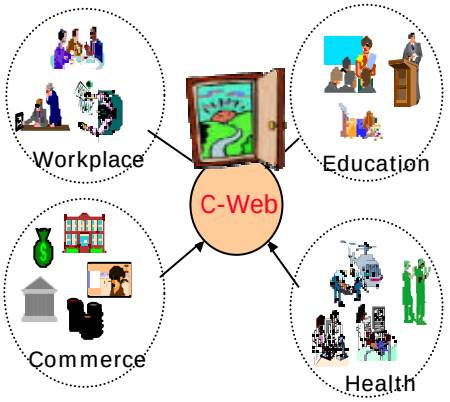
Declarative Languages for Querying Portal Catalogs

V. Christophides
D. Plexousakis
G. Karvounarakis
S. Alexaki

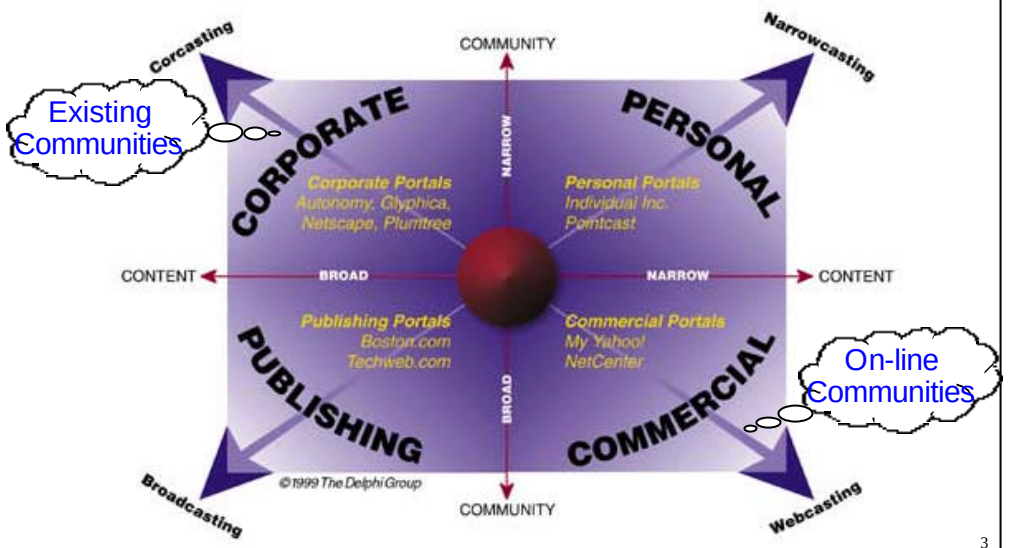
Computer Science Department, University of Crete
Institute for Computer Science - FORTH
Heraklion, Crete

What is a Community Web ?

- A group of people sharing a domain of discourse and a set of information resources (e.g., data, documents, services) and having some common interests
  - Commerce, Education, Health
- The main requirement is to provide a single point of useful, ubiquitous comprehensive, and integrated access to various information resources
  - Web Portals

