Cleaning up a legacy thesaurus to make it fit for transformation into a Semantic Web KOS – a learning process

Dr. Anna Kasprzik,
Porto, 13th September 2018
NKOS workshop at TPDL

LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK

TIB

Leibniz
Association

Agenda

- **Point of departure**
  - **The person**
  - **The thesaurus**
  - **The project**

- **A sandbox ontology for electric mobility**

- **A top-level structure for the thesaurus**

- **Relations and facets for the thesaurus**

- **Lessons learned**

The thesaurus "Technik und Management" (TEMA)

- ~ 57.000 concepts, ~ 195.000 terms

- bilingual: German and English

- domains:

  - engineering basics
  - manufacturing, production
  - electrotechnology, electronics
  - measuring and control
  - management and organisation
  - environmental engineering
  - information technology

  - vehicle contruction
  - textile engineering
  - mining
  - materials
  - power engineering
  - medical technology
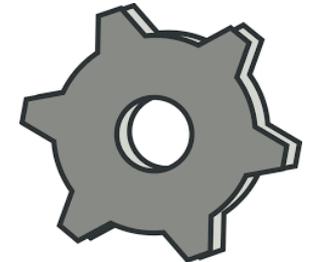
- created from six print thesauri

- currently maintained on the Averbis Terminology Platform (ATP)
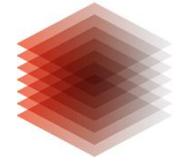
Project „Fachontologie Technik" (2013–2017)

- joint project of TIB and WTI

- <u>goal:</u> enrich the TEMA thesaurus is various ways
  to take it to the next level of digital development

  - migration to the ATP

  - introduction of English terms throughout

  - alignment with the German Integrated Authority File (GND)

  - increase its interoperability with other KOS
    by transforming it into a KOS that complies
    with Semantic Web standards,
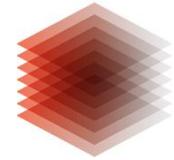   to be used for various use cases

Terminologies

☑ ⊙ Ontologie_EMOBI 1.0
☑ ⊙ TEMA 1.0
☑ ⊙ TEMA Fachordnung 1.0

… ?

A hybrid sandbox ontology for electric mobility

```
EMOBI.txt ☒
 1   066023
 2   066023    AD    1000-Grad-Celsius-Bereich
 3   131055    AE    1000 degree Celsius range
 4   707978    ND    3IFT (Messung und Prüfung t
 5   707978    NE    3IFT (Measuring/testing of
 6   627447    HD    Zuordnung: es gibt nur ein
 7   066030    BD    1000-C-Bereich
 8   066029    BD    1000-Celsius-Bereich
 9   066019    BD    1000-Grad-C-Bereich
10   066020    BE    1000 degree C range
11   066031    O     Grad-Celsius-Bereich
12   127162
13   127162    AD    1000-Kelvin-Bereich
14   127165    AE    1000 Kelvin range
15   707978    ND    3IFT (Messung und Prüfung t
16   707978    NE    3IFT (Measuring/testing of
17   627447    HD    Zuordnung: es gibt nur ein
18   127160    BD    1000-Grad-K-Bereich
19   127159    BD    1000-Grad-Kelvin-Bereich
20   127161    BD    1000-K-Bereich
21   131044    BE    1000 degree Kelvin range
22   127163    BE    1000 degree K range
23   127164    BE    1000 K range
24   127085    O     Kelvin-Bereich
```

Input: TEMA thesaurus
   as a.txt file

At the time, two goals:

- explore potential for an ontology by interpreting concepts as classes and *broader* relations as superclass relations

- enhance potential for thesaurus applications by using constructs from SKOS-XL and *iso-thes*

https://www.w3.org/2006/07/SWD/SKOS/skos-and-owl/master.html

Definitions (serialized in Turtle)

```
16   # Definitionen:
17
18   :TemaConcept rdf:type rdfs:Class.
19   :TemaTerm rdfs:subClassOf skosxl:Label.
20   :Tema rdf:type rdfs:Class;
21       owl:unionOf (:TemaConcept :TemaTerm).
22
23   skos:related rdf:type owl:ObjectProperty;
24       rdfs:domain :TemaConcept;
25       rdfs:range :TemaConcept.
26
27   skosxl:Label rdf:type rdfs:Class.
28
29   iso-thes:PreferredTerm rdfs:subClassOf skosxl:Label.
30   iso-thes:SimpleNonPreferredTerm rdfs:subClassOf skosxl:Label.
31
32   :TemaID rdf:type rdfs:Datatype;
33       rdfs:comment "TEMA-IDs als Strings, wegen eventueller fuehrender Nullen";
34       owl:equivalentClass [rdf:type rdfs:Datatype; owl:onDatatype xsd:string;
35                              owl:withRestrictions ([xsd:pattern "C_[0-9]{6}|[0-9]{6}"])].
36
37   :hasTemaID rdf:type owl:DatatypeProperty;
38       rdf:type owl:FunctionalProperty;
39       rdfs:domain :Tema;
40       rdfs:range :TemaID.
```

