

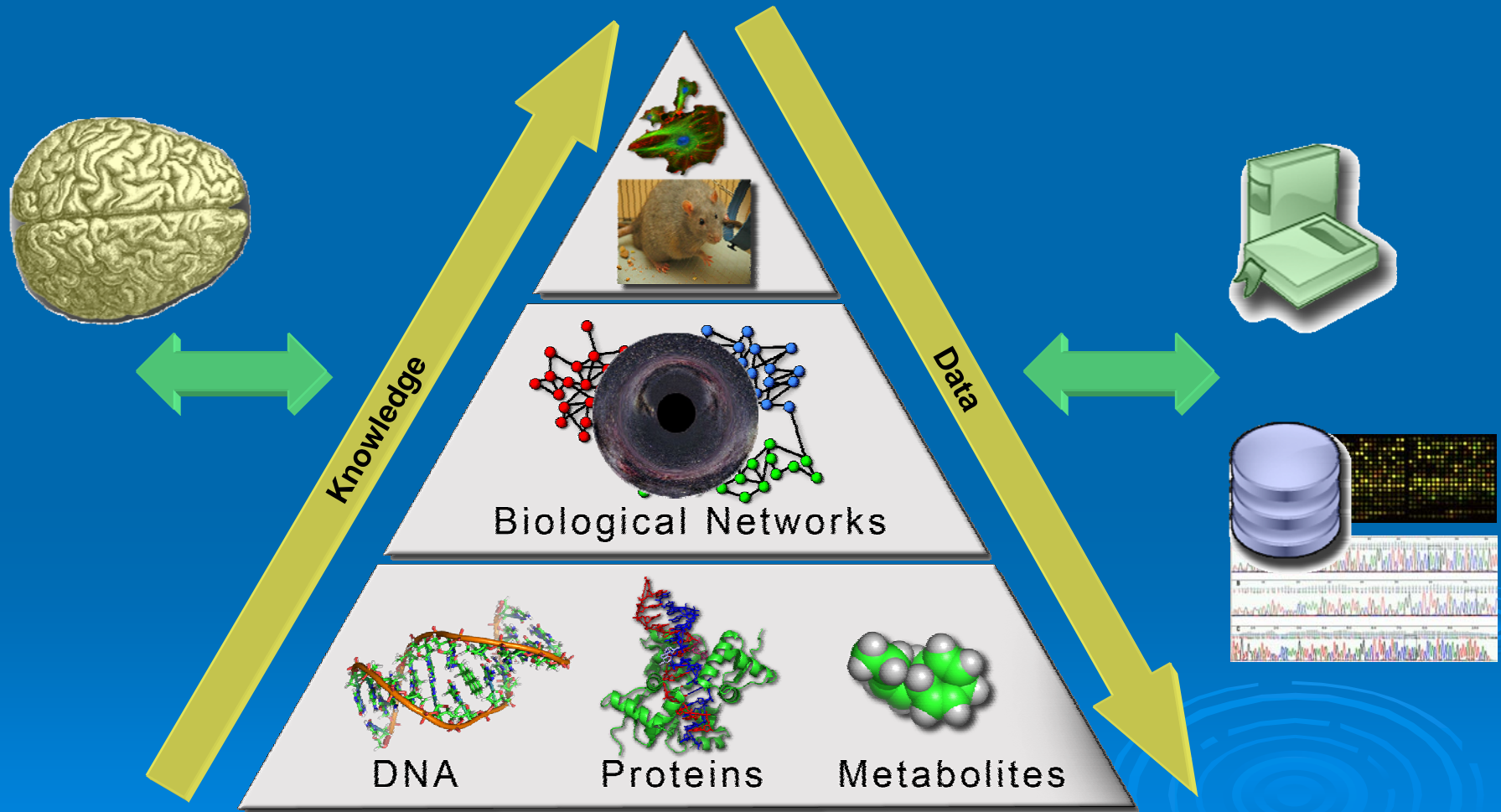
Large Scale Knowledge Representation of Distributed Biomedical Information

Volker Stümpflen

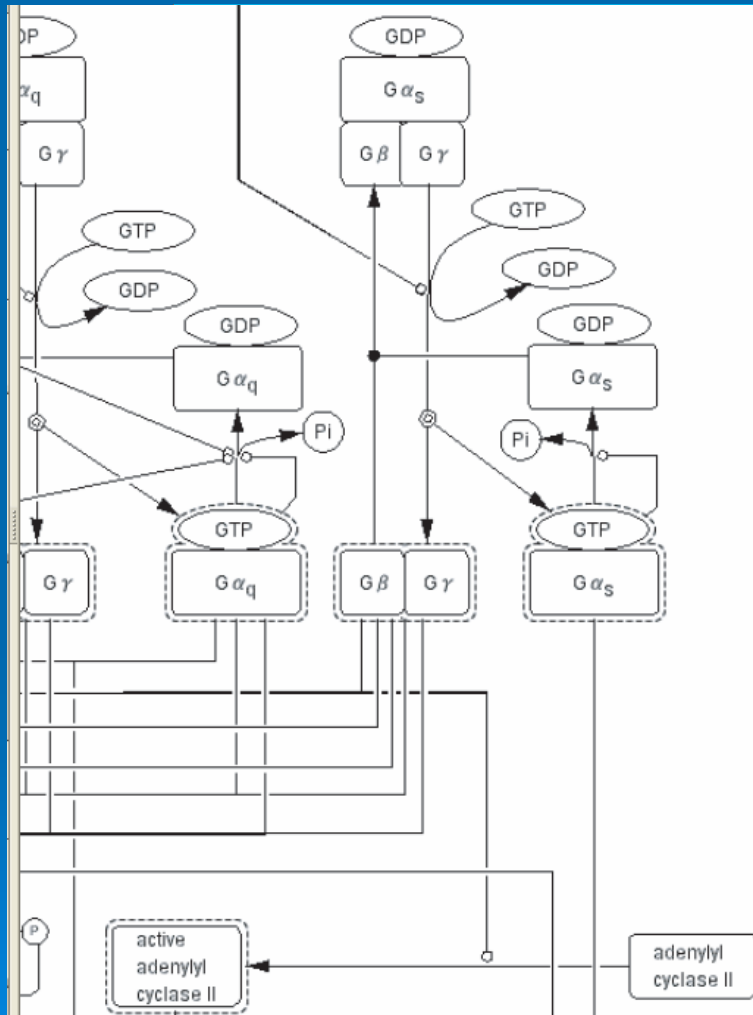
Thorsten Barnickel
Karamfilka Nenova

MIPS / Institute for Bioinformatics
GSF – National Research Center for Environment and Health

Understanding Complex Biological Systems



Systems Biology



Unknown

Phenotype

A → B
state transition

A → B
mediated transactivation

A -| B
transcriptional inhibition

A -| B
translational inhibition

ErbB family

Ca signaling

ControlPanel Single_Gene.mdl

File Edit View Simulation

Time span: 20
End Time: 20
Num. of Points: 100
Error tolerance: 1e-6
Exponent: -6

Species	Parameters	Change amount	Parameter Scan	Initial Qty
01	src	c2		0
02	waste	c2		0
03	RNAp	c2		0
04	mRNAuc	c2		9
05	mRNAuc	c2		1
06	mRNAuc	c1		1
07	mRNAuc	c1		1
08	P	c1		0
09	AA	c1		9

Concentration vs Time graph showing oscillatory behavior of species over 20 time units.

