technische universität dortmund

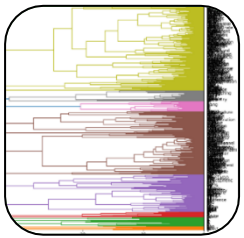
NFDI4Cat
NFDI for Catalysis-Related Sciences

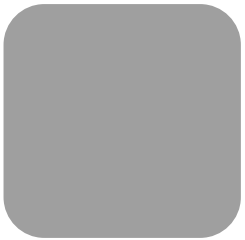
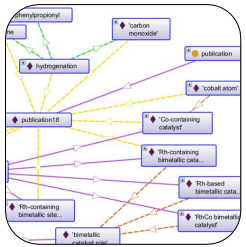
Natural Language Processing-Based Extension of the Allotrope® Foundation Ontology

Alexander S. Behr, Norbert Kockmann

TU Dortmund University

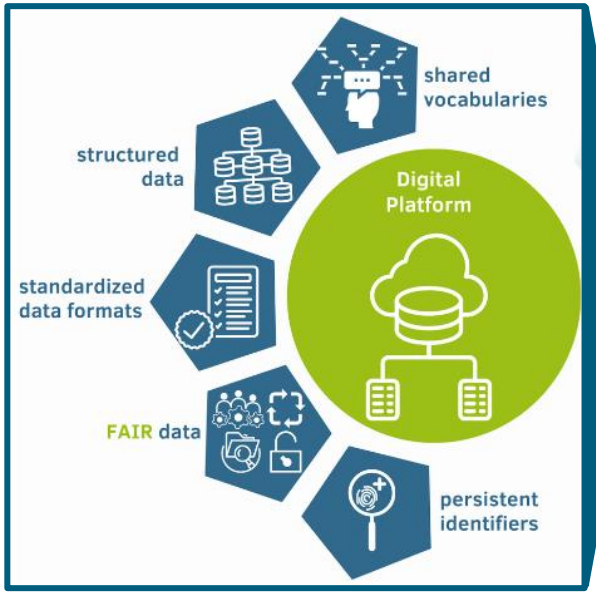
16.05.2024



NFDI4Cat for Open Science and Digitalisation

Digital repositories & collaborative platforms

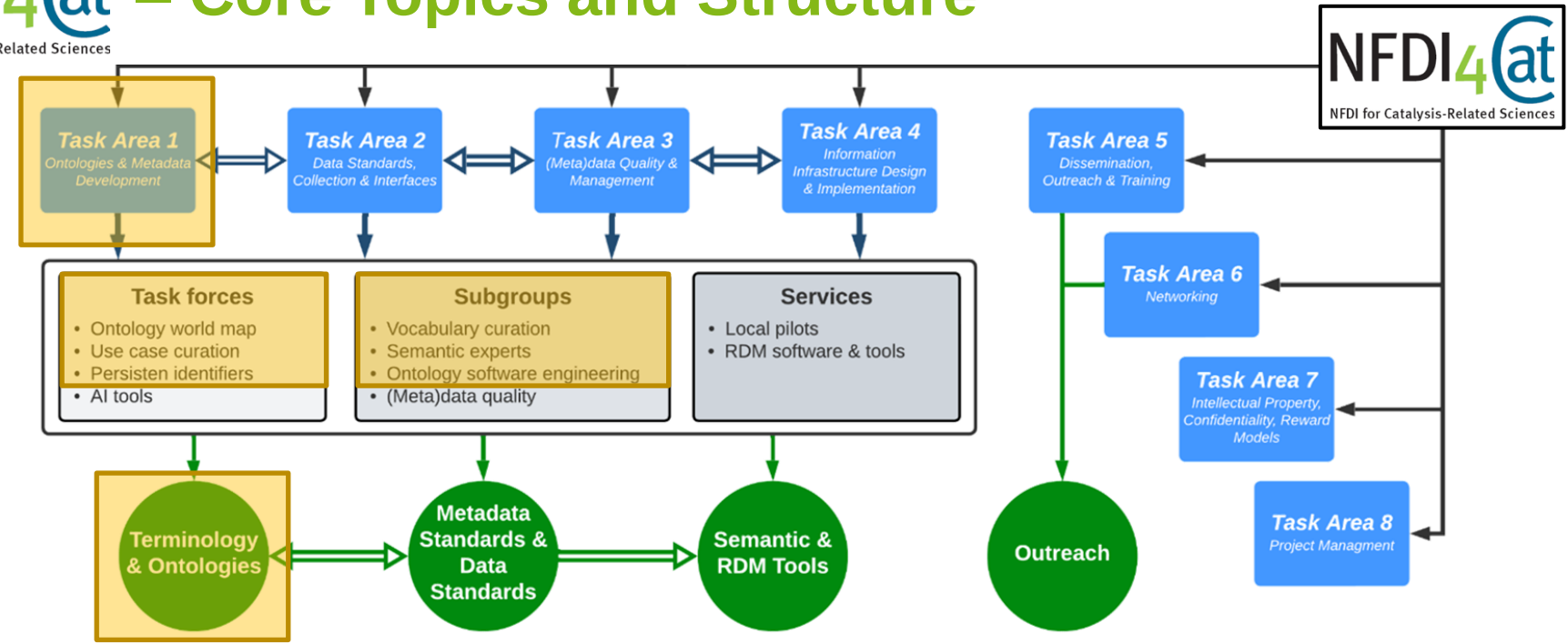


Facilitator of cooperation between academia and industry

BASF The Chemical Company
hte
CLARIANT
Linde
thyssenkrupp
covestro
EVONIK Leading Beyond Chemistry

