
# Using formal thesauri and controlled vocabulary as the interface between the unstructured data and axiomatic ontologies

**Farhad Ameri**

NIST TLP COI Event

April 2021, Gaithersburg, MD

# High-level Research questions

**How Knowledge Organization Systems (KOS) can be used in an integrated fashion to support formal knowledge extraction from unstructured data?**

- Supply Chain Use Case :
  - Data type: Manufacturing Capability data
  - Data source: company website
- Maintenance Use Case :
  - Data type: MWO records
  - Data Source: CMMS

# Maintenance Use Case Objectives

- To expose and formalize the semantics of unstructured maintenance data.
- To visualize MWO records as RDF knowledge graphs
- To use RDF knowledge graphs for maintenance diagnostics and root-cause analysis.

# Knowledge Organization Systems (KOS)

Controlled vocabulary for Information Retrieval

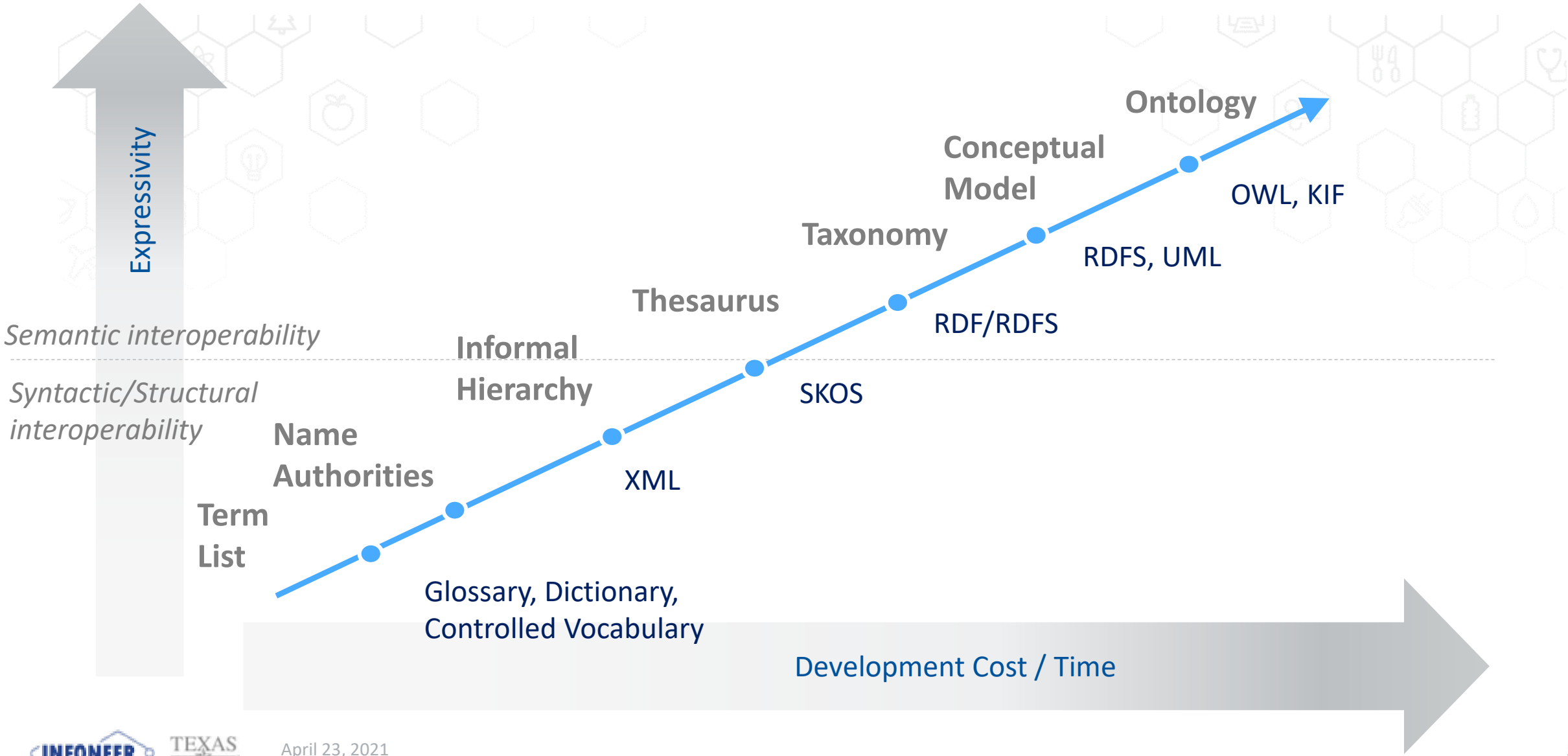
**KOS Types**

- term lists
- synonym rings
- name authorities
- taxonomies
- **thesauri**
- glossaries
- dictionaries
- gazetteers
- categorization schemes
- classification systems
- subject heading schemes
- semantic networks
- **ontologies**

- Knowledge organization system is any system of terms or **scheme** that is created to organize, manage, and retrieve information.
- They vary in complexity, structure, function, and expressivity.
- Services:
  - Abstraction & Indexing
  - Tagging
  - Information retrieval and term disambiguation
  - Navigation and translation
  - Reasoning and inference

Source: Introduction to taxonomies and other knowledge organization systems, Heather Hedden and Helmut Nagy, Semantic Web Company

