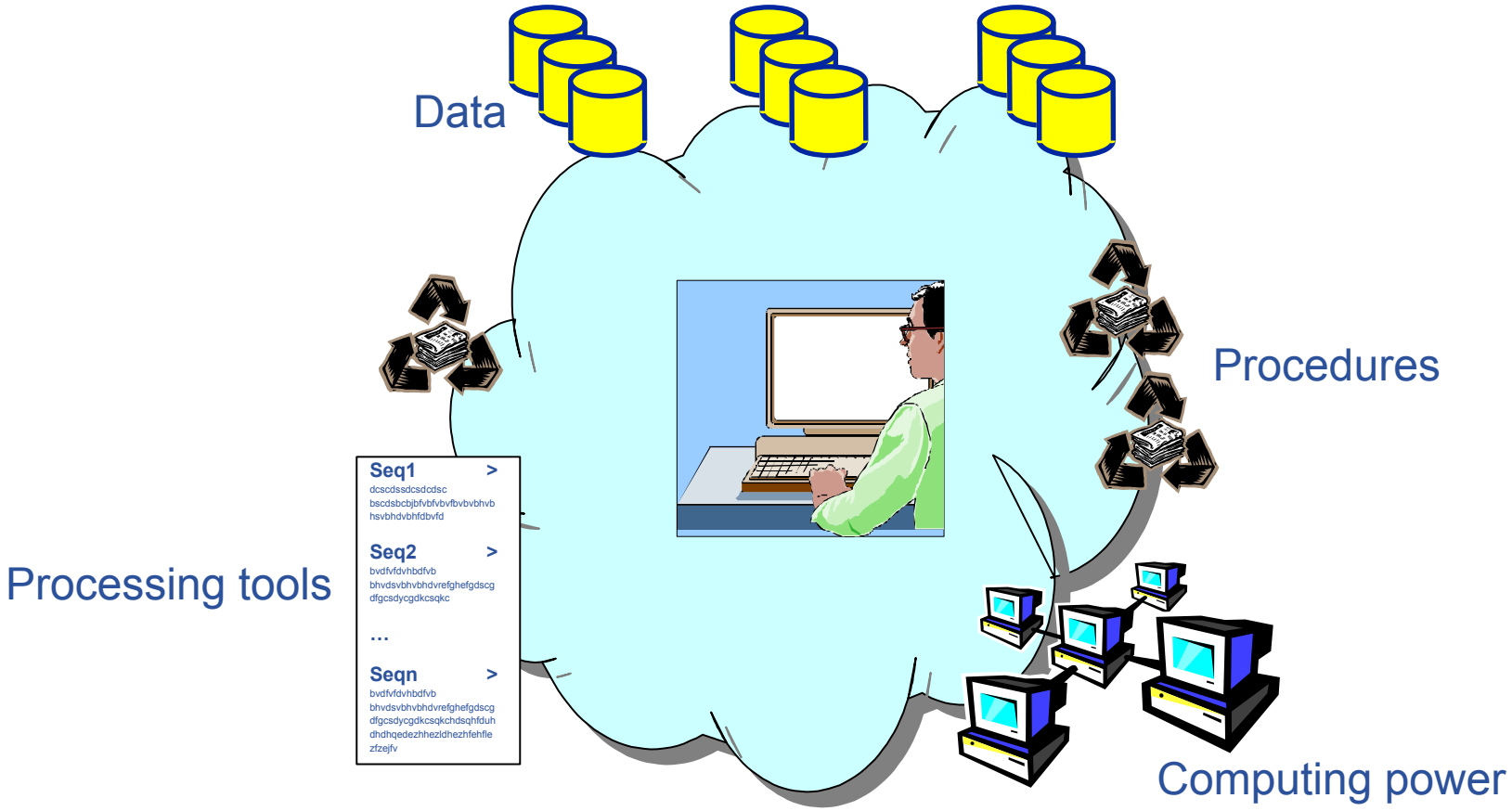
The NeuroLOG ontology-based approach to federate distributed neurodata stores

