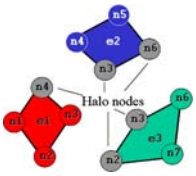
Data Services @neurIST and beyond

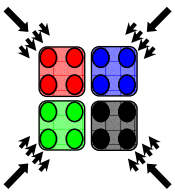
Siegfried Benkner
Department of Scientific Computing
Faculty of Computer Science
University of Vienna
http://www.par.univie.ac.at

Department of Scientific Computing

Parallel Computing / HPC

- Programming Models and Languages
- Compiler and Runtime Technologies
- Programming Environments and Tools



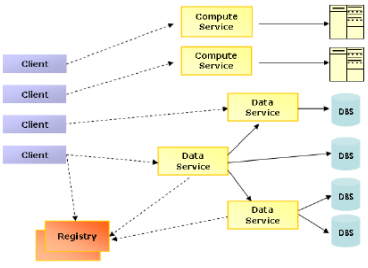


Vienna Fortran, HPF, HPF+, Hybrid Programming, Multicore…

Grid/SOA/Cloud Computing

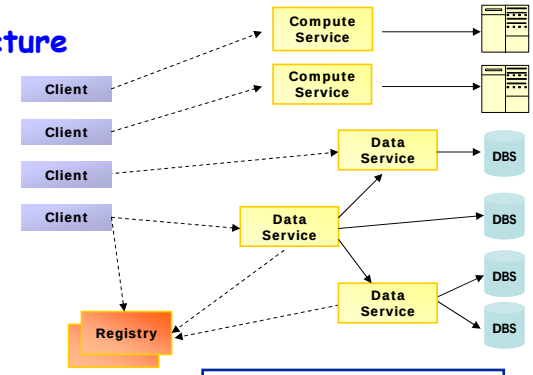
- Parallel Application Services
- On-demand supercomputing
- Data Virtualization & Integration & Mining

Grid Miner, Vienna Grid Environment, Cloud, …



Vienna Grid Environment (VGE)

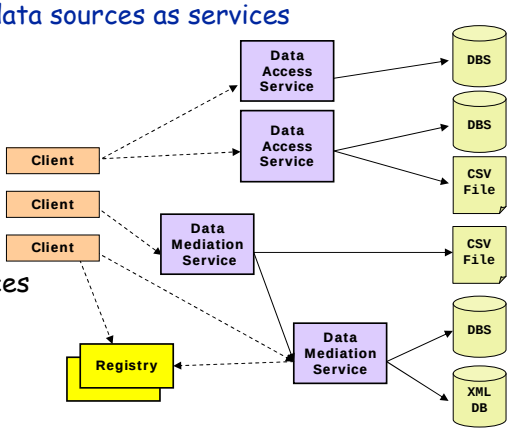
- **Service Oriented Architecture**
  - Compute Services
  - Data Services
- **Virtualization**
  - HPC-Applications-as-a-Service
  - Data-as-a-Service
- **Service Environment**
  - Service provisioning, deployment & hosting
- **Client Framework**
  - High-level client API; Workflow support



Capabilities
- Application provisioning
- Data provisioning
- Job handling
- Query handling
- Client APIs
- Security

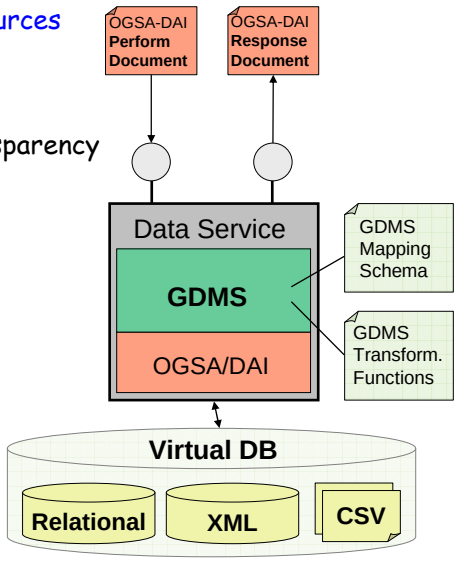
@neuInfo Data Services

- Virtualization of heterogeneous data sources as services
- **Data Access Services**
  access to single data source
- **Data Mediation Services**
  integration of multiple data sources via single virtual schema
- Based on standards
  - OGSA/DAI, OGSA/DQP
  - SQL, XML



