LahDAK: Large-scale Heterogeneous Data and Knowledge

Members of 2 former teams:
- the Database team (DB)
- the Intelligence Artificial and Inference System team (IASI)

A previous experience of collaboration (EI LEO)
Database & Oak Team
## Team Members – June 2013

<table>
<thead>
<tr>
<th>Team</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Permanents</strong></td>
<td>5 UPSud faculty members (1 PR – 3 MdC), 1 INRIA (1 DR)</td>
</tr>
<tr>
<td></td>
<td>3 HdR</td>
</tr>
<tr>
<td><strong>PhD Students</strong></td>
<td>2 in 1st year, 2 in 2nd year, 1 in 3rd year, 1 in 4th</td>
</tr>
<tr>
<td></td>
<td>4 Alloc. MESR, 1 Digiteo, 1 Cordi INRIA</td>
</tr>
<tr>
<td><strong>Post Doc</strong></td>
<td>1 INRIA</td>
</tr>
<tr>
<td><strong>Engineers</strong></td>
<td>1 INRIA</td>
</tr>
<tr>
<td><strong>Associated members</strong></td>
<td>1 INRIA (DR), 2 faculty members (1 PR Cnam, 1 MdC Cnam)</td>
</tr>
</tbody>
</table>

## Team Evolution – 2008/2013

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recruitment</strong></td>
<td>1 MdC – BQR position 2011</td>
</tr>
<tr>
<td><strong>Team changes</strong></td>
<td>2 PR → Toccata 2010 → A&amp;O 2012, 1 DR ← IASI 2012</td>
</tr>
<tr>
<td><strong>PhD Students</strong></td>
<td>6 Alloc. MESR, 3 ETR, 1 cotutelle</td>
</tr>
<tr>
<td><strong>Joint Inria team</strong></td>
<td>creation of Oak (Sept. 2009 → April 2012)</td>
</tr>
<tr>
<td></td>
<td>leader: Ioana Manolescu</td>
</tr>
<tr>
<td></td>
<td>Database Optimization and Architectures for Complex large data</td>
</tr>
</tbody>
</table>
Research Directions

Research Context

- **Complex Data**
  - Structured document or tree data (XML, JSON, DTD, XML Schema, RelaxNG, ...),
  - Graph-based data (RDF, social networks, RDF Schema, ...),
  - Digital libraries

- **Complex processing**
  - Fine-granularity search (Xquery, SPARQL, ...),
  - Updates (Xquery Update Facility, ...)
  - Complex data transformation (data integration, data cleaning), provenance
  - Data analysis, data mining

- **Optimization for large & distributed data sets**
  - Time & Space - Scalability
  - Cloud environment
  - Static analysis, View based tech, Parallelization

Main Topics of Interest / Contributions

- Static type-analysis for XML query and update languages
- XML and View based optimization
- Web Data processing in the cloud
- Data transformation Quality
- Digital Libraries
- Data mining
- Foundations of database

Labex Digicosme - Action Line DATASENSE

**Task 1**: Scalable, expressive and secure tools for large scale data
**Task 2**: Making sense of heterogeneous, complex data
Type-based optimization

- Type-based projection for **update**
- Optimizing main memory XML Processing

Memory limitations

- XML document pruning
- Streaming + no buffering
- Generic methods
- no rewriting of the engines

(150MB - 1GB)

**3-projector**

(Galax, Saxon, QizX, Zoorba, Xbase)

**specific scenario**
Static type-analysis for XML

DB (XML data management) + Programming Languages (Type systems)

Type-based optimization

- Type-based projection for **update**
- Optimizing main memory XML Processing
  - Memory limitations
    - XML document pruning
  - Streaming + no buffering
  - Generic methods
    - no rewriting of the engines

- **Temporal** XML data maintenance
  - Compact Timestamp document

(150MB - 1GB) **3-projector**

(Galax, Saxon, QizX, Zoorba, Xbase) **specific scenario**

Experiments: QuizX

Experiments with QizX

Update of documents up to 2GB

Better execution time

Wednesday, October 5, 2011
**Static type-analysis for XML**

DB (XML data management) + Programming Languages (Type systems)

**Type-based optimization**
- Type-based projection for **update**
- Optimizing main memory XML Processing

**Memory limitations**
- XML document pruning
- Streaming + no buffering
- Generic methods
- no rewriting of the engines

**Temporal XML data maintenance**
Compact Timestamp document

**Query Update Independence**
View re-materialization, Concurrency
- Independence is undecidable for XQuery
  - precision
  - finite analysis
  - polynomial complexity

**Experiments with QizX**
- Better execution time
- Reduction of execution time as well.

**Projection for XML Update Optimization**
- (150MB - 1GB) 3-projector
- (Galax, Saxon, QizX, Zoorba, Xbase)

**Independence Detection - Precision**

```
INDEPENDENCIES DETECTED (%)

100  25  50  75
     0

UI1  UA1  UI2  UA2  UA3  UA4  UA5  UA6  UA7  UA8

INSERT                                  RENAME                                  REPLACE

UI5

UA1

UA2

UA3

UA4

UA5

UA6

UA7

UA8

INSERT                                  RENAME                                  REPLACE

UI5

UA1

UA2

UA3

UA4

UA5

UA6

UA7

UA8

VIEW re materialization

Compact Timestamp document

```
Static type-analysis for XML

DB (XML data management) + Programming Languages (Type systems)

**Type-based optimization**

- Type-based projection for update [PhD : M. Sahakyan, M-A Baazizi]
- Projection for temporal XML data [PhD : M-A Baazizi]
- Query Update Independence [PhD : F. Ulliana]

**Publi:** SKDB10, EDBT11, TIME11, PVLDB12, TODS13, ...

**Software:** XUpOp, XUpTe, XUpIn

**Contrat:** ANR Codex
Static type-analysis for XML

DB (XML data management) + Programming Languages (Type systems)