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

NeuroInformatics.NL – Data sharing workshop
Amsterdam, December 14, 2012

Multi-centric studies in neurosciences

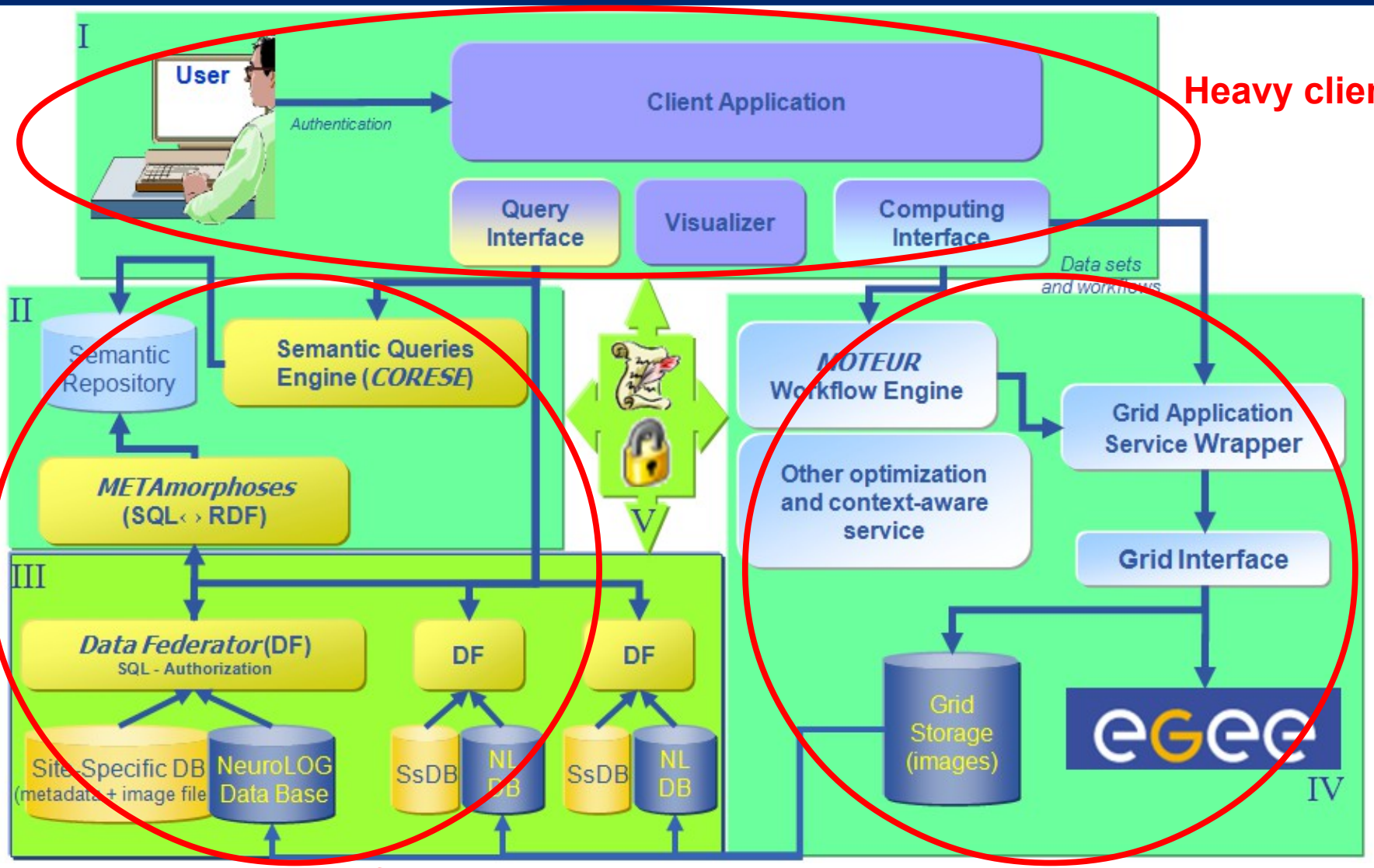
- **Sharing computing resources and algorithms**
  - Research (large data sets, statistical studies, models design...)
  - Complex analysis (compute-intensive studies, validation procedures...)



Motivations

- **Neuroscience data**
  - Increasing use of imaging biomarkers for research and diagnosis
  - Increasing number of (multi-centric) large-scale studies
    - Publicly available databases
  - Distribution of resources over acquisition sites
    - Need to consider existing site-wide legacy environments
- **Centralized approaches encounter limitations**
  - Large data volumes to transfer, archive & search
  - Data acquisition sites are distributed
    - Need to periodically synchronized with new data acquired
  - Sensitive patient data
    - Need to transfer data access control
  - Need to adopt uniform data model & format
- **Approach: federate existing resources in a distributed, collaborative platform**

NeuroLOG middleware



NeuroLOG platform

ANR TLOG (2006-2010)

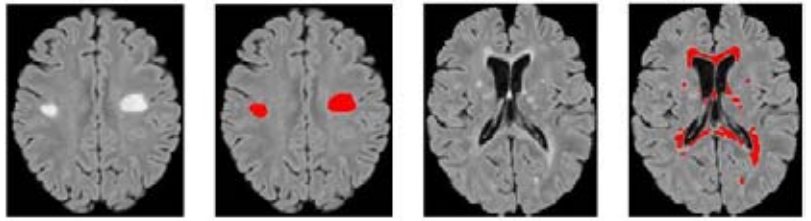
Multi-centric environment for neurosciences