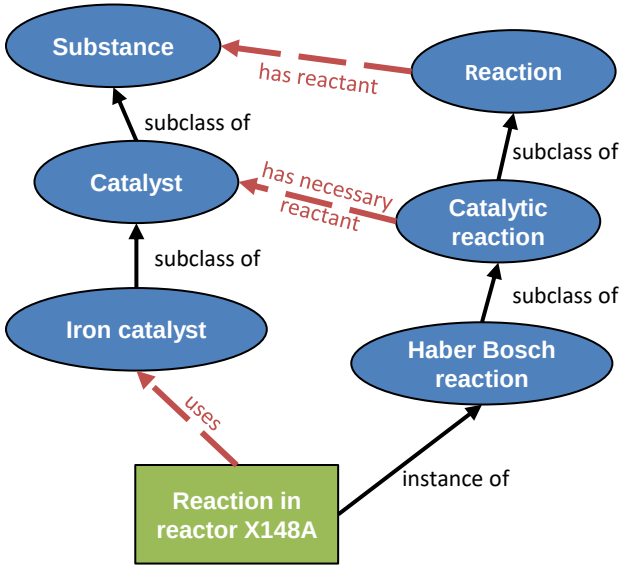
enabling sustainable production of chemicals and energy carriers

NFDI4Cat – Core Topics and Structure



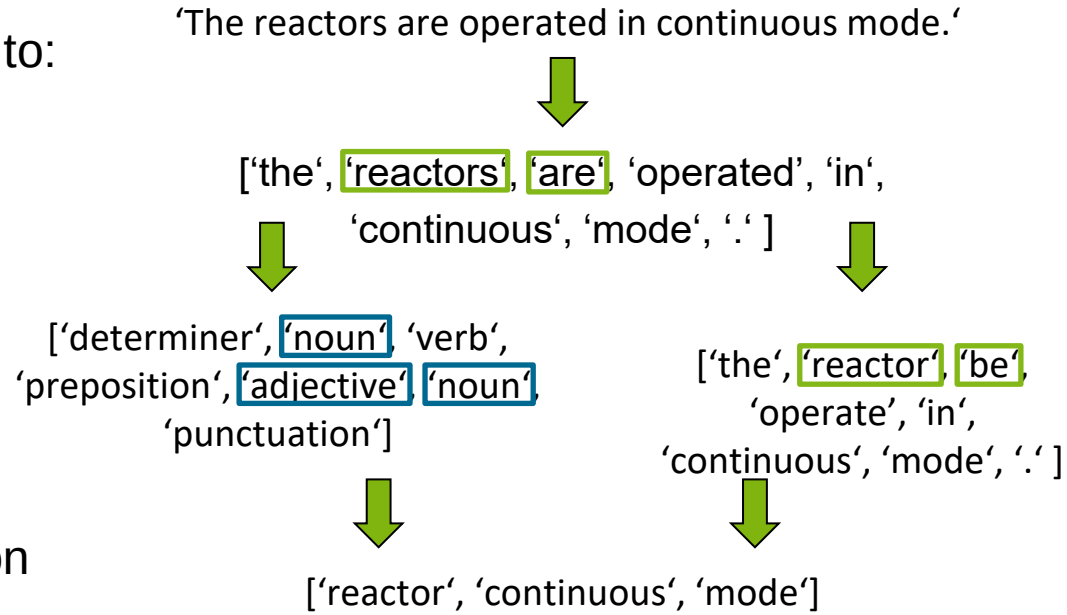
Ontologies – A simple example

- Ontologies consist of
  - **Classes** to express concepts
  - **Relations** between classes
  - **Individuals** representing real existing elements
  - **Rules**, like
    „Catalytic reactions need one or more catalyst“
- Information is stored in triplets
- Reasoning enhances data
  - “*The reaction in reactor X148A uses an iron catalyst*“
  - ➢ Inference yields: “*The reaction in reactor X148A is a Haber-Bosch reaction which in turn is a catalytic reaction and uses iron catalyst as catalyst.*“



Preparing Natural Language for Processing

- Natural Language Processing to:
  - Split sentences and words
  - Extract relevant data
- Tokenizing
- Lemmatizing
- Part-of-speech (POS) tagging
- ➢ Using Python‘s SpaCy for POS tagging, tokenization, and lemmatization

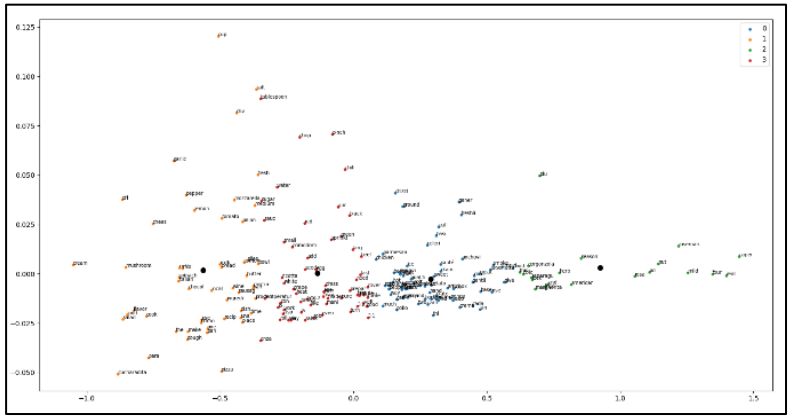


Word2Vec and min_count

- Vectorization of lemmatized tokens
- Training neural network with text
- Vector size of Word2Vec model = 300
- Similarity of words by cosine similarity

$$\cos(\varphi) = \frac{\vec{A} \cdot \vec{B}}{\left|\left|\vec{A}\right|\right| \left|\left|\vec{B}\right|\right|}$$

