Mapping Data in Peer-to-Peer Systems: Semantics and Algorithmic Issues

Anastasios Kementsietsidis, Marcelo Arenas, Renée J. Miller
ACM SIGMOD International Conference on Management of Data 2003

Rolando Blanco

CS856 – Winter 2005

Overview

- Data Sharing in P2P systems
- Mapping table approach
- Conclusions/Discussion

Data Sharing in P2P

- Between autonomous structured data sources
- Data sources may use different schemas
- Sources may not be willing to share schema
- Data and schemas overlap or are related

Different schemas → semantic issues!

Example

[Berstein02]

Peer1: Toronto General Hospital (TGHDB)
Patients(TGH#, OHIP#, Name, FamilyDr, Sex, Age, …)
Treatments(TreatID, TGH#, Date, TreatDesc, PhysID)

Peer2: Dr Davis Family Dr (DavisDB)
Patients(OHIP#, FName, LName, Phone#, Sex, …)
Events(OHIP#, Date, Description)

- Patient visits hospital → load data from DavisDB
- Patient receives treatment → update Events at DavisDB
- A pharmacist db may update Events relation at DavisDB as well

How to implement data sharing?
Note global key OHIP# and similarities between attribute names

[Berstein02] Bernstein et al, "Data management for peer-to-peer computing: A vision". Workshop on the Web and Databases, WebDB 2002

Data Sharing

- Traditional Approach: Mediated schemas
  - "semantic tree"
  - global-as-view
  - local-as-view

Mediated Schema

TGHDB DavisDB

- P2P: Schema mappings

TGHDB
map(DavisDB)

DavisDB
map(TGHDB)
map(ClinicDB)

Victoria Walking Clinic
ClinicDB
map(DavisDB)

Graph of interconnected schemas form semantic network/topology

Variations [Tatarinov03]:

TGHDB — **Mediating Peer** (TGHDB schema, DavisDB schema) — DavisDB — **Mediating Peer** (DavisDB schema, ClinicDB schema) — ClinicDB

[Tatarinov03] Igor Tatarinov et al, "The Piazza Peer Data Management System". ACM SIGMOD Record Volume 32 , Issue 3 (September 2003)

Data Sharing

More Variations [Löser03]:

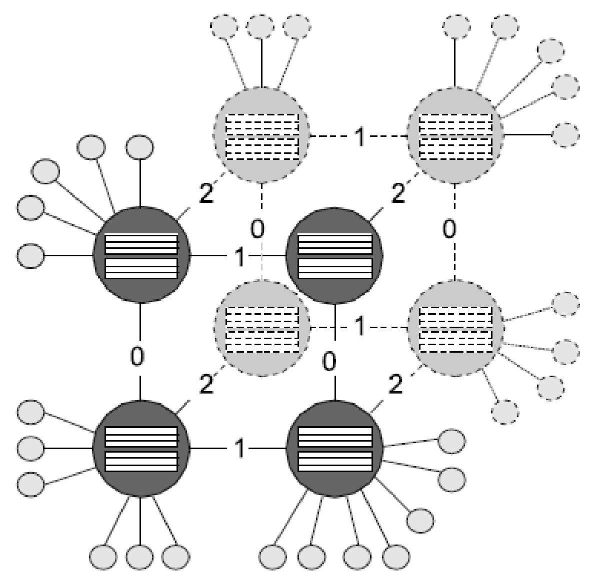


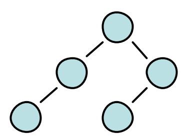
Figure 2: HyperCuP Super-peer Topology

Super-peers store schema mappings between super-peers,
and between super-peers and regular neighbour peers.