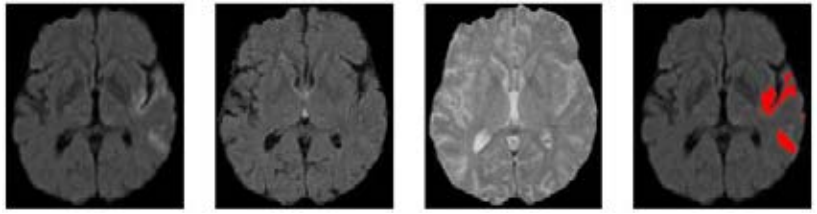
5 neuroscience centers federated

- 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

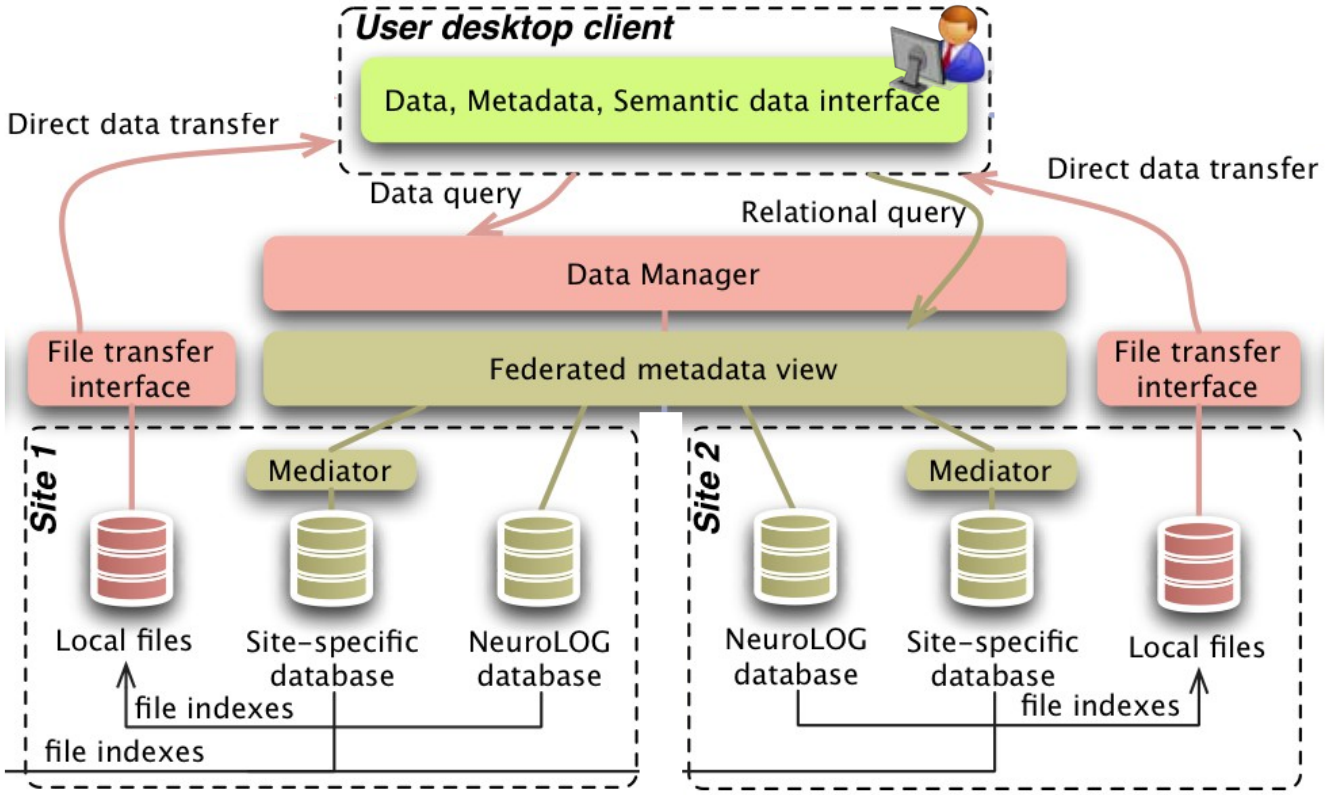




- **Data considered**
  - Imaging data
    - Various MR modalities (T1, T1 Gado, T2, Flair, Diffusion, PD)
    - Processed images (Registered, Segmented, …)
  - Associated metadata
    - Studies
    - Subjects
    - Data acquisition context and provenance
    - Neurophysiological and Neuroclinical tests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

