Semantische Suche und Visualisierung von biomedizinischen Relationsdaten

Johannes Hellrich

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

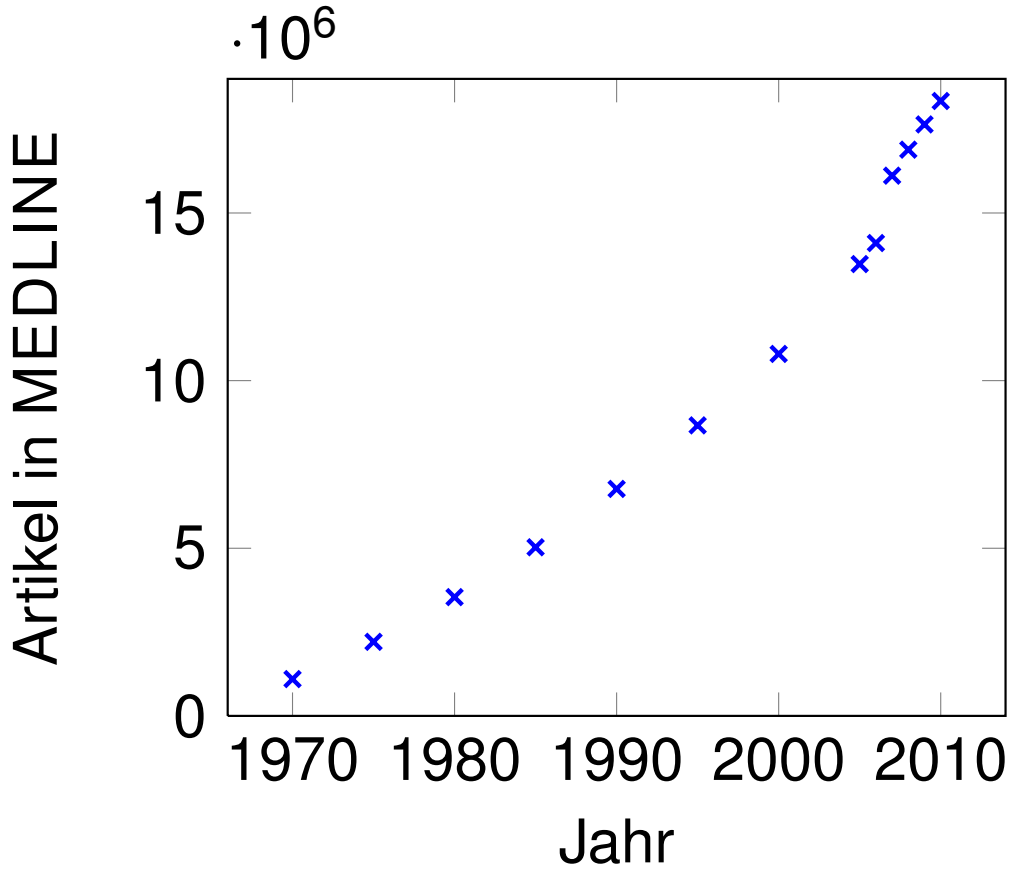
Tagung der Computerlinguistik-Studierenden, 2012

Über das JULIE Lab

- 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

Publikationsflut



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

PubMed

- 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

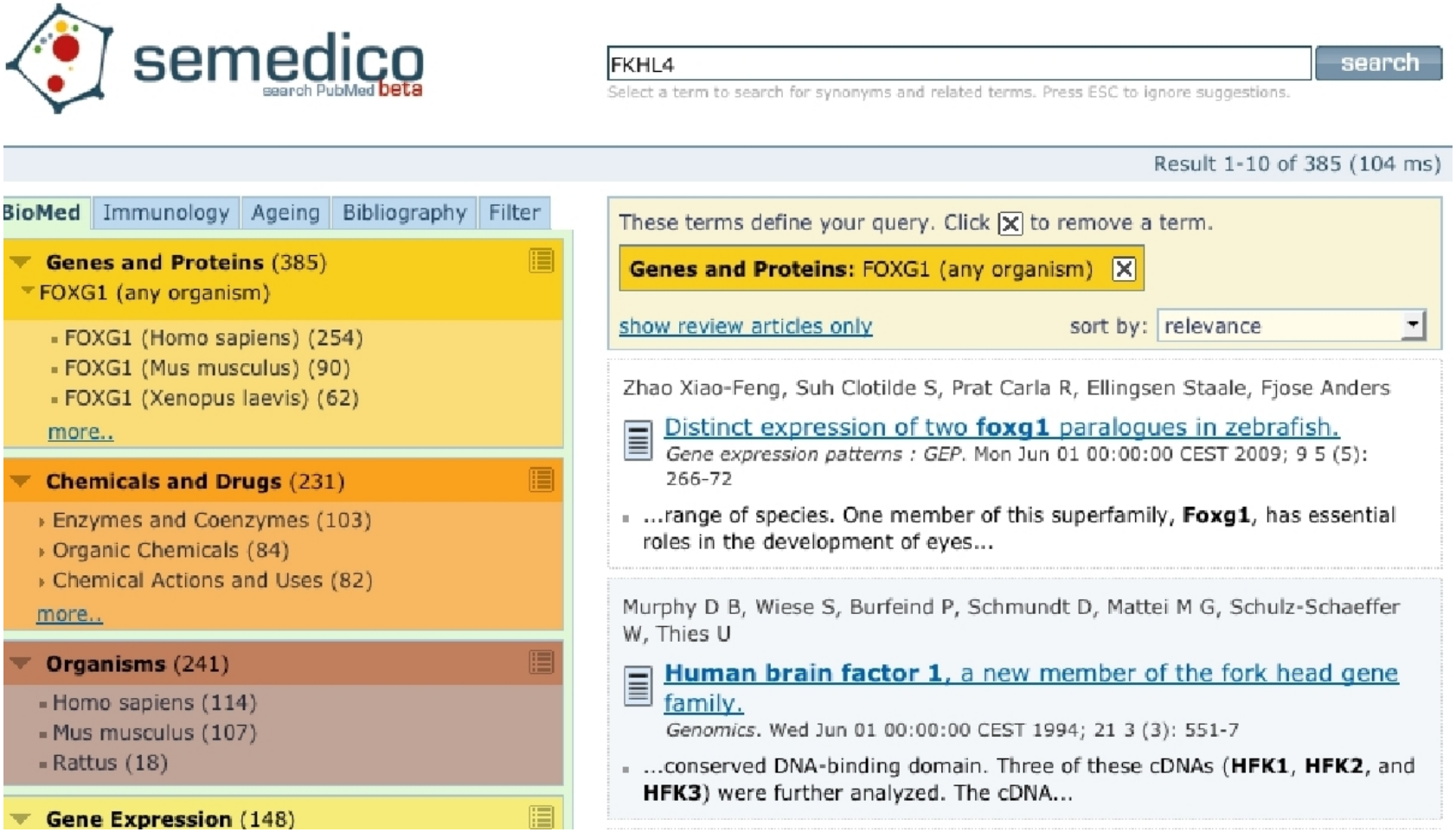
- brain-factor-1
- FOXG1
- hBF-2
- bf1
- FKHL4
- ...

