

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

The image displays a collage of various bioinformatics web interfaces and tools, illustrating the combination of different tools and data repositories for in-silico experiments. The tools shown include:

- BLAST**: National Center for Biotechnology Information's Basic Local Alignment Search Tool.
- TWINSKAN**: The New GENSCAN Web Server at MIT for identification of complete gene structures in genomic DNA.
- SignalP 3.0 Server**: A web server for predicting signal peptides.
- WWS (WWW Signal Scan)**: A web server for finding and identifying published signal sequences with the signal DNA sequence.
- RepeatMasker Web Server**: A web server for identifying and masking repetitive elements in genomic DNA.
- EMBL-EBI**: European Bioinformatics Institute.
- NCBI**: National Center for Biotechnology Information.
- AGENT**: A web server for finding and identifying published signal sequences with the signal DNA sequence.
- SUMOPit**: A web server for identifying and masking repetitive elements in genomic DNA.
- National Center for Biotechnology Information**: A web server for finding and identifying published signal sequences with the signal DNA sequence.
- EMBL-EBI**: European Bioinformatics Institute.
- EMBL-EBI**: European Bioinformatics Institute.
- EMBL-EBI**: European Bioinformatics Institute.

Red arrows indicate the flow of data and the integration of these tools. A central DNA sequence is shown:

```
aattctac caacagtga tgaggtgtt ggtctgtt  
aaattgggttt 12241 cagtcttta aatttaac  
gaag agtcatacag tcaatagcct ttttagct 12301  
ccta atagatacag agtgggtct cactgtatt ttaattga  
tctt 12361 gactattat gttgagttg ttacattta  
ttca ttagagaag tctaatatt 12421 tagtgact  
ctgtttt tttttatg gatcttaatt tttttaa tttgattg  
tggagcatt  
tggagc
```

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, ScufI, ...
- 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



