Language Technology and the Semantic Web

Dr. Günter Neumann

http://www.dfki.de/~neumann

Language Technology-Lab

DFKI, Saarbrücken

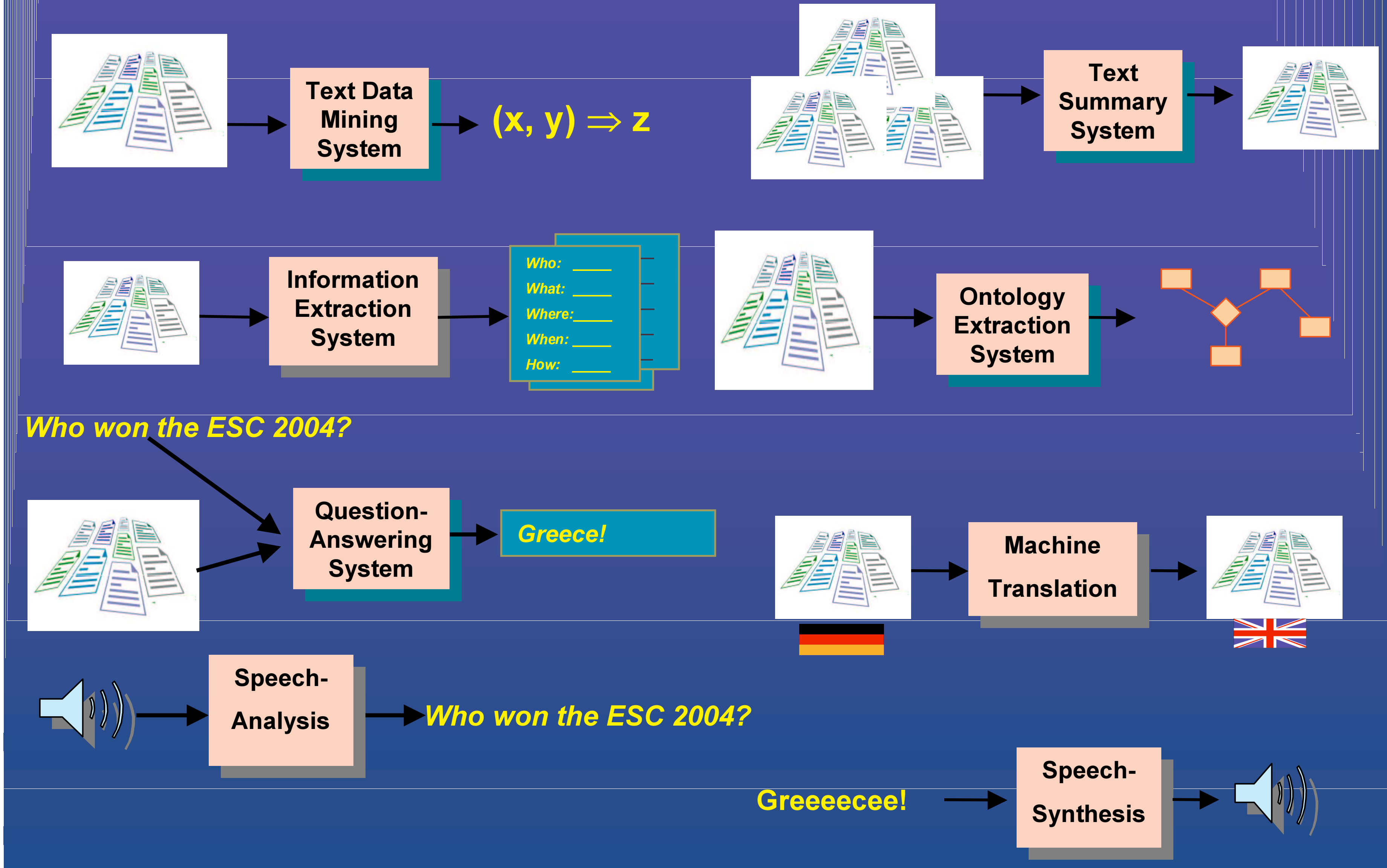
Overview

- Language Technology
- Semantic Web
- Information Extraction
- Information Access

Human Language Technology

- *Human Language Technology LT* – covers
  - The design and implementation of algorithms, data and electronic devices for processing of natural language (text and speech), and
  - Their integration into real-world applications and products
- Language Technology defines the engineering part of computational linguistic

LT-methods cover many areas



Text Data Mining System

(x, y) ⇒ z

Text Summary System

Information Extraction System

Who: ____

What: ____

Where:____

When: ____

How: ____

Ontology Extraction System

Who won the ESC 2004?

Question-Answering System

Greece!

Machine Translation

Speech-Analysis

Who won the ESC 2004?

Greeeecee!

Speech-Synthesis

Multi/cross-linguality is of great importance in all these areas!

LT as embedded part of applications

- Human-Machine Communication

Integration

•Modularity

•Multi-media

•Software-Engineering standards

- Data-oriented Knowledge Acquisition

High Performance

•Real-time

•Robustness

•Scalability

•Adaptation

•Evaluation

Language Technology

- **Core technology**
  - Efficient data structures
  - Weighted finite state automata
  - Machine learning
  - Statistical inference

- **LT-Methods**
  - Named Entity-Recognition
  - PoS/Sem-Tagging
  - Controlled Languages
  - Integration of shallow & deep NLP („text zooming“)
  - Reference-resolution
  - NL-oriented ontologies

- Already a successful technology transfer
  - Industry (Microsoft, IBM, Siemens, Telekom, ...) & Spin-offs, competence centers, ...
  - Speech-systems, MT, Editors, Text-Mining, Knowledge-Mining Content-Management, ...

- Newest Technology Hype: the Semantic Web
  - What role does it play for LT?

The Semantic Web (SW)

- Tim Berners-Lee, 1998:
  - “This document is a plan for achieving a set of connected applications for data on the Web in such a way as to form a consistent logical web of data (semantic web).”

- Tim Berners-Lee et al., 2001
  - “… an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.”

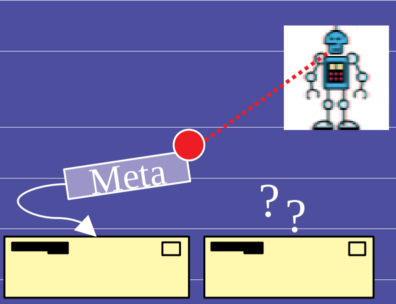
SW – illustrated

1 Extension of the Current Web

The existing web will further emerge, so that computers can understand content on-line, to better help humans to organize, search, and exchange information.

