YeastMed: An XML-Based System for Biological Data Integration of Yeast

Abdelaali Briache, Kamar Marrakchi, Amine Kerzazi, Ismael Navas-Delgado, Jose F Aldana Montes, Badr D. Rossi Hassani and Khalid Lairini

***LABIPHABE*, Department of Biology, F. S. T. Of Tangier , Morocco.**

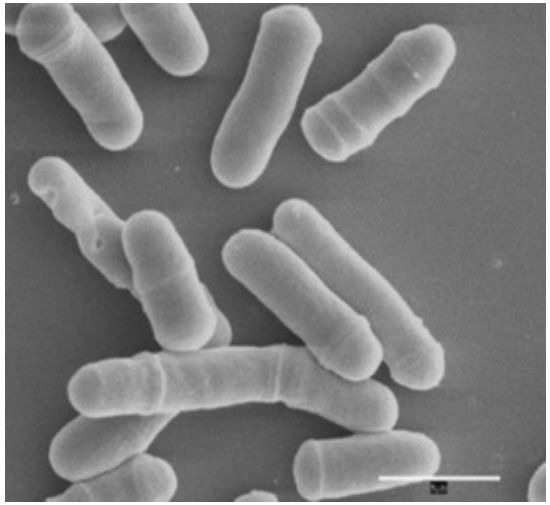
***KHAOS*, Departamento de Lenguajes y Ciencias de la Computación, Universidad de Málaga, Spain.**

Index

Introduction: Yeasts

- Tiny forms of fungi;
- Microorganisms visible only with a microscope;
- Cell cycle similar to that of Humans,
- rapid growth,
- dispersed cells,
- a well-defined genetic system,
- genome can be easily manipulated,
- ……

filament-shaped Candida albicans



Elongated schizosaccharomyces pombe

Egg-shaped Saccharomyces cerevisae

Introduction: Yeasts

Yeast-Specialized sources

AGD	*Ashbya gossypii* genome database
CandidaDB	*Candida albicans* genome database
Candida Genome	*Candida albicans* genome database
CYGD	MIPS Comprehensive yeast genome database
Génolevures	A comparison of *S.cerevisiae* and 14 other yeast species
PROPHECY	Profiling of phenotypic characteristics in yeast
SCMD	*Saccharomyces cerevisiae* morphological database: micrographs of budding yeast mutants
SCPD	*Saccharomyces cerevisiae* promoter database
SGD	*Saccharomyces* genome database
TRIPLES	Transposon-insertion phenotypes, localization and expression in *Saccharomyces*
YDPM	Yeast deletion project and mitochondria database
Yeast Intron Database	Ares lab database of splicesomal introns in *S.cerevisiae*
Yeast snoRNA Database	Yeast small nucleolar RNAs
yMGV	Yeast microarray global viewer
YRC PDR	Yeast resource center public data repository

General Sources

NCBI Protein database	All protein sequences: translated from GenBank and imported from other protein databases
Swiss-Prot	Now UniProt/Swiss-Prot: expertly curated protein sequence database, section of the UniProt knowledgebase
UniProt	Universal protein knowledgebase: merged data from Swiss-Prot, TrEMBL and PIR protein sequence databases
PROSITE	Biologically significant protein patterns and profiles
PRINTS	Hierarchical gene family fingerprints
Pfam	Protein families: multiple sequence alignments and profile hidden Markov
DDBJ—DNA Data Bank of Japan	All known nucleotide and protein sequences
EMBL Nucleotide Sequence Database	All known nucleotide and protein sequences

Introduction: Data Integration

Different data models and schemas,

Different model constructs that can be used to describe the same object, even if the same model is used,

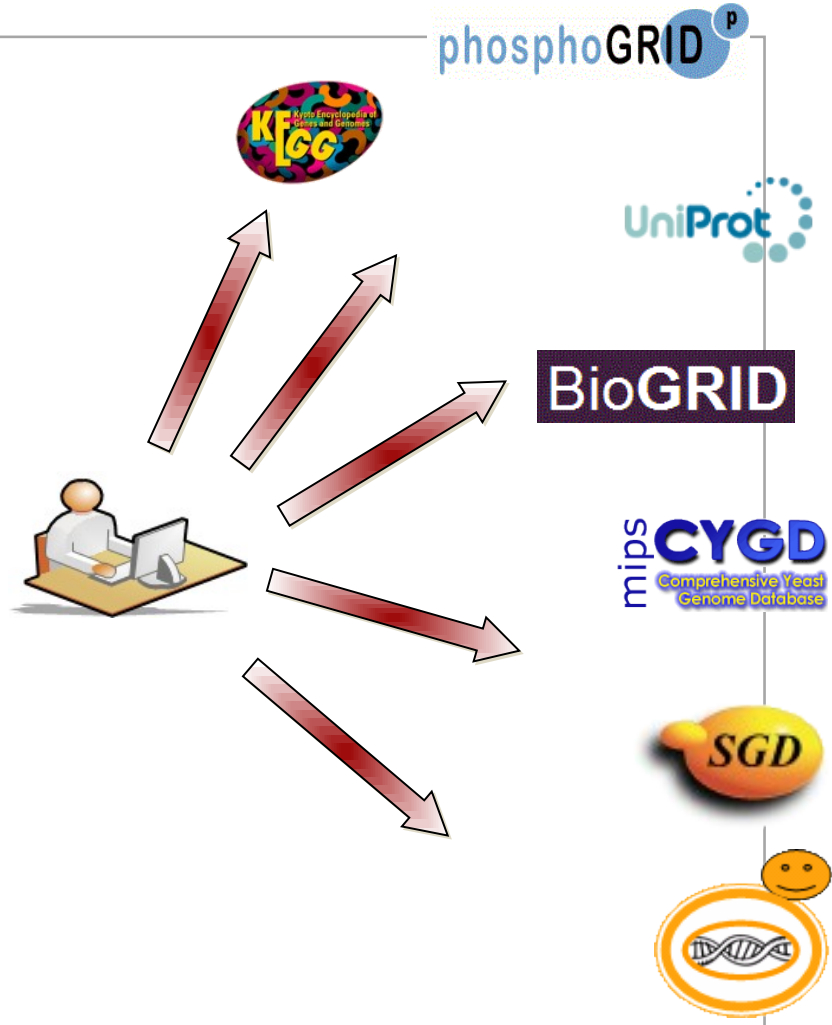
wide variety of formats that have been used for data representation : ASN.1 *(Abstract Syntax Notation One)*, XML, HTML...etc.,

