Компьютерные подходы к интеграции и систематизации знаний в области молекулярной биологии

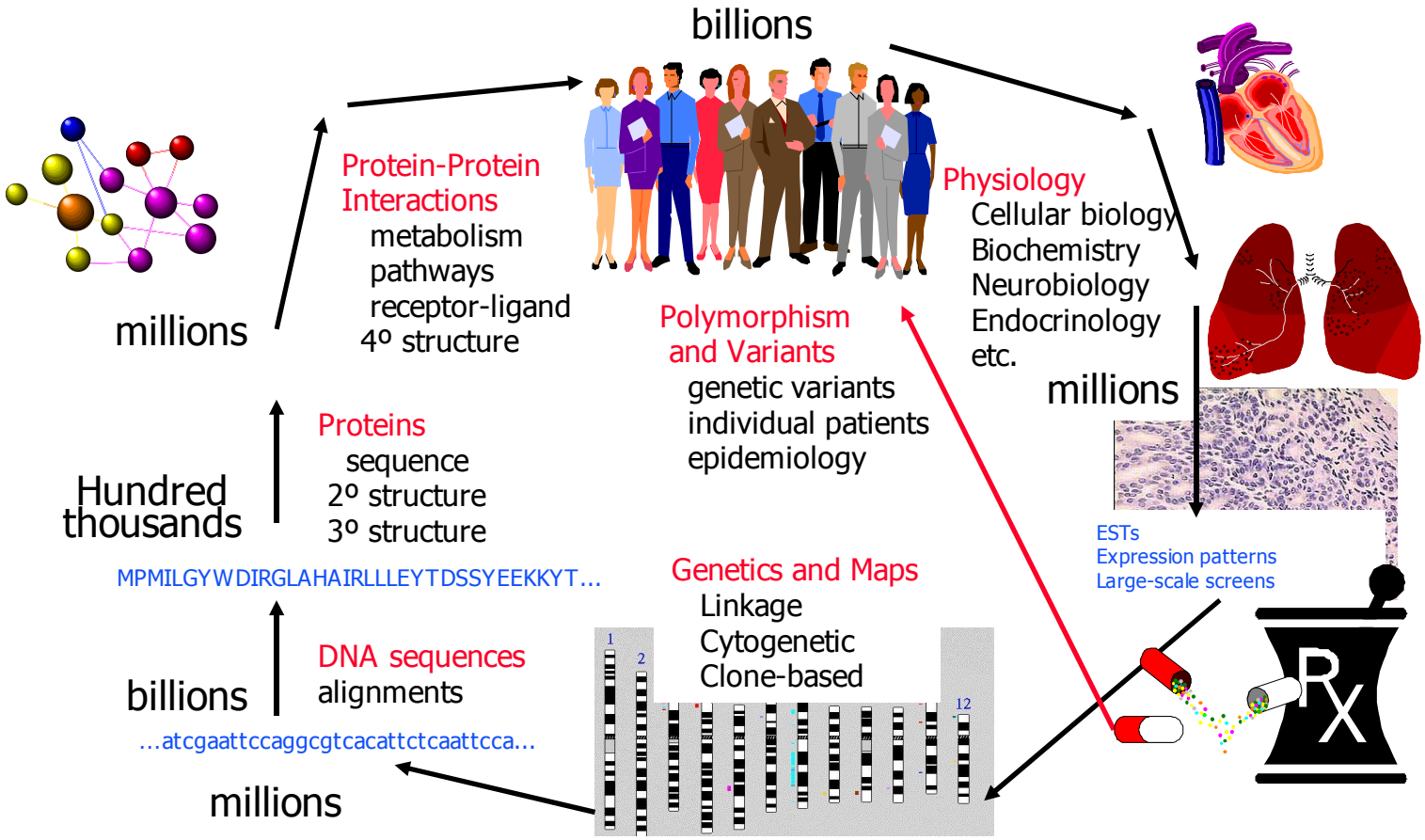
Н.Л.Подколодный

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

Содержание лекции

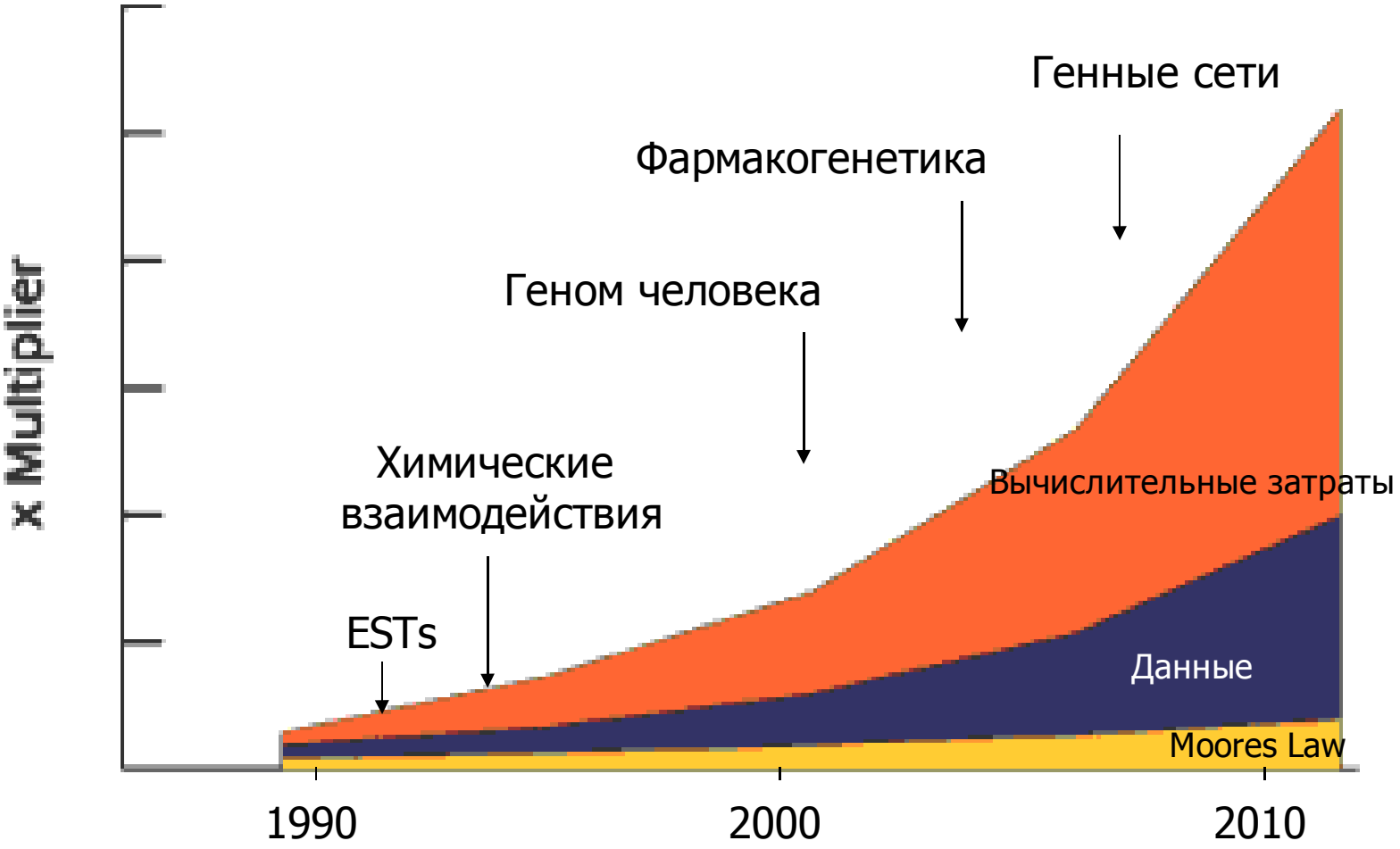
- Проблемы интеграции и основные подходы
- Средства интеграции и технологии
- Подходы к созданию единого описания предметной области. Онтология.
- GRID системы
- Примеры интегрированных систем в биоинформатике

Биомедицинские данные:
Исключительная сложность и масштабность

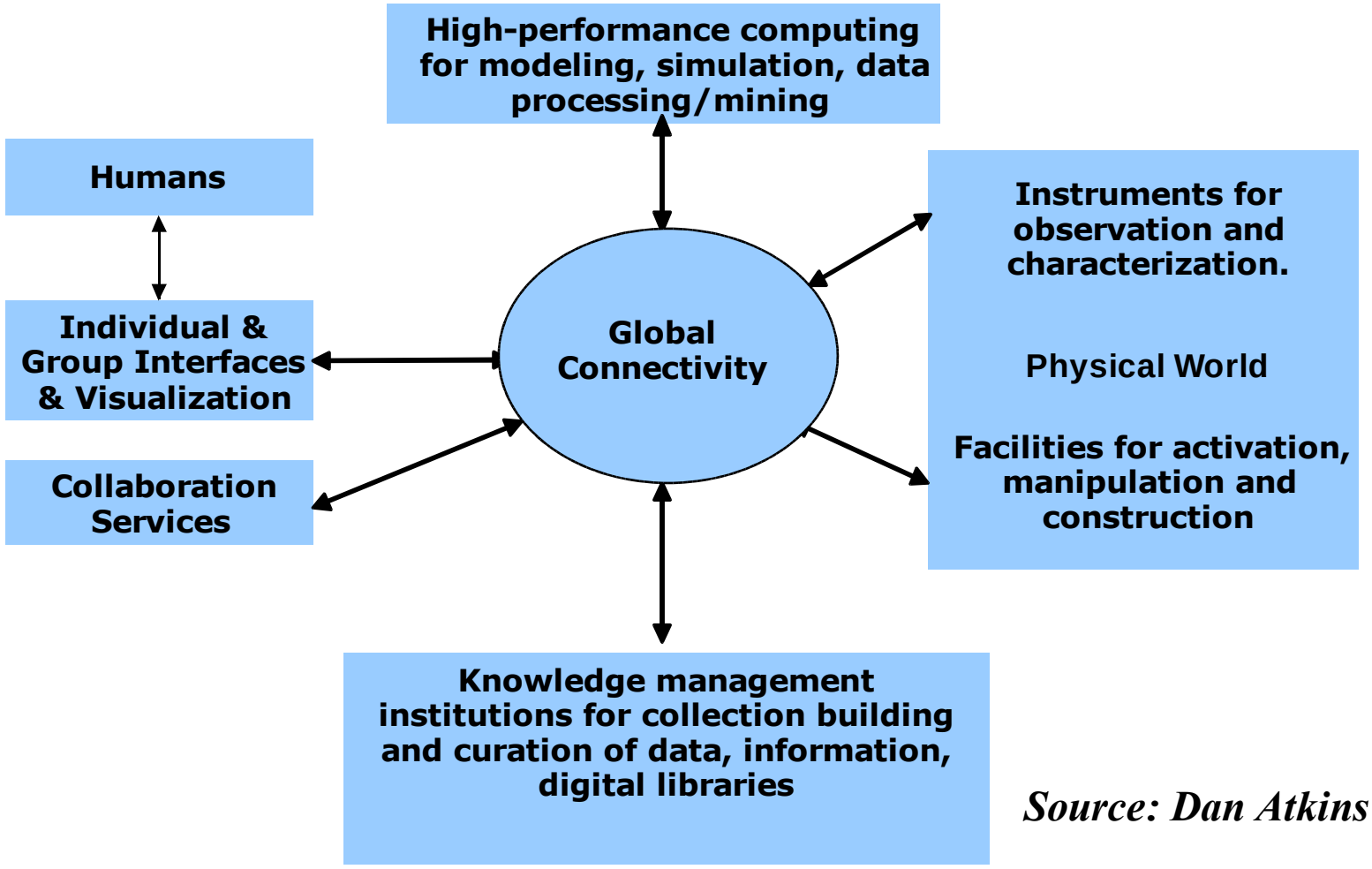


Мотивация

Сложность проблемы и большие затраты



Компоненты глобальной системы научного взаимодействия



Source: Dan Atkins

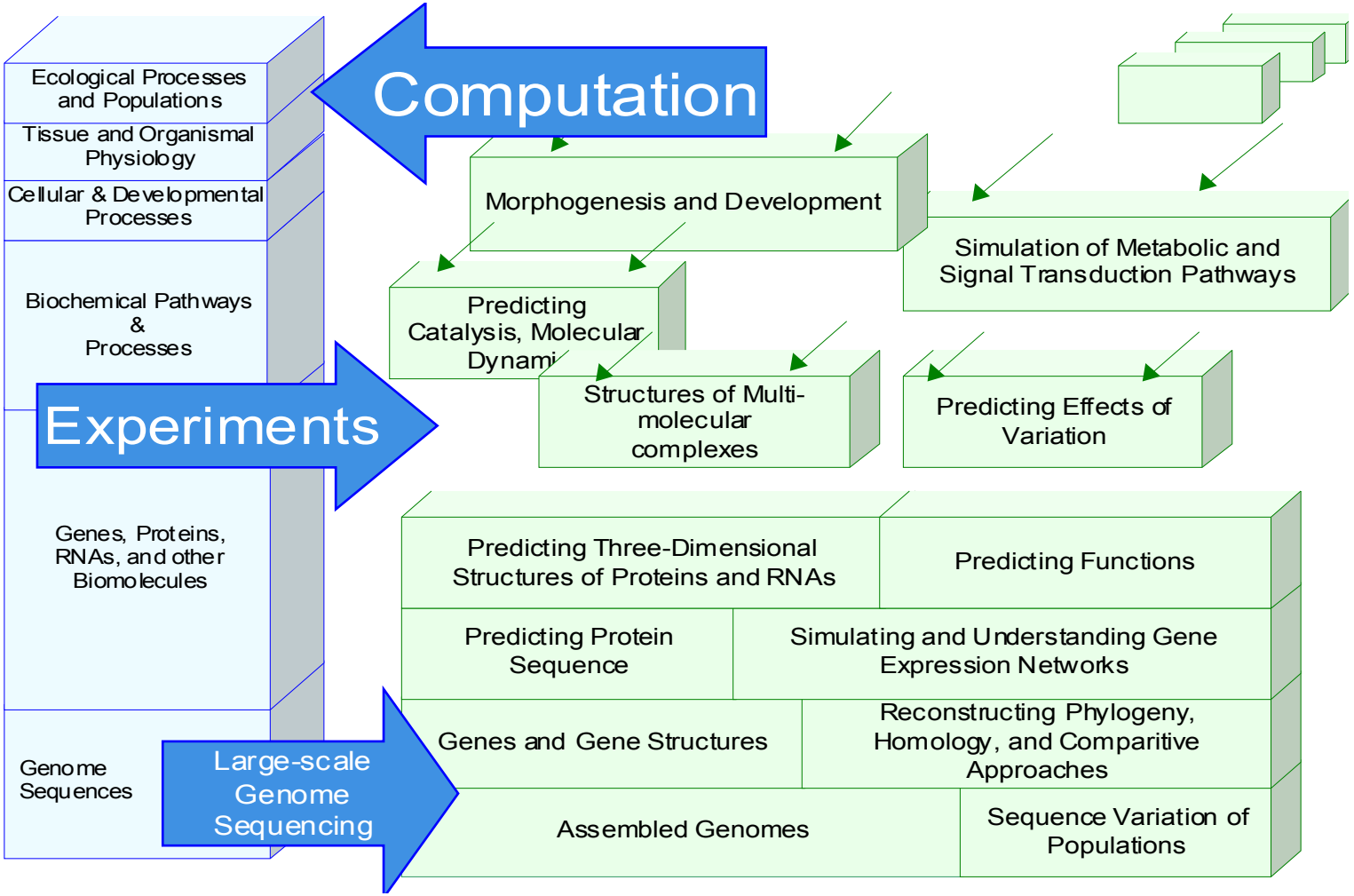
Типы ресурсов

- HTML документы
- Цифровые images
- Базы данных
- книги
- статьи
- метаданные
- коллекции
- Службы (сервера)
- программы
- люди
- понятия
- события

Что надо пользователю?

- Пользователь хочет
  - найти, идентифицировать, отобрать, получить и использовать
- Собственник / администратор
  - Описать, обеспечить доступ
  - Управлять доступом
  - администрировать
- Полезные свойства службы метаданных

Пример взаимодействия данных и программ при решении сложной задачи



Типы источников информации

- Результаты эксперимента, лабораторные журналы.
- Научные публикации, в которых описаны результаты эксперимента и их интерпретация.
- Научные публикации (обзоры), в которых обобщены результаты многих исследований в конкретной предметной области
- Базы данных, в которых представлены факты или экспериментальные данные по отдельным аспектам исследования регуляции экспрессии генов.
- Обобщения результатов экспериментов, используя методы предсказания, распознавания и т.д.
- data mining, text mining и т.д.
- Результаты моделирования и возможные интерпретации.
- .....

Трудности интеграции знаний

- Распределенность знаний.
- Гетерогенность источников данных.
- Разнородность форматов представления данных и средств доступа.
- Многоуровневая иерархическая структура данных.
- Сложная сетевая организация взаимосвязей объектов.
- Существенное различие в детальности и полноте описания.
- Различие по точности описания и противоречивость знаний.
- Разнородность понятий предметной области.
- Разнообразие взглядов на данные в зависимости от задачи.
- Разнообразие экспериментов и логики их анализа для формирования моделей ПО.
- Разнородность методов анализа данных и моделирования.

Фундаментальные проблемы в интеграции знаний

- Гетерогенные программные системы
  - hardware platforms
  - Операционные системы
  - Сетевые протоколы
  - Языки программирования & форматы данных
- Гетерогенная структура и семантика данных
  - Конфликт имен
  - Конфликт измерений
  - Конфликт представления
  - Конфликт вычислений
  - Конфликт уровней описания

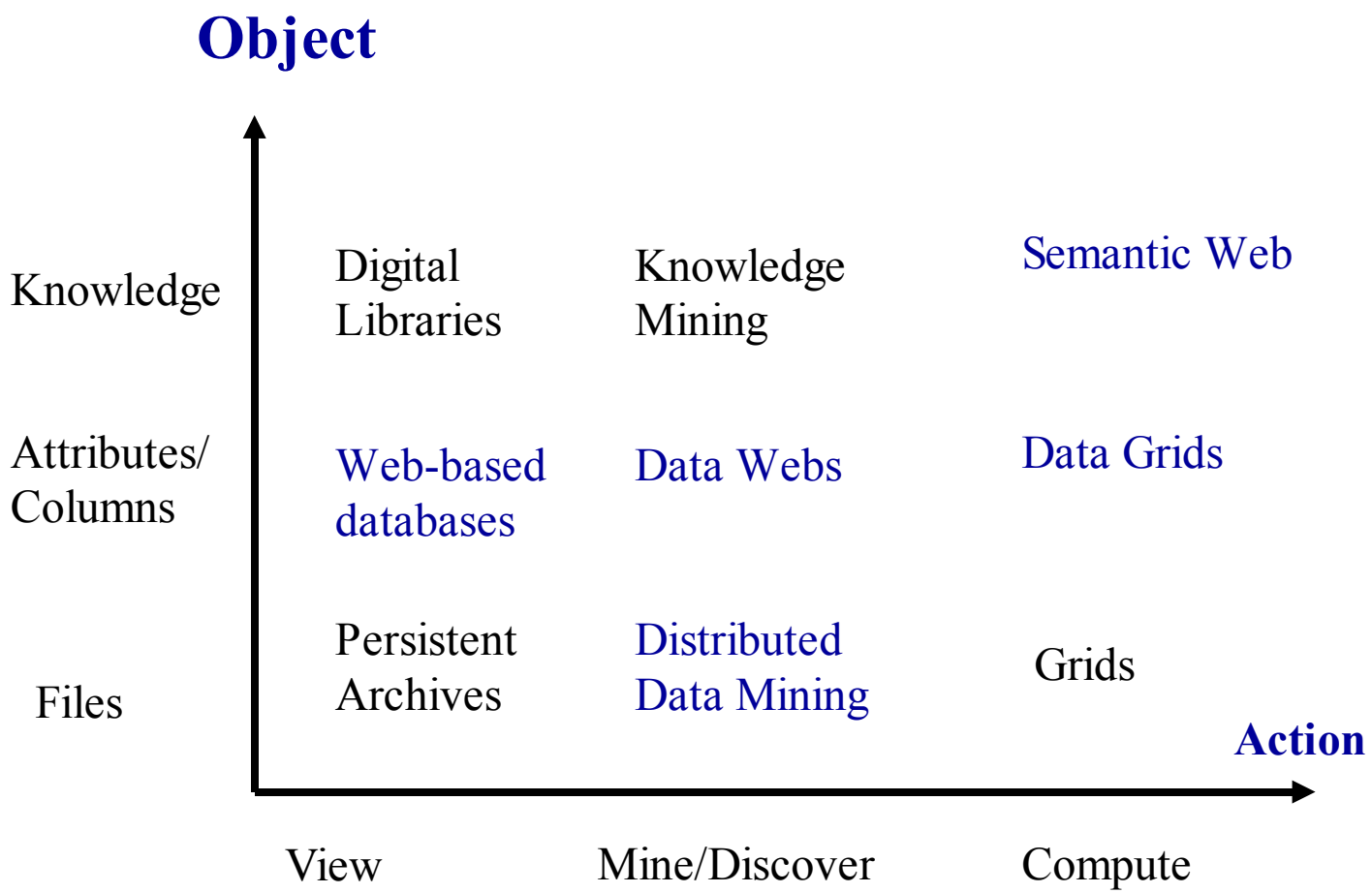
Интеграция распределенных гетерогенных данных

