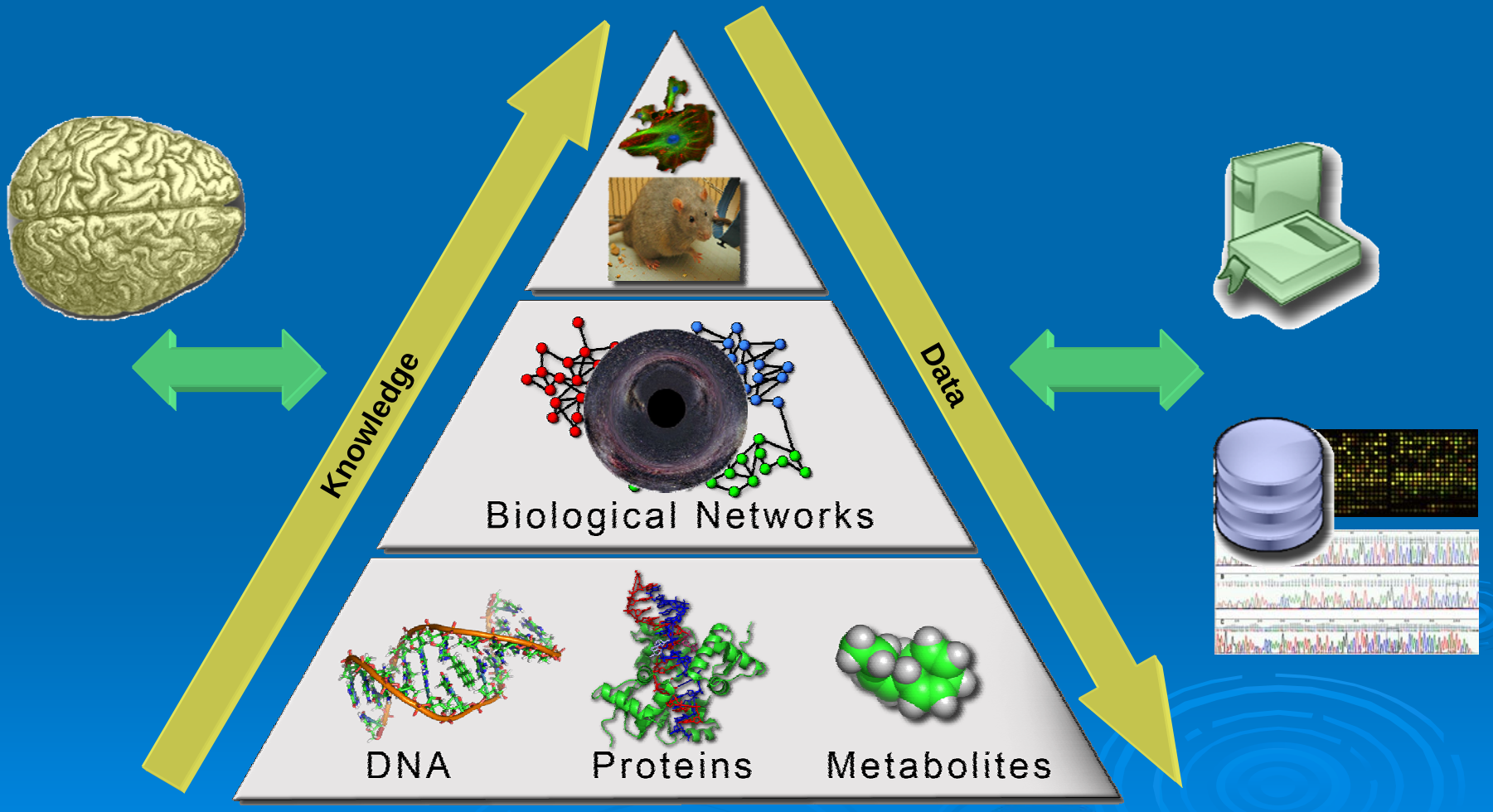
Large Scale Knowledge Representation of Distributed Biomedical Information