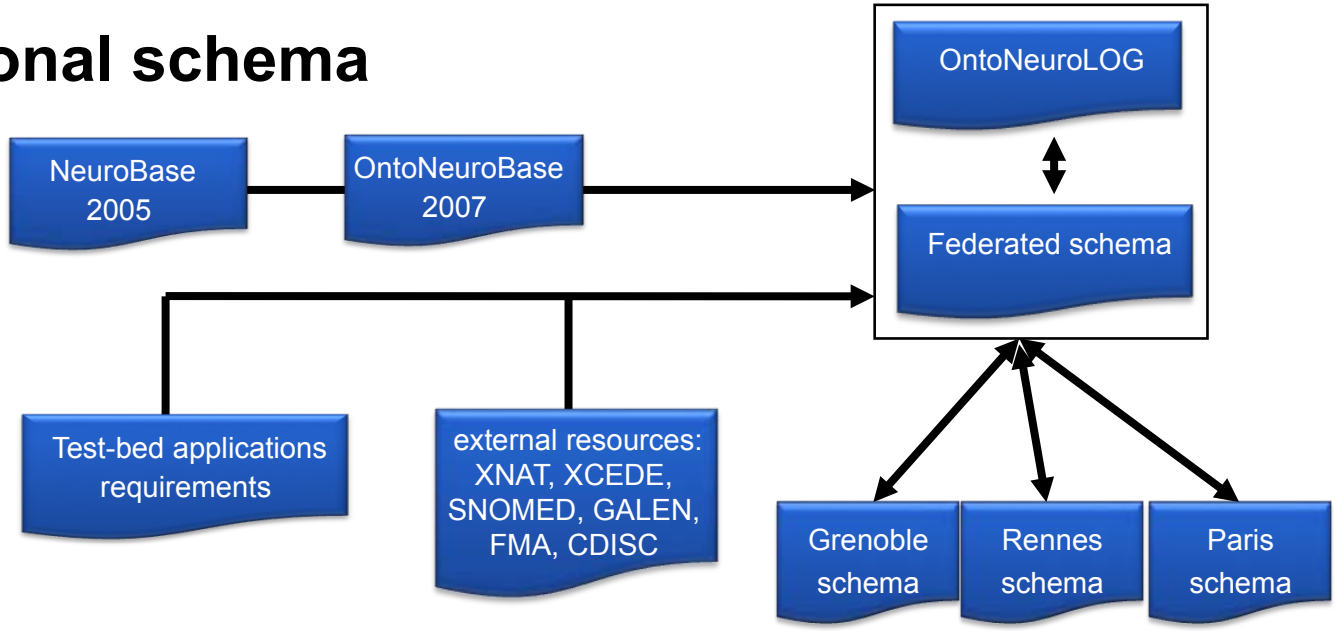
NeuroLOG data mediation & federation

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

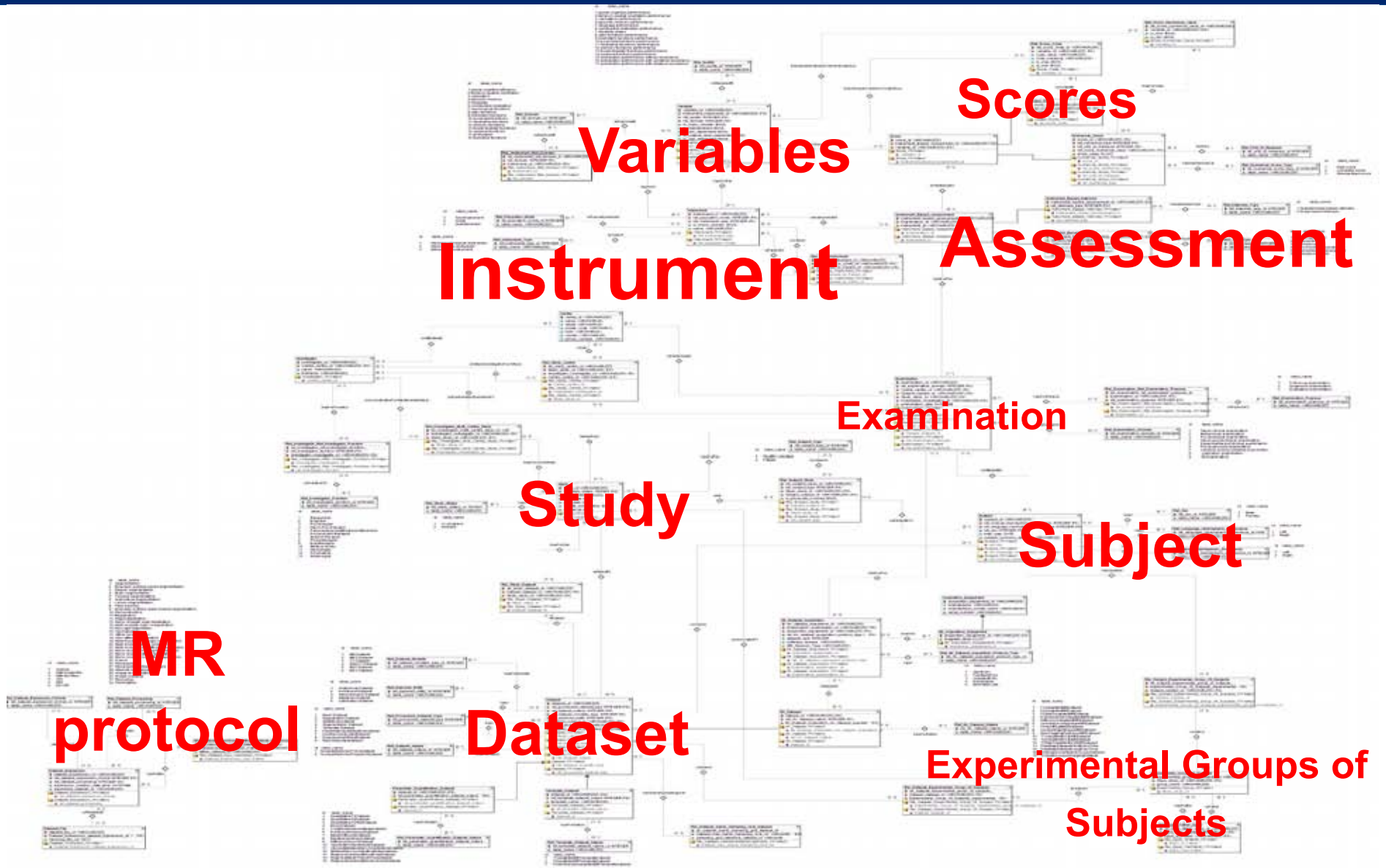


Semantic reference

- **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**

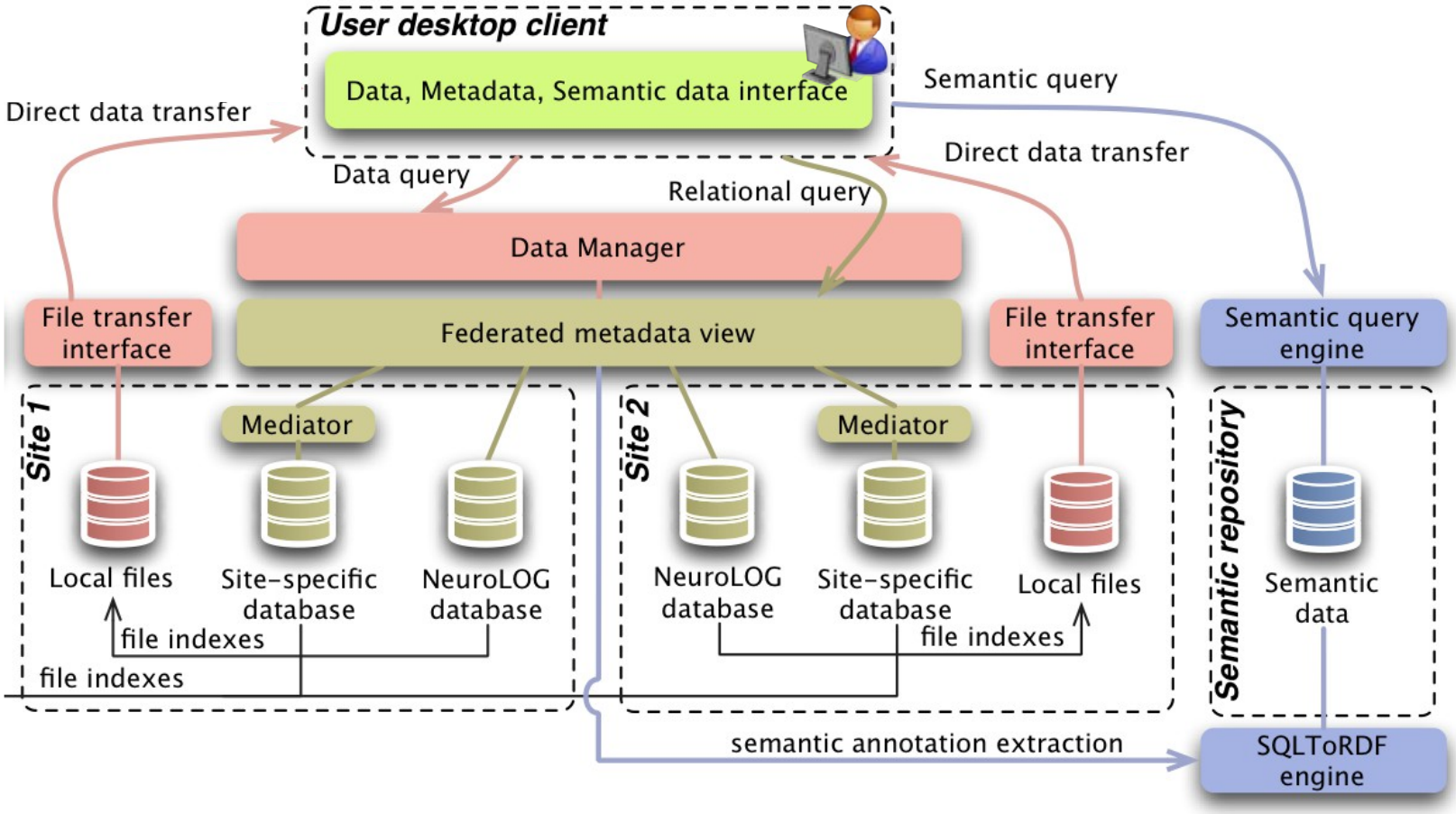


OntoNeuroLOG



NeuroLOG data mediation & federation

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

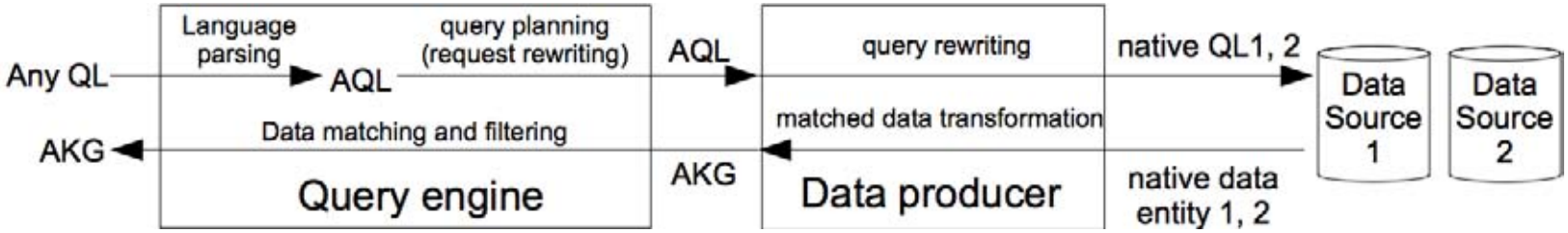


From relational to semantic datastores

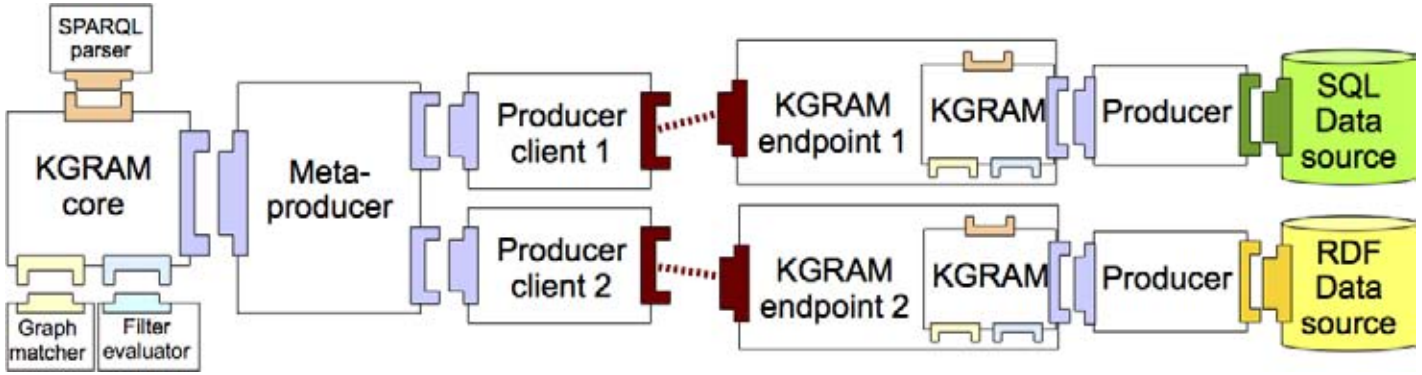
- **CrEDIBLE multi-disciplinary workshop in Sophia Antipolis (Oct. 15-17, 2012)**
