Beyond Search >>

Recognition of Chemical Entities in Scientific Literature

Charles Huot, Chief Operating Officer

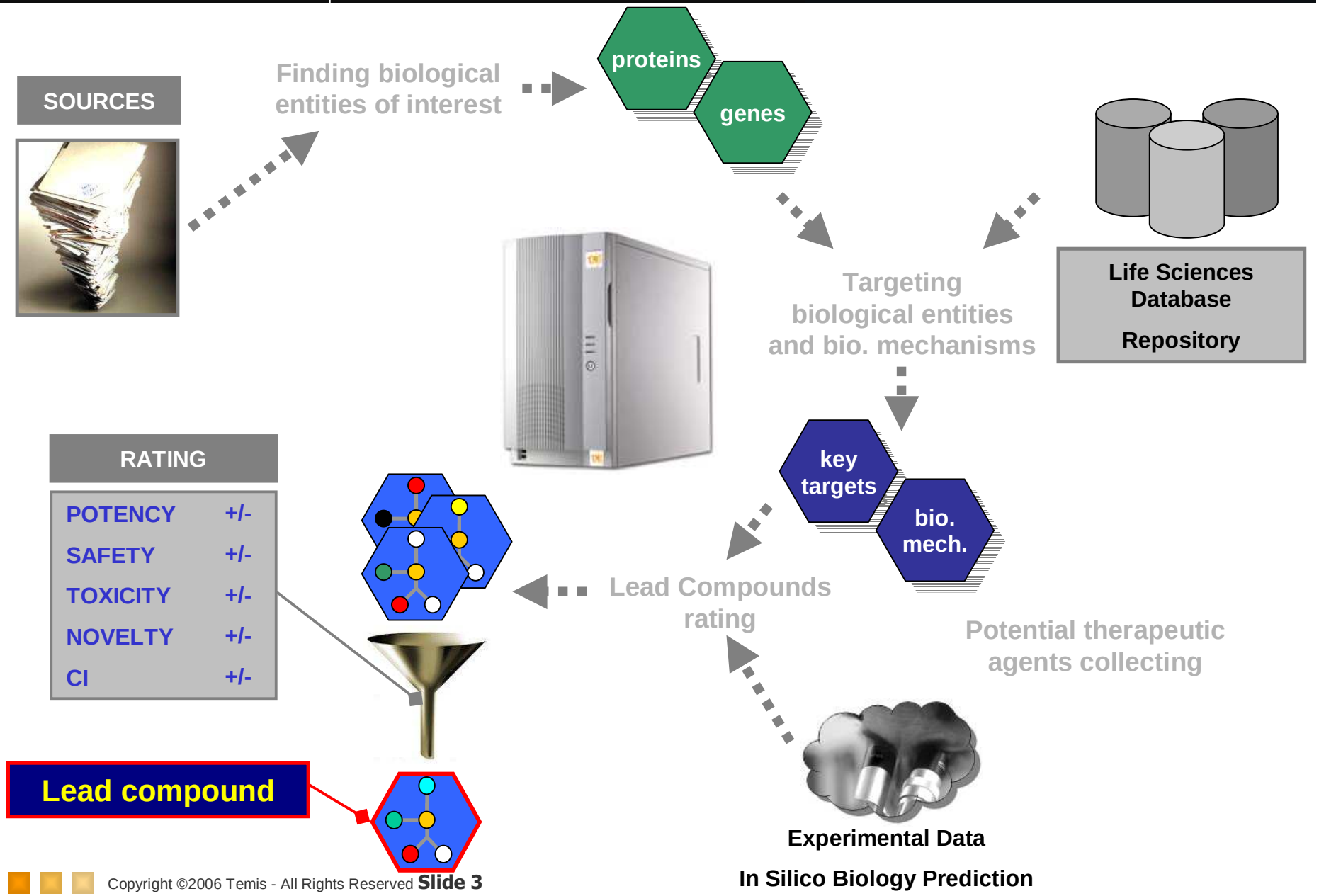
& Co-Founder

International Chemical Information Conference -

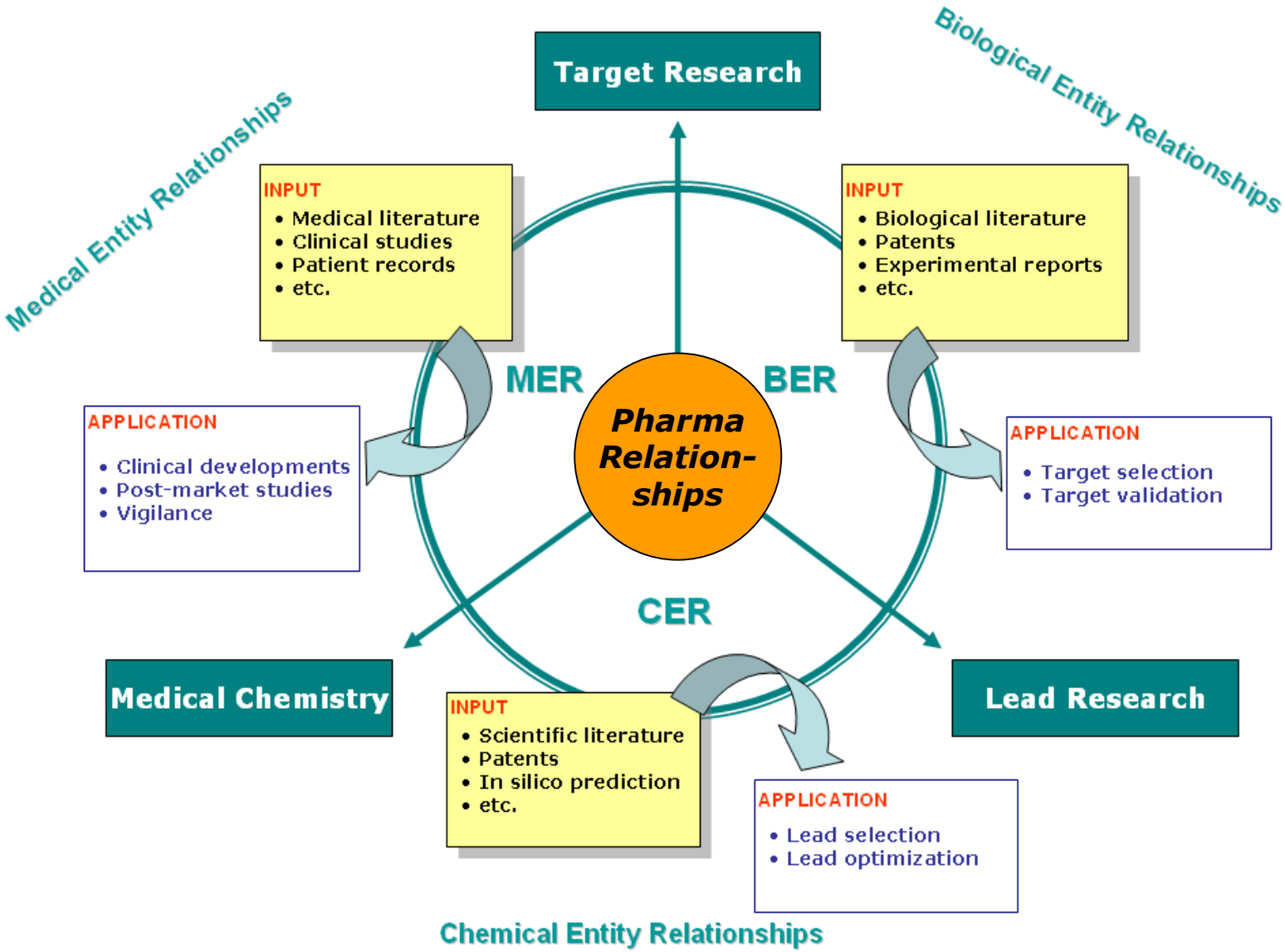
Nîmes, October 24th, 2006

Agenda

1. **Drug Discovery Process**
2. **Recognition of Chemical Entities**
3. **Case study - Scenario Highlight**
4. **TEMIS Presentation**



Life Sciences Skill Cartridges™

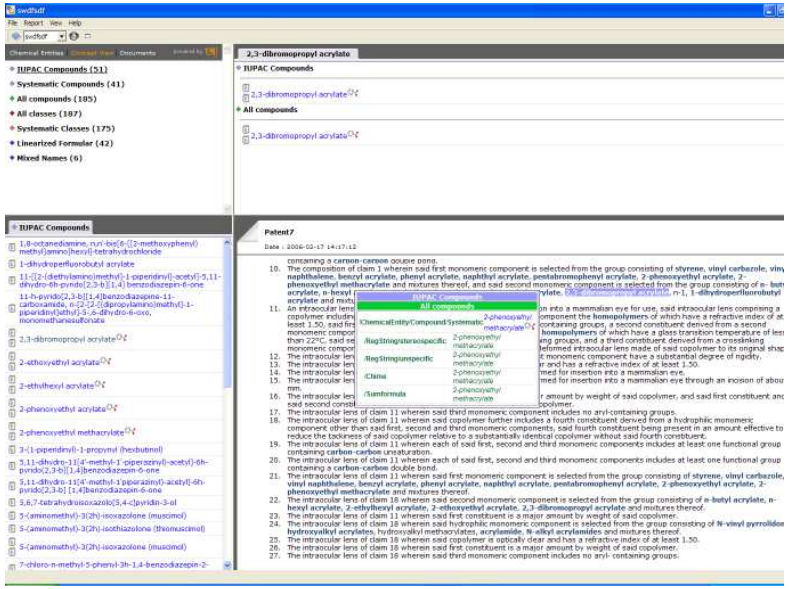


Agenda

1. **Drug Discovery Process**
2. **Recognition of Chemical Entities**
3. **Case study - Scenario Highlight**
4. **TEMIS Presentation**

Recognition of Chemical Entities

- A new information discovery weapon
  - To analyze chemistry-related documents



- Key features
  - Powerful recognition & extraction of chemical entities
  - Highlighting within text for intelligent navigation
  - Display graphical representation of the molecular structure
- Result of close cooperation between MDL & TEMIS