Volker Stümpflen

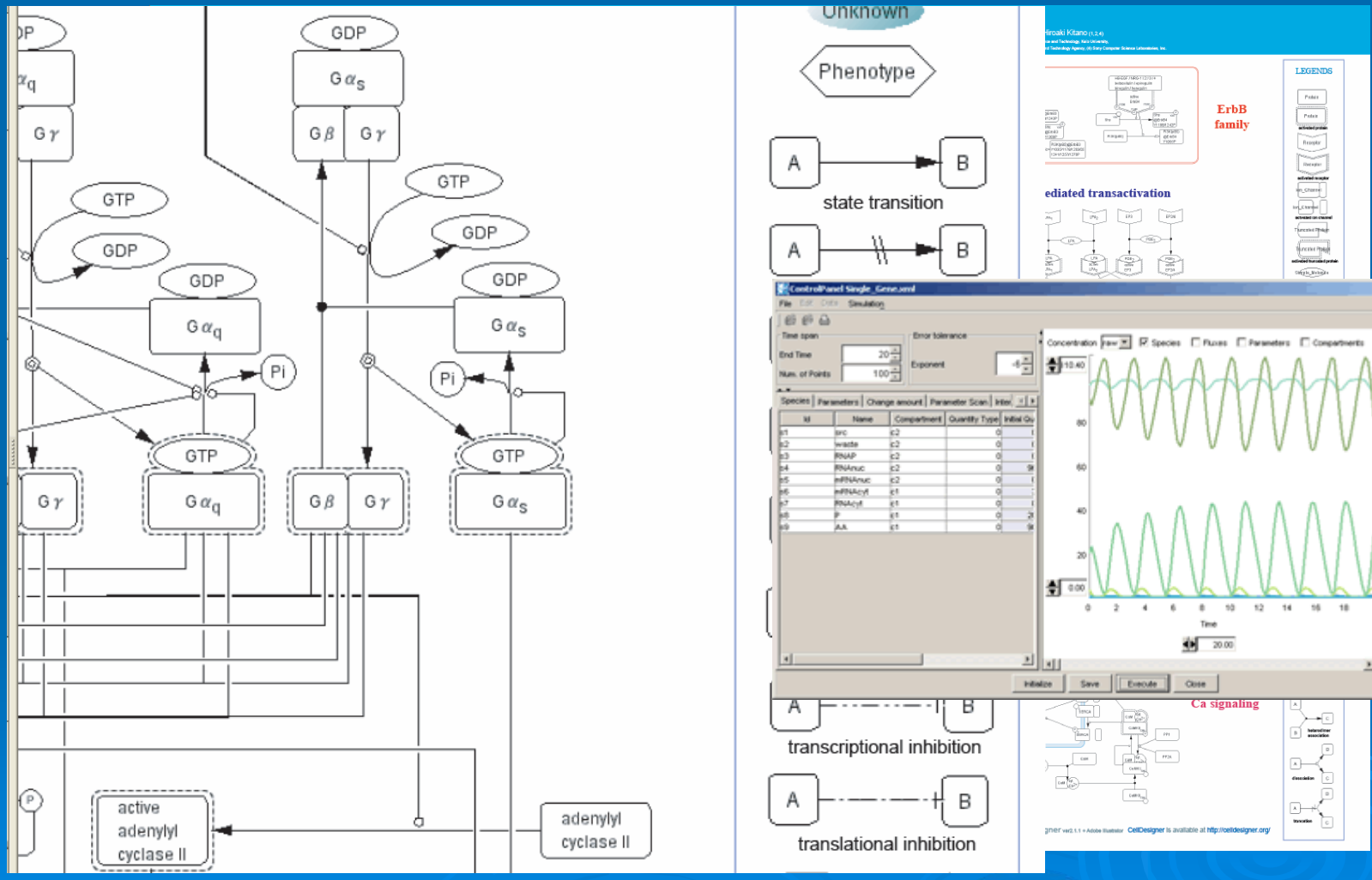
Thorsten Barnickel
Karamfilka Nenova

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

Understanding Complex Biological Systems



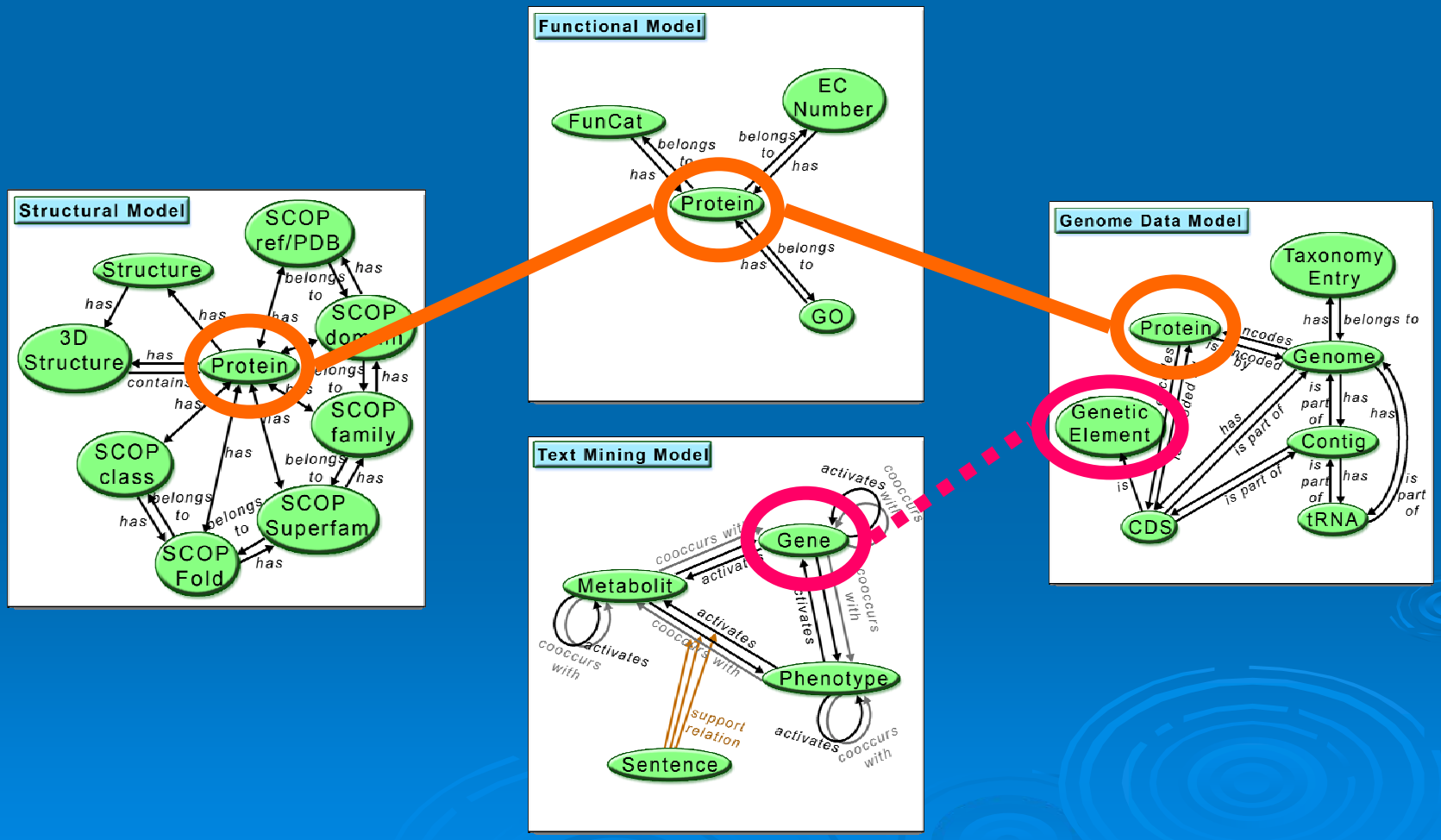
Systems Biology



Questions

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

Merging Knowledge from Different Domains

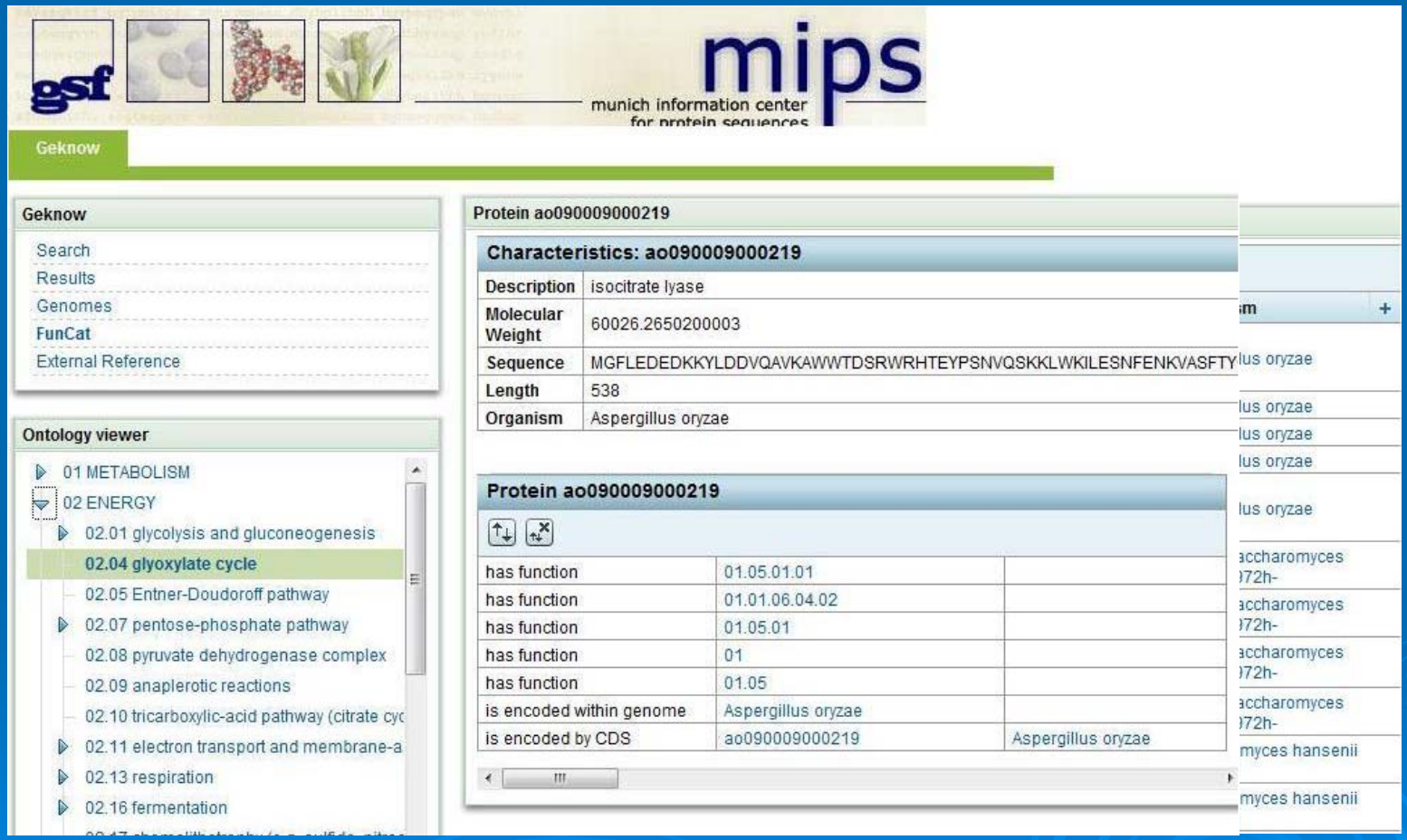


Semantic Structuring Demands for Ontologies

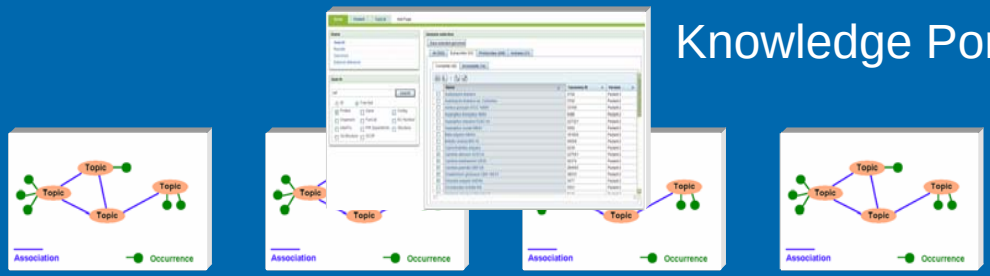
- Life sciences have a long tradition in classification …
- … various ontologies are available and in use

- 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