# Semantic Spectrum of Knowledge Organization Systems



# Thesaurus

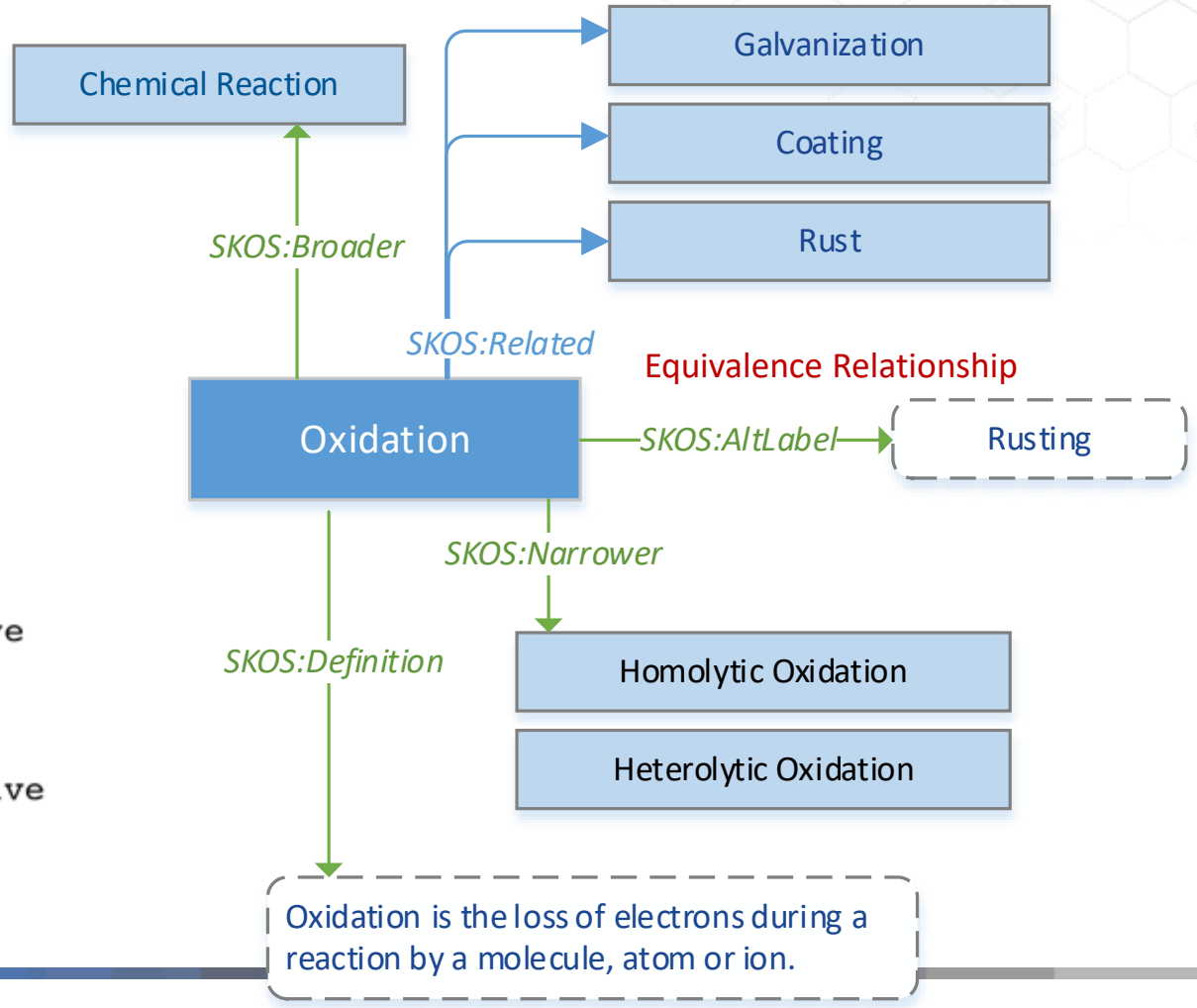
- A controlled vocabulary that represents three type of relationships between terms or concepts:
  - Hierarchical: broader term/narrower terms
  - Associative: related terms
  - Equivalence: preferred label /alternative label
- Created in accordance with standards:
- ISO 25964 (2011, 2013) and ANSI/NISO Z39.19

# SKOS model (Simple Knowledge Organization System)

SKOS is a **standard** data model for representation of knowledge organization systems.

- W3C Standard
- Web-native syntax
- Encoded using XML/RDF
- Each concept has a unique URI
- Machine-readable
- Simple semantics

```
skos:semanticRelation
 |
 +— skos:related
 |
 +— skos:broaderTransitive
 |     |
 |     +— skos:broader
 |
 +— skos:narrowerTransitive
       |
       +— skos:narrower
```



# Why Thesaurus?

| Thesaurus |
| --- |
| They can accommodate large or constantly growing vocabulary |
| They can be used to directly tag unstructured data |
| They are easier to develop and extend compared to ontologies |

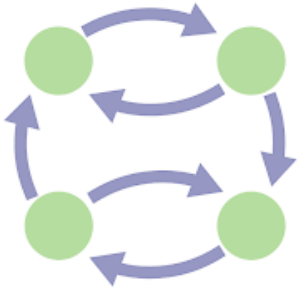
## We want to use the best of both worlds

### Ontology

**Logic-based semantics**

**High level expressivity and formality**

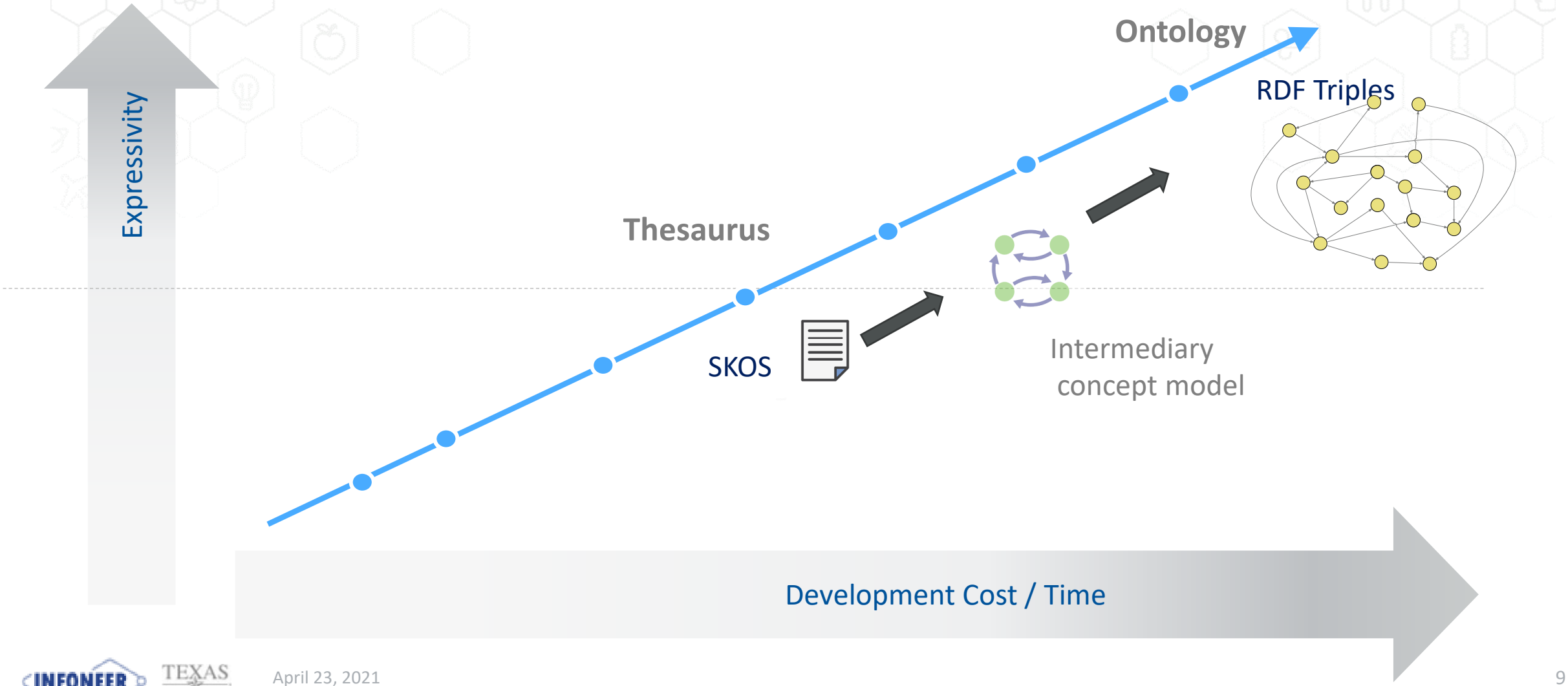
### Thesaurus



**Lexical semantics**

**Ease of extension**

# Overall Approach: Incremental structuring and semantic enrichment process



# Example

**Raw text (portion of reality) from a MWO Record**

Solenoid valve on upper lift cyl leaking. Hyd pressure reduced on Boom. Technician called. valve replaced, works normal

