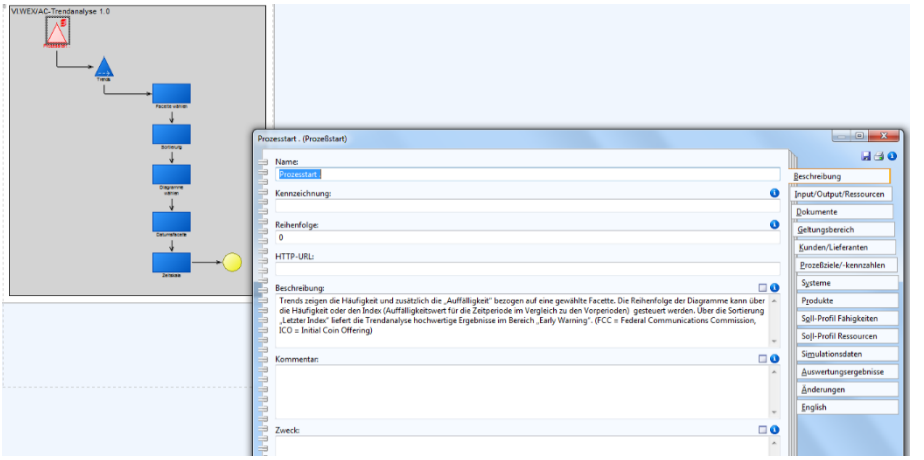


Abbildung 20: WEX/AC - Trendanalyse - Prozessbeschreibung

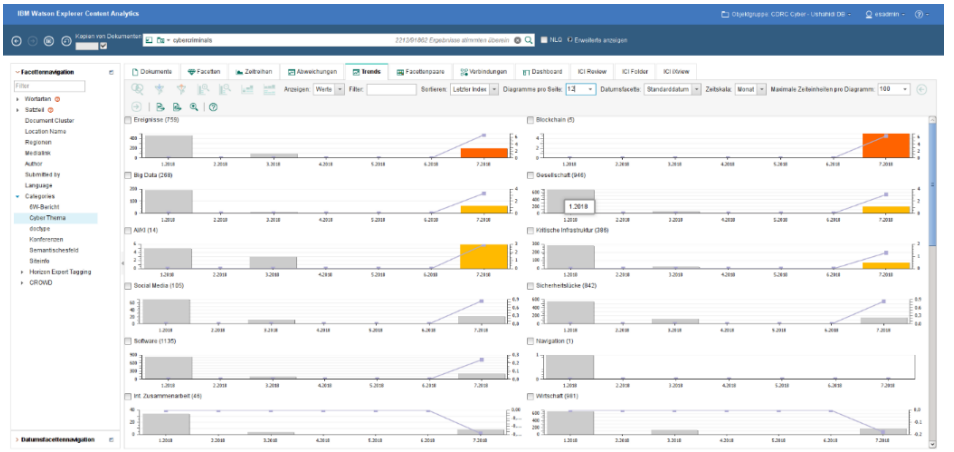


Abbildung 21: WEX/AC - Trendanalyse - Beispiel

5.5.7 Facettenpaare

Die Auswertung über „Facettenpaare“ zeigt das Potential der Vorarbeiten, die im Pre-Processing geleistet werden, eindrucksvoll auf. Nach Auswahl von zwei Facetten kann einerseits nach bestimmten Werten unterschiedlicher Facetten gefiltert, andererseits nach Häufigkeit oder Korrelation, also dem Zusammenhang dieser zwei Facettenwerte, sortiert werden. Mit einem Mausklick können sofort die Dokumente angezeigt werden, die den Zusammenhang dieser Facetten belegen.

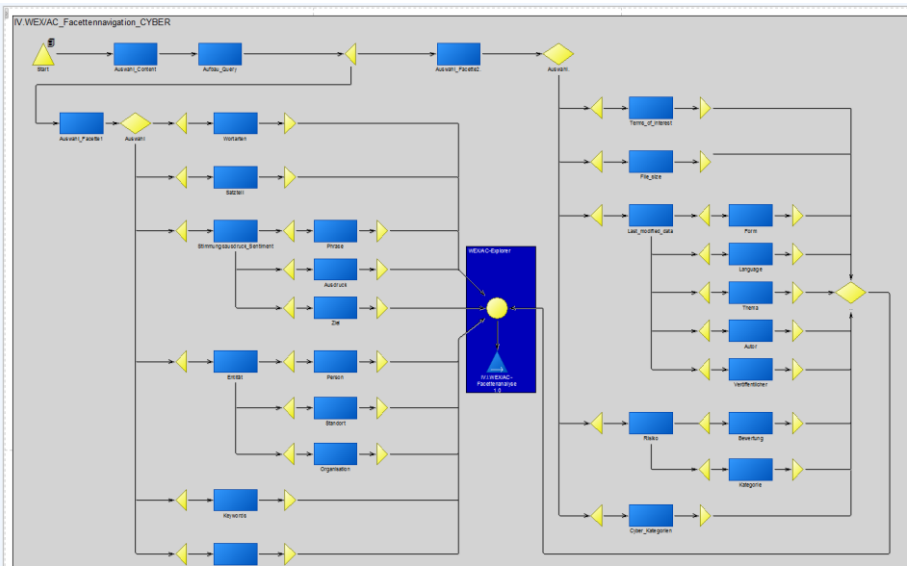


Abbildung 22: WEX/AC - Facettenpaaranalyse - Modell Beispiel „Cyber“

Hier wird der Wert des Modells besonders sichtbar. Die Kombinationsmöglichkeiten erreichen eine Komplexität, die nur über diese Visualisierung optimal genutzt werden können.

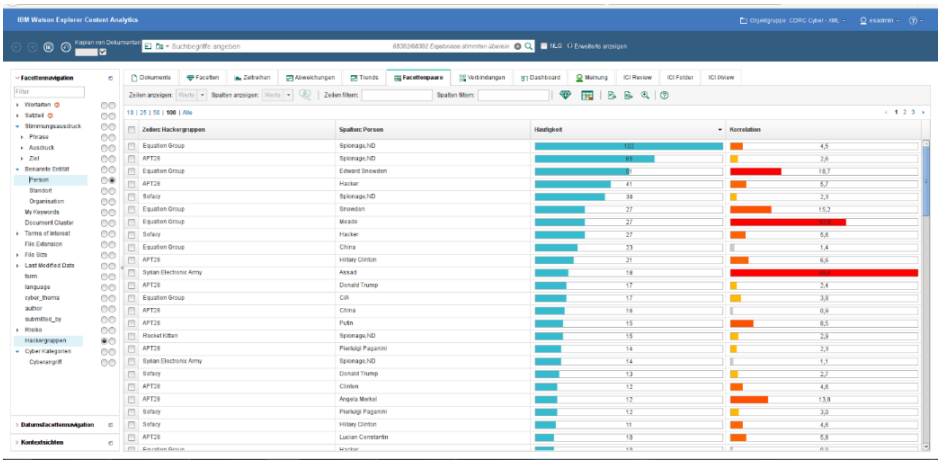


Abbildung 23: WEX/AC - Facettenpaaranalyse - Beispiel

Die Anwendungsmöglichkeiten am Beispiel „Cyber“ erlangen bereits Komplexitäten, wo beispielsweise die Auswertung der Korrelation der Facetten „Themen“ und „Ereignisse“ das Potential eindrucksvoll veranschaulichen.

5.5.8 Verbindungen

Verbindungen zeigen graphisch Zusammenhänge, Häufigkeiten und Auffälligkeiten zwischen zwei beliebigen Facetten bezogen auf die aktuelle Abfrage. Diese Visualisierung ergänzt die Facettenanalyse somit, um die Qualität der Analyse weiter zu verbessern.

Weiterführende Netzwerkanalysen oder Visualisierungen können bei Bedarf durch vorhandene Schnittstellen zu IBM i2 EIA und IBM Analyst‘s Notebook durchgeführt werden.

Das Dokument selbst oder beliebige Facetteneinträge werden zu Entitäten und die Häufigkeit oder Korrelation wird zum Link für i2 in der ELP Modellierung.

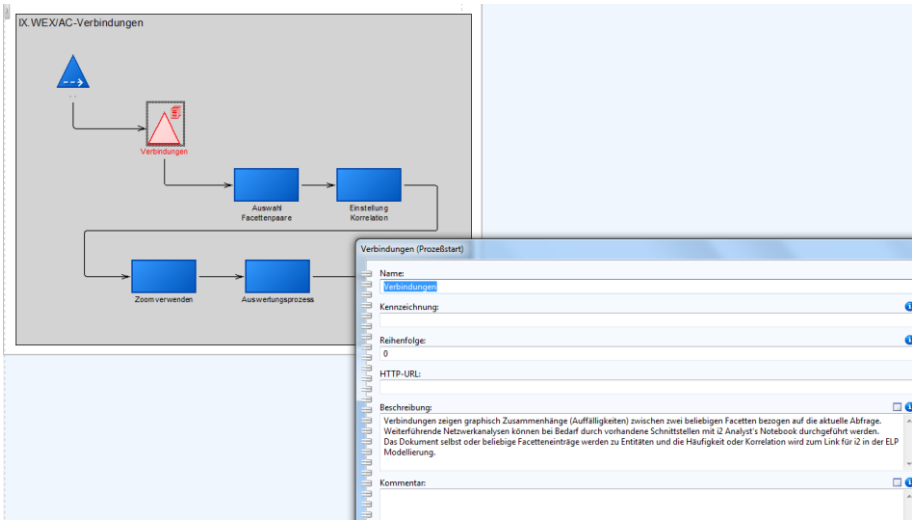


Abbildung 24: WEX/AC - Verbindungen - Prozessbeschreibung

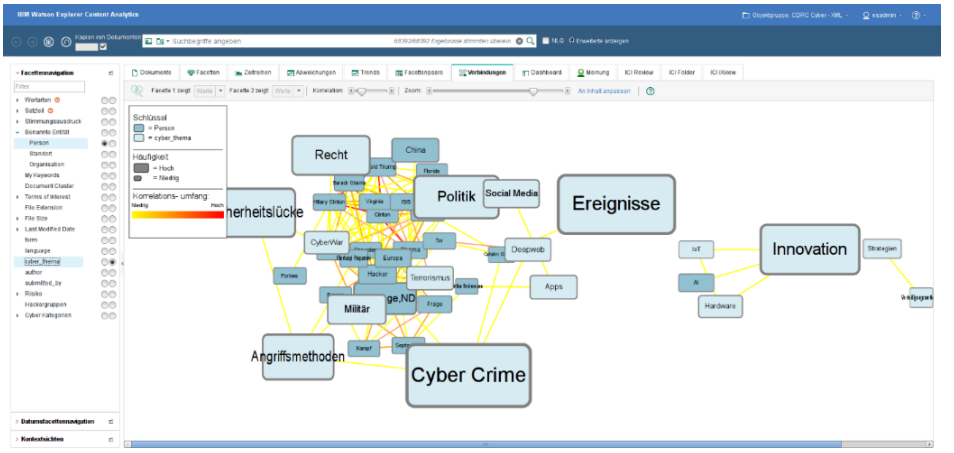


Abbildung 25: WEX/AC - Verbindungen – Beispiel

Das in Abbildung 25 gezeigte Beispiel veranschaulicht, dass die CDFZ Daten belegen, dass IoT, AI und Hardware mit Innovation zusammenhängen. Eine durchaus logisch nachvollziehbare Auswertung.

5.5.9 Dashboard

Dashboards ermöglichen sinnvolle Analysen und Auswertungen in frei wählbaren Darstellungsformen (Diagrammen) als fertige Informationssichten bereitzustellen. Die Anpassung (Entwicklung) von Dashboards kann von Power-Usern (Analysten) selbstständig über die grafische Mineroberfläche erfolgen.
Der Inhalt der Dashboards kann interaktiv (Suchabfrage / Facettenauswahl) aber auch qualitätsgesichert mit vorgebenden Parametern gefüllt werden. Dashboards bieten eine hervorragende Möglichkeit, die im Watson Explorer ausgewerteten Informationen schnell und übersichtlich aufzubereiten, um auf einen Blick eine Übersicht über die Daten zu bekommen.

