



Advancing Clinical Genomic Trials on Cancer



The ACGT Data Access Infrastructure

Luis Martín (lmartin@infomed.dia.fi.upm.es)

ACGT Workshop, EGEE 07

05/10/2007



<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.

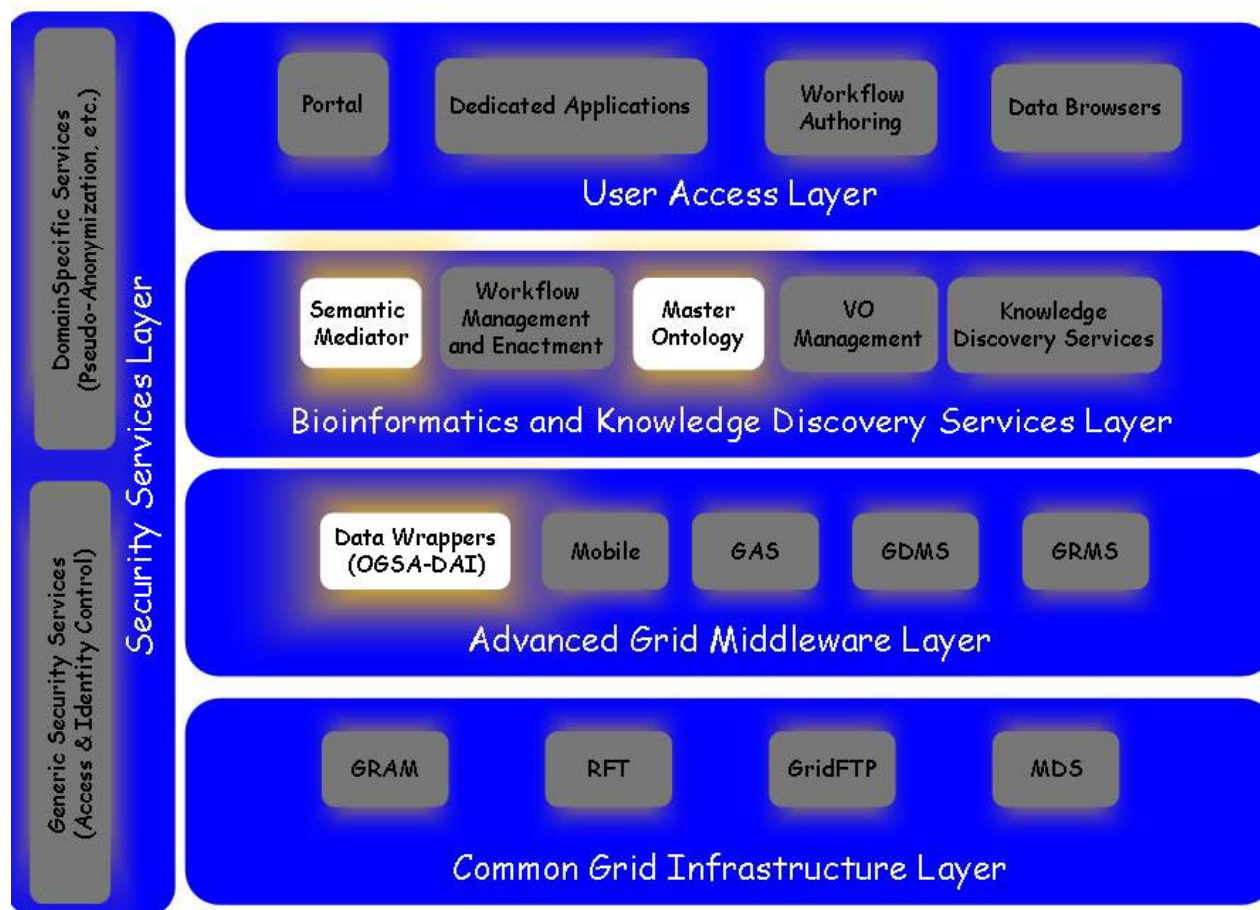
Aims

- ▶ Provide homogeneous, seamless *access to heterogeneous* (in terms of syntactics and semantics) *sources of information*.
- ▶ Provide *querying services* to both end users and data analysis tools.

