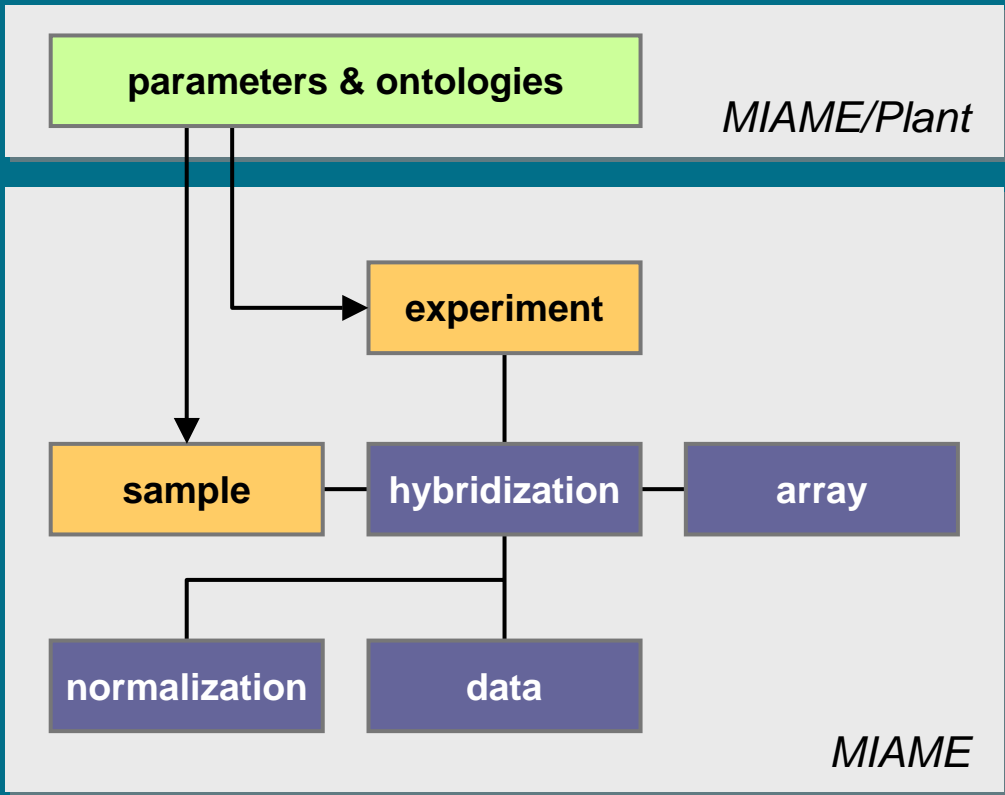


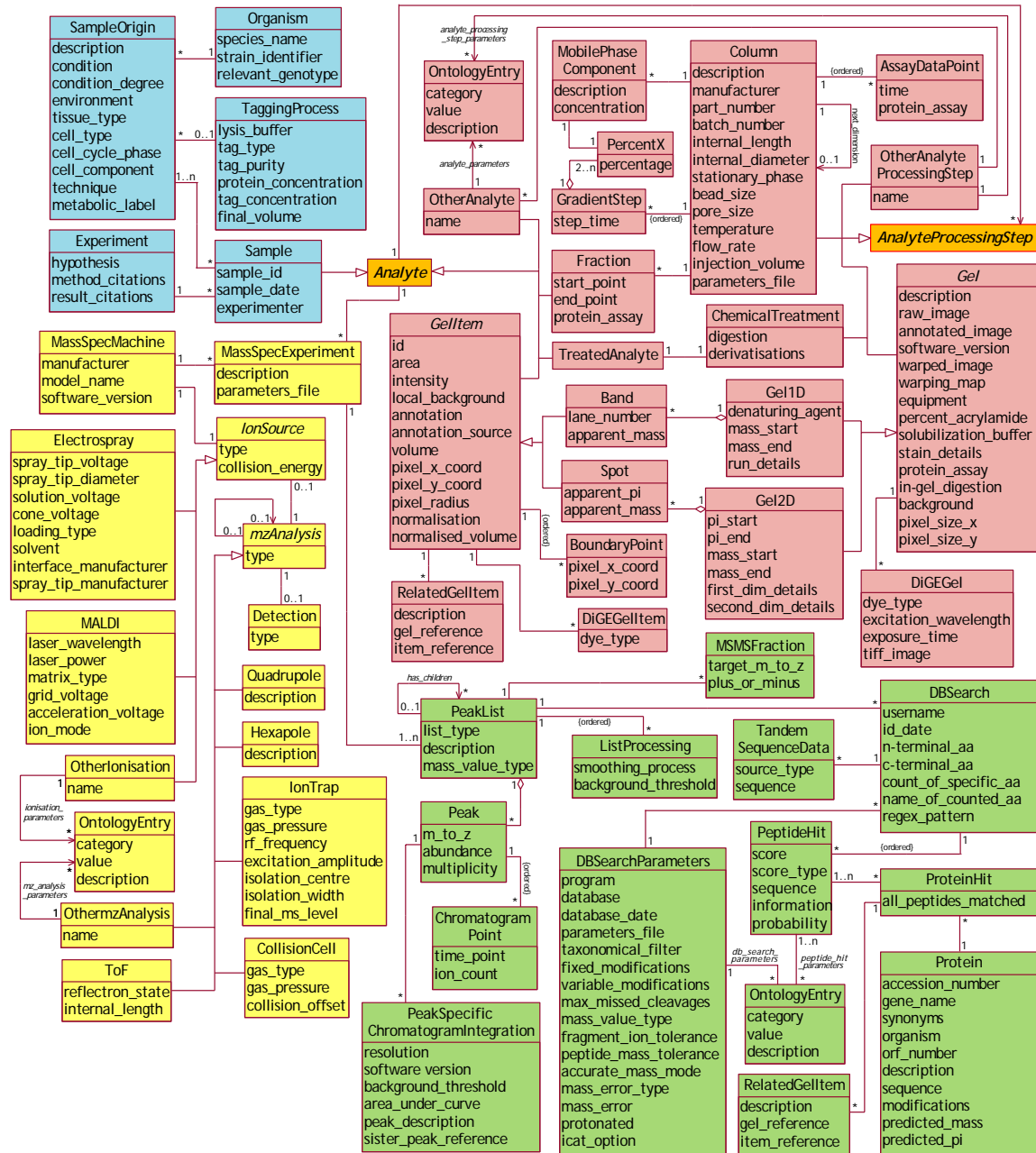
Bioinformatics & experimental practice in proteomics.

Tom Burdick

Perfection (in design) is achieved not when there is nothing more to add, but rather when there is nothing more to take away.¹

1. The Cathedral and the Bazaar, Eric Steven Raymond





Sample Preparation Technologies

Affinity Depletion/Enrichment



Antibody

Lectin

Chemistry

Chromatography



IEX (SCX, AEC)

