Gefördert durch:
Bundesministerium
für Wirtschaft
und Technologie
aufgrund eines Beschlusses
des Deutschen Bundestages

TechWatch

Technology and Market Observation

powered by SMILA

PD Dr. Günter Neumann
DFKI, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Juni 2011

© DFKI, Günter Neumann, Juni 2011

Goal - Observation of Innovations and Trends

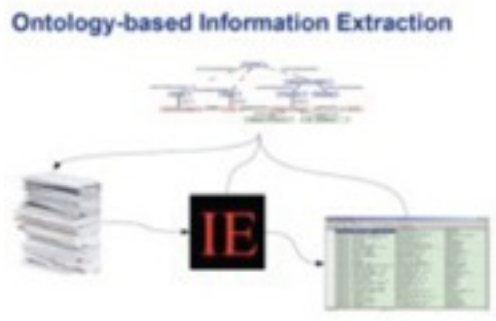
» Technologies:
  » Combine methods from Bibliometrics, Information Wrapping, Text Mining, Information Extraction, Semantic Search and Knowledge Management.

» Applications:
  » Search for publications and patents
  » The analysis of extracted relationships between themes, authors, time behavior and organizations.
  » Search for Key-Players of innovations.
  » Web-based search for trend statements.
  » Presentation of a knowledge domain as ontology, which can be used in further functionalities that improve the search and analysis

© DFKI, Günter Neumann, Juni 2011

Partner: ThyssenKrupp Steel AG

- Cooperation between DFKI and TKS since 2006
- Ontology-based search for key players in the area of welding and joining technologies
  - Authors, inventors, companies, institutes, technologies
    - Advanced patent search (EPO)
    - Publications (Google Scholar)
  - Visualization of the created innovation network and of new trends
- Software development „TechWatchTool"
- In Theseus-Ordo
  - SMILA-fication of the system architecture
  - Ontology-based information extraction
  - Trend-Monitoring and analysis (new technologies, innovation pushers)



© DFKI, Günter Neumann, Juni 2011

Important: Language Technology!

- **Extraction of relations through linguistic analysis**
  - e.g., dependency analysis
- **Information that change over time**
  - e.g., „Increase of publications about lithium batteries"
- **Near-Synonyms**
  - „Approach"; „Method"
- **Collocations**
  - „come to a decision", „make a picture"
- **Co-reference**
  - „dependency analysis", „this method"
- **Hyponyms, Part-of Relationship**
  - „is a", „consists of"
- **Meta data**
  - „is author of ", „is owner of the patent"

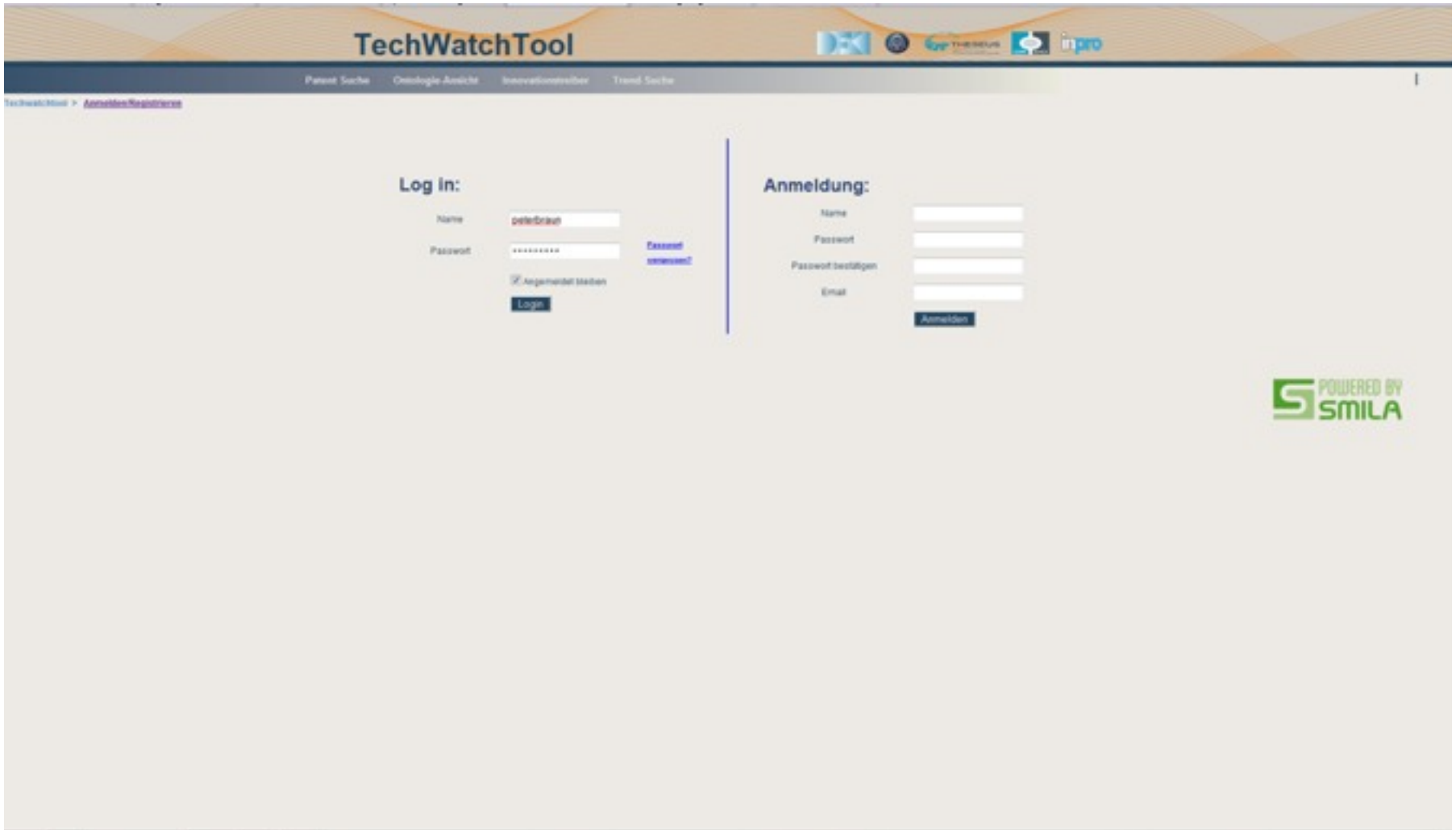
© DFKI, Günter Neumann, Juni 2011

Search and Analysis of Innovation Indicators

- Relevant Key Phrases
  - „Technology of tomorrow", „Future", „Development"
- Sentence patterns
  - „Lithium-ion batteries will power cars in the future."
  - „Lithium-ion technology is the future for hybrid and electric cars."
  - „The laser will always be good for a surprise also in the future", predicts Peter Leibinger.

