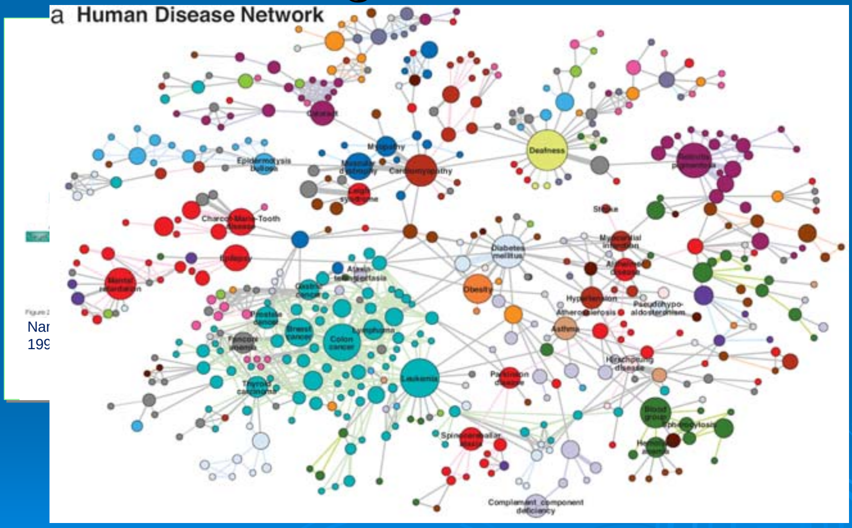
Explaining the Complexity of Life with Topic Maps?!

Volker Stümpflen

MIPS / Institute for Bioinformatics and Systems Biology

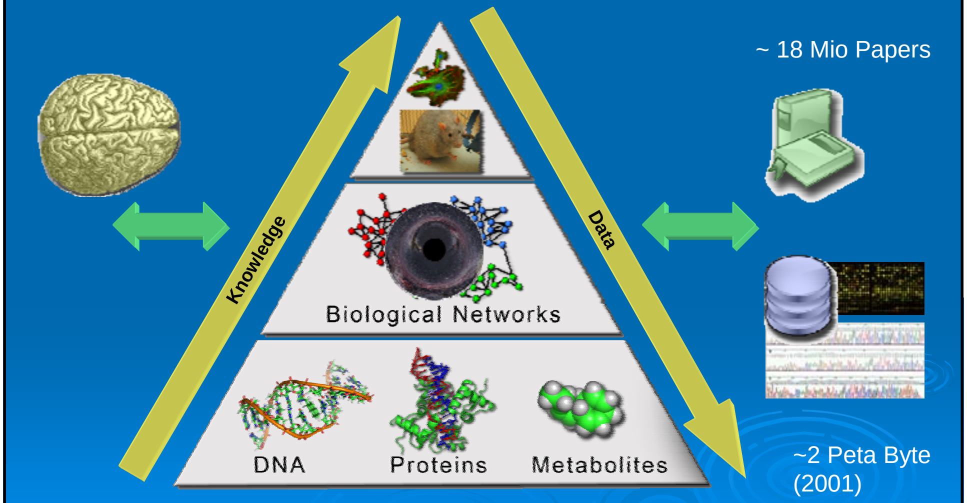
Helmholtz Zentrum München –
German Research Center for Environmental Health

Biological Context



Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabasi A-L (2007) *Proc Natl Acad Sci USA* 104:8685-8690

Understanding
Complex Biological Systems



~ 18 Mio Papers

Knowledge

Data

Biological Networks

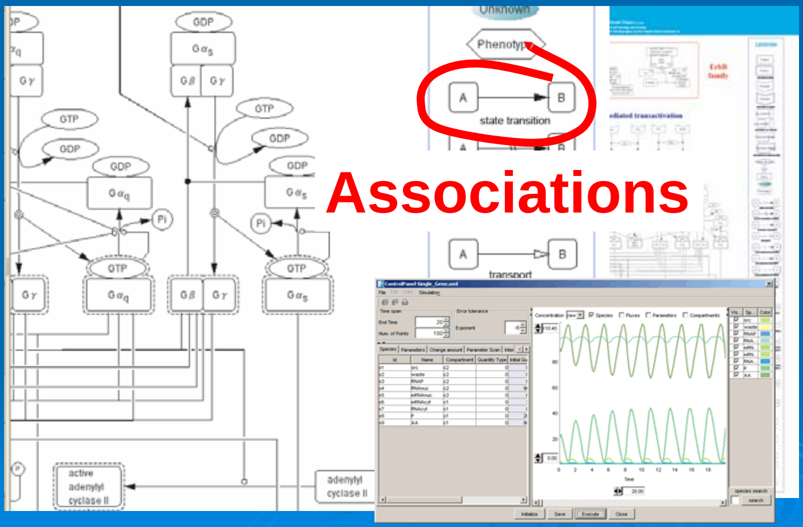
DNA

Proteins

Metabolites

~2 Peta Byte
(2001)

