



Die SEMPRIA®-Suchmaschine: Mehr erreichen als mit Stichwortsuche

Eine Kurzpräsentation, 2011-10-12

SEMPRIA GmbH, Grafenberger Allee 277–287, 40237 Düsseldorf
<http://www.sempria.de/>

SEMPRIA GmbH, 2011-10-12

SEMPRIA Entstehung



- 1992 bis heute: Forschung der Arbeitsgruppe Intelligente Informations- und Kommunikationssysteme von Prof. Dr. Hermann Helbig (FernUniversität in Hagen) auf dem Gebiet der wissensbasierten Systeme
<http://pi7.fernuni-hagen.de/forschung/>
- 2003 bis heute: erfolgreiche Teilnahmen an internationalen Wettbewerben im Bereich Suche und Information Retrieval (CLEF)
- 2009: Prof. Helbig und drei langjährige Mitarbeiter gründen die SEMPRIA GmbH

SEMPRIA GmbH, 2011-10-12



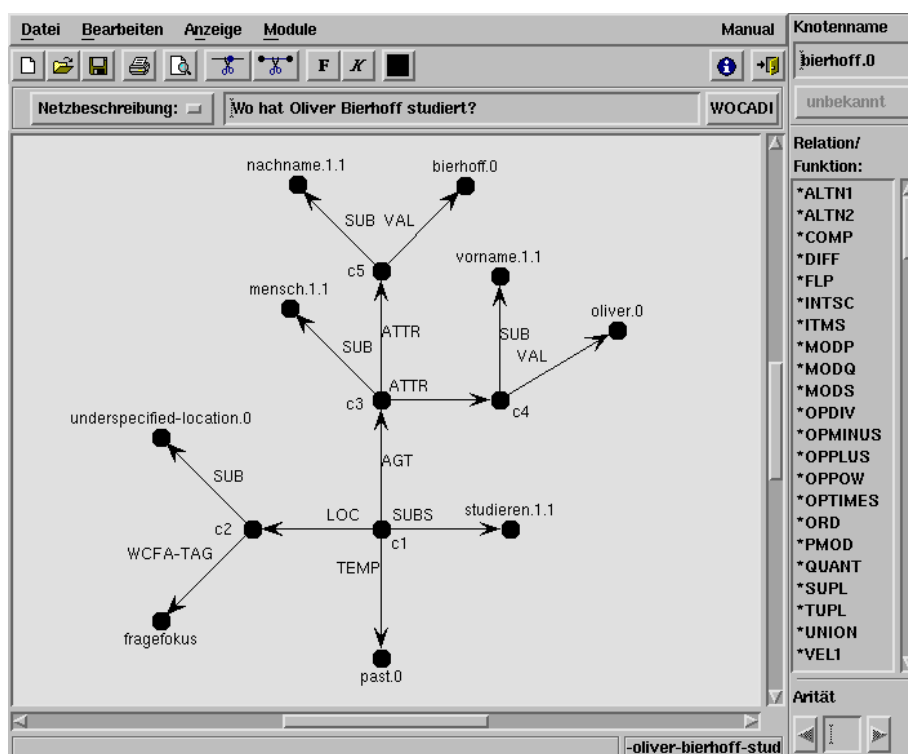
- ▶ SEMPRIA GmbH mit Sitz in Düsseldorf
- ▶ Kooperation mit der AG Intelligente Informations- und Kommunikationssysteme:
Leiter Prof. Helbig, Projektmitarbeiter, Diplomanden/Studenten
- ▶ Aktuelle Arbeitsschwerpunkte der SEMPRIA GmbH:
Vertrieb und Anpassung der bedeutungsorientierten Suchmaschine SEMPRIA® an individuelle Kundenbedürfnisse

SEMPRIA GmbH, 2011-10-12

SEMPRIA Kompetenz: Bedeutungsanalyse



Bedeutungsanalyse von Texten:
Natürliche Sprache \rightsquigarrow Bedeutungsdarstellung durch semantische Netze



