

# Semantische Suche und Visualisierung von biomedizinischen Relationsdaten

Johannes Hellrich

Jena University Language & Information Engineering Lab  
Friedrich-Schiller-Universität Jena

Tagung der Computerlinguistik-Studierenden, 2012

- 2004 an der Friedrich-Schiller-Universität Jena neu gegründet
- Leiter: Prof. Dr. Udo Hahn
- Forschungsschwerpunkte:
  - Informationsextraktion / Text Mining
  - Information Retrieval: Suchmaschinen
  - Wissensmodellierung und Ontologien (Ontology Engineering)
  - Domänenkontext: Lebenswissenschaften
- Forschungsorientiert und drittmittelstark:
  - JenAge (BMBF: nationaler Forschungskern zur Altersforschung)
  - Mantra (EU: crosslinguales biomedizinisches Lexikonlernen)
  - Ab 2013: AquaDiva (DFG-SFB: Textwissensmanagement zu Biogeosphären)
- Kontinuierliche Präsenz auf: ACL, EMNLP, COLING, LREC, ...
- Exzellenten Mastern & Doktoranden stehen bei uns alle Türen offen!

# Inhalt

## Biomedizin und Computerlinguistik

- Informationsflut
- Semantische Suche
- Relationen

## Suche

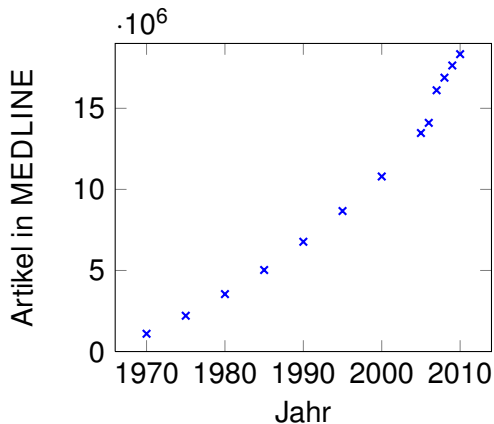
- RDF
- Einbindung in Semedico

## Visualisierung

- Visualisierung in der Biomedizin
- PPI Anzeige in Semedico

## Zusammenfassung

# Publikationsflut



Jährlich ca. 5% mehr Artikel in MEDLINE (biomedizinische Literaturdatenbank)

- Schlagwortsuche (vgl. Google)
- Basiert auf MEDLINE
- Suche über MeSH (Medical Subject Headings) Terminologie
- Artikel werden von Hand indexiert (nach MeSH)
- Freitextsuche + Suche nach MeSH Termen (Eingabe teilweise auf MeSH abgebildet, z.B. *black death* auf *plague*)

# Probleme bei der Suche

- Ambiguität
- Synonymie

# Ambiguität

- Eigennamen: Suche nach *Leber*
  - Frau Leber (Autorin)
  - Leber (Organ)
- Organismen: Suche nach *IL-6*
  - IL6\_MOUSE
  - IL6\_PIG
  - IL6\_HUMAN

# Synonymie

- brain-factor-1
- **FOXG1**
- hBF-2
- bf1
- FKHL4
- ...

⇒ 64 Möglichkeiten auf das Gen **FOXG1** zu verweisen.



- Semantische Suchmaschine des JULIE Labs
- Basiert auf MEDLINE
- Artikel werden automatisch indexiert
- Facettierung zur Unterstützung der Frageformulierung und Orientierung im Termraum (25k Facetten)

# Semedito Screenshot



FKHL4

search

Select a term to search for synonyms and related terms. Press ESC to ignore suggestions.

Result 1-10 of 385 (104 ms)

BioMed Immunology Ageing Bibliography Filter

**Genes and Proteins (385)**

## FOXG1 (any organism)

- FOXG1 (Homo sapiens) (254)
- FOXG1 (Mus musculus) (90)
- FOXG1 (Xenopus laevis) (62)

[more..](#)
**Chemicals and Drugs (231)**

- Enzymes and Coenzymes (103)
- Organic Chemicals (84)
- Chemical Actions and Uses (82)

[more..](#)
**Organisms (241)**

- Homo sapiens (114)
- Mus musculus (107)
- Rattus (18)

**Gene Expression (148)**

 These terms define your query. Click  to remove a term.

