Methodologies for Ontology-Based Semantic Translation

H. Stuckenschmidt, H. Wache, U. Visser, G. Schuster

The BUSTER Project:
TZI, Intelligent Systems Group

Outline

- A Survey of Existing Approaches
  - The Role on Ontologies
  - Used technologies
  - Conclusion
- Methodologies in the BUSTER Project
  - Integration of Data Structures
  - Integration of Catalogues

Motivation

- Interoperability problem
  - Structural and semantical heterogeneity
  - *Meaning* of the information
- Causes for semantic heterogeneity (Goh, 1997)
  - Confounding conflicts (same meaning, different context, e.g. “latest trade price”)
  - Scaling conflicts (different reference systems, e.g. currencies)
  - Naming conflicts (homonyms, synonyms)
- Using ontologies to overcome the problem
- Ontologies as key application (Uschold & Grüniger 1996)

Motivation (cont.)

- Survey of existing solutions
  - 25 approaches
- Focus:
  - Role and use of ontologies
  - Integration of information sources (not knowledge bases)

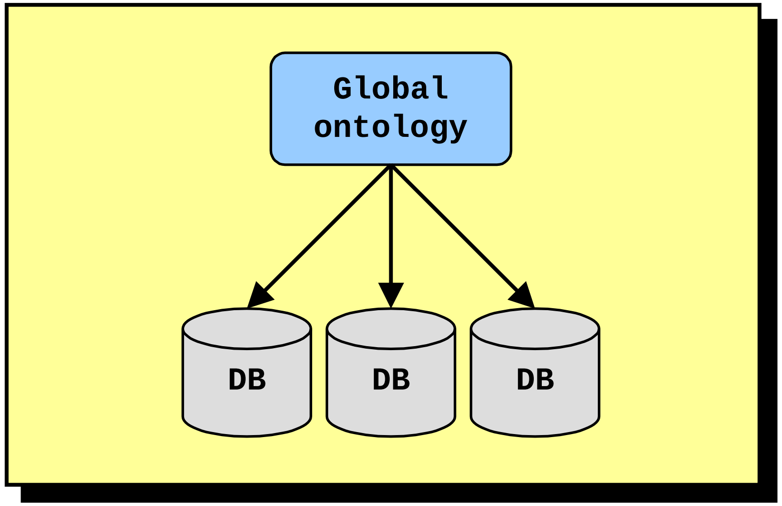
SIMS, TSIMMIS, OBSERVER, CARNOT, KRAFT, Infosleuth, PICSEL, DWQ, Ontobroker, SHOE, MECOTA, BUSTER,…

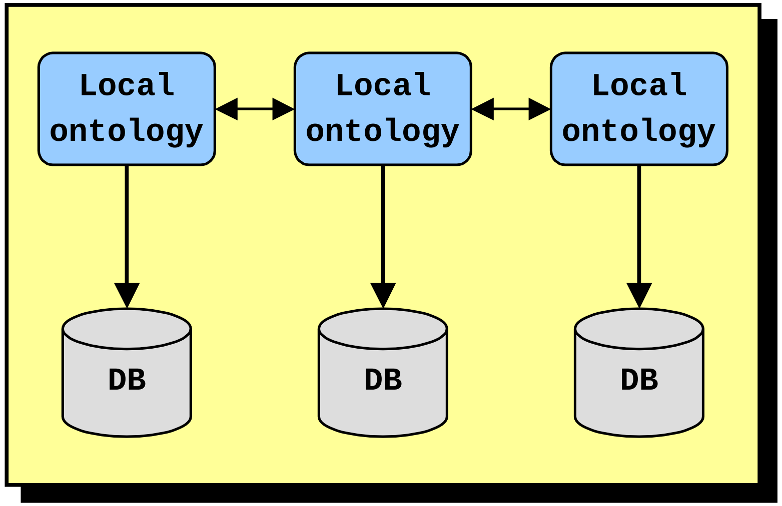
Evaluation criteria

- Use of ontologies
  - Role and architecture of ontologies influence the representation
- Ontology representation
  - Different representation capabilities
- Use of mappings
  - Ontologies linked to information sources
  - Several ontologies cause mappings between them
- Ontology engineering
  - Acquisition support and reuse

Role of ontologies

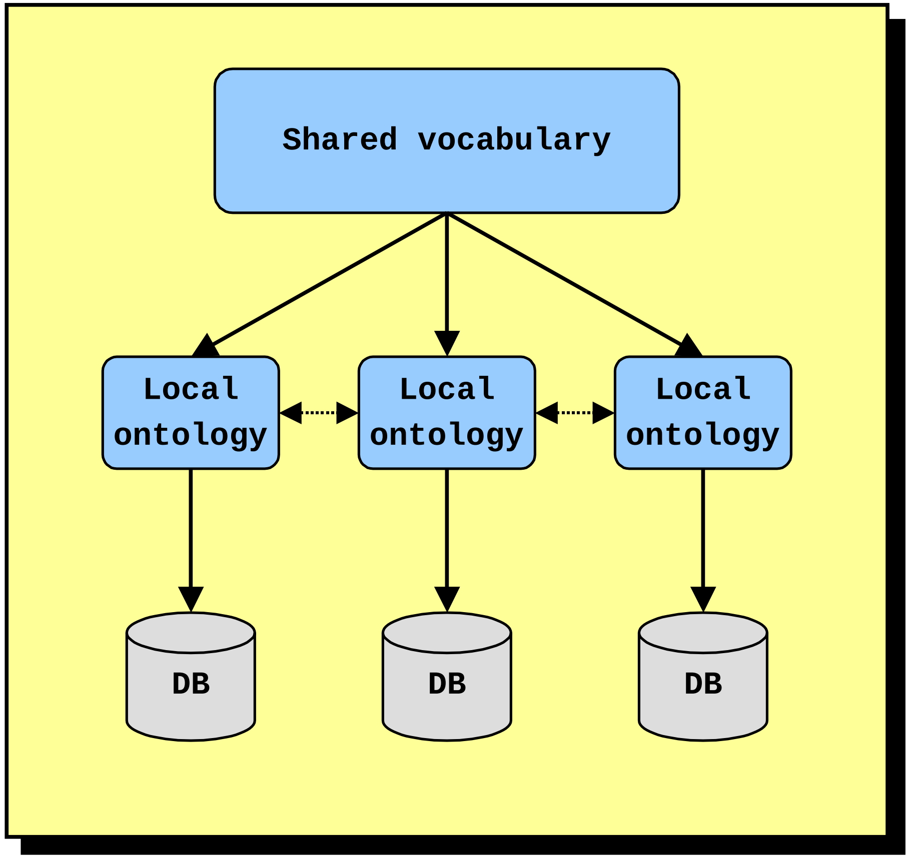
- Content explication
  - Single ontology approaches
    - Global ontology, shared vocabulary (e.g. SIMS)
    - Can be combination of several ontologies because of modularization
    - Same view on domain nessecary, susceptible when information source changes, minimal ontology commitment hard to find
  - Multiple ontology approaches
    - Information source has own ontology (e.g. OBSERVER)
    - No shared vocabulary
    - No common and minimal ontology commitment needed (about global ontology)
    - Problems with different source ontologies (inter-ontology-mapping needed)
    - Hard to define inter-ontology mappings in reality





Role of ontologies (cont.)

- Content explication
  - Hybrid approaches
    - Information source has own ontology
    - Built upon one global shared vocabulary
    - Description of local ontologies is interesting
      - COIN: context is attribute-value vector
      - MECOTA: Information source is annotated by label for the semantics, label combines primitive terms
      - BUSTER: Shared vocabulary as „general ontology" (e.g. value ranges), source ontology is refinement (values are restricted)
    - Advantages
      - New information sources easily added
      - „Comparable" ontologies due to shared vocabulary
    - Disavantage
      - Reuse of existing ontologies difficult



Role of ontologies (cont.)

	Single ontologies	*Multiple ontologies*	*Hybrid ontologies*
Implementation effort	straight-forward	costly	reasonable
Semantic heterogeneity	similar view of domain	supports heterogeneous views	supports heterogeneous views
Adding/removing sources	need for adaption in the global ontology	new source ontology; relation to other ontologies	new source ontology
Comparsion of ontologies	-	difficult due to lack of shared vocabulary	simple due to shared vocabulary

Role of ontologies (cont.)

- Additional roles
  - Query model (e.g. SIMS)
    - User formulates in terms of ontology
    - System reformulates in sub-queries of each source
    - Ontology „acts“ as global query scheme
    - User has to know structure and contents of ontology
  - Verification
    - Mapping from global schema to local source schema during integration
      - Sub-query correct w.r.t. a global query if local sub-query provides a part of the queried answers → sub-query must be contained in global query
    - DWQ
      - Sub-queries are correct if their ontology concepts are subsumed by the global query concepts
    - PICSEL
      - Also generates mapping hypotheses which are validated w.r.t the global ontology

Ontology representation

- Focus on languages and structures
  - No contents discussion
  - Restriction to object-centered knowledge representations

- Description logic variants dominant
  - Pure description logic languages
    - CLASSIC (e.g. OBSERVER, SIMS, Kayshap & Sheth)
    - GRAIL (e.g. Tambis)
    - OIL (e.g. BUSTER)
  - Extensions of description logic (incl. rule bases)
    - CARIN (e.g. PICSEL) → DL with function-free horn rules
    - *AL-log* (e.g. DWQ) → DL and datalog combination
    - *DLR* (e.g. Calvanese et al., 2001) → DL with *n*-ary relations

Ontology representation (cont.)