$\Rightarrow$ 64 Möglichkeiten auf das Gen FOXG1 zu verweisen.

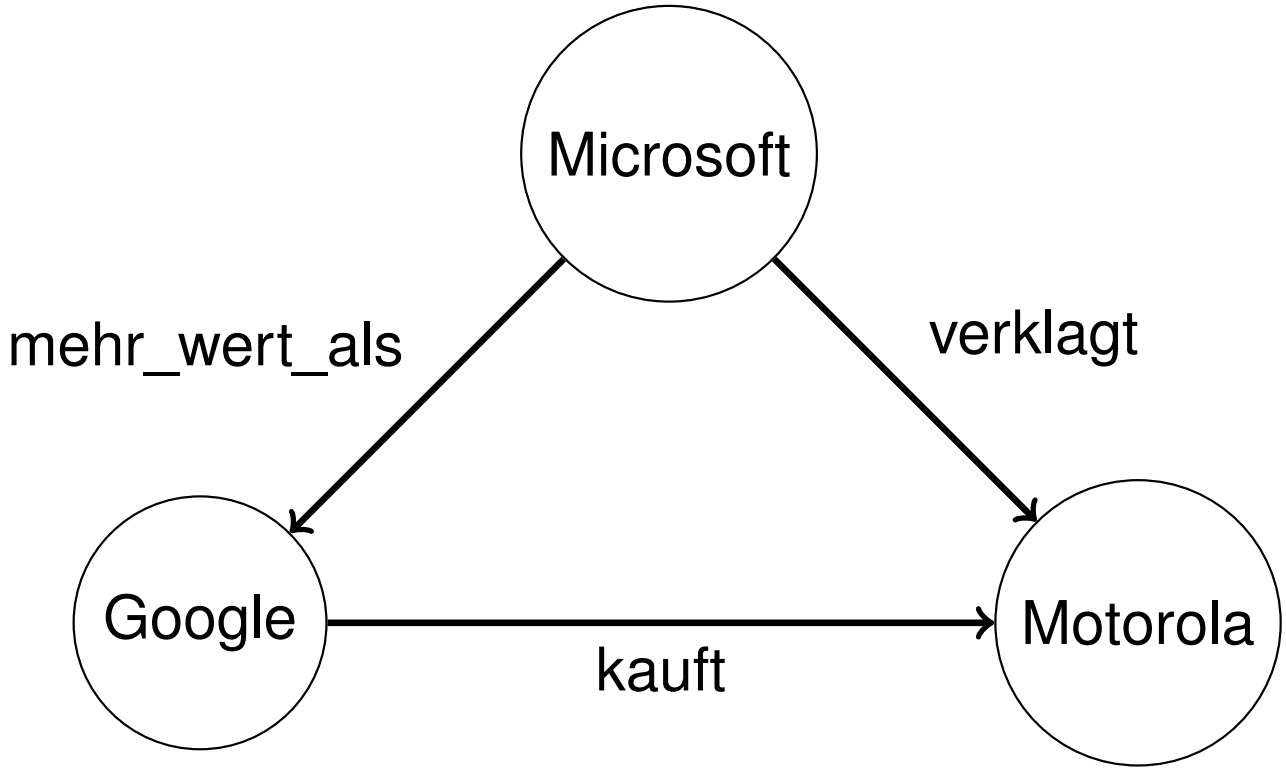
Semedico

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

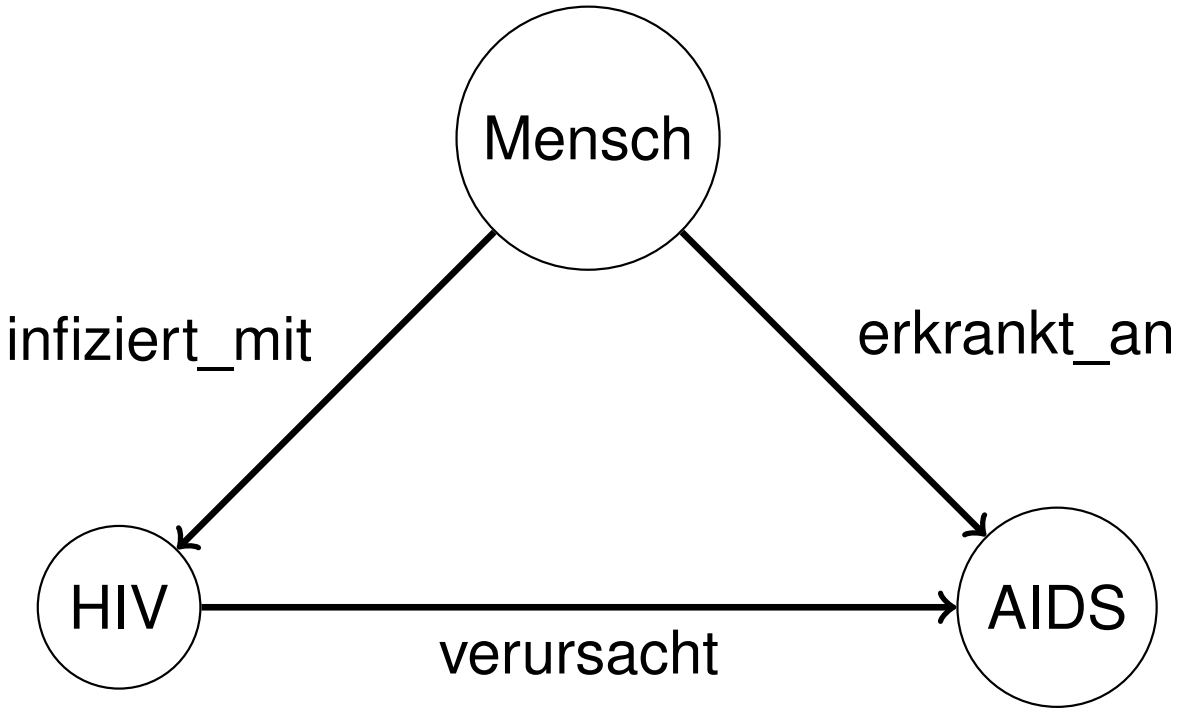
Semedico Screenshot



Relationen in Zeitungstexten



