Applying Text Analytics to the Patent Literature to Gain Competitive Insight

Gilles Montier, Strategic Account Manager, Life Sciences

TEMIS, Paris

www.temis.com

Lessons Learnt

- 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

Copyright © 2007 TEMIS - All Rights Reserved

Text Mining for Life Sciences Organizations

- Improve information discovery through the chain
  - Drug discovery & lead identification
  - Patent analysis
  - Safety & adverse event detection
  - Competitive Intelligence
  - Sentiment Analysis

Copyright © 2007 TEMIS - All Rights Reserved

Patent 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 againts 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 knew terms which are hard to find by classical means
- How Text Analytics can help?

Copyright © 2007 TEMIS - All Rights Reserved

Text Analytics Requirements

- 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

Copyright © 2007 TEMIS - All Rights Reserved

Text Analytics Requirements

- 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
    - Acecylsalicylic acid is the same thing as aspirin

Copyright © 2007 TEMIS - All Rights Reserved

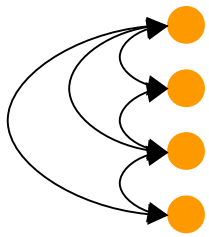
Text Analytics Requirements

- 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

Copyright © 2007 TEMIS - All Rights Reserved

Text Analytics Requirements

- 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

Copyright © 2007 TEMIS - All Rights Reserved

Text Analytics Requirements

- 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?

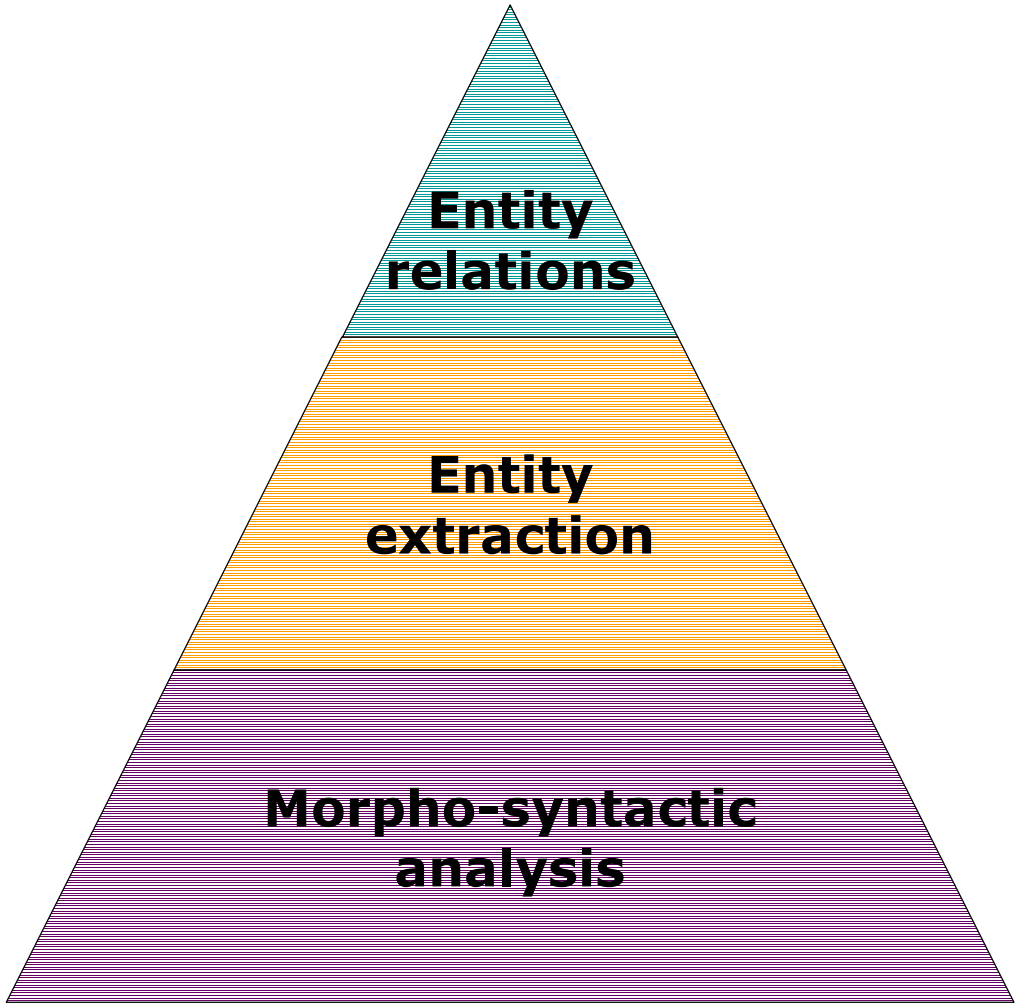
Copyright © 2007 TEMIS - All Rights Reserved

Text Analytics Requirements

- 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

Copyright © 2007 TEMIS - All Rights Reserved

Text Analytics Levels



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 ...