- Frame-based representations
  - Systems
    - COIN, KRAFT, Infosleuth, Infomaster, Ontobroker
  - Languages
    - Ontolingua, OKBC, F-Logic

Mapping

- Integration task puts ontologies into context
  - Relation ontology and their environment important
  - Two mappings are important
    - Mapping between ontology and the information they describe
    - Mapping between ontologies

- Connection to information sources
  - Structural resemblance (1-1 copy of DB-structure) (e.g. SIMS, TSIMMIS)
  - Definition of terms (only link to source) (e.g. BUSTER)
  - Structural enrichment (e.g. OBSERVER, KRAFT, PICSEL, DWQ)
    - Common approach, combines the first two approaches
    - Logical model that refers to the DB scheme, additional definitions
  - Meta-annotation
    - New approach w.r.t to the semantic web
    - Annotation resembling parts of the real information (e.g. SHOE)
    - Annotation to avoid redundancy (e.g. Ontobroker)

Mapping (cont.)

- Inter-ontology mapping
  - Defined mapping
    - E.g. KRAFT: Translation between ontologies by mediator agents
      - 1-1 mappings between classes and values
      - Flexible but fails to ensure semantic preservation
  - Lexical relations
    - Quantified inter-ontology relationships from linguistics (e.g. OBSERVER)
      - Synonym, hyponym, overlap, covering, disjoint
      - No formal semantics → subsumption is rather heuristic
  - Top-level grounding (e.g. DWQ)
    - Relate all ontologies to a top-level ontology
    - Stay inside a formal representation language
  - Semantic correspondences (e.g. MECOTA, BUSTER)
    - Find semantic correspondences, use shared vocabulary
    - FCA-approaches

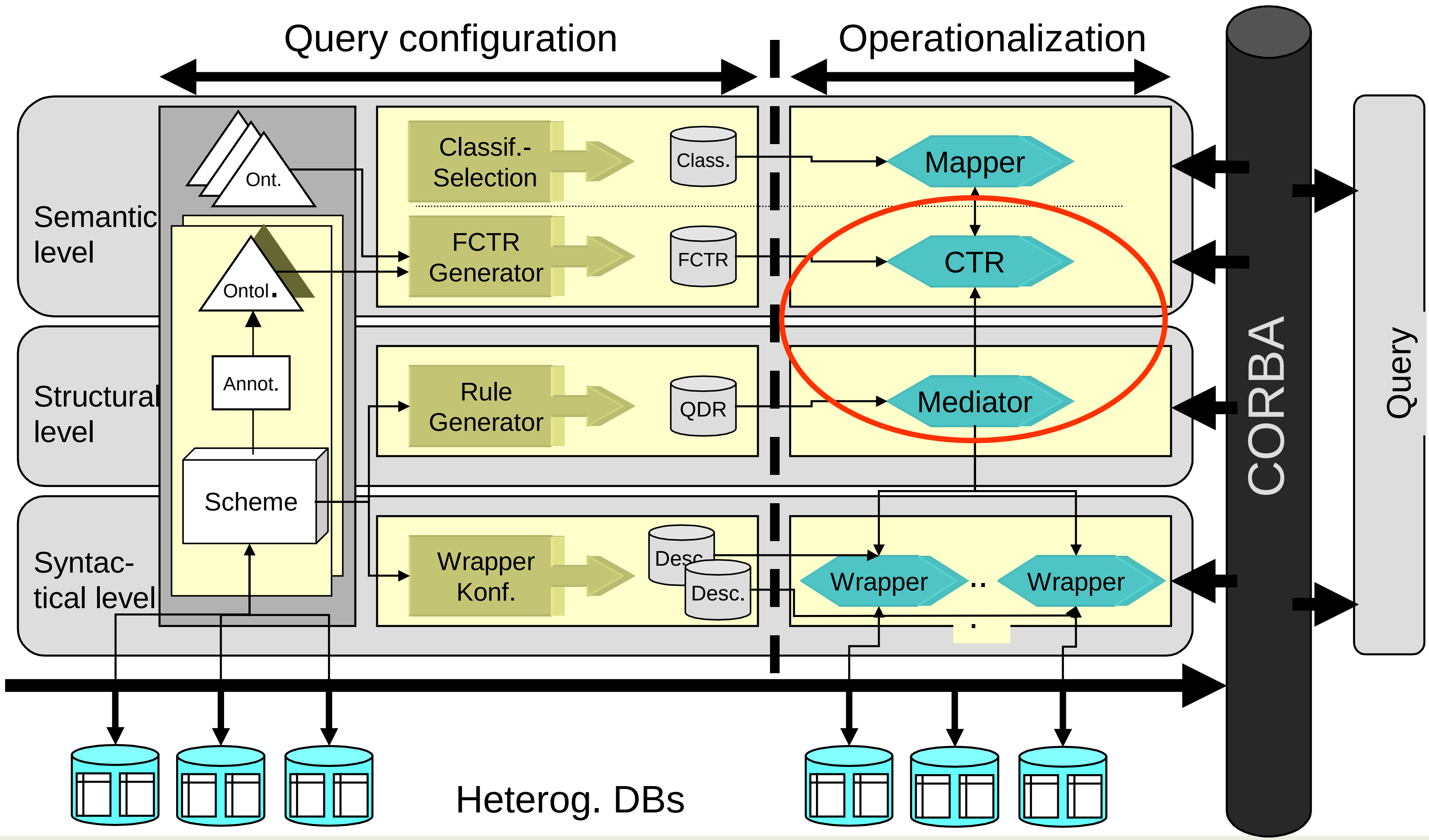
Conclusions

- State-of-the-art
  - „Typical“ information integration system
    - Use established technologies
    - Ontologies for the explication of the contents of an information source (mainly by describing the meaning of table and datafield names)
    - Each information source has ontology (resembles and extends structure of DB)
    - Integration with either common ontology or fixed mappings between ontologies
    - Ontology language based on DL
    - Subsumption reasoning for computation relations between information sources and (sometimes) for validation of the integration result
    - Specialized tools (mainly editors) support the process of building an ontology

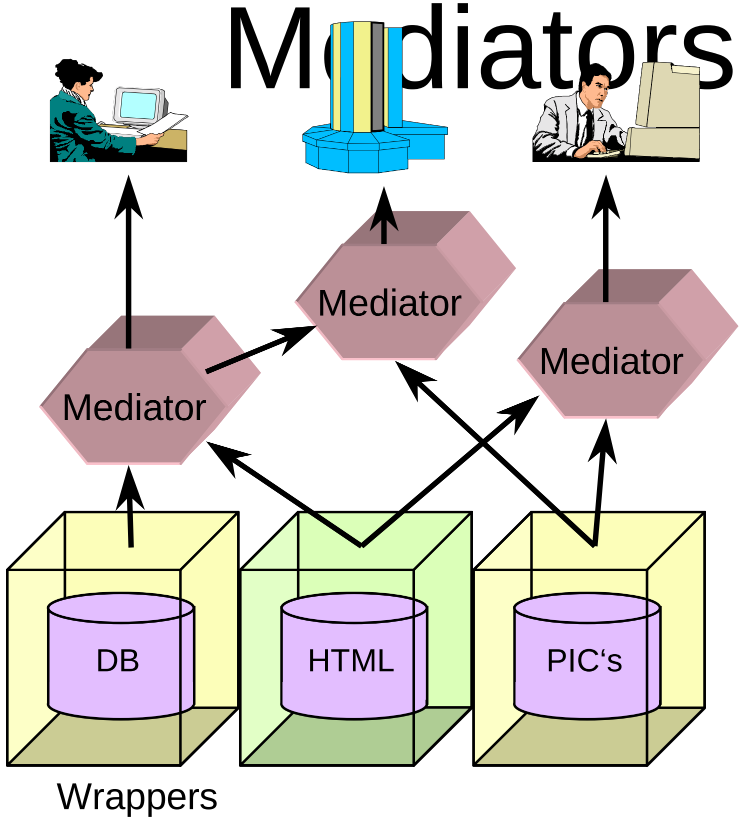
Conclusions (cont.)

- Open questions
  - Mapping between ontologies still „ad-hoc or arbitrary“ rather than well-founded
  - Need for the investigation on a theoretical and empirical basis
  - Lack of methodologies supporting the development and use of ontologies
  - Methodology should be language independent