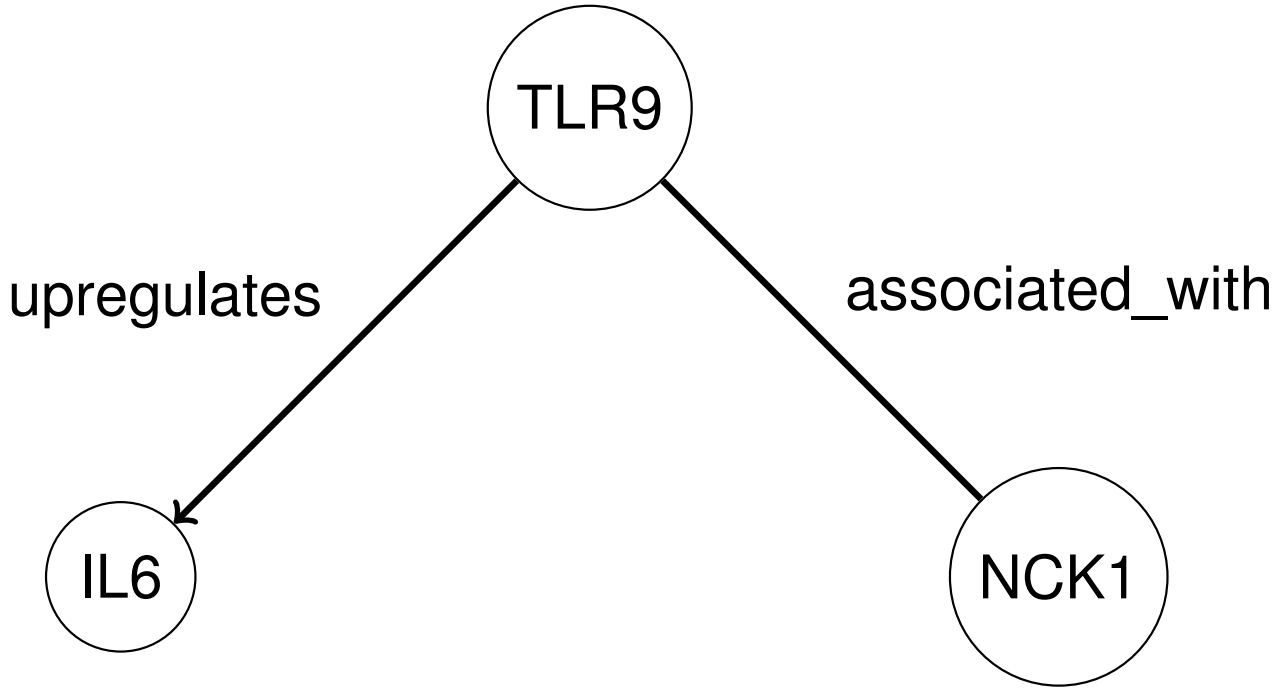
Relationen in Fachtexten



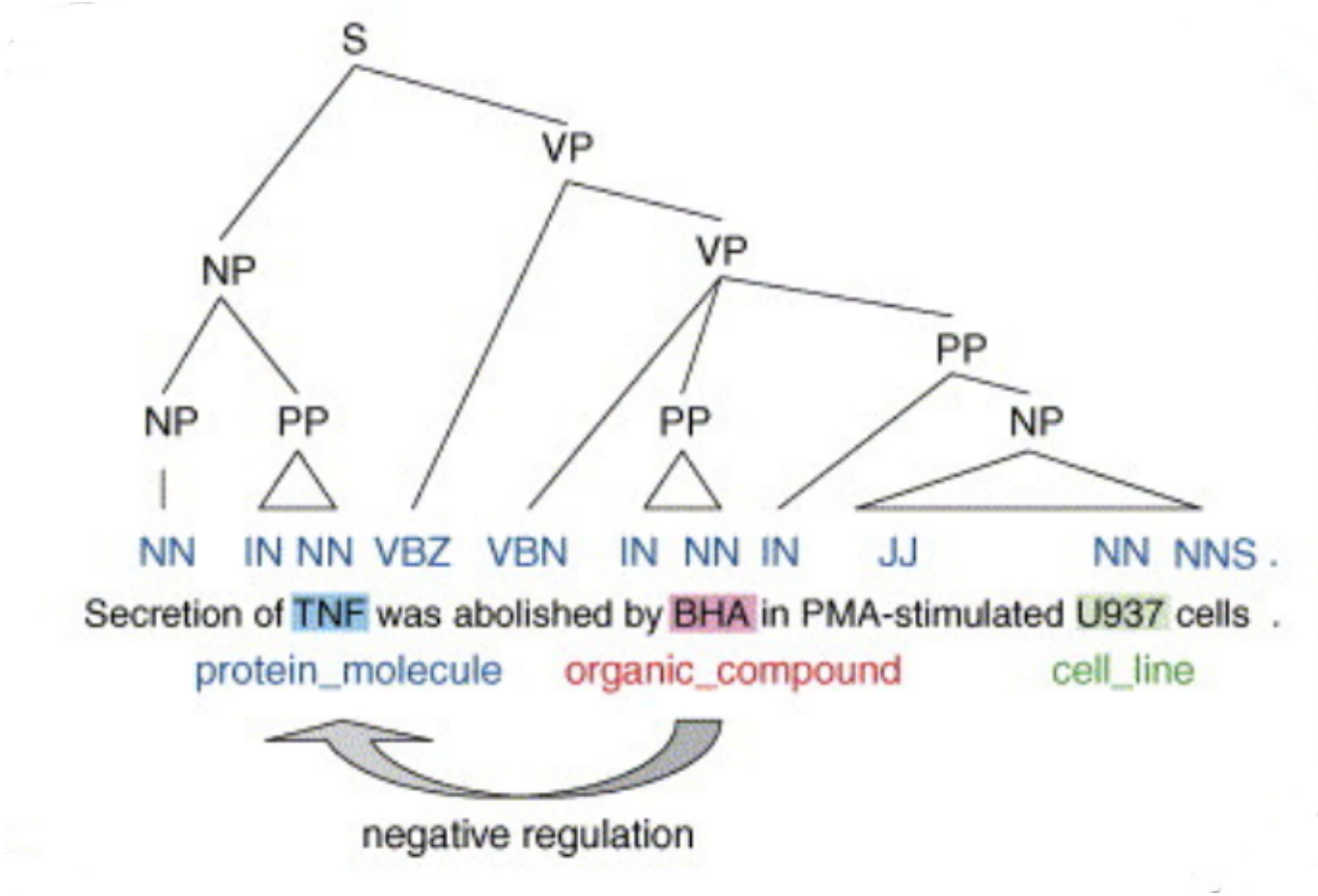
Protein-Protein-Interaktionen (PPIs)

- 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

$\Rightarrow$ 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

