Fédération et analyse de données distribuées en imagerie biomédicale

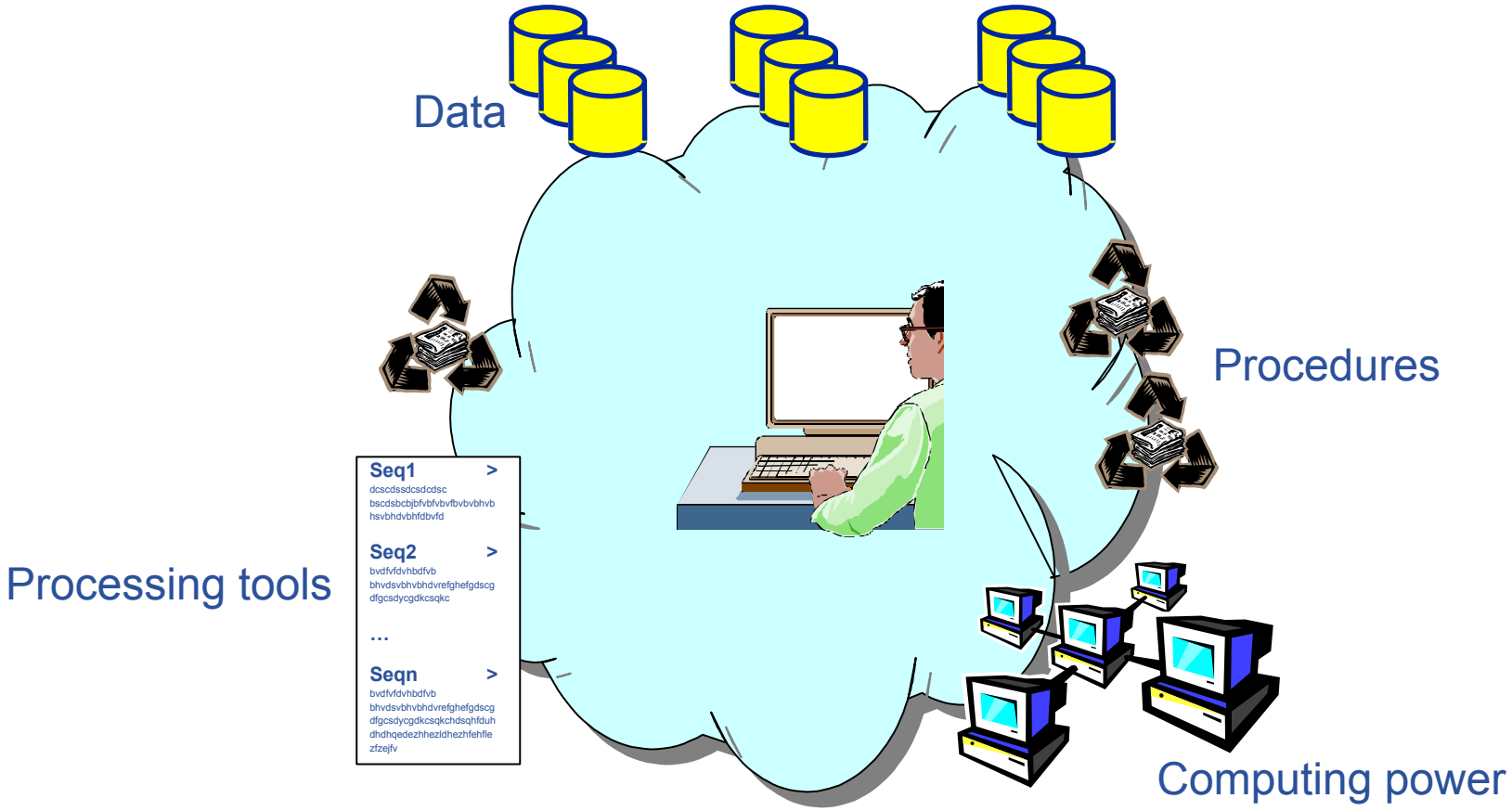
Johan Montagnat
CNRS, I3S lab, Modalis team
on behalf of the NeuroLOG and
the CrEDIBLE consortiums