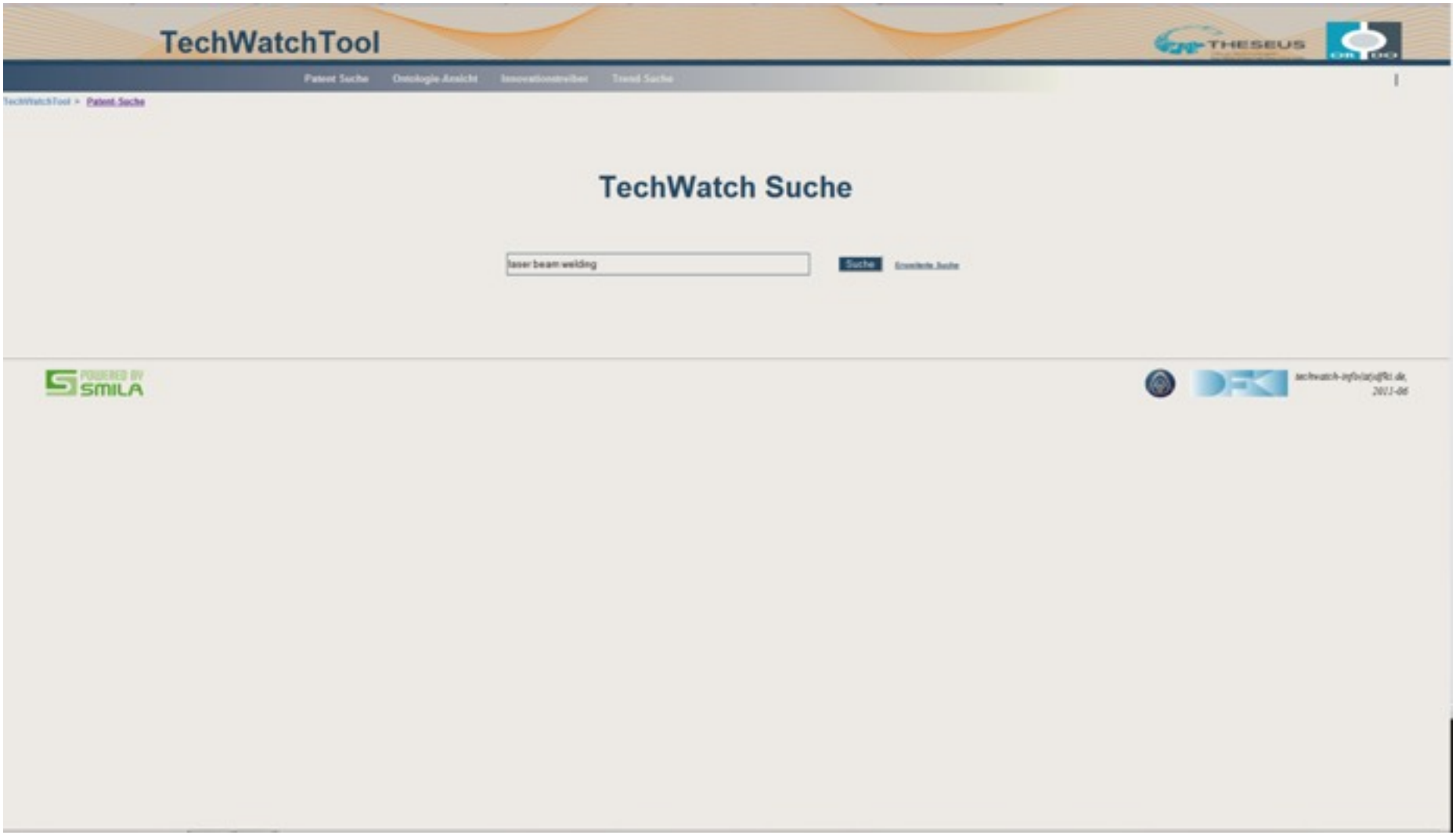
© DFKI, Günter Neumann, Juni 2011

TechWatchTool



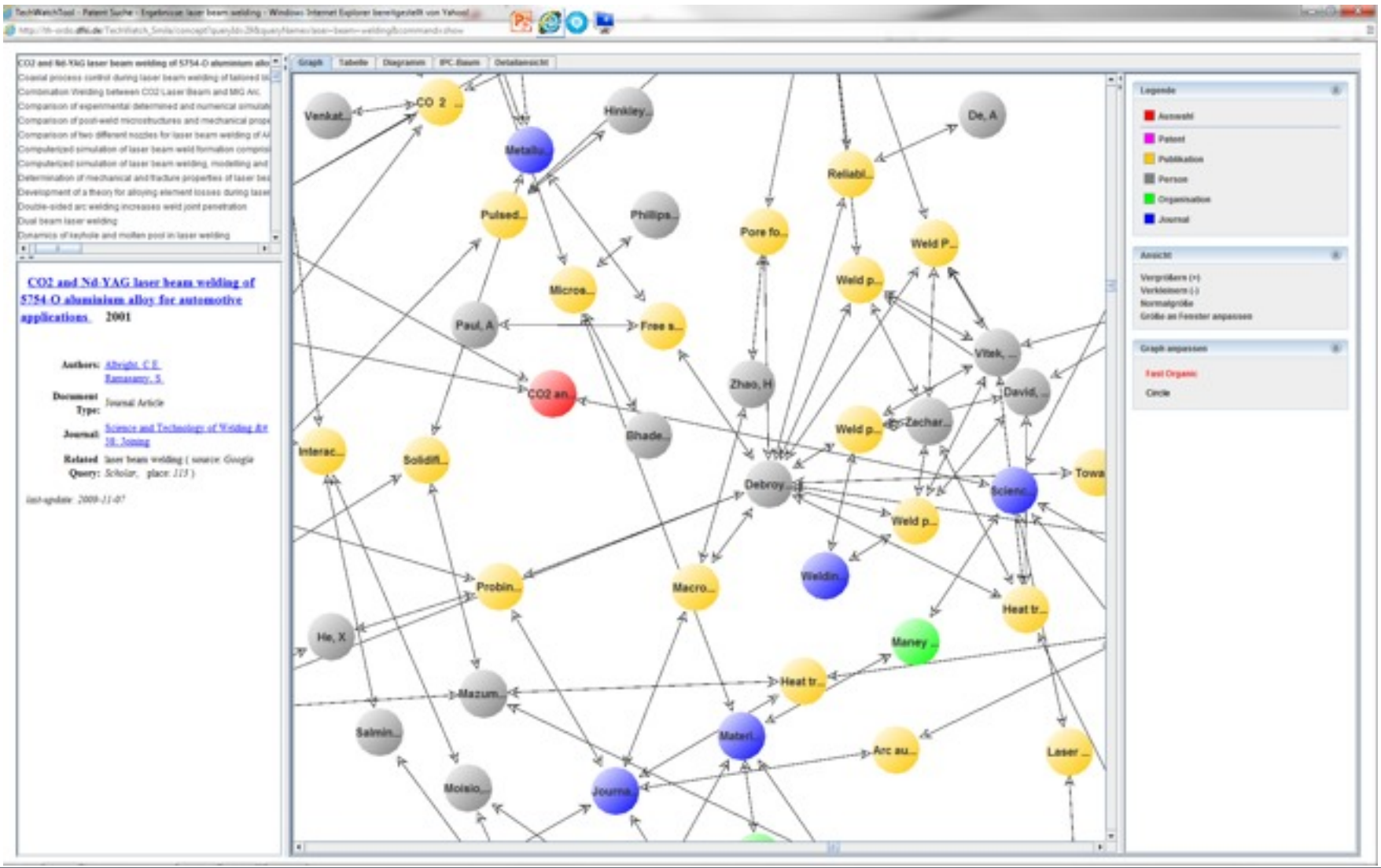
© DFKI, Günter Neumann, Juni 2011

Patent Search



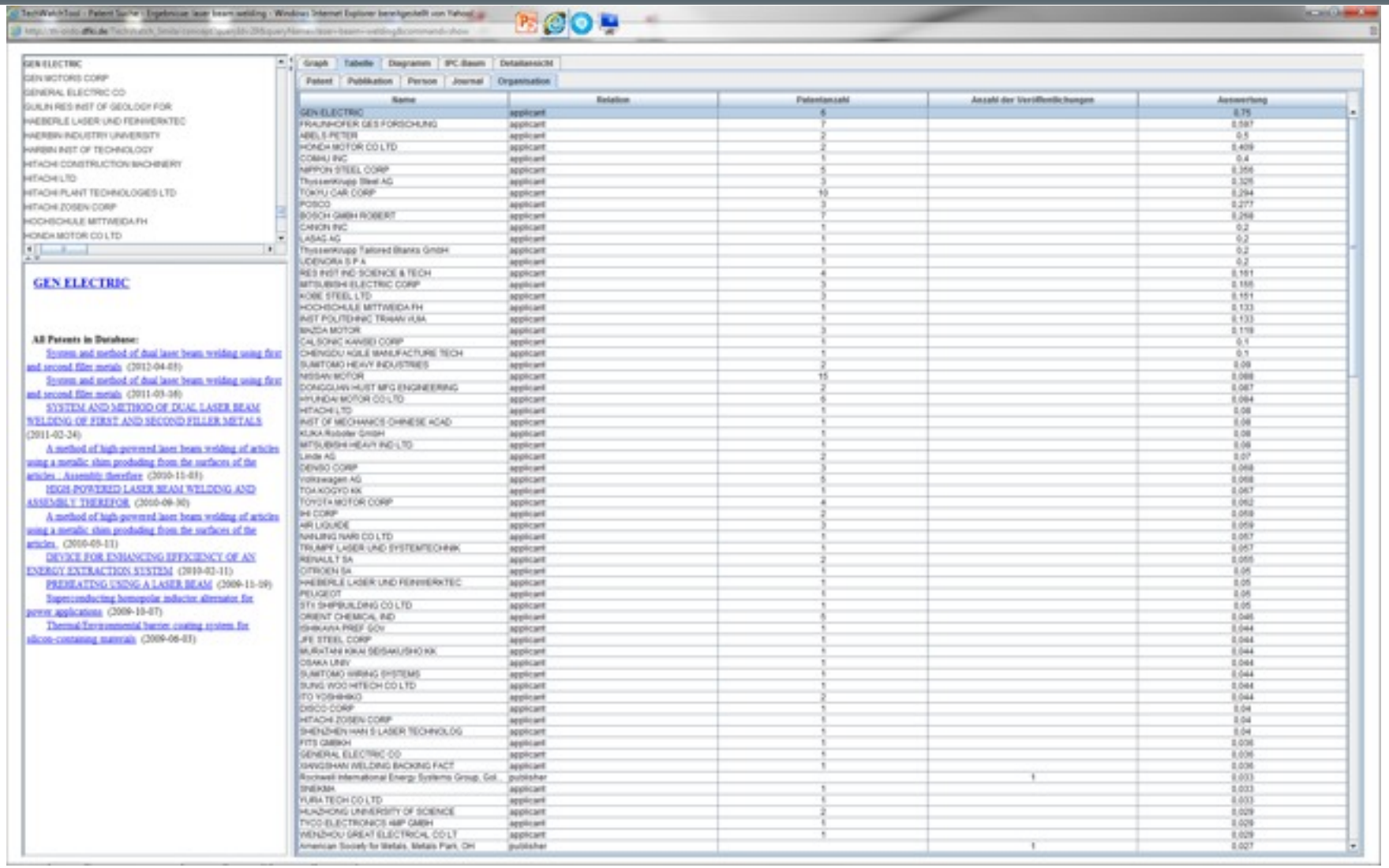
© DFKI, Günter Neumann, Juni 2011

Patent Search



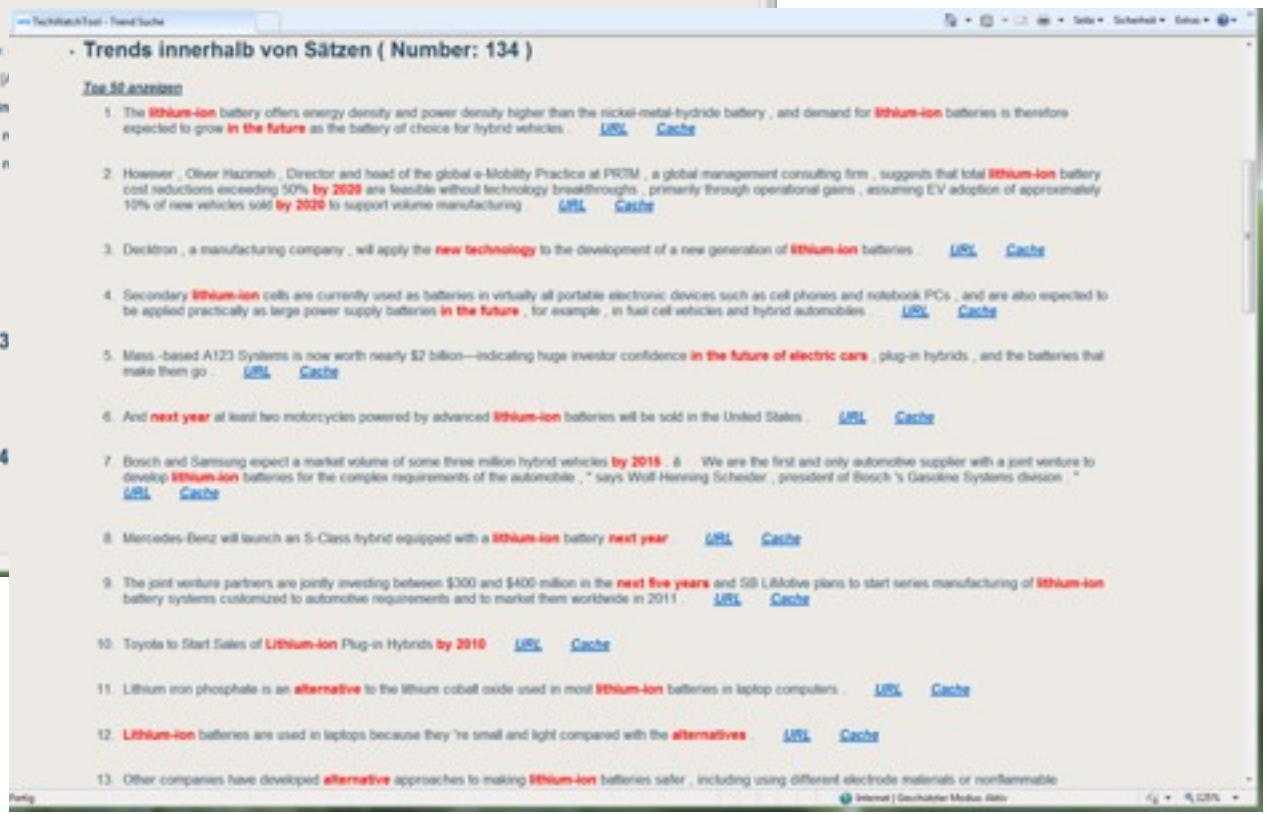
© DFKI, Günter Neumann, Juni 2011

Patent Search



