LEO

Distributed and heterogeneous data and knowledge

Ioana MANOLESCU-GOUJOT

LEO

INRIA Saclay– Île-de-France

Plan

1. Leo team
2. Leo research themes
3. Perspectives
4. Wrap-up

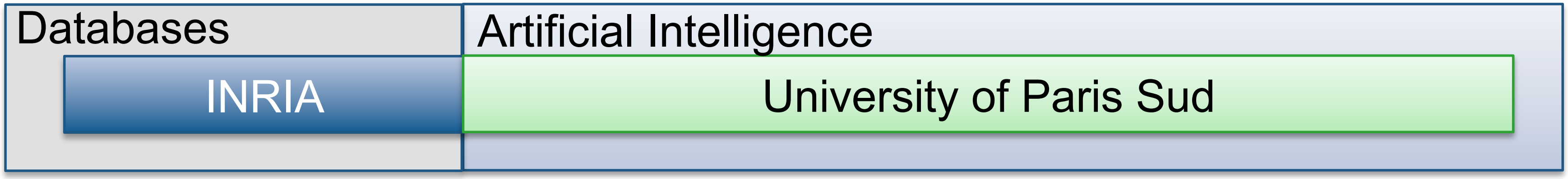
1

LEO team

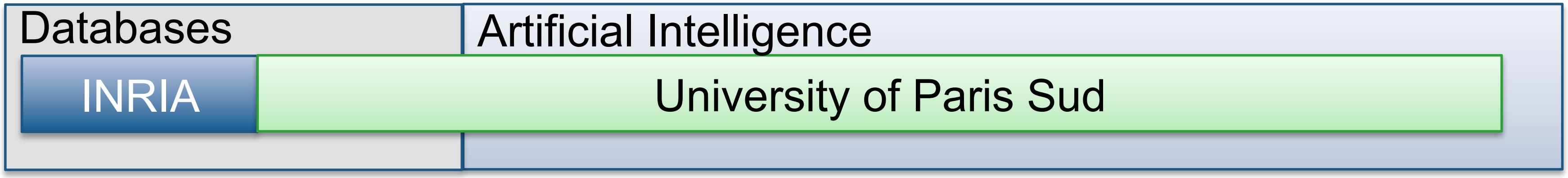
Short history

Gemo team headed by Serge Abiteboul: 2002 – mid 2009

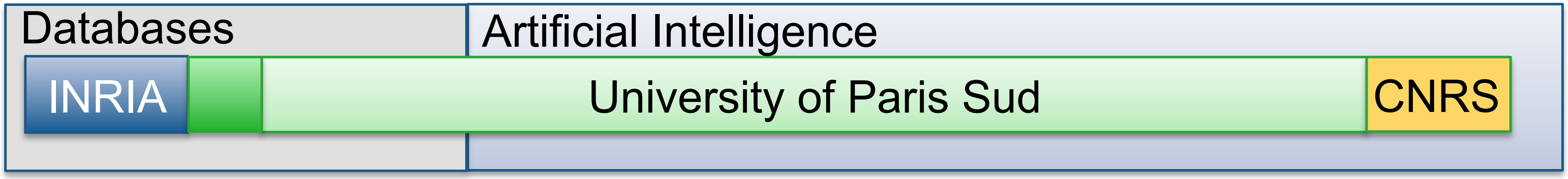
Mid 2009: Ioana takes the lead of Gemo



Leo created as a team on January 1st, 2010



Leo today



Leo as of January 2012

38 people, **13** permanent

INRIA senior researcher

Ioana Manolescu

CNRS junior researcher

Meghyn Bienvenu

U. Paris Sud professors

Nicole Bidoit, Philippe Dague,
Chantal Reynaud

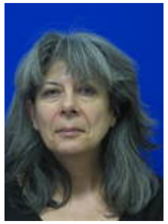
U. Paris Sud assistant professors

Philippe Chatalic, Dario Colazzo,
François Goasdoué, Melanie Herschel,
Nathalie Pernelle, Brigitte Safar,
Fatiha Sais, Laurent Simon

Post-doc: Zoi Kaoudi, Gianluca Quercini

17 PhD students

4 junior INRIA engineers



2

LEO research themes

The data wave

Volume of digital data keeps increasing

- Some areas of science are facing hundred- to thousand-fold increases in data volumes… compared to the volumes generated only a decade ago”