**Type-based optimization**
- Type-based projection for update [PhD : M. Sahakyan, M-A Baazizi]
- projection for temporal XML data [PhD : M-A Baazizi]
- Query Update Independence [PhD : F. Ulliana]

Publi : SKDB10, EDBT11, TIME11, PVLDB12, TODS13, ...
Software : XUpOp, XUpTe, XUpIn
Contrat : ANR Codex

**Mapping Maintenance in XML Data Integration Systems**
- Precise type inference, and efficient correctness checking

**Inclusion of XML schemas with interleaving and counting**

(a&d).b[1..3] XML Schema / RelaxNG
- quadratic inclusion under conflict-free restriction **types as constraints**
- quasi-linear inclusion, relying on the quadratic one

Publi : CIKM08, TOIT09, CIKM09, ICDT09, IS09, TCS13, TODS13
Collab : Univ. Pisa, Univ. Basilicata
Contrat : ANR Codex
XML View based optimization (IASI - BD)

Storage optimization
- Materialized Views (Store, Index, Views) = Rich Tree Patterns
  selection of XQuery views / multiple-views rewritings
- View based query optimization
  Query Rewriting / Minimization / …
- View Maintenance

Distributed Environment (P2P)
- Queries over distributed XML data
  250 machines / 1000 ViP2P peers deployed

PhD : K. Karanasos, A. Katsifodimos, S. Zoupanos
Publi : WWW08, ICDE08, ICDE10, ICWE10, EDBT11, ICDE11,
  SIGMOD 12, TODS13
Software : ViP2P Plateform (APP11 - demo), LiquidXML (demo)
Contrat : ANR Codex, ANR WebStand, ANR DataRing
Collab : CNR, Univ. Calabria, UCSC, Univ. of E.& B. Athens
Parallel & Cloud data management

Map-Reduce XML Processing
- Xquery and Update: from xGB to any document size
  From pruning to partitioning (static analysis again) [PhD: N. Malla]

Web Data in the Cloud
- Warehousing large-scale XML & RDF data
  Amazon Web Services / monetary cost
  indexing techniques [PhD: J. Camacho-Rodriguez] [PostDoc: F. Bugiotti, Z. Kaoudi]

Cloud warehouse

<table>
<thead>
<tr>
<th></th>
<th>Publi:</th>
<th>Software:</th>
<th>Grant:</th>
<th>Collab:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IDEAS12, EDBT 2013</td>
<td>XPIOT-Andromeda (demo)</td>
<td>KIC ICT Lab Connected Clouds &amp; Europa</td>
<td>Univ. Basilicata</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Publi:</th>
<th>Software:</th>
<th>Grant:</th>
<th>Collab:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CIKM12, DanaC12, DMC12, EDBT13, ICDE13</td>
<td>Amada (demo)</td>
<td>KIC ICT Lab Connected Clouds &amp; Europa</td>
<td>Univ. Roma</td>
</tr>
</tbody>
</table>
Map-Reduce XML processing

- Xquery and Xpath from XGB to any document size
  From pruning to partitioning (static analysis again)

Demo XPIOT Andromeda

Web Data in the Cloud

- Warehousing large-scale XML & RDF data
  Amazon Web Services / monetary cost
  indexing techniques

Cloud warehouse

Publi : CIKM12, DanaC12, DMC12, EDBT13, ICDE13
Software : Amada (demo)
Grant : KIC ICT Lab Connected Clouds & Europa
Collab : Univ. Roma
Data Quality

- Efficient and Effective Duplicate Detection in Hierarchical Data
- Optimization of a class of duplicate detection

Transformation Quality

- Transformation analysis based on data provenance
  - Why provenance explain existing answer
  - Why not provenance explain missing answer

PhD: K. Tzompanaki
Publi: TKDE'12, TKDE'13, CDMW'12, CIKM'12
Software: Nautilus Plateform (Maturation Pôle Dev.)
Contrat: WebNot, KIC ICT Lab
Collab: IST-INESC-ID (Portugal), H-P Institut (Germany)
Data & Transformation Quality

Data Quality
- Efficient and Effective Duplicate Detection in Hierarchical Data
- Optimization of a class of duplicate detection

Transformation Quality
- Transformation analysis based on data provenance
  Why prove provenance?
  Why not prove provenance?

PhD: K. Tzompanaki
Publi: TKDE, CDMM, CIKM
Software: Nautilus Platform
Contrat: WebNot
Collab: IST-INESC-ID Portugal, IFP Institut (Germany)
Further contributions

**Data Analysis and Data Mining**
- Functional model and query language for data analysis  
  [PhD: E. Simonenko]
- Mining frequent conjunctive queries in a relational database  
  PostDoc: T. Sugibuchi

  Publi: EDBTo8, ICFA12, VLDB J.13
  Contrat: KP-Lab IST-4 Project, Projet STIC-Asie GOD, Digiteo VISIR, PHC SIAM
  Collab: Franco-Japanese CNRS-JST project, MeMe Media Lab (Japan), Univ. Kasetsart (Tailand)

**Digital Libraries**
- a new model and query language supporting identification, structuring, metadata creation, reuse and discovery of digital resources

  PhD: H. Belhaj Frej, M. Nguer, J. Yang  
  PostDoc: T. Sugibuchi
  Publi: Int. J. of Digital Libraries, ECDL10, ICDEo8
  Contrat: ICT Projet ASSETS
  Collab: PICS CNR Pisa

**Foundations of Database**
- expressive power of update primitives

  Publi: PODS13
  Collab: Univ. Hasselt
Summary of contributions

Publications (101)
- 16 Journals A+/A: JACM, TCS, TKDE, 3xTODS, TOIT, 3xVLDB, WWW, ...
- 33 CI A+/A: 5xCIKM, 7xICDE, 5 EDBT, PODS, 2xPPDP, 2xSIGMOD, TIME, VLDB, ...
- 4 Invited talks

Software (13)
- 3 dépôts APP, 6 demos

Contracts & Grants (15)
- 3 Digiteo, 4 ANR, 1 Investissement d’Avenir
- 3 ICT Lab, 2 European

Collaborations
- more than 40 publications co-signed with international collaborators
- Meme Media Lab (Japan), OakSad (UCSD), ...

