



## **Semantic Search for the Enterprise**

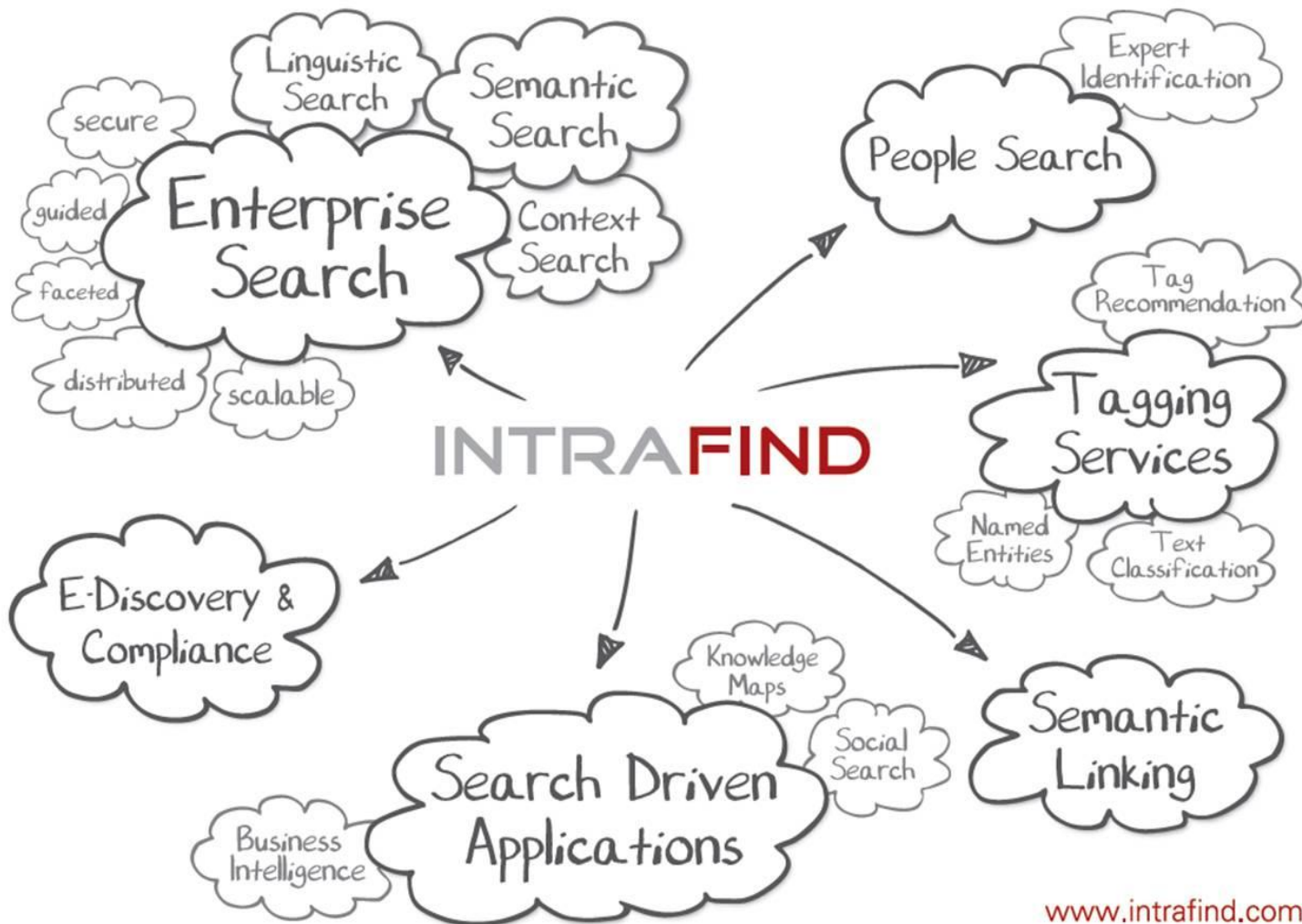
LT-Innovate Summit, June 19th, 2012,  
Dr. Christoph Goller, Director Research, IntraFind Software AG

