

Informationsextraktion

In der Chemie- und Life Science
Branche

Das Unternehmen Chemie.DE

- 1997 start als Forschungsprojekt des BMBF an der FU-Berlin
- 6/2000 Ausgründung als eigenständige GmbH
- 17 Mitarbeiter
- über 1.000 Kunden aus der Chemie- und Life Science Industrie
- 3 Standorte in Deutschland

Unternehmensbereiche

Marketing & Presse Services

Elektronische Newsletter

Banner- & Anzeigenwerbung

Produktredaktion

Personal Services

Stellenportale

Bewerberdatenbanken

Information Services

Markt- & Wettbewerbsbeobachtung

Newsfeeds für Internet & Intranet

Kundenspezifische Newsletter

Technical Services

Erstellung v. Webseiten & Portalen

Suchtechnologien

Datenbankentwicklung

XML-Technologien

Consulting Services

Online Marketing

E-Newsletter

Website Entwicklung

Marken



Fachportale für Chemie und Analytikindustrie

in D.A.CH. und Europa

1.5 Mio Seitenabrufe/Monat

260.000 Nutzer



Fachportal für Biotech- und Pharmaindustrie

280.000 Seitenabrufe/Monat

74.000 Nutzer



Spezialisierte Karriereportale für Chemie-, Pharma-, Biotech-Industrie

240.000 Seitenabrufe/Monat

62.000 Nutzer



Web 2.0 im Life-Science-Bereich

Kleine
Communities

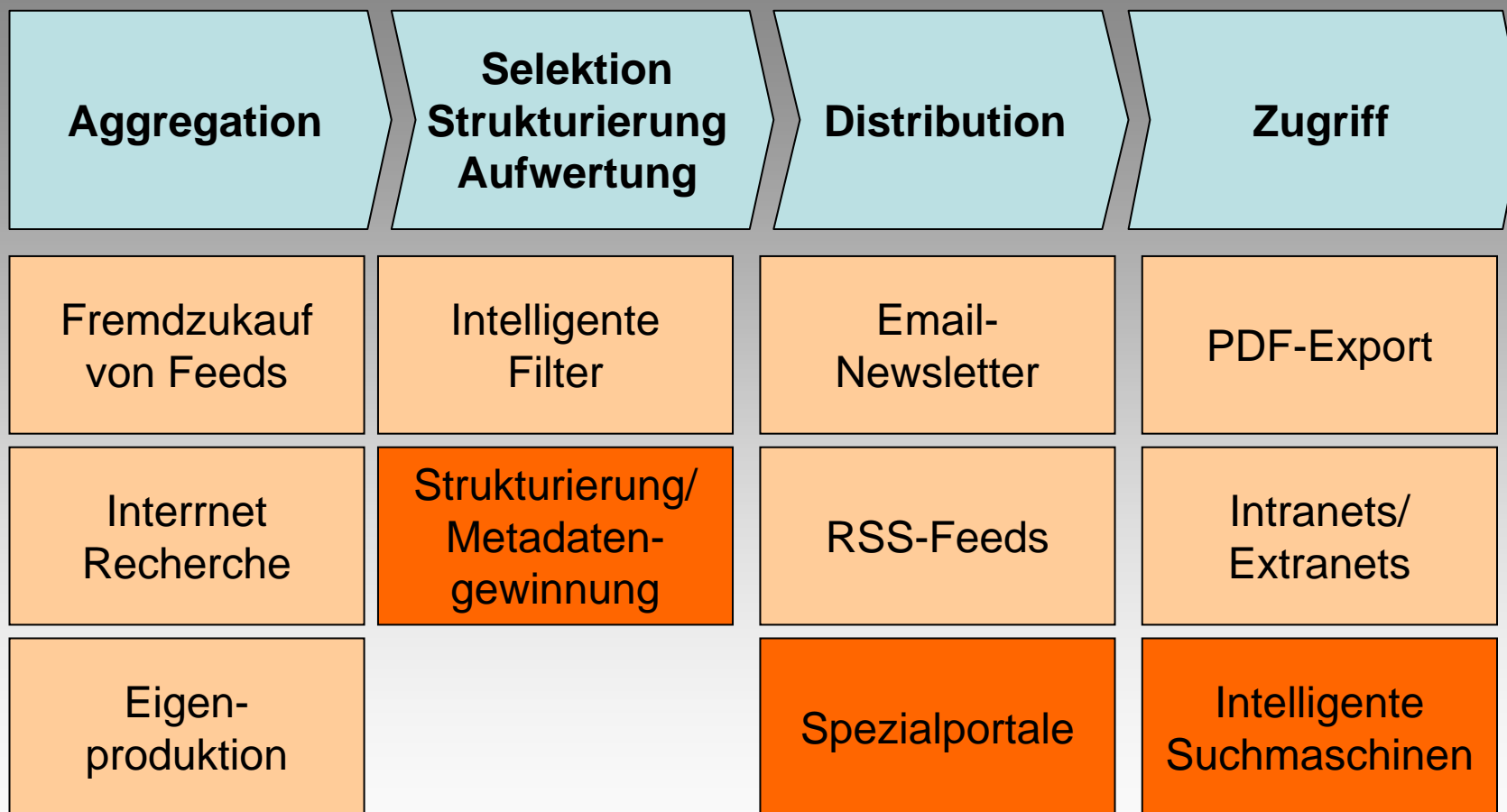
Hohe Qualitäts-
erwartungen

- Klass. Web 2.0-Ansätze scheitern an geringer aktiver Nutzerbeteiligung und Bedarf an Anonymität
- hohe Spezialisierung erfordert anwendungsorientierte Dienste
- Qualitätsanspruch und mangelnde Nutzerbeteiligung erfordern manuelle Redaktionelle Bearbeitung
- Innovative Zugriffsmöglichkeiten jenseits der gewöhnlichen Stichwortsuche
- Wirtschaftliche Interessen verhindern Preisgabe von Wissen und Identität