BUSTER: Systemarchitektur



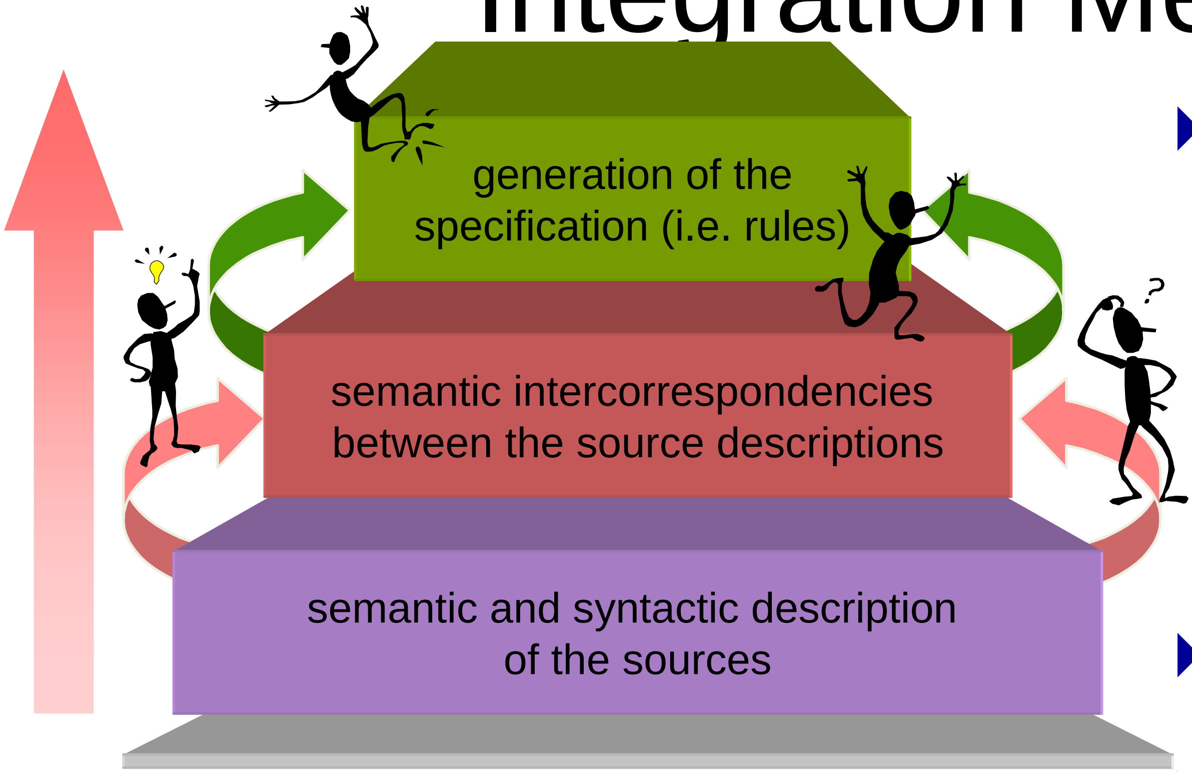
Mediators and Wrappers



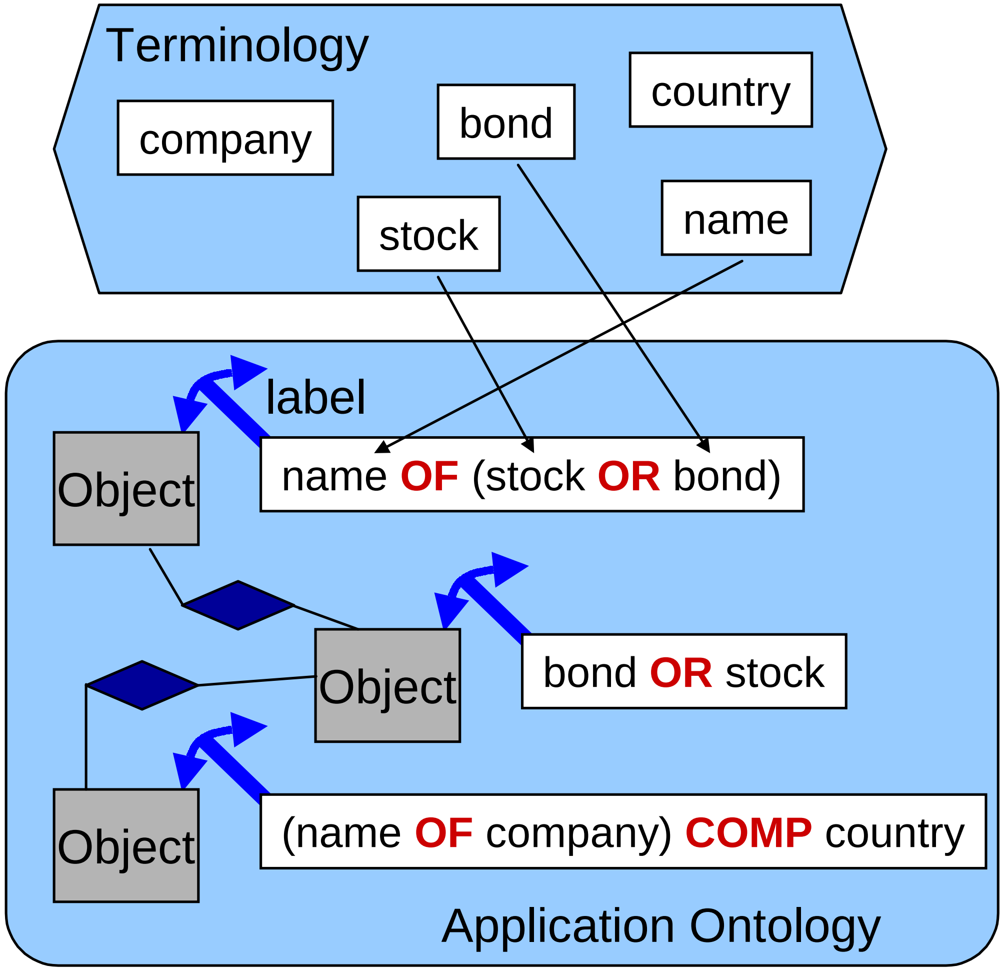
- **Wrappers** provide a uniform interface to different heterogeneous information source
- **Mediators** "combine, integrate, and abstract" [Wiederhold91] the information
- Mediators can be **specified by rules**
- **Application** in a heavy changing environment (e.g. the internet)

Problem: How to find the specification (i.e. transformation rules) for the mediator?

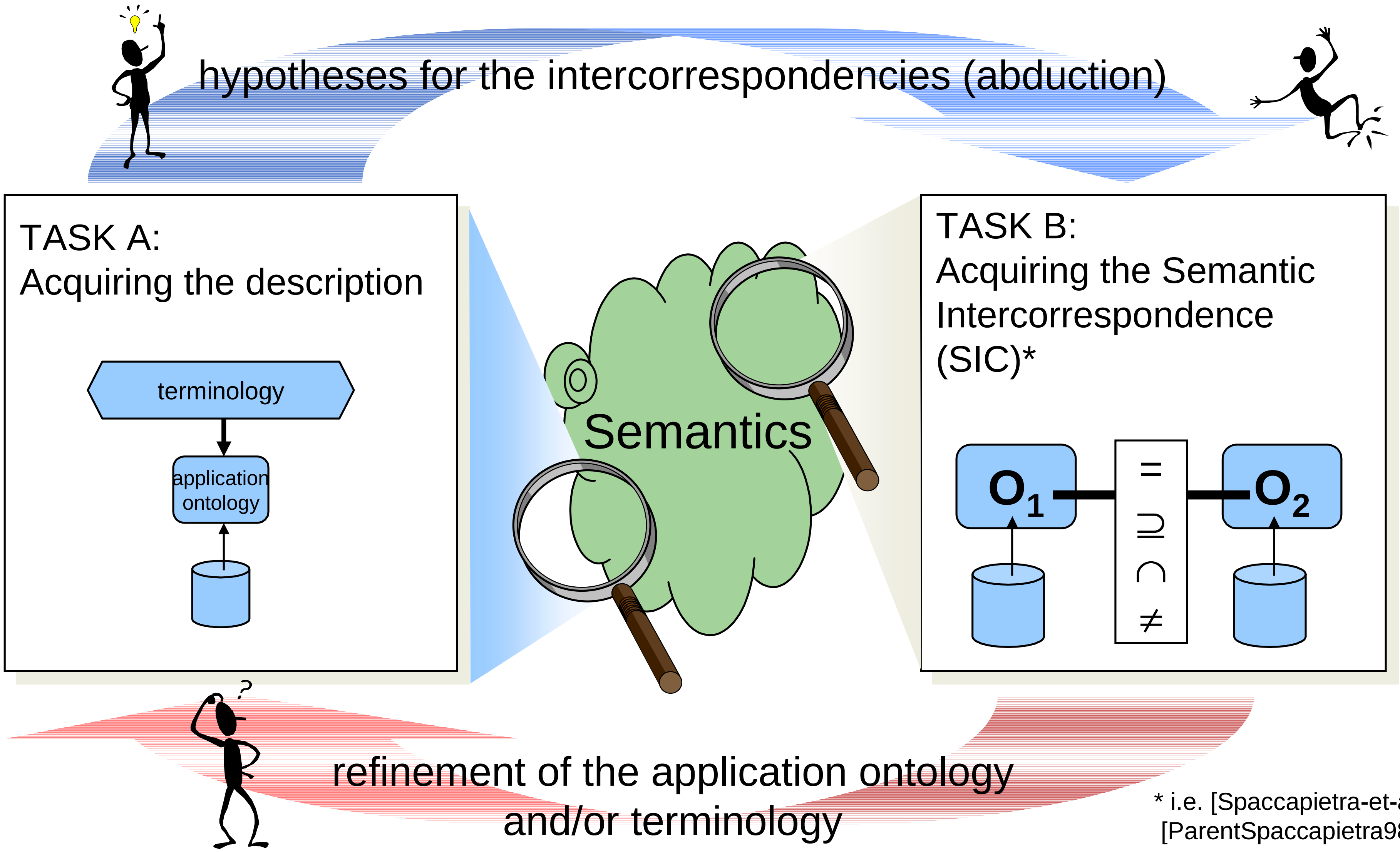
The Three Steps of the Integration Method



- **Procedure:**
  - describing each source
  - relate the source items
  - transform relationship into specification
- **Assistants** help the user in each step
- Syntactic and **semantic description** of the sources



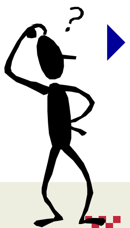
- **Terminology** = primitive domain vocabulary
- **Application Ontology (AO)** = complex terms (labels) built from primitive terms with constructors
- In AO terms are arranged according to the structure of a source
- **Constructors**
  - AND, OR, NOT
  - COMP (combination)
  - OF (specialization)
- Well-founded semantics (description logic)



