

KNOWLEDGE DISCOVERY PROCESS
AUTOMATIC EXTRACTION OF
VOCABULARY AND RELATIONSHIPS
FROM LEGACY DOCUMENTATION...
IN 10 MINUTES!



THE
REUSE
COMPANY

The Presenter



Ilyes Yousfi



ilyes.yousfi@reusecompany.com



www.linkedin.com/in/ilyesyousfi/en



THE
REUSE
COMPANY

CONTENT

Why capturing knowledge in legacy documents?

Ontologies as key support to knowledge extraction

Approach by TRC: based on NLP and semantic analysis

Demo

Q&A

Capturing knowledge in documents: why?

EXTRACT KNOWLEDGE FROM LEGACY DOCUMENTS...IN 10 MINUTES!

1. Save time when analyzing the quality of your requirements! (Domain-specific concepts)
2. Rising complexity = Rising number of documents of many different kinds!
3. Knowledge is invaluable - Knowledge Management is a key_project-enabling_process in SE

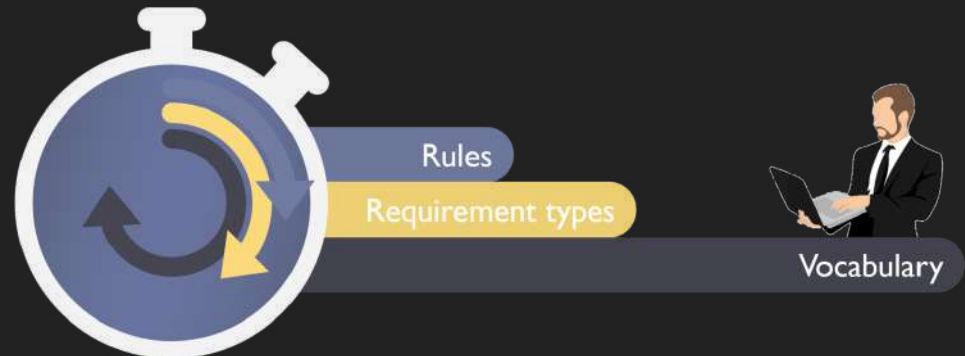


Capturing knowledge in documents: why?

EXTRACT KNOWLEDGE FROM LEGACY DOCUMENTS...IN 10 MINUTES!