- min_count to filter data



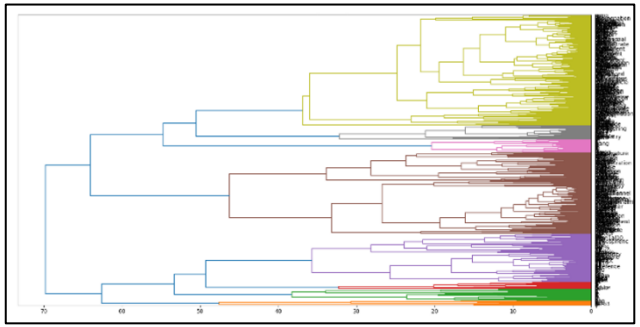
reactor, reactor, reactor, continuous, continuous, mode

min_count	
2	3
reactor	reactor
continuous	

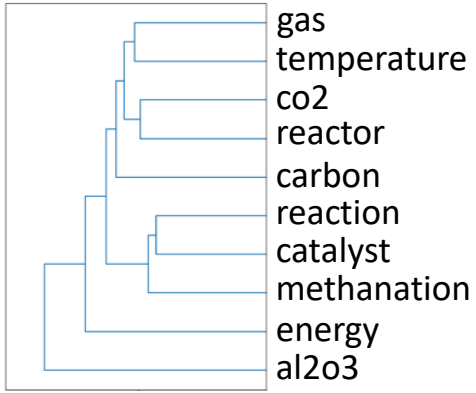
Clustering Methanation of CO_2

- Dataset: 28 papers on methanation of CO_2
- Found 535 different words that occurred more than 10 x (nouns only) in dataset (min_count = 10)

- Clustering approach not that helpful
  - Only two concepts at a time combined as siblings
  - Semantic similarity detected by Word2Vec useful only to extend



Dendrogram for a min_count of 10 (word has ≥10 repetitions in dataset)

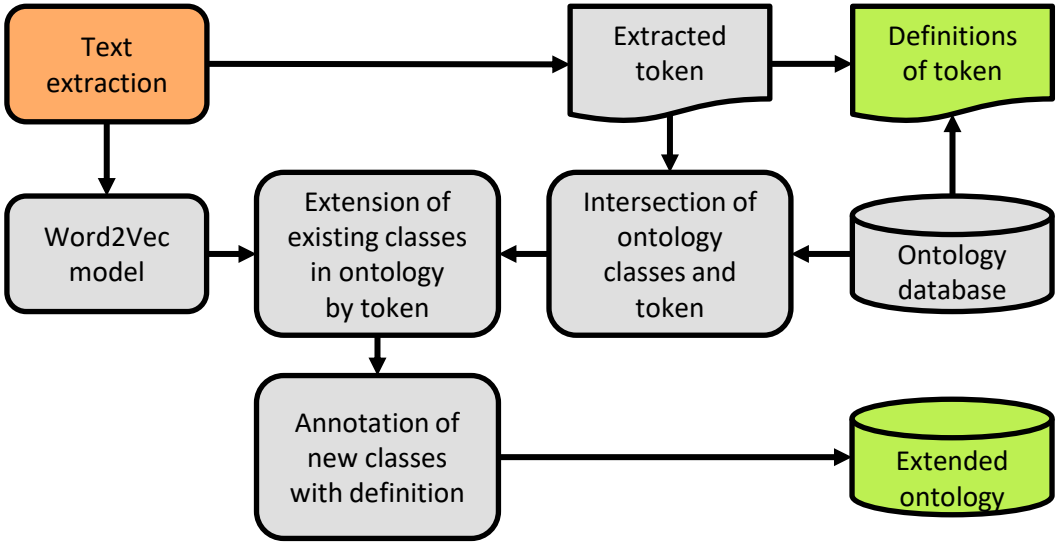


Dendrogram for a min_count of 500 (word has ≥ 500 repetitions in dataset)

Workflow for NLP-based Ontology Extension

- Including textual information into ontologies
- Textual definitions along with tokens to determine of best fitting definitions & ontologies
- Automatically extend ontologies based on text resources

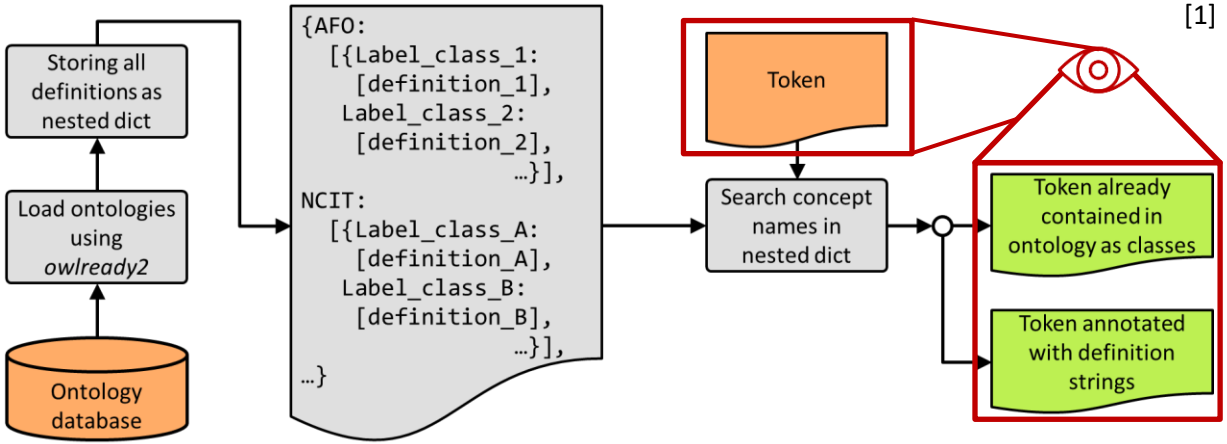
[1]



