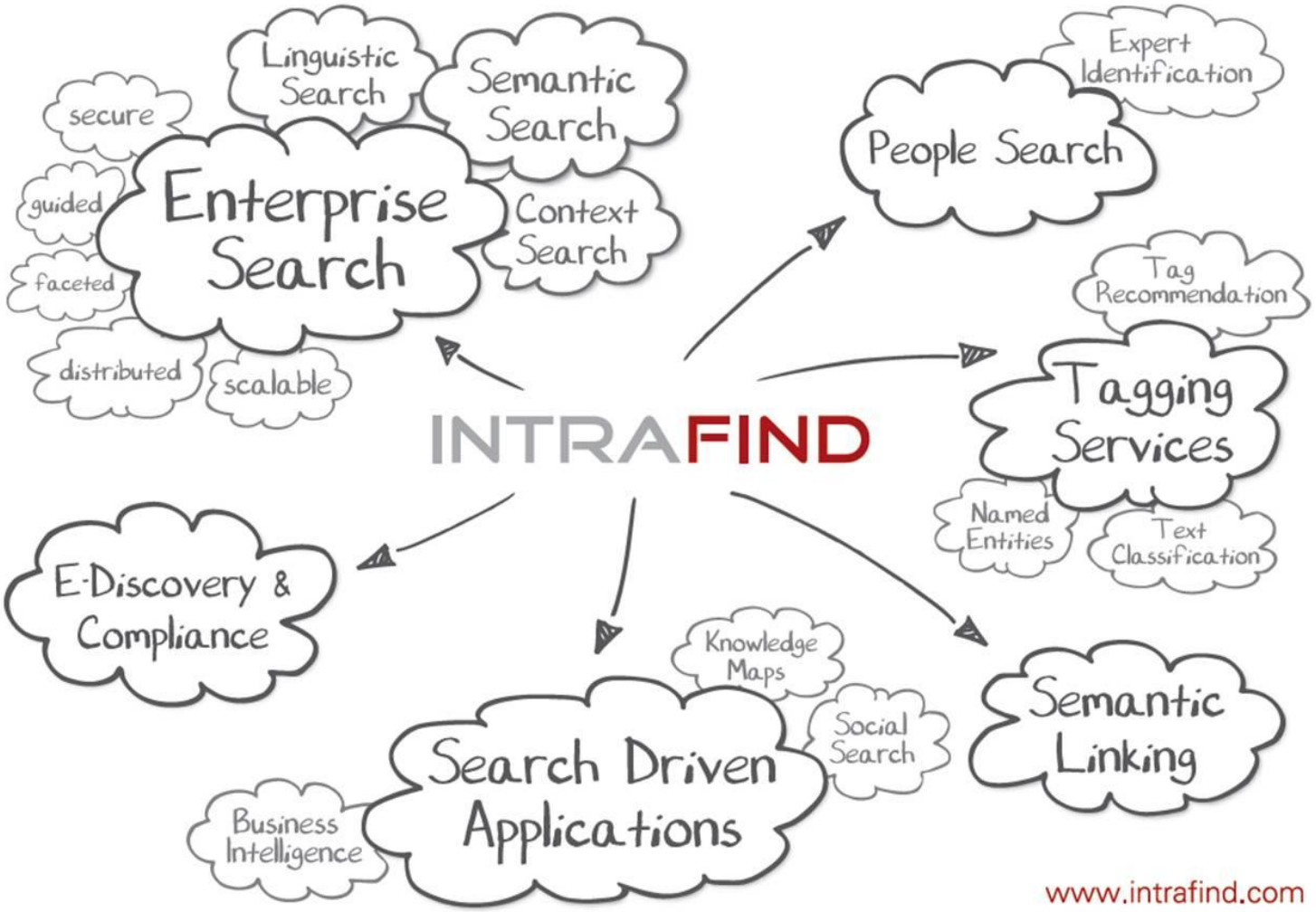
INTRAFIND

Solr-based Search & Automatic Tagging at Zeit Online – where Meta Data come from

ApacheCon Europe 2012
Dr. Christoph Goller, IntraFind Software AG

IntraFind Software AG

INTRAFIND



IntraFind Software AG

INTRAFIND

- Founding of the company: October 2000
- More than 700 customers mainly in Germany, Austria, and Switzerland
- Partner Network (> 30 VAR & embedding partners)
- Employees: 30
- Lucene Committers: B. Messer, C. Goller

Our Open Source Search Business:

- **Product Company**: iFinder, Topic Finder, Knowledge Map, Tagging Service, …
- Products are a combination of Open Source Components and in-house Development
- Support (up to 7x24), Services, Training, Stable API
- **Automatic Generation of Semantics**
  - Linguistic Analyzers for most European Languages
  - Semantic Search
  - Named Entity Recognition
  - Text Classification
  - Clustering

www.intrafind.de/jobs

Semantic Search @ Zeit Online

INTRAFIND

- “DIE ZEIT”: a German national weekly newspaper
- Free Access to 60 years of Print and Online publications, roughly 500.000 articles

Outline:

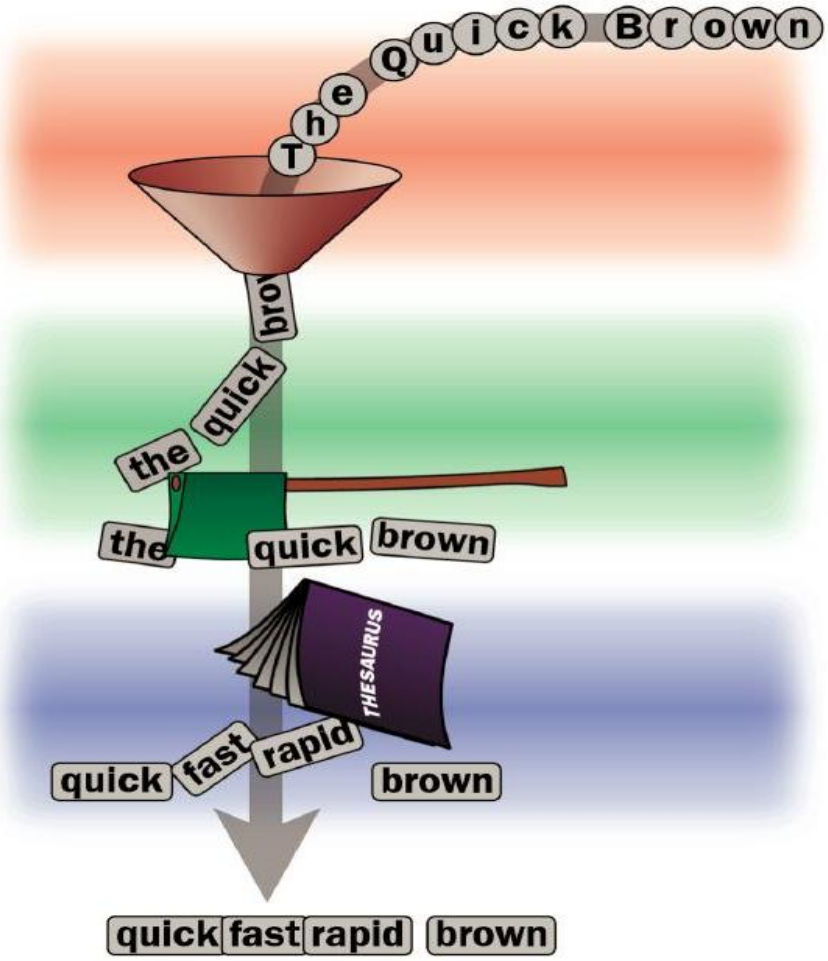
- Linguistically enhanced Search based on Solr using Intrafind Morphological Analyzers
- Automatic Tagging (Semantic Annotation) of Documents based on Intrafind Tagging Service
  - Statistical Keyword Extraction
  - Named Entity Recognition
  - Text Classification
- Future Improvements: Automatic Linking to Open Data using Apache Stanbol

Analysis / Tokenization

INTRAFIND

Break stream of characters into tokens /terms

- Normalization (e.g. case)
- Stop Words
- Stemming
- Lemmatizer / Decomposer
- Part of Speech Tagger
- Information Extraction



Morphological Analyzer vs. Stemming

INTRAFIND

Morphological Analyzer:

- **Lemmatizer**: maps words to their base forms

English			German		
going	→	go (Verb)	lief	→	laufen (Verb)
bought	→	buy (Verb)	rannte	→	rennen (Verb)
bags	→	bag (Noun)	Bücher	→	Buch (Noun)
bacteria	→	bacterium (Noun)	Taschen	→	Tasche (Noun)

- **Decomposer**: decomposes words into their compounds

  Kinderbuch (children‘s book) → Kind (Noun) | Buch (Noun)

  Versicherungsvertrag (insurance contract) → Versicherung (Noun) | Vertrag (Noun)

  Holztisch (wooden table), Glastisch (glass table)

Stemmer: usually simple algorithm

going → go king → k ???????????