**Explicit Entities**

- Artifact
- Failure
- Agent
- Process

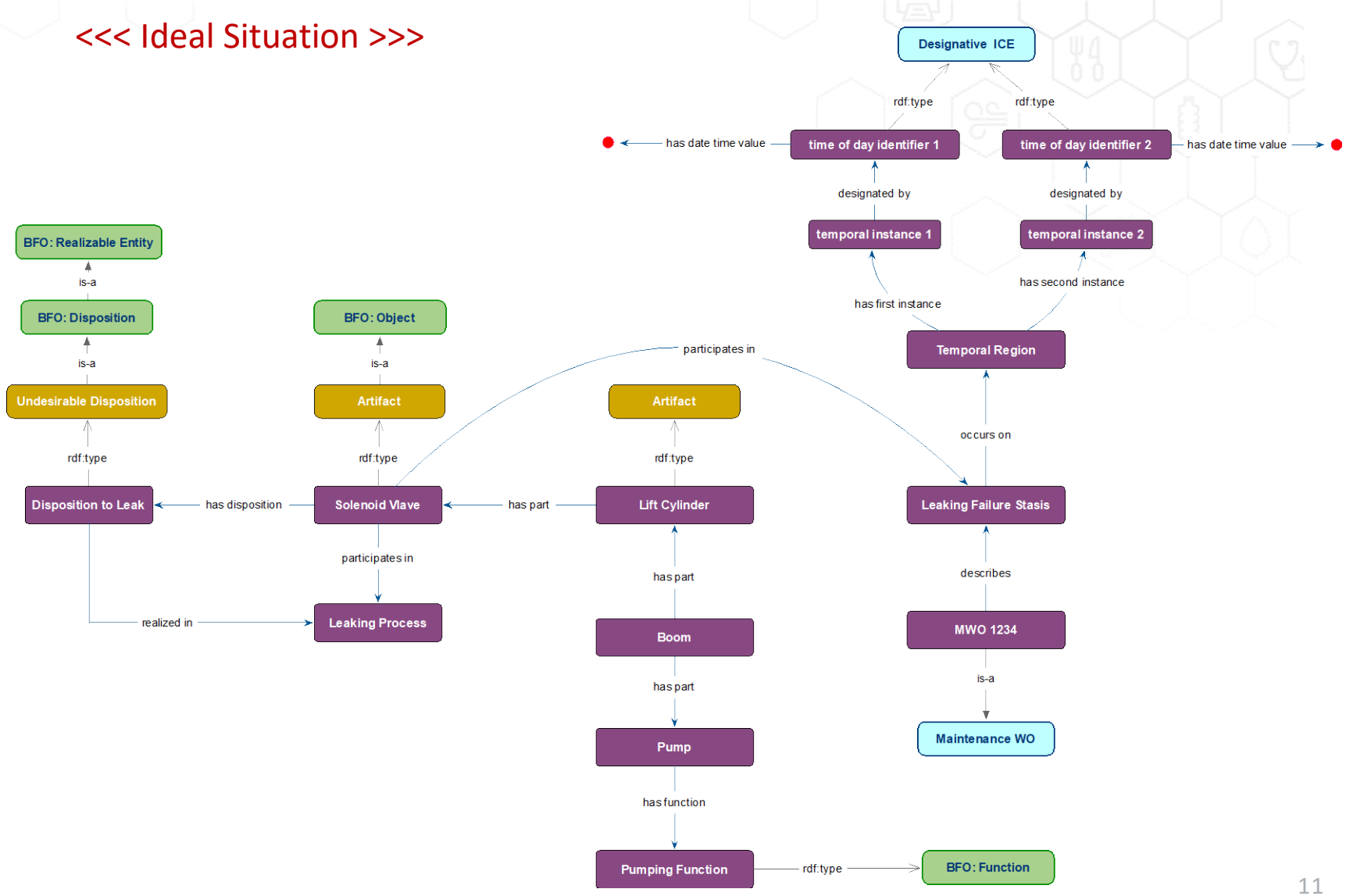
**Implicit Entities**

- States
- Events

# Ontological Representation of problem-solution description

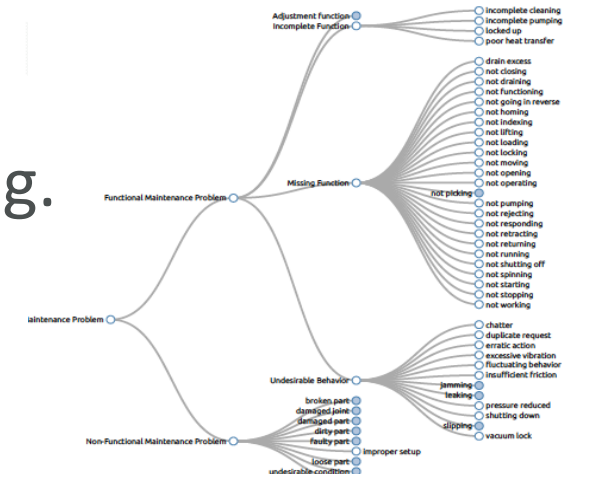
<<< Ideal Situation >>>

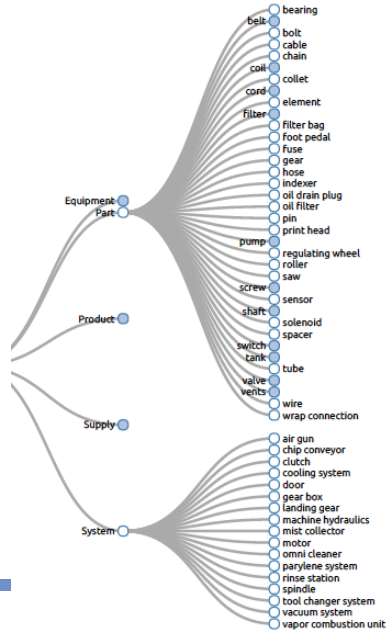
Solenoid valve on upper lift cyl leaking. Hyd pressure reduced on Boom. Technician called. valve replaced, works normal

