Welcome to the Dampierre Castle Seminar
1rst OAK seminar

Agenda of the Day

9h30 Welcome coffee + pastry

10h00 Overview of the last year (Nicole + Ioana)

Restructuring, arrivals, departures

Who does what: responsibilities within the team, Paris Sud, Inria etc.

10h15 Melanie Herschel (BD/OAK):

Foundations and Algorithms to Compute the Provenance of Missing Data

10h45 Break

11h00 Philippe Rigaux (Internet Memory):

Large-scale Web data management at InternetMemory

11h30 "Present me a problem"

12h30 Lunch

14h30 Paolo Atzeni (U. Roma Tré, Italy):

Management of Heterogeneous Data in Traditional and non Traditional Database

15h00 Visite Château de Dampierre

16h00 Coffee break

16h15 Konstantinos Karanasos:

"How to hunt for a post-doc"

16h45 Team grants and software (Nicole + Ioana)

17h00 Choose your activity:

* Work meeting (permanent Oaks, if energy left)

* Free time!

19h00 Dinner

Restructuring, arrivals, departures

OAK **is born the 1rst of April 2012 !**

Database Optimisations and Architectures for Complex large data

OAK

Short explanation/recall

- We are a joint (Inria-Paris Sud) team
- Inria teams are *created* then must become project-teams.
- (A project team lasts 4 + 4 + 4 years at most.)
- We were in the Leo team from Jan 2010 to Feb 2012 (current OAK + most of the IASI LRI team)
- In October 2011 Inria decided Leo would not become a project: too big, too little focus. They asked for a smaller alternative.
- We proposed OAK: data management

Restructuring, arrivals, departures

OAK is born the 1rst of April 2012 !

Database Optimisations and Architectures for Complex large data

OAK topics:

- Efficient techniques for XML data management
- Efficient management of Semantic Web data encoded in RDF
- Emerging data formats
- Cloud-based platforms
- Managing data transformations

Reconfiguration of the LRI Database team:

Nicolas Spyratos's retirement

Nicole is in charge of the database team since March

Restructuring, arrivals, departures

Permanent members

Nicole	Bidoit-Tollu	(PRX-UPS, Associate team leader)	
Dario	Colazzo	(MdC-UPS, HdR)	
François	Gouasdoué	(MdC-UPS, HdR)	
Melanie	**Herschel**	**(MdC-UPS)**	**since January 1rst**
Ioana	**Manolescu**	**(DR-Inria, Team leader)**	

PhD Students

Mohamed	Amine Baazizi	(Nicole, Dario)	Defense early Sept
Jesús	**Camacho Rodríguez**	(Dario, Ioana)	1rst year
Konstantinos	Karanasos	(Ioana, François)	Defense June 29th
Asterios	Katsifodimos	(Ioana)	3rd year
Julien	Leblay	(François, Ioana)	2nd year
Noor	Malla	(Nicole, Dario)	Defense mid Sept
Alexandra	**Roatis**	(Dario, François, Ioana)	1rst year
Federico	Ulliana	(Nicole, Dario)	Defense November
Stamatis	**Zampetakis**	(Ioana, François)	Starting in October

Engineers

Andrés	**Aranda Andújar**	Ioana, François	Since October 2011
Tushar	**Ghosh**	Nathalie Pernelle (IASI)	Since December 2011
André	Amorim Fonseca	Philippe Chatalic (IASI)	Until September 2012

Restructuring, arrivals, departures

Interns

Kuldeep	**Reddy**	(Zoi, Ioana)	IIT Madras
Karan	**Aggarwal**	(Asterios, Konstantinos, Ioana)	IIT Roorkee
Abishek	**Choudhary**	(Melanie)	IIT Delhi
Yulun	**Li**	(Ioana, Zoi)	Ecole Centrale & Aerospatial U. Pekin

Post-Docs

Zoi	**Kaoudi**
Marina	**Sahakyan**

External Collaborators

Serge	Abiteboul	(DR, Dahu team)
Philippe	Rigaux	(PR CNAM on leave at InternetMemory)
Marie-Christine	Rousset	(PRX, UJF)
Virginie	Thion-Goasdoué	(MdC CNAM)
Emmanuel	Waller	(MdC, Database team LRI)

Administrative assistant

Céline	Halter	(currently on sick leave, being replaced by Alexandra Merlin)

Who does What ?