Messer → mess ??????

Implementing a Lemmatizer / Decomposer

INTRAFIND

- Mapping inflected forms to base forms for all lemmas of a language
  - Finite State Techniques
  - German lexicon: about 100,000 base forms, 700,000 inflected forms
- Decomposition done algorithmically:

  Gipfelsturm – Gipfel+Sturm – Gipfel+Turm
  Staffelei – keine Zerlegung – Staffel+Ei
  Leistungen – keine Zerlegung – Leis(e)+tun+Gen
  Messerattentat – Messer+Attentat – Messe+Ratten+Tat
  Bundessteuerbehörde – Bund+Steuer+Behörde – Bund+ess+teuer+Behörde

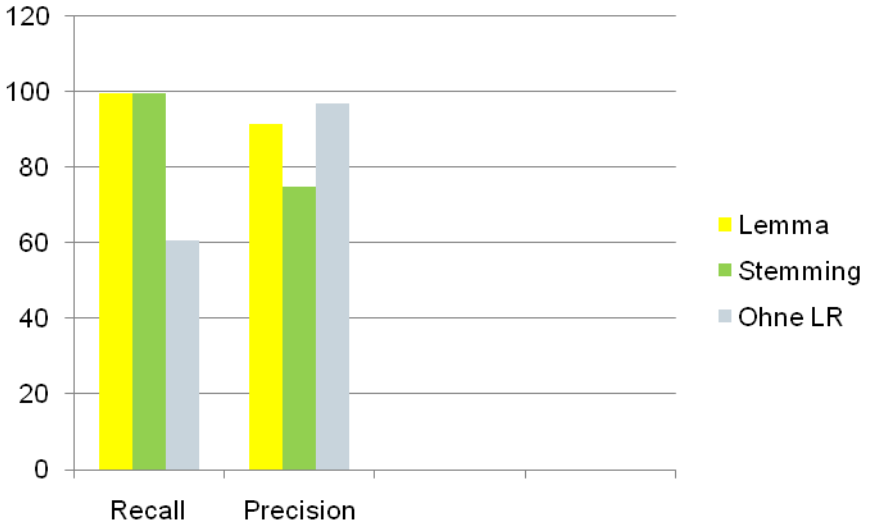
- Available Languages: German, English, Spanish, French, Italian, Dutch, Russian, Polish, Serbo-Croatian, Greek, (Chinese, Japanese, Arabian, Pasthu)

Advantages of Morphological Analysis

INTRAFIND

- Combines high Recall with high Precision for Search Applications
- Improves subsequent statistical methods
- Better suited as descriptions for faceting / clustering / autocomplete / spelling corrections than artificial stems
- Reliable lookup in lexicon resources
  - Thesaurus / Ontologies
  - Cross-lingual search

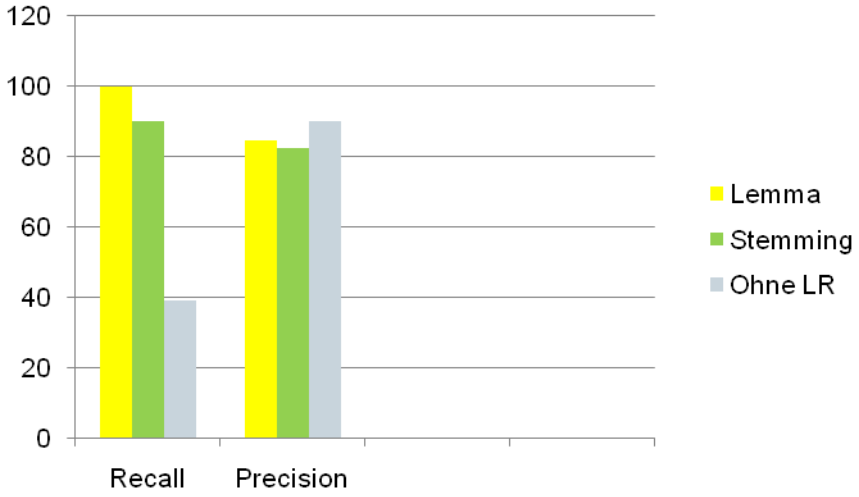
Measuring Recall and Precision



Nouns

Recall / Precision Macro Average for 40 German nouns

no compound analysis



Verbs

Recall / Precision Macro Average for 30 German verbs

no compound analysis

Bad Precision with Algorithmic Stemmer

INTRAFIND

[WEB] Apache Lucene/Solr - Who We Are

Lucene/Solr is maintained by a team of volunteer developers. Core Committers Bill Au (billa@...) Michael Busch (buschmi@...) Doron Cohen (doronc@...) Doug Cutting (cutting@...) Shai Erera (shaie@...) Erick Erickson (erick@...) Otis Gospodnetic (otis@...) Martijn van Groningen (mvg@...) Erik Hatcher (ehatcher@...) Chris Hostetter (hossman@...) Jan Høydahl (janhoy@...) Grant Ingersoll (gsingers@...) Mike McCandless (mikemccand@...) Ryan McKinley (ryan@...) Chris Male (chrism@...) Bernhard Messer (bmesser@...) Mark Miller (markrmiller@...) Robert Muir (rmuir@...) Stanislaw Osinski (stanislaw@...) Noble Paul (noble@...) Steven Rowe (sarowe@...) Uwe Schindler (uschindler@...) Shalin Shekhar Mangar (shalin@...) Yonik Seeley (yonik@...) Koji Sekiguchi (koji@...) Dawid Weiss (dweiss@...) Andi Vajda (vajda@...) Simon Willnauer (simonw@...) Emeritus Committers Josh Bloch Peter Carlson (carlson

http://lucene.apache.org/java/docs/whoweare.html

[LUCID] Information creation proliferation? Forbes looks to Solr and Lucid Imagination

2011-01-12 11:42

Information: we all love to make more of it, but it sure piles up. Forbes Magazine Online blogger Quentin Hardy takes on making sense of it in a nice post about Solr/Lucene and Lucid Imagination. Some soundbites: Our civilization may pride itself on the amount of information we create – more than every conversation, ever, in a couple of years; enough to jack The Library of Congress to the Moon on data-packed CD-Roms, take your superlative – but we've also made a holy mess of it. It is mostly, as they say, "unstructured," meaning as random as your last 25 emails, your tweets, and all those spreadsheets and documents piling up at work. That's why Solr/Lucene and Lucid Imagination are names you need to know in tech. ... That's because of the phenomenal

http://www.lucidimagination.com/blog/?p=2874

[LUCID] Information creation proliferation? Forbes looks to Solr and Lucid Imagination

Information: we all love to make more of it, but it sure piles up. Forbes Magazine Online blogger Quentin Hardy takes on making sense of it in a nice post about Solr/Lucene and Lucid Imagination. Some soundbites: Our civilization may pride itself on the amount of information we create – more than every conversation, ever, in a couple of years; enough to jack The Library of Congress to the Moon on data-packed CD-Roms, take your superlative – but we've also made a holy mess of it. It is mostly, as they say, "unstructured," meaning as random as your last 25 emails, your tweets, and all those spreadsheets and documents piling up at work. That's why Solr/Lucene and Lucid Imagination are names you need to know in tech. ... That's because of the phenomenal

http://www.lucidimagination.com/lucene-solr-blog/information-creation-prolife...

[LUCID] Estimating Memory and Storage for Lucene/Solr

2011-09-14 05:27

it for what you are actually seeing in your system. It is a DRAFT. It is likely missing a few things, but I am putting it up here and in Subversion as a means to gather feedback. I reserve the right to have messed up the calculations. I feel the values might be a little bit

http://www.lucidimagination.com/blog/?p=4000

High Recall and High Precision with Morphological Analyzer

INTRAFIND

Suchergebnisse für "buchen"

IHRE SUCHE buchen Suchen

DATUM FILTERN ALLE INHALTE | HEUTE | 24 STUNDEN | 7 TAGE | 30 TAGE

GENAUES DATUM VOM BIS

8044 ERGEBNISSE

Sortieren nach: RELEVANZ | DATUM Filter: ALLE INHALTE | NUR REZENSIONEN

STUDENTENANDRANG

Willkommen in der großen Maschine Universität

... rückt, der lieber dort als in Saarbrücken studiert usw. Darum müssen die Unis die Studiengänge "über**buchen**". Zum Semesterstart merken sie dann oft, dass sie sich geirrt haben: Das "Annahmeverhalten" war besser als geschätzt. Dann haben mehr Studierende [weiter...]