Data sources make their data available in different ways,

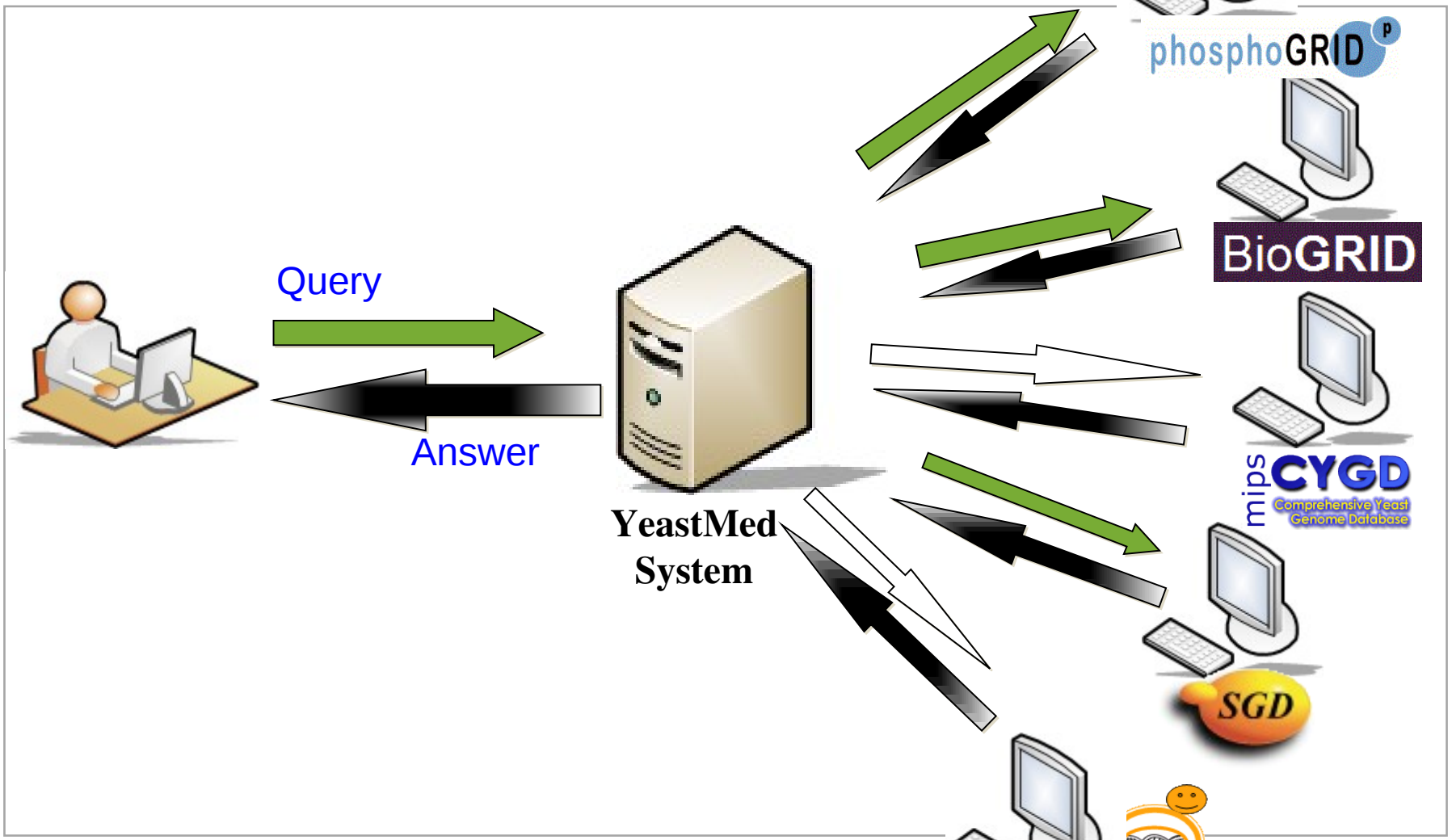
The query language used to interrogate data sources,

inconsistent use of nomenclature,

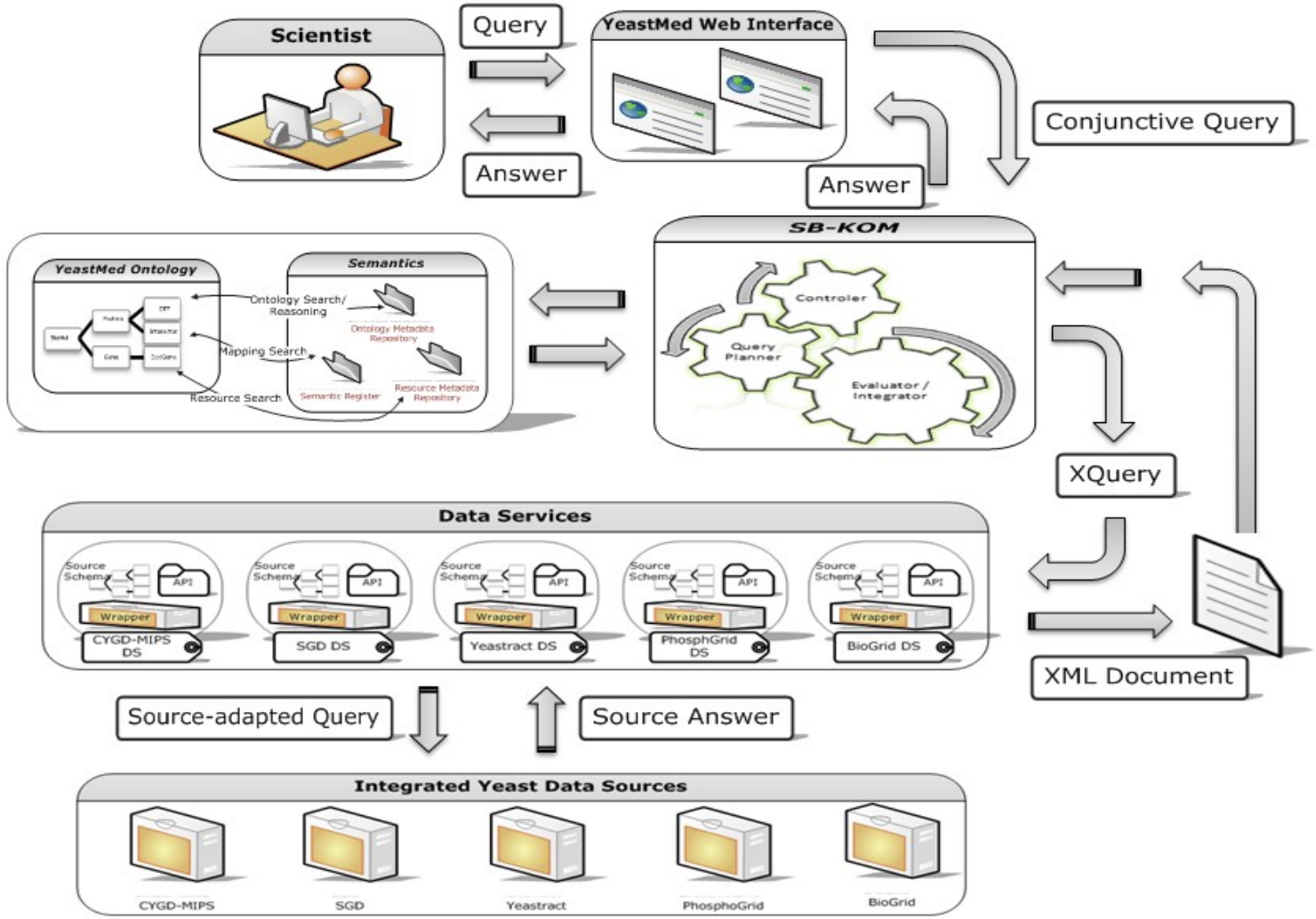
...