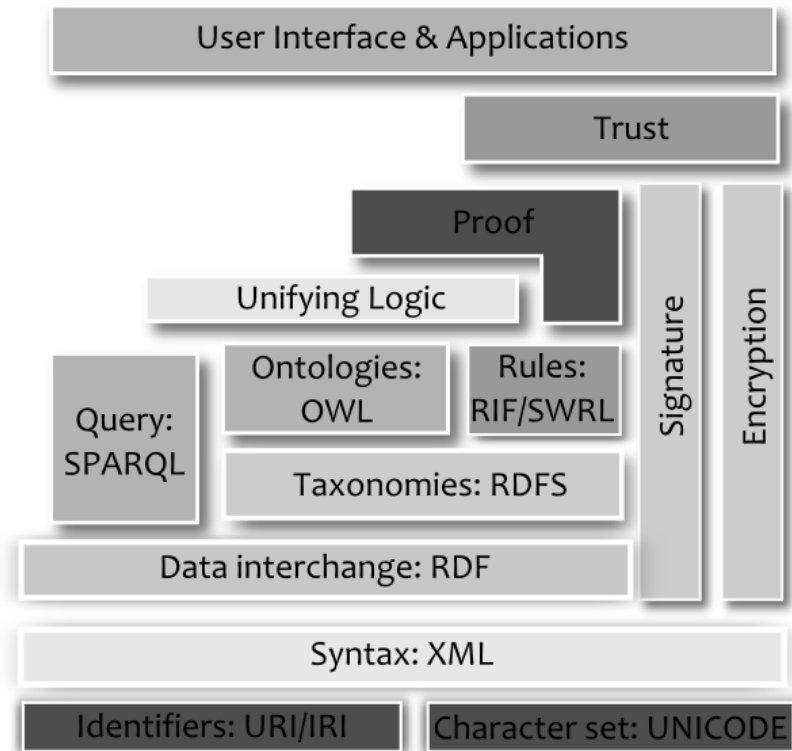
The Grid

- 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



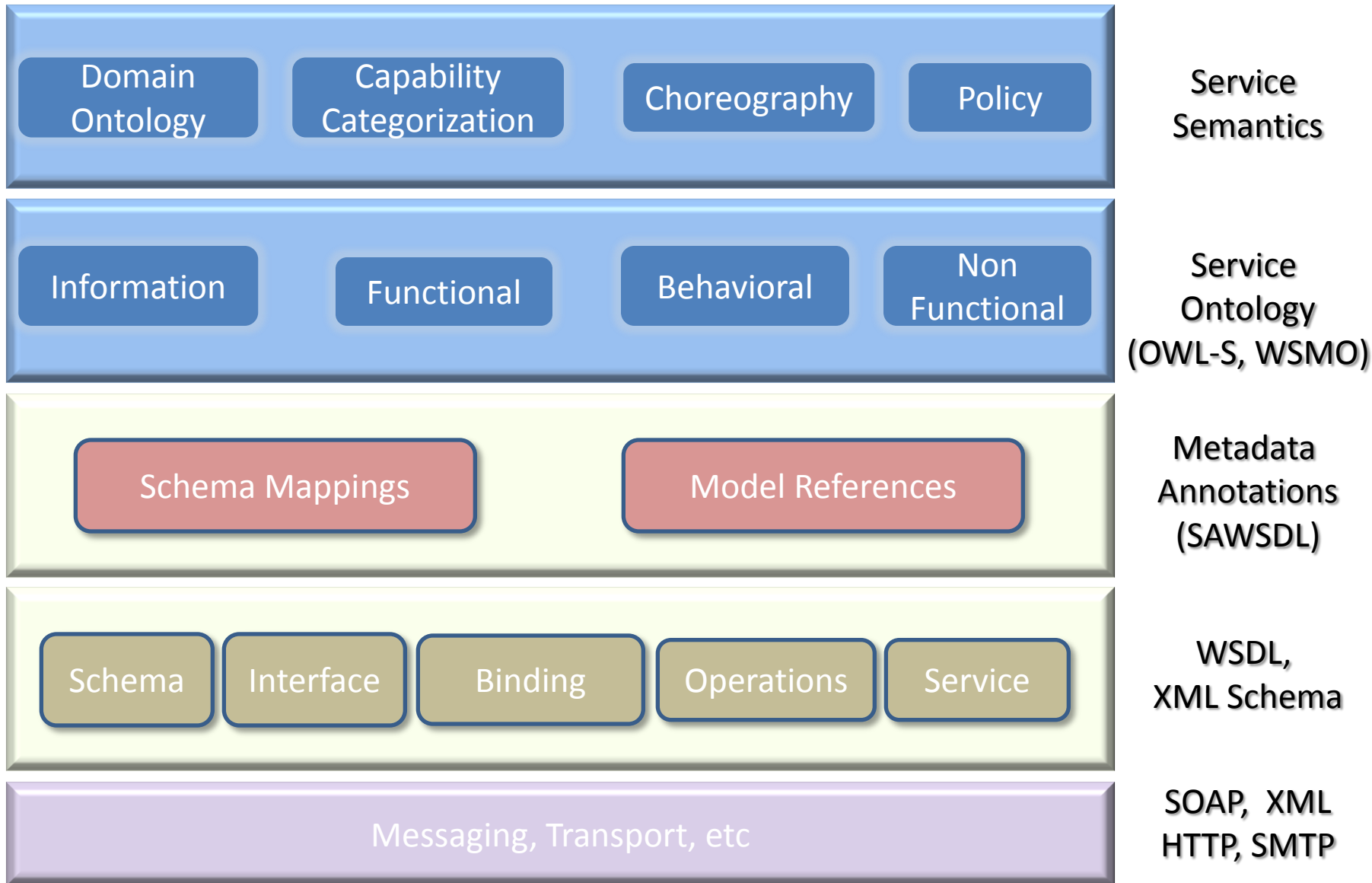
Semantic Composition of Services

- 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

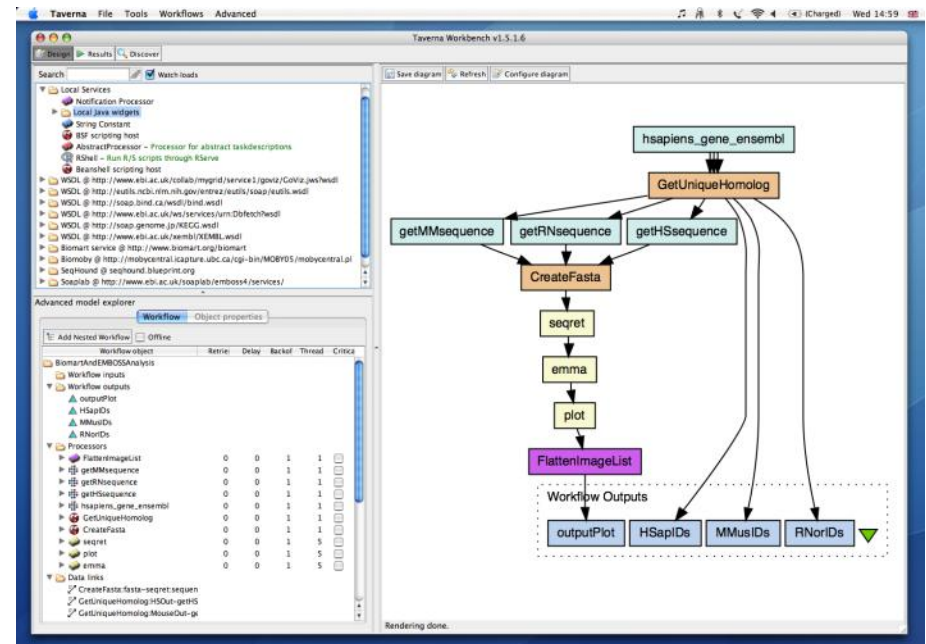
Semantics Descriptions



Scientific Workflow Environments

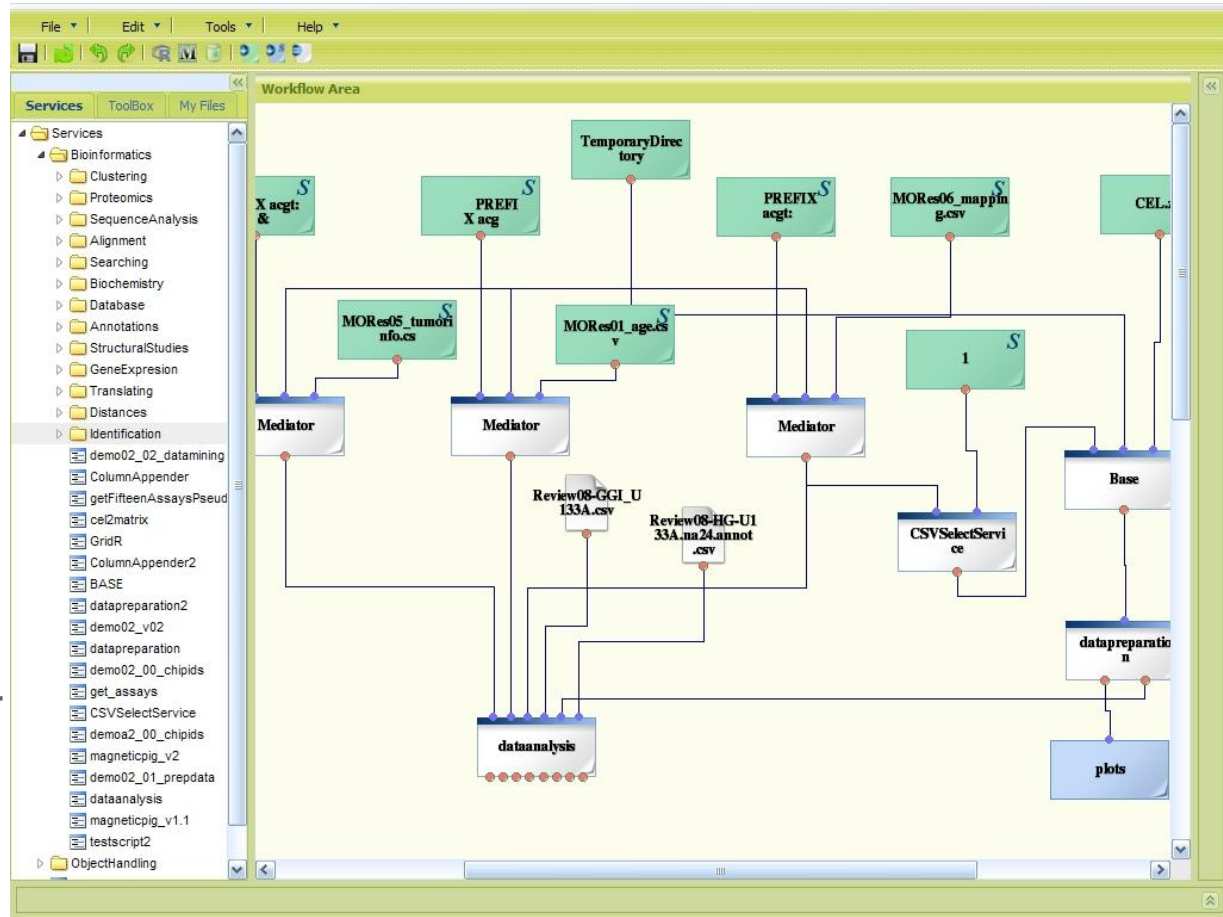
Taverna Workbench

- Desktop authoring environment and enactment engine
- Scufi custom made workflow language
In 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

Pipes: editing 'Copy of eBay Price Watch'

File Edit View History Bookmarks Window Help

bioinformatics orches... A web services chore... OWL-S: Semantic Mar... Web scraping - Wikip... Microsoft Popfly - PC... Pipes: editing 'Copy o...

