



Applying Text Analytics to the Patent Literature to Gain Competitive Insight

Gilles Montier, Strategic Account Manager, Life Sciences

TEMIS, Paris

www.temis.com



- TEMIS has been working with Life Science and other industry clients for many years
- Naturally, requests, comments and suggestions which were made in these projects led us to gradually refine and extend our approaches
- The following slides attempt to consolidate some of these aspects

➤ Improve information discovery through the chain

- Drug discovery & lead identification
- Patent analysis
- Safety & adverse event detection
- Competitive Intelligence
- Sentiment Analysis



- Questions TEMIS solutions are asked to solve:
 - **Bibliometric questions**
 - Who is active on a topic?
 - Who is rather product-oriented or process-oriented?
 - In which country is this company active?
 - ...
 - **Specific and technical questions**
 - Is this device effective against this disease?
 - Which metals from the family of rare earths are used?
 - What is the unique aspect of a given patent?
 - ...
- Patent literature deliberately uses new terms which are hard to find by classical means
- How Text Analytics can help?

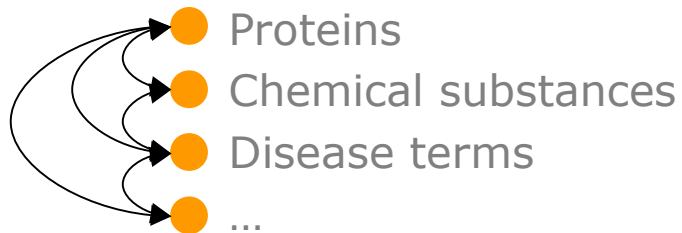
- 1. Language analysis ⇔ Domain & scenario specific
 - Recognition of the semantic type
 - *TNF is a protein,*
 - *Diabetes Mellitus type2 is a disease*
 - *Aspirin is a chemical substance*

- 1. Language analysis ⇔ Domain & scenario specific
 - Recognition of the semantic type
 - Recognition of variants
 - Tumor necrosis factor is the same thing as TNF
 - NIDDM is the same thing as Diabetes Mellitus type 2
 - Acetylsalicylic acid is the same thing as aspirin

- 1. Language analysis ⇔ Domain & scenario specific
 - Recognition of the semantic type
 - Recognition of variants
 - Linking (canonical information)
 - Proteins ⇔ database identifiers
 - Chemical substances ⇔ structures
 - Disease terms ⇔ thesaurus identifiers

- 1. Language analysis ⇔ Domain & scenario specific

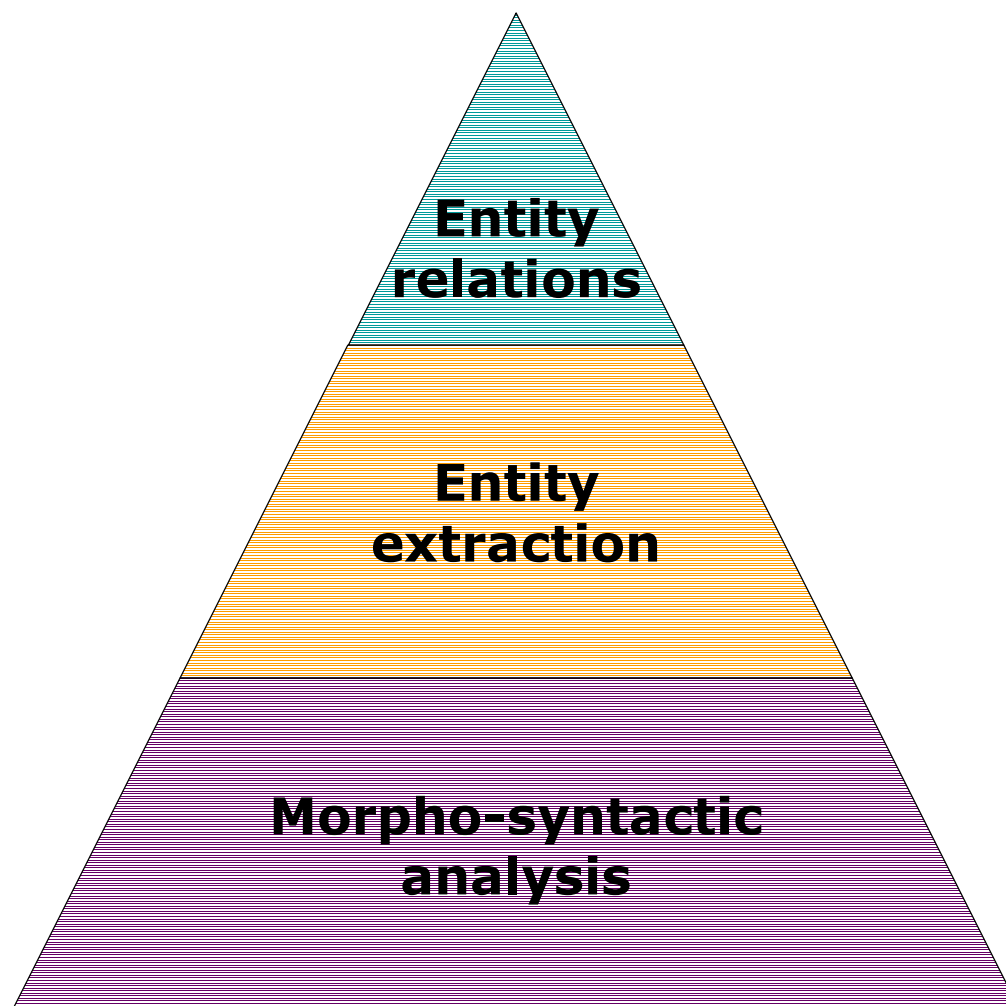
- Recognition of the semantic type
- Recognition of variants
- Linking (canonical information)
- Cross-linking between entities