Nouvelles méthodologies en imagerie du vivant

Lyon, 13 décembre 2012

Multi-centric studies in neurosciences

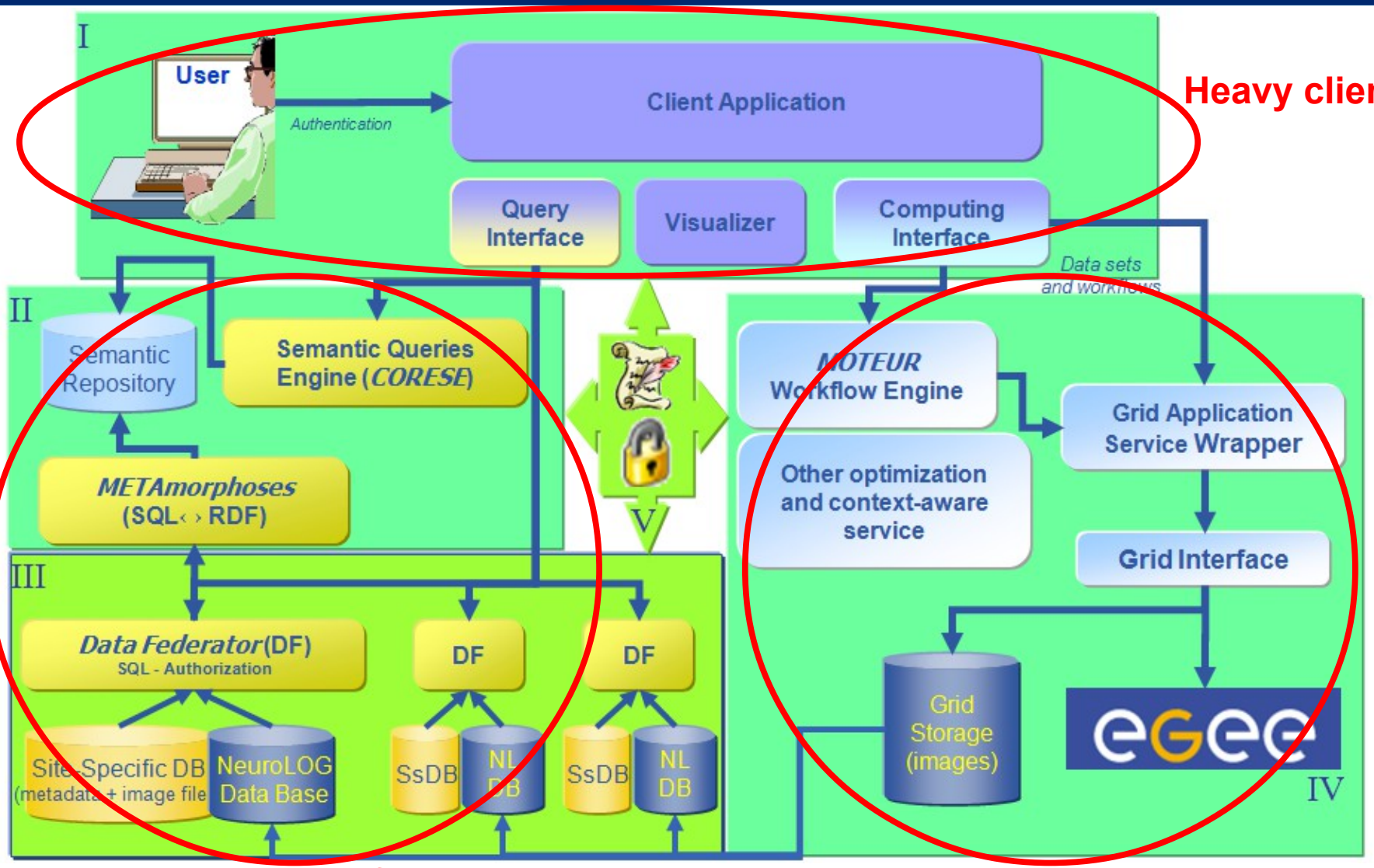
- **Sharing computing resources and algorithms**
  - Research (large data sets, statistical studies, models design...)
  - Complex analysis (compute-intensive studies, validation procedures...)



Motivations

- **Neuroscience data**
  - Increasing use of imaging biomarkers for research and diagnosis
  - Increasing number of (multi-centric) large-scale studies
  - Distribution of resources over acquisition sites
    - Need to consider existing site-wide legacy environments
- **Computing resources**
  - Heavy computing power required
  - Data analysis pipelines
    - Integrating neurodata analysis codes from different toolkits
- **Centralized approaches encounter limitations**
  - Large data volumes to transfer / archive / search
  - Sensitive patient data / complex access control policies
  - Need to adopt uniform data model & format
- **Approach: federate existing resources in a distributed, collaborative platform**

NeuroLOG middleware



NeuroLOG platform

ANR TLOG (2006-2010)

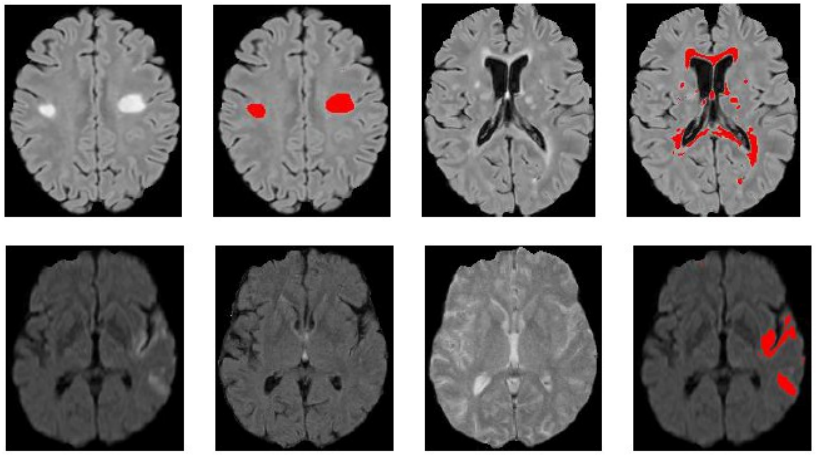
Multi-centric environment for neurosciences

5 neuroscience centers federated

- I3S (Sophia Antipolis) core technical site
- IRISA (INRIA Rennes), collaborating with the University Hospital of Rennes
- IFR49 (INSERM affiliated neuroscience group in Paris La Pitié Salpétrière Hospital)
- GIN (INSERM affiliated neurosciences institute of Grenoble, Michalon Hospital)
- INRIA Sophia Antipolis collaborating with Centre Antoine Lacassagne (Nice)

