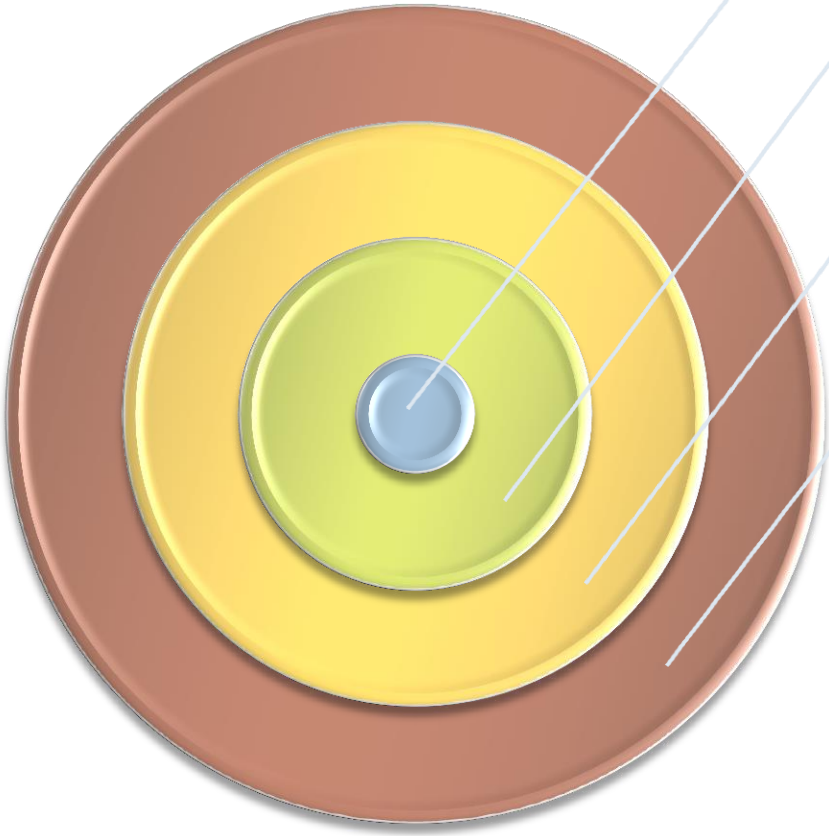
The Reuse COMPANY

Ontology Construction :Vocabulary Management

www.kr.inf.uc3m.es
www.reusecompany.com

System Knowledge Base: Ontology

What is an ontology for TRC?



Controlled vocabulary: valid terms, forbidden terms… Optionally can include a Glossary (description for every term)

Thesaurus: Relationships between terms: hierarchies, associations, synonyms…

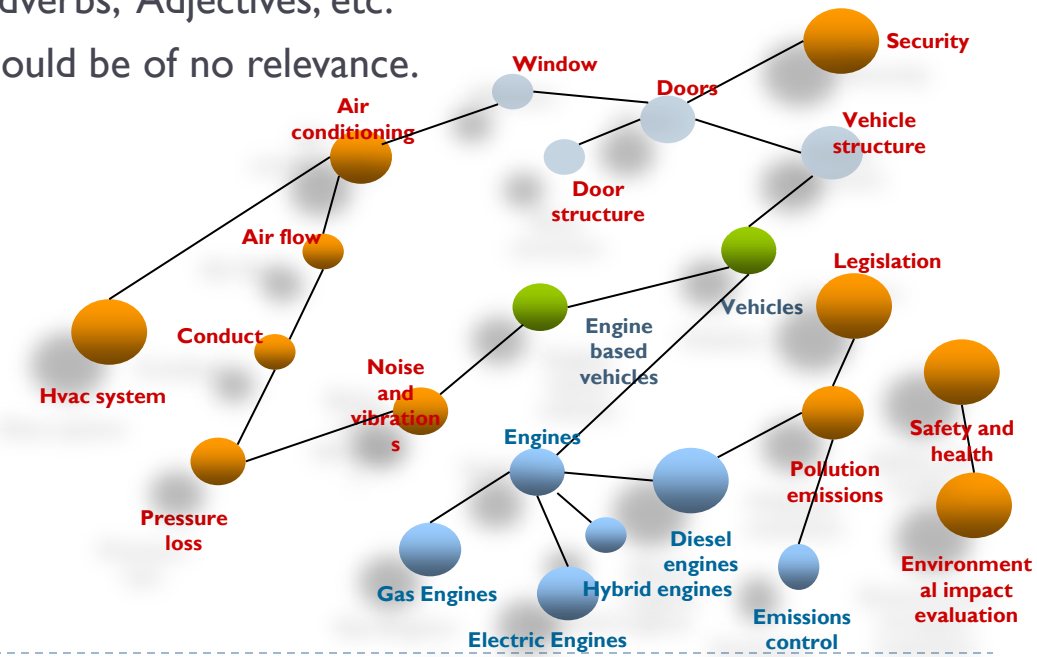
Light Ontology: syntactic and Semantic groupings for Terms and Actions (verbs). Domain terms and verbs

Patterns and Representation Schemas for Identifying (patterns) and representing (Schemas) the semantics of knowledge in Electronic Artifacts

© The REUSE Company knowledgeMANAGER v6.1
It's your knowledge, reuse it.

Controlled Vocabulary

- Needed for standardizing and normalizing the terminology used in the custom application. The input information must/should match the controlled vocabulary.
- Using a glossary with different categories of terms, the ontology may store:
  - Business related Terms : those terms central to the business area to be treated
  - General Language Terms:
  - Syntactically relevant phrases: Adverbs, Adjectives, etc.
  - Invalid terms: those terms that could be of no relevance.



© The REUSE Company knowledgeMANAGER v6.1
It's your knowledge, reuse it.

Controlled Vocabulary : Example for Requirements Authoring (RA)

UR044 : The Rad8 shall be able to identify hits at a minimum rate of 10 units per second

shall

identify

hit

minimum

unit

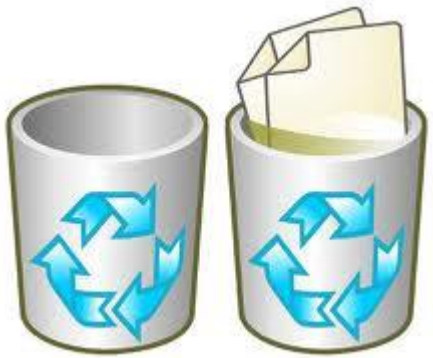
…..

second

The

to

at



Rad8

© The REUSE Company knowledgeMANAGER v6.1

It's your knowledge, reuse it.

What is knowledgeMANAGER

© The REUSE Company knowlegdeMANAGER v6.1

What is knowledeMANAGER

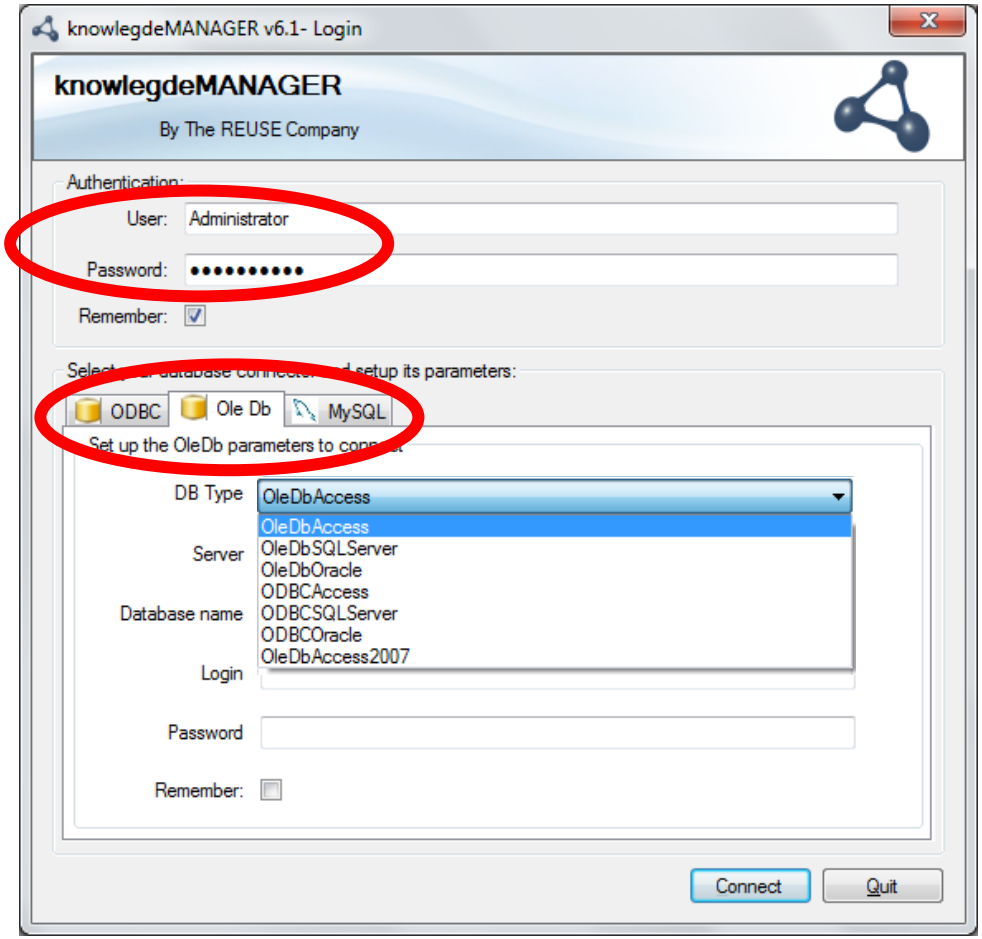
- knowledeMANAGER (kM) is a tool by TRC aimed to manage ontologies
- kM covers a bunch of different concepts:
  - Linguistic issues, vocabulary management, relationships, thesauri, patterns…
- kM (together with ontologies) represents the core of many TRC's projects
  - For RQA: managing vocabulary and relationships among concepts
  - For RAT: it allows to create boilerplates together with their formalization
  - For Semantic Search Engines and Reusable Repositories: it helps creating the semantics of the concepts

Connection screen and common interface

© The REUSE Company

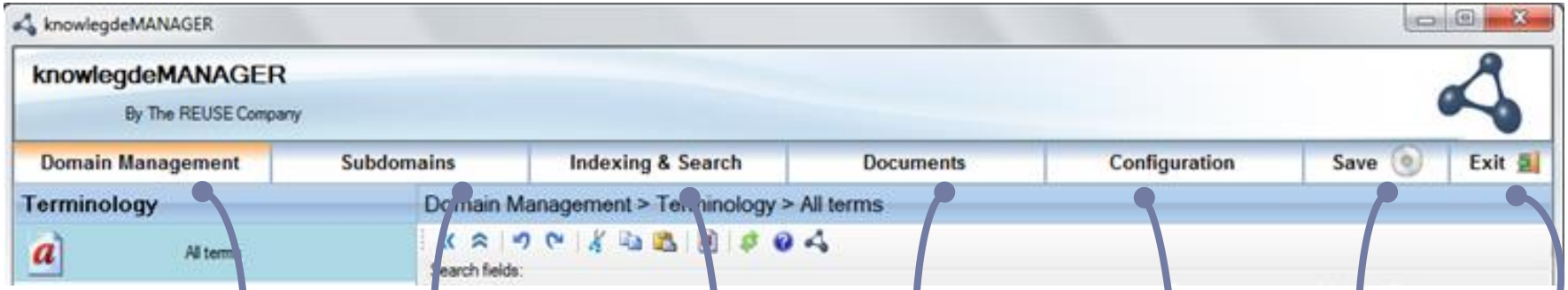
It's your knowledge, reuse it.