2. Источники могут использовать различные модели данных и предоставлять различные интерфейсы для доступа к ним (реляционные, объектные) или данные могут быть неструктурированные или слабоструктурированные (HTML, XML, текстовые, бинарные и т.д.)
3. Источники атомарные (взаимодействие только через предоставляемый интерфейс и невозможность влиять на внутренние процессы)

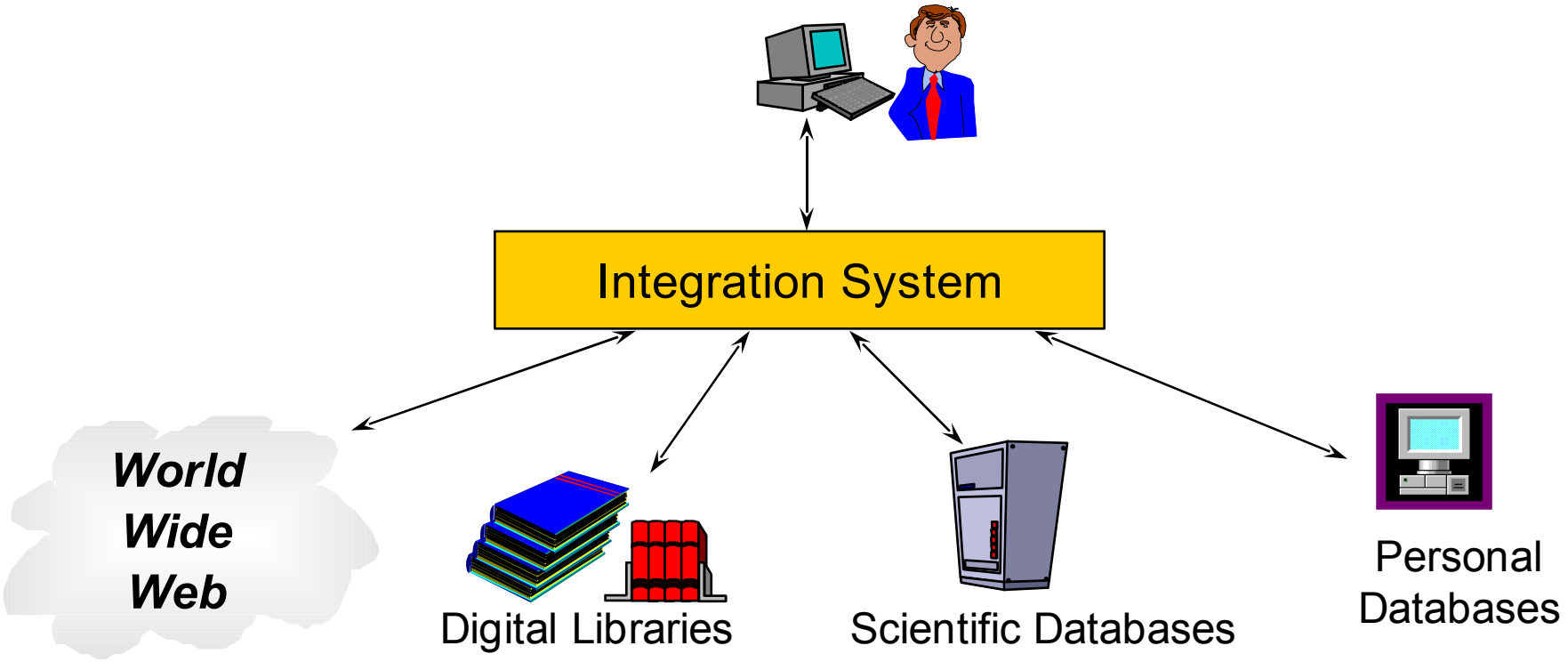
Подходы:

6. Хранилища данных
7. Виртуальные Хранилища данных

Viewpoints for Distributed Data

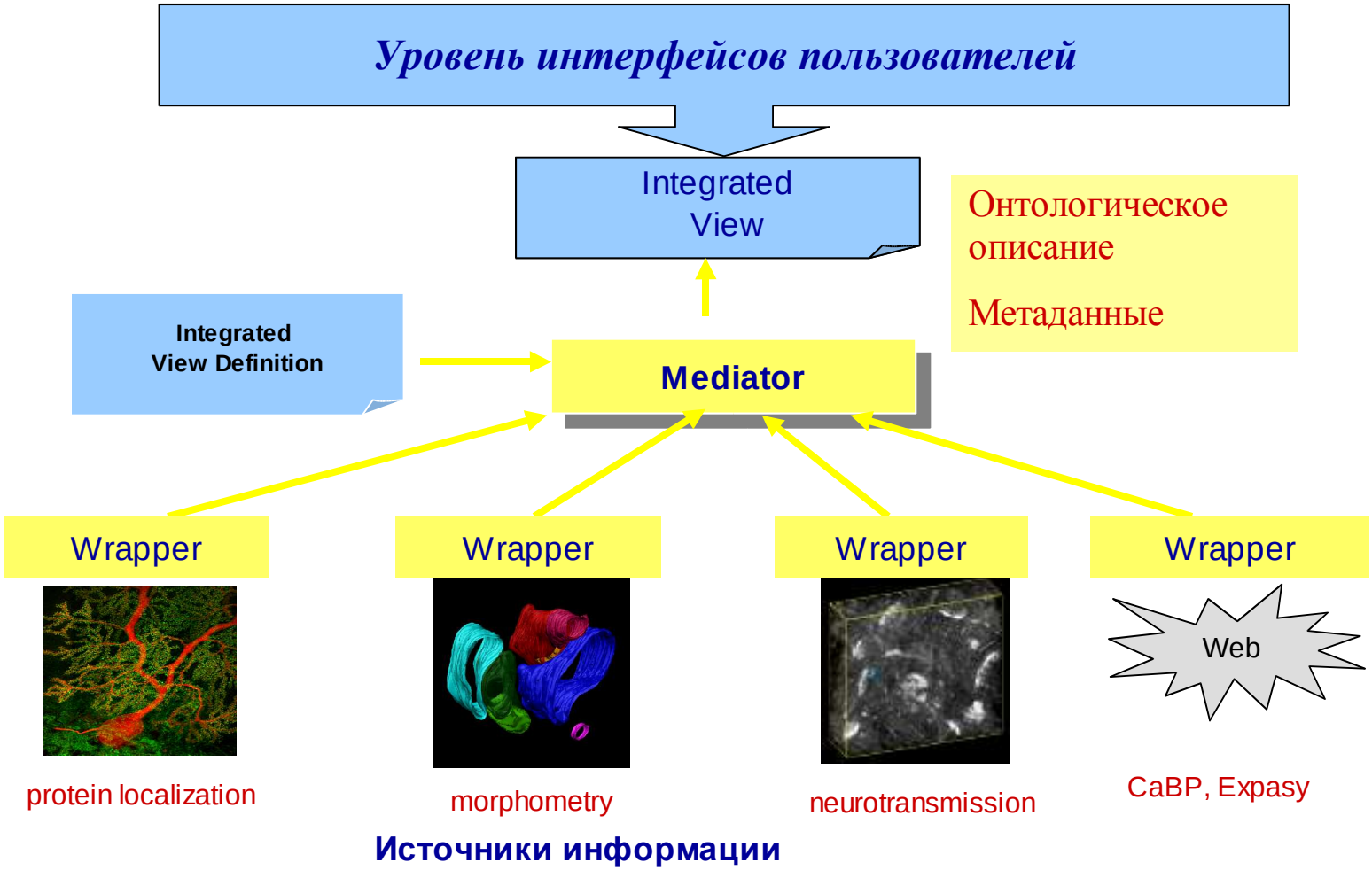


Integrated View of Heterogeneous Data



- Collects and combines information
- Provides integrated view, uniform user interface
- Supports sharing

Архитектура системы интеграции



Как нам представлять данные в интегрированной системе

- *Relational Data Model*
  - Множество столбцов и строк
  - Фиксированное множество простых типов данных
- *Data cube*
  - Специализированные системы управления хранилищами данных (warehouse)
  - Использование в качестве модели единого многомерного отношения
- *Специальные подходы для представления гетерогенных данных.*

Какое представление ?

- Нам необходимо иметь связь между репозитариями, в которых хранятся данные (data warehouse, transactional databases) и где они используются (Web interface, business application)
- Модель данных должна позволять обмениваться данными со структурой.
- Уменьшить требования к структурам в существующих высокоструктурированным базам данных.

Слабоструктурированные данные

- Почему нам необходимы слабоструктурированные данные ?
- Что такое слабоструктурированные данные ?
- Графовая модель слабоструктурированных данных

Слабоструктурированные данные

Априорная схема и апостериорное описание структуры данных

- Database: фиксируется схема, затем заполняется
- Web: создается много Web pages, затем определяется схема доступа

- Схема данных очень большая и сложная
- Схема часто игнорируется в запросах (запросы и browsing)
- Схема данных быстро меняется

Слабоструктурированные данные

Структура неправильная – данные гетерогенные

- Часть данных отсутствует
- Внешняя информация (аннотация)
- Различия в типах

- Структура может быть задана неявно. т.е. text + grammar (e.g., SGML)
- Структура может быть определена частично
  - Часть данных не имеют структуры (images)
  - Часть данных имеют плавающую структуру (text)
- Используются только индикаторные типы

Слабоструктурированные данные

- Элементы данных могут повторяться.
- Структура части информации может зависеть от точки зрения пользователя.

Например, информация о Person содержит:
name, address, phone and photo (gif file).

Example

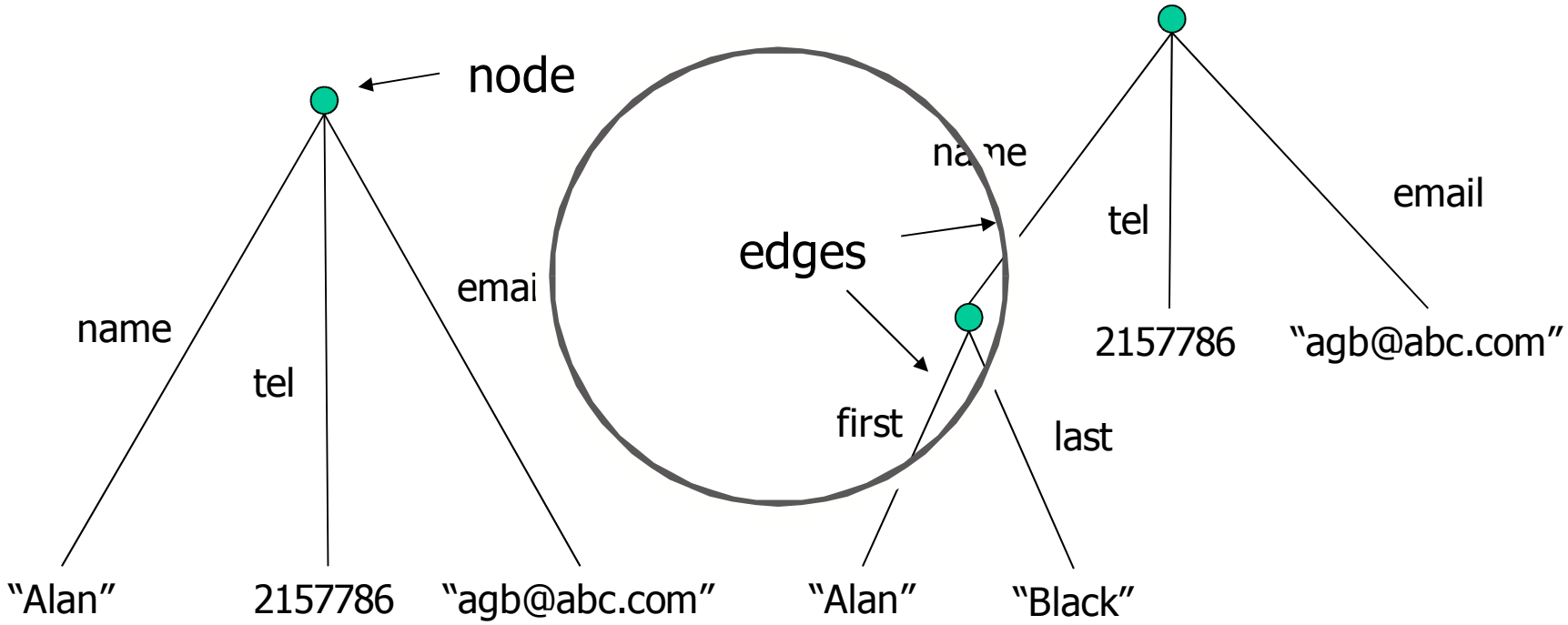
{name: “Alan”, tel: 2157786, email: “agb@abc.com”}

{name: {first: “Alan”, last: “Black”}, tel: 2157786, email: “agb@abc.com”
}

- Different from usual tuples in that we allow duplicates:

{name: “Alan”, tel: 2157786, tel: 2159989, email: “agb@abc.com”}

Graph Representation



Современные информационные технологии

- Технологии анализа и описания сложных систем

- UML – универсальный язык концептуального моделирования и инструментальные средства, поддерживающие этот язык.

- XML технологии для спецификации языка представления слабоструктурированных данных и знаний в предметной области

Что такое XML?

XML означает E**X**tensible **M**arkup **L**anguage и является подмножеством SGML.

XML – это метаязык, т.е. набор правил для создания новых языков разметки.

- XML тэги не предопределены в XML. Вы должны сами определить свои собственные тэги.
- XML использует **Document Type Definition** (DTD) или **XML Schema** для описания структуры данных
- XML с DTD или XML Schema самодостаточны для описания данных.
- XML ничего не «делает». XML создан для описания данных.

Сравнение с HTML:

XML создан для описания данных, а HTML создан для отображения данных.

Таким образом, XML не заменяет, а дополняет HTML.

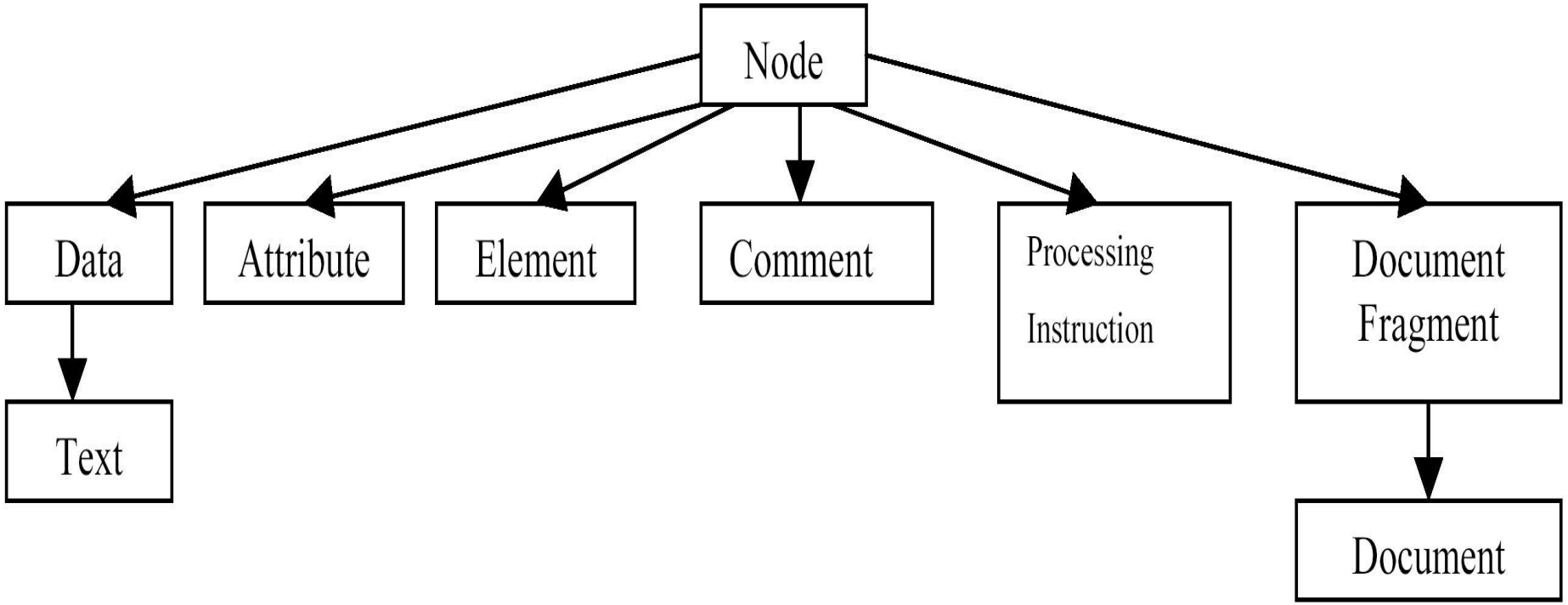
Свойства XML

- Способ стандартизации терминологии и обмена знаниями.
- XML помогает отделить синтаксис и структуру данных от способа их отображения и использования.
- Один источник данных – несколько представлений.
- Способ взаимодействия гетерогенных сетевых агентов в распределенных средах.
- Работа со слабоструктурированными данными.
- Улучшение возможности поиска слабоструктурированных данных.
- Увеличение доступности данных. Стандартизация форматов и методов представления.
- Независимый от приложения формат данных.
- Возможность отображения данных на любых устройствах.
- Более простая разработка приложений.
- Организация совместной обработки данных.
- Возможность создания новых языков.
- XML может читать как человек, так и программа.
- Возможность использования составныз документов.

Стандарты в XML

XML Document Type Definition (DTD) определяет структуру документа.
XML Schema определяет структуру и типы данных.
XML Namespaces. Обеспечивают разделение словарей элементов и атрибутов.
Объектная модель документа (DOM) ориентированный на работу с деревом документа, не зависящий от платформы и языка программирования интерфейс, позволяющий программам динамически разбирать и изменять содержимое, структуру и стиль документов
Упрощенный программный интерфейс для XML (Simple API for XML, SAX) (потоково-ориентированный независимый от платформы и языка интерфейс, позволяющий программам обрабатывать данные в формате XML)
Упрощенный протокол доступа к объектам XML (SOAP) определяет механизм удаленных вызовов процедур (RPC) с использованием синтаксиса XML, реализует клиент-серверное взаимодействие по сети.
Язык описания путей XML (XPath) система адресации, идентифицирующая одну или несколько позиций в иерархии документа
XSL (Extensible Style Language) - расширяемый язык стилей
Языки XML для ссылок и связей (XPointer и XLink) - определяют ссылки на части XML-документа или между документами.
Resource Description Framework (RDF) простая модель метаданных.
…..

Иерархия наследования интерфейсов DOM



```
<?xml version="1.0"?>
<!DOCTYPE weather SYSTEM="weather.dtd">
<weather>
        <current>
                <temp scale="F">72</temp>
                <pressure>1005</pressure>
                <humidity>43</humidity>
        </current>
        <min>
                <temp scale="F">65</temp>
                <pressure>998</pressure>
                <humidity>38</humidity>
        </min>
        <max>
                <temp scale="F">78</temp>
                <pressure>1010</pressure>
                <humidity>43</humidity>
        </max>
</weather>
```

URI – Идентификатор ресурсов
SYSTEM – используется ссылка в формате URI
PUBLIC – используется уникальный идентификатор DTD

```
<?xml version="1.0"?>
<!DOCTYPE weather [
<!ELEMENT weather ( current, ( min, max )? )>
<!ELEMENT current ( temp, pressure, humidity )>
<!ELEMENT min ( temp, pressure, humidity )>
<!ELEMENT max ( temp, pressure, humidity )>
<!ELEMENT temp ( #PCDATA )>
<!ATTLIST temp scale ( C | F ) #REQUIRED>
<!ELEMENT pressure ( #PCDATA )>
<!ELEMENT humidity ( #PCDATA )>
]>
```

Примеры языков описания молекулярно-генетических данных, основанных на XML.

BSML - Bioinformatic Sequence Markup Language
BIOML - Biopolymer Markup Language.
GAME - Genome Annotation Markup Elements
ProML - Protein Markup Language.
CML - Chemical Markup Language.
GEML – формат описания данных по экспрессии генов.
SBML - Systems Biology Markup Language для моделирования молекулярно-генетических систем и процессов.
CellML – язык описания модели клетки
AnatML - Anatomical Markup Language
MODL - Molecular Dynamics Markup Language
MSAML - An XML for Multiple Sequence Alignments
phyloML – филогенетические данные
RiboML - Ribonucleic Acid Markup Language
TML - Taxonomic Markup Language
XSIL - Extensible Scientific Interchange Language

Ontolingua
(http://www-ksl.svc.stanford.edu)

- **Предназначена для поддержки формального специфицирования задач пользователя на основе библиотеки формальных описаний фрагментов задач, моделей и понятий. Поиск, редактирование, проверка и т.д. в библиотеке Ontolingua теорий и определений.**
- **В основе языка формальных описаний и межмашинного обмена знаниями лежит язык Knowledge Interchange Format (KIF)**
- **Обеспечивает Интернет доступ пользователей-разработчиков онтологий**
- **Результат - Лисп программа описывающая**
  - **Библиотека онтологий**

Теория
Классы
Отношения
Функции
Аксиомы
Фрагменты
моделей ...
Понятия ...

Состав библиотеки онтологий

Онтология базовых знаний включает описание понятий: объект, система, состояние, поведение, событие, процесс, действие, функция

Онтология экспериментальных исследований и доказательств.

Терминологическая и информационная онтология включает тезаурусы и метаописание существующих баз данных, например, схему баз данных, описание полей, их интерпретация в терминах онтологии предметной области и т.д.