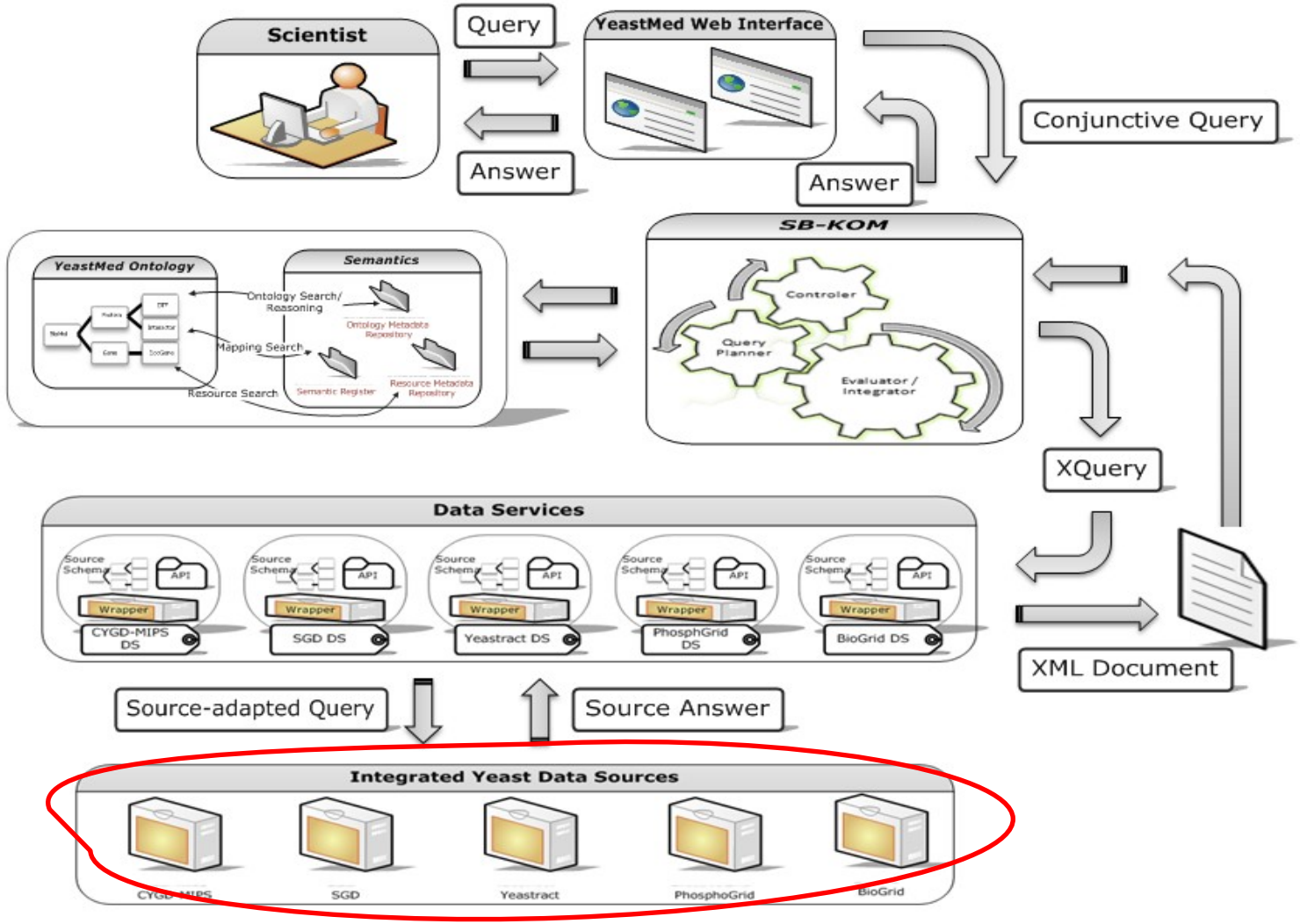
Introduction: Data Integration



YeastMed :Architecture



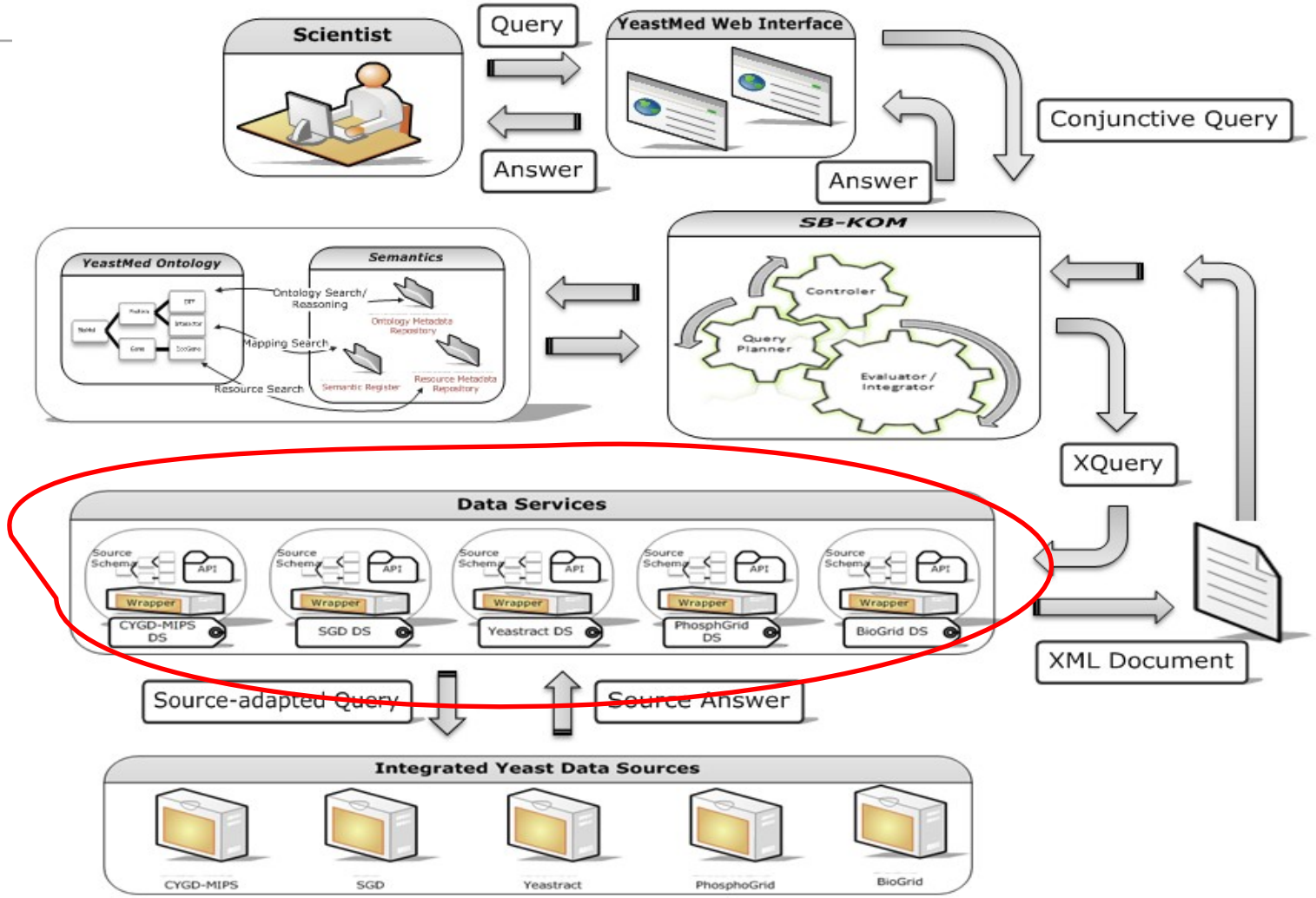
YeastMed :Architecture



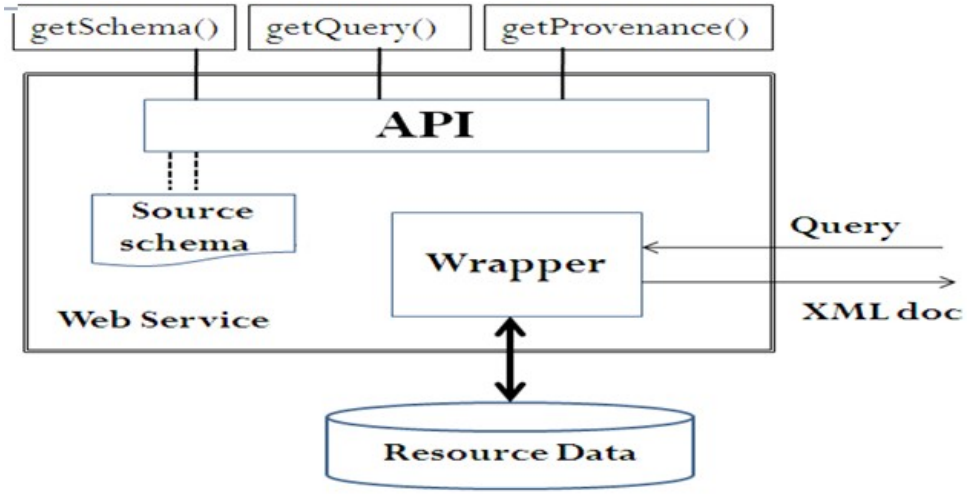
YeastMed: Data Sources

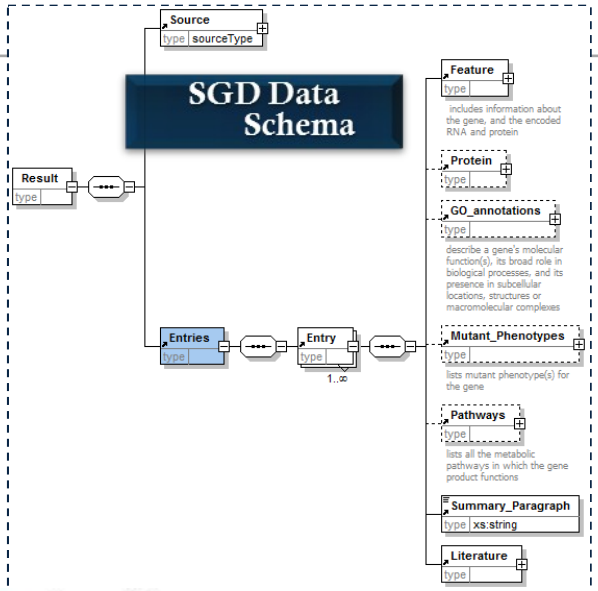
- **SGD**: Collection of genetic and molecular biological information about Sc.
- **MIPS-CYGD**: Information on the molecular structure and functional network of Sc.
- **Yeastract**: A repository of regulatory associations between transcription factors and target genes, based on experimental evidence.
- **PhosphoGrid** records the positions of specific phosphorylated residues on gene products.
- **BioGrid**: An online interaction repository with data compiled through comprehensive curation efforts.