Neuro-imaging data

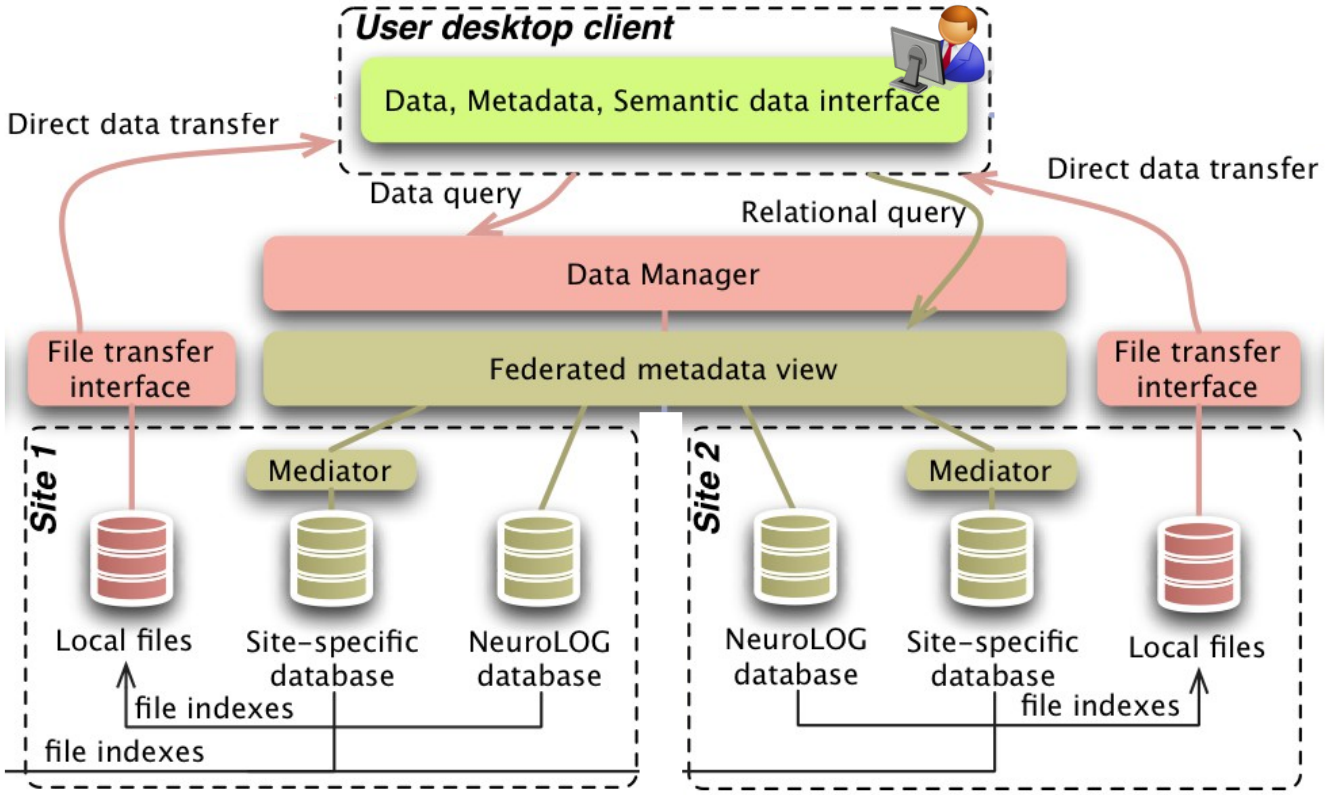
- **Pathologies**
  - Multiple Sclerosis
  - Brain strokes
  - Brain tumors
  - Alzheimer's



- **Data considered**
  - Imaging data
    - Various MR modalities (T1, T1 Gado, T2, Flair, Diffusion, PD)
    - Processed images (Registered, Segmented, …)
  - Associated metadata
    - Studies
    - Subjects
    - Data acquisition context and provenance
    - Neurophysiological and Neuroclinical tests
    - Measurements derived from image data

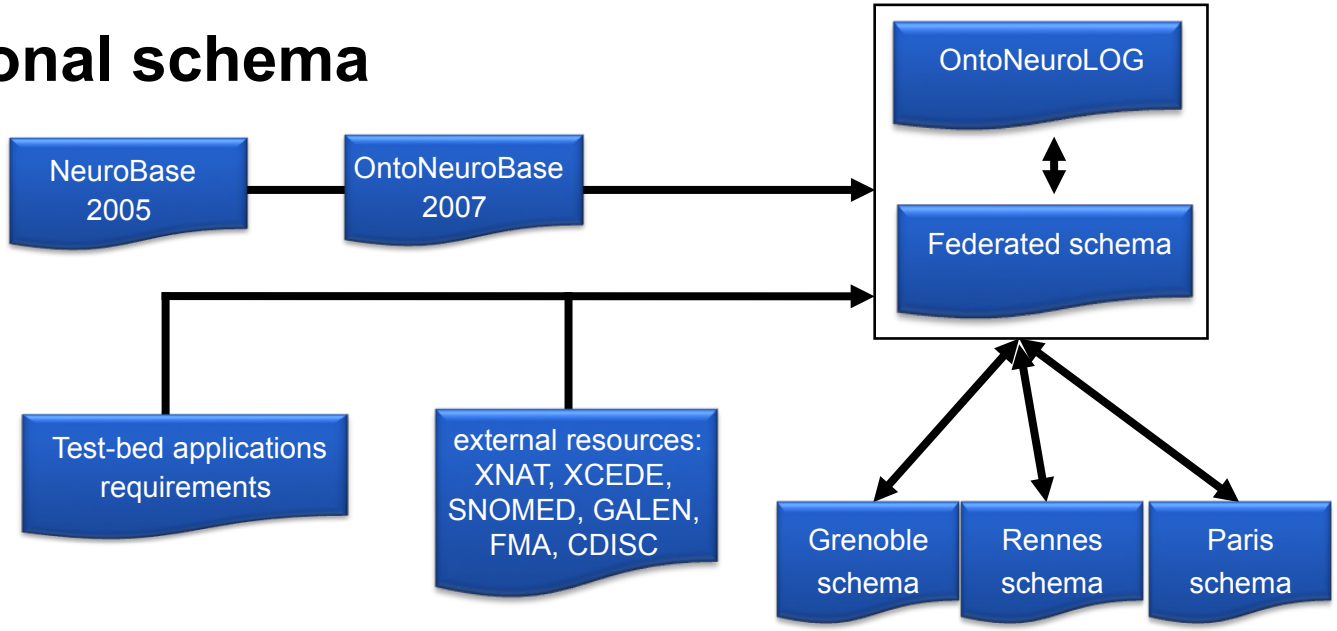
NeuroLOG data mediation & federation

- Preserve legacy environment (e.g. relational databases)
- Cope with heterogenous schemas
  - Use a relational database mediation & federation engine (BusinessObject/SAP DataFederator product)

