THESEUS
Forschungsprogramm für eine neue internetbasierte Wissensinfrastruktur

# Towards Evaluating the Core Technology Cluster of the German Research Project THESEUS


Peter Dunker
dkr@idmt.fraunhofer.de
(Deputy Evaluation Workpackage Lead)

Juan José Bosch Vicente
bsh@idmt.fraunhofer.de
Judith Liebetrau
ltu@idmt.fraunhofer.de

Fraunhofer Institute for
Digital Media Technology (IDMT),
Ilmenau, Germany


Aarhus, 16.09.2008

# Outline

THESEUS
Forschungsprogramm für eine neue internetbasierte Wissensinfrastruktur

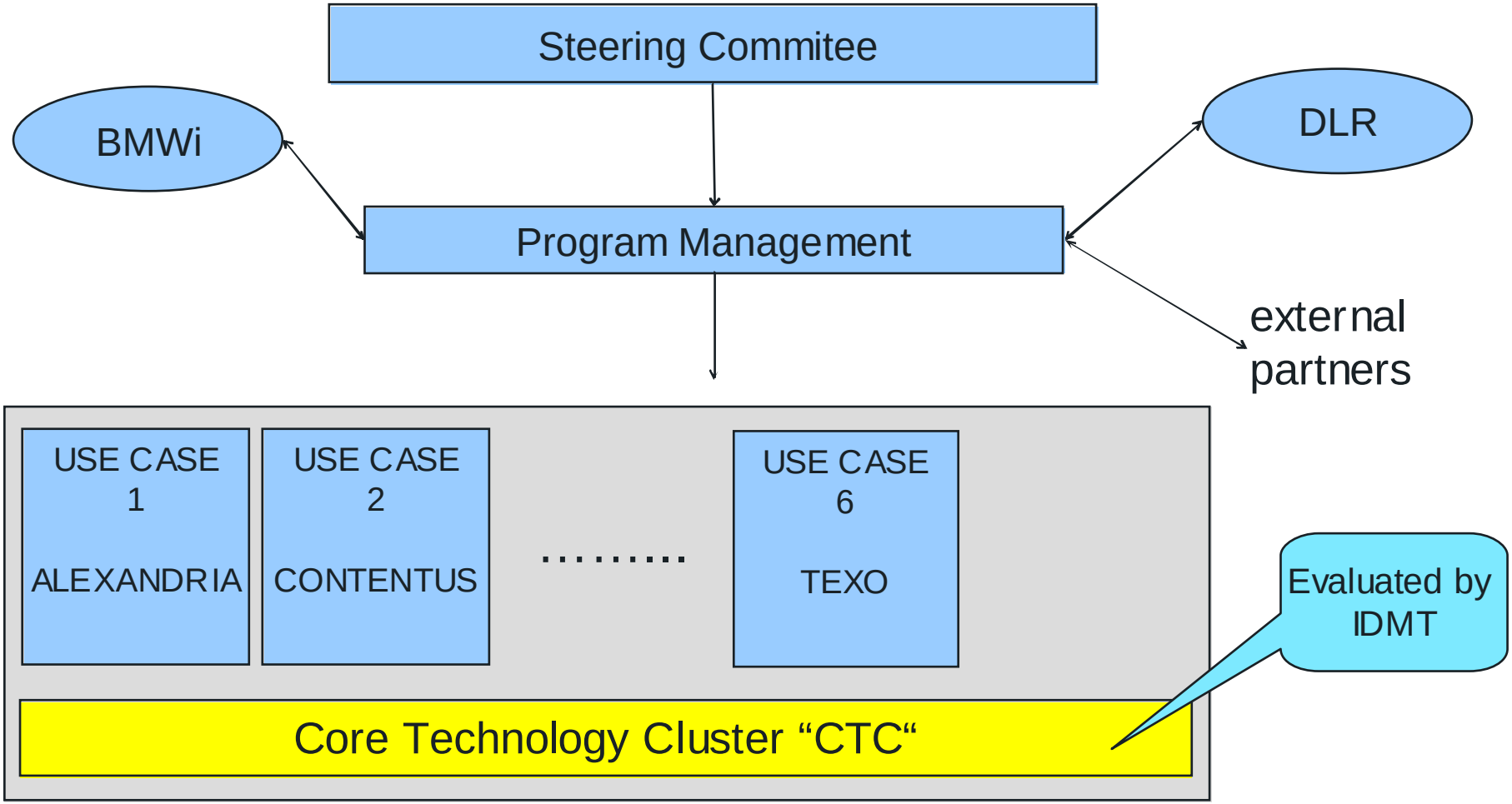
- THESEUS facts & organisation
- Core Technology Cluster – CTC

- Evaluation of CTC
  - Multimedia Analysis
  - Multimedia Quality
  - Iterative System Design and Quality in Use

- Conclusions

# THESEUS Organisation

THESEUS
Forschungsprogramm für eine neue internetbasierte Wissensinfrastruktur



# THESEUS Facts

THESEUS
Forschungsprogramm für eine neue internetbasierte Wissensinfrastruktur

## Some facts about THESEUS:

- Number of partners: 22 (30, including 9 Fraunhofer institutes)
- Start: ≈ mid 2007
- Duration: 5 years
- Budget: ≈ 180 Mio. €
- Funding: ≈ 90 Mio. €
- Web: http://theseus-programm.de

[Schäfer2007]