  - Extraction & structure recognition performed thru MDL technology embedded within a TEMIS *Skill Cartridge*™
  - Architecture, workflow & display built by TEMIS

Recognition of Chemical Entities

1. Recognition of different name representations
2. Name-To-Structure layer
   - Integration of a 3rd party tools (CambridgeSoft,…) and database look-up (internal, Beistein, ACD,…)
   - Customizable dictionary incl. pattern recognition
3. Structure-based normalization via Registration String

Recognition of Chemical Entities

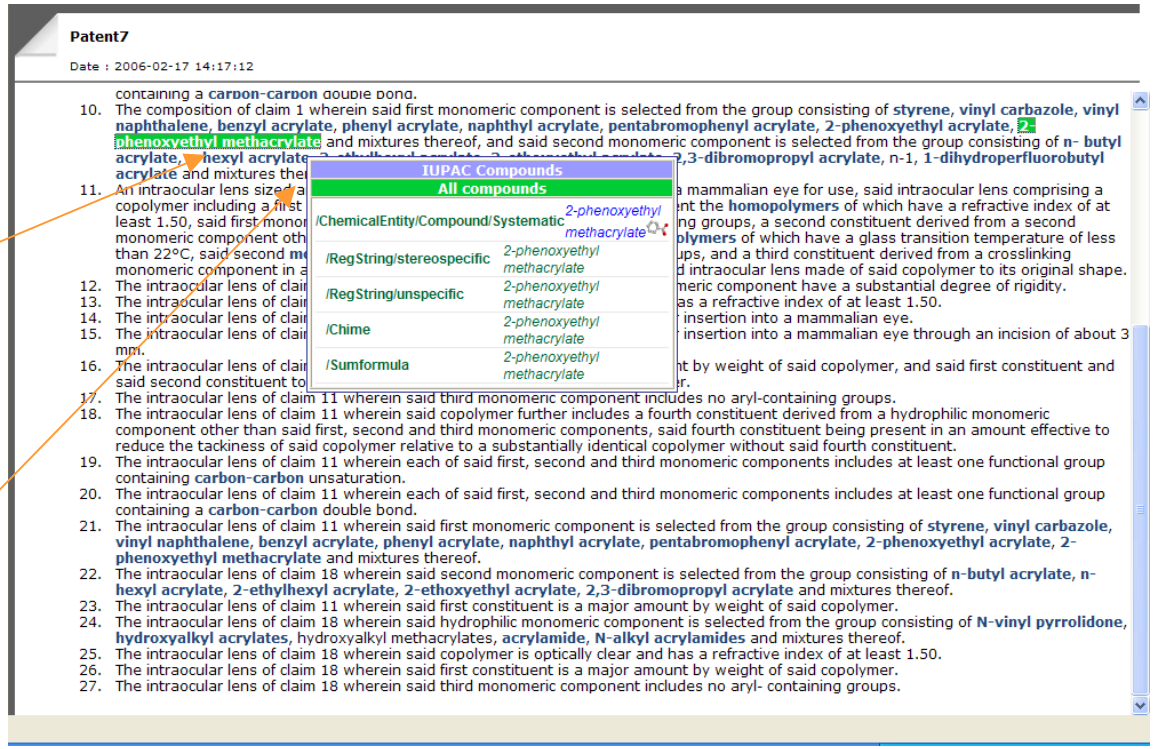
- 95% of chemical compounds extracted
  - Systematical Chemical Names according to IUPAC
  - Inverted Chemical Names according to CAS for specific stems
  - Trivial names, element names, element symbols
  - Half-systematic chemical names (with trivial fragments)
  - Explicitly allowed abbreviations
  - Chemical formula-like constructions, linearized structural formulas
  - Compound names from systematic and/or trivial fragments
  - Amino acids and nucleic acids
  - Chemical compound classes
  - Enzyme names according to IUPAC

