

Vorlesung Netzbasierte Informationssysteme
(WS 2004/05)
Semantic Web Anwendungen

Robert Tolksdorf
Freie Universität Berlin
Institut für Informatik
Netzbasierte Informationssysteme
mailto:tolk@inf.fu-berlin.de
http://www.robert-tolksdorf.de
http://nbi.inf.fu-berlin.de

[1] © Robert Tolksdorf, Berlin

Überblick

[2] © Robert Tolksdorf, Berlin

Überblick

- Beispielschemas/Ontologien
 - RSS
 - CIM
 - GO
- Beispielsysteme
- Bewertung solcher Systeme

[3] © Robert Tolksdorf, Berlin

RSS

[4] © Robert Tolksdorf, Berlin

Beispiel RSS

- RDF Site Summary (RSS) 1.0
 - Ziel: Format zur Metadatenbeschreibung und Herausgabe von Medieninhalten
 - Status: Entwicklung der RDF Site Summary 1.0 Specification Working Group
 - Quelle: <http://purl.org/rss/1.0/spec>
- RSS Dokument („channel“) beschreibt
 - Inhaltseinheiten, die durch URL zugreifbar sind
 - deren Metadaten wie Titel, Beschreibung etc.

[5] © Robert Tolksdorf, Berlin

RSS <channel>: Beschreibung des Channels

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF xmlns="http://purl.org/rss/1.0/"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">

<channel rdf:about="http://www.w3.org/2000/08/w3c-
  synd/home.rss">
<title>The World Wide Web Consortium</title>
<description>Leading the Web to its Full
  Potential...</description>
<link>http://www.w3.org/</link>
<dc:date>2002-10-28T08:07:21Z</dc:date>
<items>
  <rdf:Seq>
    <rdf:li rdf:resource="http://www.w3.org/News/2002#item164"/>
    <rdf:li rdf:resource="http://www.w3.org/News/2002#item168"/>
    <rdf:li rdf:resource="http://www.w3.org/News/2002#item167"/>
  </rdf:Seq>
</items>
</channel >
```

[6] © Robert Tolksdorf, Berlin

RSS <item>: Beschreibung eines Inhalts

```
<item rdf:about="http://www.w3.org/News/2002#item164">
<title>
  User Agent Accessibility Guidelines Become a W3C
  Proposed Recommendation
</title>
<description>
  17 October 2002: W3C is pleased to announce the
  advancement of User Agent Accessibility Guidelines 1.0
  to Proposed Recommendation. Comments are welcome
  through 14 November. Written for developers of user
  agents, the guidelines lower barriers to Web
  accessibility for people with disabilities (visual,
  hearing, physical, cognitive, and neurological). The
  companion Techniques Working Draft is updated. Read
  about the Web Accessibility Initiative. (News archive)
</description>
<link>http://www.w3.org/News/2002#item164</link>
<dc:date>2002-10-17</dc:date>
</item>
```

[7] © Robert Tolksdorf, Berlin

Beispiel CIM/XML

- EPRI (Electric Power Research Institute) Common Information Model (IEC 61970-301) (CIM) Extensions for Electrical Distribution and Asset Modeling
 - Ziel: Rahmen zur Modellierung von Stromverteilung und entsprechenden Geräten
 - Status: IEC Entwicklung
 - Quellen: <http://standards.ces.com/cim/>
<http://standards.ces.com/cim/cim7f/CIM-schema-cimu07f.xml>
- Modell aus der Domäne "Elektrik"

[8] © Robert Tolksdorf, Berlin

Klassen und Eigenschaften

```
<rdfs:Class rdf:ID="PowerSystemResource">
  <rdfs:label xml:lang="en">PowerSystemResource</rdfs:label >
  <rdfs:subClassOf rdf:resource="rdfs:Resource" />
  <rdfs:comment>"A power system component that can be either an individual
  element such as a switch or a set of elements such as a substation.
  PowerSystemResources that are sets could be members of other sets. For
  example a Switch is a member of a Substation and a Substation could be a
  member of a division of a Company"</rdfs:comment>
</rdfs:Class>

<rdfs:Class rdf:ID="Breaker">
  <rdfs:label xml:lang="en">Breaker</rdfs:label >
  <rdfs:subClassOf rdf:resource="#Switch" />
  <rdfs:comment>"A mechanical switching device capable of making, carrying,
  and breaking currents under normal circuit conditions and also making,
  carrying for a specified time, and breaking currents under specified
  abnormal circuit conditions e.g. those of short circuit. The typeName is
  the type of breaker, e.g., oil, air blast, vacuum, SF6."</rdfs:comment>
</rdfs:Class>