Editorial Board (7)
- ACM Sigmod Record, TOIT, ...

Chair (7) + PC (40)
- 6 int. Chair : ICDE, EDBT, VLDB(PhD), ...
- 31 int. PC : 4xEDBT, 3xICDE, 2xSIGMOD, 4xVLDB, WebDB, ...
- 7 BDA

Management
- VP CS Department, Dir. Doctoral School, LRI Council, ANR Scientific Committee, Master, M2R IAC, KIC EIT Lab, Labex Digicosme, CCSU27, Recruiting Committees (23)

PhD defense (10) + HdR (1)
- 1 MdC, 2 Post Doc, ...
- 1 PR

Summer School (6) + Tutorials (3)
- Masses de Données Distribuées (creation, 3 scientific organization)

Master Recherche
- UPSud (IAC), Dauphine, Univ. Beyrouth

Many other interactions
- IBM DB2 Community member
- Industrial Conf. "Big Data"
- Chair of Gilles Kahn PhD Award Committee
- Fête(s) de la Science
IASI
Artificial Intelligence and Inference Systems
Head: C. Reynaud
Group members and Evolution

• **Permanent faculty:** 9 university (2 PR – 7 MC) – 1 CNRS (CR) 4 HDR

• **Evolution**
  +1 Assistant Professor
  +1 junior CNRS researcher
  -2 Inria senior researchers ➔ LSV 2011, LRI-BD 2012
  -1 Assistant Professor (retired)

• **3 associated members**

• **15 post doctoral students in 5 years**
  Origin:
  1 Ph.D. LRI - Paris XI
  8 Ph.D. in France but not in Paris XI
  6 Ph.D. in a foreign university
  1 The Netherlands, 1 UK, 2 Italy, 1 USA, 1 Canada

• **18 defended thesis in 5 years**
  Current position:
  3 Assistant Professor,
  2 Junior researcher,
  6 Post-doc,
  5 Industry,
  2 Unknown

  *2 Ph. D. in co-tutelle with Luxembourg and Senegal*

• **9 engineers in 5 years**
  4 engineers financed by contracts
  5 ADT Inria

• **10 thesis in progress** *(October 1st, 2013)*
  Origin:
  3 M2R IAC Paris XI,
  2 M2Rs in France, but not in Paris XI
  2 M2R Franco-Hellenique,
  3 Masters in a foreign university
Objectives / Scientific themes

Objective: Developing « Intelligent » Web Information Systems

- dealing with *highly heterogeneous, numerous, incomplete, evolutive and distributed data and knowledge*
- providing *reliable and meaningful data*

Scientific Themes:

- Ontology-Based Data and Document Management
- Data and Knowledge Integration
- Scalable techniques for Semantic Web Data Management
- Reasoning over Distributed Systems
- SAT Solving
Ontology-Based Data and Document Management

1. Semantic annotation
   - Annotating more or less structured HTML documents *M. Thiam* – *Y. Mrabet Ph.D.* - SHIRI project
   - Extracting tabular data
     *Industrial project with DataPublica*
     *DataBridges - EIT ICT Labs-KIC’10 - ’11*

2. Querying adaptive ontologies *(C. Pruski Ph.D.)*
   - Managing adaptive ontologies
   - Query enrichment based on adaptive ontologies

3. Querying description logics knowledge bases
   - Containment and minimization of queries
   - Identifying first-order rewritable queries
   - Inconsistency-tolerant query answering
   - Links to other formalisms (CSP, disjunctive datalog)
   - Robust module extraction

[CAISE’10, WISE’12, ECAI’12, EDBT’13]

SHIRI-Querying

TEXAN

POI-HELPER

Collaboration with Supelec, Univ. G. Berger (Sénégal), DFKI (Germany), Univ. Southampton (UK)

[Int. Journal of Web Portals ‘09]

TARGET

[ KR’12, AAAI’12, TKDE’13, PODS’13]

Collaboration with LIG (France), UC Santa Cruz (USA), Univ. Bremen (Germany), Univ. Liverpool (UK)
Ontology-Based Data and Document Management

1. Semantic annotation
   - Annotating more or less structured HTML documents
     * M. Thiam – Y. Mrabet Ph.D. - SHIRI
   - Extracting tabular data
     * Industrial project with DataPublica
     * DataBridges - EIT ICT Labs-KIC’10 - ’11

2. Querying adaptive ontologies (C. Pruski Ph.D.)
   - Managing adaptive ontologies
   - Query enrichment based on adaptive ontologies

3. Querying description logics knowledge bases
   - Containment and minimization of queries
   - Identifying first-order rewrite rules
   - Inconsistency-tolerant query answering
   - Links to other formalisms (CSP, disjunctive datalog)
   - Robust module extraction

[CAISE’10, WISE’12, ECAI’12, EDBT’13]

SHIRI-Querying

[Int. Journal of Web Portals ‘09]

TEXAN

[ KR’12, AAAI’12, TKDE’13, PODS’13]

SHIRI-Querying

POI-HELPER

Collaboration with Supelec,
Univ. G. Berger (Sénégal)
DFKI (Germany),
Univ. Southampton (UK)

Poster

Zoom-2
Data and Knowledge Integration

In classical information management systems, Adaptive Hypermedia (AHS) and Personal Information Management System (PIMS)

1. Mapping between ontologies
   - Mapping refinement (F. Hamdi Ph.D.) – GeOnto ANR project
   - Mapping maintenance: Ongoing Ph.D. - J. C. Dos Reis Dynamo project

2. Data linking
   - Identification of links between fuzzy relation instances
   - Discovering of keys in RDF datasets
     Ongoing Ph.D. - D. Symeonidou – Qualinca ANR project

3. Integration of existing materials (N. Zemirline Ph.D.)
   - Author’s model and adaptation strategies in AHSs

4. Context-Aware Personal Information Integration
   Ongoing Ph.D. – R. Khefifi - PIMI ANR project

[EKAW’10, RIG’11, ESWC’13]
- TaxoMap Alignment (dépôt APP)
- TaxoMap Refinement (dépôt APP)
- Collaboration with IGN, CRT (Lux.)

