ACGT
Advancing Clinico Genomic Trials on Cancer

ACGT Knowledge Discovery Services and Workflows

Stelios Sfakianakis
FORTH-ICS

http://www.eu-acgt.org

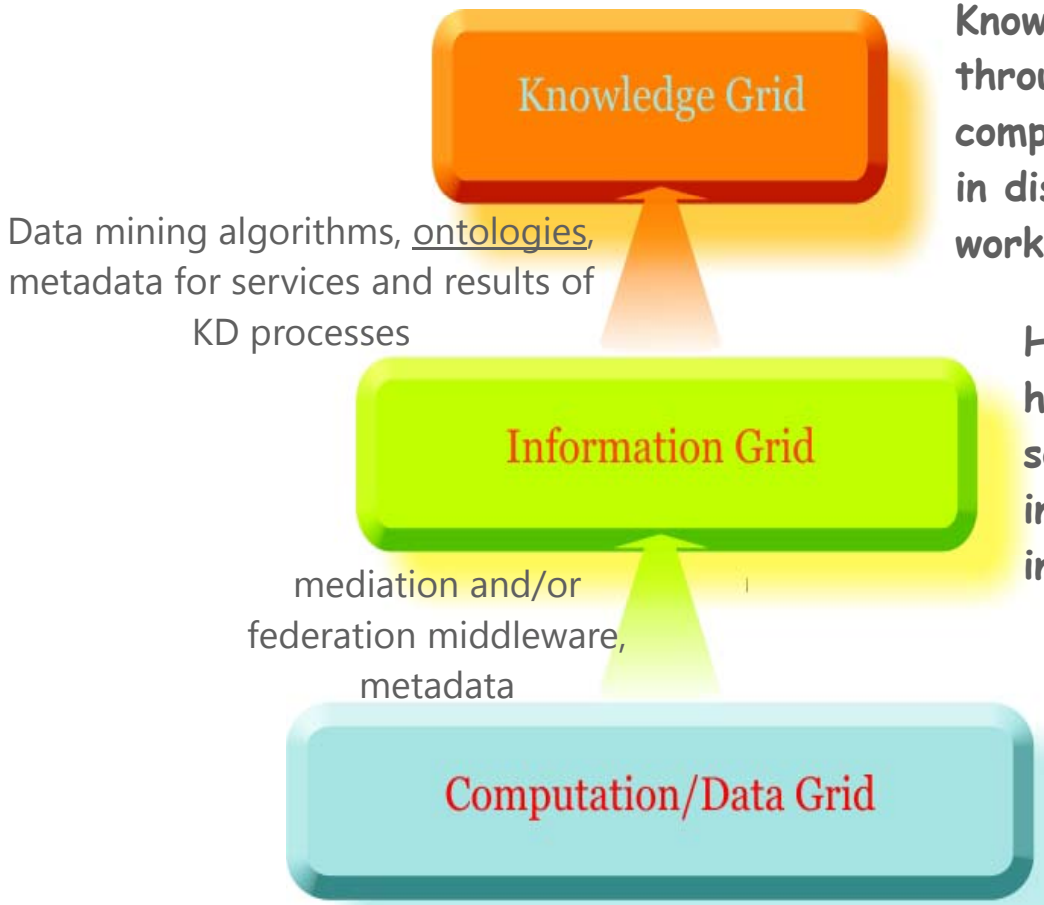
Information Society and Media

The ACGT project (FP6-2005-IST-026996) is funded by the European Commission Information Society and Media DG under the 6th Framework Programme.

Knowledge Discovery in the post genomic era

- Fact: collection of biological information at an unprecedented level of detail and in extremely large quantities
- In clinical trial research standard statistical methods do not actually fit:
  - Small number of patients in a clinical trail
  - But large quantities of genomic/proteomic data for each patient
- In many cases, instead of the analysis of huge data sets, the problem lies in the analysis of many small data sets with a plethora of possible analysis workflows

Objective: From Data to Knowledge



Knowledge extraction through semantics aware composition of services in discovery driven workflows

Data mining algorithms, ontologies, metadata for services and results of KD processes