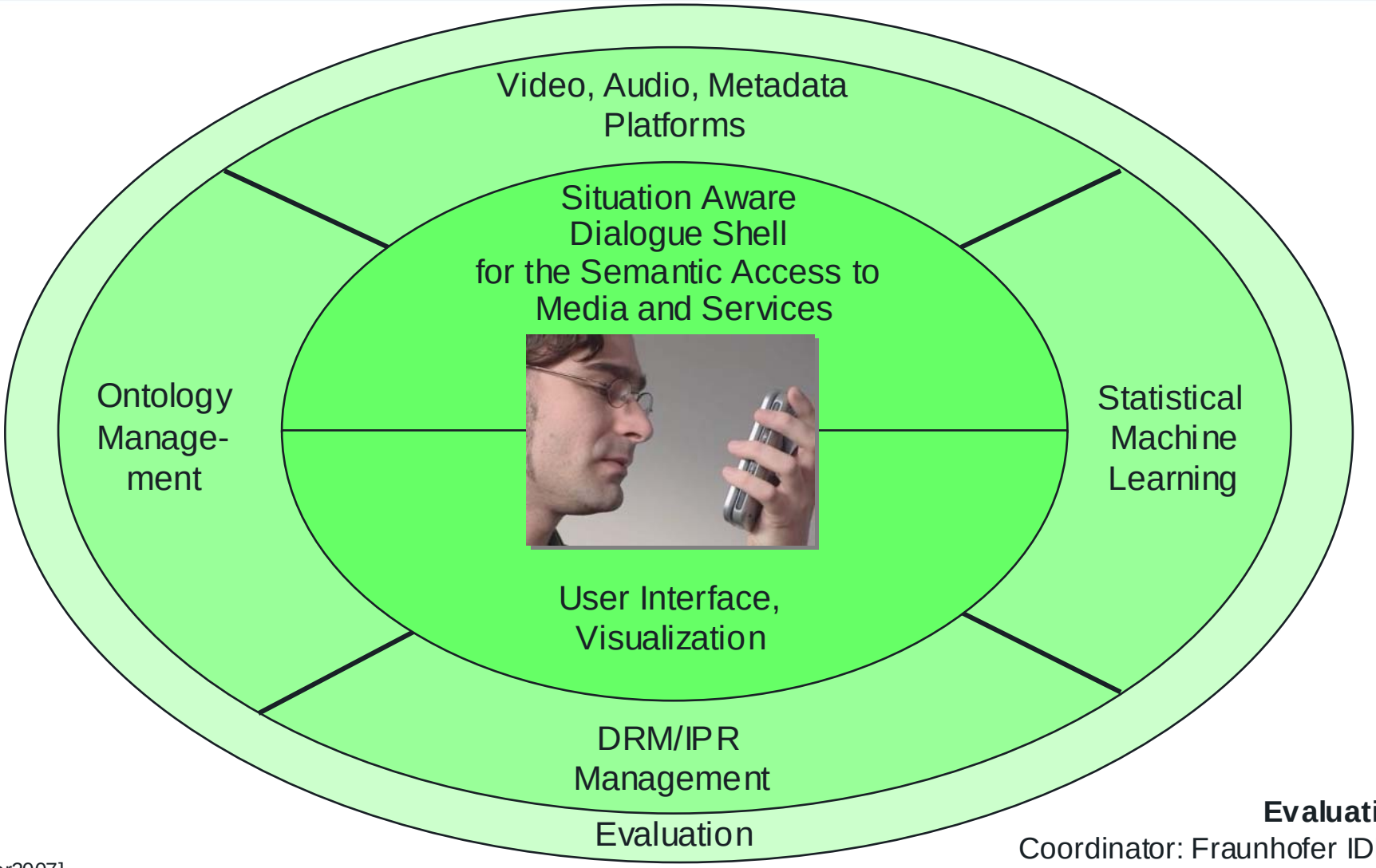
# THESEUS Facts & Partners


Deutsche Nationalbibliothek
Deutsche Thomson OHG (DTO)
Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI GmbH)
**empolis GmbH**
Festo AG
**Fraunhofer-Gesellschaft** (FIRST, HHI, IAIS, IAO, IDMT, IIS, IITB, IGD, ITWM)
Friedrich-Alexander-Universität Erlangen
FZI Forschungszentrum Informatik
Institut für Rundfunktechnik GmbH (IRT)
intelligent views gmbh
Ludwig-Maximilians-Universität (LMU)
LYCOS Europe
m2any GmbH
moresophy GmbH
ontoprise GmbH
**SAP AG**
**Siemens AG**
Technische Universität Darmstadt
Technische Universität Dresden
Technische Universität München
Universität Karlsruhe (TH)
Verband Deutscher Maschinen- und Anlagebau e.V. (VDMA)

[Schäfer2007]

# CTC Overview

THESEUS
Forschungsprogramm für eine neue internetbasierte Wissensinfrastruktur




**Evaluation**
Coordinator: Fraunhofer IDMT

[Schäfer2007]

# CTC WP8 Evaluation

THESEUS
Forschungsprogramm für eine neue internetbasierte Wissensinfrastruktur

**CTC WP8: Evaluation**

Coordinator: Fraunhofer IDMT

- Task 8.1: Databases
- Task 8.2: Text Analysis
- Task 8.3: Media Data Analysis
- Task 8.4: Picture Analysis
- Task 8.5: Audio Quality
- Task 8.6: Picture Quality
- Task 8.7: Iterative System Design and Quality in Use
- Task 8.8: Privacy&Security
- Task 8.9: Field Testing (FhG FIRST)

# 8.1 Databases

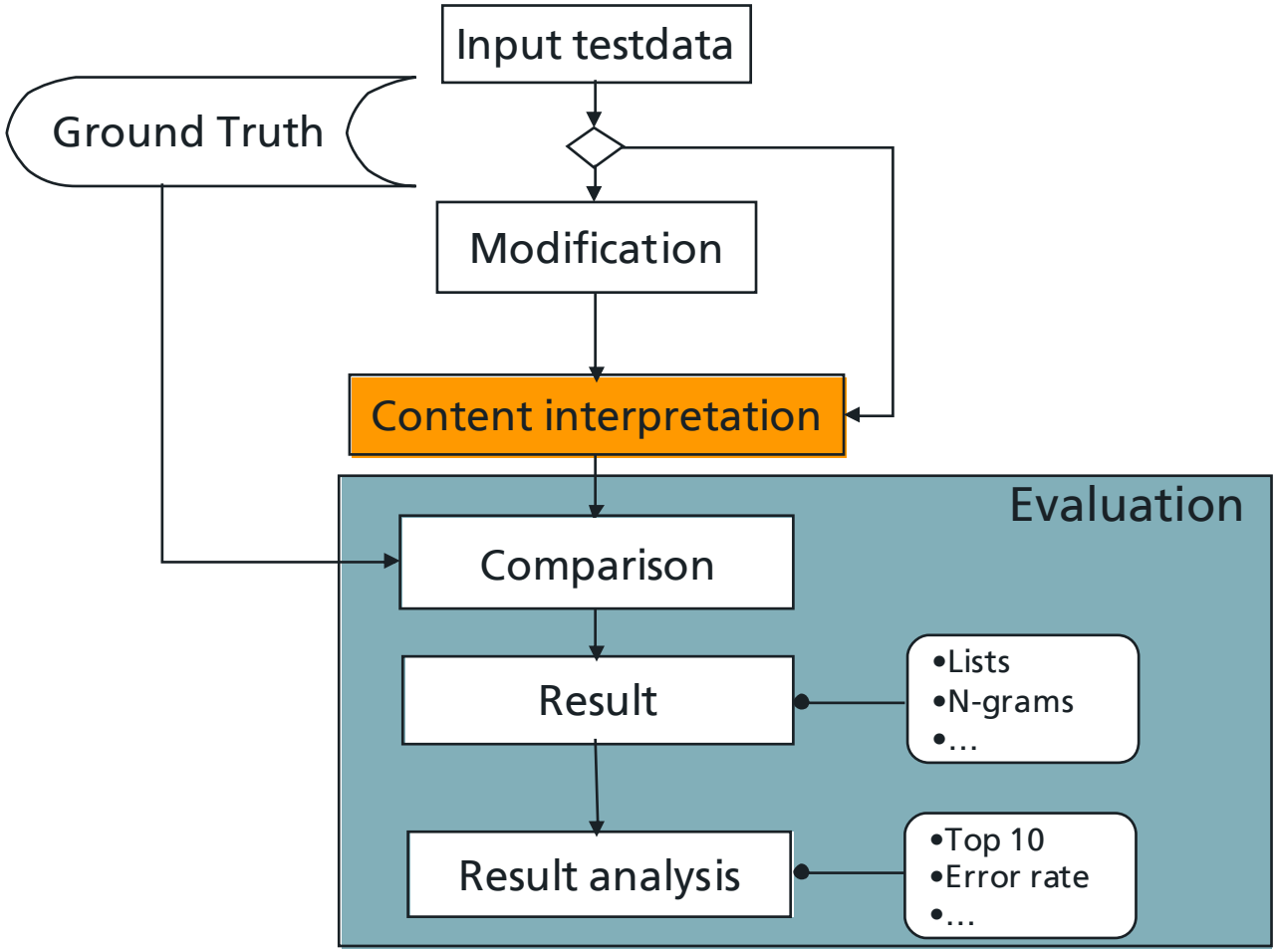
THESEUS
Forschungsprogramm für eine neue internetbasierte Wissensinfrastruktur