based on detailed syntactic analysis or 'just' proximity

- 1. Language analysis ⇔ Domain & scenario specific
 - Recognition of the semantic type
 - Recognition of variants
 - Linking (canonical information)
 - Cross-linking between entities
 - Relevance
 - So many hits! Which ones are “interesting”?
 - Which ones are new?

- 1. Language analysis ⇔ Domain & scenario specific
 - Recognition of the semantic type
 - Recognition of variants
 - Linking (canonical information)
 - Cross-linking between entities
 - Relevance
 - Openness: Black boxes won't do the job
 - No thesaurus/entity recognizer is complete: → "guessing" semantic types
 - Foresee user-defined extensions
 - Recognition of new terms



Roles and Relationships

- a company in an acquisition event
- a compound in a chemical equations

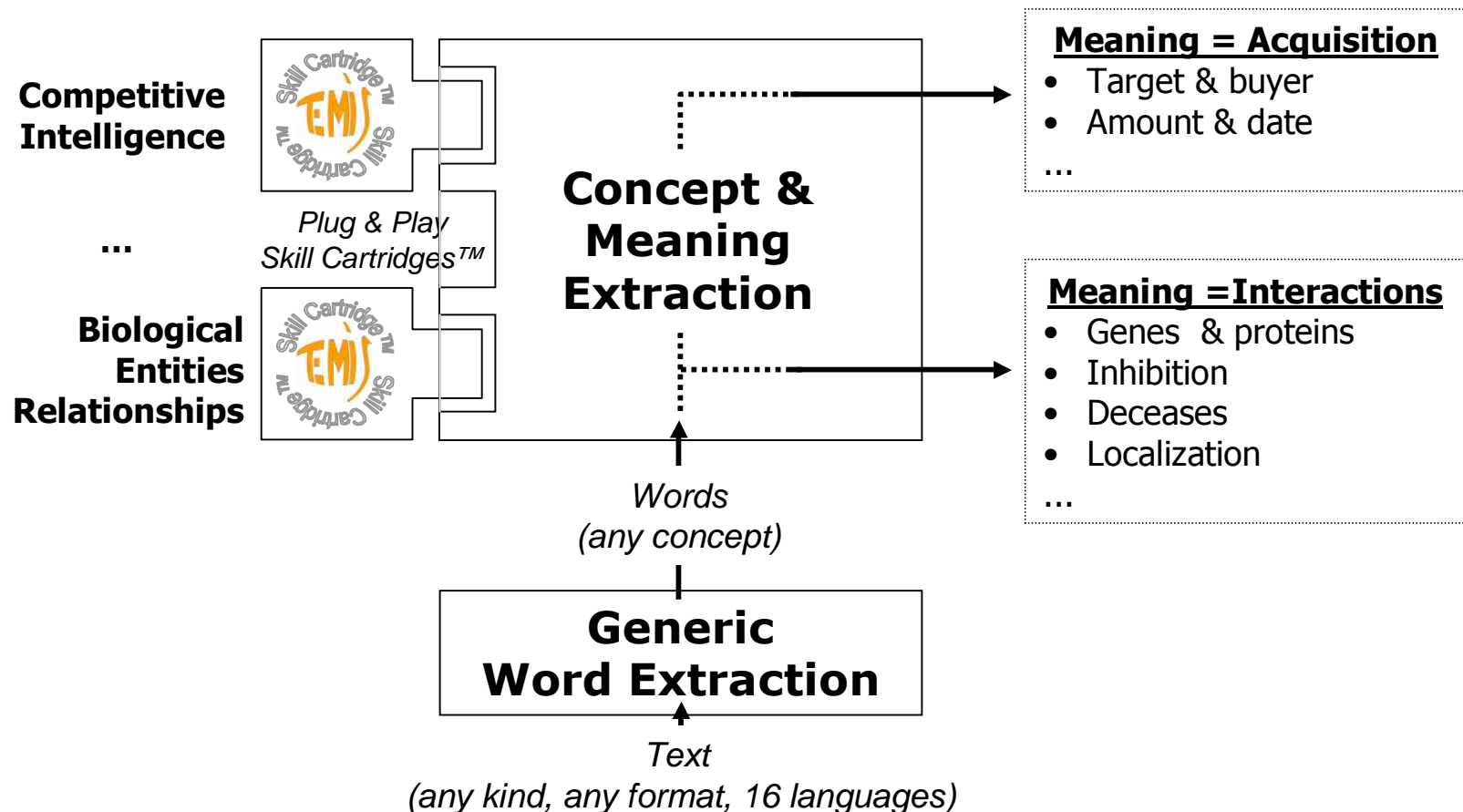
The recognition of distinct entities

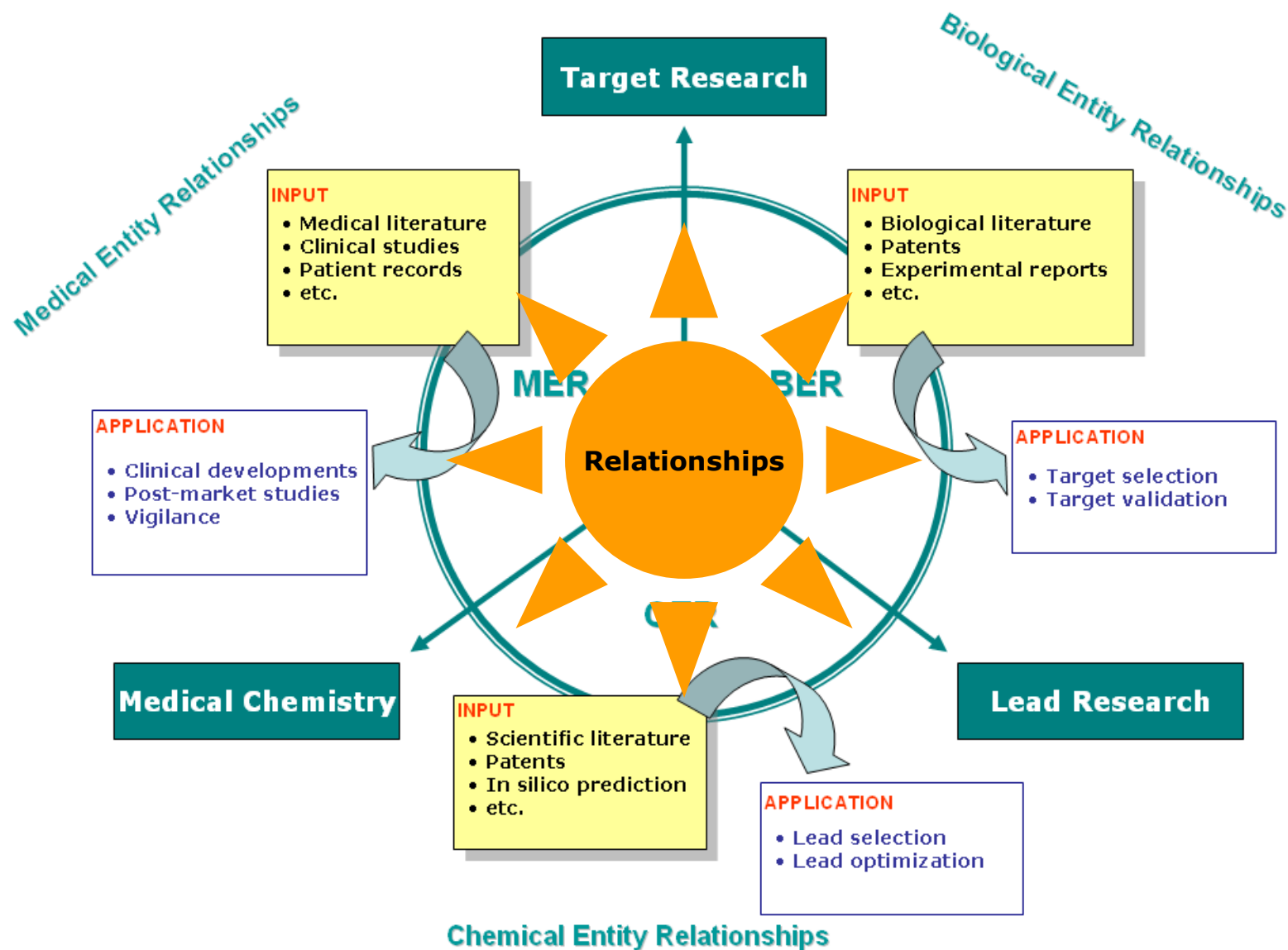
Examples: proteins, chemical compounds, diseases, companies, person names ...