SEC

RP

Chemical Labeling

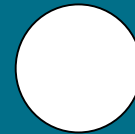


cICAT

iTRAQ

^{18}O

Chemical Modification



Chemical

Enzymatic

Chemoenzymatic

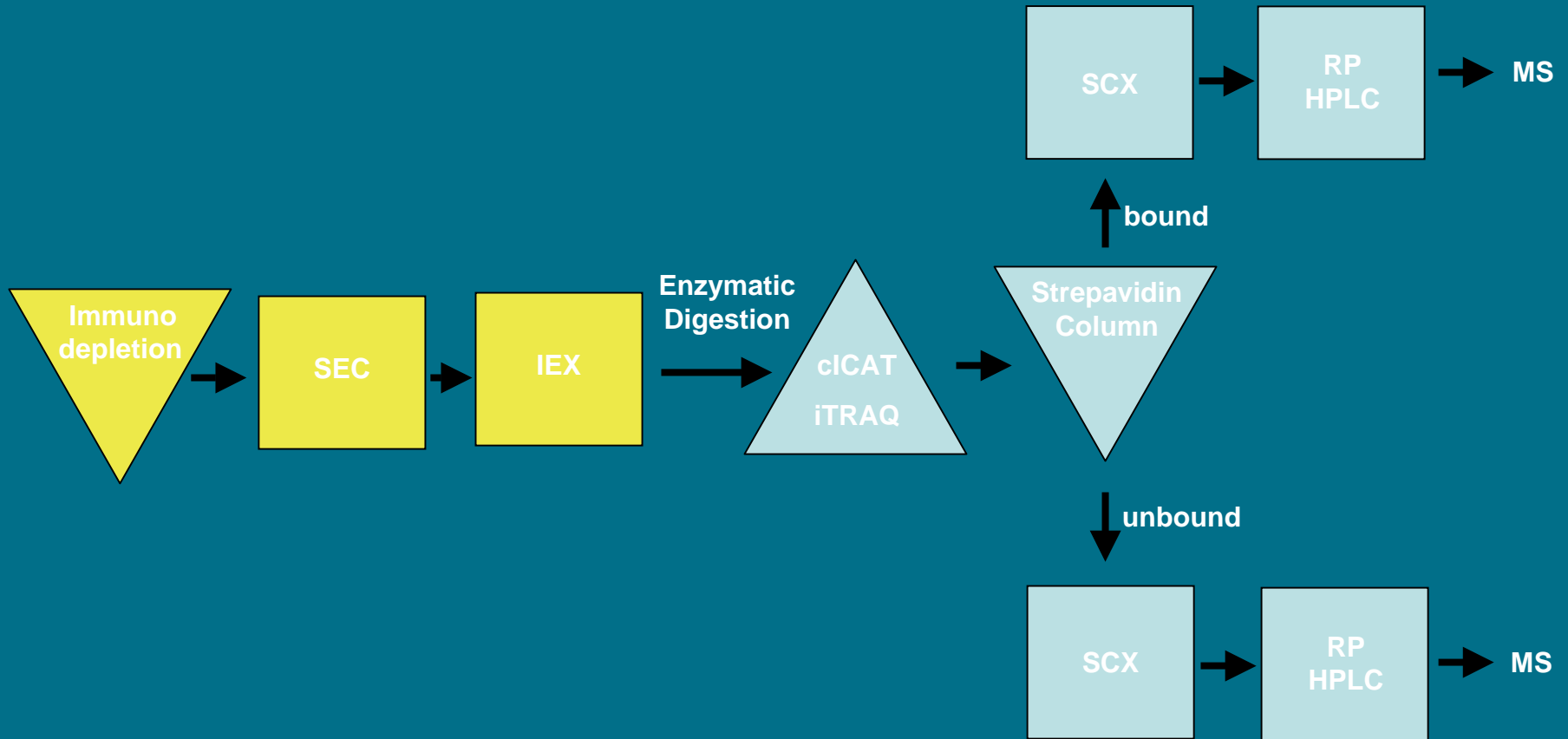
Proteins



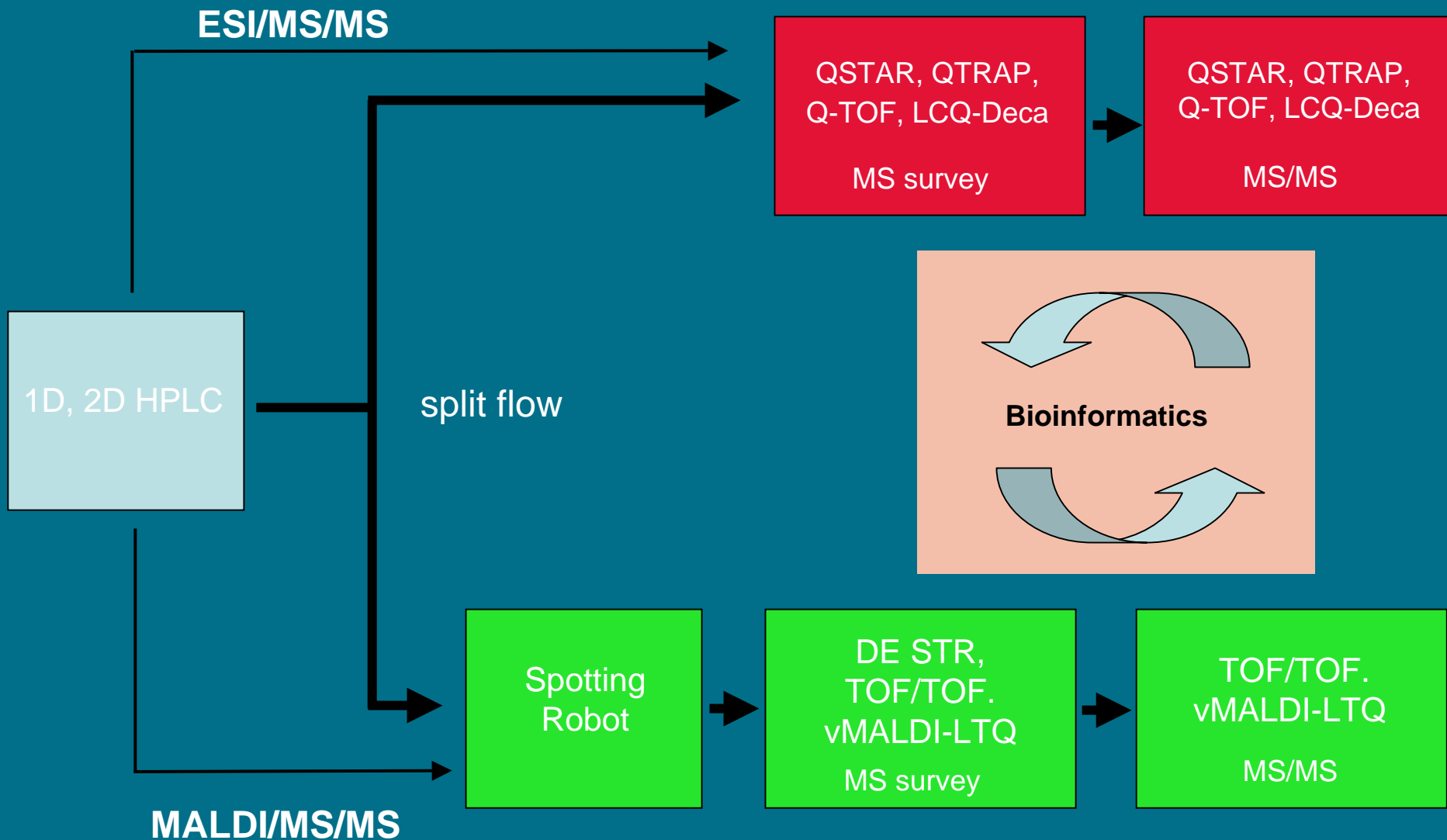
Peptides



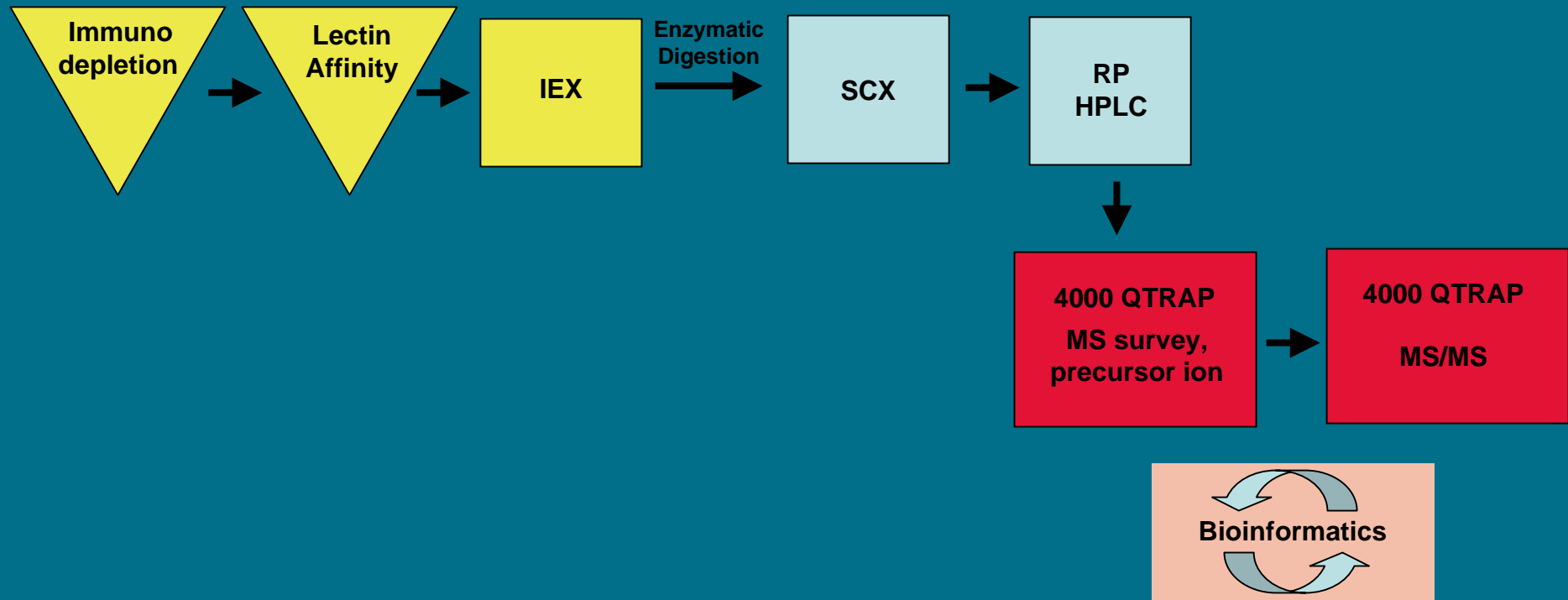
Global Sample Preparation Workflow



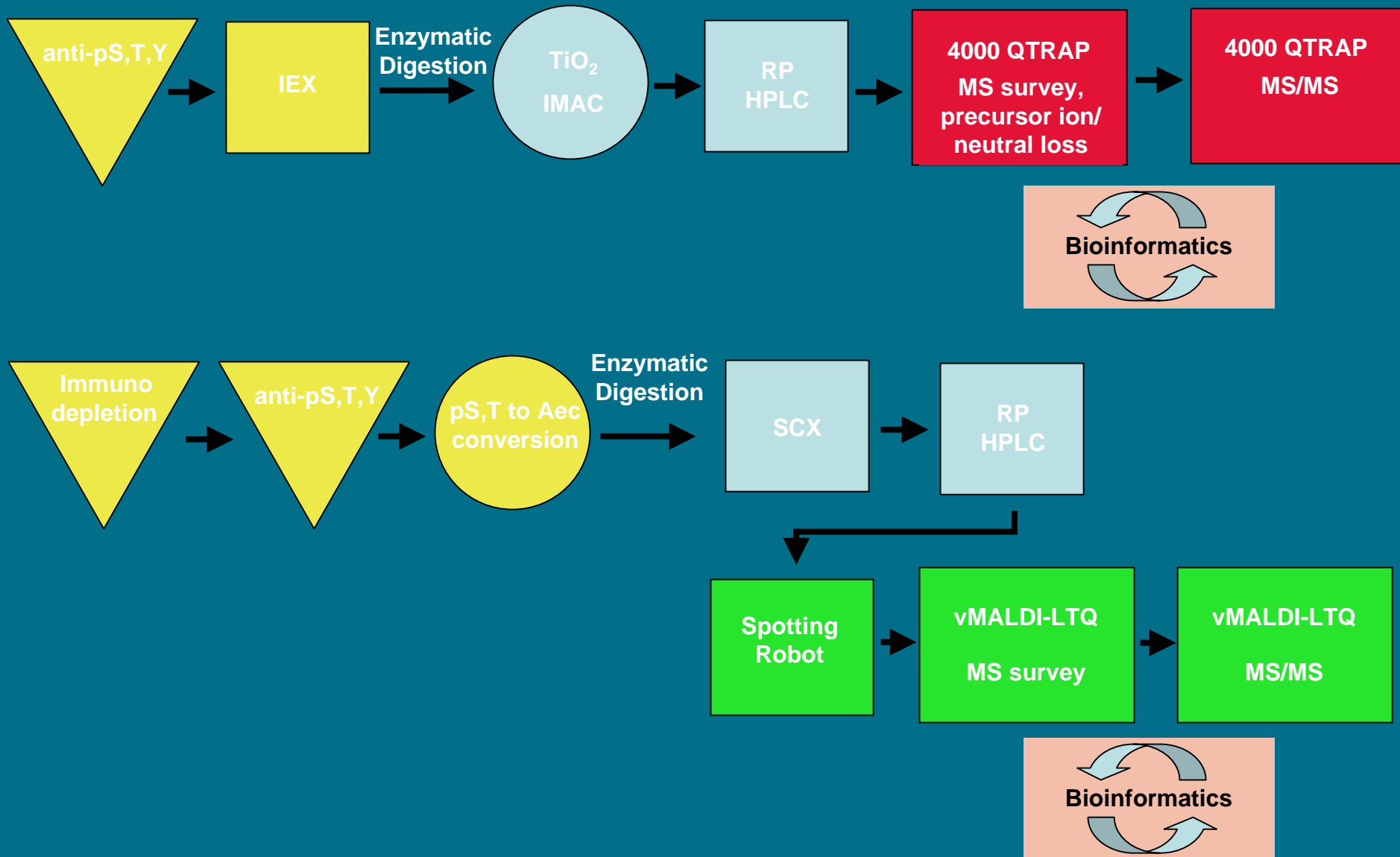
Global Mass Spectrometry Workflow



Targeted MS-based Platforms for Glycoproteins



Targeted MS-based Platforms for Phosphoproteins



1. What proteins are present?
 - IDENTITY
2. How much of each protein?
 - QUANTITY
3. How reliable are the results?
 - QUALITY

What is the desired output?

1. Study design and sample generation
2. Separations and sample handling
3. Column chromatography
4. Capillary electrophoresis
5. Mass spectrometry
6. Informatics for mass spectrometry
7. Gel electrophoresis
8. Gel image informatics