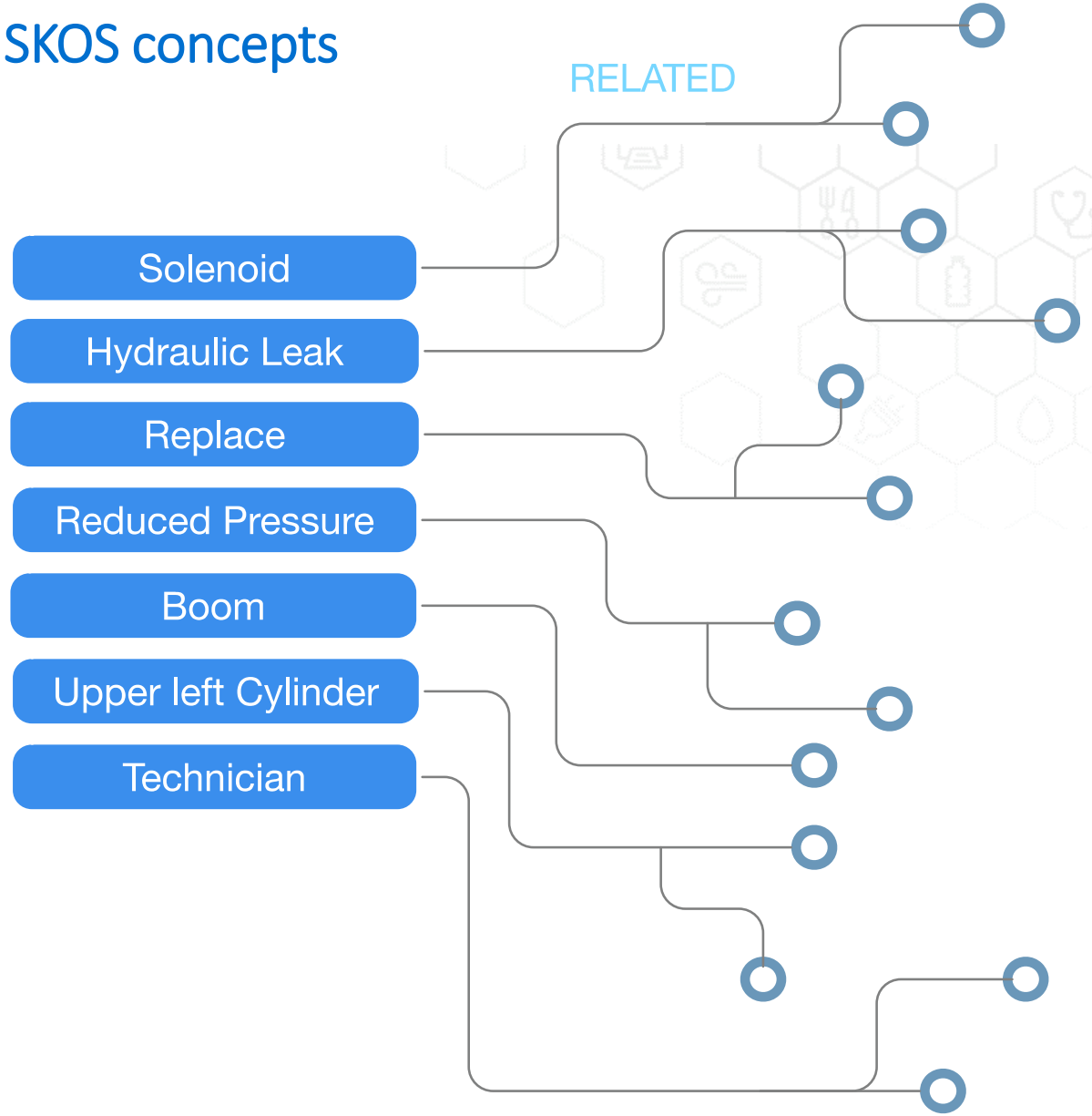


# 2-step process: First annotate the text with SKOS concepts

Solenoid valve on upper lift cyl leaking. Hyd pressure reduced on Boom. Technician called. valve replaced works normal

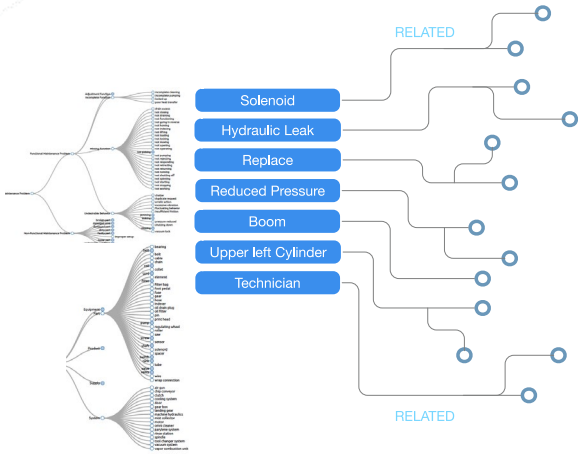


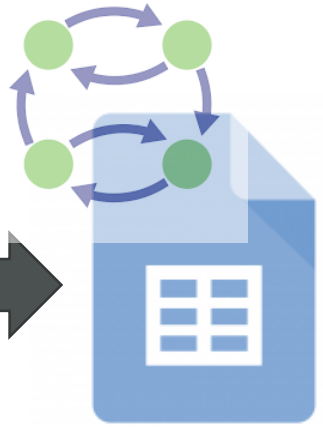



# Create an intermediary representation (structured)

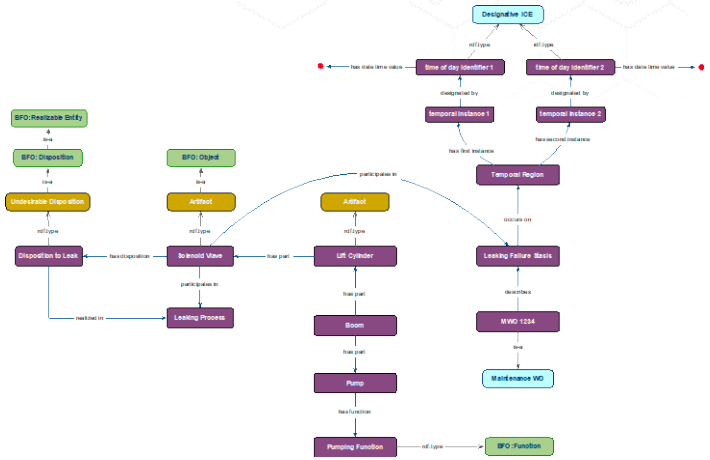
Solenoid valve on upper lift cyl leaking. Hyd pressure reduced on Boom. Technician called. valve replaced works normal