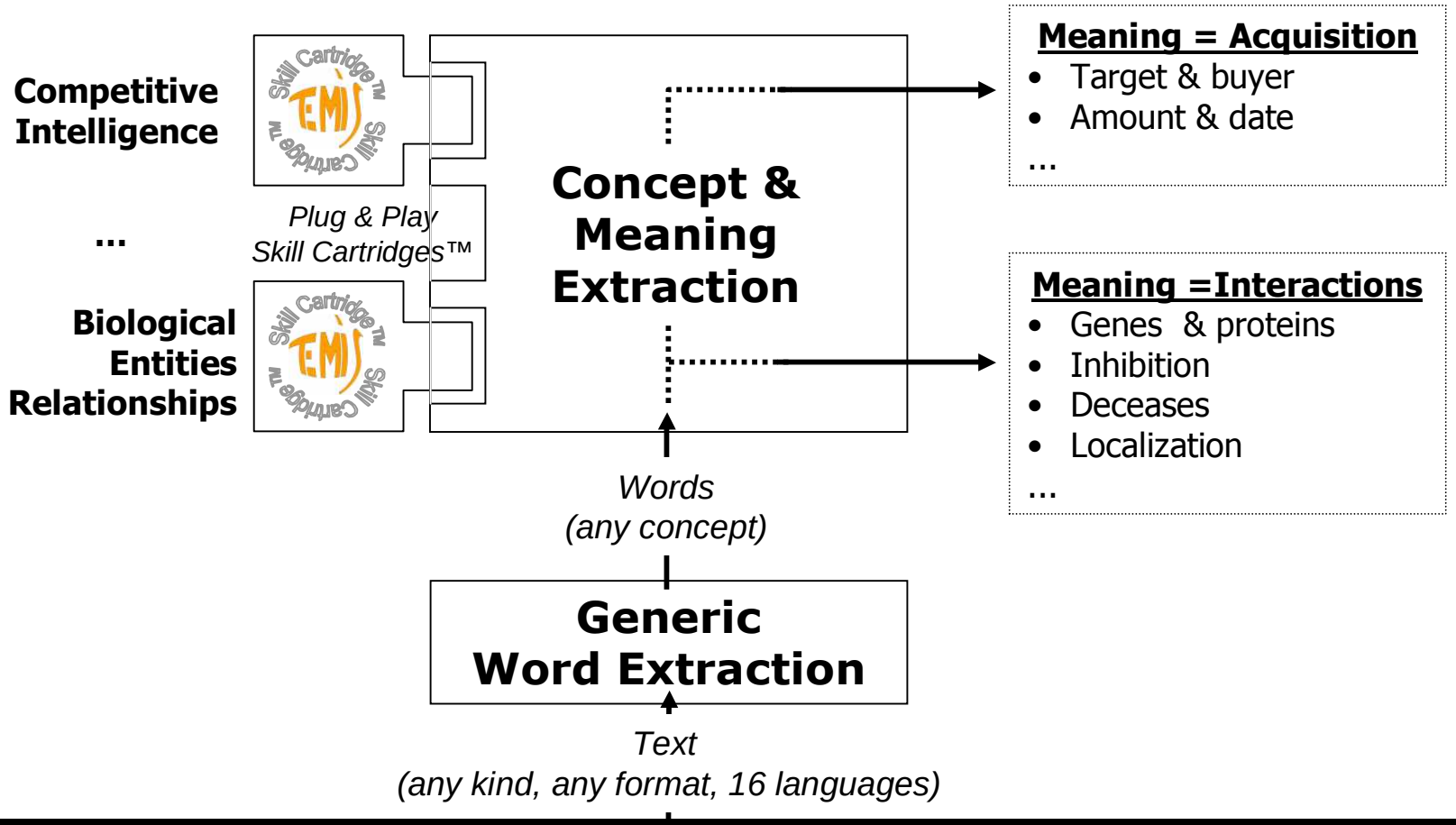
General linguistic pre-processing

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

Copyright © 2007 TEMIS - All Rights Reserved

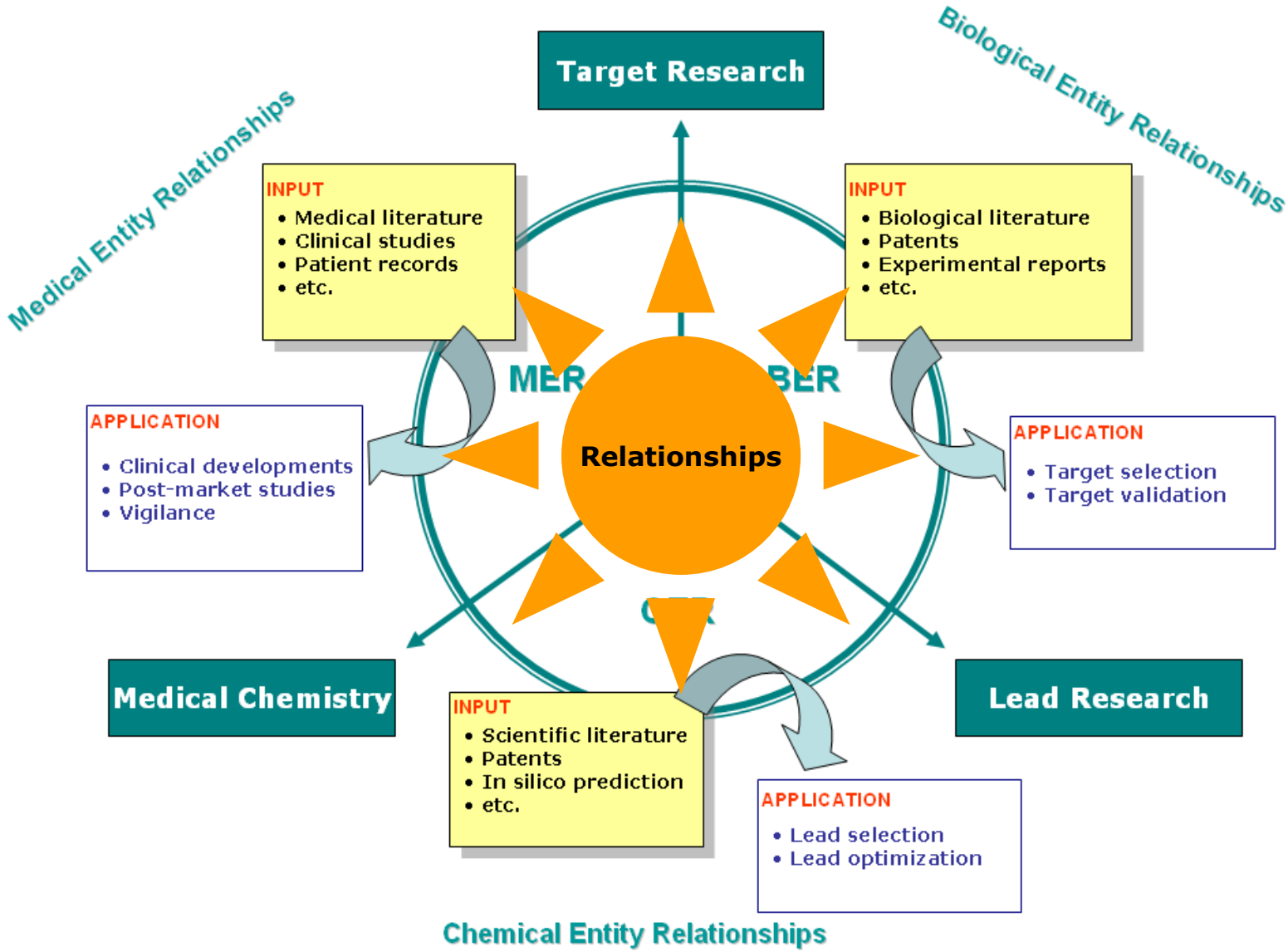
Semantic Knowledge Modeling

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



Copyright © 2007 TEMIS - All Rights Reserved

Life Sciences Skill Cartridges™