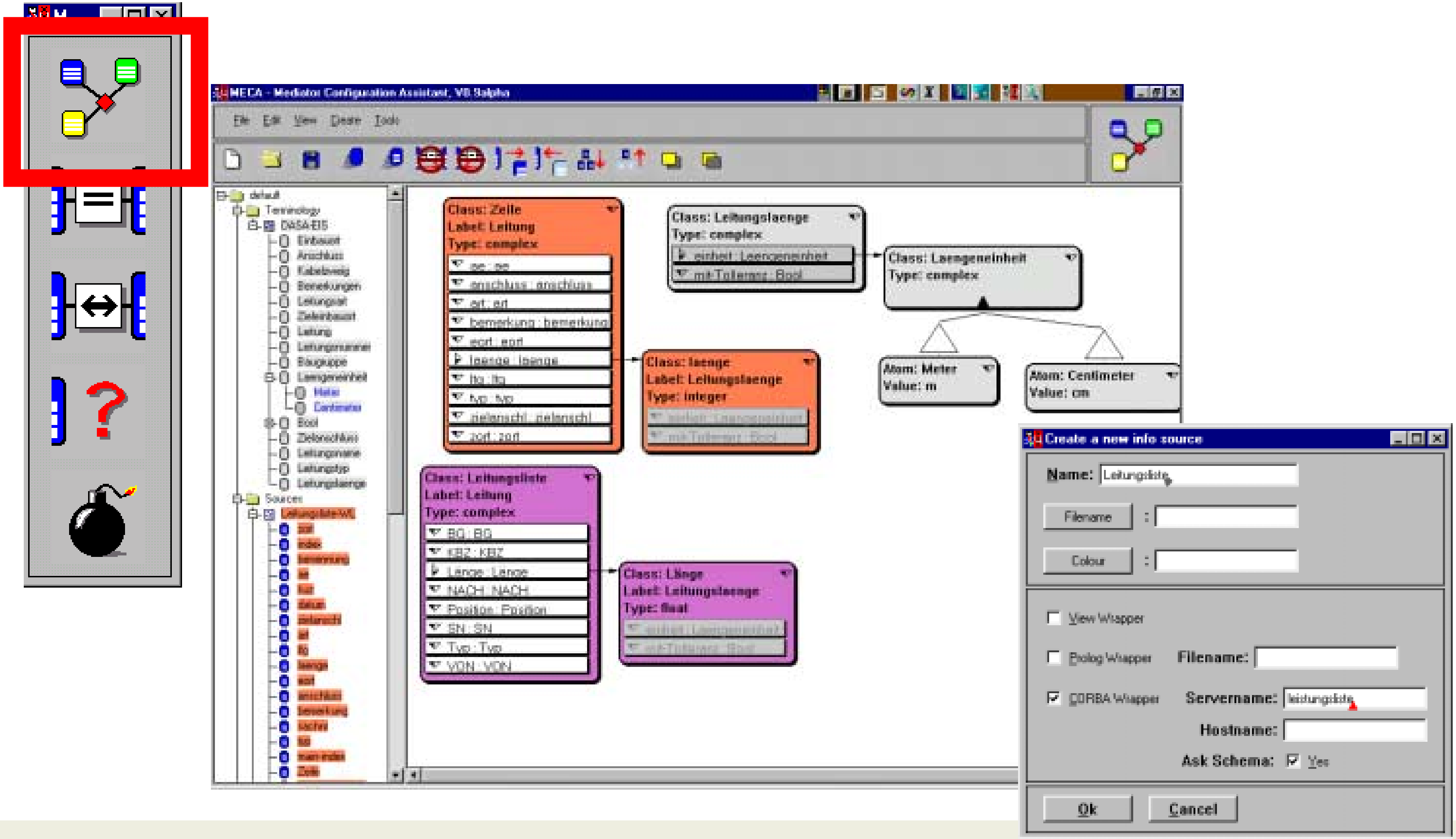
* i.e. [Spaccapietra-et-al92],
[ParentSpaccapietra98]

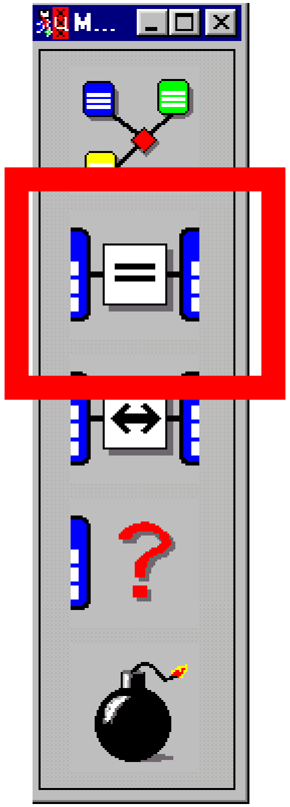
Assisting the Integration Process

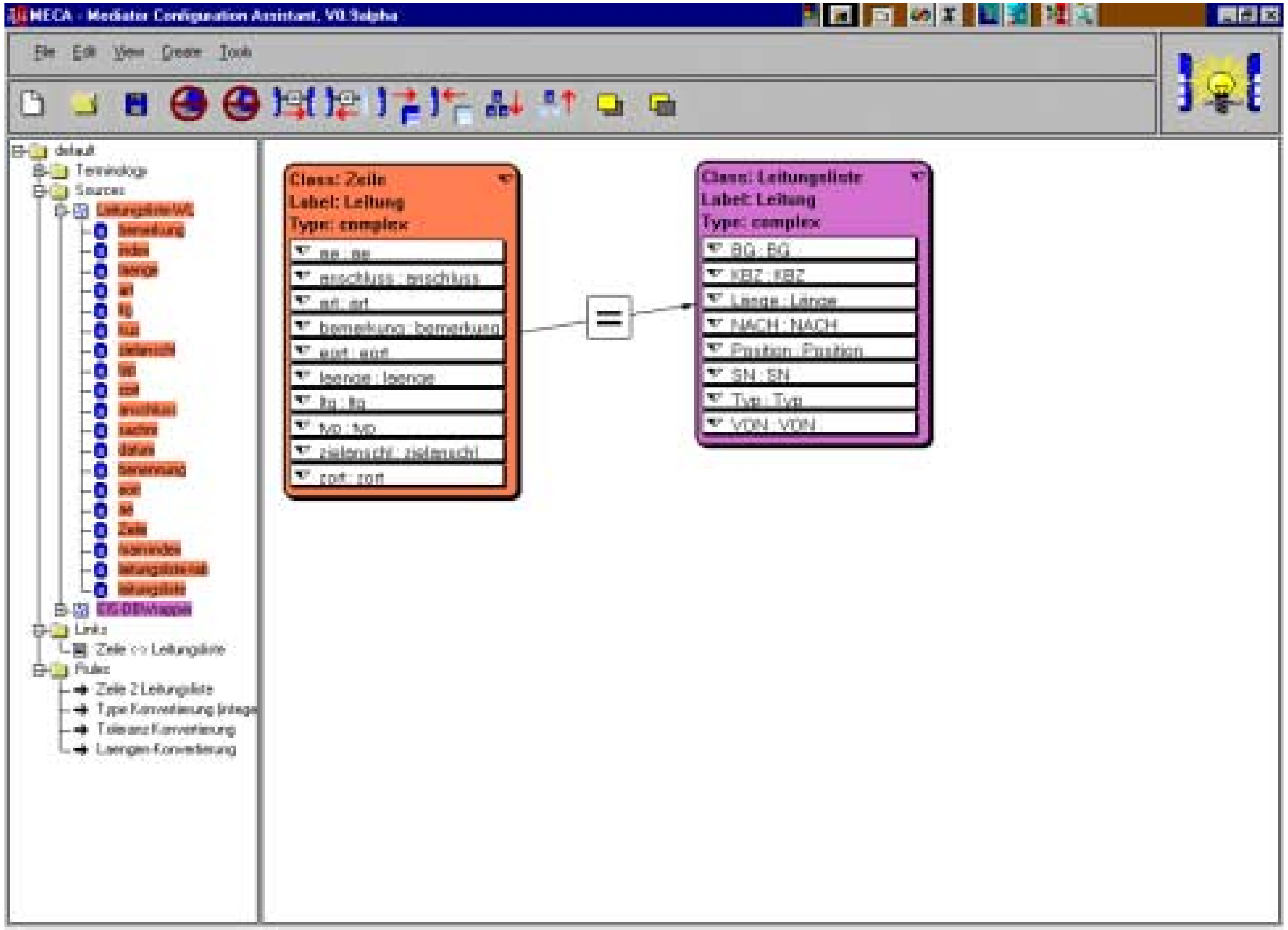
- several software assistants support the users in their tasks
- assistants only generate hypotheses validated by an user
- assistants are:
  - for the description of sources
    - case-based reasoning: (similar structure = similar semantics)
    - knowledge-based assistants (e.g. using common sense knowledge like CYC)
    - ....
  - for the semantic intercorrespondancies:
    - abduction from the semantic description of the source
    - ....