Connection screen



© The REUSE Company knowlegdeMANAGER v6.1

Interface: Main menu



Vocabulary
Linguistic rules
Light ontology
Boilerplates

Thesaurus
Suggestions
BP grouping

Indexing tests:
- Individual
- Regression

Asset management

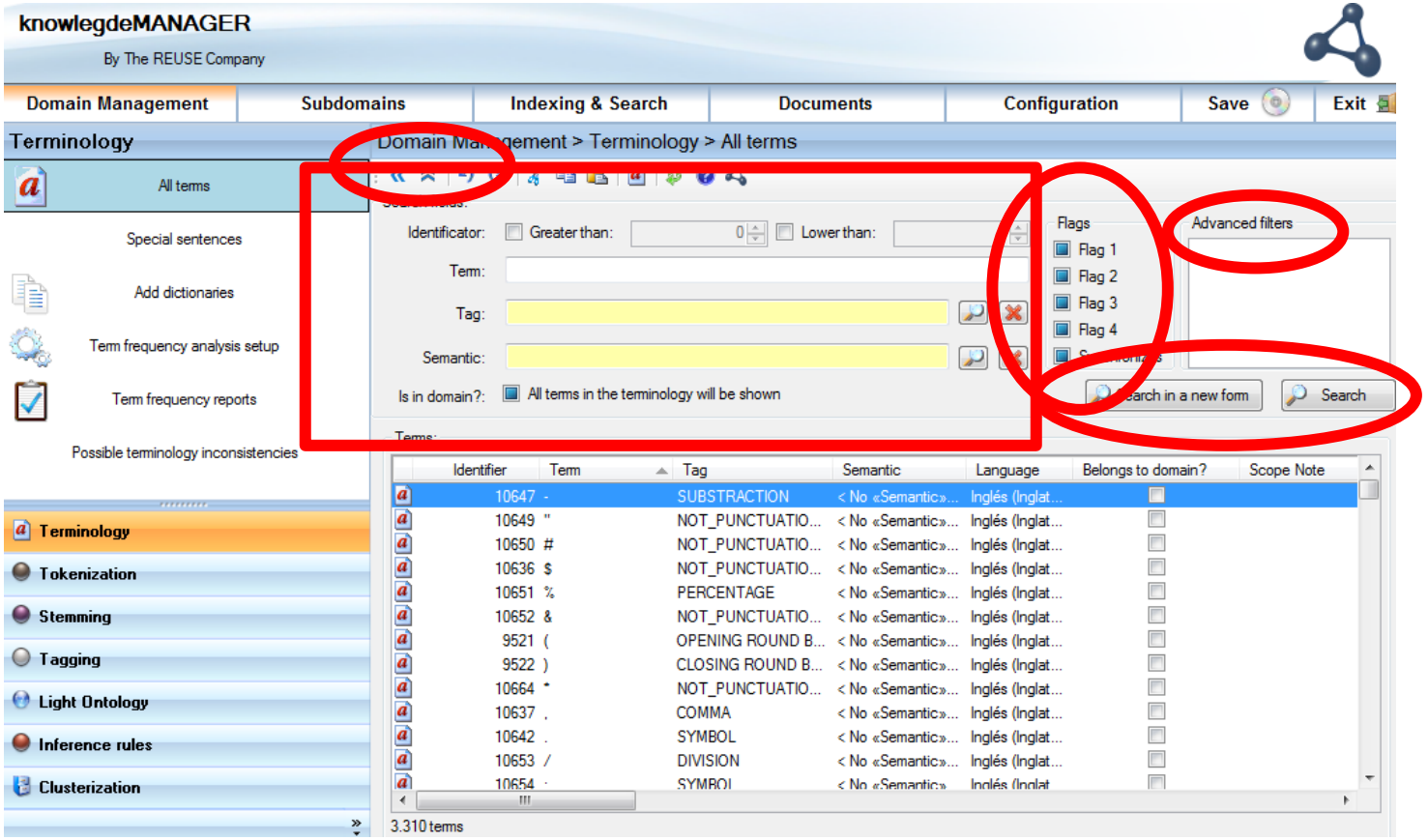
Indexing configuration
Optimization processes:
- Distances
- Copying patterns
- …

kM storage:
- A model in memory
- Concurrent

Exit

© The REUSE Company knowlegdeMANAGER v6.1

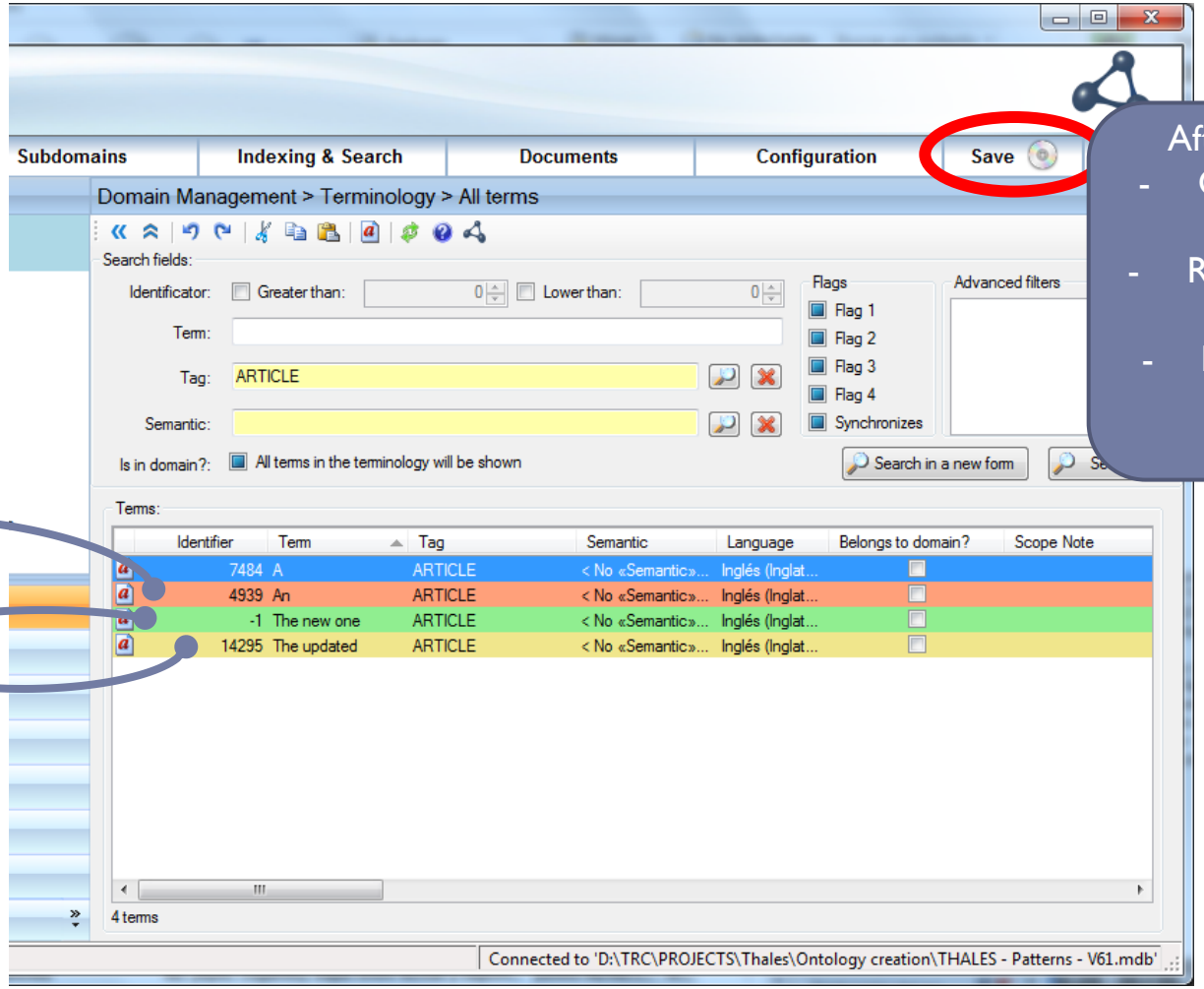
Filtering



© The REUSE Company knowledgeMANAGER v6.1

It's your knowledge, reuse it.

Common interface: new, updated, removed items



Removed item

Just added item

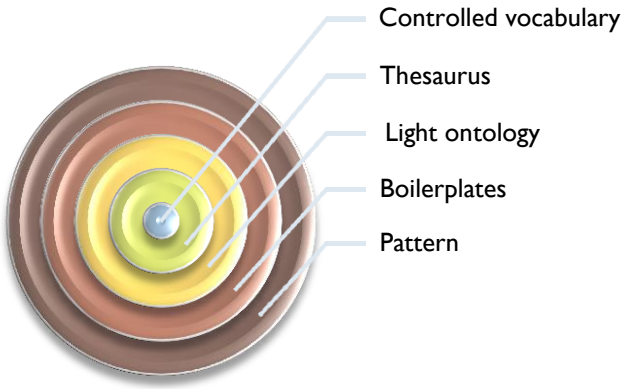
Updated item

After saving:
- Changes are stored
- Red rows will desappear
- Every other row will be white

© The REUSE Company knowledgeMANAGER v6.1

It's your knowledge, reuse it.

Vocabulary gathering



© The REUSE Company knowledgeMANAGER v6.1

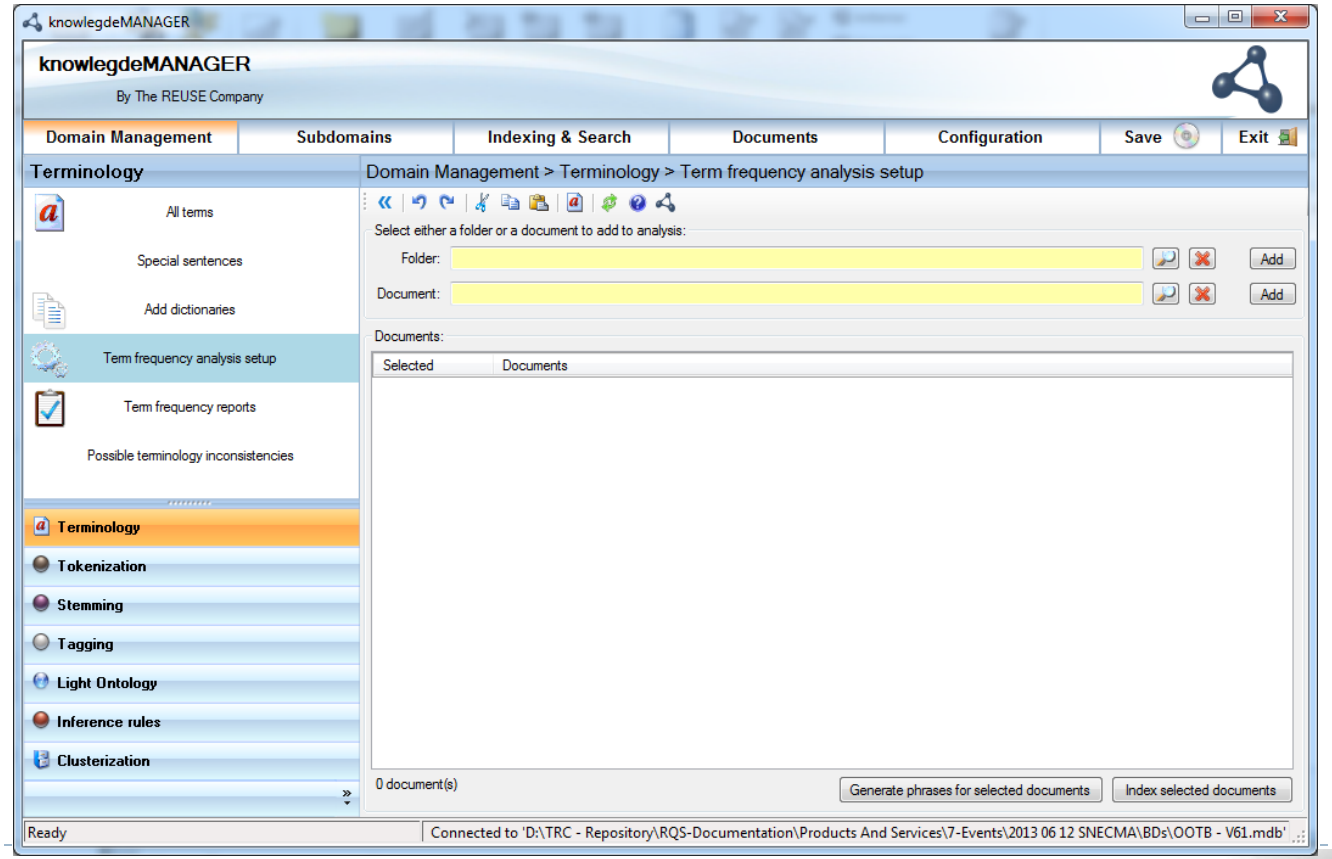
Gathering Vocabulary - To identify:

- Verbs Study and Identification
- Adjectives and Adverbs Study and Identification
- Acronyms identification
- Nouns identification
- Named Entities identification
- Simple terms identification
- Simple terms Validation
- Compound terms (phrases) Identification
- Verification and Validation of terms

© The REUSE Company knowledgeMANAGER v6.1

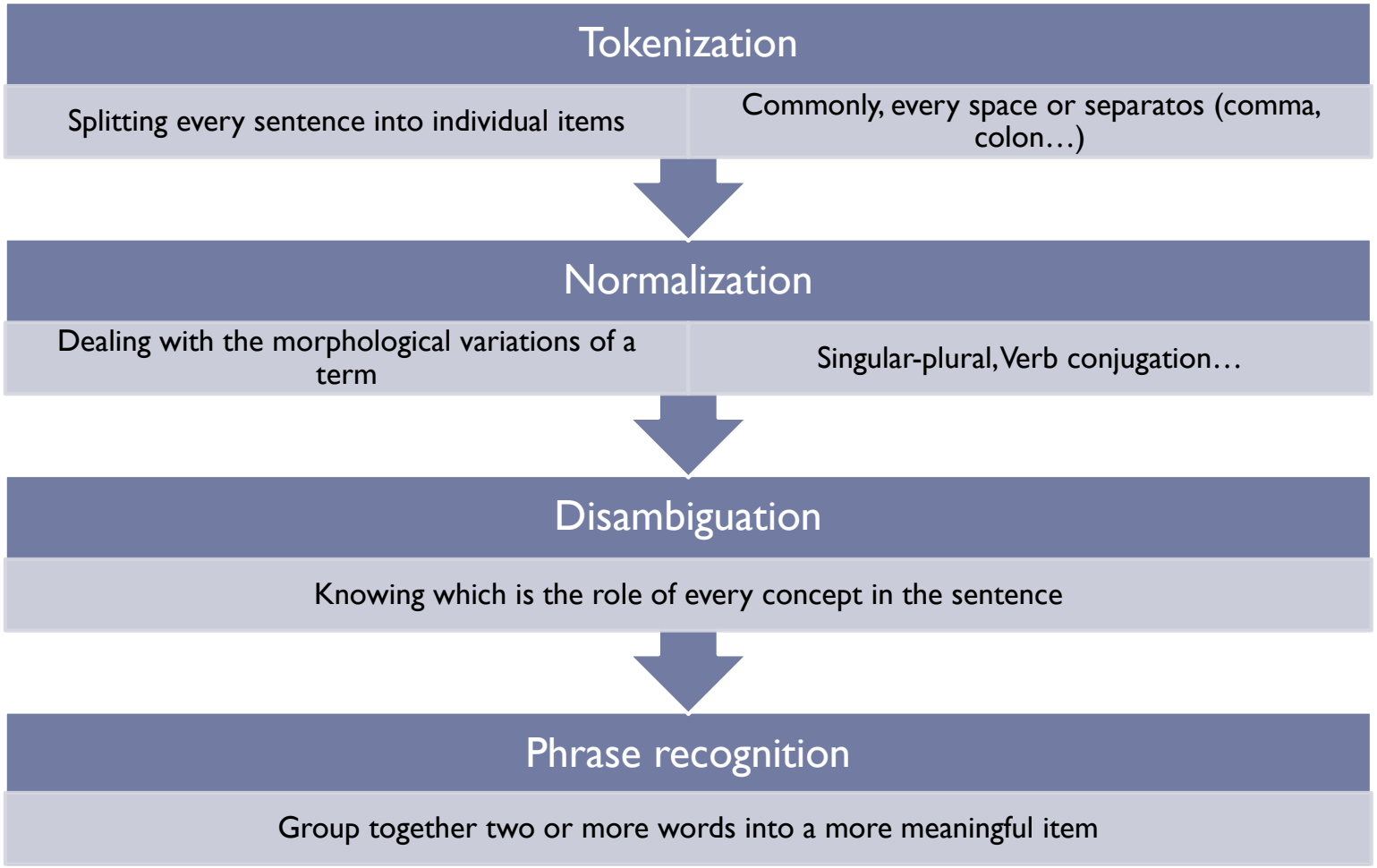
Gathering Vocabulary – How to:

- Indexing Requirements or other documents
- An iterative process (do not be afraid of repeating the process)



© The REUSE Company knowlegdeMANAGER v6.1
It's your knowledge, reuse it.

Gathering Vocabulary : How it works



© The REUSE Company knowledgeMANAGER v6.1
It's your knowledge, reuse it.

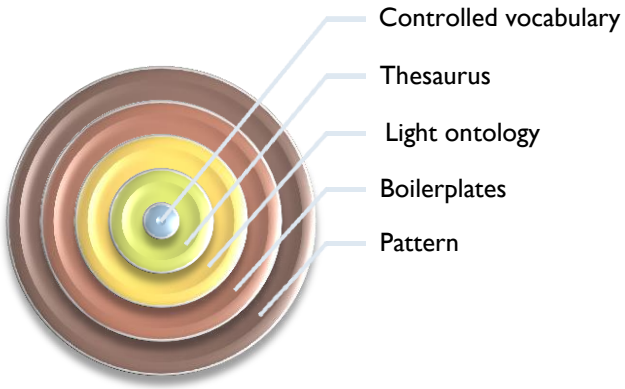
Vocabulary Gathering: How it works

All Radars shan't identify the following targeting enemy objectives:

Tokenization	[All] [Radars] [shall] [not] [identify] [the] [following] [targeting] [enemy] [objectives] [:]
Normalization	[All] [Radar] [shall] [not] [identify] [the] [following/follow] [targeting/Target] [enemy] [objective] [:]
Disambiguation	[All] [Radar] [shall] [not] [identify] [the] [following] [Target] [enemy] [objective] [:]
Phrase Identification	[All] [Radar] [shall] [not] [identify] [the] [following] [Target] [enemy objective] [:]

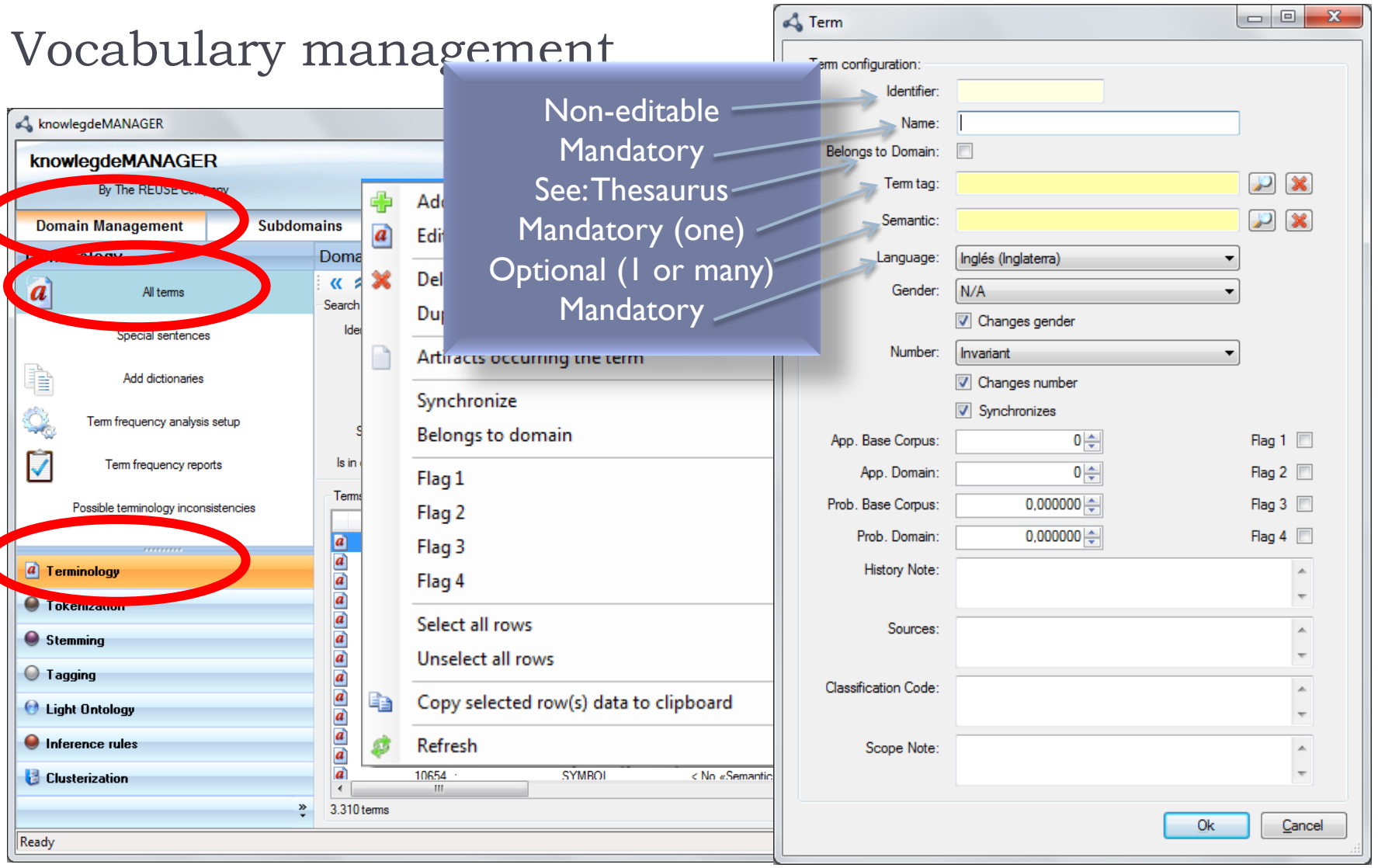
© The REUSE Company knowledgeMANAGER v6.1