  - Data modeling, data stores, mediation, (distributed) querying, users...
  - Semantic models are widely accepted. Existing systems in the biomedical community are mostly centralized. The need for multi-centric studies support is unambiguous though.
  - 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 semantic query engine

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

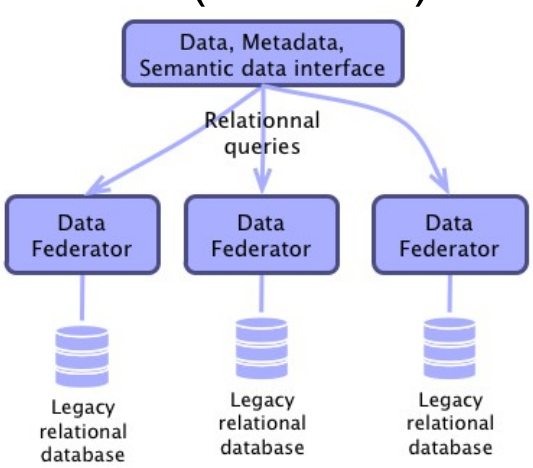


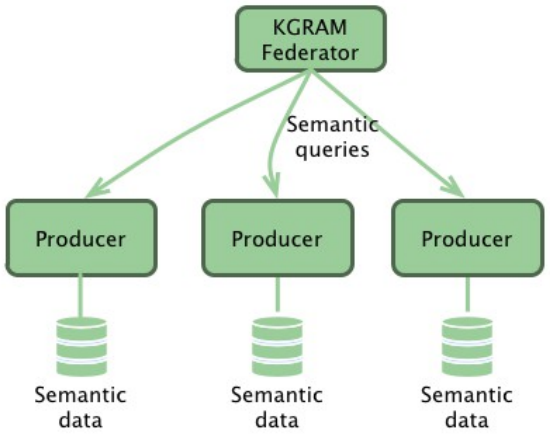
- **Deployment example**
  - Meta-producer distributes queries over multiple query endpoints
  - KGRAM endpoint interfaces with heterogeneous data stores