## Emergence of Lucene and Solr

- ▶ 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**

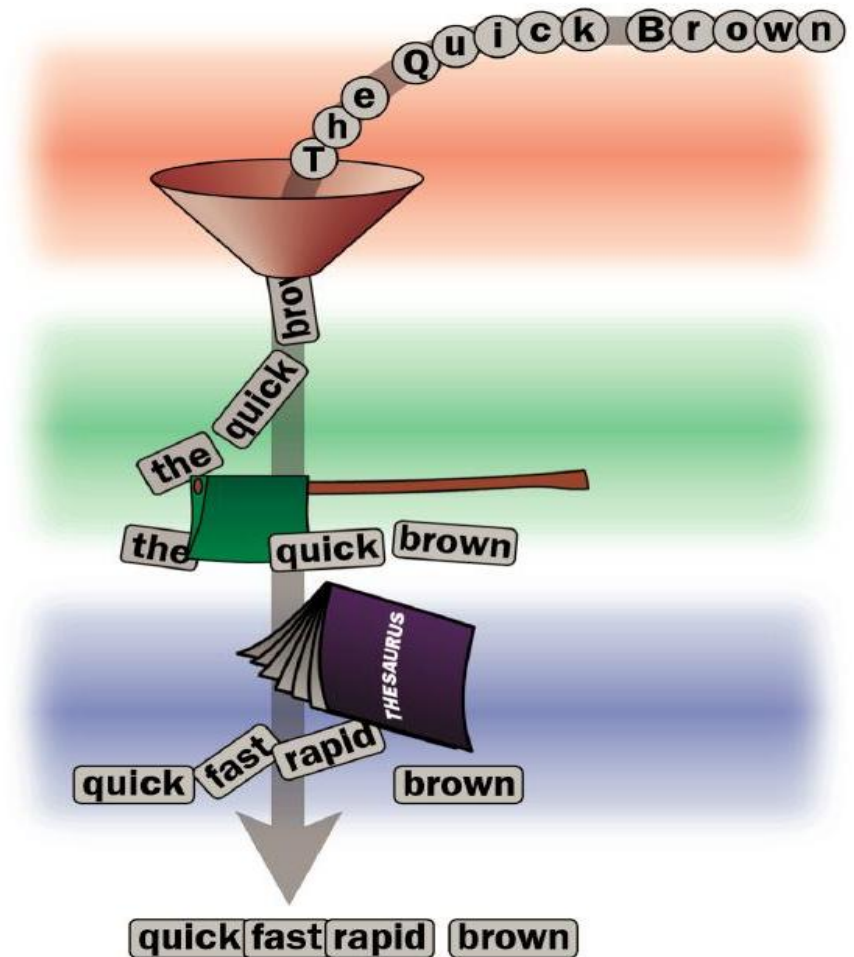


- ▶ Founding of the company: October 2000
- ▶ More than 700 customers mainly in Germany, Austria, and Switzerland
- ▶ Partner Network (> 30 VAR & embedding partners)
- ▶ Employees: 30
- ▶ Lucene Committers: B. Messer, C. Goller
  
- ▶ **Product Company:** iFinder, Topic Finder, Knowledge Map, Tagging 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**



## Break stream of characters into tokens /terms

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



## 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

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

Messer -> mess ???????

