Orchestration and composition of tools and services to support high level scientific workflows

Stelios Sfakianakis

ICS-FORTH

Bioinformatics

"the development and application of computational tools to *acquire*, *store*, *organize*, *archive*, *analyze*, and *visualize* biological data"

in silico Experiments

- Testing a hypothesis, performing a simulation, or answering a scientific question
  - Computer based information resources
  - Computational analysis
- Components
  - Objective
  - Plan
  - Methods
  - Results
  - …

Tools for Bioinformatics Analysis

- Information Repositories
  - Protein: PIR, SwissProt, PDB,…
  - Nucleotide: GenBank, DDBJ, EMBL
  - Other: Literature, Microarray Databases, …
- Services and Tools
  - BLAST, Clustal-W, InterProScan, Alibaba2, EMBOSS , InterProScan, …
- WWW-based interfaces

In-Silico Experiments

- Combination of different tools and data repositories

Problems

- Ad-hoc, user mediated and user driven integration
- Time consuming and mundane
- Much knowledge remains undocumented
  - Intermediate results
  - Failures
  - Whole process
- No easy…
  - Reuse
  - Sharing
  - Reproduction

How to support the bioinformatician?

- Make databases and tools accessible remotely
  - Through a computer
- Support "connectivity" of resources
  - In a computer assisted and automated way
- Allow the whole analysis/experiment to be stored as an *executable* artifact
  - And be annotated and discoverable
  - And be re-executed multiple times
  - And be shared, open to examination, modification, and reuse in whole or part
- Support the transcription of runs
  - user actions, final and intermediate results, lineage of data

Enabling Technologies

- Web Services
- Orchestration
- Grid
- Semantic Web

Web Services and Service Oriented Architecture

- Web Service
  - "a software system designed to support interoperable machine-to-machine interaction over a network" [W3C]
  - SOAP encoded XML Messages over HTTP(S), SMTP, …
  - Web Service Description Language (WSDL) used for description
- Service Oriented Architecture
  - A software architecture where functionality is provided in the form of *interoperable, loosely coupled services*
  - Usually Services == Web Services but needs not be the case in general
  - SOA promotes re-use and simplifies maintenance

Orchestration

- Recursive composition of services
  - Combination of different services
  - Publication of the combination as a new service
- One industry standard: WS-BPEL
- Multiple specifications: WSFL, XLANG, XPDL, Scufl, …
- Outcome: Workflows
  - reusable executable entities incorporating a series of processing activities to perform a process or answer a question

What BPEL does …

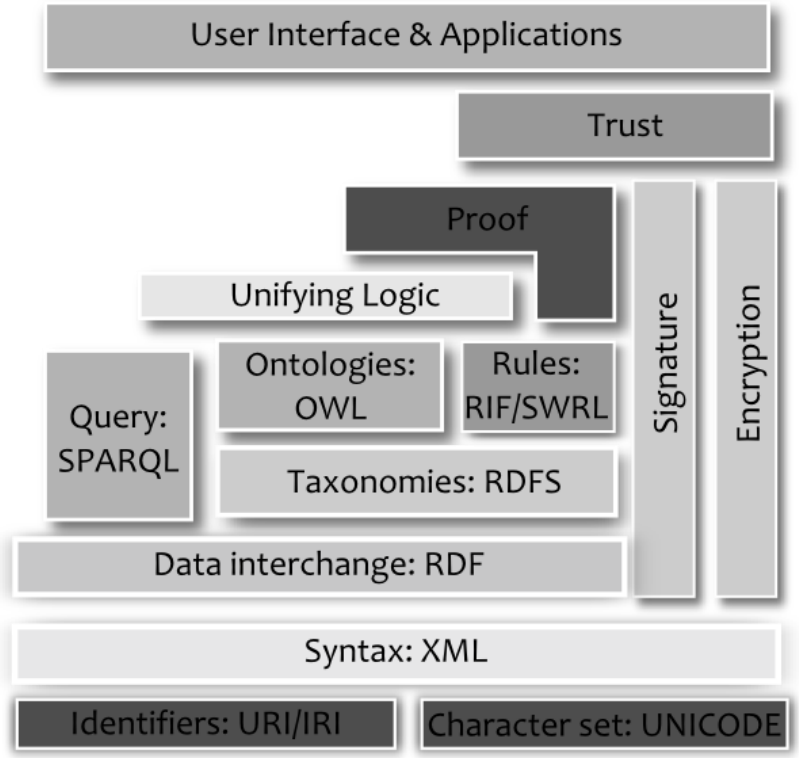
- BPEL deals with the functional aspects of business processes:
  - control flow (branch, loop, parallel)
  - asynchronous conversations and correlation
  - non-determinism (e.g. event triggering)
  - long-running, nested units of work, faults, and compensation
- An imperative programming language with XML syntax