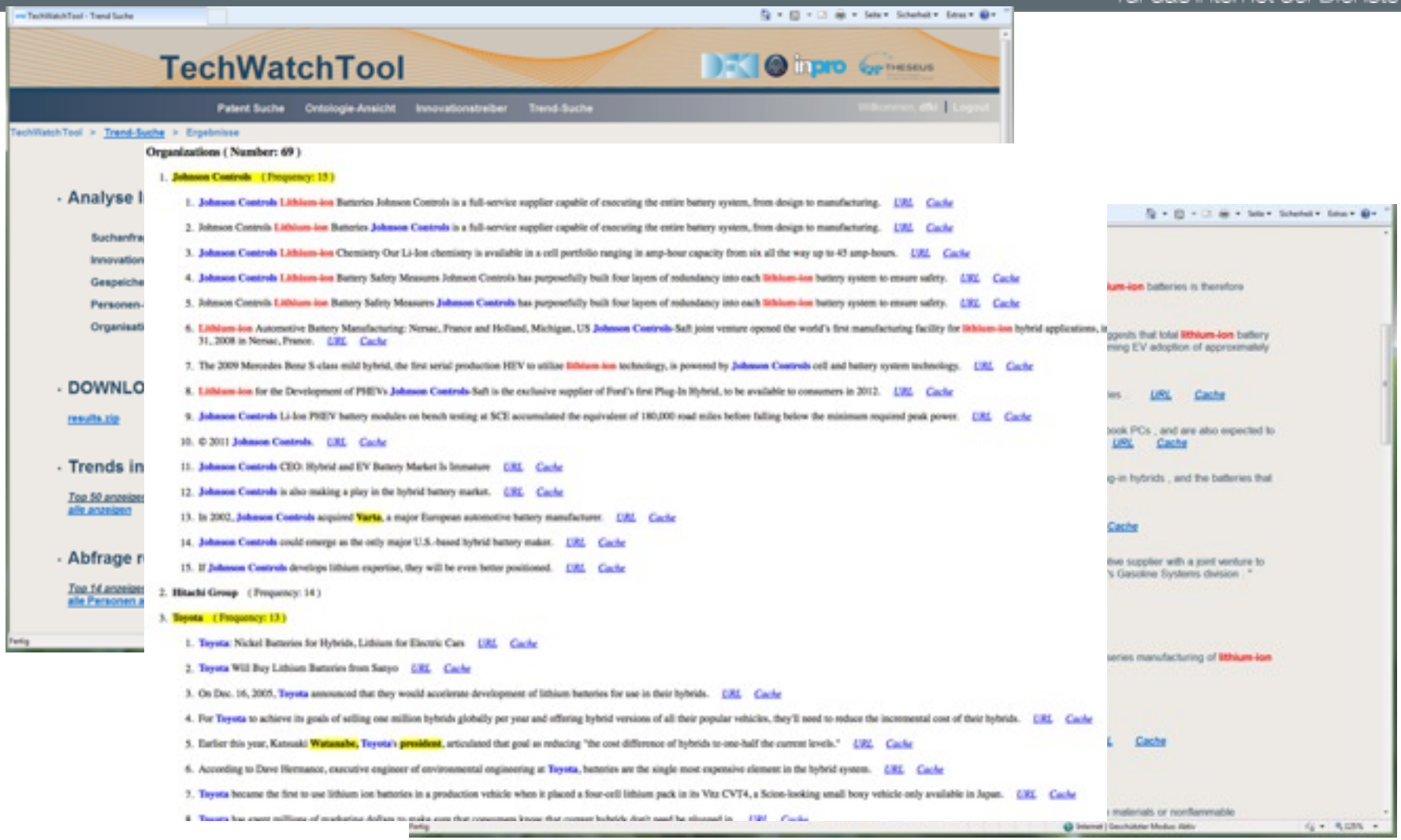
© DFKI, Günter Neumann, Juni 2011

Trend analysis



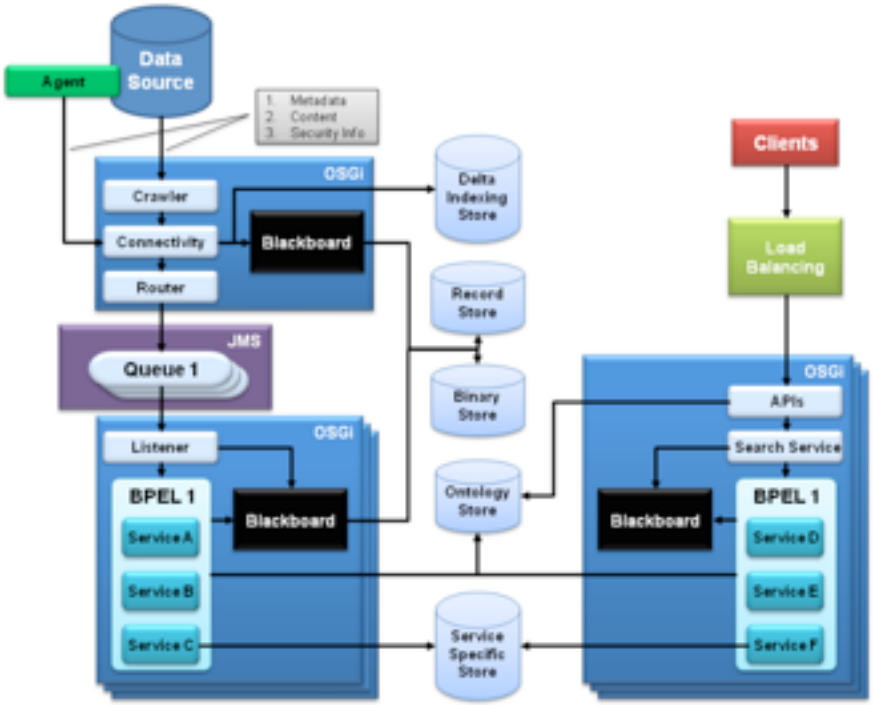
© DFKI, Günter Neumann, Juni 2011

Trend analysis



TechWatch – SMILA Integration

- The core components are integrated via wrapper methods
  - Patent search
    - Google Scholar
    - DepatisNet
    - Persistence
  - Trend search
    - Google Custom Search
    - Document processing
    - Trend patterns
    - Relation extraction
    - Named Entity recognition
    - Persistence



© DFKI, Günter Neumann, Juni 2011

TechWatch - DFKI Language Technology software

LanguageID
- Automatic language identification
- 23 languages
- Speed: ~6ms
- Accuracy: 99,8%
- Implementation: Java, OSGi