9. Molecular Interaction Experiments
10. Statistical Analysis of Data

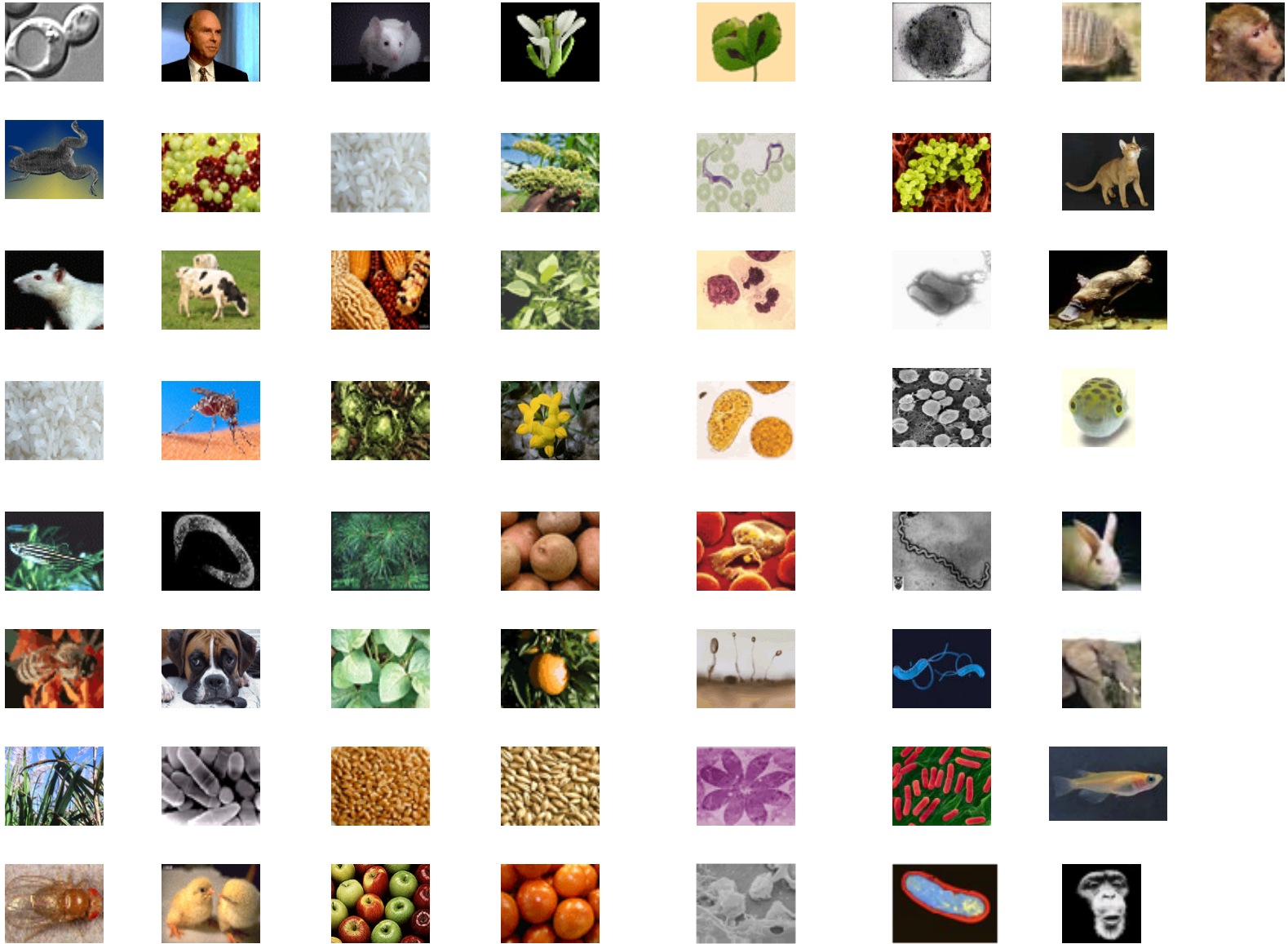
The Minimum Information About a Proteomics Experiment (MIAPE)

“The problem of legacy data sets will be significant in scale and difficult to address. Clearly, a lack of annotation does not mean that a data set is without worth ..., so the following principles should be applied when re-annotating such legacy data:

1. The data set should be re-annotated as fully as possible, with reference to the appropriate MIAPE modules; the data set should then be flagged as legacy, and an indication given of where the reporting requirements have not been met (e.g. a summary of missing items).
2. Data and metadata should never be created to supplement the real data in a file. The only allowable additions are those that serve to indicate the absence of real data”

Protein Sequence Collections (2001)