The Grid

- A distributed computing environment that supports sharing
  - Computing power
  - Data storage
- In a unified framework
  - User management
  - Virtual organizations
- To provide a virtual supercomputer!

The Semantic Web

- Extension of the current Web
- Support Machine readable and processable descriptions of Things
- RDF abstract Model
  - Directed labeled graphs
  - Subject Predicate Object triples
- Formal Ontology specification by OWL and RDF Schema



Challenges

- Data-intensive tasks
- Computation-intensive tasks
- Security & Privacy
- Heterogeneity in tools and databases
- Semantics

Data integration

- Hypertext navigation
  - The user navigates from one data source to another and performs the necessary data transformations, mappings, etc.
- Data warehouse
  - One single data source which replicates the original ones in a common (joined) schema
- Federation
  - A mediator is accepts a query in a data source neutral language and transforms it to the local schemata of each data source

Service Integration

- Syntactic Interoperability
  - "Solved" by Web Service and Grid standards
- Semantic Interoperability
  - Metadata annotations
  - Ontologies

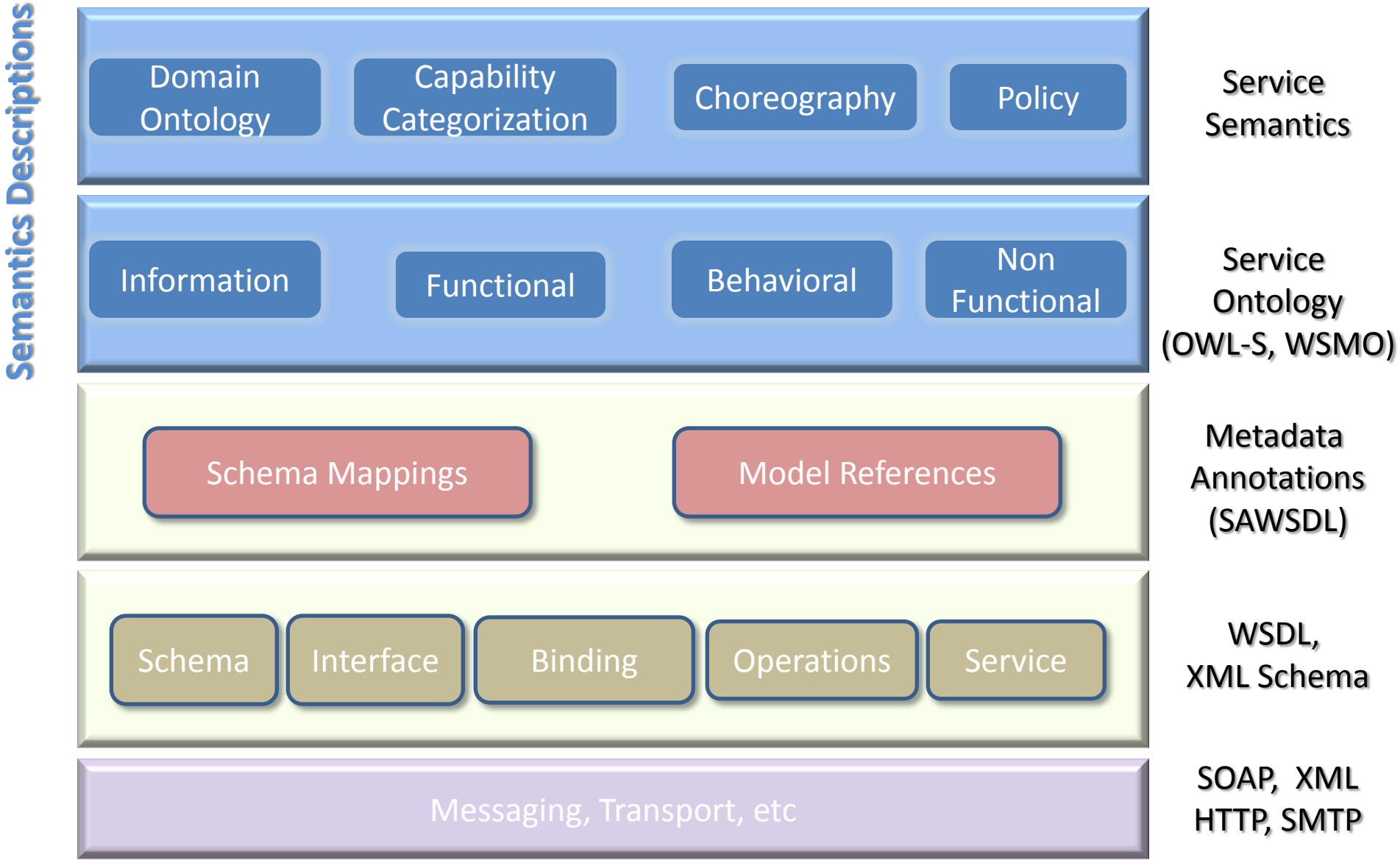
Semantics for Web Services

- Information Semantics
  - Information Model for exchanged data
  - Mediation
- Functional Semantics
  - Service Capability
  - Functionality
- Non-Functional Semantics
  - Policies
- Behavioral Semantics
  - External behavior, the protocol should a client comply with
  - Internal behavior, how the service is implemented

Semantic Composition of Services

- Service Annotation
  - Metadata descriptions of their Informational, Functional, Non-Functional, and Behavioral Semantics
- Service Publication
  - Metadata Repositories and Registries
- Service Discovery and Invocation
  - Matchmaking, Ranking, Selection,…
- Additional aspects: service recommendation, provenance & logging, etc