Онтология методов решения задач включает описания методов и средств решения задач. К этому разделу относится и метабаза, описывающая способы доступа к тем или иным программам, протоколы обращения, форматы и состав входных и выходных данных и т.д.

Онтология приложений содержит понятия необходимые для моделирования знаний при решении конкретных задач, включает в себя онтологию предметной области, онтологию базовых знаний, информационную онтологию и онтологию методов решения задач.

Примеры онтологических отношений

Общее-частное

Состав

Классификация

Происхождение

Функция

Причина-следствие

Регулятор функции

Пространственные отношения

Темпоральные отношения

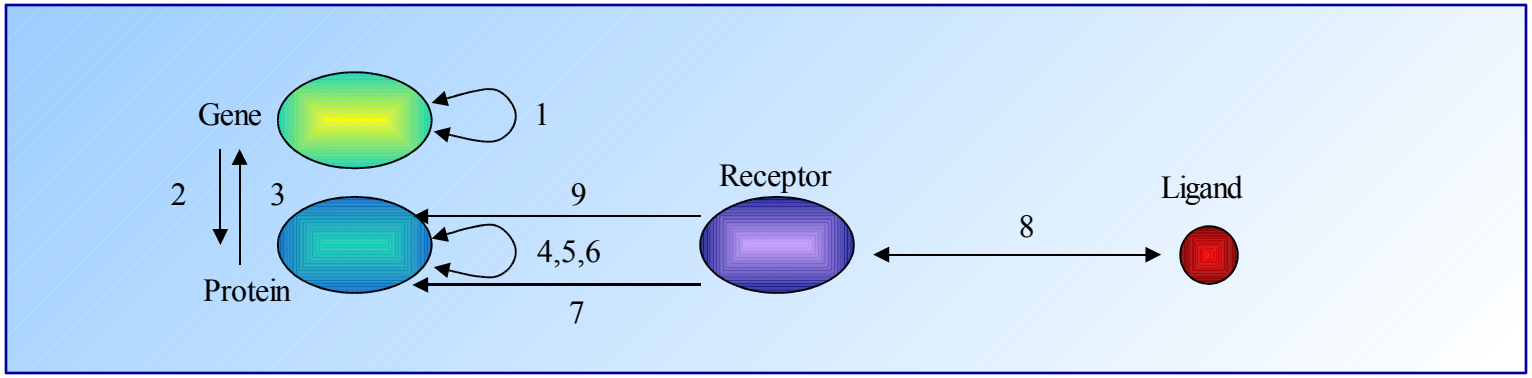
.....

Bio-Semantics

- Genes, Proteins, Interactions
- Functions and Roles
- Context-sensitive:
  - Tissue: Neural vs. epithelial
  - Developmental stages: fetal vs. adult
  - Environmental responses: Immune system
  - Compartments: Nucleus, cytoplasm, membrane
- Intercompartmental signal-transduction
- Complexity and Heterogeneity demands Formalism!

Relations Between Genes and Proteins

Part of an Ontology

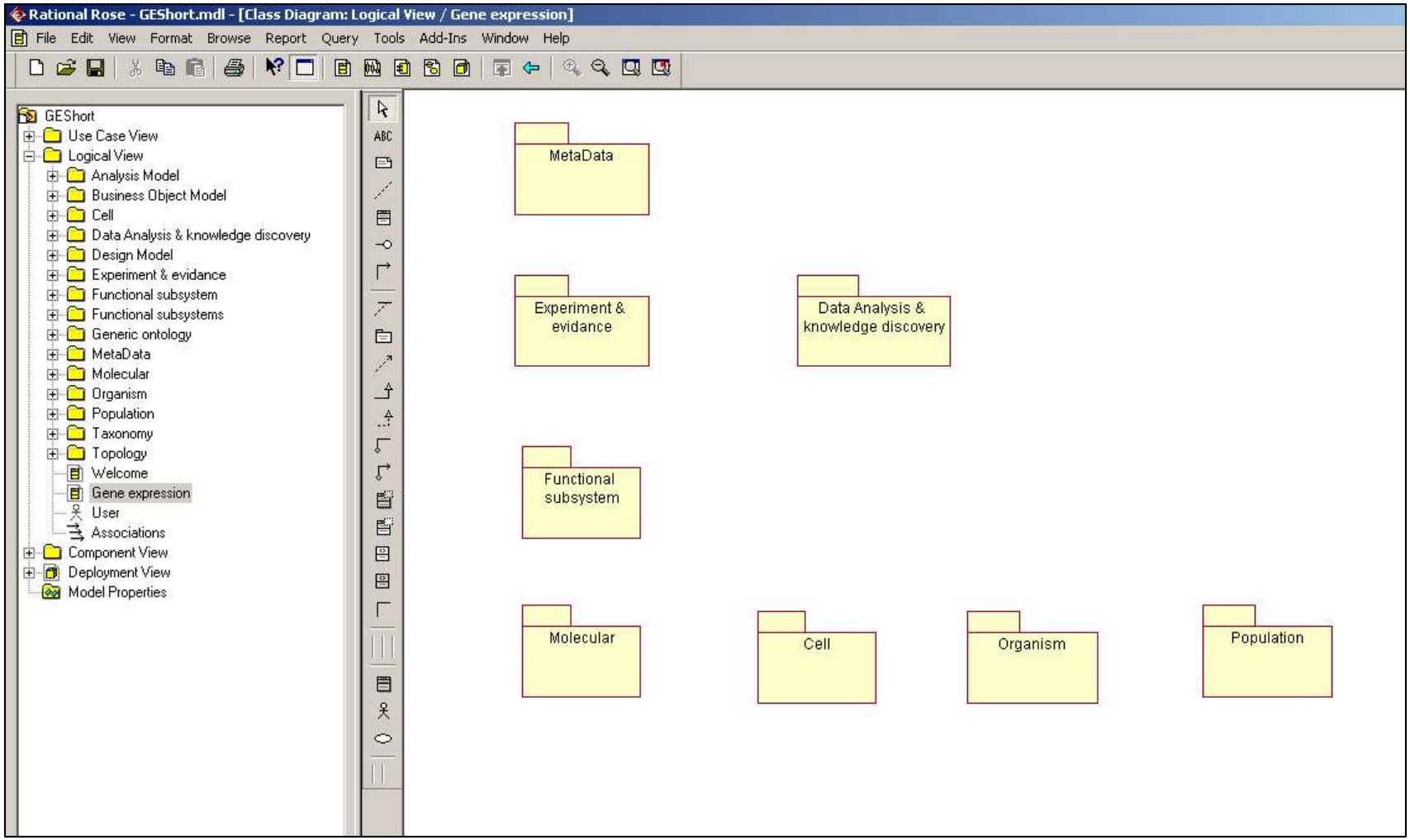


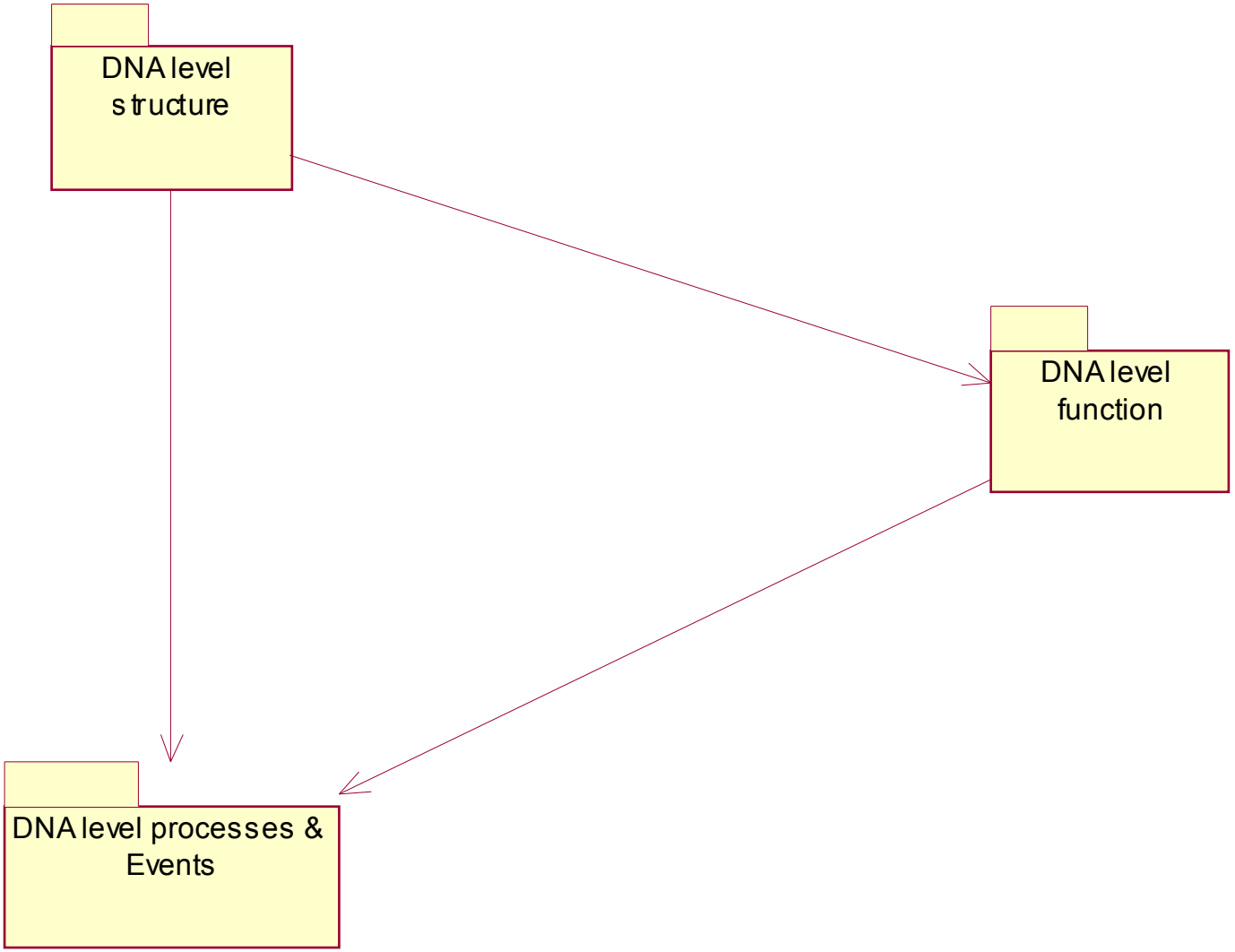
Relations

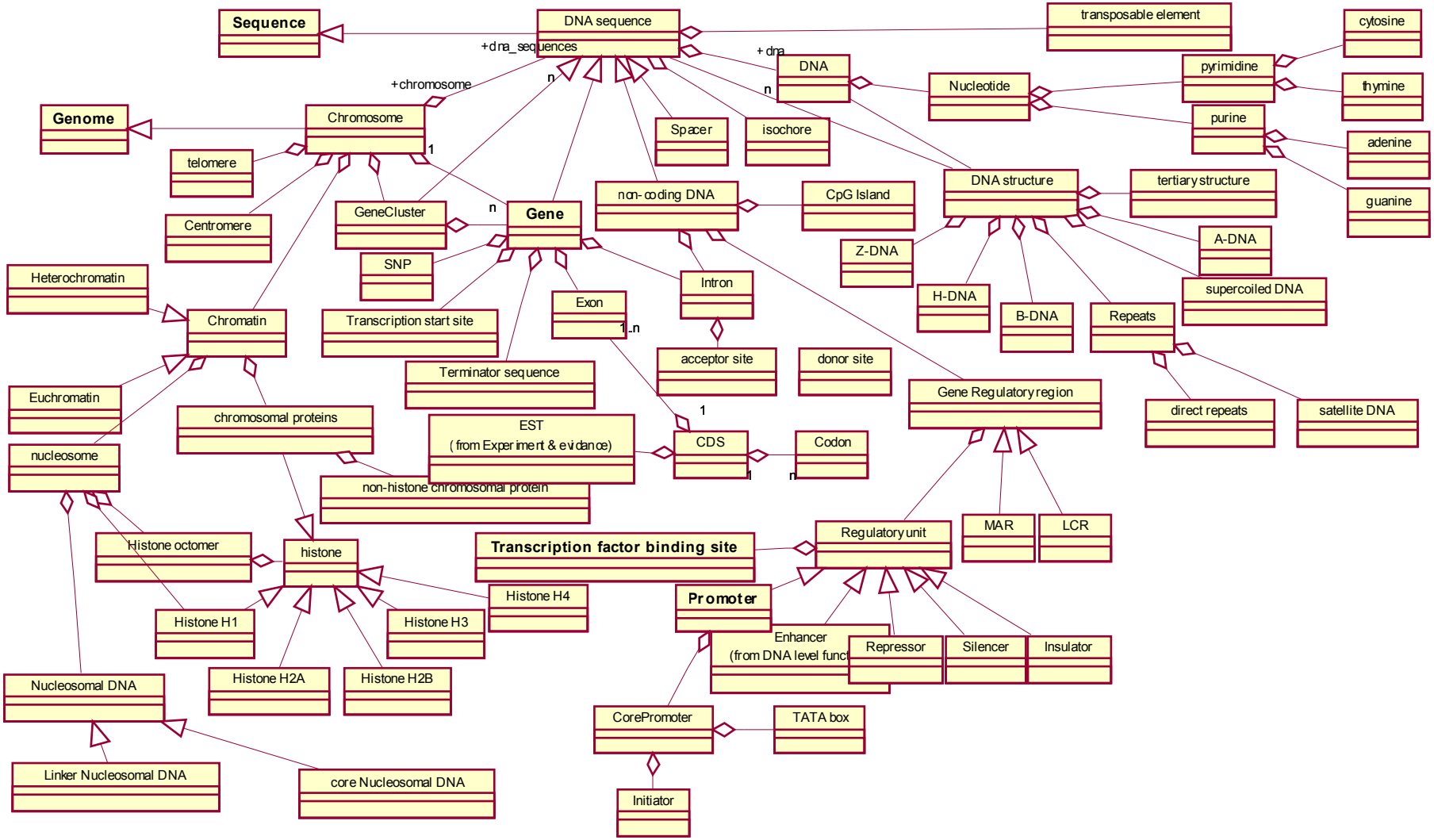
1- gene homologs
2- gene encodes a protein
3- protein can regulate the expression of a gene
4- protein phosphorylates another protein
5- protein binds to another protein
6- protein lyses another protein
7- Proteins can sometimes be receptors
8- Receptors bind a ligand
9- Receptors (if bound) activate other proteins

Объектно-ориентированные технологии

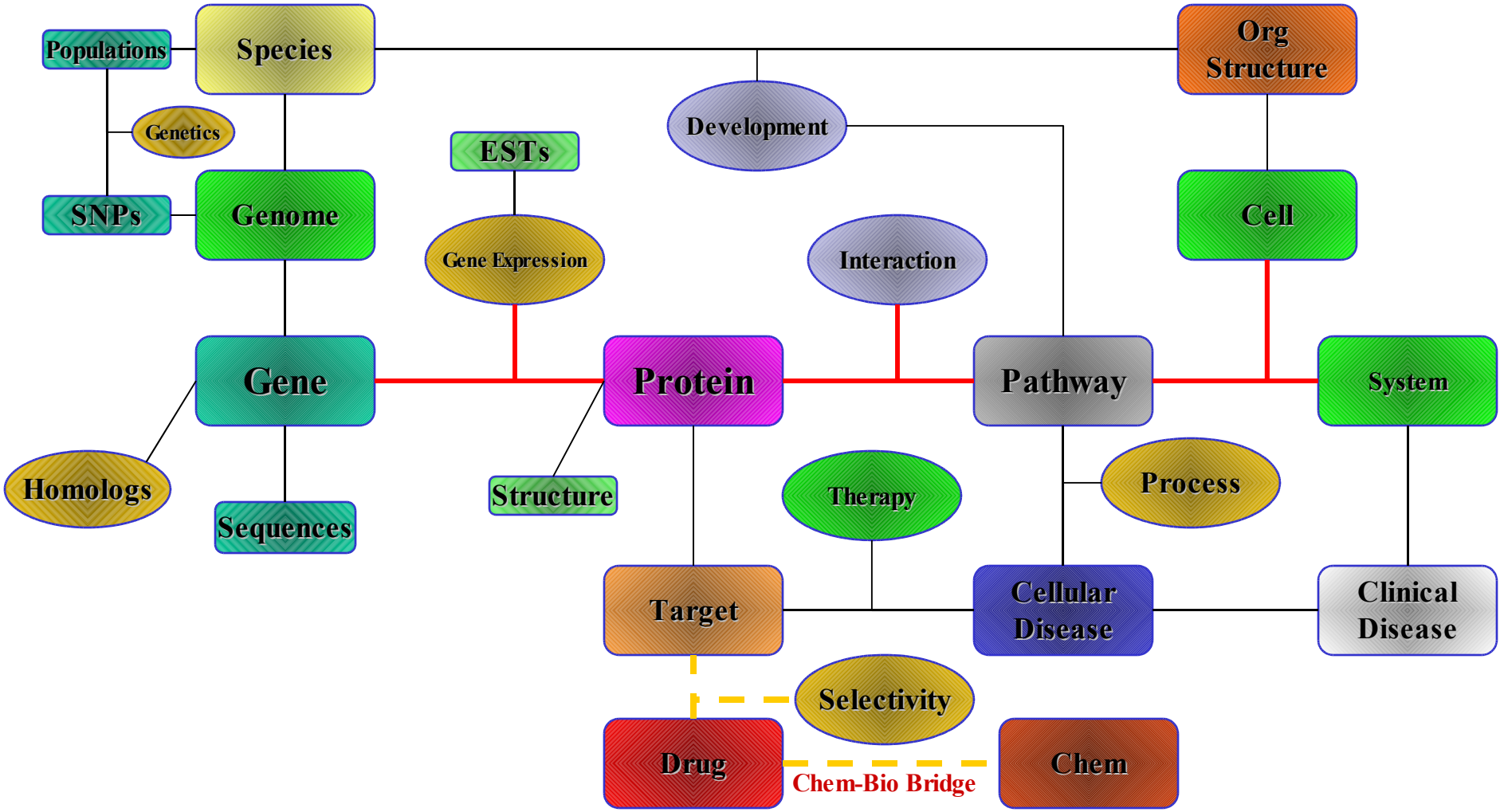
- Абстракция
- Инкапсуляция
- Модульность
- Иерархия
- Типизация


Rational Rose - GEShort.mdl - [Class Diagram: Logical View / Gene expression]
File Edit View Format Browse Report Query Tools Add-Ins Window Help
GEShort
Use Case View
Logical View
Analysis Model
Business Object Model
Cell
Data Analysis & knowledge discovery
Design Model
Experiment & evidance
Functional subsystem
Functional subsystems
Generic ontology
MetaData
Molecular
Organism
Population
Taxonomy
Topology
Welcome
Gene expression
User
Associations
Component View
Deployment View
Model Properties
MetaData
Experiment & evidance
Data Analysis & knowledge discovery
Functional subsystem
Molecular
Cell
Organism
Population


DNA level
structure
DNA level
function
DNA level processes &
Events


Sequence
DNA sequence
transposable element
cytosine
+dna_sequences
+dna
DNA
pyrimidine
thymine
+chromosome
n
n
Nucleotide
Genome
Chromosome
purine
Spacer
isochore
adenine
telomere
1
DNA structure
tertiary structure
guanine
non-coding DNA
CpG Island
GeneCluster
n
Gene
Centromere
Z-DNA
A-DNA
SNP
Heterochromatin
Intron
H-DNA
supercoiled DNA
Exon
Chromatin
Transcription start site
1..n
B-DNA
Repeats
acceptor site
donor site
Terminator sequence
Euchromatin
Gene Regulatory region
chromosomal proteins
EST
(from Experiment & evidance)
1
direct repeats
satellite DNA
CDS
Codon
nucleosome
1
n
non-histone chromosomal protein
Histone octomer
histone
Transcription factor binding site
Regulatory unit
MAR
LCR
Histone H4
Promoter
Histone H1
Histone H3
Enhancer
(from DNA level funct
Repressor
Silencer
Insulator
Histone H2A
Histone H2B
Nucleosomal DNA
CorePromoter
TATA box
Linker Nucleosomal DNA
core Nucleosomal DNA
Initiator