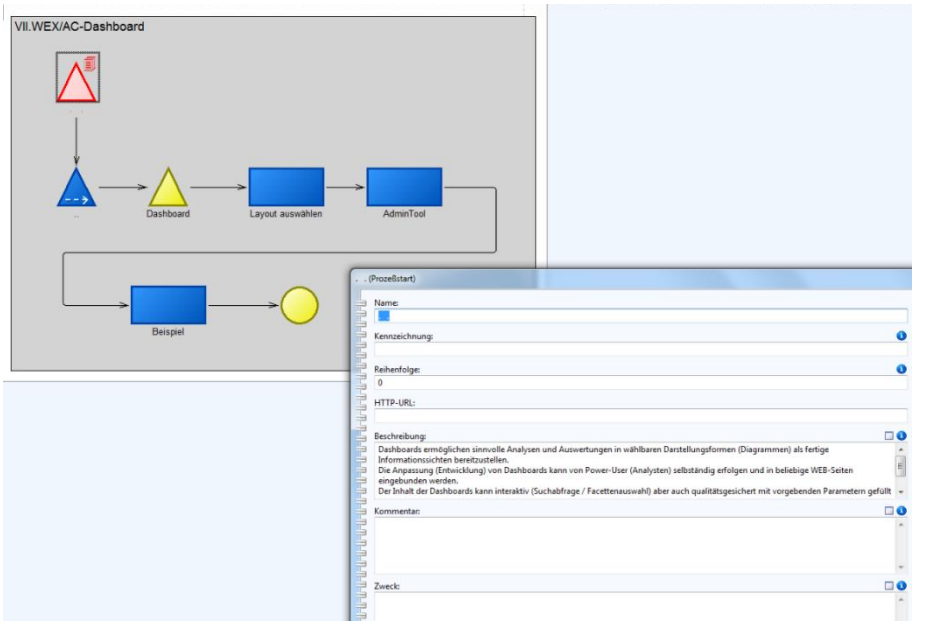


Abbildung 26: WEX/AC - Dashboard - Prozessbeschreibung

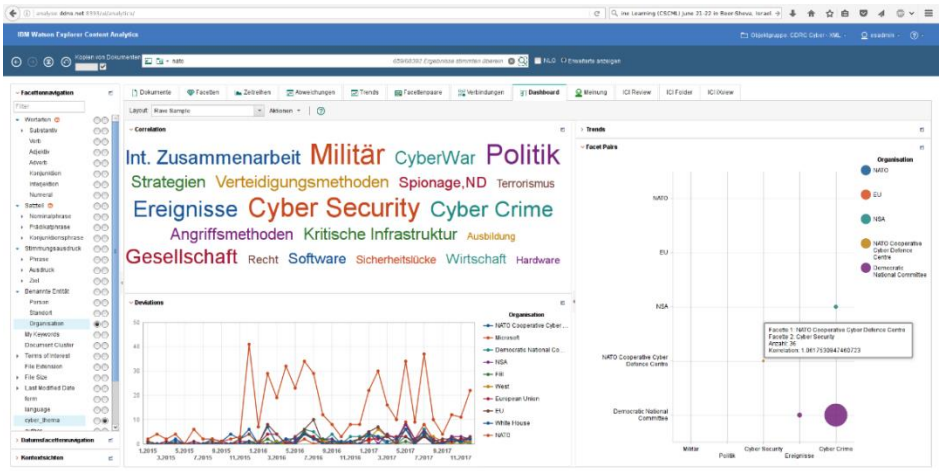


Abbildung 27: WEX/AC - Dashboard – Beispiel

5.5.10 Meinungen (Sentiment)

Die Sentiment-Analyse ermöglicht es, Dokumenteninhalte auf positive, negative und neutrale Wörter und Phrasen hin zu untersuchen. Watson Explorer erkennt dabei nicht nur das Sentiment selbst, sondern auch das Ziel, auf welches es sich bezieht. Dies ermöglicht das effiziente Analysieren von Stimmungen und Meinungen, die in Texten ausgedrückt werden beispielsweise für eine Voice of the Customer Analyse (VoC). Im Inhaltsanalyseminer werden in der Dokumentenübersicht positive und negative Tokens farblich markiert (rot und grün unterstrichen). Es steht auch ein eigenes Widget zur Verfügung, in dem eine genaue Analyse durchgeführt werden kann.
Der Watson Explorer unterstützt die Sentiment-Analyse für diverse Sprachen, es ist allerdings auch möglich, die Sentiment-Analyse im Content Analytics Studio nach eigenen Bedürfnissen anzupassen und für neue Sprachen selber zu entwickeln.

5.5.11 Berichte

Der Inhaltsanalyseminer bietet eine Schnittstelle zu Cognos Analytics, einem Business Intelligence und Reporting Tool. Durch diese ist es möglich, Berichte von Analyseergebnissen zu erstellen, statistisch zu analysieren und zu verteilen. Der Watson Explorer stellt einige Standardberichte zur Verfügung, es ist aber möglich auch eigene Berichte ganz nach Bedarf im Cognos Report Studio zu modellieren.
Der Watson Explorer bietet auch die Möglichkeit Facettenwerte in ein Datawarehouse in Form eines Starschema zu exportieren und darauf einen Cognos Cube aufzubauen um analytische Auswertungen zu tätigen und sowohl strukturierte Daten als auch unstrukturierte Daten aus einer gemeinsamen Sicht betrachten zu können.

5.6 Content Analytics Studio

IBM Watson Explorer Content Analytics Studio ist Bestandteil des Watson Explorers. Die lokale Windows-Entwicklungsumgebung auf Eclipse Basis dient zum Erstellen und Testen von benutzerdefinierten Textanalyse-Engines für die Extraktion von Entitäten:

- Sprach- und domänenspezifische Begriffe in Wörterbüchern (Dictonary Rules)
- Zeichenregeln und Zeichenmuster (Character Rules)
- Semantische Regeln (es werden diverse Standards unterstützt wie zum Beispiel Turtle, N-Triples, TriG, N-Quads, N3, RDF, OWL oder SKOS)
- Parsing-Regeln für Textmuster (Parsing Rules)
- Break Rules um die Tokenization zu beeinflussen (Break Rules)

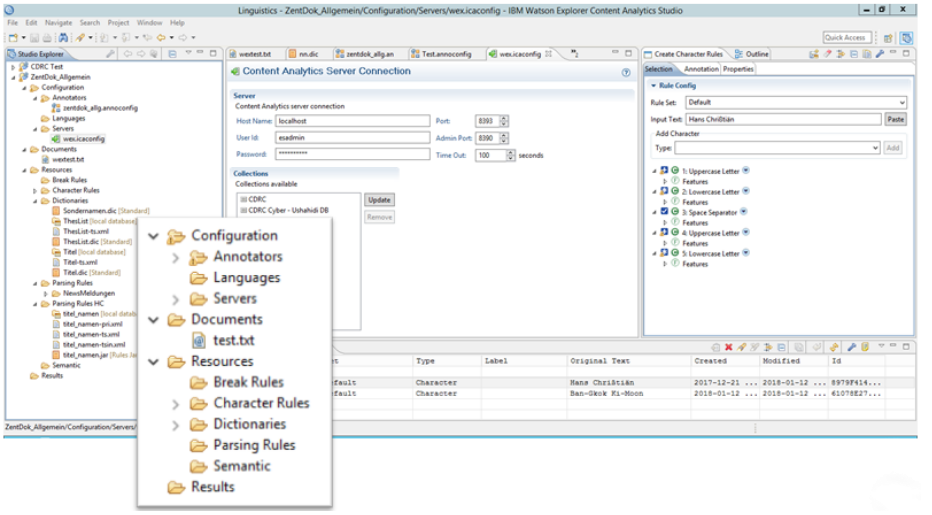


Abbildung 28: Content Analytics Studio Beispiel

Da für die Modellierung der Regeln der Open Source Standard UIMA (Unstructured Information Management Architecture) verwendet wird können allerdings auch andere Entwicklungsumgebungen verwendet werden, die es erlauben, Regeln in Form eines PEAR Files zu erstellen. Das Content Analytics Studio bietet aber gegenüber anderen Softwareprodukten vor allem den Vorteil, dass es eine Verbindung zur Watson Explorer Instanz ermöglicht. Dadurch können zum Beispiel Textdokumente zur Regelerstellung oder zum Testen von Regeln von einer WEX Collection heruntergeladen oder auch fertige Regeln direkt im Watson Explorer bereitgestellt werden. Regeln die mit dem Content Analytics Studio erstellt worden sind können mit allen Softwareprodukten verwendet werden, die den UIMA Standard unterstützen. Die Verwendung der Regeln beschränkt sich also nicht nur auf den Watson Explorer.

Die Regelerstellung mit dem Content Analytics Studio erfolgt komplett über eine grafische Oberfläche. Programmierkenntnisse sind keine erforderlich. Es ist also möglich einen Fachanwender auf die Regelerstellung einzuschulen um sein Domänenwissen effizient in die Analyse miteinfließen zu lassen. Für Spezialfälle besteht auch die Möglichkeit Programmierlogik in Regeln über das Content Analytics Studio unterzubringen (zum Beispiel Prüfsummenberechnungen oder

Umrechnung und Normalisierung von Währungen). Die Regelerstellung bzw. Annotationserstellung sind ein essentieller Teil der erheblich zur Analysequalität beiträgt.

5.6.1 UIMA Pipeline

Die Unstructured Information Management Architecture ist ein OASIS Standard zur Bearbeitung von Textinhalten der ursprünglich von IBM entwickelt wurde. Es gibt verschiedene Implementierungen dieses Standards, eine der bekanntesten ist wohl Apache UIMA, ein Projekt der Apache Software Foundation. UIMA bietet Möglichkeiten große Volumina an Text-Content analytisch auszuwerten. Dazu werden viele Funktionen und Möglichkeiten beschrieben. In dieser Publikation wird aber nur das Thema „UIMA Pipeline" behandelt, da dieses im Zusammenhang mit dem Watson Explorer und dem Content Analytics Studio besondere Bedeutung hat.
Die UIMA Pipeline besteht aus mehreren Annotatoren die hintereinander vom Watson Explorer beim Analysieren von Text automatisch abgearbeitet werden. Man kann diese Annotatoren grob in 4 aufeinander aufbauende Gruppen einteilen (Ferrucci, D. et al., 2009):

- Document language: hier wird bestimmt, in welcher Sprache der zu analysierende Text vorliegt. Nur mit dieser Information können in späteren Schritten die richtigen Regeln zum Parsen und zur Entity Extraktion ausgewählt werden.
- Lexical Analysis: in dieser Phase passiert die Tokenization. Der Text wird also in Absätze, Sätze, und Tokens geteilt. Ein weiterer Annotator teilt den einzelnen Tokens dann grammatikalische Kategorien wie zum Beispiel Nomen, Verb, Präposition, etc. zu. Dies wird auch Part of Speech (PoS) Analyse genannt. Danach werden alle im Content Analytics Studio definierten Dictionary und Charakter Rules ausgeführt um Entitäten zu extrahieren.
- Parsing Rules: In diesem Schritt werden nun aufbauend auf die Lexical Analysis die im Content Analytics Studio definierten Parsing Rules abgearbeitet.
- Clean Up: In diesem letzten Schritt werden nicht mehr benötigte oder temporäre Annotationen aus dem finalen Ergebnis entfernt.

Obwohl diese Schritte alle automatisch, ohne Zutun oder Konfiguration des Anwenders, passieren, ist es doch für das Verständnis der Watson Explorer Technologie wichtig und wurde deshalb hier kurz beschrieben. Die „Unsturctured Information Management Architecture“ ist ein sehr komplexes und tief technisches Thema. Für genauere Informationen sei hier auf die Apache UIMA Seite www.uima.apache.org und auf das IBM Knowledge Center verwiesen. Der Watson Explorer ermöglicht es, diese komplexe Technologie auch nicht IT affinen Anwendern zur Verfügung zu stellen.