[1] A. S. Behr, M. Völkenrath, N. Kockmann. Ontology Extension with NLP-based Concept Extraction for Domain Experts in Catalytic Sciences, Knowledge and Information Systems, 2023, DOI: 10.1007/s10115-023-01919-1 GitHub: github.com/TUDoAD/NLP-Based-Ontology-Extender

Annotation of Extracted Token Based on Ontologies

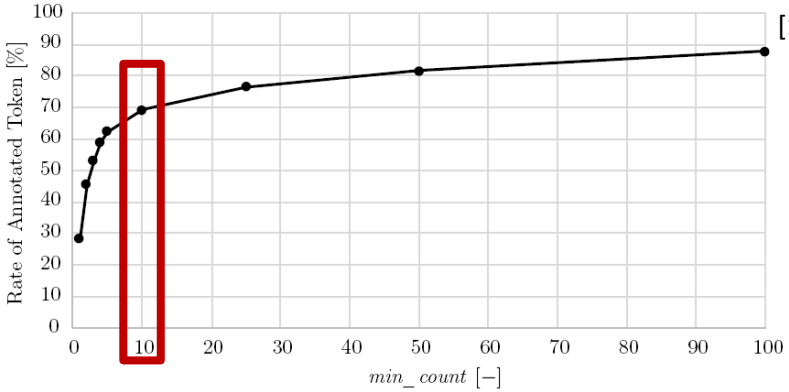
- Extracting textual definitions and classes from existing ontologies
- Compare to set of tokens extracted from text data (PDFs,…)
- ➢ Number of tokens per ontology already contained in respective ontologies
- ➢ Annotation of tokens with definition strings



[1] A. S. Behr, M. Völkenrath, N. Kockmann. Ontology Extension with NLP-based Concept Extraction for Domain Experts in Catalytic Sciences, Knowledge and Information Systems, 2023, DOI: 10.1007/s10115-023-01919-1 GitHub: github.com/TUDoAD/NLP-Based-Ontology-Extender

Results of Annotation

- min_count = number of minimal occurrence of same tokens in dataset
- Rate of annotated tokens rises with higher min_count



[1]

	min_count					
	1	5	10	25	50	100
AFO	218	130	97	62	42	27
BAO	100	56	37	25	15	9
CHEBI	107	42	27	23	16	5
CHMO	57	30	21	9	7	3
SBO	37	29	24	21	19	10
IUPAC-Goldbook	365	194	145	94	60	37
NCIT	935	440	300	172	103	54
Sum of annotated token	1178	537	364	211	125	65
Overall amount of token	4170	861	525	276	153	74
Rate of annotated token (in %)	28.25	62.37	69.33	76.45	81.70	87.84

min_count = 10 and AFO (Allotrope® Foundation Ontology) best pick for next experiments?

[1] A. S. Behr, M. Völkenrath, N. Kockmann. Ontology Extension with NLP-based Concept Extraction for Domain Experts in Catalytic Sciences, Knowledge and Information Systems, 2023, DOI: 10.1007/s10115-023-01919-1 GitHub: github.com/TUDoAD/NLP-Based-Ontology-Extender

Annotation of Tokens

- Annotation of tokens with definition strings

- min_count = 10

Token	AFO	SBO	IUPAC Goldbook
Catalyst	-	Substance that accelerates the velocity of a chemical reaction without itself being consumed or transformed. This effect is achieved by	A substance included in the solvent to increase the rate of transfer without affecting the position of equilibrium. The term accelerator may also be used but kinetic

Matches \ Ontology / Dict.	AFO	BAO	CHEBI	CHMO	SBO	IUPAC-Goldbook	NCIT	Total	Total
#	97	37	27	21	24	145	300	**364**	**535**
%	18.48	7.05	5.14	4.00	4.57	27.62	57.14	**69.33**	**100.00**

Automated annotation of tokens found by workflow with definition strings from other semantic artifacts!

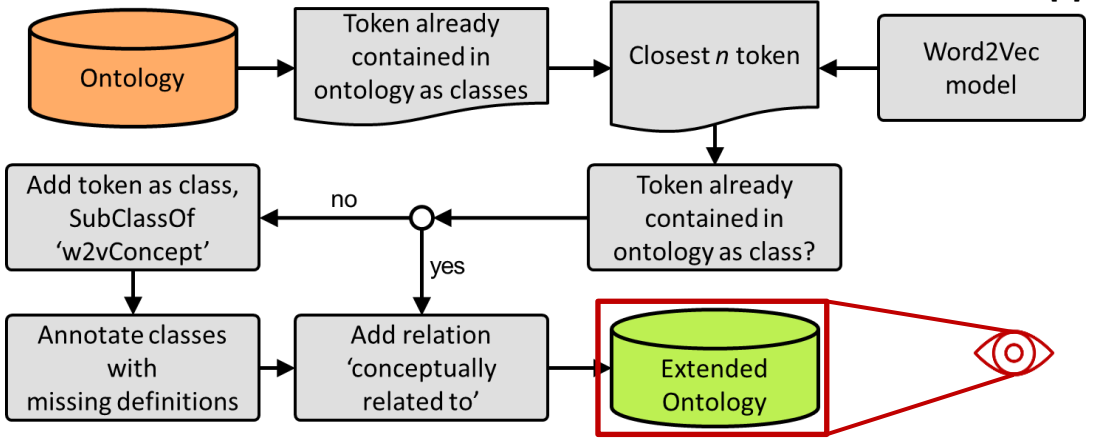
Extension of Existing Ontologies

- Extend AFO by classes suggested by Word2Vec
- Words deemed as close by Word2Vec ≠ hierarchical similarity

➢ Use token already contained in AFO as seed for Word2Vec

➢ Include new concepts and relations based on Word2Vec output

[1]