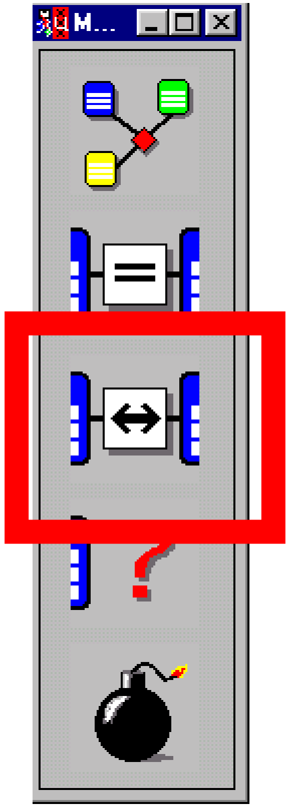


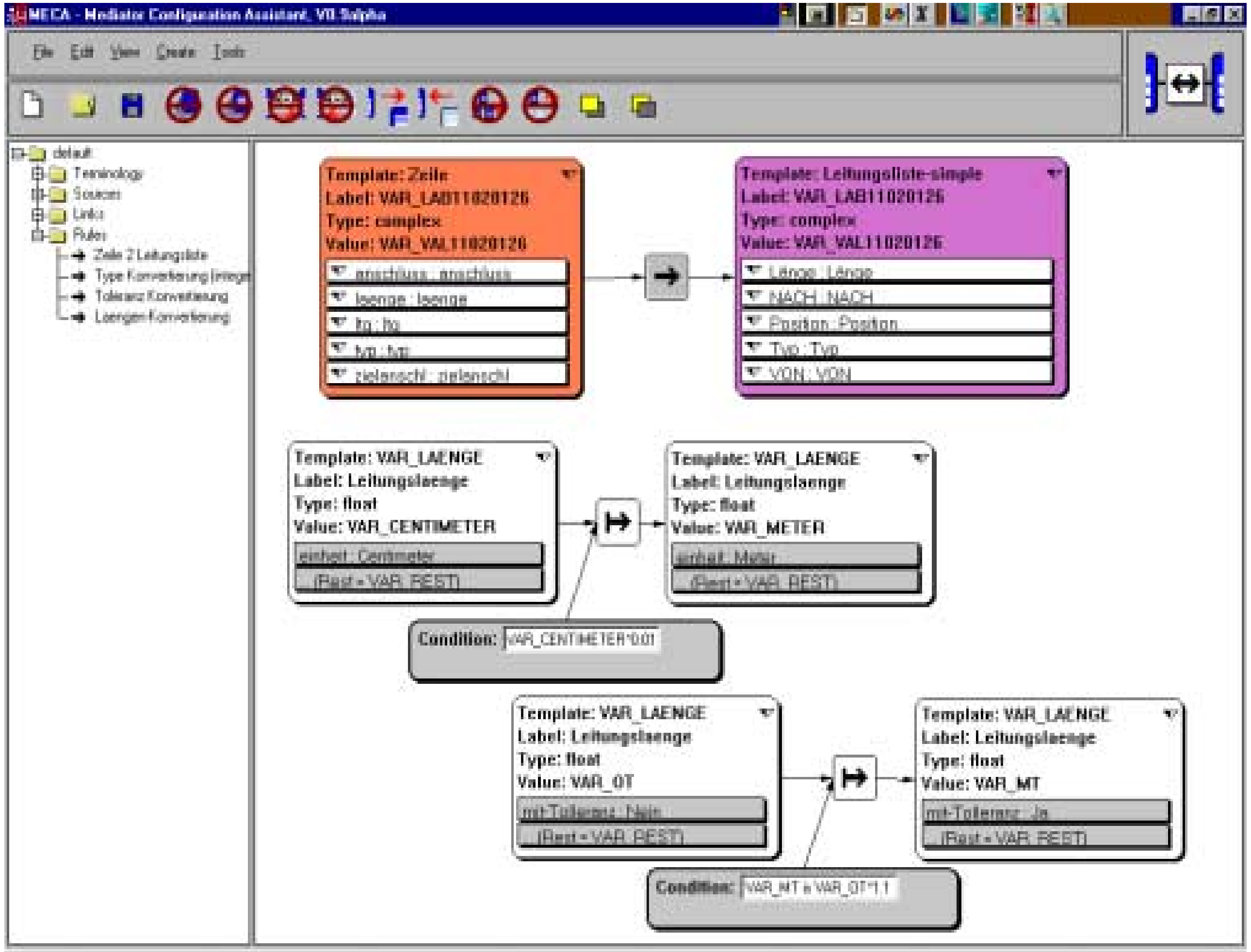
- currently under development

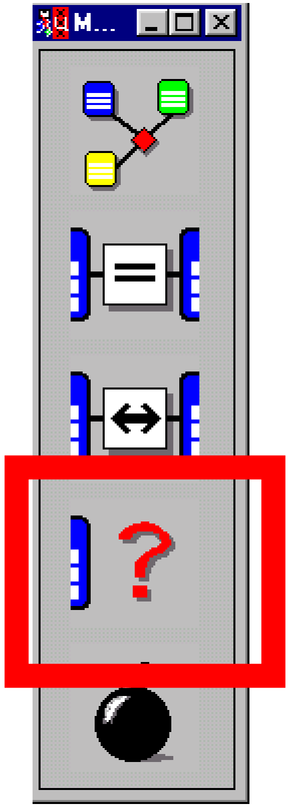

MECA - Mediator Configuration Assistant, V0.5alpha
File Edit View Create Tools
Class: Zelle
Label: Leitung
Type: complex
Class: Leitungslaenge
Type: complex
Class: Laengeneinheit
Type: complex
Class: laenge
Label: Leitungslaenge
Type: integer
Atom: Meter
Value: m
Atom: Centimeter
Value: cm
Class: Leitungsliste
Label: Leitung
Type: complex
Class: Länge
Label: Leitungslaenge
Type: float
Create a new info source
Name:
Filename
Colour
View Wrapper
Prolog Wrapper
Filename:
CORBA Wrapper
Servername:
Hostname:
Ask Schema:
Yes
Ok
Cancel

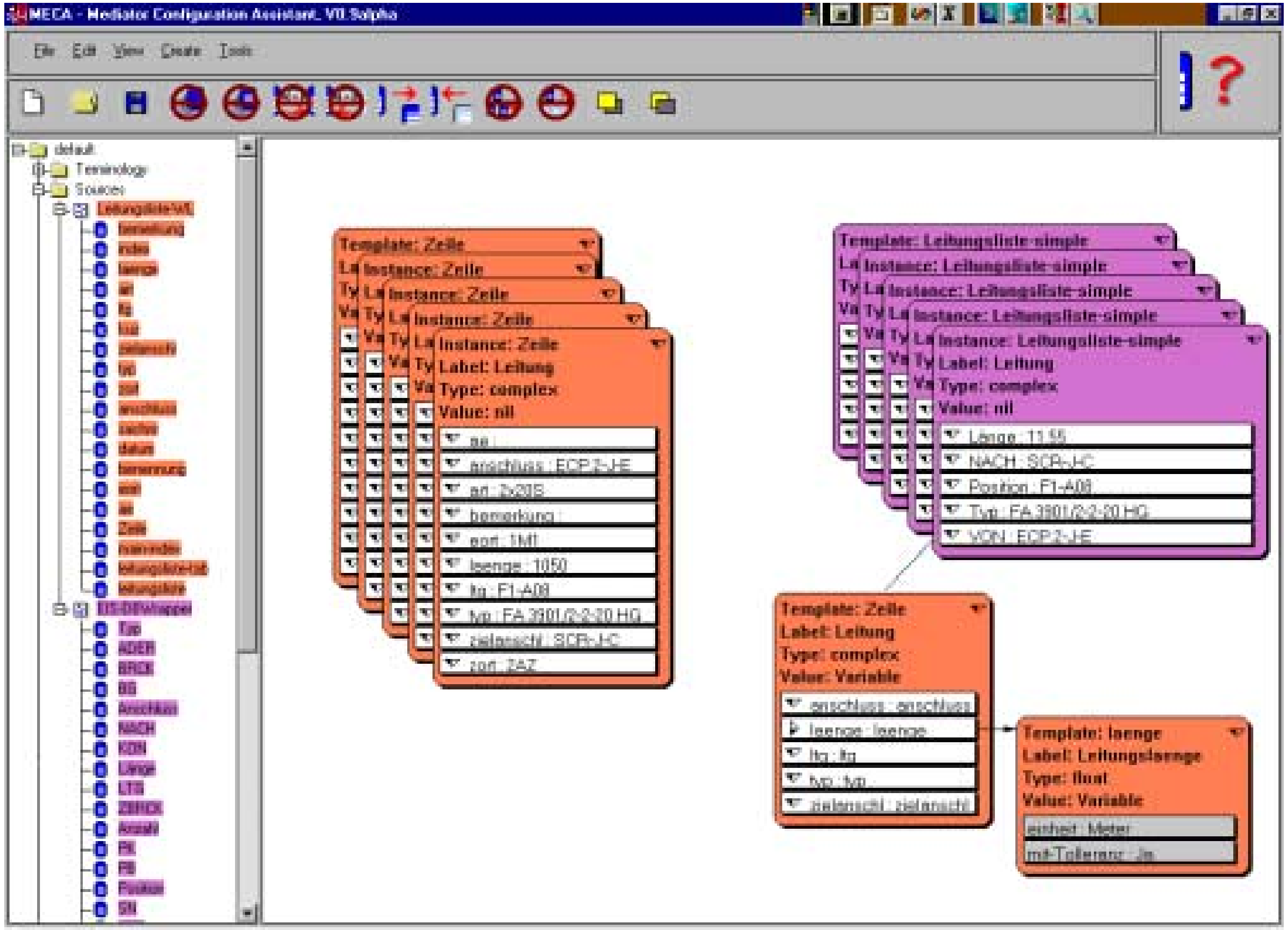



MECA - Mediator Configuration Assistant, V0.3alpha
File Edit View Create Tools
Class: Zeile
Label: Leitung
Type: complex
Class: Leitungsliste
Label: Leitung
Type: complex
BG : BG
KBZ : KBZ
Laenge : Laenge
NACH : NACH
Position : Position
SN : SN
Typ : Typ
VDN : VDN
default
Terminology
Sources
Links
Zeile <-> Leitungsliste
Rules
Zeile 2 Leitungsliste
Type Konvertierung Integer
Toleranz Konvertierung
Laengen Konvertierung