<rdf:Property rdf:ID="Breaker.ampRating">
  <rdfs:label xml:lang="en">ampRating</rdfs:label >
  <rdfs:domain rdf:resource="#Breaker" />
  <rdfs:range rdf:resource="#CurrentFlow" />
  <rdfs:comment>"Fault interrupting rating in amperes"</rdfs:comment>
</rdf:Property>
```

[9] © Robert Tolksdorf, Berlin

Eigenschaften

```
<rdf:Property rdf:ID="Breaker.OperatedBy">
  <rdfs:label xml:lang="en">OperatedBy</rdfs:label >
  <rdfs:domain rdf:resource="#Breaker" />
  <rdfs:range rdf:resource="#ProtectionEquipment" />
  <ci ms:inverseRoleName
  rdf:resource="#ProtectionEquipment.Operates"/>
  <ci ms:multiplicity rdf:resource="http://www.cim-
  logic.com/schema/990530#M:0..n"/>
  <rdfs:comment>"Circuit breakers may be operated by
  protection relays."</rdfs:comment>
</rdf:Property>

<rdf:Property rdf:ID="ProtectionEquipment.Operates">
  <rdfs:label xml:lang="en">Operates</rdfs:label >
  <rdfs:domain rdf:resource="#ProtectionEquipment" />
  <rdfs:range rdf:resource="#Breaker" />
  <ci ms:inverseRoleName rdf:resource="#Breaker.OperatedBy" />
  <ci ms:multiplicity rdf:resource="http://www.cim-
  logic.com/schema/990530#M:0..n" />
  <rdfs:comment>"Circuit breakers may be operated by
  protection relays."</rdfs:comment>
</rdf:Property>
```

[10] © Robert Tolksdorf, Berlin

Gene Ontology GO

[11] © Robert Tolksdorf, Berlin

Beispiel GO

- Gene Ontology
 - Ziel: "The goal of the Gene Ontology Consortium is to produce a dynamic controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing"
 - Status: Datensammlung, Public Domain
 - Quelle: <http://www.geneontology.org/>
- Originaldaten:

term: myosin

go id: GO:0016459

definition: A protein complex that functions as a molecular motor; uses the energy of ATP hydrolysis to move actin filaments or to move vesicles or other cargo on fixed actin filaments; has magnesium-ATPase activity and binds actin.

definition_reference: ISBN: 96235764

term: myosin ATPase

go id: GO:0008570

definition: The hydrolysis of ATP by myosin that provides the energy for actomyosin contraction.

definition_reference: NC-IUBMB: Proposed Changes to the Enzyme List concerning ATPases and GTPases

[12] © Robert Tolksdorf, Berlin

GO Oberes Modell

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE go:go>
<go:go xmlns:go="http://www.geneontology.org/xml-dtd/go.dtd#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <go:version timestamp="Wed May 9 23:55:02 2001" />

  <rdf:RDF>
    <go:term rdf:about="http://www.geneontology.org/go#GO:0003673">
      <go:accession>GO:0003673</go:accession>
      <go:name>Gene_Ontology</go:name>
      <go:definition></go:definition>
    </go:term>

    <go:term rdf:about="http://www.geneontology.org/go#GO:0003674">
      <go:accession>GO:0003674</go:accession>
      <go:name>molecul ar_functi on</go:name>
      <go:definition>The action characteristic of a gene
        product. </go:definition>
      <go:part-of rdf:resource="http://www.geneontology.org/go#GO:0003673" />
      <go:dbxref>
        <go:database_symbol>go</go:database_symbol>
        <go:reference>curators</go:reference>
      </go:dbxref>
    </go:term>
```