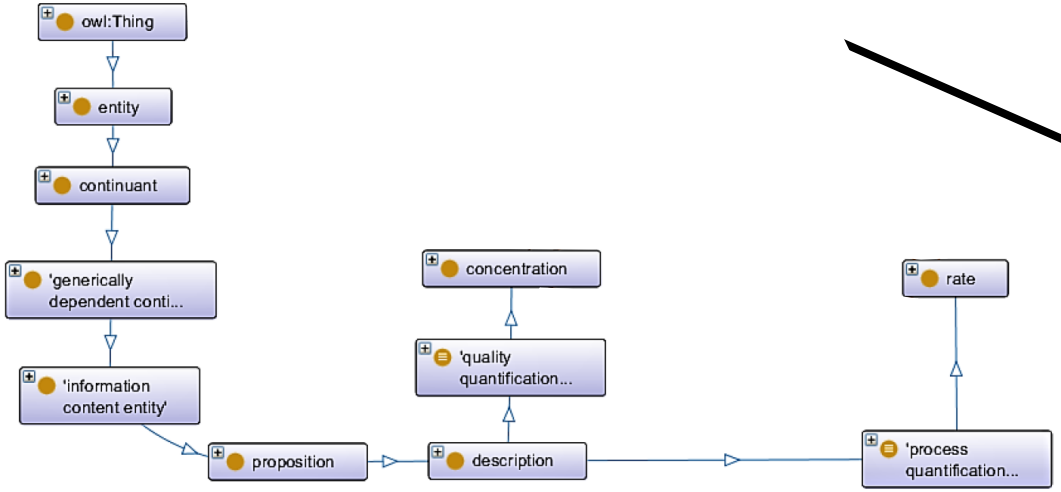


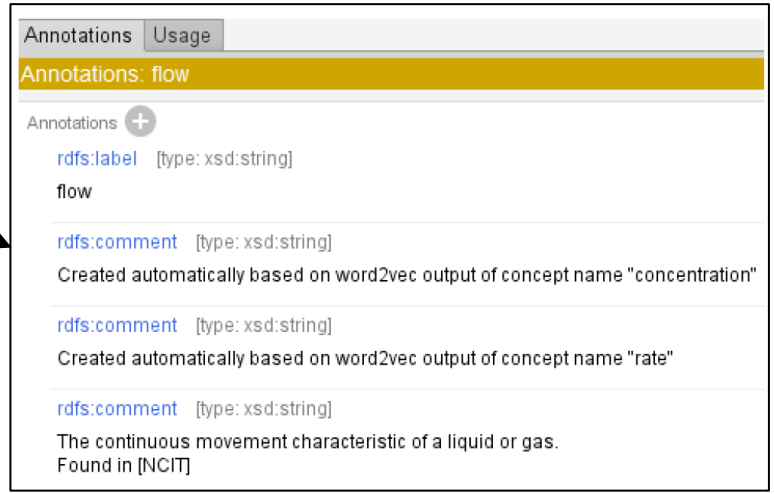
[1] A. S. Behr, M. Völkenrath, N. Kockmann. Ontology Extension with NLP-based Concept Extraction for Domain Experts in Catalytic Sciences, Knowledge and Information Systems, 2023, DOI: 10.1007/s10115-023-01919-1 GitHub: github.com/TUDoAD/NLP-Based-Ontology-Extender

Example of AFO Extension*

* *Extensions made to the AFO have not gone through the established Allotrope® Foundation governance process and are not officially endorsed or supported by Allotrope Foundation.*

- Blue: *subClassOf*
- Orange: *conceptually_related_to*



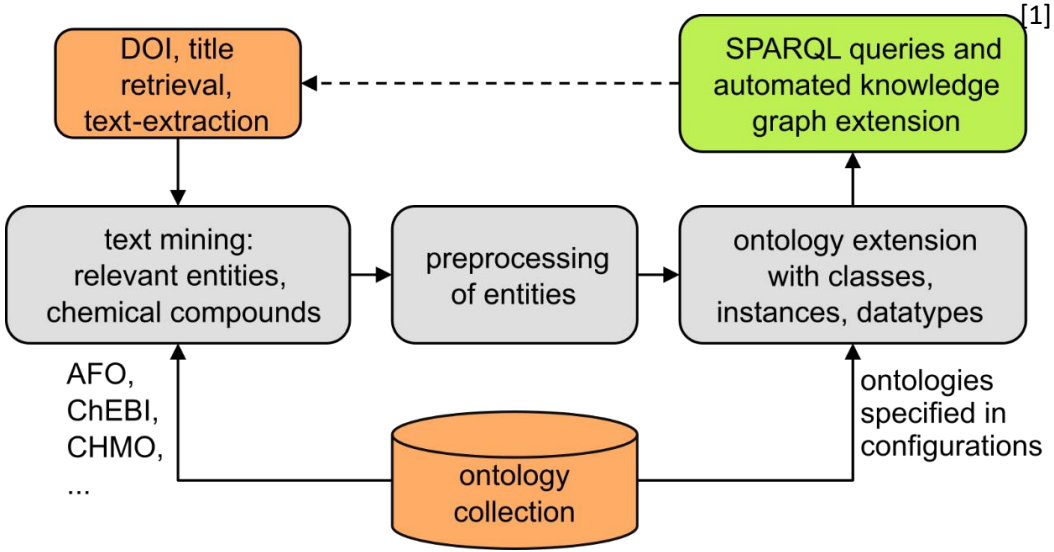


[1]

[1] A. S. Behr, M. Völkenrath, N. Kockmann. Ontology Extension with NLP-based Concept Extraction for Domain Experts in Catalytic Sciences, Knowledge and Information Systems, 2023, DOI: 10.1007/s10115-023-01919-1 GitHub: github.com/TUDoAD/NLP-Based-Ontology-Extender