Semantic Structuring and Knowledge Representation



Knowledge Portal

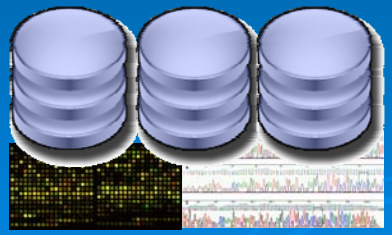
Topic Map Generation

Topic Map Generation

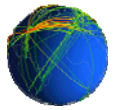
Distributed access system

Textmining

- several hundreds of biomedical resources
- distributed
- > 1-2 PetaByte



Web Service



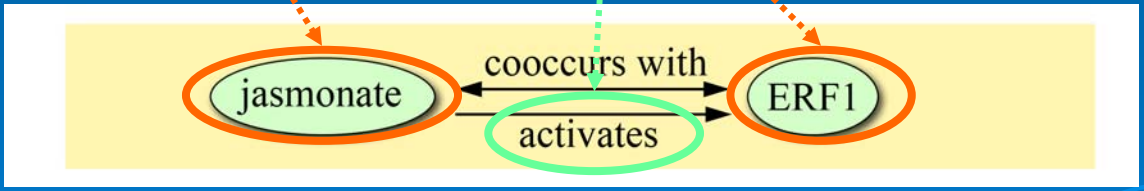
Web Service

Knowledge in Free Text

Free text

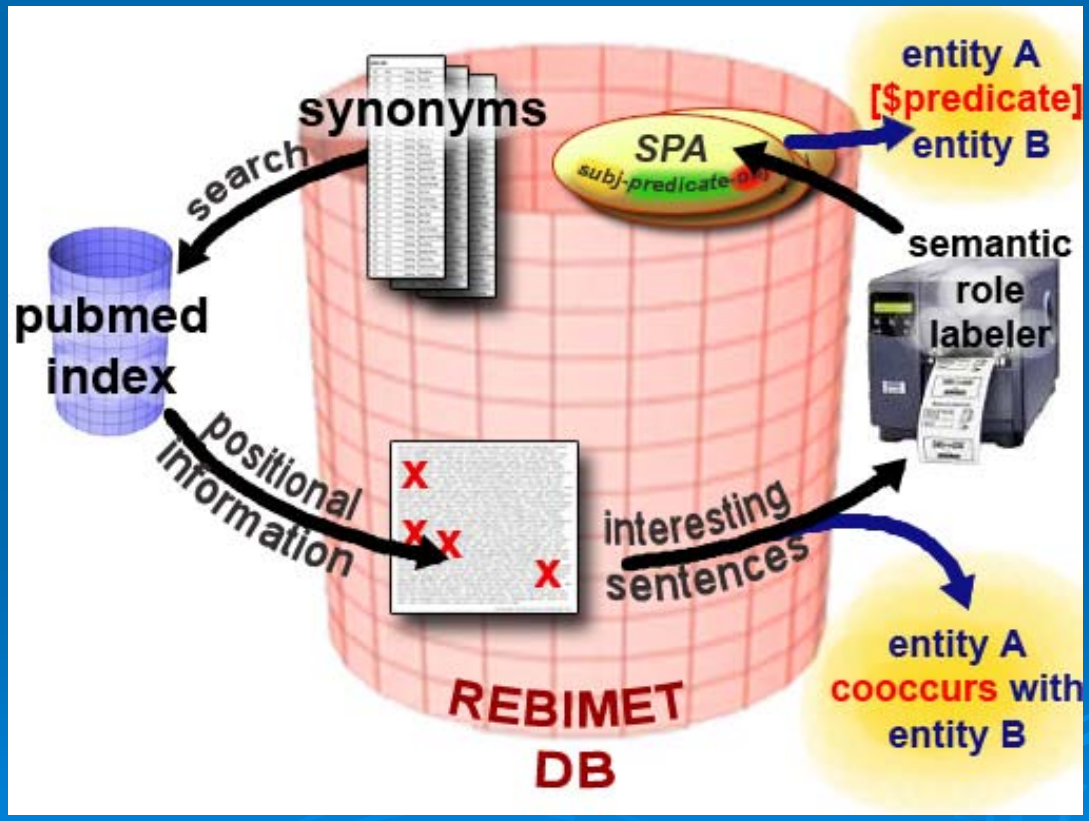
… of pathogen response genes that prevent disease progression.
The expression of ERF1 can be activated rapidly by ethylene
or jasmonate and can be activated synergistically by both hormones.
In addition, both signalling …