Legend:

- src
- waste
- RNAp
- mRNAuc
- mRNAuc
- mRNAuc
- P
- AA

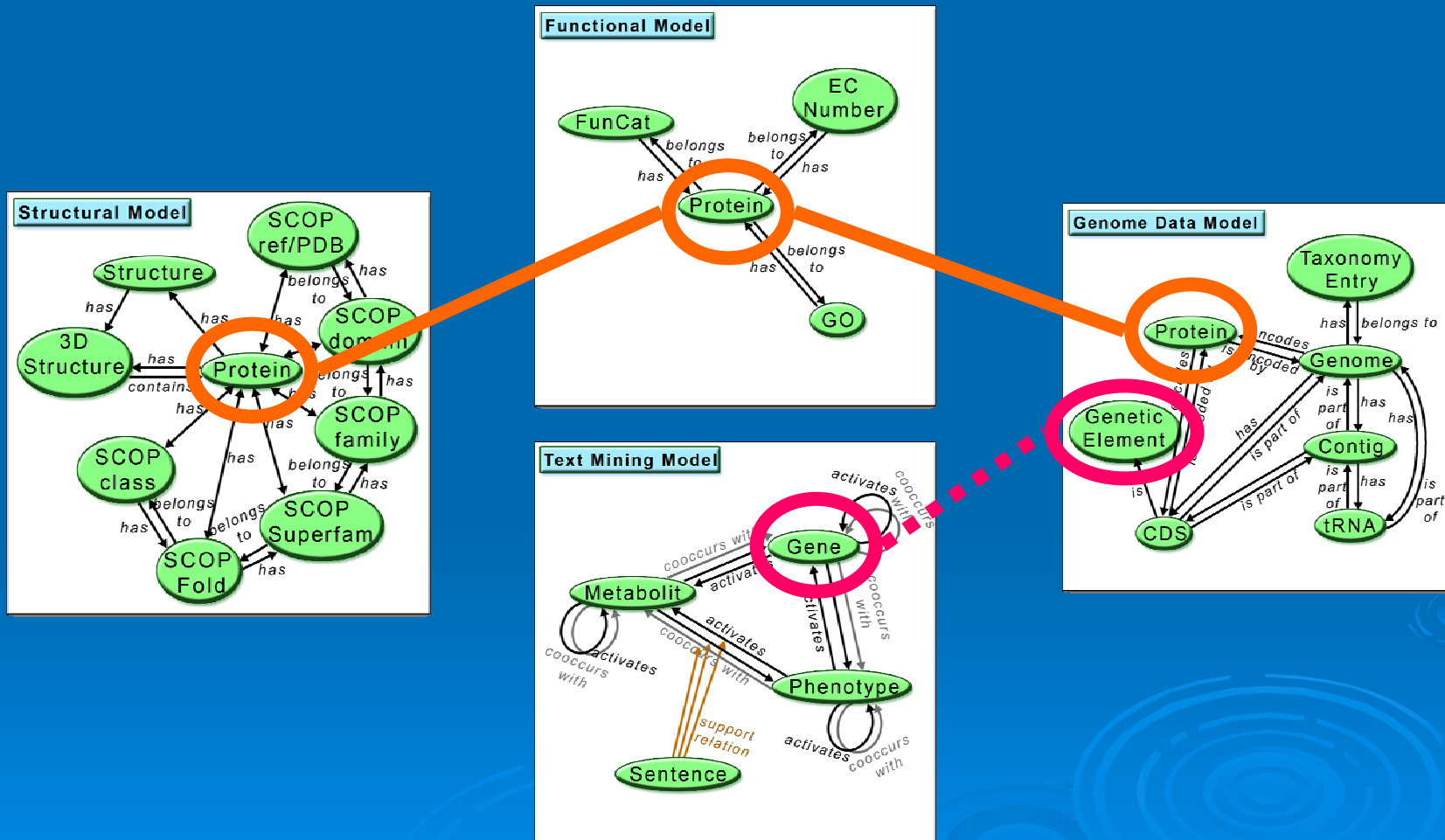
Initialize Save Execute Close

CellDesigner is available at <http://cellDesigner.org>

Questions

- Different knowledge domains ?
- Ontologies for semantic structuring ?
- Semantic structures from free text ?
- Knowledge representation from distributed resources ?

Merging Knowledge from Different Domains



Semantic Structuring Demands for Ontologies

- Life sciences have a long tradition in classification ...
- ... various ontologies are available and in use
- Ontologies (in the broadest sense):
 - Controlled vocabularies
 - Taxonomies
 - Frames
 - ...
- Examples for Ontologies:
 - MeSH terms, Gene Ontology (GO), FunCat, ...
 - Many more from e.g. Open Biomedical Ontologies (<http://obofoundry.org/>)

Example: Extending the Functional Context of Proteins

gsf **mips**
munich information center for protein sequences

Geknow

Search
Results
Genomes
FunCat
External Reference

Ontology viewer

- 01 METABOLISM
- 02 ENERGY
 - 02.01 glycolysis and gluconeogenesis
 - 02.04 glyoxylate cycle**
 - 02.05 Entner-Doudoroff pathway
 - 02.07 pentose-phosphate pathway
 - 02.08 pyruvate dehydrogenase complex
 - 02.09 anaplerotic reactions
 - 02.10 tricarboxylic-acid pathway (citrate cycle)
 - 02.11 electron transport and membrane-ATPases
 - 02.13 respiration
 - 02.16 fermentation