Metadata Extraction from Literature with NLP

- CatalysisIE for information extraction
- Categorization of extracted entities in six categories:
  - catalyst
  - reaction
  - characterization
  - reactant
  - product
  - treatment



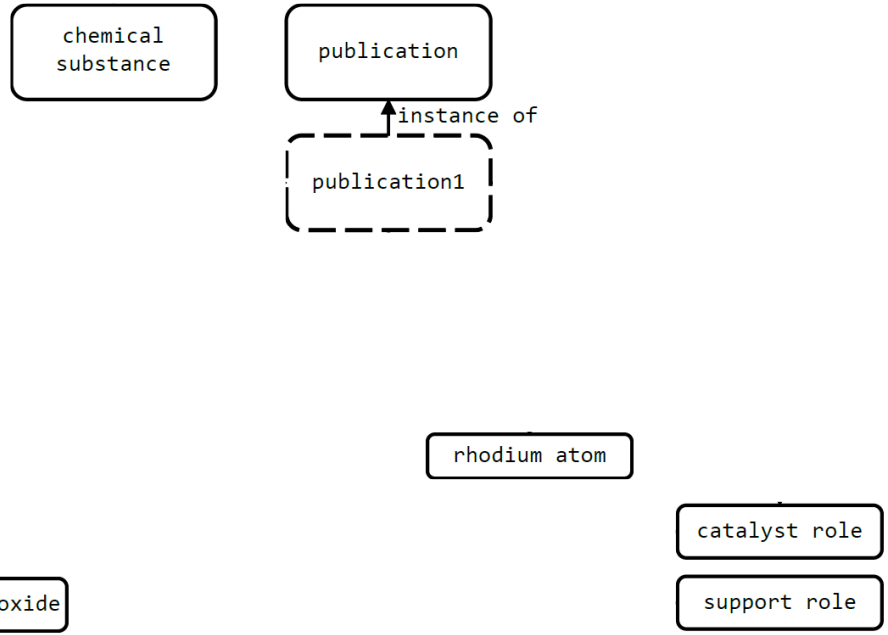
➢ Structuring of data in knowledge graphs on ontologies

[1] A. S. Behr, D. Chernenko, D. Koßmann, A. Neyyathala, S. Hanf, S. A. Schunk, N. Kockmann. Generating knowledge graphs through AI-assisted text mining of catalysis research related literature, Catalysis Science & Technology, **2024**, in review, submitted 2024-03-19

Assigned Relations

[1]

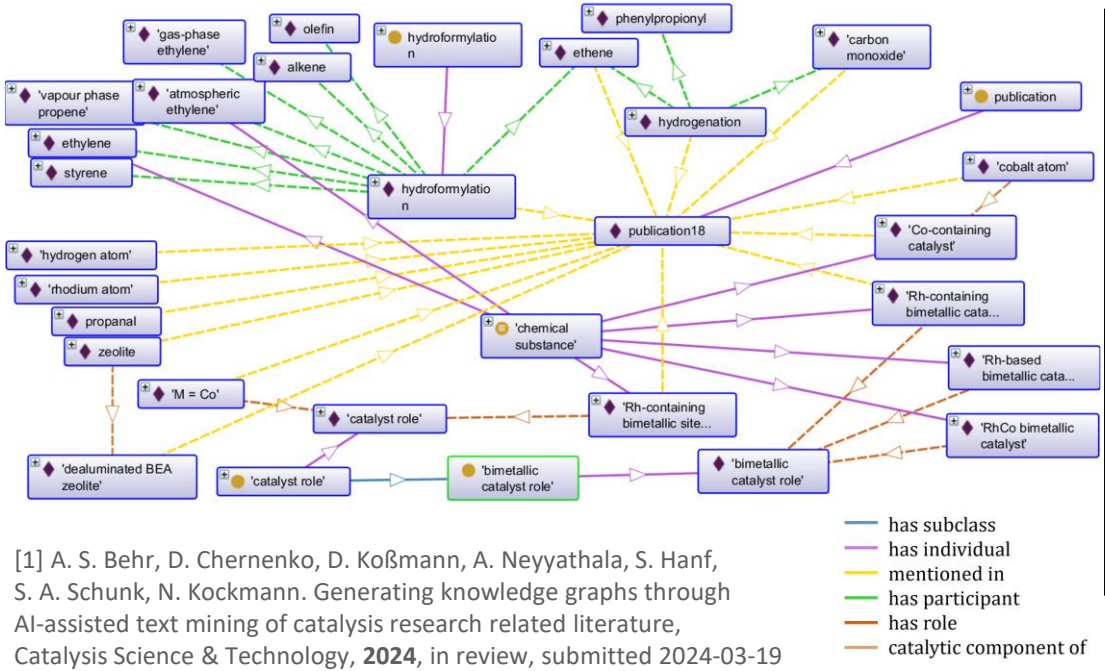
- Structured information on publications in knowledge graph
- Individuals to publication also contain more information, e.g. DOIs

