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)

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

**KOS Types**
- Controlled vocabulary for Information Retrieval
  - term lists
  - synonym rings
  - name authorities
  - taxonomies
  - thesauri
  - glossaries
  - dictionaries
  - gazetteers
  - categorization schemes
  - classification systems
  - subject heading schemes
  - semantic networks
  - ontologies

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

April 23, 2021
Semantic Spectrum of Knowledge Organization Systems

- **Term List**
  - Glossary, Dictionary, Controlled Vocabulary

- **Informal Hierarchy**
  - Thesaurus

- **Hierarchical**
  - Taxonomy

- **Conceptual Model**
  - Ontology
    - OWL, KIF
    - RDFS, UML
    - RDF, RDFS

- **Semantic Spectrum**
  - RDF/RDFS
  - SKOS
  - XML

- **Development Cost / Time**

- **Semantic Interoperability**
  - Name Authorities

- **Syntactic/Structural Interoperability**
  - Expressivity
Thesaurus

- A controlled vocabulary that represents three types 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

Oxidation is the loss of electrons during a reaction by a molecule, atom or ion.

```
skos:semanticRelation
| ← skos:related
| ← skos:broaderTransitive
|   | ← skos:broader
| ← skos:narrowerTransitive
|   | ← skos:narrower
```

**Chemical Reaction**
- Galvanization
- Coating
- Rust

**SKOS:Broader**

**SKOS:Related**

**Equivalence Relationship**

**SKOS:AltLabel**

**SKOS:Narrower**

**SKOS:Definition**

**Oxidation**

**Homolytic Oxidation**

**Heterolytic Oxidation**

**Rusting**
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

Expressivity

Development Cost / Time

Thesaurus

SKOS

Ontology

RDF Triples

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

Create an intermediary representation (structured)

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

- Condition
  - Artifact Condition (1)

- Maintenance Problem (3)
  - Defective Artifact (8)
  - Functional Maintenance Problem (3)
  - Undesirable Condition (2)

- Maintenance Treatment (8)
  - Adjust Item (5)
  - Align Item (1)
  - Other treatment (1)
  - Rebuild Treatment
  - Regular Maintenance (4)
  - Repair Item (6)
  - Replace Item (22)
  - Treatment Action (15)

- Property (3)
  - Chemical Property (2)
  - Mechanical Property (2)
  - Physical Property (7)
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
Upload text

- Paste text
- URL (case sensitive)
- Upload CSV file

B49D4E42-8994-46D8-8641-000234CDA205  GEN NOT WORKING, KEN W/ GENIE ADVISED TO CHANGE RESISTOR AND REPLACE POT SWITCH INSIDE BOX. REPLACED PARTS AS ADVISED AND RESET NEW REOSTATE. UNIT NOW WORKING.

GWS

Upload CSV file. First row must be a header row (column descriptions).

Choose File  no file selected

Select thesaurus:

- Maintenance Diagnosis Thesaurus

Include zero-occurrence concepts
Include top-level concepts
Analyze CSV file rows separately
Show URL preview page
URL depth: 1 0

Analyze
Detected Concepts

Analysis results

B49D4E42-8994-46D8-8641-000234CDA205

GEN NOT WORKING, KEN W/ GENIE ADVISE

D TO CHANGE RESISTOR AND REPLACE POT SWITCH INSIDE BOX.

REPLACED PARTS AS ADVISED AND RESET NEW REOSTATE. UNIT NOW WORKING.

GWS

Export Text  Word count: 29

Include URLs in exported text
# Exported Concept Model for the MWO record

## SKOS Tool

### Exported Table

<table>
<thead>
<tr>
<th>Concept Schema</th>
<th>Top Concept</th>
<th>Concept (preferred label)</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artifact</td>
<td>System</td>
<td>generator</td>
<td>1</td>
</tr>
<tr>
<td>Maintenance Problem</td>
<td>Functional Maintena</td>
<td>not working</td>
<td>1</td>
</tr>
<tr>
<td>Maintenance Treatment</td>
<td>Treatment Action</td>
<td>change</td>
<td>2</td>
</tr>
<tr>
<td>Artifact</td>
<td>Part</td>
<td>resistor</td>
<td>1</td>
</tr>
<tr>
<td>Maintenance Treatment</td>
<td>Treatment Action</td>
<td>replace</td>
<td>2</td>
</tr>
<tr>
<td>Artifact</td>
<td>Part</td>
<td>switch</td>
<td>1</td>
</tr>
<tr>
<td>Maintenance Treatment</td>
<td>Treatment Action</td>
<td>reset</td>
<td>1</td>
</tr>
<tr>
<td>Problem Resolved</td>
<td>Resolution Term</td>
<td>unit working</td>
<td>1</td>
</tr>
</tbody>
</table>

## 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

- Provides NLP Services
- Shallow taxonomy of terms

- More complex Taxonomy
- Lexical and associative relationships

Nestor → SKOS Tool → SKOS to OWL

We are here

MDT
Maintenance
Diagnostic
Thesaurus

CSV
CSV or rj

RDF Triples
Subject → Predicate → Object
Categorizing Items (from Nestor) in MDT

### Nestor Output (Item column)

<table>
<thead>
<tr>
<th>ID</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>battery, unit</td>
</tr>
<tr>
<td>1</td>
<td>charger, unit</td>
</tr>
<tr>
<td>2</td>
<td>boss, boss_fitting, motor, oil, ring</td>
</tr>
<tr>
<td>3</td>
<td>box, genie, pot, resistor, switch, unit</td>
</tr>
<tr>
<td>4</td>
<td>engine, fuse</td>
</tr>
<tr>
<td>5</td>
<td>limit_switch, switch, unit</td>
</tr>
<tr>
<td>6</td>
<td>exhaust, flange, gasket, tube</td>
</tr>
<tr>
<td>7</td>
<td>board</td>
</tr>
<tr>
<td>8</td>
<td>oil, ring, unit</td>
</tr>
<tr>
<td>9</td>
<td>control_valve, oil, ring, valve</td>
</tr>
</tbody>
</table>

### Candidate Terms

#### Categorize new concepts

Please select the parent of each new concept. Concepts with the same group number will be related to each other (enter 0 for no relations).

- **Group: 0**
  - **Pref. label:** control valve
  - **Select parent:** Component
  - **No relations**

- **Group: 0**
  - **Pref. label:** ring
  - **Select parent:** Component
  - **No relations**

- **Group: 0**
  - **Pref. label:** board
  - **Select parent:** Component
  - **No relations**

- **Group: 0**
  - **Pref. label:** exhaust
  - **Select parent:** Component
  - **No relations**

- **Group: 0**
  - **Pref. label:** limit switch
  - **Select parent:** Component
  - **No relations**

- **Group: 0**
  - **Pref. label:** engine,
  - **Select parent:** Functional
  - **No relations**

- **Group: 0**
  - **Pref. label:** box
  - **Select parent:** Component
  - **No relations**

**Submit**
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