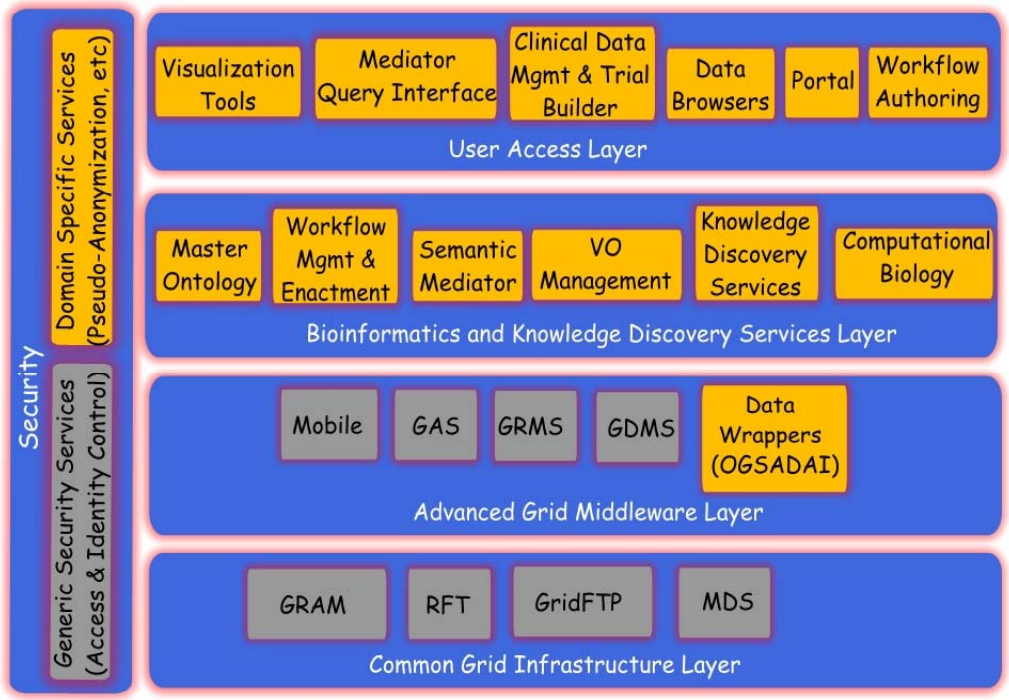
Homogeneous access to heterogeneous data sources, federation and integration of information

mediation and/or federation middleware, metadata

Data management, job scheduling and load balancing, security, etc.

The ACGT layered architecture

- Built on
  - Web Services
  - Grid
  - Semantic Web
- ✓ A *service oriented*, *ontology driven*, and *Grid enabled* platform



The major actors

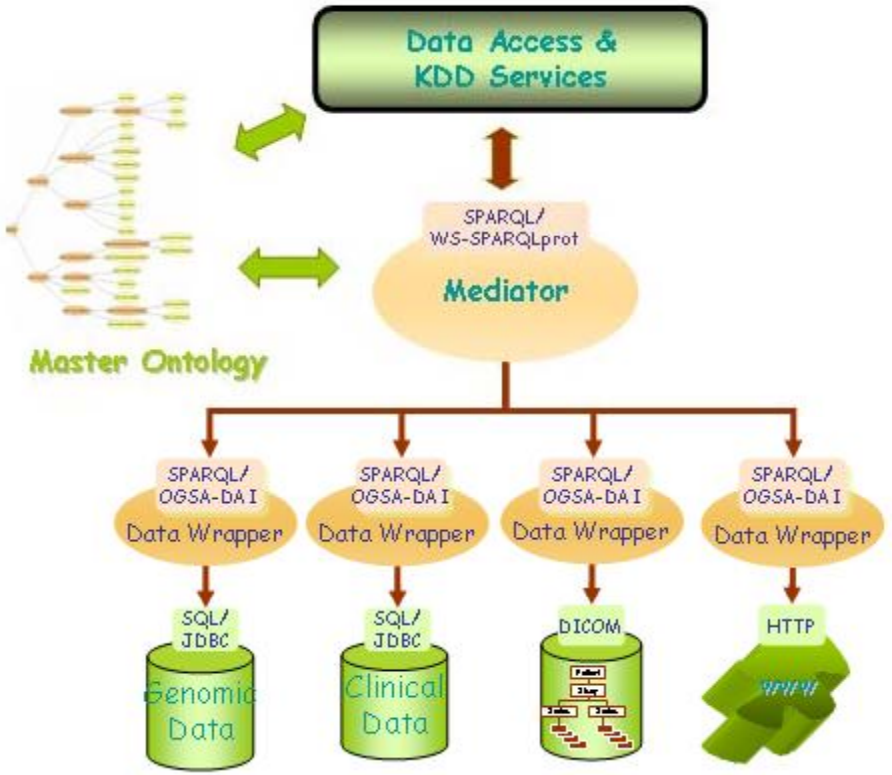
- Information integration
  - Master Ontology for Cancer
  - Semantic Mediator
- Knowledge Discovery and Data Analysis tools and services
- Service Composition layer
  - “Functional” metadata
  - Workflow Editor
  - Workflow Management and Enactment

Master Ontology for Cancer and Ontology Services

- Master Ontology for Cancer
  - The “domain of discourse”, i.e. what is available, what can be asked or reason about
  - Actively and orderly maintained
- Ontology services
  - Ontology viewers and browsers
  - Query designers

The ACGT Mediator

- Multiple Data sources:
  - Genomic, Clinical, Imaging, Public Databases
- A central issue is the data integration
  - Data warehousing vs. Query translation
  - Global as View (GaV) vs. Local as View (LaV)
- Mediation Layer
  - Supported by the ACGT Master Ontology
  - Used by other components (Data Access Services and Knowledge Discovery Services)
- "Open ended" mediation
  - Mapping format & tool



Data Analysis and Knowledge Discovery Services

- Mediator performs the data aggregation and data normalization tasks of the analyses
- Analytical services to cater for:
  - statistical analysis,
  - classification and clustering,
  - association rules mining,
  - text and literature mining,
  - domain specific analysis (e.g. sequence alignment),
  - etc.