concepts as classes

terms as classes (SKOS-XL)

use of *iso-thes*

class :TemaID for concepts and terms with restrictions on form

The result (in Turtle and in Protégé)

```
:C_Fahrersitz :hasTemaID "C_049449".
:Fahrersitz rdf:type :TemaTerm;
    rdf:type iso-thes:PreferredTerm;
    skosxl:literalForm "Fahrersitz"@de;
    :hasTemaID "049449".
:C_Fahrersitz skosxl:prefLabel :Fahrersitz.

:driving_seats rdf:type :TemaTerm;
    rdf:type iso-thes:PreferredTerm;
    skosxl:literalForm "driving seats"@en;
    :hasTemaID "077589".
:C_Fahrersitz skosxl:prefLabel :driving_seats.

:C_Fahrersitz :hasTemaNot :3NRN.

:C_Fahrersitz rdfs:comment "Zuordnung: es gibt

:Fuehrersitz rdf:type :TemaTerm;
    rdf:type iso-thes:SimpleNonPreferredTerm;
    skosxl:literalForm "Führersitz"@de;
    :hasTemaID "049450".
:C_Fahrersitz skosxl:altLabel :Fuehrersitz.

:C_Fahrersitz rdfs:subClassOf :C_Fahrzeugsitz.

:C_Fahrersitz skos:related :C_Fahrzeugfuehrer.
```
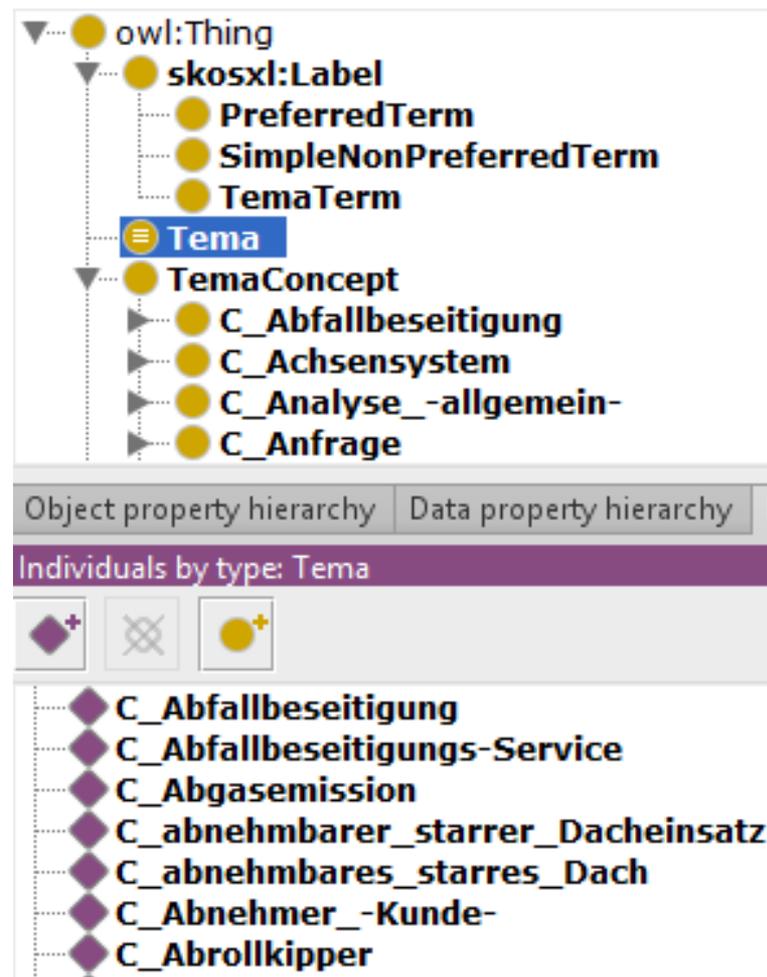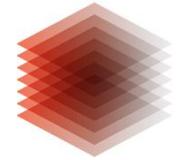
A top-level structure
for the TEMA thesaurus