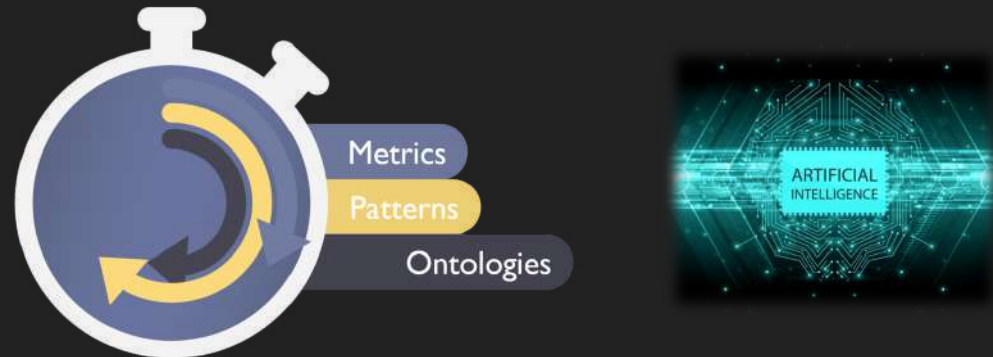
Manual Inspection

- Requires time
- Requires intensive SME support



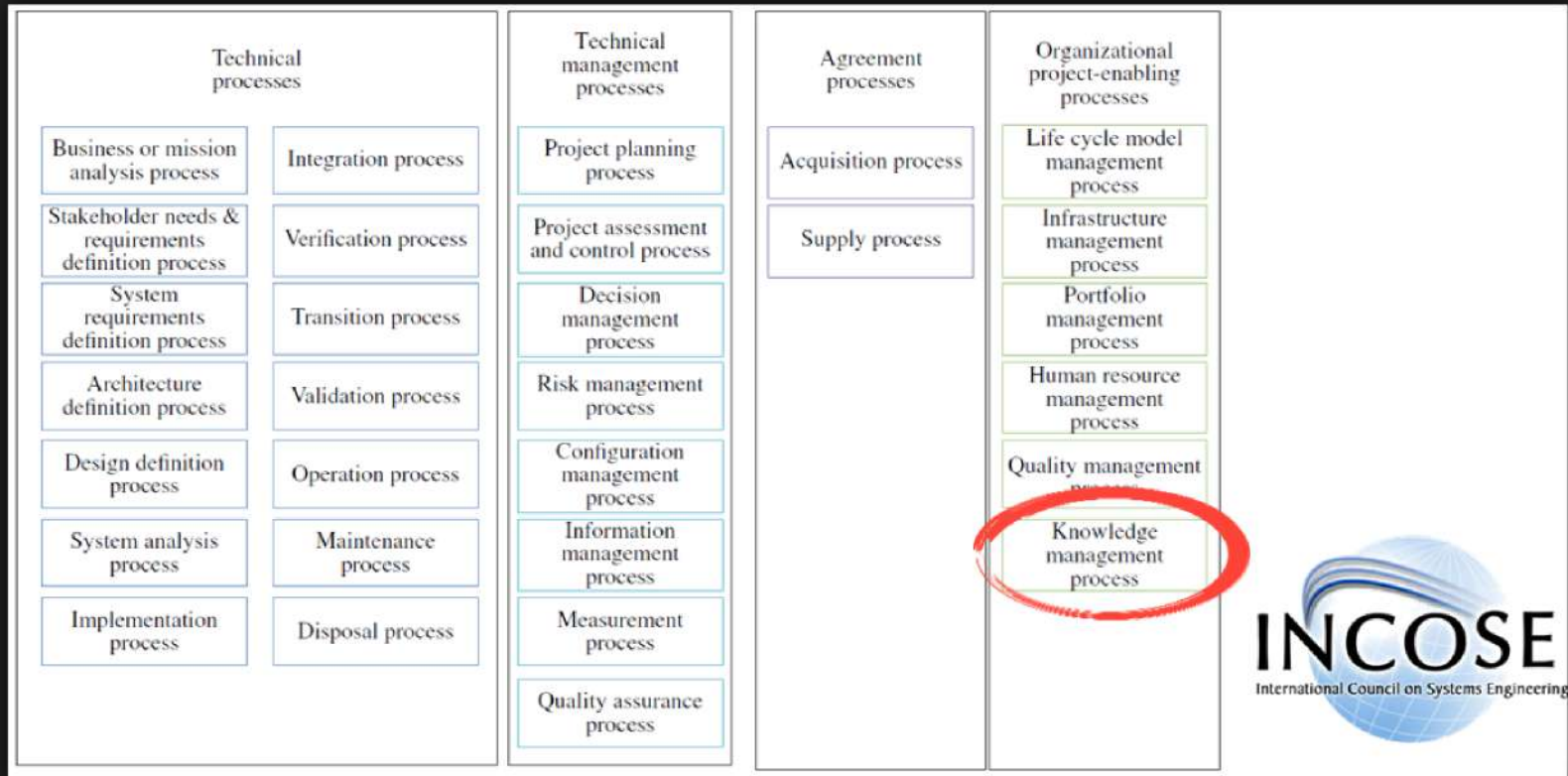
Automated Assessment (AI + NLP)

- Reduction of time
- Less support from SME



Capturing knowledge in documents: why?

EXTRACT KNOWLEDGE FROM LEGACY DOCUMENTS...IN 10 MINUTES!



Purpose : Create the capability and assets that enable the organization to **exploit opportunities to re-apply existing knowledge**

Capturing knowledge in documents: why?

EXTRACT KNOWLEDGE FROM LEGACY DOCUMENTS...IN 10 MINUTES!

Manual Inspection

- Requires time
- Requires intensive SME support



Rules

Requirement types

Vocabulary



Automated Assessment (AI + NLP)

- Reduction of time
- Less support from SME



Metrics

Patterns

Ontologies



CONTENT

Why capturing knowledge in legacy documents?

Ontologies as key support to knowledge extraction

Approach by TRC: based on NLP and semantic analysis

Demo

Q&A

Ontologies as key support to knowledge extraction

EXTRACT KNOWLEDGE FROM LEGACY DOCUMENTS...IN 10 MINUTES!

What is an ontology ?

- **Before** : philosophical concept to study the nature of the existence
- **Today** : IT concept - The list of terms or controlled and structured vocabulary (descriptors) that represents the key concepts of a given Knowledge domain

The **common ground** : Ask ourselves **what do we have** and **what do we need** to understand our world (KNOWLEDGE)

Ontologies as key support to knowledge extraction

EXTRACT KNOWLEDGE FROM LEGACY DOCUMENTS...IN 10 MINUTES!