Vocabulary management



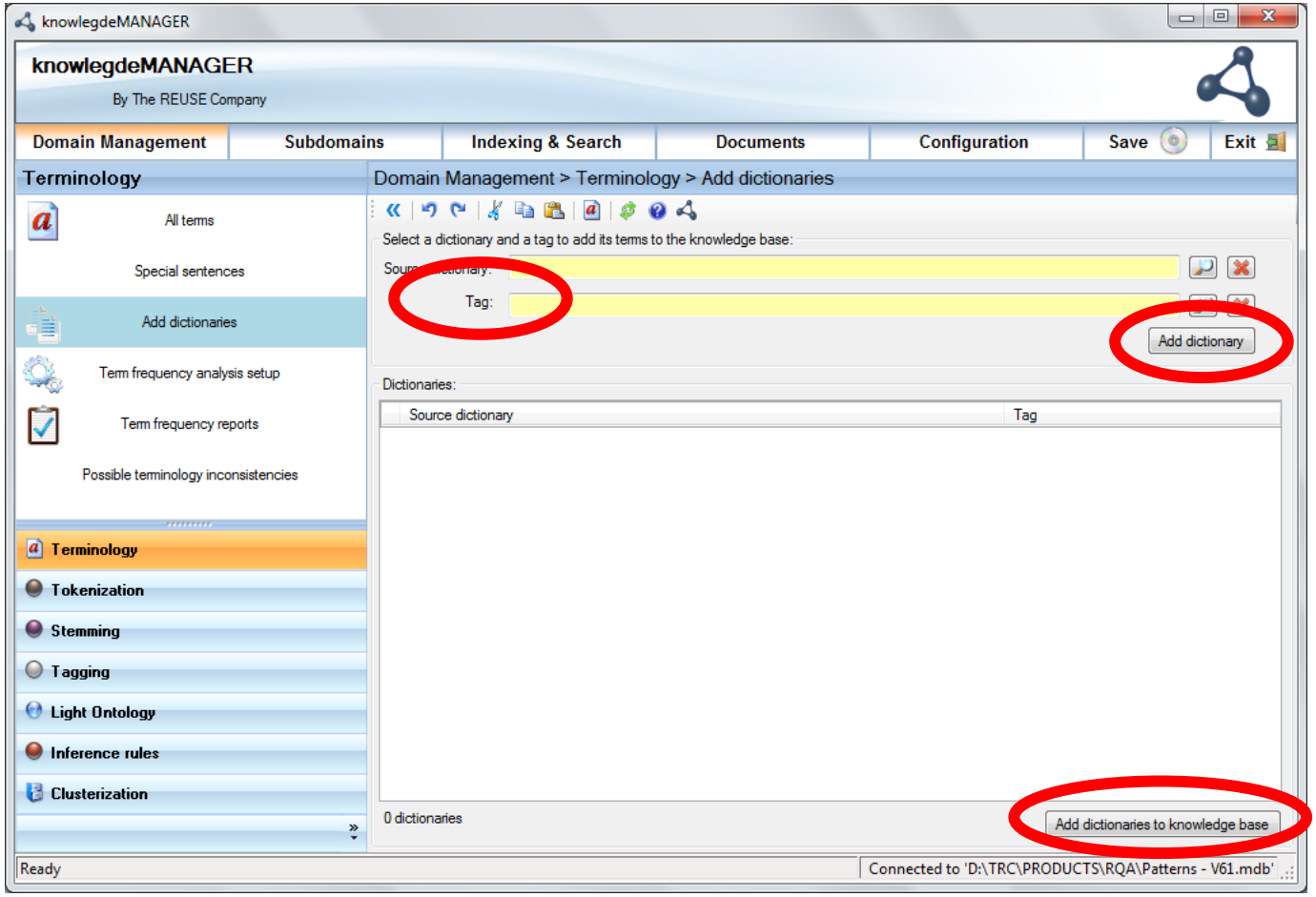
© The REUSE Company knowledgeMANAGER v6.1
It's your knowledge, reuse it.

Vocabulary management



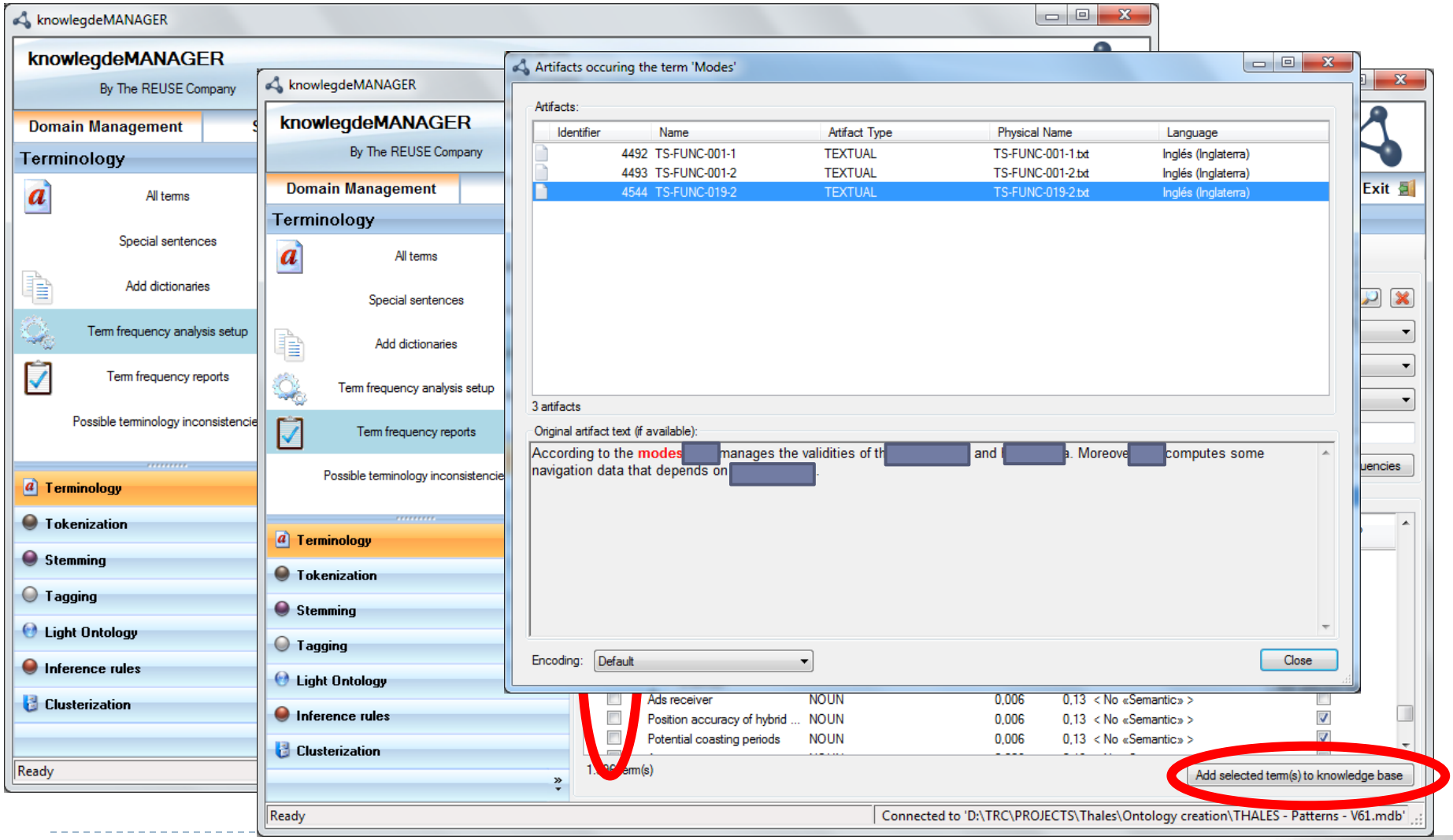
© The REUSE Company knowledgeMANAGER v6.1

Vocaburaly: add dictionaries



© The REUSE Company knowlegdeMANAGER v6.1
It's your knowledge, reuse it.

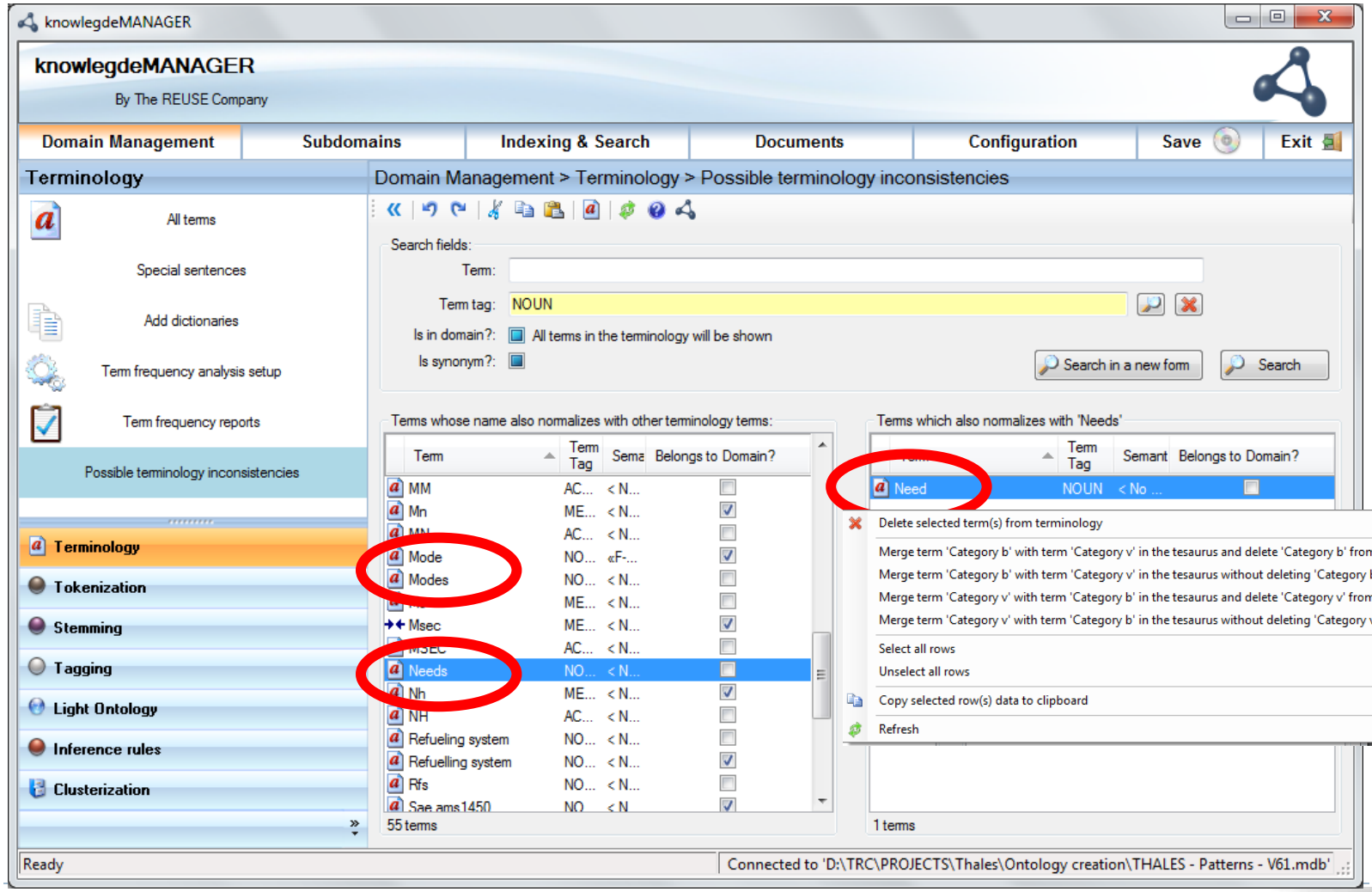
Vocaburaly: freq analysis



© The REUSE Company knowlegdeMANAGER v6.1

It's your knowledge, reuse it.

Vocabulary: inconsistencies



© The REUSE Company knowlegdeMANAGER v6.1

- This can happen when importing thesaurus
- Or normalization rules
- Or disambiguation rules

© The REUSE Company knowlegdeMANAGER v6.1

Tokenization

- Objective: to get tokens out of input text
- To correct possible misspelling problems
- To change input phrases
- To identify Named Entities:
  - Regular expressions

© The REUSE Company

It's your knowledge, reuse it.