TIB

```
TEMA 1.0
   Abschirmung
   Abtastung
   Abweichung
   Achse
   Adhäsion
   Aktivierung
   Analogschaltung
   Analysator
   Analyse (allgemein)
   Anpassung
   Anschliff
   Antenne
   Antriebstechnik
   Antwort
   Anwendung (allgemein)
   Anwendungsgebiet
   Apparat
   Arbeitsplatz
   Armatur
   Atmosphäre
   Atmung
   Atom
   Aufbereitung
   Aufhängung
   Aufzeichnung
   Materialabtragung
   {...}
```
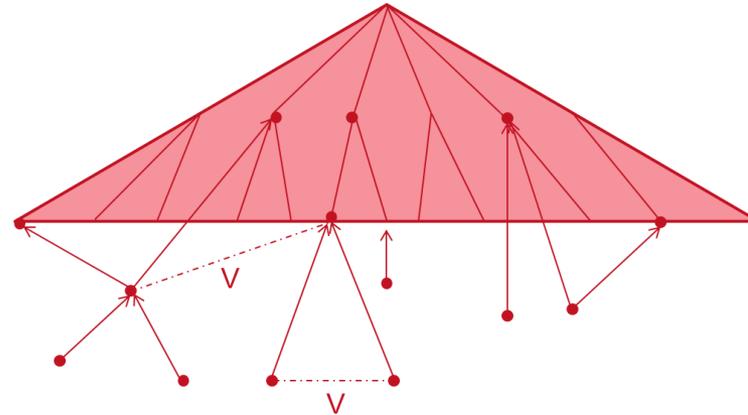

| Alles ausgeklappt | Ebene 1 ausgeklappt | Ebene 2 ausgeklappt | Ebene 3 ausgeklappt |

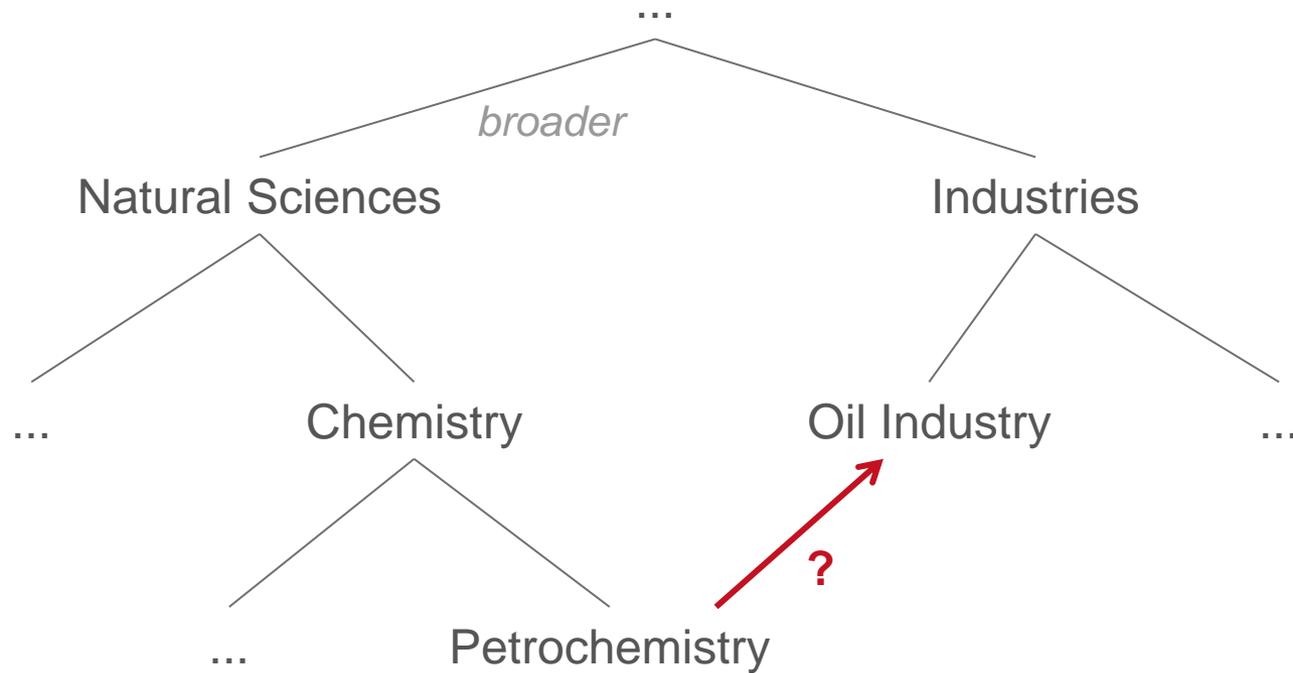Stures Mergen

Wirtschaftswissenschaften und verwandte Wissenschaften ** --> 'Organisatorische Aspekte des menschlichen Zusammenlebens' **

