

Structured and Semi- Structured Data Integration

Antonella Poggi

Joint “European Label” research doctoral thesis

Università di Roma “La Sapienza”

Advisor: Maurizio Lenzerini

Université de Paris Sud – INRIA

Advisor: Serge Abiteboul

Motivations

- Data integration relevance
 - age of the internet connectivity, and globalization (e.g. leading to mergers of companies)
 - still a challenge: foreseen huge investments on the topic
- Current data integration needs
 - serious use of (semantic) technologies for (semantic) integration
 - clear separation between an integrated virtual data layer offered to the user and the data sources themselves
 - ontology-based (structured)
 - XML-based (semi-structured)

Outline(1)

Thesis contributions

- Formalized approach to semantic data integration
 - from read-only to **write-also** Data Integration Systems (DIS)
- Read-only ontology-based data integration
 - DIS framework based on DL-Lite_A
 - novel mapping language
 - solves “**impedance mismatch**” between **values** in the data layer and **objects** at the intensional layer
 - query answering service
 - **LOGSPACE** algorithm (in the size of the data)
- Toward write-also ontology-based data integration
 - instance-level **ontology updates**
 - **PTIME** algorithm for DL-Lite_A updates

Outline(2)

Thesis contributions

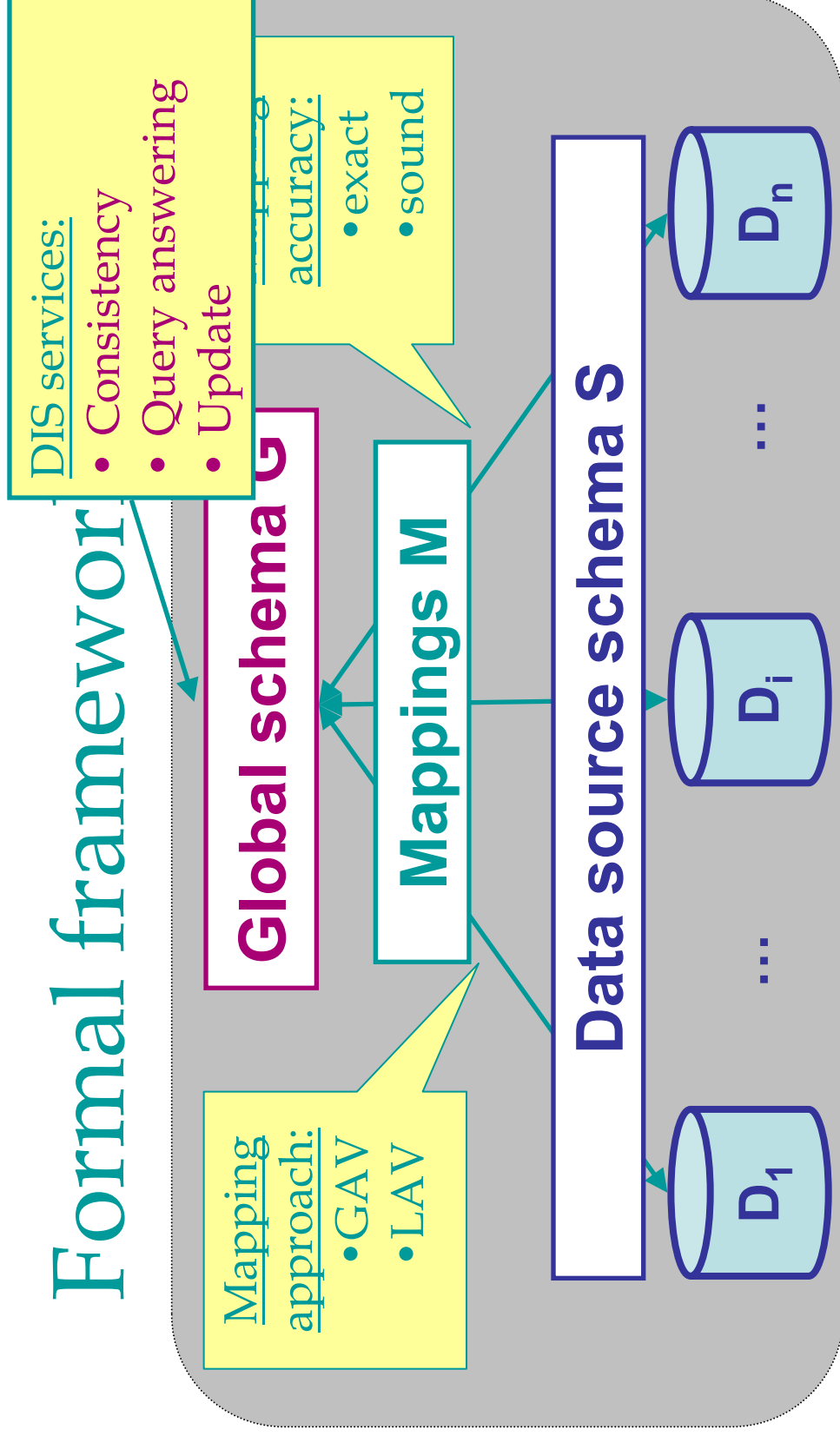
- Read-only XML-based data integration
 - logical formal DIS framework based on XML
 - definition of an identification function to address the “node identity” issue
 - consistency and query answering services
 - NP-hard consistency
 - coNP-hard query answering
 - PTIME algorithms under particular restrictions

Outline(1)

Thesis contributions

- Formalized approach to semantic data integration
 - from read-only to **write-also** Data Integration Systems (DIS)
- Read-only ontology-based data integration
 - DIS framework based on DL-Lite_A
 - novel mapping language
 - solves “impedance mismatch” between values in the data layer and objects at the intensional layer
 - query answering service
 - LOGSPACE algorithm (in the size of the data)
- Toward write-also ontology-based data integration
 - instance-level ontology updates
 - PTIME algorithm for DL-Lite_A updates

Formal framework



- Data sources schema: heterogeneous, autonomous
- **Global schema**: integrated, virtual schema
- **Mappings**: relationships between sources and global schema:
 $Q_{\mathcal{G}}(\text{query over } \mathcal{G}) \leftarrow Q_{\mathcal{S}}(\text{query over } \mathcal{S})$

Outline(1)

Thesis contributions

- Formalized approach to semantic data integration
 - from read-only to write-also Data Integration Systems (DIS)
- Read-only ontology-based data integration
 - DIS framework based on DL-Lite_A
 - novel mapping language
 - solves “impedance mismatch” between values in the data layer and objects at the intensional layer
 - query answering service
 - LOGSPACE algorithm (in the size of the data)
- Toward write-also ontology-based data integration
 - instance-level ontology updates
 - PTIME algorithm for DL-Lite_A updates

Ontology-based DIS

Partial state of the art

Data model	Constraints	Mapping approach	Mapping accuracy	Example
Ontology	Inclusions (CARIN)	LAV	Sound	Inf.Man[KirkEtAl., '95]
Ontology	Acyclic inclusions (CARIN)	GAV	Sound	PICSEL[GoasdoueEtAl., '00]
Relational	Functional, Inclusions, ...	GAV	Sound	IBIS[CaliEtAl., '02], DIS@DIS[CaliEtAl., 04], INFOMIX[LeoneEtAl., 05]
Object-oriented	Keys	LAV	Sound	STVX[Amann, '00]