YeastMed: Architecture



YeastMed: Data Services





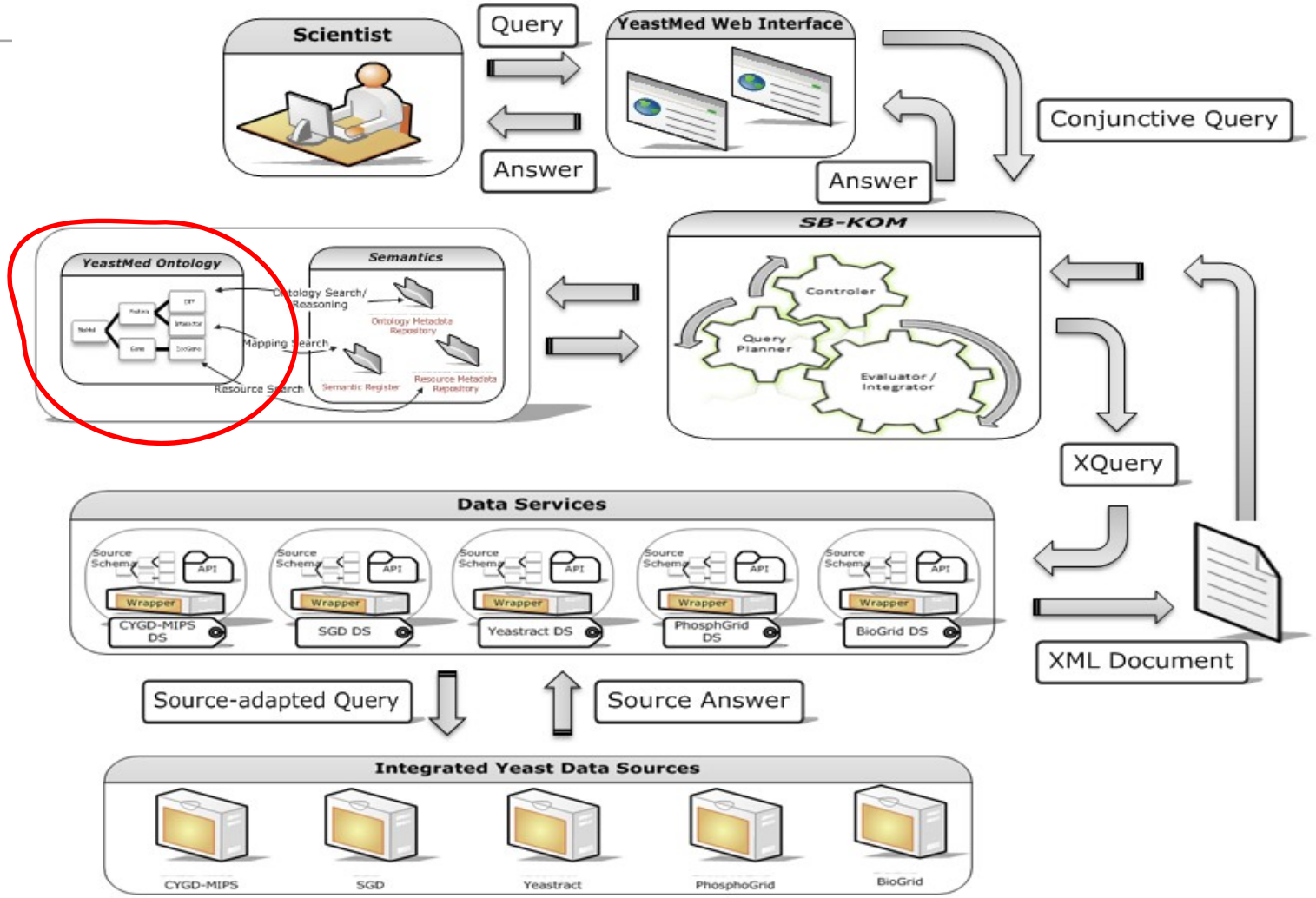
Operations

Name	
getSchema	--
getDataProvenance	--
Query	--

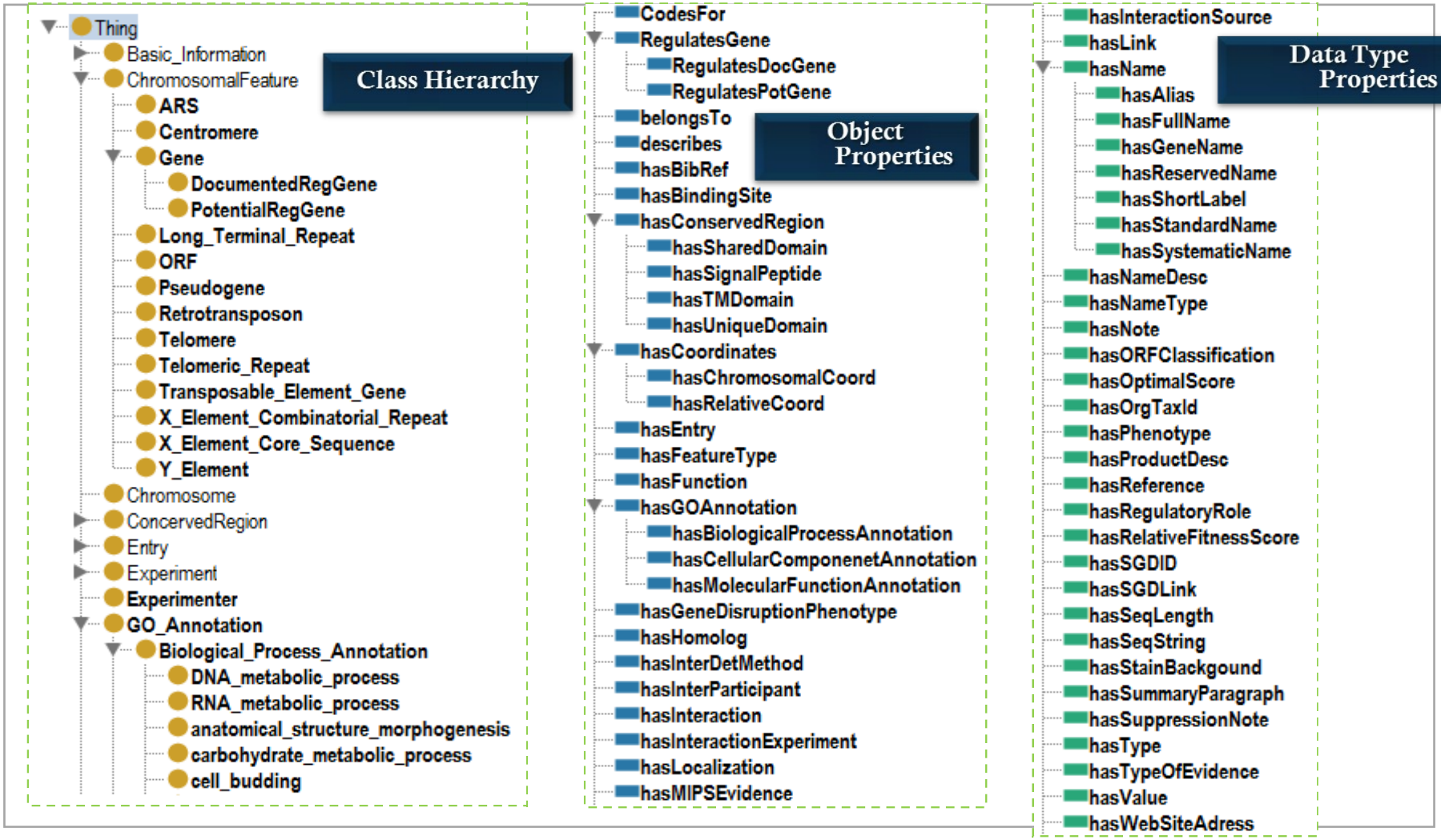
Endpoints Add Remove

☐	
☐	http://172.16.51.3:8080/SGD/services/SGD

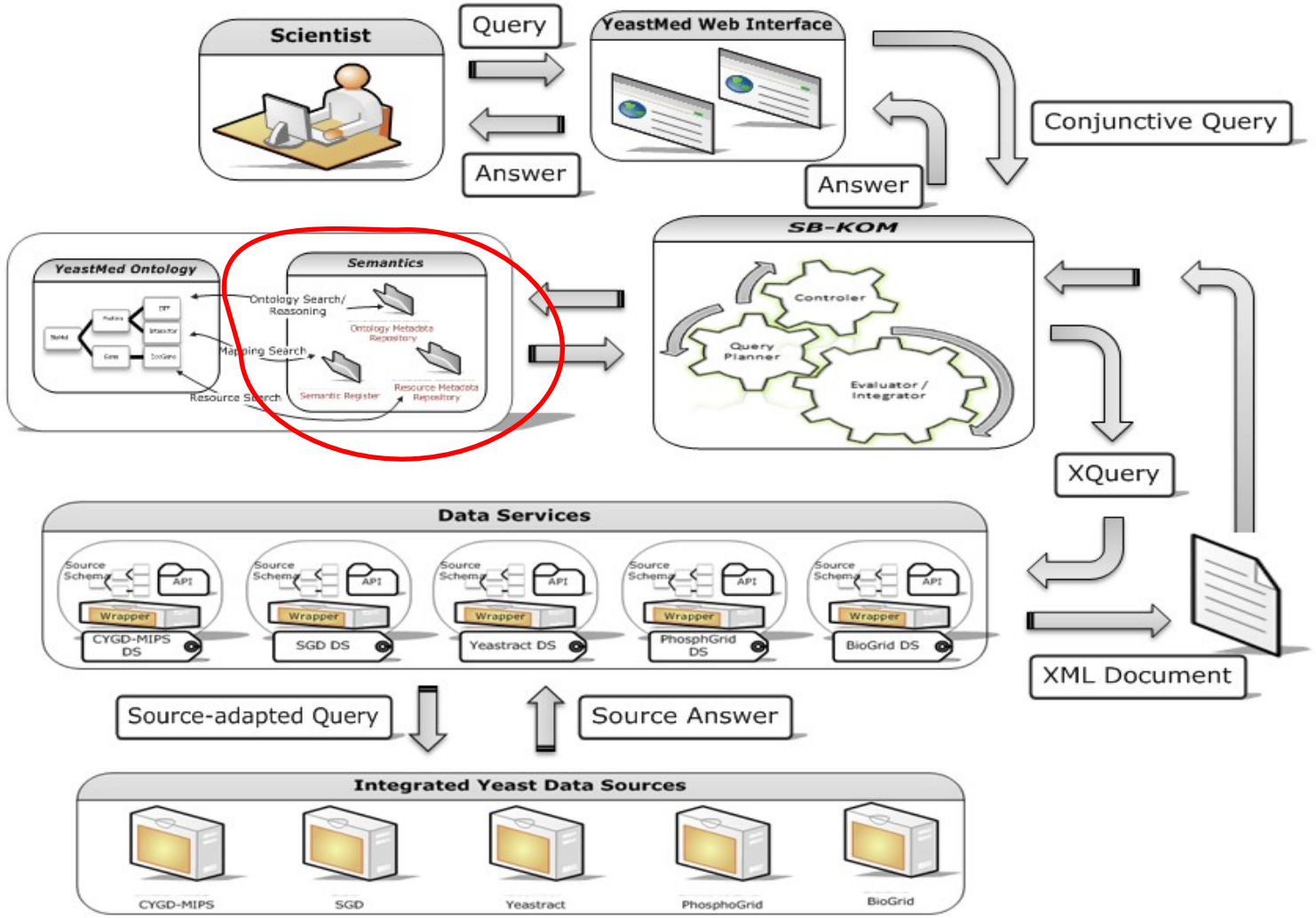
YeastMed: Architecture



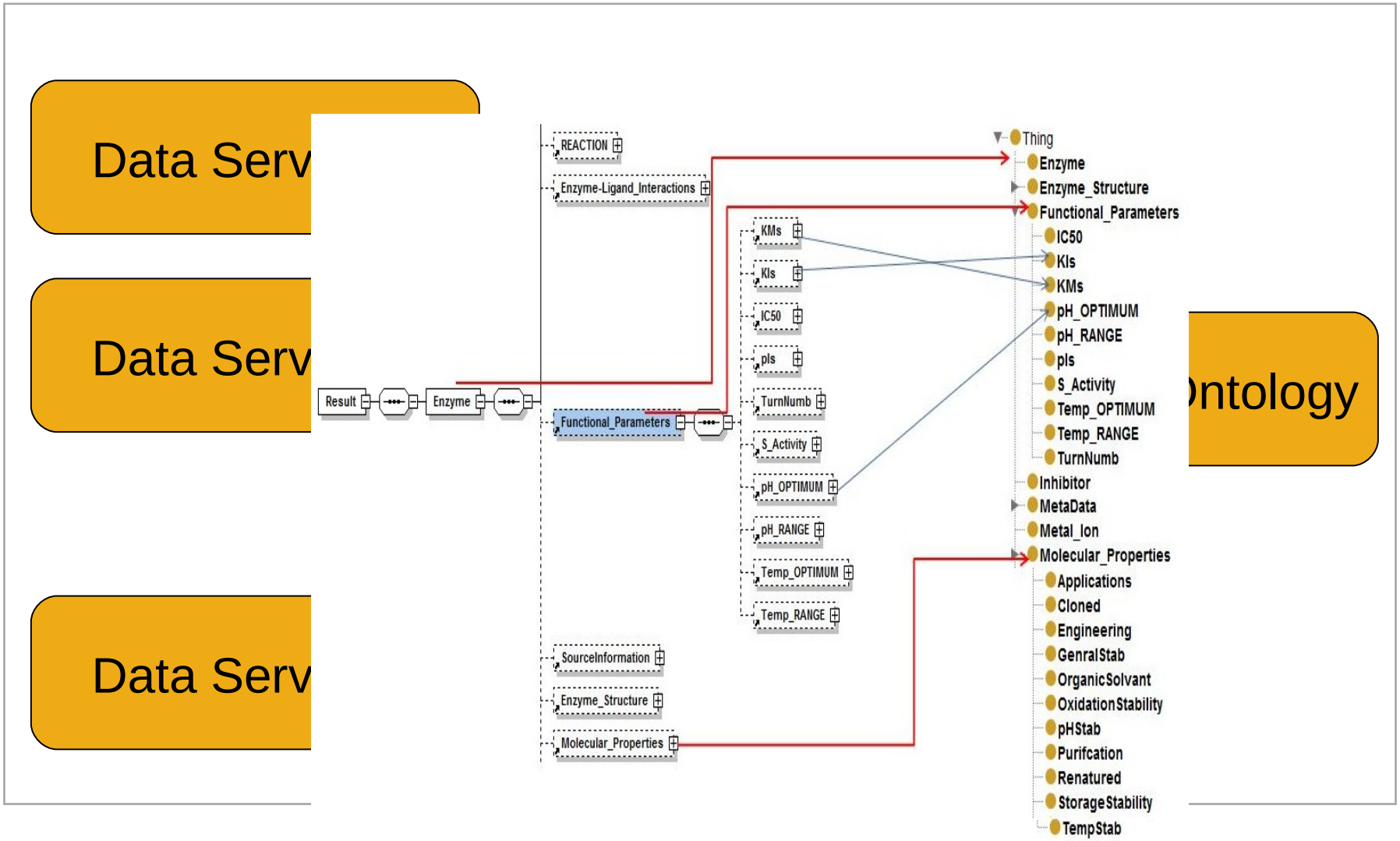