Systems Biology



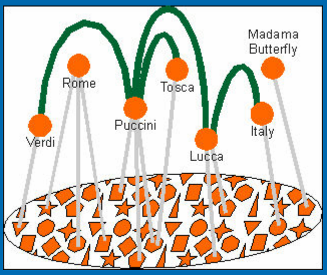
Associations

Questions

- Describing bidirectional associations ?
- Describing and merging different knowledge domains ?
- Ontologies for semantic structuring ?
- Semantic structures from free text ?
- Knowledge representation from distributed resources ?

=> Topic Maps

The Simple Reason



Bidirectional Association to Understand Extended Functional Context

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 cyc
  - 02.11 electron transport and membrane-a
  - 02.13 respiration
  - 02.16 fermentation

Protein ao090009000219

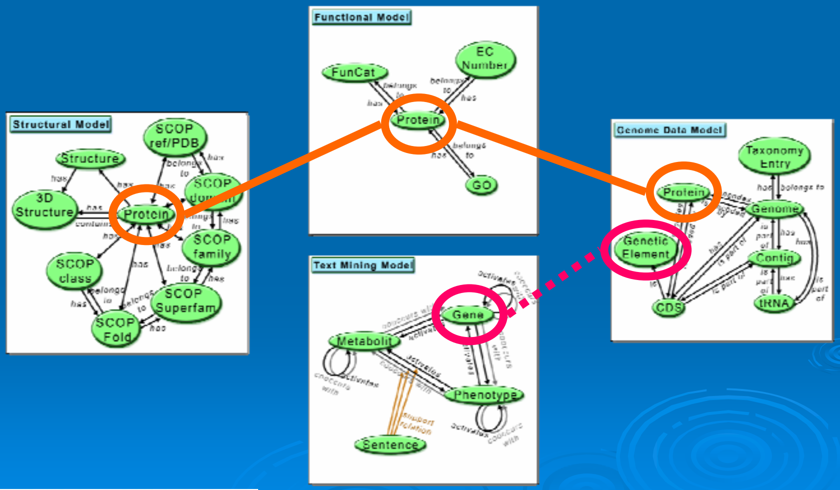
Characteristics: ao090009000219

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

Protein ao090009000219

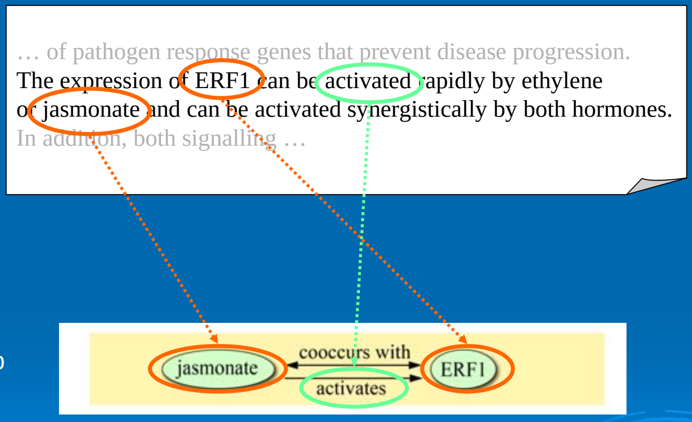
has function	01.05.01.01	
has function	01.01.06.04.02	
has function	01.05.01	
has function	01	
has function	01.05	
is encoded within genome	Aspergillus oryzae	
is encoded by CDS	ao090009000219	Aspergillus oryzae

Merging Knowledge from Different Domains



Associated Knowledge in Free Text

Free text

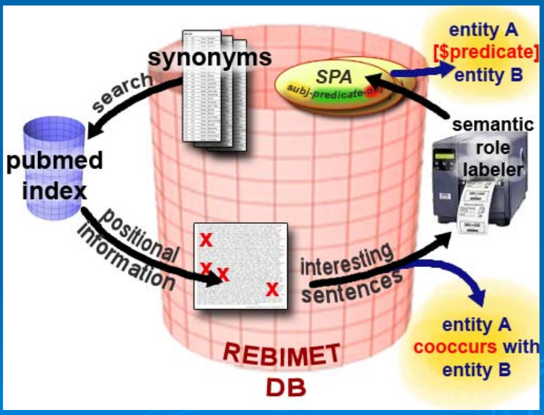


… of pathogen response genes that prevent disease progression.
The expression of ERF1 can be activated rapidly by ethylene
or jasmonate and can be activated synergistically by both hormones.
In addition, both signalling …

