

SEMPRIA[®]: Mehr als Stichwortsuche

Eine Kurzpräsentation

Stand: 2010-04-01

SEMPRIA GmbH, Grafenberger Allee 277–287, 40237 Düsseldorf
Telefon: 0211/566693-57, Fax: -58
Web: www.sempria.de, E-Mail: info@sempria.de



- ▶ Entwicklung der SEMPRIA GmbH und wissenschaftlicher Hintergrund
- ▶ SEMPRIA-Technologie
- ▶ SEMPRIA[®]: Systemmerkmale und Beispiele
- ▶ SEMPRIA[®] für unsere Kunden



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



- ▶ SEMPRIA GmbH mit Sitz in Düsseldorf:
gestartet mit 2 Mitarbeitern und 3 freien Mitarbeitern

- ▶ Kooperation mit der AG Intelligente Informations- und Kommunikationssysteme:
Leiter Prof. Helbig, 3 Projektmitarbeiter, Diplomanden/Studenten

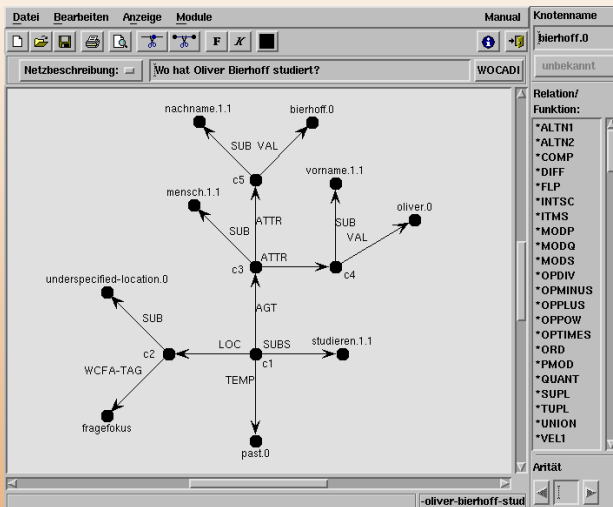
- ▶ Aktuelle Arbeitsschwerpunkte der SEMPRIA GmbH:
Anpassung und Vertrieb der bedeutungsorientierten Suchmaschine SEMPRIA[®]



SEMPRIA Kompetenz: Bedeutungsanalyse

Bedeutungsanalyse von Texten:

Natürliche Sprache \rightsquigarrow Bedeutungsdarstellung durch semantische Netze



- ▶ **Automatische Erzeugung von Wissensbasen aus Texten**
 - z.B. Wikipedia (21 Millionen Sätze)
- ▶ **Entwicklung von Werkzeugen zur Wissensakquisition**
 - Unterstützung des Wissensingenieurs
 - semi-automatischer Wissenserwerb
- ▶ **Aufbau lexikalisch-semantischer Ressourcen**
 - Synonyme, Unter-/Oberbegriffe, Nominalisierungen, Schreibvarianten: semi-automatisch aus Text-Corpora
 - 150.000 Beziehungen zwischen Begriffen
- ▶ **Aufbau logischer Regelsysteme**
 - Erfassung komplexer Beziehungen zwischen Wortbedeutungen
 - Interpretation von Funktionsverbgefügen und bildhaften Ausdrücken
 - mehrere 1000 Axiome

⇒ **Umfangreiche Erfahrung in Aufbau und Nutzung großer Wissensbestände**



Kennzahlen

- ▶ 30.000 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
- ▶ 500.000 (circa): Komposita mit Analysen



Einige wichtige Inhalte

- ▶ Zahl der Mitspieler oder Komplemente
(*exportieren.1.1*: 4, nämlich WER-WAS-WOHER-WOHIN)
- ▶ syntaktische Form der Mitspieler
(*exportieren.1.1*: Nominativ-Akkusativ-...)
- ▶ semantische Einschränkung der Mitspieler
- ▶ logische Schlussfolgerungen und semantische Verbindungen zu anderen Konzepten
- ▶ Morphologie (Wortformen)

⇒ **SEMPRIA hat eines der größten semantischen Computerlexika, das größte für Deutsch**



Universell einsetzbare elementare Sprachtechnologien (Beispiele)



Universell einsetzbare elementare Sprachtechnologien (Beispiele)

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⟩.



Universell einsetzbare elementare Sprachtechnologien (Beispiele)

- ▶ Auszeichnen von benannten Objekten

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⟩.