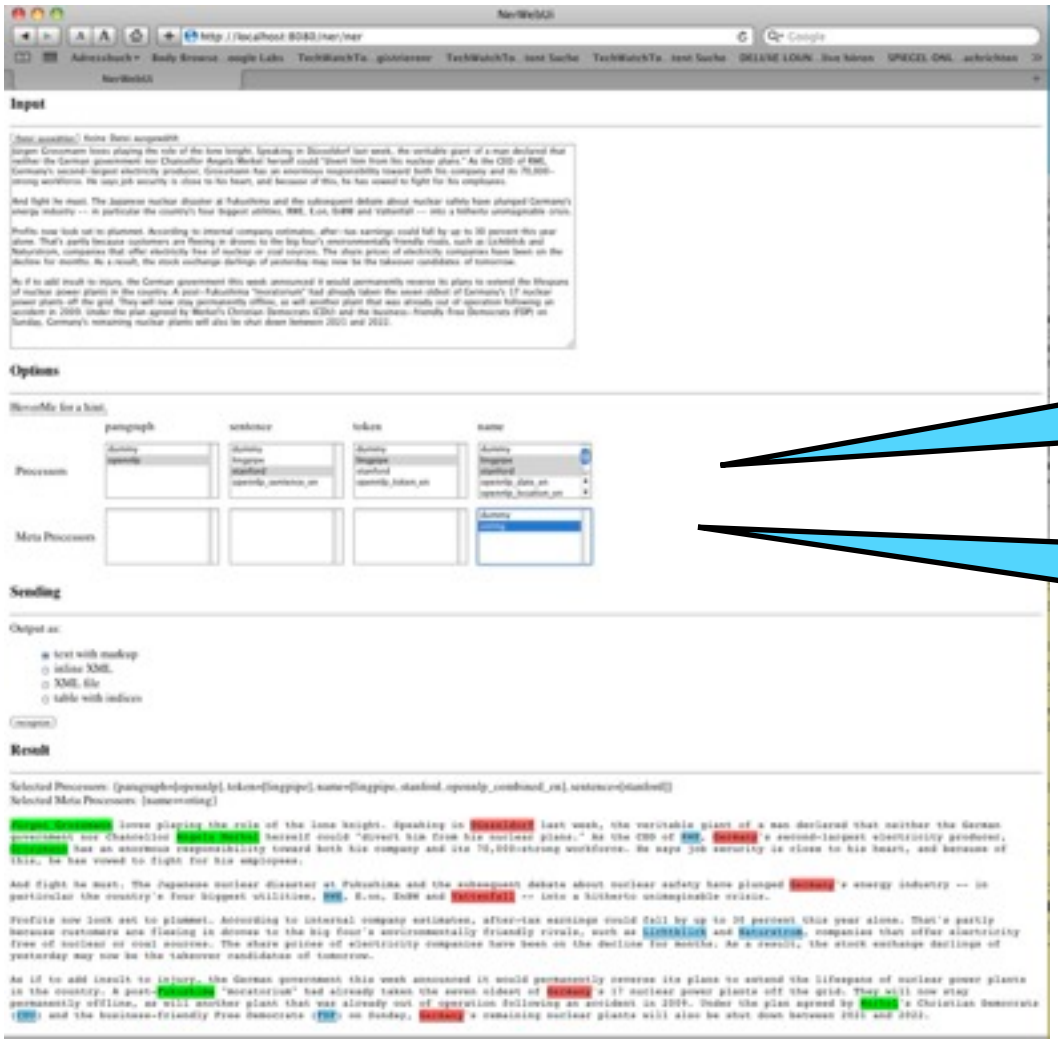
NE-Hub
- Extraction of Named Entities: Persons, Organizations, Locations, etc.
- Integrates the result of different NE-recognizers (Sprout, OpenNLP, Stanford, LingPipe, etc.)
- Languages: DE, EN
- Implementation: Java, OSGi

MDParser
- Extraction of the dependency relations of sentences
- Languages: DE, EN & more!
- Speed: 40 Sent./Sec.
- Accuracy: 86% - 88%
- Implementation: Java, OSGi

DARE
- Semi-supervised extraction of relations
- Languages: DE, EN
- Implementation: Java, OSGi

© DFKI, Günter Neumann, Juni 2011

Example - NE-Hub



Selection and orchestration of the components

Voting mechanism

© DFKI, Günter Neumann, Juni 2011

Example - NE-Hub

Options

HoverMe for a hint.

	paragraph	sentence	token	name
Processors	dummy opennlp	dummy lingpipe stanford opennlp_sentence_en	dummy lingpipe stanford opennlp_token_en	dummy lingpipe stanford
Meta Processors				dummy voting

Selection and orchestration of the components

Voting mechanism

Sending

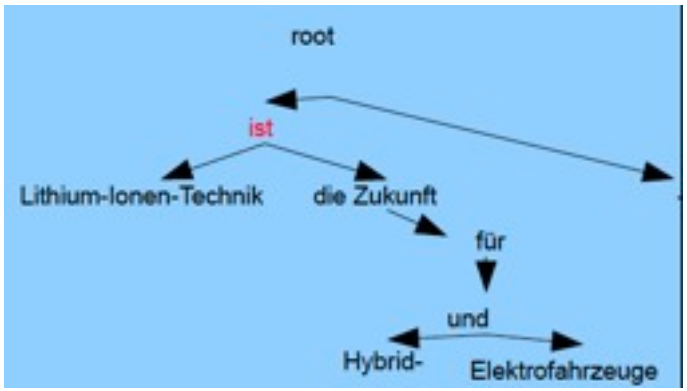
© DFKI, Günter Neumann, Juni 2011

Example
- MDParser

- A fine-grained extraction and search for entities and relations demands for a structural analysis of free text.

- Recently, dependency relations between words are getting more and more important for semantic applications

  The dependency relationships uncover a flat semantic predicate argument structure.



- Our current R&D focus
  - Automatic learning of dependency grammars and parsers from corpora
  - Multilingual dependency parsing
  - Solution: MDParser

© DFKI, Günter Neumann, Juni 2011

Example - MDParser

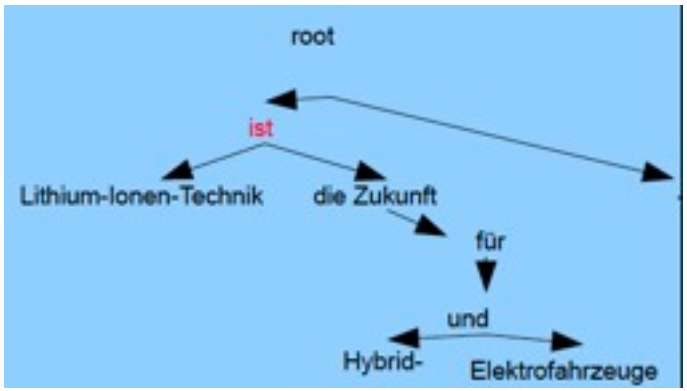
	Parsing Time	Sentences per Second	Tokens per Second
MDParser	73.188s	46.128	1015.55
MaltParser	1954.684s	1.73	38.02

RTE-6: 3376/74326 Sent./Words

- A fine-grained extraction and search for entities and relations demands for a structural analysis of free text.

- Recently, dependency relations between words are getting more and more important for semantic applications

  The dependency relationships uncover a flat semantic predicate argument structure.