[JoDS’09, ODBASE’10]
- Collaboration with INRA, LIG (France)
- LN2R–LT (APP) - KD2R

[AH’08, EKAW’08, TLT’12]
- Collaboration with Supelec, Univ. of Warwick (UK)
- MESAM – EAP framework

[EGC’11, DEXA’12]
- Collaboration with the VALS group
Data and Knowledge Integration

In classical information management systems, Adaptive Hypermedia (AHS) and Personal Information Management System (PIMS)

1. Mapping between ontologies
   - Mapping refinement (F. Hamdi Ph.D.) – GeOnto ANR project
   - Mapping maintenance: Ongoing Ph.D. - J. C. Dos Reis Dynamo project

2. Data linking
   - Identification of links between fuzzy relation instances
   - Discovering of keys in RDF datasets
     Ongoing Ph.D. - D. Symeonidou – Qualinca ANR project

3. Integration of existing materials (N. Zemirline Ph.D.)
   - Author’s model and adaptation strategies in AHSs

4. Context-Aware Personal Information Integration
   Ongoing Ph.D. – R. Khefifi - PIMI ANR project

[EKAW’10, RIG’11, ESWC’13]
- TaxoMap Alignment (dépôt APP)
- TaxoMap Refinement (dépôt APP)
- Collaboration with IGN, CRT (Lux.)

[JoDS’09, ODBASE’10]
- Collaboration with INRA, LIG (France)
- LN2R - LT (dépôt APP) - KD2R

[AH’08, EKAW’08, TL’T12]
- Collaboration with Supelec, Univ. of Warwick (UK)
- MESAM – EAP framework

[EGC’11, DEXA’12]
- Collaboration with the VALS group

27-28 Nov. 2013 Comité d’évaluation du LRI
Scalable techniques for Semantic Web Data Management

RDF data management - common with the BD group

1. **Efficient database techniques for Semantic Web query answering**
   
   *A. Roatis - J. Leblay Ph.D. - Digiteo DW4RDF project*
   
   - Reformulation- and saturation-based query answering techniques robust to updates
   - Development of models, languages and algorithms for OLAP-style analysis of RDF data

2. **Efficient automatic storage recommendation for Semantic Web databases**
   
   *K. Karanosos and J. Leblay Ph.D. - Codex ANR project*
   
   - View selection minimizing a combination of query processing, view storage and view maintenance costs
1. Efficient database techniques for Semantic Web query answering
   A. Roatis - J. Leblay Ph.D. - Digiteo DW4RDF project
   - Reformulation- and saturation-based query answering techniques robust to updates
   - Development of models, languages and algorithms for OLAP-style analysis of RDF data

2. Efficient automatic storage recommendation for Semantic Web databases
   K. Karanosos and J. Leblay Ph.D.
   Codex ANR project
   - View selection minimizing a combination of query processing, view storage and view maintenance costs
1. **Reasoning in peer-to-peer inference systems**

[ECAI’08, ODBASE’08, AICOM’09, IJCAI’09, PODS’11]

- Bayesian approach for trust modeling (*Ph.D. G.H. Nguyen*)
- Directed mappings (*Ph.D. N. Abdallah*)
- Conservative extension (*Ph.D. N. Abdallah*)
- Peer data management using DL-Lite-R
- Mapping discovery (*Ph.D. F.E. Calvier*)
- Datalog rule based distributed data management

2. **Distributed diagnosis and distributed diagnosability**

[LPAR’08, ECAI’08, JWSR’09, KES’10, CP’12]

- On-line diagnosis of distributed software processes (*Ph.D. Y. Li*)
- Distributed consistency-based diagnosis (*Ph.D. V. Armant*)
- Diagnosability analysis for distributed discrete-event systems (*Ph.D. L. Ye*)
SAT Solving

[CP’08, IJCAI’09, AAAI’10, IJCAI’11, AI Magazine’12, CP’12, SAT’12, SAT’13]

1. Practical solving of SAT problems
   - Organization of the SAT Competitions until 2009
   - Proposition of a measure of learnt clauses
   - Winner of gold medals at the SAT competition at 4 distinct events

2. Adaptation of the power of Glucose
   - To the incremental SAT solving case
   - Proposal of the basis of learning mechanisms for polynomials in GL(2)

3. Analyze of the efficiency of SAT solvers
   - Analyze of the structure of industrial problems w.r.t. essential components of modern SAT solvers

Collaboration with CRIL (France)
Summary

**Publications (216)**
- 25 in journals with 13 A+/A: AI (2), AI Magazine, KDE, JoDS, TLT, VLDB J., PVLDB, SIGMOD record, IS J.
- 175 in conferences with 43 CI A+/A: IJCAI (3), AAAI (2), ECAI (3), PODS (4), EDBT (5), ICDE (4), CP (4)
- 16 Books and chapters