Protein ao90009000219

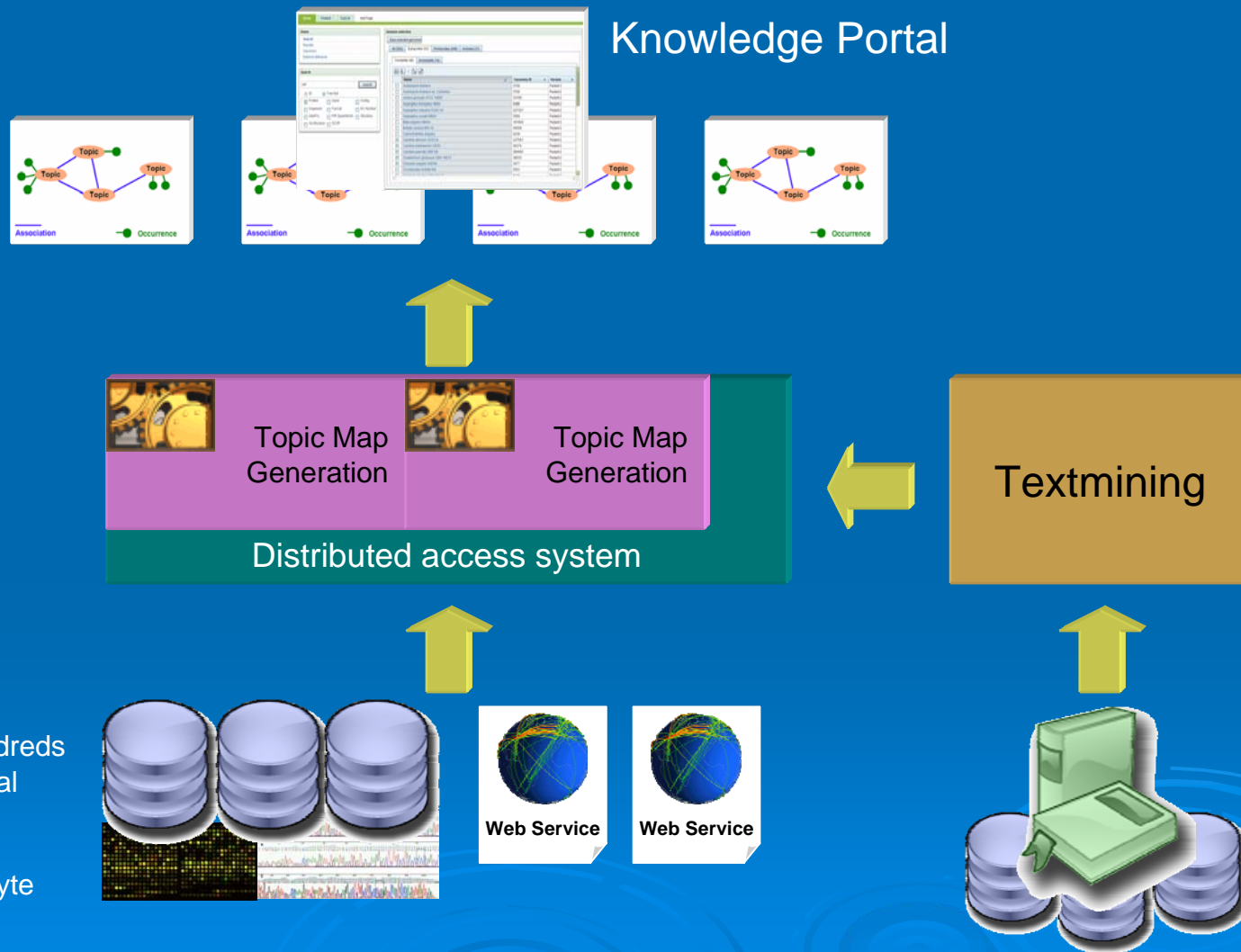
Characteristics: ao90009000219

Description	isocitrate lyase
Molecular Weight	60026.2650200003
Sequence	MGFLEDEDKKYLDDVQAVKAWWTDSDRWRHTEYPSNVQSKKLWKILESNFENKVASFTY
Length	538
Organism	Aspergillus oryzae

Protein ao90009000219

has function	01.05.01.01	
has function	01.01.06.04.02	
has function	01.05.01	
has function	01	
has function	01.05	
is encoded within genome	Aspergillus oryzae	
is encoded by CDS	ao90009000219	Aspergillus oryzae

Semantic Structuring and Knowledge Representation

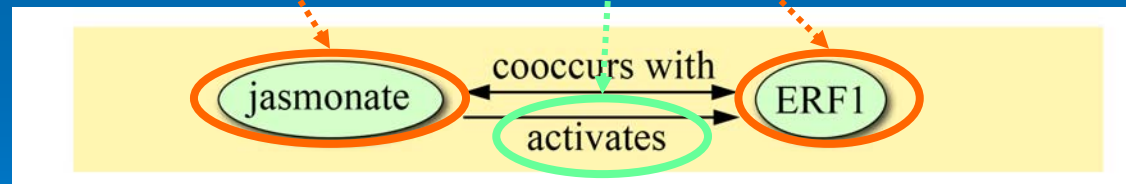


Knowledge in Free Text

Free text

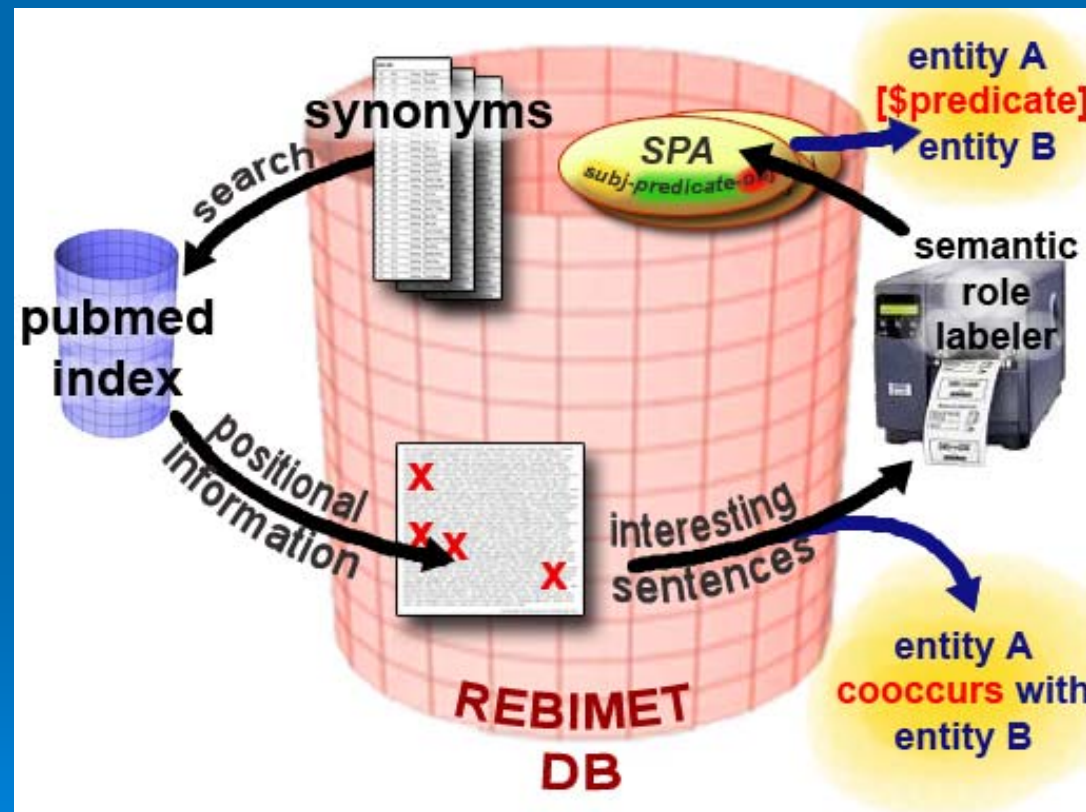
... of pathogen response genes that prevent disease progression.
The expression of ERF1 can be activated rapidly by ethylene or jasmonate and can be activated synergistically by both hormones.
In addition, both signalling ...