Our approach:

- distinguishes values from objects
- adds notable reasoning capabilities
- LOGSPACE

Ontology	Functional, inclusions	GLAV/GAV	Sound
----------	------------------------	----------	-------

Outline(1)

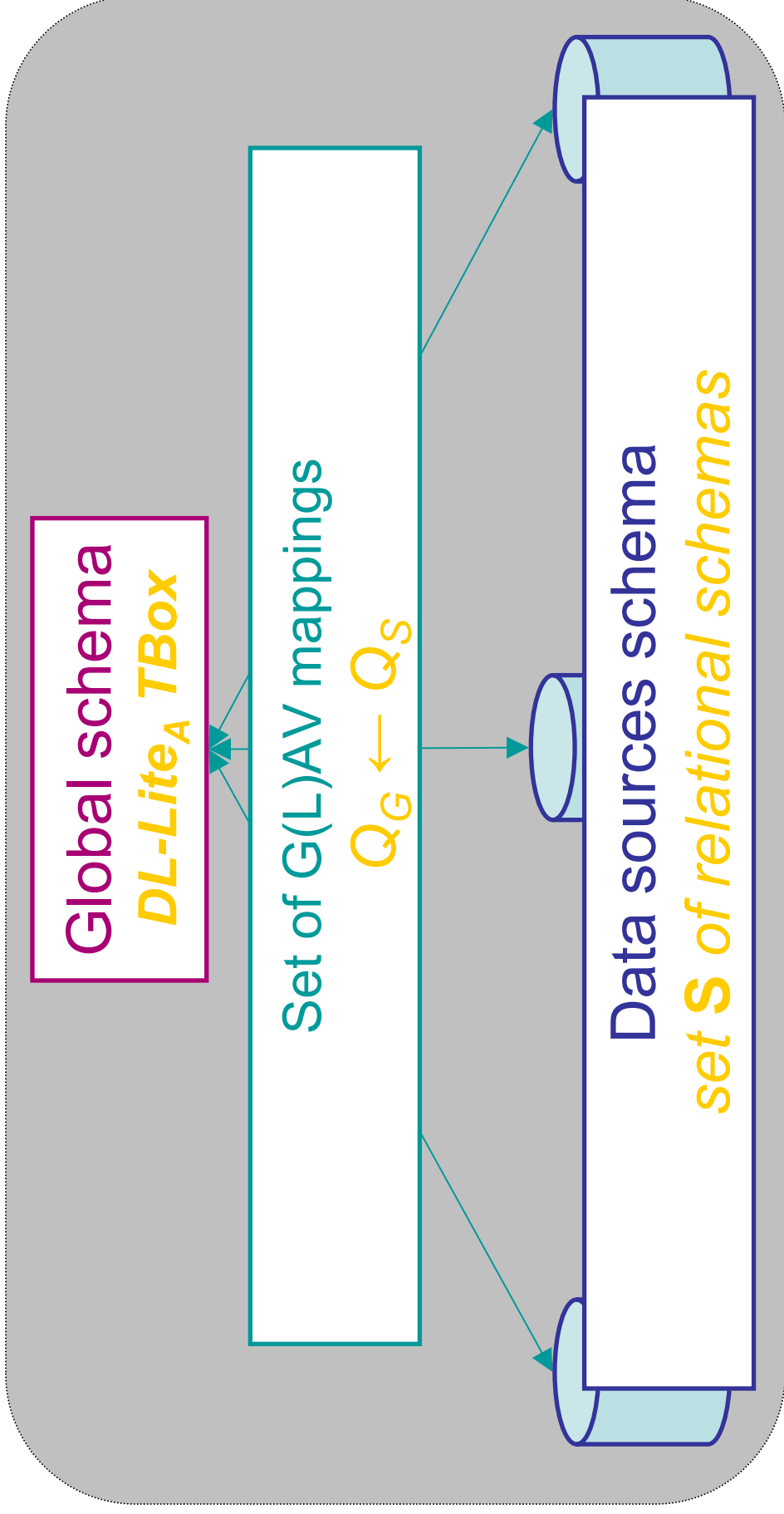
Thesis contributions

- Formalized approach to semantic data integration
 - from read-only to write-also Data Integration Systems (DIS)
- Read-only ontology-based data integration
 - DIS framework based on DL-Lite_A
 - novel mapping language
 - solves “impedance mismatch” between values in the data layer and objects at the intensional layer
 - query answering service
 - LOGSPACE algorithm (in the size of the data)
- Toward write-also ontology-based data integration
 - instance-level ontology updates
 - PTIME algorithm for DL-Lite_A updates

DL-Lite_A

- Description Logics (DLs)
 - basis of current **ontology language standard** (OWL [W3C, '04])
 - TBox (intensional level)
 - ABox (extensional level)
- Extension of DL-Lite [CalvaneseEtAl, '05]
 - distinguishes values (numbers, strings, dates, etc) from objects
 - **first one** to allow **roles to be qualified by attributes**
- Computational properties:
 - main reasoning services (subsumption, ...) can be reduced to (conjunctive) **query answering**
 - (conjunctive) query answering: **first-order query rewritable**
 - **LOGSPACE** in the size of the data
 - **PTIME** in the size of the TBox
 - computational complexity **delegated** to DBMS managing the data layer
 - allows to exploit **current DBMS optimized features**

DL-Lite_A DIS framework



Outline(1)

Thesis contributions

- Formalized approach to semantic data integration
 - from read-only to write-also Data Integration Systems (DIS)
- Read-only ontology-based data integration
 - DIS framework based on DL-Lite_A
 - novel mapping language
 - solves “impedance mismatch” between values in the data layer and objects at the intensional layer
 - query answering service
 - LOGSPACE algorithm (in the size of the data)
- Toward write-also ontology-based data integration
 - instance-level ontology updates
 - PTIME algorithm for DL-Lite_A updates

Mappings and the impedance mismatch problem

Values
Objects

tempEmp \sqsubseteq employee
 manager \sqsubseteq employee
 \exists WORKS-FOR \sqsubseteq project
 \exists WORKS-FOR \sqsubseteq employee
 ρ (Name) \sqsubseteq xsd:string
 manager(**pers(20903)**)
 Name(**pers(20903),Palmer**)

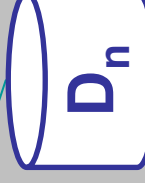
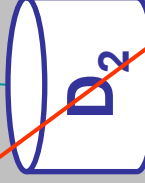
tempEmp \sqsubseteq $\exists \delta$ (until)
 δ (until) \sqsubseteq WORKS-FOR
 (funcnt until)
 ρ (until) \sqsubseteq xsd:date
 manager \sqsubseteq $\neg \exists \delta$ (until)
 manager(**mgr(Rossi,31/07/73)**)
 Name(**mgr(Rossi,31/07/73),Rossi**)

manager(**pers(s)**),Name(**pers(s),c**) \leftarrow D₁(c, n, d), D₂(s, c)
 manager(**mgr(c,d)**), Name(**mgr(c,d)**, c) \leftarrow D₁(c, n, d), $\neg \exists$ s.D₂(s, c)

Values

...(Palmer, John, 31/07/76),
 ...(Rossi, Mario, 22/08/73),...