6 IBM i2

IBM i2 ist eine umfangreiche Software Suite (Client- und Serverkomponenten) für die Analyse und Visualisierung von Datenbeständen, die in einem Entity–Link–Property Modell erfasst sind oder in ein solches übergeführt werden. IBM i2 bietet komplexe Algorithmen zur Analyse von Netzwerken oder auch zur Erkennung ähnlicher bzw. gleicher Entitäten. Im Wesentlichen resultieren daraus mehr oder weniger komplexe Netzwerkanalysen (Pfadsuche, verlinkte Elemente, Cluster, K-Core, Betweenness, Closeness, Degree, Eigenvektor, etc.) und „intelligente" Netzwerkdiagramme (Fächer, Hierarchien, Organisationen, etc.) bis hin zur Anordnung der Entitäten auf der Zeitachse oder auf Kartenmaterial (über ESRI Connector). Eine Einführung in die Grundlagen der Netzwerkanalyse kann bei Göllner et al., (2011) gefunden werden.

Analyseergebnisse aus WEX/AC können über eine Schnittstelle an IBM i2 übergeben und dort weiter analysiert und in Analyst's Notebook in diversen Diagrammen visualisiert werden. Analyst's Notebook kann mit IBM i2 Analyze sowie IBM i2 Enterprise Insight Analysis interagieren und ist somit für Big Data Anwendungen geeignet.

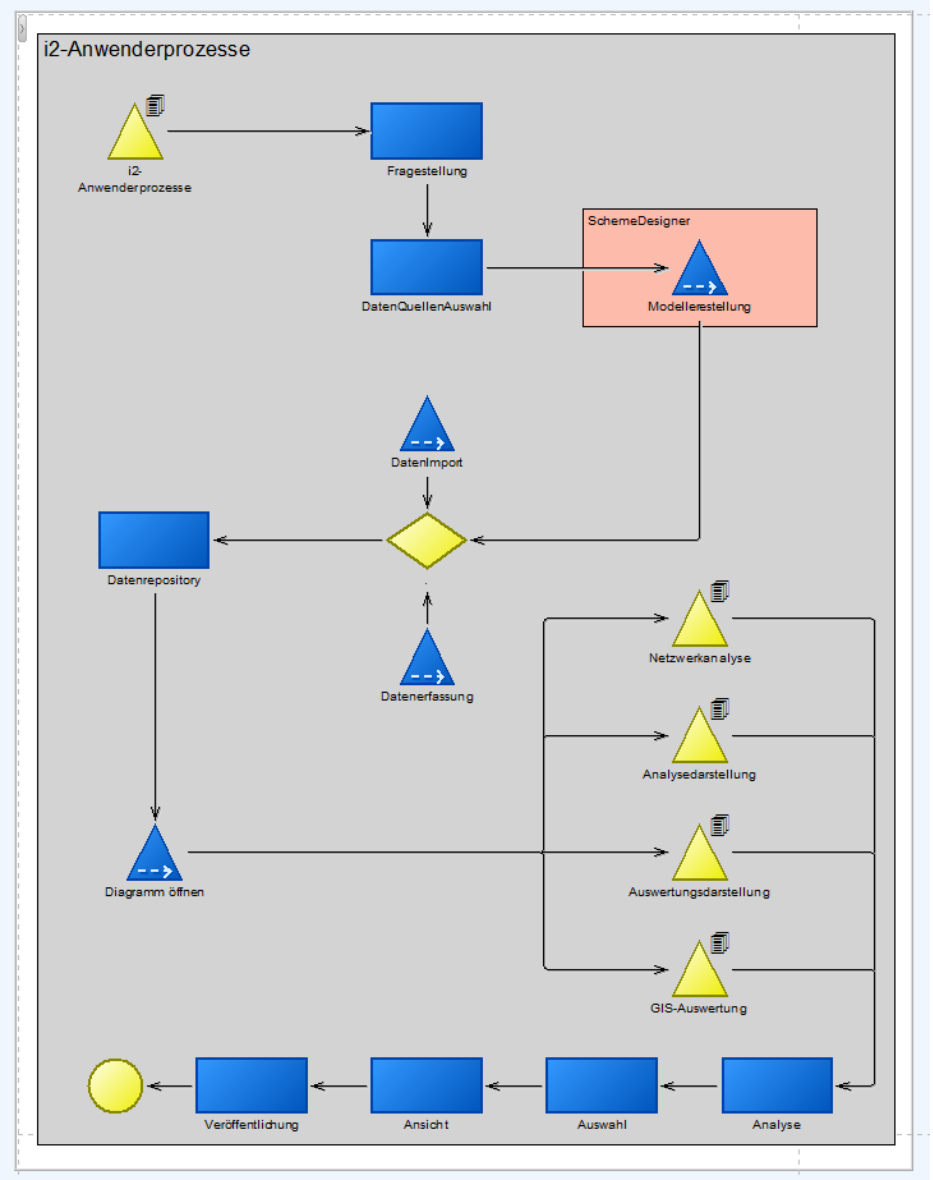


Abbildung 29: i2 Analyze - Modell Anwenderprozess

6.1 IBM i2 Analyst's Notebook Premium

Das folgende Kapitel basiert auf ibm.com, (2018d) und ibm.com, (2018b). Analyst's Notebook (ANB) ist ein Fat Client (Windows Programm) für die Modellierung, Erfassung (manuell oder per Datenimport), Visualisierung und Analyse von Informationen auf Basis von Entity-Link-Property (ELP) Modellen (Client Modellierungswerkzeug: i2 Analyze Schema Designer). Die ELP Modellierung wird genauer bei IBM, (2015) beschrieben. ANB ist ein Werkzeug für Spezialisten im Umfeld umfangreicher OSINT-Analysen, welche unter anderem dem an der ZentDok entwickelten Doppelvektoren-Modell folgen.

Dabei können ein und dieselben Daten auf sehr unterschiedliche Arten analysiert und visualisiert werden, wie die folgenden Beispiele in den nächsten Kapiteln zeigen. Durch die Möglichkeiten

- Netzwerkanalyse
- Zeitreihenanalyse
- Histogramme, Aktivitäten, Heat Maps
- Georeferenzierung

nicht nur zu berechnen, sondern auch mit den jeweils anderen Ergebnissen zu verschneiden, ergibt sich eine Vielzahl an analytischen Varianten, die auch komplexe OSINT Analysen zulassen.

IBM i2 Analyst's Notebook Premium hat zusätzlich zu den Funktionen des IBM i2 Analyst's Notebook die Möglichkeit auf ein lokales Repository, auf ein i2 Analyze Group Repository oder auf einen Information Store zuzugreifen und von dort Daten zu beziehen beziehungsweise diese auszuwerten.

6.1.1 Netzwerkanalysen

In der (sozialen) Netzwerkanalyse werden Entitäten und deren Relationen in i2 in einem ELP Schema erfasst und können, wie Abbildung 30 zeigt, in einem Graphen dargestellt werden (ibm.com, 2018e).

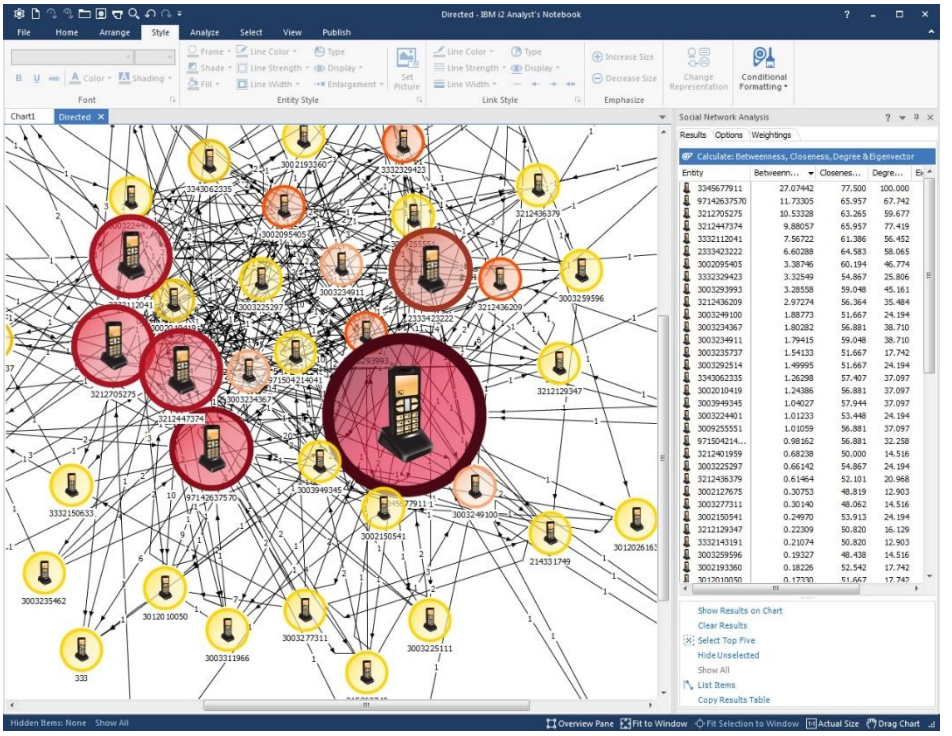


Abbildung 30: i2 Analyst's Notebook - Beispiel Netzwerkanalyse

Im folgenden Beispiel wird der Thesaurus der ZentDok visualisiert. Die Visualisierung der i2 Netzwerkanalyse ermöglicht in diesem Fall Zusammenhänge der Terminologie und der Translationsarbeit der Fachinformationsdatenbank zu erkennen. Abbildung 31 zeigt den Ausschnitt „Internet“ mit dem interessanten Ergebnis, dass sich das Internet als Führungsinstrument darstellt.

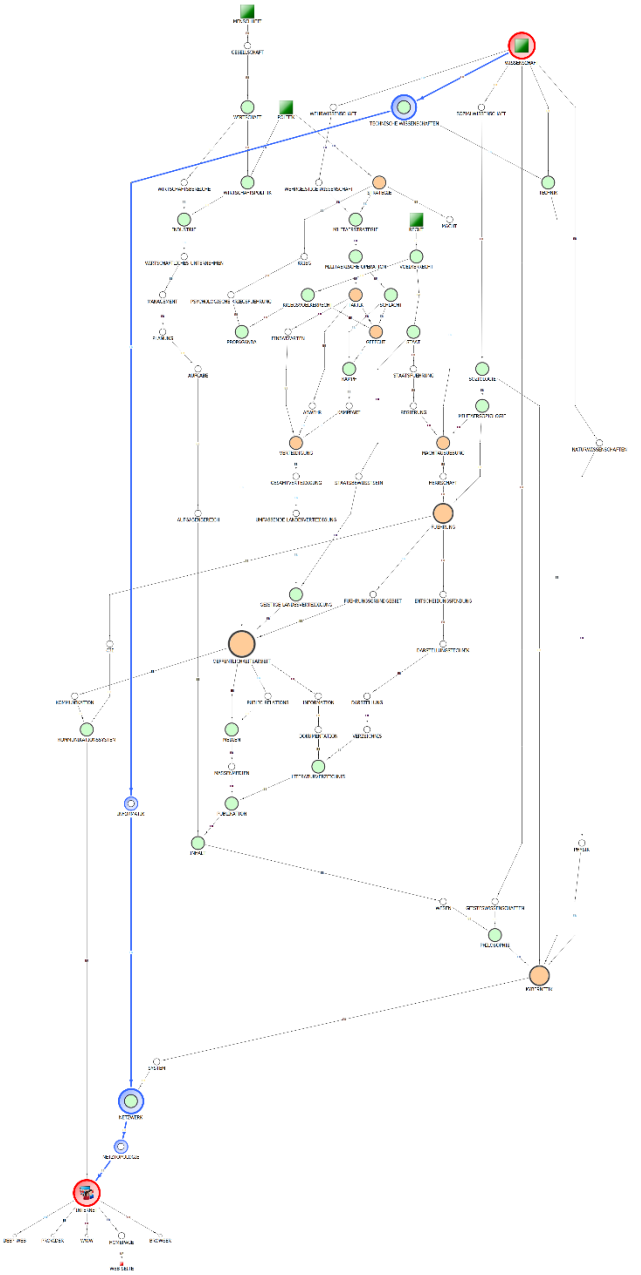


Abbildung 31: i2 Analyst's Notebook - Beispiel Netzwerkanalyse ZentDok Thesaurus Ausschnitt "Internet"

6.1.2 Zeitreihenanalysen

Zeitreihenanalysen werden immer dann verwendet, wenn die Visualisierung eines zeitlichen Ablaufes oder Verlaufes einen zusätzlichen Erkenntnisgewinn verspricht. Abbildung 32 zeigt ein Beispiel einer Ereignisdarstellung auf der Zeitachse, wobei das Ereignis selbst als Entität modelliert ist (ibm.com, 2018).