  Bell, Hey and Szalay, *Beyond the Data Deluge*, Science, 2009

Web is the world’s single biggest repository

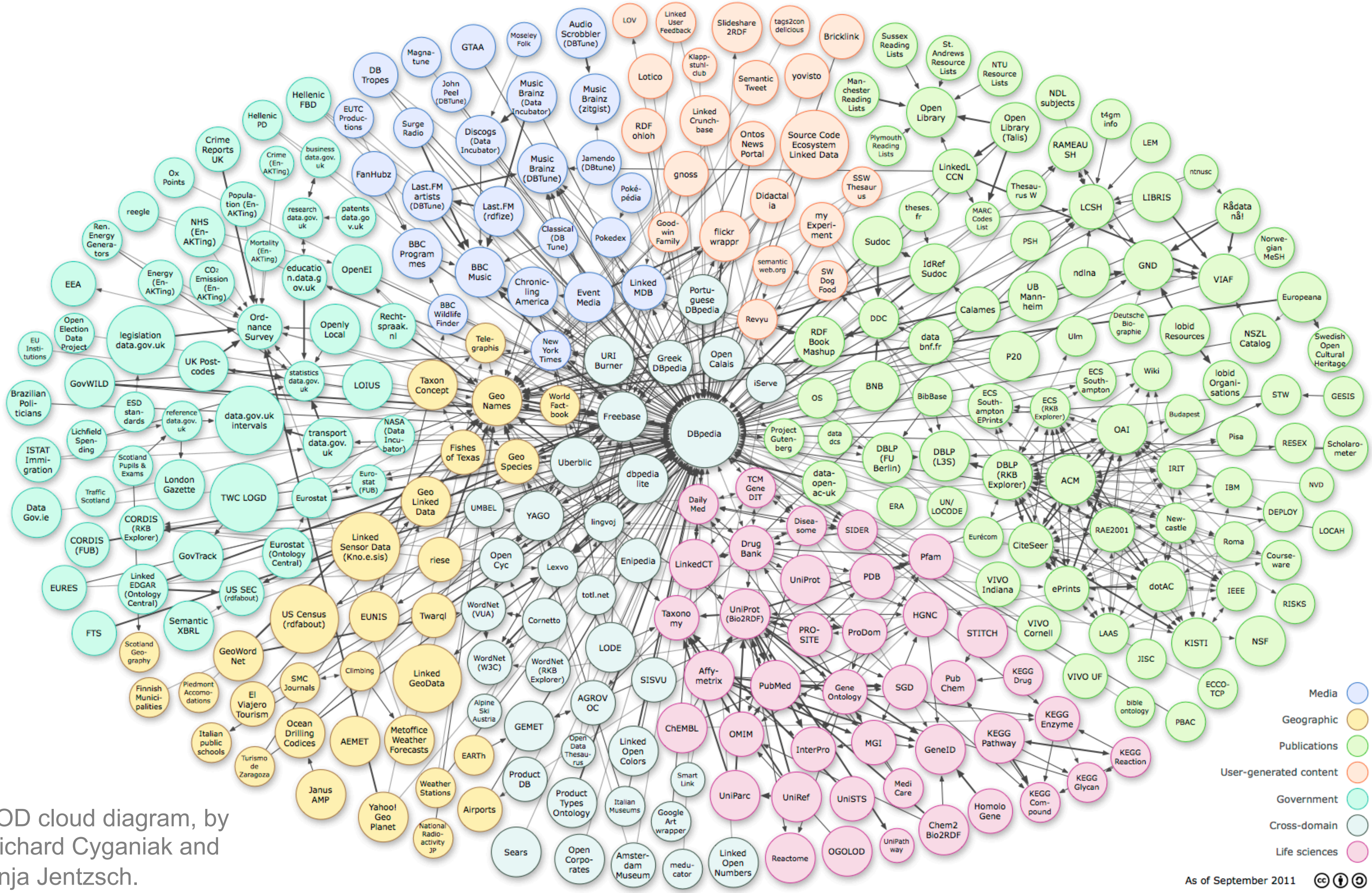
- “Every two days we create as much information as we did up to 2003. Most of it is user-generated data”

  Eric Schmidt, former CEO of Google, April 2011
- Humans memorize where to find the information (on the Web) rather than the information