The ontology for TRC



Ontologies as key support to knowledge extraction

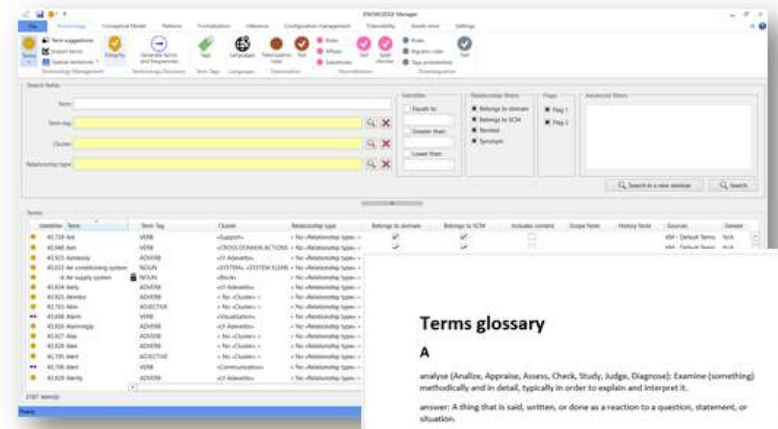
EXTRACT KNOWLEDGE FROM LEGACY DOCUMENTS...IN 10 MINUTES!



01

Vocabulary

Controlled Organizational and Project Vocabulary for a common understanding among stakeholders



Terms glossary

- A**
- analyse** (Analyze, Appraise, Assess, Check, Study, Judge, Diagnose): Examine (something) methodically and in detail, typically in order to explain and interpret it.
- answer**: A thing that is said, written, or done as a reaction to a question, statement, or situation.
- ask** (Ask for, Request): Say something in order to obtain an answer or some information.
- C**
- cancel** (Abort, Interrupt, Shut, Suspend): Decide or announce that (a planned event) will not take place.
- communicate** (Announce, Broadcast, Speak, Say, Tell, Enumerate): Share or exchange information, news, or ideas.
- customize** (Customise, Parametris, Parametrise, Personalise): Modify (something) to suit a particular individual or task.
- D**
- deny** (Refuse, Decline, Avoid, Ban, Disable, Disallow, Thule, Forbid, Inhibit, Prevent, Reject): State that one refuses to admit the truth or existence of.
- deteriorate** (Aggravate, Worsen):
- E**
- esa**: The European Space Agency (ESA; French: Agence spatiale européenne) is an intergovernmental organisation dedicated to the exploration of space.

Ontologies as key support to knowledge extraction

EXTRACT KNOWLEDGE FROM LEGACY DOCUMENTS...IN 10 MINUTES!

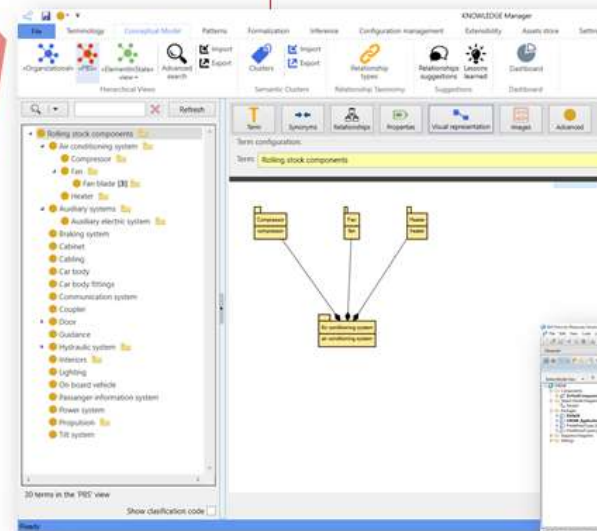
02

SCM/Architectures

Recreate and capture the system architectures represented in views and models. Stablish relationships among system and system elements.

Linking terms together and classifying terms

Useful to define lists of tags for properties



Ontologies as key support to knowledge extraction

EXTRACT KNOWLEDGE FROM LEGACY DOCUMENTS...IN 10 MINUTES!

03

Patterns

Represent requirements similarities and enable formal representation, automatic recognition and aid authors

When / After / If ...