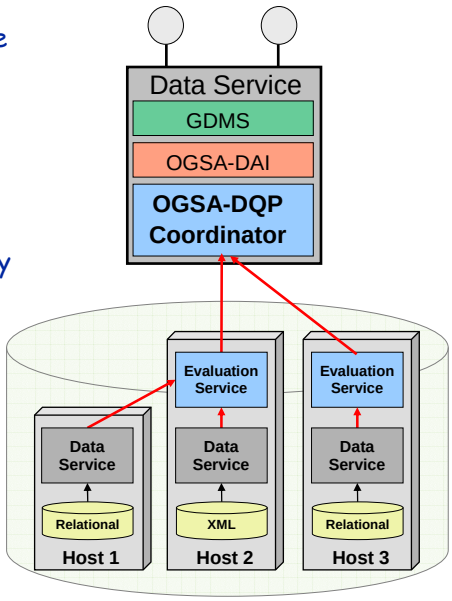
Data Mediation Services

- Transparent access to multiple data sources
  - Virtual global schema
  - Data stays where it is; always live
  - Schema, language & interface transparency
- GDMS Mapping Schema
  - Global-as-View query reformulation
  - Different views of data
- GDMS Transformation Functions
  - On-the-fly data transformation via user-defined Java methods



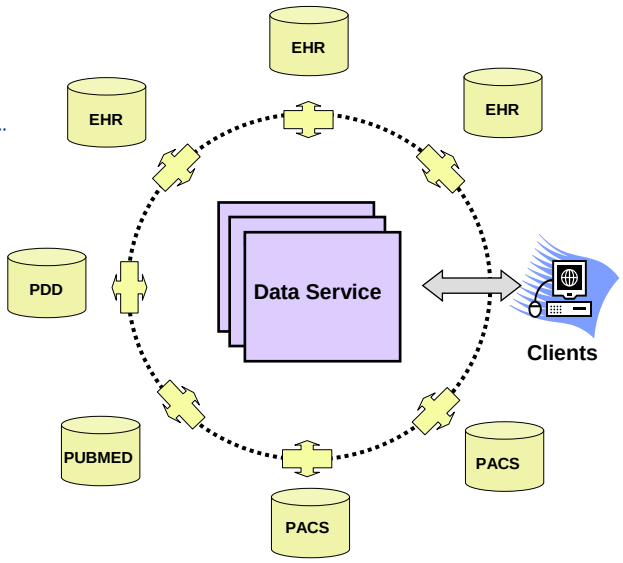
Distributed Query Processing

- Optimize complex queries using multiple evaluation services on different hosts.
- based on OGSA-DQP
- GDMS generates query plan from query against global schema
- DQP coordinator service distributes query plan onto evaluation services
- Evaluation services execute parts of query plan in parallel.



@neurIST Data Integration Scenario

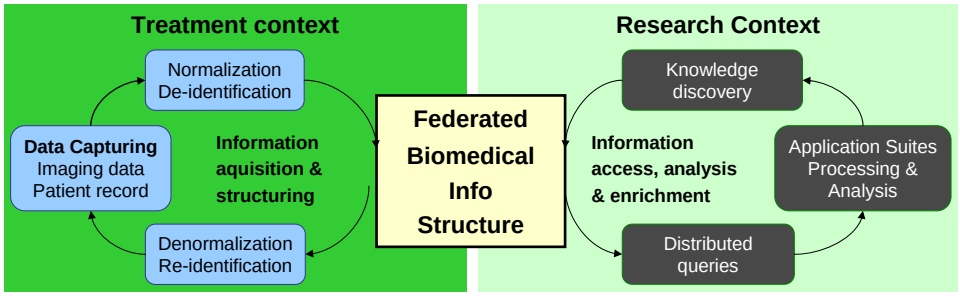
- **Approach**
  - Semantic Data Mediation
  - Federation of Services
  - **CRIM**, Ontology
  - Security, Pseudonymization, …
- **Hospital information systems**
  - Sheffield, Geneva, Rotterdam,…
  - EHR, PACS, …
- **Public databases**
  - Genetic: EBI,NCBI
  - Literature: Medline,
  - etc.
- **Product design databases**
  - COTS stents, coils, etc.