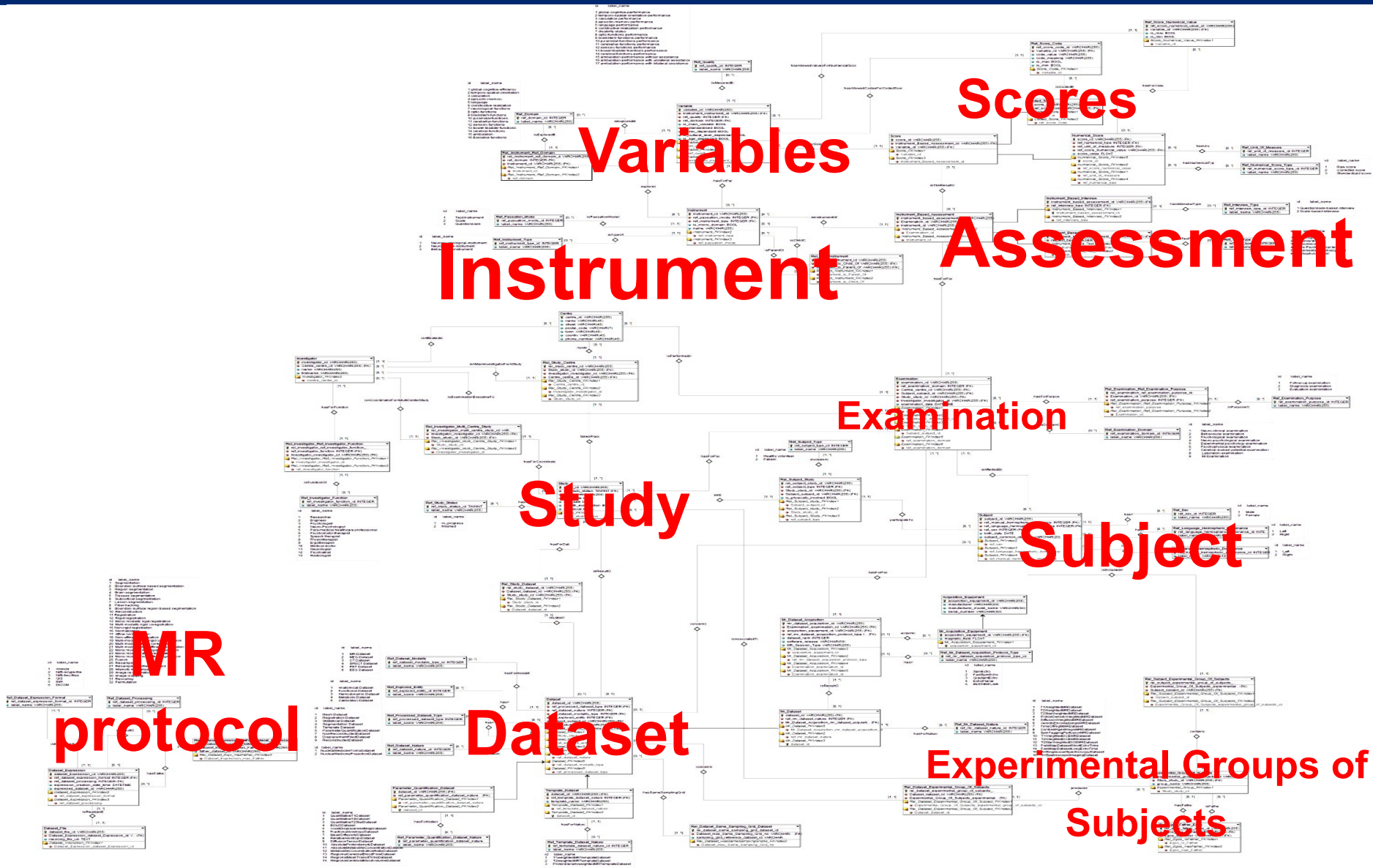


Semantic reference

- **Application ontology OntoNeuroLOG**
  - **Based on a common modeling framework**
  - **3-levels structure**
    - **one Foundational ontology: i.e. DOLCE**
    - **Several Core ontologies**
    - **Several Domain ontologies**
- **Implemented in OWL-Lite**
- **Derived relational schema**