**Genes and Proteins: FOXG1 (any organism)** 
[show review articles only](#)

 sort by: 

Zhao Xiao-Feng, Suh Clotilde S, Prat Carla R, Ellingsen Staale, Fjose Anders

**[Distinct expression of two foxg1 paralogues in zebrafish.](#)**
*Gene expression patterns* : GEP. Mon Jun 01 00:00:00 CEST 2009; 9 5 (5): 266-72

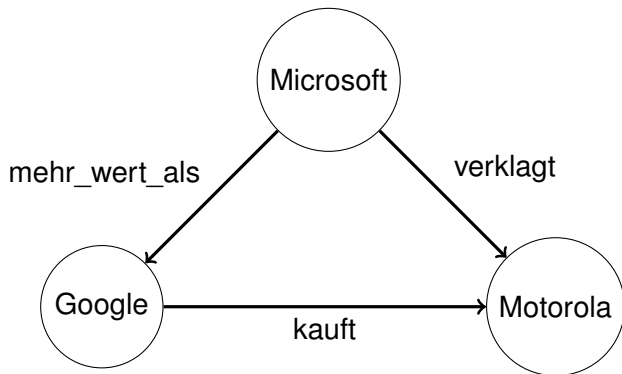
- ...range of species. One member of this superfamily, **Foxg1**, has essential roles in the development of eyes...

Murphy D B, Wiese S, Burfeind P, Schmundt D, Mattei M G, Schulz-Schaeffer W, Thies U

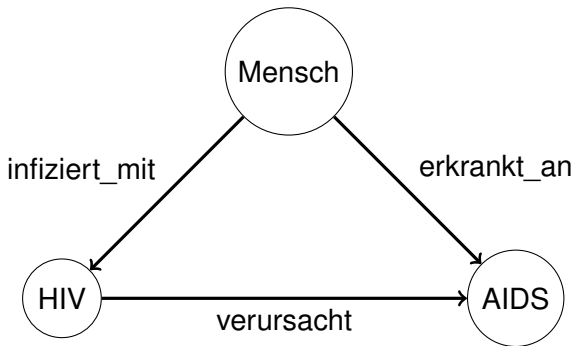
**[Human brain factor 1, a new member of the fork head gene family.](#)**
*Genomics*. Wed Jun 01 00:00:00 CEST 1994; 21 3 (3): 551-7

- ...conserved DNA-binding domain. Three of these cDNAs (**HFK1**, **HFK2**, and **HFK3**) were further analyzed. The cDNA...

# Relationen in Zeitungstexten



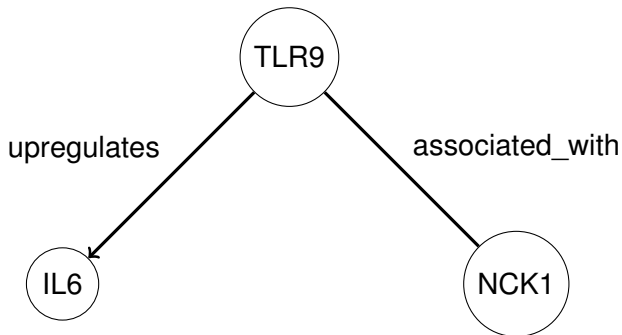
# Relationen in Fachtexten



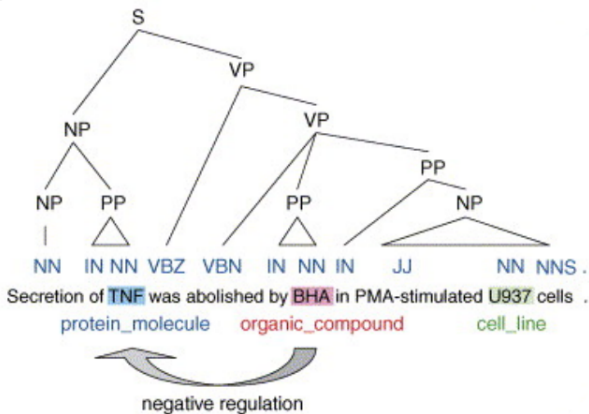
# Protein-Protein-Interaktionen (PPIs)

