INTRAFIND

Semantic Search for the Enterprise

LT-Innovate Summit, June 19th, 2012,
Dr. Christoph Goller, Director Research, IntraFind Software AG

Emergence of Lucene and Solr

INTRAFIND

- Lucene / Solr
  - Built in late 90's by Doug Cutting…. Apache release 2001
  - Wide acceptance by 2005
- **4,000+ sites** – Apple, Cisco, EMC, HP, IBM, LinkedIn, MySpace, CNET, Netflix, Salesforce, Twitter, Ebay, Immoscout, Gov, Wikipedia…,
- **Rapid Innovation, Extensible Architecture**, complete control (open source)
- **CORE TECHNOLOGY AS GOOD OR BETTER THAN ANY OTHER … AND OPEN SOURCE**
- **Fulltext Search has become a commodity**
- **Good Enterprise Search Solutions have to offer more than just full-text search**

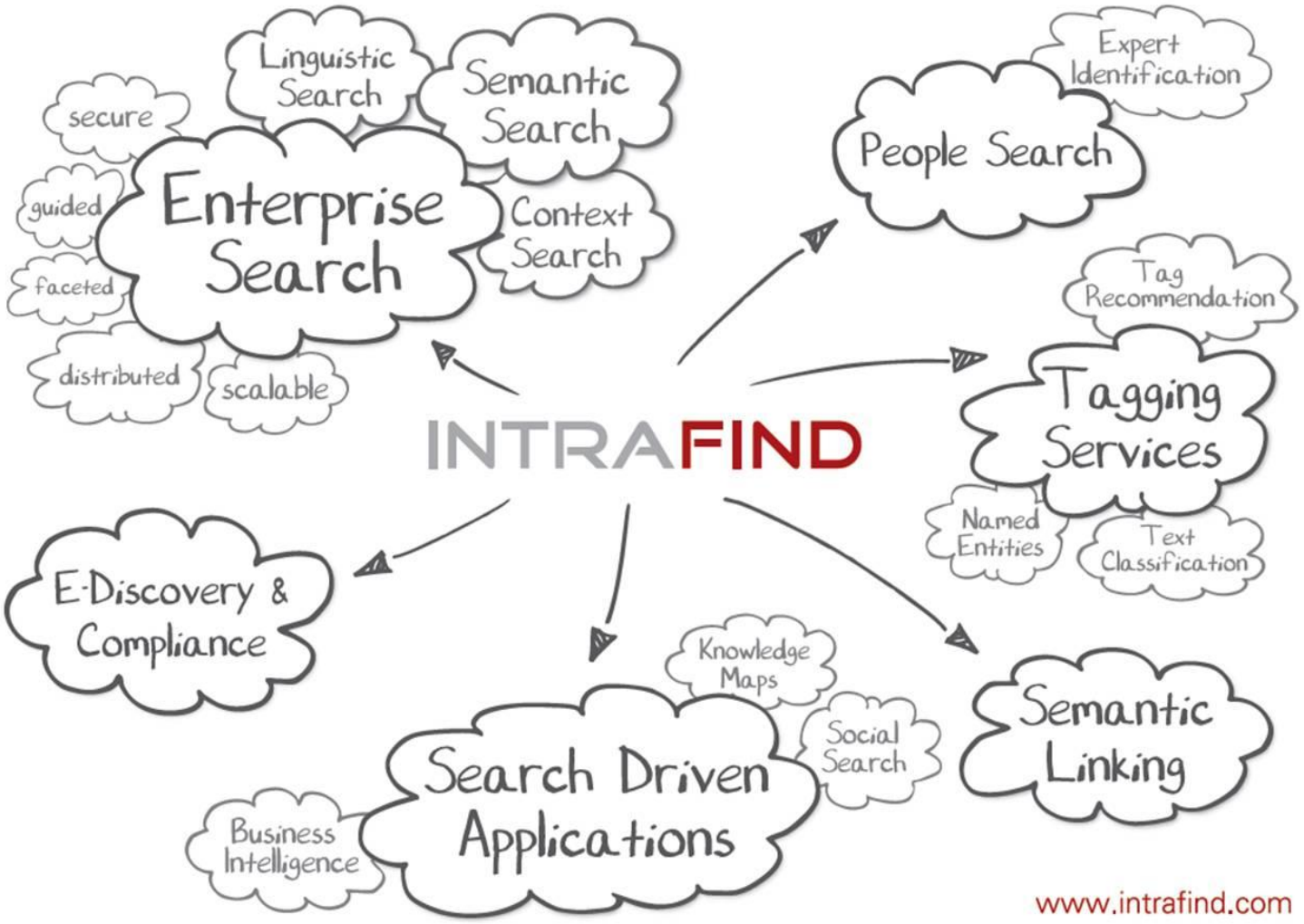
IntraFind Software AG

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

- **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
- **Text Analytics: Automatic Generation of Semantics**
  - Linguistic Analyzers for most European Languages (+ Chinese & Japanese)
  - Named Entity Recognition
  - Text Classification
  - Tagging Service
  - Clustering
  - **scalable, easily adaptable to new domains, easy to use for partners**