SEMPRIA GmbH, 2011-10-12



- ▶ **Aufbau lexikalisch-semantischer Ressourcen**
 - Synonyme, Unter-/Oberbegriffe, Nominalisierungen, Schreibvarianten: semi-automatisch aus Textarchiven
 - 150.000 Beziehungen in Begriffs-Ontologien
 - ▶ **Aufbau logischer Regelsysteme**
 - Erfassung komplexer Beziehungen zwischen Wortbedeutungen
 - Interpretation von Funktionsverbgefügen und bildhaften Ausdrücken
 - mehrere tausend Axiome
 - ▶ **Automatische Erzeugung von Wissensbasen aus Texten**
 - z.B. Wikipedia (30 Millionen Sätze)
 - ▶ **Entwicklung von Werkzeugen zur Wissensakquisition**
 - Unterstützung des Wissensingenieurs beim semi-automatischen Wissenserwerb
- **Erfahrung im Aufbau und Nutzung großer Wissensbestände**

SEMPRIA GmbH, 2011-10-12



Kennzahlen

- ▶ 32.000 semantisch umfassende Einträge (alle linguistischen Beschreibungsebenen inkl. semantische Rollen und Komplemente)
 - ▶ 50.000 unterstützende Einträge (nur Morphologie und Syntax)
 - ▶ 350.000 Eigennamen in circa 50 Klassen
 - ▶ 1.500.000 Komposita mit Analysen, z.B. ((*Erbschafts (steuer)*) reform)
- **SEMPRIA[®] verwendet eines der größten semantischen Computerlexika**

SEMPRIA GmbH, 2011-10-12



Universell einsetzbare elementare Sprachtechnologien (Beispiele)

- ▶ Auszeichnen von benannten Objekten
- ▶ Auszeichnen mit Grundformen
- ▶ Markierung der Wortarten (auch Inhalts- versus Stoppwörter)
- ▶ Zerlegung von Komposita

dpa-Meldung vom 23.11.2009 (Auszug):

<MLA lemma=leicht> **Leichter** </MLA> <MLA cat=noun> **Austritt** </MLA> **von** <MLA cat=noun>
Radioaktivität </MLA>
<NE type=city> **Washington** </NE> (<NE type=company> **dpa** </NE>) -
In dem <COMPOUND parts=kern.1.1,kraftwerk.1.1> **Kernkraftwerk** </COMPOUND> <NE type=island> **Three Mile Island** </NE> **bei** <NE type=city> **Harrisburg** </NE> **im**
US-Bundesstaat <NE type=regional_institution> **Pennsylvania** </NE> **sind geringe**
Mengen Radioaktivität <MLA lemma=austreten> **ausgetreten** </MLA>.

SEMPRIA GmbH, 2011-10-12

SEMPRIA[®] Kernfunktionen der Suchmaschine



Zwei Kernfunktionen der SEMPRIA[®]-Suchmaschine:

▶ **Bedeutungsorientierte Suche**

Nutzer: *Import von Computern, ...* (verkürzte Anfragen oder verallgemeinernde Recherchen)

System: inhaltlich passende Textstellen

Häufigkeit in Such-Logs: $\geq 95\%$

Qualität: deutliche genauer und vollständiger als Stichwortsuche durch Einsatz von Bedeutungsanalyse

▶ **Fragebeantwortung**

Nutzer: *Wer/wen/was/wo/wann hat ..., ...*

System: knappe, präzise Antwort(en)

Häufigkeit in Such-Logs: $\leq 5\%$

Qualität: 30% bis 60% korrekte Antworten

Aber: durch Anpassung an die Domäne verbesserbar und ständige Verbesserung durch Fortentwicklung des Gesamtsystems

➔ **Mit SEMPRIA[®] Verschiebung in Richtung Fragebeantwortung**

SEMPRIA GmbH, 2011-10-12



Nutzeranfrage:

Import von Computern

Gefundene Textstellen (→ Vollständigkeit):

Einfuhr an Rechnern stieg ...

importierte 2010 ... Laptops aus ...

führt im nächsten Jahr ... Workstations aus

Verworfenne Textstellen (→ Genauigkeit durch semantische Filter):

um den Import mit Computern besser zu überwachen

SEMPRIA GmbH, 2011-10-12



- ▶ **Suchfunktion (im Web oder in-house) für Textarchive von:**
 - Zeitungen, Zeitschriften
 - Radio, Fernsehen
 - Enzyklopädien, Akten, Dokumentation ...
- ▶ **Alternatives Suchmodul für Content-Management-Systeme**
- ▶ **Unternehmensweite Suche (*enterprise search*)**
- ▶ **Archivierungsunterstützung**
 - Verschlagwortung
 - Verlinkung
 - Duplikatserkennung ...
- ▶ **Sentimentanalyse (*opinion mining, issue management*)**
in Vorbereitung

SEMPRIA GmbH, 2011-10-12



- ▶ SEMPRIA® **erhöht Anteil richtiger Ergebnisse: + 119%***
→ steigert die **Zufriedenheit** der Suchenden
- ▶ SEMPRIA® findet auf intelligentem Weg Dokumente, die eine Stichwortsuche nicht finden kann:
mehr (+ 39%) Suchanfragen mit mindestens einem richtigen Ergebnis*
→ maximiert das **Verwertungspotenzial** von Texten
- ▶ SEMPRIA® platziert richtige Ergebnisse weiter oben:
spart 47% Lesezeit bei Suchanfragen mit richtigen Treffern*
→ steigert **Effizienz** der Suchenden
- ▶ SEMPRIA® bietet erstmals eine bedeutungsorientierte Suche speziell für deutschsprachige Archive
→ sichert Ihnen einen markanten **Innovationsvorsprung**

* vorläufige Studie, jeweils 10 Treffer bewertet

SEMPRIA GmbH, 2011-10-12

SEMPRIA® Merkmale



Traditionelle Suchmaschinen Nutzung von Metadaten
Flexion, Derivation, Komposita (teils unvollständig)
facettierte Suche für einige Namensklassen

SEMPRIA® (zusätzlich)
Mehrwortausdrücke
Mehrdeutigkeiten (lexikalische, strukturelle)*
Metonymie, Idiome, Funktionsverbgefüge*
Zeitangaben (absolute und relative)
Zahlen, Maße
Bezüge von Pronomen*
Beziehungen zwischen Objekten
Rolle von Beteiligten in Ereignissen
Bedeutung aus mehreren Sätzen und Dokumenten*
semantische Suchvorschläge aus den Dokumenten
Fehlerrobustheit (Rechtschreibung, Zusammenschreibung)
facettierte Suche mittels voller Semantik (in Vorb.)

* abhängig vom verfügbaren Hintergrundwissen

SEMPRIA GmbH, 2011-10-12



Gehostete Suchlösung:

- ▶ flexibel
- ▶ keine Hardwarekosten, kein Installations-Aufwand, kein Update-Aufwand
- ▶ preiswerte Monatsgebühren

Randbedingungen:

- ▶ Vereinbarung der Formate für die zu durchsuchenden Archive
- ▶ Festlegung der Internet-Schnittstelle zur Einrichtung und Aktualisierung des Suchindexes

SEMPRIA GmbH, 2011-10-12

Impressum



SEMPRIA GmbH
Grafenberger Allee 277–287
40237 Düsseldorf

Telefon: 0211/566693-57
Fax: 0211/566693-58
Web: <http://www.sempria.de/>
E-Mail: info@sempria.de

Geschäftsführer: Dr. Sven Hartrumpf
Handelsregister: Amtsgericht Düsseldorf, HRB 62168
UStID-Nr: DE268248179

SEMPRIA GmbH, 2011-10-12