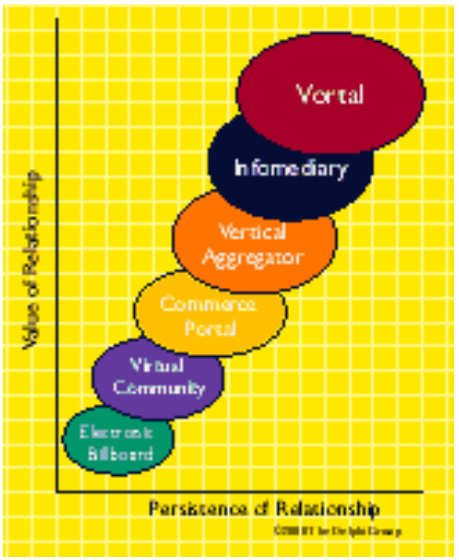


Portals Classification



3

Portals Classification



4

Knowledge Processes in Corporate Communities



Generating new knowledge







Accessing knowledge from external sources





Representing knowledge in documents and databases



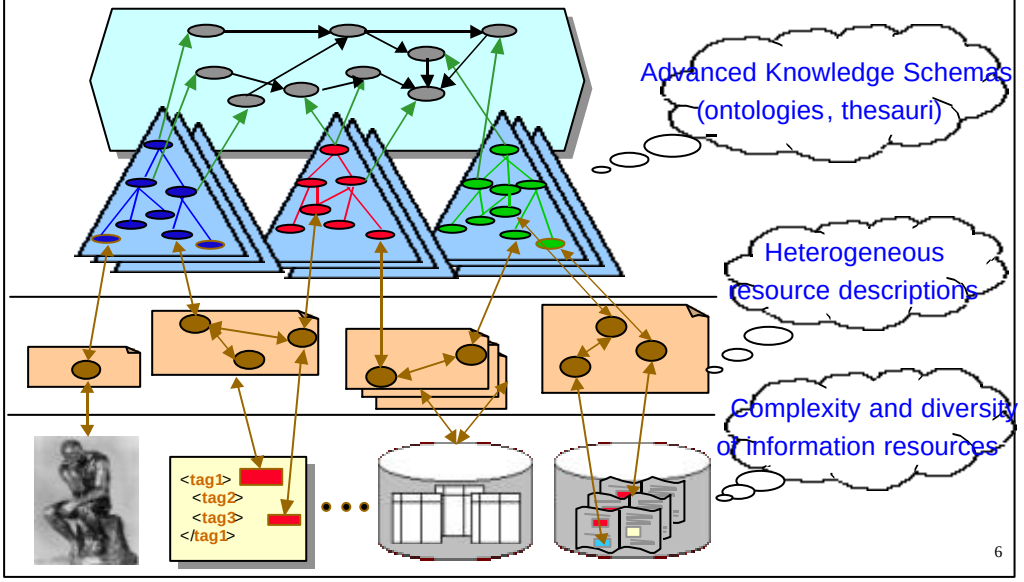
Embedding knowledge in services and processes



Dissemination of knowledge within organisation

Using knowledge in decision making

Community Web Portals: Main Features



Advanced Knowledge Schemas (ontologies, thesauri)

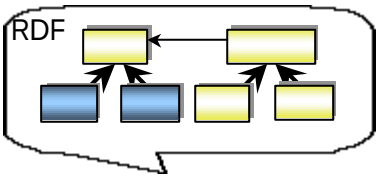
Heterogeneous resource descriptions

Complexity and diversity of information resources

What we need?

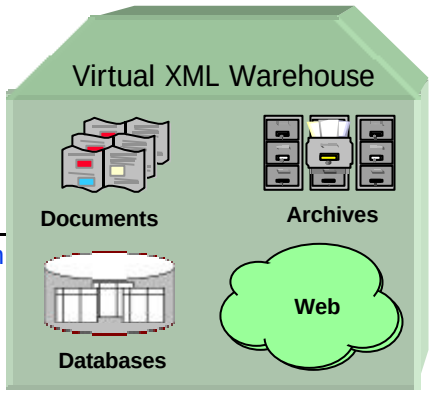
- Flexible Modeling of Community Web Portal Catalogs
  - ❶ Exploit existing forms of community knowledge
    - Ranging from simple vocabularies to formal ontologies
  - ❷ Describe in various ways community information resources
    - Administration, Classification, Content Rating, Channels, ….
- Secondary Storage Management of Portal Metadata
  - ❶ Large Schemas: e.g., 100 Mbytes of ODP Topics (the Art Hierarchy contains 20000 terms)
  - ❷ Voluminous Description Bases: e.g., 700 Mbytes of ODP indexed URLs
- Declarative Query Languages for Portal Catalogs
  - ❶ Interleave schema with data querying
  - ❷ Optimize access to Portal Catalogs