pipes Copy of eBay Price Watch*

Layout Expand All Collapse All Back to My Pipes New Save Save a copy Properties...

Sources

- Fetch CSV
- Feed Auto-Discover
- Fetch Feed
- Fetch Data
- Fetch Page
- Fetch Site Feed
- Flickr
- Google Base
- Item Builder
- Yahoo! Local
- Yahoo! Search

User inputs

- Operators
- Url
- String
- Date
- Location
- Number

URL Builder

Base:

Path elements

- text

Query parameters

- FeedName: SearchResults
- siteld: 0
- language: en-US
- output: RSS20
- satitle: text
- from: R8
- submitsearch: text [wired]

Search For (text)

Name:

Prompt: Search For

Position: 1

Default: curta calculator

Debug: curta calculator

Fetch Feed

Named Above (enter 20050 for

Name:

Prompt: Priced Above (enter 20

Position: 2

Default: 20000

Debug: 20000

Priced Below (enter 20050 for

Name:

Prompt: Priced Below (enter 20

Position: 3

Default: 100000

Debug: 100000

Debugger: Text Input (0 items)

Microsoft Popfly Alpha

Logo goes here!

Find Users | Forums | Help

Unsaved Mashup

Save Save As

Sign Out

Soum My Page My Projects

Blocks

184 matches found

- Video Player
- Virtual Earth
- Whack-A-Mole
- Yahoo! Answers
- Yahoo! Images
- Yahoo! News
- Yahoo! Term Extraction
- Yahoo! Traffic
- Yahoo! Video
- Xbox Live

Blocks Shared By Users

- GeoNamesFR
- WoW EU Data
- My Live Contacts IE

Preview Clear Add Custom HTML

Mode:

Digg flickr Whack-A-Mole

You Have a Key

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!