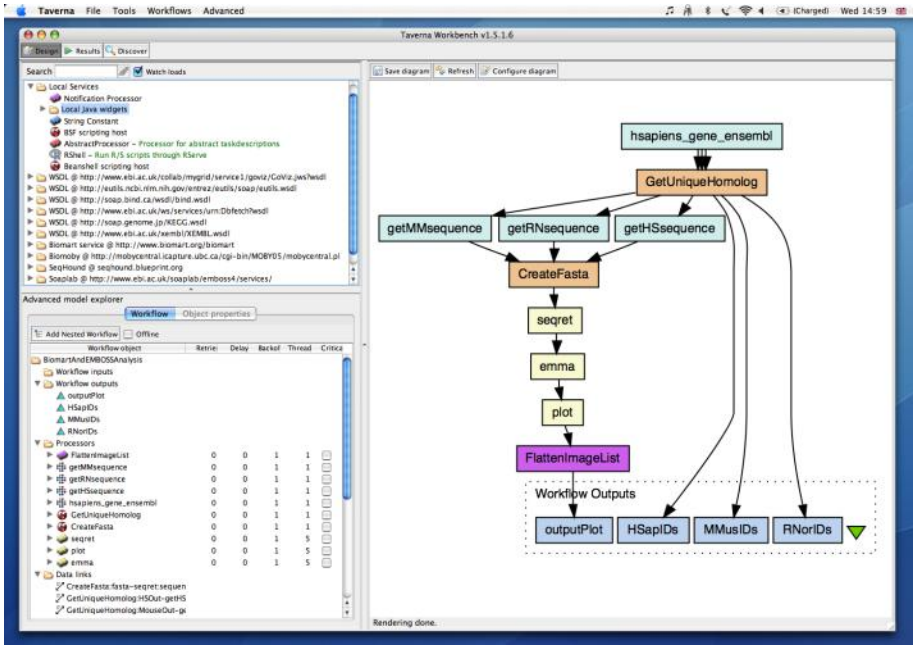
Semantic Web Services stack



Scientific Workflow Environments

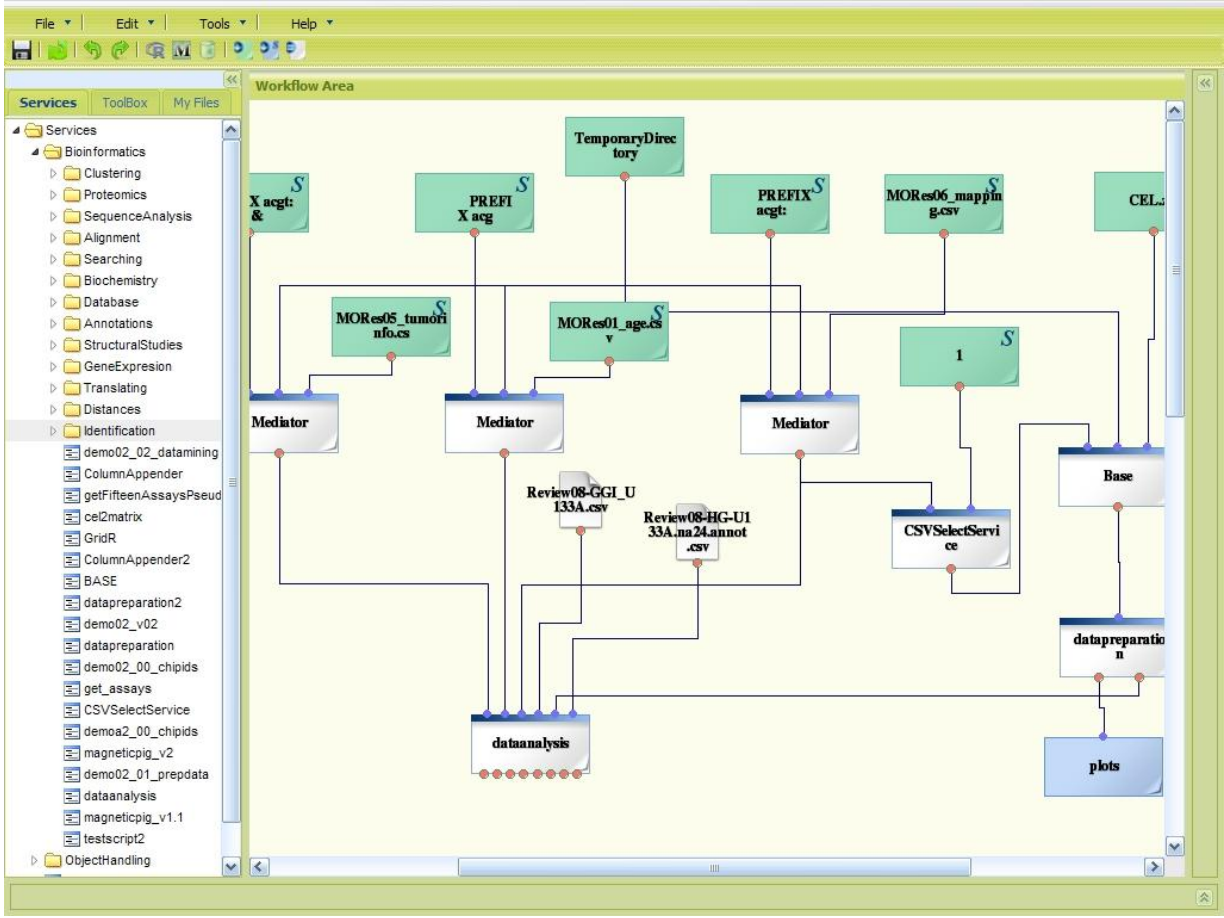
Taverna Workbench

- Desktop authoring environment and enactment engine
- Scufl custom made workflow languageIn addition to Web Services it supports SoapLab, BioMoby, and many other kind of services
- De facto standard for bioinformatics workflows