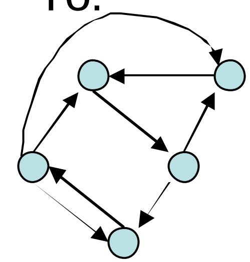
[Löser] Alexander Löser et al. “Information Integration in Schema-Based Peer-To-Peer Networks”
15th Conference on Advanced Information Systems Engineering (CAiSE'03)

"… The true novelty lies in the PDMS ability to exploit transitive relationships among peers' schemas …" [Halevy04]

From:



To:



[Halevy04] Alon Halevy et al. "Schema Mediation for Large-Scale Semantic Data Sharing", VLDB Journal, 2004.

How to create schema mappings

- Machine learning techniques: GLUE [Doan03]
  - Correspondences between taxonomies
  - “Similarity” between concepts based on probability distributions
- Gossiping [Aberer03]:
  - Propagation of queries toward nodes for which no direct mapping exists (“semantic gossiping”)
  - Analyse results and create/adjust mappings
  - Goal: incremental development of global agreement (semantics == form of agreement)
- On the fly (PeerDB [Ng03]):
  - No shared/distributed schema
  - Attributes have associated words
    - (e.g. desc → description, characteristics, features, functions)
  - Selection of candidate relations using IR techniques (flooding + TTL)
  - User confirms selections, system remembers.
- Don’t query, subscribe!

[Aberer03] Karl Aberer et al. The Chatty Web: Emergent Semantics Through Gossiping. Proceedings International WWW Conference 2003.
[Doan03] AnHai Doan, et al. Learning to Match Ontologies on the Semantic Web. VLDB journal, vol. 12, No. 4. 2003
[Ng03] Wee Siong Ng, et al. PeerDB: A P2P-based System for Distributed Data Sharing. 19th International Conference on Data Engineering 2003

Schema Mappings - Interesting Problems

- Schema composition
- Minimal composition
- Semantical redundancy
- Semantical partition

Are schema mappings enough?

<u>Peer1</u>: ABC Rentals (ABC)
ProdClasses(ProdClassID, ProdClassDesc, …)

<u>Peer2</u>: The Rental Store (TRS)
ProdGroups(ProdGroupID, ProdGroupDesc, …)

Customer of ABC Rentals wants to rent a product, ABC Rentals subrents from TRS if none available

Schema mapping:

ABC.ProdClassID ≅ TRS.ProdGroupID

ABC.ProdClassDesc ≅ TRS.ProdGroupDesc

ABC's ProdClasses

C001	"Air Compressors 2-4 CFM"
C002	"Air Compressors 5-7 CFM"
C003	"Air Compressors 8-10 CFM"

TRS's ProdGroups:

A001-31	"Air Comp. 2-6 CFM"
A001-32	"Air Comp. 7-10 CFM"

- Unless global ID, → different ID's imply different "meaning"
- Query: Customer wants air compressor of at least 5 CFM
- Assume no "capacity" column. This is a real-world example.

Data Mappings

ABC's ProdClasses
C001 "Air Compressors 2-4 CFM"
C002 "Air Compressors 5-7 CFM"
C003 "Air Compressors 8-10 CFM"

TRS's ProdGroups:
A001-31 "Air Comp. 2-6 CFM"
A001-32 "Air Comp. 7-10 CFM"

ProdClassID	ProdGroupID
C001	A001-31
C002	A001-32
C003	A001-32

- Represent knowledge, created/maintained by experts
- Semantically "richer"/more specific than schema mappings (but complementary)
- Note mapping is unidirectional (schema mapping is typically bi-directional)
- But still transitivity!
- Peer network logically defined by mappings among peers
- The way data sharing is done today in many applications
- Goals (paper's):
  (1) Specification of different semantics for data mappings
  (2) Inference/Validation of new data mappings

Definitions

Mapping Table $MP_{A \to B}$:

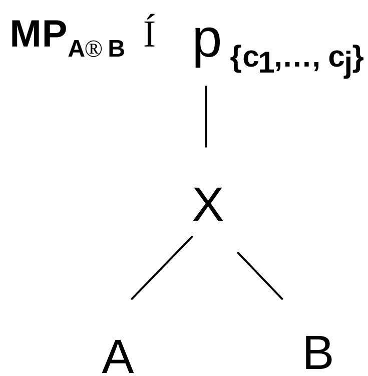
Given tables $A(a_1, a_2, \ldots, a_n)$, $B(b_1, b_2, \ldots, b_m)$, $MP_{A \to B}(c_1, \ldots, c_i, c_{i+1}, \ldots, c_j)$ with $\{c_1, \ldots, c_i\} \subseteq \{a_1, \ldots, a_n\}$ and $\{c_{i+1}, \ldots, c_j\} \subseteq \{b_1, \ldots, b_m\}$, then

$MP_{A \to B}$ is a mapping table from A to B if:

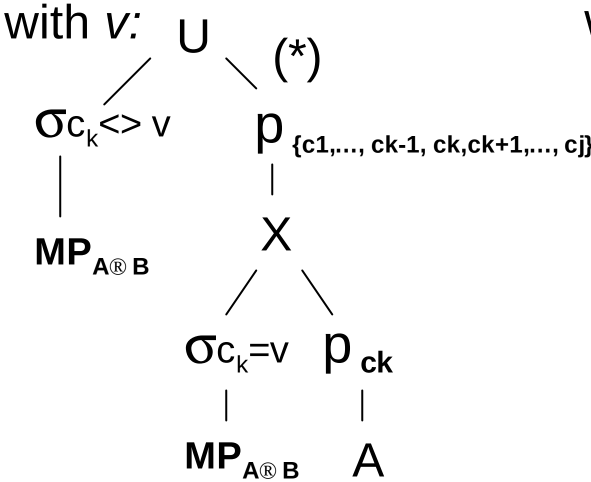
$\forall\, t \in MP_{A \to B}$: $t[c_k]$ = value in $dom(a_l)$, or *v* (variable), or *v* – subset($dom(a_l)$)
(*assuming* c_k *corresponds to* a_l)

Restriction!: v can appear one or more times in one and only one tuple of $MP_{A \to B}$

Is this definition sound?: assuming v can have values in $dom(a_l)$



with *v:*



with *v – subset(dom(a_l)):*

$subset(dom(a_l)) = \{val_1, val_2 \ldots val_z\}$

(*) $\sigma\, a_l <> val_1 \wedge a_l <> val_2 \wedge \ldots a_l <> val_z$

More definitions

What about values of $p_{\{c1,\ldots,\ ci\}}(A)$ not in $p_{\{c1,\ldots,\ ci\}}(\mathbf{MP}_{A®\ B})$?

- Closed world semantics:
  - data cannot be associated to values in B
- Open world semantics:
  - data can be associated to any value in B

    $\cong v - \{p_{\{c_w\}}(\mathbf{MP}_{A®\ B})\}$ with c_w attribute of B
  - represents partial knowledge
- Tuple satisfies mapping table:

  Given a mapping $MP_{A\to B}(c_1,\ldots, c_i, c_{i+1},\ldots, c_j)$, a tuple t with attributes $\{r_1, \ldots, r_w\} \supseteq \{c_1, \ldots, c_j\}$ satisfies $MP_{A\to B}$ if $t[c_1,\ldots, c_i, c_{i+1},\ldots, c_j] \in MP_{A\to B}$
- Mapping constraint:

  Assume attribute sets $A' = \{c_1, \ldots, c_i\}$, $B' = \{c_{i+1}, \ldots, c_j\}$ and mapping $MP_{A\to B}(c_1,\ldots, c_i, c_{i+1},\ldots, c_j)$, μ is a mapping constraint over A' U B' (represented $\mu : A' \xrightarrow{MP} B'$), from A' to B', if for every tuple t with attributes $\supseteq \{c_1,\ldots, c_i, c_{i+1},\ldots, c_j\}$, t satisfies μ, ($t \models \mu$) if $t[(c_1,\ldots, c_i, c_{i+1},\ldots, c_j] \in MP_{A\to B}$.
- Relation satisfies mapping constraint: $R \models \mu$ (R *satisfies* μ)

  A relation R with attributes $\{r_1, \ldots, r_w\} \subseteq \{c_1, \ldots, c_j\}$ satisfies μ ($R \models \mu$) if for every tuple t in t, $t \models \mu$

More definitions (almost done!)