2 Add meta-data

Data over data; Structural linkage of heterogeneous data sources

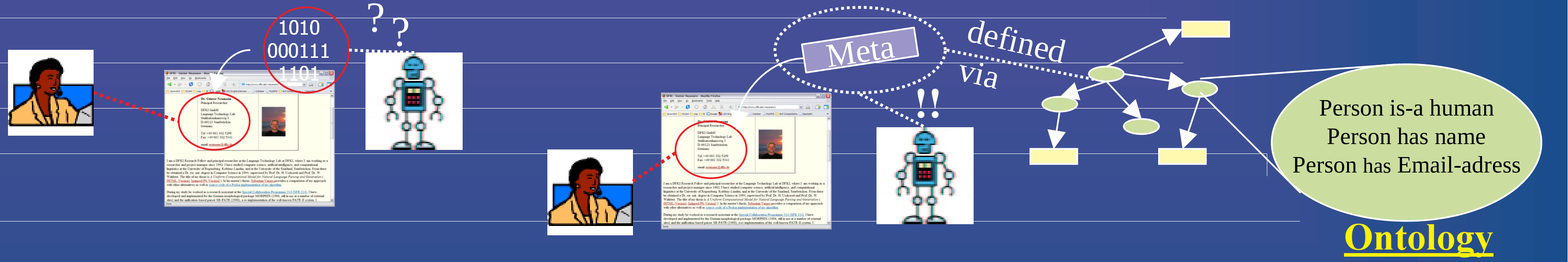


3 Ontologies associate meaning to meta-data

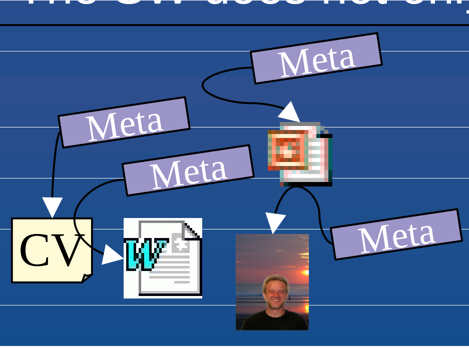
SW exists of meta-data and links to global ontolgoies, which define the meaning of terms.

An ontology serves as a structural vocabulary for the interpretation of domain-specific terms.

4 Strukturiertes Web von Daten



5 The SW does not only consider Web-pages



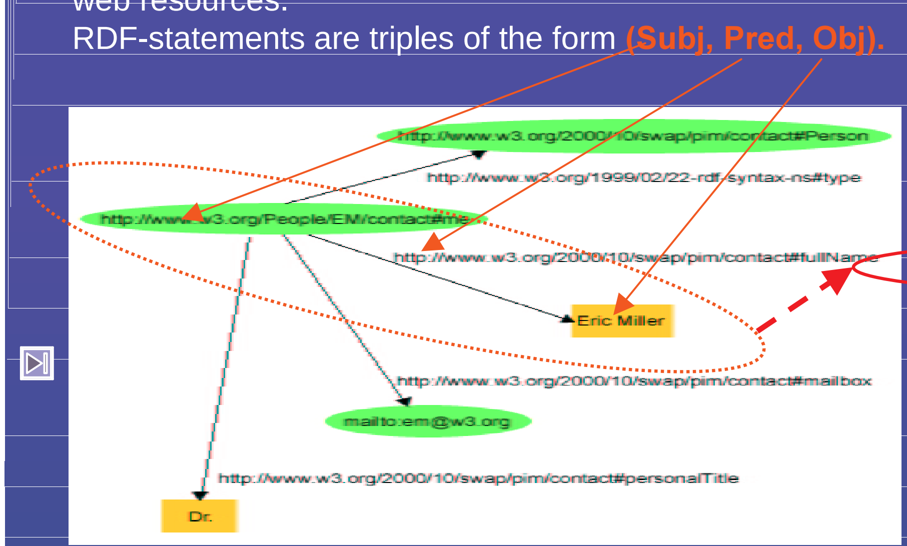
6 How will I use the SW?

- Intelligent information search;
- Automatic support for the management of my personal information on the SW

RDF and OWL: Modeling data on the SW

1 RDF: Resource Description Framework

RDF is language for the representation of meta-data over web resources.
RDF-statements are triples of the form **(Subj, Pred, Obj).**



http://www.w3.org/2000/10/swap/pim/contact#Person
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://www.w3.org/People/EM/contact#me
http://www.w3.org/2000/10/swap/pim/contact#fullName
Eric Miller
http://www.w3.org/2000/10/swap/pim/contact#mailbox
mailto:em@w3.org
http://www.w3.org/2000/10/swap/pim/contact#personalTitle
Dr.

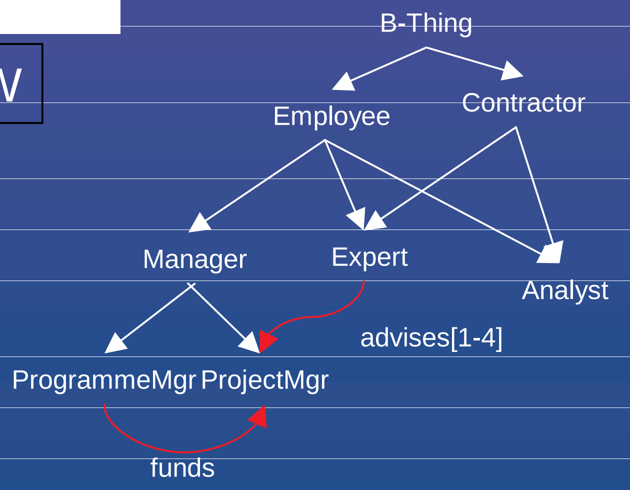
2 XML & N3 sind alternative RDF-Syntaxen

XML schematically: <Subj> <Pred> Obj </Pred> </Subj>

N3:
@prefix rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns# .
@prefix contact: http://www.w3.org/2000/10/swap/pim/contact# .
@prefix EM: http://www.w3.org/People/EM/contact# .

EM:me rdf:type contact:person .
EM:me contact:full-name “Eric Miller” .
EM:me contact:personalTitle “Dr.” .
EM:me contact:mailbox rdf:resoure “mailto:em@w3.org .

3 OWL: Web Ontology Language

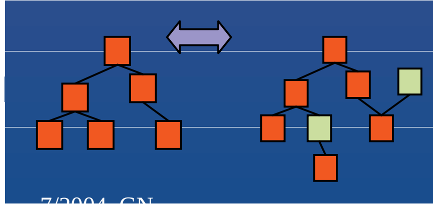


- some RDF-statements have a fix interpretation (is-a, =, inverseOf, card, ...)
- **Sharing** of information between individuals from multiple documents ⇒ Web of data from heterogeneous sources
- Semantic of OWL as basis for inference mechanism over these data structures.