Topic Map

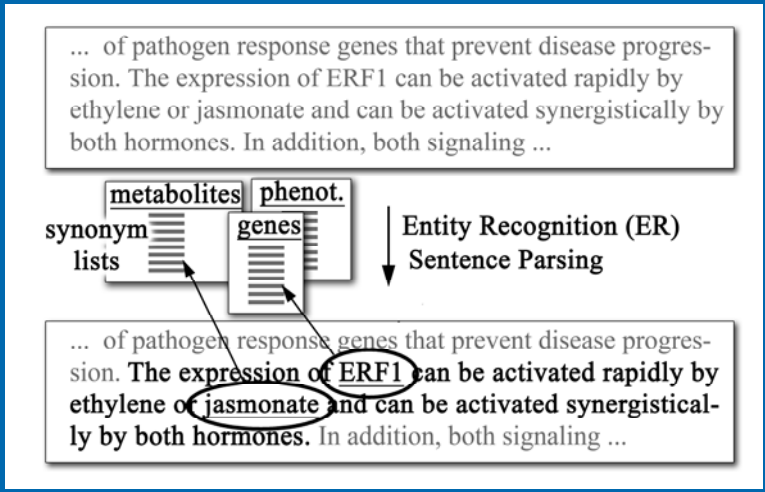


REBIMET

- 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 synergystically by both hormones.

1. Semantic Role Labeling:

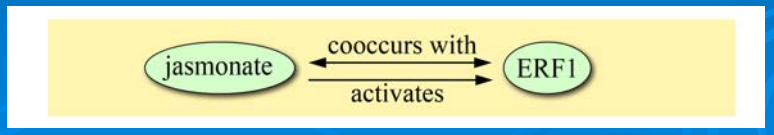
ASSERT tool
(Pradhan S. et al., 2005)

1.1 SPA structure for verb a)
1.2 SPA structure for verb b)

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

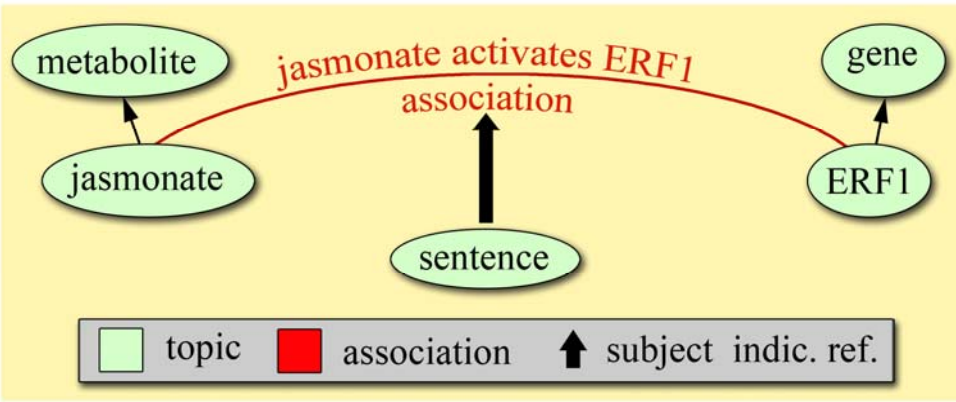
[$_{arg1}$ *The expression of ERF1*] can be activated rapidly by ethylene and jasmonate and [$_{argm\text{-}mod}$ *can*] be [$_{target}$ *activated*] [$_{argm\text{-}tmp}$ *synergystically*] [$_{arg0}$ *by both hormones*]