**Intermediary representation (Concept Model)**



- *Graph*
- *Table*

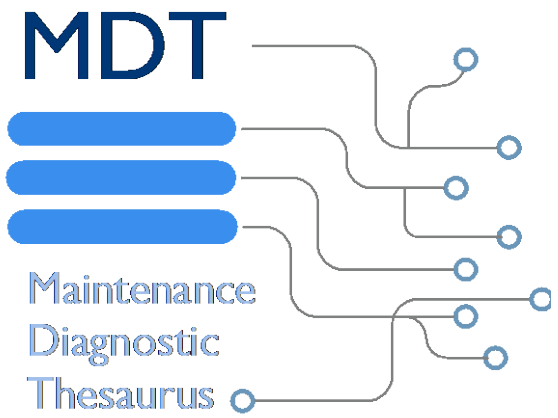


**Semantic Web Zone**

SKOS, RFD/XML, RDFS. OWL

# Maintenance Diagnostics Thesaurus

For this purpose, we developed Maintenance Diagnostics Thesaurus or  MDT

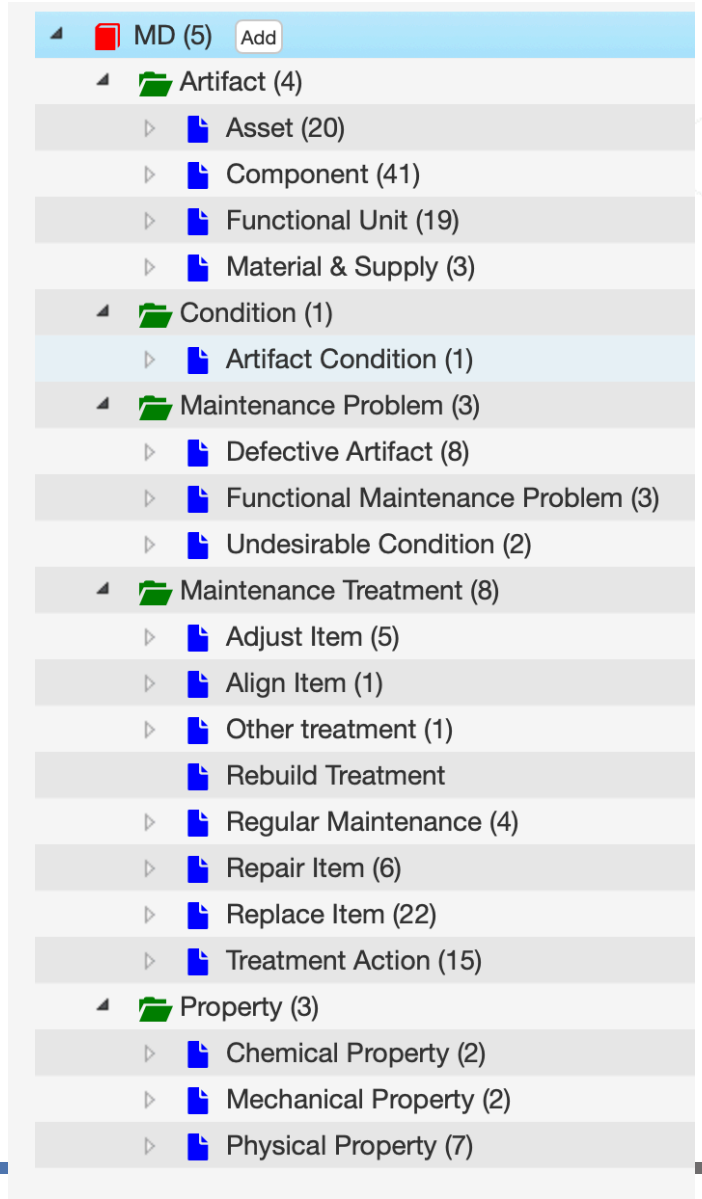


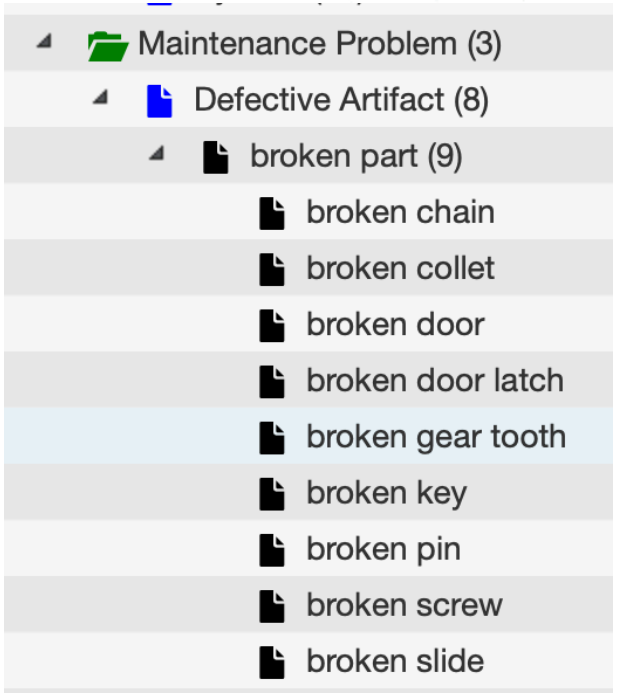

A SKOS model for representation of controlled vocabulary in the maintenance domain.