Zeit-
mangel

Anonymitäts-
anspruch

Die Wertschöpfungskette in der Informationsversorgung



**Szenario:
Extraktion von branchenrelevanter Information
aus Presstexten**

Motivation

Branchenspezifische Fragestellungen benötigen effizienten Zugriff auf Wissen:

Beispiele:

Welche Firmen sind mit Bayer assoziiert?

Welche Firmen sind mit dem Produkt "Aerosil" assoziiert?

Welche Personen sitzen bei Bayer CropScience im Vorstand?

Motivation

Gewinnung von relevanten Texten
(Crawler, Email, Fax-OCR etc.)

Gewinnung von Metadaten und Fakten

Erkennung der inhaltlichen Struktur

Kombination von Fakten und Struktur

Strukturierter Zugriff

Von: "Mummert Consulting" Joerg.Forthmann@presse.mummert.de
An: presseservice@chemie.de
Betreff: Presseinfo: Ahnungslose Mitarbeiter gefährden die IT-Sicherheit
Erstellt: 28.12.2004 10:06:37
Anlage: [file-1](#) [file-2](#)

Email-Header

Guten Morgen,

eine neue Nachricht von Mummert Consulting: Bitte beachten Sie die beigefuegte Presseinformation.

Mit freundlichen Gruessen
Joerg Forthmann

Anschreiben

Ahnungslose Mitarbeiter gefaehrden die IT-Sicherheit

Titel

Die eigenen Mitarbeiter sind ein grosses Risiko fuer die IT-Sicherheit. Versehentliche Fehler der Angestellten gefaehrden zunehmend die Informationssicherheit in deutschen Unternehmen. Nur drei von fuenf Beschaeftigten, so die Einschaeztung von IT-Verantwortlichen, wissen, wie sie [...] Diese Presseinformation basiert auf der Studie „IT-Security 2004“ der InformationWeek, die zusammen mit Mummert Consulting ausgewertet wurde. Von April bis Juni 2004 wurden dafuer 693 IT-Manager und Sicherheitsverantwortliche deutscher Unternehmen befragt.