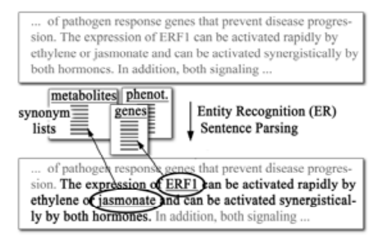
Topic Map

REBIMET

- Relation Extraction from Biomedical Texts



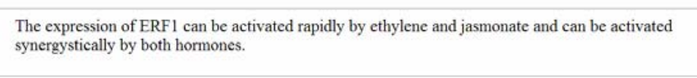
Entity Recognition


... 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 signaling ...

... 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 signaling ...

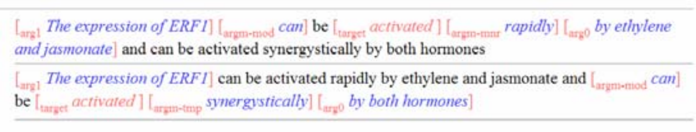
- Identification of relevant biological entities:
  - Based on synonym lists created from terms in taxonomies, gene names, ….
- Realized with Apaches Lucene

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.

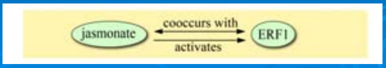
1. Semantic Role Labeling:

1.1 PAS structure for verb a)
1.2 PAS 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

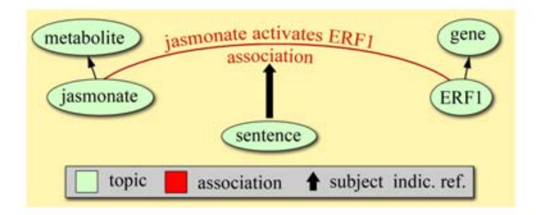
[$_{arg1}$ *The expression of ERF1*] can be activated rapidly by ethylene and jasmonate and [$_{argm\text{-}mod}$ *can*] be [$_{target}$ *activated*] [$_{argm\text{-}tmp}$ *synergystically*] [$_{arg0}$ *by both hormones*]

2. Information Extraction:

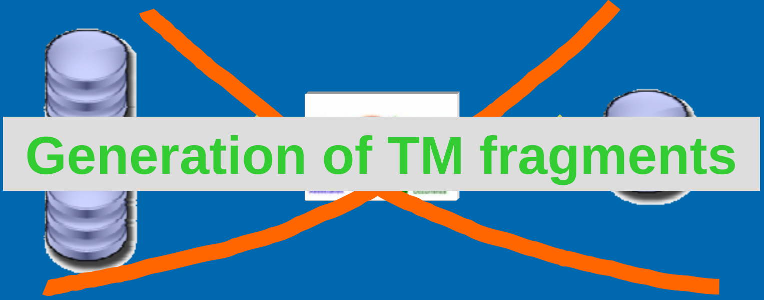


Simplified TM Representation

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



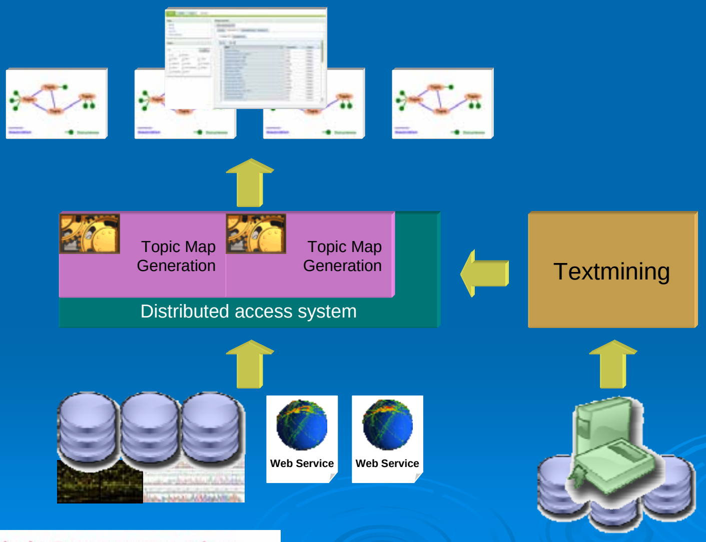
How To Generate the Topic Maps ?