Объединение онтологий



Bio-Ontologies

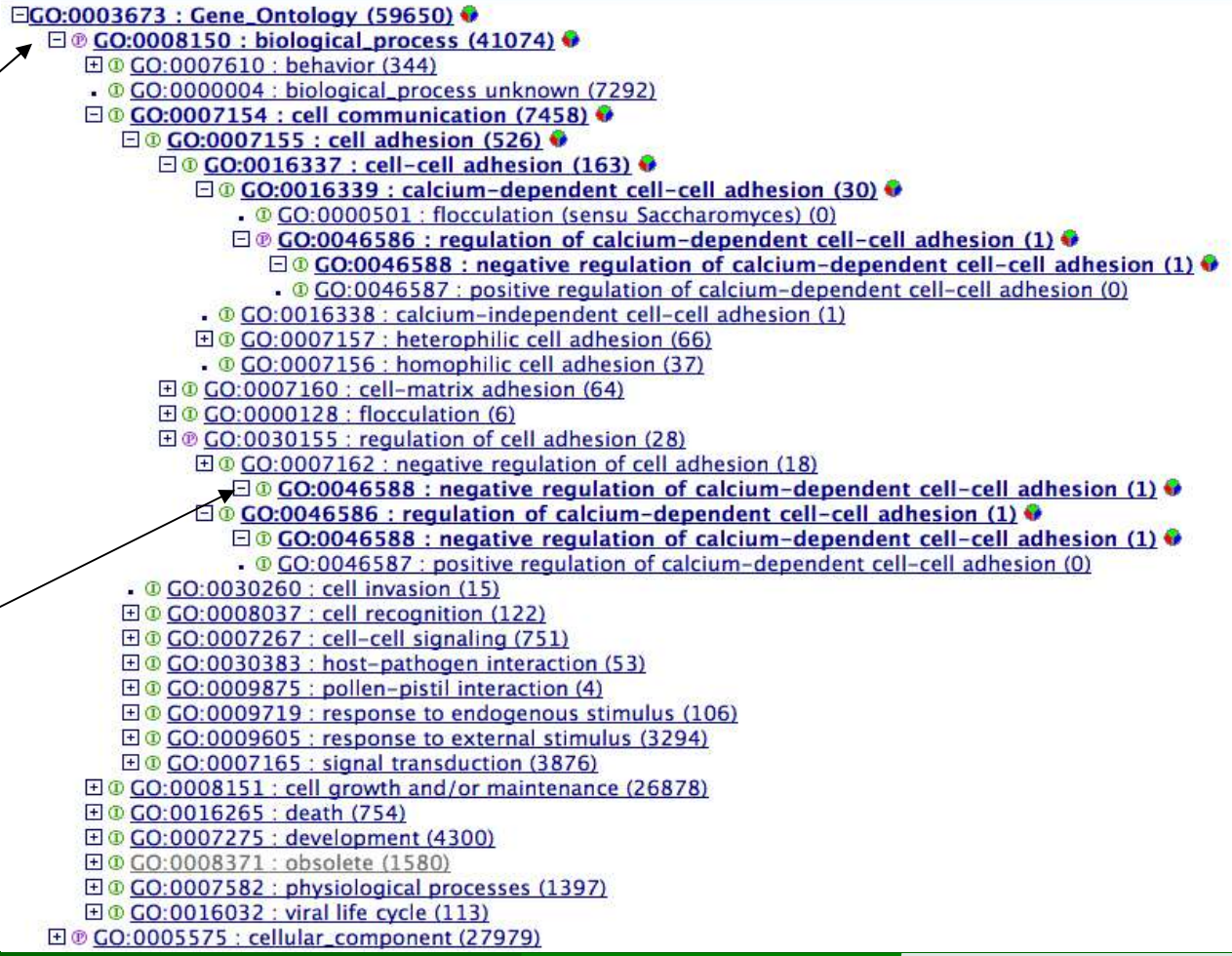
- Gene Ontology – http://www.geneontology.org
- TAMBIS - http://img.cs.man.ac.uk/tambis
- MBO - http://igd.rz-berlin.mpg.de/www/oe/mbo.html
- Riboweb – http://smi-web.stanford.edu/projects/helix/riboweb.html
- PharmaGKB - http://pharmgkb.stanford.edu/
- Interaction Ontology - http://www.ai.sri.edu/pkarp/interactions.html

Directed Acyclic Graphs (DAGs)

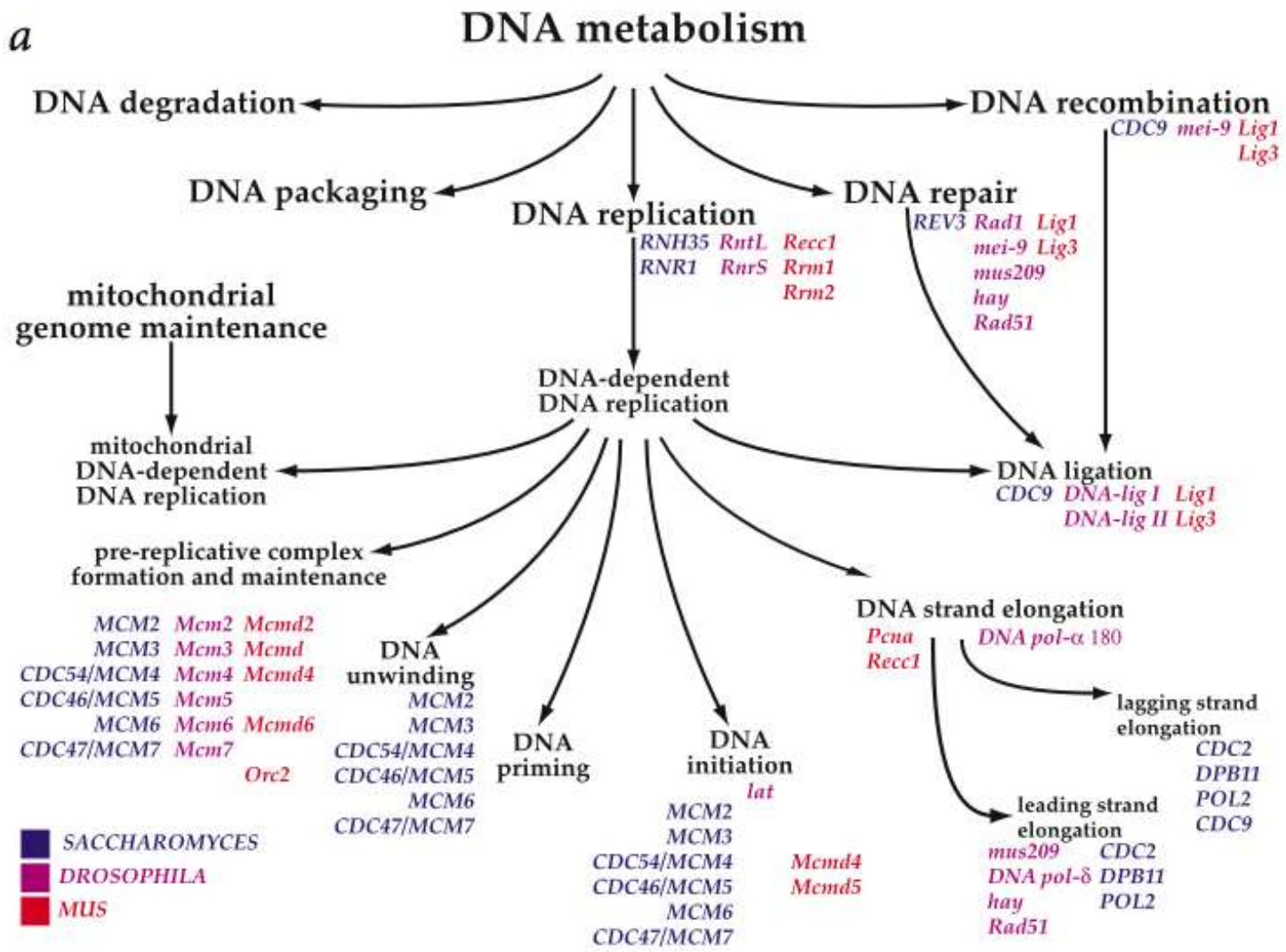
The ontologies are structured vocabularies in the form of directed acyclic graphs (DAGs) that represent a network in which each term may be a “child” of one or more than one “parent.”

“part-of”: child is a component of the parent

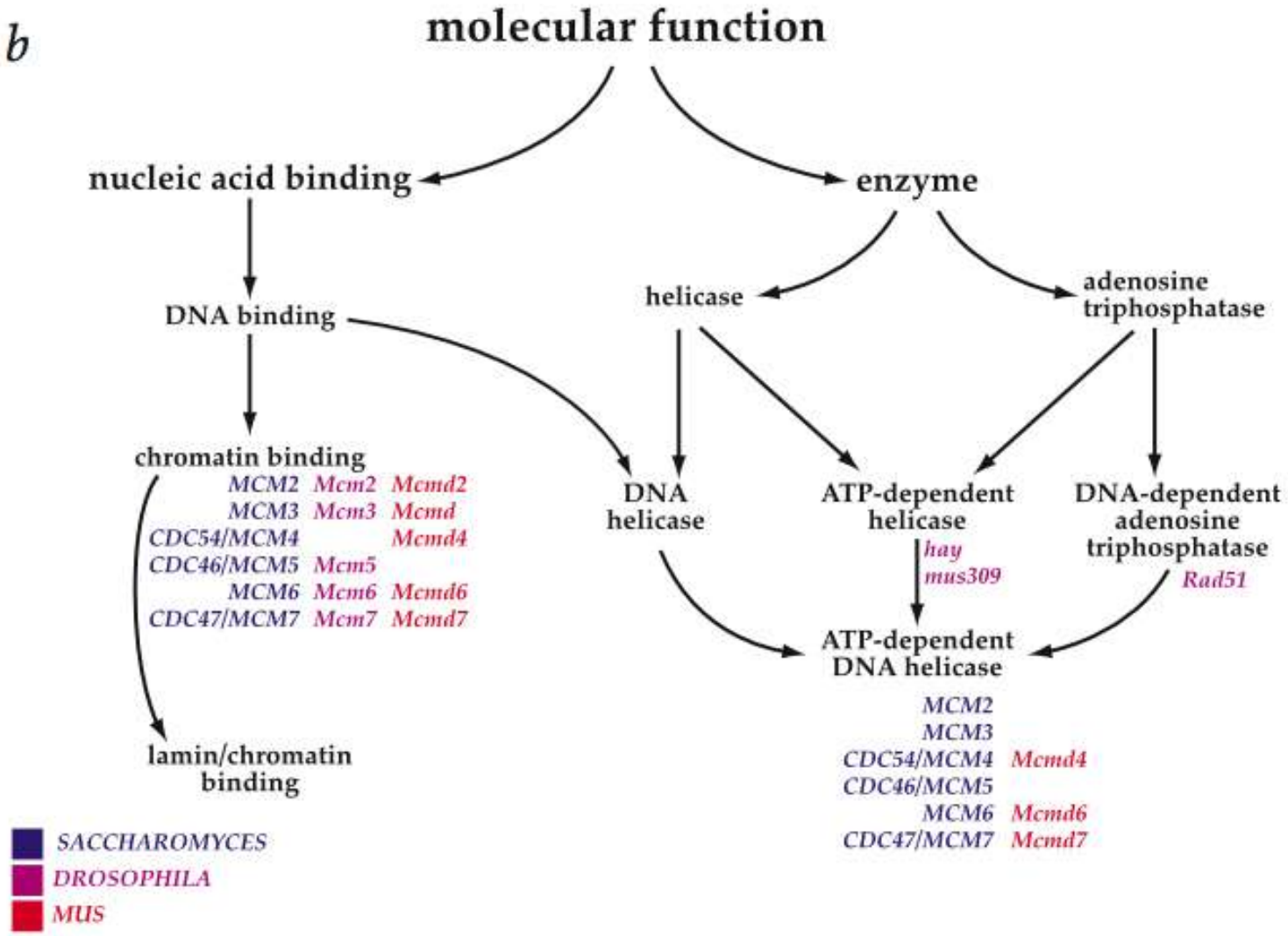
“is-a”: child is an instance of the parent



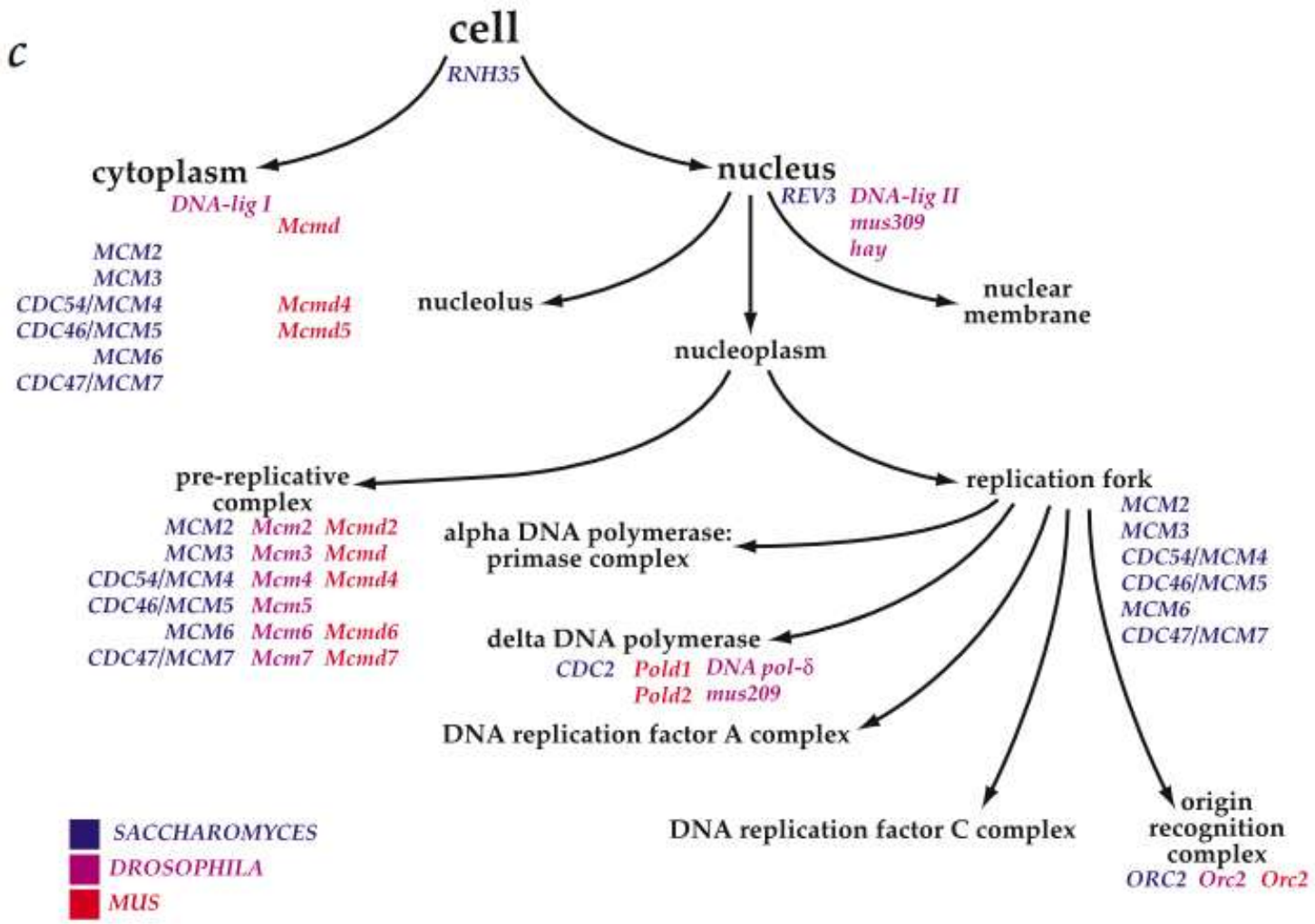
Biological Process Ontology



Molecular Function Ontology

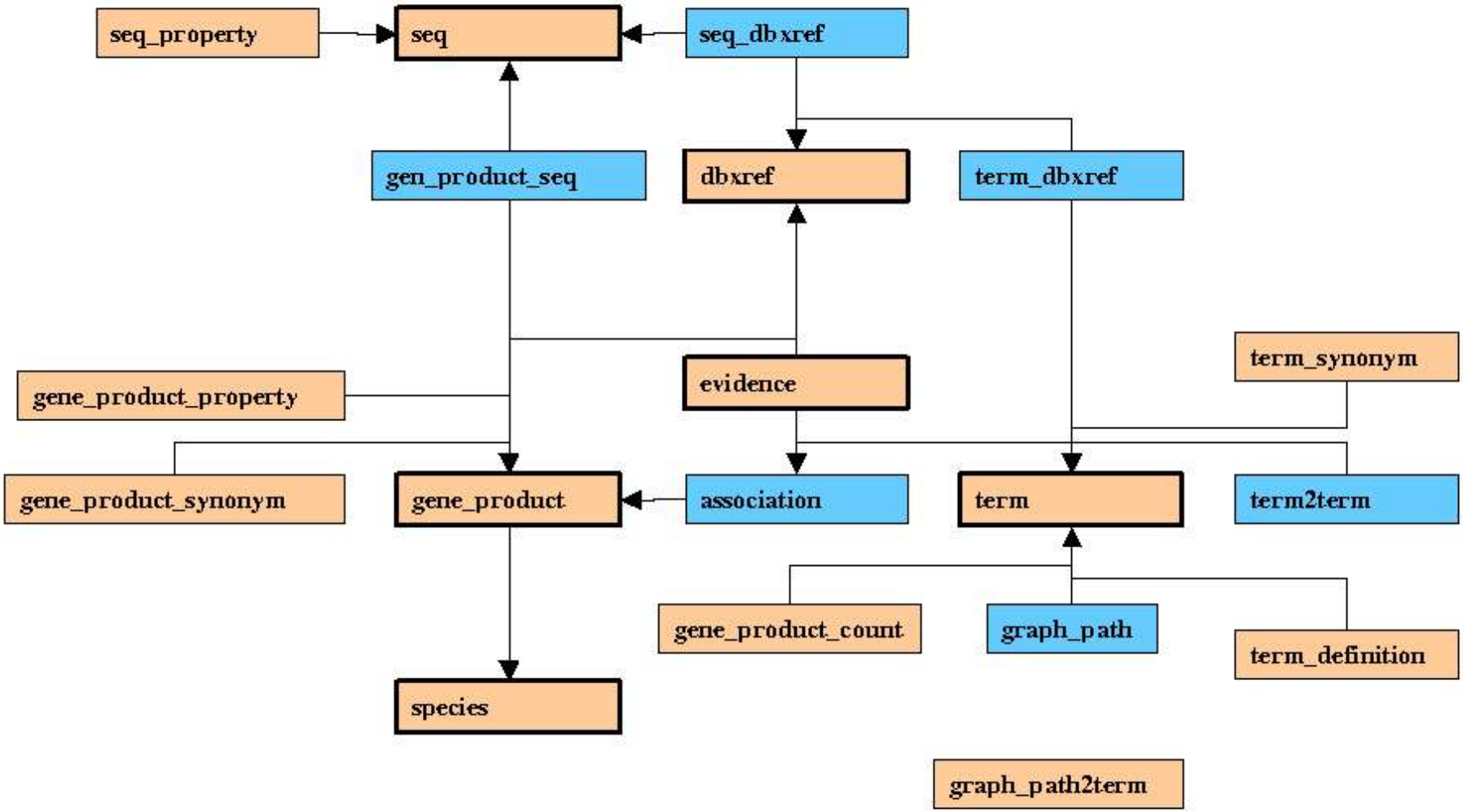


Cellular Component Ontology



The GO Relational Database in MySQL

Examples of SQL Queries: http://www.fruitfly.org/cgi-bin/wiki/view.pl/GoWeb/GoExampleQueries



Datamining & Knowledge Discovery

- Анализ сходства
- Классификация
- Регрессия
- Апроксимация
- Обобщение
- Анализ временных рядов
- Оценивание
- ...