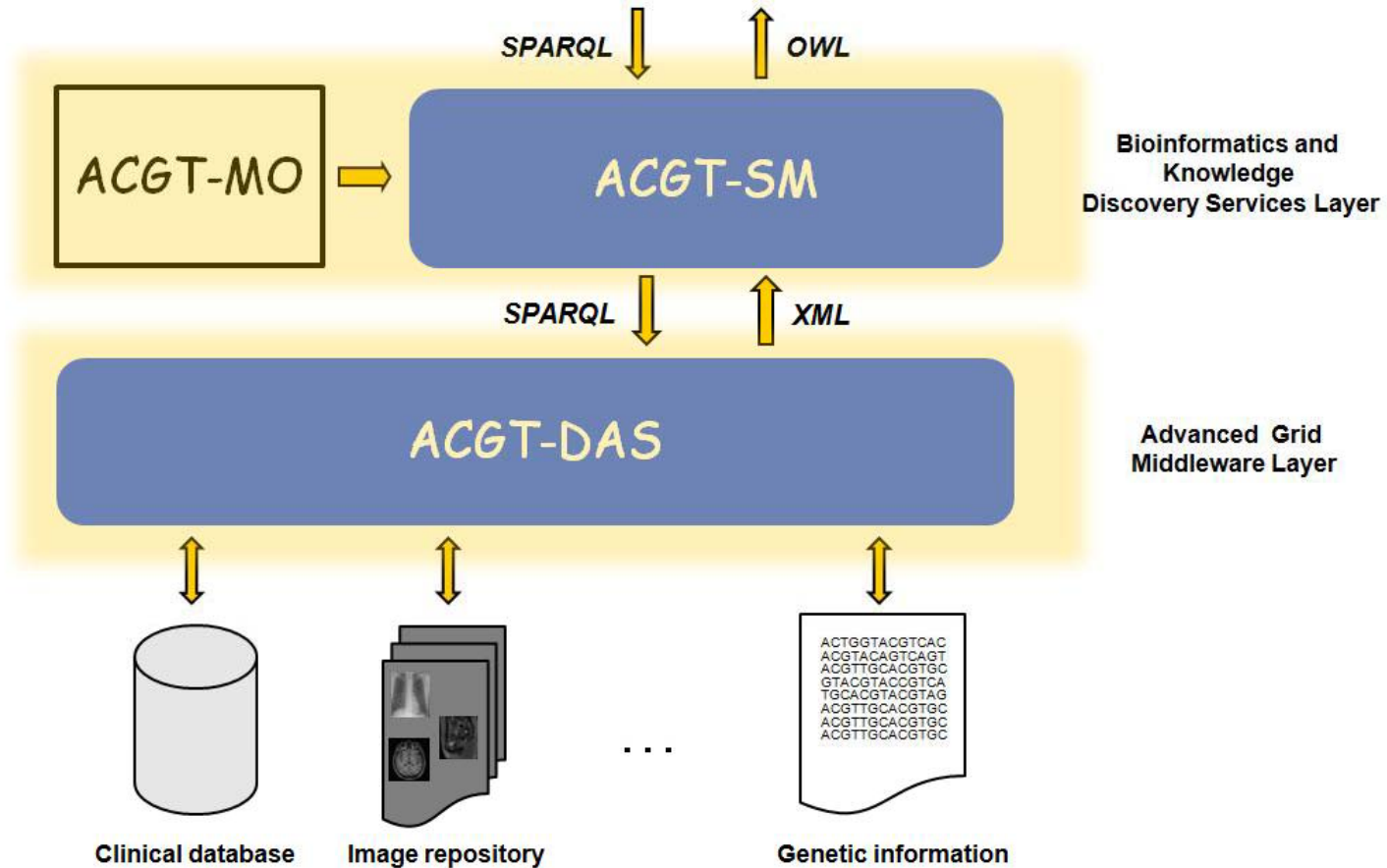
Main Resources

- ▶ Within the framework of the Data Access Infrastructure in ACGT, several tools are being developed, namely:
 - ▶ *The ACGT Master Ontology on Cancer*
 - ▶ *The ACGT Semantic Mediator*
 - ▶ *The ACGT Data Access Services*

Data Access Infrastructure within ACGT



Data Access Architecture



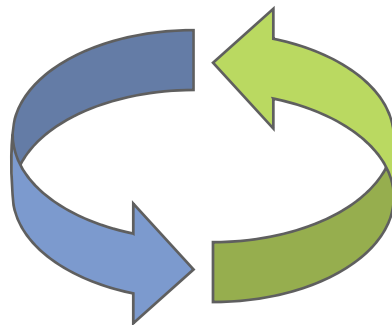
The ACGT Master Ontology

- ▶ The ACGT Master Ontology on Cancer aims at:
 - ❖ Enhancing cancer management in Europe by enabling *semantic interoperability*.
 - ❖ Meeting all necessary preconditions of the *project infrastructure*.
 - ❖ Creating an ontology that is *both philosophically and technically valid and sound*.