Generation of TM fragments

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

Large Scale Integration and Knowledge Representation



Topic Map Generation

Topic Map Generation

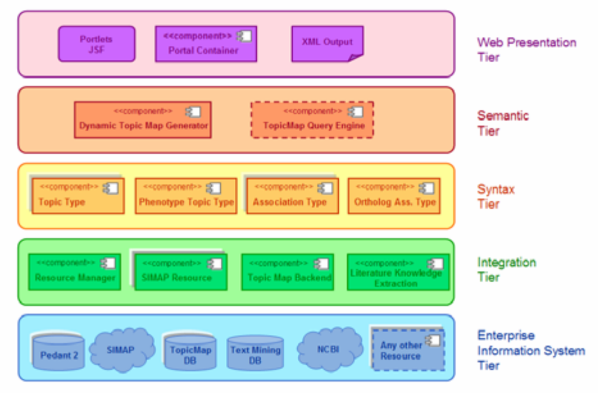
Distributed access system

Textmining

GeKnowME

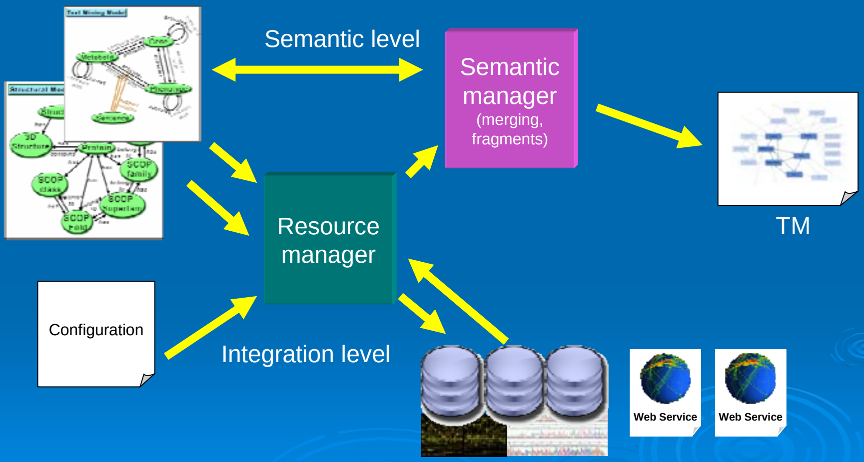
(Generic Knowledge Modeling Environment)

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



Concept:

- Independent semantic layer on top of arbitrary data sources



Semantic level

Semantic manager
(merging, fragments)