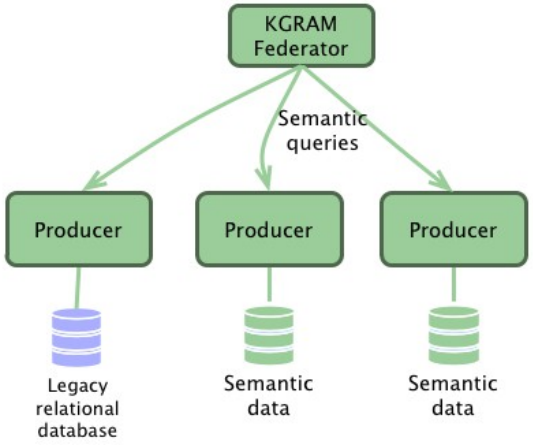


Results: distributed query processing

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





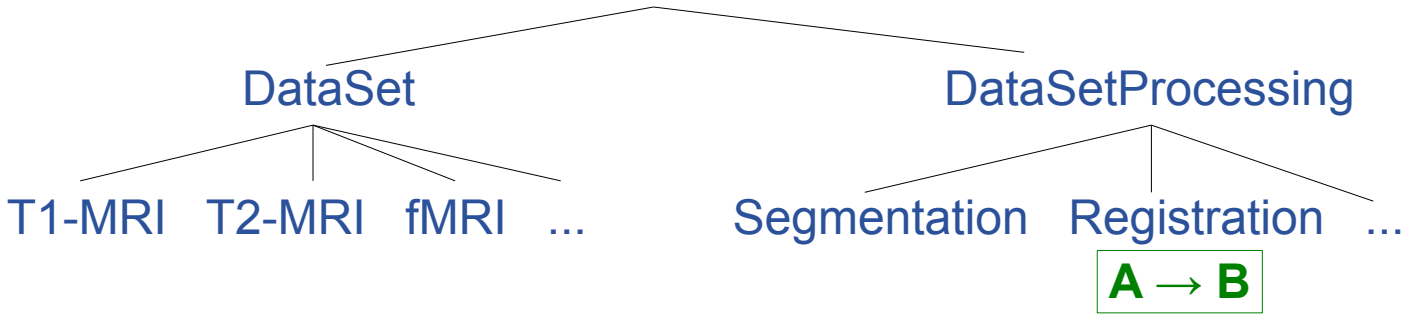


- **Performance analysis**
  - Q1 : costly evaluation (336 remote invocations)
  - Q2 : selective query (only 5 resulting datasets)