4 Relevante Aspekte für das SW

standardization, Web-globalization, distribution of resources

5 Ontology Mapping



Mapping between distributed, local ontologies

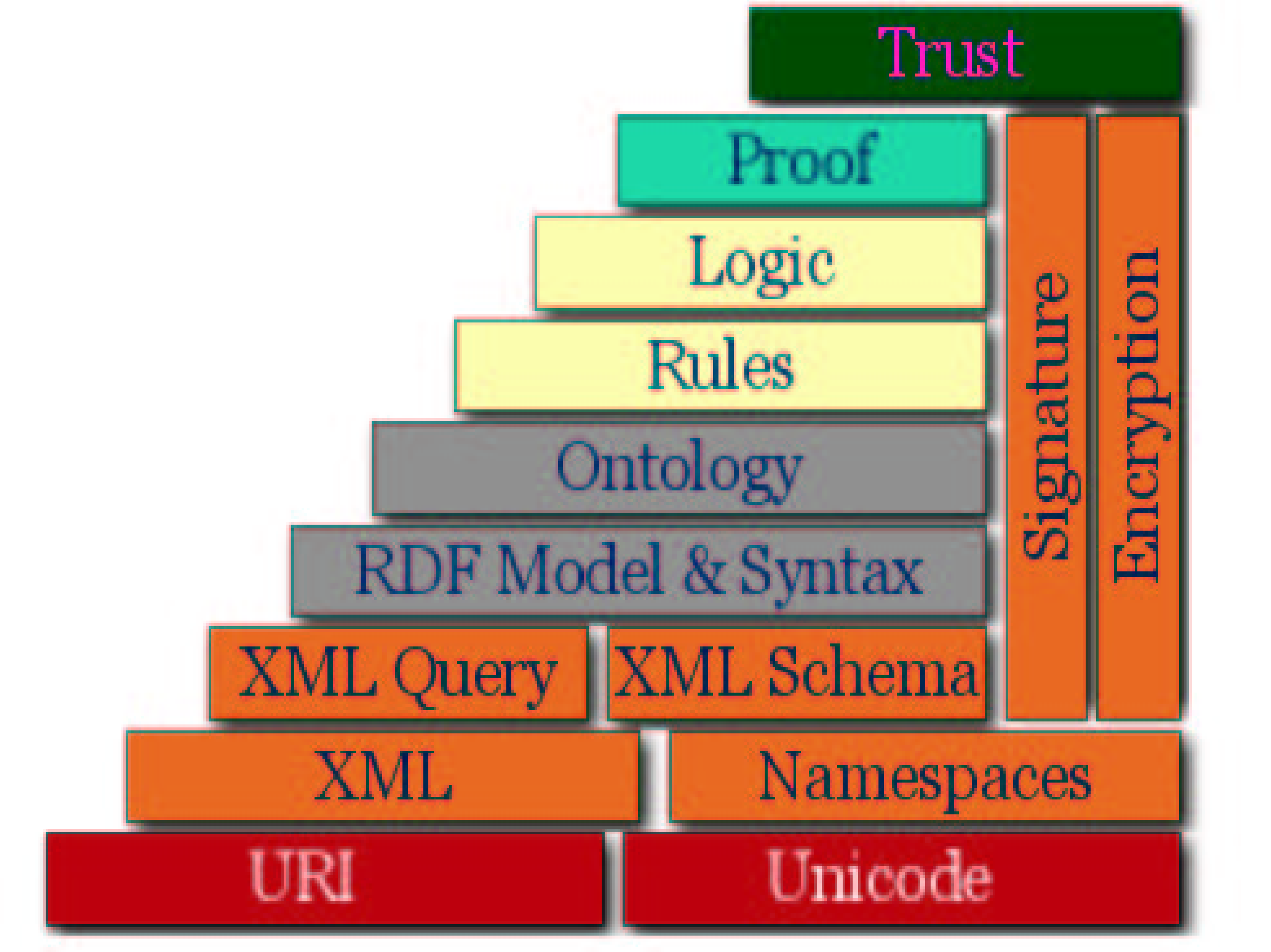
The SW-pyramid

© W3C

Basic research

Current focus of major efforts

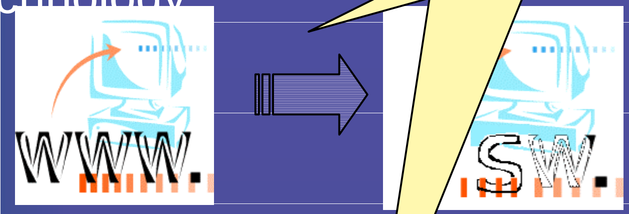
Established standards



Relevance of LT for SW

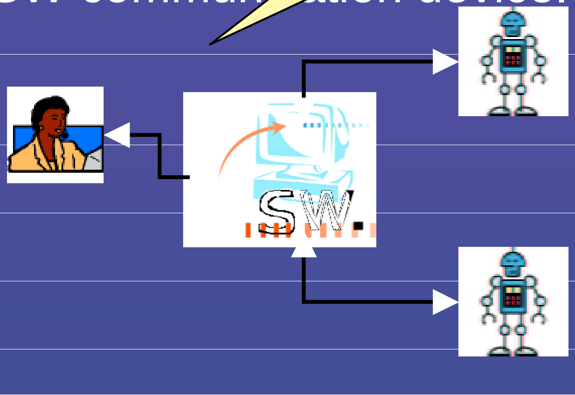
Intelligent Information Extraction

1 During the transi
WWW to the SW, L
technology

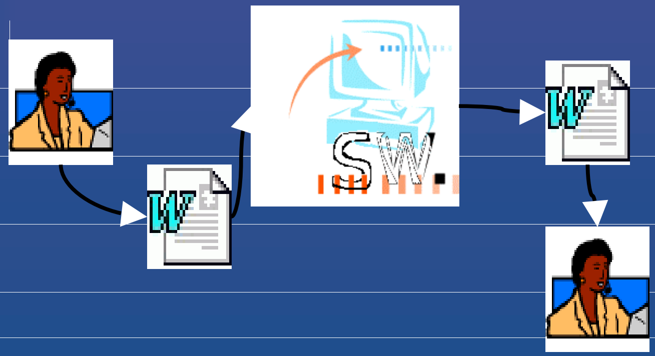


Intelligent Information Access

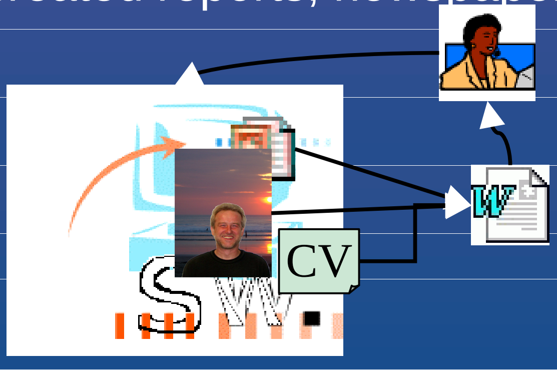
2 As long as the hum et
Loop", NL will remain to e core
Human-SW communication device.