- Extension of a mapping constraint (ext(μ)):

  μ with all variable and variable expressions instantiated
- Mapping constraint formula f:

  Built from mapping constraints plus $\neg$, $\vee$, $\wedge$ such that

  if f = μ then t|= f iff t μ

  if f = $\neg$ μ then t|= f iff not t |= μ (*remember this one*)

  if f = f1 $\vee$ f2 then t |= f iff t |= f1 or t |= f2

  if f = f1 $\wedge$ f2 then t |= f iff t |= f1 and t |= f2
- Given a set of formulas Σ, t |= Σ iff t |= f for every f in Σ

Inference/Consistency Problem

- *Inference problem*: Given a set of formulas $\sum$, can f be deduced from $\sum$ ($\sum$ |= f)?
  - Deductive calculus: prove $\neg \$t : t \mid= \sum U \{\neg f\}$ (*consistency problem*: can anything be deduced from $\sum$?)
  - Note if you have an algorithm to resolve consistency problem, then you can use it to resolve inference problem as well.

One more definition

- Cover of a set of constraints:
  – Consider semantic path P_1, ... P_n with set of attributes A_i for peer P_i. Assume $\sum$ is the set of mapping constraints in P_1, ... P_n. μ is the cover of a set of constraints $\sum$ iff:
    $\forall \mu' \; A_1 \xrightarrow{MP'} A_n : \sum \models \mu'$ iff $ext(\mu) \subseteq ext(\mu')$
  – Argument:
    - If an algorithm can compute cover μ then inference consistency problem is solved (since $\mu <> \emptyset$)
    - To show that a mapping constraint μ' can be inferred from $\sum$ we just need to show $ext(\mu) \subseteq ext(\mu')$
  – Are the arguments valid, what type of things can be shown to be deduced from $\sum$?

Cover over set of constraints - Issues

Consider relations A(x), B(y), C(z) such that A(x) = {1, 2}, B(y) = {a, b}, C(z) = {a', b', c', d', e'} and Σ = {MP1, MP2}:

MP1 ($A \rightarrow B$)

x	y
1	a
2	b

MP2 ($B \rightarrow C$)

y	z
a	a'
a	b'
b	c'

μ cover for Σ

x	z
1	a'
1	b'
2	c'

Let μ' $A \rightarrow C$ be:

x	z
1	a'
1	b'
2	c'
2	d'

$\neg\mu'$

x	z
1	d'
1	e'
2	a'
2	b'
2	e'

Note: ext(μ) $\subseteq$ ext(μ'), then according to previous arguments, $\mu' \models \Sigma$
Also note Σ U {$\neg\mu'$} is empty, then according to theory μ' is inferable from Σ. Shouldn't only data that follows the mapping constraints in Σ be inferable? Presented theory accepts as inferable something that generates *new* data not considered by the mapping constraints.

Cover over set of constraints - Issues

- Better to write?:
  - μ is a cover of Σ if:
    - (1) $\forall t, t \in ext(\mu)$: t can be deduced from Σ $(t \models \Sigma)$
    - (2) $\forall \mu', \mu' \; A_1 \xrightarrow{MP'} A_n$: $\Sigma \models \mu'$ iff $ext(\mu') \subseteq ext(\mu)$, and
      $\forall t, t \in ext(\mu')$: t can be deduced from Σ $(t \models \Sigma)$
  - Then:
    - Inference: $ext(\mu') \subseteq ext(\mu)$, and $ext(\mu')$ not empty
    - Consistency: μ exists
  - Note this guarantees that data non-deducible from Σ is not considered inferable
  - Issue: a method to decide if $t \models \Sigma$ needs to be provided