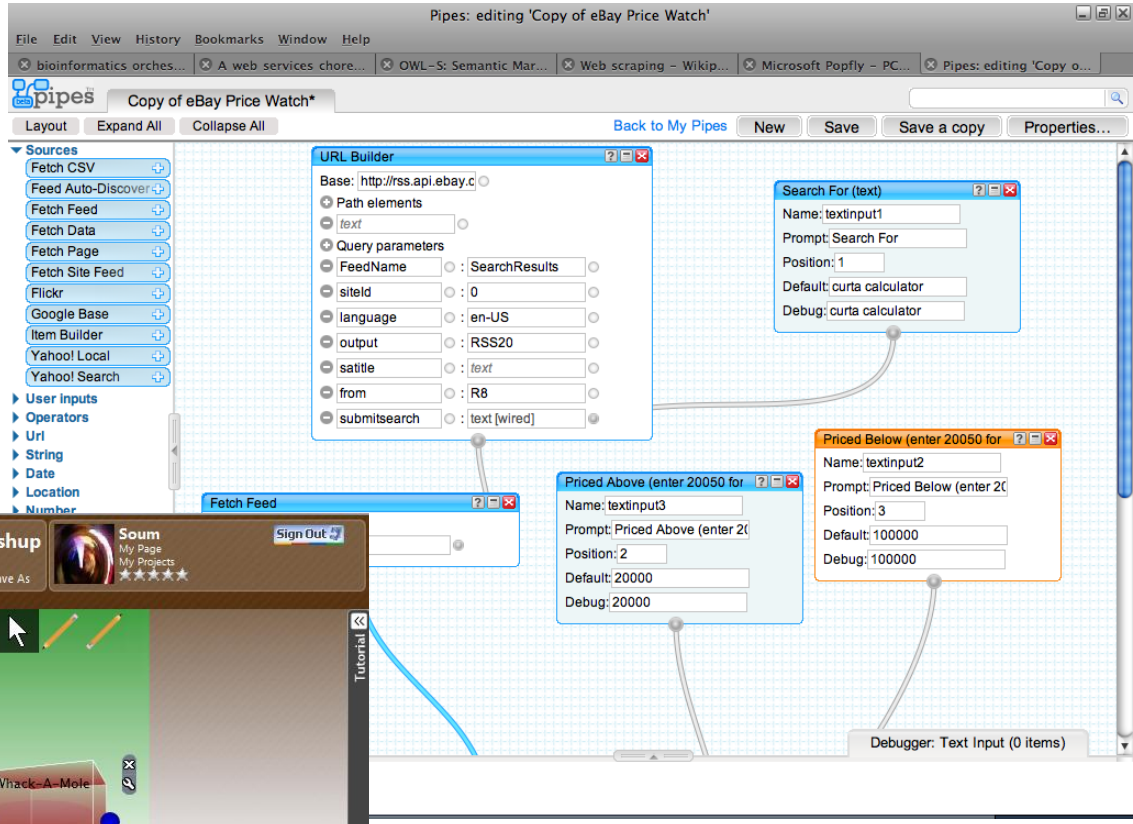
ACGT Workflow Editing and Enactment Environment

- Web Based
  - No installation, automatic update
  - Server side: persistent storage, heavy computations
  - Client side: UI, session state
- Conforms to the ACGT MO and Metadata Repositories



Composition tools for Web 2.0

- Yahoo! Pipes
- MS Popfly



Conclusions

- Bioinformatics tools abound but more discipline must be enforced
- IT offers an interesting array of technologies
  - Grid, Web Services, Semantic Web are complementary to each other
  - Evolving and converging
- Ontologies and Metadata are very important
  - Anything that is not described does not exist

Ευχαριστώ!