Mitteilung

Unter dieser Internetadresse koennen Sie Ihr Profil editieren oder loeschen:
<http://217.111.5.68/mummertnews/abonnenten/index.php?e=803&u=1>
Ihr Passwort lautet: Q8aYkbqq4n

Mummert Consulting AG
Presse- und Oeffentlichkeitsarbeit
Joerg Forthmann

Organisation

Hans-Henny-Jahnn-Weg 29
22085 Hamburg
Tel.: 040/227 03-7787
Fax: 040/227 03-7961
Mail: Joerg.Forthmann@mummert.de

Ansprechpartner

Anschrift

Von: "Mummert Consulting" Joerg.Forthmann@presse.mummert.de
 An: pressservice@chemie.de
 Betreff: Presseinfo: Ahnungslose Mitarbeiter gefährden die IT-Sicherheit
 Erstellt: 28.12.2004 10:06:37
 Anlage: [file-1](#) [file-2](#)

Guten Morgen,

eine neue Nachricht von Mummert Consulting: Bitte beachten Sie die beigefuegte
 Presseinformation.

Mit freundlichen Gruessen
 Joerg Forthmann

Ahnungslose Mitarbeiter gefaehrden die IT-Sicherheit

Die eigenen Mitarbeiter sind ein grosses Risiko fuer die IT-Sicherheit.
 Versehentliche Fehler der Angestellten gefaehrden zunehmend die
 Informationssicherheit in deutschen Unternehmen. Nur drei von fuenf
 Beschaeftigten, so die Einschaeztung von IT-Verantwortlichen, wissen, wie sie
 [...] Diese Presseinformation basiert auf der Studie „IT-Security 2004“ der
 InformationWeek, die zusammen mit Mummert Consulting ausgewertet wurde. Von
 April bis Juni 2004 wurden dafuer 693 IT-Manager und Sicherheitsverantwortliche
 deutscher Unternehmen befragt.

 Unter dieser Internetadresse koennen Sie Ihr Profil editieren oder loeschen:
<http://217.111.5.68/mummertnews/abonnenten/index.php?e=803&u=1>
 Ihr Passwort lautet: Q8aYkbqq4n

Mummert Consulting AG
 Presse- und Oeffentlichkeitsarbeit
 Joerg Forthmann
 Hans-Henny-Jahnn-Weg 29
 22085 Hamburg
 Tel.: 040/227 03-7787
 Fax: 040/227 03-7961
 Mail: Joerg.Forthmann@mummert.de



Gewinnung der Trainingsdaten

Ziele:

- Effektive Unterstützung des Benutzers beim Trainieren des Extraktionssystems
- Minimierung der Zeit pro Annotation

Features:

- Texteditor für Textnormalisierung
- Annotationen-Farbe-System
- One-Click-Annotation-Engine
- Schnellmarkierung von Wörtern, Phrasen, Textblöcken
- Lernfähiges Annotations-Vorschlagsystem
- Textsuchmaschine Lucene zur Reduzierung der Textvorauswahl für gewünschte Annotation
- Annotationsansicht pro Text
- Annotationsansicht pro Textkorporus
- Tooltips für Annotationsprüfung
- Annotation persistent in AnnotationBundleXML
- Netzwerkfähig

Textsuchmaschine
Lucene

Editor für
Annotationen

The screenshot shows a software interface with three main panes. On the left is a 'Dateien' pane with a 'Suche' dropdown and a 'FileList' containing various 'Presstext' files. The middle pane is a 'Text - Vollzugriff' editor showing a document with several highlighted text blocks. On the right is a 'Tagset' pane with a legend of colored boxes and labels for different annotation types. At the bottom, a table shows a list of annotations with columns for 'Name', 'Annotationstyp', 'Start', 'Länge', and 'AnnotationText'.

Name	Annotationstyp	Start	Länge	AnnotationText
/user/common/Pr...UBERSCHRIFT	UBERSCHRIFT	119	149	
/user/common/Pr...KERNINFORMATION	KERNINFORMATION	171	699	
/user/common/Pr...KERNINFORMATION	KERNINFORMATION	872	868	
/user/common/Pr...FIRMENINFORMATION	FIRMENINFORMATION	1743	746	
/user/common/Pr...FIRMENINFORMATION	FIRMENINFORMATION	2491	442	
/user/common/Pr...FIRMENINFORMATION	FIRMENINFORMATION	2936	832	
/user/common/Pr...LINK	LINK	3771	76	