18.10.2011, ZEIT ONLINE

SACHBUCH

Liebe und solche Sachen

..., wird nie mehr in aller Unschuld in diesem angesagten Restaurant für einen Jahrestag der Liebe den teuren Fensterplatz **buchen** oder naiv die Reise zu zweit in den Süden für eine individuelle Entscheidung halten, also blind sein gegenüber dem Ausleben von [weiter...]

15.10.2011, DIE ZEIT

ZUFLUCHTSORT DEUTSCHLAND

Das gelobte Land

... Familie schon einmal einen Kurzurlaub in Deutschland **gebucht**, auf dem Oktoberfest waren sie. Die Lebensfreude der Deutschen habe seinen Kindern gefallen, sagt er. Der Rest habe sie eher kaltgelassen, vor allem das Essen. Er lächelt. »Aber vielleicht habe [weiter...]

03.10.2011, DIE ZEIT

HANDY-FAHRSCHEIN

Von der Deutschen Bahn verfolgt

... Endpunkt jeder Fahrt einzugeben. Das System berechnet den Fahrpreis und **bucht** ihn vom Konto des Nutzers ab. Selbstverständlich traut die Bahn ihren Kunden nicht und will überprüfen, ob deren Angaben stimmen. Sie sammelt dazu Bewegungsdaten. Das ist [weiter...]

27.09.2011, ZEIT ONLINE

SPITZENGASTRONOMIE

Frankreichs Köche setzen Jean-François Piège auf Platz 1

... ständig aus**gebucht**. Die Köche als Leser des Branchenmagazins Le Chef konnten den ihrer Ansicht nach besten Kollegen selbst bestimmen. Es gibt keine Liste oder Vorauswahl. Neben Piège kürten sie Alexandre Jean zum "Sommelier des Jahres [weiter...]

27.09.2011, ZEIT ONLINE

High Recall and High Precision with Morphological Analyzer

INTRAFIND

Suchergebnisse für "Buch"

IHRE SUCHE Buch Suchen

DATUM FILTERN ALLE INHALTE | HEUTE | 24 STUNDEN | 7 TAGE | 30 TAGE

GENAUES DATUM VOM BIS

93445 ERGEBNISSE

Sortieren nach: RELEVANZ | DATUM Filter: ALLE INHALTE | NUR REZENSIONEN

POLITISCHE GEFANGENE

Magischer Realismus

... Tagen hatte der alte Report des heute 84-jährigen Schriftstellers Platz eins auf Irans Bestsellerliste erklommen, wie die News-Webseite Aftab wissen ließ. Kurz darauf war das **Buch** ausverkauft. Was kümmert die Teheraner urplötzlich das [weiter...]

18.10.2011, DIE ZEIT

WINKLERS "GESCHICHTE DES WESTENS"

Das deutsche Kapitel

...Heinrich August Winkler Geschichte des Westens Die Zeit der Weltkriege 1914-1945 Politisches **Buch** C.H. Beck München 2011 1350 38 Der Westen [weiter...]

18.10.2011, DIE ZEIT

Wirtschaftspolitik

... Praxis bedeutet. "Es gibt keine Inflationsgefahren" (DIE ZEIT, Nr. 25/2010) Jürgen Stark, Chefvolkswirt der Europäischen Zentralbank, soircht in einem Interview über die Rettung kriselnder Staaten, die Grenzen des Lehr**buch**wissens und Merkels Sparpaket [weiter...]

18.10.2011, ZEIT ONLINE

THINKTANKS

Politische Vordenker gesucht

... besänftigen konnte. Cross-over: Institut Solidarische Moderne Der Name ist trügerisch. Unter einem Institut stellt man sich etwas anderes vor: ein paar schicke Räume, **Bücher**-Regale, Konferenztische und viele [weiter...]

18.10.2011, ZEIT ONLINE

ROMAN "GRUBER GEHT"

Ein Mann, ein Krebs

... es dann eben doch: Er, das Testosteronbündel, lässt Trost und Nähe zu. Es sind vor allem die beiläufigen, absurden Momente, die das **Buch** anrührend machen. Wie Gruber, schon vom Krebs gezeichnet, in seiner bislang unberührten Edelstahlküche [weiter...]

18.10.2011, ZEIT ONLINE

KOLONIALGESCHICHTE

Schädel im Schrank

... Arzt und Rasseforscher Eugen Fischer hatte sie »bestellt«. Von der Exekution existieren Fotos; auch beschrieb der Kolonialarzt Wilhelm Wendland 1939 in seinem **Buch** Im Wunderland der Papuas, wie er die Köpfe nach der Erschießung abschnitt [weiter...]

18.10.2011, DIE ZEIT

Solr Configuration: schema.xml

INTRAFIND

```
<fieldType name="text-IF" class="solr.TextField" positionIncrementGap="100">
    <analyzer type="index"
      class="org.apache.solr.analysis.IntrafindLiSaAnalyzerDeIndex"/>
    <analyzer type="query"
      class="org.apache.solr.analysis.IntrafindLiSaAnalyzerDeSearch"/>
 </fieldType>
```

Solr Configuration: solrconfig.xml

INTRAFIND

```
<queryParser name="IntrafindQueryParser" class= "org.apache.solr.analysis.
    IntrafindQParserPlugin">
    <lst name="generalConfig">
        <float name="linguisticBoost">5.0f</float>
        <bool name="disambiguationOnCase">false</bool>
        <bool name="disambiguationOnBaseEquality">true</bool>
    </lst>
    <lst name="compositaTreatment">
        <bool name="inCompositaSearch">true</bool>
        <int name="compositaSloppyness">3</int>
        <float name="boostExact">1.5f</float>
  </lst>
 </queryParser>
```

Statistical Keyword Extraction

INTRAFIND

- Extract most important keywords of a document using TF*IDF measure
- Identify Phrases
- Use POS (part of speech) tag patterns to identify good noun phrases

Named Entity Recognition (NER)

Automated extraction of information from unstructured data

- People names
- Company names
- Brands from product lists
- Technical key figures from technical data (raw materials, product types, order IDs, process numbers, eClass categories)
- Names of streets and locations
- Currency and accounting values
- Dates
- Phone numbers, email addresses, hyperlinks

Named Entities: Applications

INTRAFIND

- Facets
- Search for „Experts“
- Additional Query Types
  - Index Structure: Additional Tokens on the same position:
    - N_PersonName
    - N_Peter Müller
  - Search for a person named „Brown“(Semantic Search)
- Question Answering / Natural Language Queries
  - Search for a company near „founded“ and „Bill Gates“
- Part of our Tagging Services

Semantic Search: Comparison with standard Search Engines

INTRAFIND

Frage: **Wo liegen Werke von Audi?**

NL-Search

Audi.
....angehört. Die Fahrzeuge der Marke Audi werden außer in den beiden deutschen Werken Ingolstadt und Neckarsulm in Győr (Ungarn), Bratislava (Slowakei), Changchun (Volksrepublik China), Brüssel...

Google

Audi – Wikipedia
Die Fahrzeuge der Marke **Audi** werden außer in den beiden deutschen **Werken** Ingolstadt und **...** 1915 wurde die „**Audi Werke** AG" gegründet. Nachdem **Audi** 1928 in **...**

Audi und BMW: Neue **Werke** in Bayern - Börse - FOCUS Online
Audi-Chef Franz-Josef Paefgen stößt mit seinen **Werken** in Ingolstadt, Neckarsulm und im ungarischen Györ an die Kapazitätsgrenzen. Wie FOCUS- MONEY erfuhr, **...**

INTRAFIND

... höre" ins Lateinische „Audi". Im Juli 1910 verließ das erste Fahrzeug mit dem Namen Audi das Zwickauer Werk. 1915 wurde die „Audi Werke AG" gegründet. Nachdem Audi 1928 ...

Semantic Search: Comparison with standard Search Engines

INTRAFIND

Frage: **Wer hat Microsoft gegründet?**

NL-Search

Microsoft.
....sein Betriebssystem Windows und seine Büro-Software Office. Das Unternehmen wurde 1975 von Bill Gates und Paul Allen gegründet. Der Name Microsoft steht für Microcomputer-Software, ursprünglich...