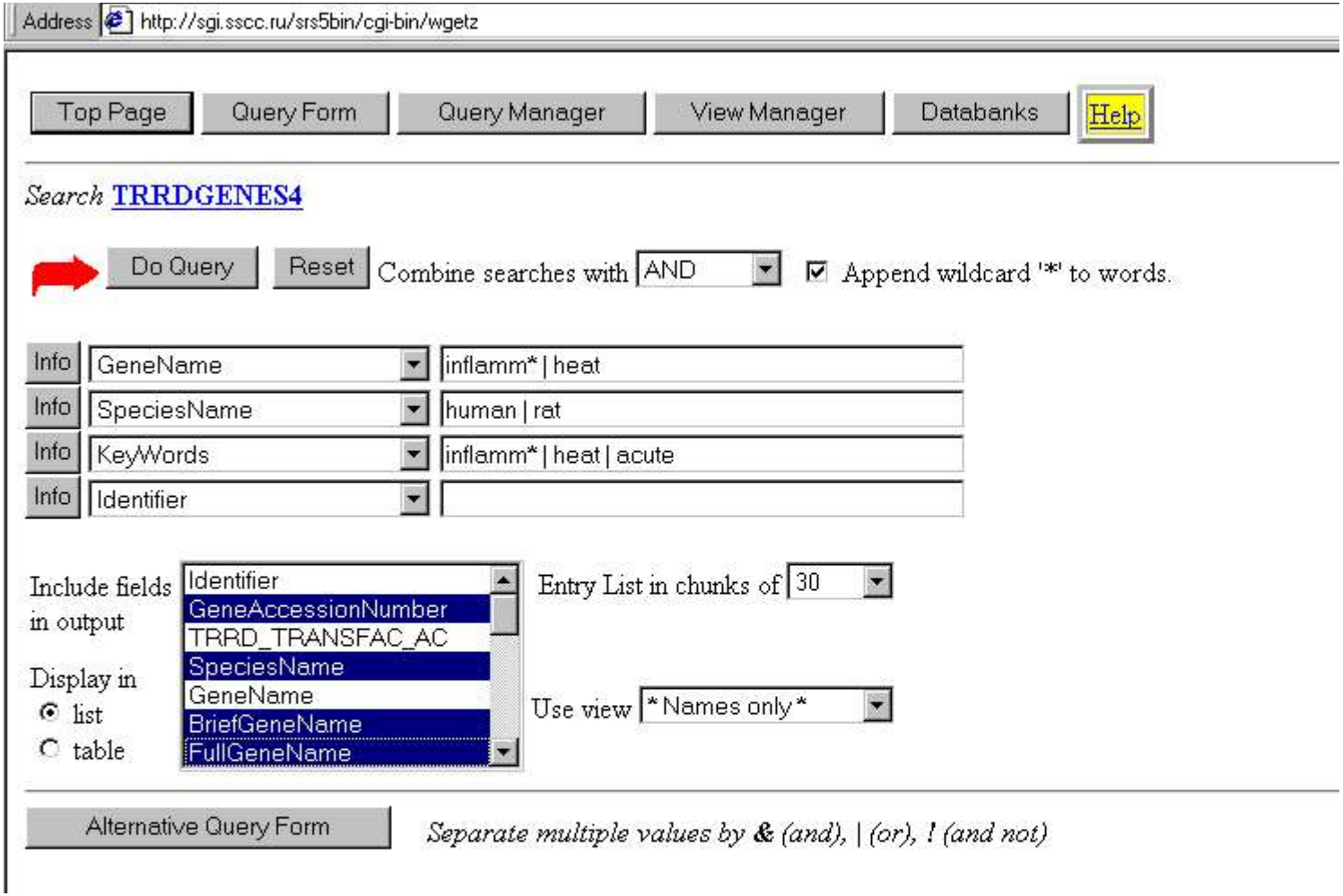
Основные способы интеграции баз данных

- Предоставление доступа к разным базам данных
- Использование гиперссылок между документами различных баз данных
- Связь между базами данных
- Автоматическая генерация запросов по аналогии
- Использование метазнаний для автоматической привязки документов
- Отображение информации в единое семантическое пространство

Примеры систем интеграции молекулярно-генетических данных

SRS - Sequence Retrieval System

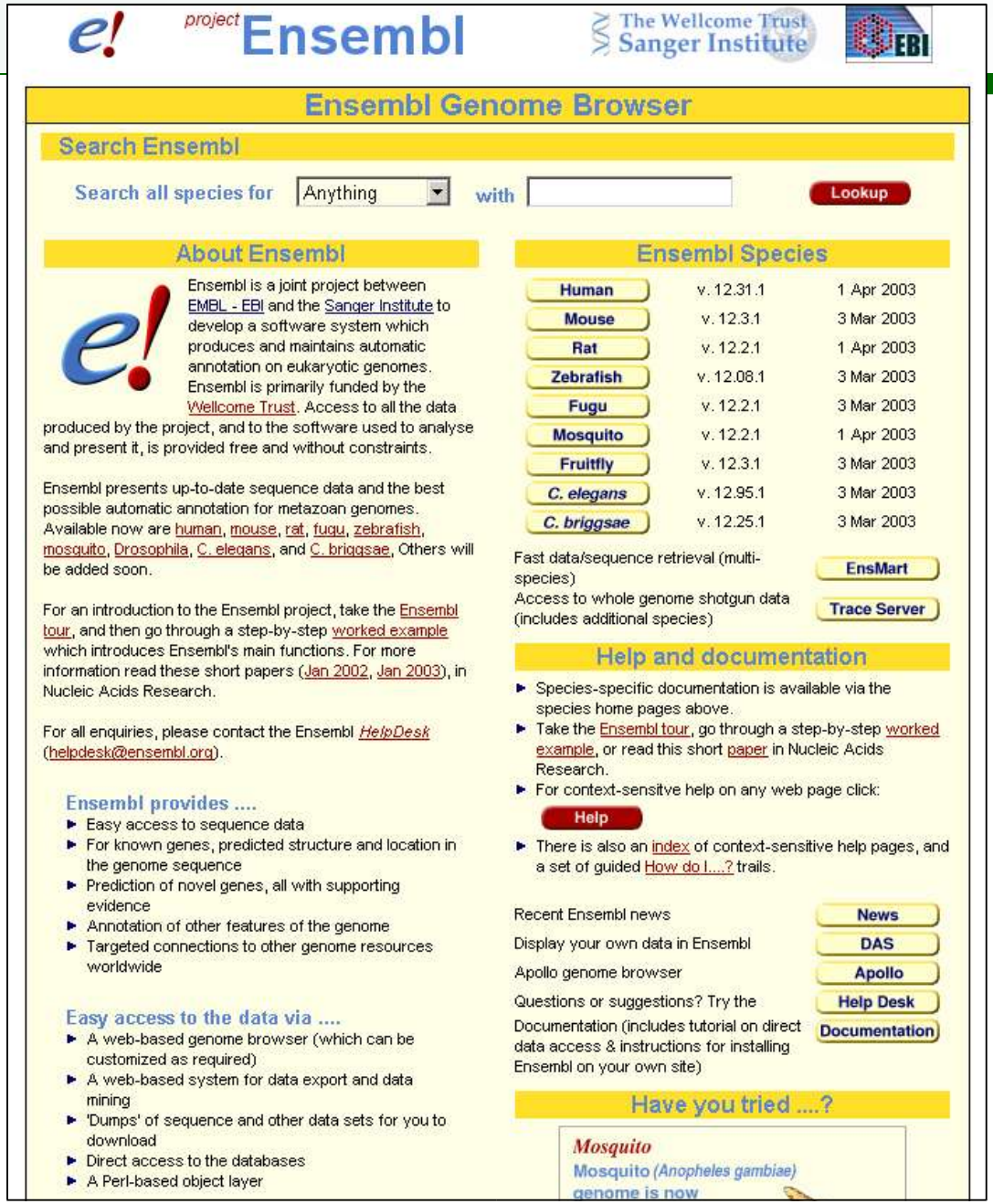
- Сетевой браузер для молекулярно-генетических баз данных
- Основой является объектно-ориентированный язык Icarus на которой описываются структура и синтаксис данных.
- Более 300 основных молекулярно-генетических баз данных уже установлены под SRS.
- Доступ к системе SRS осуществляется через WWW сервер с помощью стандартного браузера.
- Исходные данные представлены в виде текстового флэт-файла.
- Средства синтаксического анализа и индексирования данных
- Гибкие средства преобразования данных
- Стандартное представление молекулярно-генетической информации (например fasta, pir и др.).
- Возможность интеграции с другими базами данных и компьютерными системами для проведения расчетов.


Address
http://sgi.sscc.ru/srs5bin/cgi-bin/wgetz
Top Page
Query Form
Query Manager
View Manager
Databanks
Help
Search TRRDGENES4
Do Query
Reset
Combine searches with
AND
Append wildcard '*' to words.
Info
GeneName
inflamm* | heat
Info
SpeciesName
human | rat
Info
KeyWords
inflamm* | heat | acute
Info
Identifier
Include fields in output
Identifier
GeneAccessionNumber
TRRD_TRANSFAC_AC
SpeciesName
GeneName
BriefGeneName
FullGeneName
Entry List in chunks of
30
Display in
list
table
Use view
* Names only *
Alternative Query Form
Separate multiple values by & (and), | (or), ! (and not)

Address http://sgi.sscc.ru/srs5bin/cgi-bin/wgetz?-id+uUQ71BL3lv+-e+[TRRDGENES4-Identifier:'Hs:GRH']

link save view TRRDGENES4

ID Hs:GRH (TRRD_Viewer, Transcription factors, Gene expression regulation, Bibliography)
DT 30/09/98
AC 00070
GN (TransFac Link)
CR Merkulova T.I.
OS human, Homo sapiens
SN GH
NG growth hormone gene-1
SY GH-N
CG 6.1.5.3.2.
KW hormone, ES-TRRD(Medline,GenBank)
CH 17
RG 5'region
AP REGULATORY UNIT: P00385
PR Promoter; ST; -289 to +1; 213, 1038, 214, 215, 216, 2183, 2184, 2066, 2830,
2067
PQ Site:(2066) -289 to -267; AP-1 bs; activator protein-2 binding site
Site:(2830) -289 to -267; NF-1 bs; nuclear factor-1 binding site
Site:(2067) -266 to -256; USF/MLTF bs; upstream stimulatory factor/major late transcription factor bs
Site:(213) -224 to -208; GRE (1); glucocorticoid responsive element
Site:(2183) -187 to -183; dCRE; distal cAMP response element
CE HOM$C2H2_001; C00038; -139 to -105; 1038, 214;[Lemaigre F.P. et al., 1990]
HOM$HOM_002; C00046; -130 to -65; 214, 215 ;[Lemaigre F.P. et al., 1990]
PQ Site:(1038) -139 to -115; Sp1 bs; Sp1 binding site
Site:(214) -130 to -105; Pit-1 (1); Pit-1 binding element (1)
Site:(2184) -99 to -95; pCRE; proximal cAMP response element
Site:(215) -92 to -66; Pit-1 (2) ; Pit-1 binding element (2)
Site:(216) -32 to -26; TATA box;
RG intron 1
AP REGULATORY UNIT: P00386
PR Enhancer; ST; +87 to +115; 217
PQ Site:(217) +87 to +114; GRE (2); glucocorticoid responsive element
AL Human genes for growth hormone and placental lactogen display 95%
homology [Eliard P.H. et al., 1985]


project Ensembl
The Wellcome Trust Sanger Institute
EBI
Ensembl Genome Browser
Search Ensembl
Search all species for
Anything
with
Lookup
About Ensembl
Ensembl is a joint project between EMBL - EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on eukaryotic genomes. Ensembl is primarily funded by the Wellcome Trust. Access to all the data produced by the project, and to the software used to analyse and present it, is provided free and without constraints.
Ensembl presents up-to-date sequence data and the best possible automatic annotation for metazoan genomes. Available now are human, mouse, rat, fugu, zebrafish, mosquito, Drosophila, C. elegans, and C. briggsae, Others will be added soon.
For an introduction to the Ensembl project, take the Ensembl tour, and then go through a step-by-step worked example which introduces Ensembl's main functions. For more information read these short papers (Jan 2002, Jan 2003), in Nucleic Acids Research.
For all enquiries, please contact the Ensembl HelpDesk (helpdesk@ensembl.org).
Ensembl provides ....
Easy access to sequence data
For known genes, predicted structure and location in the genome sequence
Prediction of novel genes, all with supporting evidence
Annotation of other features of the genome
Targeted connections to other genome resources worldwide
Easy access to the data via ....
A web-based genome browser (which can be customized as required)
A web-based system for data export and data mining
'Dumps' of sequence and other data sets for you to download
Direct access to the databases
A Perl-based object layer
Ensembl Species
Human v. 12.31.1 1 Apr 2003
Mouse v. 12.3.1 3 Mar 2003
Rat v. 12.2.1 1 Apr 2003
Zebrafish v. 12.08.1 3 Mar 2003
Fugu v. 12.2.1 3 Mar 2003
Mosquito v. 12.2.1 1 Apr 2003
Fruitfly v. 12.3.1 3 Mar 2003
C. elegans v. 12.95.1 3 Mar 2003
C. briggsae v. 12.25.1 3 Mar 2003
Fast data/sequence retrieval (multi-species)
EnsMart
Access to whole genome shotgun data (includes additional species)
Trace Server
Help and documentation
Species-specific documentation is available via the species home pages above.
Take the Ensembl tour, go through a step-by-step worked example, or read this short paper in Nucleic Acids Research.
For context-sensitve help on any web page click:
Help
There is also an index of context-sensitive help pages, and a set of guided How do I...? trails.
Recent Ensembl news
News
Display your own data in Ensembl
DAS
Apollo genome browser
Apollo
Questions or suggestions? Try the
Help Desk
Documentation (includes tutorial on direct data access & instructions for installing Ensembl on your own site)
Documentation
Have you tried ....?
Mosquito
Mosquito (Anopheles gambiae) genome is now

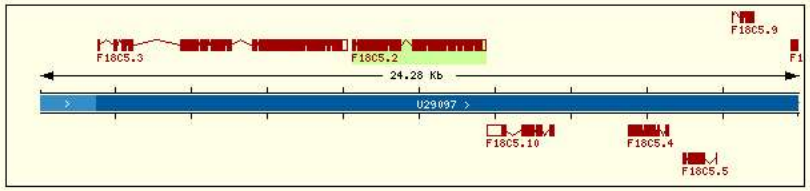
Ensembl C. elegans GeneView

The Wellcome Trust Sanger Institute | EBI

Home ▸ C.elegans ▲ What's New ▲ BLAST ▲ SSAHA ▲ MartView ▲ Export Data ▲ Download ▲ Docs ▲

Find Gene F18C5.2 Lookup [e.g. M6.1, W06F12.2] Help

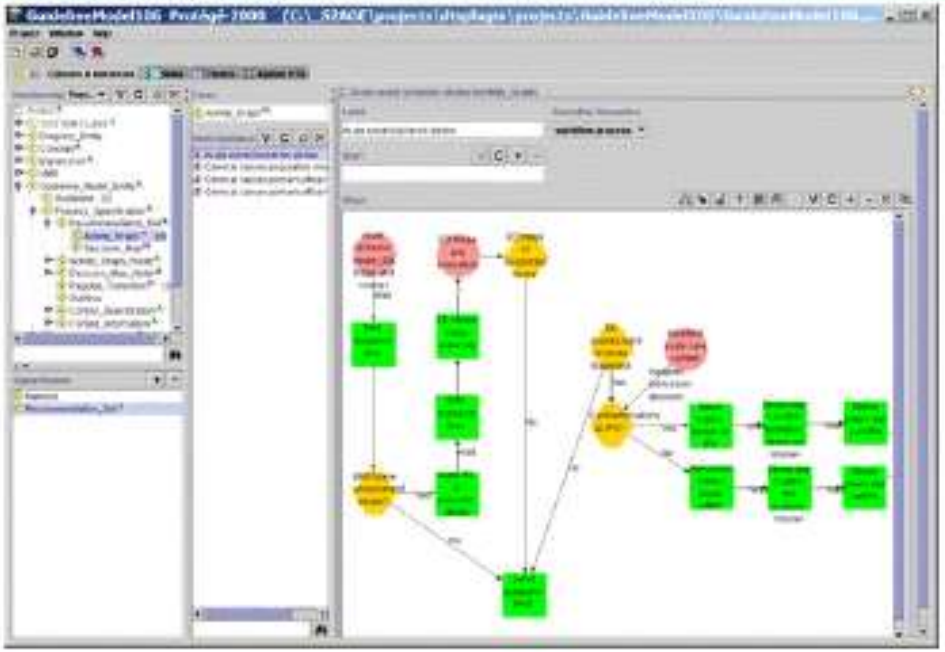
Ensembl Gene Report

Gene	**F18C5.2** (wormbase_gene ID)
Ensembl Gene ID	**F18C5.2**
Genomic Location	**View gene in genomic location:** 6554916 - 6559196 bp (6.6 Mb) on chromosome II **This gene is located in sequence:** U29097.1.1.29095
Description	HUMAN WRN (WERNER'S SYNDROME) RELATED PROTEIN 1. [Source: SPTREMBL (AAA68410)]
Prediction Method	This gene was annotated by Wormbase through a process of automatic and manual curation.
Predicted Transcripts	1: F18C5.2 - [View transcript info] [View exon info] [View protein info] (F18C5.2) **Transcript Neighbourhood** 
Homology Matches	**These gene(s) have been identified as putative homologues by reciprocal BLAST analysis:** ***Caenorhabditis briggsae*** ENSCBRG00000006152 (CBG02689) No description
Export Data	Export gene data in EMBL, GenBank or FASTA

Transcripts/Translation Summary

F18C5.2	**Stable ID:** F18C5.2 **Exons:** 16 **Transcript length:** 3281 bp **Translation length:** 1057 residues [View transcript information] [View exon information] [View protein information]

Welcome to the Protégé Project



Protégé-2000 is an ontology editor and a knowledge-base editor.

Protégé-2000 is also an open-source, Java tool which provides an extensible architecture for the creation of customized knowledge-based tools.

Release 1.7	April 10, 2002
Beta 1.8	November 12, 2002

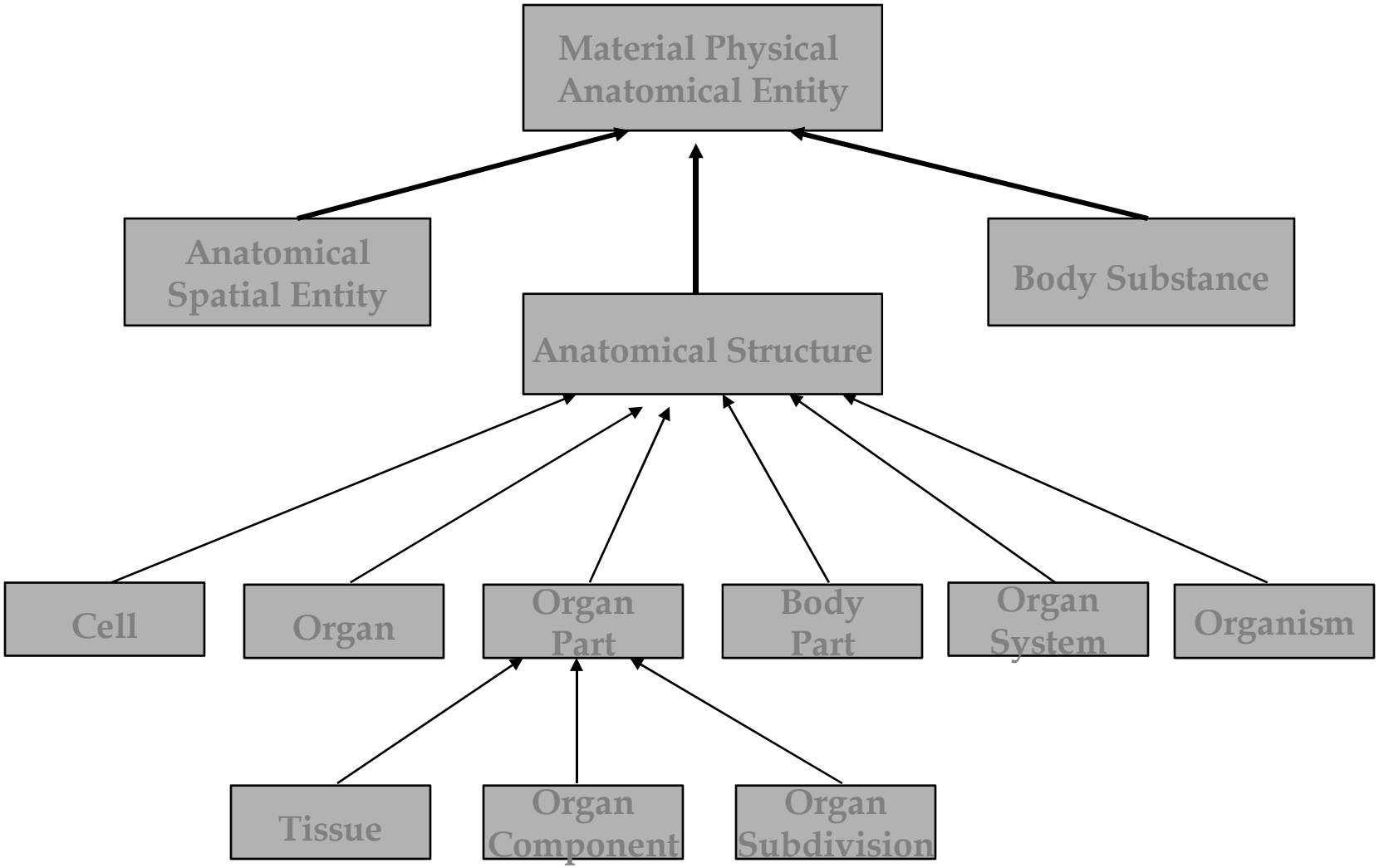
Protégé Community Statistics	
Registered Users	5554
users list members	3214
discussion list members	1111
discussion list messages	3053
Plug-ins	35

Updated November 15, 2002

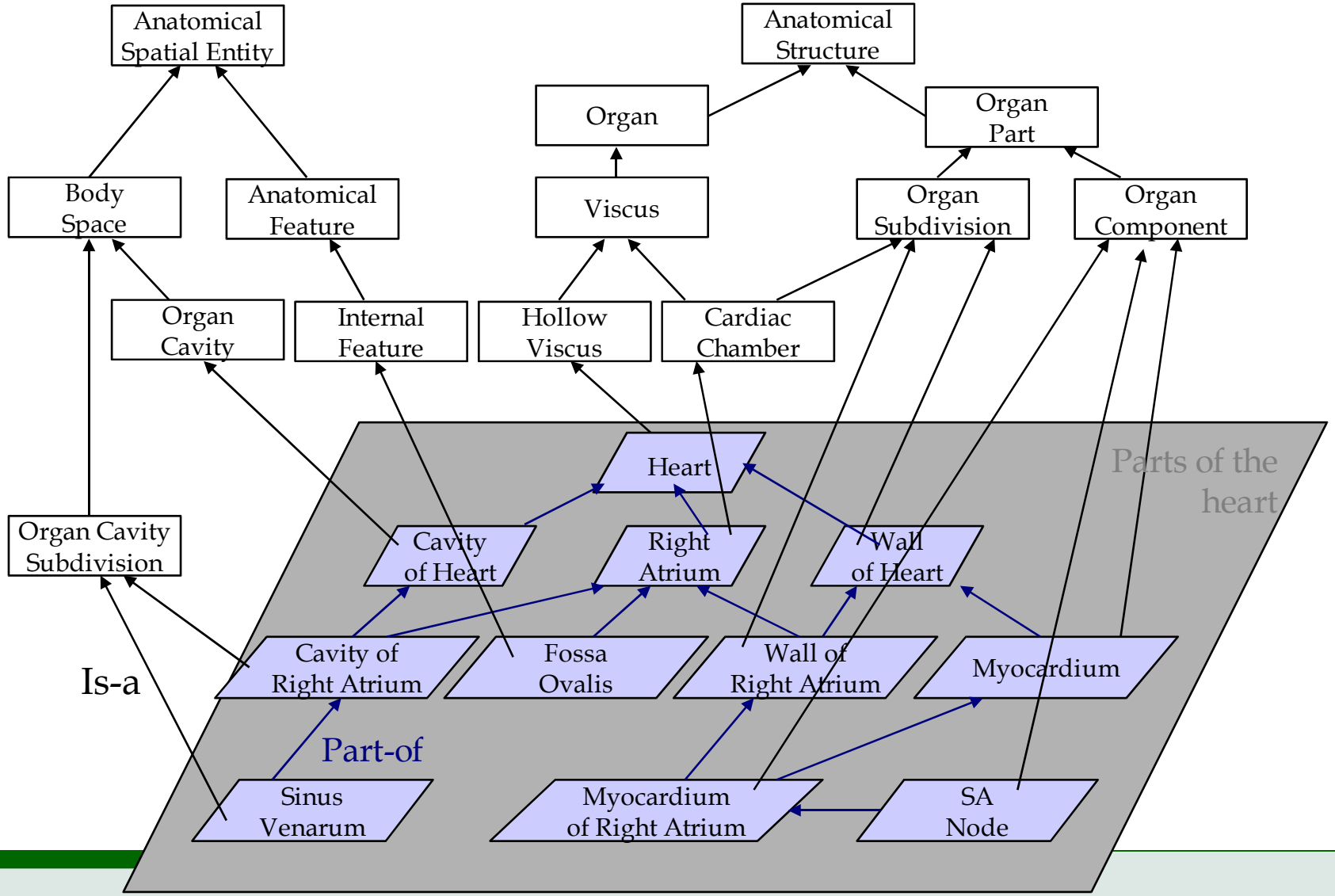
What Protégé-2000 offers

1. Ontology editor for modeling of concepts, attributes, and relationships
2. Automated generation of tools that instantiate concepts defined in ontologies to build knowledge bases
3. Systems to visualize both ontologies and knowledge bases
4. Ability to archive ontologies and knowledge bases in a variety of formats, such as ODBC, XML, RDF, DAML+OIL
5. Lots of user-contributed “plug ins”
6. A world-wide community of active users