Topic Map

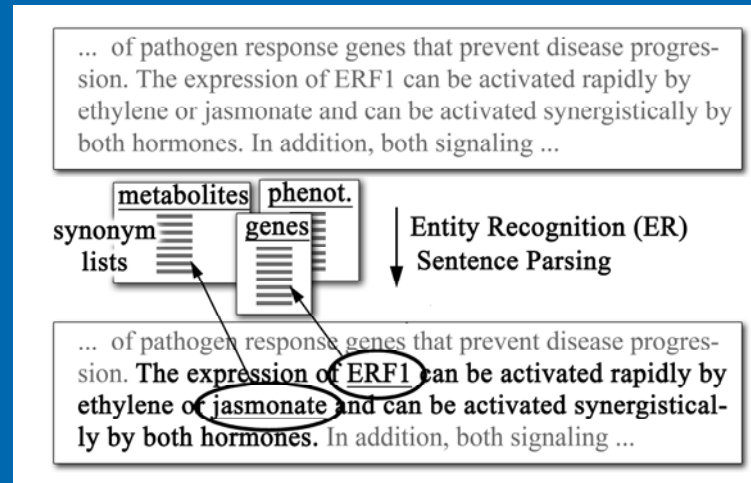


REBIMET

➤ Relation Extraction from Biomedical Texts



Entity Recognition



- Identification of relevant biological entities:
 - Based on synonym lists created from terms in taxonomies, gene names,
- Realized with Apaches Lucene

Information Extraction with Semantic Role Labeling and Cooccurrence

The expression of ERF1 can be activated rapidly by ethylene and jasmonate and can be activated synergistically by both hormones.

1. Semantic Role Labeling:



ASSERT tool
(Pradhan S. et al., 2005)

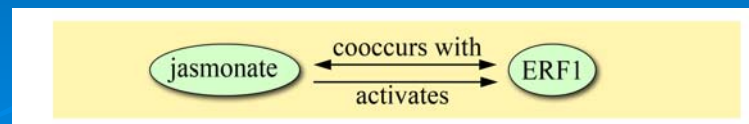
1.1 SPA structure for verb a)

[_{arg1} *The expression of ERF1*] [_{argm-mod} *can*] be [_{target} *activated*] [_{argm-mnr} *rapidly*] [_{arg0} *by ethylene and jasmonate*] and can be activated synergistically by both hormones

1.2 SPA structure for verb b)

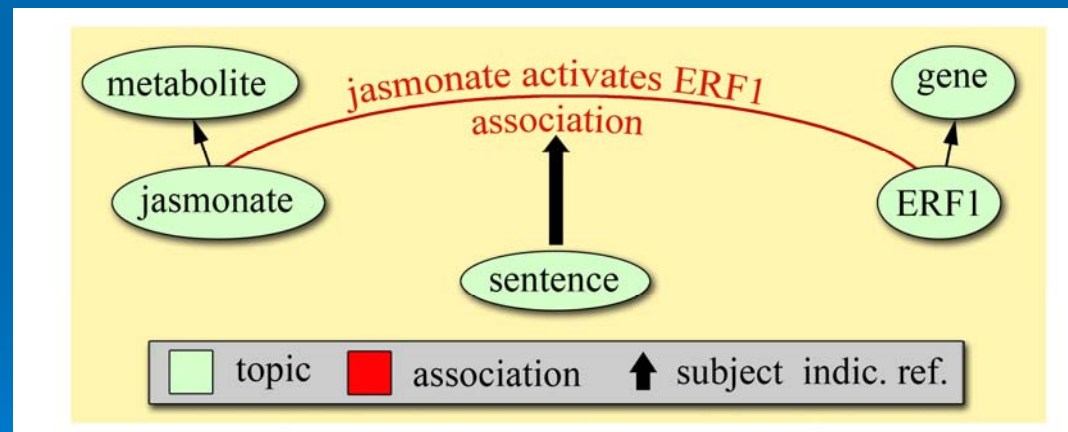
[_{arg1} *The expression of ERF1*] can be activated rapidly by ethylene and jasmonate and [_{argm-mod} *can*] be [_{target} *activated*] [_{argm-tmp} *synergistically*] [_{arg0} *by both hormones*]

2. Information Extraction:



Simplified TM Representation

- Generation of Topic Map fragments
- Connection to evidence in text by reification



Screenshot Portal

- PSI based merging of textmining model with genome model

Coding Sequence gi_19115202



Characteristics: gi_19115202

Database: ...

Results

Name	Type	Description	Organism
PID: 15508018	2005	Mutation in the 5' alternatively spliced region of the XNP/ATR-X gene causes Chudley-Lowry syndrome.	

Sign

munich information center for protein sequences

GeKnow

Search



Results

Genomes

FunCat

External Reference

http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=retrieve&db=pubmed&uid=15508018

A service of the National Library of Medicine and the National Institutes of Health

My NCBI [Sign In] [Register]


All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search PubMed for Go Clear

Limits Preview/Index History Clipboard Details

Display AbstractPlus Show 20 Sort By Send to

All: 1 Review: 0

1: Eur J Hum Genet. 2005 Feb;13(2):176-83.  Links

Mutation in the 5' alternatively spliced region of the XNP/ATR-X gene causes Chudley-Lowry syndrome.

Abidi FE, Cardoso C, Lossi AM, Lowry RB, Depetris D, Mattèi MG, Lubs HA, Stevenson RE, Fontes M, Chudley AE, Schwartz CE.

JC Self Research Institute, Greenwood Genetic Center, SC 29646, USA.