...(20903, Palmer)...



...

Results on the mapping language

- use of **object terms** built from values by applying suitable (Skolem) functions
 - inspired by work done on object-based data manipulation languages which provide mechanisms for explicit creation of object identifiers [AbiteboulEtAl., '88][HullEtAl., '90]
- solves to the “**impedance mismatch**” between **values** in the data layer and **objects** at the intensional layer
- provides means to **reconcile the identity** of occurrences of the same object in different relational databases

Outline(1)

Thesis contributions

- Formalized approach to semantic data integration
 - from read-only to write-also Data Integration Systems (DIS)
- Read-only ontology-based data integration
 - DIS framework based on DL-Lite_A
 - novel mapping language
 - solves “impedance mismatch” between values in the data layer and objects at the intensional layer
 - query answering service
 - LOGSPACE algorithm (in the size of the data)
- Toward write-also ontology-based data integration
 - instance-level ontology updates
 - PTIME algorithm for DL-Lite_A updates

Approach naïve

SQL query

Answers
Palmer, Rossi

Conjunctive query

$q(x) :- \text{employee}(y), \text{Name}(y, x)$

$\text{tempEmp} \sqsubseteq \text{employee}$
 $\text{manager} \sqsubseteq \text{employee}$
 $\exists \text{WORKS-FOR} \sqsubseteq \text{project}$
 $\exists \text{WORKS-FOR} \sqsubseteq \text{employee}$
 $\rho(\text{Name}) \sqsubseteq \text{sd:string}$

$\text{tempEmp} \sqsubseteq \exists \delta(\text{until})$

FO query rewriting

$q(x) :- \text{employee}(y), \text{Name}(y, x)$
 $q(x) :- \text{tempEmp}(y), \text{Name}(y, x)$
 $q(x) :- \text{manager}(y), \text{Name}(y, x)$
 $\text{WORKS-FOR}(y, z), \text{Name}(y, x)$
 $\text{until}(y, z, w), \text{Name}(y, x)$

PTIME in the size of the data!

$\text{manager}(\text{pers}(s, \underline{c}), \text{Name}(\text{mgr}(s, \underline{c}), \underline{d}), D_2(s, \underline{c}))$
 $\text{manager}(\text{mgr}(c, \underline{d}), \underline{c}) \leftarrow D_1(\underline{c}, n, \underline{d}), \neg \exists D_2(s, \underline{c})$

Materialized ABox (by evaluating mappings)

$\text{manager}(\text{pers}(20903))$
 $\text{Name}(\text{mgr}(20903), \text{Palmer})$

$\text{manager}(\text{mgr}(\text{Rossi}, 22/08/73))$

$\text{mgr}(\text{Rossi}, 22/08/73), \text{Rossi}$

D₁

D₂

D_n

...(Palmer, John, 31/07/76),
 ...(Rossi, Mario, 22/08/73),...

...(20903, Palmer)...

Rewriting approach

Answers

Palmer, Rossi

Conjunctive query

$q(x) :- \text{employee}(y), \text{Name}(y,x)$

tempEmp :- employee
 manager :- employee
 WORKS-FOR :- project
 WORKS-FOR :- employ
 p(Name) :- sd:string

tempEmp $\equiv \exists \delta(\text{un})$

FO rewriting

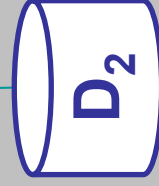
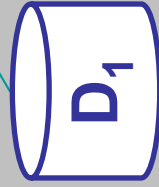
$q(x) :- \text{employee}(y), \text{Name}(y,x)$
 $q(x) :- \text{tempEmp}(y), \text{Name}(y,x)$
 $q(x) :- \text{manager}(y), \text{Name}(y,x)$
 $\neg \text{FOR}(y,z), \text{Name}(y,x)$
 $(y,z,w), \text{Name}(y,x)$

LOGSPACE in the size of the data!

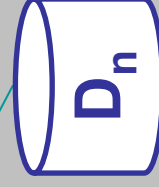
manager(pers(s), Name(pers(s)), Name(mgr(c,d)), Name(mgr(c,d)))

SQL query

$q(c) :- D_1(c,n,d), D_2(s,c)$
 $q(c) :- D_1(c,n,d), \neg D_2(s,c)$



...



...(Palmer, John, 31/07/76),
 ...(Rossi, Mario, 22/08/73),...

...(20903, Palmer)...

Results on DL-Lite_A query answering

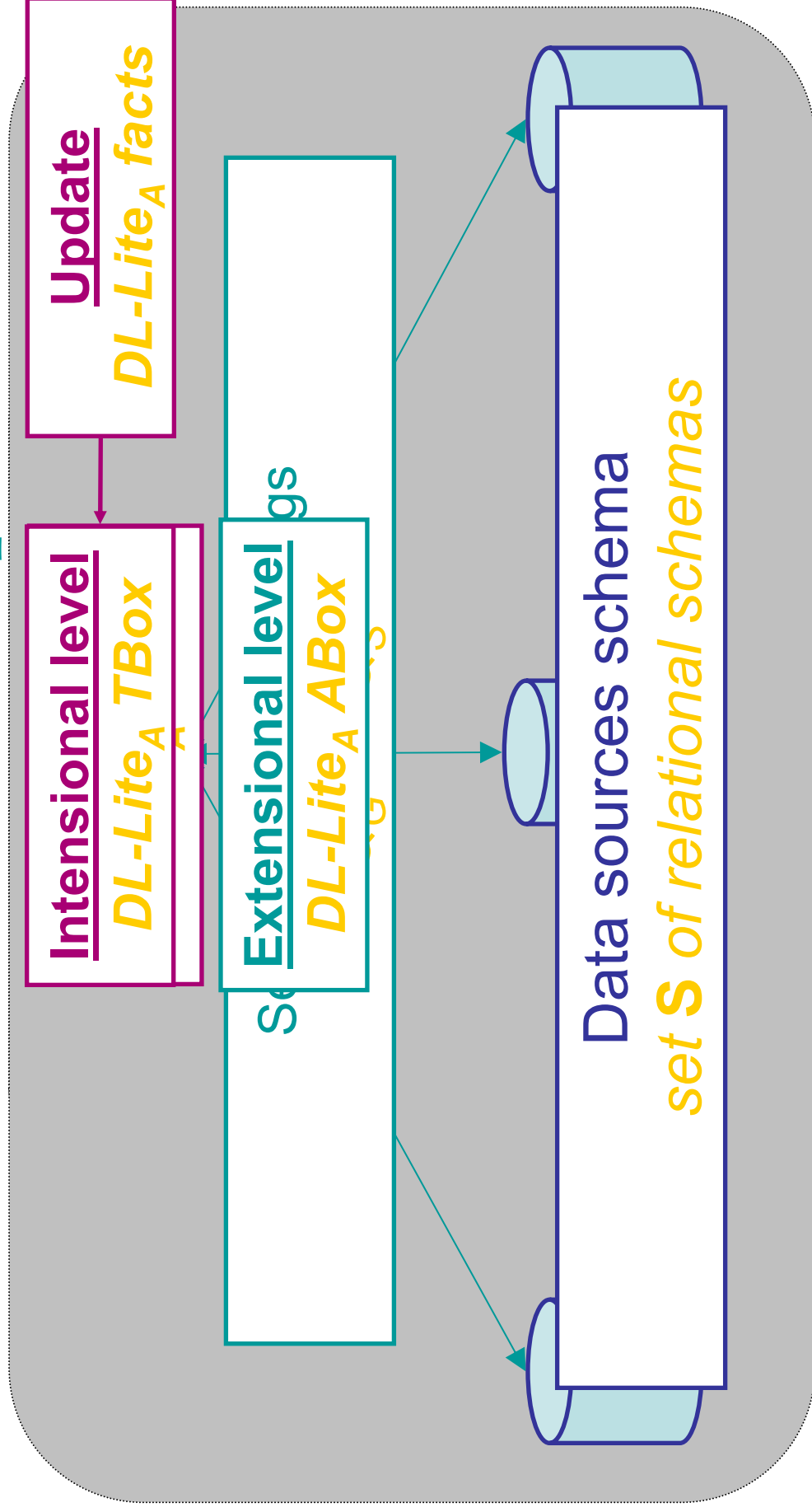
- Query answering algorithm
 - sound and complete
 - LOGSPACE in the size of the data
 - requires using techniques from partial evaluation in logic programming
 - exploits the underlying optimized DBMS technology
 - under implementation as part of the QuOnto ontology management system
 - first to provide conjunctive query answering sound and complete
 - first to address the mismatch between values and objects