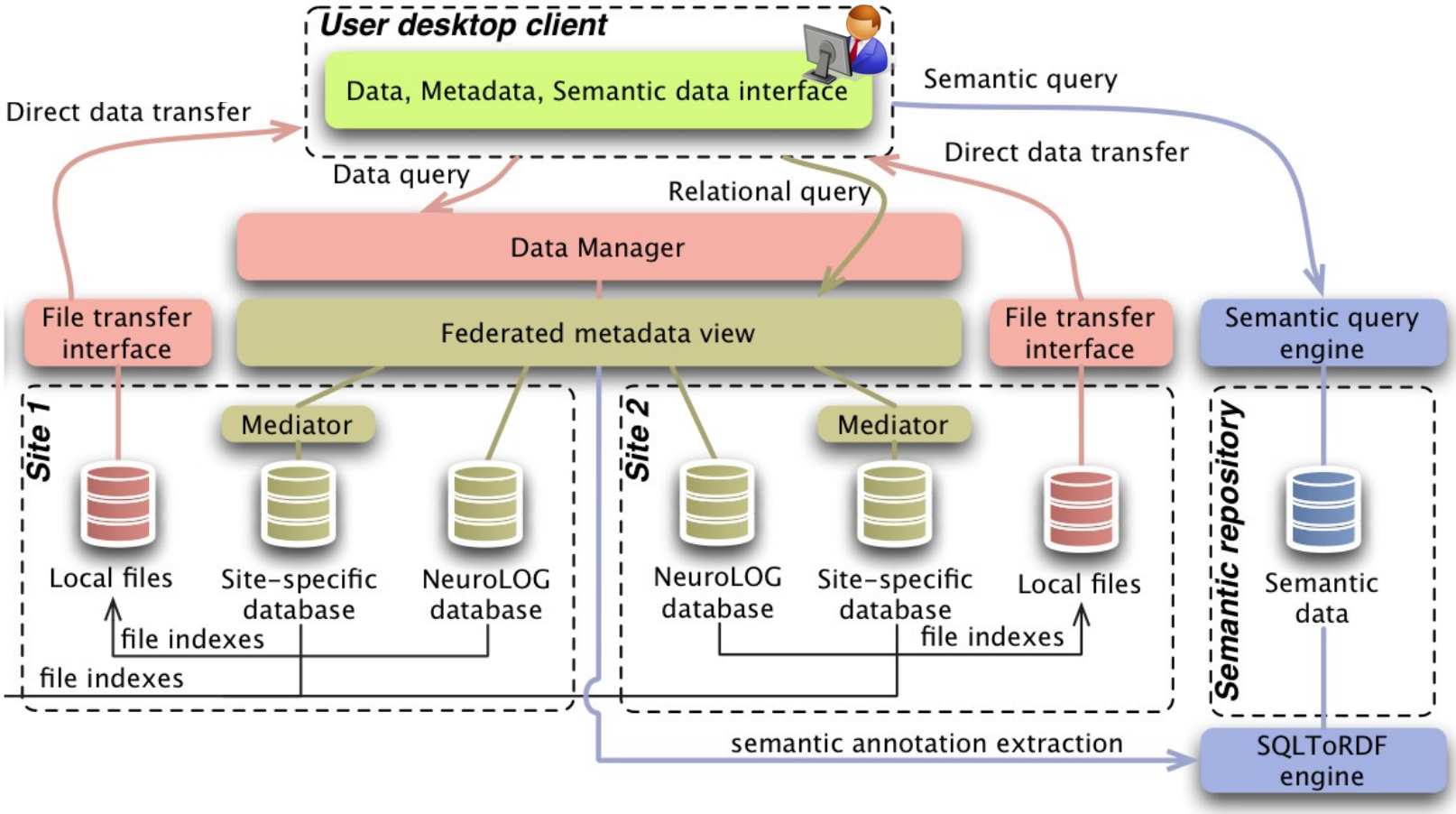


OntoNeuroLOG



NeuroLOG data mediation & federation

- Experimenting both relational and semantics technologies
  - METAMorphoses conversion of (federated) relational databases into a semantic annotations store

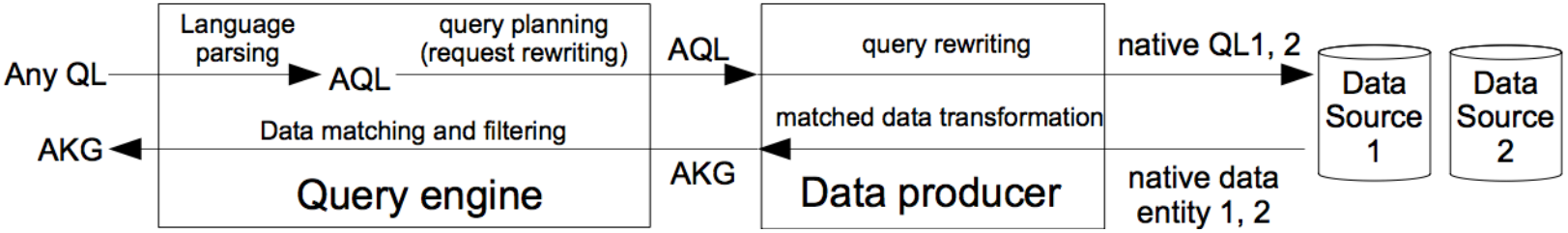


From relational to semantic datastores

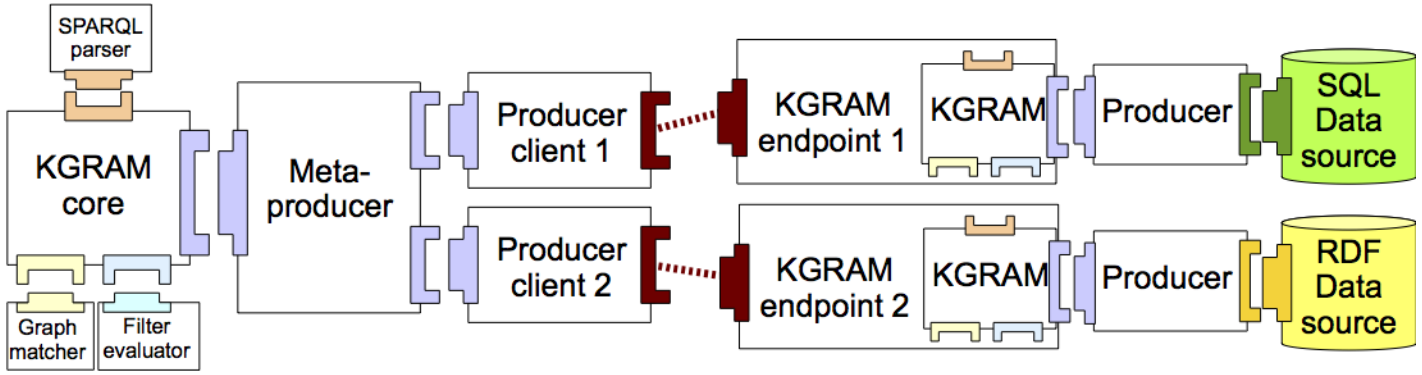
- **CrEDIBLE multi-disciplinary workshop in Sophia Antipolis (Oct. 15-17, 2012)**
  - Semantic models are widely accepted
  - Existing systems in the biomedical community are mostly centralized
  - The need for multi-centric studies support is unambiguous
  - Exploiting / reusing data in a multi-disciplinary context is still preliminary, and ontological resources are not sufficient
- **Approach**
  - Semantic reference design
  - RDF triples-based knowledge bases
    - Semantic alignment for heterogeneous data sources
    - Data sources mapping
  - Distributed semantic query engine
    - SPARQL v1.1 compliant

Distributed query engine

- **Based on KGRAM (Knowledge Graph Abstract Machine)**
  - Full support of SPARQL v1.1
  - Flexible software architecture adaptable to many use cases