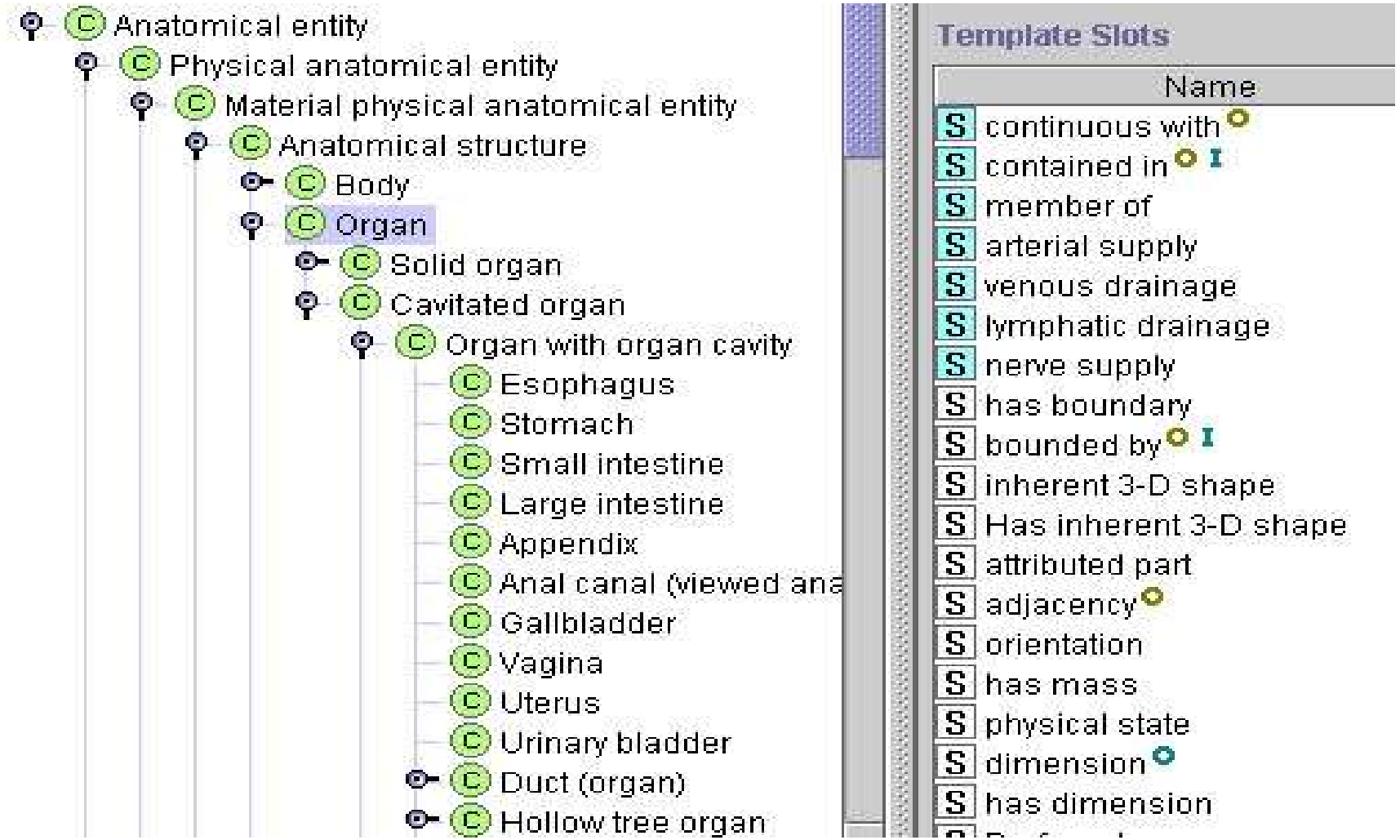
Digital Anatomist Foundational Model of Anatomy

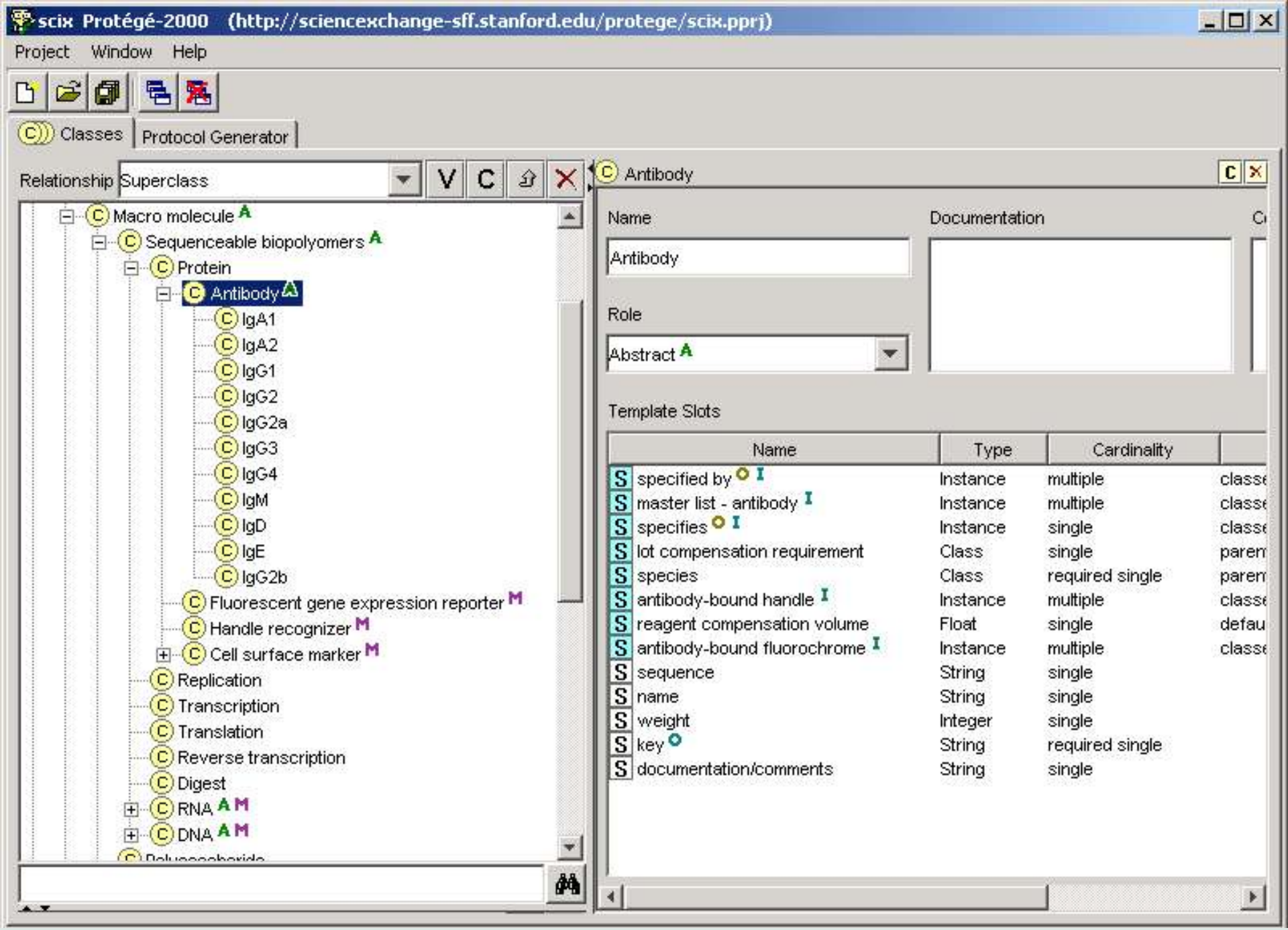


Classes of anatomical structures



The Ontology in Protégé-2000

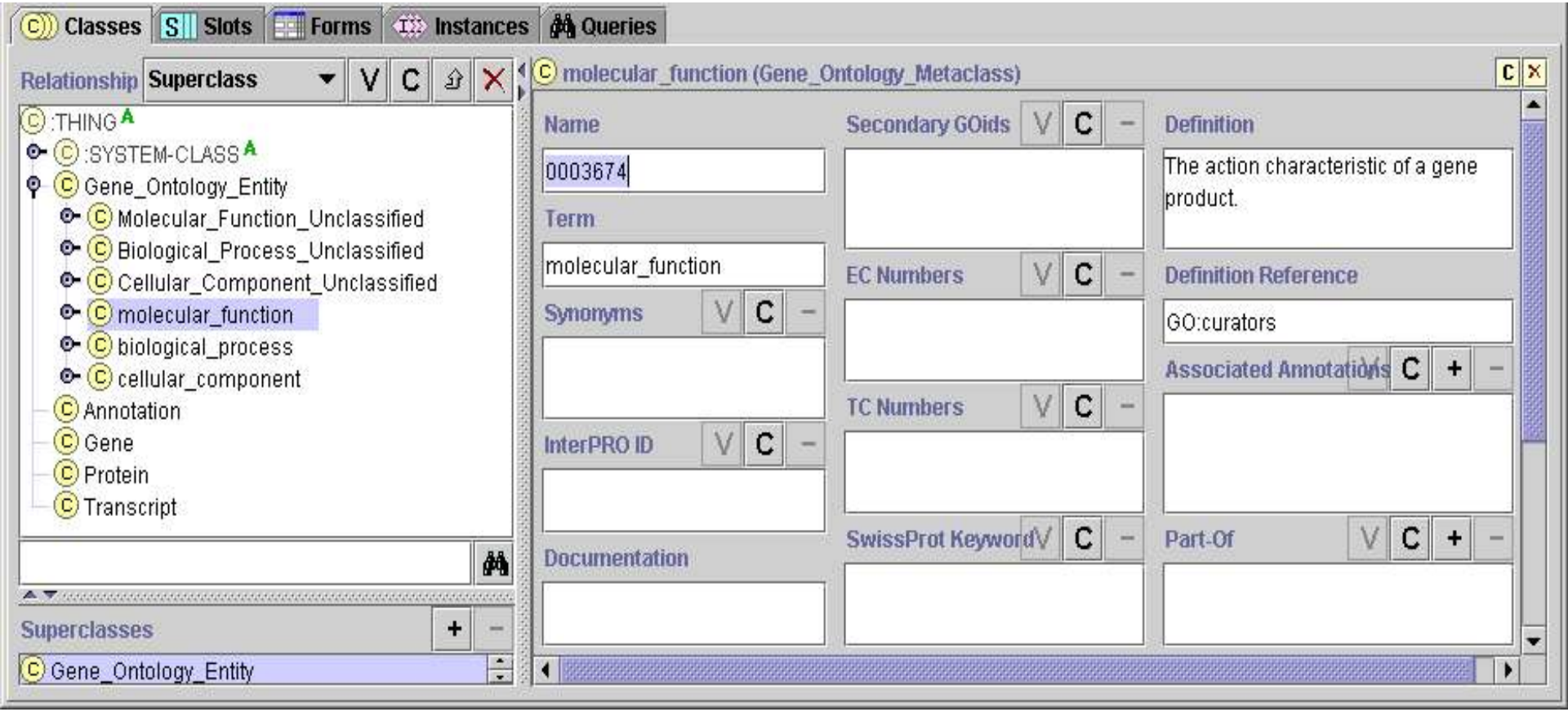



scix Protégé-2000 (http://sciencexchange-sff.stanford.edu/protege/scix.pprj)
Project Window Help
Classes
Protocol Generator
Relationship Superclass
Macro molecule A
Sequenceable biopolyomers A
Protein
Antibody A
IgA1
IgA2
IgG1
IgG2
IgG2a
IgG3
IgG4
IgM
IgD
IgE
IgG2b
Fluorescent gene expression reporter M
Handle recognizer M
Cell surface marker M
Replication
Transcription
Translation
Reverse transcription
Digest
RNA A M
DNA A M
Antibody
Name
Antibody
Documentation
Role
Abstract A
Template Slots
Name
Type
Cardinality
specified by I
Instance
multiple
master list - antibody I
Instance
multiple
specifies I
Instance
single
lot compensation requirement
Class
single
species
Class
required single
antibody-bound handle I
Instance
multiple
reagent compensation volume
Float
single
antibody-bound fluorochrome I
Instance
multiple
sequence
String
single
name
String
single
weight
Integer
single
key
String
required single
documentation/comments
String
single

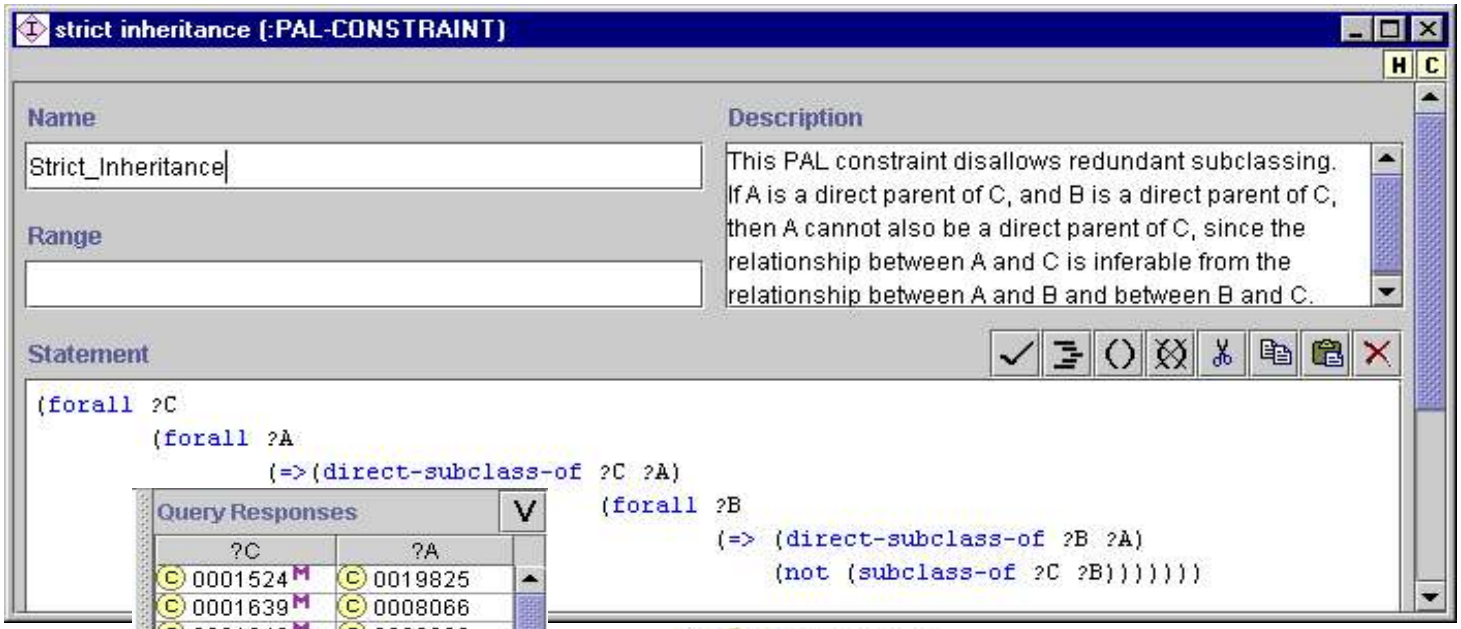
Additional Protégé features that aid building ontologies

- Metaclasses allow developers to define special-purpose base classes that are “instances” of the metaclasses
- Protégé axiom language (PAL) allows developers to specify complex semantic constraints using logic

Metaclasses allow developers to define new template slots
(e.g., as in Gene Ontology)