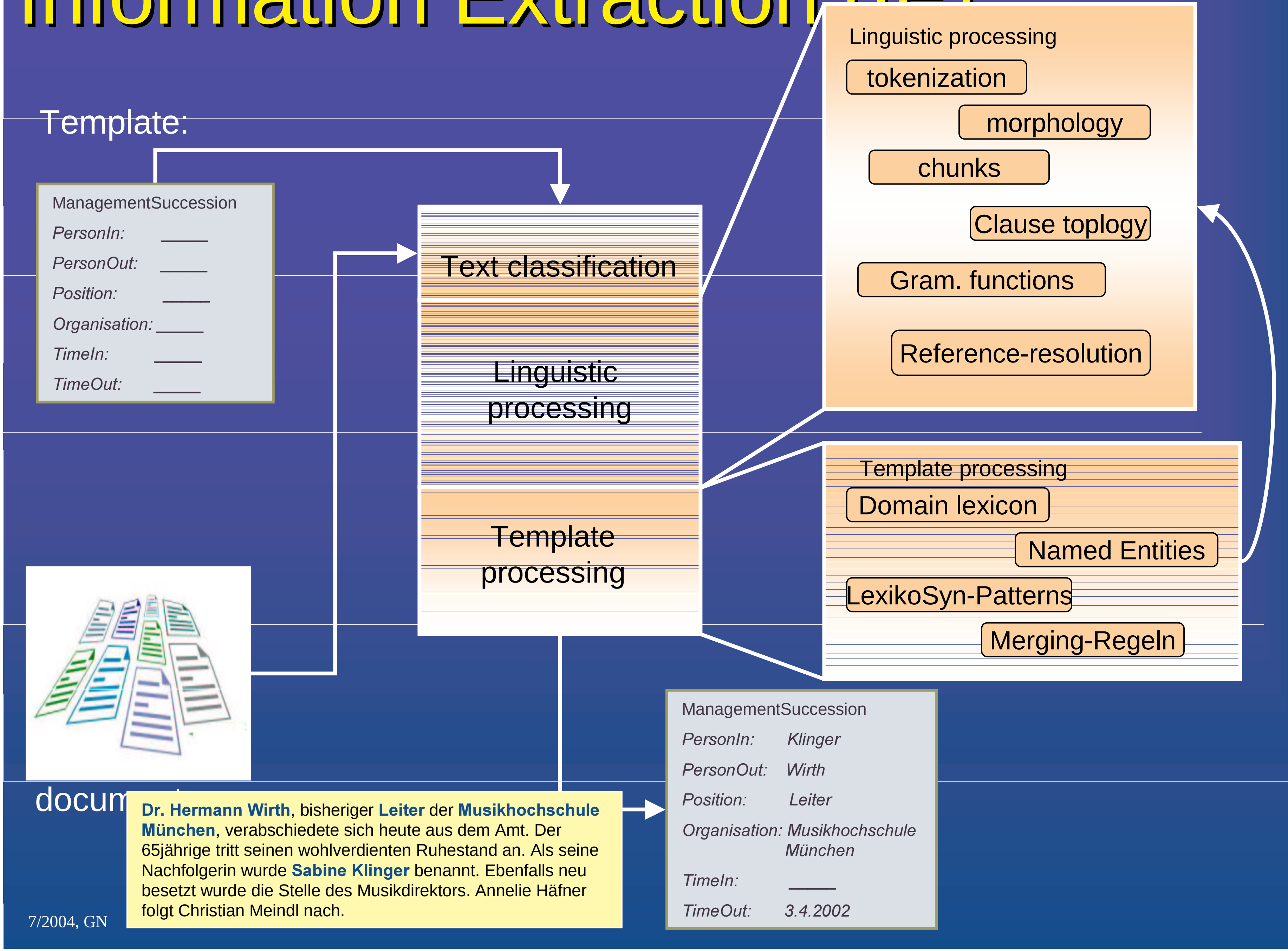
3 Humans will also in the future exchange knowledge via NL documents: Semantically annotated documents as Human-SW interface



4 NL-generation of information in form of NL-Text, e.g., heterogeneous resources, dynamically created reports, newspapers, …



Information Extraction (IE)



Template:

ManagementSuccession
PersonIn: _____
PersonOut: _____
Position: _____
Organisation: _____
TimeIn: _____
TimeOut: _____

Dr. Hermann Wirth, bisheriger **Leiter** der **Musikhochschule München**, verabschiedete sich heute aus dem Amt. Der 65jährige tritt seinen wohlverdienten Ruhestand an. Als seine Nachfolgerin wurde **Sabine Klinger** benannt. Ebenfalls neu besetzt wurde die Stelle des Musikdirektors. Annelie Häfner folgt Christian Meindl nach.

ManagementSuccession
PersonIn: *Klinger*
PersonOut: *Wirth*
Position: *Leiter*
Organisation: *Musikhochschule München*
TimeIn: _____
TimeOut: *3.4.2002*

IE for semantic annotation

Identification of IE-sub-tasks:

- basic entities (e.g., proper names)
- binary relations between entities
- n-ary relations/events

Machine learning!

Automatic Content Extraction (ACE)

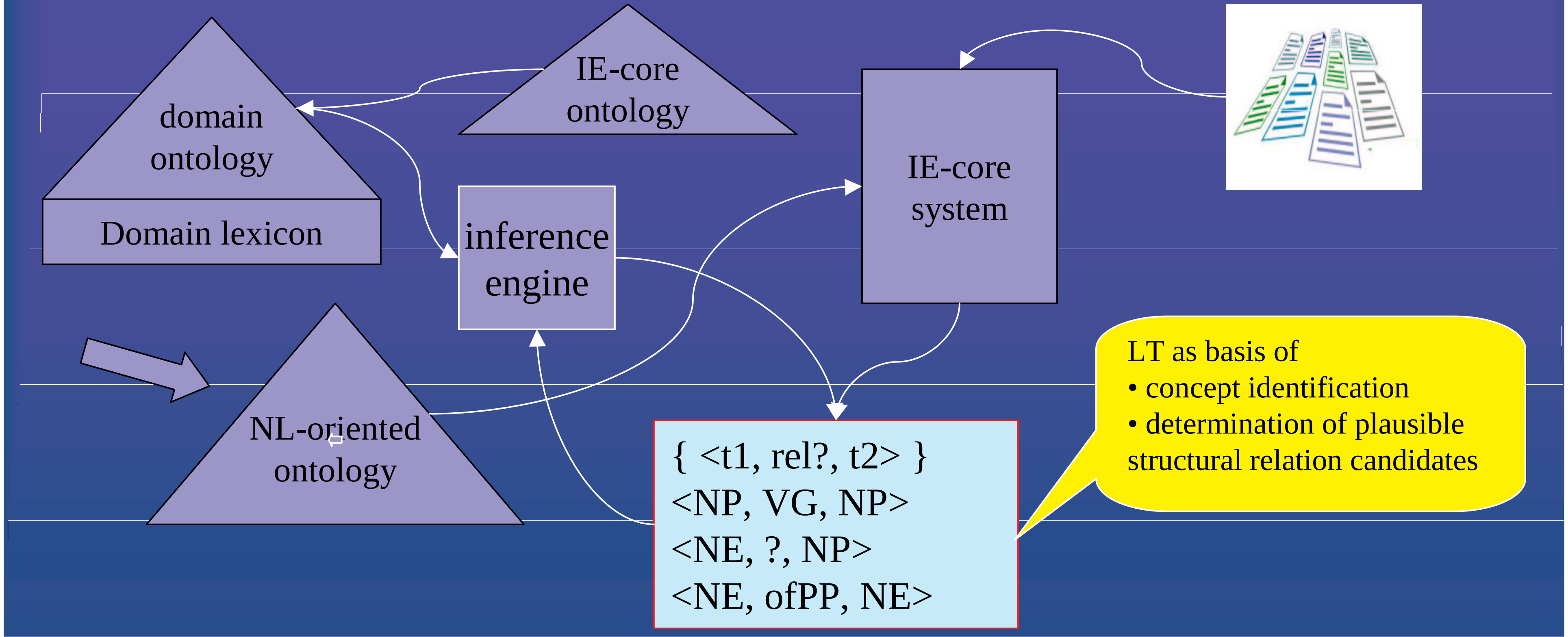
- Spezification of an IE-core-ontology
- Annotation-specification & -tools
- Templates as specializations of the IE-core-ontology (also multi-templates)

IE as core for semantic annotation

- identification
- discovery
- validation
- evaluation

of semantic relationships & as basis for the automatic creation of meta data

IE for semantic annotation



domain ontology

Domain lexicon

IE-core ontology

inference engine

IE-core system

NL-oriented ontology

{ <t1, rel?, t2> }
<NP, VG, NP>
<NE, ?, NP>
<NE, ofPP, NE>

LT as basis of
- concept identification
- determination of plausible structural relation candidates

Example for entities & their mentions

[COLOGNE, [Germany]] (AP) _ [A [Chilean] exile] has filed a complaint against [former [Chilean] dictator Gen. Augusto Pinochet] accusing [him] of responsibility for [her] arrest and torture in [Chile] in 1973, [prosecutors] said Tuesday.