YeastMed: Ontology



YeastMed: Architecture



YeastMed: Mappings



YeastMed: Mappings

- ***Class Mapping:*** *it maps an Ontology class to the source schema.*

  XPath-Element-Location, Ontology-Class-Name, Correspondence-index

- ***Datatype Property Mapping:*** *it maps an Ontology datatype property to the source schema.*

  Example: Result/Entries/Entry/Protein, Protein,100

  XPath-Domain-Location; XPath-value-Location, Ontology-Domain-Name; Property-Name, correspondence-index

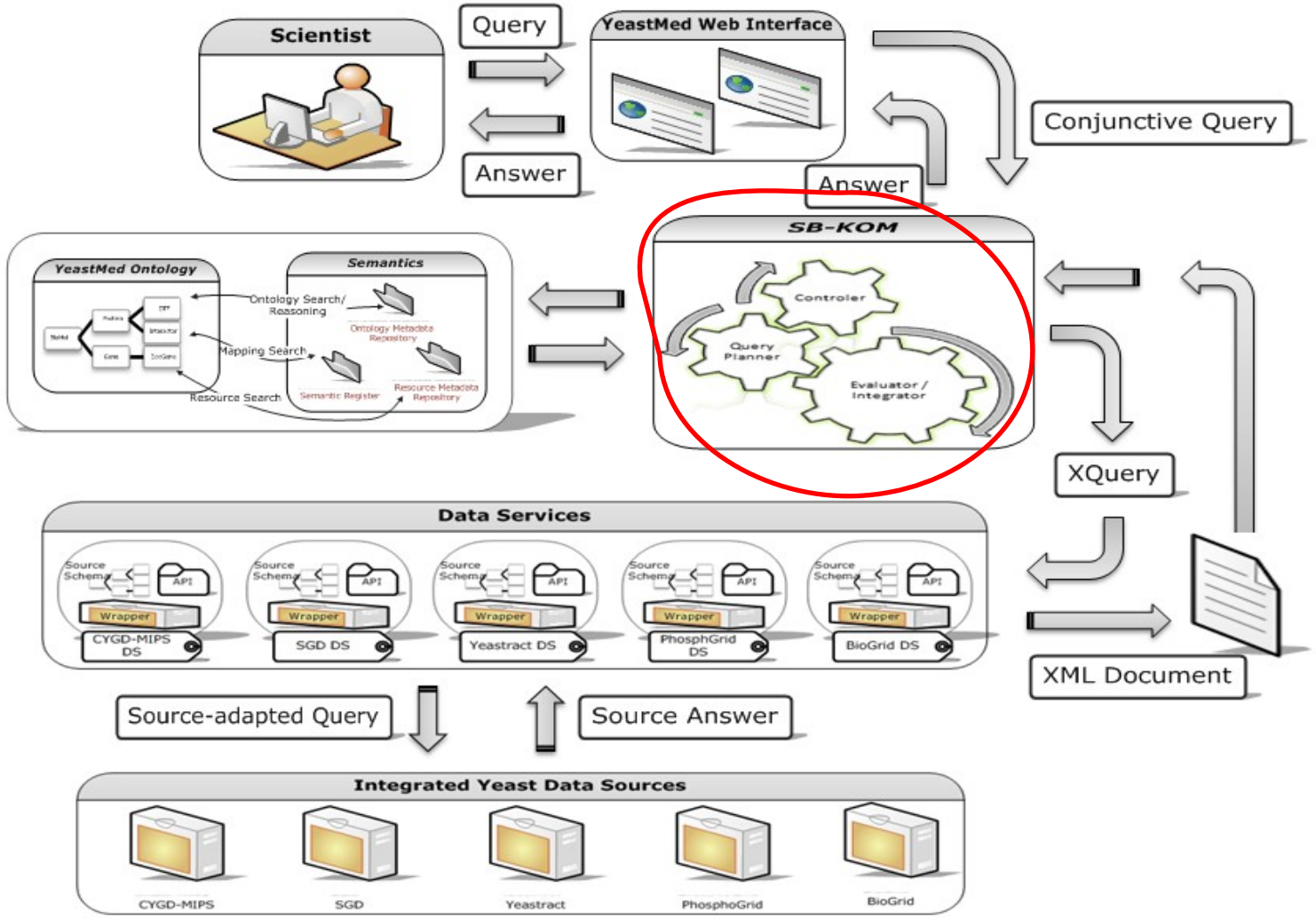
- ***Object Property Mapping:*** *it maps an Ontology object property to the source schema.*

  XPath-Domain-Location;XPath-Range-Location, Ontology-Domain-Name; Ontology-Range-Name; Property-Name,correspondence-index