[Condition]

<Subject>

Shall

<Action>

<Object>

[Constraint]

Name:
[METRIC - System Component Definition (Completeness & Consistency)]

Description:
N/A

Pattern group(s):
• METRIC - System Component Definition Requirements (Completeness & Consistency) (150)

Example:
N/A

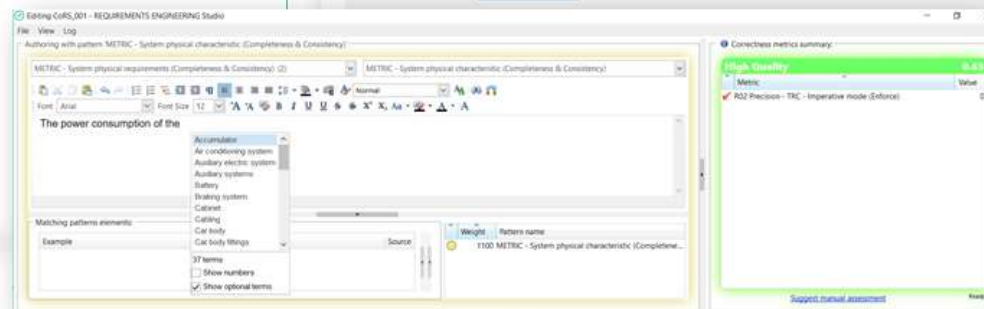
Indexable: Yes **Enabled:** Yes **Weight:** 1,200

Syntax:

DEFINITE ARTICLE * <SYSTEM ELEMENT> * <MODAL COMPULSORY> * VERB TO HAVE * NUMBER * <SYSTEM ELEMENT>

or

<SYSTEM>



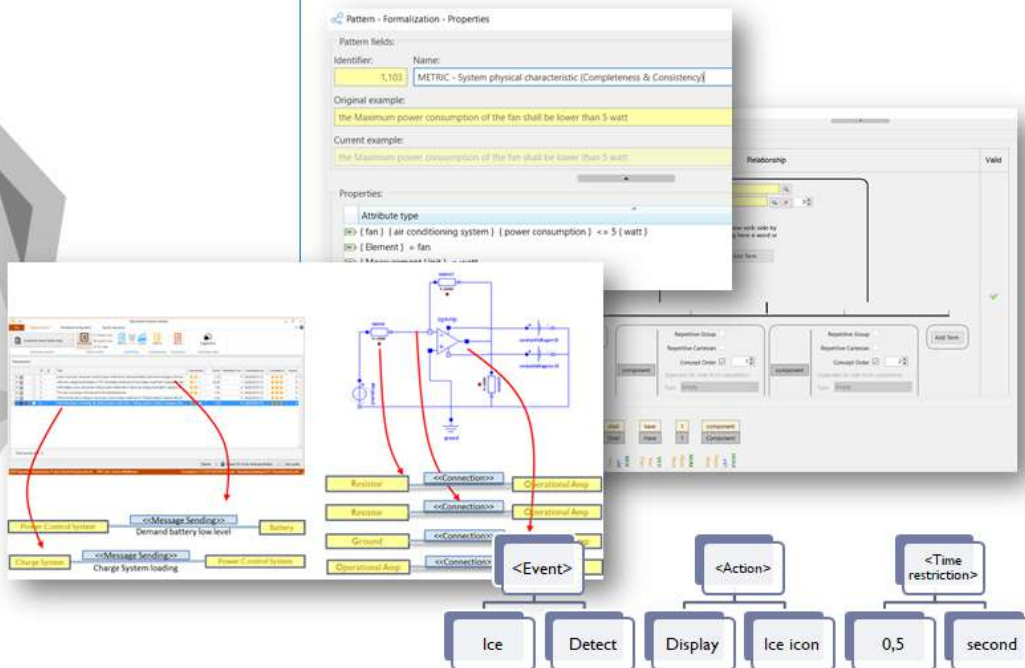
Ontologies as key support to knowledge extraction

EXTRACT KNOWLEDGE FROM LEGACY DOCUMENTS...IN 10 MINUTES!