# The scope of Maintenance Diagnostic Thesaurus (MDT)

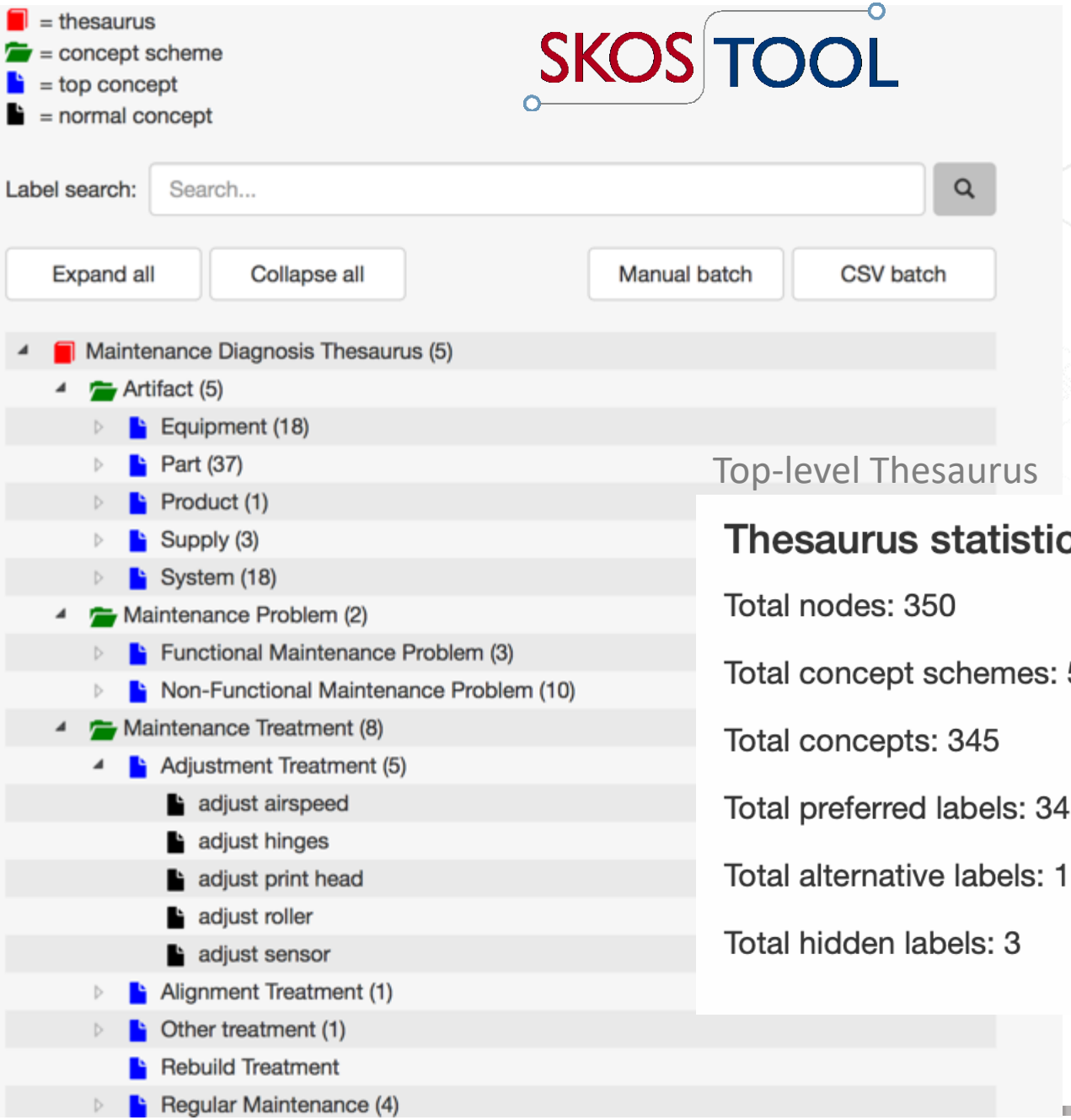
**Scope of MDT:**

- Failures
- Treatment
- Quality (color, temperature, viscosity, lumen)
- Feature (crack, hole, breakage)
- States
  - Working, not working
- Process
  - Malfunctioning (e.g., leaking, rotating slowly)
    - Including not functioning or reduced functioning
- Items
  - Component
  - Equipment/Machine
  - Sub-system