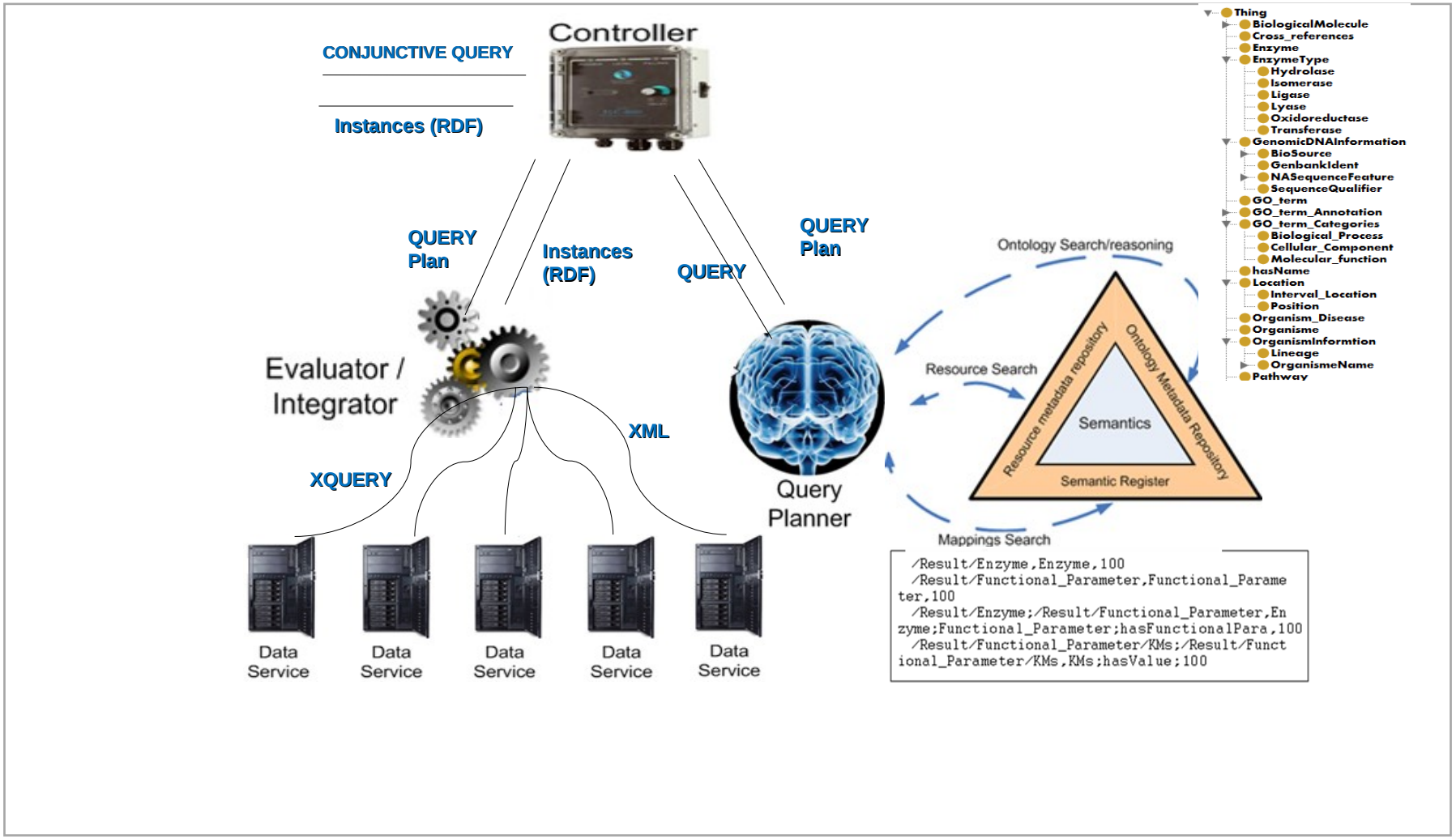
  Example: Result/Entries/Entry/Protein;

  Example: Result/Entries/Entry/Protein/SysName,TranscriptionFactor;hasName,100
  Result/Entries/Entry/Protein;Result/Entries/Entry/Literature,Protein;BibRef;hasBibRef,100

YeastMed: Architecture

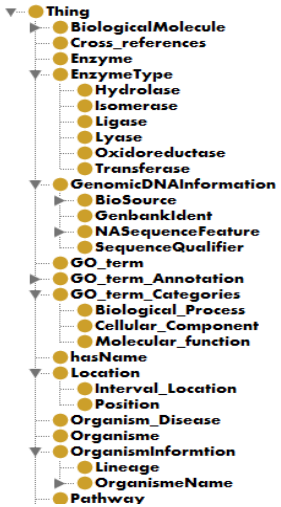


YeastMed: SB-KOM



YeastMed: SB-KOM

"find all the bibliographic References of topoisomerase III and also all the information about the phosphorylation sites that contain the transcription factors of DNA Topisomerase III, and specially the one (or ones if exist) whose gene is located on the Chromosome XVI."



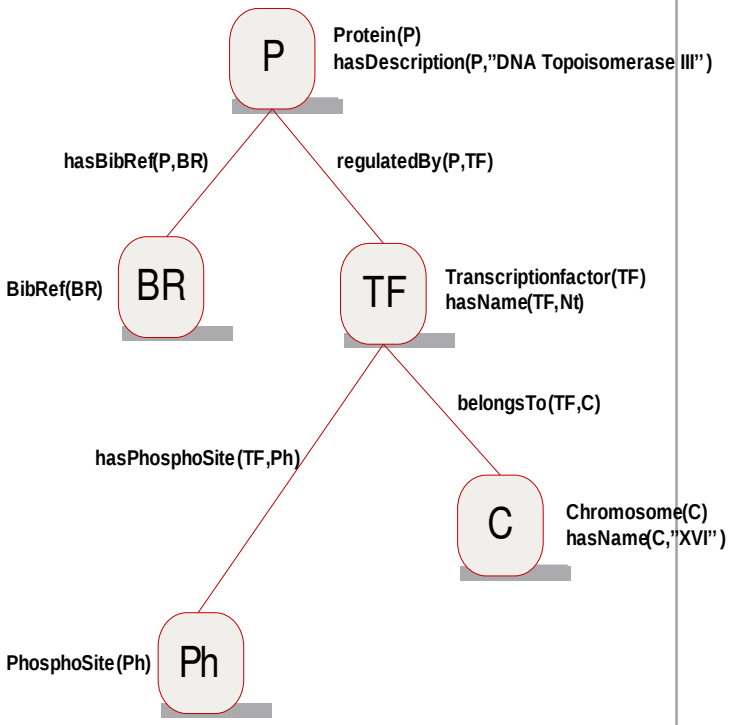
- ***Classes*** : *Protein*, *BibRef*, *TranscriptionFactor*, *Chromosome* and *PhosphoSite.*
- ***Datatype properties:*** *hasDescription, hasSystematicName and hasName*
- ***Object properties:*** *hasBibRef, regulatedBy, belongsTo and hasPhosphoSite*

Ans(BR,Ph) := Protein(P), hasDescription(P,"DNA Topoisomerase III"), BibRef(BR), hasBibRef(P,BR), hasSystematicName(P,SN), regulatedBy (P,TF), hasName(TF,Nt), TranscriptionFactor(TF), Chromosome(C), hasName(C,"XVI"), BelongsTo(TF,C), PhosphoSite(Ph), hasPhosphoSite(TF,Ph);