Betriebswirtschaftslehre

Allgemeine Betriebswirtschaftslehre

Anlagenwirtschaft

Beschaffung

Controlling

Finanzwirtschaft

Informationswirtschaft

Logistik ** existiert auch noch in 'Maschinenbau'! **

https://ontologie.tib.eu/toplevelkand/sturmerge.html

Relations and facets for the TEMA thesaurus

```
                              ...
                       broader
        Natural Sciences              Industries

     ...      Chemistry          Oil Industry      ...

          ...     Petrochemistry      ?
```

Relations and facets for the TEMA thesaurus

⊞ **Wissenschaft**

⊟ **Maschinenteil** „machine part"

⊞ Feder (Federungselement)

⊞ Lager (Maschinenelement)

⊞ mechanische Kupplung

⊟ Anker (elektrische Maschine)

⊞ Bremse

⊟ Polschuh

⊟ Ständer (elektrische Maschine)

⊞ Stift (Maschinenelement)

⊞ Spindel

⊟ Lagerschale

⊞ Nocken

⊞ Nabe

⊟ Siebwechsler

⊟ Angussentnahmegerät

⊞ Joch

⊟ Ausgleichselement

⊟ Kupplungsteil

⊟ Federungsbauteil

⊟ Griffschale

⊟ Übertragungselement

⊞ **Quantisierung**

⊟ **Quelle** „source"

⊞ Energiequelle

⊟ Fehlerquelle

⊟ Rohstoffquelle

⊟ Schallquelle

⊟ Störquelle

⊞ Strahlungsquelle

⊟ Emissionsquelle

⊟ Speicherquelle

⊟ Datenquelle

⊟ Informationsquelle

⊟ Zündquelle

⊞ Nährstoffquelle

⊟ Stickstoffquelle

⊟ Source (FET)

⊟ Quelle (Vektorfeld)

⊟ Rauschquelle

⊞ heiße Quelle

⊟ Kältequelle

⊞ Quelle (Wasser)

⊟ Ölquelle

⊟ Frank-Read-Quelle

⊟ Versetzungsquelle

⊞ **Antriebstechnik**

⊞ **Antwort**

⊟ **Anwendung (allgemein)** „application (general)"

⊞ Laseranwendung

⊞ Rechneranwendung

⊟ Multimedia-Anwendung

⊟ Infrastrukturanwendung

⊟ Büroanwendung

⊟ Kompositanwendung

⊟ Altsystemanwendung

⊟ Netzwerkanwendung

⊟ Sicherheitsanwendung

⊟ verteilte Anwendung

⊟ virtuelle Anwendung

⊟ Radaranwendung

⊟ Unterwasseranwendung

⊟ optische Anwendung

⊟ kommerzielle Anwendung

⊟ Kunststoffanwendung

⊞ **Anwendungsgebiet**

Facetting by SKOS collections

```
_:b0 rdf:type skos:Collection;
        skos:prefLabel "source (general)"@en;
        skos:member :EnergySource;
        skos:member :InformationSource;
        skos:member :WaterSource;

        ...


_:b0 rdf:type skos:Collection;
        skos:prefLabel "applications of organic chemistry"@en;
        skos:member :ProductionOfPlastic;
        skos:member :Production of Detergents;
        skos:member :PharmaIndustry;

        ...
```