# MD Thesaurus


Top-level Thesaurus

## Thesaurus statistics

Total nodes: 350

Total concept schemes: 5

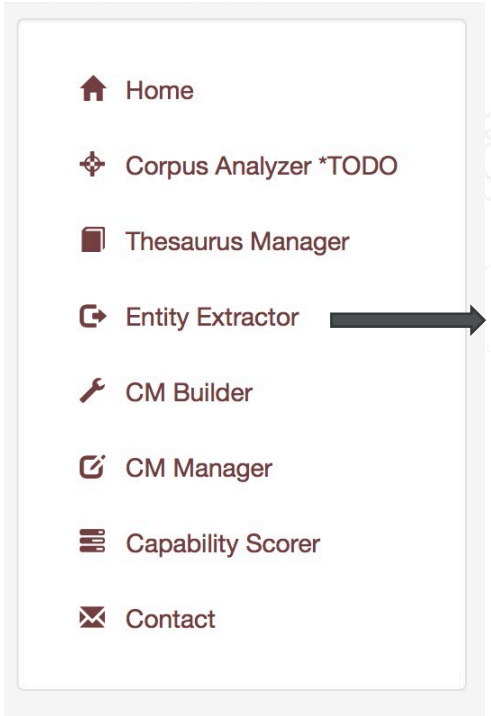
Total concepts: 345

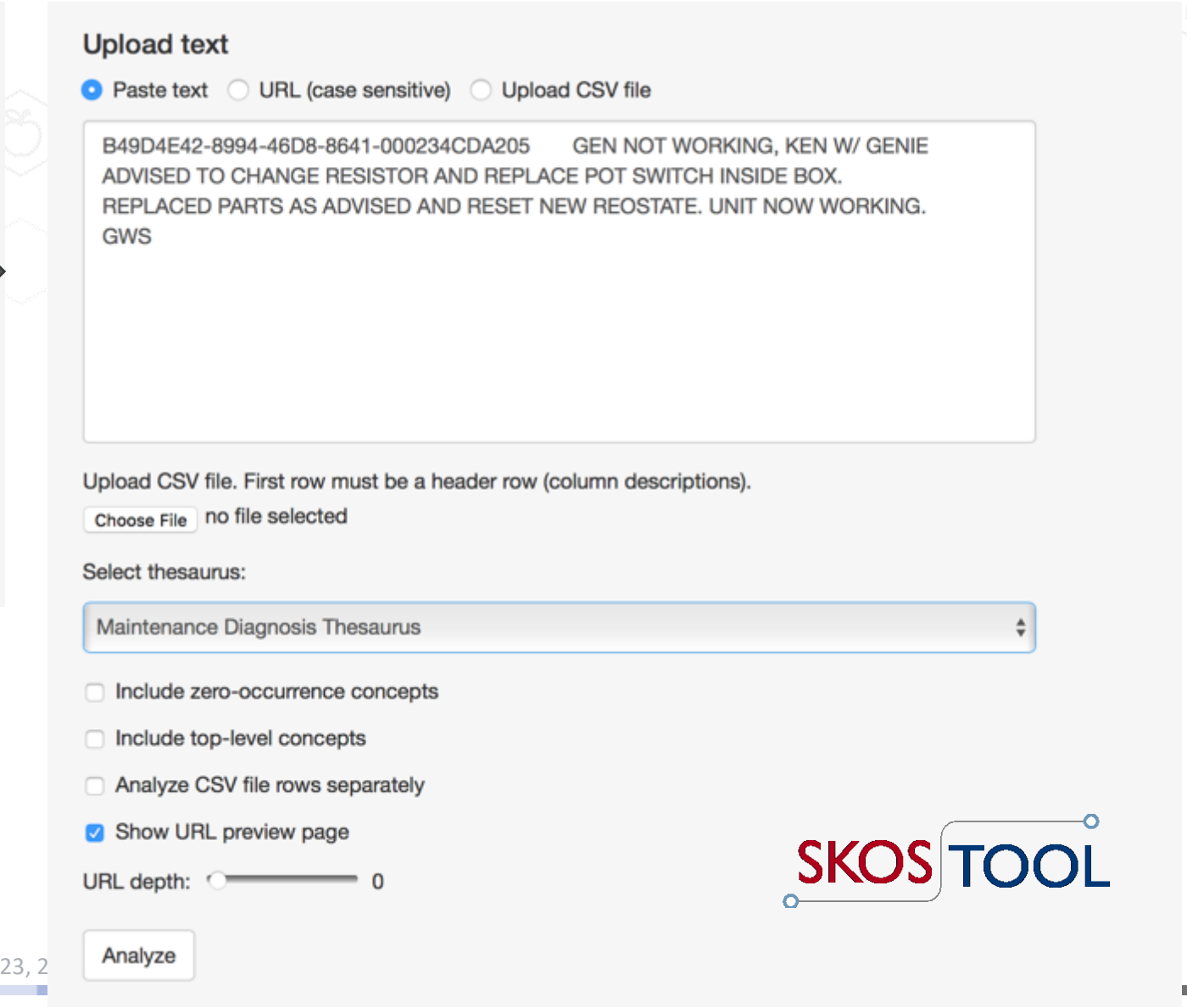
Total preferred labels: 345

Total alternative labels: 101

Total hidden labels: 3

# Entity Extractor



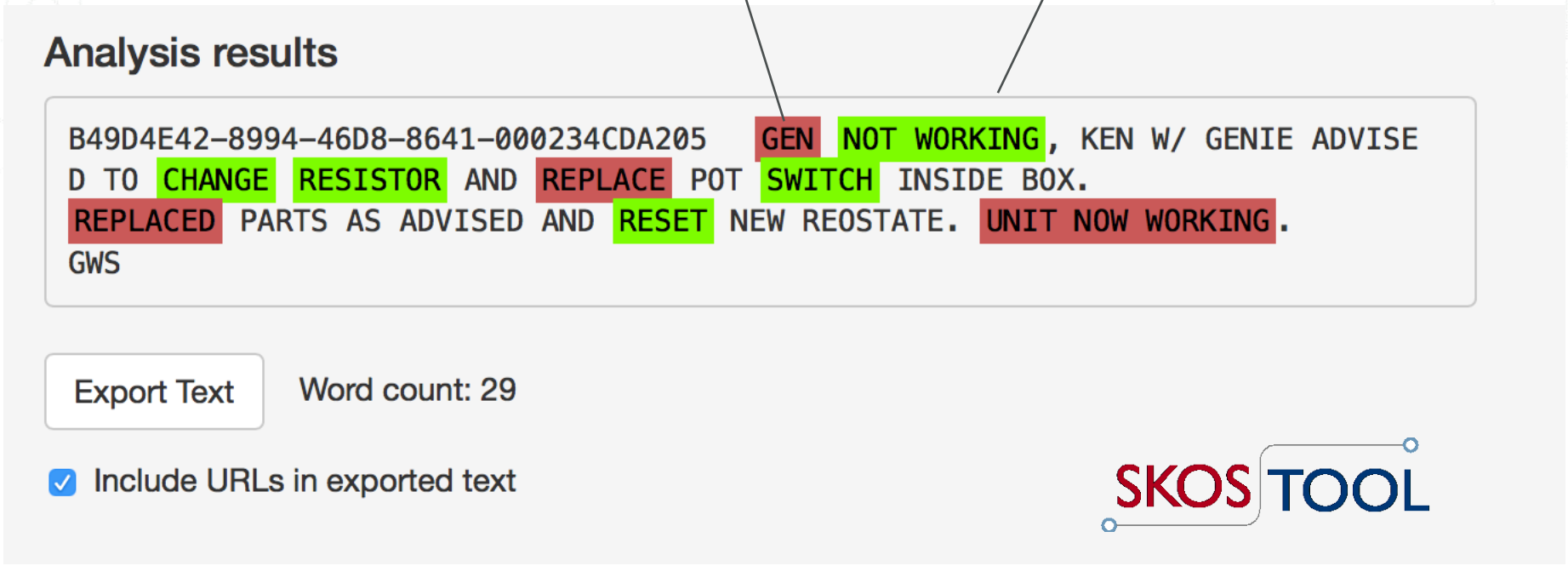




SKOS TOOL

# Detected Concepts

Detected as skos:prefLabel

Detected as skos:altLabel



# Exported Concept Model for the MWO record

Sort table:

Alphabetical Occurrences

SKOS TOOL

| Concept (preferred label) | Occurrences |
|---|---|
| change | 2 |
| generator | 1 |
| not working | 1 |
| replace | 2 |
| reset | 1 |
| resistor | 1 |
| switch | 1 |
| unit working | 1 |

Export type:

2-column 4-column

Export Table

Exported Table

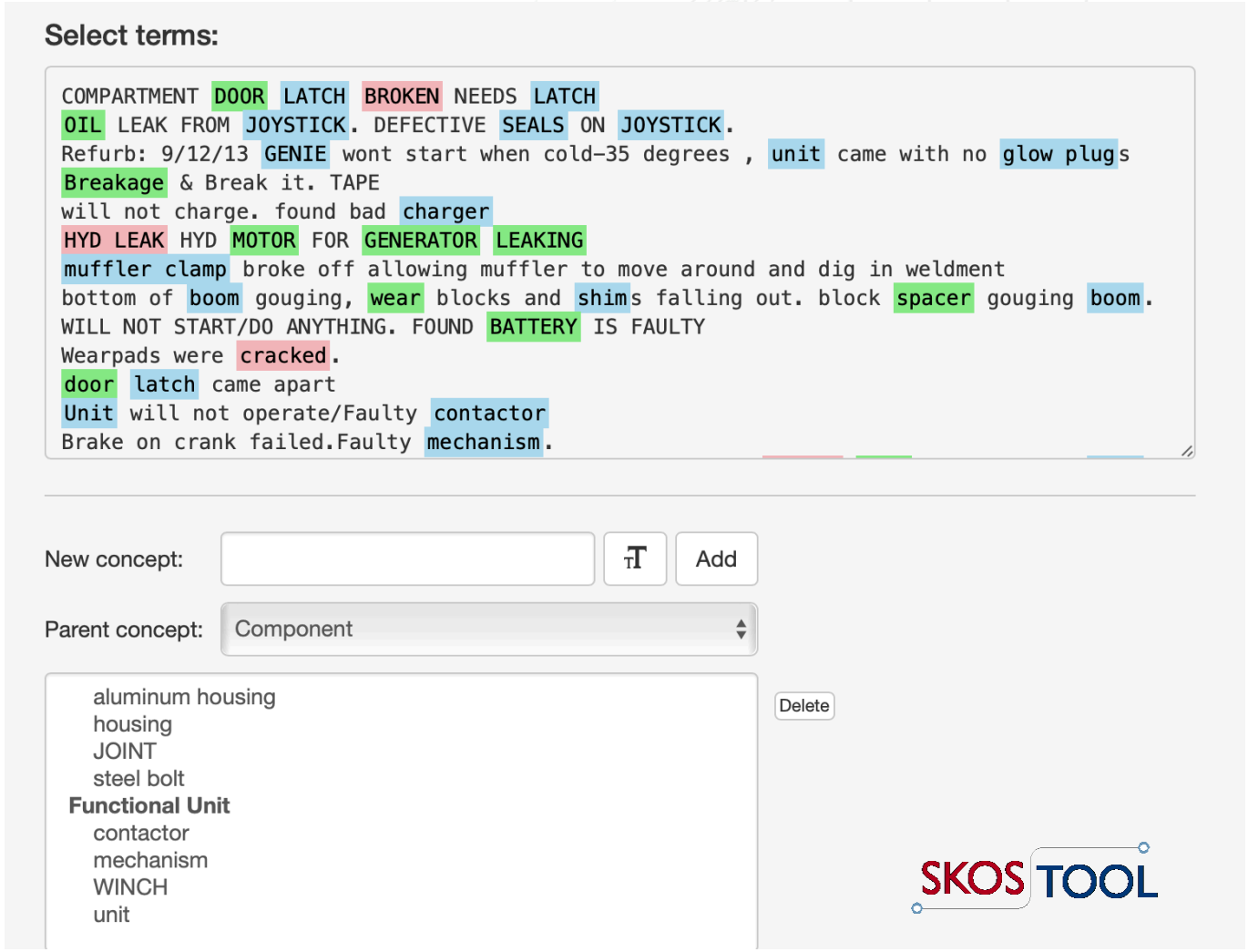
| Concept Schema | Top Concept | Concept (preferred label) | Occurrences |
|---|---|---|---|
| Artifact | System | generator | 1 |
| Maintenance Problem | Functional Maintena | not working | 1 |
| Maintenance Treatment | Treatment Action | change | 2 |
| Artifact | Part | resistor | 1 |
| Maintenance Treatment | Treatment Action | replace | 2 |
| Artifact | Part | switch | 1 |
| Maintenance Treatment | Treatment Action | reset | 1 |
| Problem Resolved | Resolution Term | unit working | 1 |