04

Formalization

Representation of assets semantic through SRL – System Representation Language



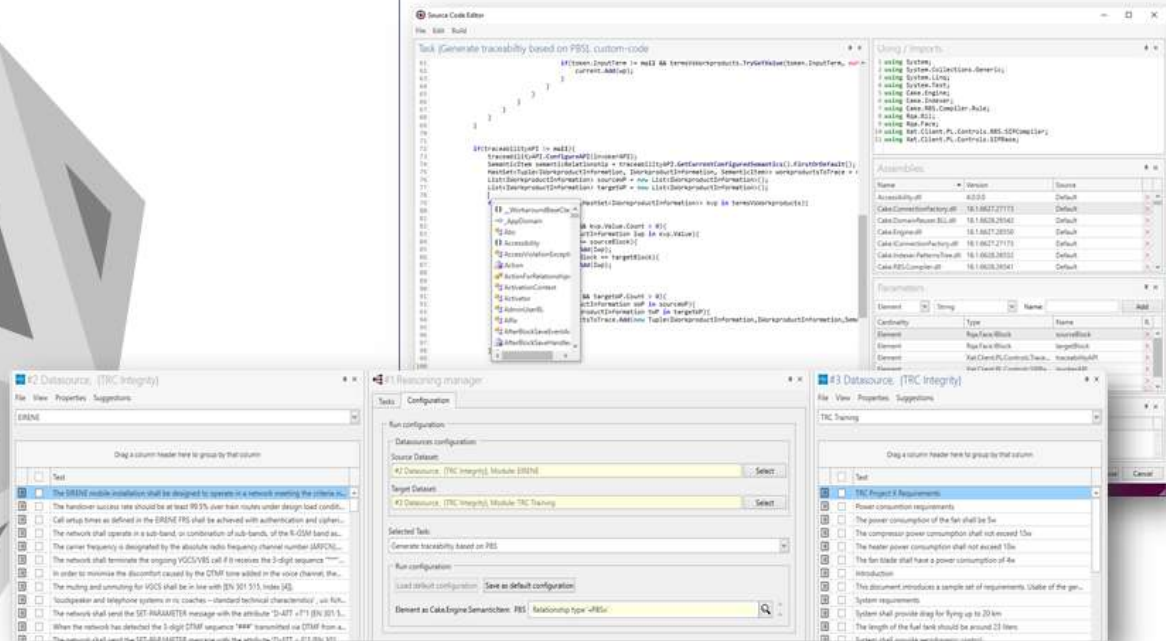
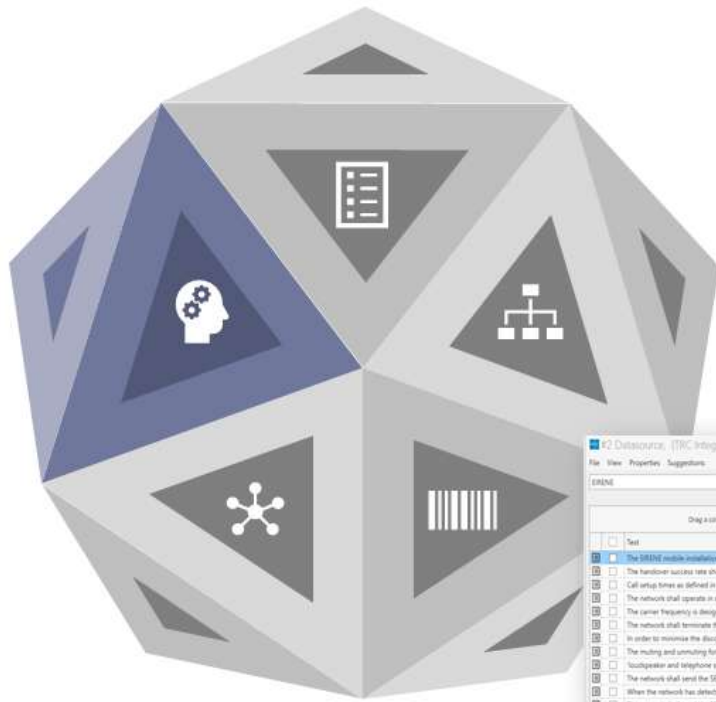
Ontologies as key support to knowledge extraction

EXTRACT KNOWLEDGE FROM LEGACY DOCUMENTS...IN 10 MINUTES!

05

Reasoning

A combination of rules, tasks and groups to infer information from valuable assets



CONTENT

Why capturing knowledge in legacy documents?

Ontologies as key support to knowledge extraction

Approach by TRC: based on NLP and semantic analysis

Demo

Q&A

Knowledge Discovery Process by TRC

EXTRACT KNOWLEDGE FROM LEGACY DOCUMENTS...IN 10 MINUTES!

1. Problem definition
2. Knowledge Source Understanding
3. Knowledge Processing and Model Development (ONTOLOGY)
4. Knowledge Selection
5. Rating the results

Knowledge mining is iterative!

Knowledge Discovery Process by TRC

EXTRACT KNOWLEDGE FROM LEGACY DOCUMENTS...IN 10 MINUTES!