  Sparrow, Liu and Wegner, Science, 2011

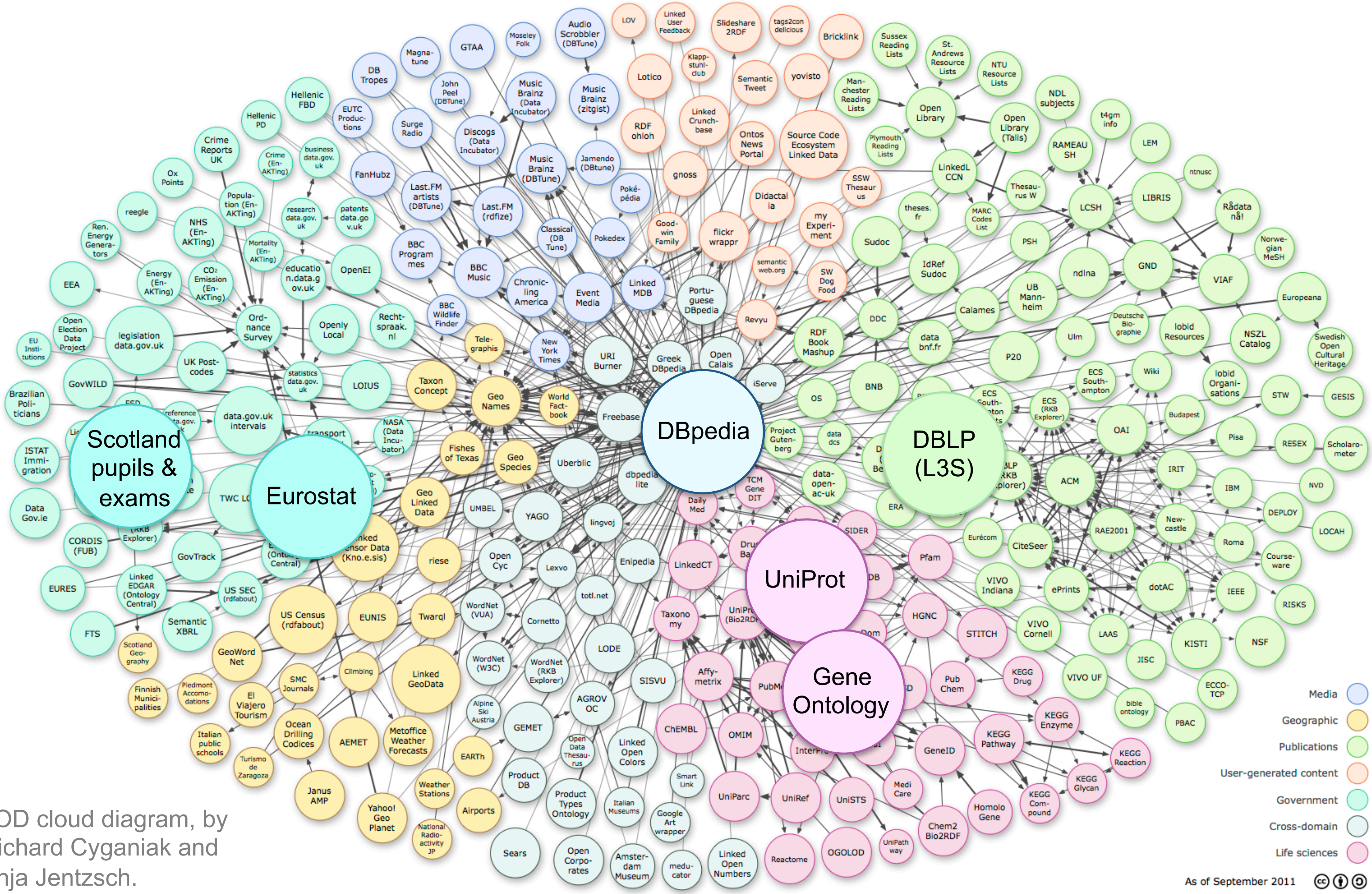
We focus on the scale and heterogeneity of data and knowledge

Large heterogeneous data: Linked Open Data



LOD cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

Large heterogeneous data: Linked Open Data



LOD cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

Leo research objectives

1. Expressive models for data and knowledge

2. Taming data and knowledge heterogeneity

3. Efficient platforms for data and knowledge
   - Large volumes of data
   - Distribution: P2P → cloud

3

LEO theme 1: expressive models for data and knowledge

Expressive models for data and knowledge

Formally modeling, understanding and reasoning about heterogeneous and distributed data