Copyright © 2007 TEMIS - All Rights Reserved

Current situation

- 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

Copyright © 2007 TEMIS - All Rights Reserved

Solution n°1: Open Terms

- 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

Copyright © 2007 TEMIS - All Rights Reserved

Solution n°1: Open Terms with RTF

- 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

Copyright © 2007 TEMIS - All Rights Reserved

Solution n°1: Example with RTF

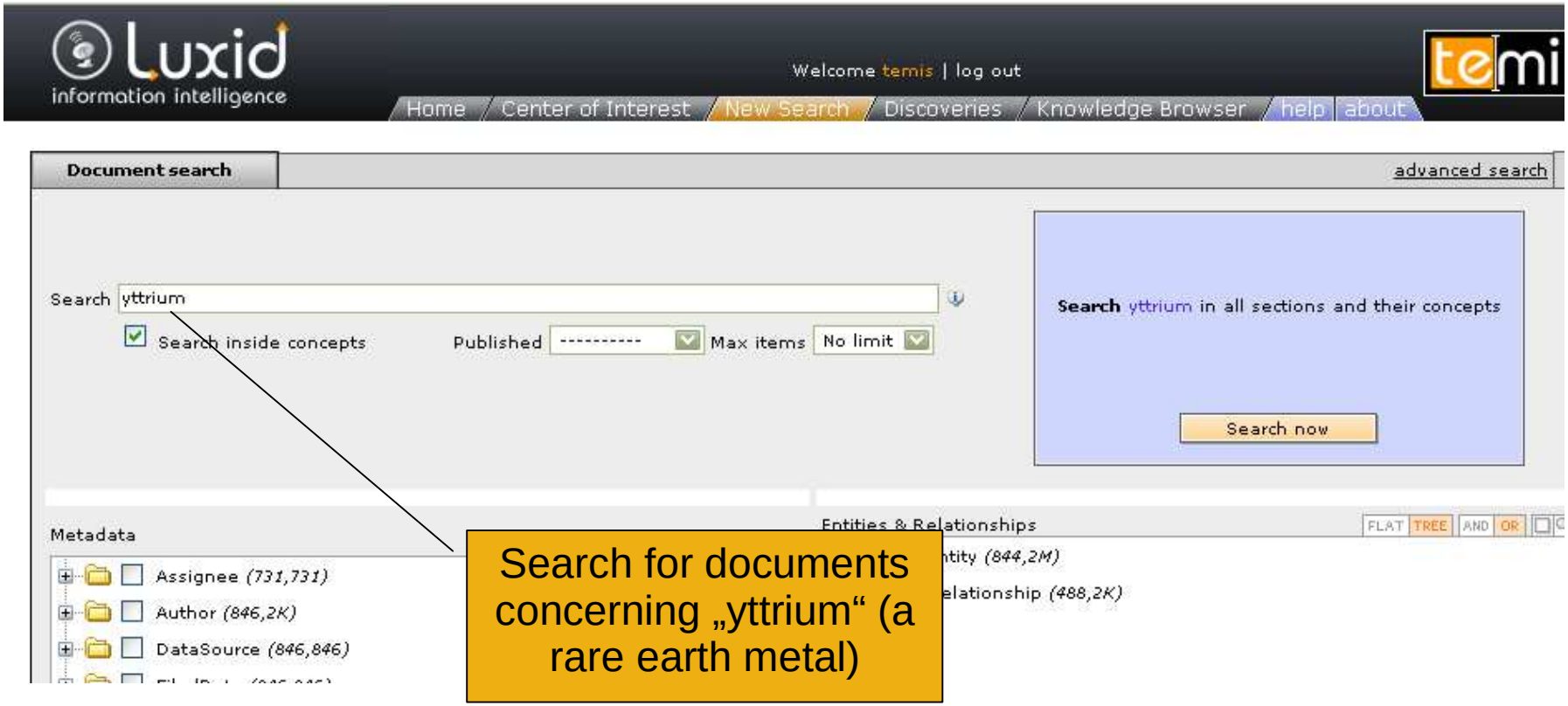
- One real-world example on patent data
  (there are many others…)

- Scenario
  - Corpus of ~2000 Patents on « Stents » (bioresorbing 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?

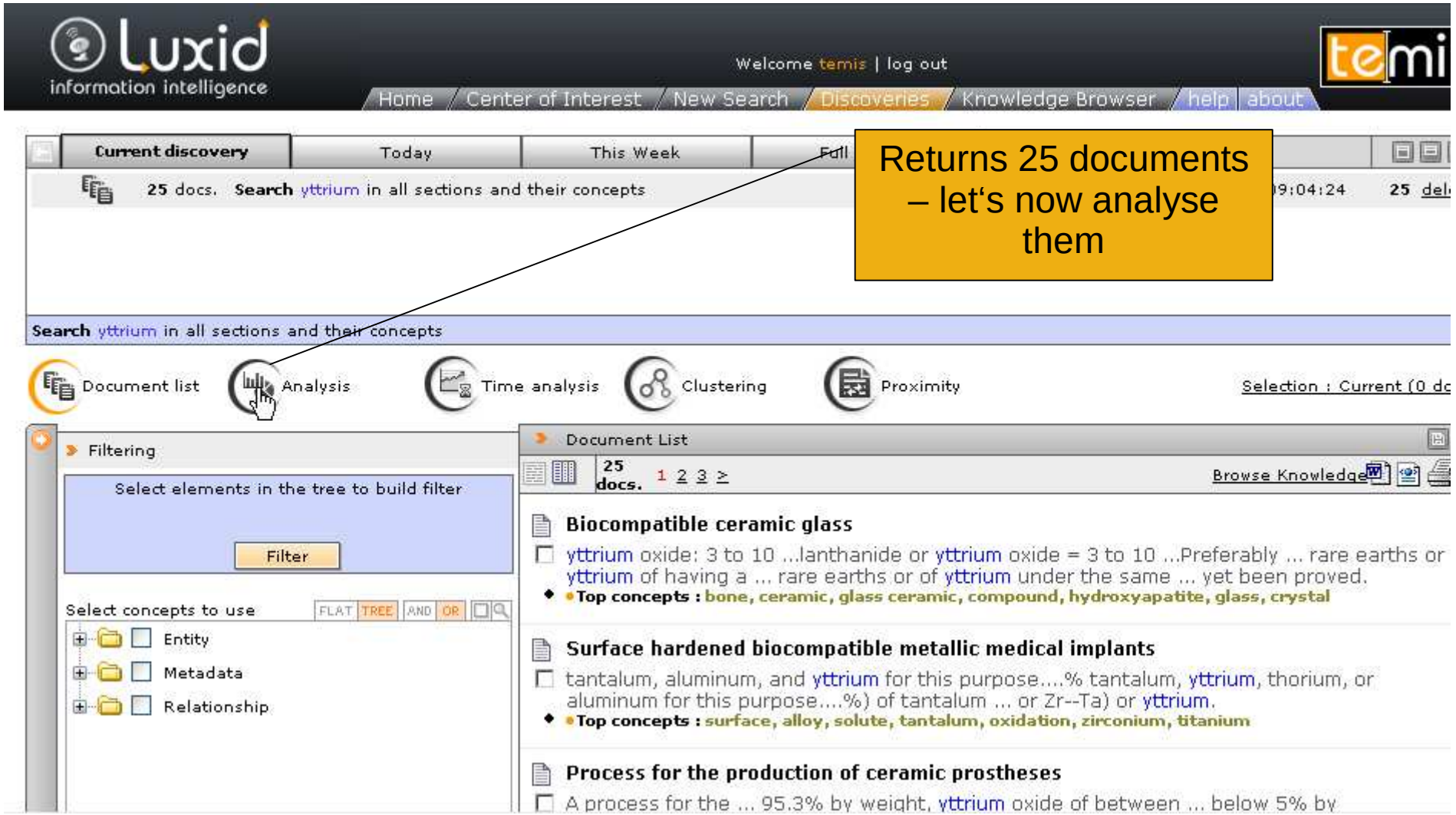
Copyright © 2007 TEMIS - All Rights Reserved