Google

Microsoft – Wikipedia
Das Unternehmen wurde 1975 von Bill Gates und Paul Allen **gegründet**. **.....** Juli 2004 **hat Microsoft** bekanntgegeben, dass es nach der nun erfolgten Beilegung **...**

INTRAFIND

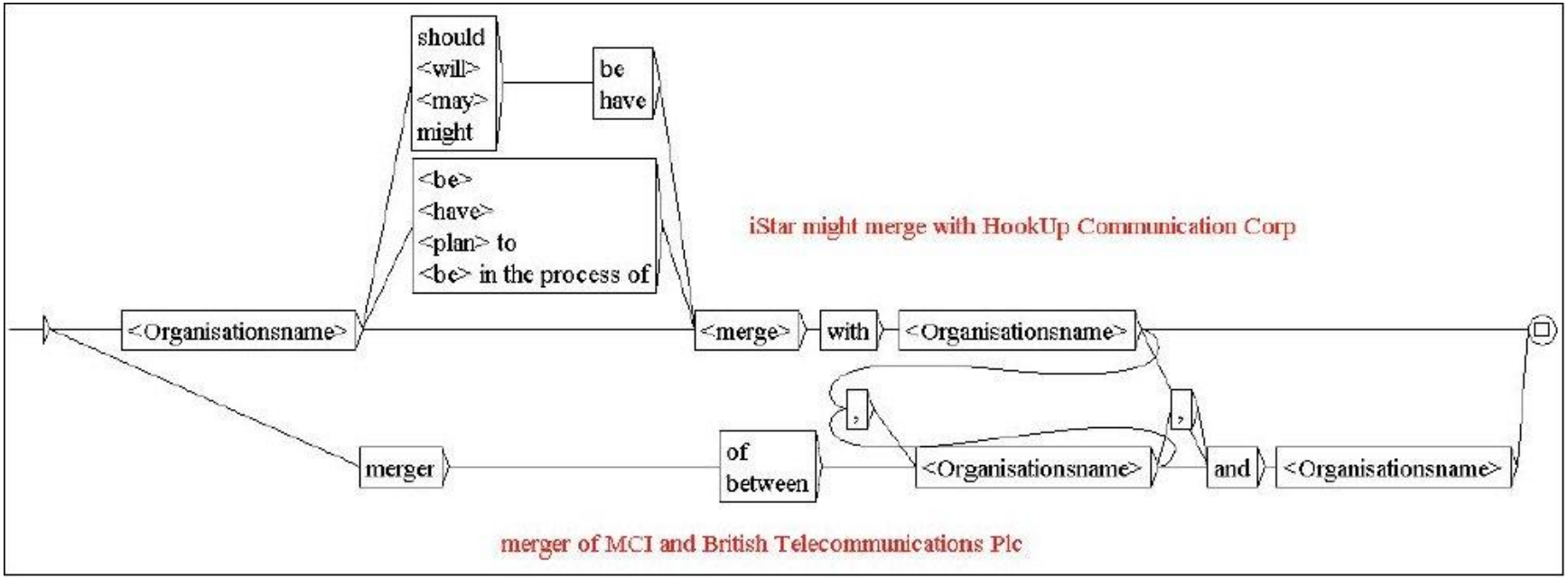
Microsoft Dynamics NAV ist eine Standardsoftware für ERP-Systeme. 2002 übernahm Microsoft den dänischen Hersteller und integrierte es in seinen Geschäftsbereich Microsoft Business Solutions. Seitdem wird Microsoft Dynamics ...

Implementing Named Entity Recognition

INTRAFIND

- Technology: gazetteers, local grammars (rule based), regular expressions
- Gate: open source platform for NLP (gate.ac.uk)

  GUI and Jape Grammars (all other components substituted due to stability issues)



Namer – Orthomatcher

INTRAFIND

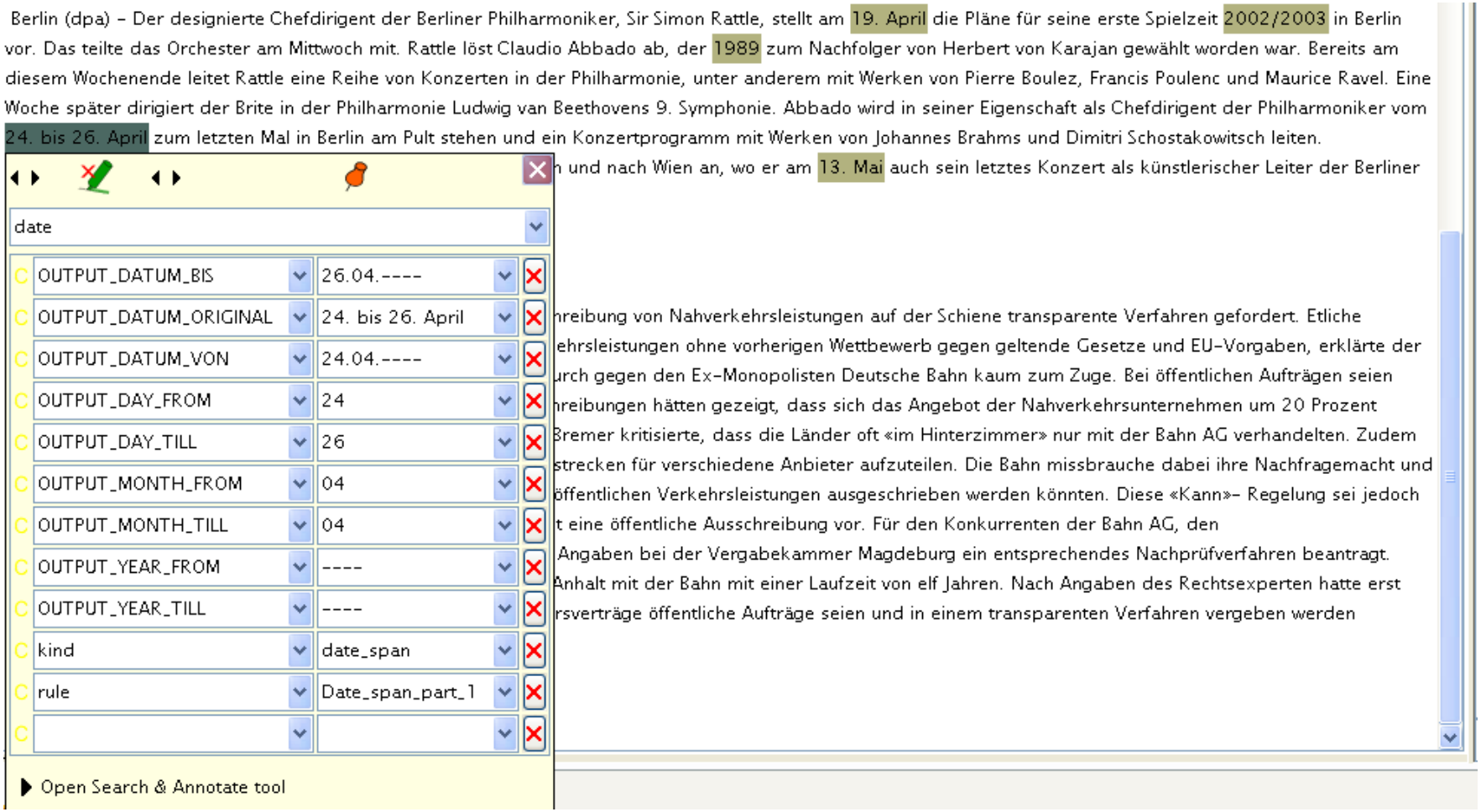
Tel Aviv (dpa) – Die Illusion ist schnell zusammengebrochen: Der blutige Einmarsch israelischer Truppen in die Palästinensergebiete hat dem Selbstmord-Terror gegen israelische Zivilisten kein Ende gesetzt. Ein zerfetzter Bus auf der Autobahn südlich von Haifa, Leichenteile an den Metallresten, Verletzte auf der Fahrbahn, neun Tote und 14 Verletzte – die Israelis wachten am Mittwochmorgen mit einem mulmigen Gefühl auf. Die zentrale Bushaltestelle in Haifa, wo der Attentäter 20 Minuten vor der Explosion eingestiegen war, galt als eine der sichersten im Lande. Dutzende von Wächtern kontrollieren dort die Reisenden und überprüfen ihre Gepäckstücke. Nach fast täglichen Selbstmord-Attacken über die Ostertage mit Dutzenden von Toten – Juden wie Moslems – war das Grauen zehn Tage lang zumindest auf den Straßen der israelischen Städte ausgeblieben. Doch vielen war klar: Früher oder später würde der alltägliche Kreislauf der Gewalt zwischen Israelis und Palästinensern wieder in Gang kommen. In den vergangenen Tagen meldete die Polizei immer häufiger Terror-Warnungen. Am Mittwoch ließ das Bekennerschreiben der radikal-islamischen Hamas-Bewegung nicht lange auf sich warten. Die israelische Regierung reagierte mit einem «jetzt erst recht». Die «Operation Schutzwall» werde fortgesetzt, beschloss das Sicherheitskabinett. Ministerpräsident Ariel Scharon weicht keinen Deut von seiner Grundhaltung ab, die palästinensische Intifada mit brachialer Gewalt zu zerschlagen. Doch der Busanschlag von Haifa liefert ihm nicht nur neue Argumente, sondern setzt ihn gleichzeitig innenpolitisch stärker unter Erfolgsdruck. Denn seine Unterstützung durch die Bevölkerung maß sich bisher am Sicherheitsgefühl der Israelis. Kurz vor Ostern auf dem Höhepunkt der Selbstmord-Serie sanken Scharons Sympathie-Werte auf etwa 40 Prozent. Dann gab er am 29. März den Befehl zum Einmarsch ins Westjordanland. Drei Tage später endete mit den Selbstmord-Anschlägen auch Scharons Stimmungstief. In jüngsten Umfragen unterstützen ihn mehr als 60 Prozent der Bevölkerung. Noch am Mittwochmorgen konnten die Israelis in der frisch gedruckten Ausgabe der Zeitung «Jediot Achronot» lesen: «Die Tatsache, dass es in zehn Tagen keinen Terroranschlag mehr gegeben hat, nährt die Illusion, dass der Krieg seine Ziele erreicht hat.» Wenn Scharon dies nicht einlöst, muss er mit wachsender Kritik am Preis der Besatzung rechnen. Angesichts Hunderter toter Palästinenser und der humanitären Katastrophe bringen Friedensorganisationen inzwischen Tausende von Menschen auf die Straße – statt weniger hundert wie noch vor zwei Wochen. Am Dienstag musste die Armee zudem mit 13 Toten die bisher größten Verluste durch einen Hinterhalt palästinensischer Kämpfer in Dschenin einräumen. Die umfangreichste Einberufung von Reservisten seit 20 Jahren kostet die Wirtschaft täglich zwei Millionen Euro. Schon zeichnet sich ab, dass Sozialleistungen angesichts der hohen Verteidigungsausgaben gekürzt werden müssen. Der Bombenattentäter von Haifa bringt Scharon in einem weiteren Punkt in Erklärungsnot. Bisher hatte er stets Jassir Arafat für alle Anschläge verantwortlich gemacht. Inzwischen glaubt kaum noch jemand, dass der seit Tagen in Ramallah eingeschlossene Palästinenserführer alle Aktionen der Palästinenser mit einem Mobiltelefon kontrollieren kann. Der Widerstand hat sich längst verselbstständigt, während die Autonomiebehörde zusammengebrochen ist. Trotzdem lässt Scharon keine weichere Haltung gegenüber Arafat erkennen. Den ausdrücklichen Wunsch des US-Außenministers Colin Powell, Arafat am Freitag zu besuchen, bezeichnete Scharon verschnupft als «tragischen Fehler». Israelische Journalisten vermuten, Scharon habe ohnehin mehr als die jetzige «Operation Schutzwall» im Hinterkopf. Nicht zufällig habe der Ministerpräsident am Montag in der Knesset den Plural benutzt, als er vom Militäreinsatz sprach. «Es wird noch viele «Schutzwälle» geben», befürchtet ein Kommentator.