We mostly focus on Web data

- Structured objects: XML (XML Schema, XQuery); JSON
- Semantic Web data: RDF (RDF Schema, Description Logics)
- Documents annotated with semantics: XML and RDF brought together
- Associated computation models, backed by logic formalisms

Core competencies:

- Nested object algebras
- XML type systems
- Logical models for (distributed) semantic inference (OWL2)
- Formal methods for monitoring and diagnosing distributed systems

Expressive models for data and knowledge

Formally modeling, understanding and reasoning about heterogeneous and distributed data

We mostly focus on Web data

- Structured objects: XML (XML Schema, XQuery); JSON
- Semantic Web data: RDF (RDF Schema, Description Logics)
- Semantically annotated documents: XML and RDF brought together
- Associated computation models, backed by logic formalisms

Amine Baazizi, Nicole Bidoit, Dario Colazzo:
Efficient encoding of temporal XML documents. TIME 2011

National grant CODEX (“Coordination, Dynamicity and Efficiency for XML”), I. Manolescu

Dario Colazzo, Giorgio Ghelli, Carlo Sartiani:
Schemas for safe and efficient XML processing. IEEE ICDE 2011 (tutorial)

Serge Abiteboul, Meghyn Bienvenu, Alban Galland, Emilien Antoine*: A rule-based language for web data management*. ACM PODS 2011

4

LEO theme 2:
Taming heterogeneity

Taming heterogeneity

Schema / ontology heterogeneity: different viewpoints, organizations

Data heterogeneity: different representation formats, errors in the data

Diffferent scenarios require different techniques to handle heterogeneity

- Identify and exploit relationships
  - Ontology alignment → mappings → mapping refinement → ontology evolution
- Eliminate heterogeneity by unifying data and ontologies
  - Entity reconciliation (*aka* data cleaning) iterating over the data and schema

Taming heterogeneity

Schema / ontology heterogeneity: different viewpoints, organizations

Data heterogeneity: different representation formats, errors in the data

Diffferent scenarios require different techniques to handle heterogeneity

- Identify and exploit relationships
  - Ontology alignment → mappings → mapping refinement → ontology evolution
- Eliminate heterogeneity by unifying data and ontologies
  - Entity reconciliation (*aka* data cleaning) iterating over the data and schema

Y. Mrabet, N. Bennacer, N. Pernelle, M. Thiam. *Supporting Semantic Search on Heterogeneous Semi-structured Documents*. CAISE 2010

F. Hamdi, C. Reynaud and B. Safar. *Pattern-based Mapping Refinement*. EKAW 2010

National grant GeOnto ("Geographic Ontology Alignment"), C. Reynaud

Regional grant SHIRI ("Heterogeneous Information Retrieval"), N. Pernelle

Contract with Thales (N. Pernelle, F. Sais)

5

LEO theme 3: efficient platforms for Web data and knowledge

Efficient platforms for Web data management

1. Maximize the performance of the storage: materialized views
   - Expressive XML views: multiple returned data items
   - Recommend RDF views based on workload and implicit data
2. Reduce memory and processing costs by focusing on the necessary subset of the input: type projectors

Efficient platforms for Web data management

1. Maximize the performance of the storage: materialized views
   - Expressive XML views: multiple returned data items
   - Recommend RDF views based on workload and implicit data
2. Reduce memory and processing costs by focusing on the necessary subset of the input: type projectors