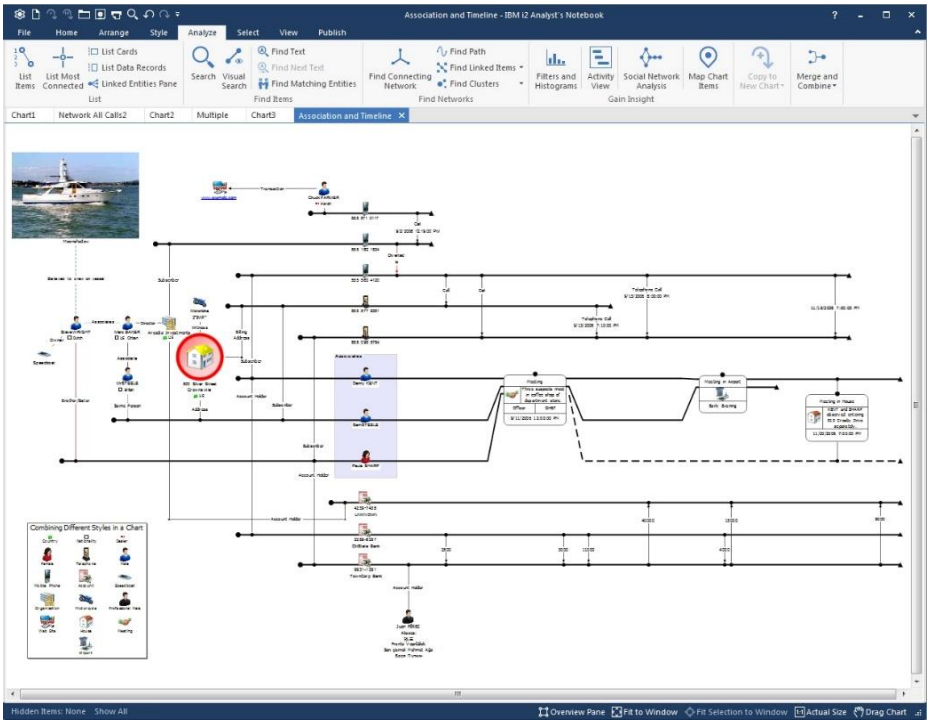


Abbildung 32: i2 Analyst's Notebook - Beispiel Zeitreihenanalyse

Diese Darstellung eignet sich gut, um die Historie und mögliche zukünftige Entwicklungen um ein Ereignis herum darzustellen. Gerade für die taktische aber vor allem auch für strategische Analysen ist die Darstellung im Zeitverlauf eine der wichtigsten Visualisierungen. Die im Foresight häufig verwendeten Darstellungen von Horizon Scanning Analysen beziehen sich auf eine zeitliche Einordung der Ergebnisse aus unterschiedlichen Quellen. Mit einer zeitlichen Zuordnung der Quellen

lassen sich robuste Aussagen über mögliche zukünftige Entwicklungen machen, wenn bekannt ist, wie lange die Wissensdiffusion von einer zur nächsten Quelle dauert.

6.1.3 Histogramme und Heatmaps

Histogramme und Heatmaps sind eine simple aber effiziente Methode, um einen schnellen Überblick über die Häufigkeitsverteilungen innerhalb der zu analysierenden Daten zu bekommen. Abbildung 33 zeigt das Beispiel einer Auswertung von Netzwerkprotokollen. In diesem sind die Häufigkeiten von Verbindungen beispielsweise zusätzlich als Histogramm und Heatmap visualisiert (ibm.com, 2018e).

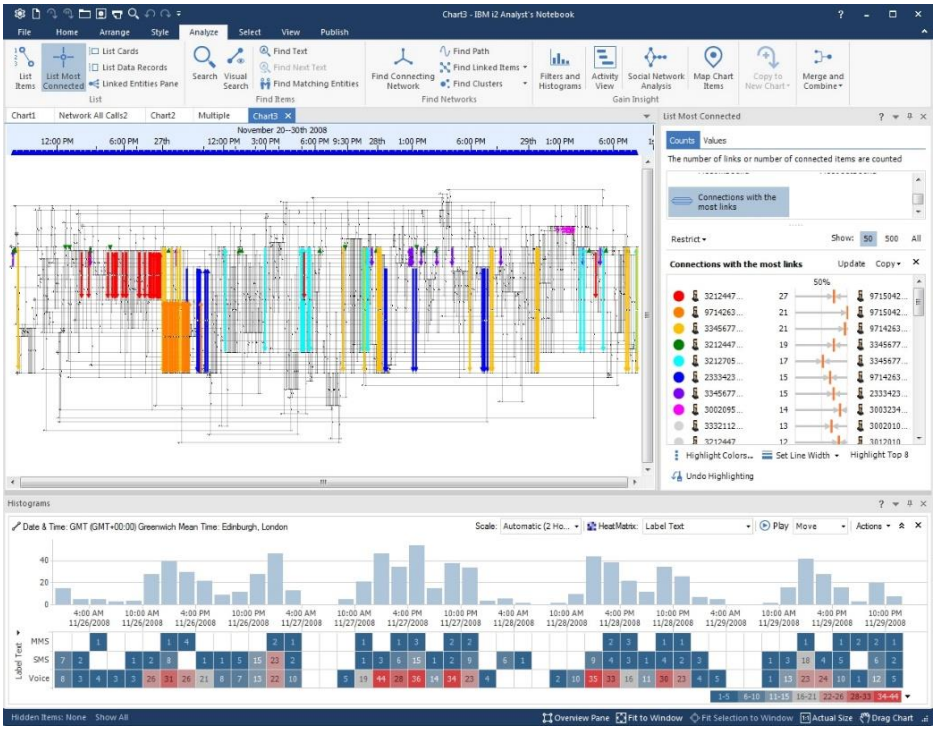


Abbildung 33: i2 Analyst's Notebook - Beispiel Histogramme und Aktivitäten

Histogramme können zudem verwendet werden, um Daten bezüglich ihrer Häufigkeit zu gewichten, womit eine Stärke von Analyst Notebook zum

Tragen kommt. Alle einmal eingepflegten Daten und alle daraus errechneten analytischen Parameter lassen sich in weiteren Analysen verknüpfen. In Bezug auf Histogramm Daten heißt das, dass nachfolgende Analysen z.B. mit der Häufigkeit in der Grundgesamtheit gewichtet werden können, welches wichtig ist, um mit veränderlichen Grundgesamtheiten zu belastbaren Aussagen zu kommen.

6.1.4 Georeferenzierung

Mit der Funktion der Georeferenzierung ist es möglich Daten auf einer Karte darzustellen. Zusammen mit den anderen schon erklärten Funktionen wird deutlich, wie die unterschiedlichen funktionalen Analysen zu einem gemeinsamen Bild beitragen, welches durch die Georeferenzierung besonders nutzbringend ist. Wie Abbildung 34 zeigt, können Entitäten auch über das i2 ESRI Plugin (GIS mapping Software, ArcGIS) als GIS Auswertung auf Kartenmaterial georeferenziert werden. Dabei können auch Entitäten bezogen auf Umkreis und Pfadbereich vom System automatisch selektiert werden (ibm.com, 2018e).

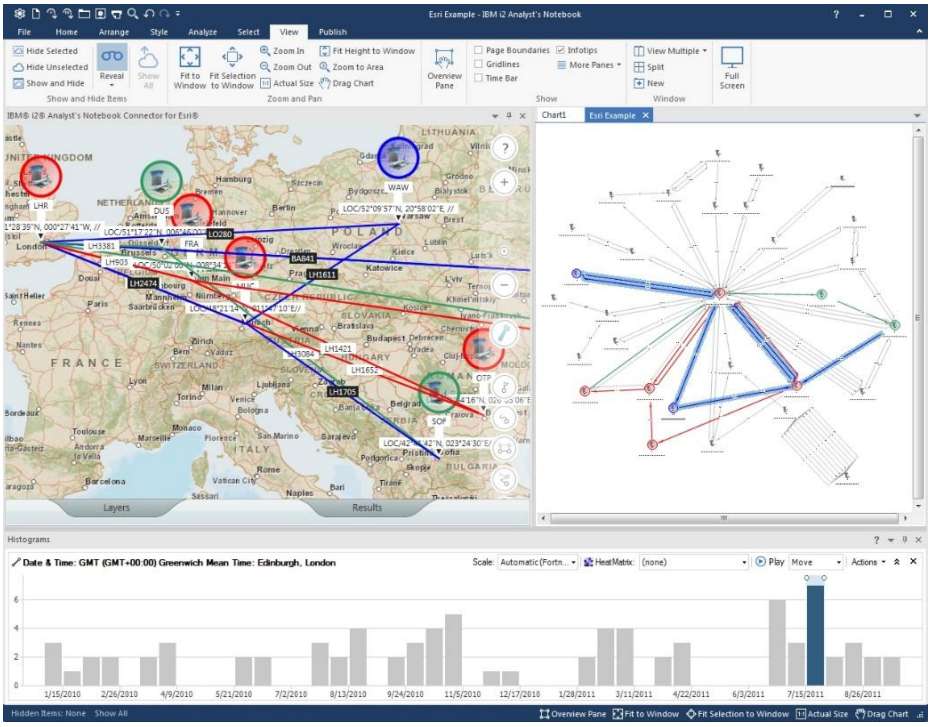


Abbildung 34: i2 Analyst's Notebook - Beispiel Georeferenzierung

Damit lassen sich nicht nur einzelne Datenpunkte, sondern z.B. auch die Ergebnisse der Netzwerkanalyse auf einer Karte darstellen. Räumliche Zusammenhänge werden so direkt ersichtlich.

Durch die Kombination der unterschiedlichen Funktionen, die hier vorgestellt wurden, lassen sich nun räumliche Zusammenhänge, zeitliche Zusammenhänge und inhaltliche Zusammenhänge in einer graphischen Visualisierung kombiniert darstellen.

6.2 IBM i2 Analyze

Bei IBM i2 Analyze handelt es sich um zahlreiche Serverkomponenten für die zentrale Erfassung und Verwaltung (Schema, Datenquellanbindungen, Zugriffe und Sicherheit) von Inhalten in einem i2 Analysis Group Repository (zentrale Datenbank für die Speicherung von Entitäten, Links und Diagrammen). Für die Verwaltung, Datenerfassung und Suche steht

ein WEB-Frontend, das Intelligence Portal, zur Verfügung. Weitere Informationen zu i2 Analyze können unter ibm.com, (2018a) und ibm.com, (2018b) gefunden werden.

6.3 IBM i2 Enterprise Insight Analysis (EIA)

IBM i2 Enterprise Insight Analysis ist ein Software Bundle aus i2 Analyze plus weiterer Software zum Datenhandling. Es werden auch fertige Modelle (zum Beispiel für den Defence oder Cybercrime Bereich) mit ausgeliefert. EIA erweitert die Datenhaltung von i2 mit dem IBM Information Store (DB2). Dieser kann mit den bei EIA mitgelieferten Software Komponenten auch „Massendaten" (Big Data) laden und diese über ein Web Frontend mit entsprechenden Filtern zur Weiterverarbeitung in Analyst's Notebook bereitstellen. Weitere Informationen hierzu können unter ibm.com, (2018a) gefunden werden.

6.4 Intelligence Portal

Das „Intelligence Portal" enthält Funktionen zum Suchen, Organisieren und Analysieren von Informationen und unterstützt die analytischen Aufgaben. Neben dem Erstellen und Ändern von Objekten kann nach interessanten Objekten gesucht und die Verbindungen zwischen diesen und anderen Objekten visualisiert werden. Intelligence Portal kann von den i2 Produkten Analyst's Notebook Premium oder i2 Analyze aufgerufen werden. Bei großen Datenmengen wird das Intelligence Portal zur Vor-Filterung für eine Weiterverarbeitung in Analyst's Notebook verwendet.