Search



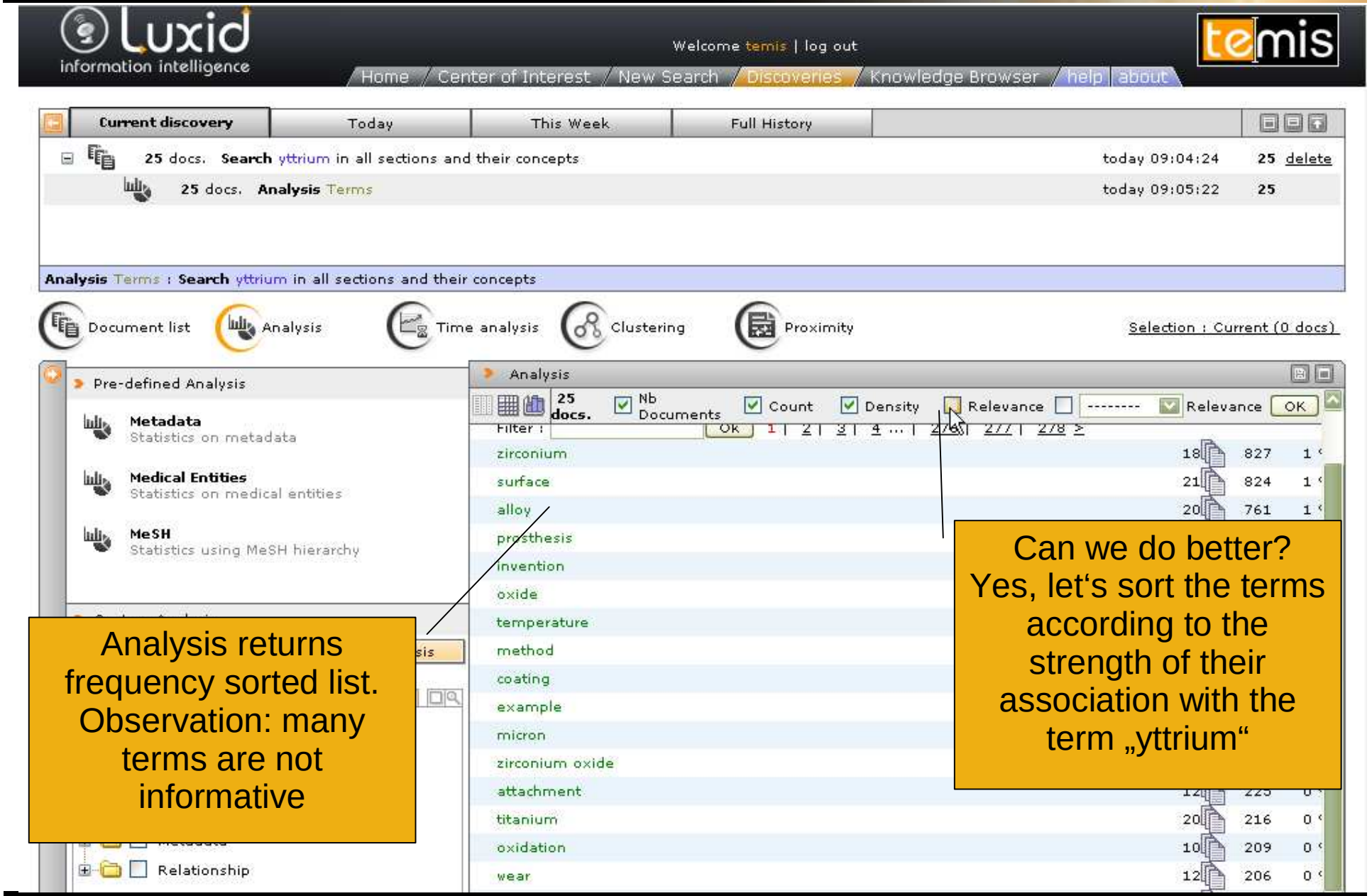
Copyright © 2007 TEMIS - All Rights Reserved