Clinical Reference Information Model (CRIM)

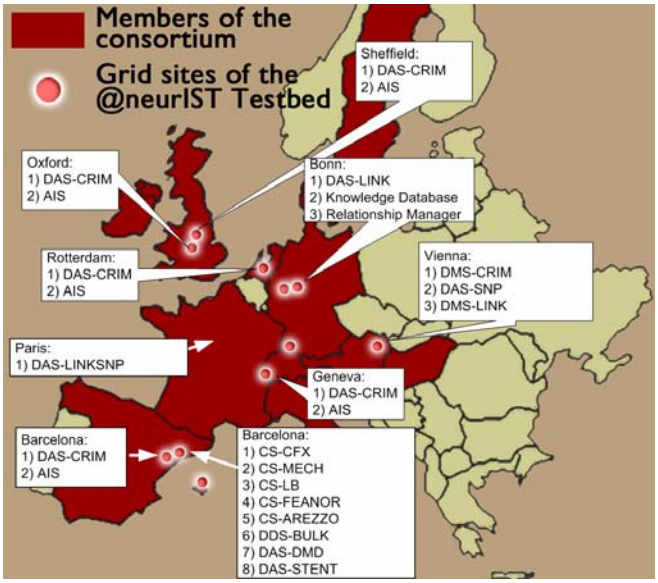
Defines all information to be captured for a patient

- **clinical information (imaging, diagnostic and treatment data, …)**
- **administrative information**
- **research results produced (indicators)**

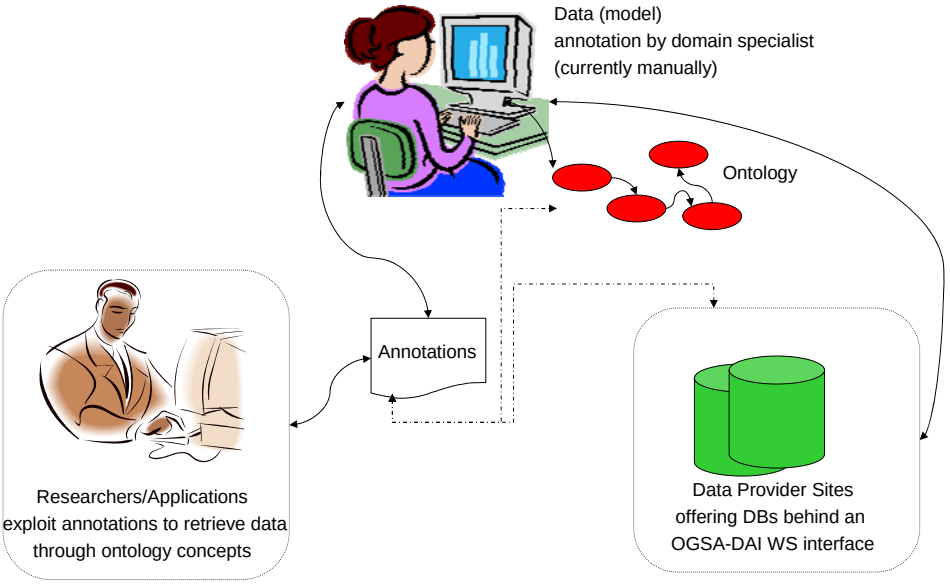


- **Biomedical data infostructure** – two different architectures
  - ANO: CIS → anonymized DB
  - OTF: on-the-fly access to CIS

@neurIST Testbed

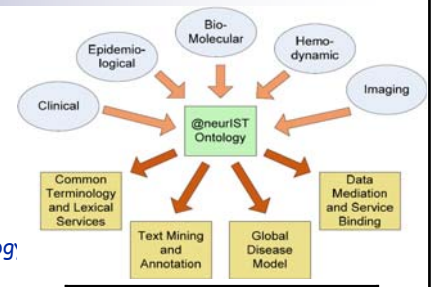


Semantic Support for Data Retrieval



@neurIST Semantic Technologies

- @neurIST Ontology
  - Global "schema" of the disease
  - Implemented in OWL-DL
  - Incorporates existing ontologies
    - » FMA (Foundational Model of Anatomy)
    - » GO (Gene Ontology), DOLCE as Upper Ontology
    - » Concepts mapped to UMLS (Unified Medical Language System)



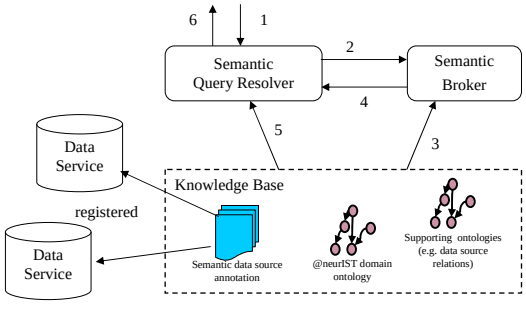
Classes	Relationship Types	UMLS Map
2319	**85**	**1183**

- Semantic support in @neurIST
  - Semantic annotation of services
  - Semantic broker (semantic service discovery)
  - Semantic query resolver (reduce relational complexity)
  - Semantic mediation between data sources (generation of mapping files)