Recognition of Chemical Entities

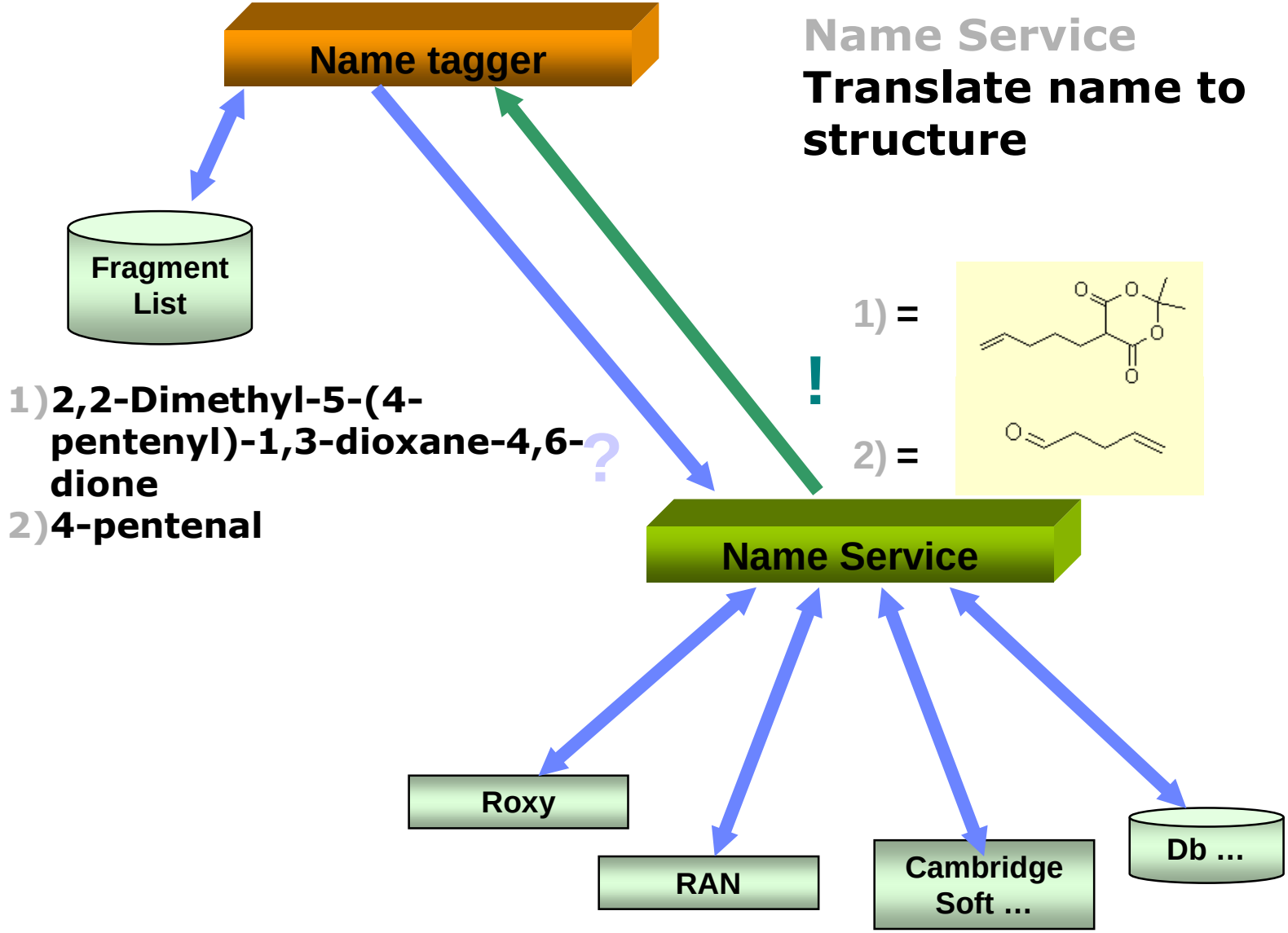
- 95% of chemical components extracted
- Highlighting of chemical entities within text
  - Enable instant & rapid visualization of most-valuable information within large documents

Highlighting

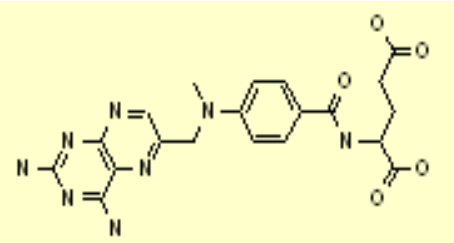
On-demand explanations



Name-To-Structure layer



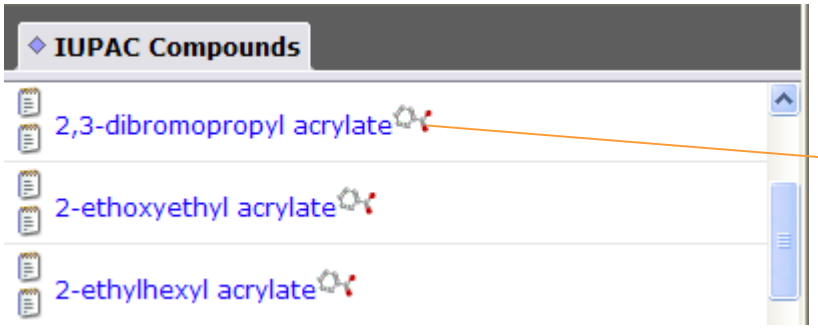
Structure-based normalization

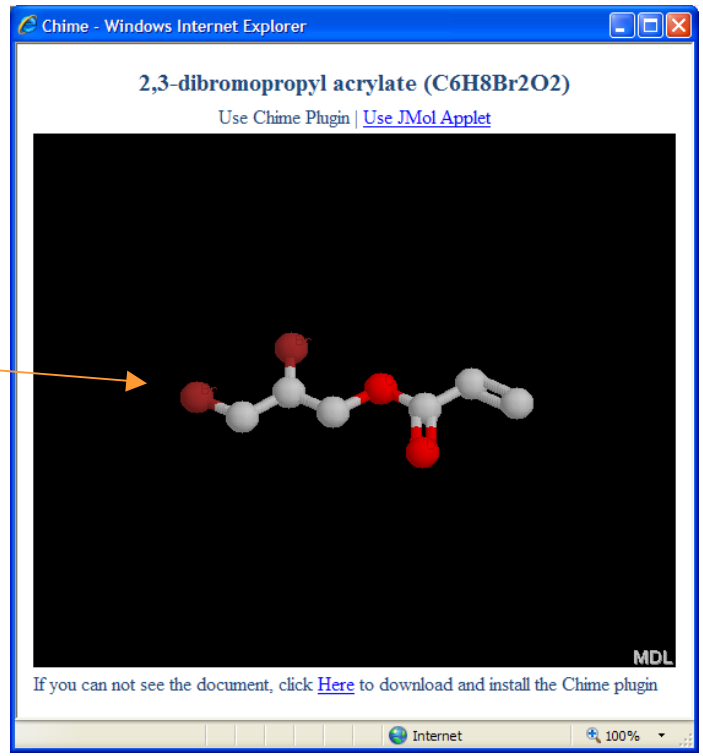
Chemical name	structure
Amethopterin 4-amino-N10-methylpteroylglutamic acid MEXATE METHOTREXATE 2-{4-[(2,4-Diamino-pteridin-6-ylmethyl)-methyl-amino]-benzoylamino}-pentanedioic acid	