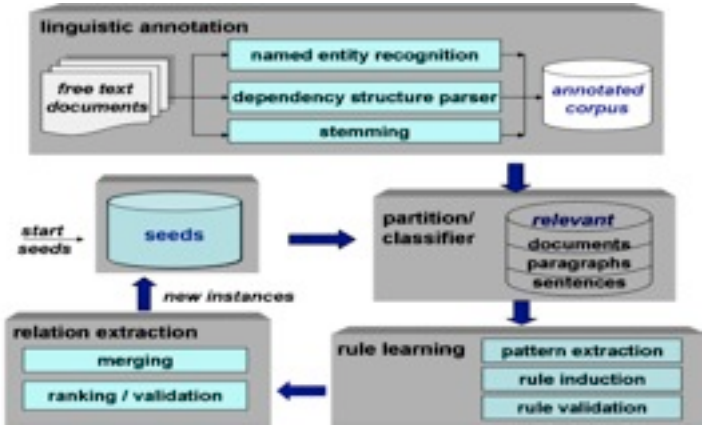
- Our current R&D focus
  - Automatic learning of dependency grammars and parsers from corpora
  - Multilingual dependency parsing
  - Solution: MDParser

© DFKI, Günter Neumann, Juni 2011

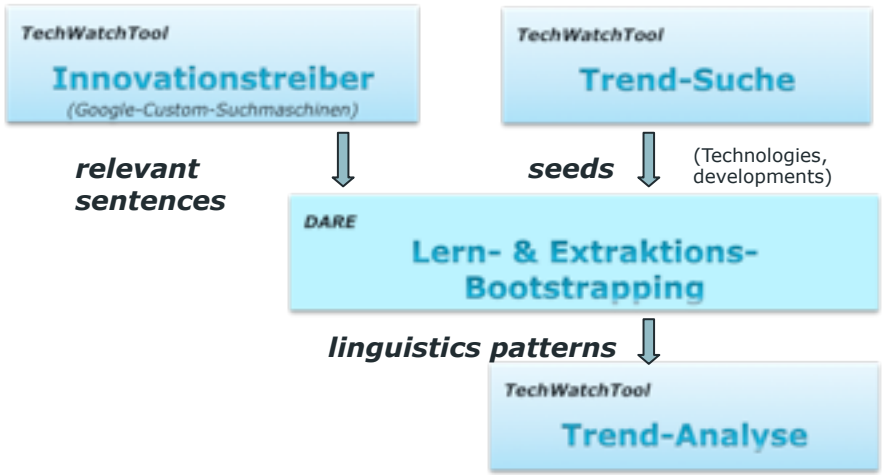
Example - DARE

THESEUS
Neue Technologien für das Internet der Dienste

- Extraction of relations of different complexities.
- Automatic learning of linguistic patterns for the extraction of new instances and their projections.
- **Bootstrapping** between pattern learning and relation extraction which is initialized by means of a small set of seed instances



Integration into *TechWatchTool*



- Learning of linguistic patterns for the automatic recognition of new technologies and developments for a future trend analysis.
- Use of the found results (technologies & developments) of the trend analysis as new seeds
- Search of relevant sentences by means of innovation indicators.

TechWatchTool – DFKI advanced LT components

uKeyWe
- Unsupervised Machine Learning algorithm for the extraction of key phrases
- Languages: language independent
- Implementation: Java, OSGi

ConExt
- Extraction of concept hierarchies from free texts
- Languages: DE, EN
- Implementation: Java, OSGi

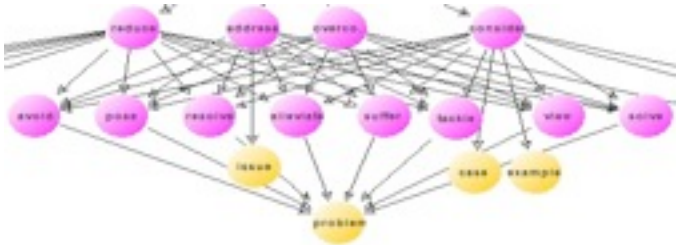
TERA (Textual Entailment)
- Decides for two text fragments (phrases, sentences, paragraphs), whether the one semantically entails the other
- Languages: DE, EN
- Accuracy: best results at international competitions (cf. Text Analysis Conference - TAC)
- Implementation: Java, OSGi

© DFKI, Günter Neumann, Juni 2011

Example - ConExt

- **Extraction of concept hierarchies** (based on Cimiano, Hotho, Staab; 2005)
- a method for extracting domain specific concept hierarchies from free text which has been syntactically analysed
- **Core idea:** words, which occur with same syntactic functions and context are semantically related.
- Can also be used to determine the terminology of a domain.
- semantically related words can be easier identified.
- Information extraction can benefit (e.g., by means of Query Expansion, Clustering)
- The strategy has been tested with several corpora:
  - business reports (3.964.925 sentences)
  - scientific publications (884.904 sentences)
  - Wikipedia, technical articles (628.560 Sentences)