Nicole	Head of the LRI DB team, CS Depart. vice-president (until 30 January 2012), LR-Lab Committee, Head of the Master program, IAC, ICT-Lab Master DSS, Doctoral School Committee, CCSU, Labex and Idex Working Group, Recruiting Committees, IEF Junior Committee, Chair of the G. Kahn PhD Price Co-chair of the Database Summer School
Dario	CCSU (Board), LRI Hardware Committee, Codex grant (LRI), KIC Europa
François	LRI Website Committee, Miage student coordinator
Melanie	LRI Software Committee; KIC DataBridges 2013, it special issue editor "Data Integration"
Ioana	Head of the Inria Oak team; 7 (small ☹) grants to manage; team mailing lists Editor in chief, ACM SIGMOD Record; associate editor ACM TOIT 4 (co-)chairing of workshops/conference [tracks] in 2012 Member of Inria Saclay ADT committee; comités de projet 1/month Scientific coordination from Inria for SAP collaboration
Emmanuel	Office allocation Committee (Céline needs to know!)
Andrés & Jesús	(Inria) hardware
Zoi	Help Ioana managing KIC activities (Europa, DataBridges 2012)
Konstantinos	Oak Seminars; help Ioana managing ANR Codex and DataBridges; LaTeX guru
Asterios	Help Ioana managing ANR DataRing, ConnectedCities; software guru (ruby prophet)
Julien	Oak Website (http://team.inria.fr/oak)
Céline	Member list on Oak Web site; mailing lists

Research Project : who sponsors our hard work?

Ongoing

ANR Codex (finishing end of June)

ANR DataRing (finishing end of September)

ANR DataBridges, ANR ConnectedCities (finishing end of September)

KIC Cloud Europa (finishing end of 2012, re-submitted)

KIC DigitalCities DataBridges (finishing end of 2012, re-submitted)

Digiteo DW4RDF

Submitted

ANR ODIN, PAI Pessoa, PAI Tournesol

KIC DataBridges, KIC Europa, KIC Massive Shared Data Applications

Under construction

EDF Time series & Cloud

SAP DaSL

AAP Cloud 1 -- Grenoble

AAP Cloud 2

RAPID grant with DSA/DGSR

FET Flagship (?)

ACM SIGMOD 2012

KIC Cloud "Europa"

- Project started in 2012, in the context of the KIC Cloud action line.
- Involved European sites : TU Berlin (global coordinator),, PSUD-INRIA Saclay, Swedish Institute of Computer Science, TU Deft, University of Trento.
- Aim: building a data intensive computing infrastructure based on the Stratosphere platform (developed @TU Berlin).
- Local coordination: Dario and Ioana.
- Our 2012 task: distributed indexing for XML and RDF, XQuery to PACTL compilation
- Our 2013 task (if renewal accepted): work load optimization, parallel evaluation of SPARQL queries via RDF Schemas
- Active collaboration with the Stratosphere team at TU Berlin.

Open Data INtelligence

Expected funding (3 years)

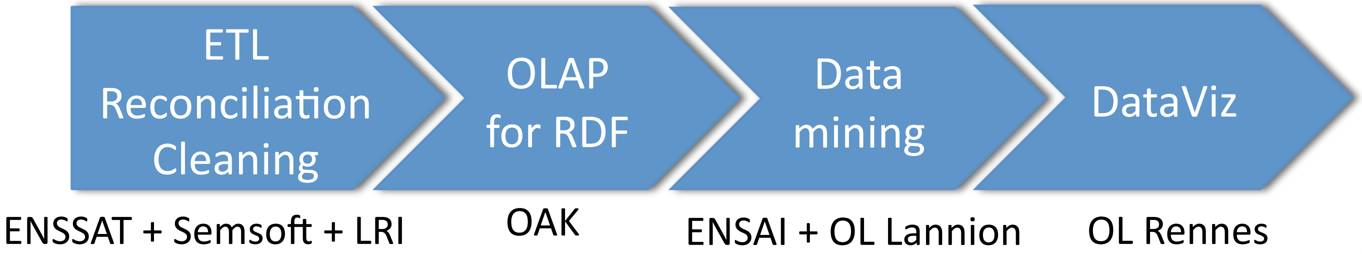
- ANR *Modèles Numériques*

Consortium

- ENSAI (A. Bidault et al.)
- ENSSAT (H. Jaudoin et al.)
- LRI/INRIA (Alexandra, Dario, François, Ioana et al.)
- Orange Labs Lannion & Rennes (J. Royan et al.)
- SemSoft (F. Paulus et al.)