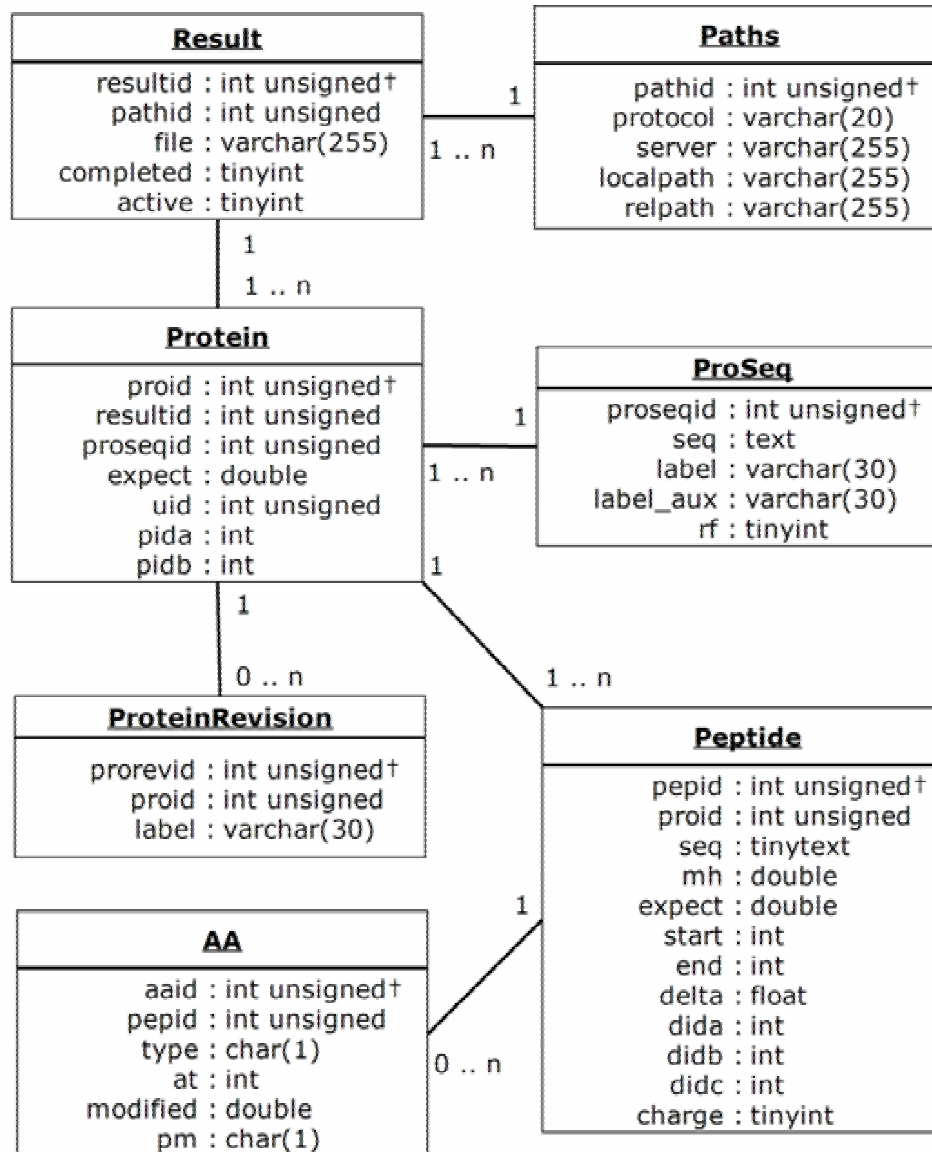
Collection	Annotations
PIR	Good (public)
SWISS-PROT	Good (private)
GenPept	Some (public)
TREMBL	Some (public)
NR	Good (public)
OWL	n/a (public)
dbEST	n/a (public)
HGP	progressing (both)
YGP	Good (both)

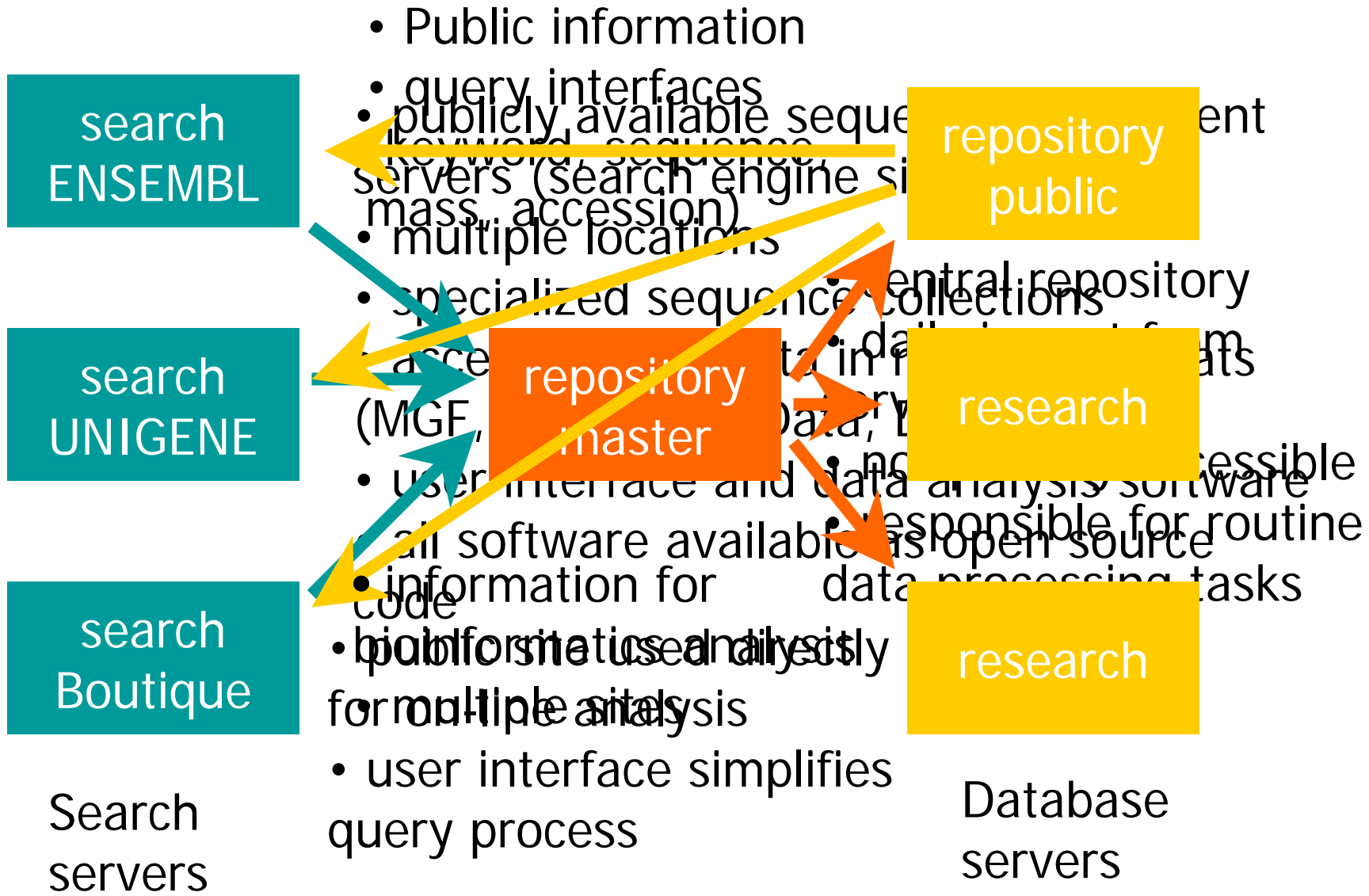


Genomes/Unigene collections

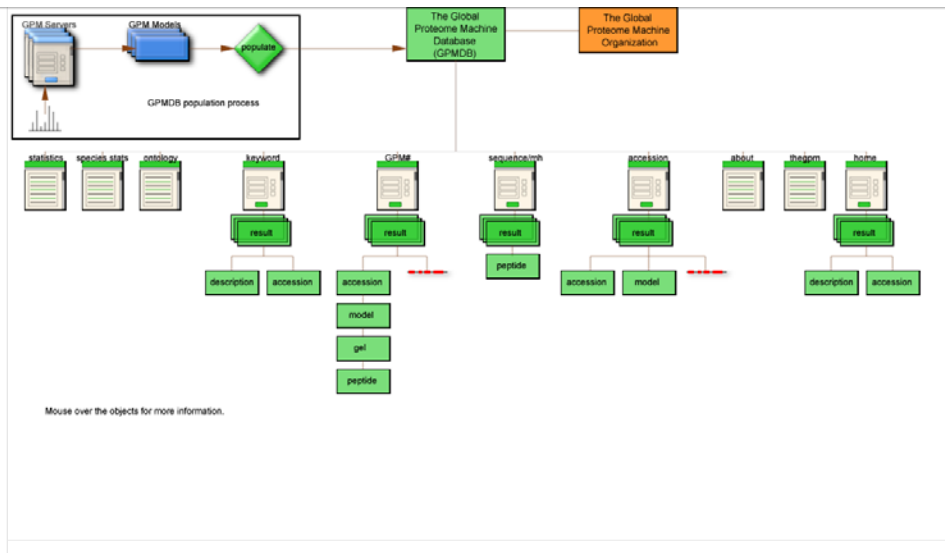
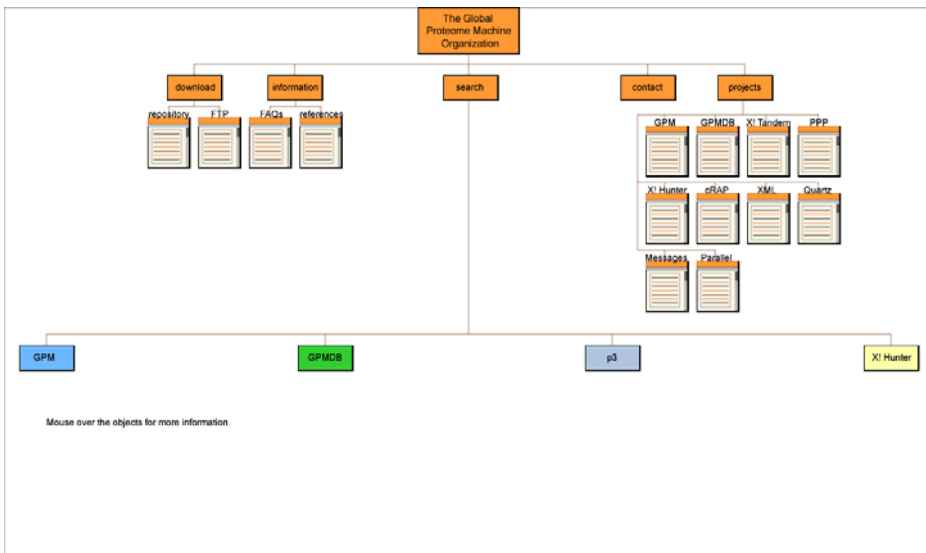
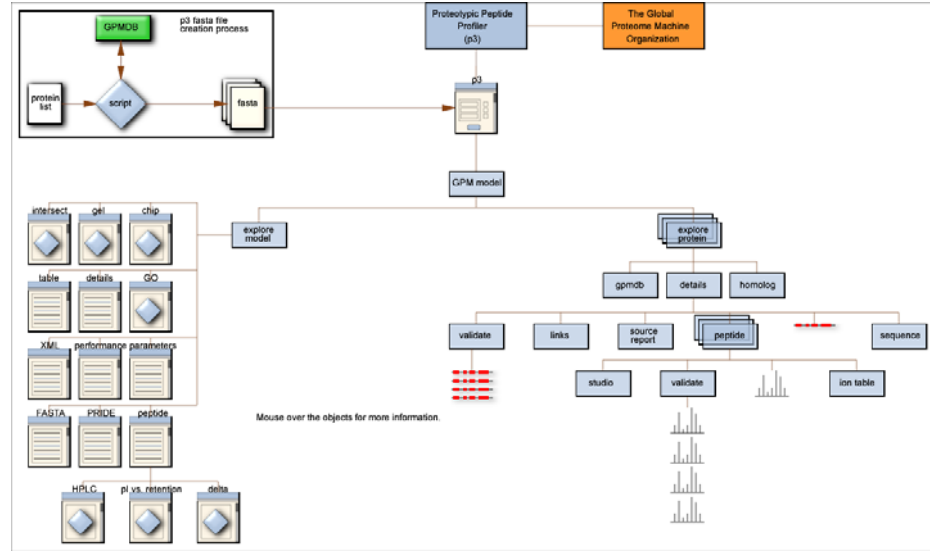
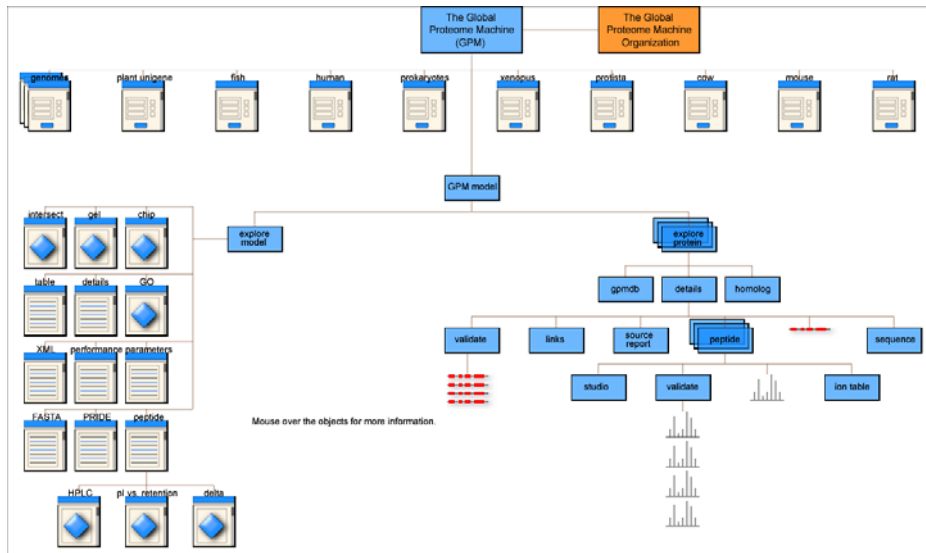
"Biologists would rather share their toothbrush than share a gene name," says Michael Ashburner, ... "Gene nomenclature is beyond redemption."