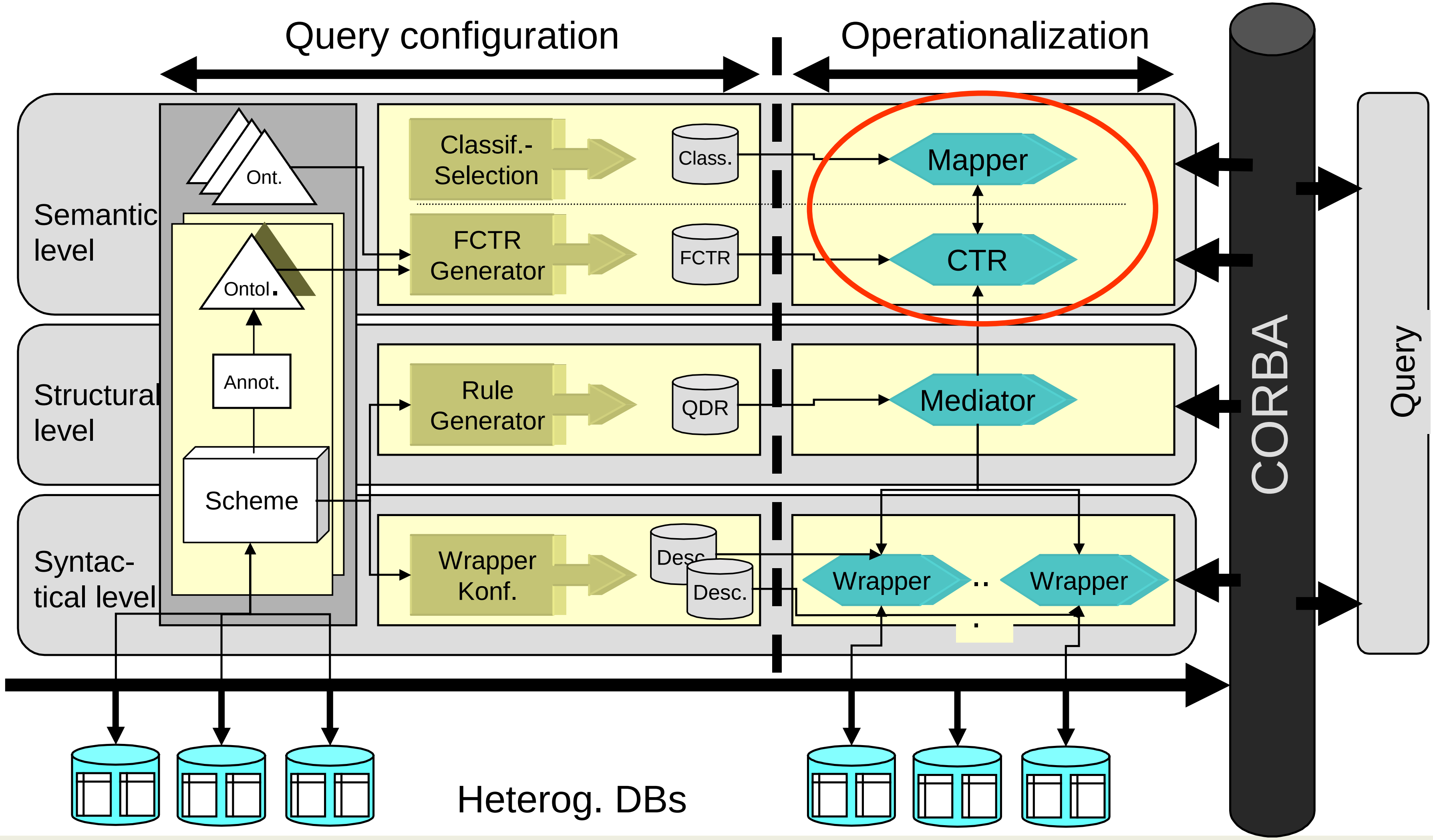



MECA - Mediator Configuration Assistant, V0.5alpha
File Edit View Create Tools
default
Terminology
Sources
Links
Rules
Zeile 2 Leitungsliste
Type Konvertierung Integer
Toleranz Konvertierung
Laengen Konvertierung
Template: Zeile
Label: VAR_LAB11020126
Type: complex
Value: VAR_VAL11020126
anschluss : anschluss
laenge : laenge
lfg : lfg
typ : typ
zielanschl : zielanschl
Template: Leitungsliste-simple
Label: VAR_LAB11020126
Type: complex
Value: VAR_VAL11020126
Länge : Länge
NACH : NACH
Position : Position
Typ : Typ
VON : VON
Template: VAR_LAENGE
Label: Leitungslaenge
Type: float
Value: VAR_CENTIMETER
einheit : Centimeter
(Rest = VAR_REST)
Template: VAR_LAENGE
Label: Leitungslaenge
Type: float
Value: VAR_METER
einheit : Meter
(Rest = VAR_REST)
Condition: VAR_CENTIMETER*0.01
Template: VAR_LAENGE
Label: Leitungslaenge
Type: float
Value: VAR_OT
mit-Toleranz : Nein
(Rest = VAR_REST)
Template: VAR_LAENGE
Label: Leitungslaenge
Type: float
Value: VAR_MT
mit-Toleranz : Ja
(Rest = VAR_REST)
Condition: VAR_MT is VAR_OT*1.1



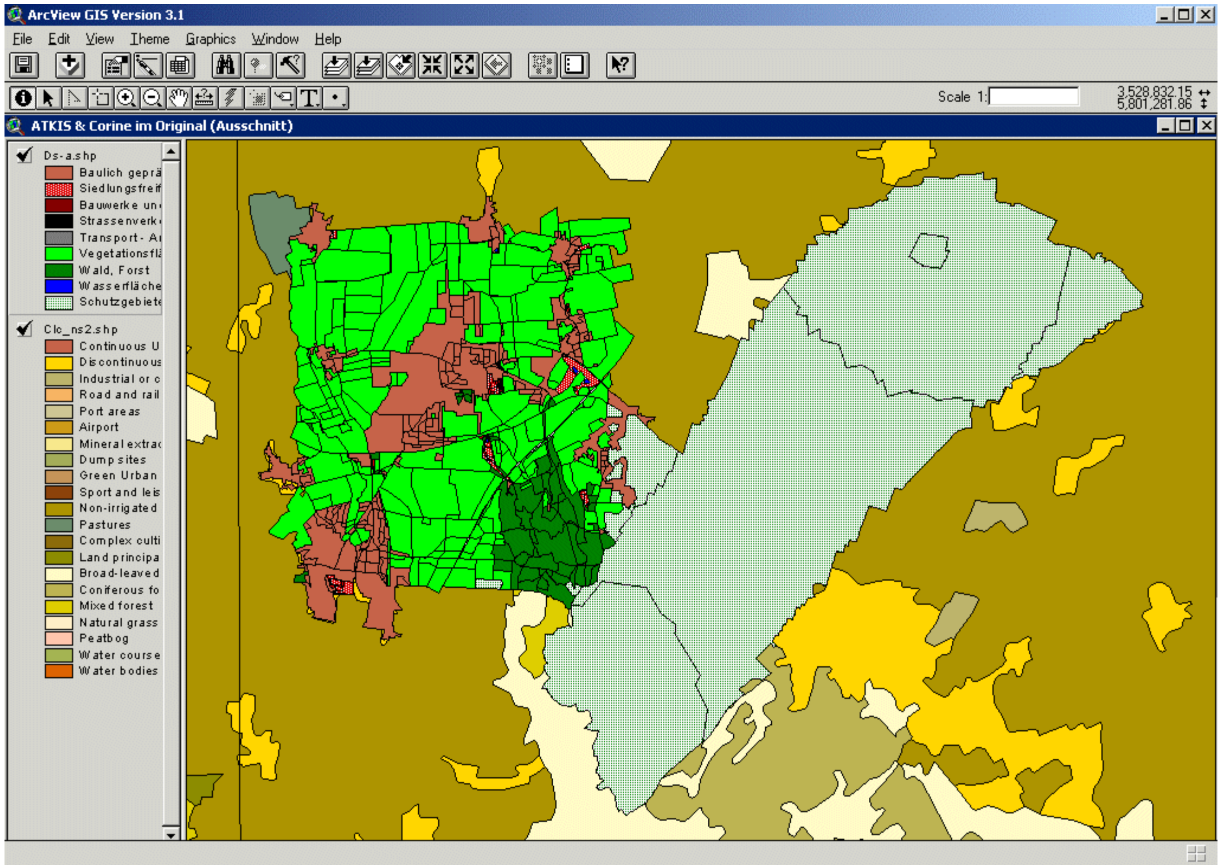

MECA - Mediator Configuration Assistant, V0.9alpha
File Edit View Create Tools
default
Terminology
Sources
Template: Zelle
Instance: Zelle
Label: Leitung
Type: complex
Value: nil
Template: Leitungsliste-simple
Instance: Leitungsliste-simple
Value: Variable
Template: laenge
Label: Leitungslaenge
Type: float