Tokenization Process

- 0 - Misspelling
  - Always first
- 1- Sorted sequence of Regular Expressions
  - Order defined by the user

© The REUSE Company knowledgeMANAGER v6.1

Misspelling Types

- ' PRE-PROCESSING
- MS_PRE_Contiguous_Double_Chars = 1
- MS_PRE_Word_Endings = 2
- MS_PRE_Whole_Word = 3

- ' PROCESSING
- MS_PROC_Contiguous_Double_Chars = 8

- ' IDENTIFICATION NUMBERS WRITTEN LETTERS
- Identification_Words_To_Numbers = 9

© The REUSE Company knowledgeMANAGER v6.1

Regular Expressions Types

- RE_SimpleReplace = 10
- RE_Groups_Replace = 11

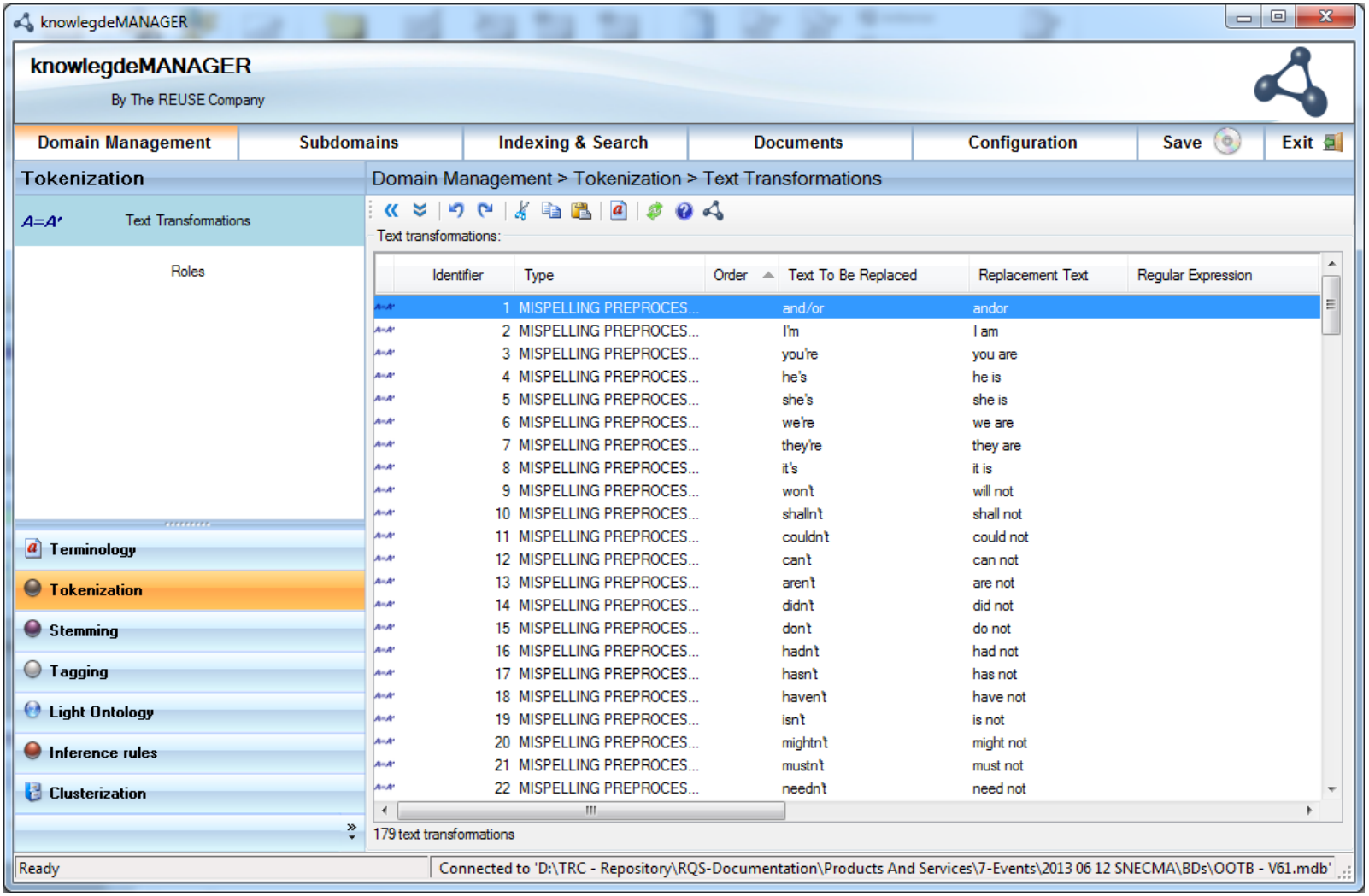
- Remove_Symbols = 12
- Remove_Multiple_Spaces = 13

- Separate_Symbols = 14
- Separate_Symbols_From_Entities = 15

- RE_Named_Entities_Identification = 20
- RE_Named_Entities_Identification_With_Delimiters = 21

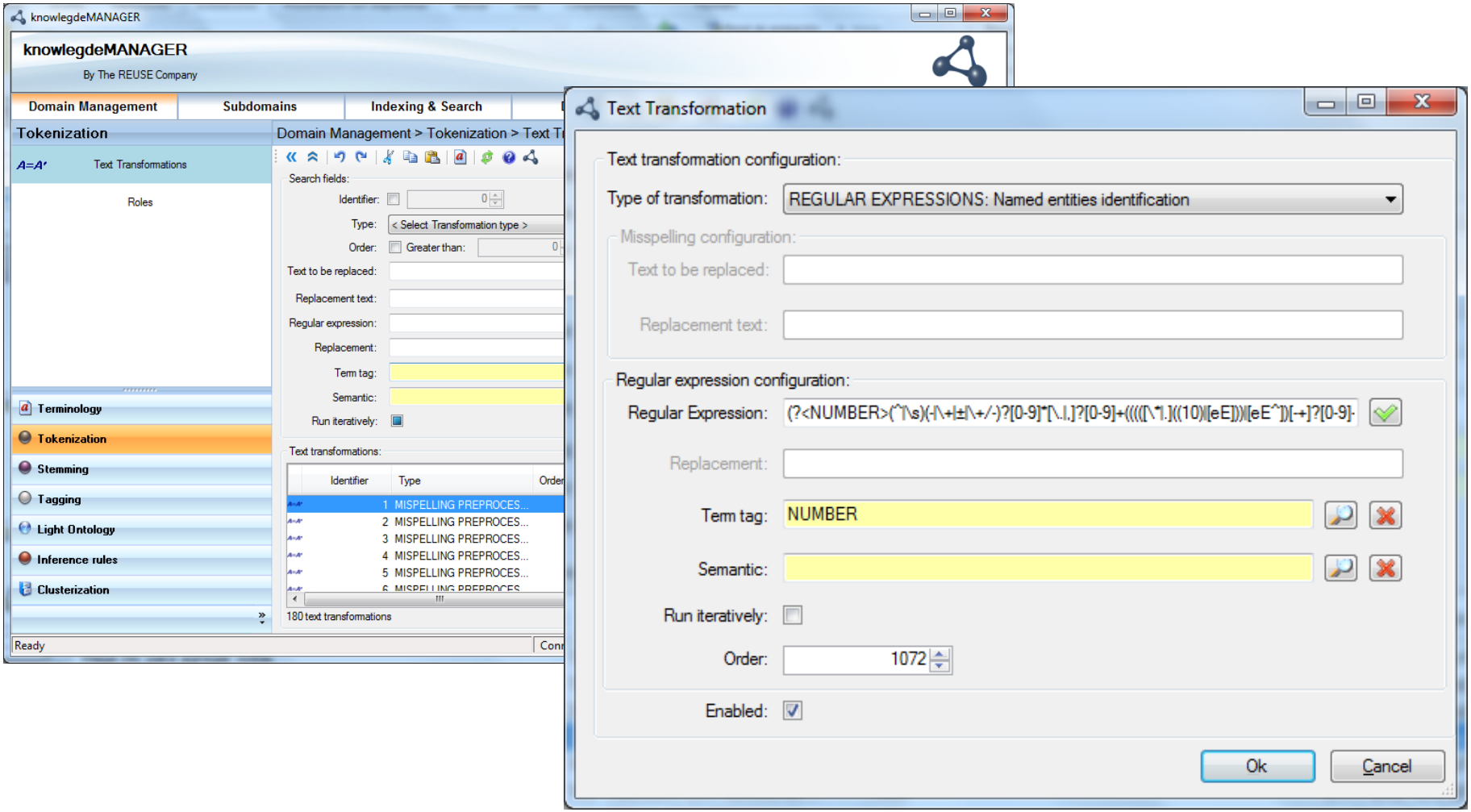
© The REUSE Company knowlegdeMANAGER v6.1
It's your knowledge, reuse it.

Tokenization



© The REUSE Company knowlegdeMANAGER v6.1
It's your knowledge, reuse it.

Tokenization

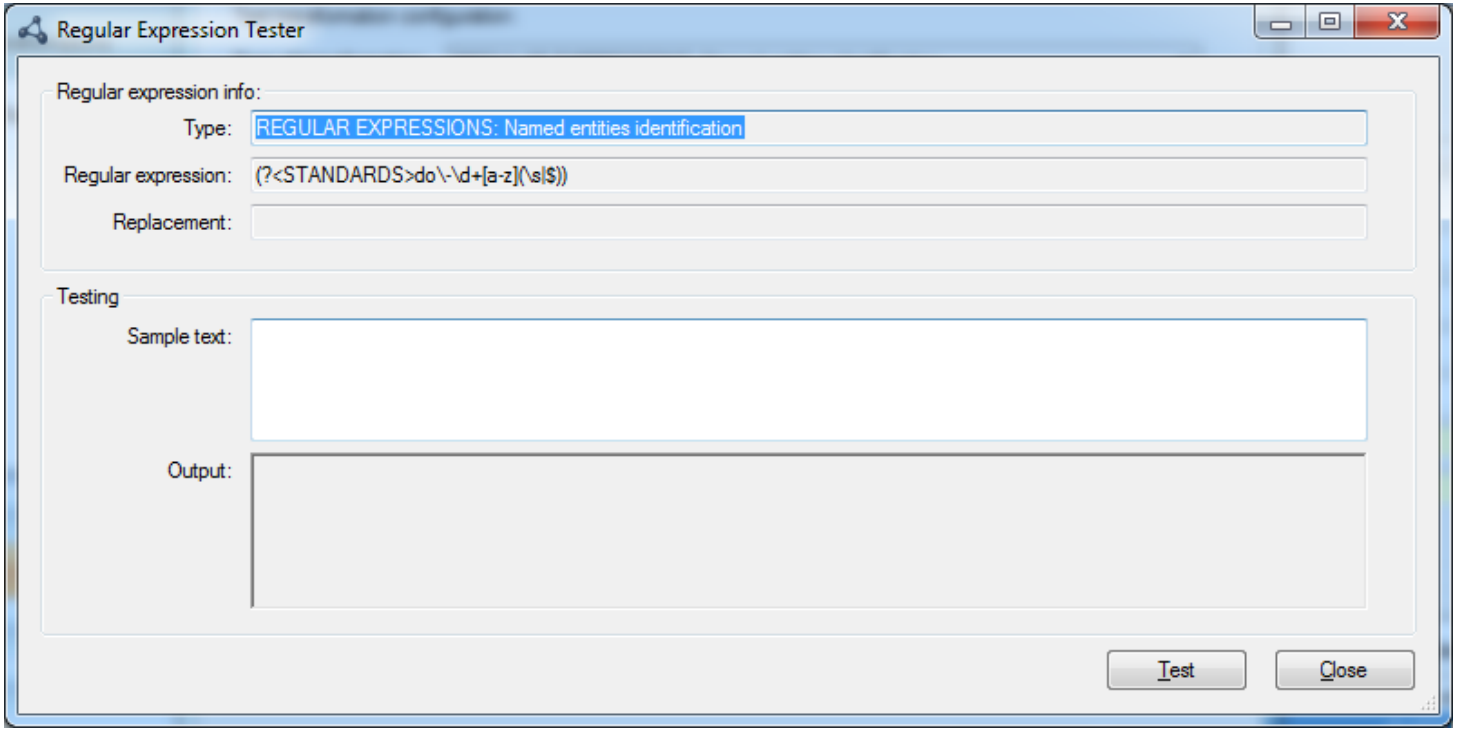


© The REUSE Company knowlegdeMANAGER v6.1

It's your knowledge, reuse it.

Tokenization : Testing Regular Expressions

- NOT valid regular expressions presuming Upper case input
  - The indexer transforms everything to LOWER



© The REUSE Company knowledgeMANAGER v6.1
It's your knowledge, reuse it.