- Ein aktueller Schwerpunkt der biomedizinischen Forschung
- Biologische Datenbanken: Lassen menschliche Experten ausgewählte Zeitschriften nach Neuentdeckungen durchsuchen...
- Aus CL-Sicht: Relationen zwischen Named Entities

# PPIs als Relationen



# Relationenerkennung



Entnommen aus Ananiadou et al.: Trends in Biotechnology, 2006;24(12).

# Jena Relation Extractor (JREX)

- Nutzt Dependenzparsebäume
- State of the Art System: Platz 2 von 24 bei der BioNLP'09 Challenge

⇒ Automatische Informationsextraktion statt Experten!



# Inhalt

## Biomedizin und Computerlinguistik

- Informationsflut
- Semantische Suche
- Relationen

## Suche

- RDF
- Einbindung in Semedico

## Visualisierung

- Visualisierung in der Biomedizin
- PPI Anzeige in Semedico

## Zusammenfassung

# Anforderungen an die Relationsrepräsentation

- Suche nach nur teilweise spezifizierten Zusammenhängen, z.B.:  
Welches Virus steht in der Relation *verursacht* zu *AIDS*?
- Verknüpfung mit biomedizinischen Datenbanken
- Integration mit Terminologien/Ontologien wäre wünschenswert

# RDF als Lösung

- W3C Standard für Speicherung/Austausch semantischer Informationen
- Besteht aus Tripeln (Subjekt - Prädikat - Objekt)
- Die Elemente der Tripel sind URIs
- Kann mit Ontologien interagieren

⇒ Ermöglicht Suche nach Zusammenhängen zwischen Proteinen

# RDF Beispiel

Wikipedia  
was created by  
Jimmy Wales

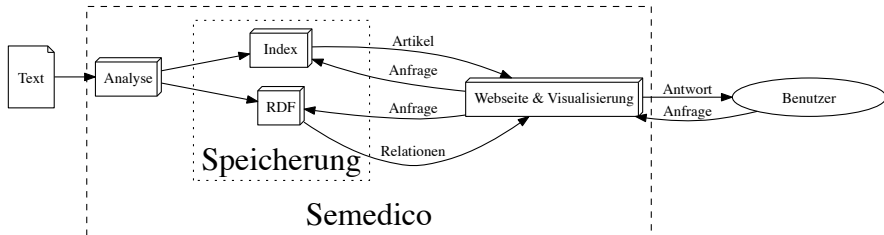
```
<http://en.wikipedia.org/wiki/Wikipedia>  
<http://purl.org/dc/elements/1.1/creator>  
<http://en.wikipedia.org/wiki/Jimmy_Wales>
```

# Indexbasierte Suche

- Bisher wurde Solr (Suchserver) zur Indexierung genutzt
- Solr speichert nun auch die Relationen in jedem Dokument
- Termbasiert – ungeeignet zur Suche nach nur teilweise spezifizierten Zusammenhängen

⇒ Ermöglicht Suche nach Publikationen zu einer Relation

# Architektur des Prototypen



- Analyse: UIMA
- Speicherung: Solr & Virtuoso
- Webseite: Tapestry

# Verbindung mit dem bisherigen Semedico



il-2

search

Select a term to search for synonyms and related terms. Press ESC to ignore suggestions.

Result 1-10 of 48 (158 ms)

Med Immunology Bibliography Aging

## Genes and Proteins (48)

### IL2 (any organism)

- IL2 (Homo sapiens) (46)
- IL2 (Mus musculus) (5)
- IL2 (Rattus norvegicus) (2)

## Chemicals and Drugs (28)

- Chemical Actions and Uses (11)
- Organic Chemicals (10)
- Enzymes and Coenzymes (10)

[more..](#)

These terms define your query. Click  to remove a term.