Goal

- Building a complete framework for Open/RDF Data analysis



Managing annotated Web data

Expected funding (2~3 years)

- DGA/DGRI/DGSE *RAPID*

Consortium

- LRI-OAK (François, Ioana, Julien et al.)
- IRD (Laure Berti-Equille et al.)
- SemSoft (François Paulus et al.)

Goal

- Managing XML documents with RDF annotations

Building annotations based on ontologies and quality assessment

Management of XR documents

LRI + IRD

OAK

DataBridges
Creating, Enriching, and Exploiting Data for Digital Cities
(KIC Activity Proposal 2013)

@OAK

(1) Incorporate provenance (Why- & Why-Not) into RDFGears

(2) Apply models for warehouse-style analysis of RDF data on LOD

Open Data Generation

from multimedia

from text

from data streams

from relations

Data Integration

data linking

data fusion

data quality

data provenance

context data

Data Exploitation

data analytics

faceted search

visualization

recommendation

KIC EIT ICT Labs activity "Massive Shared Data Applications"

New activity, submitted for 2013

Within the "Digital Cities" action line

Coordinated by Sandro Battisti (Trento RISE)

Partners:

- Telecom Italia, Politecnico di Milano (Italy)
- SICS (Sweden)
- Alcatel, DataPublica, Inria (OAK), Institut Telecom (France)
- Novay (Netherlands),

Goal: foster development of new services based on shared data

Data to be archived on a Web-based platform

"*The Activity will set the backbone of a Europe-based ecosystem for Shared Data, involving the EIT ICT Labs nodes in Italy, France, Sweden and the Netherlands, thus enabling radical service innovations based on a semantic enrichment of data*"

"Datalyse" project (formerly known as: "User Big Data")

National French grant proposal, to be submitted by July 13, 2012

Answers the "Appel à Projets Cloud 3 – Big Data" call initiated by the French government

- The lead must be a company
- There may or may not be research labs
- The company must create wealth and/or jobs and it must reimburse or otherwise financially interest the state

Datalyse lead: Business & Decision (BI company from Grenoble)

Other partners: LIG, Inria Grenoble + Lille (OS/Data Mining/IA/DB)

Sihem Amer-Yahia, Marie-Christine Rousset, Alex. Termier

Purpose: enrich B&D's business intelligence suit for **data** acquisition, storage, indexing, integration, reconciliation, mining, and exploitation typically for marketing purposes

"Big Data Lab" project

National French grant proposal, to be submitted by July 13, 2012

Answers the "Appel à Projets Cloud 3 – Big Data" call initiated by the French government

"Big Data Lab" lead: Exalead (part of Dassault Systèmes)

Other partners: Inria Lille, Sophia, Rennes, Bordeaux (semantics / image / NLP / graph visualization…)

