# Data Semantics Revisited: Databases and the Semantic Web


Alex Borgida
Rutgers University

John Mylopoulos
University of Toronto


SEBD'05, Bressanone
June 19, 2005



## Next Generation Database Systems Won't Work Without Semantics!

Panel Discussion
ACM SIGMOD'98,
Seattle, June 4, 1998

…1998…

## Data Semantics Can't Fail This Time!

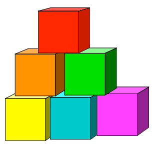
Panel Discussion
CAiSE'98, Pisa
June 11, 1998



## The Panelists

**SIGMOD'98 Panel:**
Philip Bernstein (Microsoft)
Umesh Dayal (HP Laboratories)
Sham Navathe (Georgia Tech)
Marek Rusinkiewicz (MCC)
**Panel chair:**
John Mylopoulos (Univ. of Toronto)

**CAiSE'98 Panel:**
Michael Brodie (GTE Labs)
Stefano Ceri (Politecnico di Milano)
Arne Solvberg (Univ. of Trondheim)
**Panel chair:**
John Mylopoulos (Univ. of Toronto)





*"…The three most important problems in Databases used to be* ***Performance, Performance and Performance****;*
*in the future, the three most important problems will be* ***Semantics, Semantics and Semantics****…"*

*(paraphrase) Stefano Ceri*
*June 11, 1998*

## This Tutorial

- *Data Semantics: The problem and its history*
- *The Semantic Web: The vision and the challenges*
- *Towards richer theories of data semantics*



## Data Semantics

- The problem of establishing and maintaining the correspondence between an information source, hereafter a ***model***, and its intended subject matter.
- The model may be
  - ✓ A database storing data about employees in a company;
  - ✓ A database schema describing parts, projects and suppliers;
  - ✓ A website presenting information about a university;
  - ✓ A plain text file describing the battle of Waterloo;
  - ✓ An equation relating observables.

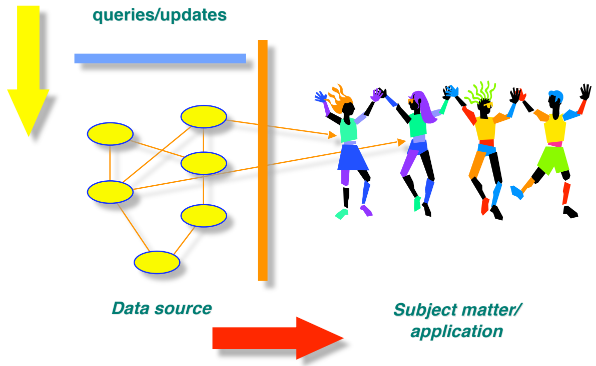


## Possible Solutions: The Good, the Bad and the Ugly

- Have a person build the data source and maintain it throughout its lifetime, answering questions and doing updates (..the bad…)
- Make the data source a rich theory in First Order Logic that represents the subject matter, answers queries and checks consistency for updates (…the good…)
- Capture a variety of information about the data source, including how were the data obtained, what were the intentions of the designers of the data source, and how trustworthy was it found to be in the past (…the ugly…)



## Machine State vs World Semantics



## Semantic Data Models

- Data models that attempt to capture "more world knowledge" [Codd79] than their logical counterparts.
- Make ***ontological assumptions*** about the subject matter, offer primitives accordingly.
- Are more expressive than their logical cousins, i.e., can represent constraints other than keys and foreign keys.
- For example, the Entity-Relationship model "assumes that the world consists of entities and relationships" [Chen76].
- The Relational Model makes no ontological assumptions [Codd70].




## Data Semantics -- take 1

- The semantics of data is captured through a rich conceptual schema in an expressive semantic data model.
- Emphasis on expressiveness over tractability/ performance.