$\Rightarrow$ Ermöglicht Suche nach Zusammenhängen zwischen Proteinen

RDF Beispiel

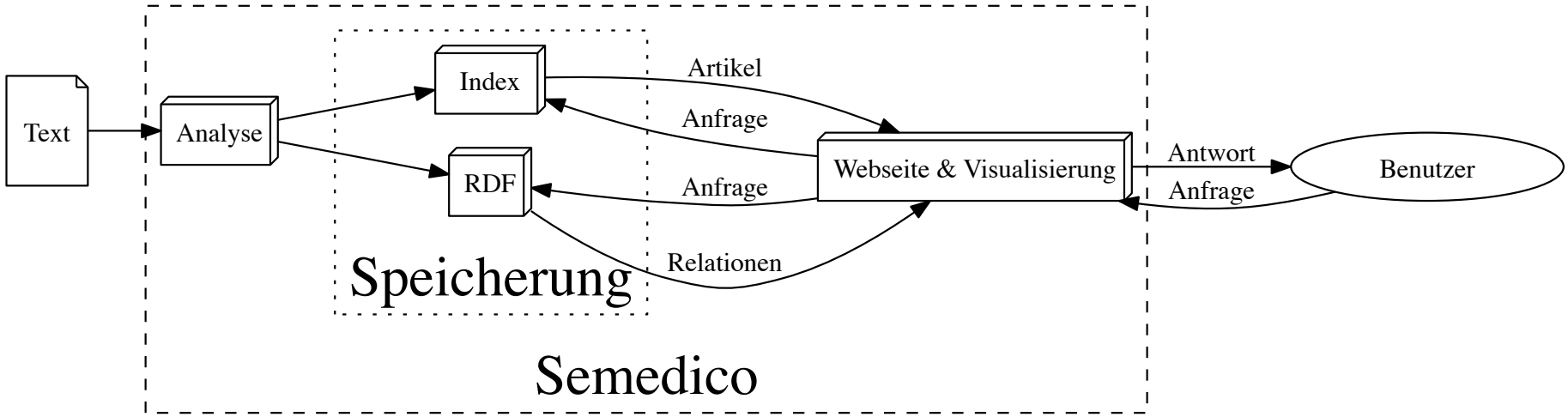
Wikipedia	<http://en.wikipedia.org/wiki/Wikipedia>
was created by	<http://purl.org/dc/elements/1.1/creator>
Jimmy Wales	<http://en.wikipedia.org/wiki/Jimmy_Wales>