Tokenization Problems

- The tokenization process can produce strange results when a domain changes
  - Ex: the following RegEX: (?<1>[a-z]+)(?<2>[\-/])(?<3>[0-9a-z]+)
    - works well in English language to separate units and words by /, like in Km/h, or Input/Output , but if it is used in French, the Arrêt/Démarrage, would produce a strange result Arrê t / D émarrage
    - Solution:
  - (?<1>[a-záéíóúàèìòùâêîôûäëïöü]+)(?<2>[\-/])(?<3>[0-9a-záéíóúàèìòùâêîôûäëïöü]+)

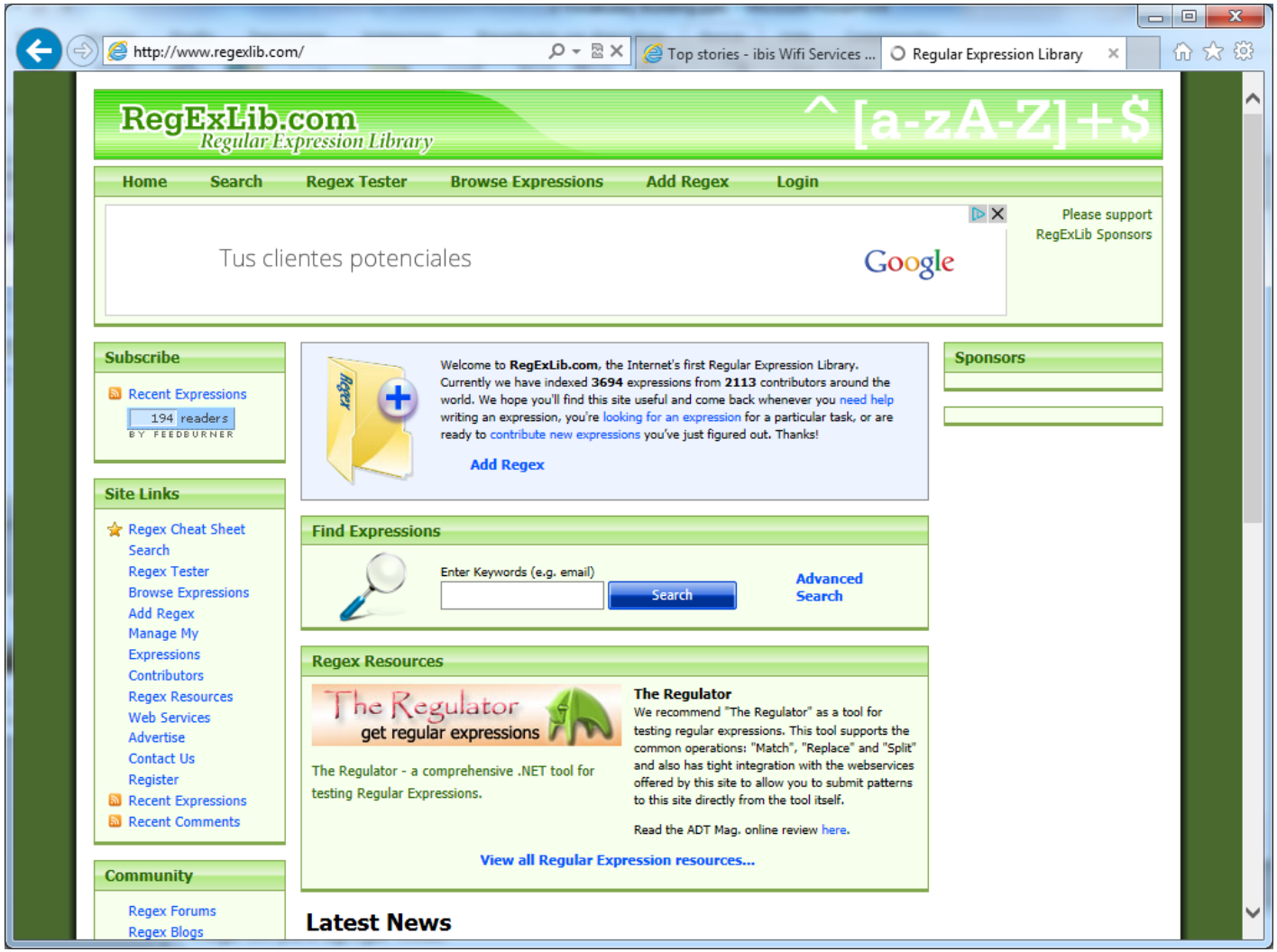
- The tokenization process can produce strange results, depending on the order of application of the regular expressions

© The REUSE Company knowledgeMANAGER v6.1

External tools:

© The REUSE Company knowledgeMANAGER v6.1

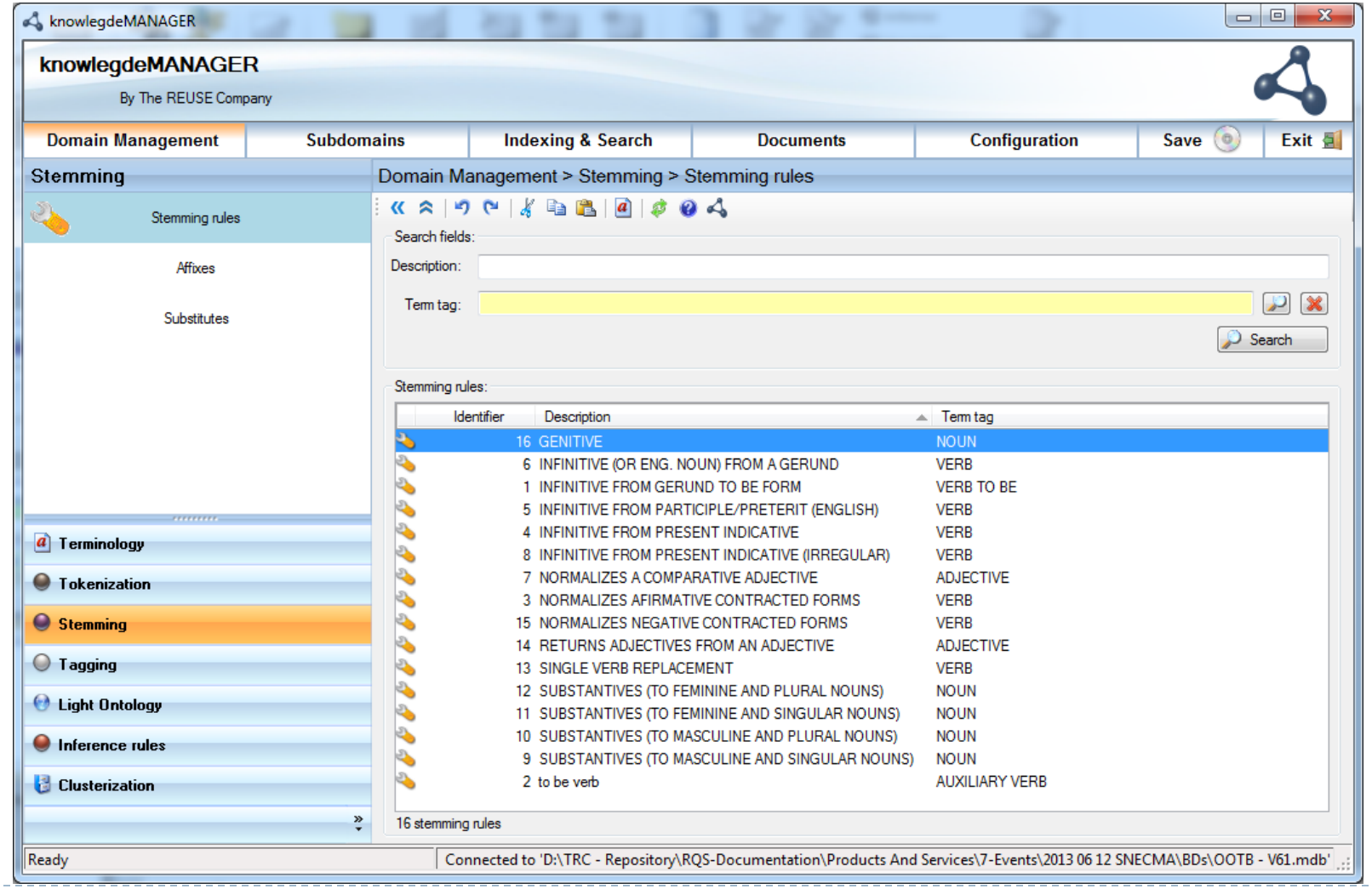
External tools:



© The REUSE Company knowledgeMANAGER v6.1

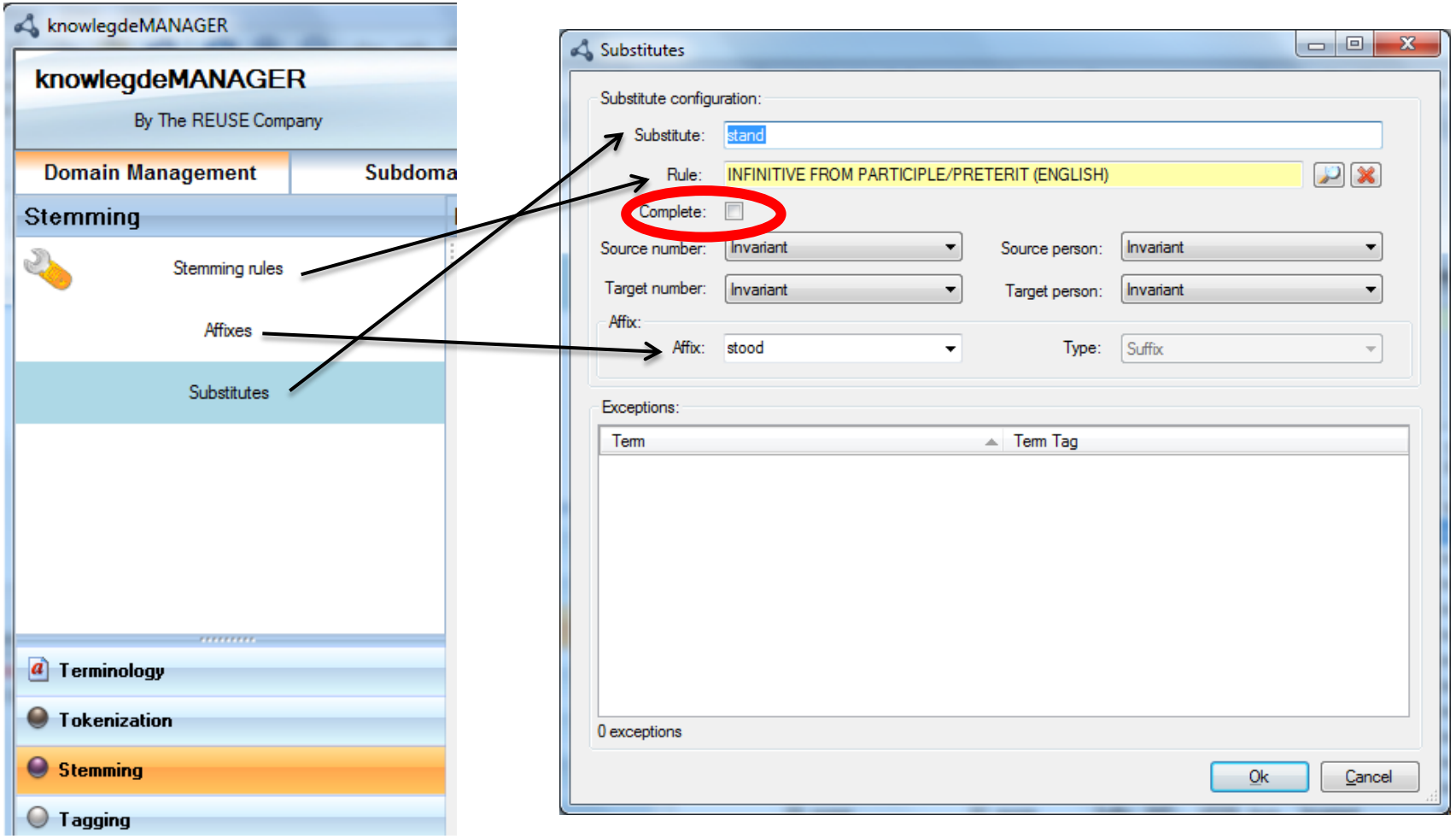
It's your knowledge, reuse it.