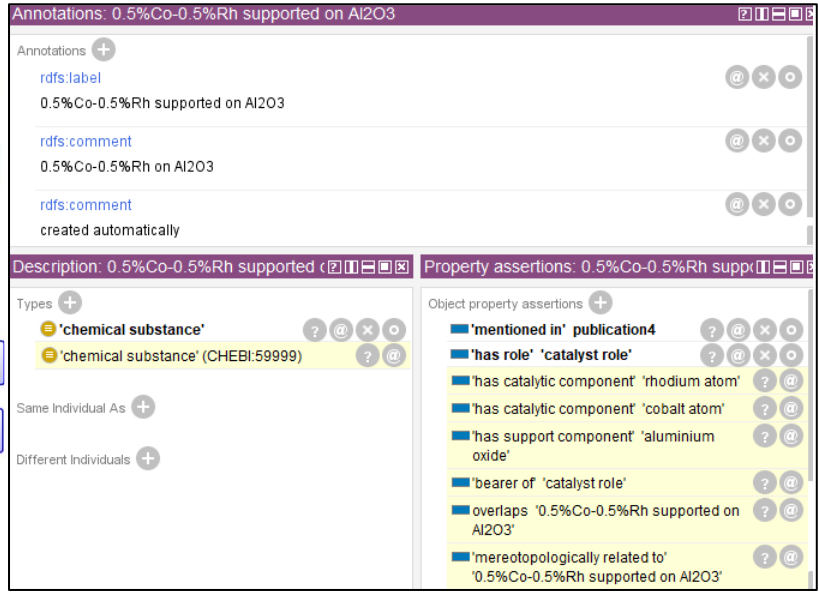


[1] A. S. Behr, D. Chernenko, D. Koßmann, A. Neyyathala, S. Hanf, S. A. Schunk, N. Kockmann. Generating knowledge graphs through AI-assisted text mining of catalysis research related literature, Catalysis Science & Technology, **2024**, in review, submitted 2024-03-19

Extracted and Inferred Relations

[1]

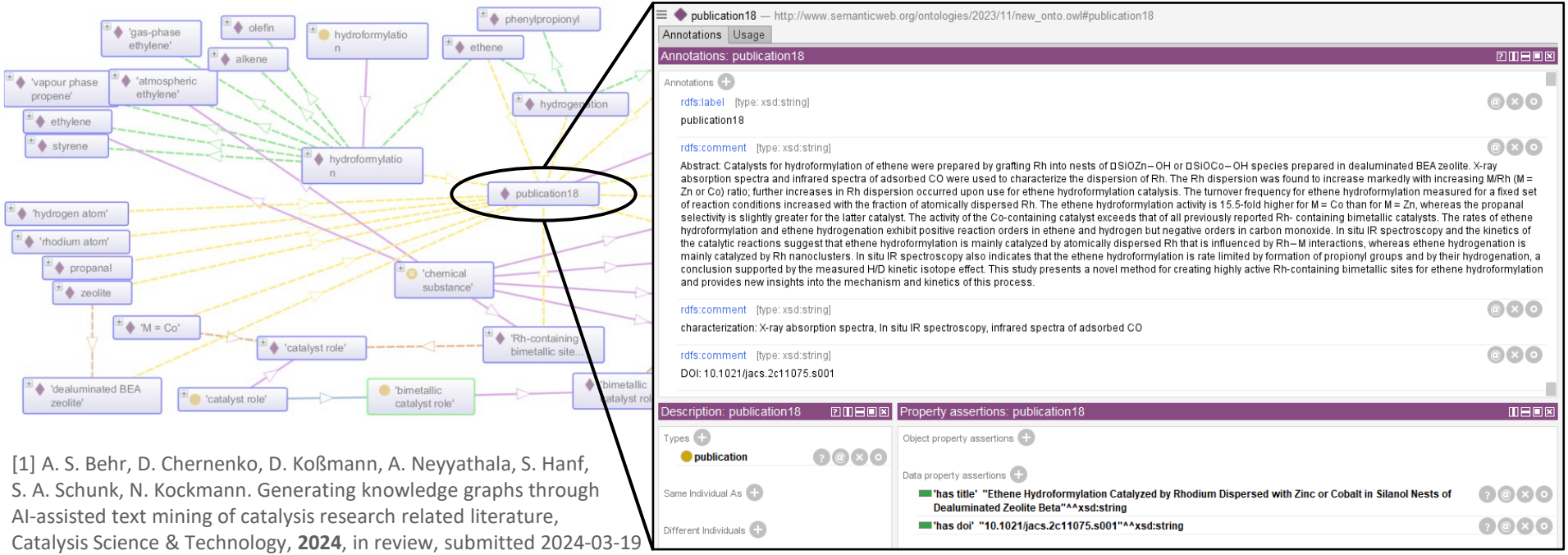




[1] A. S. Behr, D. Chernenko, D. Koßmann, A. Neyyathala, S. Hanf, S. A. Schunk, N. Kockmann. Generating knowledge graphs through AI-assisted text mining of catalysis research related literature, Catalysis Science & Technology, **2024**, in review, submitted 2024-03-19

Further Information Stored within Publication Individuals

[1]



[1] A. S. Behr, D. Chernenko, D. Koßmann, A. Neyyathala, S. Hanf, S. A. Schunk, N. Kockmann. Generating knowledge graphs through AI-assisted text mining of catalysis research related literature, Catalysis Science & Technology, **2024**, in review, submitted 2024-03-19

Querying Knowledge

- Jupyter notebooks for predefined SPARQL queries

1. Show me the abstract of a specific publication
2. I need all publications that mention the same reactions as in this specific publication
3. Give me the list DOIs of publications, that mention the specific reaction
4. …

1.

```
In [4]: doi_1=r'10.1021/acscatal.1c04359'
        abstract=get_abstr(doi_1)
```