## History

- First semantic data models were proposed in 1974:
  - ✓ Jean-Raymond Abrial;
  - ✓ Giampio Bracchi, Paolo Paolini, Giuseppe Pelagatti;
  - ✓ Jean-Luc Hainaut and Alain Pirotte;
  - ✓ Hans Schmid and Richard Swenson;

Then in 1975
  - ✓ Peter Chen;
  - ✓ Sham Navathe;
  - ✓ Nick Roussopoulos and John Mylopoulos;
  - ✓ ….many others…




## Where do Semantic Data Models Fit?

- …in database practice?
- Several possibilities, actually:
  - ✓ They are part of the technology --> Semantic DBMSs;
  - ✓ They are used during design --> Constitute an integral component of the design process;
  - ✓ They are part of the user interface to a database.
- Note that database research -- then and now -- deals mostly with technologies, not methodologies or user interfaces

## Data Semantics -- take 2

- Option 2 prevailed. Semantics were to be dealt with during ***design-time***, rather than ***run-time***, for performance reasons.
- That is, the database system processes at run-time tables with no regard for their meaning.
- But how does one use a database where semantics have been factored out?

***Rely on a stable environment of users and application programs to know the semantics!***



## …There is a down side to this data management practice: Legacy data…



## What did the Panel Experts See?

- Factoring out the semantics of data won't work in ever-changing, distributed, open environments, such as the web.
- In such settings, access to the data can not be restricted to a small set of users and applications programs.
- This means that new users will have no clue what the data means; and the application programs that process the data may not have been designed specifically for these data.

***Hence the need to bundle together data and its semantics!***



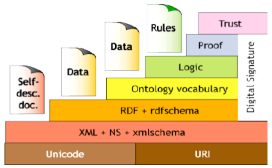
## …1999… The Semantic Web

- Unlike databases, hypertext data ***are*** designed for human consumption. However, these data are not machine processable.
- Hence the call for the Semantic Web [Berners-Lee01].
- Machine-processable web data has come to mean "…having semantic metadata and ontologies for web content to enable information access, integration, interoperation and consistency…"

Katia Sycara
ODBASE'03, November 7, 2003

## The Layered Cake Architecture

- From bottom to top…
  - ✓ Unicode, URI;
  - ✓ XML data, XML schema;
  - ✓ RDF/OWL data, schema;
  - ✓ Ontology, vocabulary;
  - ✓ Logic;
  - ✓ Proof;
  - ✓ Trust.

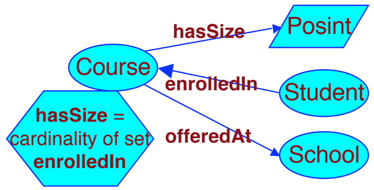



***…but, who uses what, when??***



## Data Semantics - take 3

- ***Formal ontology*** helps clarify the meaning of terms
- Web page, in XML, is ***annotated*** using terms from the ontology
- Annotation is used by browsers, search engines, e-service composition agents..




**... The <concept superclass= UnivOnt:Course ...> seminar </concept> will cover a lot of material about the Greek philosophers in a short time ...**



## Progress

- A lot of work on formal languages for expressing ontologies, and reasoning with them -- OWL

***Class(univ:course partial***
  ***restriction(univ:offeredAt***
    ***allValuesFrom(univ:school))***
  ***...***
***DisjClasses(univ:course univ:student univ:school)***

- Substantial advances over semantic data models.
- Very little work on annotation and use, which are intimately related.



## Some Concerns

- Hard to develop technologies for computationally demanding tasks, e.g., theorem provers, model checkers, deductive databases,…
- Scalability??
- Practitioners tend to not use logical formal languages, e.g., Z, Datalog,…

**From bottom to top...**
- ✓ **Unicode, URI;**
- ✓ **XML data, XML schema;**
- ✓ **RDF data, RDF schema;**
- ✓ **Ontology, vocabulary;**
- ✓ **Logic;**
- ✓ **Proof;**
- ✓ **Trust.**

***Take 3 looks a lot like take 1, but progress has been made wrt expressiveness***

## Towards Novel Theories of Data Semantics

- Basic premise: If we are going to tackle the problem of data semantics -- *again* -- we better have some new angles on the problem:

  ***I. The mapping continuum and semantic encapsulation***

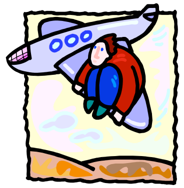
  ***II. Intentional aspects of data semantics***




## I. The Mapping Continuum & Semantic Encapsulation

- *Look more carefully at the notion of modeling*

(Joint work with Yuan An)






## What is a Model?

***"M is a model of subject S for purpose P"*** [Ladkin97]

- Usually, the purpose is *"answering certain kinds of questions* (about the subject)"
- For example, a model airplane at Boeing can answer questions about the ***relative*** dimensions and aerodynamics of an aircraft; but not questions about its absolute dimensions, color, interior decoration, etc.
- (Note that the model usually introduces details that are not relevant to the subject: E.g., material from which the model airplane is made)




## Building a Model

- Hence, for every model, we need
  - ✓ methods for building/updating the model;
  - ✓ a way of asking questions (viz tell/ask interface)
  - ✓ a mapping to help translate applicable queries about the subject matter into queries about the model;
  - ✓ conversely, a way to translate results of the query on the model to answers about the subject
- For example, a student database as a model of the (actual) university.

## Typology of Models

- ***E-models*** *(extensional):*
  - ✓ have a set-theoretic structure;
  - ✓ Query-answering uses set-theoretic relationships.
  - ✓ e.g., logical model theory: Tarski, Kripke models, dbs
- ***I-models*** *(intentional):*
  - ✓ consist of sets of formulas in some language
  - ✓ question answering based on an entailment relationship |=
  - ✓ e.g., ontologies, schemas
- ***C-models*** (*computational*):
  - ✓ query answering is produced by running software
  - ✓ e.g., a simulation program, ADTs




## Other Aspects of Models

- ***Terminological** vs. **Assertional***

  Generic/schema vs. Specific/individuals,facts
- ***Faithfullness***
  - ✓ (in)completeness (nulls, disjunction, OWA/CWA)
  - ✓ (un)certainty (probabilities, fuzzy sets)
  - ✓ (in)accuracy (errors, bounds)
- ***Query answering:*** choice of query languages, tractability/expressiveness tradeoffs.
- These have been the concerns of traditional research in databases and knowledge representation.




## Typology of Subjects

- Physical reality
- Human's perception of reality
- Another model !!!

  One can think of a database schema as being a model of an ER schema (of some domain); or, of an ontology for that domain.

  This makes it possible to specify precise the mapping.

**_db_:CLASS(cname,size,...) ⇔ _ont_:COURSE(e) ∧ _ont_:hasName(e,cname) ∧ _ont_:hasSize(e,size) ...**

**_db_:ENROLMENT(cname, studentId,...) ⇔ _ont_:enroledIn(c,s) ∧ _ont_:hasName(c,cname) ∧ ...**




## Mapping Languages

- Form of mapping specification
  - ✓ correspondences (drawn graphically?) [Pottinger03]
  - ✓ Queries over the subject and model: GAV, LAV, GLAV,... [Lenzerini02, Madhavan02]
- Language of query specification
  - ✓ close to First Order Logic: Datalog, SQL, Description Logics
  - ✓ more complex: higher order logics (Hylog), dealing with nesting (XQuery)

## Data Semantics - take 4

- Suppose that the contents of a relational database are published on the web in an XML document, with schema

```
<!ELEMENT catalog (course *) >
<!ELEMENT course ( name (student *) ) >
<!ELEMENT student (studtNumber (email l phone)) >
```

- Document might be populated using an XQuery:

```
<catalog>
  {for $c in $db/Course/Tuple
     return <course>
        <name> $c/cname </name>
         {for $e in $db/Enrolment/Tuple <student>
  ...
```

- Accordingly, the meaning of the XML doc is (***should be***) defined in terms of that of the database, its predecessor!