- **Starting Point**
  - Reliable evaluation needs
    - Manifold testdata
    - Defined standard distortions
    - Defined test environments
- **Task**
  - Collection of multimedia test data
  - Collection of standard distortions
  - Annotation of test data
  - Documentation of origin and rights of use
  - (Distribution of training database to other CTC Tasks)

# 8.2 Text Analysis



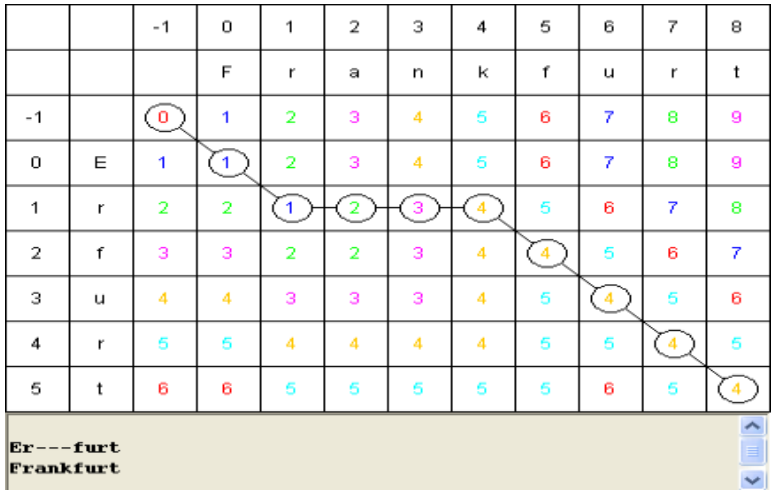
» **Workflow**



# 8.2 Text Analysis



- **Input test data**
  - Generation of test data sets from unused/unlabeled data
  - Distortion/modification of the data
- **Result analysis**
  - Calculation of recognition / error rates of the system
  - Similarity analysis, sequence comparison




[http://www-igm.univ-mlv.fr/~lecroq/seqcomp/node2.html]

$$\text{recognition\_rate} = \frac{\text{recognized\_results}}{\text{all\_truthed\_results}}$$

$$\text{error\_rate} = \frac{\text{incorrect\_recognized\_results}}{\text{all\_truthed\_results}}$$

# 8.3 Media Data Analysis

THESEUS
Forschungsprogramm für eine neue internetbasierte Wissensinfrastruktur

» **Starting Point**

» Watermarking :
  » Process of embedding information into multimedia signals
  » Used for protection of copyrights

» Important :
  » Reproduction-, encoding- and transmission process should not influence the detectability of watermark

» **Evaluation Task**

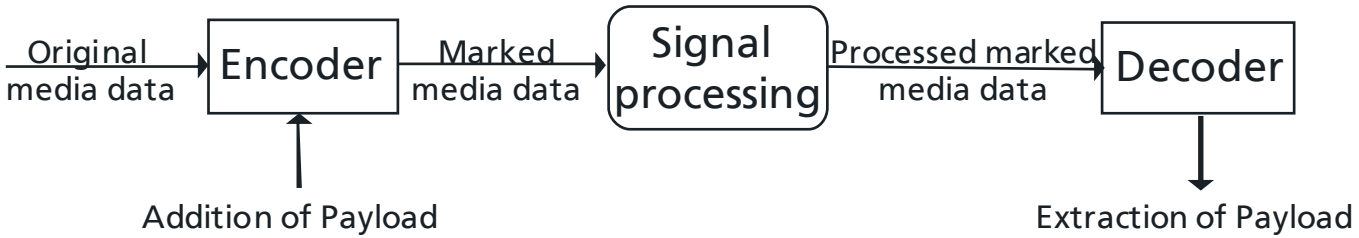
» Evaluation of robustness of watermarking technologies

# 8.3 Media Data Analysis



» **Evaluation Procedure**

- » Embedding watermark/label in data (supporting medium)
- » „Manipulation" of the supporting medium
  - » Reproduction of signal processing like mastering, on air broadcasting, down-mixing, equalization (user-performed enhancements) and data reduction
  - » Signal processing adjusted an adopted to the special needs of the Theseus project
- » Measurement of detect ability of watermark/label after signal processing



# 8.4 Picture Analysis

THESEUS
Forschungsprogramm für eine neue internetbasierte Wissensinfrastruktur

» **Starting Point**

- » Various kinds of CTC algorithms and approaches
  - » Shot/Subshot/Scene Detektion
  - » Video Genre Classification
  - » Image and Video Identification
  - » Video Analysis and Understanding
  - » Video Event Detection
  - » Machine Learning Algorithms for Optimization of various Technologies
  - » Still Image and Spatio-Temporal Segmentation
  - » Image Classification and Fast Indexing
    - » New Image Representations
    - » New Classification Schemes
    - » New Indexing Methods
  - » Face Detection

# 8.4 Picture Analysis



» **Evaluation Procedure**

A generic evaluation framework will be developed to handle and measure various Image and Video Analysis Technologies

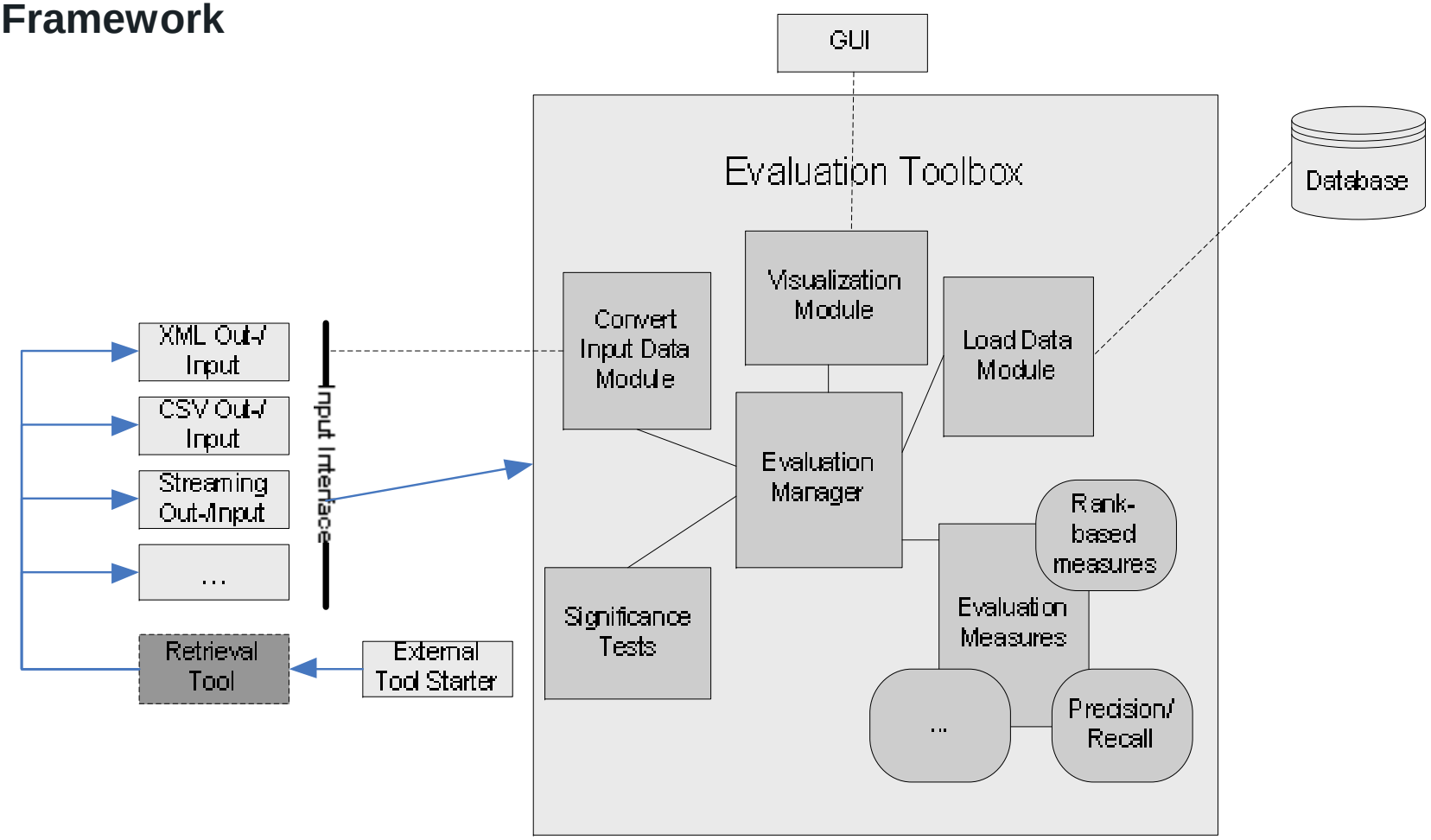
The key features of the framework are:

- easy extension to new formats and measures
- storing previous test results for comparison and measurement of improvements
- sophisticated visualizations for interactive reviewing and generation of descriptive test results.

# 8.4 Picture Analysis

THESEUS
Forschungsprogramm für eine neue internetbasierte Wissensinfrastruktur

» **Framework**



# 8.5 Audio Quality

THESEUS
Forschungsprogramm für eine neue internetbasierte Wissensinfrastruktur

- **Starting Point**
- **Evaluation Task**
  - Watermark should be inaudible
  - Common measurement methods:
    - Detection and analysis possible
    - Influence of the perceptual quality impossible
- **Evaluation Procedure**
  - Evaluation of perceived quality of watermarked content
  - Subjective listening tests according to standardized test procedures
    - ITU-R BS.1116 „Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems“
    - ITU-R BS.1534 „MUlti Stimulus test with Hidden Reference and Anchor “
    - A-B-X Method

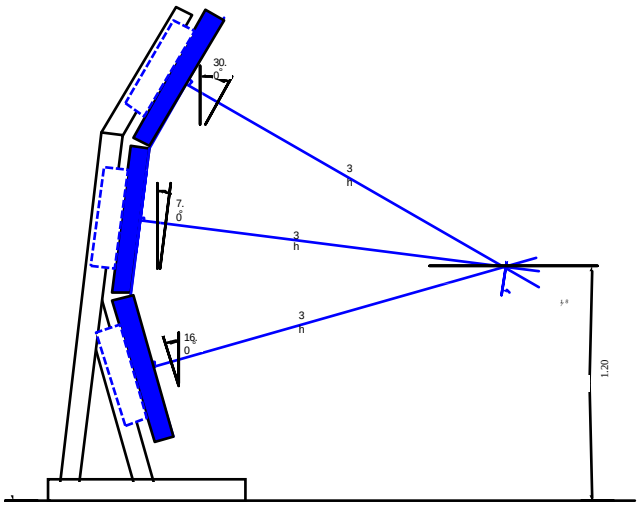
# 8.6 Picture Quality

THESEUS
Forschungsprogramm für eine neue internetbasierte Wissensinfrastruktur

- **Starting Point**
  - Watermark/label should be invisible
  - Picture/Video Compression:
    - High data reduction desired
    - Perceptible quality lost
- **Evaluation Tasks**
  - Measurement of Quality
    - Objective methodes are not precise as subjective
    - Subjective methods time consuming
  - Evaluation of perceived quality of coded, watermarked or labeled content and
  - Comparison of objective measurement with subjective methods

# 8.6 Picture Quality

» **Evaluation Procedure**



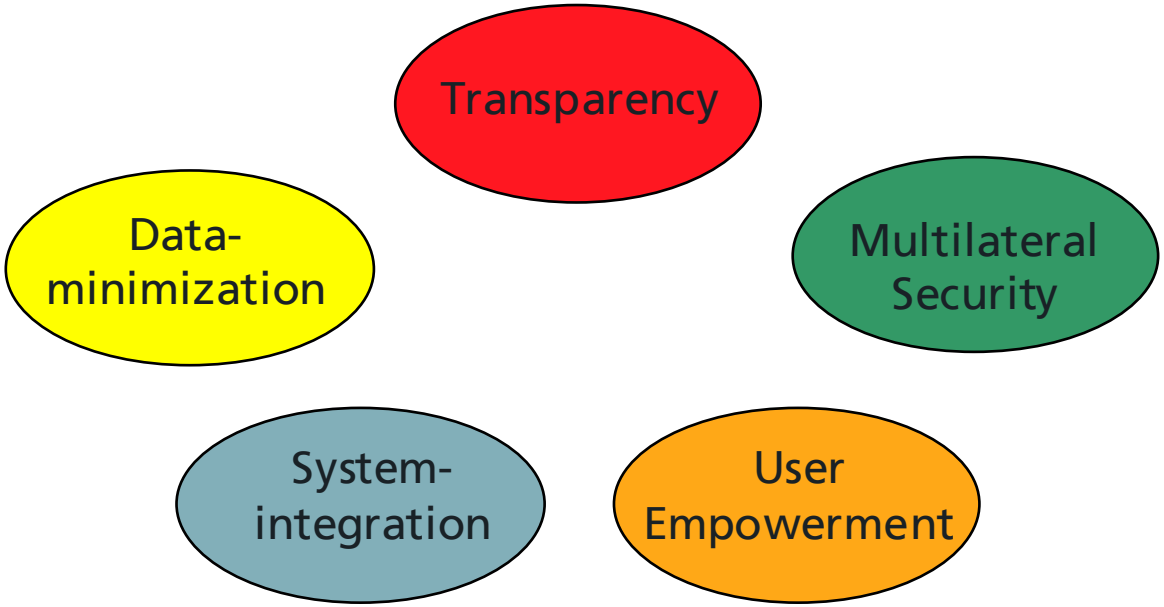

[ITUR08]

» Picture/Video Quality:
  » Subjective visual tests according to standardized test procedures
    » ITU-R BT.500 „Methodology for the subjective assessment of the quality of television pictures“
    » TSCES „Triple Stimulus Continuous Evaluation Scale Method“
    » SAMVIQ „Subjective Assessment Methodology for Video Quality“
» Measurement of Quality:
  » Performance of visual tests according to standardized test procedures
    » Full reference methods
    » No reference methods
  » Performance of measurements with same testdata
  » Comparison of results

# 8.8 – Privacy & Security

THESEUS
Forschungsprogramm für eine neue internetbasierte Wissensinfrastruktur

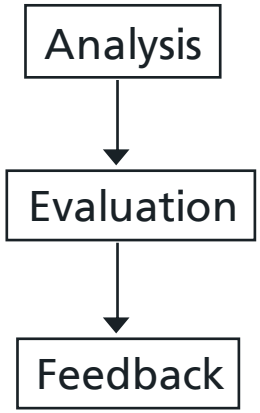
- Privacy & Security Evaluation:
  - Combination of
    - Legal, Technical
    - Economic & Organizational aspects
  - Analysis of Data flow / Data traces
  - Privacy-Criteria:



# 8.8 – Privacy & Security

THESEUS
Forschungsprogramm für eine neue internetbasierte Wissensinfrastruktur

1. Analysis
- Figure out system-modules / create component diagram
- Produce UseCase chart
- Create activity diagram
- Determine all data flows / create sequence diagram
- Optional: create class charts
- Analyse all data traces of the system and communication flows




2. Evaluation
- Evaluate the system / model by defined privacy criteria

3. Feedback Mechanisms
- Feedback will given within defined procedures after finishing the evaluation

Furthermore:
- Select adequate “Privacy Enhancing Technologies”
- Recommend organizational and technical methods for Privacy Enhancement

# THESEUS WP8: Evaluation
# CTC-Task 8.7:
# Iterative system design and quality in use

Juan José Bosch Vicente
bsh@idmt.fraunhofer.de

Dr. Fanny Klett
klt@idmt.fraunhofer.de

# Task 8.7: Iterative system design and quality in use

THESEUS
Forschungsprogramm für eine neue internetbasierte Wissensinfrastruktur

»This task refers to the evaluation of the partners' developments towards:

- » certain ***Evaluation criteria*** depending on the particular task and the ***Context of use***
- » determined ***References***

»by following an ***Evaluation process*** that involves:

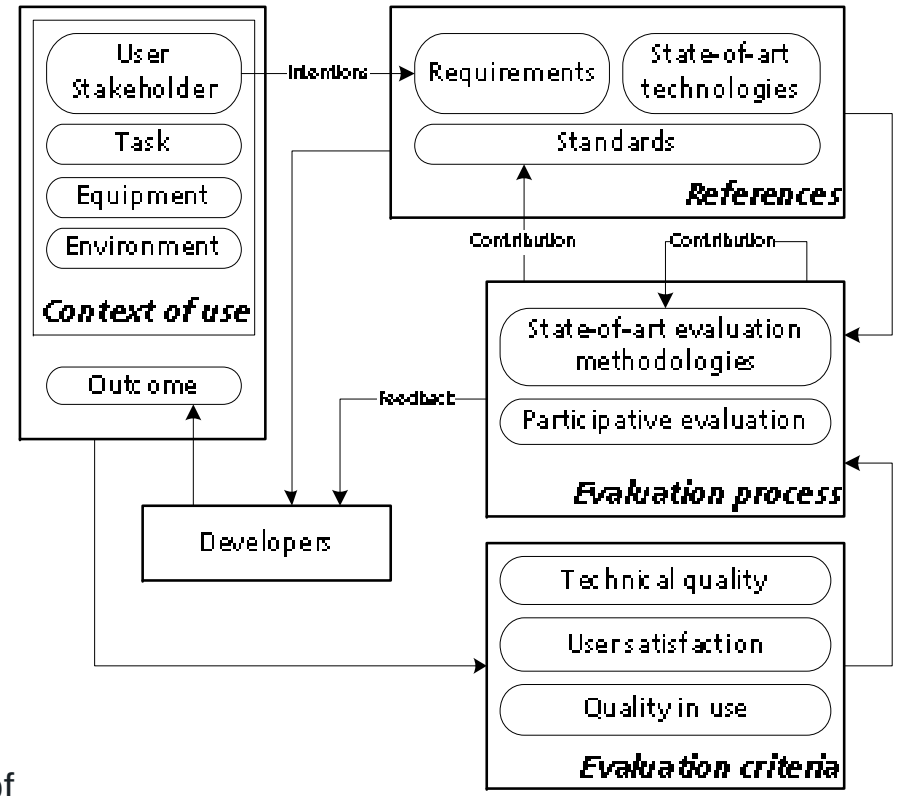
- » State of the art methodologies
- » Participative evaluation

»***References*** include:

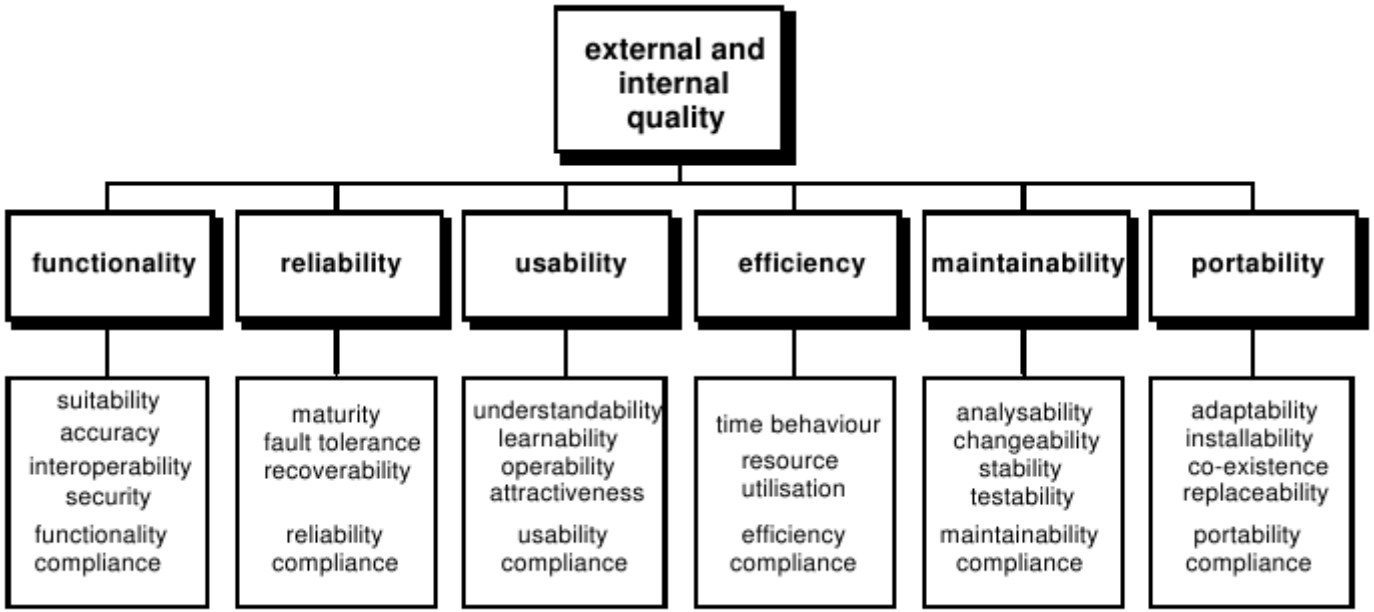
- » Standards and recommendations
- » State-of-the-art technologies
- » Requirements specified by the Use Cases

»***Evaluation process*** will provide:

- » Feedback to developers to be considered in the next iterative stage of the system design
- » Contribution to standards, and state-of-art developments



- **References for the evaluation**
  - The references will include:
    - Standards for: User-Centred Design, Software Quality, Accessibility
    - Initiatives (WAI, OAEI), Campaigns, Conferences (MUC)
  - Use of external or quality in use measures [ISO9126]:



# Task 8.7: Iterative system design and quality in use

THESEUS
Forschungsprogramm für eine neue internetbasierte Wissensinfrastruktur

- **Ontology Management**

- Infrastructure for handling ontologies and semantic meta data
- **Ontology design** and **evolution** evaluation:
  - Technical performance criteria
    - Speed, scalability (ability to store and work with big ontologies)
  - Pragmatic criteria: possibility of collaboration