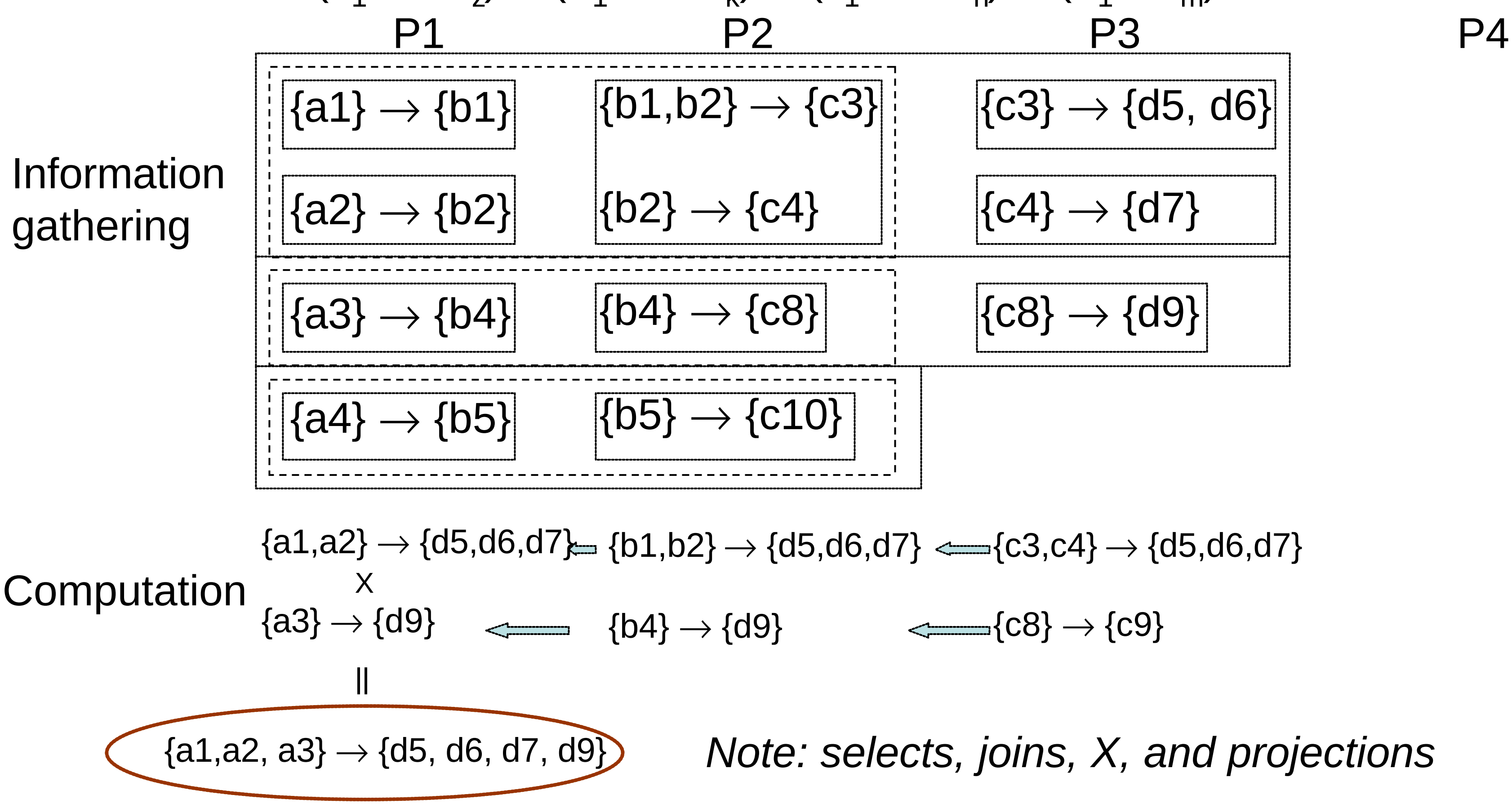
Algorithm

- Restrictions:
  - Number of peers in path $\rightarrow$ assumed small
  - Number of mapping constraints $\rightarrow$ fixed to a maximum per peer
  - Number of rows in each mapping $\rightarrow$ no restrictions
  - Number of columns in each constraint $\rightarrow$ to a max per mapping constraint
- Input:
  - Σ set of mapping constraints form path P_1 ... P_n
  - Sets A_1 and A_n with A_1 subset if attributes of mappings in P_1, A_n subset of attributes of mappings in P_n
- Output:
  - μ, cover of Σ for attribute sets A_1 and A_n ($A_1 \xrightarrow{MP} A_n$)
- Complexity: polynomial on input

Algorithm

- Goals:
  - Distribute computation
  - Stream results (first row optimisation?)