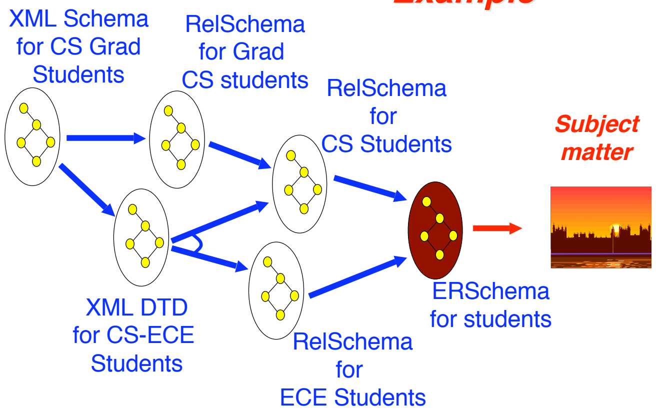



## More generally: The Correspondence Continuum

- Consider:
  - ✓ A photo of a landscape is a model of that landscape (the photo's subject matter);
  - ✓ A photocopy of the photo is a model of a model of the landscape;
  - ✓ A digitization of the photocopy is a model of the model of the model of the landscape….etc.
- Meaning is rarely a simple mapping from symbol to object; instead, it often involves a ***continuum of (semantic) correspondences*** from symbol to (symbol to)* object [Smith87]




## Example



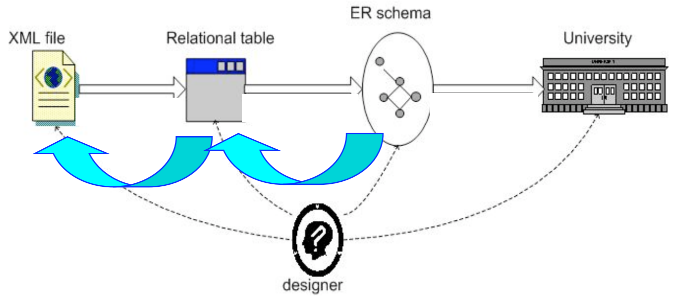




## Mapping Graphs

- The graph associated with each mapping continuum is acyclic and has one or more "roots"
- One would expect roots to be "closer" to the actual domain, and hence capture more of its semantics, acting like semantic models, ontologies.
- Some mappings are ***lineal***, defining the meaning of a schema, in the data semantics sense. Let's call them ***semantic mappings.***

# Lineal/Semantic Mappings






# Data Semantics - take 5

- Require "semantic encapsulation":

***Every (non-root) model comes with an explicit semantic mapping to some other model(s)***

- The meaning of a model is given in terms of the composition of the semantic mappings that relate it to its root.



# Related Work

- Heterogeneous Database Integration
- Ontology Integration
- Model management
- Data provenance

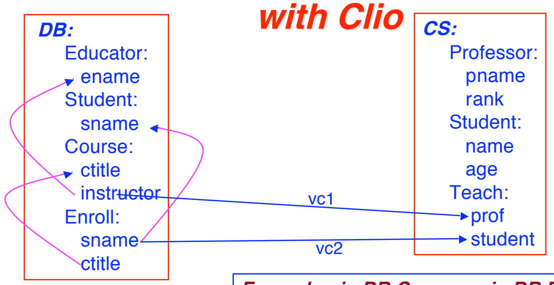
Usually, these involve models at the same level of "semantic abstraction" (e.g., two ontologies, or two XML schemas). The proposal integrates these into a single framework accounting for data semantics.



# Whence Mappings?

- The ***lineal mappings*** should be saved during design. Other mappings must be ***discovered*** and specified.
- Tools may be needed to do the later:
  - ✓ Clio [Miller00] is a schema mapping tool that discovers mappings between two schemas, given a set of (value) correspondences between their elements (attributes).
  - ✓ [Velegrakis03, Velegrakis04] extends Clio to support mapping evolution, and schema and mapping management.