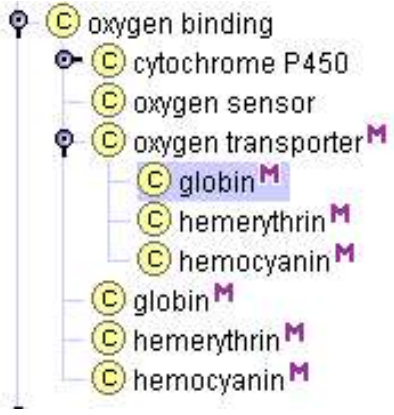


Evaluation of constraints can point out semantic errors



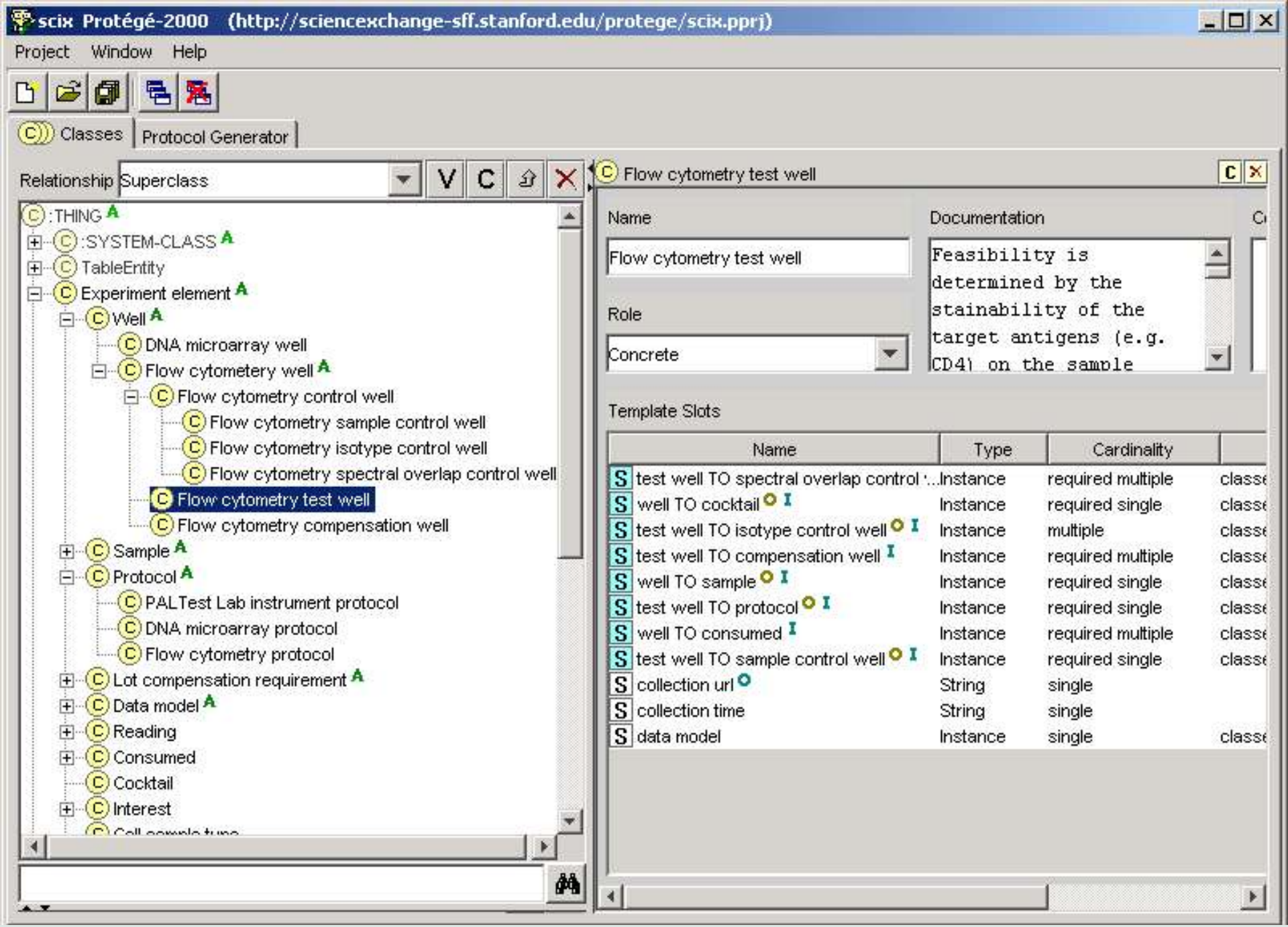
Query Responses

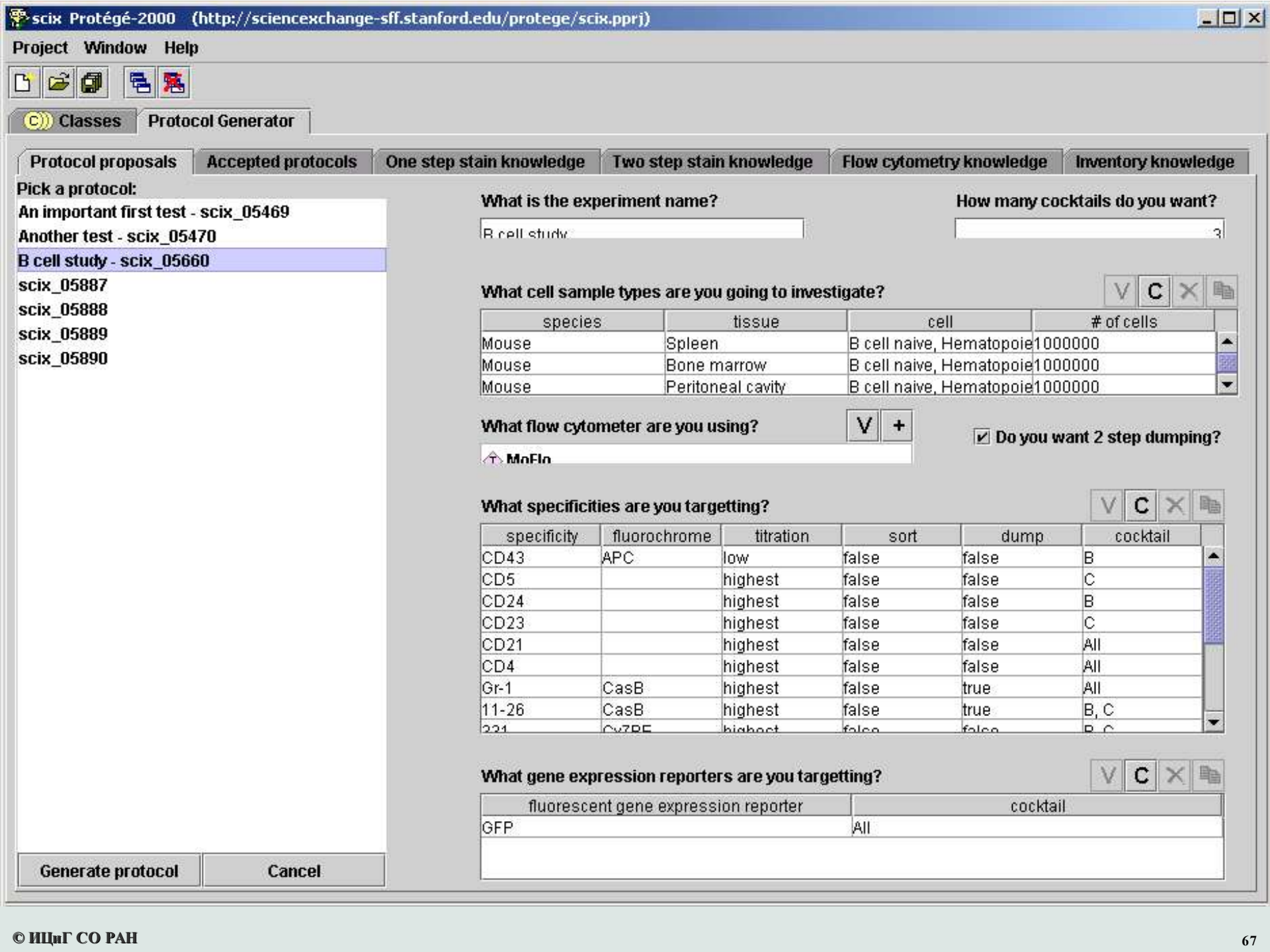
?C	?A
0001524 M	0019825
0001639 M	0008066
0001640 M	0008066
0003850 M	0016302
0004001 M	0016301
0004001 M	0016773
0004005 M	0005386
0004005 M	0016820
0004005 M	0016887 M
0004012 M	0016887 M
0004017 M	0016301
0004054 M	0016301
0004072 M	0016301
0004127 M	0016301
0004127 M	0016773
0004136 M	0016301
0004136 M	0016773
0004137 M	0016301
0004137 M	0016773
0004138 M	0016301



2. Automated generation of tools for building intelligent systems

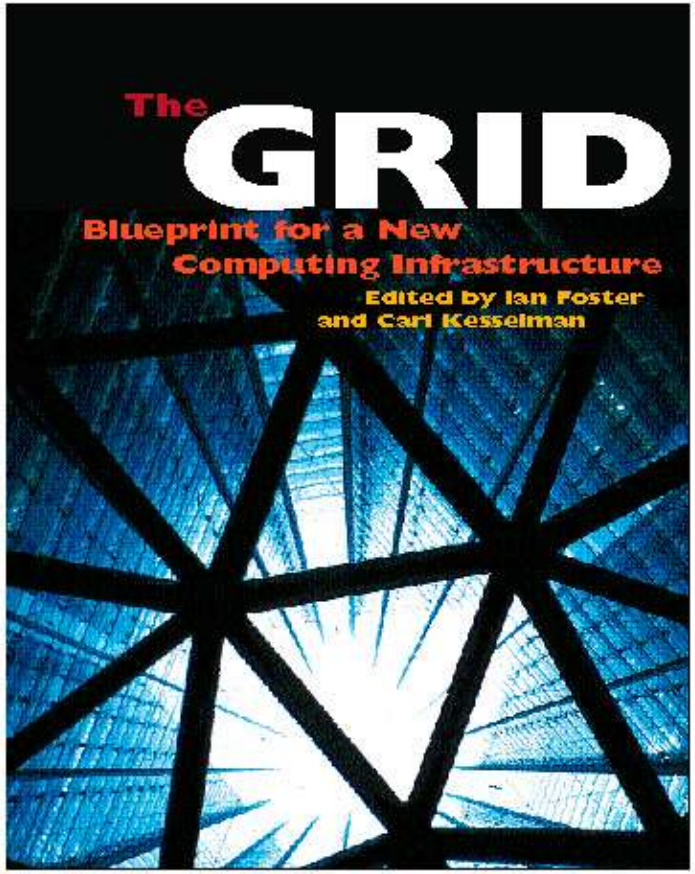
- From the beginning, Protégé was built as part of a comprehensive methodology for developing intelligent systems
- Ontologies in Protégé guide the acquisition of content knowledge from subject-matter experts
- Protégé generates automatically tools to acquire content knowledge and to automate problem solving


scix Protégé-2000 (http://sciencexchange-sff.stanford.edu/protege/scix.pprj)
Project Window Help
Classes
Protocol Generator
Relationship Superclass
:THING
:SYSTEM-CLASS
TableEntity
Experiment element
Well
DNA microarray well
Flow cytometery well
Flow cytometry control well
Flow cytometry sample control well
Flow cytometry isotype control well
Flow cytometry spectral overlap control well
Flow cytometry test well
Flow cytometry compensation well
Sample
Protocol
PALTest Lab instrument protocol
DNA microarray protocol
Flow cytometry protocol
Lot compensation requirement
Data model
Reading
Consumed
Cocktail
Interest
Flow cytometry test well
Name
Flow cytometry test well
Documentation
Feasibility is determined by the stainability of the target antigens (e.g. CD4) on the sample
Role
Concrete
Template Slots
Name
Type
Cardinality
test well TO spectral overlap control ... Instance required multiple
well TO cocktail Instance required single
test well TO isotype control well Instance multiple
test well TO compensation well Instance required multiple
well TO sample Instance required single
test well TO protocol Instance required single
well TO consumed Instance required multiple
test well TO sample control well Instance required single
collection url String single
collection time String single
data model Instance single


scix Protégé-2000 (http://sciencexchange-sff.stanford.edu/protege/scix.pprj)
Project Window Help
Classes
Protocol Generator
Protocol proposals
Accepted protocols
One step stain knowledge
Two step stain knowledge
Flow cytometry knowledge
Inventory knowledge
Pick a protocol:
An important first test - scix_05469
Another test - scix_05470
B cell study - scix_05660
scix_05887
scix_05888
scix_05889
scix_05890
Generate protocol
Cancel
What is the experiment name?
B cell study
How many cocktails do you want?
3
What cell sample types are you going to investigate?
species | tissue | cell | # of cells
Mouse | Spleen | B cell naive, Hematopoie | 1000000
Mouse | Bone marrow | B cell naive, Hematopoie | 1000000
Mouse | Peritoneal cavity | B cell naive, Hematopoie | 1000000
What flow cytometer are you using?
MoFlo
Do you want 2 step dumping?
What specificities are you targetting?
specificity | fluorochrome | titration | sort | dump | cocktail
CD43 | APC | low | false | false | B
CD5 | | highest | false | false | C
CD24 | | highest | false | false | B
CD23 | | highest | false | false | C
CD21 | | highest | false | false | All
CD4 | | highest | false | false | All
Gr-1 | CasB | highest | false | true | All
11-26 | CasB | highest | false | true | B, C
What gene expression reporters are you targetting?
fluorescent gene expression reporter | cocktail
GFP | All

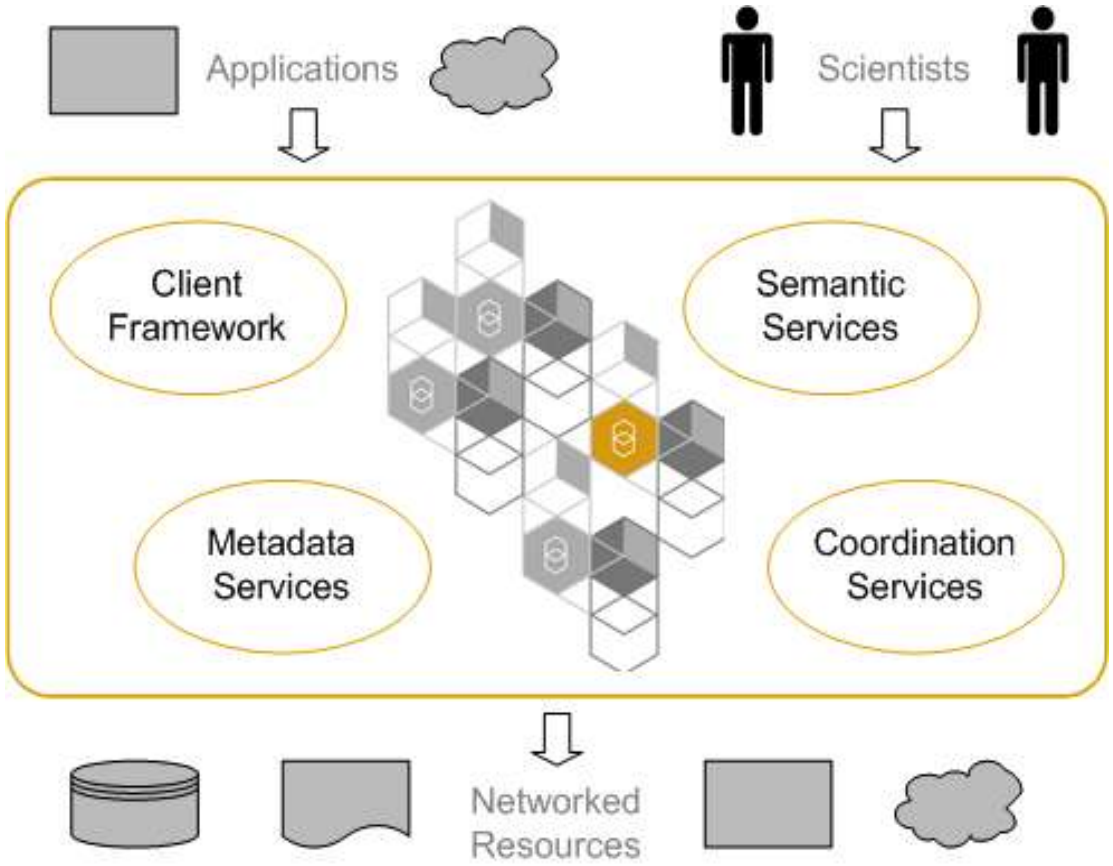
Five Emerging Models of Networked Computing From *The Grid*

- Distributed Computing
  - || synchronous processing
- High-Throughput Computing
  - || asynchronous processing
- On-Demand Computing
  - || dynamic resources
- Data-Intensive Computing
  - || databases
- Collaborative Computing
  - || scientists

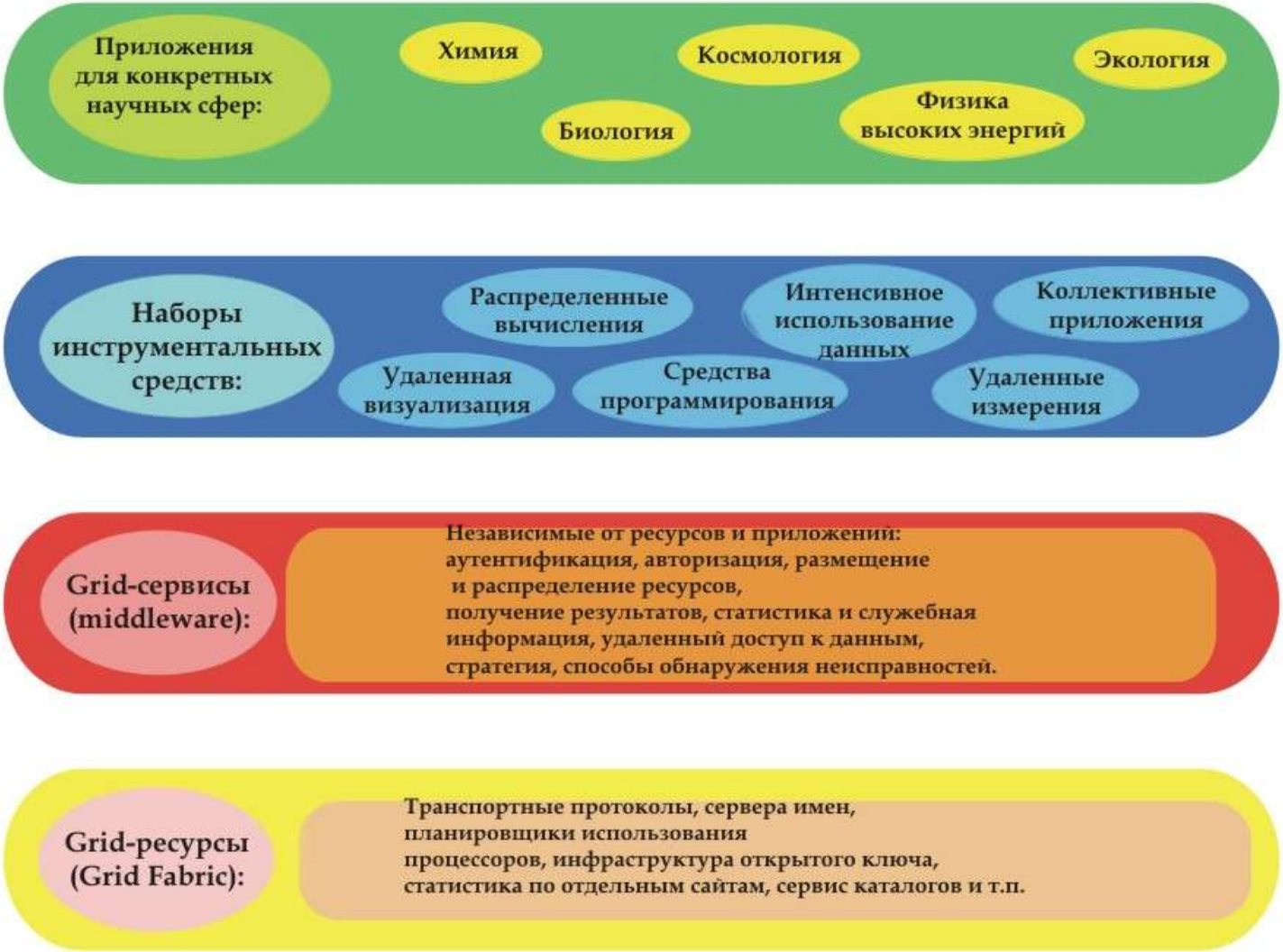


Ian Foster and Carl Kesselman, editors, "The Grid: Blueprint for a New Computing Infrastructure," Morgan Kaufmann, 1999, http://www.mkp.com/grids