[The woman, [[a Chilean] who has since gained [German] citizenship]], accused [Pinochet] of depriving [her] of personal liberty and causing bodily harm during [her] arrest and torture.

Person
Organization
Geopolitical Entity

LT-challenges

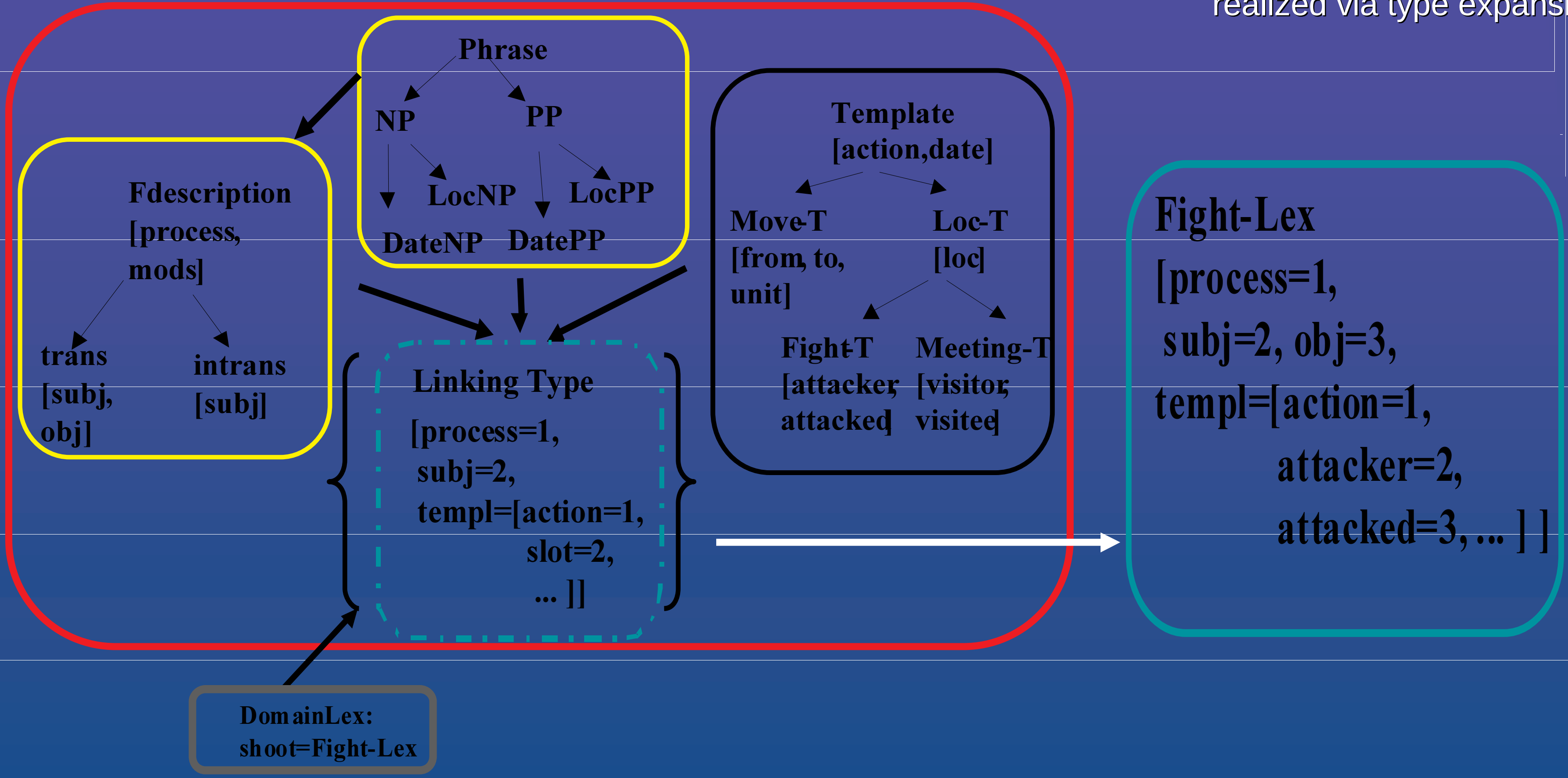
Identification of verbalizations/mentioning of concepts/instances

- Linking of domain ontology and NL-oriented ontology (e.g., WordNet)
- Paraphrasing
- Metonymy ("Peking orgainzes the Olympic Games 2008.")
- Reference identification ("Chancellor Schröder, Schröder, the German chancellor, he, …")
- Analysis of sublanguages as basis for adaptive IE (cf. Grishman, 2001)

Domain modeling in DFKI system SMES is realised using typed feature structures

- Domain modeling via hierarchy of templates (black box), using the formalism TDL, which is also used to model hierarchies of linguistic objects (yellow boxes).

- The interface between domain knowledge and linguistic entities is specified via *linking types* (green box), which represent a close connection between concepts of the different layers, and which are accessible via the domain lexicon (brown & green box). Template-filling is then realized via type expansion.



7/2004, GN

NL-annotations for the SW

Starting point: START multi-media QA system, by Boris Katz et al

T-expression
<subject relation object>

Central issues

1. Sentence-based N Analysis
2. NL-annotations fo multi-media information segr ents

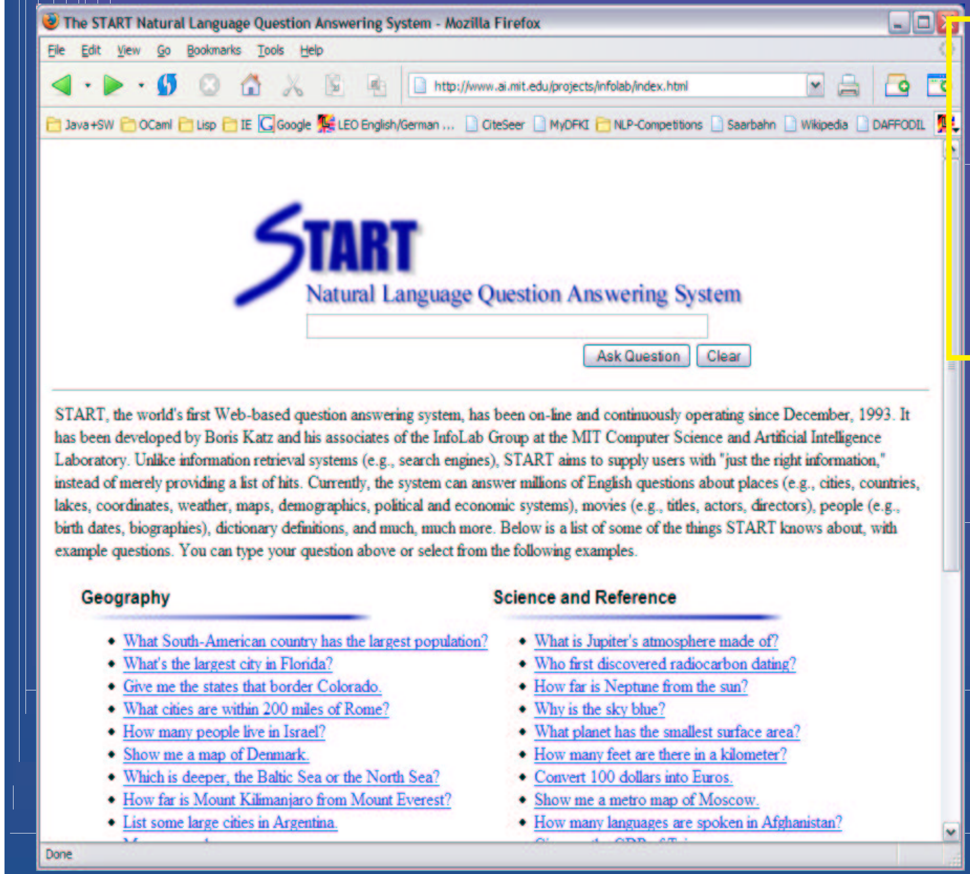
Bill surprised Hillary with his answer

<<Bill surprise Hillary> with answer>
<answer related-to Bill>

Processing of huge text collections:

1. Extraction of relevant sentences from texts.
2. Syntax analysis
3. Annotation of the texts with syntax

NL-Question
Whose answer surprised Hillary?
<answer surprise Hillary>
<answer related-to ***whom***>


The START Natural Language Question Answering System - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.ai.mit.edu/projects/infolab/index.html
Java+SW OCaml Lisp IE Google LEO English/German ... CiteSeer MyDFKI NLP-Competitions Saarbahn Wikipedia DAFFODIL
START
Natural Language Question Answering System
Ask Question
Clear
Done

START, the world's first Web-based question answering system, has been on-line and continuously operating since December, 1993. It has been developed by Boris Katz and his associates of the InfoLab Group at the MIT Computer Science and Artificial Intelligence Laboratory. Unlike information retrieval systems (e.g., search engines), START aims to supply users with "just the right information," instead of merely providing a list of hits. Currently, the system can answer millions of English questions about places (e.g., cities, countries, lakes, coordinates, weather, maps, demographics, political and economic systems), movies (e.g., titles, actors, directors), people (e.g., birth dates, biographies), dictionary definitions, and much, much more. Below is a list of some of the things START knows about, with example questions. You can type your question above or select from the following examples.

Geography

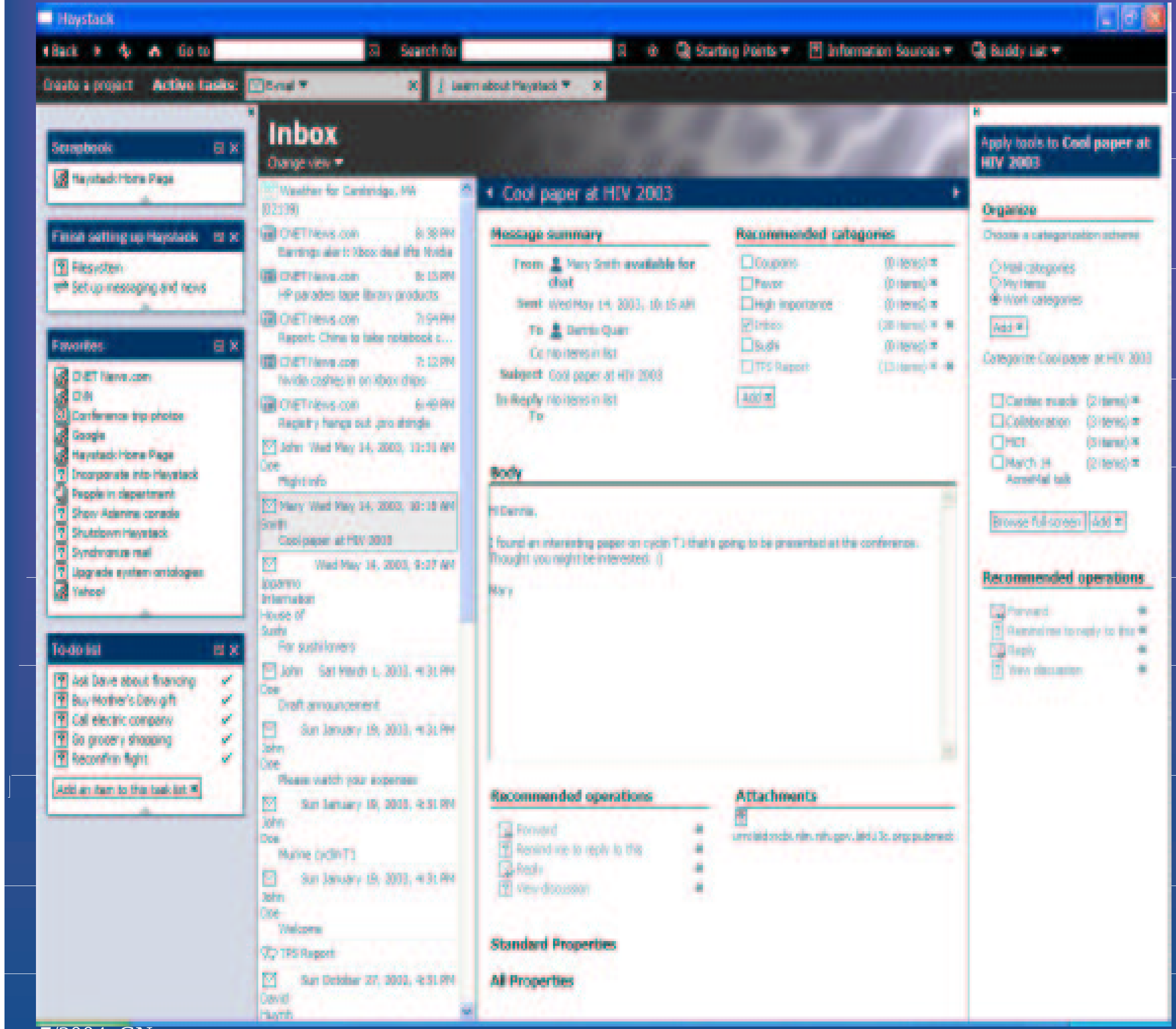
- What South-American country has the largest population?
- What's the largest city in Florida?
- Give me the states that border Colorado.
- What cities are within 200 miles of Rome?
- How many people live in Israel?
- Show me a map of Denmark.
- Which is deeper, the Baltic Sea or the North Sea?
- How far is Mount Kilimanjaro from Mount Everest?
- List some large cities in Argentina.

Science and Reference

- What is Jupiter's atmosphere made of?
- Who first discovered radiocarbon dating?
- How far is Neptune from the sun?
- Why is the sky blue?
- What planet has the smallest surface area?
- How many feet are there in a kilometer?
- Convert 100 dollars into Euros.
- Show me a metro map of Moscow.
- How many languages are spoken in Afghanistan?

Haystack: the universal information client

http://haystack.lcs.mit.edu/



Haystack

Back · Go to · Search for · Starting Points · Information Sources · Buddy List

Create a project · Active tasks: E-mail · Learn about Haystack

Scrapbook
- Haystack Home Page

Finish setting up Haystack
- Files/often
- Set up messaging and news

Favorites
- CNET News.com
- CNN
- Conference trip photos
- Google
- Haystack Home Page
- Incorporate into Haystack
- People in department
- Show Adenine console
- Shutdown Haystack
- Synchronize mail
- Upgrade system ontologies
- Yahoo!