## Discovering Schema Mapping with Clio

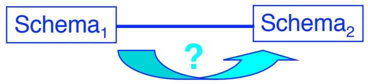



*Teach(prof,student)*
*:-Course(ct, prof),*
*Enroll(ct, student)*

*Foreach c in DB.Course, e in DB.Enroll*
*where c.Course.ctitle = e.Enroll.ctitle*
*exists t in CS.Teach*
*with c.Course.prof = t.Teach.prof*
*and e.Enroll.sname = t.teach.student*



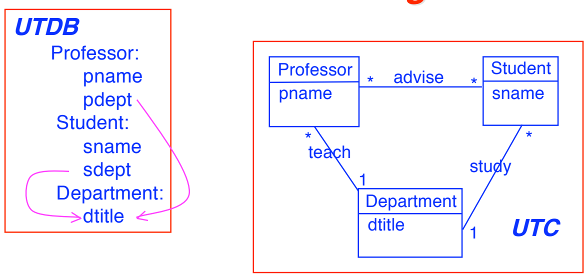
## The Clio Problem




- Clio's philosophy: it is easier for users to specify correspondences, and have tool discover the actual, more complex, mappings.
- Clio relies on co-occurrence of attributes and foreign key constraints to discover "reasonable" mappings.



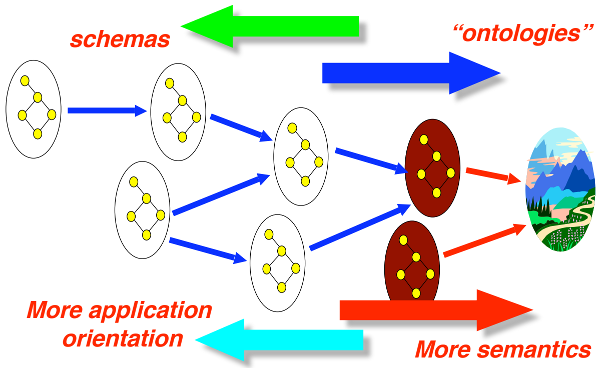
## Extending Clio




*Foreach path p:Professor----teach----d:Department in UTC*
*exists r in UTDB.Professors*
*where r.Professor.pname=p.Professor.pname*
*and r.Professor.pdept=d.Department.dtitle.*



## The Punchline



# *II. An Intentional Dimension for Data Semantics*

- Traditionally, data semantics has dealt with "***what*** *(**when**)*" questions: objects, inter-relationships, groupings and constraints on them.
- But to achieve real understanding, humans also rely on "***how***" and especially "***why***" questions: ***How/Why were the data gathered?***
- Several conceptual modeling languages (e.g., Taxis, RML) already integrated description of data semantics with the processes involving them (transactions and workflows).
- Recent progress in *Requirements Engineering* has also shown how ***goals*** and (organizational) ***actors*** fit in.




# *Introduction to Tropos*

- Our aim here is to illustrate briefly the notation and process by which conceptual data (and process) design is carried out, and mention a few possible applications.
- Although the formalisms described here have been studied for years, their application to data semantics is novel, and highly speculative!!!




# *A Motivating Example*

- Consider the relational schema

  Student(<u>st#,</u>nm,addr,faculty,dept,degree)

  Course(<u>crs#,</u>crsname,instr,dept,yr,sem,size)

  Signup(<u>st#,crs#,yr,sem</u>,date)

  Payment(<u>st#,crs#,yr,sem</u>,date,amount)

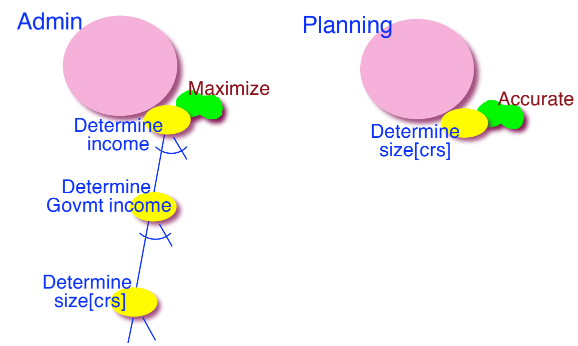
  where 'size' refers to the number of students enrolled for a particular section.
- We assume that the enrolment process consists of students enrolling for courses at the end of one term (e.g., May), and paying for them at the beginning of the nex term (e.g., September).