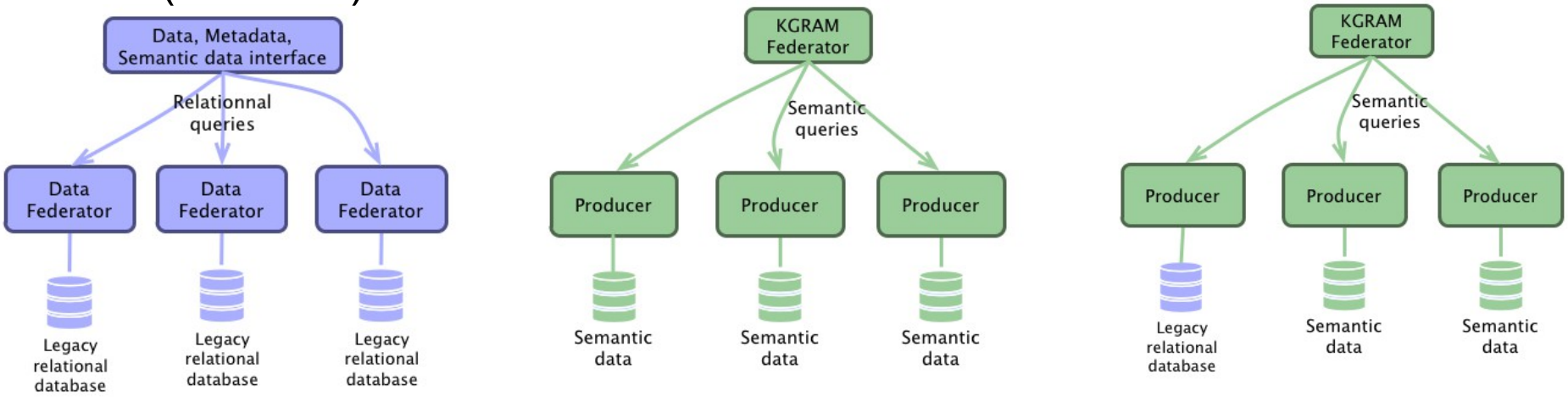


- **Deployment example**
  - Meta-producer distributes queries over multiple query endpoints
  - KGRAM endpoint interfaces with heterogeneous data stores



Results: distributed query processing

- **Multiple neuroscience data stores querying**
  - Relational stores (DF), semantic bases (KGRAM) or both (KGRAM)



- **Performance analysis**
  - Q1 : costly evaluation (336 remote invocations)
  - Q2 : selective query (only 5 resulting datasets)

Query	Relational (SAP DF)	Semantic	Semantic+Relational
Q1	**3.03 s ± 0.25**	6.13 s ± 0.05	11.76 s ± 0.05
Q2	1.52 s ± 0.62	**0.60 s ± 0.03**	1.53 s ± 0.14

Workflow enactor

- **Workflow language + enactor**
  - Abstract description of data analysis pipelines
  - Interface to computing infrastructure
    - Data processing code wrapper component (jigsaw)
    - Interface to grid(s) workload management API
- **MOTEUR workflow manager**
  - http://modalis.polytech.unice.fr/moteur2
  - GWENDIA language
    - Data-driven, implicit-parallelism oriented language
    - Graphic interface to design workflows
  - MOTEUR enactor
    - Transparent exploitation of parallelism
    - Interfaced to multiple submission interface
      - *Local resources, EGI, Grid5000...*