IntraFind Software AG

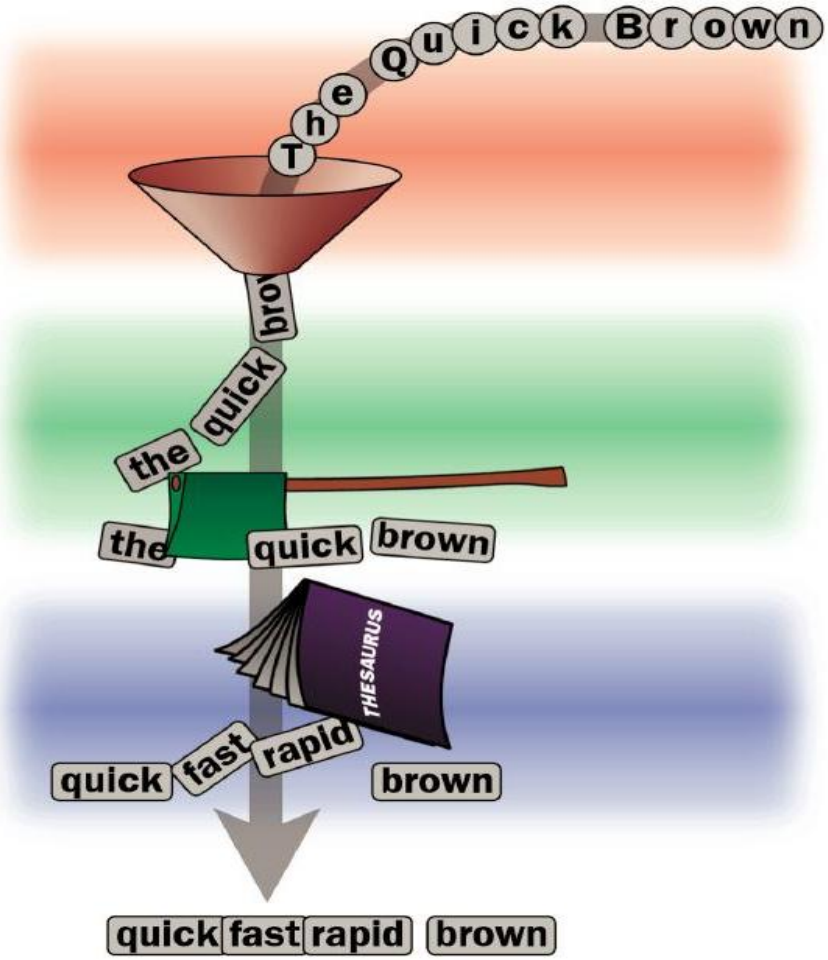


Analysis / Tokenization

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 (contract of insurance) -> Versicherung (Noun) | Vertrag (Noun)

  Holztisch (wooden table), Glastisch (table made of glass)

- **Stemmer:** usually simple algorithm

  going -> go

  Messer -> mess ??????

  king -> k ???????????

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

High Recall and High Precision with Morphological Analyzer

INTRAFIND

Suchergebnisse für "Buche"

IHRE SUCHE Buche Suchen

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

GENAUES DATUM VOM BIS

3991 ERGEBNISSE

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

KREUZWORTRÄTSEL (DRUCKVERSION)

Um die Ecke gedacht Nr. 2089

... Genossen-Gegenspieler, wie er in italienischem **Buche** steht 23 Wohl strebt die 32 waagerecht ihm entgegen, erreicht es aber nicht 27 Perfekte beeindrucken mit Plangenauigkeit 28 Verabredung mit Treffabsicht 30 Bisweilen noch wird zu Füllern gegriffen, wo's die zu [weiter...]

13.10.2011, DIE ZEIT

EXISTENZGRÜNDER

Hoch hinaus

... seiner Eltern geholt und sie wild in eine Rot**buche** getrieben. Dann, Jahre später, während der Semesterferien, baute sich Schelle ein zweites, größeres Baumhaus, eines zum Übernachten. Irgendwann übermannte ihn die Lust am Baumhausbauen, und so setzte er [weiter...]

08.10.2011, DIE ZEIT

VERBRAUCHERSCHUTZ

Was sollen wir essen?

... einige Tausend Liter Milch. Ihr Name ist Programm: Jeden zweiten Tag fährt der Milchsammelwagen mit den glänzenden Edelstahltanks entlang der Birken und **Buchen** zum Hof der Familie Haak. Er rollt an zwei Wohnhäusern vorbei und an dem Stall mit [weiter...]

04.10.2011, ZEIT ONLINE

Eurokrise: vor allem die Banken sind schuld

... Bundesbank mit 335 Milliarden Euro oder 13,4 Prozent des Sozialprodukts zu **Buche**. Wie war es zu der Krise gekommen? Vor allem die Banken der USA und Westeuropas hatten ihre Bilanzen vollgeladen mit Wertpapieren angeblich bester Bonität. Sie waren durch [weiter...]