TM

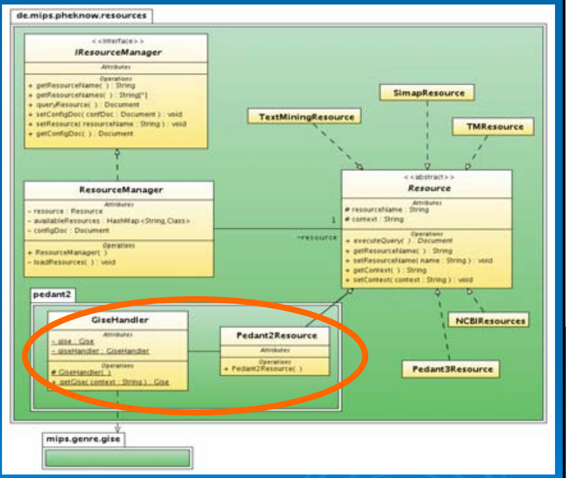
Resource manager

Configuration

Integration level

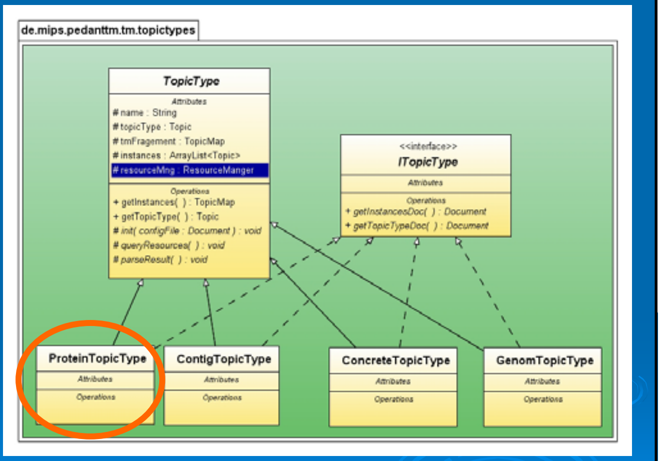
Integration Tier

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



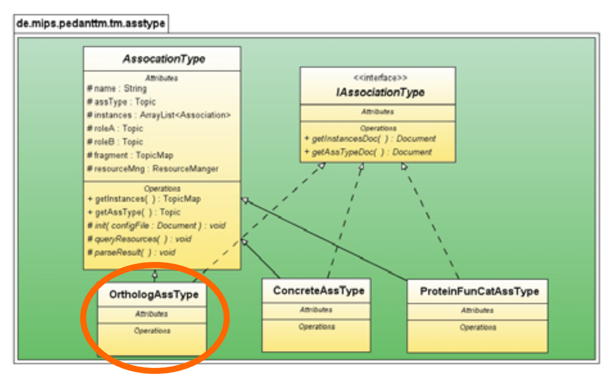
Syntax Tier – Topic Types

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



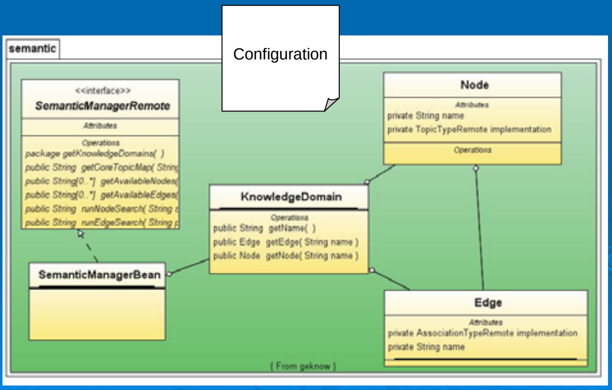
Syntax Tier – Association Types

- 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)



GeKnowME: 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 and their similarities
- NCBI
  - Taxonomy information (some thousands)
- Textmining from PubMed
  - 16 Mio. abstracts, 65 Mio Hits, 15 Mio. Sentences, 13 Mio. PAS structures

- Integration of these data on the fly
- Semantic linking of PEDANT databases with SIMAP and NCBI Taxonomy
- No redundant data

Screenshot Portal

- PSI based merging of textmining model with genome model

Coding Sequence gi_19115202

Characteristics: gi_19115202

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.	

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

Search: PubMed

Display: AbstractPlus

Show: 20

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

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, Mattéi 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 lea [Am J Hum Genet. 1998]
- ATR-X mutations cause impaired nuclear location ... [J Med Genet. 2003]
- Determination of the genomic structure of the XN [Genomics. 1997]
- Prenatal diagnosis of ATR-X syndrome in a h [Prenat Diagn. 2001]
- A novel splicing mutation of the ATRX gene in ATR- [Brain Dev. 2006]

See all Related Articles...

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

Proudly we went to the bench biologists and succesfully …

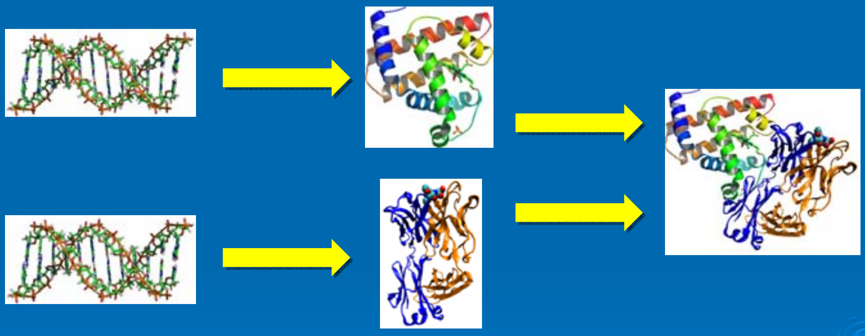
… we failed

Why ?