[13] © Robert Tolksdorf, Berlin

GO Unteres Modell

```
<go:term rdf:about="http://www.geneontology.org/go#GO:0016209">
  <go:accession>GO:0016209</go:accession>
  <go:name>anti oxidant</go:name>
  <go:definition></go:definition>
  <go:isa rdf:resource="http://www.geneontology.org/go#GO:0003674" />
  <go:association>
    <go:evidence evidence_code="ISS">
      <go:dbxref>
        <go:database_symbol>fb</go:database_symbol>
        <go:reference>fbrf0105495</go:reference>
      </go:dbxref>
    </go:evidence>
    <go:gene_product>
      <go:name>CG7217</go:name>
      <go:dbxref>
        <go:database_symbol>fb</go:database_symbol>
        <go:reference>FBgn0038570</go:reference>
      </go:dbxref>
    </go:gene_product>
  </go:association>
  ...
</go:term>
...
```

[14] © Robert Tolksdorf, Berlin

Top Level Ontologien

- Top Level Ontologien
 - Domänenüberschreitend
 - Allgemeine Konzepte
 - Person, Mensch, Tätigkeit, Artefakt
- Domain Ontologien
 - Auf bestimmten Bereich bezogen
 - Dozent, Veranstaltungsbesuch, Schein
- Es gibt hinreichende Ansätze zu Top-Level Ontologien
 - Cyc/OpenCyc (<http://www.opencyc.com>)
 - 6000 Konzepte, 60000 Aussagen
 - WordNet (<http://www.cogsci.princeton.edu/~wn>)
 - 146350 Lexeme
 - IEEE P1600.1 Standard Upper Ontology (SUO) Working Group (<http://suo.ieee.org>)

[15] © Robert Tolksdorf, Berlin

Domänenontologien

- Es gibt hinreichend Domänenontologien, oft aber nur Klassifikationsschemata
 - UMLS (Medizinische Klassifikationen, Semantische Netze, <http://www.nlm.nih.gov/research/umls>)
 - ICD10 (Krankheiten)
 - Produktschemata (eCl@ss, <http://www.eclass.de>)
 - Indexierungsschemata (Bibliotheken)
- DAML Ontology Library (<http://www.daml.org/ontologies>)

[16] © Robert Tolksdorf, Berlin

Example Semantic Web Project: *Organizing Knowledge in a Semantic Web for Pathology*

Robert Tolksdorf, Elena Paslaru

[17] © Robert Tolksdorf, Berlin

Digital Pathology

- Typical diagnostics procedure:
 - generate and analyze tissue sample on glass slide
 - generate medical report
 - store text and image data
- Extended usage of digital images for diagnostics support and educational purposes in everyday pathology




[18] © Robert Tolksdorf, Berlin

Digital Pathology

- Pathology data (Institute for Pathology, Charité Berlin):
 - 15.000 cases annually
 - per medical case up to 5 pathology reports
 - per pathology report up to 10 digital images (15GB)
- Problems:
 - textual and image-based data is stored separately
 - image-based retrieval is restricted to structural image characteristics
 - text-based retrieval is restricted to string matching
 - expert knowledge can not be shared or reused (for diagnostics or teaching purposes) without technical know-how

[19] © Robert Tolksdorf, Berlin

A Semantic Web for Pathology

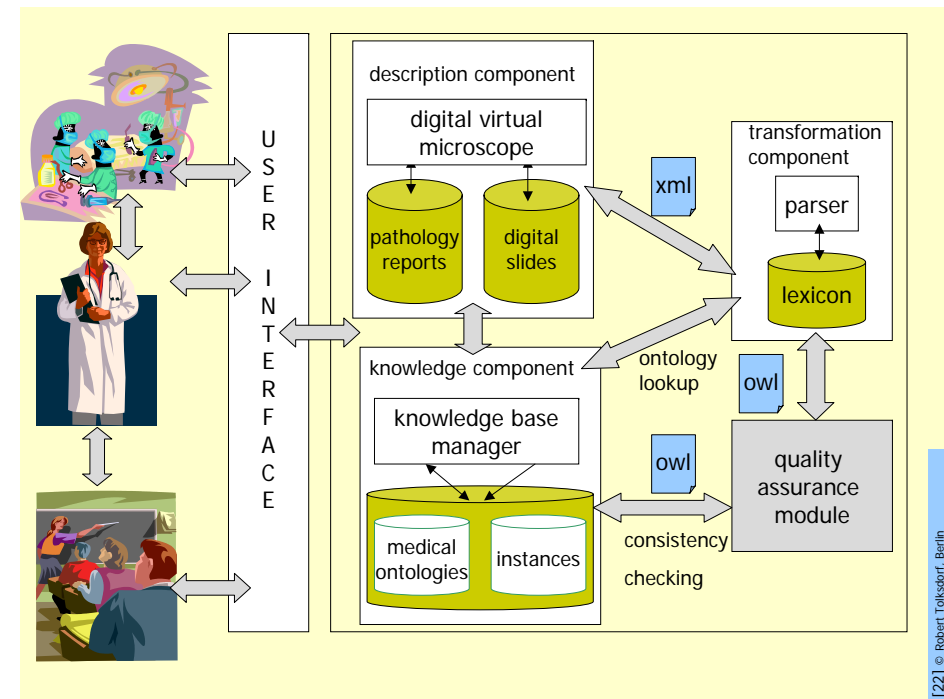
- Semantic Web based Information System for Lung Pathology 
- Objectives
 - re-organization of the available and future expert knowledge for efficient diagnostics and differential diagnostics tasks
 - reuse text and image data for case-based teaching materials
 - minimal invasive usage
 - integration in the available technical infrastructure
 - good precision values (under-diagnosed cases)
- Improve retrieval capabilities of the pathology archive
 - pathology reports are textual representations of the digital images
 - the content of the text and image-based data is represented explicitly
 - use medical ontologies to refine search features
 - use rules to describe diagnostics processes
 - use ontology-based NLP algorithms to extract and represent the content of the pathology reports (semantic annotation)