BUSTER: Systemarchitektur



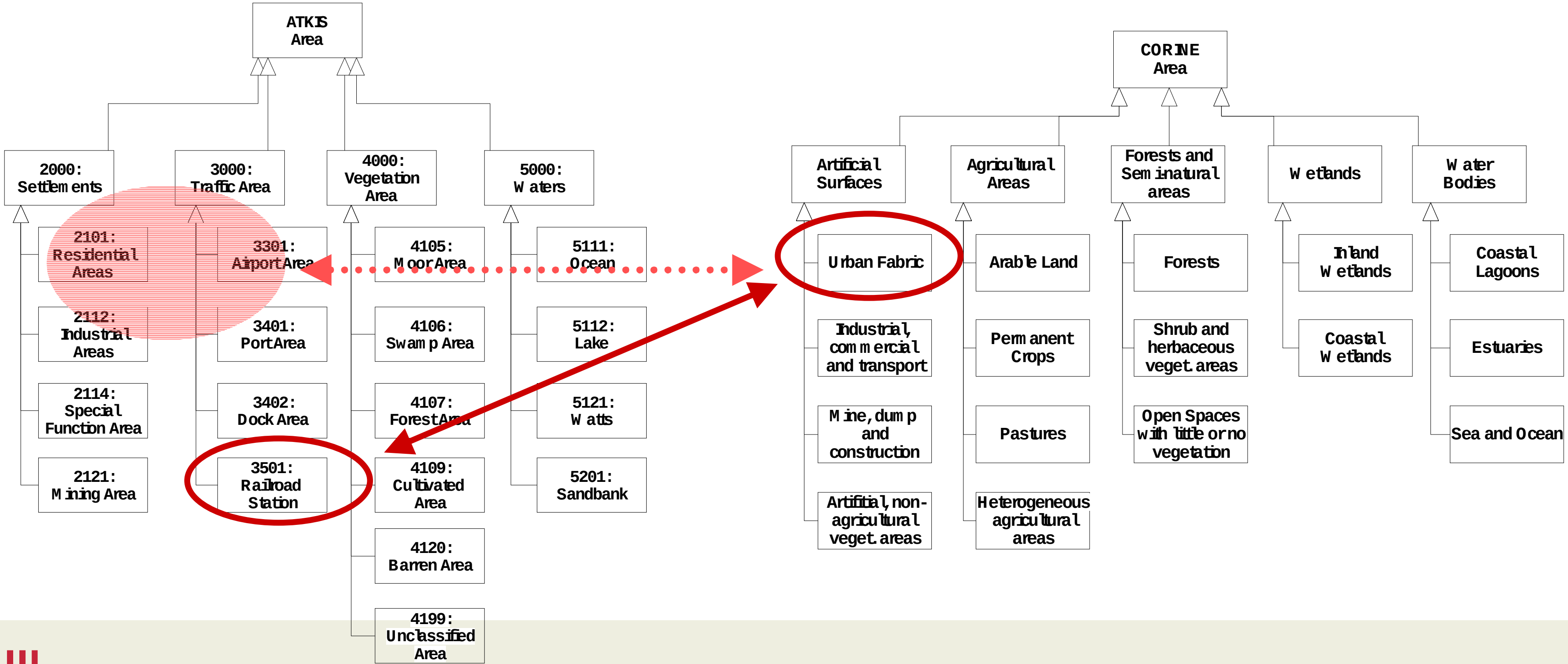
Motivation

- Semantic Heterogeneity
- Example:
  - Sharing geographic information
  - Integration of land-use classes from different catalogues