F. Goasdoué, K. Karanasos, J. Leblay and I. Manolescu.
View Selection in Semantic Web Databases, PVLDB 2011

I. Manolescu, K. Karanasos, V. Vassalos and S. Zoupanos.
Efficient XQuery Rewriting using Multiple Views, IEEE ICDE 2011

A. Baazizi, N. Bidoit, D. Colazzo, N. Malla, M. Sahakyan.
Projections for XML Update Optimization, EDBT 2011

A. Bonifati, M. Goodfellow, I. Manolescu and D. Sileo.
Algebraic Incremental Maintenance of XML Views, EDBT 2011

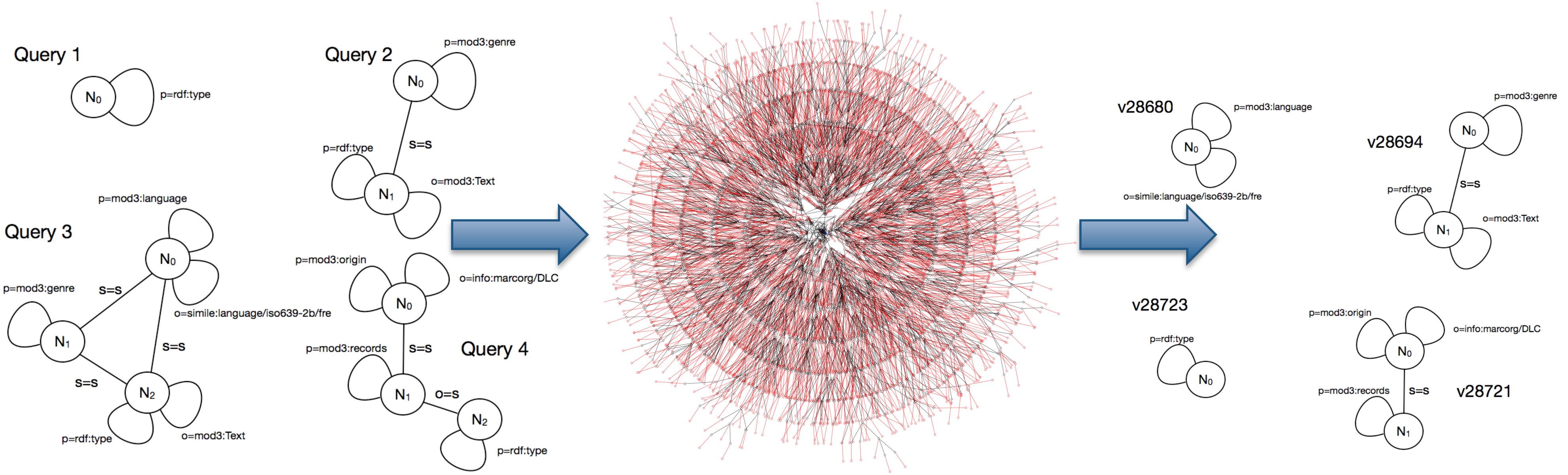
National grant
CODEX

View selection in Semantic Web databases (PVLB 2011)

Input: RDF database D, RDF Schema S, workload $\{Q_1, Q_2, \ldots, Q_n\}$

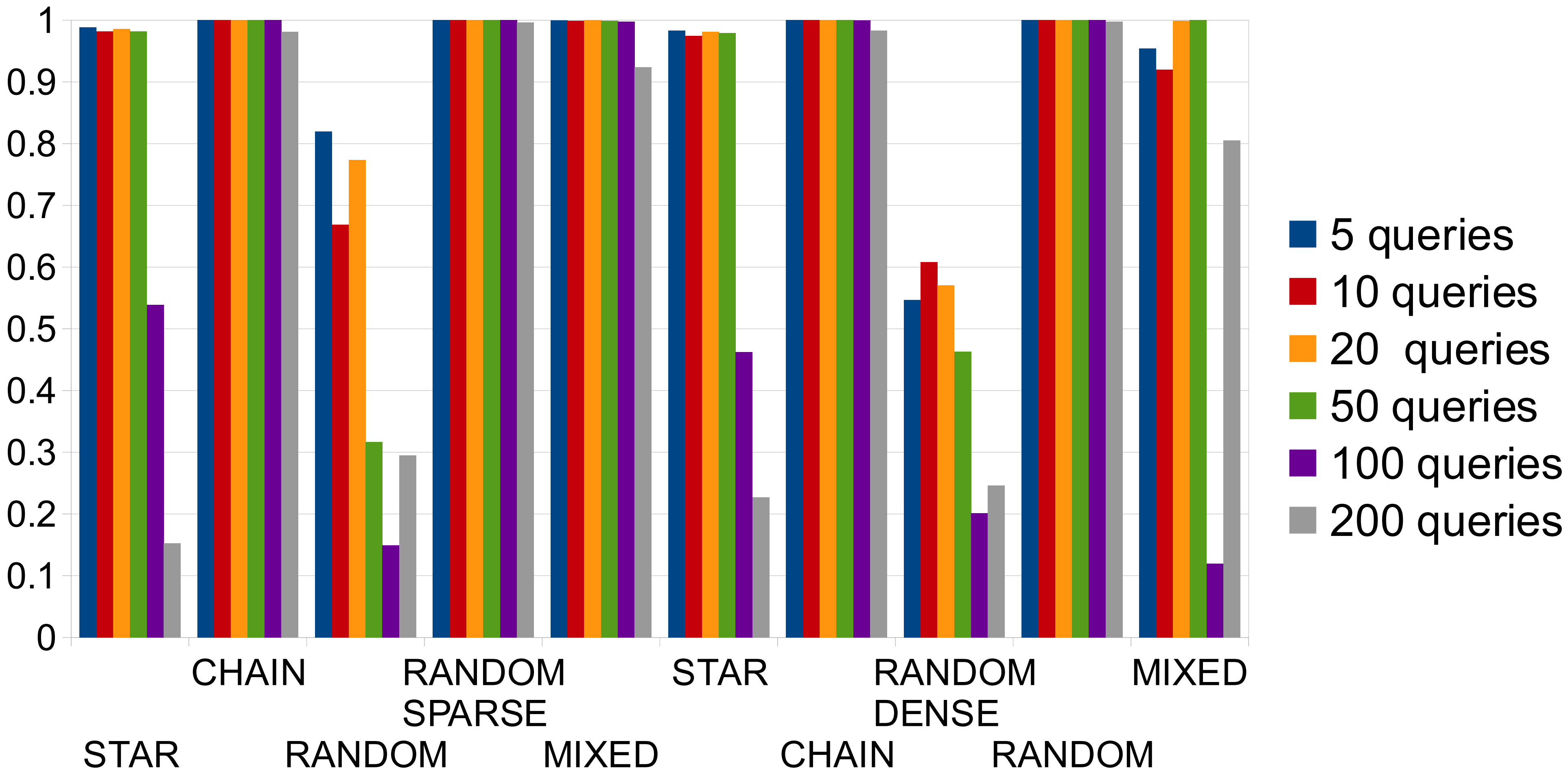
Output: Set of views $\{V_1, V_2, \ldots, V_k\}$ to materialize in order to minimize cost (workload processing + view storage and maintenance)