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

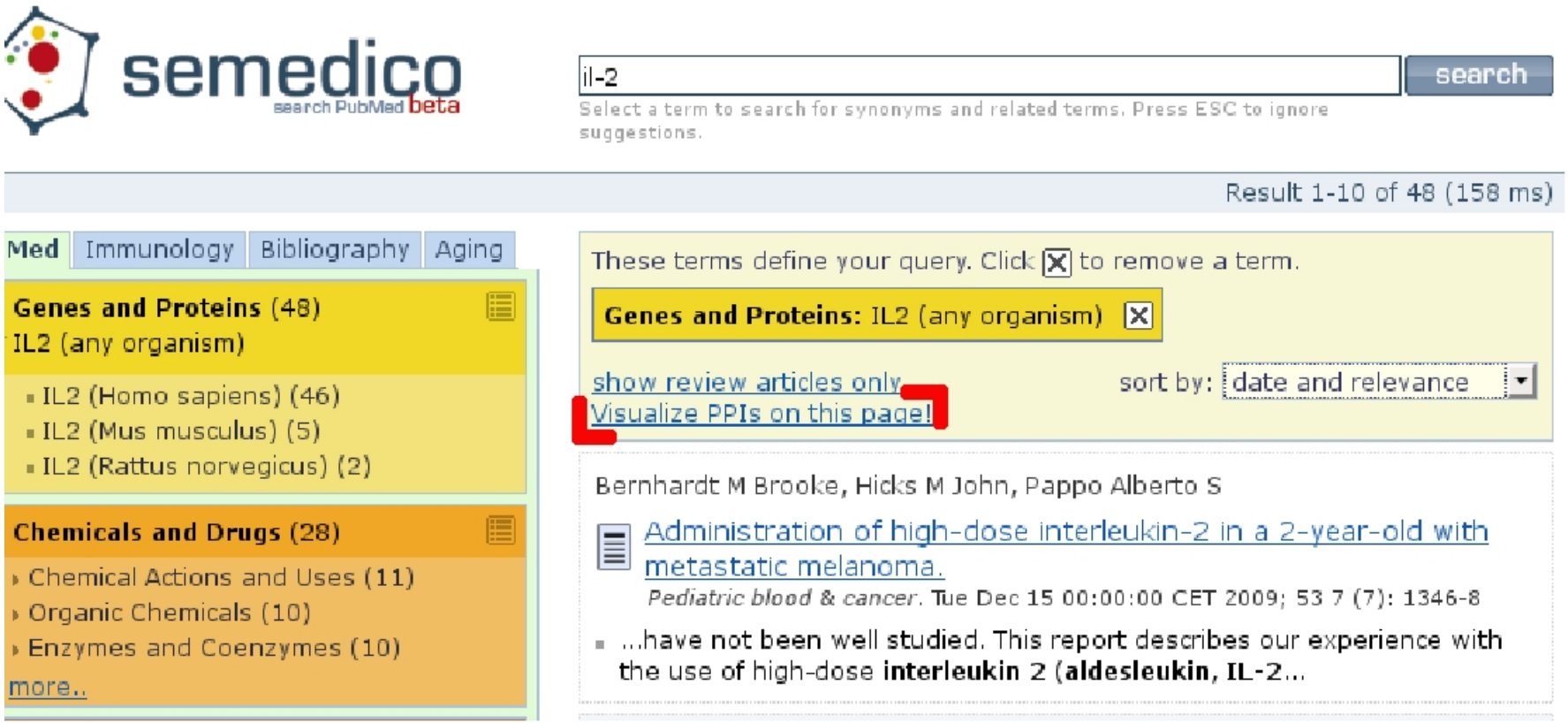
$\Rightarrow$ Ermöglicht Suche nach Publikationen zu einer Relation