- Reuse of existing tools and know-how
  - E.g. R environment for statistical computing
- Taking advantage of the Grid resources
  - Grid-R, both as a client UI and as a "gridified" server side component
- Leveraged by Semantic Descriptions and Ontologies
  - Statistical and Data Mining specific metadata descriptions for KD services and data sets

Service composition

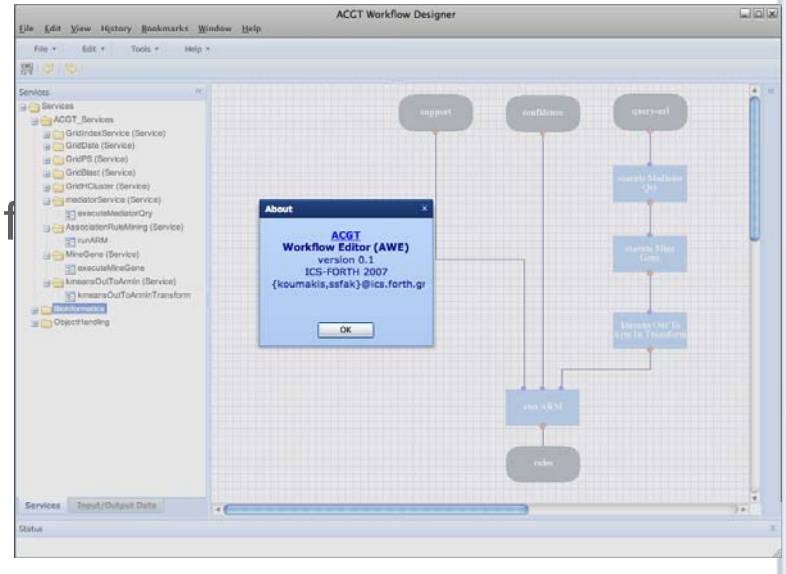
- New analysis and KD scenarios can be tested by combining existing services
- ACGT aims to support the “ad hoc” composition of different data access and knowledge extraction and analytical services into complex workflows
- The users extend functionality of ACGT by creating and *sharing* their workflows

Workflow management

- Necessary components:
  - Workflow Editor, used for the definition of new workflows
  - Workflow Engine, used for the execution of the workflows
  - Metadata repository, used for the persistence and maintenance of services and workflows metadata descriptions

ACGT Workflow Editor

- Web based (integrated into the ACGT web portal)
- Efficient browsing and searching of available services and workflows
- Syntactic and Semantic validation of workflows
- Multiple views and abstraction layers
  - E.g. supports the use of R scripts as Web Services
- WS-BPEL export



Workflow Engine

- Separate of the Workflow Editor to take advantage of the Grid infrastructure
  - Support long running workflows
  - Utilize the Grid Data Management
- WS-BPEL v2.0 compliant
- Workflows as composite Services
  - Workflows are services that can be combined even further into more higher level workflows

Metadata

- "*Metadata is structured, encoded data that describe characteristics of information-bearing entities to aid in the identification, discovery, assessment, and management of the described entities*" [American Library Association, Task Force on Metadata Summary Report, June 1999]
- Necessary for the semantic discovery, interoperability, and composition of services and tools
- The need of sophisticated model for data provenance and the usage of workflows presents also the need for complex metadata about workflows

"Functional" metadata

- Metadata for services and KD "operators" are necessary
  - Classification of functionality
  - Input/output parameters data types
  - Domain specific classification of "kinds" of data irrespective of representation (e.g. DNA sequence)
  - Service composition (e.g. what are the constituent services of a workflow)
- Provenance
  - E.g. Back-linking a data set to the service/operator that created it and other relevant information
- Metadata Repository: the metadata guardian

Conclusions

- "Grid intelligence": The convergence of web service, grid and semantic web technologies using ontologies and metadata
- Semantic integration in ACGT relies on metadata publishing and ontologies
- Knowledge Discovery in Grid as an extension of KD in Databases is greatly facilitated by the Grid infrastructure but it also presents new challenges
- ACGT aims to build an open architecture and the necessary infrastructure to enable KD in Grids


Thank You