Outline(1)

Thesis contributions

- Formalized approach to semantic data integration
 - from read-only to write-also Data Integration Systems (DIS)
- Read-only ontology-based data integration
 - DIS framework based on DL-Lite_A
 - novel mapping language
 - solves “impedance mismatch” between values in the data layer and objects at the intensional layer
 - query answering service
 - LOGSPACE algorithm (in the size of the data)
- Toward write-also ontology-based data integration
 - instance-level ontology updates
 - PTIME algorithm for DL-Lite_A updates

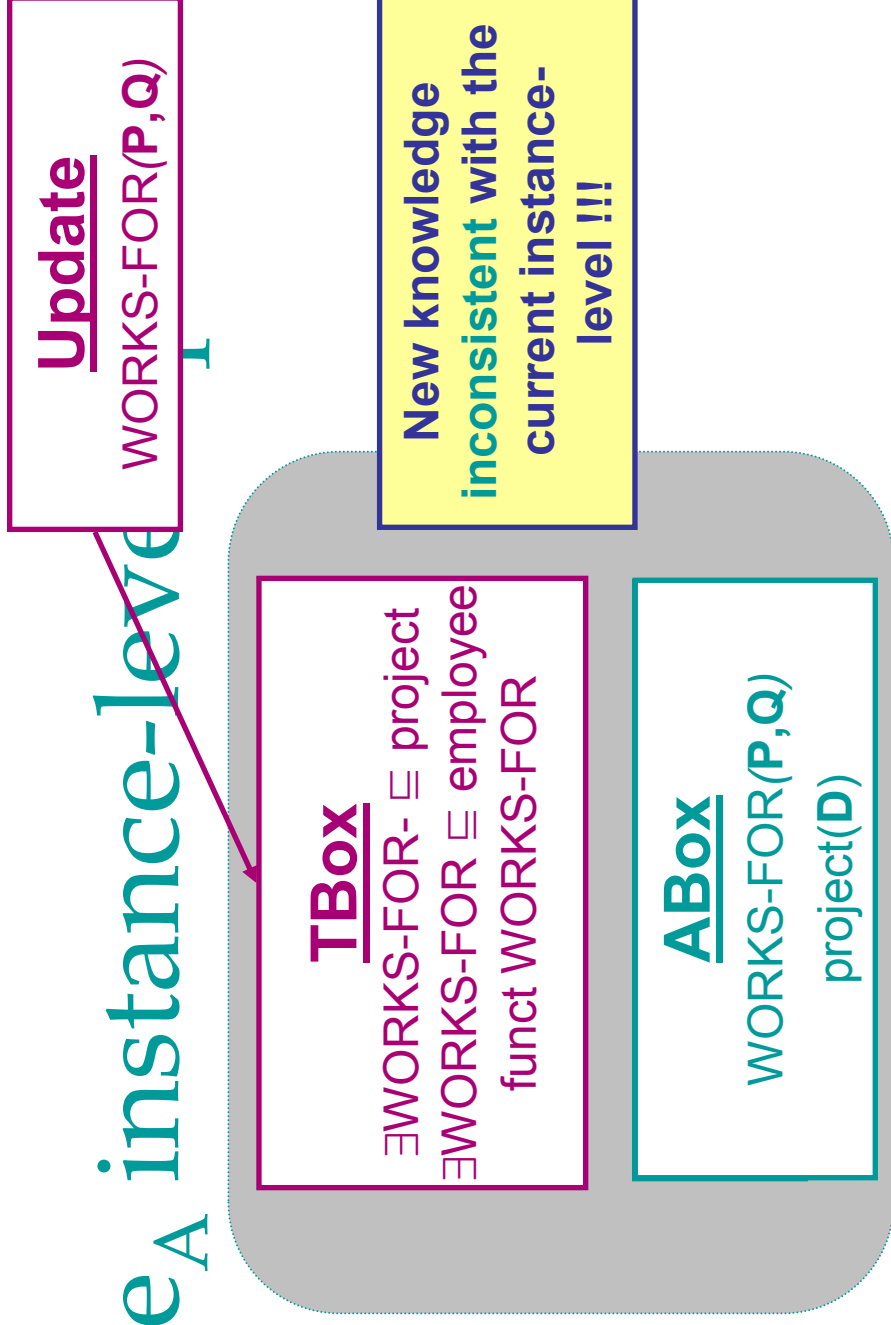
DL-Lite_A ontology instance-level update



Instance-level ontology update

- **Challenge:** how to react to the case where the update is inconsistent with the current knowledge
- Updates: an “old” problem
 - **knowledge base** general context: several classical results
 - [FaginEtAl, '83]: based on minimal syntactic theory change
 - [Winslett, '90]: based on minimal mutilation of models
 - related to **view update problem** [e.g. BancilhonEtAl, '81]
- ≠ **ontology** context: only few recent results
 - [Liu *et al.*, '06]: updates **not expressible** for several standard DLs
- **Novelty: ontology instance-level update**
 - the TBox is invariant
 - the ABox resulting from the update must be consistent with it

DL-Lite_A instance-level



- update result:
- TBox unchanged
 - asserted membership satisfied
 - minimally changed semantics

Results on DL-Lite_A ontology instance-level update

- Updates always **expressible** in DL-Lite_A
- Update algorithm
 - idea
 1. infer everything that we can from the update
 2. remove from the original ABox assertions that are contradicted
 3. leave behind implicates of them
 - **sound** and **complete**
 - **P**TIME in the size of the ABox
 - under **implementation** as service provided by the QuOnto ontology management system
 - **first** to provide update sound and complete

Outline(2)

Thesis contributions

- Read-only XML-based data integration
 - logical formal DIS framework based on XML
 - definition of an identification function to address the “node identity” issue
 - consistency and query answering services
 - NP-hard consistency
 - coNP-hard query answering
 - PTIME algorithms under particular restrictions