Architektur des Prototypen



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

Verbindung mit dem bisherigen Semedico



Inhalt

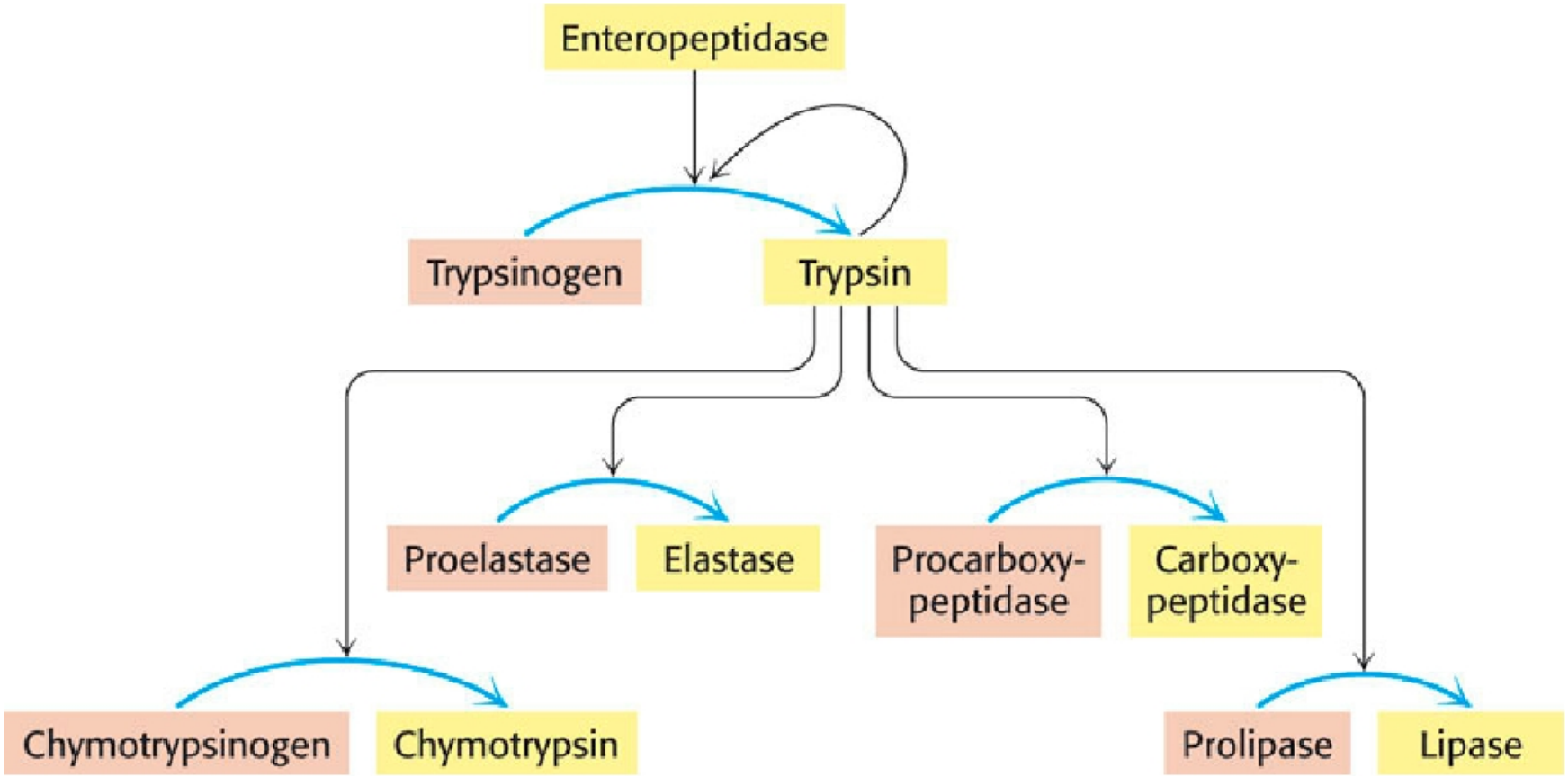
Biomedizin und Computerlinguistik
- Informationsflut
- Semantische Suche
- Relationen