Semantic Query Resolver

Goal: simplify access to distributed data sources utilizing ontology concepts

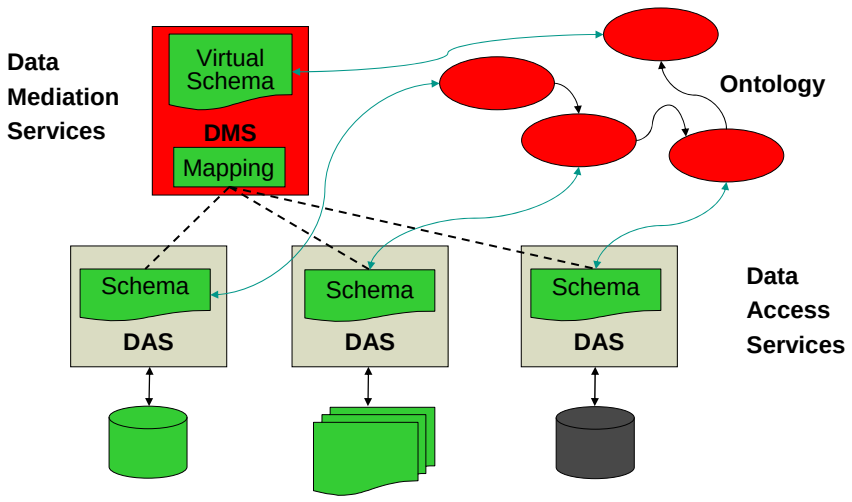
- Semantic Broker: „What data to combine?"
- Semantic Query Resolver: „How to combine?"
  - reduces relational complexity
  - (semi-)automatic generation of mapping schemes



Not fully realized within @neurIST!

SQR based on UNITY framework by University of British Columbia.

Semantic Data Integration



Developmemts beyond @neurIST

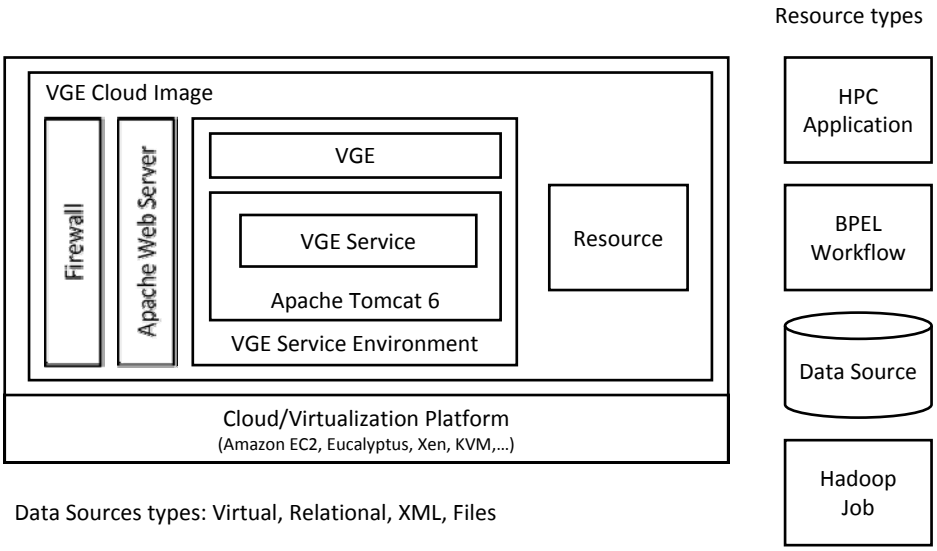
- **Optimized Download Mechanisms for Data Services**
  - SOAP Attachments (standard mechanism)
  - Data blocks (speed-up up to 5X)
  - via HTTP URL (speed-up up to 10X)

- **Support for Cloud Computing**
  - deployment of compute services and data services within Cloud
  - Ubuntu, Eucalyptus

- **Workflow Services**
  - Based on WEEP workflow engine; WS-BPEL v. 2.0 compliant

- **Large-Scale Data services**
  - based on Hadoop HDFS; Map/Reduce framework
  - installation on 64 core cluster at Vienna

Cloud-enabled VGE



Data Sources types: Virtual, Relational, XML, Files

Large-Scale Data Services

- Hadoop Installation on (virtual machine) cluster
- Name Node
  - Start Hadoop job
  - Distribute Map and Reduce Tasks
- Data Nodes
  - Execute Map and/or Reduce tasks
- HDFS file system
  - Replicate Files
  - Partition Files