$A(a_1,\ldots, a_z), B(b_1, \ldots, b_k), C(c_1, \ldots, c_n), D(d_1 \ldots d_m)$



Note: selects, joins, X, and projections

Experimental results

- Six biological dbs (G, H, L, M, S, U). 11 mapping tables, seven paths:

  H → L → G → S → M

  H → L → G → M

  H → S → M

  H → L → U → S → M

  H → L →M

  H → G → S → M

  H → G → M

- 13,000 avg mappings per table

Path	Length	Computed Mappings	New Mappings	Time (in secs)
1	5	6163	927	16.00
2	4	6193	11	15.00
3	3	9334	543	22.00
4	3	8704	10	22.00
5	3	6525	64	10.00
6	5	3276	397	26.00
7	4	8813	24	23.00

Figure 10: Inferred mappings

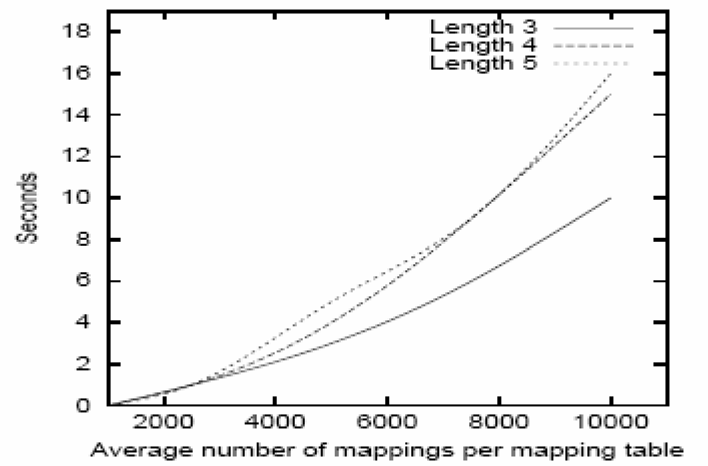


Figure 11: Scalability in path and table size

Experimental Results

- 3 peers
- Multi-attribute constraints
- Use of variables
- Synthetically generated mappings