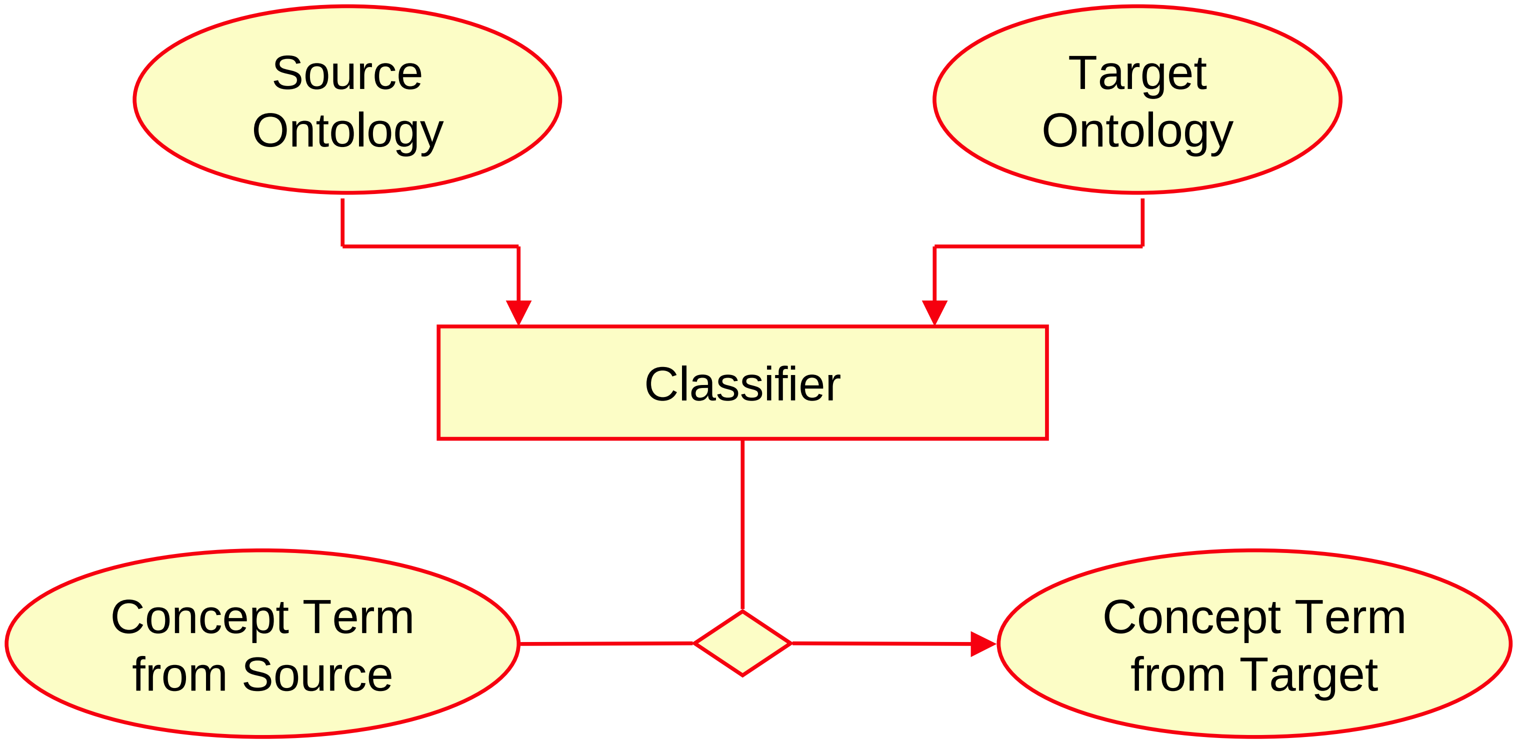


The Problem: Different Catalogues

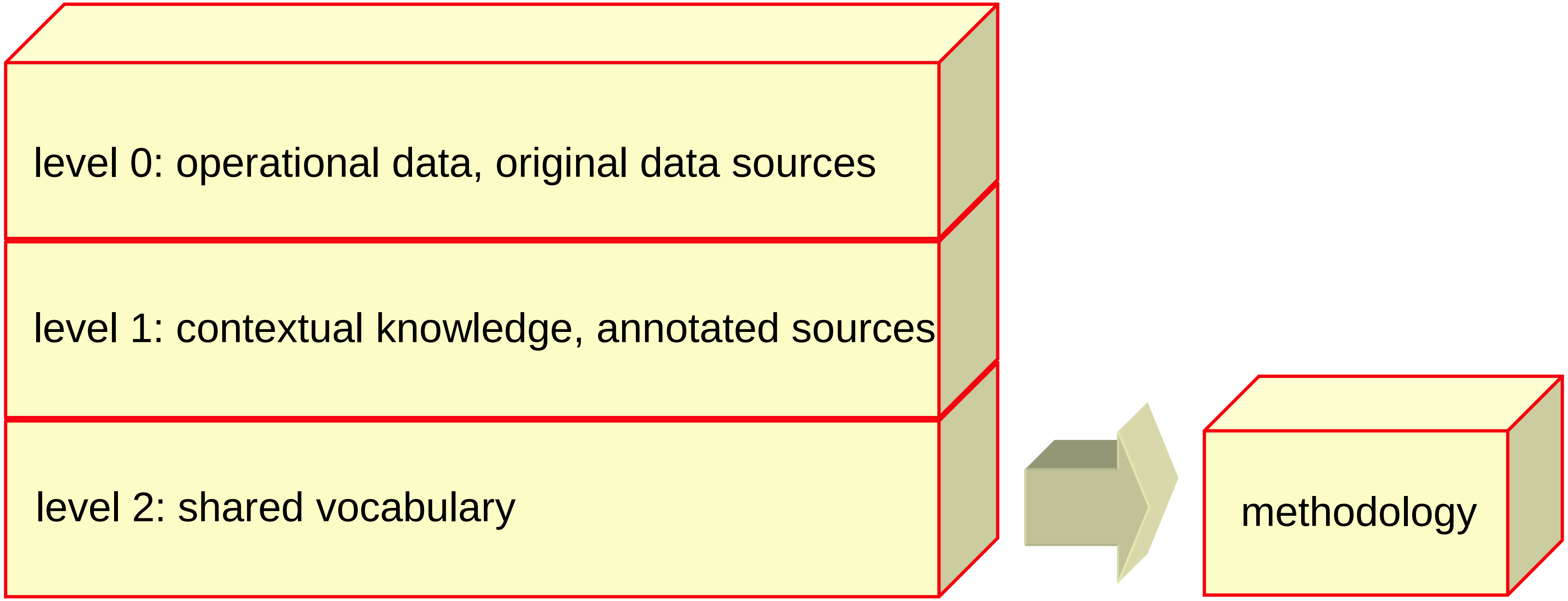
- ATKIS-OK-1000
- CORINE Landcover



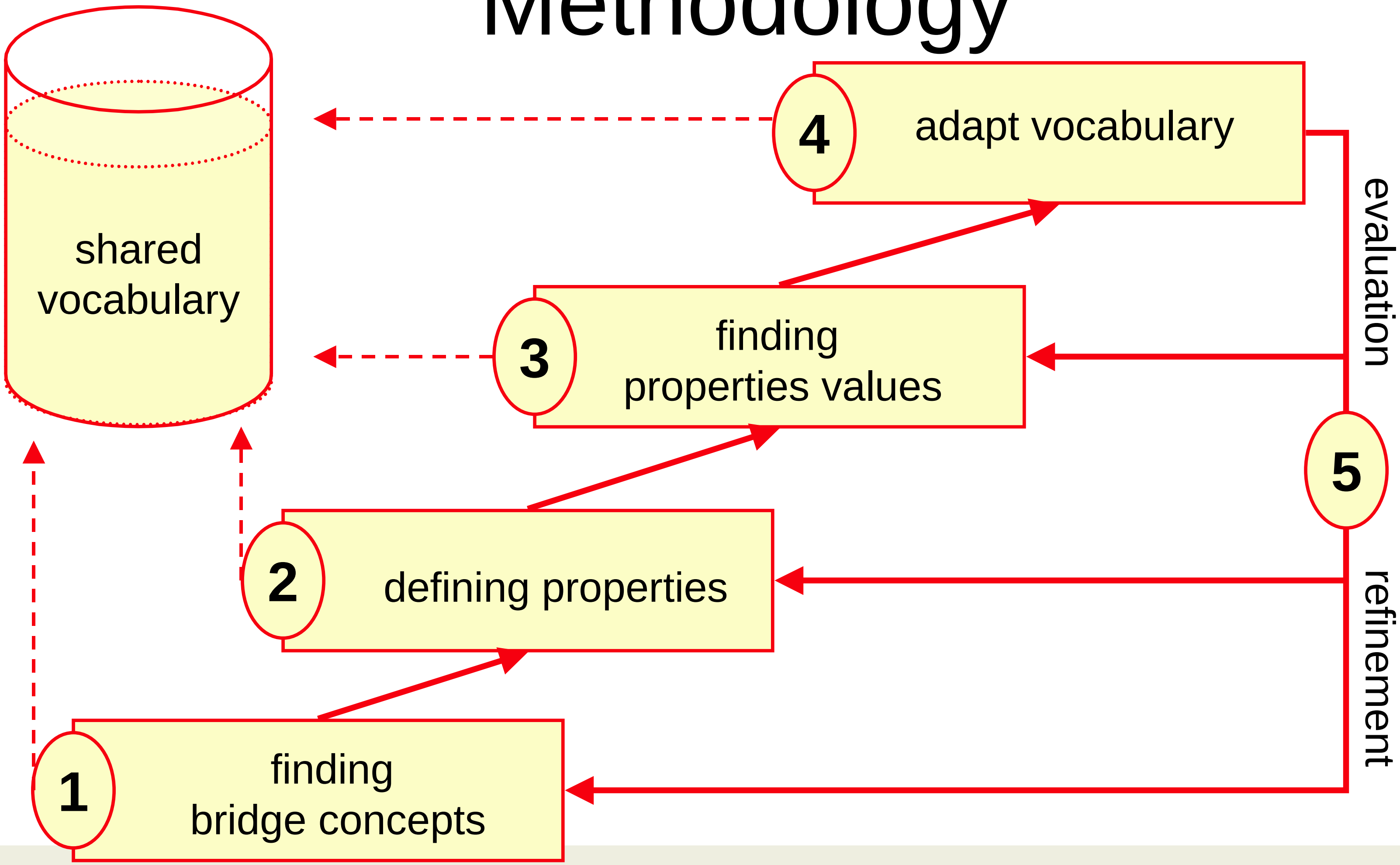
Semantic Translation of Information Entities



Role of Ontologies



Methodology



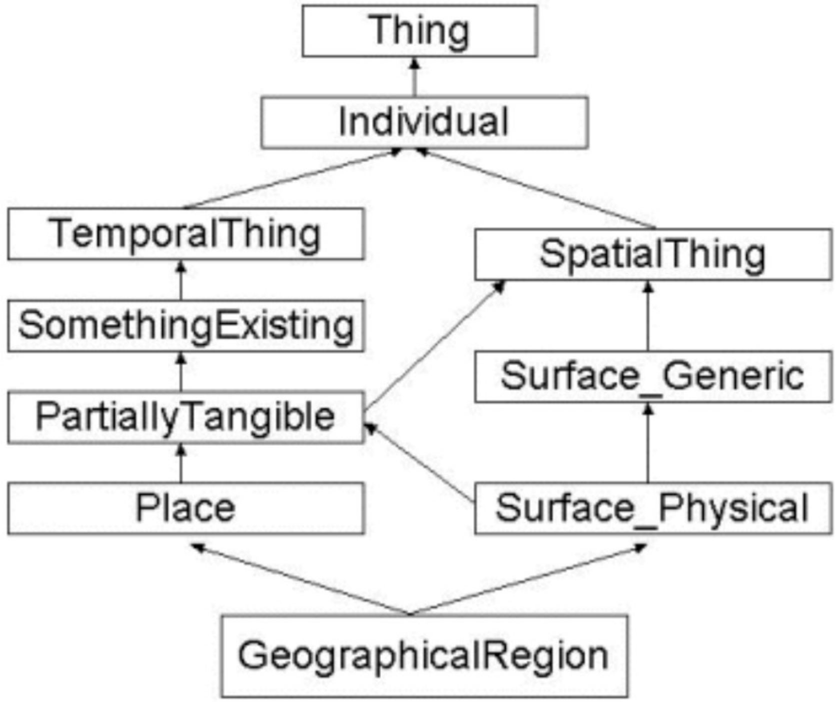
Sources of Information

- Data Catalogues
  - Task specific
- Upper-Level Ontologies
  - Upper-Cyc [Lenat/Guha1990], Pangloss [Knight/Luk1994] ...
- Scientific Classifications
  - Classification of plant life, ...
- Domain Thesauri
  - Task specific thesauri, like UDK, GEMET, ...
- Linguistic Thesauri
  - WordNet, ...

1. Finding bridge concepts

- Need for concept like „area“
  - subsums all land-use classes
- Search in Upper-CYC
  - results in GeographicalRegion
- OIL-Notation:

class-def Geographical-Region



2. Defining Properties

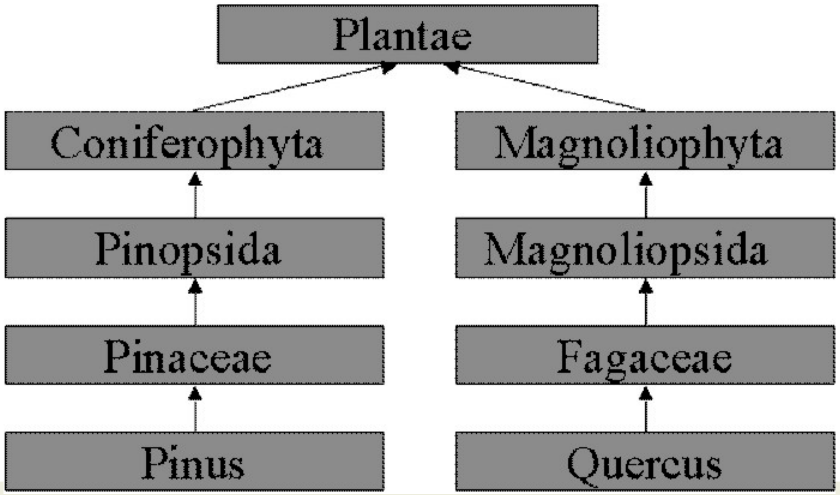
- Search in Gemet:
  - **Geography**: *The study of the natural features of the earth‘s surface, comprising topography, climate, soil, **vegetation**, etc. and man‘s response to them.*
  - **Region:** *A designated area or an administrative division of a city, county or larger geographical territory that is formulated according to some **biological**, political, economic or demographic criteria.*

- OIL-Notation:

slot-def vegetation Domain Geographical-Region
class-def Geographical-Region

3. Finding property values

- Search for „vegetation“ in Gemet:
  - *The plants of an area considered in general or as communities [· · ·]; the total plant cover in a particular area or on the Earth as a whole.*
  - **WordNet:** *The plant life characterizing a specific geographic region or environment.*
- Integration of standard scientific taxonomies
  - GoogleWebdirectory (plants)

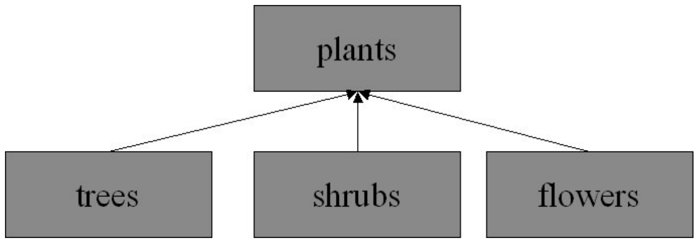


4. Adapt shared vocabulary

- Annotated concept → Problem: **current vocabulary not specific enough**

class-def c-Broad-leaved-forest
subclass-of Geographical-Region
slot-constraint vegetation **value-type** Magnoliophyta

- Enhance shared vocabulary:



class-def c-Broad-leaved-forest
subclass-of Geographical-Region
slot-constraint vegetation **value-type** Magnoliophyta and (trees or shrubs)

5. Evaluation / Refinement

- Evaluation through re-classification
  - Try to annotate all concepts from data catalogues with shared vocabulary
  - Classify by reasoning mechanisms (FaCT, Racer)
- Examine results
- Iterative Refinement if needed
  - Return to Step 1 to 4

Summary

- Semantic interoperability is an important problem
  - Data Warehouses and distributes
  - World-Wide Web, Intranets
- Ontologies are a key technology
  - Many integration approaches rely on them
  - New interest in connection with the World Wide Web

Summary

- Technical Solutions exist
  - Many Systems, some products
  - Well founded in formal logics and still applicable
- Modeling is the Bottleneck
  - Ontologies have to be built
  - Information has to be annotated

Conclusion

- There is a need for
  - **methodologies,**
  - ...that are partially automated
  - ...and supported by tools.

- Reserach on this Issue must go hand in hand with applications, because we have to learn from the users.