- REGSTRING
  - Unique fingerprint
  - Easy duplicate check

Recognition of Chemical Entities

- 95% of chemical components extracted
- Highlighting of chemical entities within text
- Link chemical name to associated structure
  - Thru structure graphical display





Value Proposition

Enables the user to search, navigate, analyze the scientific literature …

- … interactively …
  - From full-text documents, abstracts and database text fields
- … with a Life Sciences specific view …
  - Detection of Medical, Biological and Chemical entities (e.g. diseases, proteins, chemical names)
  - Translation of the entities to their native representations (e.g. chemical structures & bio-sequences)
  - Recognition of entity relationships (e.g. drug-disease or protein-ligand)
- & Support decisions with facts
  - Scientists, information professionals, patent officers are supported to backup their intuition with factual information in a project.

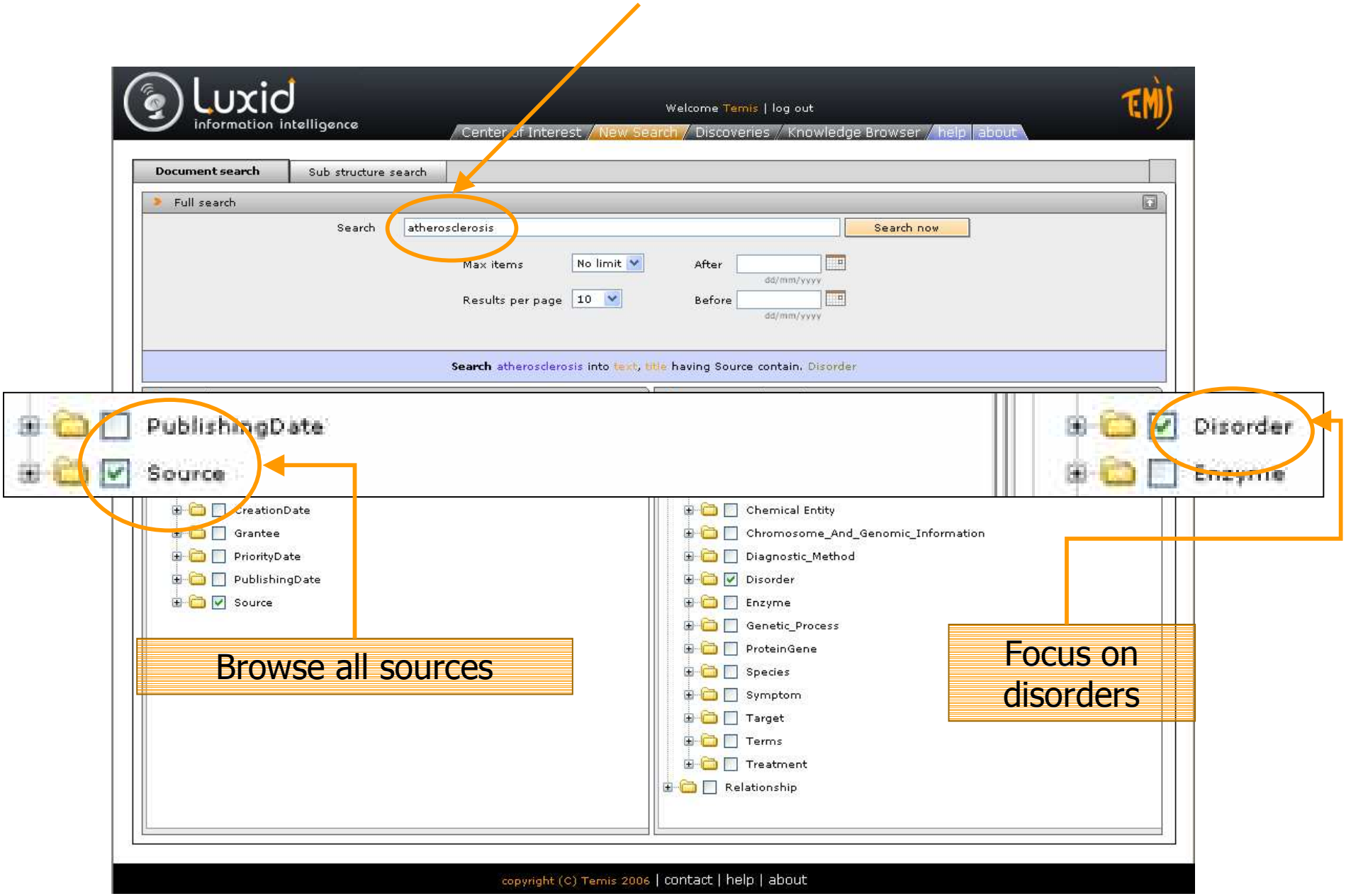
Agenda

1. **Drug Discovery Process**
2. **Recognition of Chemical Entities**
3. **Case study - Scenario Highlight**
4. **TEMIS Presentation**