# *Answers and Intentions*

- Consider now a query, such as
  "Find the sizes of all courses in a particular year"
- Suppose a value is assigned to the "size" attribute by counting signups.
- The answer that we get back actually tells us how many students signed up for courses; presumably, some of them changed their minds, didn't pay and didn't take the course(s) they signed up for.
- The answer to the query is "optimistic", OK for a University reporting its enrolments to a Government, not OK for a University planning department that wants accurate answers.

## Actors, Goals and Softgoals



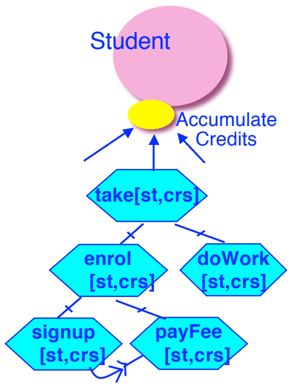



## From Goals to Tasks and Objects

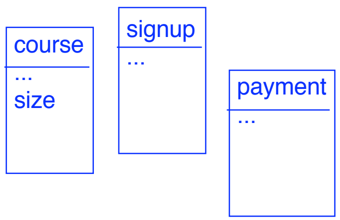
- Start with actors (stakeholders), their intentions/goals and the criteria (softgoals) by which they select among alternatives for fulfilling their goals;
- Analyze goals and softgoals; find alternative ways of fulfilling each goal; evaluate these with respect to softgoals; for each alternative define tasks to be performed and delegate these;
- Identify (concerned) objects for each goal; these are objects about which information is required in order to fulfill the goal; there may be alternative ways of defining these objects; each of these leads to a different database design.



## Task Analysis and Database Design

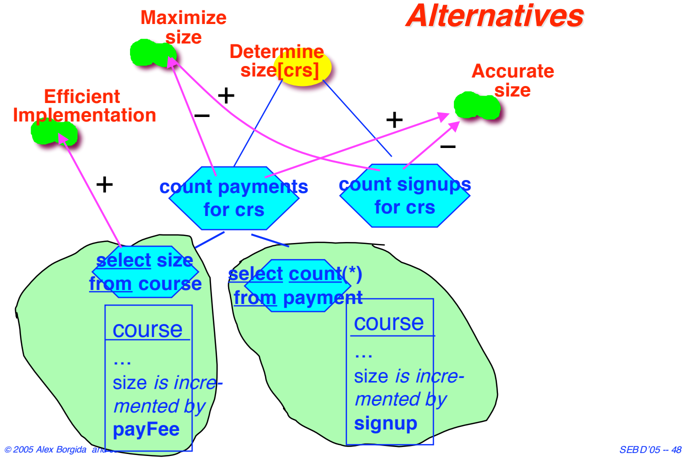



- Although this seems hand-wavy, there are ways to make it ***precise, translatable into formal (temporal) logic***






## Implementation Alternatives



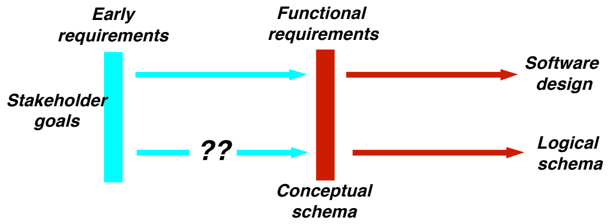
## Data Semantics - take 6

- A database schema is selected from a design space of alternative schemas on the basis of criteria (softgoals).
- The suitability of the database to answer a given query depends on how well its schema matches the softgoals behind the query.
- To answer queries whose intentions don't match those of the database, we need to keep around multiple schemas and their design history; we also need mappings amongst them.
- Note that this kind of data semantics relates to trust, I.e., "how much can I trust the data I get out of a particular database?"