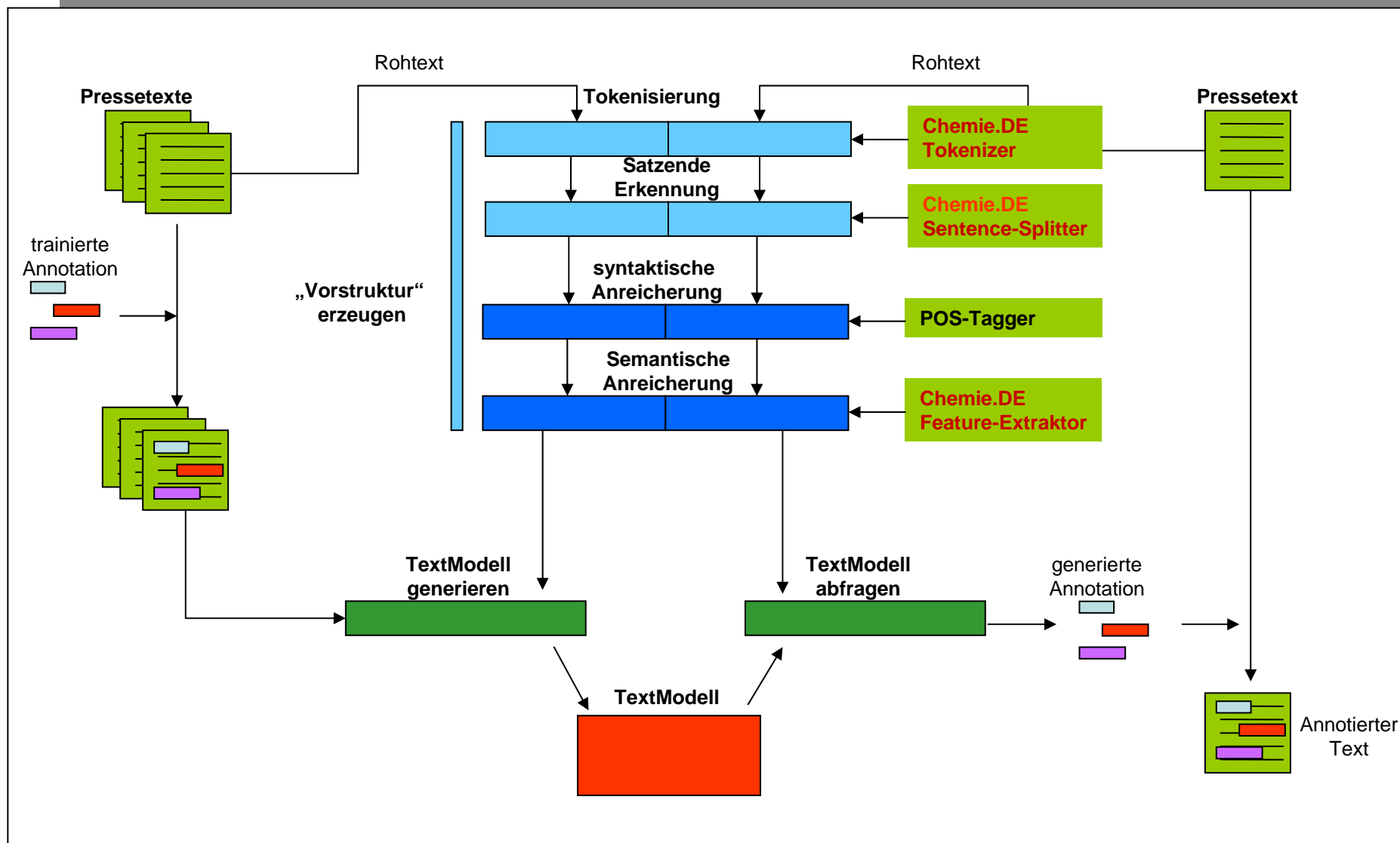
Ansicht für Dokumente
im Trainingskorporus