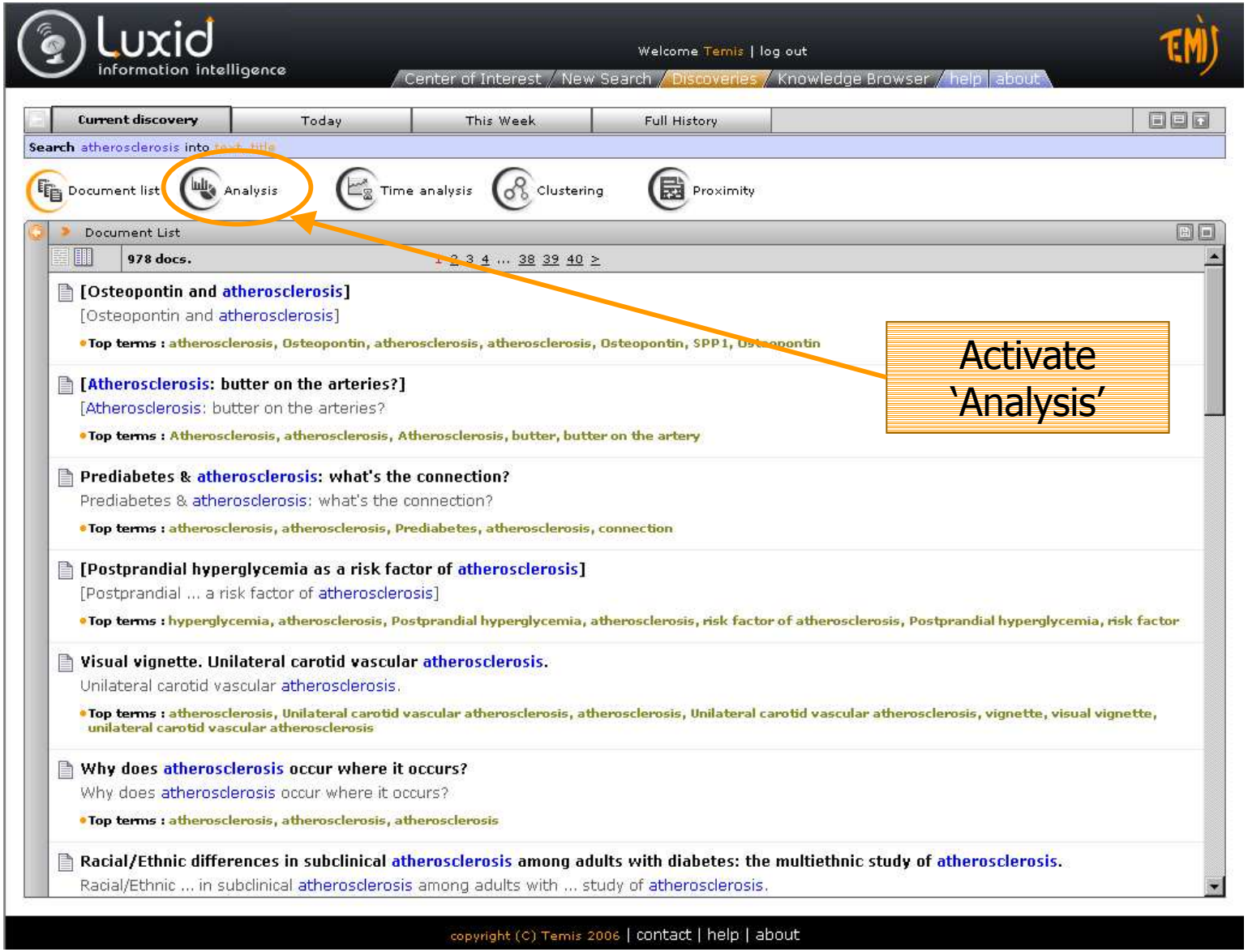
Scenario Highlight

- You have 2 hours to produce a synthetic report for your manager and give your colleagues first elements in an crisis management meeting
- The topic? atherosclerosis
  - which potential targets are mentioned ?
  - what are the latest studied ones
  - which associated diseases fall in our therapeutic areas ?
  - which treatments were successfully used on these associated diseases
  - can we find potential new compounds similar to these known treatments ?)

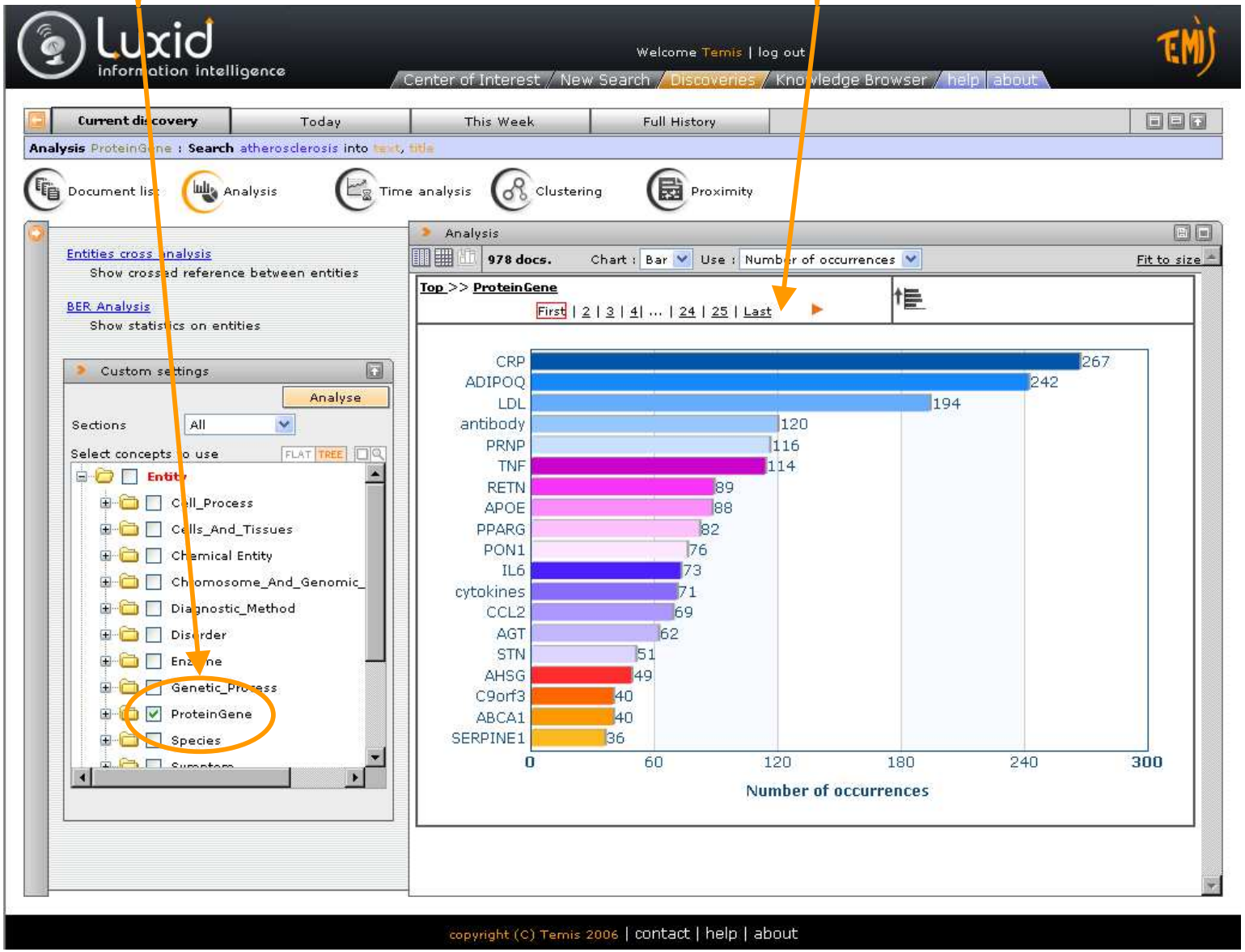
First, perform a search on « atherosclerosis » to locate relevant documents



I receive a first set of 978 documents with what to start my discovery. I can't read them all so look for Proteins and Genes frequency distribution to identify potential targets and biomarkers



The distribution gives you an idea of strongly associated proteins/genes (potential targets) : CRP, ADIPOQ, LDL…



Now we need to look for the "hottest" in the last two years of literature

Activate 'Time Analysis'