Normalization / Stemming



© The REUSE Company knowlegdeMANAGER v6.1

Normalization (Stemming)



© The REUSE Company knowlegdeMANAGER v6.1

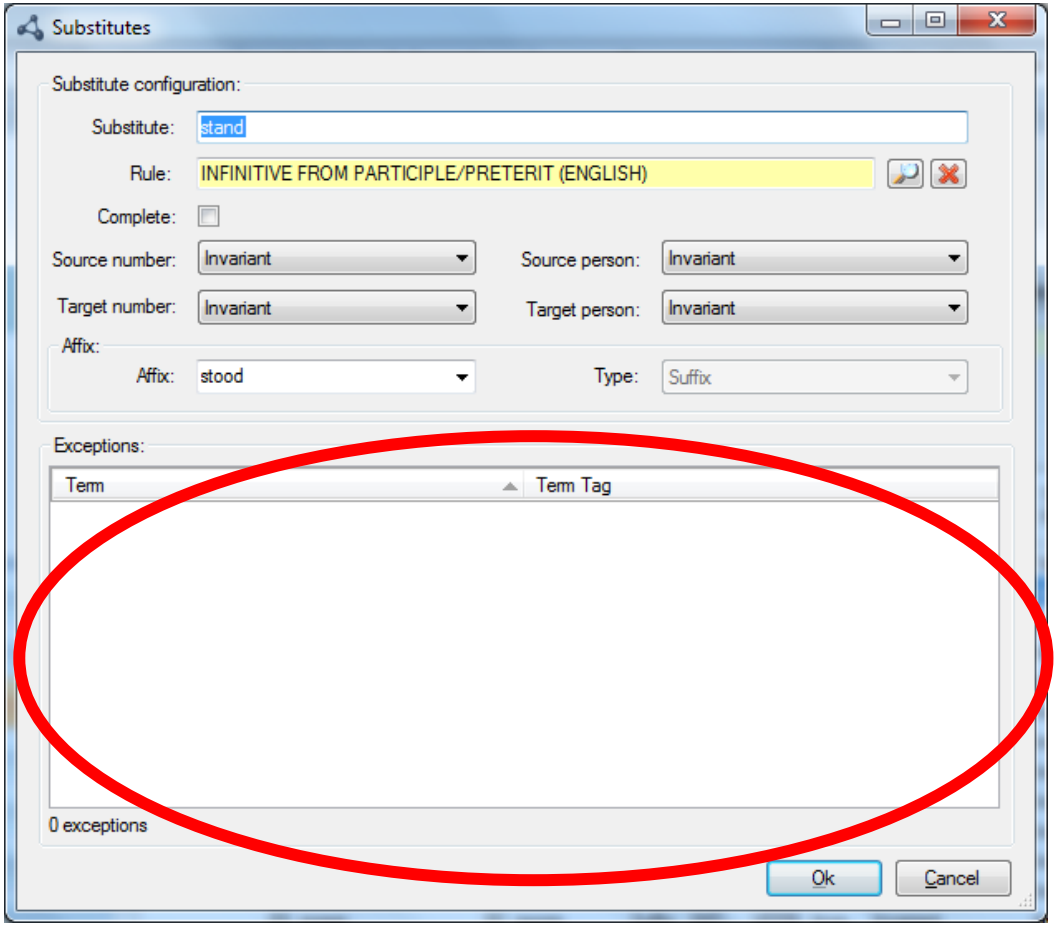
It's your knowledge, reuse it.

Normalization to Plural

- The Normalization rule should be “nothing” to “s” or “es” or…
- No support to “empty” suffix
- So particular rules should be created
- Eg. For Interface -> Suffix “ace” to “aces”
- Etc.

© The REUSE Company knowledgeMANAGER v6.1

Normalization (Stemming) - Exceptions



- Exceptions are applied to terms before the affix work
- If a term is the exception => do not apply the substitution.
- For Latin Languages, Exceptions are also applied after reflective word separations
  - Eg. Retirarse –
    - Retirar SE

© The REUSE Company knowledgeMANAGER v6.1

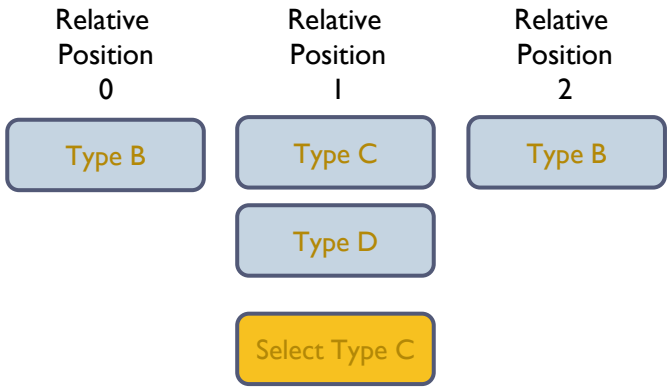
Disambiguation / Tagging

- Process needed to solve multiple candidates created by the normalization process.
- Based on Context
  - At both sides of the ambiguous term
- Based on Disambiguation rules + Bigrams Statistics + Rules of thumb
- Application Order:
  - Disambiguation rules
  - Bigrams Statistics
  - Rules of Thumb
    - The tag of the candidate that matches with the input text (if not unclassified)
    - The tag with more probability of occurrence by itself
    - The Noun tag if a candidate is of that type
    - The tag of the first candidate

© The REUSE Company knowledgeMANAGER v6.1

Disambiguation / Tagging : (Rules)

- Disambiguation rules



- If In Position 0 we have a candidate of type B and in Position 1 we have a candidate of Type C and a Candidate of Type D and in position 2 we have a candidate of type B => select the candidate of type C in position 1 as the right one

© The REUSE Company knowledgeMANAGER v6.1

Disambiguation / Tagging : (Rules)

- Example:

Initial Sentence

He	is	as	intelligent	as	my	father
Pronoun	Be Verb	Conjunction	Noun	Conjunction	Determiner	Noun
			Adjective			

Rule

	Relative Position 0	Relative Position 1	Relative Position 2
	Type B	Type C	Type B
		Type D	
		Select Type C	

Final Sentence

He	is	as	intelligent	as	my	father
Pronoun	Be Verb	Conjunction	Adjective	Conjunction	Determiner	Noun

© The REUSE Company knowledgeMANAGER v6.1
It's your knowledge, reuse it.

Disambiguation / Tagging : (Rules)

- Rules are applied in Order of weight
- All the valid candidates at the moment of the application of the rule are considered
  - Which can imply possible problems
- Eg. If the sentence is:

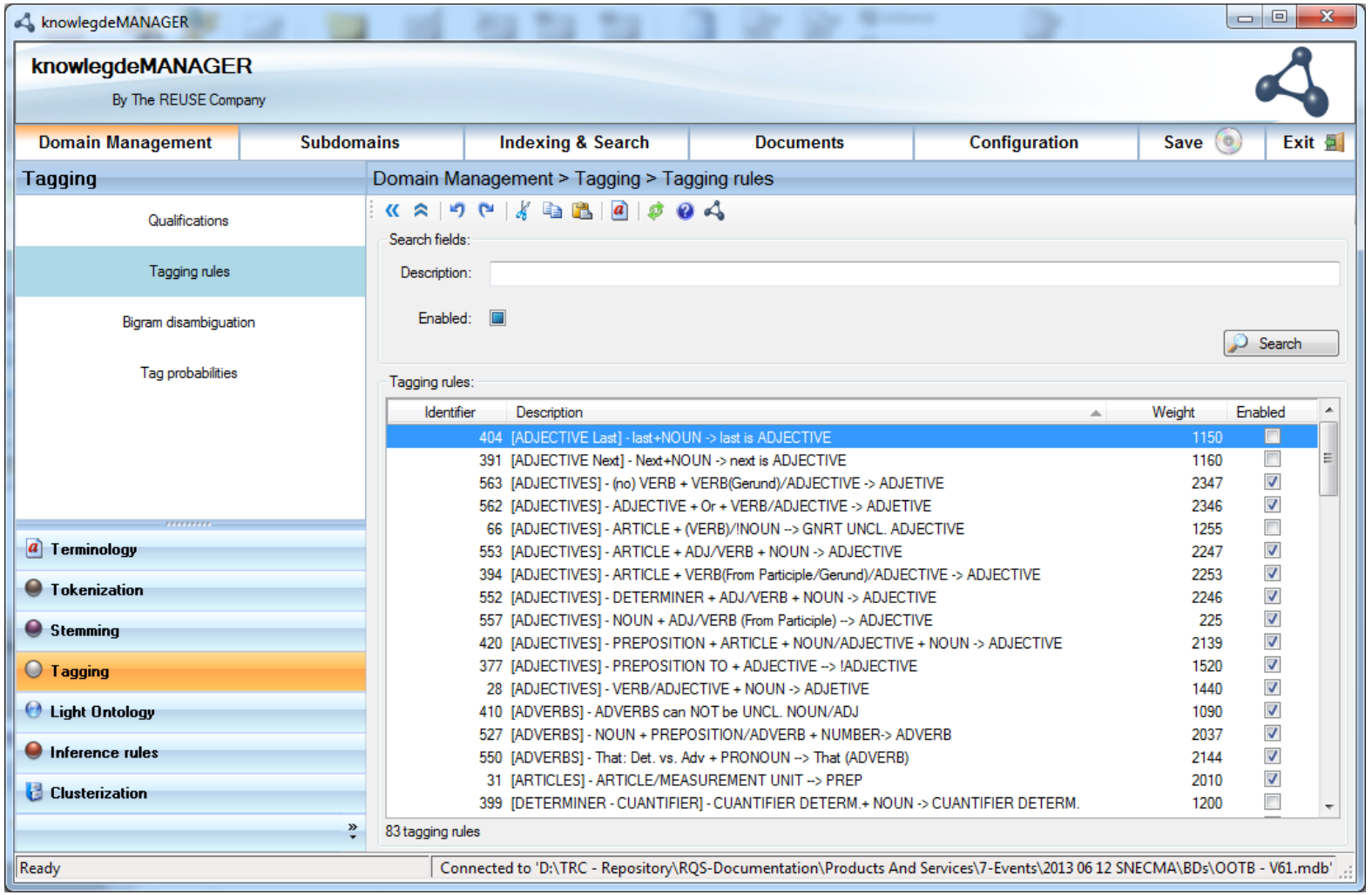
He	is	as	intelligent	as	my	father
Pronoun	Be Verb	Conjunction	Noun	Conjunction	Determiner	Noun
		Adverb	Adjective	Adverb		

- And the rule of the previous slide is applied first => the rest of the rules should work with the following sentence:

He	is	as	intelligent	as	my	father
Pronoun	Be Verb	Conjunction	Adjective	Conjunction	Determiner	Noun
		Adverb		Adverb		

© The REUSE Company knowledgeMANAGER v6.1
It's your knowledge, reuse it.