"Without the umbrella of HUPO, hopes for standardization in proteomics would have been bleak, with researchers being more inclined to use their rivals' toothbrushes than their protocols."





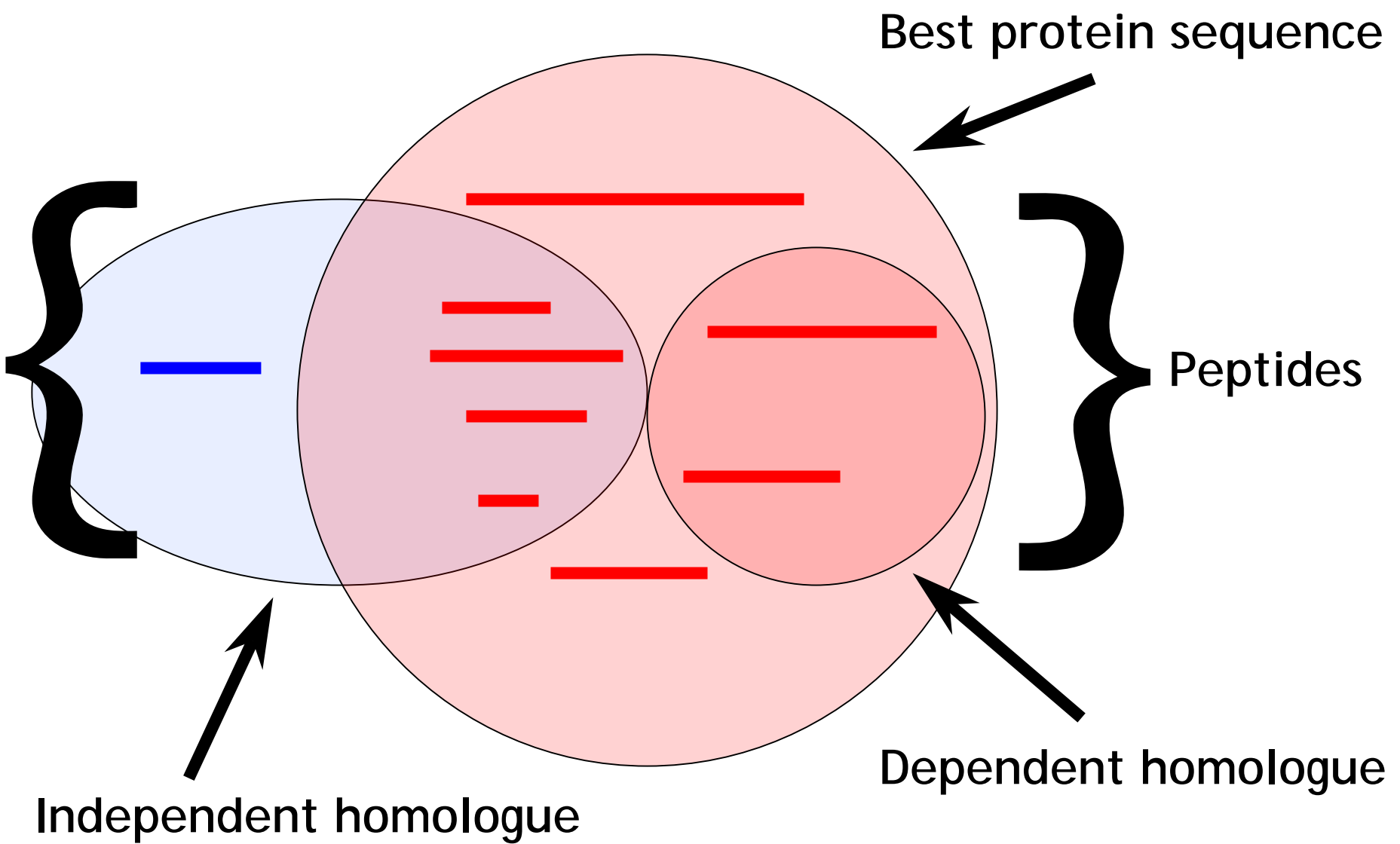
Practical systems lead to complications



Minimum user interface?

1. What does homologue mean if you only have a bunch of peptides?
2. How do you resolve privacy issues?
3. What data formats should be allowed, both for input and output?
4. Which computer operating systems should be supported? Which computer languages should be used?
5. How much detail about each experiment has to be recorded to make the data useful?