Luxid information intelligence

Welcome Temis | log out

Center of Interest / New Search / Discoveries / Knowledge Browser / help / about

Current discovery | Today | This Week | Full History

Analysis ProteinGene, 2005, 2006 : **Search** atherosclerosis into text, title

Document list | Analysis | Time analysis | Clustering | Proximity

Analysis

978 docs. Min. docs. : 1 Columns : ProteinGene OK

Filter ProteinGene ▶ ▼ 2005 ▼ 2006

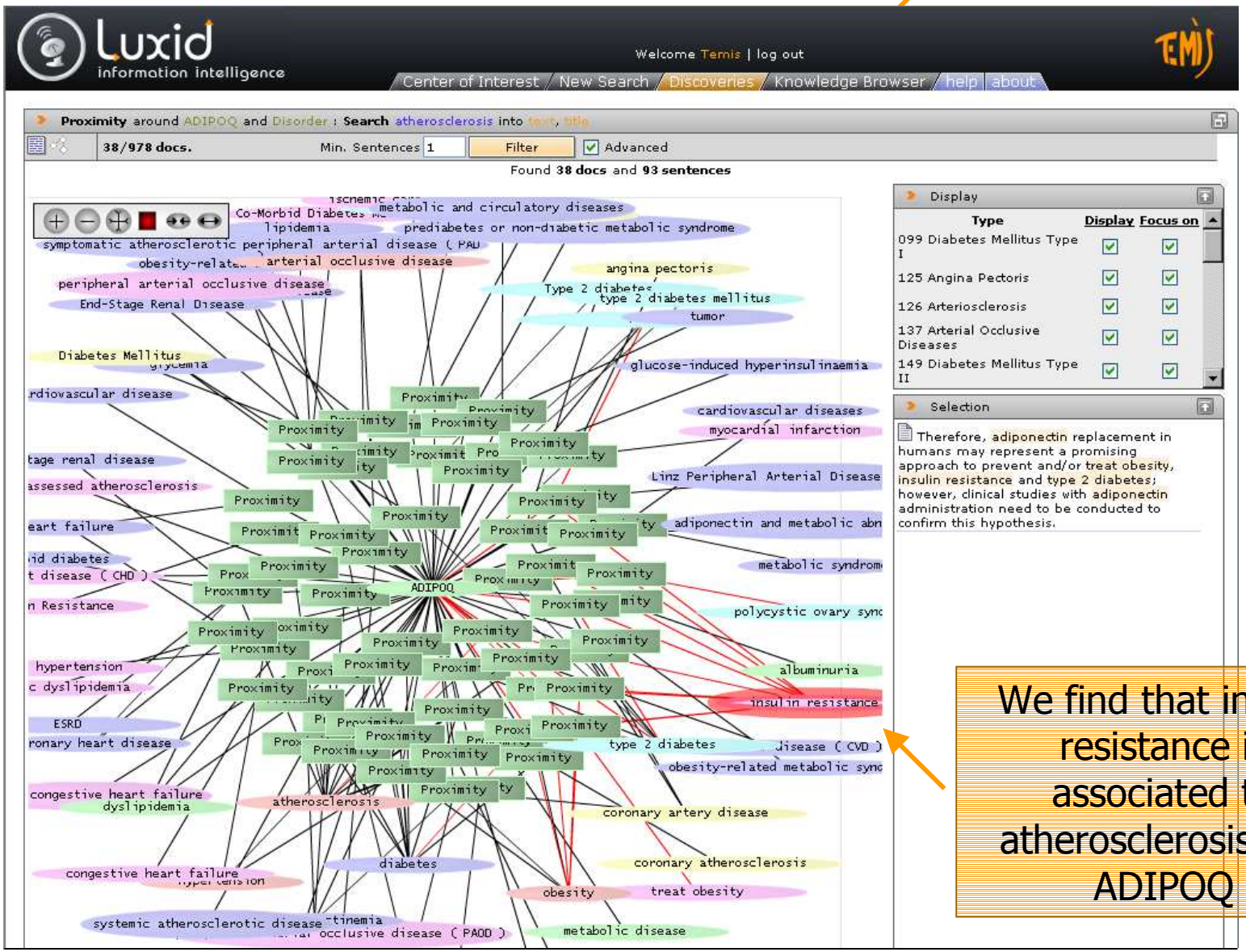
	ADIPOQ	CD36	CRP	cytokine	IL6	LDL	LDLR	STN	AGT	ALB	anticoagulant	antioxidant	APLN	C3	CCL2	chemokine
2005-03	1	1		1	1	1		1	2			1				
2005-04	2		2		1				1							
2005-05	2		3	1	1	6			2			2			1	
2005-06	2		4		1	3		3	3	1	1					
2005-07	1		5		2	5		5	5	1		1				
2005-08	2	1	6		1	7	2		3			2			1	2
2005-09	1		2			3	1	1				2				
2005-10	2	1	4	1	1	4		5	1			2				1
2005-11	3	1	8	3		4			1						1	1
2005-12			6			2						2		1		
2006-01			6		2	4		2	1	1		1				
2006-02	4		3													
2006-03	6	1	5		2	5	1		2			3			1	1
2006-04	6		5		3				2							1
2006-05	1	1	9			3	3	2	1	1					3	1
2006-06	3		3		1	2	2	1	1						1	
2006-07	4	2	6			2	5	2	2	2		3	1			
2006-08	7		10			2	6	1	1	1		3			3	1
2006-09	4	2	2	2	2	2	2	2	1	1	1	1	1	1	1	1

Notice the strong trends about ADIPOQ, CRP and LDL in the last two years

copyright (C) Temis 2006 | contact | help | about

I now want to visualize all the diseases in the vicinity of ADIPOQ (Adiponectin)