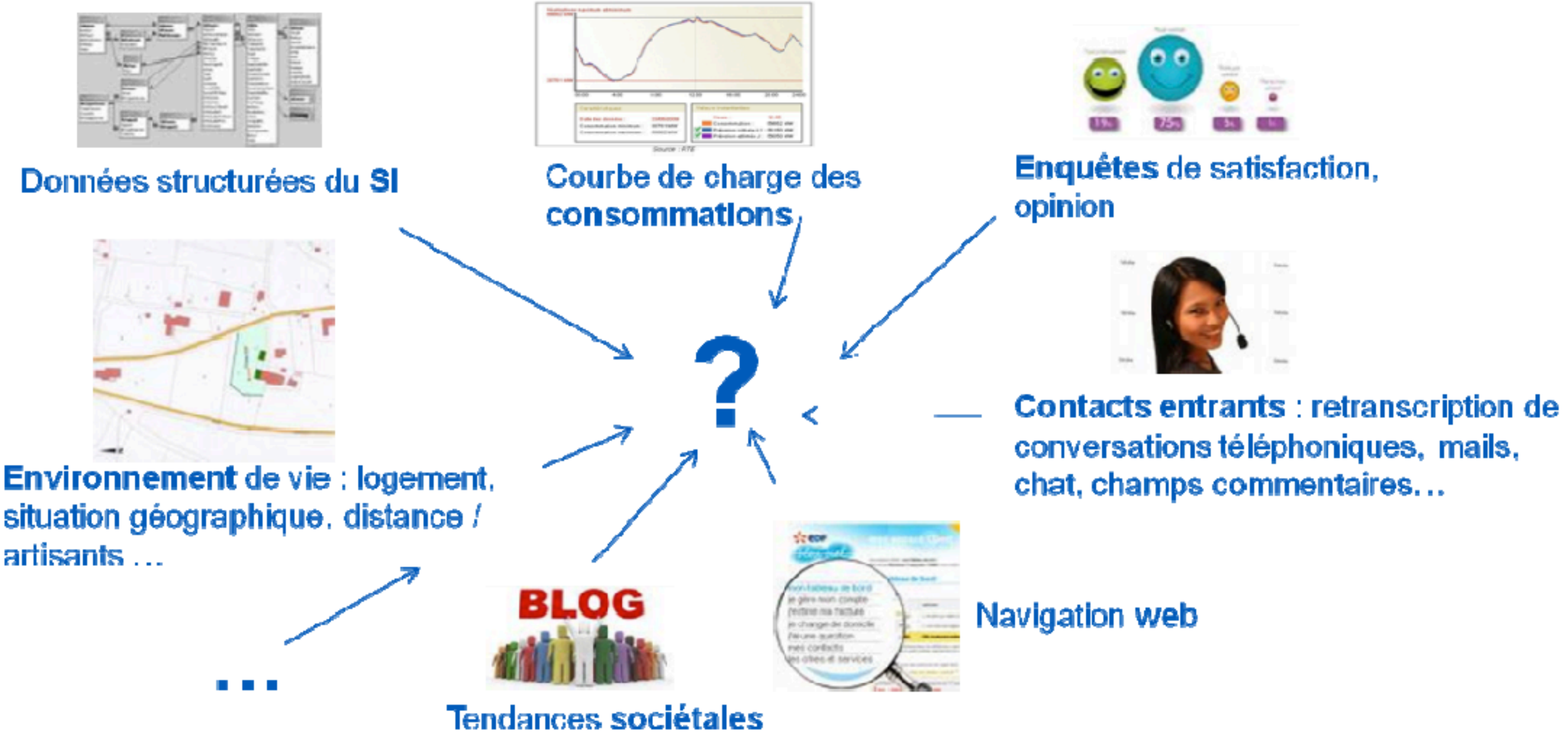
Purpose: enrich Exalead's data acquisition & processing pipe, in exchange for getting to play with their data "at Web scale"

- They start with Web data (pages) → crawler from former Gemo work
- They analyze the text, identify entities, classify, annotate, index, draw etc.

Discussions with EDF

- Context: Inria partnership with EDF
- EDF data:
  - Electricity usage data = time series (Linky smart counter…)
  - User (customer) data: age, revenue, history, address, … stuff they could take from the Internet
- Problems they consider
  - Time series data management
    - Find specific patterns in the consumption ("who has an electric water heater")
    - Deal with periodicity over a day, week, season, year
  - Customer data integration – very early on in the thinking process

Discussions with EDF



Données structurées du SI

Courbe de charge des consommations

Enquêtes de satisfaction, opinion

Contacts entrants : retranscription de conversations téléphoniques, mails, chat, champs commentaires…

Navigation web

Tendances sociétales

Environnement de vie : logement, situation géographique, distance / artisans …

…

EDF R&D ICAME

Discussions with EDF

- EDF data:
  - Electricity usage data = time series (Linky smart counter…)
  - User (customer) data: age, revenue, history, address, … stuff they could take from the Internet
- Problems they consider
  - Time series data management
    - Find specific patterns in the consumption ("who has an electric water heater")
    - Deal with periodicity over a day, week, season, year
  - Customer data integration – very early on in the thinking process
- Status
  - We may be weekly involved in co-supervising an intern in 2012
  - Possibly surveys etc. on data integration, provenance, lineage

Discussions with SAP

- Context: Inria partnership with SAP
- They have
  - Big Business Intelligence (BI) operations: try to make it easy for the "end user" (business user, not the engineer) to specify a query / a cube
  - (A new in-memory database called Hana…)
  - An in-house language called DaSL (Data Structured Language) for performing complex schema-to-schema transformations
    - As opposed to SQL which is schema-to-table
- We could do for them ☺
  - Formal semantics and optimization for DaSL
- Status: hope to hire a post-doc for fall 2012, discussions ongoing