Namer – Orthomatcher

INTRAFIND

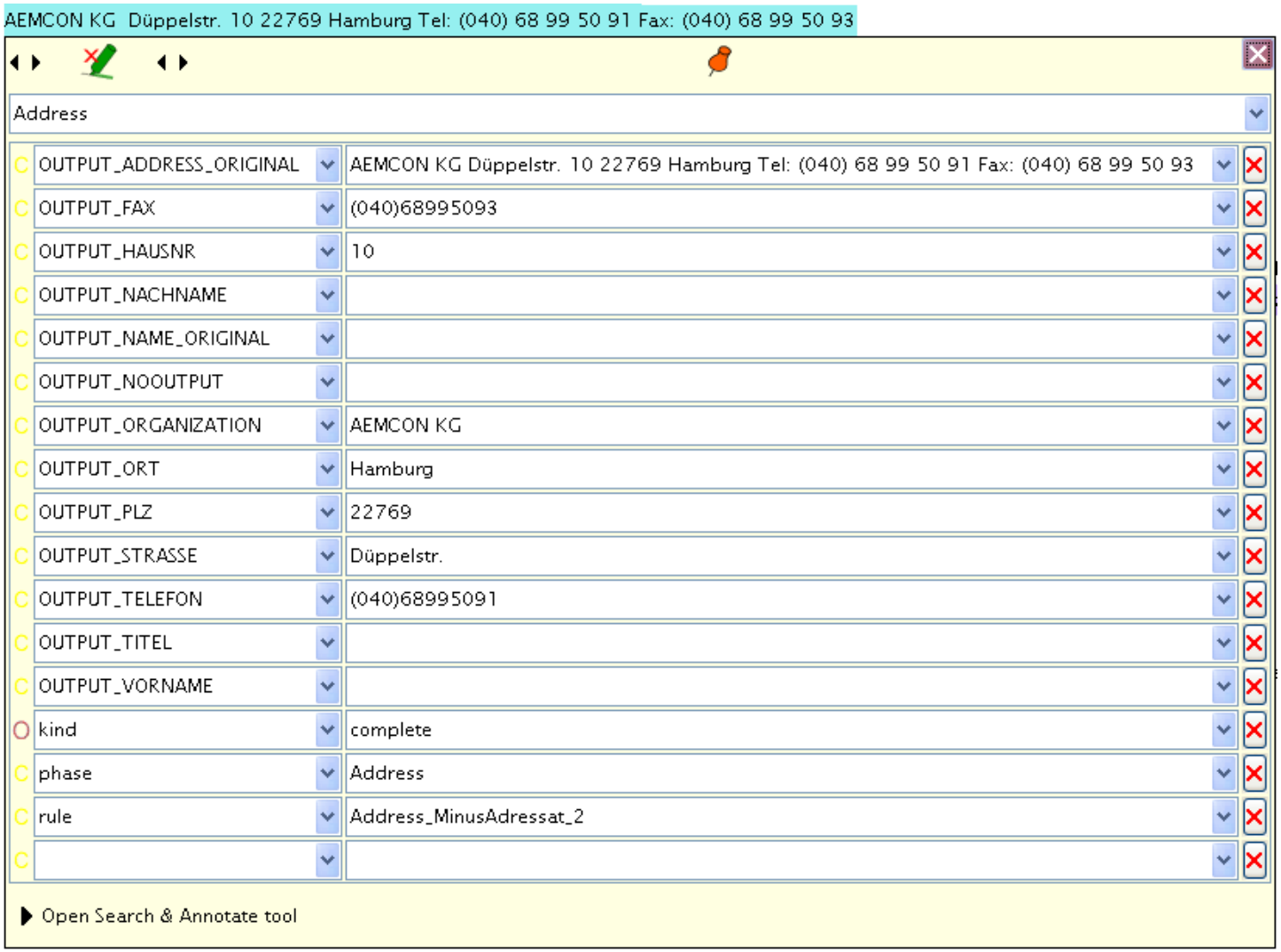
Tel Aviv (dpa) – Die Illusion ist schnell zusammengebrochen: Der blutige Einmarsch israelischer Truppen in die Palästinensergebiete hat dem Selbstmord-Terror gegen israelische Zivilisten kein Ende gesetzt. Ein zerfetzter Bus auf der Autobahn südlich von Haifa, Leichenteile an den Metallresten, Verletzte auf der Fahrbahn, neun Tote und 14 Verletzte – die Israelis wachten am Mittwochmorgen mit einem mulmigen Gefühl auf. Die zentrale Bushaltestelle in Haifa, wo der Attentäter 20 Minuten vor der Explosion eingestiegen war, galt als eine der sichersten im Lande. Dutzende von Wächtern kontrollieren dort die Reisenden und überprüfen ihre Gepäckstücke. Nach fast täglichen Selbstmord-Attacken über die Ostertage mit Dutzenden von Toten – Juden wie Moslems – war das Grauen zehn Tage lang zumindest auf den Straßen der israelischen Städte ausgeblieben. Doch vielen war klar: Früher oder später würde der alltägliche Kreislauf der Gewalt zwischen Israelis und Palästinensern wieder in Gang kommen. In den vergangenen Tagen meldete die Polizei immer häufiger Terror-Warnungen. Am Mittwoch ließ das Bekennerschreiben der radikal-islamischen Hamas-Bewegung nicht lange auf sich warten. Die israelische Regierung reagierte mit einem «jetzt erst recht». Die «Operation Schutzwall» werde fortgesetzt, beschloss das Sicherheitskabinett. Ministerpräsident Ariel Scharon weicht keinen Deut von seiner Grundhaltung ab, die palästinensische Intifada mit brachialer Gewalt zu zerschlagen. Doch der Busanschlag von Haifa liefert ihm nicht nur neue Argumente, sondern setzt ihn gleichzeitig innenpolitisch stärker unter Erfolgsdruck. Denn seine Unterstützung durch die Bevölkerung maß sich bisher am Sicherheitsgefühl der Israelis. Kurz vor Ostern auf dem Höhepunkt der Selbstmord-Serie sanken Scharons Sympathie-Werte auf etwa 40 Prozent. Dann gab er am 29. März den Befehl zum Einmarsch ins Westjordanland. Drei Tage später endete mit den Selbstmord-Anschlägen auch Scharons Stimmungstief. In jüngsten Umfragen unterstützen ihn mehr als 60 Prozent der Bevölkerung. Noch am Mittwochmorgen konnten die Israelis in der frisch gedruckten Ausgabe der Zeitung «Jediot Achronot» lesen: «Die Tatsache, dass es in zehn Tagen keinen Terroranschlag mehr gegeben hat, nährt die Illusion, dass der Krieg seine Ziele erreicht hat.» Wenn Scharon dies nicht einlöst, muss er mit wachsender Kritik am Preis der Besatzung rechnen. Angesichts Hunderter toter Palästinenser und der humanitären Katastrophe bringen Friedensorganisationen inzwischen Tausende von Menschen auf die Straße – statt weniger hundert wie noch vor zwei Wochen. Am Dienstag musste die Armee zudem mit 13 Toten die bisher größten Verluste durch einen Hinterhalt palästinensischer Kämpfer in Dschenin einräumen. Die umfangreichste Einberufung von Reservisten seit 20 Jahren kostet die Wirtschaft täglich zwei Millionen Euro. Schon zeichnet sich ab, dass Sozialleistungen angesichts der hohen Verteidigungsausgaben gekürzt werden müssen. Der Bombenattentäter von Haifa bringt Scharon in einem weiteren Punkt in Erklärungsnot. Bisher hatte er stets Jassir Arafat für alle Anschläge verantwortlich gemacht. Inzwischen glaubt kaum noch jemand, dass der seit Tagen in Ramallah eingeschlossene Palästinenserführer alle Aktionen der Palästinenser mit einem Mobiltelefon kontrollieren kann. Der Widerstand hat sich längst verselbstständigt, während die Autonomiebehörde zusammengebrochen ist. Trotzdem lässt Scharon keine weichere Haltung gegenüber Arafat erkennen. Den ausdrücklichen Wunsch des US-Außenministers Colin Powell, Arafat am Freitag zu besuchen, bezeichnete Scharon verschnupft als «tragischen Fehler». Israelische Journalisten vermuten, Scharon habe ohnehin mehr als die jetzige «Operation Schutzwall» im Hinterkopf. Nicht zufällig habe der Ministerpräsident am Montag in der Knesset den Plural benutzt, als er vom Militäreinsatz sprach. «Es wird noch viele «Schutzwälle» geben», befürchtet ein Kommentator.