Development Procedure



- ▶ Continuous iterative development process that includes domain experts via face-to-face meetings, online telcos and e-mail discussions
- ▶ At all times feedback is highly encouraged and integrated in the development

ontology
developers



clinicians
researchers

Introduction

- ▶ Following examples are taken from the clinical trial forms of the TOP trial on breast cancer
- ▶ Another source are the forms from clinical trials on nephroblastoma done by  and 

The Trial of Principle (TOP trial)

Case Report Form

The Trial of Principle

PROSPECTIVE EVALUATION OF TOPOISOMERASE II ALPHA GENE AMPLIFICATION AND PROTEIN OVEREXPRESSION AS MARKERS PREDICTING THE EFFICACY OF EPIRUBICIN IN THE PRIMARY TREATMENT OF PATIENTS WITH BREAST CANCER

Registration Number: _____

Patient initials: _____

Institution number: _____

Responsible physician: _____

Instructions: Please fill out the CRF and fax it to the J. Bordet Institute at the fax no. +32 2 5413090

Top Trial #: 23 Aug 2006 Page 1

The Trial of Principle (TOP trial)

Registration Number: |_|_|_|_|_|_| Patient initials: |_|_|_|

Patient's Characteristics

- **Height** |_|_|_|_| (cm)
- **Weight** |_|_|_|_|_| (Kg)
- BSA |_|_|_| (m²)
- Menopausal status:
 - ₁ premenopausal (< 6 months since last menstrual period (LMP) and no prior ovariectomy and no estrogen replacement therapy)
 - ₂ postmenopausal (prior bilateral ovariectomy, or > 12 months since LMP with no prior hysterectomy and not receiving LH-RH analog)
 - ₃ above category not applicable and < 50
 - ₄ above category not applicable and ≥ 50
- Significant medical history:
 - ₁ No
 - ₂ Yes, please specify below

Disease	Date started (day/month/year)	Date ceased	
		(day/month/year)	or Ongoing
	--/--/----	--/--/----	<input type="checkbox"/>
	--/--/----	--/--/----	<input type="checkbox"/>
	--/--/----	--/--/----	<input type="checkbox"/>
	--/--/----	--/--/----	<input type="checkbox"/>

Primary Breast Cancer

- Date of **Trucut Biopsy** --/--/----
(day/month/year)
- Trucut Biopsy identification number: _____
- Side of lesion ₁ Left ₂ Right



Ontology as Black Box

- ▶ Ontology has a heavily complex interface that is not transparent to the user
- ▶ End users interact with applications via specifications that are not understood by the consortium