**Genes and Proteins: IL2 (any organism)**

[show review articles only](#)

sort by:

[visualize PPIs on this page!](#)

Bernhardt M Brooke, Hicks M John, Pappo Alberto S

[Administration of high-dose interleukin-2 in a 2-year-old with metastatic melanoma.](#)

*Pediatric blood & cancer.* Tue Dec 15 00:00:00 CET 2009; 53 7 (7): 1346-8

- ...have not been well studied. This report describes our experience with the use of high-dose **interleukin 2 (aldesleukin, IL-2...**

# Inhalt

## Biomedizin und Computerlinguistik

- Informationsflut
- Semantische Suche
- Relationen

## Suche

- RDF
- Einbindung in Semedico

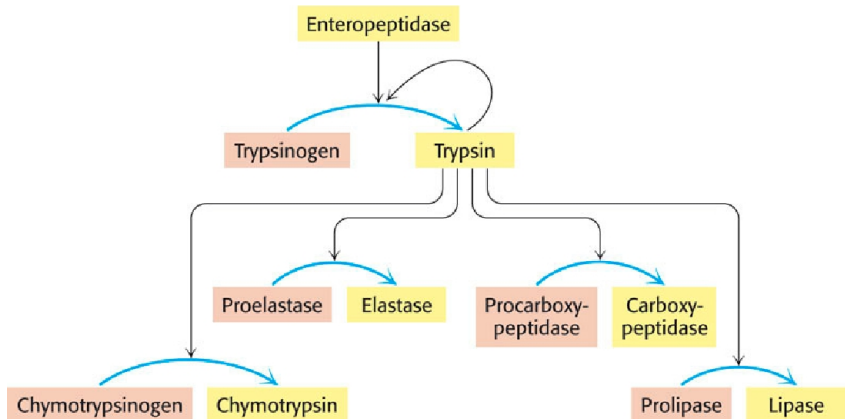
## Visualisierung

- Visualisierung in der Biomedizin
- PPI Anzeige in Semedico

## Zusammenfassung

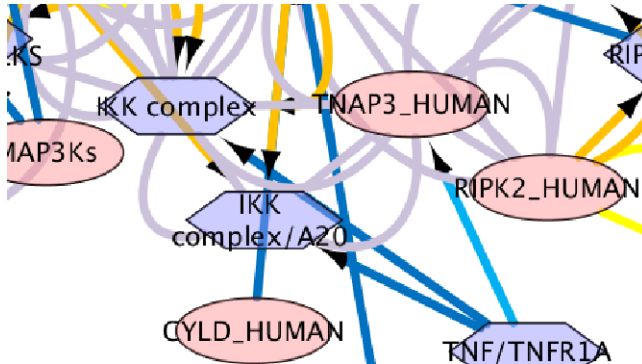


# Visualisierung in Lehrbüchern



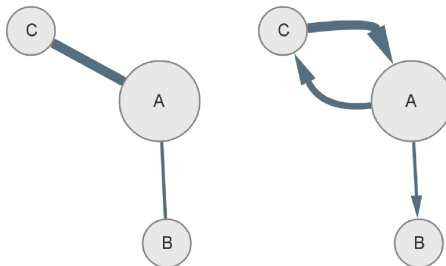
Aus: Berg/Tymoczka/Stryer, *Biochemie*, 6. Aufl., © 2007 Elsevier GmbH

# Visualisierung mit Cytoscape



Cytoscape – das Standardprogramm für biomedizinische Visualisierung

# Visualisierung mit Cytoscape Web



Online Version: Cytoscape Web

# Vergleich der Cytoscapeversionen

	Cytoscape	Cytoscape Web
max. Knoten:	10k+	ca. 200
Sprache:	Java	Flash & JS
Formate:	überlappend, u.a. XGMML	
Ausgeführt:	Rechner des Benutzers	

# Visualisierung in Semedico: Übersicht



Edges merged?  yes  no  
 Layout:  default  circle  radial  
 Darstellungsoptionen



Powered by [Cytoscape Web](#)

References:

UniProt: [HSP74\\_HUMAN](#)  
 Beleg

# Visualisierung in Semedico: Detailansicht

semedico  
search PubMed beta

Edges merged?  yes  no  
Layout: default  circle  radial

Powered by [Cytoscape Web](#)