Namer – Normalization



Namer – Aggregation

INTRAFIND



Text Classification

INTRAFIND

Goal:

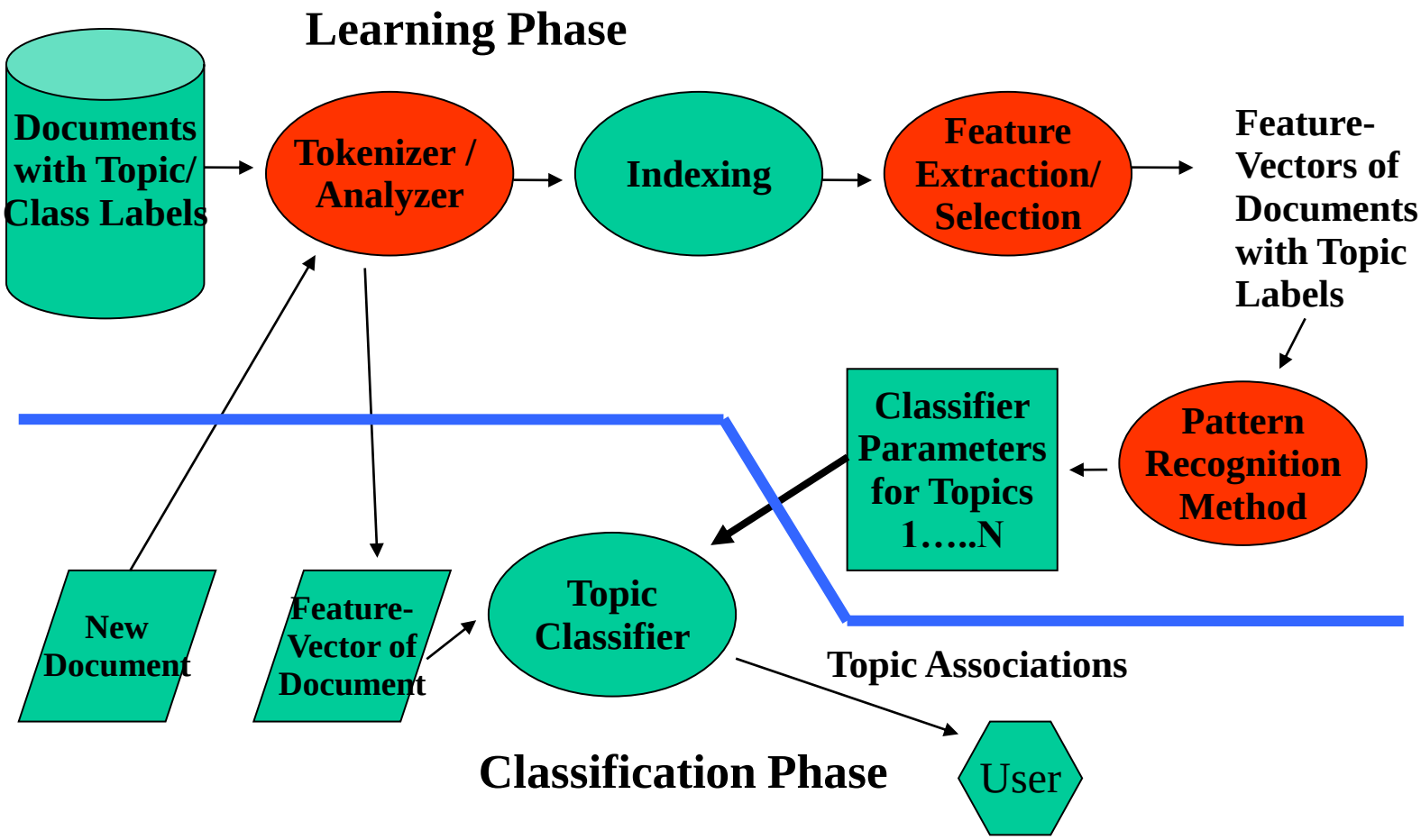
- Automatically assign documents to topics based on their content.
- Topics are defined by example documents.