XML-based DIS

State of the art

Data model	Constraints	Mapping approach	Mapping accuracy	Example
XML	DTD	LAV	Sound	Agora[Manolescu, '01]
XML	XML Schema types, functional	LAV	Sound	[PopaEtAl., '04]

Our approach:

- expressive global schema (DTD + keys constraints)
- consistency and query answering computational complexity issues
- algorithms sound and complete

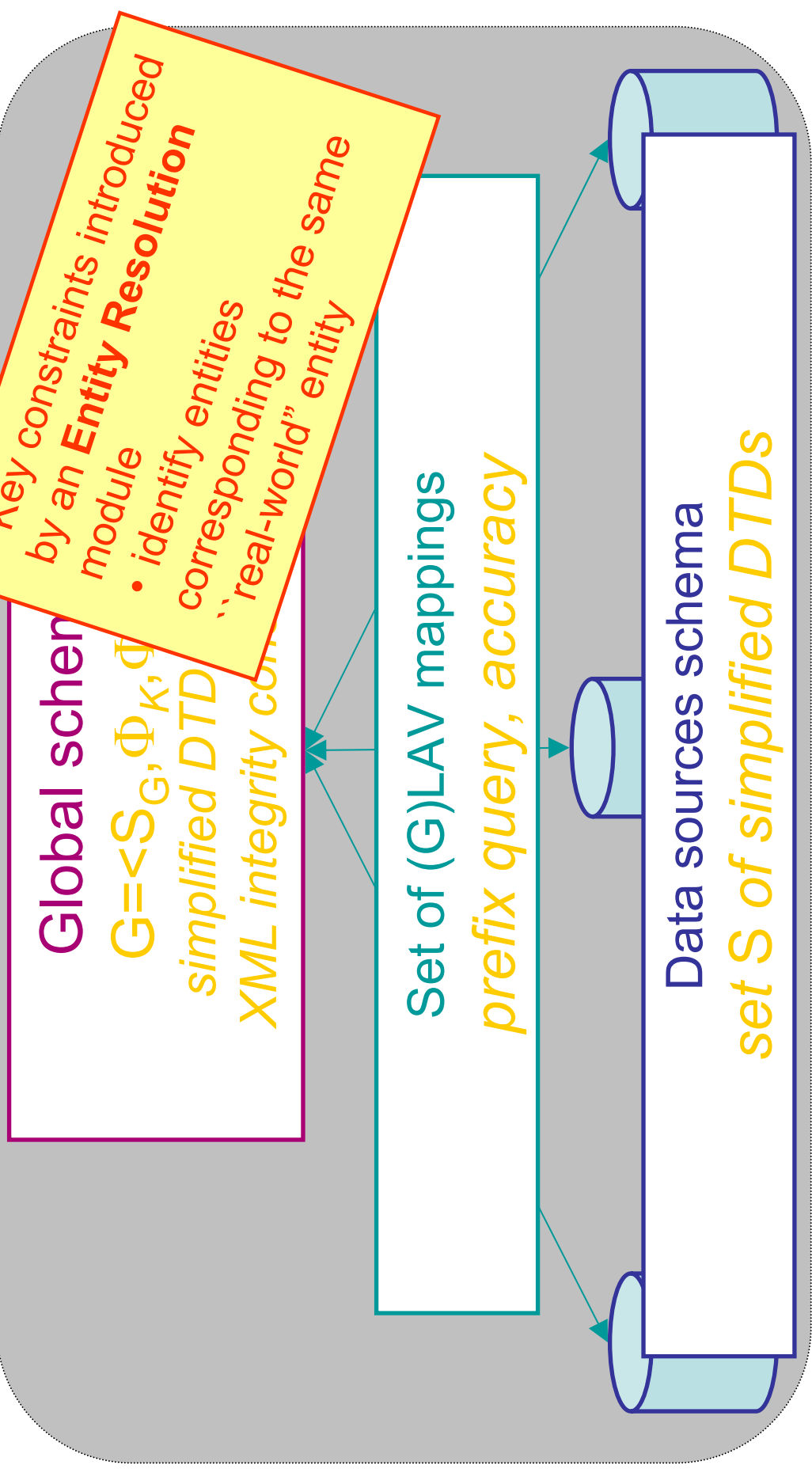
XML	DTD, XML keys and foreign keys	GLAV	Sound/exact
-----	--------------------------------	------	-------------

Outline(2)

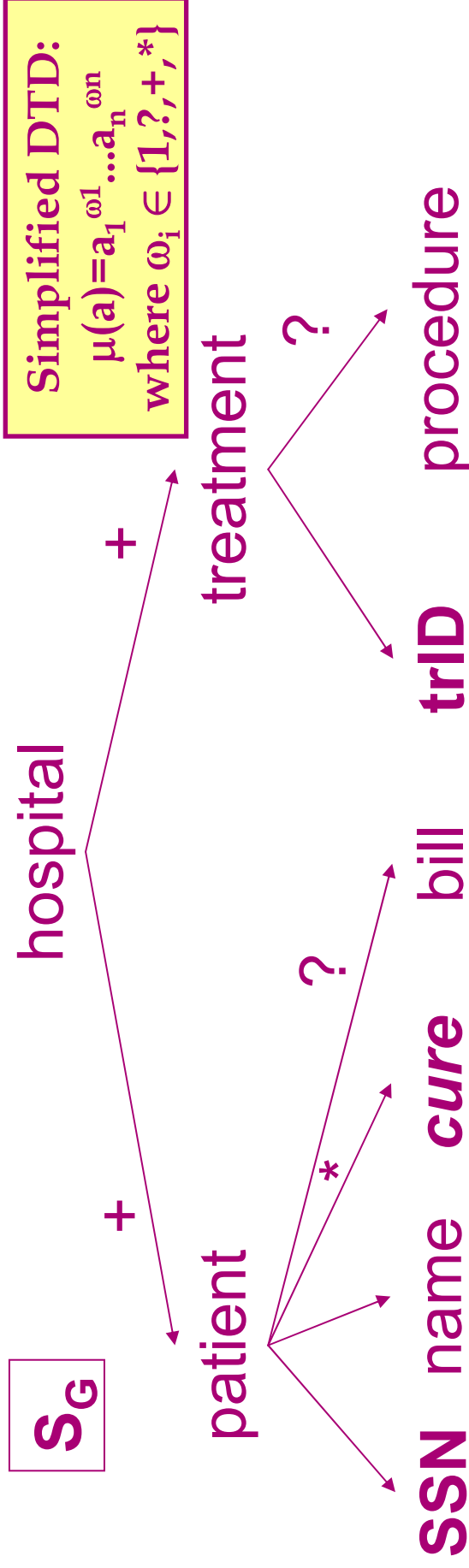
Thesis contributions

- Read-only XML-based data integration
 - logical formal DIS framework based on XML
 - definition of an identification function to address the “node identity” issue
 - consistency and query answering services
 - NP-hard consistency
 - coNP-hard query answering
 - PTIME algorithms under particular restrictions