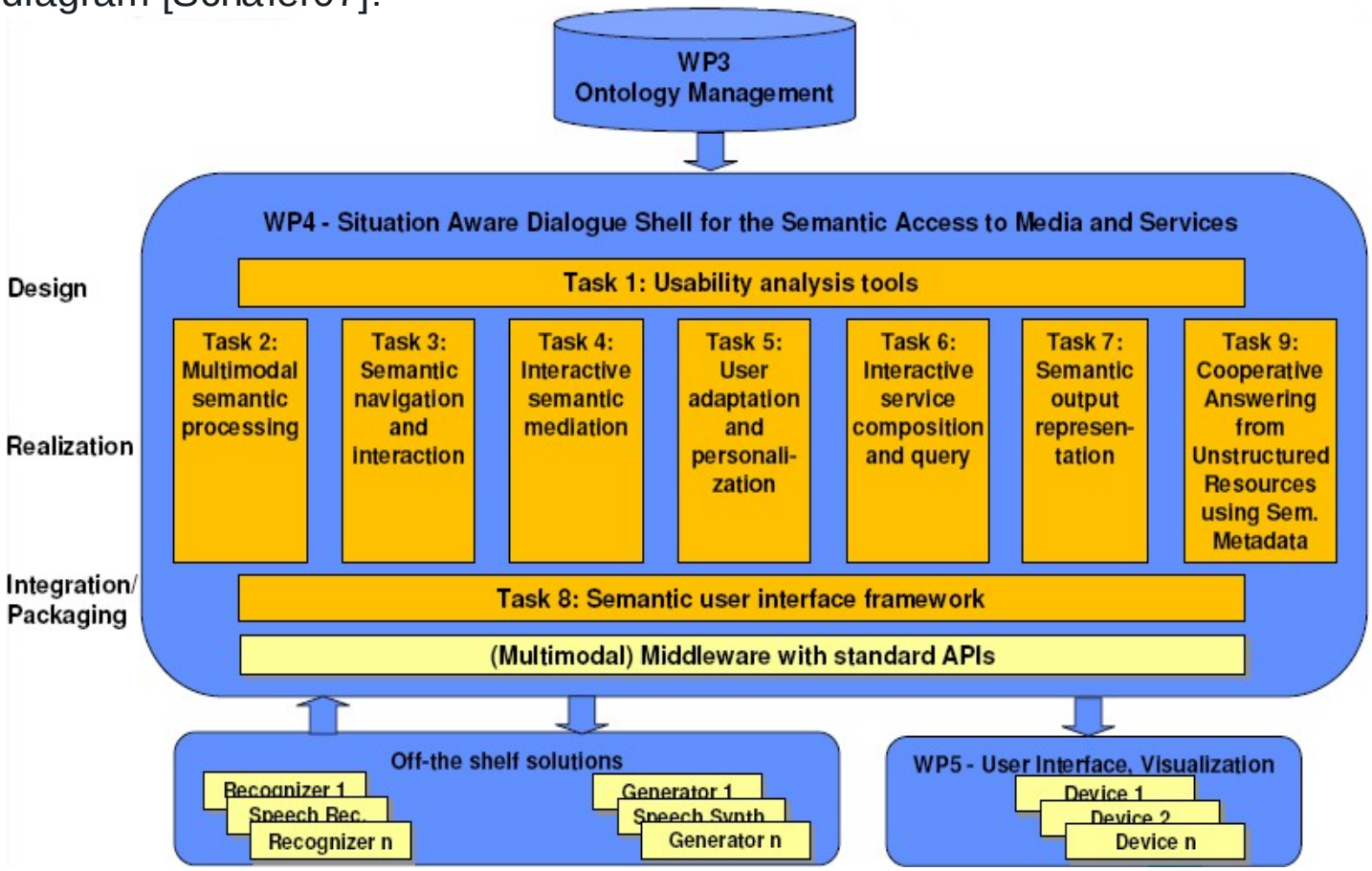
- **Ontology Mapping** evaluation:
  - Use of a “gold standard” to compare to the mapping result
  - Precision, Recall and F-measure
  - Ontology Alignment Evaluation Initiative Campaign (OAEI)

- **Ontology Reasoning** evaluation
  - Reasoning used for validation and deduction
  - Correctness, Performance (execution time, memory consumption, scalability)
  - Use of different data, ontologies, queries, reasoners

# Task 8.7: Iterative system design and quality in use

THESEUS
Forschungsprogramm für eine neue internetbasierte Wissensinfrastruktur

- **Situation Aware Dialogue Shell for the Semantic Access to Media and Services**
- Tasks evaluated unitarily, and also end-to-end evaluation
- Overview diagram [Schäfer07]:



# Task 8.7: Iterative system design and quality in use

THESEUS
Forschungsprogramm für eine neue internetbasierte Wissensinfrastruktur

- **Situation Aware Dialogue Shell for the Semantic Access to Media and Services**

- Goals for the design and evaluation of dialogue systems [ISO9241-110]:
  - Suitability for the task, controllability, error tolerance, etc.

- Multimodal semantic processing & Semantic navigation and interaction evaluation:
  - Response time
  - Mutual disambiguation rate (error handling)
  - Possible interaction types
  - Easy-to-use end device adaptation (mobile, desktop, etc.)

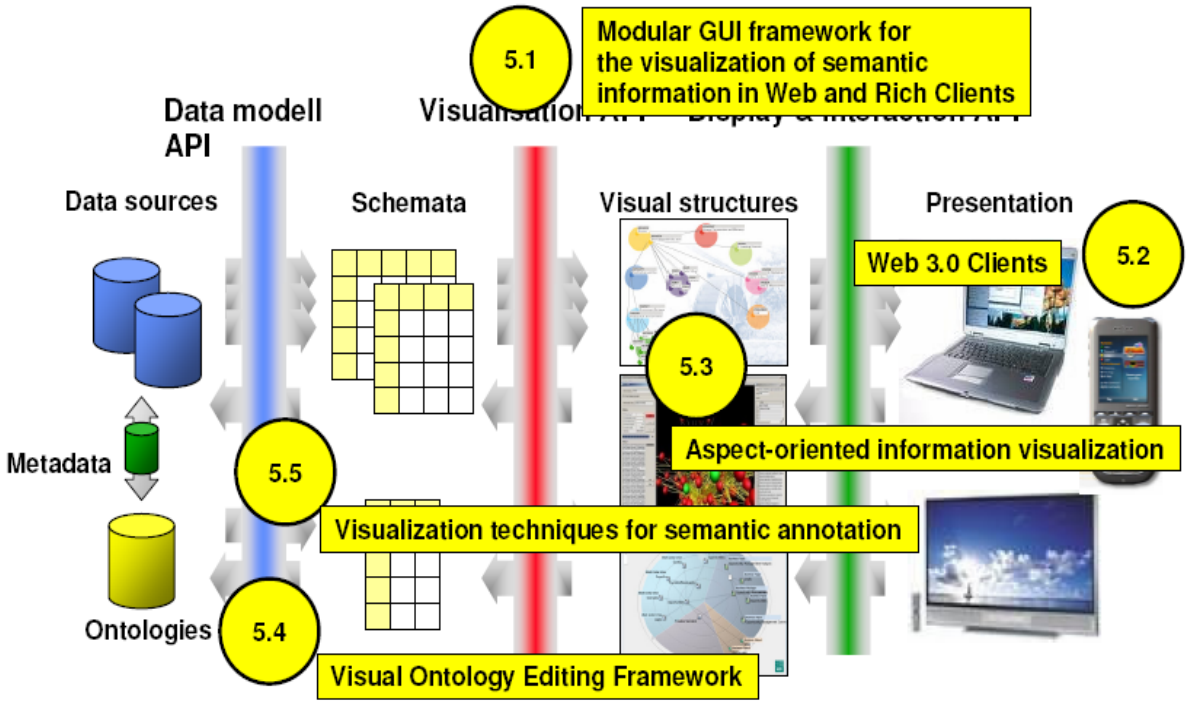
- User adaptation and personalization evaluation:
  - Scalability (number of user profiles)
  - Adequate inference of user preferences
  - Privacy issues (authentication, restrictions)