Applications:

- News: Newsletter-Management System
- Spam-Filtering; Mail / Email Classification
- Product Classification (Online Shops), ECLASS /UNSPSC
- Subject Area Assignment for Libraries & Publishing Companies
- Opinion Mining / Sentiment Detection
- Part of our Tagging Services

Text Classification Workflow

INTRAFIND



Lessons Learned

INTRAFIND

- Analysis / Tokenization:
  - Normalization (e.g. Morphological Analyzers) and Stopwords improve classification
- Feature Selection:
  - TF*IDF, Mutual Information, Covariance / Chi Square, ...
  - Multiword Phrases, positive & negative correlation
- Machine Learning:
  - Goal: Good Generalization
  - Avoid Overfitting: „entia non sunt multiplicanda praeter necessitatem“ (Occam´s Razor)
  - SVM: linear is enough
- **Don’t trust blindly in**
  - **Manual Classification by Experts**
  - **Statistics / Machine Learning Results: Test !**

Required Features

INTRAFIND

- Training & Test GUI needed
- Automatically identify inconsistencies in training & test data
  - Duplicates detection
  - Similarity Search (More Like This)
- Automatic Testing: Cross-Validation (Multi-Threaded!)
- **Classification Rules have to be readable**
- **False Positive and (False Negative) Analysis,**
  - **Iterative Training**
  - **Clustering of False Positive / False Negative**

Product Classification: Example Rules

INTRAFIND

- **Server:**
  einbauschächte^24.7 | speicherspezifikation^22.1 | tastatur^-0.7 | monitortyp^21.5 | socket^-9.2 -1.15
- **Workstation:**
  monitortyp^28.8 | arbeitsstation^38.8 | cpu^0.1 | tower^8.9 | barebone^35.8 | audio^3.7 | eingang^5.2 | out^6.5 | core^9.0 | agp^5.2 -2.1
- **PC:**
  kleinbetrieb^7.9 | personal^18.3 | db-25^2.2 | technology^5.6 | cache^10.0 | arbeitsstation^-28.1 | dynamic^7.4 | bereitgestelltes^25.7 | dmi^5.5 | ata-100^13.7 | socket^6.2 | wireless^2.5 | 16x^10.0 | 1/2h^13.1 | nvidia^1.0 | din^4.6 | tasten^13.4 | international^7.2 | 802.1p^8.1 | level^-4.4 -1.5
- **Notebook:**
  eingabeperipheriegeräte^64.0 – 1.3
- **Tablet PC:**
  tc4200^16.4 | tablet^6.9 | konvertibel^10.6 | multibay^4.6 | itu^3.3 | abb^2.7 | digitalstift^8.5 | flugzeug^1.8 – 1.75
- **Handheld:**
  bildschirmauflösung^39.8 | smartphone^8.1 | ram^0.29 | speicherkarten^0.53 | telefon^0.35 - 1.4

Pharmaceutical Newsletter: Highlighting Example

INTRAFIND

Effects of vascular endothelial growth factor receptor inhibitor SU5416 and prostacyclin on murine lung metastasis Angiogenesis Weekly via NewsEdge Corporation : 2007 MAR 23 - (NewsRx.com) -- A report, "Effects of vascular endothelial growth factor receptor inhibitor SU5416 and prostacyclin on murine lung metastasis," is newly published data in Anti-Cancer Drugs. "The majority of patients with a diagnosis of cancer die from metastatic disease. Targeting specific steps in the metastatic process has the potential to improve patient outcomes," investigators in the United States report. "In this study, a novel lung metastasis model was developed by injecting DiI (1,1'-dioctadecyl-3,3,3',3'-tetramethylindocarbo-cyanine perchlorate)-labeled Lewis lung carcinoma cells into the tail vein of mice. The temporal development of tumor metastases was studied in the lung, liver and spleen. Additionally, the effects of vascular endothelial growth factor receptor inhibitor SU5416 and platelet activation inhibitor prostacyclin were tested in this metastasis model. Systemically injected Lewis lung carcinoma cells present in the lung at 15 min slowly accumulated in the liver and spleen reaching a peak at 4 days. After 8 days, tumor development was only evident in the lung. Use of SU5416 or prostacyclin lowered the initial density of Lewis lung carcinoma-labeled cells in the lung by a factor 1.8 and 2.3, respectively ($p < 0.05$). Furthermore, treatment with prostacyclin or SU5416 decreased lung weight by over 50% and the number of visible metastatic nodes by over 90% ($p < 0.05$). Combined treatment resulted in grossly normal lung tissue. Additionally, systemic treatment with prostacyclin reduced harvested metastatic cell adherence to endothelial cells by a factor of 10 and treatment with SU5416 attenuated vascular formation ($p < 0.001$)," wrote K.C. Cuneo and colleagues, Vanderbilt University, Department of Radiation Oncology. The researchers concluded: "SU5416 and prostacyclin effectively attenuated metastasis formation in this model. DiI labeling is an effective technique to monitor the temporal and spatial distribution of metastatic cells." Cuneo and colleagues published their study in Anti-Cancer Drugs (Effects of vascular endothelial growth factor receptor inhibitor SU5416 and prostacyclin on murine lung metastasis. Anti-Cancer Drugs, 2007;18(3):349-55). For additional information, contact K.C. Cuneo, Vanderbilt University School of Medicine, Dept. of Radiation Oncology, Nashville, Tennessee USA. This article was prepared by Angiogenesis Weekly editors from staff and other reports. Copyright 2007, Angiogenesis Weekly via NewsRx.com.
<<Angiogenesis Weekly -- 03/16/07>>

Implementation Details

INTRAFIND

- Training- and Test Documents are stored in a Lucene Index
- Information about topics is stored in a separate untokenized field
- Feature Selection simply consists of comparing posting lists of topics and terms form the text-content
- Consistency of manual topic-assignement can be checked by
  - using MD5-Keys for duplicates checks
  - Lucene's Similarity Search for checking for near duplicates
- Feature vectors are generated from Lucene posting lists
- Training is completely done by LibSVM / LibLinear
  - www.csie.ntu.edu.tw/~cjlin/libsvm
  - www.csie.ntu.edu.tw/~cjlin/liblinear
- Instead of storing support vectors, hyperplanes are stored directly

Tagging Service: Semantic Linking

INTRAFIND

Tagging Service: Generates semantic tags automatically

combines:

- Simple Statistical Tagging (TF*IDF) with Noun Phrase Identification
- Named Entity Recognition
- Text Classification

allows:

- Blacklists / Whitelists / BoostingLists

- Example: Semantic Linking for Zeit Online
  - http://www.zeit.de/schlagworte
  - http://www.zeit.de/schlagworte/themen
  - http://www.zeit.de/schlagworte/personen
  - http://www.zeit.de/schlagworte/organisationen
  - http://www.zeit.de/schlagworte/orte

Tagging Service at ZEIT Online

INTRAFIND

Abo | Shop | Audio | Apps | E-Paper | Newsletter | Archiv | Spiele | Blogs | Fotostrecken | Video | Schlagzeilen ZEITmagazin | ZEITCampus | ZEITGeschichte | ZEITWissen

ZEIT ONLINE | DEUTSCHLAND

Suchen

Partnersuche Immobilien Automarkt Jobs Reiseangebote

STARTSEITE POLITIK WIRTSCHAFT MEINUNG GESELLSCHAFT KULTUR WISSEN DIGITAL STUDIUM KARRIERE LEBENSART REISEN AUTO SPORT

Deutschland | Ausland

Anmelden | Registrieren

FINANZMARKTSTEUER

Kabinettsbeschluss soll Zweifler auf Linie bringen

SPD und Grüne trauen nicht der Zusage der Regierung, eine Finanzmarktsteuer einzuführen. Während Kanzlerin und Finanzminister beschwichtigen, bleibt die FDP skeptisch.

© Michael Kappeler/dpa

Finanzminister Wolfgang Schäuble im Kabinett (Archivbild)

Im Streit mit der Opposition um eine Finanztransaktionssteuer hat Bundesfinanzminister Wolfgang Schäuble auf einen Kabinettsbeschluss von

DATUM 11.06.2012 - 11:03 Uhr
SEITE 1 | 2 | Auf einer Seite lesen
QUELLE ZEIT ONLINE, Reuters, dpa, AFP
KOMMENTARE 19
VERSENDEN E-Mail verschicken
EMPFEHLEN Facebook, Twitter, Google+
ARTIKEL DRUCKEN Druckversion | PDF
SCHLAGWORTE Horst Seehofer | Wolfgang Schäuble | Angela Merkel | Transaktionsteuer | Finanzmarktregulierung | Finanzmarkt