Ontology Overview

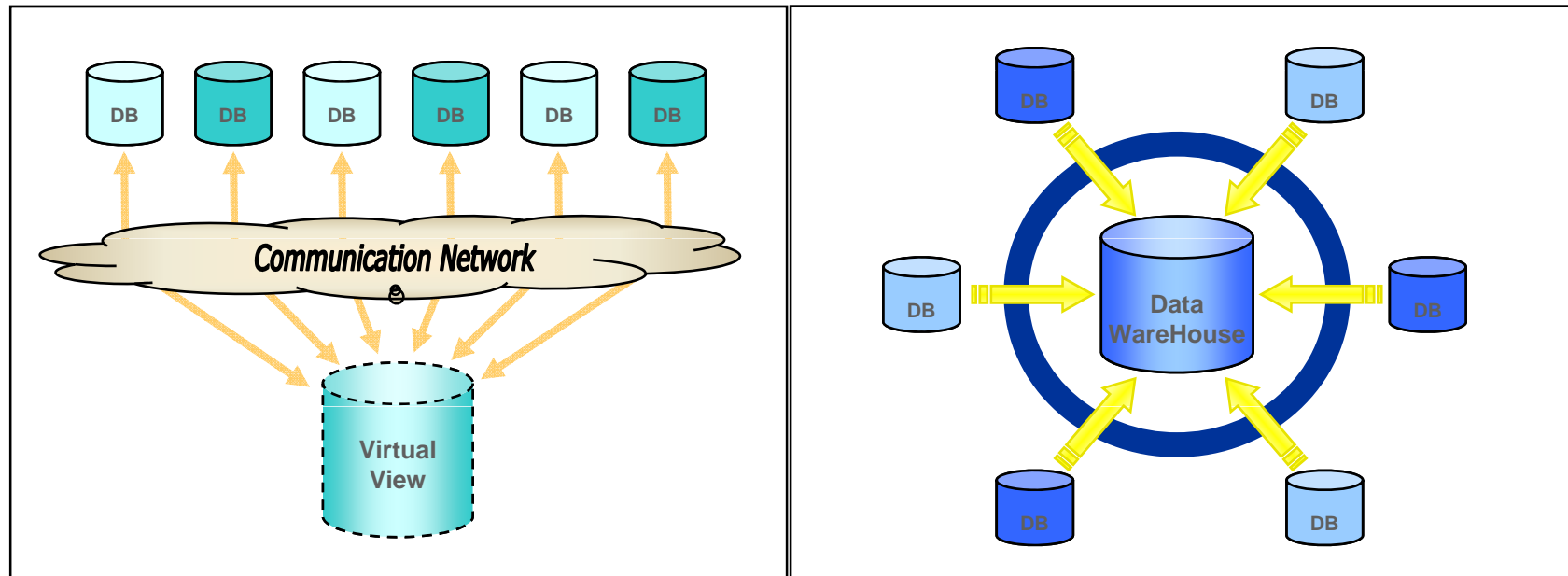
- ▶ Domain ontology that is oriented on clinical needs and integrates multiple subdomains.
- ▶ Implemented in OWL-DL to allow for automatic consistency checking and classification by reasoner applications.
- ▶ Formally defined „is-a“ and other restriction relations between classes (e.g. that each „Person has-a Name“ or that „Male“ is disjoint from „Female“).

The ACGT Semantic Mediator

- ▶ The ACGT Semantic Mediator aims at:
 - ❏ Providing *access to integrated repositories* of semantically heterogeneous databases.
 - ❏ Offering users a *friendly interface* to query these data.

Scientific Foundations of the Semantic Data Integration Approach (I)