Ansicht Trainingskorporus-
annotationen

Ansicht Dokument-
annotationen

Ansicht für zur Verfügung
stehende Annotationen

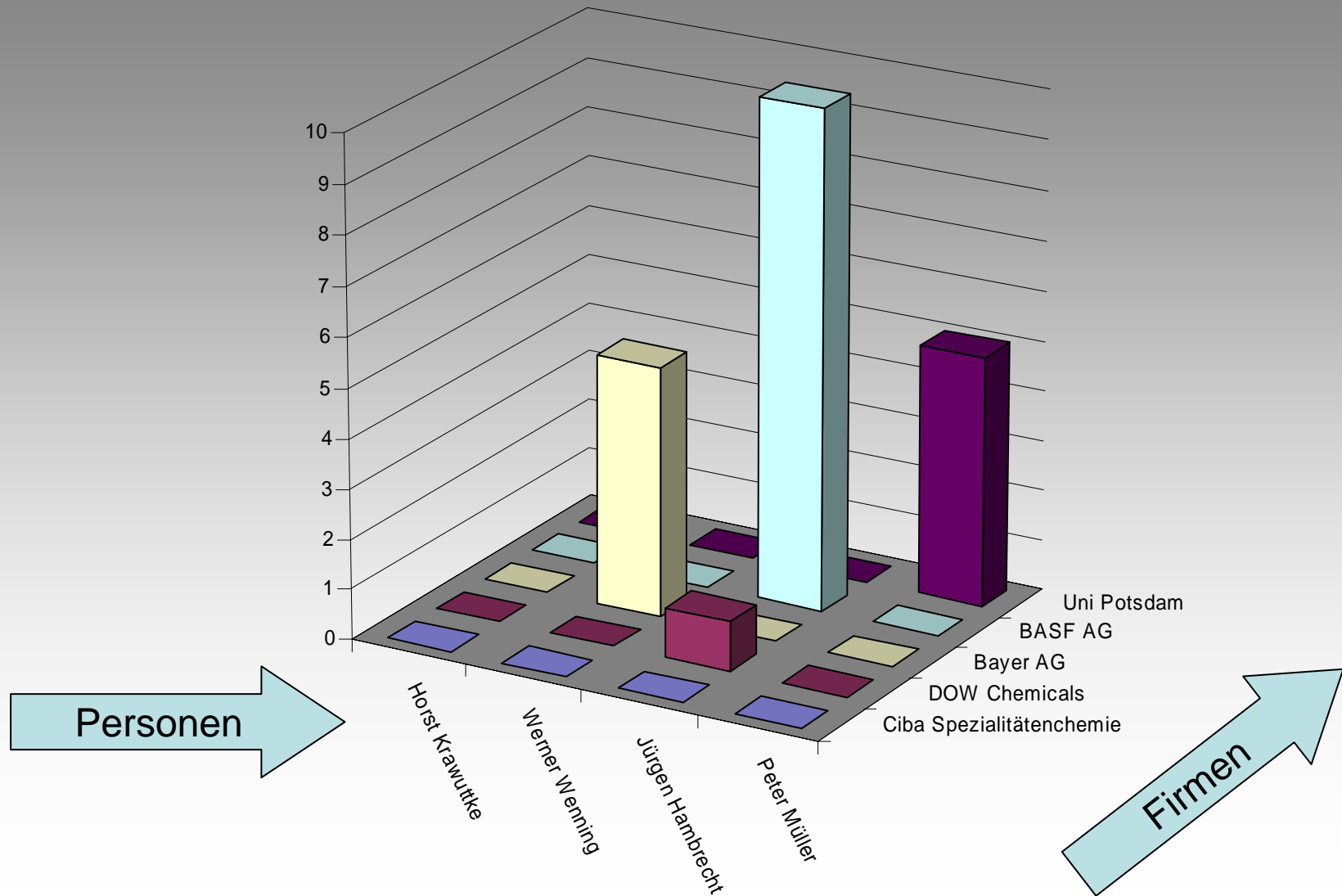
Das Strukturierungsverfahren



Live Demo

<http://pink.chemie.de>

Faktenanalyse auf großen Textbeständen



Vorteile der Extraktion

- **Struktur:** gezielter Zugriff auf inhaltlich relevante Teile eines Textes
- **Fakten:** Kombination zu anwendungsorientierten Diensten
- Training auch auf andere Materialien möglich
- **Ergonomie:** Höhere Nutzerfreundlichkeit als Stichwortsuchen
- **Zeitersparnis:** Minimierung des redaktionellen Aufwandes
- Verfügbarkeit aller Komponenten als Webservice