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

Semantic extensions

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

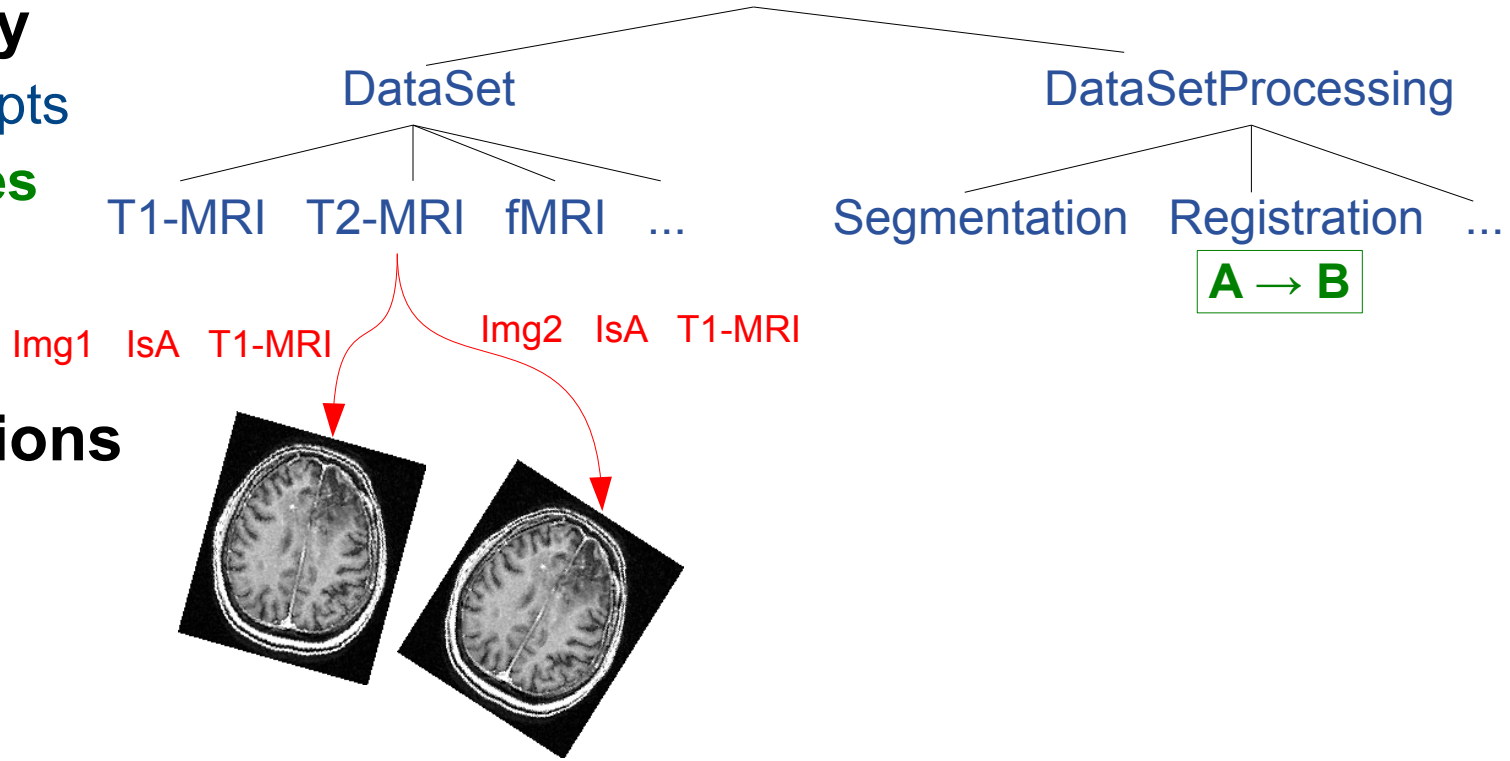


- **Annotations**

- **Processing**

Semantic extensions

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

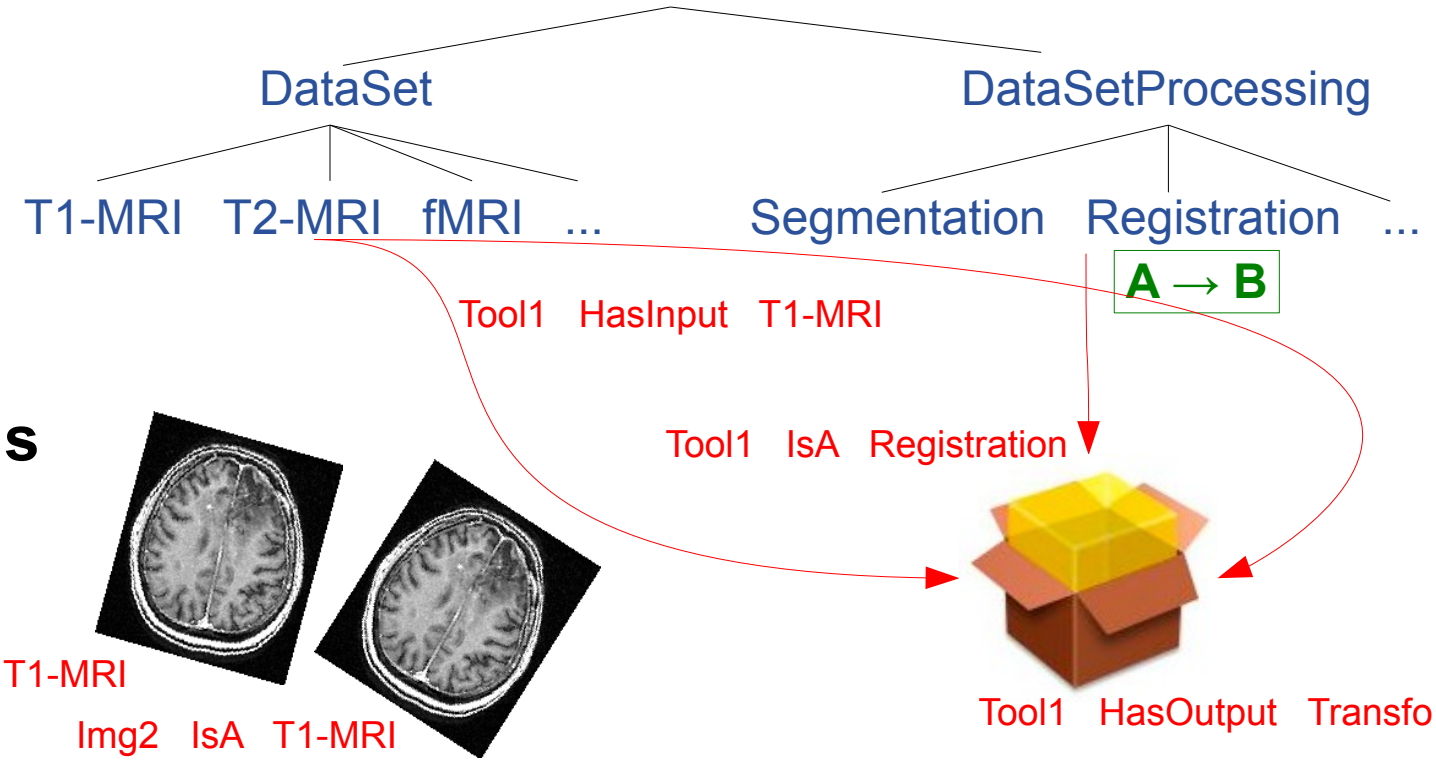


- **Annotations**

- **Processing**

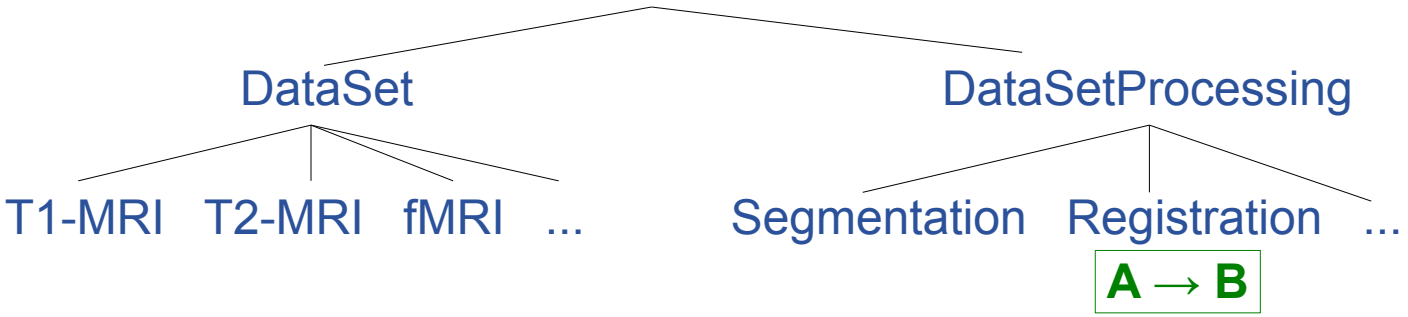
Semantic extensions

- **Ontology**
  - Concepts & **Rules**
- **Annotations**
- **Processing**



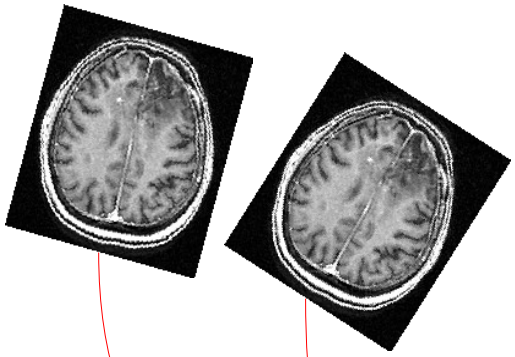
Semantic extensions

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



- **Annotations**

Img1 IsA T1-MRI
Img2 IsA T1-MRI



Tool1 HasInput T1-MRI
Tool1 HasOutput Transfo
Tool1 IsA Registration

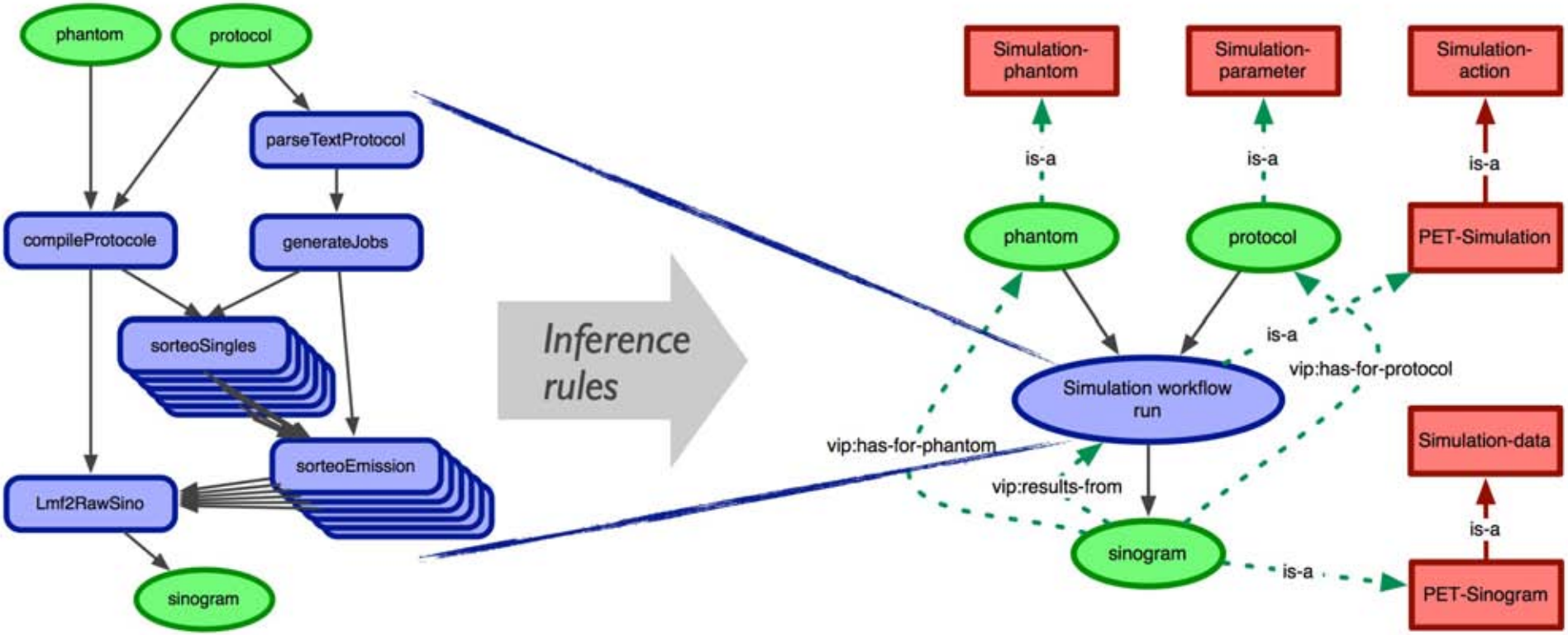
- **Processing**

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



(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
    - Leverage semantic information and infer new knowledge
  - Semantic querying and inference capability are foreign to users
    - Non-trivial user interface to be defined to query the federation

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

Bonus 1: ready for data sharing?

- **Nobody can analyze my data as well as I can / Others will wrongly interpret it / Others can't possibly understand what happened at the recording session**
  - Mostly true. This is why we are trying to produce a complete and formally documented data schema
- **It adds complexity in my life**
  - Definitely true in a short term. You should think of the future now.
- **My archive is too big**
  - Maybe it is. It is one of the reason why we will not copy it.