# Task 8.7: Iterative system design and quality in use

THESEUS
Forschungsprogramm für eine neue internetbasierte Wissensinfrastruktur

» **User Interfaces and Visualization**

» Appropriate and intuitive interface to the user
» Overview diagram [Schäfer07]:



# Task 8.7: Iterative system design and quality in use

THESEUS
Forschungsprogramm für eine neue internetbasierte Wissensinfrastruktur

- **User Interfaces and Visualization**

- **Semantic information visualisation** evaluation based on:
  - Scalability
  - Appropriate use of interaction and navigation techniques [Hearst1999]
  - Usability( efficient navigation, user satisfaction), accessibility
  - Personalisation, collaboration, role views

- **Ontology editing framework** evaluation:
  - Ontology schema and instance editing
  - Versioning system (change management)

- **Visualisation techniques for semantic annotation** evaluation:
  - Support several document formats for annotation: HTML, XML, images, etc.
  - Use of automation (or semi-automation)
  - Support for privileges, trust, access rights

# Task 8.7: Iterative system design and quality in use

THESEUS
Forschungsprogramm für eine neue internetbasierte Wissensinfrastruktur

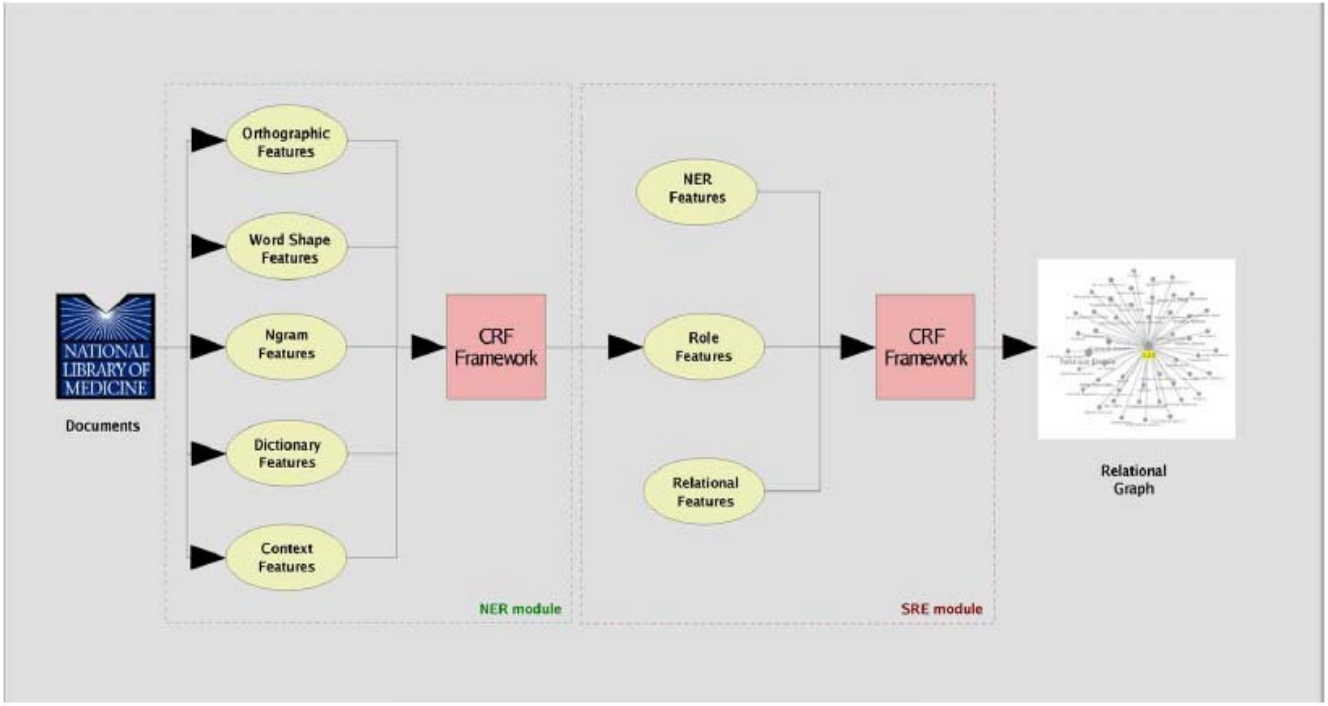
- **Statistical Machine Learning**

- **Learning with Relational Data and Ontologies** evaluation based on:
  - Extraction of ontology from text data
  - Scalability
  - Ontology evaluation [Brewster2004][Navigli2004]:
    - Quantitative: performance of the algorithms (precision, recall or F-measure)
    - Expert evaluation: assess the relations discovered between concepts

- **Learning Semantic Annotation in Textual Data and Web Services**
  - Text format predominant on the web
  - Ontology based semantic annotation approaches [Diallo2006] or [Khelif2004]:
    - ontology instantiation: Detect terms considered as instances of ontology concepts and relations
    - annotation generation: Extract relevant information for describing the content

- **Statistical Machine Learning**

- **Learning Semantic Annotation in Textual Data and Web Services**
  - Use of Named Entity Recognition and Semantic Relation Extraction
  - Possible cascaded Workflow [Bundschus2008]:



# Task 8.7: Iterative system design and quality in use

THESEUS
Forschungsprogramm für eine neue internetbasierte Wissensinfrastruktur

- **Statistical Machine Learning**

- **Learning Semantic Annotation in Textual Data and Web Services** evaluation:
  - Precision, Recall, F-measure, ROC, AUC score [Huang2005]
  - Possible human based evaluation (to validate the annotations)

- Importance of the degree of matching required [Tsai2006]:
  - left match, right match, partial match, approximate match, etc.