2. Information Extraction:



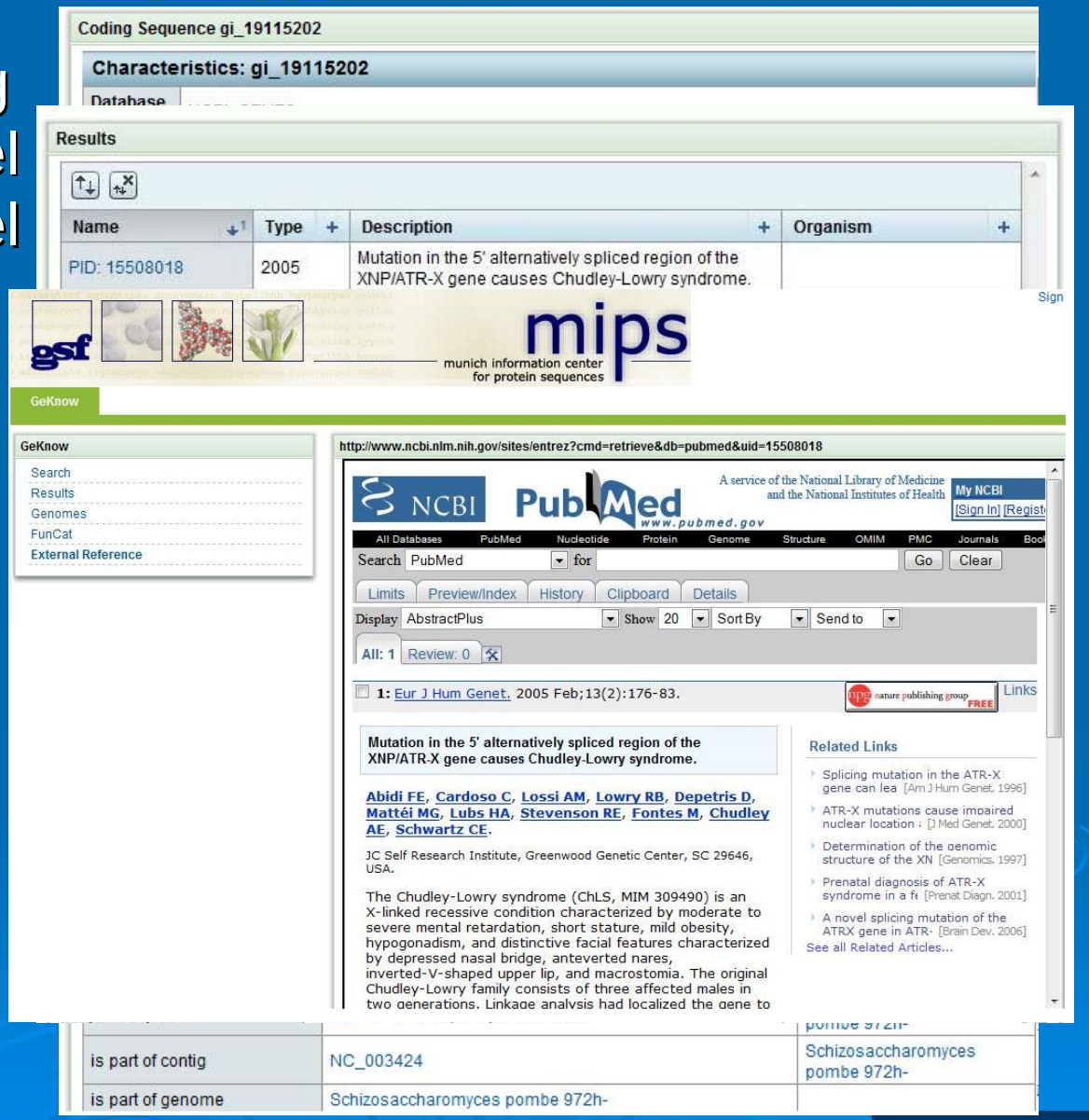
Simplified TM Representation

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

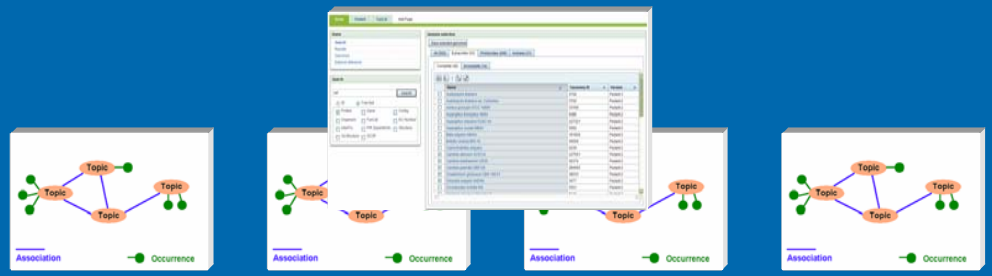


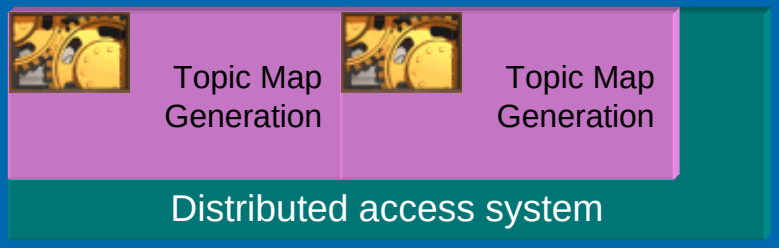
Screenshot Portal

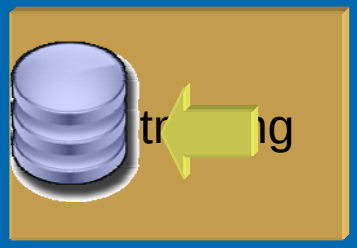
- PSI based merging of textmining model with genome model

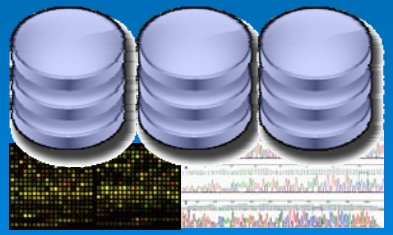


Large Scale Integration and Knowledge Representation







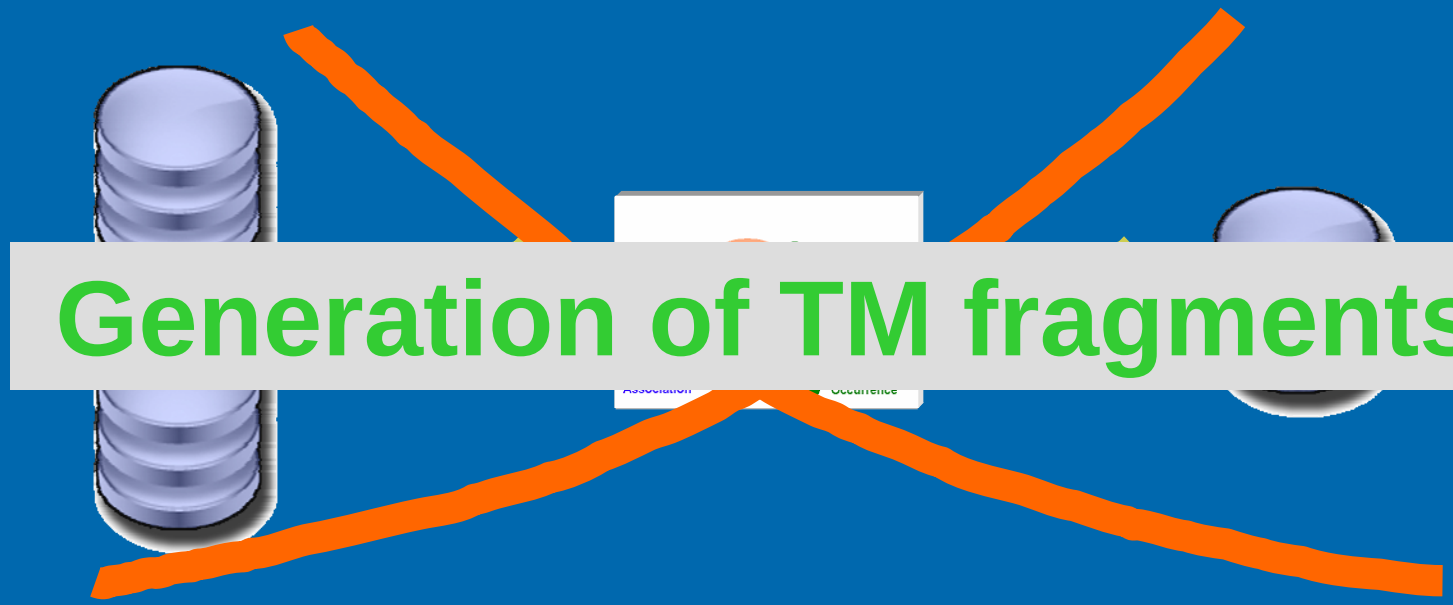


GeKnow: Integration of PEDANT, SIMAP, NCBI data, NCBI PubMed

- 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 ?