Our Approach





High-level Access to community information

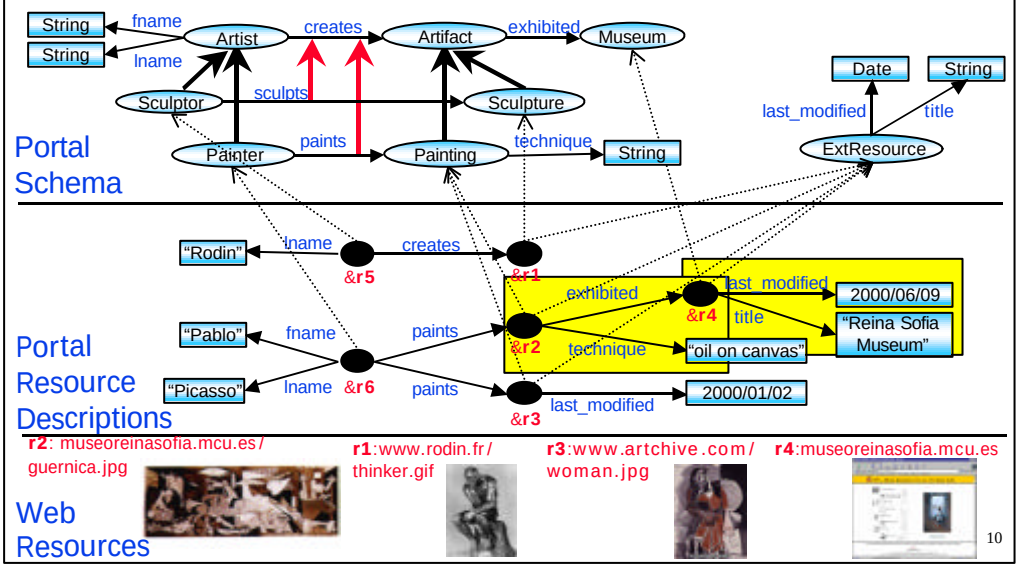


- Use W3C Standards to describe (RDF/S) & exchange (XML) information
- Our Contribution: Declarative Languages for Browsing & Querying

Outline

- Example of a Portal Catalog for Cultural Communities
- Describing and Querying Community Resources
  - A Formal Data Model for RDF/S
  - The RDF Query Language (RQL)
- Architecture
  - Core Middleware (RDF Store, Parser/Loader, Query Interpreter)
- Build Community Web Portals
  - Create/Extract Resource Metadata
  - Develop Specific Applications for Communities
  - Optimize RDF Storage and Queries
- Summary

Building a Cultural Community Web Portal using RDF



Resource Description Framework (RDF/S)

- RDF: Resource Descriptions
  - Data Model: Directed Labeled Graphs
    - Nodes: Resources (URIs) or Litterals
    - Edges: Properties – Attributes or Relationships
    - Labels: Nodes (Class names) and Edges (Property names)
    - Statement: assertion of the form *resource, property, value*
    - Description: collection of statements concerning a resource
  - XML syntax
- RDF Schema (RDFS): Schema Vocabularies
  - Specialization of both classes & properties (simple & multiple)
  - Multiple classification under several classes
  - Unordered, optional, and multi-valued properties
  - Domain and range polymorphism of properties

RDF/XML Serialization: Data

```
<rdf:RDF xml:lang="en"
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:rdfs="http://www.w3.org/TR/2000/PR-rdf-schema-20000327#"
     xmlns="">
....
<Painter rdf:id="picasso132">
   <fname>Pablo</fname>
   <lname>Picasso</lname>
   <paints>
     <Painting rdf:about="http://museoreinasofia.mcu.es/guernica.gif">
       <exhibited>
         <Museum rdf:about="http://museoreinasofia.mcu.es"/>
       </exhibited >
       <technique>oil on canvas</technique>
     </Painting>
   </paints>
   <paints>
     <Painting rdf:about="http://www.artchive.com/woman.jpg"/>
   </paints>
</Painter>
 <ExtResource rdf:about="http://museoreinasofia.mcu.es">
   <title>Reina Sophia Museum</title >
   <lastmodified>2000/06/09</lastmodified>
<Sculptor rdf:id="rodin424" lname="Rodin">
   <creates>
     <Sculpture rdf:about="http://www.rodin.fr/thinker.gif"/>
   </creates>
</Sculptor>
</rdf:RDF>
```