[20] © Robert Tolksdorf, Berlin

Use Cases

- Use cases
 - Diagnostics and differential diagnostics:
 - similar cases
 - cases with certain patient or morphological criteria
 - cases with similar appearance, but alternative diagnosis
 - Quality assurance
 - Case-based teaching materials
 - Information exchange among health care organizations
- Implementation
 - Web-based architecture (remote consultation, data exchange among health-care organizations, access of teaching materials)
 - usage of established medical data formats (exchange of patient data)
 - usage of standard ontology representation languages (ontology share and reuse)

[21] © Robert Tolksdorf, Berlin



[22] © Robert Tolksdorf, Berlin

Description Component

- formalization of the pathology reports and metadata for digital slides in XML (SVG, XML-HL7)
- management of the original medical data
- report editor, image annotation tool
- new text reports are forwarded for annotation to the transformation component
- integration in the current environment

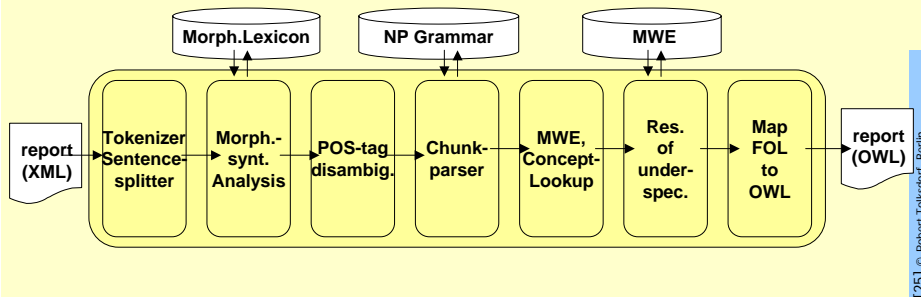
[23] © Robert Tolksdorf, Berlin

The screenshot shows a web browser window with a pathology report editor. The report text includes sections for 'Makroskopie', 'Mikroskopie', and 'Kommentar'. A histological image is displayed with a red box highlighting a specific area. Below the image, there is a code editor showing XML markup for the report content. The XML includes sections for 'Befund', 'Mikroskopie', 'Kritischer_Bericht', and 'Kommentar'. The 'Befund' section contains a paragraph describing the findings: 'Zwei Gewebszylinder von 15 und 4 mm Laenge'. The 'Kritischer_Bericht' section contains a paragraph describing the microscopic findings: 'Stanzbiopsate aus Lungengewebe mit deutlicher Stoerung der alveolaren Textur, soweit noch nachweisbar deutlich Verbreiterung der Alveolarsepten, stellenweise Nachweis von Bronchialepithelregeneraten'. The 'Kommentar' section contains a paragraph describing the findings: 'Stanzbiopsate aus der Lunge mit Zeichen der organisierenden Pneumonie (klin. Mittellappen)'.