**Software (27)**
- 3 dépôts APP
- Competition: 4 gold medals at the SAT competition

**Contracts & Grants (33 – 3235,47 K€)**
- 15 ANR (2193,37 K€), 7 Industry (464,12 K€), 2 Digiteo (199,78 K€), 1 Subv. (13 K€)
- 4 ICT Labs (141,54 K€), 2 European (202,66 K€), 1 FNR Lux. (9 K€), 1 Subv. (13 K€)

**Collaborations**
- more than 90 publications co-signed with collaborators (33 with international collaborators)
- 16 visiting professors and students (74 months)

**Editorial Board (18)**
- ACM SIGMOD, ACM TOIT, DKE, JoDS, IS J., JSAT, PVLDB

**Chair (19) + PC International (79) + PC National (98)**
- 16 int. (Co-)Chair: CIKM, SIGMOD, ECAI, ICDE, EDBT, SAT, VLDB, WWW
- 49 int. PC: AAAI, CIKM, SIGMOD, ECAI, ECDL, EDBT, EKAW, ESWC, IJCAI, KR, PODS, SAT, VLDB, WWW
- 9 RFIA, 9 IC, 9 EGC, 6 BDA

**Evaluation committee**
- 2 AERES, 16 ANR, 1 CAPERS-COPECUB
- 1 CEI INRA, 5 CIFRE, 1 CPER, 2 ERC

**Interactions with socio-economic and cultural environment**
- Fête de la science
- Participant in the competition for the recognition of research results in Paris-Sud, OMTE Digiteo IASI-Tools
- Invited talks and demos (5 at iMatch INRIA 2011)
- Expertise for the tax administration, consulting activities

**Management**
- Dir. LRI, VP CS Department, Dir. CS Department IFIPS, M2R IAC
- In concil of: Ecole Doctorale EDIPS, Conseil de la Recherche UFR, LRI, IUT Sceaux, Member of CCSU, Recruiting Committee (31)
- KIC EIT Lab
- Teaching responsibilities: apprenticeship L3-Miage, Standard M2 Miage, apprenticeship DUT in CS.

**Master Internships (19)**
- 8 Master IAC Paris-Sud
- 6 Master in France but not in Paris-Sud
- 2 Master Franco-Hellenique
- 3 Foreign Master or Graduate School Degree

**Courses**
- Graduate courses in UPSud (I3 - IAC), MIDO Dauphine, Univ. Beyrout, ENS Lyon
- Courses in Univ. Thies (Senegal), Ecole ibn Sina (Ramallah) in Palestine
- Design of an educational tool for teaching programming languages
Merci
Zoom 1: Database techniques for the Semantic Web

François Goasdoué (now U. Rennes 1), Konstantinos Karanasos (now IBM Almaden), Julien Leblay (now U. Oxford), Ioana Manolescu, Alexandra Roatiş
Database techniques for the Semantic Web

Data management research has produced highly scalable and robust DBMSs

• Data stores (transactions, …)
• Query evaluation: $Q(\text{DB}) = ?$
• What about query answering: $Q(\text{DB}, \text{KB}) = ?$
Database techniques for the Semantic Web

Data management research has produced highly scalable and robust DBMSs

• Data stores (transactions, …)
• Query evaluation: $Q(\text{DB}) = ?$
• What about query answering: $Q(\text{DB}, \text{KB}) = ?$

Semantic Web: W3C’s RDF; KB in RDF Schema

• Wide range of applications (government, science, economy, culture, …)
• Some data stores and query evaluation tools ignoring semantics/knowledge
Database techniques for the Semantic Web

Data management research has produced highly scalable and robust DBMSs

- **Data stores** (transactions, …)
- **Query evaluation**: $Q(\text{DB}) = ?$
- **What about query answering**: $Q(\text{DB}, \text{KB}) = ?$

**Semantic Web**: W3C’s RDF; KB in RDF Schema

- Wide range of applications (government, science, economy, culture, …)
- Some data stores and query evaluation tools ignoring semantics/knowledge

**Goal**: database techniques for efficient and expressive Semantic Web data management
Given: RDF query workload + RDF Schema

Find: set of relations to materialize which minimizes combination of:
view storage + maintenance + workload evaluation cost
Given: RDF query workload + RDF Schema
Find: set of relations to materialize which minimizes combination of:
view storage + maintenance + workload evaluation cost

1. Characterization of the search space
2. Complete set of state transformations
3. Efficient and effective search heuristics, also considering the KB

ACM CIKM 2010  PVLDB 2011  ANR CODEX
**Experimental results**: heuristics very effective reducing cost (up to $10^4$)

![Chart showing relative cost reduction for different Commonality levels and query sizes.](chart.png)

- **5 queries**
- **10 queries**
- **20 queries**
- **50 queries**
- **100 queries**
- **200 queries**

**Academic Conference**

- ACM CIKM 2010
- PVLDB 2011
- ANR CODEX
DB techniques for the Semantic Web:
2. Query answering

How to implement *reasoning* when answering queries?

- **Saturation** = reason on the data
  
  **DB**: Julie rentsAFlatIn Paris .
  **KB**: rentsAFlatIn rdfs:subPropertyOf livesIn .
    
    livesIn rdfs:subPropertyOf knows .

  Saturation adds:
  
  to the **KB**: rentsAFlatIn rdfs:subPropertyOf knows .
  to the **DB**: Julie knows Paris .

+ simplicity  
- maintenance
DB techniques for the Semantic Web:

2. Query answering

How to implement reasoning when answering queries?

- **Saturation** = reason on the data
  
  **DB**: Julie rentsAFlatIn Paris.
  
  **KB**: rentsAFlatIn rdfs:subPropertyOf livesIn.
  
  livesIn rdfs:subProperty knows.

  Saturation adds:
  
  to the **KB**: rentsAFlatIn rdfs:subPropertyOf knows.
  
  to the **DB**: Julie knows Paris.