To-do list
- Ask Dave about financing ✓
- Buy Mother's Day gift ✓
- Call electric company ✓
- Go grocery shopping ✓
- Reconfirm flight ✓

Add an item to the task list

Inbox

Change view

- Weather for Cambridge, MA (02139)
- CNET News.com 8:38 PM Earnings alert: Xbox deal lifts Nvidia
- CNET News.com 8:15 PM HP parades tape library products
- CNET News.com 7:54 PM Report: China to take notebook c...
- CNET News.com 7:12 PM Nvidia cashes in on Xbox chips
- CNET News.com 6:49 PM Registry hangs out .pro shingle
- John Doe Wed May 14, 2003, 11:31 AM Flight info
- Mary Smith Wed May 14, 2003, 10:15 AM Cool paper at HIV 2003
- Wed May 14, 2003, 9:27 AM For sushi lovers
- John Doe Sat March 1, 2003, 4:31 PM Draft announcement
- John Doe Sun January 19, 2003, 4:31 PM Please watch your expenses
- John Doe Sun January 19, 2003, 4:31 PM Murine cyclin T1
- John Doe Sun January 19, 2003, 4:31 PM Welcome
- TPS Report
- David Huynh Sun October 27, 2002, 4:31 PM

Cool paper at HIV 2003

Message summary

From: Mary Smith available for chat

Sent: Wed May 14, 2003, 10:15 AM

To: Dennis Quan

Cc: no items in list

Subject: Cool paper at HIV 2003

In Reply To: no items in list

Recommended categories

□ Coupons	(0 items)
□ Favor	(0 items)
□ High importance	(0 items)
☑ Inbox	(28 items)
□ Sushi	(0 items)
□ TPS Report	(13 items)

Add

Body

Hi Dennis,

I found an interesting paper on cyclin T1 that's going to be presented at the conference. Thought you might be interested. :)

Mary

Recommended operations
- Forward
- Remind me to reply to this
- Reply
- View discussion

Attachments

Standard Properties

All Properties

Apply tools to Cool paper at HIV 2003

Organize

Choose a categorization scheme

○ Mail categories

○ My items

● Work categories

Add

Categorize Cool paper at HIV 2003

□ Cardiac muscle	(2 items)
□ Collaboration	(3 items)
□ HCI	(3 items)
□ March 14 AcrossMail talk	(2 items)

Browse full-screen · Add

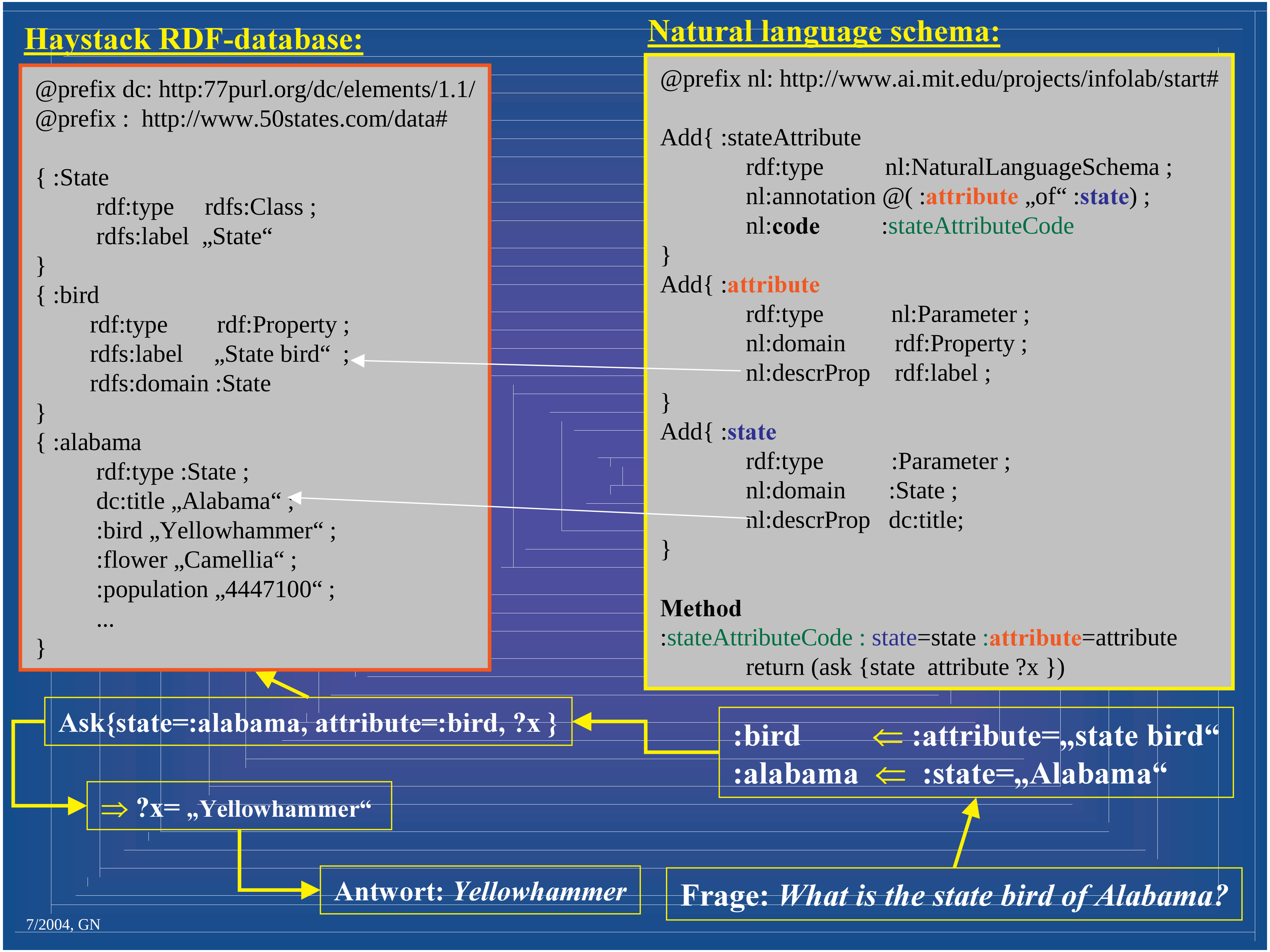
Recommended operations
- Forward
- Remind me to reply to this
- Reply
- View discussion

Motivation:
semantic annotation should be a side-effect of daily use of computer.

Idea:
Personalized information portal for all relevant services, like email, documents, calender, Web-pages, ...

Collection of all data uniformly via RDF-database

Programming language Adenine for the manipulation of frequent (i.e., as support for the implementation of specific service programs).



Haystack RDF-database:

```
@prefix dc: http:77purl.org/dc/elements/1.1/
@prefix :  http://www.50states.com/data#

{ :State
        rdf:type     rdfs:Class ;
        rdfs:label  „State“
}
{ :bird
       rdf:type        rdf:Property ;
       rdfs:label      „State bird“  ;
       rdfs:domain :State
}
{ :alabama
        rdf:type :State ;
        dc:title „Alabama“ ;
        :bird „Yellowhammer“ ;
        :flower „Camellia“ ;
        :population „4447100“ ;
        ...
}
```

Natural language schema:

```
@prefix nl: http://www.ai.mit.edu/projects/infolab/start#

Add{ :stateAttribute
           rdf:type          nl:NaturalLanguageSchema ;
           nl:annotation @( :attribute „of“ :state) ;
           nl:code          :stateAttributeCode
}
Add{ :attribute
           rdf:type           nl:Parameter ;
           nl:domain         rdf:Property ;
           nl:descrProp    rdf:label ;
}
Add{ :state
           rdf:type           :Parameter ;
           nl:domain        :State ;
           nl:descrProp   dc:title;
}

Method
:stateAttributeCode : state=state :attribute=attribute
           return (ask {state  attribute ?x })
```