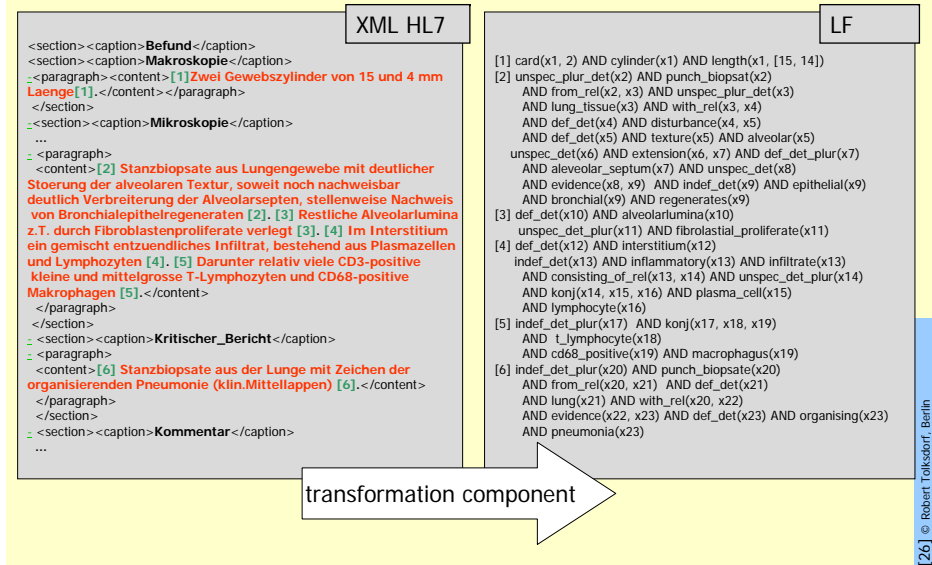
[24] © Robert Tolksdorf, Berlin

Transformation Component

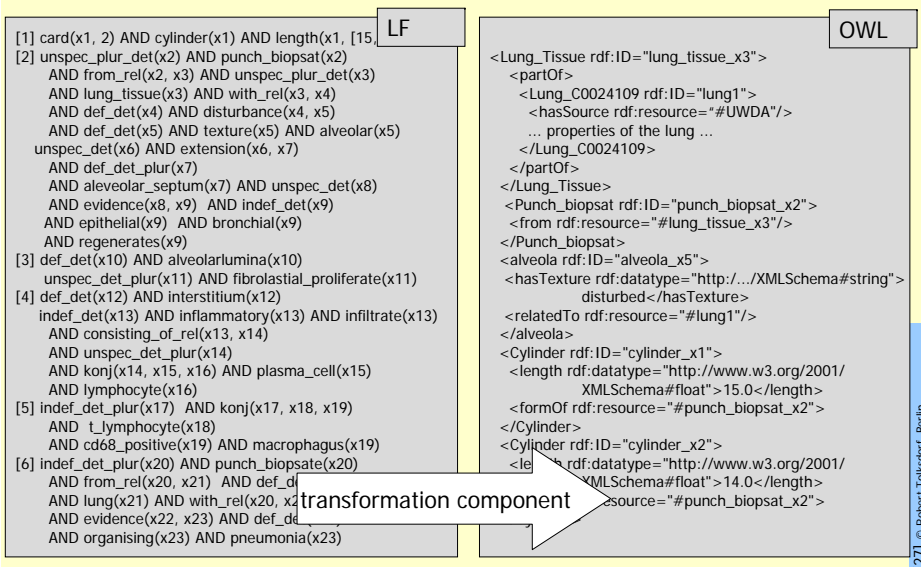
- recognizes concept instances from textual pathology reports and image metadata
- generates semantic representation of pathology reports and forwards it to the knowledge component
- suggests ontology extensions for frequent terms



Transformation Component



Transformation Component



Knowledge Component

- Knowledge base
 - medical ontologies
 - ontology of lung anatomy (UMLS)
 - ontology of lung diseases (UMLS)
 - model of pathology reports
 - immunohistology ontology
- generic ontologies
 - semantic network (UMLS)
- rules (to be done)
 - Tumor Node Metastasis (tumor classification system)
- instances of the ontology concepts from real pathology reports and digital slides
- reasoner

UMLS

- UMLS (Version 2003AC)
 - Unified Medical Language System (National Library of Medicine)
 - 100 medical libraries (1,5 billion concepts)
 - integrates libraries into a common data format (UMLS Semantic Network, UMLS Metathesaurus)
 - UMLS Semantic Network: upper level + medicine core concepts
 - UMLS Metathesaurus:
 - library-specific concepts
 - terms are grouped to single concept id
 - translation of terms