References:  
UniProt: [IL6\\_HUMAN](#)  
**top**: [Increased frequency of immunoglobulin \(Ig\)A-secreting cells following Toll-like receptor \(TLR\)-9 engagement in patients with Kawasaki disease.](#)  
[Beleg](#)

# Interaktivität

- Netzwerke um einzelne Knoten können neu geladen werden
- Alle Proteinnamen verweisen auf eine biomedizinische Datenbank (UniProt)
- Die als Beleg angegebenen Sätze verweisen auf den relevanten Artikel

# Inhalt

## Biomedizin und Computerlinguistik

- Informationsflut
- Semantische Suche
- Relationen

## Suche

- RDF
- Einbindung in Semedico

## Visualisierung

- Visualisierung in der Biomedizin
- PPI Anzeige in Semedico

## Zusammenfassung



# Zusammenfassung

- **Semantische Suche** und **Visualisierung** helfen bei der Informationssuche
- **RDF** eignet sich zur Speicherung von (biomedizinischen) **Relationen**
- Mein Prototyp erweitert die Semantische Suchmaschine **Semedico** um die **Suche und Visualisierung von Relationen**
- **Ausblick:**
  - Usability Studie
  - Stresstest

# Literatur I



Lewandowski, D.  
*Handbuch Internet-Suchmaschinen*  
Heidelberg, 2. Auflage 2011



Buyko, E. et al.  
Syntactic Simplification and Semantic Enrichment – Trimming  
Dependency Graphs for Event Extraction  
*Computational Intelligence*, 27(4):610–644, 2011



Lu, Z.  
PubMed and beyond: a survey of web tools for searching  
biomedical literature  
*Database*, 2011: article ID baq036, 2011

# Links I

▶ **JULIE Lab**

<http://www.julielab.de/>

▶ **Medline Statistik**

[http://www.nlm.nih.gov/bsd/index\\_stats.html](http://www.nlm.nih.gov/bsd/index_stats.html)

▶ **PubMed**

<http://www.ncbi.nlm.nih.gov/pubmed/>

▶ **Semeditco**

<http://www.semedico.org>

▶ **BioNLP'09**

<http://www.nactem.ac.uk/tsujii/GENIA/SharedTask/index.shtml>

▶ **RDF**

<http://www.w3.org/RDF/>

## Links II

▶ **Cytoscape Web**

<http://cytoscapeweb.cytoscape.org/>

▶ **Cytoscape**

<http://www.cytoscape.org/>

▶ **UIMA**

<http://uima.apache.org/>

▶ **Lucene**

<http://lucene.apache.org/solr/>

▶ **Virtuoso**

<http://virtuoso.openlinksw.com/>

▶ **Tapestry**

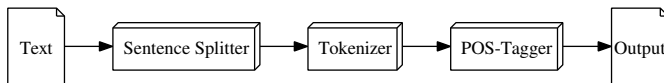
<http://tapestry.apache.org/>

# Noch Fragen?

# Danke für eure Aufmerksamkeit!

# Fragen?

# CL-Pipeline



- Ursprünglich IBM, jetzt Apache Projekt
- Java (und C++)
- Zentraler Speicher für alle Annotationen (CAS)
- Typensystem für Annotationen
- Hohe Kombinierbarkeit von Pipelinekomponenten

- Apache Projekt
- Java
- Basiert auf Lucene (Inverser Index: Speichert die für ein Schlagwort relevanten Dokumente)

⇒ DIY Suchmaschine



- Kommerzielle Software, eingeschränkte Open Source Version
- Triplestore auf Basis einer Relationalen Datenbank
- Aus Java über JDBC per SPARQL abfragbar
- Gute Benchmarkergebnisse (Berlin SPARQL Benchmark)

# Tapestry

- Apache Projekt
- Java
- Webseiten bestehen aus TML und Java Klasse
  - TML: XHTML mit Skripten zum Einfügen von Werten
  - Java Klasse: Stellt Werte zur Verfügung (Datenbankanbindung)