Decisions needed to create a repository

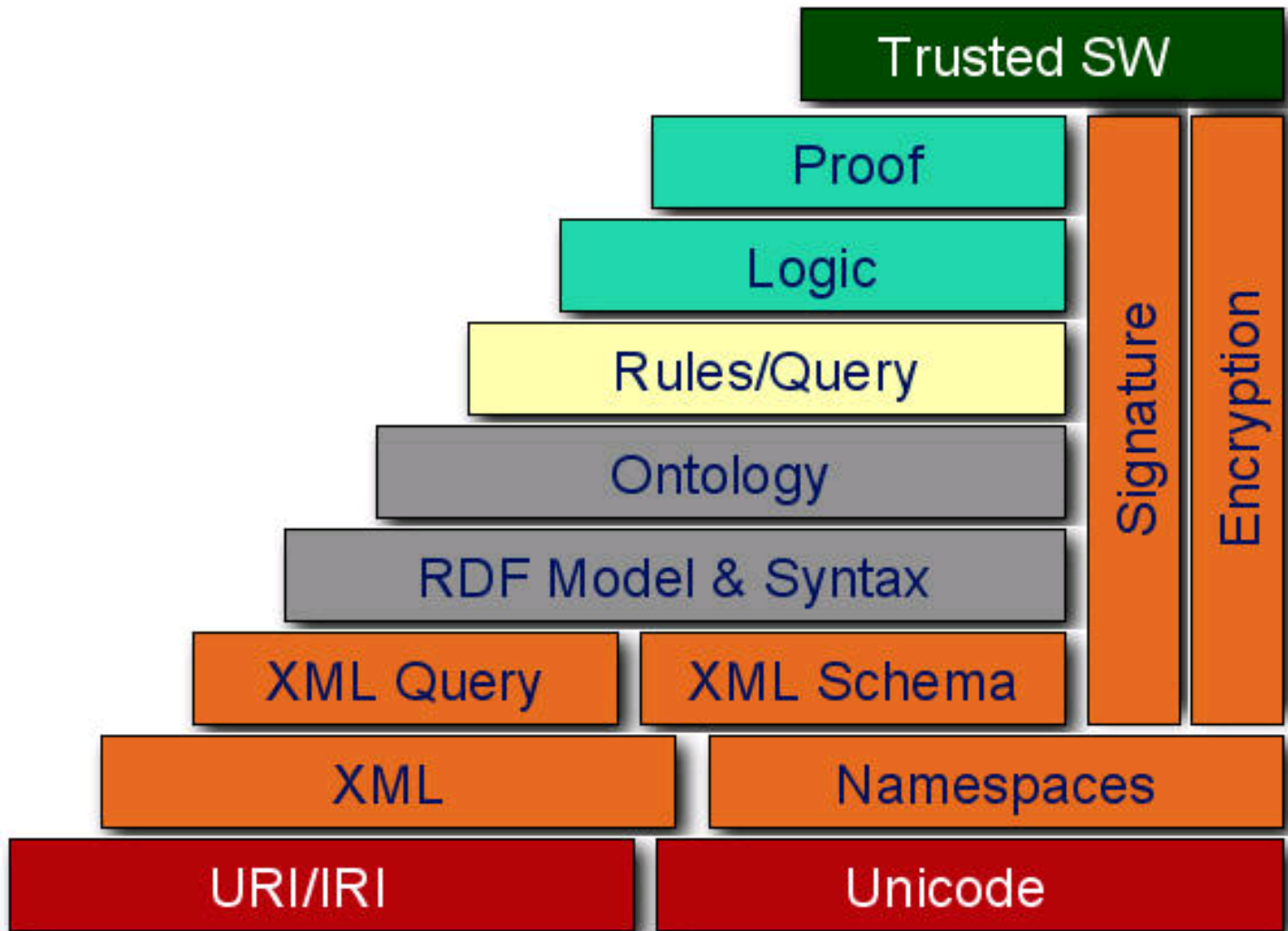


What does homologue mean if you only have a bunch of peptides?

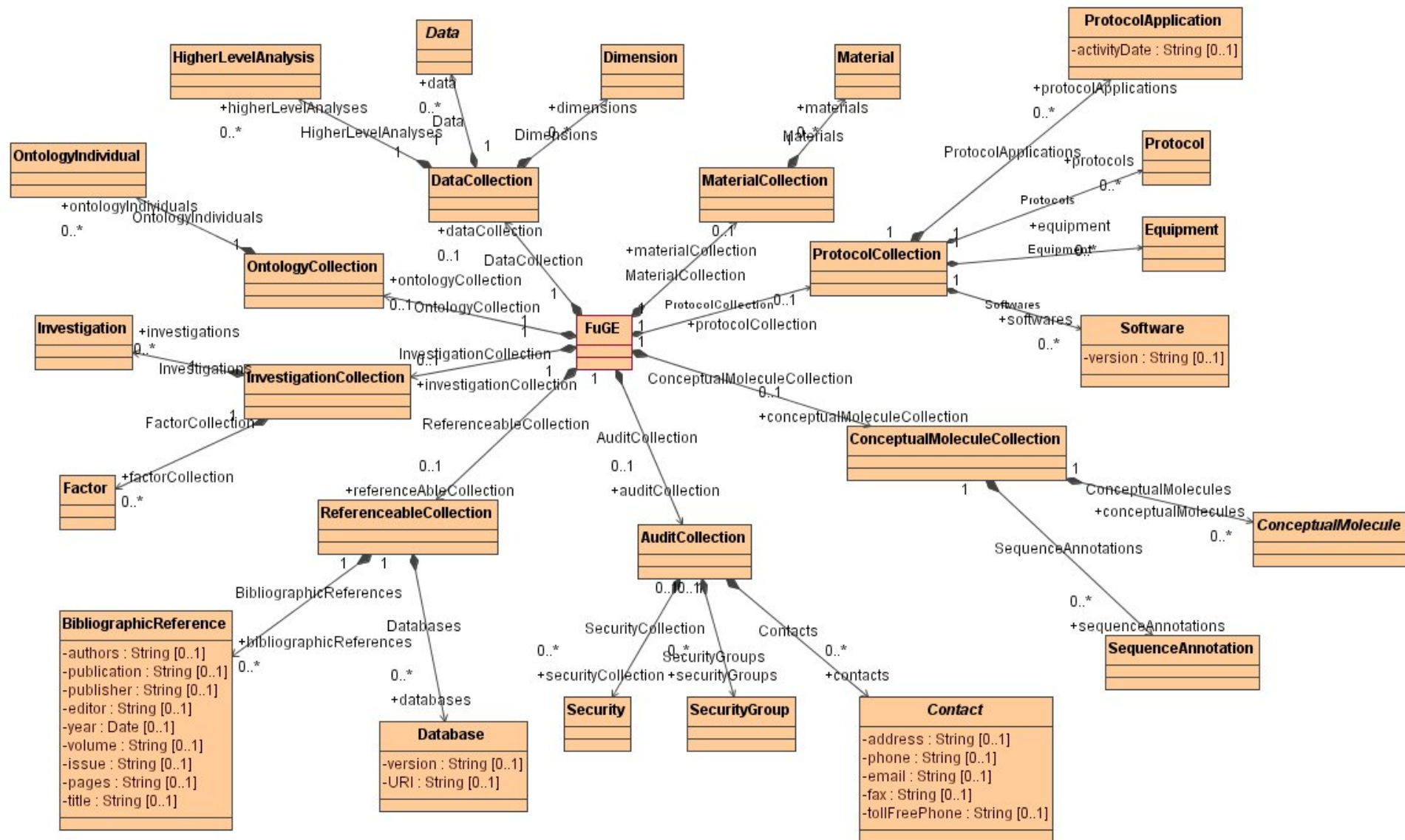
	RDB	XML (Input)	XML (Archive)
PRIDE (EBI)	MySQL	PRIDE XML	PRIDE XML mzData
PeptideAtlas (ISB)	MySQL	mzXML	pepXML mzXML
GPMDDB (UBC+RU)	MySQL	bioML GAML	bioML GAML

Current proteomics repositories

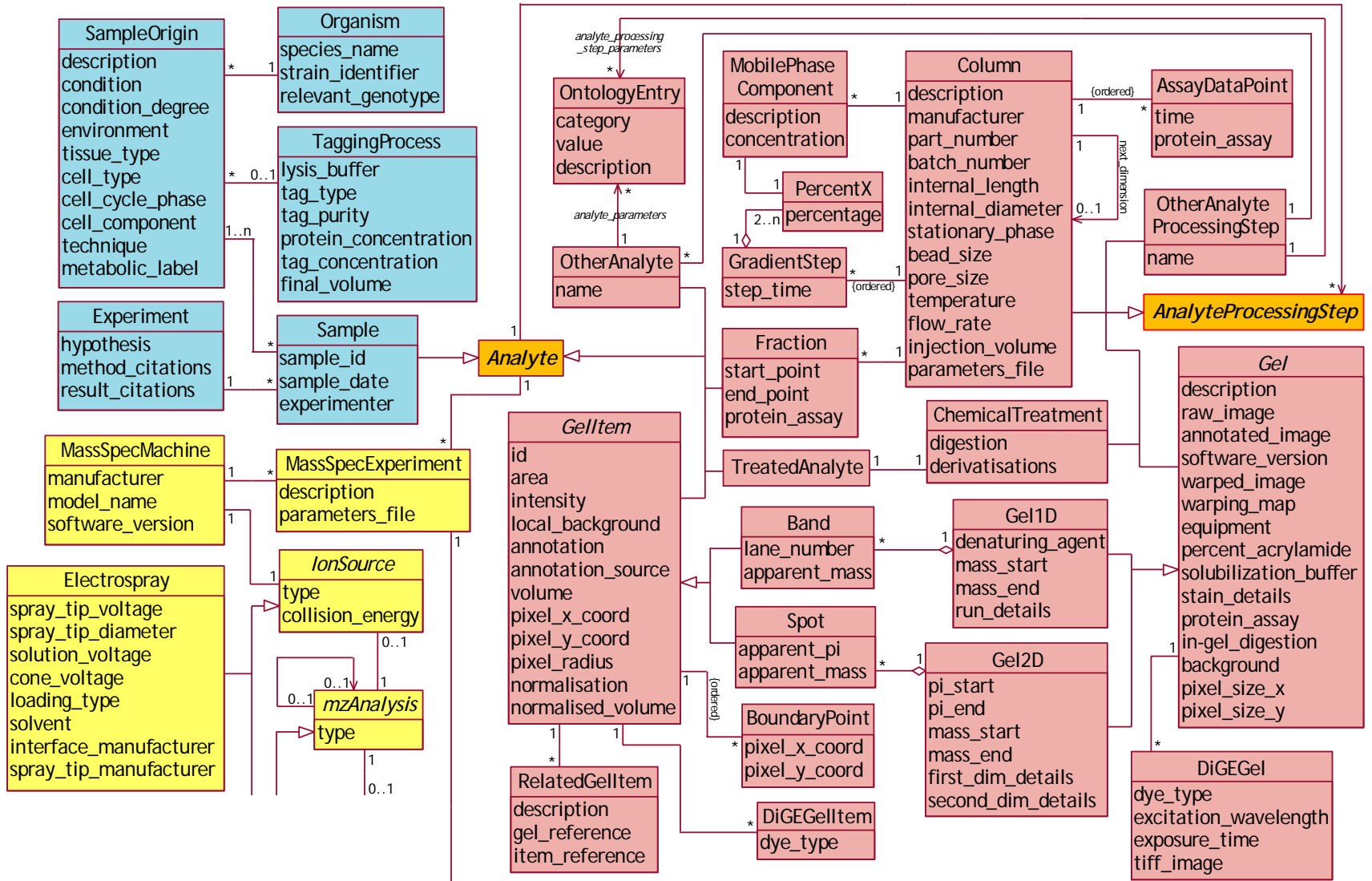
1. mzXML
2. mzData
3. analysisXML
4. PRIDE XML
5. protXML
6. pepXML
7. bioML
8. GAML
9. MI XML
10. Mascot Search Results XML