XML data integration framework



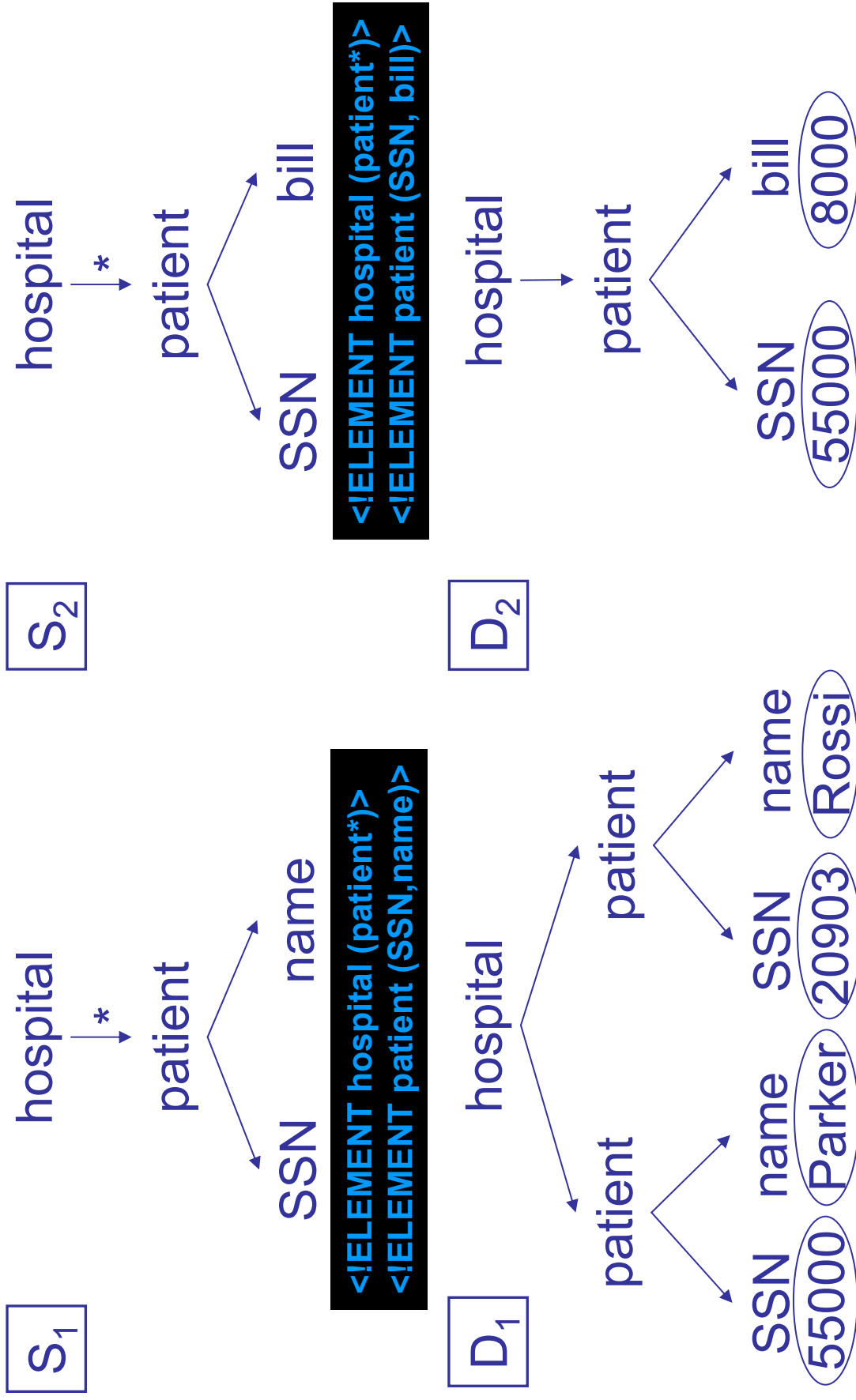
Global schema



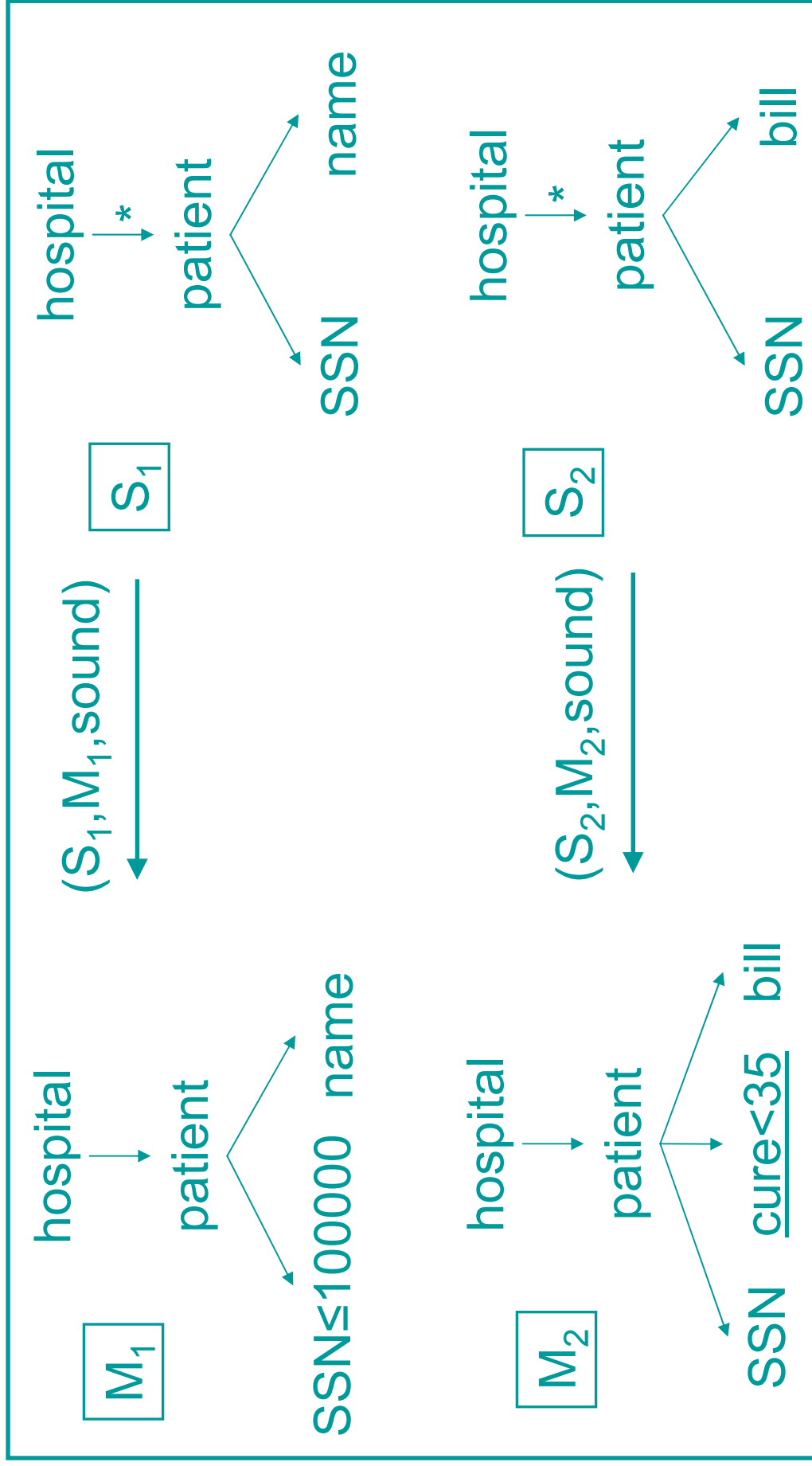
XML (absolute) keys and foreign keys [FanLibkinJACM02]:

- Φ_K : {patient.SSN \rightarrow patient
 treatment.trID \rightarrow treatment}
- Φ_{FK} : {patient.cure \subseteq treatment.trID}
- foreign keys restriction: the referenced key should be uniquely localizable!