[29] © Robert Tolksdorf, Berlin

Ontology Generation and Evaluation

- Top-down approach:
 - identify relevant UMLS libraries → 50% (700.000 concepts)
 - Map relevant libraries to archive vocabulary → ranking of 10 most application relevant UMLS libraries → 250.000 concepts
- Bottom-up approach:
 - start with 5 application relevant keywords
 - consider neighbored concepts in Metathesaurus → 1000 concepts
- Ontology evaluation
 - Check inconsistencies (reasoner) → 5%
 - Add German translations → 5%
 - Compare archive vocabulary to the ontology vocabulary:
 - add pathology-specific knowledge
 - add generic knowledge (spatial relationships, part-whole ontology)

[30] © Robert Tolksdorf, Berlin

UMLS Issues

- Not intended for automatic integration in Semantic Web applications:
 - no precise semantic definition of relationships (part-of, narrower, broader, related_to, associated_with)
 - error-prone modeling decisions:
 - no consistent upper-level ontology
 - cyclic concept definitions
 - erroneous usage of part-of and is-a relationships (right lobe of lung is-a lung)
 - meaning of concepts is encoded in concept names ("ARF-smaller-then-2", "Unspecified injury of lung with open wound into thorax")
- <http://nbi.inf.fu-berlin.de/research/swpatho/>

[31] © Robert Tolksdorf, Berlin

Reisewissen

- „Hotelbewertungseingine“ für Geschäftsreisen
- „Passendes“ Hotel finden
- Ziele:
 - Optimierung der Hotelauswahl
 - Optimierung der gesamten Geschäftsreise
 - Höhere Qualität der Reisedienstleistung
 - Zeitersparnis
 - Senkung indirekter Reisekosten
 - Signifikante Senkung der direkten Reisekosten
- <http://nbi.inf.fu-berlin.de/research/reisewissen/>

[32] © Robert Tolksdorf, Berlin

Use cases studied in KnowledgeWeb NoE

- Deliverable D1.1.2: Prototypical business use cases
 - Recruitment
 - Multimedia content analysis and annotation
 - Peer-to-peer eScience Portal
 - News aggregation service
 - Product lifecycle management
 - Data warehousing in healthcare
 - B2C marketplace for tourism
 - Digital photo album management
 - Geosciences Project Memory
 - R&D Support for Coffee
 - Co-ordination of Real Estate Management
- <http://knowledgeweb.semanticweb.org/>

[33] © Robert Tolksdorf, Berlin

Wissensnetze

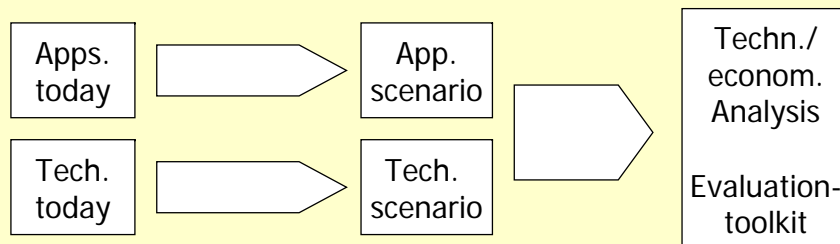
Robert Tolksdorf, Malgorzata Mochol et al.

[34] © Robert Tolksdorf, Berlin

KnowledgeNets: Approach

Business Viewpoint

Predict impacts on markets and value chains
Estimate realization chances based on participant's interests and business models



Technological Viewpoint

Derive requirements for the technological infrastructure
Evaluate current technological developments

[35] © Robert Tolksdorf, Berlin

Case 1: B2C eCommerce with Semantic Web

- Analysis of today's applications:
 - Most data relevant for purchasing decisions is available online: Product descriptions, offers, product and vendor ratings
 - Challenges for a customer
 - Find all relevant information sources for a specific product
 - Find similar products
 - Integrate the information available to compare products
 - Problem: Missing semantic clearness of the information published
- Analysis of today's technologies:
 - Uniform Resource Identifiers (URIs) as global identification mechanism for products and market participants
 - Web Ontology Language (OWL) for the definition of common terms and concepts needed to describe products and market participants
 - Resource Description Framework (RDF) as data model with its XML-based serialization syntax for the direct publication of data on the Web
- Scenario for tomorrow's technologies:
 - Clear technological path towards Semantic Web enhanced Business to Consumer e-Commerce