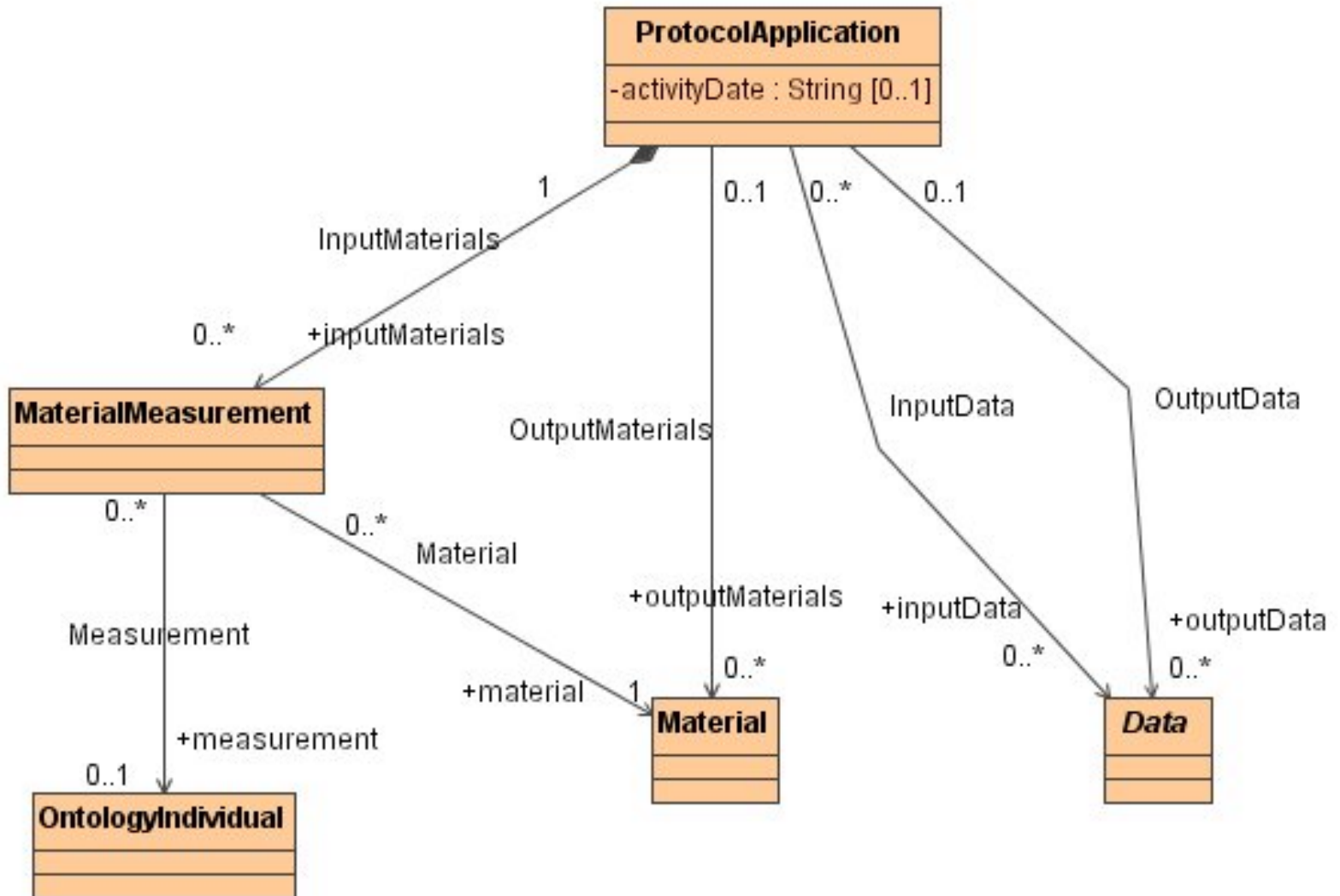


The Semantic Web to the Rescue?