PAI Pessoa DaCleaC

- PAI: Programme d'Action Intégrée (small mobility-only EU grant)
- PAI Pessoa: France-Portugal
- DaCleaC: Data Cleaning in the Cloud
  - Dario, Melanie, Ioana; Jesús, Andrés
  - Portugal:Instituto Superior Técnico, Technical University of Lisbon Helena Galhards, Bruno Martins, Emanuel
- Status: submitted in May, results in December
- If it is granted it will last 2 years
- Purpose: compiling declarative data cleaning programs (on XML) for cloud-based execution

Software Prototypes

Amada	Jesús
RDF-Hadoop	Zoi
RDFViewS	Julien
ViP2P	Asterios
XUpQDep	Federico
XUpOp	Marina
XUpIn	Amine

AMADA: Web Data Repositories in the Cloud

- **Scalable Web data store based on off-the-shelf commercial cloud services (in particular, using the Amazon Cloud)**
  - Elastic XML and RDF data storage
  - Efficient querying operations
  - Minimize resource usage ➡ Total work translates into monetary costs
- Indexes help to decide which documents are concerned by a given query
  - No need to run the query processor over the complete data set
  - Multiple indexing strategies with different levels of detail
- System fully implemented in Java 6
  - Amazon Web Services SDK for Java v1.2.14
  - XML query processor from our ViP2P project (http://vip2p.saclay.inria.fr)
  - Standard RDF-3X query processor (http://code.google.com/p/rdf3x/)
- People involved: A. Aranda-Andújar, F. Bugiotti, J. Camacho-Rodríguez, D. Colazzo, F. Goasdoué, Z. Kaoudi, and I. Manolescu

RDF Hadoop

A scalable RDF repository in Hadoop

People involved:

Francois Goasdoue,
Zoi Kaoudi,
Ioana Manolescu,
Jorge Quiane (Saarland University) and
Kuldeep Reddy

Libraries used:

Apache Hadoop (http://hadoop.apache.org/)

Implementation in embryo stage

RDFViewS

François Goasdoué, Julien Leblay, Konstantinos Karanasos, Ioana Manolescu

Giving DBAs a tool to fine-tune the storage of RDF data for fixed SPARQL workloads

Written in Java 6.0

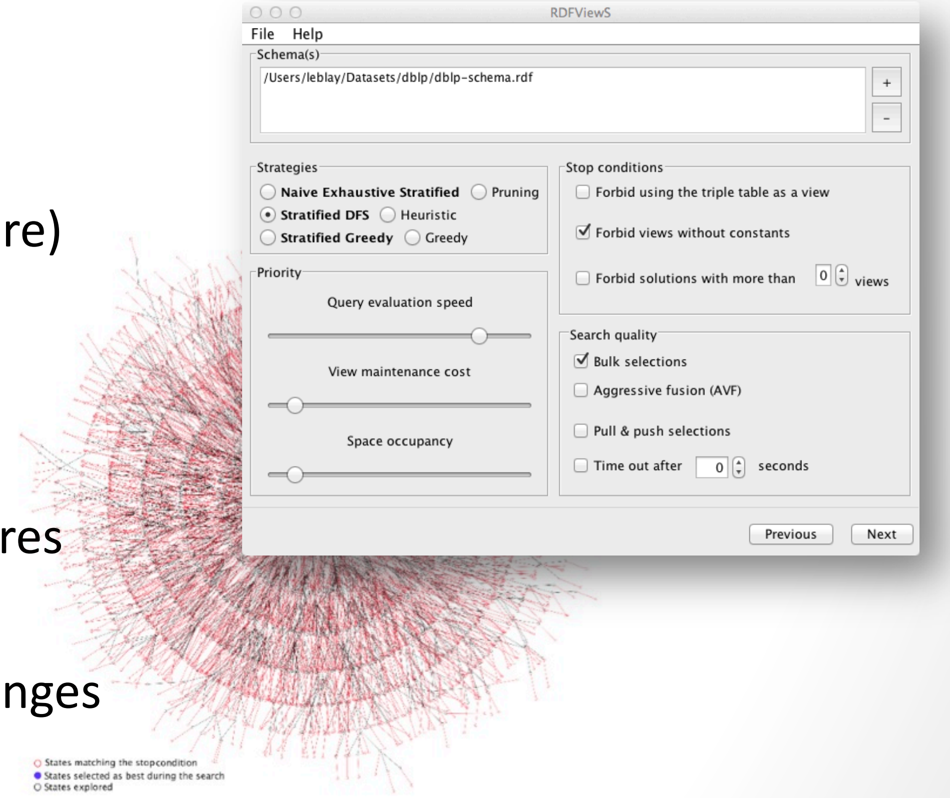
~45K lines of code

Postgresql 8.0 (underlying store)