The Chudley-Lowry syndrome (ChLS, MIM 309490) is an X-linked recessive condition characterized by moderate to severe mental retardation, short stature, mild obesity, hypogonadism, and distinctive facial features characterized by depressed nasal bridge, anteverted nares, inverted-V-shaped upper lip, and macrostomia. The original Chudley-Lowry family consists of three affected males in two generations. Linkage analysis had localized the gene to

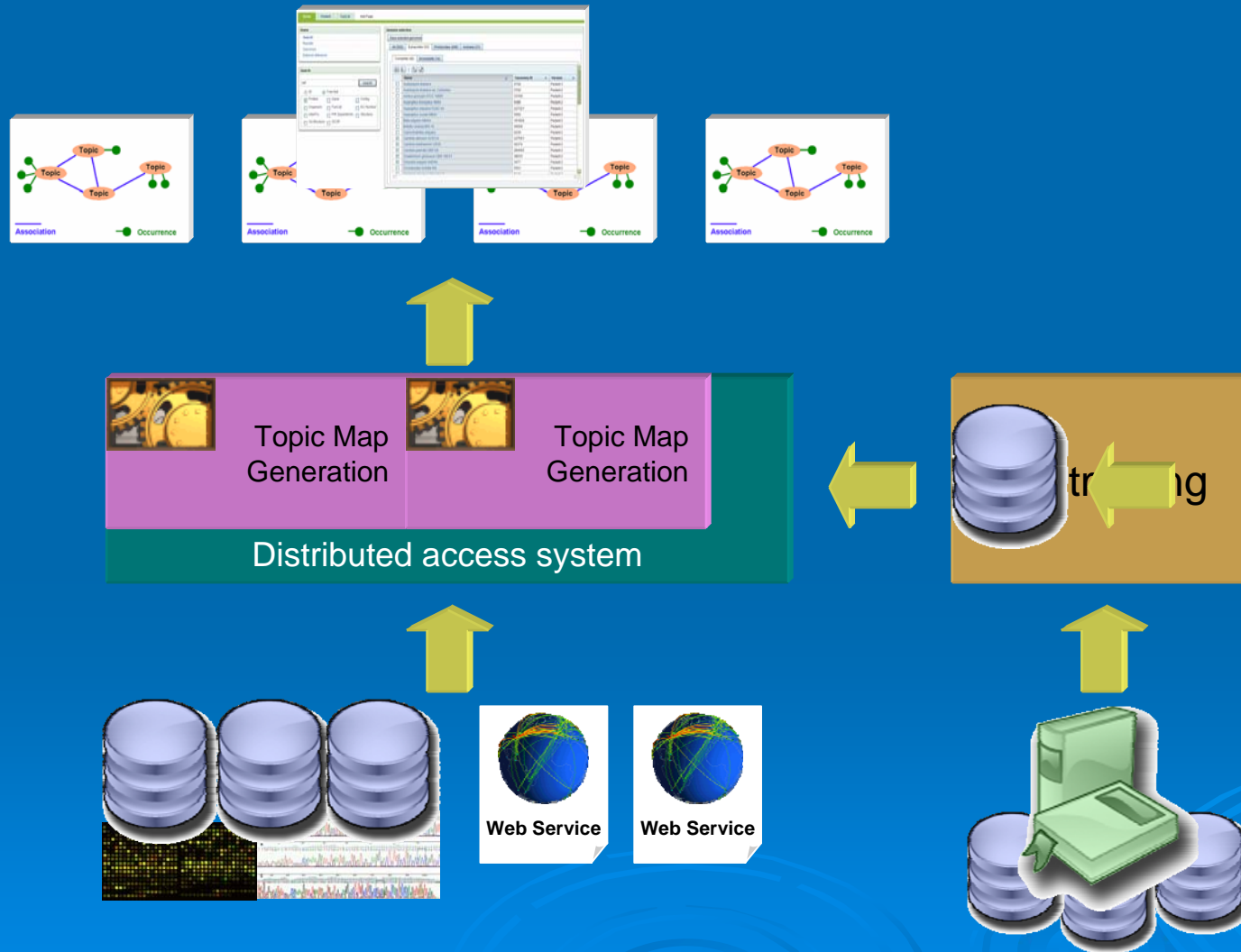
Related Links

- ▶ Splicing mutation in the ATR-X gene can lead to Chudley-Lowry syndrome. [Am J Hum Genet. 1996]
- ▶ ATR-X mutations cause impaired nuclear location. [J Med Genet. 2000]
- ▶ Determination of the genomic structure of the XNP gene. [Genomics. 1997]
- ▶ Prenatal diagnosis of ATR-X syndrome in a fetus. [Prenat Diagn. 2001]
- ▶ A novel splicing mutation of the ATRX gene in ATR-X syndrome. [Brain Dev. 2005]

See all Related Articles...

is part of contig	NC_003424			Schizosaccharomyces pombe 972h-
is part of genome	Schizosaccharomyces pombe 972h-			

Large Scale Integration and Knowledge Representation



GeKnow: Integration of PEDANT, SIMAP, NCBI data, NCBI PubMed

- PEDANT 3 ~ 600 GB
 - contains 450 genomes each stored in a single MySQL database
 - no possibilities for simultaneous cross genome comparison
- SIMAP ~ 540 GB compressed
 - contains over 7 Mio. unique protein sequences
- NCBI
 - Taxonomy information (some thousands)
- Textmining from PubMed
 - 16 Mio. abstracts, 65 Mio Hits, 15 Mio. Sentences, 13 Mio. SPA structures
- Integration of these data on the fly
- Semantic linking of PEDANT databases with SIMAP and NCBI Taxonomy
- No redundant data

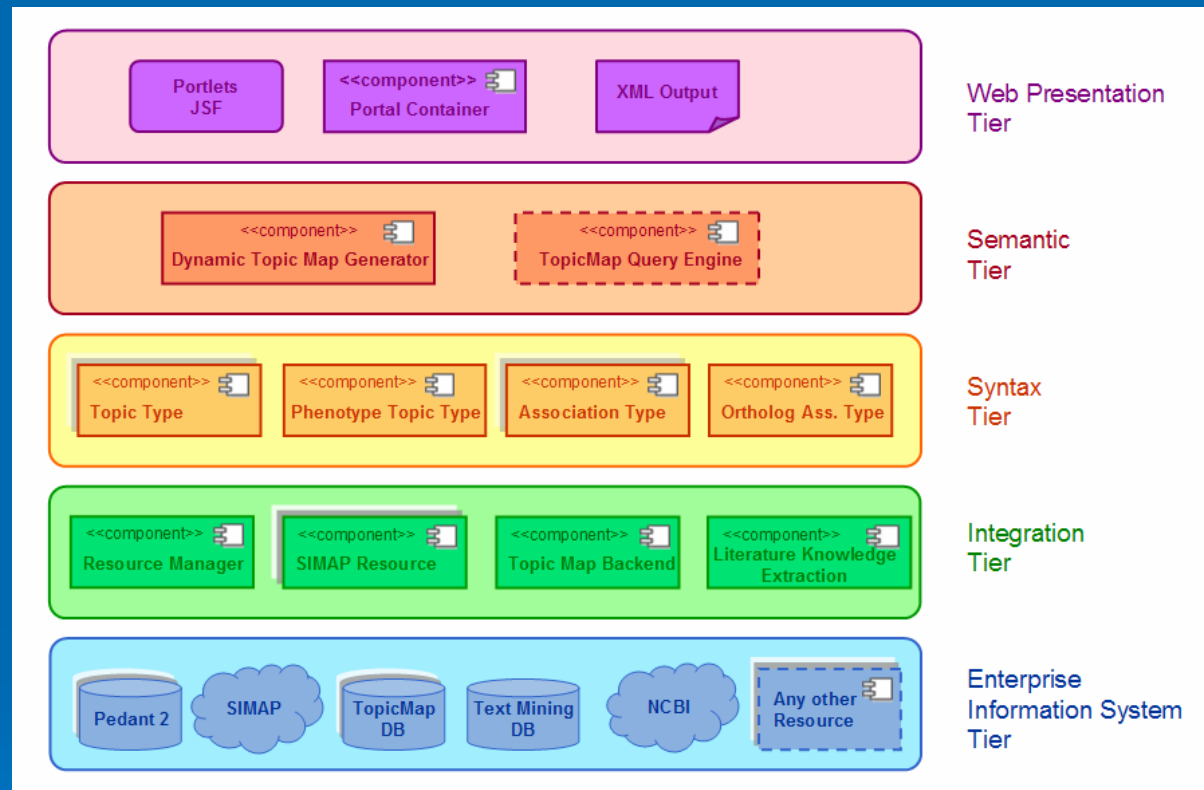
How To Generate the Topic Maps ?