Analysis



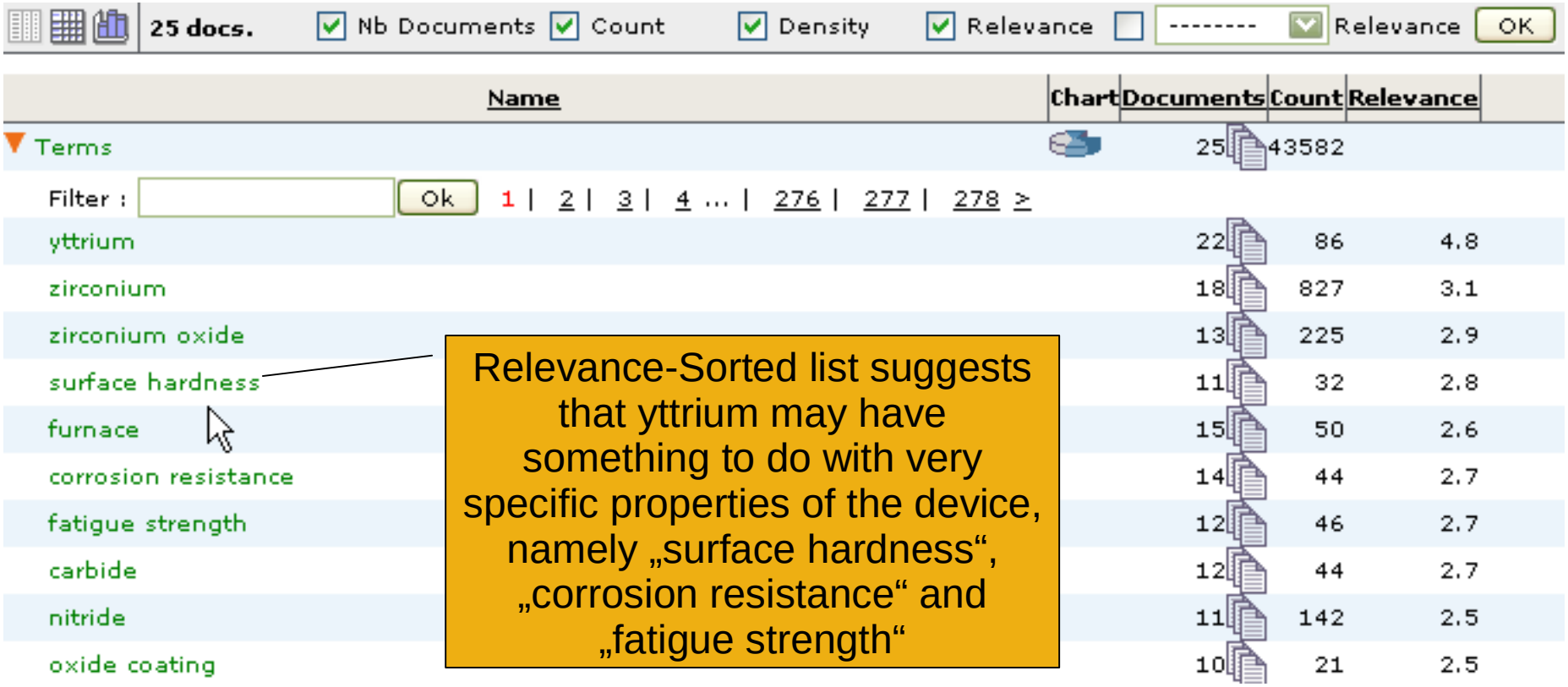
Copyright © 2007 TEMIS - All Rights Reserved

From Frequency to Relevance



Copyright © 2007 TEMIS - All Rights Reserved

Discovering relevant topics



Copyright © 2007 TEMIS - All Rights Reserved

RTF applications:

- 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

Copyright © 2007 TEMIS - All Rights Reserved

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®

Copyright © 2007 TEMIS - All Rights Reserved

Solution n°2: Easing the Skill Cartridge™ Model

- 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

Copyright © 2007 TEMIS - All Rights Reserved

Solution n°2: Easing the Skill Cartridge™ Model

- 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®

Copyright © 2007 TEMIS - All Rights Reserved

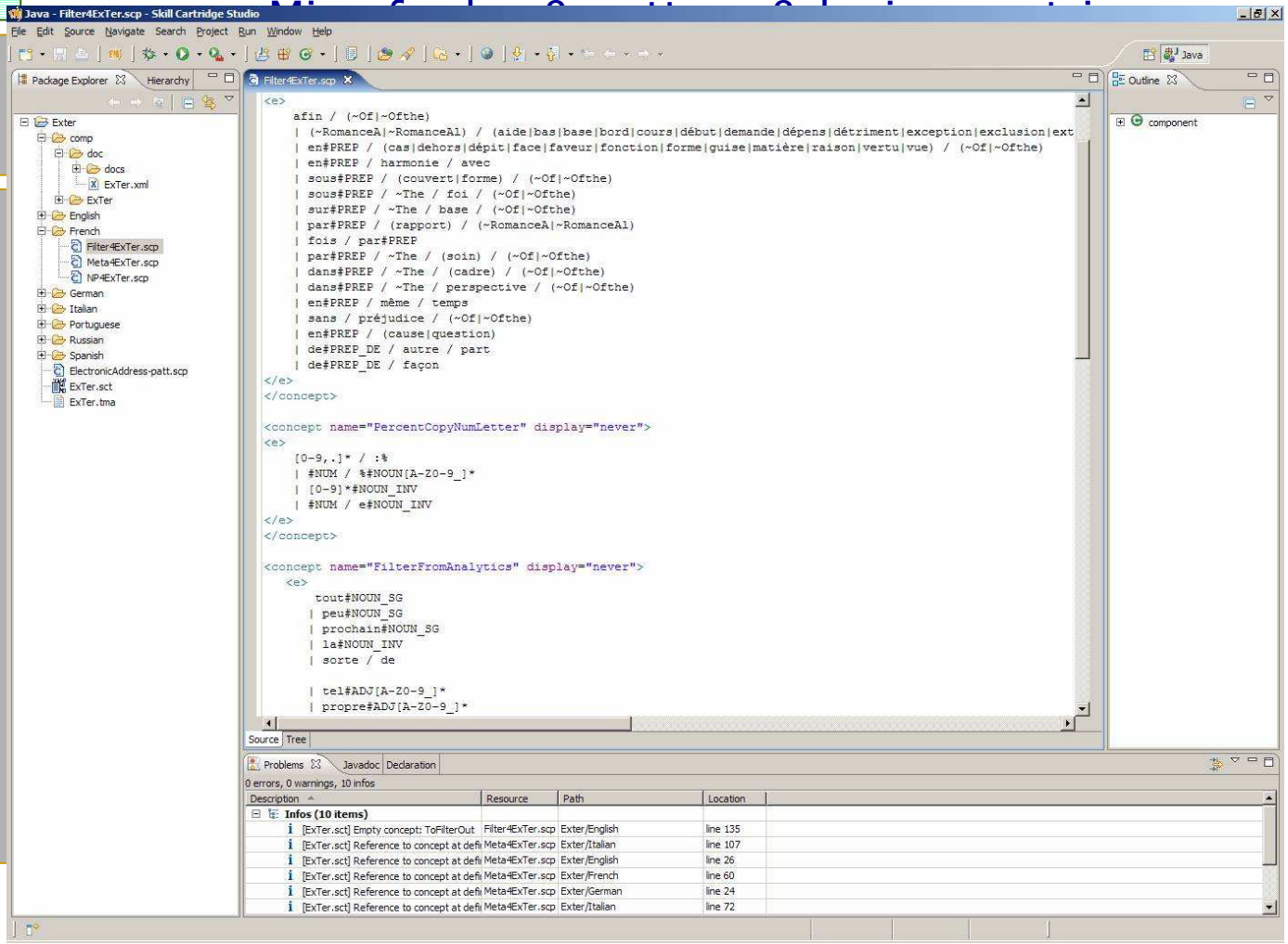
Skill Cartridge™ Builders

Skill Cartridge™ Builders

Luxid® Administrators

Luxid® Users

- Build Skill Cartridges™