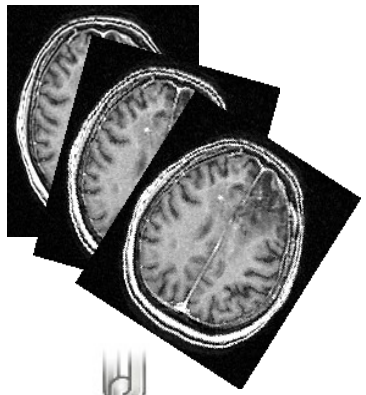
Codes packaging and enactment

Data analysis
code bundling

Bundles
deployment

Enactment

data sets



descriptor

.aar

code
+
dependencies

site server

deploy

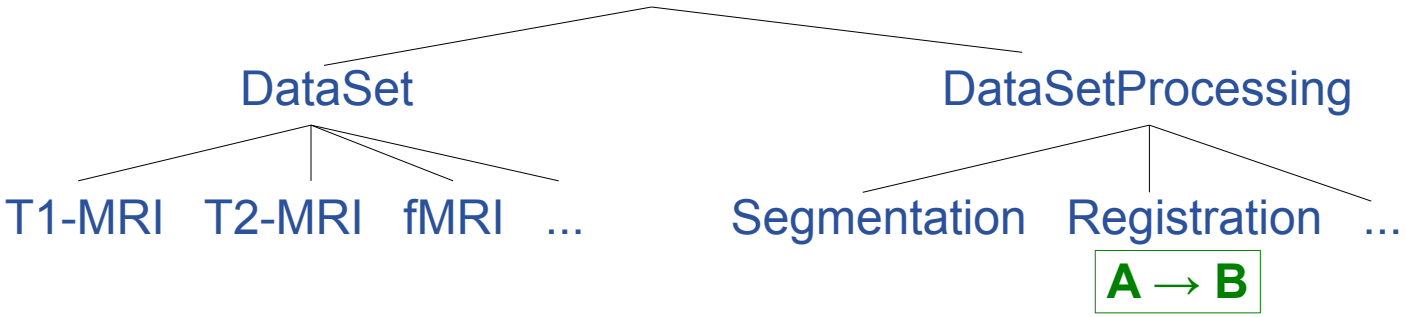
invoke

MOTEUR 2

workflow engine

Semantic extensions

- **Ontology**
  - Concepts & **Rules**

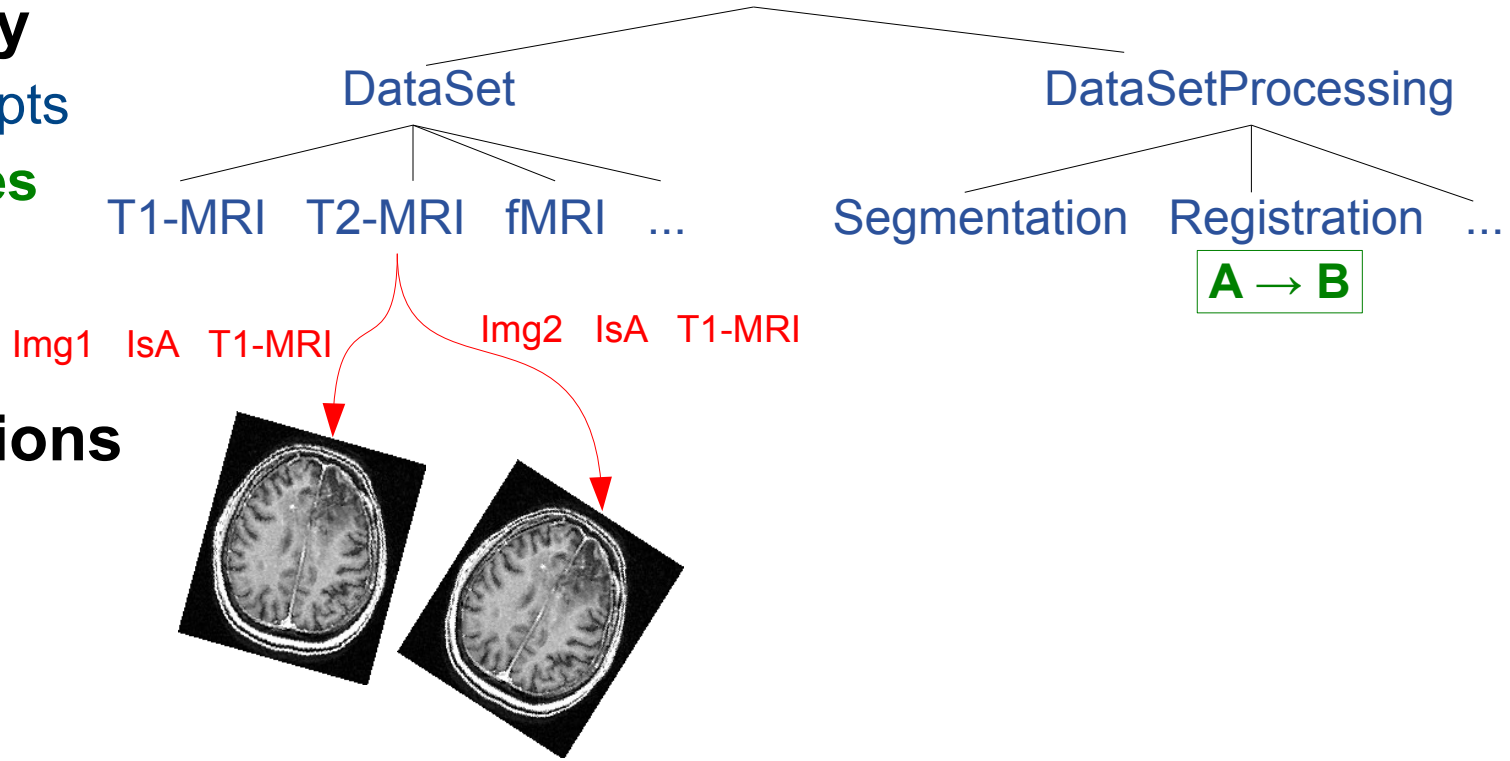


- **Annotations**

- **Processing**

Semantic extensions

- **Ontology**
  - Concepts & **Rules**

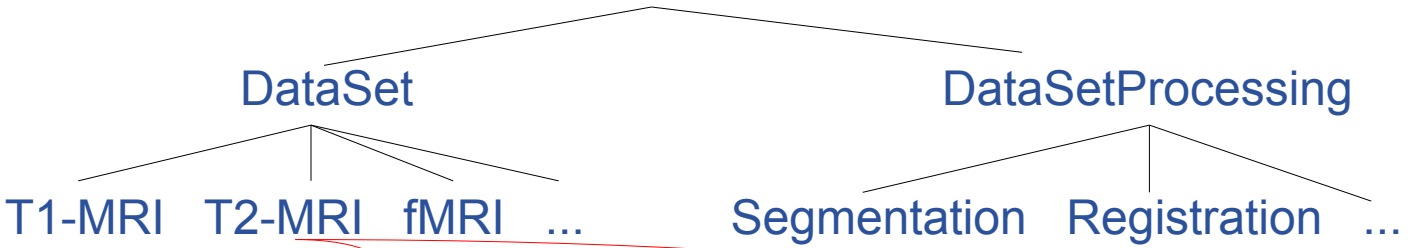


- **Annotations**

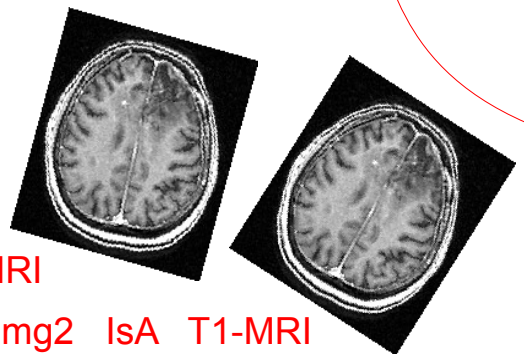
- **Processing**

Semantic extensions

- **Ontology**
  - Concepts & **Rules**



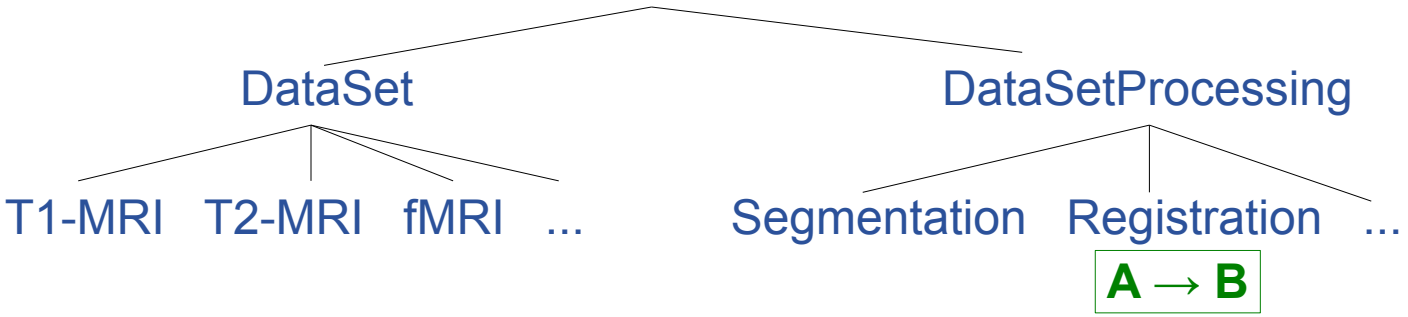
- **Annotations**



- **Processing**

Semantic extensions

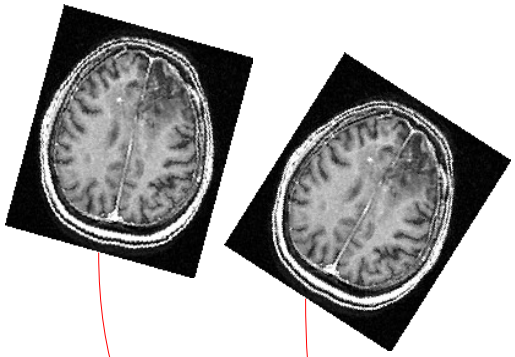
- **Ontology**
  - Concepts & **Rules**



- **Annotations**

Tool1 HasInput T1-MRI
Tool1 HasOutput Transfo
Tool1 IsA Registration

Img1 IsA T1-MRI
Img2 IsA T1-MRI



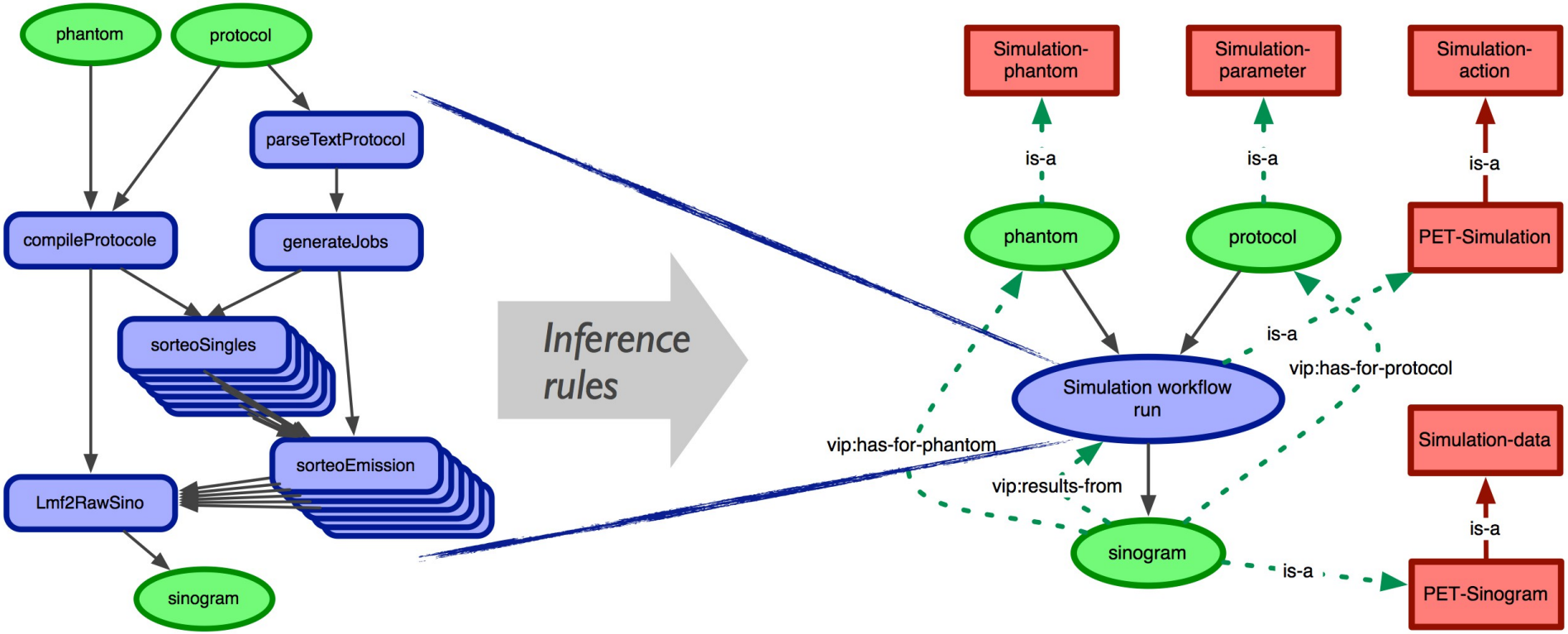
- **Processing**

Img1 IsProcessedBy Tool1
Img2 IsProcessedBy Tool1

Tool1 Produced Transfo1
Transfo1 IsA GlobalTransfo

Semantic knowledge use in workflows

- Fine-grained annotation traces generated at run-time
- Summary generated by inference rules application
  - Produce relevant and human-tractable experiment summaries



- Example in the context of the VIP (Virtual Imaging Platform) project

(Some) lessons learned

Data federation feasibility

- Relational data federation requires semantic reference
- Dual relational / semantic data view is confusing for end users
  - Mapping to a semantic, well-documented data model
- Need to cope with site failures
- Data access control is a tough problem