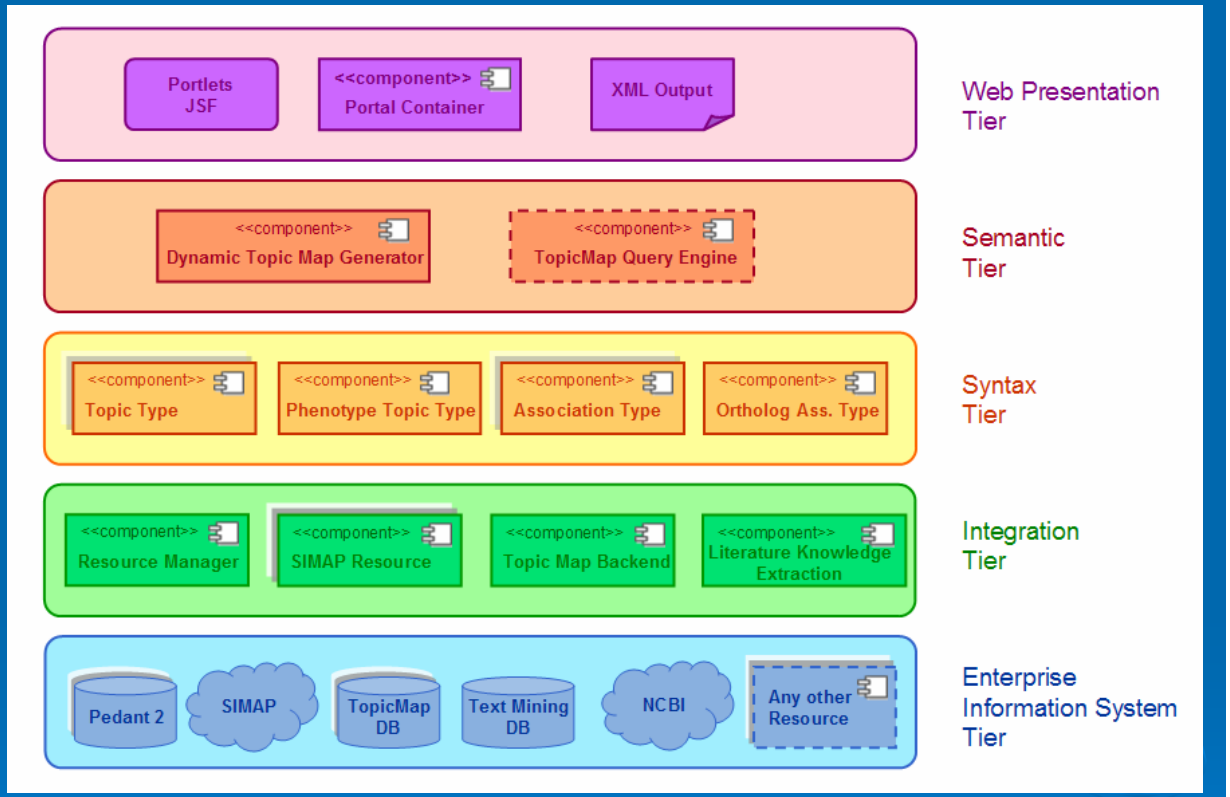


Generation of TM fragments

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

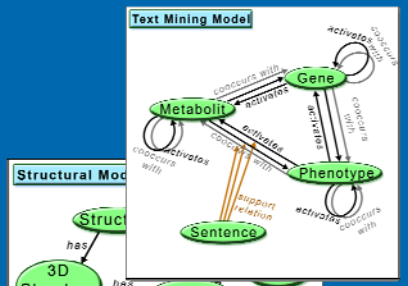
System Architecture (GeKnow)

- 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



Semantic level

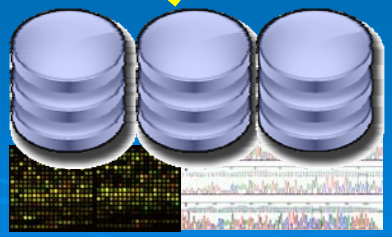
Semantic manager (merging, fragments)