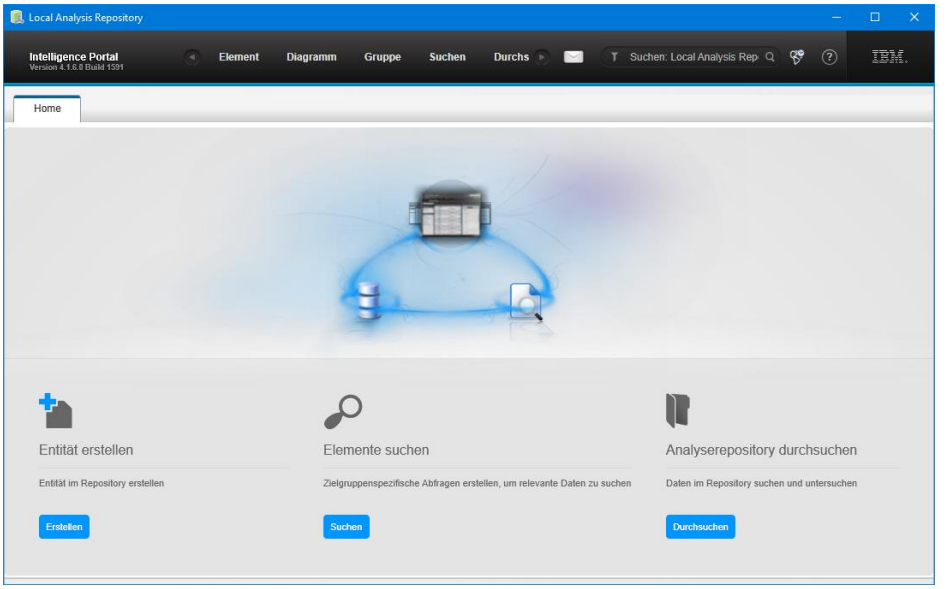


Abbildung 35: i2 Intelligence Portal

6.5 Szenarien für die Datenhaltung

Das i2 Software Paket bietet viele Möglichkeiten um Daten abzuspeichern und Anwendern zur Verfügung zu stellen. Im folgenden Kapitel soll ein Überblick über diese gegeben werden. Allgemein kann gesagt werden, dass die Art der Datenhaltung sehr stark von den Fragestellungen und Anforderungen, die Anwender an das System stellen, abhängt. Eine „beste“ Version gibt es hier nicht. Für die folgenden vorgestellten Pattern werden jeweils verschiedene Komponenten von i2 verwendet, eine Kombination der gezeigten Szenarien ist ebenfalls möglich.

6.5.1 IBM i2 ANBP mit Local Analysis Repository (LAR)

SZENARIO 1: Bei diesem Szenario werden keine Serverkomponenten benötigt. Die Datenhaltung erfolgt ausschließlich offline am Client. Die Charts werden im Local Analysis Repository oder direkt im anb-File gespeichert. Diese Architektur ist der einfachste und schnellste Weg mit dem Analyst's Notebook zu arbeiten. Bei großen Datenmengen stößt man allerdings hier schnell an die Grenzen. Auch Kollaboration zwischen mehreren Anwendern wird durch andere Architekturen, die später vorgestellt werden, besser unterstützt.

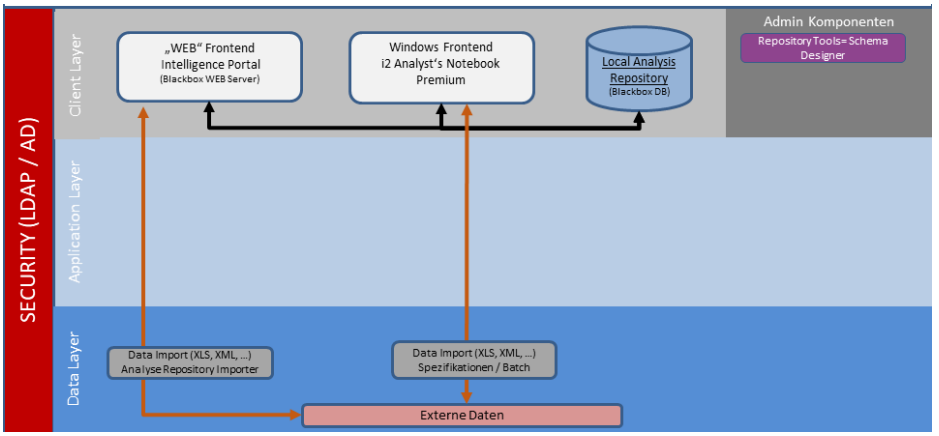


Abbildung 36: Szenario 1 (Konzept, 2017) für die Datenhaltung – Local Analysis Repository

6.5.2 IBM i2 ANBP mit Group Analysis Repository (GAR)

SZENARIO 2: Bei diesem Szenario werden die Daten in einem zentralen Datastore, dem sogenannten Group Analysis Repository auf einem Server gespeichert. Sie können nun entweder über das Intelligence Portal oder über ein verknüpftes Analyst's Notebook ausgewertet und bearbeitet werden. Das Group Analysis Repository ermöglicht eine zentrale Datenspeicherung und bietet Funktionen wie Protokollierung von Änderungen und Multi-User-Support.

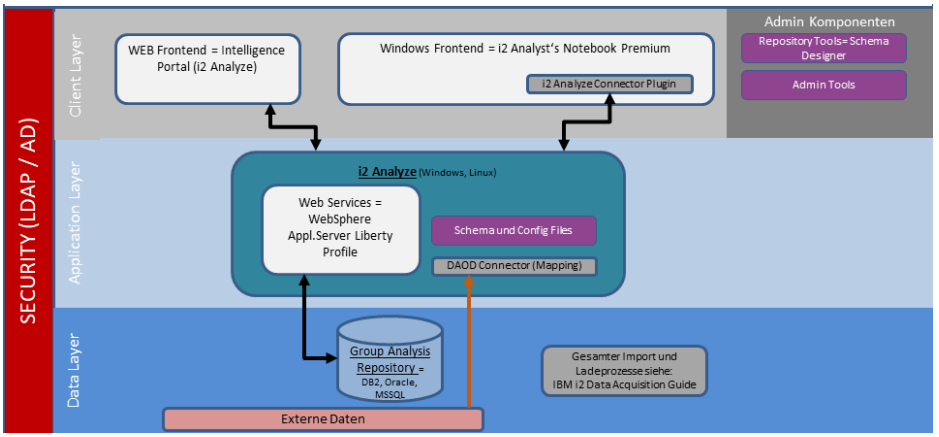


Abbildung 37: Szenario 2 (Konzept, 2017) für die Datenhaltung – Group Analysis Repository

6.5.3 IBM ANBP mit Group Analysis Repository (GAR) mit Information Store und ESRI

SZENARIO 3: Dieses Szenario zeigt die vollständige Installation des i2 EIA im Opal Deployment. Es ist das komplexeste der vorgestellten Szenarien, bietet aber auch die meisten Möglichkeiten und funktioniert selbst unter hoher Datenlast noch schnell und zuverlässig.

Der hier dargestellte ESRI Server für die Geoinformationen kann auch in den Szenarien 1 und 2 angebunden werden.

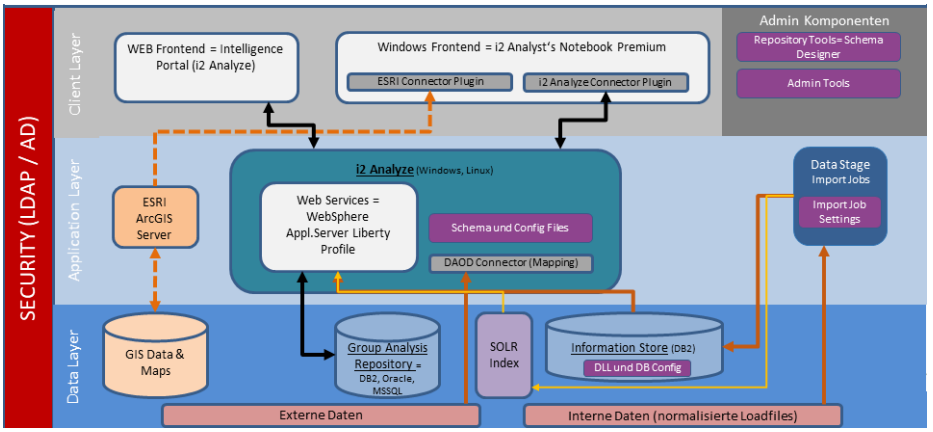


Abbildung 38: Szenario 3 (Konzept, 2017) für die Datenhaltung – i2 EIA im Opal Deployment

7 IBM Cloud

Die IBM Cloud ist die Cloud Plattform von IBM und bietet zurzeit etwa 170 Services, die sofort genutzt werden können. Grob gliedern sich die Services der IBM Cloud zum Zeitpunkt der Erstellung dieser Publikation in folgende Bereiche:

- Compute
- Containers
- Networking
- Storage
- AI
- Analytics
- Databases
- Developer Tools
- Integration
- Internet of Things
- Security and Identity
- Starter Kits
- Web and Mobile Application Services

Der Bereich AI wird im nächsten Unterkapitel genauer beleuchtet, da dieser wichtige Services bietet, die im Zusammenhang mit Textmining verwendet werden können.
Allgemein kann gesagt werden, dass die Services in unterschiedlichen Security Ebenen provisioniert werden können. Von einer Instanz in der „Public Cloud“ über eine Instanz auf einem „dedicated Server“ bis hin zu einer Installation auf eigener Hardware im eigenen Rechenzentrum mittels der „IBM Cloud Private“ (Gilozzo et al., 2017), (Jaatun, Zhao and Rong, 2009) und (Zuhu et al., 2009). Die IBM Cloud und seine Services ist im ständigen Wandel und Ausbau.

7.1 IBM Cloud - AI Services Überblick

Die Watson Services, als ein Teil der IBM Cloud, bieten Services aus dem Bereich Artificial Intelligence, auf die hier besonders eingegangen wird, zumal einige davon im Rahmen des Projekts an der ZentDok zum Einsatz kommen oder getestet werden.

- Watson Assistant = ein Chatbot Service, welches es Endusern selbst ohne Programmierkenntnisse erlaubt einen Chatbot zu entwickeln der mit Personen zu verschiedenen Themen interagieren kann.
- Watson Discovery = eine Cloud-native Insight-Engine die die Speicherung und Datenanreicherung mithilfe von Natural Language Processing kombiniert, um Erkenntnisse aus strukturierten und unstrukturierten Daten mit AI-basierten Abfragen zu extrahieren.
- Watson Knowledge Catalog = ermöglicht intelligente, Self-Service-Erkennung von Daten, Modellen und mehr. Es wird die Möglichkeit für die Anwendung von künstlicher Intelligenz, maschinellem Lernen und deep learning geboten. Der Knowledge Catalog ermöglicht einfachen Zugriff, kategorisieren und teilen von Daten, Wissensbeständen und deren Beziehungen.
- Watson Knowledge Studio = WKS, bietet die Möglichkeit für Subject Matter Experts Machine Learning Modelle für Textmining zu erstellen die dann entweder in anderen Cloudservices oder offline im Watson Explorer verwendet werden können. Auf dieses Service wird im folgenden Kapitel noch genauer eingegangen.
- Language Translator = dieses Service erlaubt es, Text von einer Sprache in eine andere zu übersetzten. Übersetzungsmodelle für spezielle Domänen können mittels Machine Learning antrainiert werden.
- Natural Language Classifier = dieses Service kann Input Text zu antrainierten Domänen zuordnen (klassifizieren).
- Natural Language Understanding = NLU, extrahiert Metadaten, Entitäten, Konzepte, Sentimente, Emotionen und Relationen aus Text. Modelle aus dem Watson Knowledge Studio können neben der Standardfunktionalität zur Erweiterung verwendet werden.