RDF/XML Serialization: Schema

```
<rdf:RDF xml:lang="en"
    xmlns:rdf="http://www.w3.org/1999/02/
22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/TR/2000/
PR-rdf-schema-20000327#">
<rdfs:Class rdf:ID="Artist"/>
<rdfs:Class rdf:ID="Artifact"/>
<rdfs:Class rdf:ID="Style"/>
<rdfs:Class rdf:ID="Sculptor">
<rdfs:Class rdf:ID="Museum">
    <rdfs:subClassOf rdf:resource="#Artist"/>
</rdfs:Class>
<rdfs:Class rdf:ID="Painter">
    <rdfs:subClassOf rdf:resource="#Artist"/>
</rdfs:Class>
<rdfs:Class rdf:ID="Sculpture">
    <rdfs:subClassOf rdf:resource="#Artifact"/>
</rdfs:Class>
<rdfs:Class rdf:ID="Painting">
    <rdfs:subClassOf rdf:resource="#Artifact"/>
</rdfs:Class>
<rdf:Property rdf:ID="creates">
    <rdfs:domain rdf:resource="#Artist"/>
    <rdfs:range rdf:resource="#Artifact"/>
</rdf:Property>
<rdf:Property rdf:ID="paints">
    <rdfs:domain rdf:resource="#Painter"/>
    <rdfs:range rdf:resource="#Painting"/>
    <rdfs:subPropertyOf rdf:resource="#creates"/>
</rdf:Property>
<rdf:Property rdf:ID="sculpts">
    <rdfs:domain rdf:resource="#Sculptor"/>
    <rdfs:range rdf:resource="#Sculpture"/>
    <rdfs:subPropertyOf rdf:resource="#creates"/>
</rdf:Property>
<rdf:Property rdf:ID="exhibited">
    <rdfs:domain rdf:resource="#Painting"/>
    <rdfs:range rdf:resource="#Museum"/>
</rdf:Property>
<rdf:Property rdf:ID="technique">
    <rdfs:domain rdf:resource="#Painting"/>
    <rdfs:range rdf:resource="http://www.w3.org/
TR/1999/PR-rdf-schema-19990303#Literal"/>
</rdf:Property>
<rdf:Property rdf:ID="title">
    <rdfs:domain rdf:resource="#ExtResource"/>
    <rdfs:range rdf:resource="http://www.w3.org/
TR/1999/PR-rdf-schema-19990303#Literal"/>
</rdf:Property>
....
</rdf:RDF>
```


RDF/S vs. Well-Known Formalisms

- Relational or Object Database Models (ODMG, SQL)
  - Classes don't define table or object types
  - Instances may have associated quite different properties
  - Collections with heterogeneous members
- Semistructured or XML Data Models (Ï Å Ì, UnQL, YAT, XML Schema)
  - Labels only on nodes or edges
  - Class and property subsumption is not captured
  - Heterogeneous structures reminiscent to SGML exceptions
- KR Languages (Telos, DL, F-Logic)
  - Absence of complex values and n-ary relationships (bags, sequences)
  - Inadequate reasoning support

A Formal Data Model for RDF

- *Ç(Í,<)* is a *well-defined* hierarchy of classes/properties iff:
  - c ∈ **C** => c < Class
  - p ∈ **P** => p < Property
  - $p_1, p_2 \in$ **P** and $p_1 < p_2$ => *domain*(p_1) ≤ *domain*(p_2) and *range*(p_1) ≤ *range*(p_2)
- Type System:

  $t = t_L \mid t_U \mid \{t\} \mid [t] \mid (1{:}t + 2{:}t + \ldots + n{:}t)$
- Interpretation Function:
  - Literal types, $[[\ t_L\ ]] = \text{dom}(t_L)$
  - Bag types, $[[\ \{t\}\ ]] = \{\nu_1, \nu_2, \ldots, \nu_n\}$, $\nu_1, \nu_2, \ldots, \nu_n \in$ **V** are values of type t
  - Seq types, $[[\ [t]\ ]] = \{\nu_1, \nu_2, \ldots, \nu_n\}$, $\nu_1, \nu_2, \ldots, \nu_n \in$ **V** are values of type t
  - Alt types, $[[\ (1{:}t_1 + 2{:}t_2 + \ldots + n{:}t_n)\ ]] = \nu_l$, $\nu_i \in$ **V**, 1<i<n is a value of type t_i
  - c ∈ **C,** [[c]] = {ν | ν ∈ π(c)}∪{π(c') | c' < c}
  - p ∈ **P,** [[p]] = {[ν_1, ν_2] | ν_1 ∈ [[*domain*(p)]], ν_1 ∈ [[*range*(p)]]}∪{π(p') | p' < p}