- The CBE (Cost-Based Evaluation) model [Sassone1987], stems from the economics field:
  - flexible with the different possible requirements from different users
  - complex definition of weights (can be simplified)
  - used in [Olsson2002], and [Maynard2005]

# Task 8.7: Iterative system design and quality in use

THESEUS
Forschungsprogramm für eine neue internetbasierte Wissensinfrastruktur

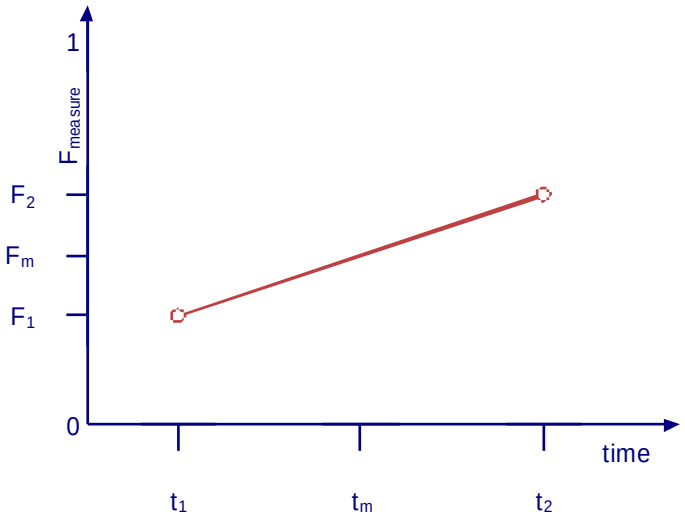
- **Statistical Machine Learning**

- **Large-Scale Self-Learning Textual Archives**

- New machine learning techniques for information extraction from textual documents
- Information visualization components will also be developed

- Evaluation
  - Confusion matrix for each kind of structure to be identified (e.g. address) [DeSitter2004]
  - Correctness depends on the required accuracy [Freitag1998]
    - exact rule, contain rule, overlap rule
  - Partially correct results in the calculation of True Positives, weight of ½
  - Possibility of using MUC (Message Understanding Conference) scoring framework

- **Statistical Machine Learning**
- **Large-Scale Self-Learning Textual Archives**
- The information visualization evaluation:
  - Improvement rates when involving the user
  - Effectiveness: accuracy and completeness with which users achieve specified goals
    → F-measure
  - Efficiency: resources expended in relation to the accuracy and completeness




$$\text{slope} = \frac{\Delta F_{measure}}{\Delta t} = \frac{F_2 - F_1}{t_2 - t_1}$$

where:

$t_1$ is the time needed to process the document without the user involvement,
$t_2$ is the time needed to process the document with the user involvement,
$F_1$ is the F-measure achieved without the user involvement,
$F_2$ is the F-measure achieved with the user involvement

# Conclusions

THESEUS
Forschungsprogramm für eine neue internetbasierte Wissensinfrastruktur

- Why evaluation in THESEUS is important [Schäfer2007]:
  - Goals must be defined precisely
  - Developers can experiment and validate their ideas, and keep only those leading to improvements
  - Continuous evaluation over 5 years to measure the improvements and project success
  - Continuous feedback to developers to improve quality
- Research and development on new evaluation technologies
  - Planned contribution to state-of-the-art evaluation technologies
- Evaluation tests not yet started
  - Work on test specifications and corpora aggregation
  - Dynamic adaption of evaluation plan if needed within 5 years period

# The End

THESEUS
Forschungsprogramm für eine neue internetbasierte Wissensinfrastruktur

Thank you




http://theseus-programm.de/

# References


- [Bradley1997] Bradley, A., The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition Volume 30, Issue 7, Pages 1145-1159, 1997
- [Brewster2004] Brewster, C., Alani, H., Dasmahapatra, S., and Wilks, Y., Data Driven Ontology Evaluation. LREC, 2004.
- [Bundschus2008] Bundschus, M., Dejori, M., Stetter, M., Tresp, V., Kriegel, HP., Extraction of semantic biomedical relations from text using conditional random fields., BMC Bioinformatics 2008, 9:207, 2008
- [Card1983] Card, S. K.; Moran, T. P.; & Newell, A. The psychology of human-computer interaction. Hillsdale, NJ: Erlbaum. 1983
- [DeSitter2004] De Sitter, A., Calders, T. and Daelemans, W., A Formal Framework for Evaluation of Information Extraction, University of Antwerp, Dept. of Mathematics and Computer Science, 2004
- [Diallo2006] Diallo, G., Simonet, M., Simonet, A.: An Approach to Automatic Ontology-Based Annotation of Biomedical Texts. IEA/AIE, 1024-1033, 2006
- [Freitag1998] Freitag, D., Machine learning for information extraction in informal domains. In Phd thesis, Carnegie Mellon University, Pittsburgh PA., 1998[Gomez-Perez2003] Gomez-Perez A., Manzano-Macho D.: A Survey of Ontology Learning Methods and Techniques. Deliverable 1.5, OntoWeb Project, 2003
- [Gomez-Perez2003] Gomez-Perez A., Manzano-Macho D.: A Survey of Ontology Learning Methods and Techniques. Deliverable 1.5, OntoWeb Project, 2003
- [Hearst1999] Hearst, M.A., User Interfaces and Visualization, In: Beaza-Yates, Ricardo, A.; Ribeiro-Neto, Berthier (Eds.): Modern Information Retrieval, Addison Wesley Longman, Reading, MA, 1999
- [Huang2005] Huang,J., Ling, C., Using AUC and Accuracy in Evaluating Learning Algorithms, IEEE Transactions on Knowledge and Data Engineering ,vol. 17, no. 3, pp. 299-310, March, 2005.
- [ISO9126] ISO/IEC 9126, Information technology — Software product evaluation —Quality characteristics and guidelines for their use
- [ISO9241-11] ISO 9241-11, Ergonomic requirements for office work with visual display terminals (VDTs) -Part 11: Guidance on usability
- [ISO9241-110] ISO 9241-110, Ergonomics of human-system interaction — Part 110: Dialogue principles
- [ITUR08] ITU-R: “BASIC PRINCIPLE OF A NEW SUBJECTIVE QUALITY EVALUATION METHOD, EBU II” Document 6G/35-E, 2008
- [ITUR97] ITU-R: "Methods for the Subjective Assessment of small Impairments in Audio Systems including Multichannel Sound Systems", Recommendation ITU-R, BS.1116-1* , 1994-1997

# References


- [Khelif2004] Khelif, K. Dieng-Kuntz, R, Ontology-Based Semantic Annotations for Biochip Domain, Lecture notes in Computer Science, 483-484, 2004
- [Maynard2005] Maynard, D. Benchmarking ontology-based annotation tools for the Semantic Web. UK e-Science Programme All Hands Meeting (AHM2005) Workshop "Text Mining, e-Research and Grid-enabled Language Technology", Nottingham, UK, 2005
- [Navigli2004] Navigli, R., Velardi, P., Cucchiarelli, A., and Neri, F., Quantitative and Qualitative Evaluation of the OntoLearn Ontology Learning System. ECAI Workshop on Ontology Learning and Population, 2004.
- [Olsson2002] F. Olsson, G. Eriksson, K. Franzn, L. Asker, and P. Lidn., Notions of Correctness when Evaluating Protein Name Taggers. In Proceedings of COLING 2002, Taipei, Taiwan, 2002
- [Oviat1999] Oviatt, S. L., “Mutual disambiguation of recognition errors in a multimodal architecture,” in Proc. Conf. Human Factors Computing Systems (CHI’99), 1999, pp. 576–583.
- [Oviat2002] Oviatt, S. L., “Breaking the robustness barrier: Recent progress on the design of robust multimodal systems,” in Advances in Computers, M. Zelkowitz, Ed. New York: Academic, 2002, vol. 56, pp. 305–341[Sampson2005] Sampson, J; Measuring the quality of ontology mappings: A multifaceted approach, 2005
- [Sassone1987] Sassone, P., Cost-benefit methodology for office systems. ACM Transactions on Office Information Systems, 5(3):273–289, 1987
- [Schäfer2007] Schäfer, R., THESEUS Project - Structure and Activities of Core Technology Cluster (CTC), 2007
- [Tsai2006] Tsai, R. T.-H., Wu, S.-H., Chou, W.-C., Lin, Y.-C., He, D., Hsiang, J., et al. Various criteria in the evaluation of biomedical named entity recognition. BMC Bioinformatics, 7(92), 2006