- Personality Insights = leitet Erkenntnisse aus Transaktions- und Social-Media-Daten ab, um psychologische Merkmale zu identifizieren, die Kaufentscheidungen, Absichts- und Verhaltensmerkmale bestimmen.
- Speech to Text = transkribiert performant gesprochene Sprache zu Text.
- Text to Speech = wird verwendet um Text in einer natürlichen Sprache ausgeben zu können. Die Aussprache des Service kann dabei an den Use Case angepasst werden.
- Tone Analyzer = Menschen zeigen verschiedene Emotionen, wie Freude, Traurigkeit, Wut und Verträglichkeit, in der täglichen Kommunikation. Tone Analyzer nutzt kognitiv-linguistische Analyse, um eine Vielzahl von Tönen sowohl auf Satz- als auch auf Dokumentebene zu identifizieren. Es erkennt drei Arten von Tönen, darunter Emotionen (Wut, Ekel, Angst, Freude und Trauer), soziale Neigungen (Offenheit, Gewissenhaftigkeit, Extrovertiertheit, Verträglichkeit und emotionale Reichweite) und Sprachstile (analytisch, selbstbewusst und vorläufig) aus Text.
- Visual Recognition = dieses Service analysiert Bilddaten um Inhalte von Bildern wie Objekte, Nahrungsmittel, Personen oder Text zu extrahieren. Spezielle image recognition kann vom User antrainiert werden.
- Watson Studio = eine Plattform für maschinelles Lernen, AI und Data Science. Es bietet eine Reihe von Tools und Umgebungen für Data Scientists, Entwickler und Domänenexperten um mit strukturierten und unstrukturierten Daten umgehen zu können.

Siehe Console.bluemix.net, (2018) und Gilozzo et al., (2017).

7.2 IBM Watson Knowledge Studio (WKS)

Das Watson Knowledge Studio ist ein Service zum Erstellen von Machine Learning Models das über die IBM Cloud angeboten wird. Das Modell wird in der Cloud entwickelt und kann dann komplett offline im Watson Explorer verwendet werden.
Knowledge Studio stellt benutzerfreundliche Tools zum Annotieren unstrukturierter Domänenliteratur bereit und verwendet diese

Anmerkungen um ein benutzerdefiniertes maschinelles Lernmodell zu erstellen, das die Sprache der Domäne versteht. Die Genauigkeit des Modells wird durch iteratives Testen verbessert, was letztlich zu einem Algorithmus führt, der aus den Mustern, die er sieht, lernen und diese Muster in großen Sammlungen neuer Dokumente erkennen kann. Das Lernmodell kann dann auf diversen Watson Plattformen verwendet werden um Mentions von Relationen und Entitäten zu finden und zu extrahieren, einschließlich Entitäts-Ko-Referenzen.

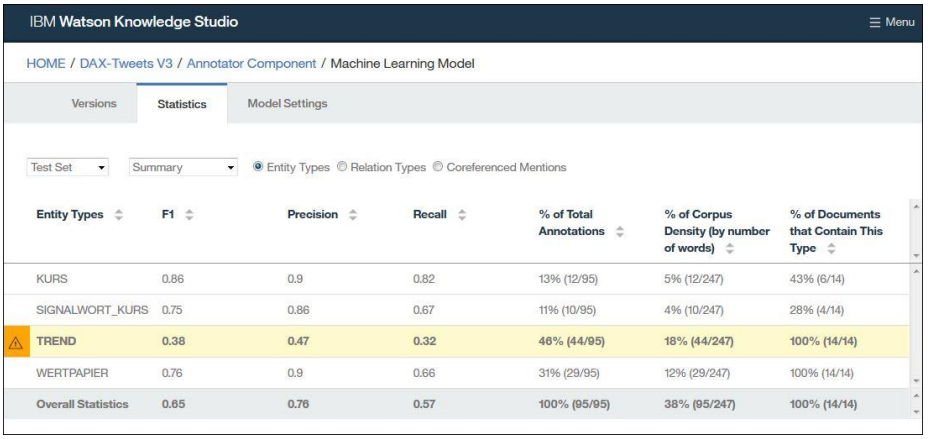
Kurz gesagt, Watson Knowledge Studio ist eine Cloud basierte Anwendung für die Erstellung von Textanalysemodellen durch Domain Experten für den Mensch-Maschine-Wissenstransfer mittels Natural Language Processing (NLP) und Machine Learning (KI). Das Anlernen von Watson funktioniert über Referenzdokumente in denen Domain Experten relevante Textstellen (Wörter, Phrasen) markierten (z.B. nach Wortart, Thema, Semantik, Bedeutung, Wertigkeit, Sentiment/Stimmungsanalyse, …). Damit weiß Watson, welche Begriffe zu welcher Annotation gehören und lernt somit. Dieses angelernte Wissen kann dann in WEX/AC für automatische Annotationen weiterverwendet werden. Abbildung 39 zeigt die WKS Oberfläche beim Annotieren (Console.bluemix.net, 2018).

Abbildung 39: Watson Knowledge Studio Beispiel

7.2.1 WKS-Prozess

Das Erstellen eines Machine Learning Modells, welches im vorhergehenden Kapitel bereits angerissen wurde, läuft nach dem folgenden Prozess ab.

Basierend auf einer Reihe domänenspezifischer Quelldokumente erstellt ein Team ein Type System, das Entitätstypen und Beziehungstypen für die Informationen definiert, die für den Anwender, die das Modell verwenden soll, von Interesse sind. Diese Quelldokumente sollen einerseits unkritisch sein, also keine geheimen Dokumente, andererseits sollen sie auch den Textdokumenten, für die die Annotationen schlussendlich verwendet werden nahekommen.
Nach der Definition des Type Systems und der Auswahl der Trainings oder Quelldaten kommentiert eine Gruppe von zwei oder mehr menschlichen Annotatoren (Domain Experts oder Subject Matter Experts) einen kleinen Satz von Quelldokumenten um Wörter zu kennzeichnen die Entitätstypen darstellen, um Relationstypen zwischen Entitäten zu identifizieren und um Ko-Referenzen zu definieren. Das sind verschiedene Erwähnungen, die sich auf dieselben beziehen. Jegliche Inkonsistenz in den Annotationen werden am Ende aufgelöst um einen Satz optimal annotierter Dokumente zum Trainieren von Watson zu haben. Diese Dokumente werden auch als die Grundwahrheit (Ground truth) bezeichnet.
Das Watson Knowledge Studio verwendet nun die Grundwahrheit, um ein Machine Learning Modell zu trainieren. Am Ende des Trainigsprozesses kann dieses Modell durch verschiedene statistische Kennzahlen wie Precision, Recall und F1 Score evaluiert und bei Bedarf adaptiert werden. Abbildung 40 zeigt diese Auswertung. Dieses Evaluieren und Adaptieren geschieht in einer Schleife solange, bis das Ergebnis zufriedenstellend ist. Genauer beschrieben wird der Watson Knowledge Studio Prozess bei Console.bluemix.net, (2018) und Gilozzo et al., (2017).



Entity Types	F1	Precision	Recall	% of Total Annotations	% of Corpus Density (by number of words)	% of Documents that Contain This Type
KURS	0.86	0.9	0.82	13% (12/95)	5% (12/247)	43% (6/14)
SIGNALWORT_KURS	0.75	0.86	0.67	11% (10/95)	4% (10/247)	28% (4/14)
⚠ **TREND**	**0.38**	**0.47**	**0.32**	**46% (44/95)**	**18% (44/247)**	**100% (14/14)**
WERTPAPIER	0.76	0.9	0.66	31% (29/95)	12% (29/247)	100% (14/14)
Overall Statistics	**0.65**	**0.76**	**0.57**	**100% (95/95)**	**38% (95/247)**	**100% (14/14)**

Abbildung 40: Watson Knowledge Studio Statistics

Das trainierte Modell wird dann in diversen Softwareprodukten wie zum Beispiel dem Watson Explorer, Watson Discovery Service oder Natural Language Understanding Service verwendet, um Entitäten, Relationen und Ko-Referenzen in neuen, noch nie gesehenen Dokumenten zu finden.

Neben dem Machine Learning Ansatz besteht auch noch die Möglichkeit regelbasierte Annotationen aus RegEx oder Wörterbüchern zu erstellen. Damit bietet das Watson Knowledge Studio ähnliche Funktionalität wie das Content Analytics Studio mit dem Zusatz des Machine Learnings. Wesentliche Unterschiede sind, dass sich mit dem Content Analytics Studio Annotationen offline auf UIMA Basis entwickeln lassen während das Watson Knowledge Studio ein Cloud Produkt mit IBM proprietärem Format ist. Das Content Analytics Studio bietet weiters als Teil der Watson Explorer Deep Analytics Edition eine bessere Integration mit dem Watson Explorer (beispielsweise um PEAR Files direkt zu exportieren oder Annotationen gleich mit Dokumenten von einem WEX Server zu testen). Außerdem bietet das Watson Knowledge Studio (noch) keine so mächtigen linguistischen Möglichkeiten wie das Content Analytics Studio. Das Content Analytics Studio bietet allerdings kein Machine Learning. Aus diesem Grund werden beide Verfahren in der ZentDok kombiniert eingesetzt um die Vorteile der einzelnen Technologien am besten nutzen zu können.

8 Zukünftige Entwicklungsschritte

Die Generierung von Wissen mit Informationen aus offenen Quellen hat sich in den letzten Jahren ganz erheblich geändert. Die Darstellung der Wissensentwicklung mit IBM Watson und i2 hat gezeigt, dass der flexible Umgang mit einem Methodenmix auf einer integrierenden Plattform zu ganz neuen Einsichten führt. Die Entwicklungen stehen allerdings nicht still. Vielmehr ist damit zu rechnen, dass die Datenverfügbarkeit, die Datenmengen und die zur Verfügung stehenden Methoden auch in Zukunft weiter zunehmen. Die generische Aufbereitung von Datenbeständen, mit schnellen anwendungsspezifischen ad-hoc Auswertungen wird zunehmend wichtiger um zeitnahe Entscheidungen zu unterstützen. Zwar verbessert sich die allgemeine Datenverfügbarkeit, die für eine Analyse zur Verfügung steht, aber die Datenmengen, die es zu analysieren gilt nehmen in gleichem Maße zu.

Neben den klassischen Quellen, mit 70 Mio. wissenschaftlichen Publikationen (Web of Science, n.d.) 12 Mio. Patenten (Patstat, n.d.), 300 Mio. Firmen (Orbis, n.d.) stehen derzeit auch 250 Mio. Pressemeldungen zu diplomatischen Events (GEDELT, n.d.), ungefähr 1,7 Milliarden Websites (von denen 25% aktive sind und ca. 75% geparkte Domains sind), davon ca. 170 Mio. aktiv genutzt und social Media Daten in dem in Abbildung 41 gezeigtem Umfang für Analysen zur Verfügung (We Are Social, 2018).

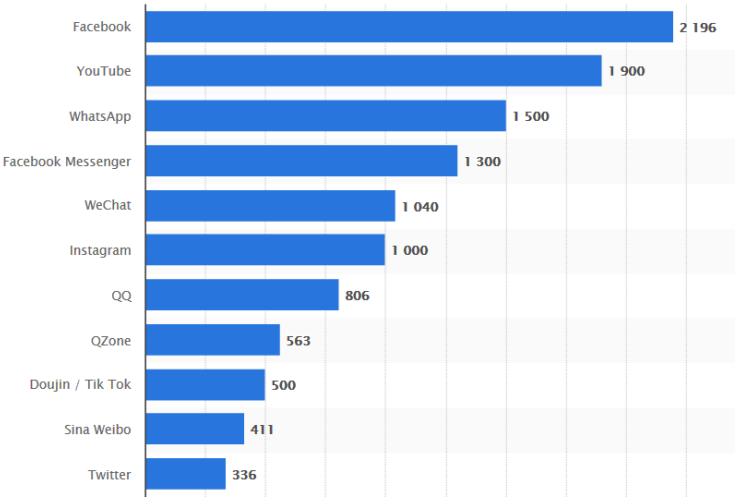


Abbildung 41: Most famous social network sites worldwide, July 2018

Zusätzlich finden sich auf Seiten wie Wikileaks oder an anderen Stellen des Darknets zunehmend relevante Datensätze mit Emails, Daten im Millionenbereich mit geleakten Userdaten oder mit anderen relevanten Informationen. Viele der genannten Quellen bieten Multimediadaten in verschiedenen Sprachen. Damit wird deutlich, dass schon aktuell die performante Analyse von Multimedia Quellen im 100 Mio. Bereich zunehmend relevant wird.