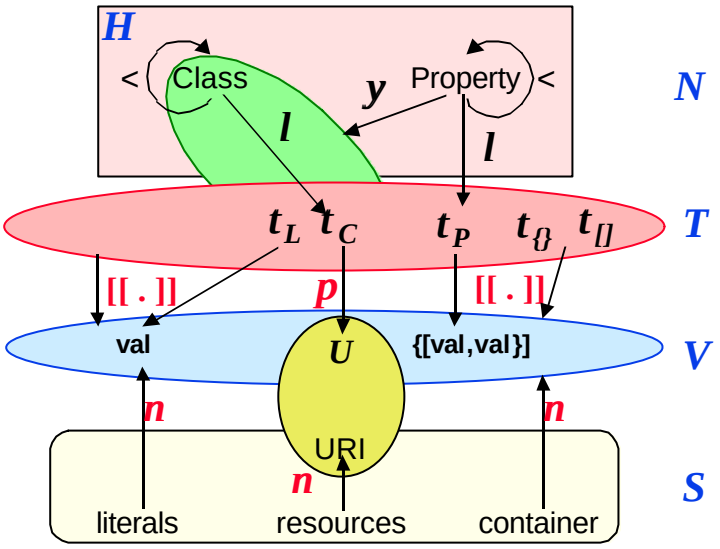
15

A Formal Data Model for RDF

- An RDF schema is a 5-tuple: RS = (V_S, E_S, H, ψ, λ)
  - V_S a set of nodes
  - E_S a set of edges
  - Ç = (Í,<) a well-formed hierarchy of names
  - λ a labeling function: $V_S \cup E_S \rightarrow$ Í ∪Ô
  - ψ an incidence function: $E_s \rightarrow V_s \times V_s$
- An RDF description base, instance of a schema RS, is a 5-tuple: RD = (V_D, E_D, ψ, ν, λ)
  - V_D a set of nodes
  - E_D a set of edges
  - ψ an incidence function: $E_D \rightarrow V_D \times V_D$
  - ν a valuation function: $V_D \rightarrow$ **V**
  - λ a labeling function: $V_D \cup E_D \rightarrow 2^{\text{Í}\cup\text{ô}}$:
    - ∀ u ∈ V_D, λ → n ∈ **C**∪**T:** ν(u) ∈[[n]]
    - ∀ e ∈ E_D [u,u'], λ → p

16

A Formal Data Model for RDF



The RDF Query Language

- Declarative query language for RDF description bases
  - relies on a typed data model (literal & container types + union types)
  - follows a functional approach (basic queries and filters)
  - adapts the functionality of semistructured or XML query languages to RDF, but also:
    - treats properties as self-existent individuals
    - exploits taxonomies of node and edge labels
    - allows querying of schemas as semistructured data
- Relational interpretation of schemas & resource descriptions
  - Classes (unary relations)
  - Properties (binary relations)
  - Containers (n-ary relations)

Browsing Portal Catalogs with RQL

- Simple set queries on class and property extents:
  - *Find the resources in the extent of the property creates*

    creates

    Includes paints & sculpts
    - {{ [www.portal.gr/rodin424, www.rodin.fr/thinker.gif], [www.portal.gr/picasso132, museoreinasofia.mcu.es/guernica.gif], [www.portal.gr/picasso132, www.artchive.com/woman.jpg] }}
  - *Find the resources of type painter and sculptor*

    ExtResource intersect Sculpture

    Multiply classified resources
    - {{ www.rodin.fr/thinker.gif }}
- Schema constructs used as query terms & support for automatic query expansion (similar to thesauri-based IRS)
  - Useful to query resources with minimal schema knowledge

Personalizing Portal Catalogs with RQL

- Navigational queries on semistructured resource descriptions
  - *Find the Museum resources that have been modified in year 2000.*

```
select  x
from    Museum{x}.last_modified{y}
where y >= 2000/01/01
```

    Data paths not foreseen in the schema
    - {{museoreinasofia.mcu.es}}
- Similar functionality to semistructured or XML query languages (Lorel, UnQL, XQL, XML-QL, XML-GL)
  - Useful in the absence of schema information or when multiple schemas are used to describe resources

Querying Portal Catalogs with Large Schemas

- Filtering both resource descriptions and schemas
  - *Find the paintings having as technique "oil on canvas" that have been created by a neo-impressionist painter*

Schema Filtering on Class hierarchies

Data filtering with schema information

```
select  y
from    {:$X}creates{y:Painting}.technique{z}
where $X <= neo-impressionist and z = "oil on canvas"
```


Querying Portal Schemas with RQL

- Pure schema queries
  - *Find the properties which specialize the property creates and may have as domain the class Painter along with their corresponding range classes*