Suche
- RDF
- Einbindung in Semedico

Visualisierung
- Visualisierung in der Biomedizin
- PPI Anzeige in Semedico

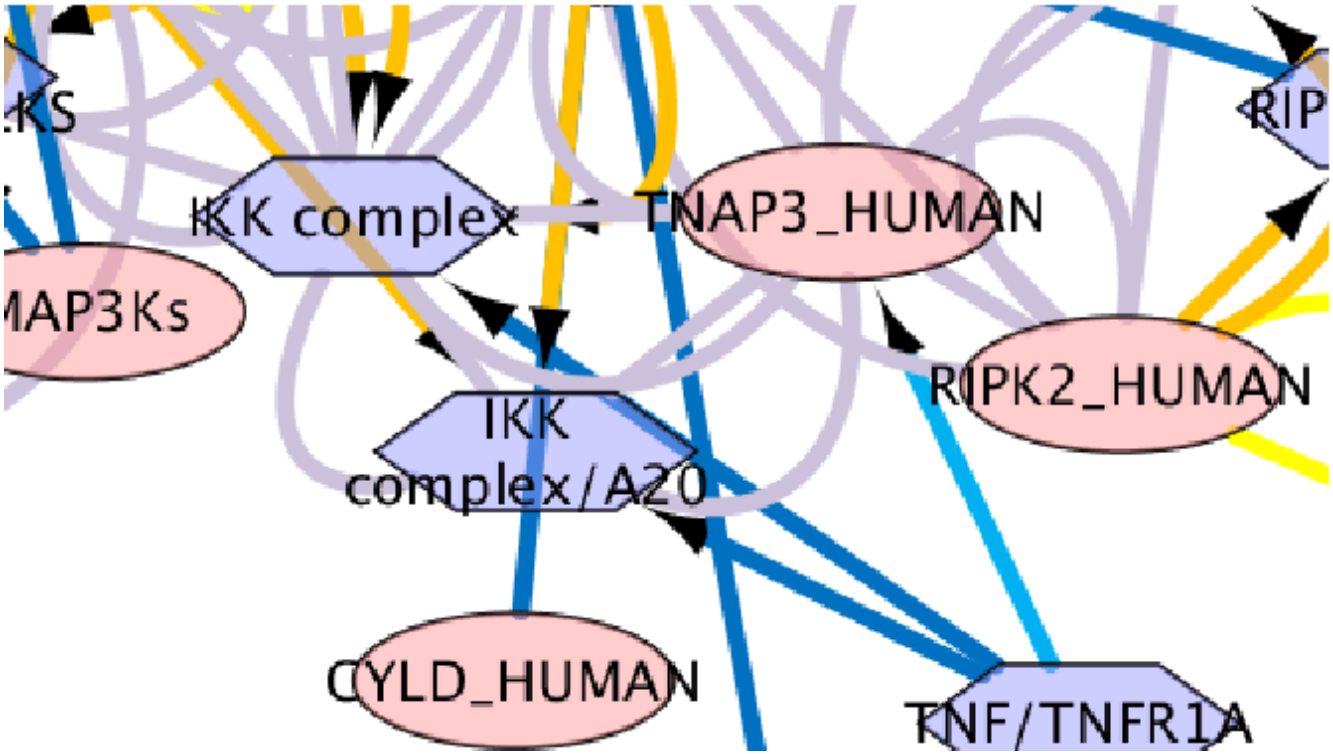
Zusammenfassung

Visualisierung in Lehrbüchern



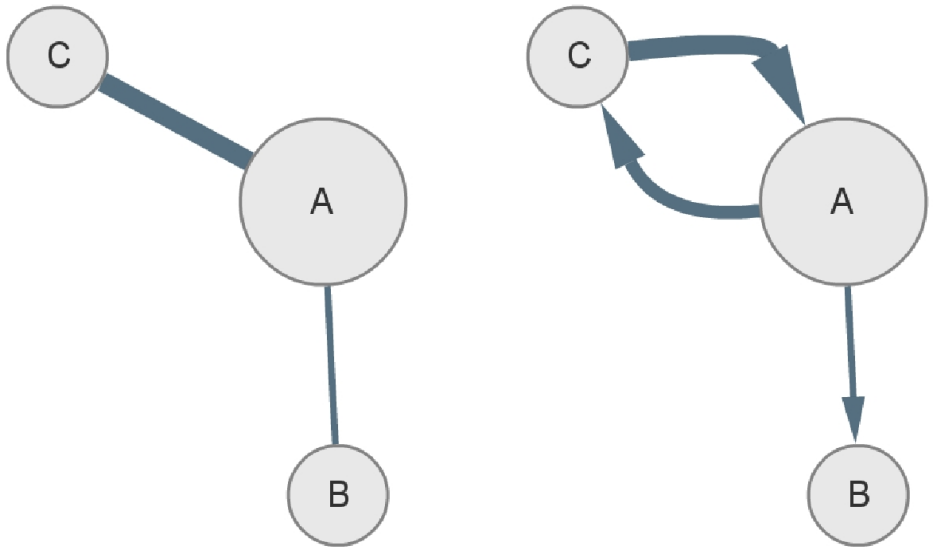
Aus: Berg/Tymoczko/Stryer, *Biochemie*, 6. Aufl., © 2007 Elsevier GmbH