Stable (but brittle)

Possible extension

- support for more RDF stores
- parallelization
- adapting to workload changes



Schema(s)

/Users/leblay/Datasets/dblp/dblp-schema.rdf

Strategies

○ **Naive Exhaustive Stratified** ○ Pruning
● **Stratified DFS** ○ Heuristic
○ **Stratified Greedy** ○ Greedy

Priority

Query evaluation speed

View maintenance cost

Space occupancy

Stop conditions

☐ Forbid using the triple table as a view
☑ Forbid views without constants
☐ Forbid solutions with more than 0 views

Search quality

☑ Bulk selections
☐ Aggressive fusion (AVF)
☐ Pull & push selections
☐ Time out after 0 seconds

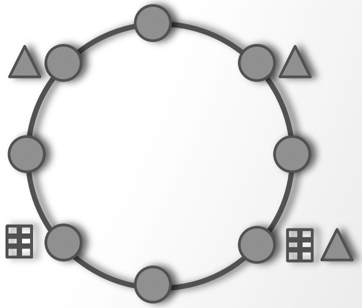
○ States matching the stopcondition
● States selected as best during the search
○ States explored

ViP2P: Views in P2P

Scalable & Efficient XML Management using Distributed Materialized Views over DHT Networks

- **Written in Java (+ some bash & ruby ☺)**
  - 70.000 lines of code (and growing!)
- **Depends (mainly) on**
  - BerkeleyDB, Apache Commons, Log4j, FreePastry, Piccolo…
- **Status:** Under development/extension
- **More than a platform!**
- **Serves as codebase for**
  - XQuery Rewriting Using Views
  - Annotated XML documents
  - Algebraic View Maintenance
  - Automatic XML View Selection
  - XML/RDF hybrid data models
  - Cloud-based XML Management
  - Multi-level views based pub/sub
  - …

- **Proven scalability**
  - Hundreds of peers
  - Hundreds of GBs of data
  - Real deployment on Grid5000
- **People coded for ViP2P**
  - Ioana Manolescu
  - Spyros Zoupanos
  - Alin Tilea
  - Jesús Camacho-Rodríguez
  - Konstantinos Karanasos
  - Asterios Katsifodimos
  - Julien Leblay
  - Alexandra Roatis
  - Stamatis Zampetakis
  - Martin Goodfellow
  - Domenica Sileo
  - Silviu Julean
  - Varunesh Mishra
  - Karan Agrawal
  - …



XML Query-Update Independence Checker (aka XUpIn ☺)

- XUpIn is **a static analyzer** for detecting XML **query update independence**, in the presence of schema

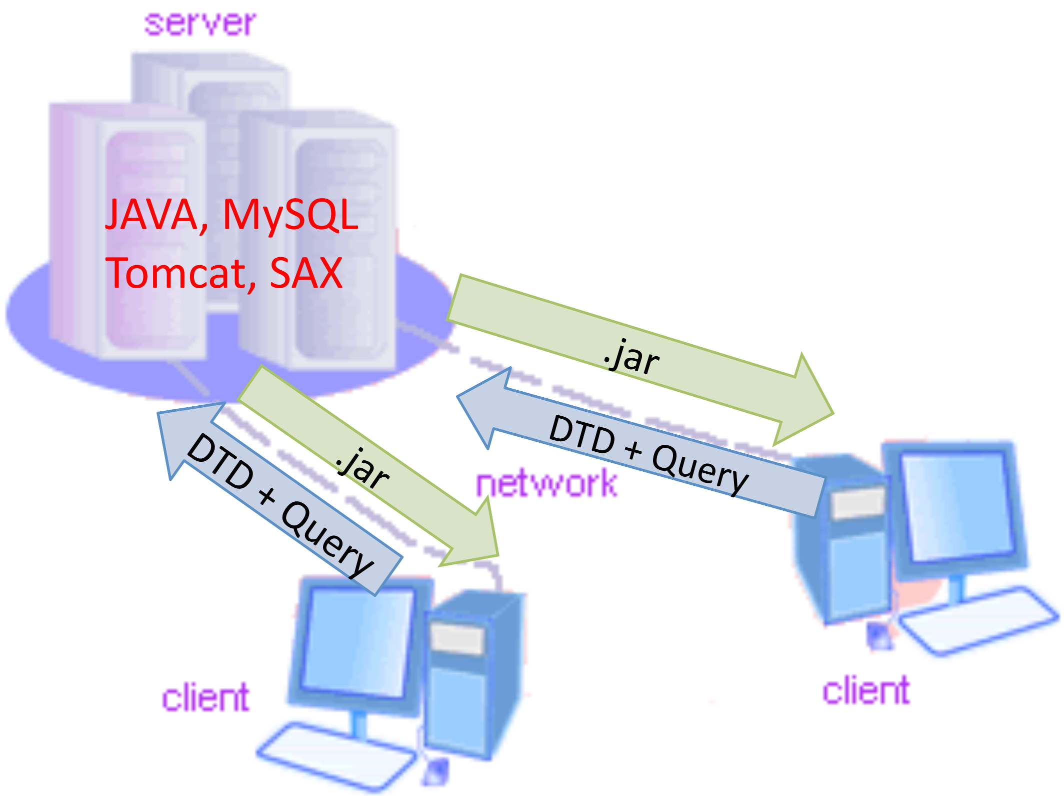
- Handles XML Schema, core XQuery + XPath + updates