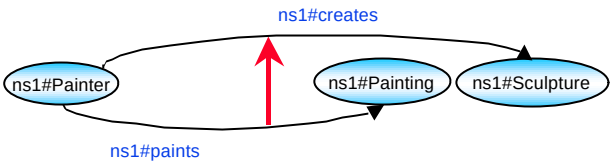
```
select  @P, $Y
from    {:Painter}@P{:$Y}
where @P <≡ creates
```

All Properties defined or inherited in class Painter

- {{ [**creates**, **Artifact**],
  [**creates**, **Painting**],
  [**creates**, **Sculpture**],
  [**paints**, **Painting**] }}

Schema filtering on property hierarchies

RQL: Examples



- Similar functionality to DBMS schema QLs (SchemaSQL, XSQL)
  - Useful for large schemas (integrating ontologies and thesauri)

Putting it all Together

- Nested schema and data queries
  - *Find the resources modified after 2000/01/01 which can be reached by a property applied to the class Painting and its subclasses*

```
select  R, y
from    (select @P
         from {:$X}@P
         where $X <= Painting){R}.{y}last_modified{z}
where z >= 2000/01/01
```

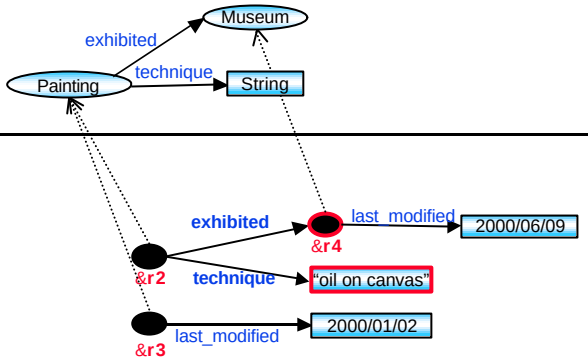
R ranges over the labels of type property

  - {{ [**exhibited**, museoreinasofia.mcu.es] }}

- Subcommunities may use different schemas while sharing the same description base

RQL:Examples

Portal Schema



Portal Resource Descriptions

Putting it all Together

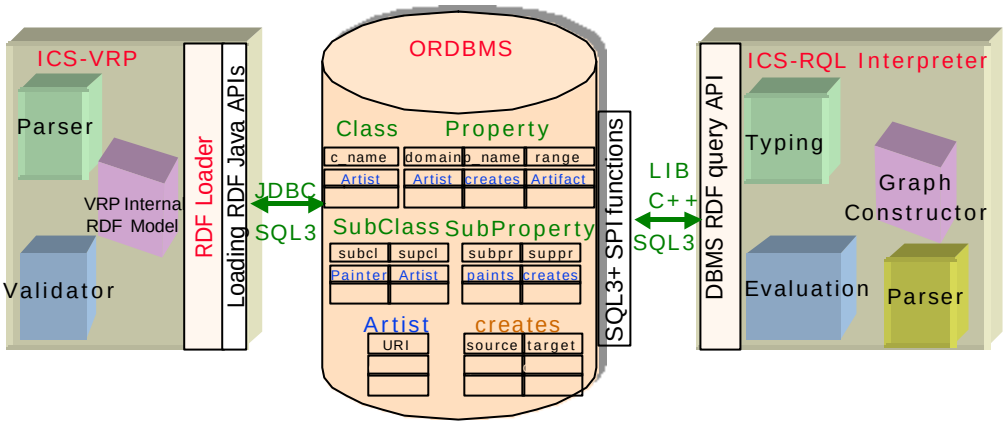
- Schema and data queries
  - *Find all metadata about the resources of the site museoreinasofia.mcu.es*

```
select  x,$$Y,$P,z,$$W
from    {x:$$Y}$P{z:$$W}
where x like "*museoreinasofia.mcu.es*" or
      y like "*museoreinasofia.mcu.es*"
```

URLs' pattern matching

  - {{[www.portal.gr/picasso132, **Painter**, **paints**, museoreinasofia.mcu.es/guernica.gif, **Painting**],
    [museoreinasofia.mcu.es/guernica.gif, **Painting, exhibited**, museoreinasofia.mcu.es, **Museum**],
    [museoreinasofia.mcu.es/guernica.gif, **Painting, technique**, "oil on canvas", **string],**
    [museoreinasofia.mcu.es, **ExtResource, title**, "Reina Sophia Museum", **string],**
    [museoreinasofia.mcu.es, **ExtResource, last_modified**, 2000/06/09, **date],** ….}}
- Subcommunities may use both different schemas and description bases