General linguistic pre-processing

Results: nouns, verbs, adjectives, noun phrases, etc.

- Building Skill Cartridges
 - Each of one define a specific domain of interest
 - Syntactic and semantic rules





- The Skill Cartridge™ concept is a powerful and successful model, allowing TEMIS to communicate and apply a known flexible approach to a wide range of scenarios

- However:
 - Skill Cartridge of the BER- and CER-type is costly to build and very specific
 - What about if you want to allow the users to analyze and explore content not only with predefined known terms but with “open” terms in a cross domain
 - Potentially any term can be of interest but the most frequent item is not always the most interesting

- Need to allow the user to analyze and explore content not only with predefined known terms but with “open” terms.
 - Important information can be discovered using the simple extraction of domain independent term candidates
 - Open terms doesn't mean simple terms

- Doing a little math we can restrict the analysis to only the (presumably) relevant information
 - Make a statistically guided guess about which terms are relevant (keep only a small number)
 - Assign confidence score to each term
 - Set of parameters to customize the results

- RelevantTermFinder (RTF)
 - Allows to work cross-domain
 - Without the need to manually adapt to new domains
 - Separates important from unimportant information

- Advantages:
 - Approach is very replicable
 - Can be applied in many different contexts with minor or no variation
 - Approach technically very simple
 - RTF is fast, multilingual
 - Approach allows to address exploratory scenarios
 - Finding “new” information – on issues that were not modeled before

- One real-world example on patent data
(there are many others...)

- Scenario
 - Corpus of ~2000 Patents on « Stents » (bioresorbable stents)

 - Answer one specific question:
why in a set of patent documents people used « Yttrium » in their products?
 - There is no specific Cartridge for this question and there will never be one, because the question came up ad hoc

- Question: Can Luxid® guide me towards the really interesting issues?

Search



Welcome **temis** | log out

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

Document search advanced search

Search

Search inside concepts Published: ----- Max items: No limit

Search yttrium in all sections and their concepts

Metadata

- Assignee (731,731)
- Author (846,2K)
- DataSource (846,846)
- ...

Entities & Relationships FLAT **TREE** AND OR

- Entity (844,2M)
- Relationship (488,2K)

Search for documents concerning „yttrium“ (a rare earth metal)

Analysis

The screenshot displays the Luxid information intelligence interface. At the top, the Luxid logo and navigation menu are visible. The main content area shows search results for 'yttrium' with 25 documents. A yellow callout box highlights the text 'Returns 25 documents - let's now analyse them'. Below the search bar, there are icons for 'Document list', 'Analysis', 'Time analysis', 'Clustering', and 'Proximity'. The 'Analysis' icon is selected. On the left, a 'Filtering' panel allows users to select concepts to use, including Entity, Metadata, and Relationship. On the right, a 'Document List' panel shows the first three results:

- Biocompatible ceramic glass**
 yttrium oxide: 3 to 10 ...lanthanide or yttrium oxide = 3 to 10 ...Preferably ... rare earths or yttrium of having a ... rare earths or of yttrium under the same ... yet been proved.
♦ **Top concepts : bone, ceramic, glass ceramic, compound, hydroxyapatite, glass, crystal**
- Surface hardened biocompatible metallic medical implants**
 tantalum, aluminum, and yttrium for this purpose....% tantalum, yttrium, thorium, or aluminum for this purpose....%) of tantalum ... or Zr-Ta) or yttrium.
♦ **Top concepts : surface, alloy, solute, tantalum, oxidation, zirconium, titanium**
- Process for the production of ceramic prostheses**
 A process for the ... 95.3% by weight, yttrium oxide of between ... below 5% by

From Frequency to Relevance



Welcome temis | log out



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

Current discovery Today This Week Full History

25 docs. Search yttrium in all sections and their concepts today 09:04:24 25 delete

25 docs. Analysis Terms today 09:05:22 25

Analysis Terms : Search yttrium in all sections and their concepts

Document list Analysis Time analysis Clustering Proximity Selection : Current (0 docs)

Pre-defined Analysis

- Metadata Statistics on metadata
- Medical Entities Statistics on medical entities
- MeSH Statistics using MeSH hierarchy

Analysis

25 docs. Nb Documents Count Density Relevance Relevance OK

Filter : OK 1 2 3 4 ... 276 277 278 Σ

zirconium	18	827	1
surface	21	824	1
alloy	20	761	1
prosthesis			
invention			
oxide			
temperature			
method			
coating			
example			
micron			
zirconium oxide			
attachment	12	225	0
titanium	20	216	0
oxidation	10	209	0
wear	12	206	0

Analysis returns frequency sorted list. Observation: many terms are not informative

Can we do better? Yes, let's sort the terms according to the strength of their association with the term „yttrium“

Discovering relevant topics

25 docs. Nb Documents Count Density Relevance ----- Relevance

Name	Chart	Documents	Count	Relevance
▼ Terms		25	43582	
Filter : <input type="text"/> <input type="button" value="Ok"/> 1 2 3 4 ... 276 277 278 ≥				
yttrium		22	86	4.8
zirconium		18	827	3.1
zirconium oxide		13	225	2.9
surface hardness		11	32	2.8
furnace		15	50	2.6
corrosion resistance		14	44	2.7
fatigue strength		12	46	2.7
carbide		12	44	2.7
nitride		11	142	2.5
oxide coating		10	21	2.5

Relevance-Sorted list suggests that yttrium may have something to do with very specific properties of the device, namely „surface hardness“, „corrosion resistance“ and „fatigue strength“

- Similar documents:
 - Show standards similar documents to one document

- Deduplication
 - Remove real duplicates and near duplicates

- Categorization
 - Classify documents automatically according to ontologies

- Clustering:
 - Classify documents automatically

Conclusion:



- RTF is a Skill Cartridge™
 - Use and deploy like any other Skill Cartridge™
 - „Self-contained“, internal DB, no external dependency

- Working with open terms is a very useful complement to our existing Skill Cartridges™
 - Combined with appropriate sorting it allows to reach high relevant results
 - It improves replicability and broadens the range of context in which Luxid® can be used

- RTF is available
 - Luxid®

- Goals
 - Better support Patent Analysts in setting vocabularies
 - Ease the customization

- How?
 - Improve our Skill Cartridge™ Development Studio
 - Develop new customization tools/products

- “Social Tagging”
 - Make Knowledge Workers contribute to Skill Cartridge™ development
 - Keep centralized control & monitoring



- 3 profiles
 - Skill Cartridge™ Builders
 - Solution administrators & customizers
 - Business Users
- 3 environments
 - Development
 - Customization & test
 - Production
- 3 product stacks
 - Development Studio®
 - Lexicon Manager®
 - Dynamic Mapping Editor®



Skill Cartridge™
Builders

Build Skill Cartridges™

Lucid®
Administrators

Lucid®
Users

The screenshot shows the Skill Cartridge Studio interface. The main editor displays XML code for a concept named "FilterFromAnalytics". The code includes a list of French prepositions and their corresponding parts of speech, such as "afin / (~Of|~Ofthe)", "en#PREP / harmonie / avec", "sous#PREP / (couvert|forme) / (~Of|~Ofthe)", etc. The interface also shows a Package Explorer on the left with a tree view of the project structure, including folders for different languages and files like "Filter4ExTer.scp". At the bottom, a table lists 10 items of information:

Description	Resource	Path	Location
[ExTer.sct] Empty concept: ToFilterOut	Filter4ExTer.scp	Exter/English	line 135
[ExTer.sct] Reference to concept at defini	Meta4ExTer.scp	Exter/Italian	line 107
[ExTer.sct] Reference to concept at defini	Meta4ExTer.scp	Exter/English	line 26
[ExTer.sct] Reference to concept at defini	Meta4ExTer.scp	Exter/French	line 60
[ExTer.sct] Reference to concept at defini	Meta4ExTer.scp	Exter/German	line 24
[ExTer.sct] Reference to concept at defini	Meta4ExTer.scp	Exter/Italian	line 72



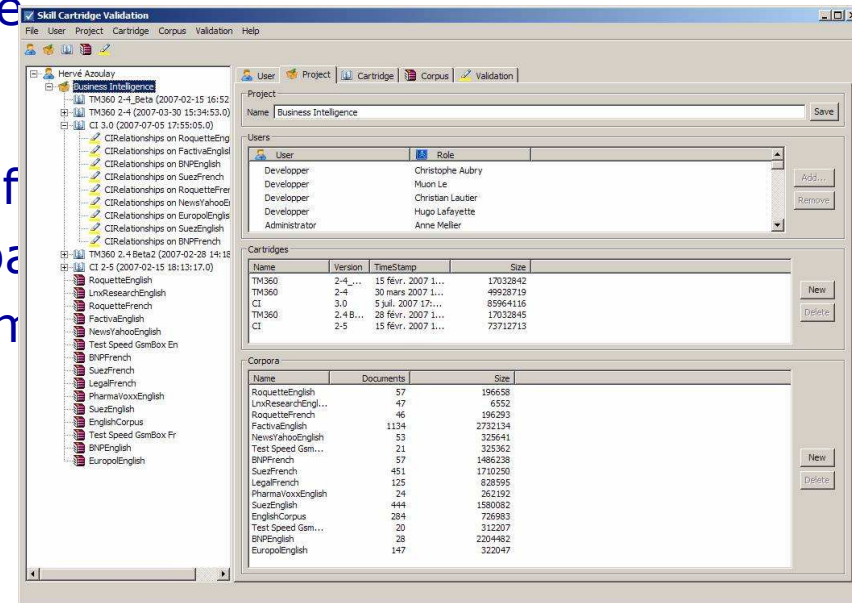
Skill Cartridge™
Builders

Luxid®
Administrators

Luxid® Users

- Enrich Skill Cartridge™
 - Tailor with project/customer taxonomy
- How? Lexicon Manager®
 - Import taxonomy
 - Adjust lexicons (add/remove/edit entries)
 - Check consistency
 - Re-compile

- Who?
 - TEMIS Prof
 - Certified pa
 - IT & Inform



Search **company** in section all sections and their concepts 2/3214

Quick Highlight
[company](#)

Highlight
FLAT TREE

- Entity (3)
 - Company (3)
 - Astra Zeneca (2)
 - Zeneca (1)
- Metadata (2)
 - Date (1)
 - Source (1)

Selection : Merge Delete Move

Notes
Selection --- All ---

Entity	Text	
Name	Status	
Zeneca	Valid	✗
Astra Zeneca	Valid	✗

Inside a swift mega-merger.
Date 2001 Jan 01
Source Pharmaceutical Times
Top concepts Astra Zeneca, Zeneca

Inside a swift mega-merger.

Pharmaceutical Times - 2001-01-01

Zeneca

- Export into Entity list
- Export into Company list
- Delete in this discovery
- Delete in this document
- Delete in all knowledge
- Move...
- Advanced edit...

Zeneca Rename

(c) 2001 Roy



- Patent Literature uses terms which are hard to find by classical means
- The powerful approach of building Skill Cartridges™ needs to be completed by new approaches and tools:
 1. Using “Open Terms” like RTF allows to discover specific information and answer open questions
 2. Allowing Knowledge workers and Patent Analysts to easily set up new vocabularies increases productivity and serendipity

Thank You

Beyond Search >>
Luxid® for Life Sciences



Gracias !

WWW.TEMIS.COM