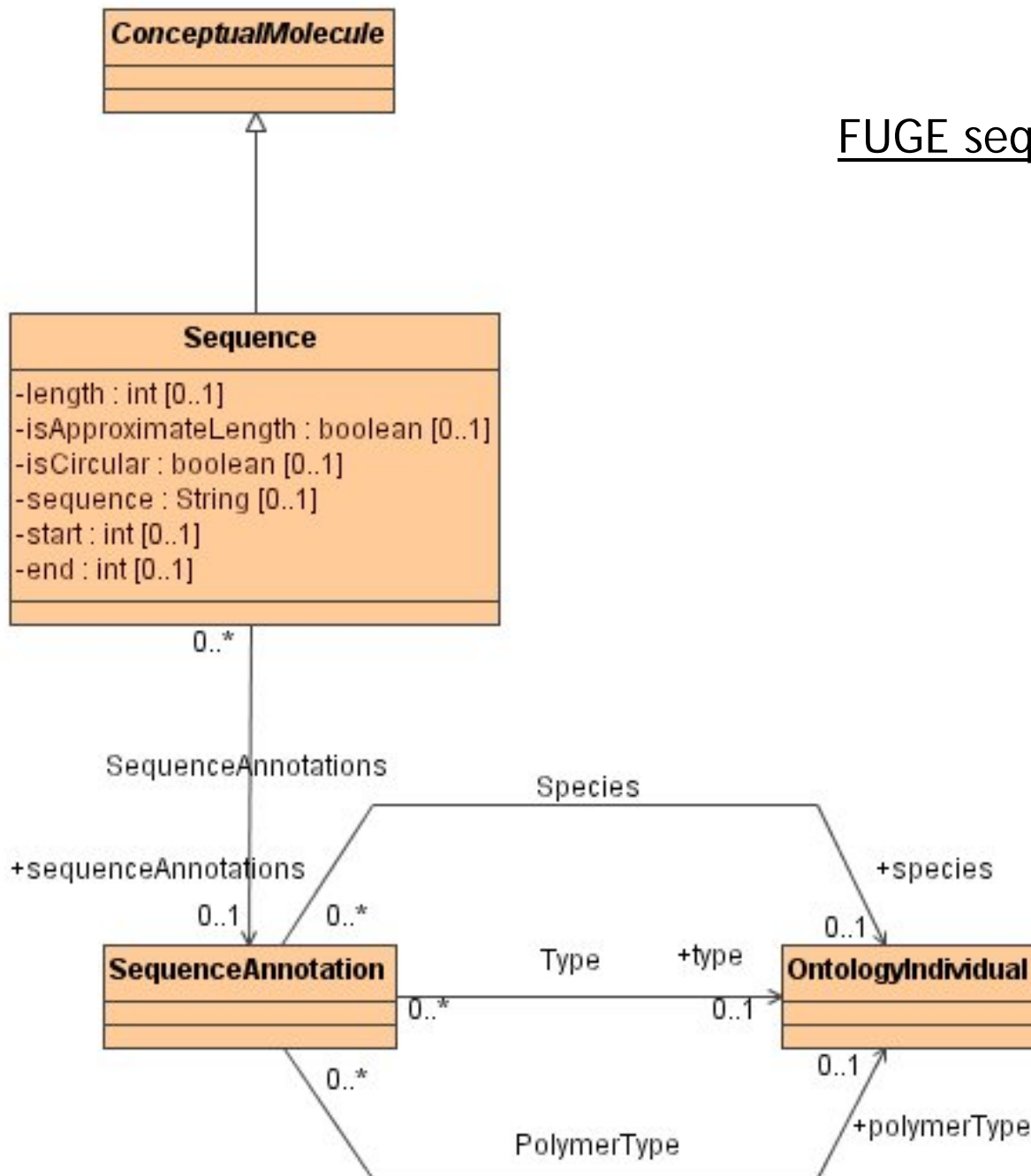
Functional Genomics (FUGE) object model

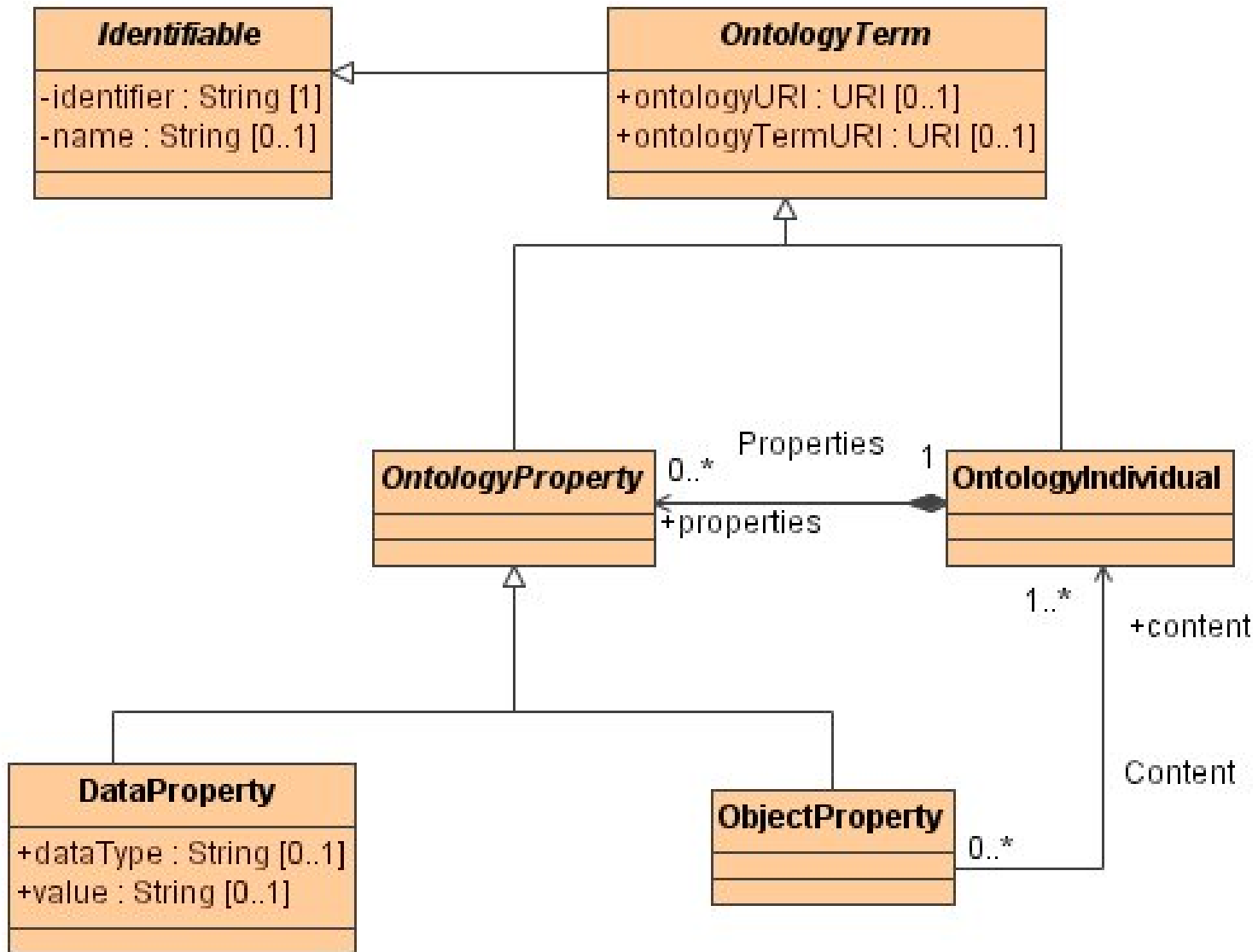




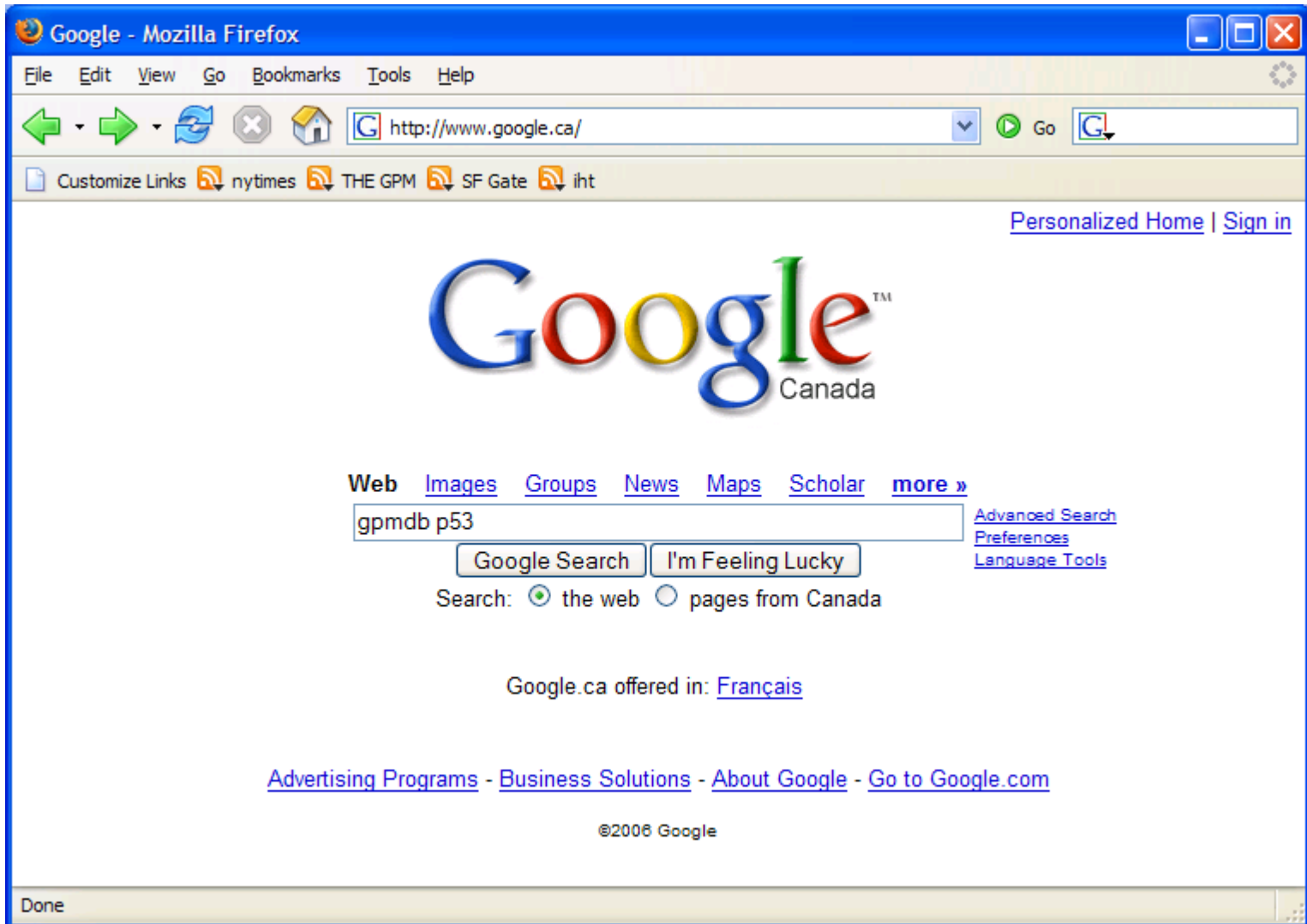
FUGE protocol model

FUGE sequence model





FUGE ontology model



Google to the rescue?

ENSP00000269305: Cellular tumor antigen p53 (Tumor suppressor p53) (Phosphoprotein p53)
log(e) = -36.6 (Antigen NY-CO-13). Source: Uniprot/SWISSPROT P04637

Annotated domains:

IPR011615 p53, DNA-binding
 IPR010991 p53, tetramerisation
 IPR002117 p53 tumor antigen
 IPR001472 Bipartite nuclear localization signal

1 meeqsdpsvepplsgetfsdlwkl1lpennvlsplpsqamddlmlspddieqwftedpgp 60
 MEEPQSDPSVEPPLSQETFSDLWKL1LPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGP
 61 deaprmpeaapvavapapaaptaapapapswplsssvpsqktyqgsygfrlgflhsgtak 120
 DEAPRMPEAAPVAVAPAPAAPTAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAK
 121 svtctyspalnkmfcqlaktcpvqlwvdstpppgtrvraramaiykqsqhmtevvrrcphe 180
 SVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAAMAIYKQSQHMTTEVVRRCPHHE
 181 rcsdsdglappqhlirvegnlrveylddrntfrhsvvvpyppevgsdcttihynymcns 240
 RCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCCTTIHYNMCMNS
 241 scmggmnrpiltiitledssgnllgrnsfevrvcacpgrdrerteenlrkkgephhelp 300
 SCMGGMNRPILTIITLEDSSGNLLGRNSFEVRVCACPGRDRRTEENLRKKGEPHHELP
 301 pgstrkralpnntssspqpkkkpldgeyftlqirgrerfemfrelnealelkdaqagkepg 360
 PGSTRKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELEALELELKDAQAGKEPG
 361 gsrahsshlskkgqstsrhkkmlfktegpdsd 393
 GSAHSSHLKSKKGQSTSRHKKMLFKTEGPDSD

(validate)

spectrum log(e) log(I) m+h delta z sequence

267.1	-1.3	1.33	984.579	-0.021	2	trvr ¹⁵⁹ AMAIYK ¹⁶⁴ qsqh
359.1	-1.1	0.19	931.548	-0.030	2	sdgl ¹⁸⁹ APPQHLIR ¹⁹⁶ vegn
458.1	-1.0	0.71	1721.894	-0.035	3	hlir ¹⁹⁷ VEGNLRVEYL DDR ²⁰⁹ ntfr
1017.1	-7.9	2.16	2212.279	-0.085	3	gmnr ²⁴⁹ RPIITITLE DSSGNLLGR ²⁶⁷ nsfe
520.1	-2.0	0.24	1346.777	-0.032	2	emfr ³⁴³ ELNEALEL ³⁵¹ daqa

Sectors, segments and products of the new biology

key:

- pharma segments
- biotech segments
- pharma/biotech segment overlap