The Core Portal Middleware

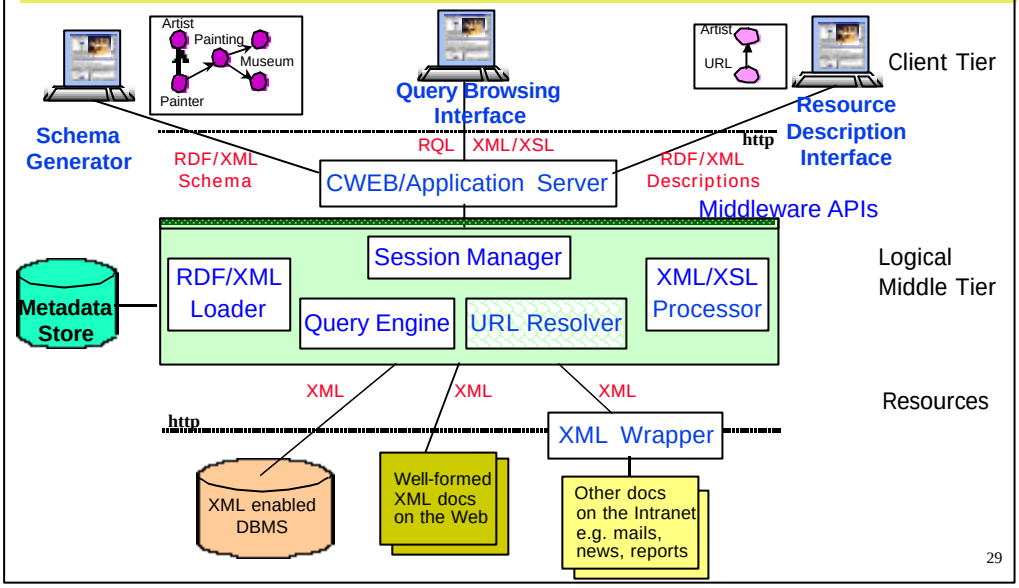


The ICS-FORTH R&D Activities

http://www.ics.forth.gr/proj/isst/RDF

- The Validating RDF Parser (VRP): Karsten Tolle Diploma Thesis
  - The First RDF Parser supporting semantic validation of both resource descriptions and schemas
- The RDF Schema Specific DataBase (RSSDB): Sophia Alexaki Ms. Thesis
  - The First RDF Store using schema knowledge to automatically generate an Object-Relational (SQL3) representation of RDF metadata and load resource descriptions
- The RDF Query Language (RQL): Greg Karvournarakis Ms. Thesis
  - The First Declarative Language for uniformly querying RDF schemas and resource descriptions

Architecture of Community Web Portals



Describing Community Information Resources

Metadata Type / Resources Nature	Individual	Collection	
		Static	Dynamic
Semantics Content Descriptive (unstructured, semistructured, structured)	Manual	Automatic	Query Mediation
Syntax Structure Descriptive (semistructured, structured)	Manual	Manual	Manual
System Management (all kinds)	Automatic	Automatic	Automatic

Related Work

- Information Integration Systems
  - YAT (INRIA Verso) www-rocq.inria.fr/~simeon/YAT
  - MIX (San Diego) www.db.ucsd.edu/projects/MIX
  - TSIMMIS (Stanford) www-db.stanford.edu/tsimmis
  - GARLIC (IBM) www.almaden.ibm.com/cs/garlic
  - DISCO (INRIA Caravel) rodin.inria.fr/Eprototype_Disco.html
  - DIOM (Alberta, Canada) ugweb.cs.ualberta.ca/~diom
  - Nomenclator (Bell Labs) cm.bell-labs.com/cm/cs/what/nomenclator
  - MANIFOLD (AT&T) www.research.att.com/~levy/imhome.html
  - Picsel (ORSAY) www.lri.fr/~picsel
  - CoBase (California) cobase-www.cs.ucla.edu
  - Momis (Modena, Italy) sparc20.dsi.unimo.it/Momis
  - InfoHarness & InfoQuilt (Georgia) lsdis.cs.uga.edu/proj/proj.html
  - Infomaster (Stanford) infomaster.stanford.edu/infomaster-info.html
  - InfoSleuth (MCC,Texas) www.mcc.com:80/projects/infosleuth
  - Ariadne (Southern California) www.isi.edu/ariadne
  - SIMS (Southern California) www.isi.edu/sims
  - HERMES (Maryland, College Park) www.cs.umd.edu/projects/hermes
  - Carnot (MCC,Texas) www.mcc.com:80/projects/carnot
- Main Focus on structured and semi-structured data integration