Universell einsetzbare elementare Sprachtechnologien (Beispiele)

- ▶ Auszeichnen von benannten Objekten
- ▶ Auszeichnen mit Grundformen

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⟩.**



Universell einsetzbare elementare Sprachtechnologien (Beispiele)

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

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⟩.



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>.`



- ▶ Bedeutungsorientierte Suche (Dokument-Retrieval)

- ▶ Fragebeantwortung





Nutzeranfrage:

Wo hat Oliver Bierhoff studiert ?



Nutzeranfrage:

Wo hat Oliver Bierhoff studiert?



Nutzeranfrage:

Wo hat *Oliver Bierhoff* studiert ?



Nutzeranfrage:

Wo hat Oliver Bierhoff studiert ?



Nutzeranfrage:

Wo hat Oliver Bierhoff studiert ?

(auch: *Oliver Bierhoffs Studium*)

Gefundene Belegstelle:

*Bierhoff hat ein wirtschaftswissenschaftliches Studium
an der FernUniversität in Hagen nach 26 Semestern
erfolgreich als Diplom-Kaufmann abgeschlossen.*



Nutzeranfrage:

Wo hat Oliver Bierhoff studiert ?

(auch: *Oliver Bierhoffs Studium*)

Gefundene Belegstelle:

Bierhoff hat ein wirtschaftswissenschaftliches Studium an der FernUniversität in Hagen nach 26 Semestern erfolgreich als Diplom-Kaufmann abgeschlossen .



Nutzeranfrage:

Wo hat Oliver Bierhoff studiert ?

(auch: *Oliver Bierhoffs Studium*)

Gefundene Belegstelle:

*Bierhoff hat ein wirtschaftswissenschaftliches Studium
an der FernUniversität in Hagen nach 26 Semestern
erfolgreich als Diplom-Kaufmann abgeschlossen .*



Nutzeranfrage:

Wo hat Oliver Bierhoff studiert ?

(auch: *Oliver Bierhoffs Studium*)

Gefundene Belegstelle:

Bierhoff hat ein wirtschaftswissenschaftliches Studium an der FernUniversität in Hagen nach 26 Semestern erfolgreich als Diplom-Kaufmann abgeschlossen .



Nutzeranfrage:

Wo hat Oliver Bierhoff studiert ?

(auch: *Oliver Bierhoffs Studium*)

Gefundene Belegstelle:

Bierhoff hat ein wirtschaftswissenschaftliches Studium an der FernUniversität in Hagen nach 26 Semestern erfolgreich als Diplom-Kaufmann abgeschlossen .

Ermittelte Antwort:

An der FernUniversität in Hagen.



- ▶ Natürlichsprachliche Anfrage statt Stichwortliste
- ▶ Logischer Strukturvergleich statt Stichwortvergleich
- ▶ Qualitätsüberprüfung statt Häufigkeitssortierung
- ▶ Präzise Resultate statt Listen von URLs

⇒ **SEMPRIA[®] leistet deutlich mehr als jede Stichwortsuche**

⇒ **Nutzer kann genauer und vollständiger suchen**



- ▶ **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)**
- ▶ **Optional: Sentiment Analysis (Entwicklungsarbeit erforderlich)**



Traditionelle
Suchmaschinen

Flexion, Derivation, Komposita (in einfacher, unvollständiger Form)
Nutzung von Metadaten

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 *
semantische Suchvorschläge aus den Dokumenten

* abhängig vom verfügbaren Hintergrundwissen



- ▶ SEMPRIA[®] erhöht den Anteil richtiger Ergebnisse
⇒ steigert die **Zufriedenheit** der Suchenden
- ▶ SEMPRIA[®] findet auf intelligentem Weg Dokumente, die eine Stichwortsuche nicht finden kann
⇒ maximiert das **Verwertungspotenzial** von Texten
- ▶ SEMPRIA[®] bietet erstmals eine bedeutungsorientierte Suche für deutschsprachige Archive
⇒ sichert Ihren **Innovationsvorsprung**



Eine gehostete, flexible Lösung mit preiswerten Monatsgebühren.

Randbedingungen:

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



SEMPRIA GmbH
Grafenberger Allee 277–287
40237 Düsseldorf

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

Geschäftsführer: Dr. Sven Hartrumpf
Handelsregister: Amtsgericht Düsseldorf HRB 62168; USt-IdNr: DE268248179