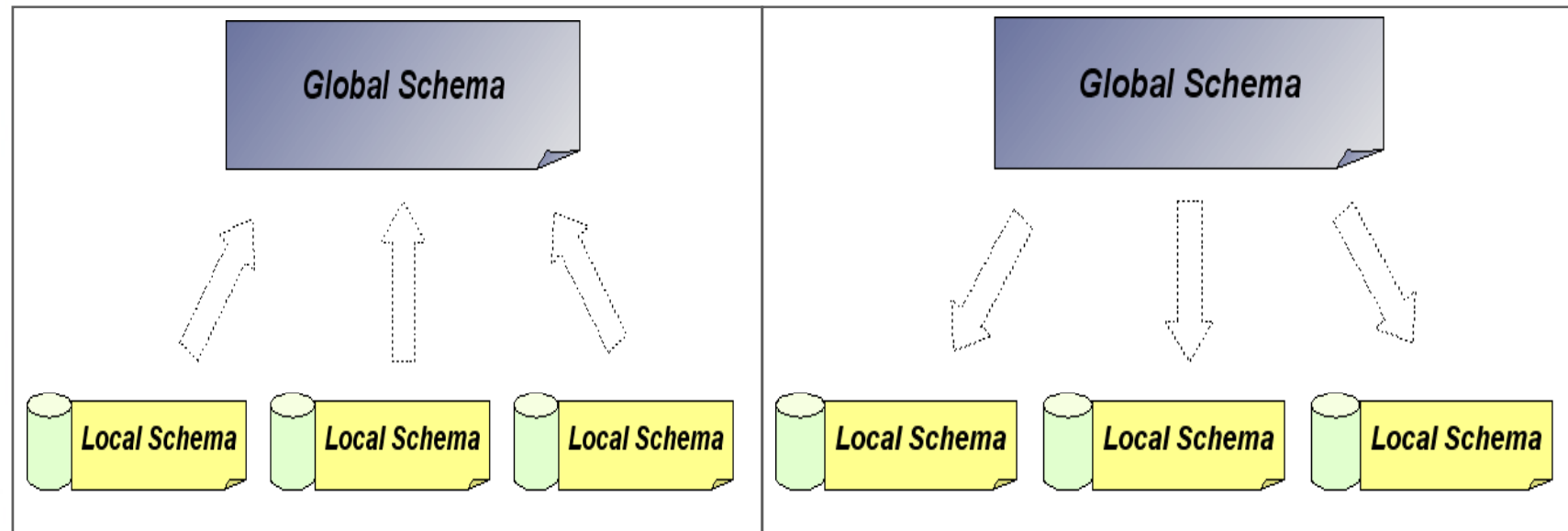
▶ Query Translation vs. Data Warehouses



▶ Given the nature of the data in ACGT, a query translation based approach was selected

Scientific Foundations of the Semantic Data Integration Approach (II)

▶ Global as View vs. Local as View



▶ A LaV based approach has been selected. Master Ontology will act as Global Schema.

The ACGT Semantic Mediator

- ▶ Data Integration using the mediator:
 - ❏ A query is performed using the interface.
 - ❏ The query is split, and different queries for the underlying databases are generated.
 - ❏ Queries are performed in the databases.
 - ❏ The results are returned and integrated.

Mediator: SIOP + Dicom query

```
SELECT
?PatientIdentifier.ClinicalTrialPatientNumber,
?PatientIdentifier.pnr,
...
WHERE
( ?a, rdf:type, h:PatientIdentifier ),
...
( ?a, h:PatientIdentifier.hasStudy.Study, ?b )
USING ...
```

```
PREFIX h: <http://gridnode.ehv-campus.philips.com/dicom/>;
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>;
SELECT ?PatientID ?PatientsName
WHERE {
OPTIONAL { ?a h:PatientID ?PatientID . }
OPTIONAL { ?a h:PatientsName ?PatientsName . }
}
```

```
SELECT DISTINCT patient.siopnr, patient.pnr,
...
FROM patient;
```


Results

```
<rdf:RDF
  xmlns:j.0="http://infomed.dia.fi.upm.es/SIOPDicom#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:j.1="http://www.w3.org/2001/XMLSchema#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  ...
  <owl:Class
    rdf:about="http://infomed.dia.fi.upm.es/SIOPDicom#PatientIdentifier"/>
    <owl:DatatypeProperty
      rdf:about="http://infomed.dia.fi.upm.es/SIOPDicom#PatientIdentifier.Clinical
      TrialPatientNumber">
      ...
      <j.0:PatientIdentifier
        rdf:about="http://infomed.dia.fi.upm.es/SIOPDicom#PatientIdentifier13">
          <j.0:PatientIdentifier.HospitalIdentifier>
            <j.1:string>
              <rdf:value>Without Information</rdf:value>
            </j.1:string>
          </j.0:PatientIdentifier.HospitalIdentifier>
          <j.0:PatientIdentifier.FirstName>
          ...
```

Issues identified

- ▶ The main technological issues identified were:
 - *Performance* issues associated to the LaV approach.
 - *Adoption of our own LaV approach to solve this issue.*
 - The ontology is near completion, and the nature of the ACGT domain suggests to maintain and update it regularly.
 - *For the second phase of ACGT, a task regarding technical and ontological refinement and the maintenance of the Master Ontology has been designed.*
 - Current query transformation methods (as known from a state of the art study) deal only with *relational databases*. This is not the case of some databases potentially useful for ACGT.
 - *An alternative mapping format that can tackle new structures was designed.*

Implementation experiences and challenges (I)

- ▶ Several implementation issues identified:
 - ❖ The *interaction with data access wrappers* is not easy. The Mediator works with RDQL, and the wrappers with SPARQL.
 - ❖ *Solution: Homogenization of the query language (Mediator to SPARQL)*
 - ❖ The *mapping format* does not fully exploits the power of the language used for representing the Master Ontology (OWL-DL).
 - ❖ *Solution: Define a new mapping format to cover all the power of OWL-DL*

Implementation experiences and challenges (II)