- **Others will be cited for my work / publish faster than me**
  - Didn't you sign it fore research? Didn't know it was competitive?
  - Scientific treachery is not knew and should always be fought.
- **My subjects didn't sign for it / My industrial partner doesn't agree**
  - Of course only authorized data can be shared

Bonus 1: ready for data sharing?

- **Others will try to reproduce my results**
  - Hopefully they can! This is likely to become mandatory BTW.
- **Nobody will visit my lab to collaborate anymore**
  - My bet is the opposite. When others can see / use their data, they will show an interest.

Bonus 2: key issues

- **Will inclusion of other institutes' data boost your research?**
  - Most probably, as witnessed from many intl data sharing initiatives
- **What about loosing your competitive edge?**
  - Research is a competition-collaboration arena
- **What about patient privacy?**
  - This is a though problem that has to be seriously addressed
- **What about publications and citations?**
  - The sharing policy does not imply wiping all rights out. Open source software developers have been practicing for a long time.
- **What about people deriving wrong claims from your data?**
  - This will happen, just like some are deriving wrong claims from their own data. In both cases, science is all about stating right facts and arguing against false ones.

Bonus 2: key issues

- **Should public money only be spent on Open Access projects?**
  - Funding administrations certainly aspire to a rational use of public money
- **How to share massive amounts of data?**
  - Distribution is the key