## Related Ideas

- Hippocratic Databases [Agrawal02]
- "Why" questions in data provenance [Buneman]
- Data semantics based on workflows and processes.
- Early requirements engineering [Dardenne93]







## Discussion

- Data semantics is and will remain a core problem for databases with/without web technologies.
- Current research directions on the Semantic Web address the data semantics problem with an emphasis on expressiveness (of the modelling language used). Other directions are also worth exploring.
- Research on models and mappings will be critical in answering the challenge of the Semantic Web in the next decade.
- Ultimately, the meaning of data needs to be tied down to the processes that generated it, also the intentions of its designers and users (and how these correlate).




## *There is a lot more to Data Semantics than descriptions of things and relationships in ever-richer modelling languages!!*

# References


[Zelazny85] Zelazny, R.: 24 views of Mount Fuji. Isaac Asimov's Science Fiction Magazine **7** (1985).

[Codd70] Codd, E.: A relational model for large shared data banks. Communications of the ACM **13** (1970) 377–387.

[Abrial74] Abrial, J.R.: Data semantics. In Klimbie, Koffeman, eds.: Data Management Systems, North-Holland (1974).

[Chen75] Chen, P.: The entity-relationship model: Towards a unified view of data. In: Proceedings International Conference on Very Large Databases (VLDB'75), (1975).

[Berners-Lee99] Berners-Lee, T., Fischetti, M.: Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor. Harper, San Francisco (1999)

[Berners-Lee01] Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American (2001)

[W3C03] W3C: Web ontology language (owl) version 1.0, http://www.w3.org/tr/2003/wd-owl-ref-20030331 (2003)

[Ladkin97] Ladkin, P.: Abstraction and modeling, res. report RVS-Occ-97-04, University of Bielefeld, 1997; http://www.rvs.uni-bielefeld.de/publications/abstracts.html#AbsMod. (Technical report).

[Levesque84] Levesque, H.: Foundations of a functional approach to knowledge representation. Artificial Intelligence **23** (1984).

[Borgida95] Borgida, A.: Description logics in data management. IEEE Transactions on Knowledge and Data Engineering **7** (1995) 671–682.




# References


[Reiter84] Reiter, R.: Towards a logical reconstruction of relational database theory. In M. Brodie, J. Mylopoulos, J.S., ed.: On Conceptual Modelling, Springer-Verlag (1984) 191–233

[Miller00] Miller, R., Haas, L., Hernadez, M.: Schema mapping as query discovery. In: Proceedings International Conference on Very Large Databases (VLDB00), Cairo. (2000)

[Pottinger03] Pottinger, R., Bernstein, P.: Merging models based on given correspondences. In: Proceedings International Conference on Very Large Databases (VLDB03), Berlin, 826–873.

[Levy96] Levy, A., Rajaraman, A., Ordille, J.: Querying heterogeneous information sources using source descriptions. In: Proceedings International Conference on Very Large Databases (VLDB96), Mumbay,. (1996) 251–262

[Lenzerini02] Lenzerini, M.: Data integration: A theoretical perspective. In: Proceedings International Conference on Principles of Database Systems (PODS02). (2002) 233–246

[Friedman99] Friedman, M., Levy, A., Millstein, T.: Navigational plans for data integration. In: Proceedings National Conference on Artificial Intelligence (AAAI99). (1999) 67–73

[Madhavan03] Madhavan, J., Halevy, A.: Composing mappings among data sources. In: Proceedings International Conference on Very Large Databases (VLDB03), Berlin. (2003) 572–583

[Fagin04] Fagin, R., Kolaitis, P., Popa, L., Tan, W.C.: Composing schema mappings: Second-order dependencies to the rescue. In: Proceedings International Conference on Principles of Database Systems (PODS04). (2004) 83–94

[Borgida03] Borgida, A., Serafini, L.: Distributed description logics: Assimilating information from peer sources. Journal of Data Semantics (2003) 153–184




# References


[SemIntegration03] Proceedings of semantic integration workshop, at ISWC03, Sanibel Island, October 2003, http://ceur-ws.org/vol-82 (2003).