Die Kernprozesse der Open Source Information (OSInfo) Verarbeitung, wie Informationsbeschaffung, Aufbereitung, Strukturierung, Harmonisierung und Anreicherung sowie das Analysieren, die Daten den Analysten zugänglich zu machen und die Qualitätssicherung zu gewährleisten, werden von den IBM Produkten WEX und i2 unterstützt. Ein besonderer Schwerpunkt liegt in zeit- und raumbezogenen Darstellungen um so gut wie möglich die zeitlichen Abhängigkeiten zwischen Quellen, Zeitpunkt der Veröffentlichung und Themen zu verstehen und übersichtlich darzustellen. Die Automatisierung der wichtigsten Schritte, gerade in der Datenbeschaffung und im Pre-Processing, obliegt in weiten Teilen noch dem Anwender. In wissenschaftlichen Communities entstehen in diesem Kontext neue Anwendergruppen, die ergänzende Software zur auftragsspezifischen

Verwendung, Weiterentwicklung, Effizienzsteigerung und Automatisierung der wichtigsten Schritte in der Open Source Analyse weiterentwickeln. Es ist damit zu rechnen, dass in den nächsten Jahren eine ganze Reihe an Services aus den Forschungsprojekten heraus entstehen und dann über den Open Source Kanal auch i2 Anwendern zur Verfügung stehen. In dem EU Projekt „ASGARD“ z.B. werden in einer dem Prinzip nach „geschlossenen Open Source Community“ Services entwickelt, die in der Anwendung Daten produzieren, die in WEX und i2 aufgegriffen und weiterverarbeitet werden können. Die Entwickler von i2 sind dem Projekt beigetreten, um die entstehenden Services für i2 Anwender zugänglich zu machen.

Im EU-Projekt „DANTE“ z.B. sind Services vor allem zur Wissens-anreicherung von Multimedia Daten entwickelt worden, die dazu verwendet werden können, um zusätzliche Metadaten zu generieren die dann in den WEX und i2 Prozess einfließen. Einer der Hintergründe ist, dass sich die Möglichkeiten der Fälschung von Multimedia Daten in den letzten Jahren dramatisch verändert haben. Inzwischen ist es möglich mit AI basierten Algorithmen sowohl Bild, als auch Ton und Video so zu fälschen, dass Verwechslungen zunehmend wahrscheinlicher werden. Lyrebird und Adobe haben z.B. Stimmenimitatoren entwickelt, die nach kurzer Trainingszeit von 1-5 Minuten erkennbar personenspezifische Texte sprechen können. Mit Deep Fakes Fakeapp lassen sich z.B. Videos produzieren, bei denen Personen durch Morphing in das Video hinein gerechnet werden, so dass es erscheint, als würde eine andere Person im Video spielen. In der Graphik weiter unten sind Beispiele dargestellt (Brundage et al., 2018).

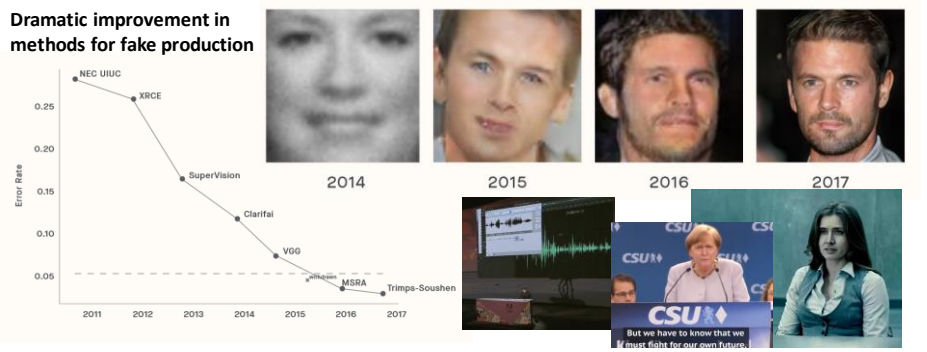


Abbildung 42: Advances in producing disinformation with AI, from different sources

Durch die Beteiligung der ZentDok an vielen aktuellen Forschungsprojekten, entweder direkt oder indirekt über Kooperationspartner wie z.B. über das AIT bei ASGARD[1], DANTE[2], ANITA[3] und DRIVER[4] werden frühzeitig neue Technologien identifiziert und können bei entsprechendem Reifegrad aufgegriffen und in den Prozess der Wissensentwicklung integriert werden. So ist die Zukunftssicherheit der analytischen Prozesse der Wissensentwicklung an der ZentDok gewährleistet.

In DANTE wurde, um ein Beispiel zu nennen, ein Algorithmus entwickelt, mit dem es möglich ist innerhalb eines Bildes die Bereiche zu markieren, die gefälscht, kopiert oder nachträglich verändert wurden. Ein Beispiel dazu zeigt das untenstehende Bild (DANTE Key Technologien, 2018).

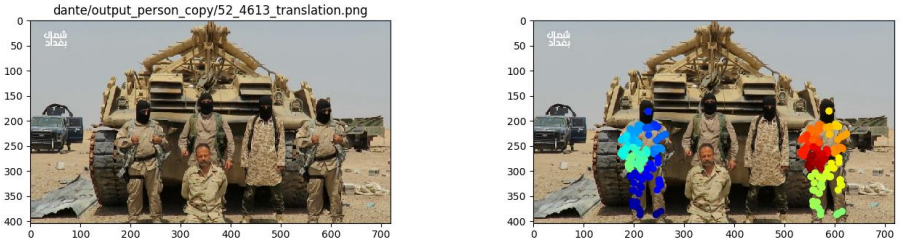


Abbildung 43: DANTE Beispiel - Identification of Miss- and Disinformation

Die identifizierten Bilder können über einen Meta-Tag klassifiziert werden, so dass innerhalb eines Datensatzes mit der Facettensuche des Watson Explorers nach allen veränderten Bildern gesucht werden kann. Diese Zusatzfunktion lässt sich derzeit über eine API ansprechen und benötigt noch Trainingsmaterial um in ausreichender Qualität Ergebnisse zu liefern. Sollte sich diese Klassifizierung im operativen Einsatz als sinnvoll erweisen, kann diese auch über eine offline Version in das Pre-Processing von Watson Explorer eingebunden werden.

[1] http://www.asgard-project.eu/
[2] https://www.h2020-dante.eu/
[3] http://www.anita-project.eu/
[4] http://www.driver-project.eu/

Weitere Funktionen, die in der Wissensanreicherung zum Einsatz kommen könnten, wäre ein Algorithmus zur Erkennung von Personen, der ähnliche Personen auf einem Bild identifiziert, wie die folgende Graphik (Abbildung 44) zeigt.

Abbildung 44: Beispiel - Erkennung und Identifizierung von Personen

Dieser Algorithmus liefert Fotos mit Personen, die von ihrer Erscheinung und Aussehen ähnlich sind. Damit lassen sich Datensätze klassifizieren, die mit hoher Wahrscheinlichkeit Personen vom gleichen Typ enthalten. Ein typischer Anwendungsfall wäre bei Pressemeldungen alle Datensätze zu identifizieren die in einem Bild Personen mit Tarnkleidung enthalten. Der gleiche Algorithmus kann auch verwendet werden, um Bilder mit Panzer vom gleichen Typ, oder Bilder mit spezifischen Waffengattungen zu erkennen. Damit bietet der Algorithmus in der Wissensanreicherung die Möglichkeit die Konzepterkennung bei Bildern zu automatisieren.

Ein letztes Beispiel für einen Algorithmus, der die Wissensentwicklung mit dem Watson Explorer und i2 unterstützen kann zeigt Abbildung 45. Der Algorithmus zur Video-Indexierung liefert eine Rangliste ähnlicher Bilder, vergleichbar der Google Bildsuche. Dieser Algorithmus ist noch in einer sehr experimentellen Phase und liefert immer wieder auch überraschende Ranking Listen. Wie bei Google Bildsuche getestet werden kann funktioniert die Identifikation gleicher Bilder, auch bei unterschiedlicher Auflösung und bei Ausschnittveränderung, in zufriedenstellender Qualität. Die in der Rangfolge gereihten möglichst ähnlichen Bilder entsprechen allerdings manchmal nicht der Nutzererwartung, was daran liegt, dass der Algorithmus alle Bildinformationen für den Vergleich heranzieht,

wohingegen der menschliche Vergleich verschiedene „wichtige“ Bildelemente höher gewichtet als die künstliche Intelligenz.

Abbildung 45: Ranked list of retrieved images

Ein Beispiel dafür ist, dass bei der menschlichen Beurteilung der Ähnlichkeit von Bildern Bilder mit gleichen Menschen als ähnlich empfunden werden auch wenn der Hintergrund, der einen Großteil des Bildes ausmacht, unterschiedlich ist. Der Mensch wird dabei höher gewichtet. Der Algorithmus dagegen liefert Bilder mit gleicher Farbkomposition. Auch wenn diese Erfahrungen nicht direkt zu brauchbaren operativen Algorithmen führen, lassen sich diese Erkenntnisse jedoch nutzen, um die Wissensentwicklung zu verbessern.

Insgesamt liegt der Vorteil der Einbindung in aktuelle Forschungsprojekte auf der Hand. Aus diesen Projekten können bei Bedarf Algorithmen in den analytischen Prozess übernommen werden, so dass der operative Einsatz nahe am aktuellen Stand der Technik unabhängig von einzelnen Plattformen gewährleistet wird.

Zurzeit gibt es einen Wettbewerb um die besten Algorithmen im Bereich der künstlichen Intelligenz. IBM hat dabei mit Watson eine gute Position. Allerdings setzten auch andere große Software Entwicklungsfirmen auf eine ähnliche Strategie. Sowohl Google mit Tensor Flow, als auch Microsoft und Amazon engagieren sich in diesem Wettbewerb. Auch kleinere sehr spezifische Firmen und Forschungsinstitute nehmen an diesem Wettbewerb teil. Die daraus entstehenden Innovationen gilt es zu nutzen und in die eigenen analytischen Prozesse einzubinden aber die Abhängigkeiten zu vermeiden.

Als staatliche Organisation profitiert die ZentDok indem sie offen für neue Entwicklungen, aber gleichzeitig kritisch in Bezug auf einen möglichen Informationsabfluss bei der Verwendung von Modulen aus verschiedenen Cloud Umgebungen ist. Darin liegt eine große Herausforderung der Zukunft in der softwaregestützen Wissensentwicklung. Nachdem gerade die leistungsfähigsten Algorithmen auf Trainingsmodellen basieren, die von einer großen Zahl an Daten profitiert, sind in der Zukunft Lösungen gefragt, die die Verwendung der leistungsfähigen cloudbasierten Algorithmen ermöglicht aber jeden Informationsabfluss vermeidet.