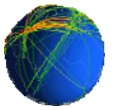
TM

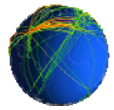
Resource manager

Configuration

Integration level

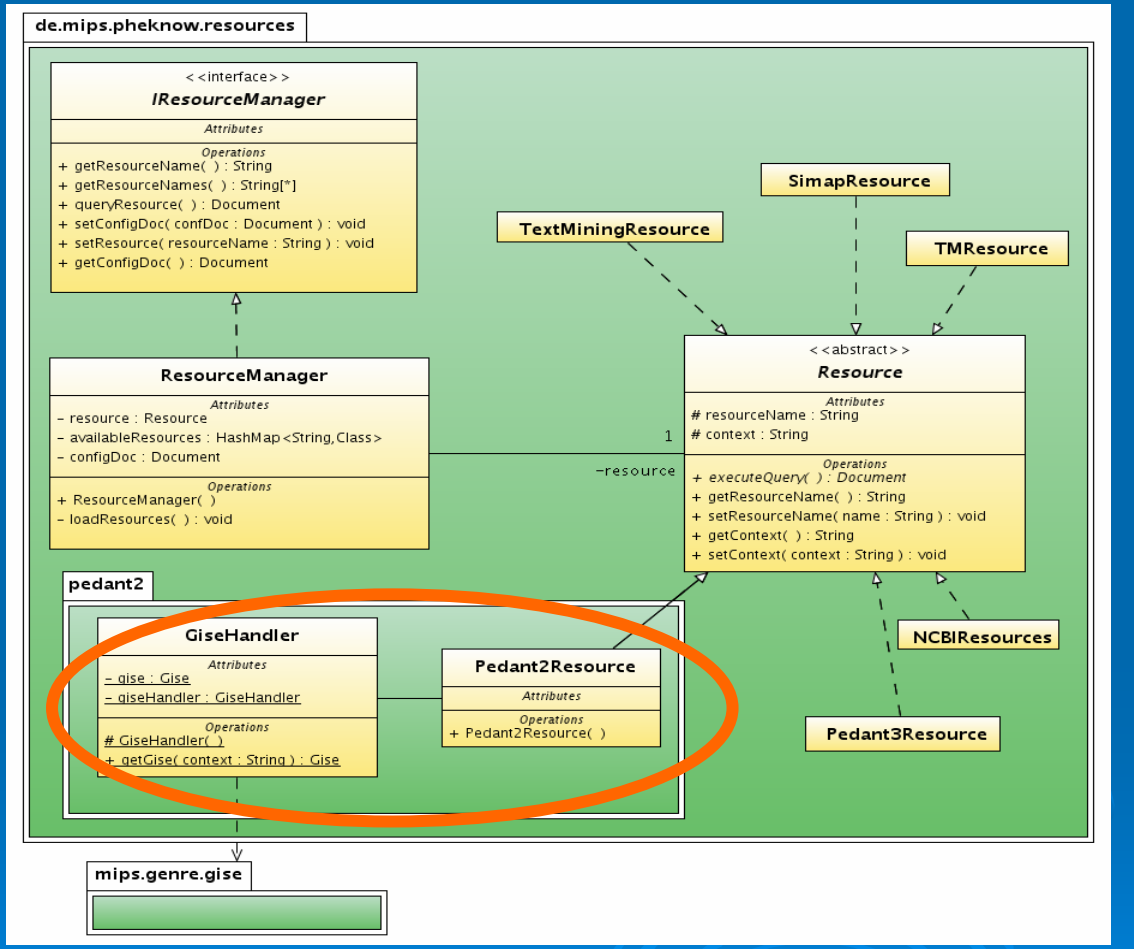



Web Service


Web Service

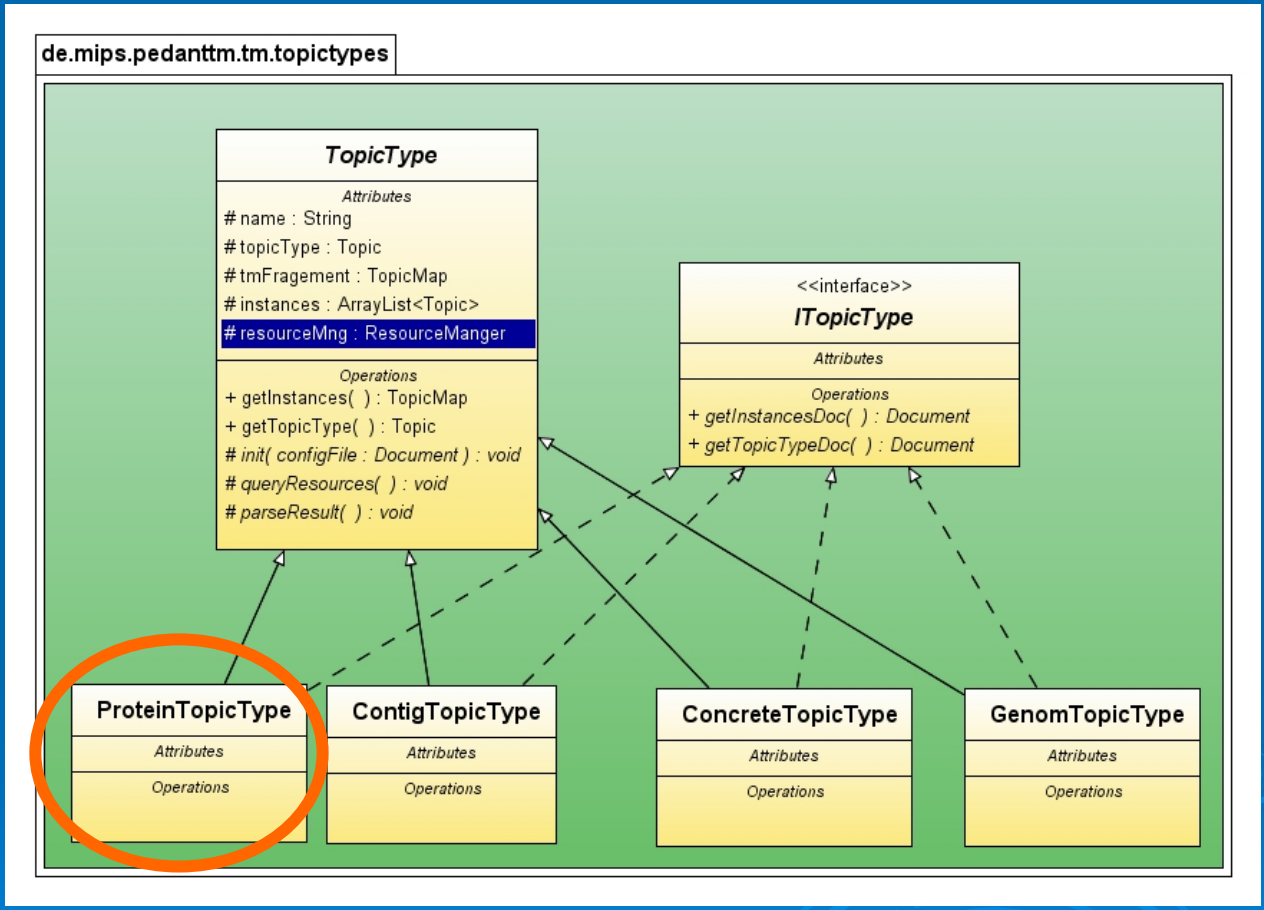
Integration Tier

- Resource:
  - Aware of mapping between topic / association types and methods from data source
- Handler:
  - Proxy
  - Manages connections
  - Execute query methods