Visualisierung mit Cytoscape



Cytoscape – das Standardprogramm für biomedizinische Visualisierung

Visualisierung mit Cytoscape Web

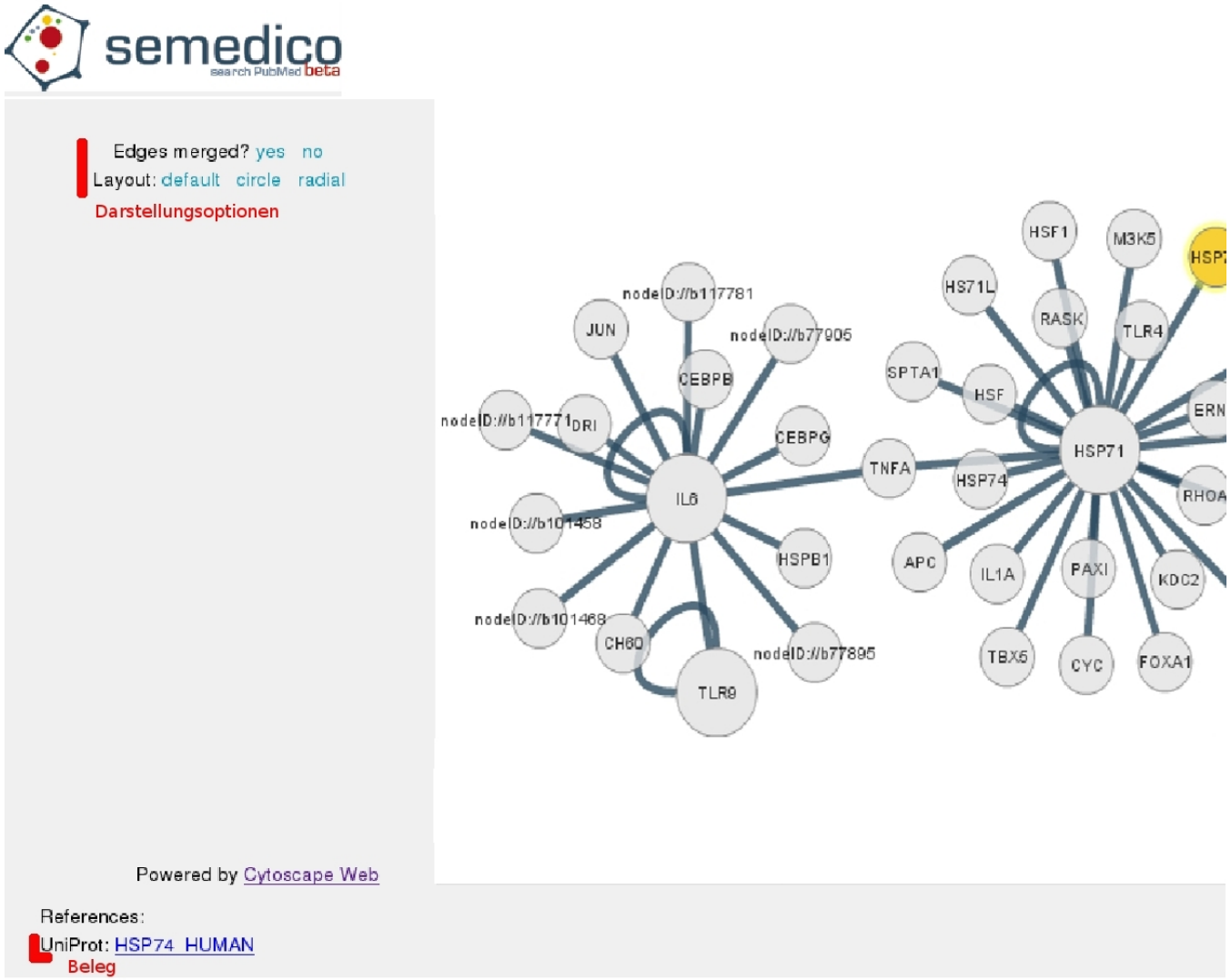


Online Version: Cytoscape Web

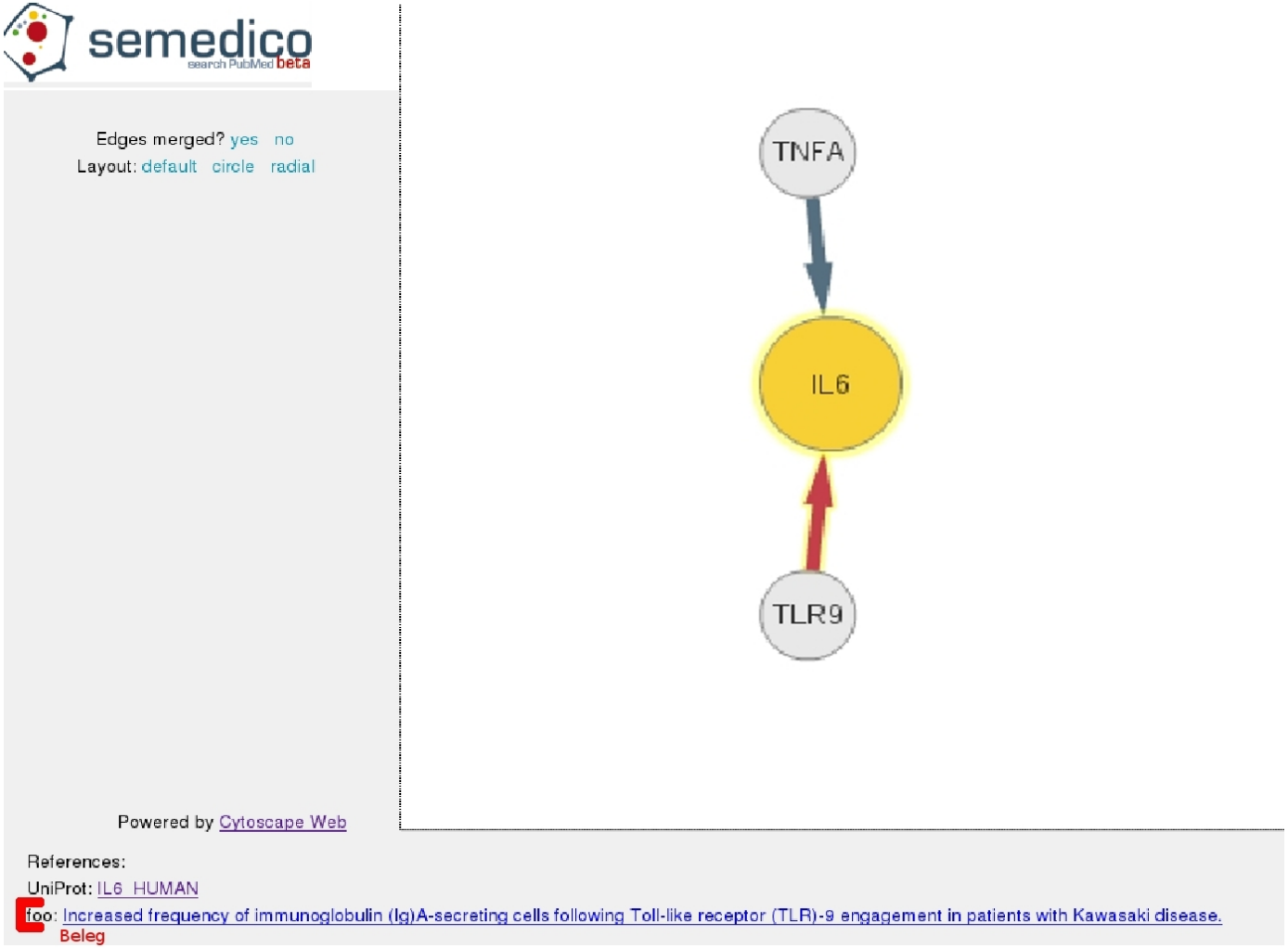
Vergleich der Cytoscapeversionen

	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



Visualisierung in Semedico: Detailansicht



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-Suchmachinen
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
- PubMed
  http://www.ncbi.nlm.nih.gov/pubmed/
- Semedico
  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/`

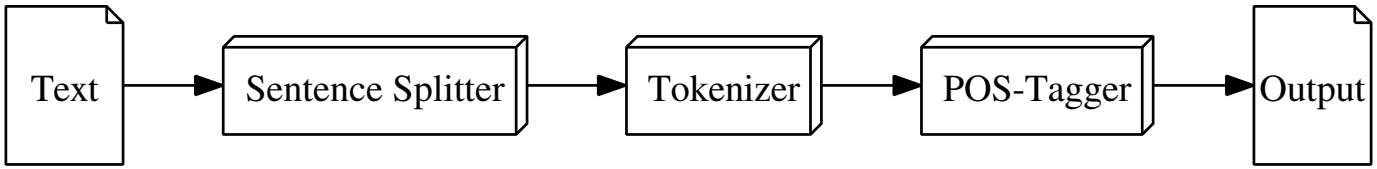
Noch Fragen?

Jena University Language & Information Engineering Lab

Danke für eure Aufmerksamkeit!

Fragen?

CL-Pipeline



UIMA

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

Solr

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

$\Rightarrow$ DIY Suchmaschine

Virtuoso

- 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)