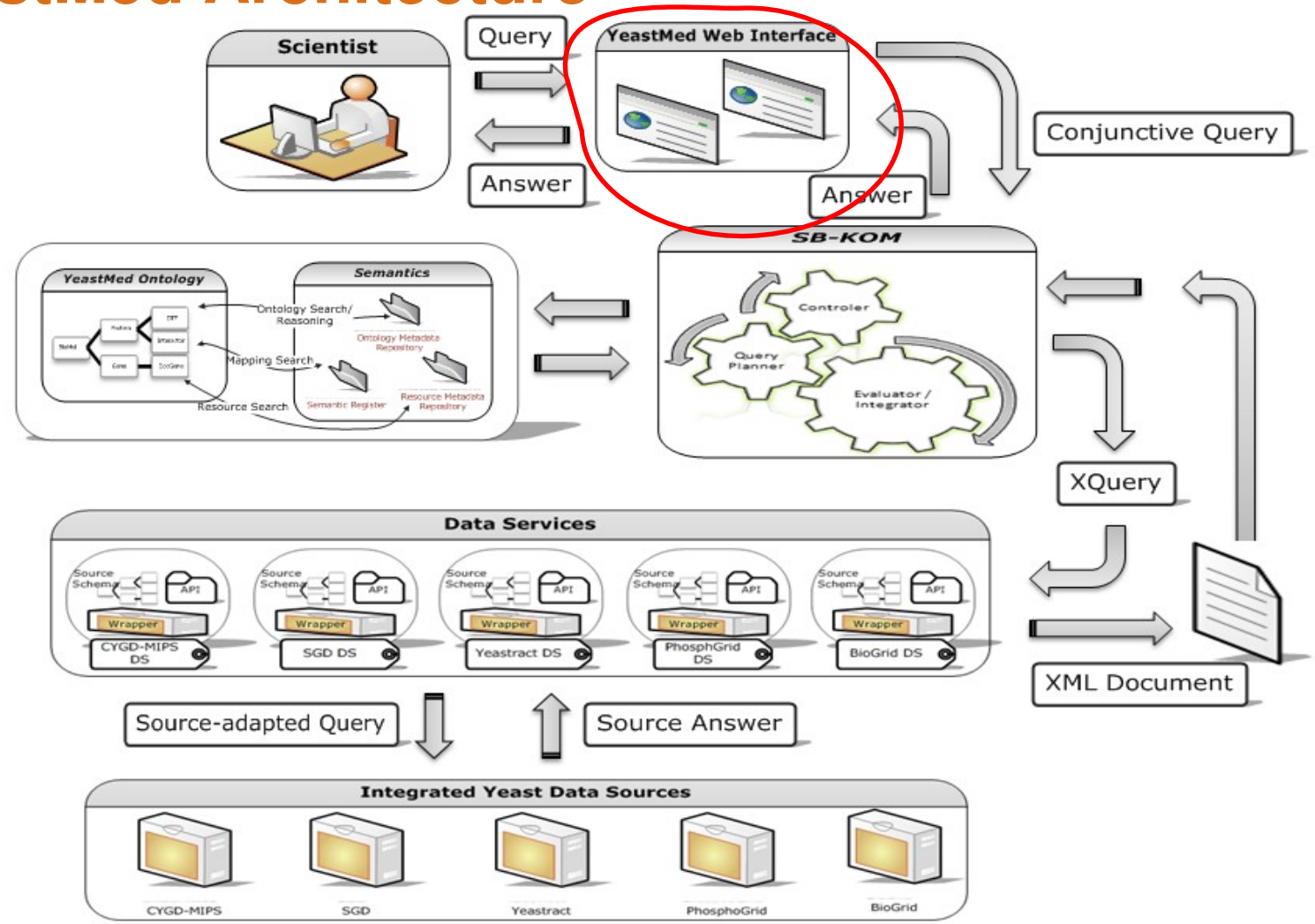
YeastMed: SB-KOM

Involved Data Sources : SGD, Yeastract and PhosphoGrid

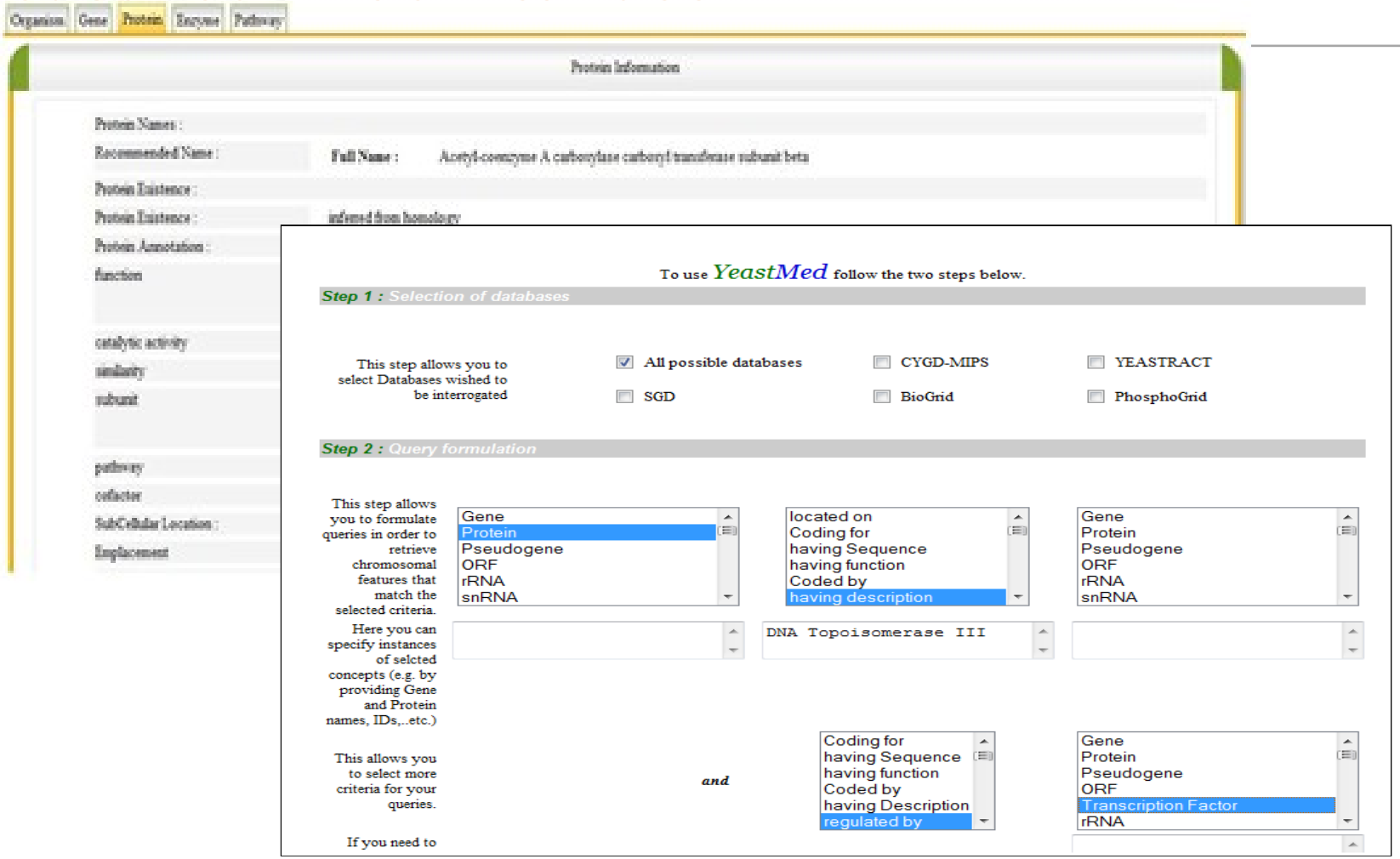
Group	Query	Mapping source
G1	Protein(P), hasBibRef(P,BR)	SGD
G2	Protein(P),hasDescription(P,"DNA Topoisomerase III")	SGD
G3	Protein(P), hasSystematicName(P, SN)	Yeastract
G4	Protein(P), RegulatedBy(P, TF)	Yeastract
G5	TranscriptionFactor(TF), hasName(TF, Nt)	Yeastract
G6	TranscriptionFactor(TF), belongsTo(TF,C)	Yeastract
G7	TranscriptionFactor(TF), hasPhosphorylationSite(TF, Ph)	PhosphoGrid
G8	Chromosom(C), hasName(C,"XVI")	Yeastract
G9	regulatedBy(P,TF)	Yeastract
G10	hasBibRef(P,BR)	SGD
G11	belongsTo(TF,C)	Yeastract
G12	hasPhosphoSite(TF,Ph)	PhosphoGrid
G13	Protein(P)	SGD Yeastract PhosphoGrid
G14	TranscriptionFactor(TF)	Yeastract PhosphoGrid
G15	BibRef(BR)	SGD
G16	Chromosome(C)	Yeastract
G17	PhosphoSite(Ph)	PhosphoGrid



YeastMed Architecture



YeastMed: Web Interface



Conclusions

- Allowing a transparent and simultaneous access to several autonomous and heterogeneous Yeast sources;
- Helping biologists to find convenient data to interpret results of their experiments;
- Avoiding biologists to confront technical and structural problems in data retrieving process.
- We're working on extending the set of the integrated sources.

Thanks

- Contact a_briache@lcc.uma.es
- Available soon on
  http://www.yeastmed.uma.es