Generation of TM fragments

- Problems with generation of one large TM
 - Very large data collections (storage problems)
 - Distributed
 - Update problems

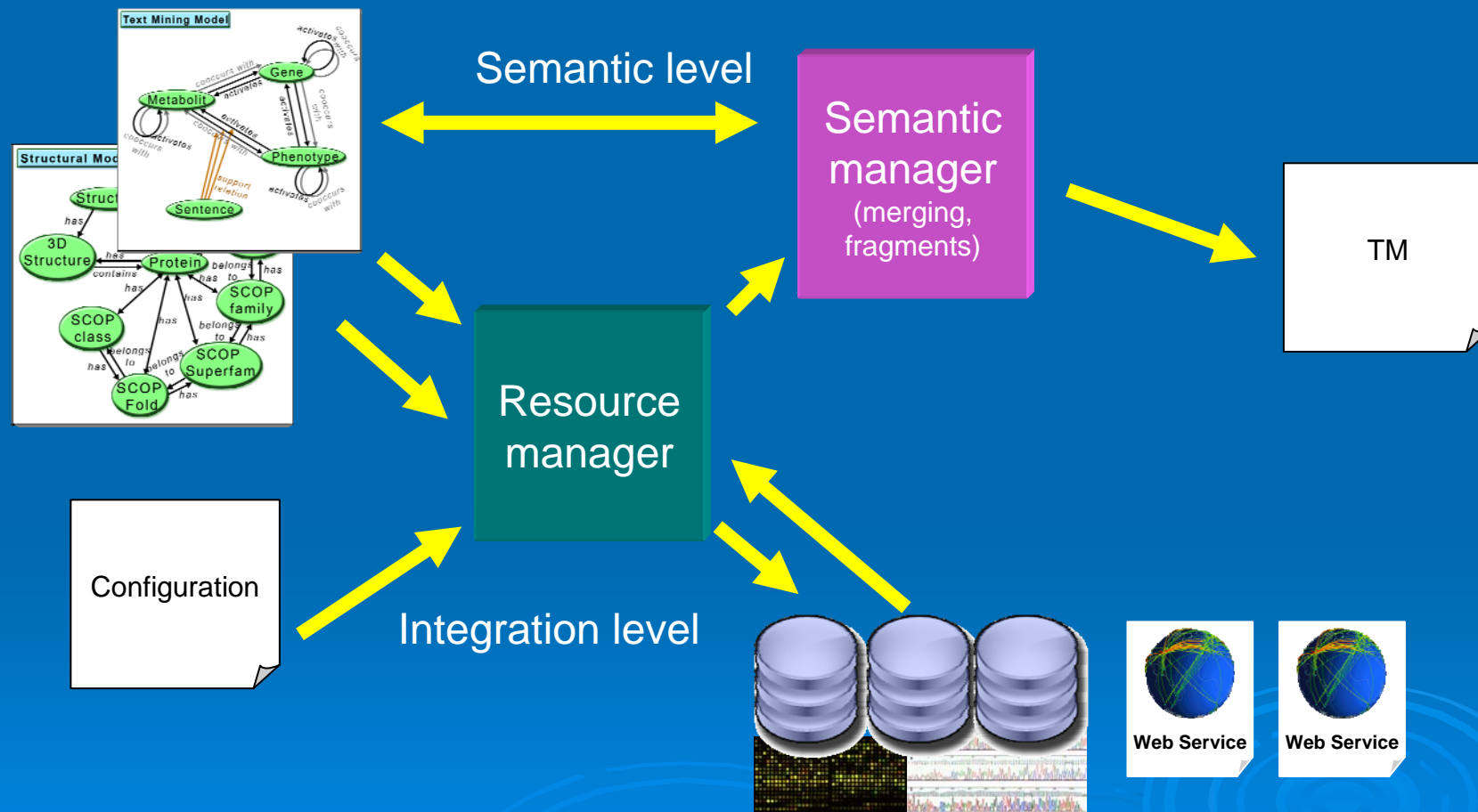
System Architecture (GeKnow)

- Extension of our n-Tier J2EE based component and service oriented architecture (EJBs and Web Services)
- Simply by adding some semantic components ..
- .. and one semantic Tier



Concept:

- Independent semantic layer on top of arbitrary data sources



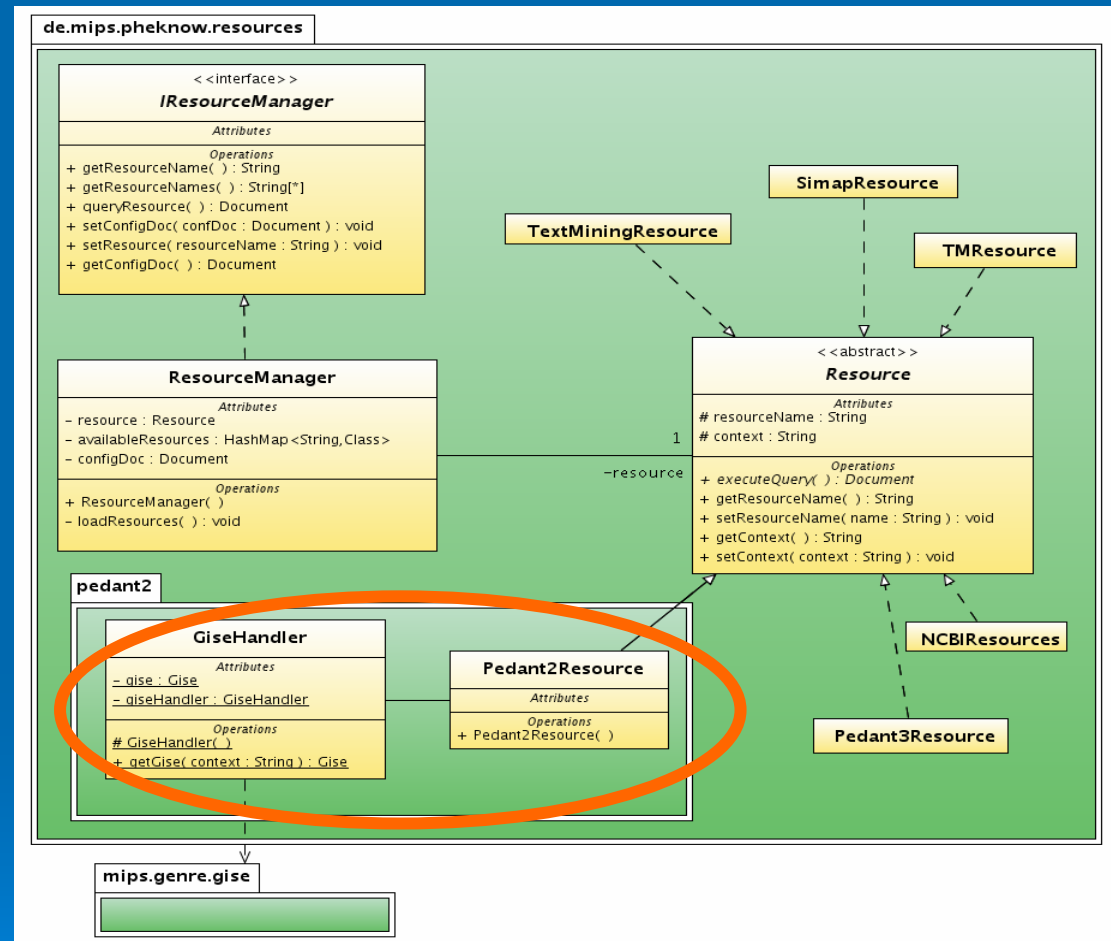
Integration Tier

➤ Resource:

- Aware of mapping between topic / association types and methods from data source

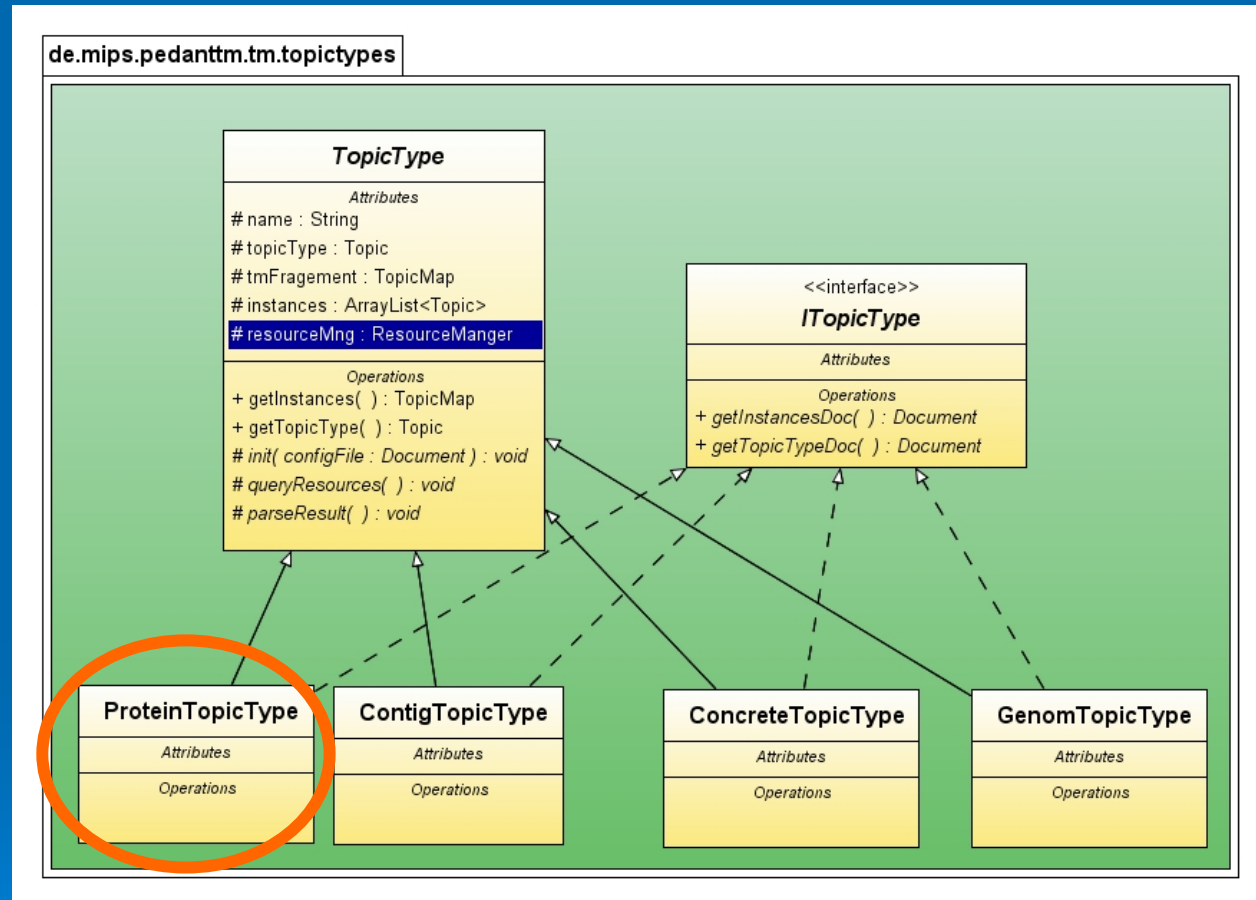
➤ Handler:

- Proxy
- Manages connections
- Execute query methods



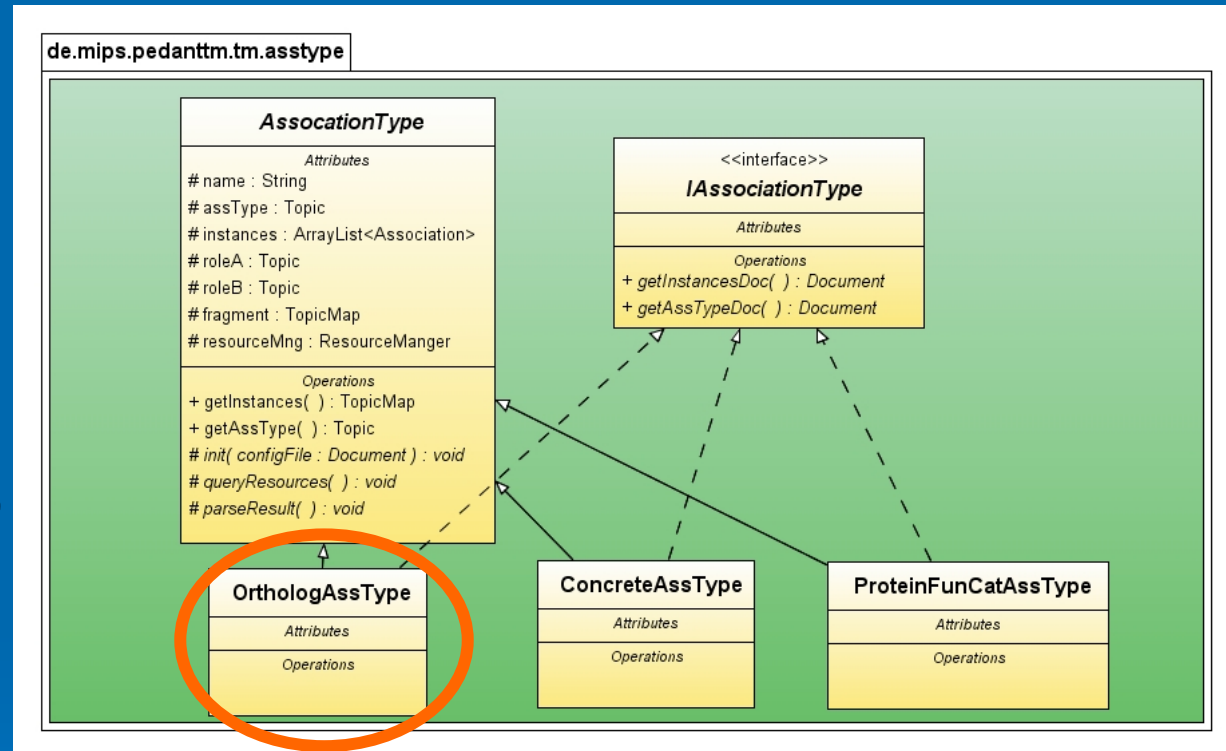
Syntax Tier – Topic Types

- Converts resource specific format into TM fragments
- May access multiple resources (handled by Resource Manager)



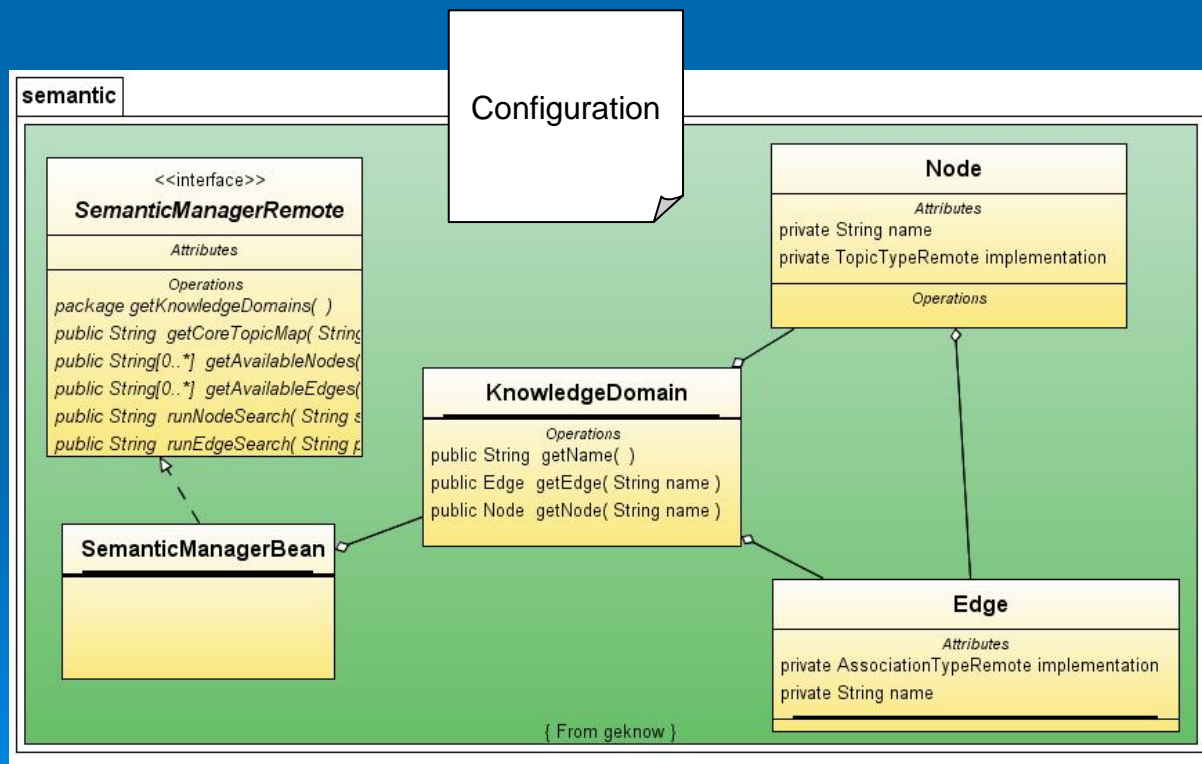
Syntax Tier – Association Types

- Converts resource specific format into TM fragments
- May access multiple resources (handled by Resource Manager)



Semantic Tier

- Responsible for
 - fragment generation
 - Merging
- No programming required (only configuration)



Portal / Portlets (JSR-168)

Protein ao090009000219

Characteristics: ao090009000219	
Description	isocitrate lyase
Molecular Weight	60026.2650200003
Sequence	MGFLEDEDKKYLDDVQAVKAWWTD...SRWRHTEYPSNVQSKKLWKILESNFENKVASFTY
Length	538
Organism	Aspergillus oryzae

Protein ao090009000219

has function	01.05.01.01	
has function	01.01.06.04.02	
has function	01.05.01	
has function	01	
has function	01.05	
is encoded within genome	Aspergillus oryzae	
is encoded by CDS	ao090009000219	Aspergillus oryzae

Portal

- Currently JSF based
 - Caused several problems
- Migration to more generic portlets (XSLT based)

What's Left ?

- GeKnow dedicated to be Open Source
- Visualization ?
- Topic Maps
 - Query language ?
 - Constraint language ?
 - OWL ?
 - XTM fragment exchange ?
- Where are we ?
Just before the killer application
- Where are Topic Maps in Life Sciences
 -
 - (German) National level :
Helmholtz Society funded Systems Biology Initiative
 - Technology platform across Helmholtz centers will use Topic Maps

Conclusion

- Aim: Solving complex biomedical questions
- Semantic knowledge representation
- Textmining
- Integration of heterogeneous distributed data on the fly
(fits well to existing enterprise information systems)
- Representation within JSR-168 portal/portlet solution
- Topic Maps are suited to represent even some 100 millions of topics / associations

Acknowledgements

➤ Filka Nenova
Thorsten Barnickel
Richard Gregory

Matthias Oesterheld

Roland Arnold
Minh-Duc Truong

...

➤ Thomas Rattei

➤ Ulrich Güdener
Martin Münsterkötter

➤ Andreas Ruepp and the
Annotation Group

➤ Funding
Impuls- und
Vernetzungsfonds der
Helmholtz-Gemeinschaft
Deutscher
Forschungszentren e.V.