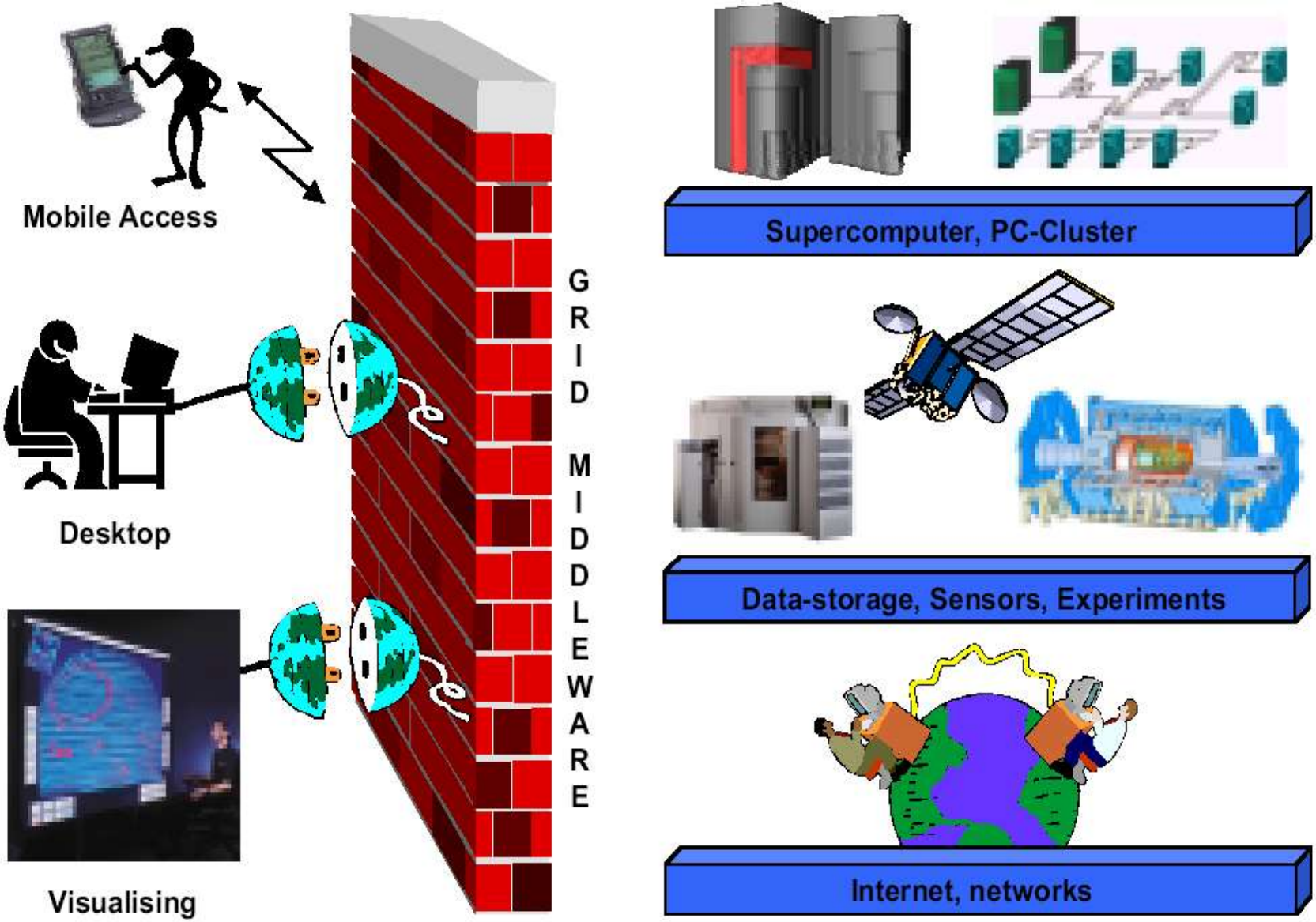
myGrid as a collection of services

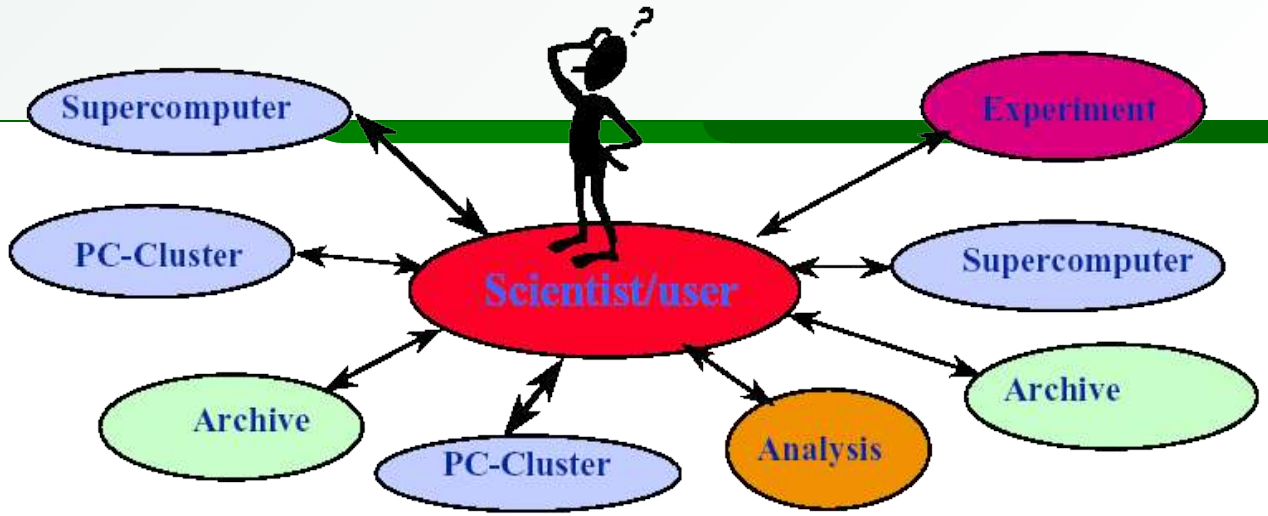


Grid – архитектура с точки зрения программного обеспечения



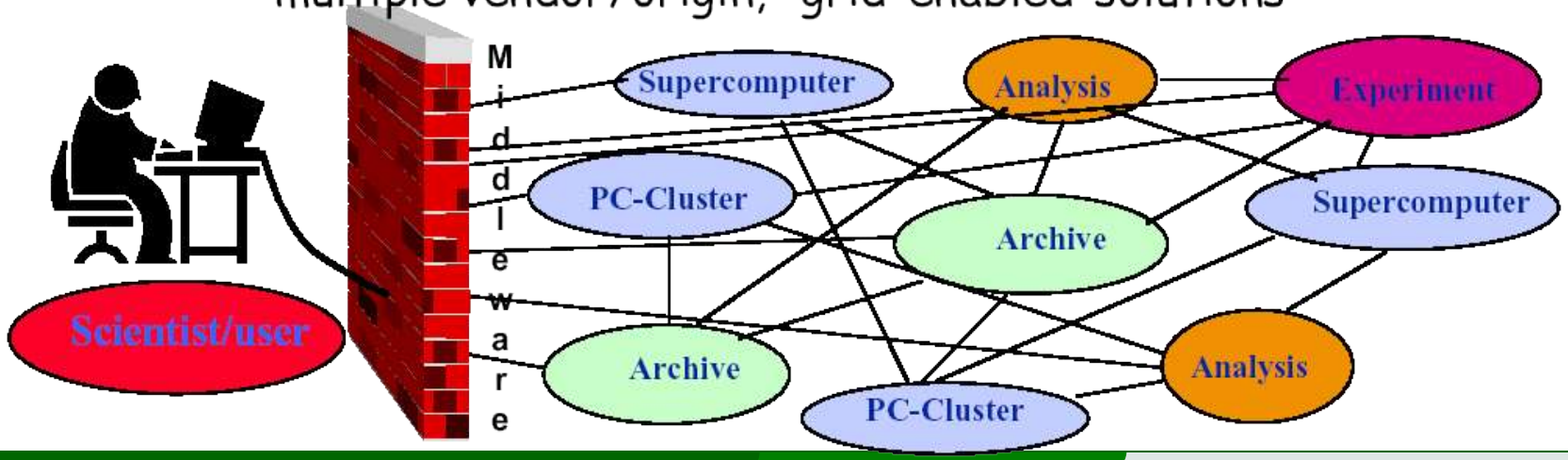
Grids


Mobile Access
Desktop
Visualising
GRID MIDDLEWARE
Supercomputer, PC-Cluster
Data-storage, Sensors, Experiments
Internet, networks
Hoffmann, Reinefeld, Putzer



Often monolithic, "vertical", proprietary solutions

Through open, standard interfaces: flexible, adaptable, interchangeable, multiple vendor/origin, "grid-enabled"solutions



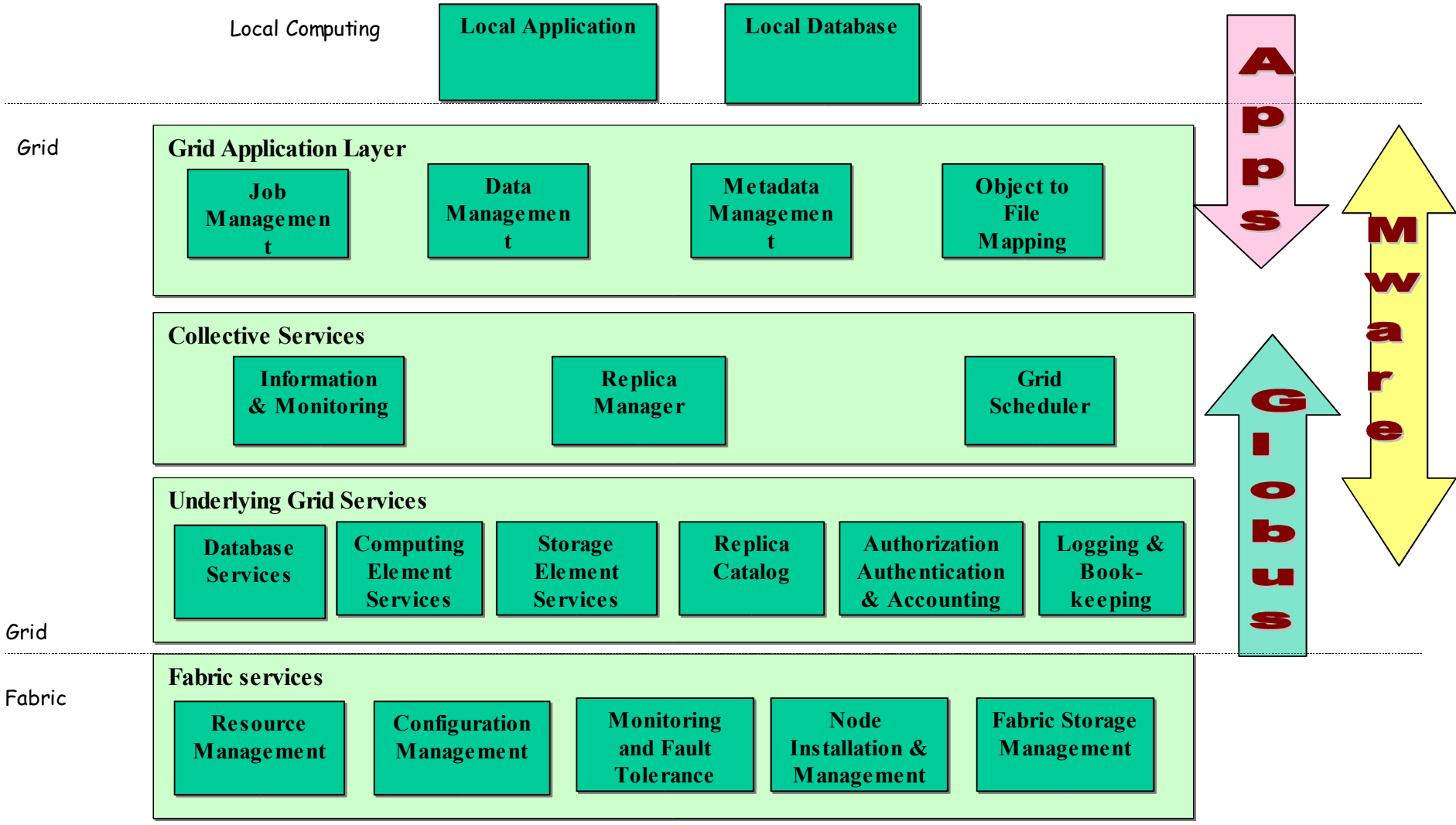
Некоторые Grid Проекты

Name	URL/Sponsor	Focus
European Union (EU) DataGrid	**www.eu-datagrid.org** European Union	**Создание реальной Grid для различных приложений в области Физики Высоких Энергий, Биоинформатики** и ООС.
EU DataTAG Project	www.datatag.org	**Interoperability between European and US Grids**
CrossGrid	European Union	
EuroGrid, Grid Interoperability (GRIP)	www.eurogrid.org European Union	Создание технологий для удалённого доступа к суперкомпьютерам и их приложениям
Globus Project™	**globus.org DARPA, DOE, NSF, NASA, Msoft**	**Исследование в области Grid технологий; создание и тех. поддержка Globus Toolkit™; приложения.**
GridPP	**gridpp.ac.uk U.K. eScience**	**Создание реальной Grid в Англии для исследований в области Физики Элементарных Частиц.**

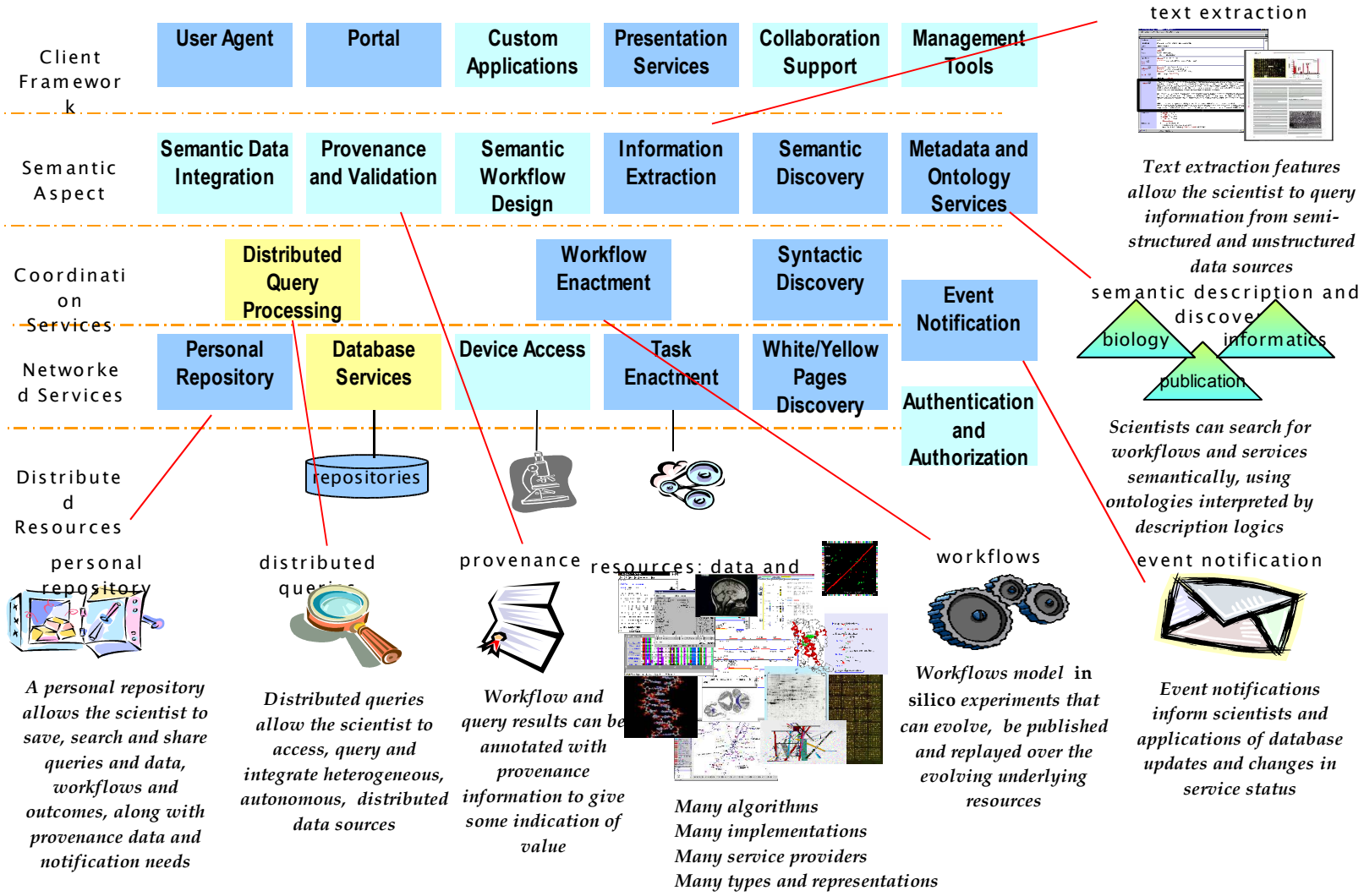
Некоторые Grid Проекты

Name	URL/Sponsor	Focus
Grid Physics Network	**griphyn.org** **NSF**	**Создание технологий для анализа данных в физике: ATLAS, CMS, LIGO, SDSS**
International Virtual Data Grid Laboratory	ivdgl.org NSF	Создание реальной международной Grid для экспериментов над Grid технологиями и приложениями
TeraGrid	**teragrid.org** **NSF**	**Научная инфраструктура в США, связывающая 4 организации 40 Gb/s**
Particle Physics Data Grid	**ppdg.net** **DOE Science**	**Создание реальной Grid для анализа данных в Физике Высоких Энергий и Ядерной физике**

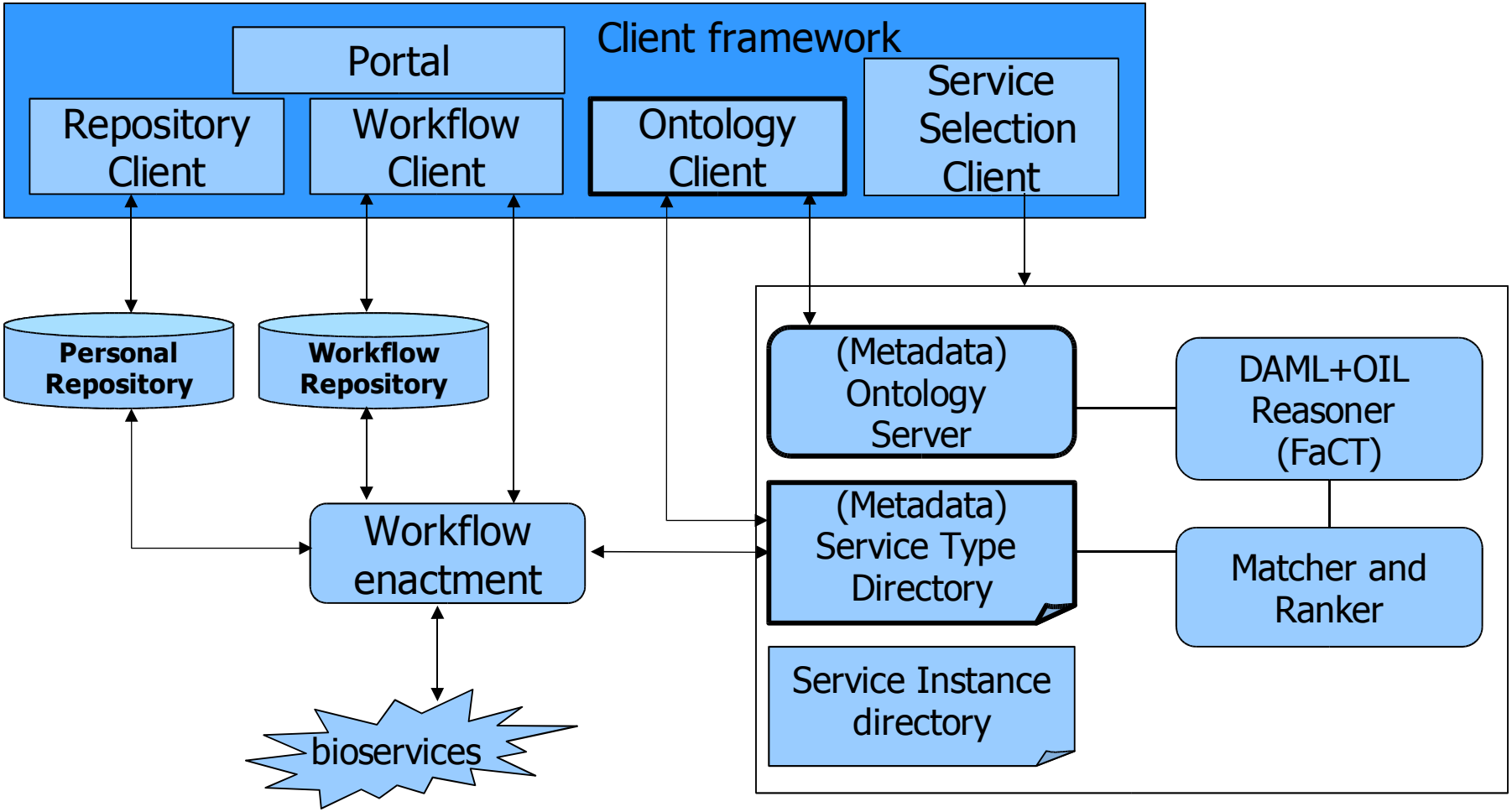
DataGrid Architecture



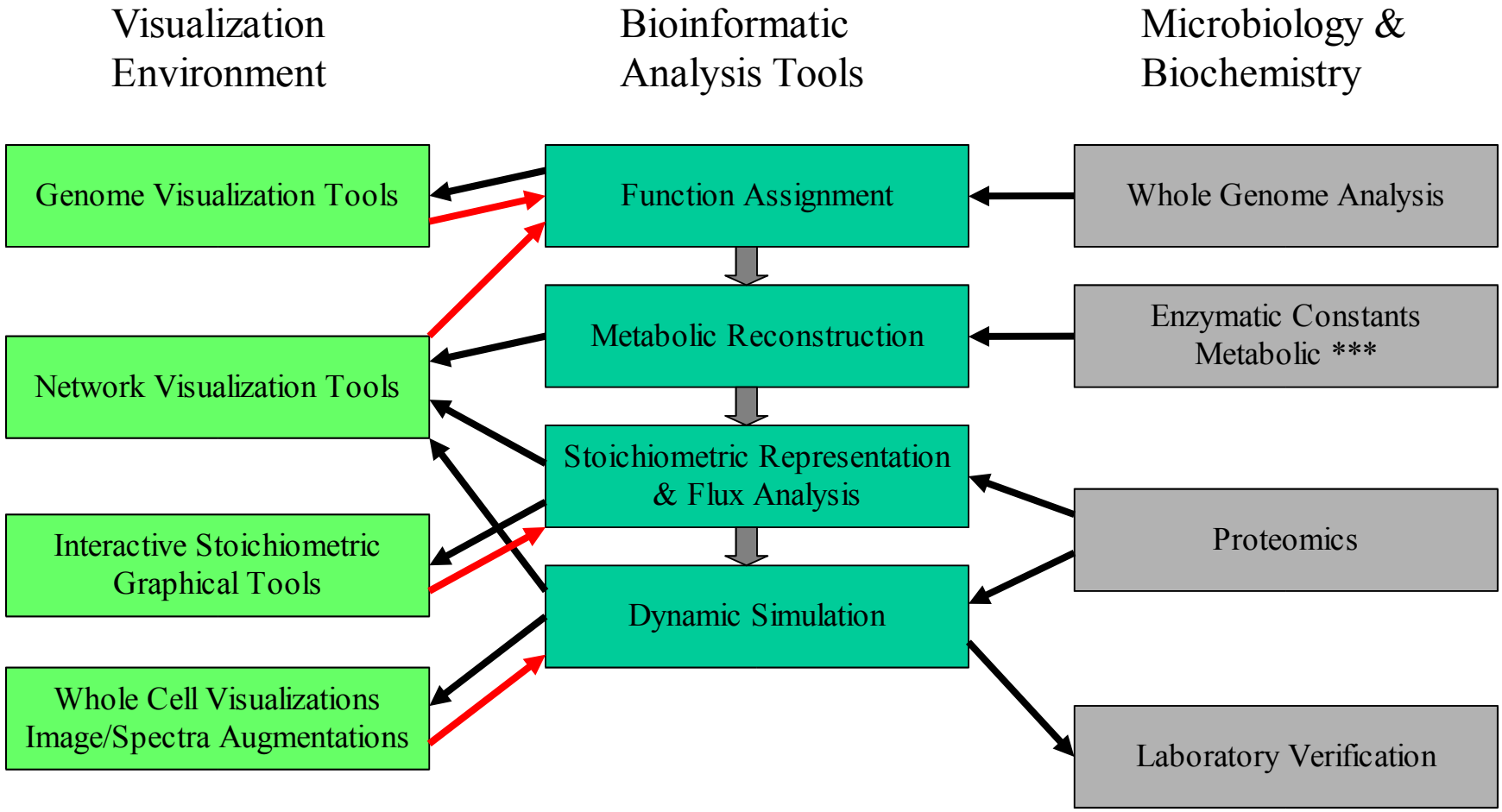
myGrid layered services



myGrid 0.0, 10.02



Visualization + Bioinformatics



Data Grids, Data Mining & Data Webs

	Data Grid	Distributed Data Mining	Data Web
Goal	distributed computation	distributed data mining	data explor. & mining
Services	authorization, security, resources	building models, transforming data, etc.	publishing, merging, & correlating columns
Protocol	TCP, GridFTP	TCP	DWTP, …
Platform	dist. clusters	server	dist. cluster

Semantic Web vs. Data Web

	Document Web	Semantic Web	Data Web
Protocol	HTTP	HTTP, SOAP	DWTP, SOAP
Languages	HTML, XML	XML, RDF	XML, PMML …
Action	keyword search	RDF inferences	correlate and mine
Platform	server	server	server, cluster