[Smith87] Smith, B.C.: The correspondence continuum, TR CSLI-87-71, Stanford University. Technical report (1987).

[Halevy03] Halevy, A., Ives, Z., Suciu, D., Tatarinov, I.: Schema mediation in peer data management systems. In: Proceedings International Conference on Data Engineering (ICDE03). (2003)

[Bernstein02] Bernstein, P., Giunchiglia, F., Kementsietsidis, A., Mylopoulos, J., Serafini, L., Zaihrayeu, I.: Data management for peer-to-peer computing: A vision. In: Proceedings SIGMOD WebDB Workshop. (2002) 89–94.

[Buneman01] Buneman, P., Khanna, S., Tan,W.C.: Why and where: A characterization of data provenance. Proceedings International Conference on Database Theory (ICDT01), 316–330.

[Calvanese01] Calvanese, D., De Giacomo, G., Lenzerini, M., Nardi, D., Rosati, R.: Data integration in data warehouses. Journal of Cooperative Information Systems **10** (2001) 237–271

[Velegrakis05] Velegrakis, Y., Miller, R., Mylopoulos, J.: Representing and querying data transformations. In: Proceedings International Conference on Data Engineering (ICDE05), to appear. (2005)

[An04] An, Y., Borgida, A., Mylopoulos, J.: Refining mappings from relational tables to ontologies. Proceedings VLDB Workshop on the Semantic Web and Databases (SWDB04), Toronto, Aug.2004.

[Mylopoulos80] Mylopoulos, J., Bernstein, P.,Wong, H.: A language facility for designing database-intensive applications. ACM Transactions on Database Systems **5** (1980) 185–207

[Barron82] Barron, J.: Dialogue and process design for interactive information systems using Taxis. Proceedings ACM SIGOA Conference on Office Information Systems, Philadelphia, 12–20.




# References


[Greenspan82] Greenspan, S., Mylopoulos, J., Borgida, A.: Capturing more world knowledge in the requirements specification. In: Proceedings International Conference on Software Engineering, (ICSE82), Kyoto. (1982) 225–235.

[King02] King, R., "The Story of Civilization and the Rubicon of Smart Data", Proceedings 28th International Conference on Very Large Databases (VLDB'02), Hong Kong, August 2002.

[Yu93] Yu, E.: Modeling organizations for information systems requirements engineering. In: Proceedings IEEE International Symposium on Requirements Engineering (RE93), San Diego, IEEE Computer Society Press. (1993) 34–41

[Castro02] Castro, J., Kolp, M., Mylopoulos, J.: Towards requirements-driven software development methodology: The tropos project. Information Systems **27** (2002) 365–389

[Dardenne93] Dardenne, A., van Lamsweerde, A., Fickas, S.: Goal-directed requirements acquisition. Science of Computer Programming **20** (1993) 3–50

[Bernstein00] Bernstein, P., Halevy, A., Pottinger, R., "A Vision for Management of Complex Models, *ACM SIGMOD Record 29(4)*, 2000

# More References


- [Ladkin97] Ladkin, P., "Abstraction and Modeling" Research Report RVS-Occ-97-04, University of Bielefeld, 1997; http://www.rvs.uni-bielefeld.de/publications/abstracts.html#AbsMod.
- [Miller00] Miller, R., Haas, L., Hernadez, M., "Schema Mapping as Query Discovery", Proceedings International Conference on Very Large Databases (VLDB'00), Cairo, September 2000.
- [Velegrakis03] Velegrakis, Y., Miller, R., Popa, L., "Mapping Adaptation Under Evolving Schemas", Proceedings International Conference on Very Large Databases (VLDB'03), Berlin, September 2003.
- [Velegrakis04] Velegrakis, Y., Miller, R., Mylopoulos, J., "Representing and Querying Data Transformations", (to appear.)