More work necessary

Suggestion by one of the reviewers for future work:
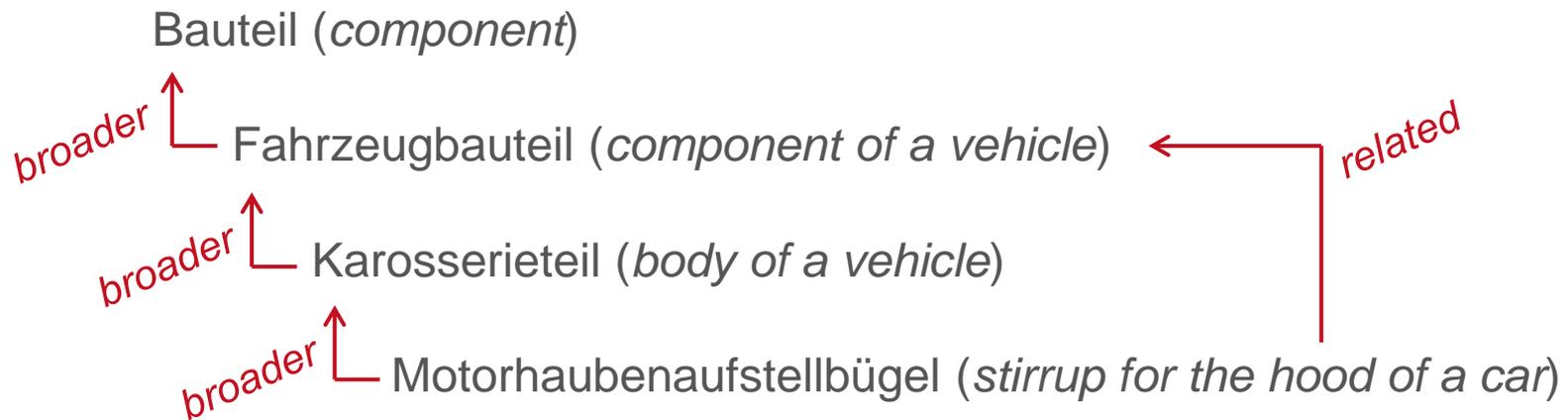
„complete the conversion of TEMA to RDF (if not done yet),

validate it with some SKOS quality tools (eg qSKOS),

and publish the result for inspection"

Empty Labels: OK (no potential problems found)

Omitted or Invalid Language Tags: OK (no potential problems found)

Incomplete Language Coverage: OK (no potential problems found)

> Undocumented Concepts: FAIL (2023)

No Common Languages: OK (no potential problems found)

> Missing Labels: FAIL (4038)

Overlapping Labels: OK (no potential problems found)

> Orphan Concepts: FAIL (2019)

> Disconnected Concept Clusters: FAIL (9)

Cyclic Hierarchical Relations: OK (no potential problems found)

> Valueless Associative Relations: FAIL (284)

Solely Transitively Related Concepts: OK (no potential problems found)

Omitted Top Concepts: OK (no potential problems found)

Top Concepts Having Broader Concepts: OK (no potential problems found)

Hierarchical Redundancy: OK (no potential problems found)

Mapping Relations Misuse: OK (no potential problems found)

Reflexively Related Concepts: OK (no potential problems found)

> Ambiguous Notation References: FAIL (121)

Unprintable Characters in Labels: OK (no potential problems found)

> Missing Out-Links: FAIL (4038)

Undefined SKOS Resources: OK (no potential problems found)

> Unidirectionally Related Concepts: FAIL (1922)

HTTP URI Scheme Violation: OK (no potential problems found)

> Relation Clashes: FAIL (6)

Mapping Clashes: OK (no potential problems found)

Inconsistent Preferred Labels: OK (no potential problems found)

Disjoint Labels Violation: OK (no potential problems found)

Relation clashes: FAIL (6)

Bauteil (*component*)

broader ↑
⌐ Fahrzeugbauteil (*component of a vehicle*) ← *related*

broader ↑
⌐ Karosserieteil (*body of a vehicle*)

broader ↑
⌐ Motorhaubenaufstellbügel (*stirrup for the hood of a car*)

047144		Motorhaubenaufstellbügel

Terms/Synonyms (2) | Relatives (5 / 3) | Attributes (1) | Xr

| Type	Label	Concept ID	
V	**Stab**	010028	⚲
V	**Motorhaube**	501424	⚲
V	**Fahrzeugbauteil**	013002	⚲
V	**Stütze**	013979	⚲
O	**Karosserieteil**	049469	⚲

Conclusions

- explore possibilities towards more structural quality
  for the thesaurus and towards smaller, sharper ontologies
  – in the future, keep that separate

- define concrete objectives and practical applications,
  then develop along the lines of those

- Comment by one of the reviewers about the work done here:
  „[The TEMA thesaurus] when completed will contribute
   a valuable thesaurus to the LOD space"

LEIBNIZ INFORMATION CENTRE
FOR SCIENCE AND TECHNOLOGY
UNIVERSITY LIBRARY
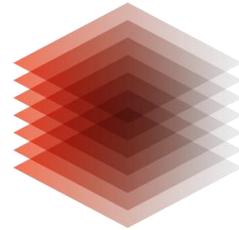
Thank you

Contact
Dr. Anna Kasprzik
T +49 511 762-14219, anna.kasprzik@gmx.de