9 Abbildungsverzeichnis

Abbildung 1: Projektkonfiguration 1 13
Abbildung 2: News Erfassung und Visualisierung in Ushahidi 14
Abbildung 3: Übersicht über die verwendeten Kategorien von Metadaten in IBM Watson Explorer 16
Abbildung 4: Annotationsarten 19
Abbildung 5: Entwicklungsschritte der Suche 20
Abbildung 6: ProOSINT Prozess 25
Abbildung 7: ProTerm Textanalyse - Modul NewTerm – Beispiel „cyber“ 27
Abbildung 8: WEX/AC - Anwenderprozess 30
Abbildung 9: WEX/AC - Analysemodule im Inhaltsanalyseminer 32
Abbildung 10: WEX/AC – Analysemodule 33
Abbildung 11: WEX/AC – Volltextsuche mit Treffer Highlighting und Sentiment 34
Abbildung 12: WEX/AC - Erweiterte Suche - Modell Suchprozess 35
Abbildung 13: WEX/AC - Erweiterte Suche - Beispiel 36
Abbildung 14: WEX/AC – Facettennavigation, Beispiele unterschiedlicher Inhaltsanalyseminer-Objektgruppen 37
Abbildung 15: WEX/AC - Facettenanalyse - Modell 37
Abbildung 16: WEX/AC - Facettenanalyse - Beispiel 38
Abbildung 17: WEX/AC - Zeitreihenanalyse - Prozessbeschreibung 39
Abbildung 18: WEX/AC - Zeitreihenanalyse - Beispiel 39
Abbildung 19: WEX/AC - Abweichungen - Beispiel 40
Abbildung 20: WEX/AC - Trendanalyse - Prozessbeschreibung 41
Abbildung 21: WEX/AC - Trendanalyse - Beispiel 41
Abbildung 22: WEX/AC - Facettenpaaranalyse - Modell Beispiel „Cyber“ 42
Abbildung 23: WEX/AC - Facettenpaaranalyse - Beispiel 43
Abbildung 24: WEX/AC - Verbindungen - Prozessbeschreibung 44
Abbildung 25: WEX/AC - Verbindungen – Beispiel 44
Abbildung 26: WEX/AC - Dashboard - Prozessbeschreibung 45
Abbildung 27: WEX/AC - Dashboard – Beispiel 46
Abbildung 28: Content Analytics Studio Beispiel 48
Abbildung 29: i2 Analyze - Modell Anwenderprozess 52
Abbildung 30: i2 Analyst's Notebook - Beispiel Netzwerkanalyse 54

Abbildung 31: i2 Analyst's Notebook - Beispiel Netzwerkanalyse ZentDok Thesaurus Ausschnitt "Internet" 55
Abbildung 32: i2 Analyst's Notebook - Beispiel Zeitreihenanalyse 56
Abbildung 33: i2 Analyst's Notebook - Beispiel Histogramme und Aktivitäten 57
Abbildung 34: i2 Analyst's Notebook - Beispiel Georeferenzierung 59
Abbildung 35: i2 Intelligence Portal 61
Abbildung 36: Szenario 1 (Konzept, 2017) für die Datenhaltung – Local Analysis Repository 62
Abbildung 37: Szenario 2 (Konzept, 2017) für die Datenhaltung – Group Analysis Repository 63
Abbildung 38: Szenario 3 (Konzept, 2017) für die Datenhaltung – i2 EIA im Opal Deployment 64
Abbildung 39: Watson Knowledge Studio Beispiel 68
Abbildung 40: Watson Knowledge Studio Statistics 70
Abbildung 41: Most famous social network sites worldwide, July 2018 72
Abbildung 42: Advances in producing disinformation with AI, from different sources 73
Abbildung 43: DANTE Beispiel - Identification of Miss- and Disinformation 74
Abbildung 44: Beispiel - Erkennung und Identifizierung von Personen 75
Abbildung 45: Ranked list of retrieved images 76

10 Glossar

AIT	Austrian Institute of Technology
ANB	Analyst's Notebook
ANBP	Analyst's Notebook Premium
CAS	Content Analytics Studio
CAS	Common Analysis Structure
CDFZ	Cyber Dokumentations- und Forschungszentrum
EIA	Enterprise Insight Analysis
ELP	Entity-Link-Property
ESRI	Environmental Systems Research Institute
KI/AI	Künstliche Intelligenz / Artificial Intelligence
ML	Machine Learning
NER	Named Entity Recognition
NLP	Natural Language Processing
NLQ	Natural Language Query
NLU	Natural Language Understanding
OASIS	Organization for the Advancement of Structured Information Standards
OSInfo	Open Source Information
OSINT	Open Source Intelligence
OWL	Web Ontology Language
PaaS	Platform as a Service
PEAR	Processing Engine Archive
POS	Part of Speech
QMS	Qualitätsmanagement System
RDF	Ressource Description Framework
SKOS	Simple Knowledge Organisation System
UIMA	Unstructured Information Management Architecture
WEX	Watson Explorer
WEX/AC	Watson Explorer Analytical Components
WEX/FC	Watson Explorer Foundational Components
WKS	Watson Knowledge Studio
WM	Wissensmanagement

11 Stichwortverzeichnis

A

Annotation 18, 36, 69
Artificial Intelligence 66, 81

C

CDFZ 14
Cloud 8, 9, 31, 65, 66, 67, 68, 77, 87
Cognitive Computing 29, 87
Cognos 47
Content Analytics Studio 18, 19, 31, 36, 46, 47, 48, 49, 81
Crawler 30

E

ESRI 51, 58, 64

F

Facetten 16, 17, 18, 19, 23, 33, 36, 42, 43

G

Georeferenzierung 53, 58, 59

I

i2
- Analyst's Notebook 81
- Enterprise Insight Analysis 7, 9, 81
- Group Analysis Repository 63, 64
- i2 Analyze 43, 51, 52, 53, 59, 60
- Information Store 53, 60, 64

Index 31, 38, 40

M

Machine Learning 11, 19, 23, 31, 36, 68, 70, 81

N

Natural Language Processing 23, 29, 66, 68, 81
Netzwerkanalyse 51, 53, 54, 55

O

Ontologien 7, 18, 31
OSINT 53

P

ProTerm 23, 26, 27

T

Trendanalyse 40, 41

U

UIMA 31

W

Watson
- Analytical Components 29, 81
- Foundational Components 29, 81
- oneWEX 29
- Watson Explorer 7, 9
- Watson Knowledge Studio 18, 19, 31, 36, 66, 67, 68, 70, 81
- Watson Services 8, 9, 66
- WEX 68

Z

Zeitreihenanalyse 39, 53, 56
ZentDok 7, 11, 14, 16, 23, 26, 54, 55, 74, 77

12 Literaturverzeichniss

Bossert, B. (2014). *Metadaten & Struktur - nestor - Deutsche Nationalbibliothek - Wiki*. [online] wiki.dnb.de. Available at: https://wiki.dnb.de/pages/viewpage.action?pageId=95651769 [Accessed 21 Sep. 2018].

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P. and Garfinkel, B. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*.

Chen, W., Adams, B., Dean, C., Naganna, S., Nandam, U. and Thorne, E. (2014). *Building 360-degree information applications*. Poughkeepsie, NY: IBM Corp., International Technical Support Organization.

Console.bluemix.net. (2018). *IBM Cloud Docs - Watson Knowledge Studio*. [online] Available at: https://console.bluemix.net/docs/services/knowledge-studio/index.html#wks_overview_full [Accessed 27 Aug. 2018].

DANTE Key Technologien. (2018). In: MidTerm Review. Brüssel.

Ferrucci, D., Lally, A., Verspoor, K., Nyberg, E. (2009). *Unstructured Information Management Architecture (UIMA) Version 1.0*.

GEDELT. (n.d.). [online] Available at: http://www.gdeltproject.org [Accessed 23 Sep. 2018].

Gliozzo, A., Ackerson, C., Bhattacharya, R., Goering, A. and Jumba, A. (2017). *Building cognitive applications with IBM Watson Services*.

Göllner, J., Mak, K. and Woitsch, R. (2010a). *Grundlagen zum Wissensmanagement im ÖBH (Teil 1: Ein WM-Ramenwerk aus der Sicht praktischer Anwendungen)*. Wien: Landesverteidigungsakademie / Zentraldokumentation (ZentDok).
Göllner, J., Mak, K. and Woitsch, R. (2010b). *Grundlagen zum Wissensmanagement im ÖBH (Teil 2: Wissensbilanz als Steuerungsinstrument im*

ÖBH: Ein Evaluierungs-Rahmenwerk aus der Sicht praktischer Anwendungen). Wien: Landesverteidigungsakademie / Zentraldokumentation (ZentDok).

Göllner, J., Meurers, C., Peer, A. and Povoden, G. (2011). *Einführung in die Soziale Netzwerkanalyse und exemplarische Anwendungen*.

Göllner, J., Meurers, C., Peer, A., Langer, L. and Karmmerstetter, M. (2014). *Bedeutung des Risikomanagements für die Sicherheit von Smart Grids*. In: 13. Symposium Energieinnovation.

High, R. (2012). *The Era of Cognitive Systems: An Inside Look at IBM Watson and How it Works*. IBM.

IBM (2015). *IBM i2 Analyze Data Model White Paper Version 4*. [online] Available at: https://www-01.ibm.com/support/docview.wss?uid=swg27042357&aid=1 [Accessed 23 Sep. 2018].

ibm.com. (2018a). *i2 EIA - IBM Knowledge Center*. [online] Available at: https://www.ibm.com/support/knowledgecenter/en/SSXVXZ_2.2.0/com.ibm.i2.landing.doc/eia_welcome.html [Accessed 27 Aug. 2018].

ibm.com. (2018b). *i2 Intelligence Analysis - IBM Knowledge Center*. [online] Available at: https://www.ibm.com/support/knowledgecenter/en/SS3J58_9.1.0/com.ibm.i2.welcome.doc/portfolio_welcome.html [Accessed 27 Aug. 2018].

ibm.com. (2018c). *Watson Explorer - IBM Knowledge Center*. [online] Available at: https://www.ibm.com/support/knowledgecenter/en/SS8NLW_12.0.0/com.ibm.swg.im.infosphere.dataexpl.welcome.doc/doc/watsonexplorer_12.0.0.html [Accessed 27 Aug. 2018].

ibm.com. (2018d). *i2 Analyst's Notebook - IBM Knowledge Center*. [online] Available at: https://www.ibm.com/support/knowledgecenter/en/SS3J58_9.1.0/com.ibm.i2.anb.doc/analysts_notebook_welcome.html [Accessed 27 Aug. 2018].

ibm.com. (2018e). *IBM i2 Analyst's Notebook - Überblick - Deutschland*. [online] Available at: https://www.ibm.com/de-de/marketplace/analysts-notebook [Accessed 5 Sep. 2018].

Jaatun, M., Zhao, G. and Rong, C. (2009). *Cloud Computing*. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg, pp.621-625.

Mak, K. and Woitsch, R. (2005). *Der Einsatz des prozessorientierten Wissensmanagementwerkzeuges PROMOTE® in der Zentraldokumentation der Landesverteidigungsakademie*. Wien.

Mak, K., Klerx, J., Pilles, H. and Göllner, J. (2015). *Wissensentwicklung mit "Crowd OSInfo"*.

Orbis. (n.d.). [online] Available at: https://www.bvdinfo.com/de-de/our-products/company-information/international-products/orbis [Accessed 23 Sep. 2018].

Patstat. (n.d.). [online] Available at: http://www.epo.org/searching-for-patents/business/patstat_de.html [Accessed 23 Sep. 2018].

We Are Social. n.d. *Most famous social network sites worldwide as of July 2018, ranked by number of active users (in millions)*. Statista. Accessed September 5, 2018. Available from https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Web of Science. (n.d.). [online] Available at: http://www.webofknowledge.com [Accessed 23 Sep. 2018].

Zhu, J. et al. (2009) *IBM Cloud Computing Powering a Smarter Planet*. In: Jaatun M.G., Zhao G., Rong C. (eds) Cloud Computing. CloudCom 2009. Lecture Notes in Computer Science, vol 5931. Springer, Berlin, Heidelberg

Zhu, W., Foyle, B., Gagné, D., Gupta, V., Magdalen, J., Mundi, A., Nasukawa, T., Paulis, M., Singer, J. and Triska, M. (2014). *IBM Watson Content Analytics: Discovering Actionable Insight from Your Content*. Poughkeepsie, NY: IBM Corp., International Technical Support Organization.

13 Autoren

Ing. Mag. Klaus MAK
Oberst des höheren militärfachlichen Dienstes
Leiter der Zentraldokumentation an der Landesverteidigungsakademie
klaus.mak@bmlv.gv.at

Hans Christian PILLES, ADir RgR
Leiter Technische Dokumentation
an der Zentraldokumentation/Landesverteidigungsakademie
hans.pilles@bmlv.gv.at

Dr. Joachim KLERX
Forscher im Innovation System Department des Austrian Institute of Technology (AIT)

Markus BERTL, BSc.
IT Consultant für Cognitive Computing & Cloud

Band 15 / 2018

An der Zentraldokumentation der Landesverteidigungsakademie (ZentDok/LVAk) werden seit mittlerweile 50 Jahren Anstrengungen unternommen, um der Gesamtorganisation des ÖBH qualitativ hochwertige offene Fachinformationen zur Verfügung zu stellen. Diese ständige operative und anwendungsorientierte Entwicklungsarbeit findet in den letzten Jahren ihren vorläufigen Höhepunkt in der Auseinandersetzung mit dem anspruchsvollen Programmpaket IBM Watson. Es wurden und werden alle Anstrengungen unternommen und in dieser Publikation beschrieben, um die Wissensentwicklung und bereits realisierte und zukünftige Anwendungsmöglichkeiten dieser Zukunftstechnologie für die Bereitstellung von offenen Fachinformationen im ÖBH nutzbar zu machen.

ISBN: 978-3-903121-55-3