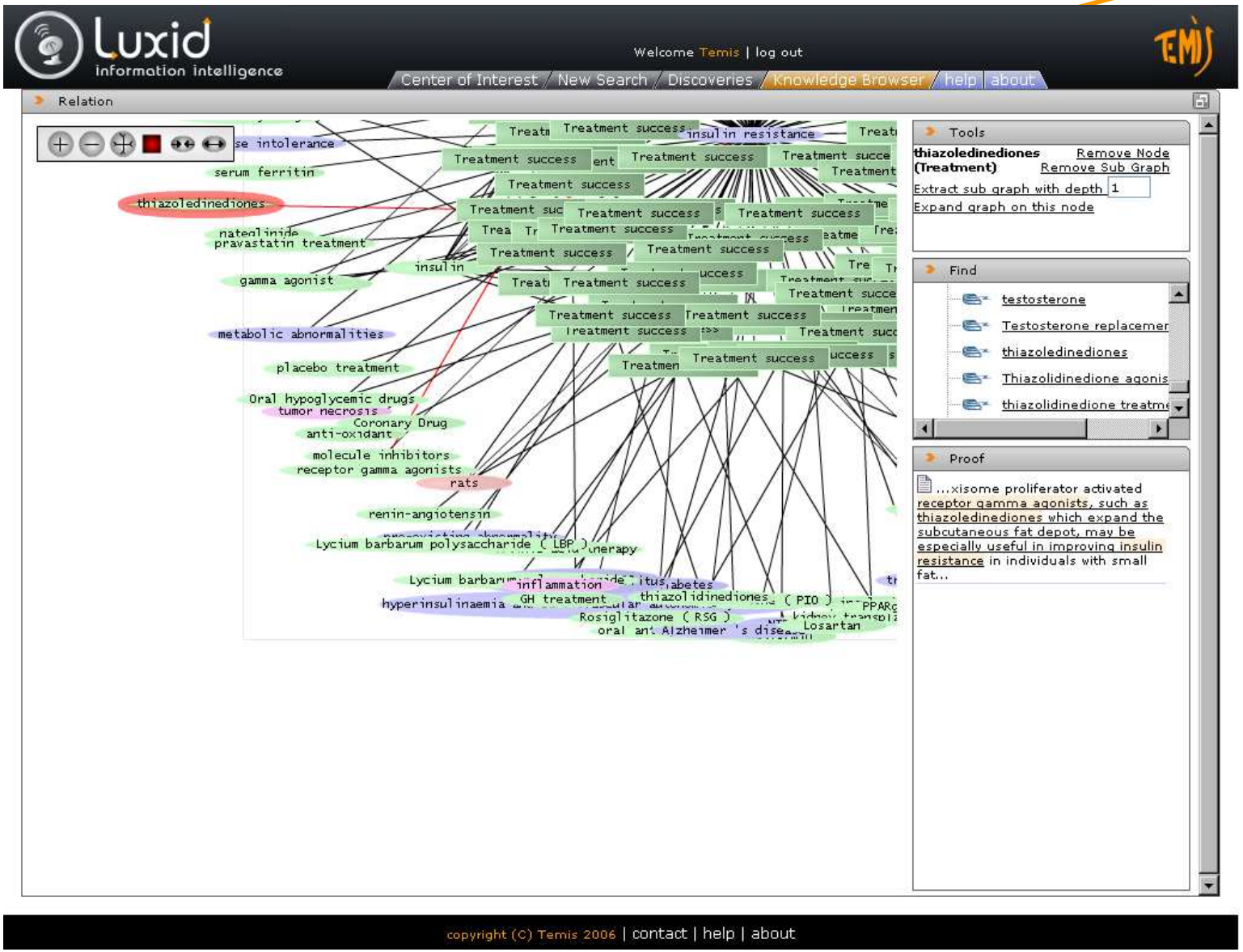
Activate 'Proximity Analysis'



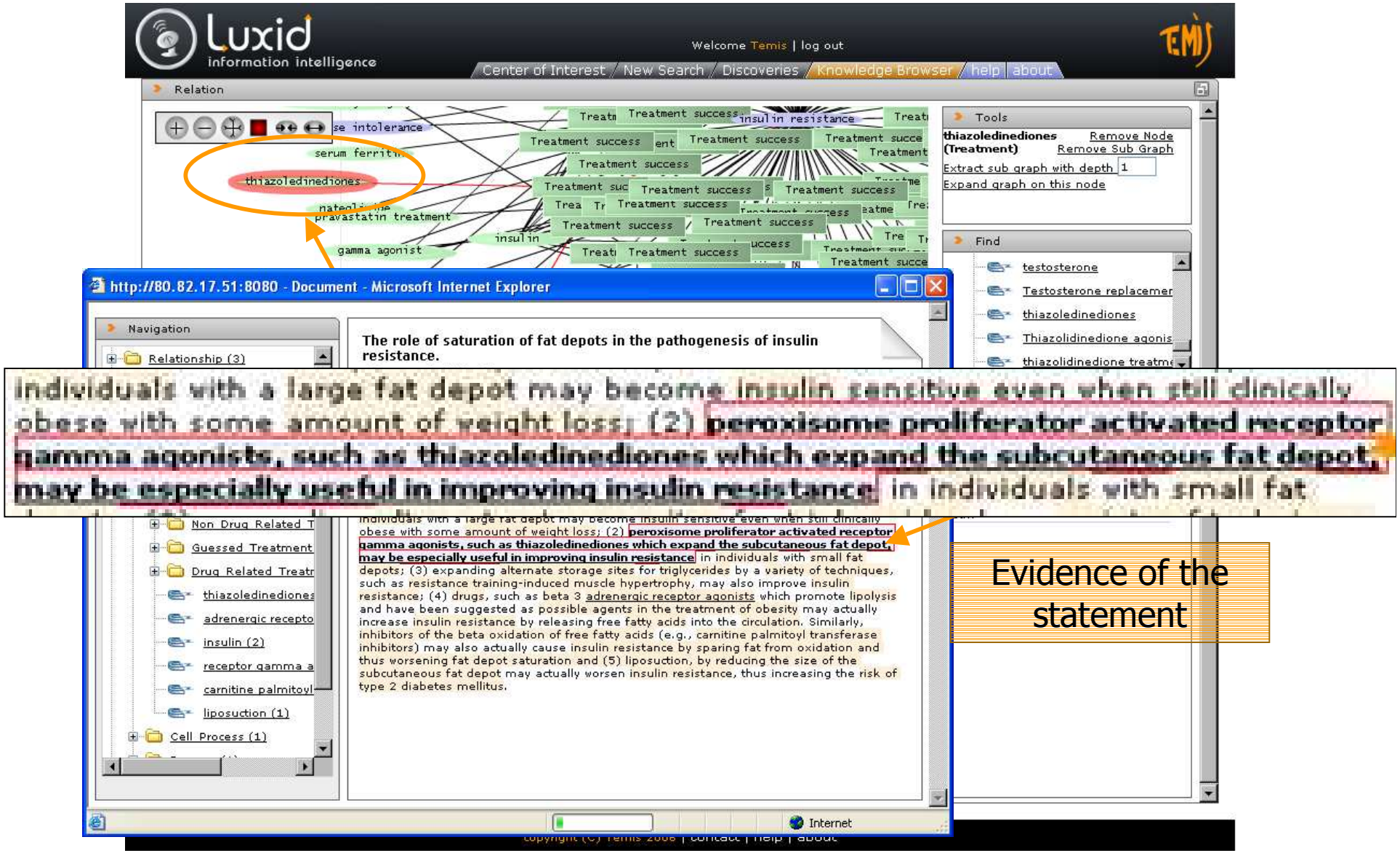
We find that insulin resistance is associated to atherosclerosis and ADIPOQ

Now which compounds are used to treat insulin resistance? We can use a semantic search to look for treatments that have a positive effect on the disease "insulin resistance"…

Activate 'knowledge browser'