Abstract: The reaction mechanisms of heterogeneous hydro- formylation of ethylene and propylene were compared at 413−453 K using RhCo3/MCM-41 as catalysts. The reaction rates of propylene for both hydroformylation and the undesired side reaction of hydrogenation were found to be about one order of magnitude lower than those for ethylene in flow reactor studies. The difference in the kinetic behavior between ethylene and propylene was investigated by measuring the reaction orders and apparent activation energies, and these macrokinetic observables were analyzed using the degree of rate control (DRC) method. In situ diffuse reflectance infrared Fourier transform spectroscopy (DRIFTS) experiments were performed to characterize the surface intermediates formed during the reactions. When the reactant was changed from ethylene to propylene, the IR peak corresponding to adsorbed CO exhibited a significant increase, while the IR peaks of the alkyl group decreased in magnitude. Combined with the DRIFTS results, DRC analysis indicates that the first step of olefin hydroformylation, the formation of an alkyl group on the catalyst surface, plays a key role in the difference between ethylene and propylene. This step is kinetically nonrelevant when ethylene is the reactant, but it is one of the rate-controlling steps for propylene. The low concentration of the adsorbed propyl group, which is a common intermediate shared by both hydroformylation and hydrogenation of propylene, decreases the rates of both reaction pathways as compared to ethylene. KEYWORDS: hydroformylation, ethylene, propylene, kinetics, degree of rate control

2.

```
In [5]: list_reac_doi,_ = get_reaction(reac=None,doi=doi,include_all=False) #doi=None (if from all publications)
        reac_all = [*set([i[0].lower() for i in list_reac_doi])]
        print(reac_all)
```

```
['heterogeneous hydroformylation', 'hydroformylation', 'hydrogenation']
```

3.

```
In [6]: list_reac_doi,_ = get_reaction(reac = None,doi = doi) #get list of all reactions mentioned in given doi (doi should be part
        same_reac_doi = []
        for i in list_reac_doi:
            reac_doi,_ = get_reaction(reac = i[0],doi = None)
            for c in reac_doi:
                if c not in same_reac_doi and c[0] != doi:
                    same_reac_doi.append(c)     #output example: [['10.1016/0304-5102(93)87113-m'],['10.1016/1381-1169(96)00243-9']]

        print(list_reac_doi)
        print(same_reac_doi)
```

```
[['hydroformylation'], ['hydrogenation'], ['heterogeneous hydroformylation']]
[['10.1021/acscatal.1c02014.s001'], ['10.1021/acscatal.0c04684.s001'], ['10.1021/acscatal.1c00705.s002'], ['10.1021/acscatal.7b00499.s001'], ['10.1021/acscatal.9b02111.s001'], ['10.1021/jacs.1c09665.s001'], ['10.1016/j.apcata.2018.02.019'], ['10.1021/jacs.2c11075.s001'], ['10.1016/0304-5102(93)87113-m'], ['10.1016/1381-1169(96)00243-9'], ['10.1016/s0920-5861(00)00261-3'], ['10.1016/j.apcata.2013.10.019'], ['10.1016/s1381-1169(97)00035-6'], ['10.1021/acs.iecr.0c03437.s001'], ['10.1021/acs.iecr.9b03598.s001'], ['10.1021/acsami.0c21749.s001']]
```

Results – NLP-based Knowledge Extraction

- Tool for information extraction from publications of catalysts developed
  - Tested on two datasets (DS1: 19 publications, DS2: 26 publications)
  - New instances of “chemical substance”: DS1: 53, DS2; 55

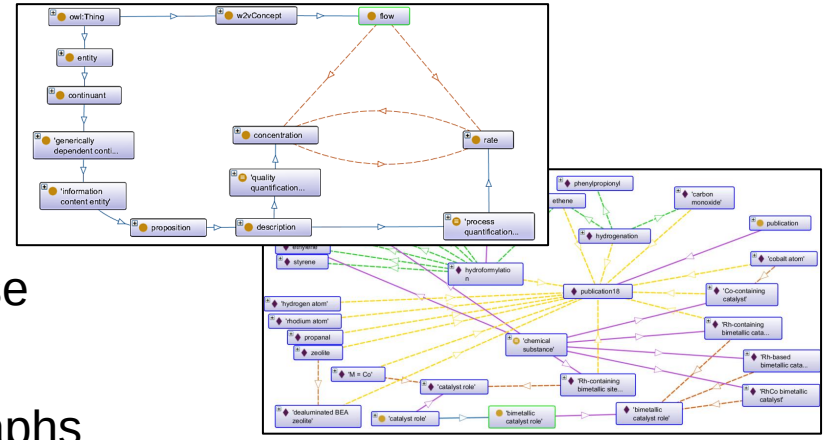
[1]

Metric	Initial ontology	Extended ontology Dataset 1	Extended ontology Dataset 2
classes	3116	3447	3338
instances	47	203	178
logical axioms	5755	6936	6596
`SubClassOf`	4823	5372	5174
Equivalent Classes	178	188	185

- Automated generation of catalysis knowledge graph based on AFO
- Automated search for similar publications found 731 publications similar to DS1

[1] A. S. Behr, D. Chernenko, D. Koßmann, A. Neyyathala, S. Hanf, S. A. Schunk, N. Kockmann. Generating knowledge graphs through AI-assisted text mining of catalysis research related literature, Catalysis Science & Technology, **2024**, in review, submitted 2024-03-19

Summary and Outlook



- Automated extension of AFO based on scientific literature on catalysis research
  - Extraction of concepts from text data base
  - Annotation of classes
  - Generation of AFO-based knowledge graphs
  - ➢ Governance process not integrated
- Further extension of workflow possible with *chemdataextractor*, e.g. for TOF:

```
doc = Document(
      Paragraph('For 0.6Rh/SiO2, the catalyst productivity was 44 mol/(molRh·h)
                 and the TOF was 2511 h-1 at the time on stream(TOS) of 5 h.'),
    models=[TurnOverFrequency])
```

```
output:
[{'TurnOverFrequency': {
   'raw_value': '2511',
   'raw_units': 'h-1',
   'value': [2511.0],
   'units': 'Hour^(-1.0)',
   'specifier': 'TOF',
   'compound': {'Compound': {'names':
    ['0.6Rh/SiO2']}}}}]
```

Thank you for your attention!
Questions?

Alexander Behr