  + simplicity  - maintenance

- **Reformulation** = reason on the query

  \[ q(x) :- x \text{ knows Paris} \]

  \[ q(x) :- x \text{ knows Paris} \rightarrow x \text{ livesIn Paris} U x \text{ rentsAFlatIn Paris} \]

  + data unchanged  - syntactically large queries \rightarrow inefficiency
DB techniques for the Semantic Web:

2. Query answering

How to implement reasoning when answering queries?

- **Saturation** = reason on the data
  
  **DB**: Julie rentsAFlatIn Paris .
  
  **KB**: rentsAFlatIn rdfs:subPropertyOf livesIn .
  
  livesIn rdfs:subProperty knows .

  Saturation adds:
  
  to the **KB**: rentsAFlatIn rdfs:subPropertyOf knows .
  
  to the **DB**: Julie knows Paris .

  + simplicity  
  
  - maintenance

- **Reformulation** = reason on the query

  q(x) :- x knows Paris .

  →

  q(x) :- x knows Paris . U x livesIn Paris U x rentsAFlatIn Paris .

  + data unchanged  
  
  - syntactically large queries → inefficiency
DB techniques for the Semantic Web:
2. Query answering

How to implement *reasoning* when answering queries?

- **Saturation** = reason on the data
  - DB: Julie rentsAFlatIn Paris.
  - KB: rentsAFlatIn rdfs:subPropertyOf livesIn.
    - livesIn rdfs:subProperty knows.
  Saturation adds:
    - to the KB: rentsAFlatIn rdfs:subPropertyOf knows.
    - to the DB: Julie knows Paris.
  + simplicity - maintenance

- **Reformulation** = reason on the query
  - q(x) :- x knows Paris.
  - q(x) :- x knows Paris. U x livesIn Paris U x rentsAFlatIn Paris.
  + data unchanged - syntactically large queries \(\rightarrow\) inefficiency
DB techniques for the Semantic Web:
2. Query answering

How to implement reasoning when answering queries?

- **Saturation** = reason on the data
  + simplicity - maintenance
- **Reformulation** = reason on the query
  + data unchanged - syntactically large queries → inefficiency

Our contribution:
1. Novel database fragment of RDF
   - RDF Schema
   - Blank nodes in schema and data
   - Joint querying of schema and data
2. Algorithms for
   - Incremental saturation (and saturation maintenance)
   - Reformulation + non-standard evaluation relying on off-the-shelf RDBMS
DB techniques for the Semantic Web:
2. Query answering: in the long run?

Saturation
Once per database, then maintenance

Reformulation
Once per query

Running $q$ $n$ times with saturation: $Sat + n \cdot \text{exec}(q, \text{Sat}(DB, KB))$
Running $q$ $n$ times with reformulation: $n \cdot \text{exec}(\text{Ref}(q, KB), DB, KB)$
DB techniques for the Semantic Web:
2. Query answering: in the long run?

Saturation

Once per database, then maintenance

Reformulation

Once per query

Running q n times with saturation: \( \text{Sat} + n \cdot \text{exec}(q, \text{Sat}(DB, KB)) \)

Running q n times with reformulation: \( n \cdot \text{exec}(\text{Ref}(q, KB), DB, KB) \)

Saturation threshold of q:
How many q runs until Sat amortizes?
DB techniques for the Semantic Web: 2. Query answering: in the long run?

Saturation

Once per database, then maintenance

Reformulation

Once per query

Running q n times with saturation: \( \text{Sat} + n \times \text{exec} \left( q, \text{Sat}(DB,KB) \right) \)

Running q n times with reformulation: \( n \times \text{exec}(\text{Ref}(q, KB), DB, KB) \)

Saturation threshold of q:
How many q runs until Sat amortizes?

Insertion / deletion threshold of q:
How many q runs until Sat maintenance cost for 1 added / deleted triple amortizes?
DB techniques for the Semantic Web:
2. Query answering: in the long run?

Saturation

Once per database, then maintenance

Reformulation

Once per query

Running $q$ $n$ times with saturation: $Sat + n \times exec(q, Sat(DB,KB))$

Running $q$ $n$ times with reformulation: $n \times exec(Ref(q, KB), DB, KB)$

Saturation threshold of $q$:

How many $q$ runs until $Sat$ amortizes?

Insertion / deletion threshold of $q$:

How many $q$ runs until $Sat$ maintenance cost for 1 added / deleted (DB | KB) triple amortizes?
DB techniques for the Semantic Web:  
2. Query answering: in the long run?

Saturation

Once per database, then maintenance

Reformulation

Once per query

Running q \( n \) times with saturation:
\[ \text{Sat} + n \ast \text{exec} (q, \text{Sat}(DB, KB)) \]

Running q \( n \) times with reformulation:
\[ n \ast \text{exec} (\text{Ref}(q, KB), DB, KB) \]

1. Saturation threshold varies significantly: 1… 10.000.000
2. Instance deletion: \( \downarrow \); instance insertion: \( \uparrow \uparrow \)
3. Schema insertion: \( \uparrow \); schema deletion: \( \uparrow \uparrow \)

New optimizations of query reformulation since.

<table>
<thead>
<tr>
<th>WWW 2012</th>
<th>EDBT 2013</th>
<th>Digiteo DW4RDF</th>
</tr>
</thead>
</table>

27-28 Nov. 2013  
Comité d’évaluation du LRI
Zoom 2: Querying description logic knowledge bases

Meghyn Bienvenu

In collaboration with: Balder ten Cate, Carsten Lutz, Riccardo Rosati, and Frank Wolter
Description logic ontologies

- *Expressive DLs (ALC, SHIQ, …)*  
  OWL 2
- *Lightweight DLs (DL-Lite, EL, …)*  
  OWL 2 profiles

 Conjunctive queries, atomic queries, path queries…
Query reformulation:
• reduce reasoning to evaluation of FO (~SQL) queries in databases
• main technique used for query answering in DL-Lite (~OWL 2 QL)

Drawback: limited applicability
• for most DLs, reformulations need not exist!
Query reformulation:
• reduce reasoning to evaluation of FO (~SQL) queries in databases
• main technique used for query answering in DL-Lite (~OWL 2 QL)

Drawback: limited applicability
• for most DLs, reformulations need not exist!