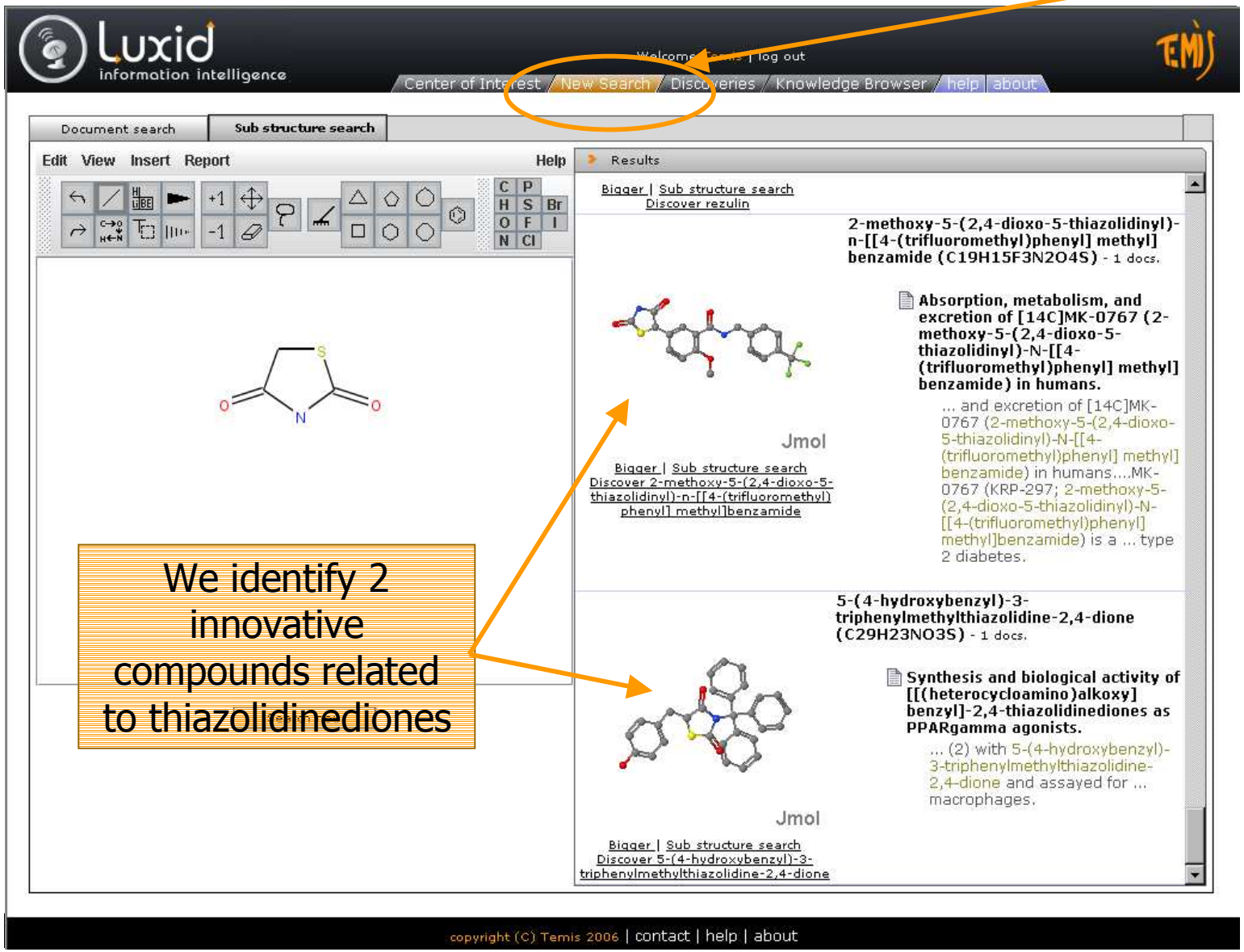


Now which compounds are used to treat insulin resistance ?
We can use a semantic search to look for treatments that have a positive effect on the disease "insulin resistance"…



We need to find all thiazolidinediones to spot therapeutic opportunities

Activate 'Sub-Structure Search'



We identify 2 innovative compounds related to thiazolidinediones

It is now time to share all my discoveries with colleagues and managers. The center of interest has kept the most relevant proofs of my analysis.

Activate 'Centers of Interest'

Luxid information intelligence

Welcome Temis | log out

Center of Interest / New Search / Discoveries / Knowledge Browser / help / about

my center of interest

atherosclerosis

978 docs about atherosclerosis

978 docs. 1 2 3 4 ... 38 39 40 >

[Osteopontin and atherosclerosis]
[Osteopontin and atherosclerosis]
Top terms : atherosclerosis, Osteopontin, atherosclerosis, atherosclerosis, Osteopontin, SPP1, Osteopontin

[Atherosclerosis: butter on the arteries?]
[Atherosclerosis: butter on the arteries?
Top terms : Atherosclerosis, atherosclerosis, Atherosclerosis, butter, butter on the artery

[Postprandial hyperglycemia as a risk factor of atherosclerosis]
[Postprandial ... a risk factor of atherosclerosis]
Top terms : hyperglycemia, atherosclerosis, Postprandial hyperglycemia, atherosclerosis, risk factor of atherosclerosis, Postprandial hyperglycemia, risk factor

Protein Distribution Charts

978 docs. Chart : Bar Use : Number of occurrences Fit to size

Top >> ProteinGene

First | 4 | ... | 38 | 39 | Last

Number of occurrences

Analysis ProteinGene, 2005, 2006 : Search atherosclerosis into

978 docs. Min. docs. : 1 Columns : ProteinGene OK

Filter ProteinGene ▼ 2005 ▼ 2006

	ADIPOQ	CD36	CRP	cytokine
2005-03	1	1		1
2005-04	2		2	
2005-05	2		3	1
2005-06	2		4	
2005-07	1		5	
2005-08	2	1	6	
2005-09	1		2	
2005-10	2	1	4	1
2005-11	3	1	8	3
2005-12			6	
2006-01			6	2

Proximity around ADIPOQ and Disorder : Search atherosclerosis into

38/978 docs. Min. Sentences 1 Filter Advanced

Agenda

1. **Drug Discovery Process**
2. **Recognition of Chemical Entities**
3. **Case study - Scenario Highlight**
4. **TEMIS Presentation**

TEMIS – Presentation

- TEMIS = **TE**xt **MI**ning **S**olutions
  - Software company founded in Sept 2000
- Strong background & experience in Text Mining
  - Seasoned engineers and managers from IBM
  - Acquisition of Xerox Linguistics in 2003 (XeLDA® = 20 years of R&D, 26 worldwide patents)
- Largest European solution provider
- Blue chip customers
  - Life Sciences
  - Publishing
  - Industry

TEMIS - Some references

Life Sciences >>

NOVARTIS
Pfizer
Roche
sanofi aventis
Because health matters
Lilly
FOURNIER PHARMA
A member of the Solvay Group
IPSEN
Innovation for patient care

Publishing >>

THOMSON
ims INTELLIGENCE. APPLIED.
ELSEVIER
MDL
HACHETTE FILIPACCHI MEDIAS
LexisNexis
Springer
the language of science
JOUVE

Industry >>

Ford
EADS
DAIMLERCHRYSLER
ALCATEL
TOTAL
EDF
PSA PEUGEOT CITROËN

Luxid® for Life Sciences

- A new TEMIS Software

… Your information make you lucid …

… Enlight your ideas (lux) …

…The new generation of TEMIS Insight Discoverer™ (ID)…

Beyond Search >>

Recognition of Chemical Entities in Scientific Literature

Luxid
information intelligence

THANK YOU!

TEMIS