Ask{state=:alabama, attribute=:bird, ?x }

:bird ⇐ :attribute=„state bird“
:alabama ⇐ :state=„Alabama“

⇒ ?x= „Yellowhammer“

Antwort: *Yellowhammer*

Frage: *What is the state bird of Alabama?*

7/2004, GN

Example:
Linking of t-expressions & RDF

```
@prefix nl: http://www.ai.mit.edu/projects/infolab/start#

Add{ :Person
        rdf:type          rdfs:Class ;
}

Add{ :homeAddress
        rdf:type           rdf:Property ;
        rdfs:domain    :Person ;

        nl:annotation  @(nl:subj „lives at“ nl:obj) ;
        nl:annotation  @(nl:subj „‘s home adress is“ nl:obj) ;
        nl:annotation  @(nl:subj „‘s apartment“ nl:obj) ;

        nl:generation  @(nl:subj „‘s home address is“ nl:obj) ;

}
```

Remarks:

- NL-annotations as a means for controlling the paraphrasing potential of NL expressions
- Richer linguistic annotations are possible (e.g., fine-grained grammatical functions, agreement)
- Also relevant for user-oriented adaptation of service programs

Natural language annotations for the SW

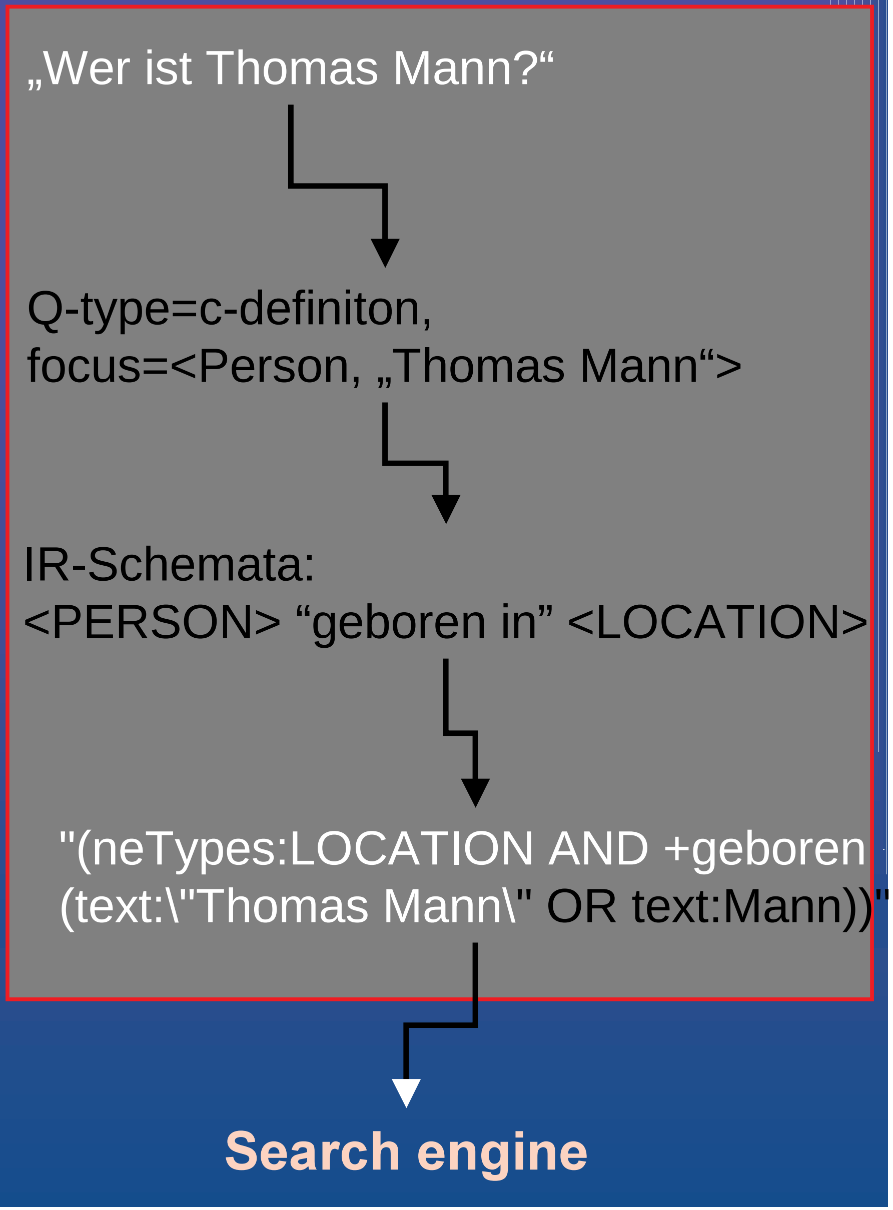
- NL used as meta-data
  - Readability of RDF
  - Supports transition from WWW to SW
  - NL-annotation specifies which kind of (NL)-question a meta-data is able to answer
⇒ controlled question-answering systems
- Information access (IA) within SW
  - Development of programs, which help a user to locate, to collect, to compare and to link information
- NL is the most natural way for user to perform IA
  - SW should support in the same way IA using specialized languages/exchange formats & NL

Relevance

- Approach is open for future extensions:
  - statistical-based models (add weight to the NL-annotations)
  - Machine Learning of NL-annotations on basis fo ontology-oriented IE (cf. Hovy et al. 2002)
- The current mechanism of NL-annotations is idiosyncratic, however at DFKI we plan the following:
  - Exploration of a linking mechanism between dependency structure and RDF/OWL
  - Foundation for novel template-based QA-strategies

Example for the processing of complex questions

- Approach:
  - Select templates via Q-Type & Q-Focus:
    - Definition question, list-question
    - Person: born-where, born-when, business-what ⇒ Ontology
  - Pro property P, select IR-Schema:
    - NL-based query-pattern
  - P might be:
    - From the set of known NE-types (person, location, date, …)
      ⇒ answer-type
    - NL-Phrase, which “describes” P, in case no a-type can be determined
- Compute for each P für jede P one/several IR-Query-terms, e.g.,
  - NE-type:person & text:<query term>



„Wer ist Thomas Mann?“

Q-type=c-definiton,
focus=<Person, „Thomas Mann“>

IR-Schemata:
<PERSON> “geboren in” <LOCATION>

"(neTypes:LOCATION AND +geboren (text:\"Thomas Mann\" OR text:Mann))"

Search engine

IE-based question answering

- Approach can also be used for template-based questions:
  - let t ∈ T, set of templates, which are known to the system – via IE-Ontology – e.g., “management-sucession-Template”
  - for all properties E of t, combine E with NL-schema
    - E.g., “Person-In” ⇒ (<PERS> “is_successor_of” <PERS>)
- Answering of complex questions
  - As composition of the answering of – relative to the conceptual description – simple questions
  - Implementation of this approach as part of the DFKI project Quetal (prototype as part of DFKI’s qa@clef-2004 system)
  - Interactive online IE through close integration of IE & IA

Concluding remarks

- LT is a key technology for the construction of the Semantic Web
- Very high requirements on
  - Performance
  - Modularity & integration
  - scalability & on-demand availability
  - Domain & user adaptation
- Systematic evaluation of LT-methods
  - Driving power & revisions of futuer developments
- In the future, cognitive-based methods will be considered
  - as inspiration for more effectiv LT-methods