XML data sources

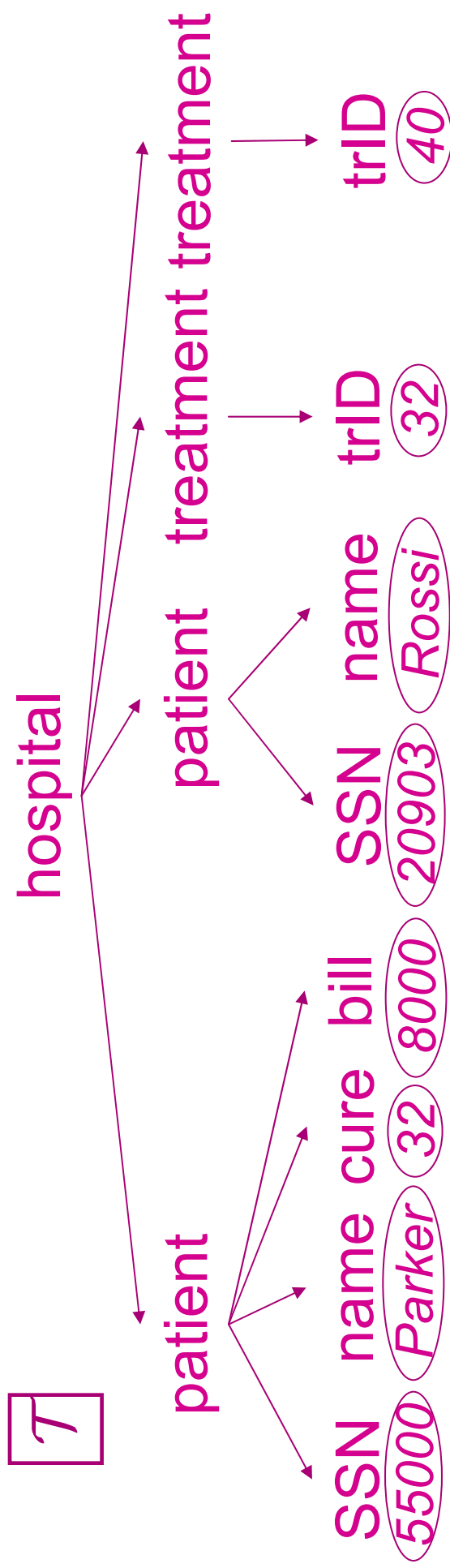


(G)LAV mappings



DIS semantics

Legal data tree



$\mathcal{T} \models \mathcal{G}$, i.e.:

- $\mathcal{T} \models S_G$
- $\mathcal{T} \models \Phi_K$
- $\mathcal{T} \models \Phi_{FK}$

\mathcal{T} satisfies the sound mappings:

- $D_1 \leq M_1(\mathcal{T})$
- $D_2 \leq M_2(\mathcal{T})$

Results on the logical XML-based DIS framework

- **Extension** of the prefix-selection query language of [AbiteboulEtAl., '01]
 - existential subtree patterns
 - absence of node identifiers
- **Formalization based on homomorphism** relation between trees
 - to specify the semantics of mappings
 - to specify the semantics of **certain answers**

Outline(2)

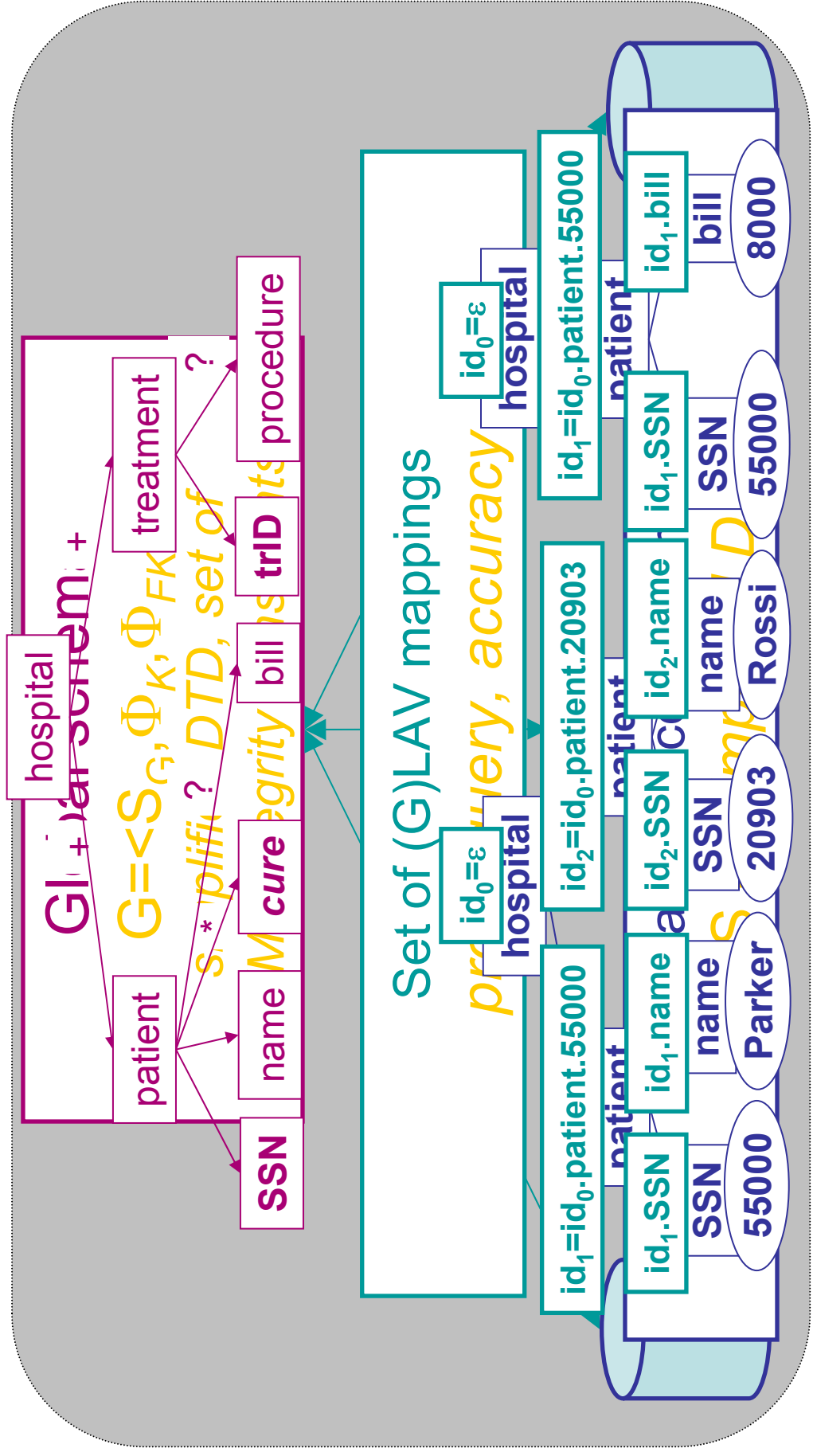
Thesis contributions

- Read-only XML-based data integration
 - logical formal DIS framework based on XML
 - definition of an **identification function** to address the “node identity” issue
 - consistency and query answering services
 - NP-hard consistency
 - coNP-hard query answering
 - PTIME algorithms under particular restrictions