- If you can‘t transport within 5 seconds your message you‘re gone
(independent of the quality of the content)

- In our specific case the context will not be clear by providing just text with hyperlinks

Crash Course Biochemistry

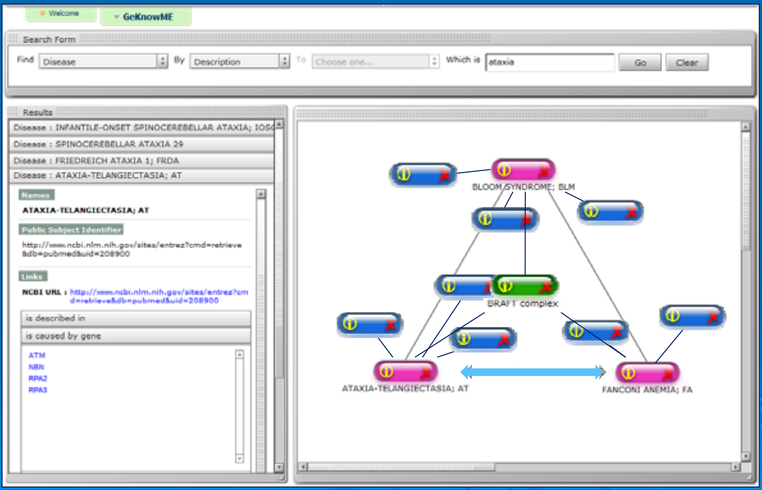


Gene

Protein

Protein Complex

Context From Web Based Graphical Interface



Find: Disease
By: Description
To: Choose one...
Which is: ataxia

Disease : INFANTILE-ONSET SPINOCEREBELLAR ATAXIA; IOS

Disease : SPINOCEREBELLAR ATAXIA 29

Disease : FRIEDREICH ATAXIA 1; FRDA

Disease : ATAXIA-TELANGIECTASIA; AT

Names

ATAXIA-TELANGIECTASIA; AT

Public Subject Identifier

http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=retrieve&db=pubmed&uid=208900

Links

NCBI URL : http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=retrieve&db=pubmed&uid=208900

is described in

is caused by gene

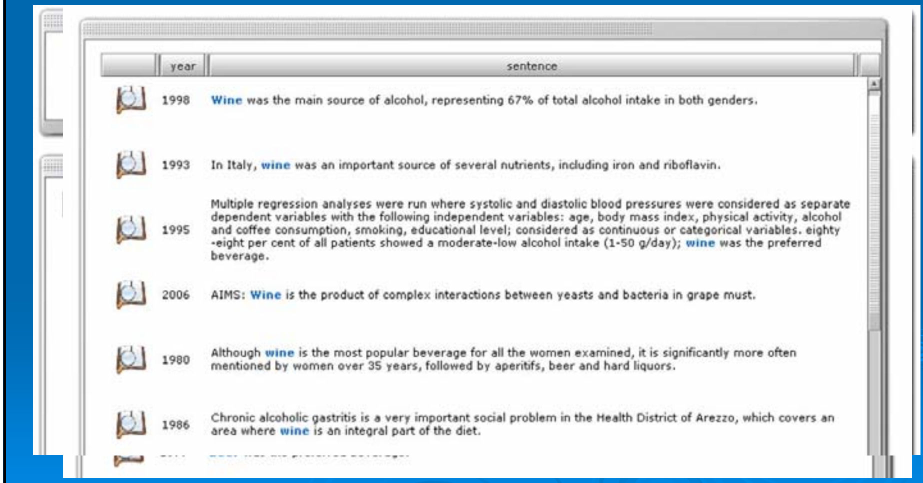
ATM
NBN
RPA2
RPA3

Conclusion

- Can we explain the complexity of life ?
  - ….
- However:
  - TMs help us to model and associate information …
  - … in a way we WANT and NEED
  - We can utilize existing and open technologies to work with them
  - Topic Maps are suited to represent even some 100 millions of topics / associations
- Topic Maps will help us to understand at least the next level of complexity

A Final One:

Do it for the user, not the technology



year	sentence
1998	Wine was the main source of alcohol, representing 67% of total alcohol intake in both genders.
1993	In Italy, wine was an important source of several nutrients, including iron and riboflavin.
1995	Multiple regression analyses were run where systolic and diastolic blood pressures were considered as separate dependent variables with the following independent variables: age, body mass index, physical activity, alcohol and coffee consumption, smoking, educational level; considered as continuous or categorical variables. eighty -eight per cent of all patients showed a moderate-low alcohol intake (1-50 g/day); wine was the preferred beverage.
2006	AIMS: Wine is the product of complex interactions between yeasts and bacteria in grape must.
1980	Although wine is the most popular beverage for all the women examined, it is significantly more often mentioned by women over 35 years, followed by aperitifs, beer and hard liquors.
1986	Chronic alcoholic gastritis is a very important social problem in the Health District of Arezzo, which covers an area where wine is an integral part of the diet.

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.