- ▶ Several implementation issues identified:
 - ❏ The *query interface* used in the first version of the mediator did not fully cover the cases of semantic and structural heterogeneity.
 - ❏ *Solution: Design and develop a new query interface covering the new requirements.*
 - ❏ RDQL and SPARQL do not fully cover the *expressiveness* of SQL (lack of aggregation functions).
 - ❏ *Solution: Perform the aggregation after receiving the results.*
 - ❏ There exist *performance* constraints associated to the LaV approach, as well as to the transformation of results to an OWL-compliant format.
 - ❏ *Solution: Restrict the mapping format just to cover user requirements. Development of a customized library to deal with OWL results.*

The ACGT Data Access Services

- ▶ The ACGT Data Access Services aim at:
 - Provide uniform interface
 - uniform transport protocol
 - uniform message syntax
 - uniform query syntax
 - uniform data format
 - Hide query peculiarities of data source
 - Hide query limitations of data source
 - Export data model of data source



Uniform interface

- ▶ Uniform transport protocol
 http
- ▶ Uniform message syntax
 XML + OGSA-DAI
- ▶ Uniform query syntax
 SPARQL

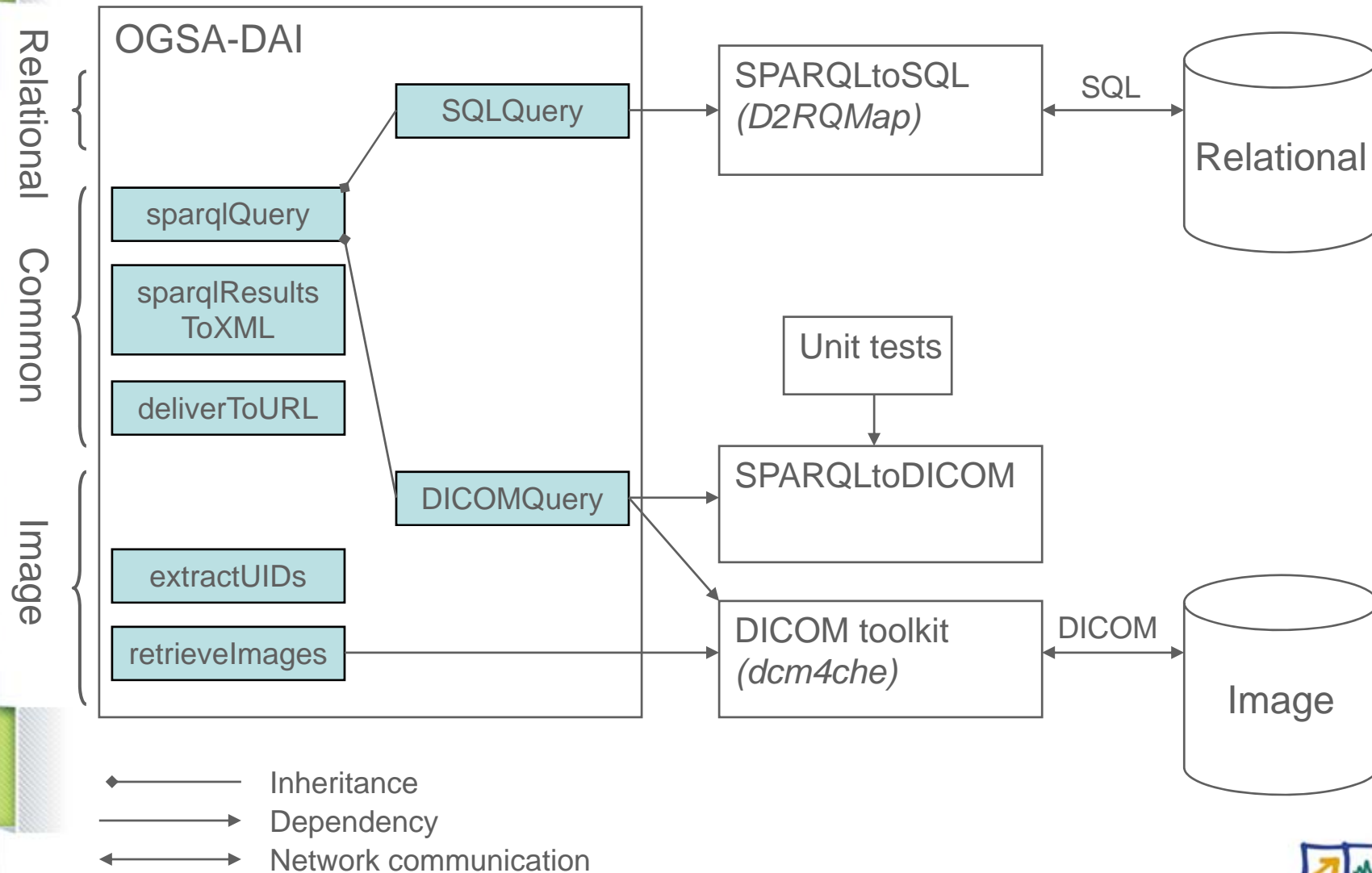
Main types of data sources

- ▶ Relational databases
 - 📁 CRF data, microarray data
- ▶ DICOM servers
 - 📁 Medical image data
- ▶ Public web databases
 - 📁 Gene and protein sequence databases
- ▶ Files in various formats
 - 📁 Excell, XML, comma separated

Technology choices

- ▶ OGSA-DAI
 - ▢ The standard web services framework for Data Access Interfaces
 - ▢ Supports activity framework for efficient and flexible services invocation
- ▶ SPARQL
 - ▢ Modern RDF query language
 - ▢ Fits needs of mediator
 - ▢ Intermediate level of expressiveness
 - ▢ E.g. more expressive than DICOM query capabilities, less expressive than SQL
 - ▢ Suitable as an initial query language for wrappers

Data services implementation



Uniform interface

SQL: SimpleQuery.xml

sparqlQuery
Statement

```