Difficulties: (a) implicit RDF data (b) scale (large queries, huge space)



View Selection in Semantic Web Databases (PVLDB 2011)

Cost reduction (%) achieved by our view search algorithm



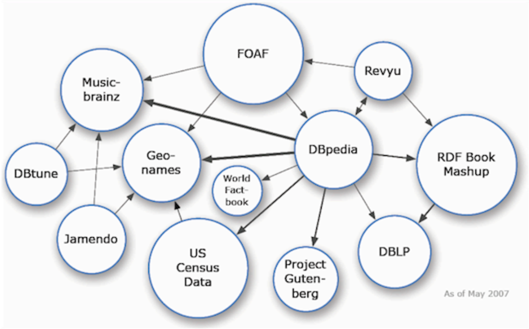
6

LEO outlook

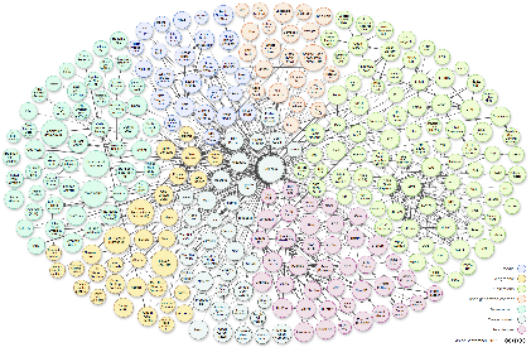
Working with heterogeneous data and knowledge

Heterogeneous systems evolve

The LOD diagram in 2005



The LOD diagram in 2011



Evolution impacts existing mappings.

Automate heterogeneous knowledge systems **evolution** & mapping **maintenance**

Data and schema (ontology) matching:

Use alignments in one to inform the other

Applied to RDF and LOD

EIT ICT Labs
"DataBridges"
I. Manolescu

With DataPublica and Zenith,
N. Pernelle

DATA PUBLICA

Web data management in the cloud

"MapReduce changed the world" (remember parallel databases? ☺)

- Really large-scale distributed hardware a commodity
- Parallel processing frameworks zoo: Hadoop, Hive, Pig, PACTs, Asterix, Hyracks…
- *Shared-nothing large-scale paralellism* + *concrete monetary costs*