- Written by myself (Federico) in Java, small project
  - < 20 classes, few K lines of code

- **In the future**, may add other features
  - attributes, arithmetic, negation, type-switch, function calls
  - update transform expressions
  - also, try to replace schemas with dataguides

Web based XQuery update evaluation

A. Baazizi, N. Bidoit, D. Colazzo, N.Malla, M. Sahakyan

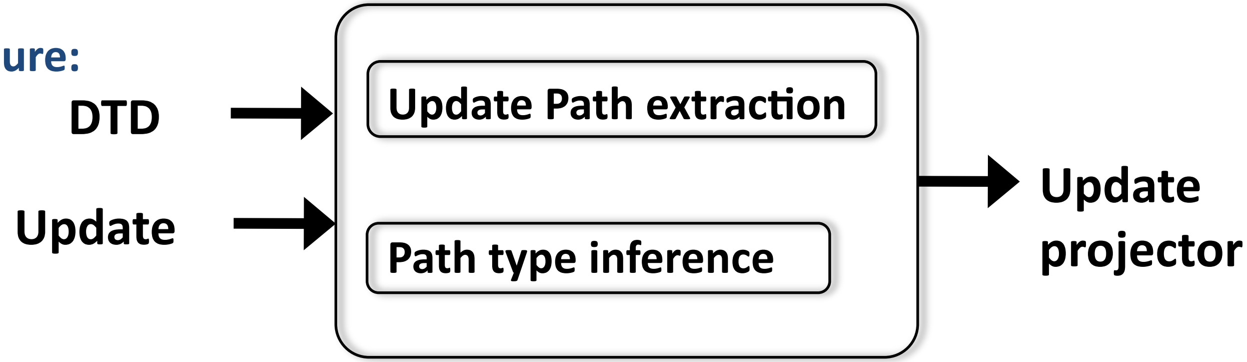


XUPIn: XML Update Projector Inference

Participants: Alessandro Solimando (intern july-sep 2011),
Mohamed-Amine Baazizi, Dario Colazzo, Nicole Bidoit

Purpose: Static analysis of XUF updates using schema – inference of update projectors for *XUpOp* (cf. M. Sahakyan's slide)

Architecture:



Features: Java 1.6, ~30 K lines of code,
fully compatible with W3C standards, stable version

Extras: visualization of path-type inference results,
dealing with XSDs

Wrap-up of the 2011-2012 year

Work has been hard (1)

Work has been hard (2)

Work has been hard (2)

Work has been hard (3)

Summer school BDA 2012

We will keep working very hard in the short term

VLDB 2012

We will keep working very hard in the near future

BDA 2012

Ioana: program chair!

We only ask to work harder

Let the hard day of work start!

Groups for "Present me a problem"

1. Ioana/Tushar/Noor/Amine/Philippe
2. Nicole/Julien/Alexandra/Paolo
3. François/Abishek/Marina/Jesús
4. Dario/Karan/Konstantinos/Francesca
5. Melanie/Federico/Andrés/Zoi