Identification

- **Problem:** do two nodes correspond to the same node in every legal data tree?
- **Idea:** define an **identification function**
 - assigns to each data source node a “semantic identifier” such that:
 - two nodes with the same id are always mapped to the same node (**sound**)
 - two nodes having different ids are mapped to distinct nodes at least in one legal data tree (**complete**)
- **Our solution:** the identification function Id_g based on global schema **key constraints**

Identification function Id_G



Results on the identification function $\text{Id}_{\mathcal{G}}$

- always **sound**
- computation: **P**TIME
- **not complete** in general
 - there exists no **P**TIME identification function **sound** and **complete**, under the general setting
- complete under **Visible Key Restriction (VKR)**
 - for each collection of nodes, there exists a key constraint in \mathcal{G}
 - for each exact mapping, each selected element is selected together with its key

Outline(2)

Thesis contributions

- **Read-only XML-based data integration**
 - logical formal DIS framework based on XML
 - definition of an identification function to address the “node identity” issue
- **consistency and query answering services**
 - NP-hard consistency
 - coNP-hard query answering
 - PTIME algorithms under particular restrictions

Results on XML-based consistency

- Algorithm assuming to have a sound and complete identification function Id (e.g. Id^g under VKR)
 - idea
 1. apply Id to all data source nodes
 2. check whether there are key violations
 - two nodes with the same id have different labels and/or different values
 - two nodes with the same key value have different parent ids
 - **sound** and **complete**
 - **PTIME** in the size of the data
- Complexity lower-bound
 - **NP-hard** in the size of the data
 - reduction from 3-colorability

Results on XML-based query answering

- Algorithm assuming to have a sound and complete identification function Id (e.g. Id^g under VKR)
 - idea
 1. apply Id to all data source nodes
 2. build an incomplete tree [AbiteboulEtAl., '01]
 3. use it as a **representation system** [Imielinski, Lipski`84] from which derive certain answers
 - **sound** and **complete**
 - **PTIME** algorithm
- Complexity lower-bound
 - **coNP-hard** in the size of the data→ by reduction from variant of 3-colorability

Conclusion

Open problems

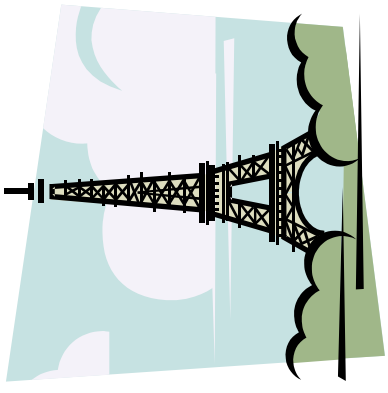
- Read-only ontology-based data integration [OWLED,'06]
 - can we extend DL-Lite_A and keep query answering LOGSPACE?
- Write-also ontology-based data integration [AAAI,'06]
 - mappings?
 - other approaches to updates (e.g. belief revision)?
- Read-only XML-based data integration [DBPL,'06]
 - algorithms for consistency and query answering under general setting (upper-bounds)?
- Relationships between ontology-based and XML-based data integration
 - node identity/object identifier?
 - XML query rewriting in the same spirit of the approach followed in DL-Lite_A?

Other activities and publications

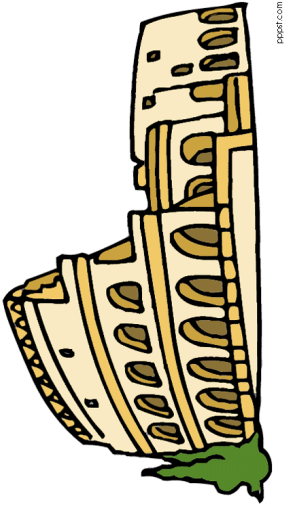
- Research on Personal Information Management
 - Vivi Katifori, Antonella Poggi, Monica Scannapieco, Tiziana Catarci, and Yannis Ioannidis. **OntoPIM: how to rely on a personal ontology for Personal Information Management.** In *Proceedings of the First Workshop on The Semantic Desktop*. November 2005.
 - Vivi Katifori, Antonella Poggi, Monica Scannapieco, Tiziana Catarci, and Yannis Ioannidis. **Managing personal data with an Ontology.** In *Proceedings of the Second Italian Research Conference on Digital Library Management Systems*. January 2006.
 - Alan Dix, Tiziana Catarci, Benjamin Habegger, Yannis Ioannidis, Azrina Kamaruddin, Vivi Katifori, Giorgos Lepouras, Antonella Poggi, Devina Ramduny-Ellis (2006). **Intelligent context-sensitive interactions on desktop and the web.** In *Proceedings of the International AVI'2006 Workshop on Context in Advanced Interfaces*. May 2006.
 - Tiziana Catarci, Benjamin Habegger, and Antonella Poggi. **Intelligent User Task Oriented Systems.** In *Proceedings of the Second SIGIR Workshop on Personal Information Management (PIM)*, 2006.
 - Tiziana Catarci, Xin Luna Dong, Alon Halevy and Antonella Poggi. *Personal Information Management*, chapter **Structure Everything**. University of Washington Press (UW Press). To Appear.
- Study of the current data federation commercial tools
 - Antonella Poggi and Marco Ruzzi. **Filling the gap between data integration and data federation.** In *Proceedings of the Twelfth Italian Symposium on Advanced Database Systems (SEBD)*. June 2004.
- Involved in the management of the QuOnto reasoner system, developed within the Data & Knowledge Bases group of the University of Rome “La Sapienza”

Thesis related publications

- Antonella Poggi and Serge Abiteboul. **XML data integration with identification.** *In Proceedings of the Tenth International Symposium on Database Programming Languages (DBPL)*. August 2005.
- Antonella Poggi and Serge Abiteboul. **XML data integration with identification (Extended Abstract).** *In Proceedings of the Thirteenth Italian Symposium on Advanced Database Systems (SEBD)*. June 2005.
- Giuseppe De Giacomo, Maurizio Lenzerini, Antonella Poggi, and Riccardo Rosati . **On the Update of Description Logic Ontologies at the Instance Level.** *In Proceedings of the 21st National Conference on Artificial Intelligence*. July 2006.
- Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, Antonella Poggi, and Riccardo Rosati. **Linking Data to Ontologies: The Description Logic DL-LiteA.** *In Proceedings of the Workshop OWLED*. November 2006.
- Antonella Poggi, Domenico Lembo , Diego Calvanese , Giuseppe De Giacomo , Maurizio Lenzerini , and Riccardo Rosati. **Linking Ontologies to Data.** Submitted to an international journal, 2007.



Grazie



Merci