1. Problem definition

Correctness metrics

Metric Identifier	Custom Metric	Name	Rationale	Weight	Enabled	Correctness type
24.361	N/A	Bigram detection/Bigram	To identify names for...	1	✓	Parameterized - Pattern matching
24.362	N/A	Bigram detection/Single term	To identify names for...	1	✓	Parameterized - Term tag
24.363	N/A	Detect system structure relationships		1	✓	Parameterized - Relationships not SCM compliant
24.364	N/A	Detection of derived units	Units like km/h	1	✓	Parameterized - Pattern matching
24.365	N/A	Extract missing actions	Using the pattern: Sh...	1	✓	Parameterized - Term tag
24.366	N/A	Extract standard names		1	✓	Parameterized - Pattern matching
24.367	N/A	Extract states and modes		1	✓	Parameterized - Pattern group matching
24.368	N/A	Extract system/component names		1	✓	Parameterized - Term tag
24.373	N/A	Trigram detection/Single term	To identify names for...	1	✓	Parameterized - Term tag
24.374	N/A	Trigram detection/Trigram	To identify names for...	1	✓	Parameterized - Pattern matching

No. of metrics: 10, Enabled: 14

☒ Starts with([Name], 'extract') Or Contains([Name], 'detect')

Quality function for selected metric

Range	Mandatory	Quality Level	Summary
[0]	No	★ ★ ★	
(0, ∞]	No	★ ★ ★	Do not use any of specified term tags

No. of ranges: 2

System names
States/Transitions
Standards
Actors...

Bar chart showing quality levels (High, Medium, Low) for ranges [0] and ∞.

Knowledge Discovery Process by TRC

EXTRACT KNOWLEDGE FROM LEGACY DOCUMENTS...IN 10 MINUTES!

2. Knowledge Source Understanding

Document purpose
Level of system abstraction
Authors...

ID	Title	Description
SysR1	PSysR-01	The Power System shall have a...
SysR2	PSysR-01	The Power System shall have a...
SysR3	PSysR-02	The Power System shall have a...
SysR4	PSysR-03	The Power System shall have a...
SysR5	PSysR-04	The Power System shall have a...
SysR6	PSysR-05	The Power System shall have a...
SysR7	PSysR-06	The Power System shall have a...
SysR8	PSysR-02	The Power System shall have a...
SysR9	PSysR-07	The Power System shall have a...
SysR10		While the Temperature Warrior is in Combat Mode, the Temperature Warrior shall...
SysR11		While the Temperature Warrior is in Combat Mode, the Temperature Warrior shall...
SysR12		While the Temperature Warrior is in Combat Mode, the Temperature Warrior shall...
SysR13		While the Temperature Warrior is in Combat Mode, the Temperature Warrior shall...
SysR14		While the Temperature Warrior is in Combat Mode, the Temperature Warrior shall...
SysR15		While the Temperature Warrior is in Combat Mode, the Temperature Warrior shall...

Knowledge Discovery Process by TRC

EXTRACT KNOWLEDGE FROM LEGACY DOCUMENTS...IN 10 MINUTES!

3. Knowledge Processing and Model Development (Ontology)

Knowledge Manager

File Terminology Conceptual Model Patterns Formalization Inference Configuration management Extensibility Assets store Settings

Term suggestions Import terms Special sentences Integrity Generate terms and frequencies Tags Languages Multi-language Configuration Tokenization Rules Affixes Substitutes Test Spell checker Rules Bigrams rules Tags probabilities Disambiguation

Search fields:

Terms:

Term tag:

Cluster:

Relationship type:

Language:

Identifier:

Relationship filters: ☐ Equals to:
☐ Greater than:
☐ Lower than:

Flags: ☐ Flag 1 ☐ Flag 2

Advanced filters:

Search in a new window Search

Terms:

Identifier	Term	Term Tag	Cluster	Relationship type	Belongs to domain	Belongs to SCM	Scope Note
66.830	[OPENING ROUND BRACKETS	< No «Cluster» >	< No «Relationship type» >	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
66.829	*	SYMBOL	< No «Cluster» >	< No «Relationship type» >	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
66.828	^	SYMBOL	< No «Cluster» >	< No «Relationship type» >	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
66.827	~	SYMBOL	< No «Cluster» >	< No «Relationship type» >	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
66.826	+	SYMBOL	< No «Cluster» >	< No «Relationship type» >	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
66.825	&	SYMBOL	< No «Cluster» >	< No «Relationship type» >	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
66.824	#	SYMBOL	< No «Cluster» >	< No «Relationship type» >	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
66.823	MTTF	NOUN	«QUALITY PROPERTY»	< No «Relationship type» >	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
66.822	MTBF	NOUN	«QUALITY PROPERTY»	< No «Relationship type» >	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
66.821	MTTR	NOUN	«QUALITY PROPERTY»	< No «Relationship type» >	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
66.820	Cubic feet	MEASUREMENT UNIT	«MEASUREMENT UNIT REQ»	< No «Relationship type» >	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
66.819	Gpm	MEASUREMENT UNIT	< No «Cluster» >	< No «Relationship type» >	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
66.818	Gal/day	MEASUREMENT UNIT	< No «Cluster» >	< No «Relationship type» >	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
66.817	Gal/min	MEASUREMENT UNIT	< No «Cluster» >	< No «Relationship type» >	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
66.816	GMP	MEASUREMENT UNIT	< No «Cluster» >	< No «Relationship type» >	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
66.815	WH/g	MEASUREMENT UNIT	< No «Cluster» >	< No «Relationship type» >	<input checked="" type="checkbox"/>	<input type="checkbox"/>	

Knowledge Discovery Process by TRC

EXTRACT KNOWLEDGE FROM LEGACY DOCUMENTS...IN 10 MINUTES!

3. Knowledge Processing and Model Development (Ontology)

The screenshot displays the TRC Knowledge Discovery Process interface, which is used for extracting knowledge from legacy documents and developing an ontology. The interface is divided into several sections:

- Searching fields:** This section at the top allows users to filter results by Cluster (set to 'COM*'), Identifier (set to '0'), and KM Code (set to '0'). There is also a checkbox for 'Clusters with terms'.
- Clusters:** On the left, a list of clusters is shown, including «COMBINATORS», «Communication», «COMPLIANCE : ACTION», and «COMPONENT». The «COMPONENT» cluster is currently selected.
- Terms:** In the center, a list of terms associated with the selected cluster is displayed. The terms include: Connection plate, Control laptop, Control subsystem, Control system, Laptop, Lcd, Lcd-s301c31tr, Lumex, Lumex lcd, Management system, Micro usb cable, Motherboard netduino, Netduino, Power source, Power system, Protection equipment, Regulator, Temperature actuation system, Temperature registration system, Temperature regulation software, and Temperature warrior. The 'Control laptop' term is highlighted.
- Right Panel:** This panel shows a detailed view of the selected term, 'Control laptop'. It lists various units and measurements associated with this term, such as Absorbed dose of ionizing radiation, Altitude, Amount of substance, Angle, Area, Attoradian, Centiradian, Crad, Deciradian, Drad, Femtoradian, Frad, Microradian, Milliradian, Mrad, Nanoradian, Nrad, Picoradian, Prad, Rad, Radian, Vectoradian, Vrad, Zeptoradian, Zrad, Capacity, µl, µm³, Ål, Åm³, Attoliter, Attolitre, Barrel, Bbl, Board-foot, Bu, and Bushel. The 'Angle' unit is currently selected.
- Top Menu:** The top of the interface features a menu bar with options like File, Terminology, Conceptual Model, Patterns, Formalization, Inference, Configuration management, Extensibility, Assets store, and Settings. Below the menu bar, there are icons for various functions, including Hierarchical Views, Advanced search, Import, Export, Clusters, and Semantic Clusters.

Knowledge Discovery Process by TRC

EXTRACT KNOWLEDGE FROM LEGACY DOCUMENTS...IN 10 MINUTES!

4. Knowledge selection

	A	B
1	TERM	TERM_TAG
2	authentication mode	NOUN
3	bidirectional services	NOUN
4	configuration mode	NOUN
5	configuration parameters	NOUN
6	control laptop	NOUN
7	control system	NOUN
8	management system	NOUN
	<ul style="list-style-type: none"> control laptop validation mode ready mode physical environment temperature actuation configuration parameters temperature registration management system temperature of the sensor one configuration algorithm of the control temperature regulation temperature threshold configuration of the temperature screen the total screen the temperature competition of the temperature 	
	Total: 51	

Import terms from Excel

Configuration:

File:

Sheet: Reading start row: 1

Term:

Term tag source:

☐ Excel Column ☒ Fixed

Term tag:

Cluster source:

☐ Excel Column ☒ Fixed

Cluster:

Properties source:

Attribute source:

☐ Excel Column ☒ Fixed

Attribute:

Value source:

Initial value:

Final value:

Value Qualifier:

☐ Excel Column ☒ Fixed

Qualifier:

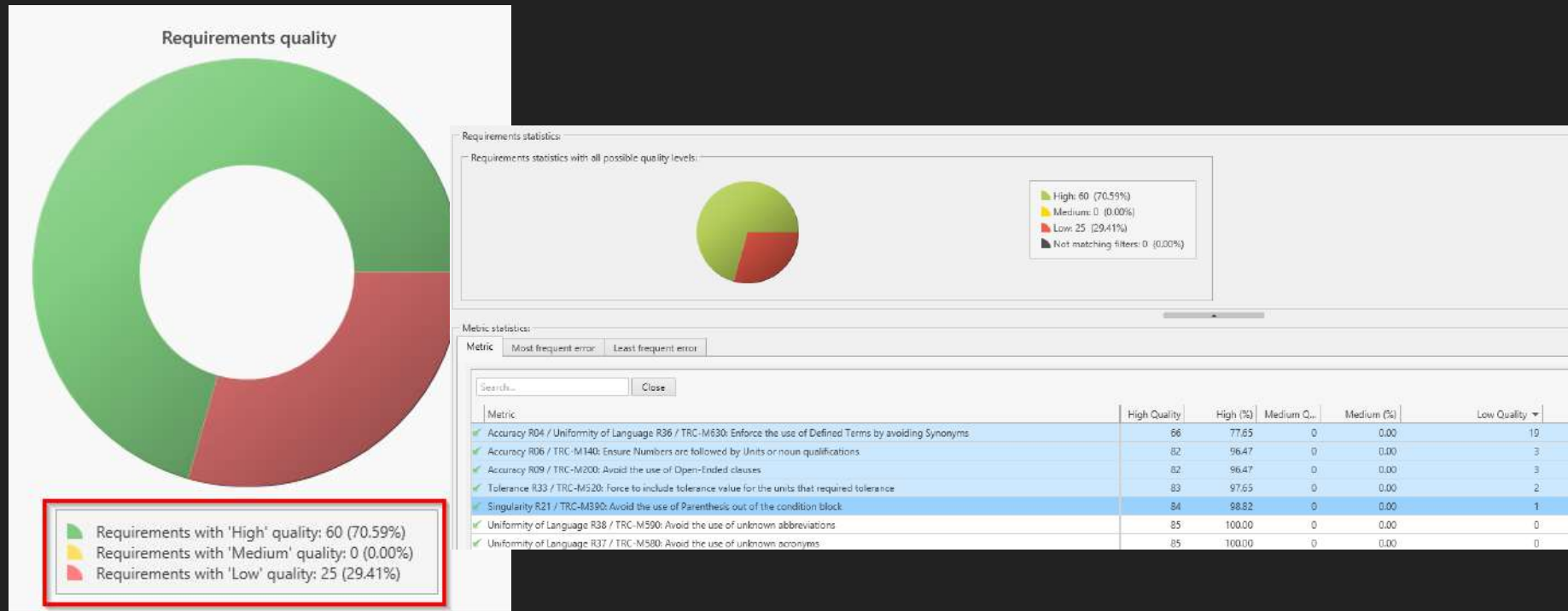
☒ Prioritize content from Excel

OK Cancel

Knowledge Discovery Process by TRC

EXTRACT KNOWLEDGE FROM LEGACY DOCUMENTS...IN 10 MINUTES!

5. Rate the results



Not satisfactory ? = New iteration!

Live Demo



THE
REUSE
COMPANY

Live demo

EXTRACT KNOWLEDGE FROM LEGACY DOCUMENTS...IN 10 MINUTES!

[illegible]

Q&A



THE
REUSE
COMPANY

COMING UP NEXT...

15' EXPRESS WEBINAR

KNOWLEDGE EXTRACTION FROM MODELS

TO ENHANCE DOMAIN-SPECIFIC

QUALITY ASSESSMENT

DECEMBER 14TH, 2021 - 05:00pm CET

DECEMBER 16TH, 2021 - 09:00am CET

SAVE THE DATE!



THE
REUSE
COMPANY

COMING UP NEXT...

ANY IDEA?

**SUGGEST YOUR TOPIC
FOR A NEXT
EXPRESS WEBINAR**



Thanks for your attention!



ilyes.yousfi@reusecompany.com



www.reusecompany.com



THE
REUSE
COMPANY