Ontology contains:
\[
\text{cardiacVentricle} \sqsubseteq \text{PartOfHeart} \\
\exists \text{partOf}.\text{PartOfHeart} \sqsubseteq \text{PartOfHeart}
\]

no reformulation of query \( \text{PartOfHeart}(x) \)!
When is query reformulation possible?

**Query reformulation:**
- reduce reasoning to evaluation of FO (~SQL) queries in databases
- main technique used for query answering in DL-Lite (~OWL 2 QL)

**Drawback:** **limited applicability**
- for most DLs, reformulations need not exist!

**Ontology contains:** 
- \( \text{cardiacVentricle} \sqsubseteq \text{PartOfHeart} \)
- \( \exists \text{partOf}. \text{PartOfHeart} \sqsubseteq \text{PartOfHeart} \)

**no reformulation of query** \( \text{PartOfHeart}(x) \)!

**Goal:** *identify those queries which admit an FO reformulation*
Testing existence of reformulations

Clear picture of complexity landscape for atomic queries

For the popular lightweight logic EL:
• PSPACE-complete / EXPTIME-complete
  • sharp contrast to tractability of standard reasoning in EL!

For Horn DLs like ELI and Horn-SHI:
• EXPTIME-complete (same complexity as standard reasoning)
• upper bound via tree automata

For expressive DLs like ALC and SHI:
• NEXPTIME-complete (slight increase over standard reasoning)
• holds also for datalog reformulations
• upper bound proof exploits recent results on CSPs
Handling data inconsistencies

Errors can cause data to be inconsistent with the ontology

- **standard semantics useless** (everything is entailed!)
Errors can cause data to be inconsistent with the ontology

- **standard semantics useless** (everything is entailed!)

ontology $\text{Prof} \sqsubseteq \text{Staff}, \text{Postdoc} \sqsubseteq \text{Staff}, \text{Prof} \sqsubseteq \neg \text{Postdoc}$

data $\text{Prof}(\text{Kim}), \text{Postdoc}(\text{Kim}), \text{Student}(\text{Joe})$

reasonable to infer $\text{Staff}(\text{Kim})$ but **not** $\text{Staff}(\text{Joe})$
Errors can cause data to be inconsistent with the ontology

- **standard semantics useless** (everything is entailed!)

  ontology \( \text{Prof} \sqsubseteq \text{Staff}, \text{Postdoc} \sqsubseteq \text{Staff}, \text{Prof} \sqsubseteq \neg \text{Postdoc} \)

  data \( \text{Prof}(\text{Kim}), \text{Postdoc}(\text{Kim}), \text{Student}(\text{Joe}) \)

  reasonable to infer \( \text{Staff}(\text{Kim}) \) but **not** \( \text{Staff}(\text{Joe}) \)

**Solution**: adopt inconsistency-tolerant semantics
Errors can cause data to be inconsistent with the ontology

- **standard semantics useless** (everything is entailed!)

\[
\begin{align*}
\text{ontology} & \quad \text{Prof} \sqsubseteq \text{Staff}, \text{Postdoc} \sqsubseteq \text{Staff}, \text{Prof} \sqsubseteq \neg \text{Postdoc} \\
\text{data} & \quad \text{Prof}(\text{Kim}), \text{Postdoc}(\text{Kim}), \text{Student}(\text{Joe})
\end{align*}
\]

reasonable to infer \(\text{Staff}(\text{Kim})\) but **not** \(\text{Staff}(\text{Joe})\)

Solution: **adopt inconsistency-tolerant semantics**

**Repair**: maximal subset of data consistent with the ontology

- **CQA semantics**: query holds irrespective of which repair chosen
- **IAR semantics**: query holds in the intersection of repairs
Inconsistency-tolerant querying

Focus: DL-Lite ontologies

Complexity analysis

• **Strong intractability results for CQA semantics**
  - also identified some tractable cases
• **Positive result for IAR semantics**
  - query reformulation always possible

New family of parameterized semantics

• Interpolate between IAR and CQA semantics
• Nice computational properties
  • same complexity as classical semantics
• Also: family of upper approximations w/ same properties
LaHDAK

Large-scale Heterogeneous Data and Knowledge

Heads:
C. Reynaud
N. Bidoit

Oak: joint team (LRI-INRIA)
Group members

Permanent faculty LaHDAK (September 1st, 2013)
11 university (4 PR – 7 MC) – 1 INRIA (DR) - 1 CNRS (CR)

Permanent faculty BD + IASI (June 30, 2013)
13 university (3 PR – 10 MC) - 1 INRIA (DR) - 1 CNRS (CR)

- 6 associated members
- 2 post-doctoral students

Origin:
- Ph.D. in Greece
- Ph.D. in Italy

- 3 engineers
  1 Digiteo: OMTE IASI-Tools
  2 INRIA : ADT Inria ODMap & CLOAK

- 13 thesis in progress

Grants:
6 French Department of Research
1 Digicosme
4 Fixed-term contracts
2 Foreign fundings

Origin:
3 M2R Paris XI
2 Master Franco-Hellenique
3 Other French M2Rs
5 Foreign university
**Objective:**
Data and knowledge management with
- complex-structured,
- semantically heterogeneous,
- massive
- uncertain,
- missing
- evolving data and knowledge

**Scientific Themes:**
- Integration of data and knowledge (N. Pernelle)
- Automated reasoning (M. Bienvenu)
- Web data management (B. Cautis)
- Massively distributed algorithms for complex data (I. Manolescu)
- Complex and evolving data transformations (M. Herschel)
1. Key discovery for Data Interlinking

Approaches to discover keys when RDF datasets contain erroneous data or data that are poorly described (i.e. almost keys, conditional keys or keys that involve chains of properties).