- **Semantic technologies**
  - Powerful semantic query and inference engine
    - Trade-off between query language expressivity and performance
  - Coupling data and processing semantics
  - Semantic querying and inference capability are foreign to users
  - Non-trivial user interface to be defined to query the federation
- **Application to other domains**
  - VIP Virtual Imaging Platform
  - http://www.creatis.insa-lyon.fr/vip

References

- **Reports & publications available on-line**
  - http://credible.i3s.unice.fr & http://neurolog.i3s.unice.fr
- **Publications**
  - O. Corby, A. Gaignard, C. Faron-Zucker, J. Montagnat.
    KGRAM Versatile Inference and Query Engine for the Web of Linked Data
    IEEE/WIC/ACM International Conference on Web Intelligence, Macao, China, Dec. 2012.
  - A. Gaignard, J. Montagnat, C. Faron-Zucker, O. Corby.
    Semantic Federation of Distributed Neurodata
    MICCAI Workshop on Data- and Compute-Intensive Clinical and Translational Imaging Applications, pages 41-50, Nice, France, October 2012.
  - B. Gibaud, G. Kassel, M. Dojat, B. Batrancourt, F. Michel, A. Gaignard, J. Montagnat
    NeuroLOG: sharing neuroimaging data using an ontology-based federated approach
    AMIA, vol. 2011, pages 472–480, Washington DC, USA, October 2011.
  - F. Michel, A. Gaignard, F. Ahmad, C. Barillot, B. Batrancourt, M. Dojat, B. Gibaud, *et al.*
    Grid-wide neuroimaging data federation in the context of the NeuroLOG project
    HealthGrid'10, pages 112-123, IOS Press, Paris, France, 28-30 June 2010.
- **Research reports**
  - CrEDIBLE-12-1-v1: multi-disciplinary workshop report
  - CrEDIBLE-12-2-v1: distributed semantic query engines
  - CrEDIBLE-12-3-v1: sémantique des données de l'observation