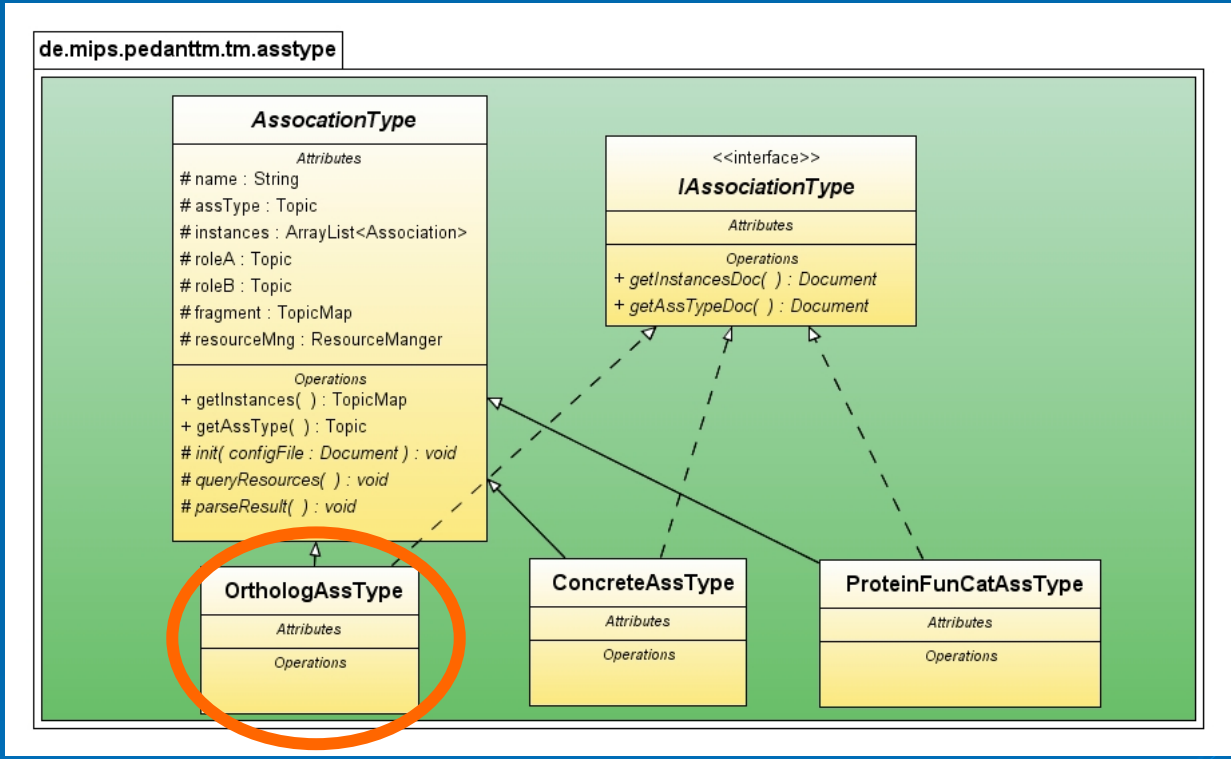
Syntax Tier – Topic Types

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



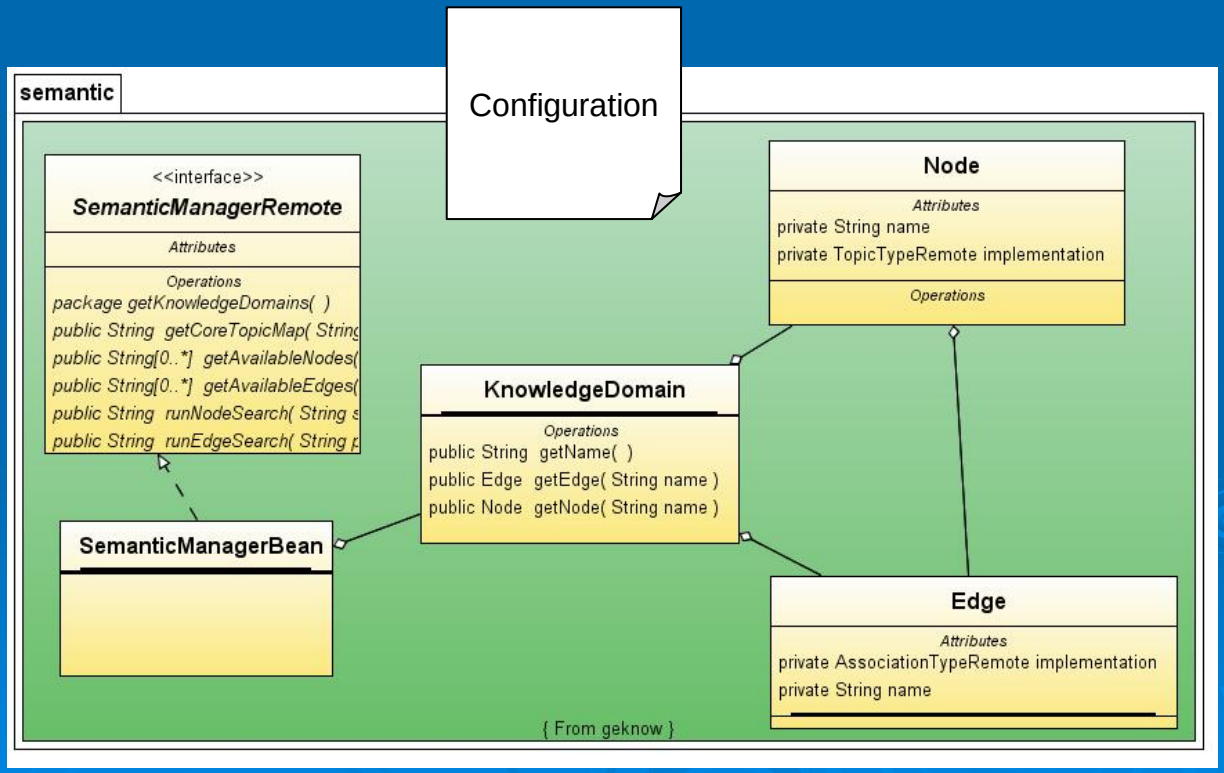
Syntax Tier – Association Types

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

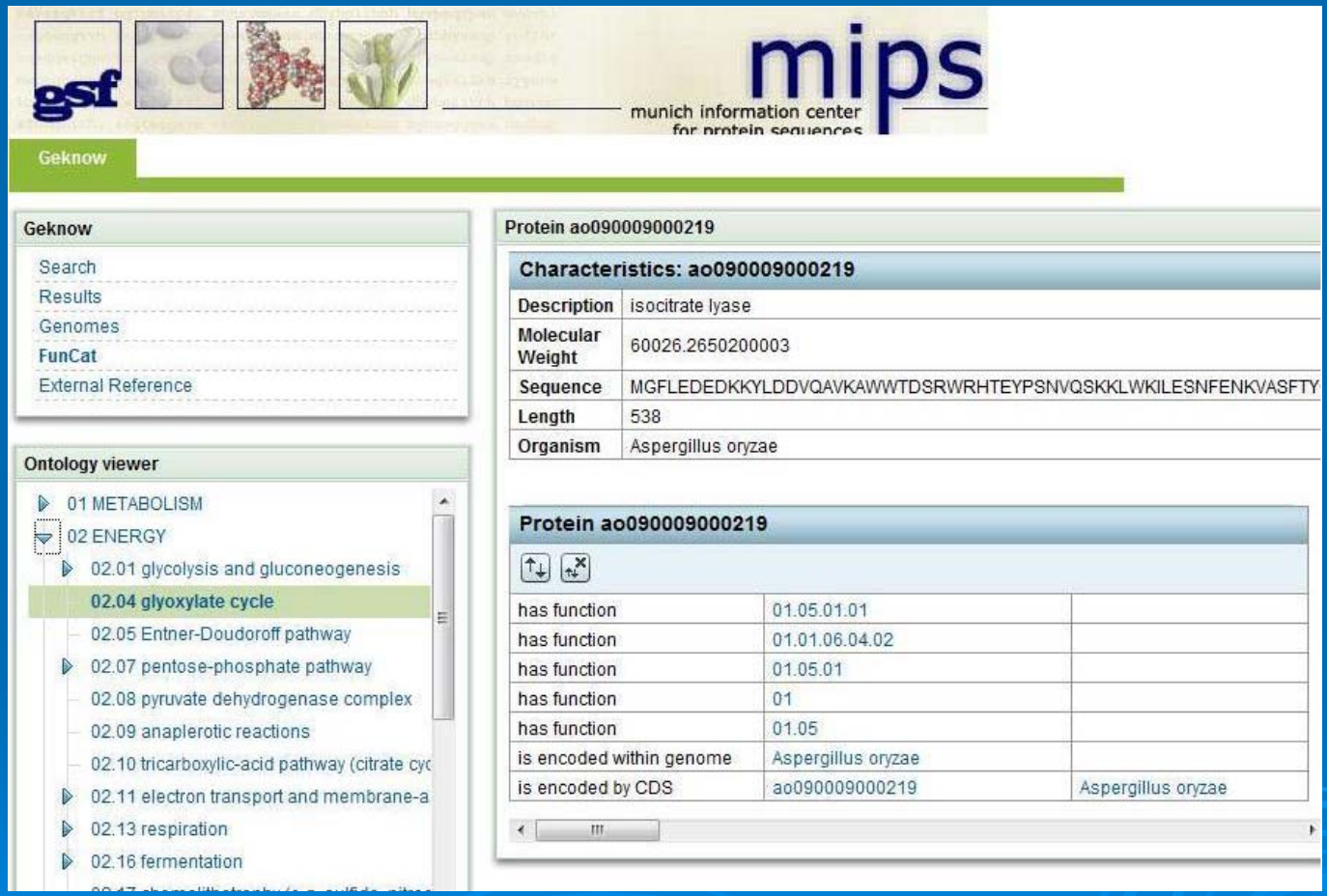


Semantic Tier

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



Portal / Portlets (JSR-168)



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
- Textming
- Integration of heterogenous 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 Güldener
  Martin Münsterkötter
- Andreas Ruepp and the Annotation Group
- Funding
  Impuls- und Vernetzungsfonds der Helmholtz-Gemeinschaft Deutscher Forschungszentren e.V.