PREFIX vocab: <http://.../TestSqlResource>
SELECT ?name
WHERE {
  ?patient vocab:patient_patientsname ?name ;
  vocab:patient_patientid "100111" .
}
```

sparqlResults
ToXML

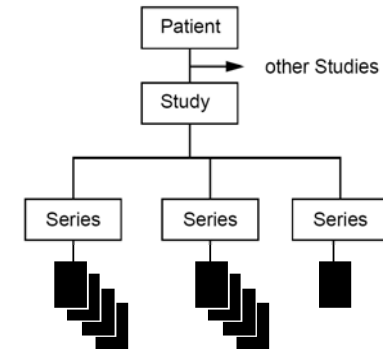
```
<head>
<variable name="name"/>
</head>
<results ordered="false" distinct="false">
<result>
<binding name="name">Rita Valencia</binding>
</result>
</results>
```

SPARQL for querying DICOM

- ▶ Uniform query syntax
 - 📁 Any DICOM query can be expressed as SPARQL
 - 📁 SPARQL does not impose any limitations
- ▶ Hide query limitations of data source
 - 📁 SPARQL filters can be used to create queries that cannot be expressed as DICOM queries
 - 📁 However, not all SPARQL queries can be efficiently converted to DICOM queries
 - 📁 Therefore, the data access service does not accept all queries
 - 📁 This is unavoidable, for performance reasons

Using the DICOM levels

Information Model



DICOM: LevelQuery.xml

sparqlQuery Statement

```

SELECT ?patientId ?studyId ?seriesId
WHERE {
  ?patient dicom:PatientID ?patientId ;
           dicom:PatientsName "Huge, Lurch" .
  ?study   dicom:Patient ?patient ;
           dicom:StudyInstanceUID ?studyId .
  ?series  dicom:Study ?study ;
           dicom:SeriesNumber "3" ;
           dicom:SeriesInstanceUID ?seriesId .
}
  
```

sparqlResults ToXML

```

...
<result>
<binding name="patientId">200650</binding>
<binding name="studyId">1.3.46.670589.5.2.12.
2158432007.1002671691.401594</binding>
<binding name="seriesId">1.3.46.670589.5.2.12.
2158432007.1002671552.91561</binding>
</result>
...
  
```



Ongoing Work

- ▶ New version of the Semantic Mediator (and related tools)
 - ▣ SPARQL language
 - ▣ Enhanced Mapping format
 - ▣ Mapping API and Tool

- ▶ Master Ontology viewer
 - ▣ Making ontology navigation friendly for end users

Ongoing Work

- ▶ New Data Access Services
 - 📁 Web public databases
 - 📁 File databases
 - 📁 Microarray data





Advancing Clinical Genomic Trials on Cancer



Thank you



<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.