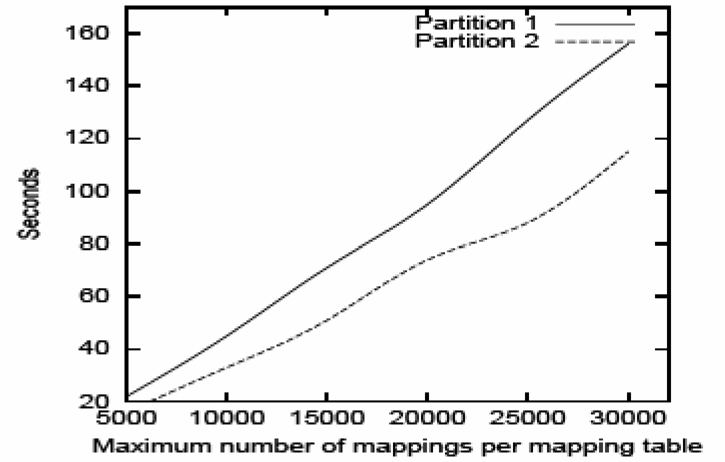


Figure 12: Per partition execution time

Conclusions

- Mapping tables semantically more precise than mapping schemas
- Formal presentation of mapping tables
- Algorithm to compute cover for a semantic path
- More recent work:
  - Data coordination: triggers (event-condition-action) to enforce mapping *expressions* (Hyperion Project [Arenas03, Tasos03, Tasos04])
  - Query translation based on data mappings

Comments/Discussion

- Notational issues and use of math formalisms
- Why deductive calculus and not relational calculus?
  - In VLDB04 “Data Query Through Query Translation in Autonomous Sources” [Arenas04], use of relational calculus (“Example 6, Definition 7” numbering still there though!)
- Not clear formal presentation is complete (consider definition in section 6)
- Poor description of algorithm
- Minimal experimentation
- Caching. Unable to comment from information in the paper (Buffer?)
- Clear improvements to algorithm not addressed (consider A $\rightarrow$ B $\rightarrow$ C with mappings in A being the most restrictive)

Comments/Discussion

- Applicability:
  - Maintenance of data mappings
  - Length of semantic paths
  - Types of queries

References

- [Aberer03] Aberer, Karl and Cudre-Mauroux, Philippe and Hauswirth, Manfred. The Chatty Web: Emergent Semantics Through Gossiping. Proceedings International WWW Conference 2003
- [Arenas03] Marcelo Arenas, Vasiliki Kantere, Anastasios Kementsietsidis, Iluju Kiringa, Renée J. Miller, John Mylopoulos. The Hyperion Project: From Data Integration to Data Coordination. In SIGMOD Record, Special Issue on Peer-to-Peer Data Management, 32(3):53-58, 2003
- [Bernstein2002] Bernstein, P.A., Giunchiglia, F., Kementsietsidis, A., Mylopoulos, J., Serafini, L., Zaihrayeu, I.: Data management for peer-to-peer computing: A vision. In: Workshop on the Web and Databases, WebDB 2002
- [Doan03] AnHai Doan, Jayant Madhavan, Robin Dhamankar and Alon Halevy. Learning to Match Ontologies on the Semantic Web. VLDB journal, vol. 12, No. 4. 2003
- [Halevy04] Alon Halevy et al. "Schema Mediation for Large-Scale Semantic Data Sharing", VLDB Journal, 2004.
- [Löser03] Alexander Löser, Wolf Siberski, Martin Wolpers, Wolfgang Nejdl. Information Integration in Schema-Based Peer-To-Peer Networks. The 15th Conference on Advanced Information Systems Engineering (CAiSE'03), Klagenfurt/Velden, Austria, June 2003
- [Ng03] Wee Siong Ng, Beng Chin Ooi, Kian-Lee Tan and Ao Ying Zhou. PeerDB: A P2P-based System for Distributed Data Sharing. 19th International Conference on Data Engineering 2003
- [Tasos03] Anastasios Kementsietsidis, Marcelo Arenas, Renée J. Miller. Managing Data Mappings in the Hyperion Project. In Proceedings of the International Conference on Data Engineering (ICDE) 2003, pages 732-73
- [Tasos04] Anastasios Kementsietsidis and Marcelo Arenas. Data Sharing Through Query Translation in Autonomous Sources. In Proceedings of the International Conference on Very Large Data Bases (VLDB) , September 2004.
- [Tatarinov03] Igor Tatarinov et al, “The Piazza Peer Data Management System”. ACM SIGMOD Record Volume 32 , Issue 3 (September 2003)