NEU AUF ZEIT ONLINE

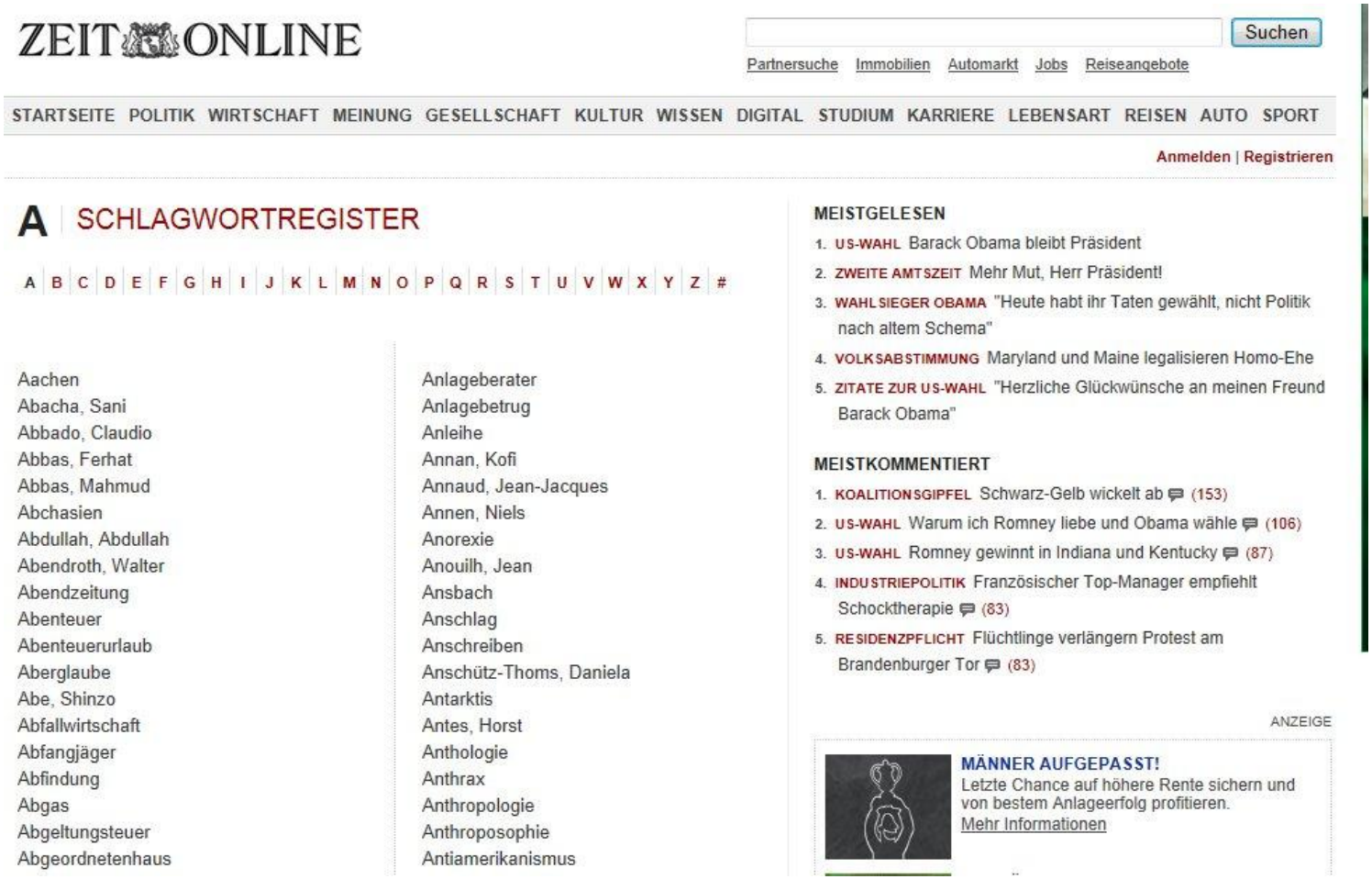
1. FINANZMARKTSTEUER Kabinettsbeschluss soll Zweifler auf Linie bringen
2. FRANKFURTS OBERBÜRGERMEISTERIN Die Liberale unter den Erzkonservativen
3. MOSKAU Razzia bei russischen Oppositionsführern
4. SPD BERLIN Wowereit verliert den Draht zur Partei
5. EUROPAMEISTERSCHAFT Kiews Frauen entscheiden die EM

NEU IM RESSORT

1. FINANZTRANSAKTIONSSTEUER Regierung will Zweifler mit Kabinettsbeschluss besänftigen
2. FRANKFURTS OBERBÜRGERMEISTERIN Die Liberale unter den Erzkonservativen
3. SPD BERLIN Wowereit verliert den Draht zur Partei
4. STUDIE Betreuungsgeld schadet laut OECD der Integration
5. FRANKREICH Linkes Lager gewinnt erste Wahlrunde deutlich

Tagging Service at ZEIT Online

INTRAFIND



Semantic Web, RDF Stores & Linked Data

INTRAFIND

Semantic Web proposed by Tim Berners-Lee, the founder of the WWW

Idea: Computers should be able to

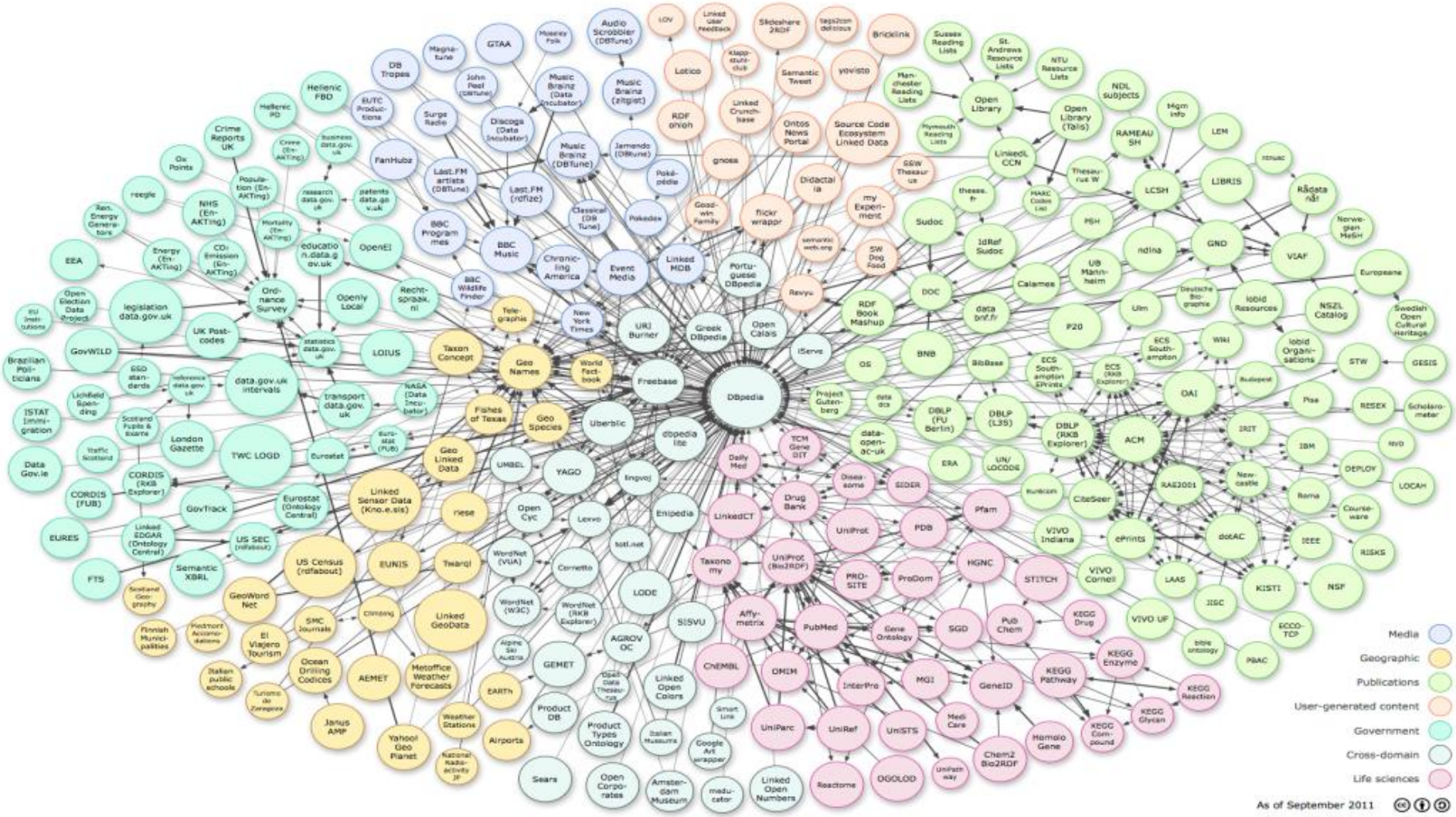
- evaluate information according to its meanings
- connect information
- reason with it (inference), generate new information

- RDF Stores:
  - Originally designed as Meta-Data model: machine-readable information
  - Triples: subject-predicate-object
  - General Data Model for Knowledge Representation
  - Query and Inference Languages: SPARQL

- Linked Open Data:
  - method of publishing structured data so that it can be interlinked
  - Uses RDF

Linked Open Data

INTRAFIND



Semantic Lifting

- Enhance content
  Stanbol Enhancer

- Manage entities
  Entityhub

- Store & search
  Contenthub

Questions?

INTRAFIND

Dr. Christoph Goller
Director Research

Phone: +49 89 3090446-0
Fax: +49 89 3090446-29
Email: christoph.goller@intrafind.de
Web: www.intrafind.de

IntraFindSoftware AG
Landsberger Straße 368
80687 München
Germany

www.intrafind.de/jobs