Formats available:

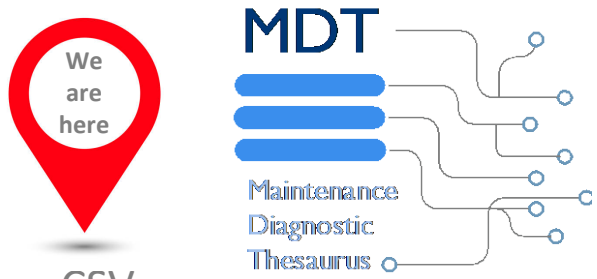
- CSV
- RDF/JSON

# Thesaurus Development challenges

- To adequately annotated all WO records even in a single company, we need a large thesaurus.
- Thesaurus extension is a bottleneck
- Mainly manual process
  - One concept at a time
  - Batch import



# Semi-automated Extension of MDT




CSV

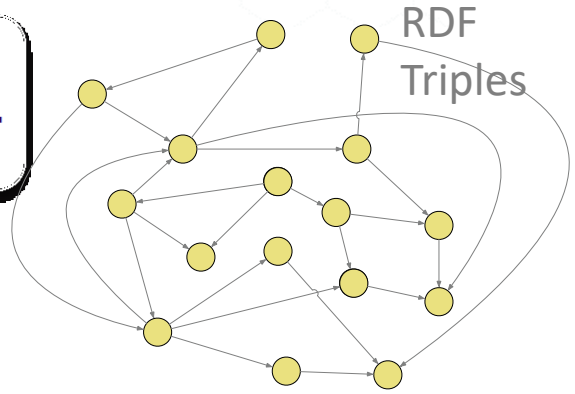
CSV or rj



**Nestor**

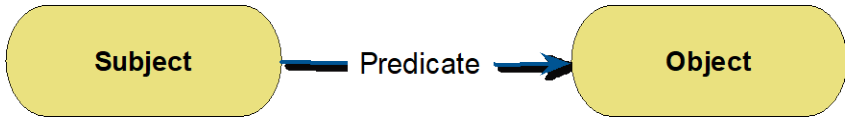
**SKOS Tool**

**SKOS to OWL**




- Provides NLP Services
- Shallow taxonomy of terms

- More complex Taxonomy
- Lexical and associative relationships

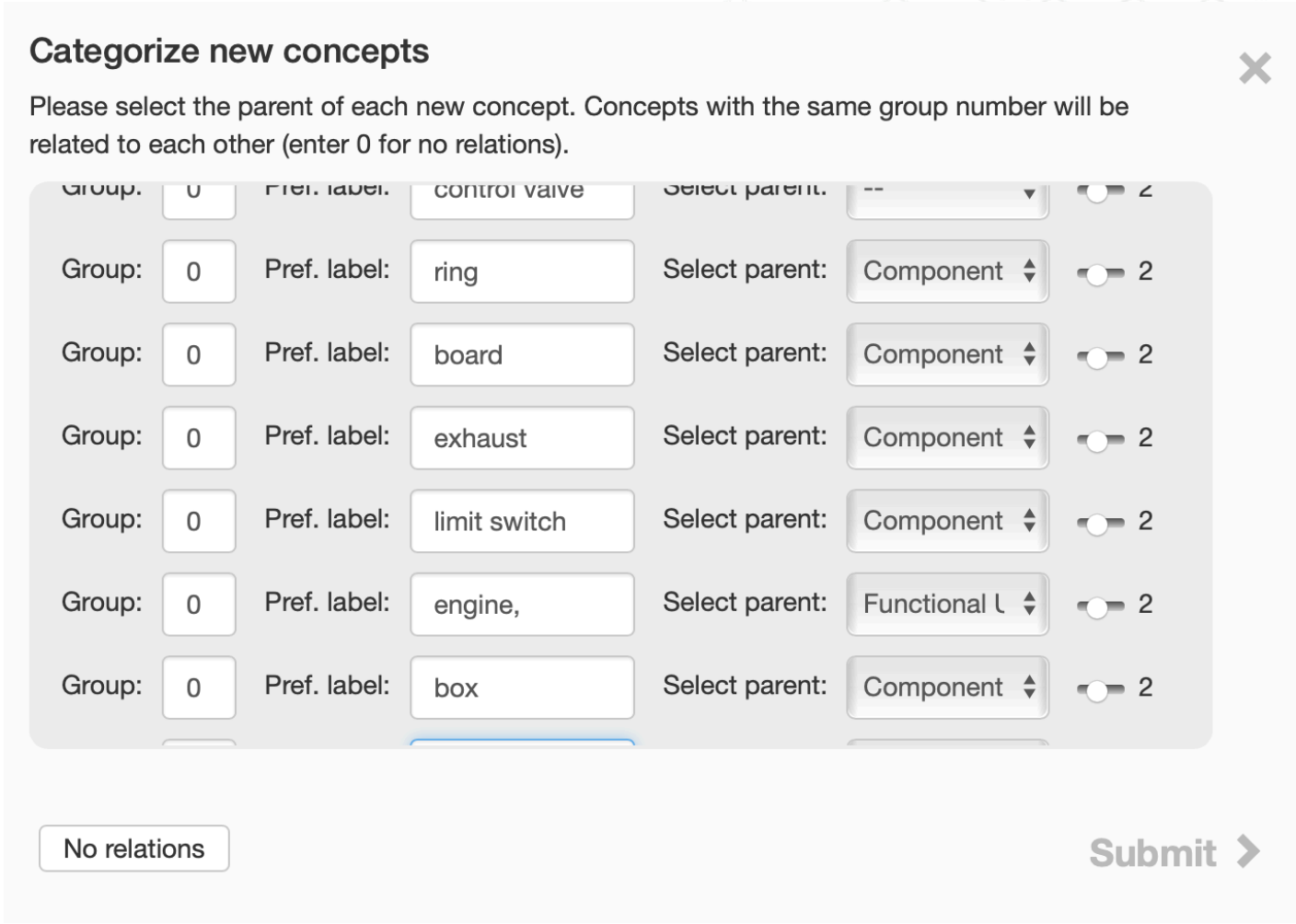


# Categorizing Items (from Nestor) in MDT

**Nestor Output (Item column)**

| | I | N |
|---|---|---|
| 0 | battery, unit | |
| 1 | charger, unit | |
| 2 | boss, boss_fitting, motor, oil, ring | |
| 3 | box, genie, pot, resistor, switch, unit | |
| 4 | engine, fuse | |
| 5 | limit_switch, switch, unit | |
| 6 | exhaust, flange, gasket, tube | |
| 7 | board | |
| 8 | oil, ring, unit | |
| 9 | control_valve, oil, ring, valve | |

Candidate Terms



# Next Steps

- Extending the thesaurus and ontology in tandem
  - Better integration with Nestor
  - More diverse dataset is needed to feed the thesaurus extension process
- Tool Development :
  - SKOS to OWL : Receives SKOS Concept Model and converts it to RDF graph
    - Reasoning to extract inferred knowledge (hidden / implicit)
  - Diagnosis Tool: Provides diagnostic support by querying the RDF graph