[36] © Robert Tolksdorf, Berlin

Case 1: Tomorrows applications

- Scenario for tomorrows applications:
 - Information providers publish their data according to a set of domain ontologies using RDF
 - All portals operate on the same information basis
- Analysis of tomorrows applications:
 - All portals allow product searches and comparisons based on all market data published
 - Increased market transparency
 - Querying instead of browsing
 - Query per article or product category and not per shop
 - Domain knowledge supported queries
 - Query extension based on concept relations
 - Dissolving vague searches with domain knowledge
 - Result rendering according to user preferences
 - Level of detail, device, language

[37] © Robert Tolksdorf, Berlin

Case 1: Economic Implications

- The usage of Semantic Web technologies would lead to an increased market transparency.
- Central Questions: Who would benefit? Is there a win-win situation for all participants?
 - ✓ Customers: Benefit from the increased market transparency.
 - ✗ Manufacturers: Face higher pressure of competition
 - ✗ Merchants: Face higher pressure of competition
- Realization chances depend on the characteristics of the market.
 - Advantage of being found might be more important
 - Call for tenders, labor market

[38] © Robert Tolksdorf, Berlin

Case 2: Employment market

- Analysis of todays applications:
 - Current prognoses over half of future employment procurement will occur online
 - Online personnel marketing is increasingly used with cost cutting results and efficacy
 - many websites and online portals financed by publishing fees (www.jobpilot.de, www.monster.de)
 - different business websites
 - portal set up by the state job centre (www.arbeitsagentur.de, www.ams.se)
- Analysis of todays technologies:
 - Overview on all portals nearly impossible by manual visits
 - Many different existing taxonomies for the classification of job posting
 - Job postings as free text using uncontrolled vocabularies
 - The meta-search engines available conduct searches on a full text basis and as a result are limited in their ability to provide offers that match the precise needs of their clients

[39] © Robert Tolksdorf, Berlin

Case 2: Employment market – Integration

- Scenario for tomorrows technologies:
 - Integrated job database – first efforts
 - Bundesanstalt für Arbeit (Germany)
 - National Labour Market Board (Sweden)
 - solutions based on an HR-XML (>75 „XML-Schemas“)
 - HR-BA-XML – www.arbeitsagentur.de
 - HR-XML-SE – www.ams.se
 - Information could be represented as RDF using controlled vocabularies in the form of taxonomies:
 - Free-text
 - required skills and qualifications
 - contract details
 - parts of the description of the organisation
 - Applications
 - The RDF descriptions of open positions could be matched with the profiles of applicants using background knowledge represented in rich ontologies

[40] © Robert Tolksdorf, Berlin

Case 2: Employment market – Target state (I)

- Scenario for tomorrows applications:
 - They aim to create greater visibility to the job market (applicant and employment offers) by cross-linking employment offers from private job exchanges, state employment agencies and medium-sized businesses to provide a highly efficient balance of supply and demand
 - The classical closed 1:n communication links between an organisation and a fixed set of job portals would change into open n:m communication
 - Some international and national classification standards and specification will be used:
 - Standard Occupational Classification (SOC) System
 - Berufskennziffern System (BKZ)
 - Klassifikation der Wirtschaftszweige (WZ2003)
 - Identification of organisation (D&B DUNS Number)
 - HR-XML (also the extension – HR-BA-XML)

[41] © Robert Tolksdorf, Berlin

Case 2: Implications

- ✓ The use of concepts for describing jobs
 - makes the descriptions language independent
 - simplifies the global exchange of postings
- ✓ Transparency rises: all job postings accessible trough all portals
- ✓ Employers would minimize their publishing costs
- ✓ The use of controlled vocabularies allows better automatic matchmaking between requirements and applications
- ✓ Better machine processability of semantically described postings
- ✗ Job portals would have to change their business models (fees)
- ✗ Additional effort for writing semantic descriptions
- ✗ Applicants have to add RDF descriptions to their application

[42] © Robert Tolksdorf, Berlin

Zusammenfassung

[43] © Robert Tolksdorf, Berlin

Zusammenfassung

- Beispielschemas/Ontologien
 - RSS
 - CIM
 - GO
- Beispielsysteme
- Bewertung solcher Systeme

[44] © Robert Tolksdorf, Berlin