Query expansion

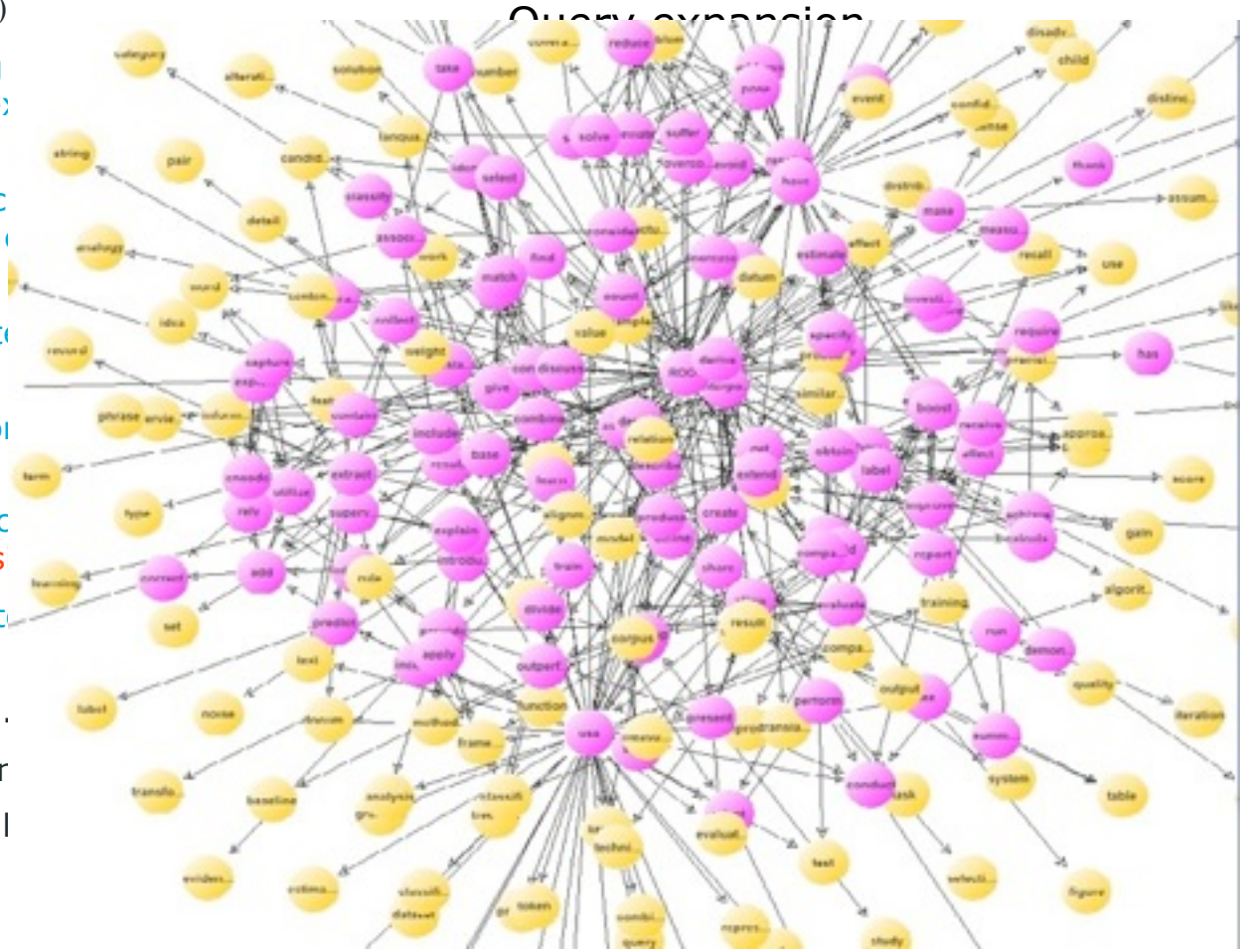


Clutsering

© DFKI, Günter Neumann, Juni 2011

Example - ConExt

- **Extraction of concept hierarchies** (based on Cimiano, Hotho, Staab; 2005)
- a method for extracting hierarchies from free te syntactically analysed
- **Core idea:** words, whic syntactic functions and related.
- Can also be used to det of a domain.
- semantically related wo identified.
- Information extraction c means of Query Expans
- The strategy has been t corpora:
  - business reports (3.
  - scientific publicatior
  - Wikipedia, technical Sentences)



© DFKI, Günter Neumann, Juni 2011

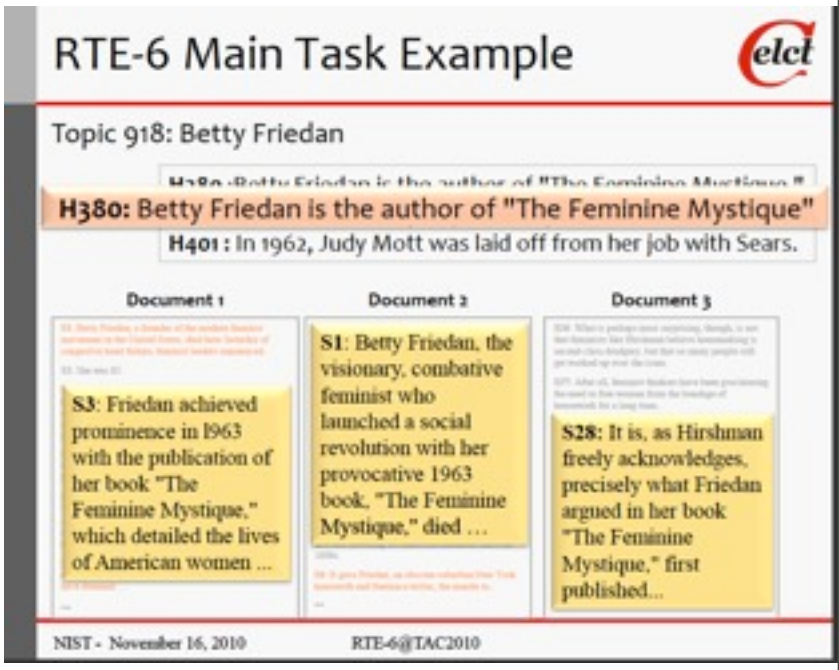
Example
- Textual Inference

- Textual Inference
  - Determine semantic inference relationships between texts (Recognition of text variants/paraphrases)
  - Entailment: Does Text T entails Text H ?
  - Needed in many text analytics applications (e.g., information extraction, question answering, text summarization)
  - Q: „Wo acquired Overture?“
    A: „Googles buying of Overture supports ...“
    ⇒: „X buys Y“ ≃ „X acquire Y“

© DFKI, Günter Neumann, Juni 2011

Example
- Textual Inference

- Textual Inference
  - Determine semantic inference relationships between texts (Recognition of text variants/paraphrases)
  - Entailment: Does Text T entails Text H ?
  - Needed in many text analytics applications (e.g., information extraction, question answering, text summarization)
  - Q: „Wo acquired Overture?“
    A: „Googles buying of Overture supports ...“
    ⇒: „X buys Y“ ≃ „X acquire Y“



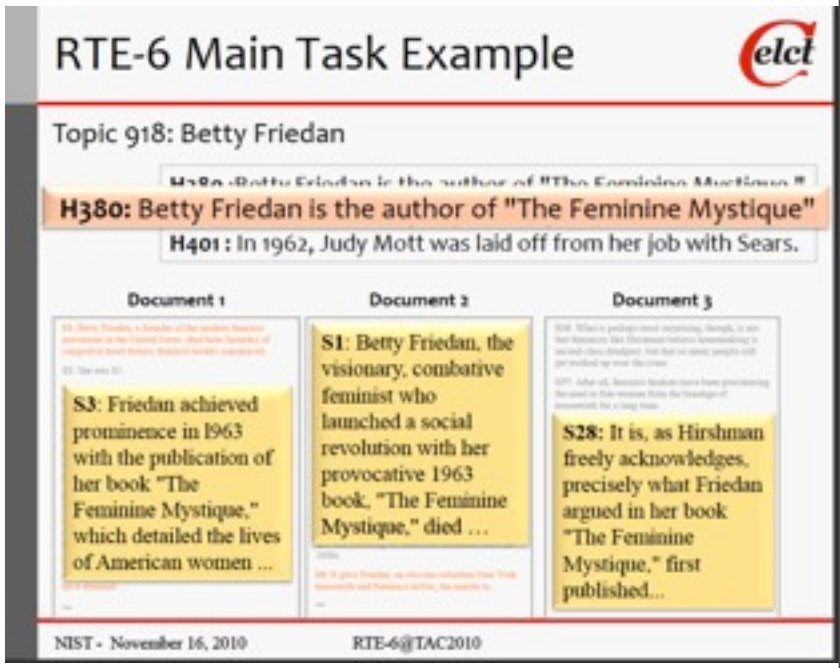
© DFKI, Günter Neumann, Juni 2011

Example
- Textual Inference

- Textual Inference
  - Determine semantic inference relationships between texts (Recognition of text variants/paraphrases)
  - Entailment: Does Text T entails Text H ?
  - Needed in many text analytics applications (e.g., information extraction, question answering, text summarization)
  - Q: „Wo acquired Overture?"
    A: „Googles buying of Overture supports ..."
    ⇒: „X buys Y" ≃ „X acquire Y"

- Textual inference in TechWatch Ordo
  - semantic search
  - concept extraction/ontology learning

- Textual inference (Entailment) is still a very young and dynamic R&D area !
  - Progress through international competitions (NIST, USA)
  - DFKI with very good results
  - EU project Excitement -> prob. from 2012 -> ~ 3.4 million EUR funding (DFKI is one of the 4 core scientific partners, 3 industrial partners; analysis of customer interaction data)