Copyright © 2007 TEMIS - All Rights Reserved

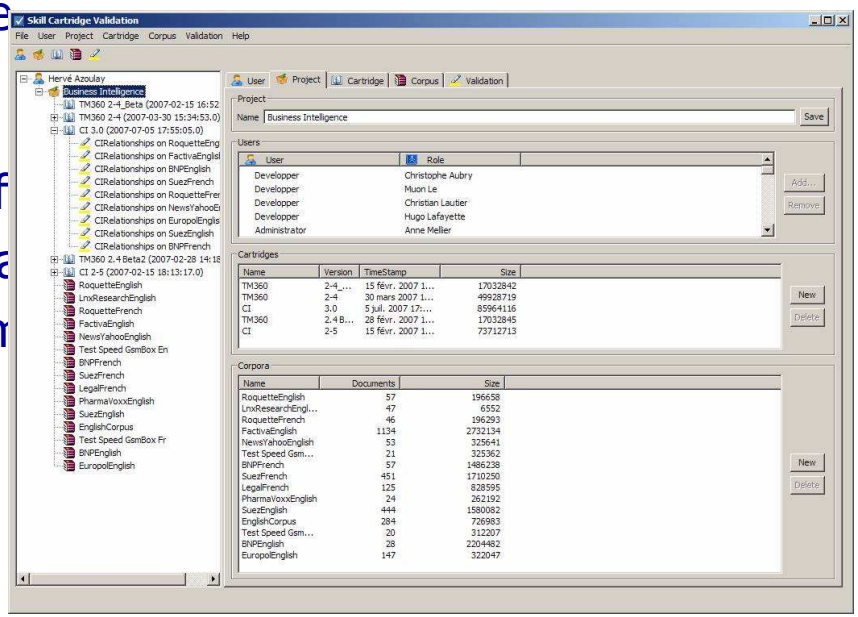
Luxid® Administrators & Customizers

Skill Cartridge™ Builders

Luxid® Administrators

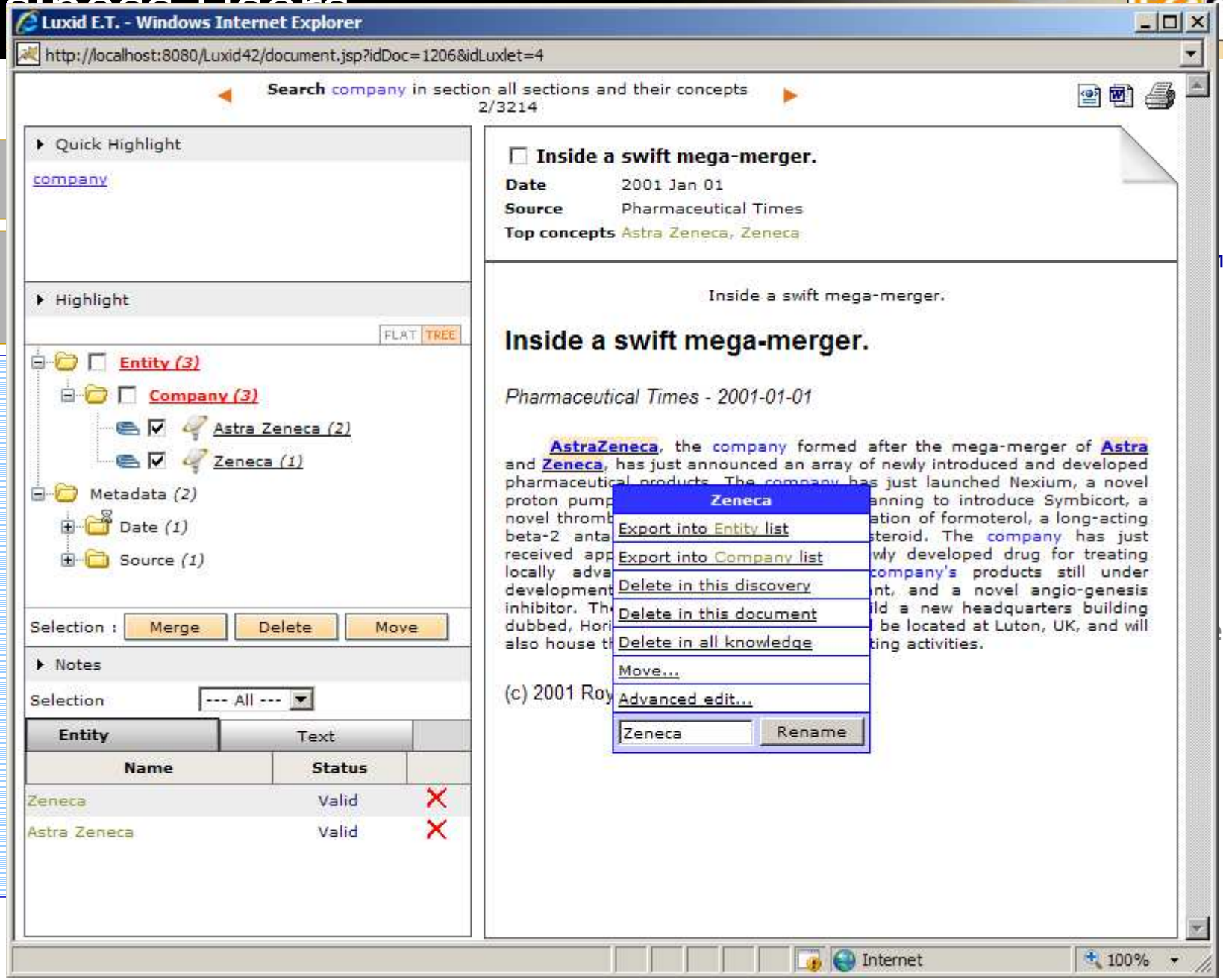
Luxid® Users

- 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



Copyright © 2007 TEMIS - All Rights Reserved

Business Users



Copyright © 2007 TEMIS - All Rights Reserved

Conclusion

- Patent Literature uses terms which are hard to find by classical means
- The powerful approach of building Skill Cartridges™ needs to 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

Copyright © 2007 TEMIS - All Rights Reserved

Beyond Search >>

Luxid® for Life Sciences

Gracias !

WWW.TEMIS.COM