23.09.2011, ZEIT ONLINE

NPD-Funktionär droht Verfahren in Tschechien

... Siener als Redner auftrat. Außerdem wurde in den Jahren 2010 und 2011 ein sogenannter „Tag der Freundschaft" in **Buchen**hof und Neustadt an der Waldnaab veranstaltet, an dem Neonazis aus beiden Ländern teilgenommen hatten. [weiter...]

23.09.2011, ZEIT ONLINE

TÜRKISCHE MIGRANTEN

Heimatsuche hin und zurück

...Wenn Emine Sahin Heimweh hat, geht sie in den Wald. Nördlich von Istanbul, unter den Eichen und **Buchen** des sogenannten Belgrader Waldes, lässt sie die hyperventilierende Metropole hinter sich. Sie fühlt sich hier ihrer unterfränkischen Heimat nah [weiter...]

21.09.2011, ZEIT ONLINE

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

- Search query: Which people / companies strongly correlate with each other?

Question Answering: 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 ...

Question Answering: 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...

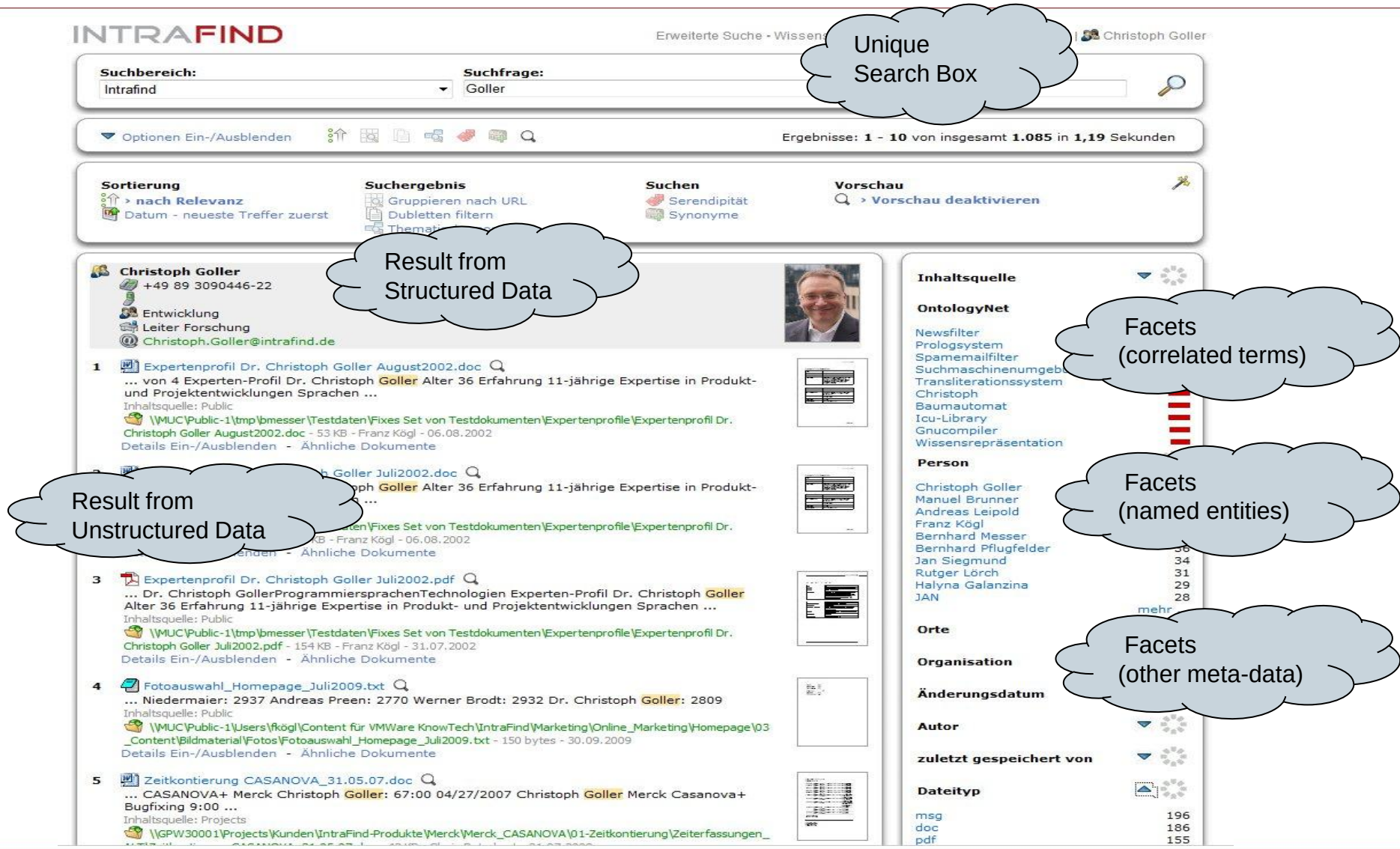
Google

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 ...

iFinder: Semantic Search and Navigation



Questions?

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