© DFKI, Günter Neumann, Juni 2011

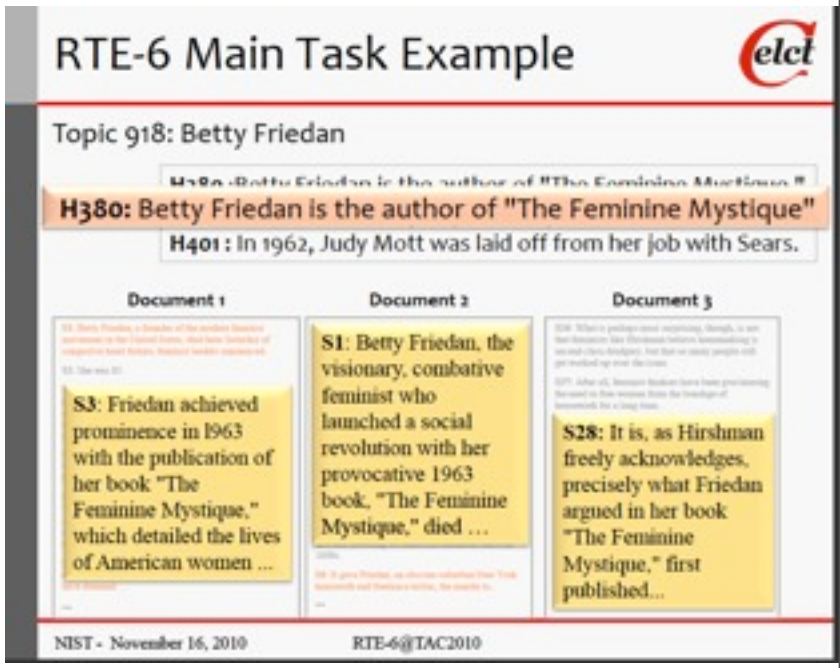
Example
- Textual Inference

THESEUS
Neue Technologien
für das Internet der Dienste

- Textual Inference
  - Determine semantic inference relationships between texts (Recognition of text variants/paraphrases)
  - Entailment: Does Text T entails Text H ?
  - Needed in many text analytics applications (e.g., information extraction, question answering, text summarization)
  - Q: „Wo acquired Overture?"
    A: „Googles buying of Overture supports ..."
    ⇒: „X buys Y" ≃ „X acquire Y"

- Textual inference in TechWatch Ordo
  - semantic search
  - concept extraction/ontology learning

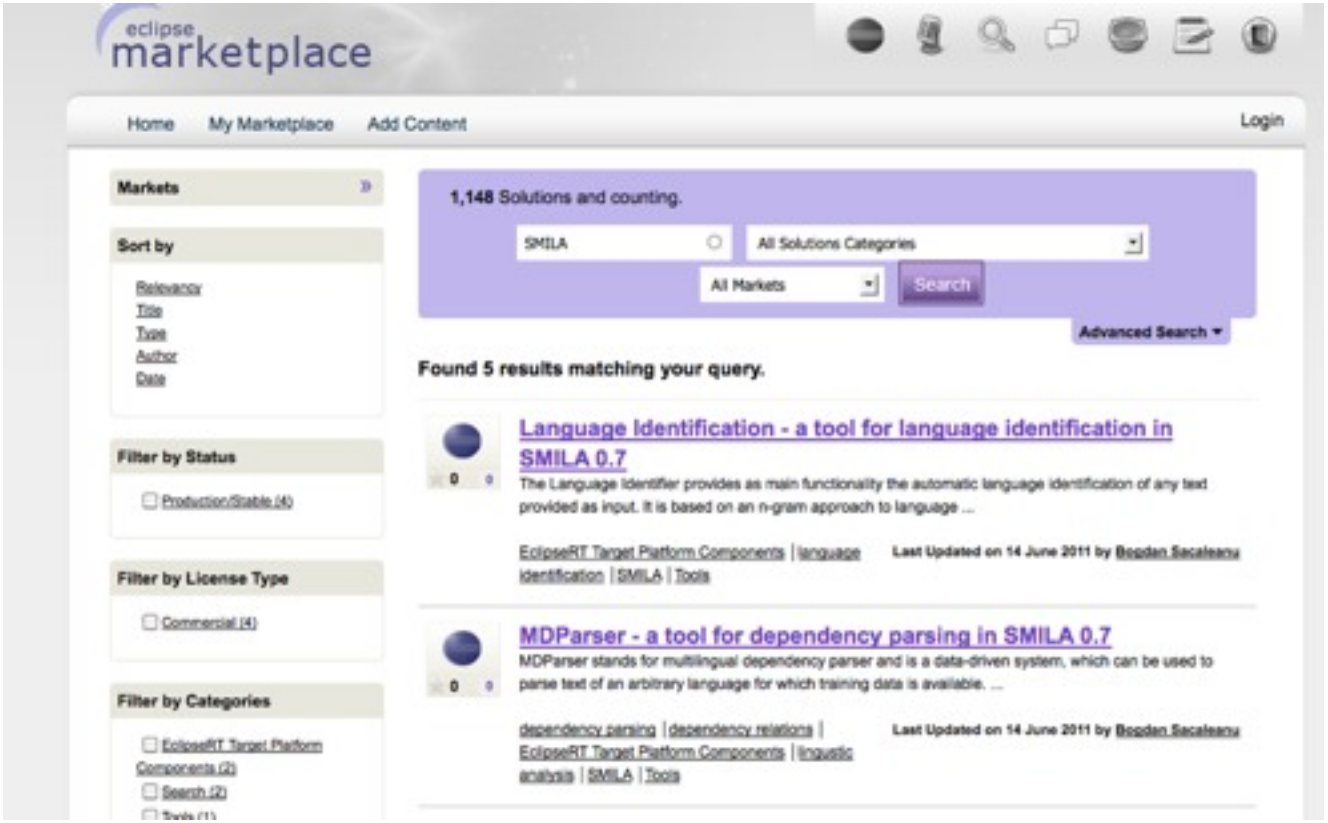
- Textual inference (Entailment) is still a very young and dynamic R&D area !
  - Progress through international competitions (NIST, USA)
  - DFKI with very good results
  - EU project Excitement -> prob. from 2012 -> ~ 3.4 million EUR funding (DFKI is one of the 4 core scientific partners, 3 industrial partners; analysis of customer interaction data)



Focus:
- robustness, scalability
- organized by NIST, USA
- 2010: 19 Unis/Institute (world wide), DFKI, 4. best result in Novelty Subtask (with new team) (2009: 2. best result in Main Task.)

© DFKI, Günter Neumann, Juni 2011

Eclipse Marketplace



Current

- Implementation: SMILA Piplets (Java-OSGi web service)
- Direct integration into SMILA-process
- License: Research & commercial

Planned

- NE-Hub – Multilingual architecture for NE recognition

© DFKI, Günter Neumann, Juni 2011

Experience of our SMILA developers

Pros

- Rapid entry point for End-User by means of:
  - detailed documentation
  - many good examples
  - Out-of-the-box runnable system
- Parallelism of workflow through BPEL
- Fast composition of new apps by means of existing components and BPEL
- Rapid and competent support by means of the forum

Cons

- Complicated data structures Version 0.7 – solved in Version 0.8
- Restricted extendability of the index algorithms in order to support several dynamically changing indices per user (a very specific requierement)

© DFKI, Günter Neumann, Juni 2011