# Bad Precision with Algorithmic Stemmer

## [\[WEB\] 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 (markmiller@...) 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

DATUM FILTERN

GENAUES DATUM VOM   BIS

8044 ERGEBNISSE

Sortieren nach:

Filter:



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 **ausgebucht**. 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

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

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 waagrecht ihm entgegen, erreicht es aber nicht 27 Perfekte beeindrucken mit Plangenaugigkeit 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 Edelstahl tanks 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)

Die **Beiersdorf AG** ist als Dachmarke Hersteller zahlreicher Markenprodukte, darunter Marken wie Nivea, Labello, Hansaplast, Futuro, Eucerin, Florena oder Tesa. Außerdem gehören zur Kosmetiksparte die Marken Juvena, la prairie of Switzerland, 8x4, atrix und die Haarpflegeprodukte von Marlies Möller. Weiterhin stellt **Beiersdorf** verschiedene Körperpflegeprodukte (Basis PH, Doppel Dusch, Gammon) her.

Der Hauptsitz befindet sich in Hamburg, weitere deutsche Standorte sind Baden-Baden, Berlin, Emmerich am Rhein, Hannover, Heistersheim, Offenburg und Waldheim. Der Standort in Wien wird weiter als Zentrale für Mittel- und Osteuropa ausgebaut.

Nivea ist eine geschützte Marke der **Beiersdorf AG**. Die **1911** auf den Markt gekommene Hautpflegecreme NIVEA Creme ist das bekannteste Produkt der **Beiersdorf AG**. Den Namen leitete **Oscar Troplowitz** vom lateinischen Adjektiv niveus (zu nix, nivis, Schnee) ab, er bedeutet „die Schneeweiße“. Zuvor gab es bereits seit **1906** eine ebenfalls weiße Seife.

Zusammensetzung der **Hautpflegecreme**  
Grundlage war die Entdeckung von Eucerit, einem aus Schafswollfett gewonnenen Emulgator, dem ersten Wasser-in-Öl-Emulgator. **1911** entwickelte der Besitzer von **Beiersdorf**, **Oscar Troplowitz**, eine **Hautcreme** in enger Zusammenarbeit mit dem Chemiker **Isaac Lifschütz** und dem Dermatologen **Paul Gerson Unna**. Im Dezember desselben Jahres kam die erste **Hautcreme** der Welt mit langanhaltender Wirkung auf den Markt. Die Rezeptur ist seit den Anfangstagen nahezu unverändert geblieben: unter anderem **Glycerin**, **Panthenol**, **Zitronensäure**, **Wasser**, **Emulgator** (Eucerit) und **Duftstoffe**.

- Cat\_ADJECTIVE
- Cat\_NOUN
- Cat\_VERB
- City
- Organization
- Produkt
- PersonName
- Location
- OrgCompanyTemp
- Org\_Company
- ProduktMarke
- Rohstoff
- date

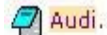
## Automated extraction of information from unstructured data

- ▶ 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

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

NL-Search



....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...



[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 Győr an die Kapazitätsgrenzen. Wie FOCUS- MONEY erfuhr, ...

INTRAFIND

... höre<sup>®</sup> 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

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...



[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

# INTRAFIND

INTRAFIND

Erweiterte Suche - Wissen

Christoph Goller

Suchbereich:

Intrafind

Suchfrage:

Goller

Unique Search Box

Optionen Ein-/Ausblenden

Ergebnisse: 1 - 10 von insgesamt 1.085 in 1,19 Sekunden

Sortierung

nach Relevanz

Datum - neueste Treffer zuerst

Suchergebnis

Gruppieren nach URL

Dubletten filtern

Themen

Suchen

Serendipität

Synonyme

Vorschau

Vorschau deaktivieren

Result from Structured Data

**Christoph Goller**  
+49 89 3090446-22  
Entwicklung  
Leiter Forschung  
Christoph.Goller@intrafind.de



1 Expertenprofil Dr. Christoph Goller August2002.doc  
... von 4 Experten-Profil Dr. Christoph Goller Alter 36 Erfahrung 11-jährige Expertise in Produkt- und Projektentwicklungen Sprachen ...  
Inhaltsquelle: Public  
\\MUC\Public-1\ntp\bmesser\Testdaten\Fixes Set von Testdokumenten\Expertenprofile\Expertenprofil Dr. Christoph Goller August2002.doc - 53 KB - Franz Kögl - 06.08.2002  
Details Ein-/Ausblenden - Ähnliche Dokumente