Ph.D. thesis of D. Symeonidou

2. Data fusion

Algorithms for data fusion that consider data quality and knowledge (e.g. disjunction, functionality) to solve the problem of possible conflicting values.

3. Evaluation of the quality of the results obtained by integration approaches

Define models and algorithms to characterize the quality of existing links.

Post-doc Qualinca hired for the topics 2 and 3
3. Mapping adaptation between biomedical Knowledge Organization Systems

An efficient approach which does not re-compute the whole set of mappings each time a KOS evolves

Ph.D. thesis of J. C. Dos Reis

4. Reasoning with semantic annotations

Objective: automatically populate an ontology by exploiting annotations generated by current annotation tools

Ph.D. thesis of C. Alec

5. A framework grouping innovative tools to integrate information from numerous, highly heterogeneous data sources

An engineer hired for 1 year
1. Diagnosability / Predictability of Discrete Event Systems

- Distributed diagnosability and testability of faulty systems
  
  **Ph.D. thesis of H. Ibrahim**, co-supervision with L. Simon (LABRI)
- SAT-based approaches to systems biology
  
  **Ph.D. thesis of M. Morterol**, co-supervision with S. Peres (Bio-Info)
  
  L. Simon (LABRI)

2. Ontology-Based Data Access

- Lightweight description logics and two main challenges
  
  - Scalability of query answering algorithms
  - Robustness to data inconsistencies

  **Ph.D. thesis of C. Bourgaux**, co-supervision with F. Goasdoue (IRISA)
1. Optimization for Web data

Queries and updates

- RDF, temporal aspects, evolution (constraints)

- Query optimization (view indexing, large set of views) – XML & relational data

  **Ph.D. thesis of I. Ileana**

- Semantic-rich semi-structured data

  **Ph.D. thesis of A. Roatis**, co-supervision with F. Goasdoue (IRISA) and D. Colazzo (Paris-Dauphine)

Collaboration with UCSD

OAKSaD

Database Group

Collaboration with UCSD

OAKSaD

Database Group

Collaboration with UCSD

OAKSaD

Database Group

Collaboration with UCSD

OAKSaD

Database Group

Collaboration with UCSD

OAKSaD

Database Group

Collaboration with UCSD

OAKSaD

Database Group

Collaboration with UCSD

OAKSaD

Database Group

Collaboration with UCSD

OAKSaD

Database Group

Collaboration with UCSD

OAKSaD

Database Group

Collaboration with UCSD

OAKSaD

Database Group
2. Social media and crowdsourcing

- Network-aware search in social media: multi-faced data, community-centric search, rich social data (semantics – XR)

  **Ph.D. thesis of R. Bonaque, cosupervision with F. Goasdoue (IRISA)**

- Discovering and refining profiles in user centric applications

  **Ph.D. thesis and post-doctoral students in 2014**

Collaboration with HKU and Univ. Singapore

Collaboration with XRCE, LIG

SocialXRSearch

KEYSTONE

ALICIA

Comité d’évaluation du LRI
Massively distributed algorithms for complex data (1)

1. Xquery processing through translation to the PACT framework (generalization of MapReduce)

   Ph.D. thesis of J. Camacho-Rodriguez, co-supervision with D. Colazzo (Paris-Dauphine)
   Post-doctoral student: F. Bugiotti

2. Efficient RDF query evaluation on Hadoop (impl. open source of MapReduce)

   Ph.D. thesis of S. Zampetakis, co-supervision with F. Goasdoue (IRISA)
   Engineer: ADT Inria CLOAK 2013-2015

3. Cloud-based warehouses for big Data Analytics

   Ph.D. thesis of D. Bursztyn and Y. Li
   co-supervision with F. Goasdoue & D. Colazzo
Massively distributed algorithms for complex data (2)

4. Workflow optimization through common sub-expression identification

Post-doctoral student: S. Roy Chowdhury

5. Advanced cloud-Based Services for the Semantic Web

Cloud-based Semantic Web storage, reasoning, and complex processing (workflows, static analysis)

Preposa ANR project

Collaboration with Univ. Rennes and Hong Kong

6. Business models for the cloud

Pluri-disciplinary research – in collaboration with researchers in social sciences at Paris-Sud
1. Provenance-enabled transformation evolution

- RDF data and queries based on RDF data
  Support to the analyze-fix-test cycles in a transformation lifecycle

**Ph.D. student : K. Tzompanaki (2nd year)**

- Semi-automatic support for transformation development and maintenance

Planned for submission as an ERC Starting grant

2. Provenance for data-centric business process verification

Instrumentation and exploitation in resolving conflicts
Complex and evolving data transformations

3. Semi-automatic scientific workflow design and configuration

Provenance for:
- auto-completing workflows
- repairing workflows that do not work any more.

Collaboration with the BioInfo group

Planned for submission as an ANR-DFG project
**Conclusion (1)**

- **Big Data**
- **Internet**
- **LOD**
- **Social Web**
- **Cloud**

- **Scalability**
- **Interoperability**
- **Relevance**
- **Efficiency**
- **Crowdsourcing**

- **Distributed systems**
- **Semantic Web**
- **Workflow**

- **Massively distributed computing**
- **Optimization**

- **27-28 Nov. 2013**
- **Comité d’évaluation du LRI**
Conclusion (2)

- PAGODA
- DynaMO
- Datalyse
- Europa
- SocialXRS
- Qualinca
- DW4RDF
- KEYSTONE
- ALICIA
- IASI-Tools
- DATE
- CBOD
LaHDAK: an integration of data management and Semantic Web techniques