Current work: scalable XML and RDF stores on Amazon

- Index in SimpleDB: models, performance?...
- Store in S3

Planned: indexing and high-level management on top of PACTs (TU Berlin); JSON

EIT ICT Labs "ConnectedCities", with Telecom SudParis

EIT ICT Labs "Europa" with Volker Markl (TU Berlin), SICS, TU Delft, KTH…

Regional grant DW4RDF, F.Goasdoué

7

Wrap-up

Main LEO publications

Main publications	2010	2011
Journal	2	6
International conferences and workshops	22	17
National conferences	13	4
Books		1
Book chapters	2	

Not including: WebDam members having left LEO; M. Herschel (to join in early 2012)

Main LEO software

- **Glucose** (G. Audemard, L. Simon): SAT solver 1.000 lines C
    - *Ranked 1st in the Applications category at SAT competition 2011*
- **TaxoMap** (C. Reynaud, B. Safar), 20.000 lines Java
    - Ontology (taxonomy) alignment, mapping refinement
- **LN2R** (N. Pernelle, F. Sais), 8.000 lines Java
    - Data reconciliation
    - *To be integrated with schema mapping (2011-2013 engineer)*
- **SomeWhere** (P. Chatalic, F. Goasdoué): distributed reasoning*, currently being re-structured by an engineer*
- ViP2P (I. Manolescu), 70.000 lines
    - Real-scale XML sharing in structured P2P networks
    - *Scalability 2010: 200 GB XML data, 200+ peers in Grid5000*
- ***RDFViewS** (F. Goasdoué, I. Manolescu), 26.000 lines Java*
- ***XUpOp** (N. Bidoit, D. Colazzo): XML update optimization through type projection*
- ***XUpIn** (N. Bidoit, D. Colazzo): XML query-update independence tester*

LEO in its INRIA environment

- Zenith: scientific data management, heterogeneity, scale
  - Continuous close collaboration continues (DataRing, DataPublica, cloud…)
  - LEO: Semantic Web, Zenith: scientific data
- Mostrare
  - Close interests in XML processing, collaborations within CODEX, with A. Bonifati
  - LEO: databases / Semantic Web / querying, integrating
  - Mostrare: tree automata, machine learning / XML / information extraction
- Exmo
  - Close interests in ontology alignment, reconciliation, open data (+LIRMM)
  - LEO: dynamic aspects, ontology evolution, databases
- Dahu: common background on formal models for databases
- ERC WebDam: M. Bienvenu and I. Manolescu collaborate with S. Abiteboul
- Aviz : joint interest in scalable frameworks, Open Data

LEO environment: France and abroad

Main collaborations in France

- University of Paris XI: Fortesse and BioInfo teams
- INRA Paris and Toulouse
- University of Grenoble
- CNAM

Collaborations abroad:

- UC San Diego: view management for the Web
- Athens University of Economics and Business
- TU Berlin, TU Delft: cloud-based data management
- CRP Henri Tudor (Luxembourg): reconciliation of dynamic medical knowledge organization systems
- U. Pisa, U. Genova: XML types, XQuery updates
- U. Bremen: knowledge representation

More competition

Max-Planck Institut fur Informatik (Saarbrucken): Gerhard Weikum

We (send them | compete for) post-docs

TU Munchen, U. Mannheim: Guido Moerkotte, Thomas Neumann

Great RDF query optimization work

Hasso-Plattner Institut

Massive RDF management / Open Data

UC San Diego (A. Deutsch, Y. Papakonstantinou), UC Irvine (M. Carey)

Views for the Web, cloud-based XML data management

U. Roma 1 et U. Bolzano (R. Rosati, D. Calvanese), U. Liverpool,

U. Manchester, U. Oxford, U. Dresden, U. Karlsruhe…

Knowledge representation, description logics, semantic Web

LEO visibility

Conference chairs:

- SAT 2011 (Theory and Application of Satisfiability Testing)
- EDBT 2010 (Extending Database Technology)
- IEEE ICDE (Data Engineering) XML and semantic Web (2012), experiments & applications (2011), demos (2010)

Editorial:

I. Manolescu editor-in-chief of ACM SIGMOD Record, editor of ACM TOIT, editorial board of PVLDB

N. Bidoit co-organizer of national database summer school (2010, 2012)

Conference PC members: ACM SIGMOD, VLDB, ICDE, IJCAI, DX, …

Project coordinators: F. Goasdoué, I. Manolescu, N. Pernelle, C. Reynaud, L. Simon merci / questions?

Inria

INVENTEURS DU MONDE NUMÉRIQUE