2 Expertenprofil Dr. Christoph Goller Juli2002.doc  
... von 4 Experten-Profil Dr. Christoph Goller Alter 36 Erfahrung 11-jährige Expertise in Produkt- und Projektentwicklungen Sprachen ...  
Inhaltsquelle: Public  
\\MUC\Public-1\ntp\bmesser\Testdaten\Fixes Set von Testdokumenten\Expertenprofile\Expertenprofil Dr. Christoph Goller Juli2002.doc - 154 KB - Franz Kögl - 06.08.2002  
Details Ein-/Ausblenden - Ähnliche Dokumente



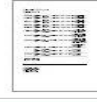
3 Expertenprofil Dr. Christoph Goller Juli2002.pdf  
... Dr. Christoph Goller Programmiersprachen Technologien Experten-Profil Dr. Christoph Goller Alter 36 Erfahrung 11-jährige Expertise in Produkt- und Projektentwicklungen Sprachen ...  
Inhaltsquelle: Public  
\\MUC\Public-1\ntp\bmesser\Testdaten\Fixes Set von Testdokumenten\Expertenprofile\Expertenprofil Dr. Christoph Goller Juli2002.pdf - 154 KB - Franz Kögl - 31.07.2002  
Details Ein-/Ausblenden - Ähnliche Dokumente



4 Fotoauswahl\_Homepage\_Juli2009.txt  
... Niedermaier: 2937 Andreas Preen: 2770 Werner Brodt: 2932 Dr. Christoph Goller: 2809  
Inhaltsquelle: Public  
\\MUC\Public-1\Users\Kögl\Content für VMware KnowTech\IntraFind\Marketing\Online\_Marketing\Homepage\03\_Content\Bildmaterial\Fotos\Fotoauswahl\_Homepage\_Juli2009.txt - 150 bytes - 30.09.2009  
Details Ein-/Ausblenden - Ähnliche Dokumente



5 Zeitkontierung CASANOVA\_31.05.07.doc  
... CASANOVA+ Merck Christoph Goller: 67:00 04/27/2007 Christoph Goller Merck Casanova+ Bugfixing 9:00 ...  
Inhaltsquelle: Projects  
\\GPW30001\Projects\Kunden\IntraFind-Produkte\Merck\Merck\_CASANOVA\01-Zeitkontierung\Zeiterfassungen\_31.05.07.doc - 150 bytes - 30.09.2009  
Details Ein-/Ausblenden - Ähnliche Dokumente



Inhaltsquelle

OntologyNet

Newsfilter  
Prologsystem  
Spamemailfilter  
Suchmaschinenumgeb  
Transliterationssystem  
Christoph  
Baumautomat  
Icu-Library  
Gnucompiler  
Wissensrepräsentation

Person

Christoph Goller  
Manuel Brunner  
Andreas Leipold  
Franz Kögl  
Bernhard Messer  
Bernhard Pflugfelder  
Jan Siegmund  
Rutger Lörch  
Halyna Galanzina  
JAN

Orte

Organisation

Änderungsdatum

Autor

zuletzt gespeichert von

Dateityp

msg	196
doc	186
pdf	155

Facets (correlated terms)

Facets (named entities)

Facets (other meta-data)

Result from Unstructured Data

# Questions?

**Dr. Christoph Goller**  
**Director Research**

Phone: +49 89 3090446-0

Fax: +49 89 3090446-29

Email: [christoph.goller@intrafind.de](mailto:christoph.goller@intrafind.de)

Web: [www.intrafind.de](http://www.intrafind.de)

IntraFindSoftware AG  
Landsberger Straße 368  
80687 München  
Germany