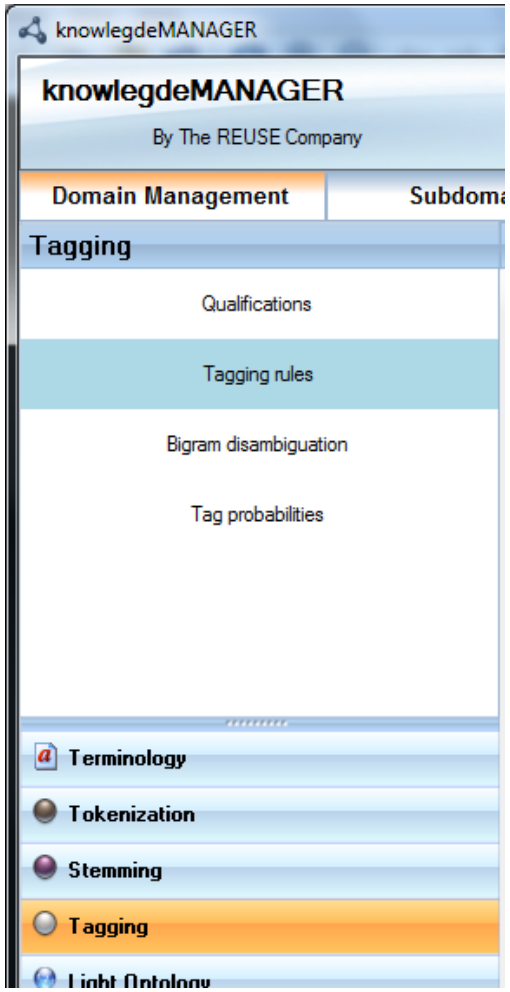
Disambiguation / Tagging : (Rules)

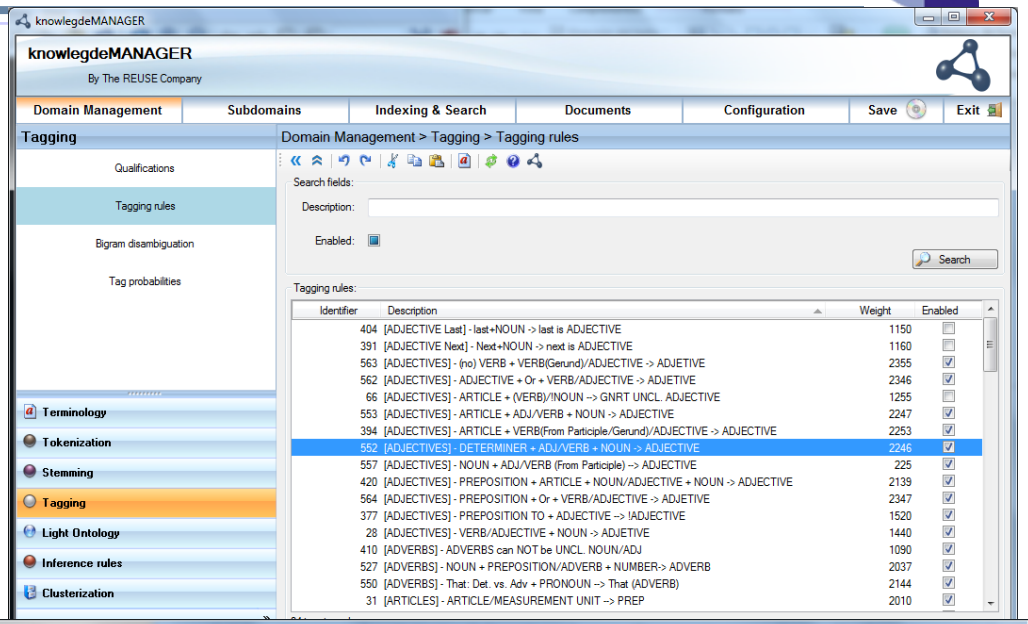


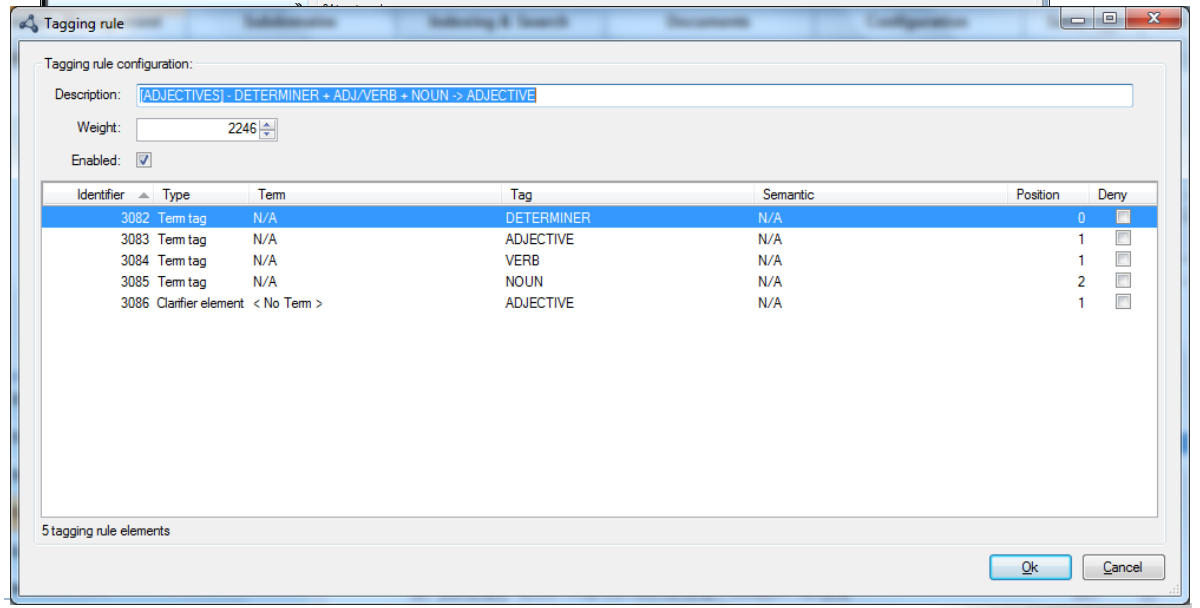
© The REUSE Company knowlegdeMANAGER v6.1

It's your knowledge, reuse it.

Tagging : (Rules)







© The REUSE Company knowlegdeMANAGER v6.1

It's your knowledge, reuse it.

Disambiguation / Tagging : (Bigrams)

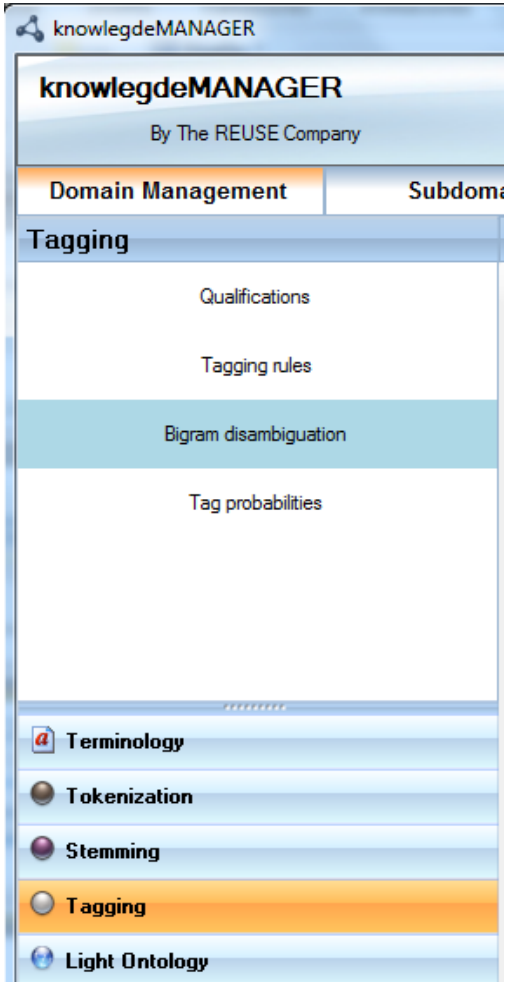
- Eg. If the sentence is:

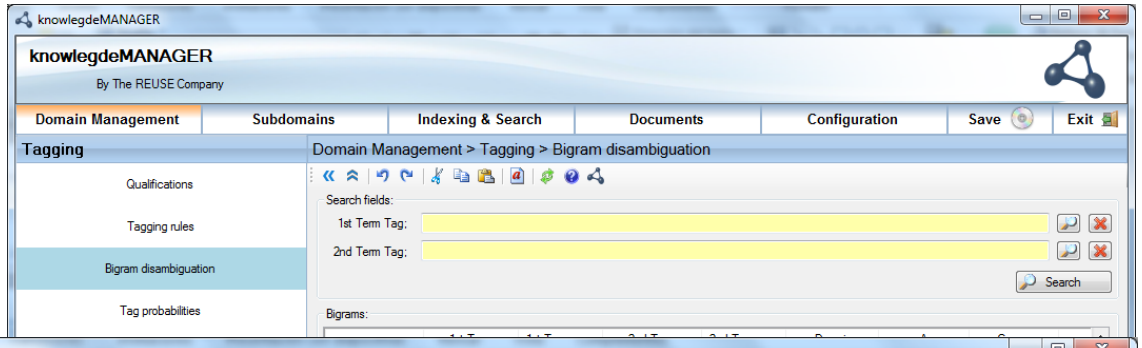
- The following statistics will be checked:
  - Be Verb – Conjunction
  - Be Verb – Adverb
  - Conjunction – Adjective
  - Adverb – Adjective

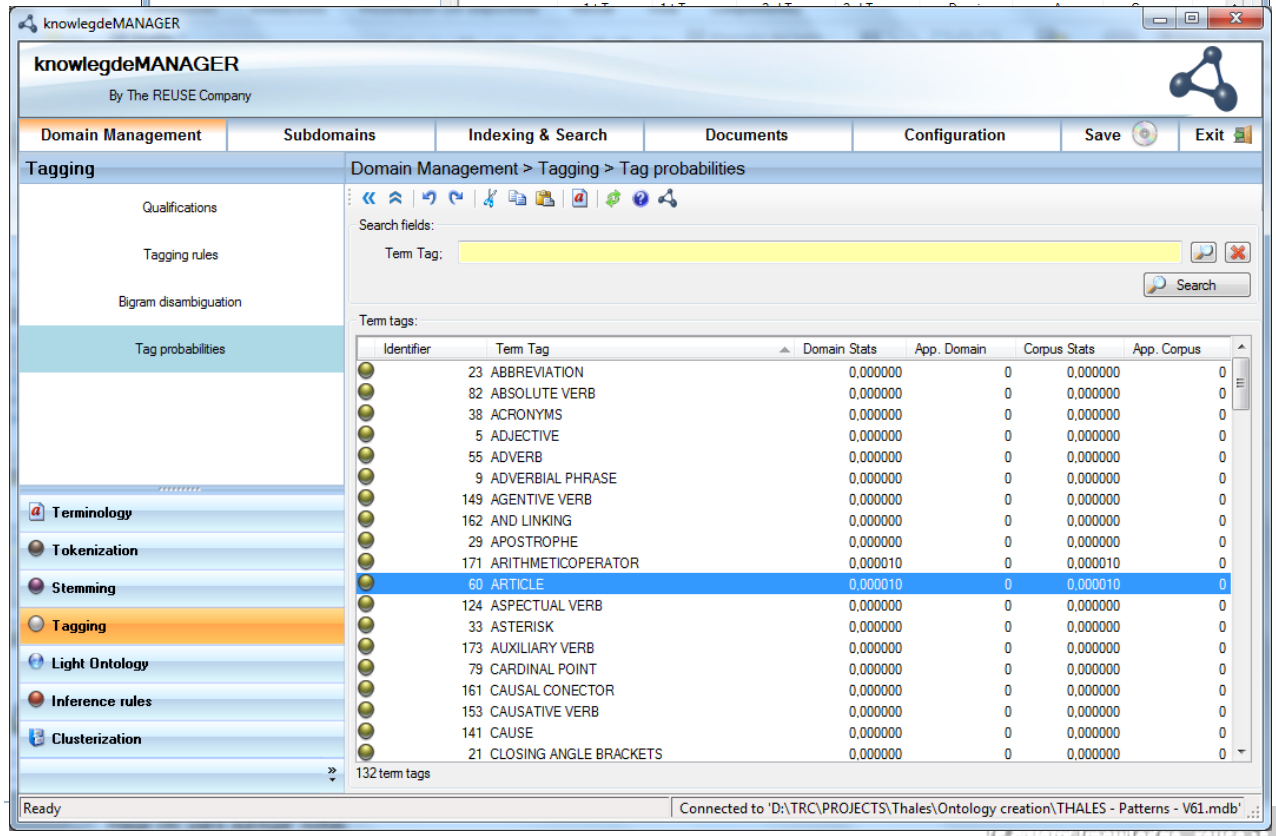
  - To decide which is the right tag for the first "AS"

© The REUSE Company knowledgeMANAGER v6.1
It's your knowledge, reuse it.

Tagging : (Bigrams)







© The REUSE Company knowlegdeMANAGER v6.1

It's your knowledge, reuse it.

The Reuse COMPANY

Margarita Salas, 16 2nd Floor
Innovation Center
LEGATEC Technology Park
28919 Leganés – Madrid
SPAIN – EU

Tel: (+34) 912 17 25 96
Fax: (+34) 916 80 98 26

http://www.reusecompany.com

@ReuseCompany

contact@reusecompany.com