31

ICS-FORTH & Univ. of Crete
Christophides Vassilis December 2000

Related Work

- Ontology-based Information Systems
  - On-To-Knowledge (IST-1999-10132) www.ontoknowledge.org
  - IBROW (IST-1999-19005) www.swi.psy.uva.nl/projects/ibrow/home.html
  - OntoSeek (Padova, Italy) www.ladseb.pd.cnr.it/infor/ontology/ontology.html
  - Observer (Basque Country, Spain) siul02.si.ehu.es/~jirgbdat/OBSERVER
  - Ontological Mediation (New York at Buffalo) www.cs.buffalo.edu/~aec/OM
- Main Focus on knowledge integration
- Ontology-driven Document Annotation Systems
  - Ontobroker (Karlsruhe, Germany) ontobroker.aifb.uni-karlsruhe.de
  - SHOE (Maryland, College Park) www.cs.umd.edu/projects/plus/SHOE
  - WebKB (Griffith, Australia) meganesia.int.gu.edu.au/~phmartin/WebKB
  - OHTML, XAL
- Main Focus on intelligent document access

32

Related Work

- Corporate Knowledge Management and Ontological Engineering
  - CommonKADS (ESPRIT P5248) www.swi.psy.uva.nl/projects/CommonKADS
  - Cyc (MCC,Texas) www.mcc.com:80/projects/cyc
  - TOVE (Toronto) www.eil.utoronto.ca/eil.html
  - Enterprise (DTI, ISIP Programme) www.aiai.ed.ac.uk/~entprise
  - Kactus (ESPRIT Project 8145) www.swi.psy.uva.nl/projects/Kactus/home.html
  - KRAFT (Liverpool , U.K.) www.csc.liv.ac.uk/~kraft
  - O-Plan (Edinburgh , U.K.) www.aiai.ed.ac.uk/~oplan
  - Generalized Upper Model (GMD) www.darmstadt.gmd.de/publish/komet/gen-um/newUM.html
  - ONTOINT (CNR, Rome, Italy) saussure.irmkant.rm.cnr.it/onto/ontoint.html
  - PlanetOnto (Open University, U.K.) kmi.open.ac.uk/projects/planetonto
- Main Focus on domain modeling and knowledge representation

Related Work

- Digital Libraries
  - Electronic Environmental DL (Berkeley) elib.cs.berkeley.edu
  - Informedia (Carnegie-Mellon) www.informedia.cs.cmu.edu
  - Interspace (Illinois) dli.grainger.uiuc.edu
  - UMDL (Michigan) www.si.umich.edu/UMDL
  - DL Technologies (Stanford) www-diglib.stanford.edu
  - Alexandria (Santa Barbara) alexandria.sdc.ucsb.edu
  - ADEPT (Georgia) lsdis.cs.uga.edu/~adept
  - ERCIM DELOS DL (Europe) www.ercim.org/delos
  - UKOLN (U.K.) www.ukoln.ac.uk
  - CIDL (Canada) www.nlc-bnc.ca/cidl
  - Waikato (New Zealand) www.cs.waikato.ac.nz/cgi-bin/nzdlbeta/gw
- Main Focus on system infrastructure for global information sharing

Summary and Future Work

- Our contribution:
  - First declarative language for uniformly querying RDF schemas and resource descriptions
  - Formal model for RDF/S (+ typed, operational semantics)
- Functionality evaluation:
  - Developed in the context of the EC project C-Web (http://cweb.inria.fr)
  - Accepted for use in the ongoing EC projects:
    - MesMuses (http://aquarelle.inria.fr/mesmuses)
    - Ontoknowledge (http://www.ontoknowledge.org)
- Performance evaluation:
  - Testbed: Open Directory RDF dump
    - 321,988 Topics (~100M), 2,123,382 Sites (~700M)
  - Optimization opportunities:
    - Schema and Data Indexing Techniques (transitive closure queries)
    - Heuristics for Algebraic Transformations (schema and data queries)

35