Voc4Cat

Vocabulary Guidelines for NFDI4Cat

Version: 1.0 beta

February 2023

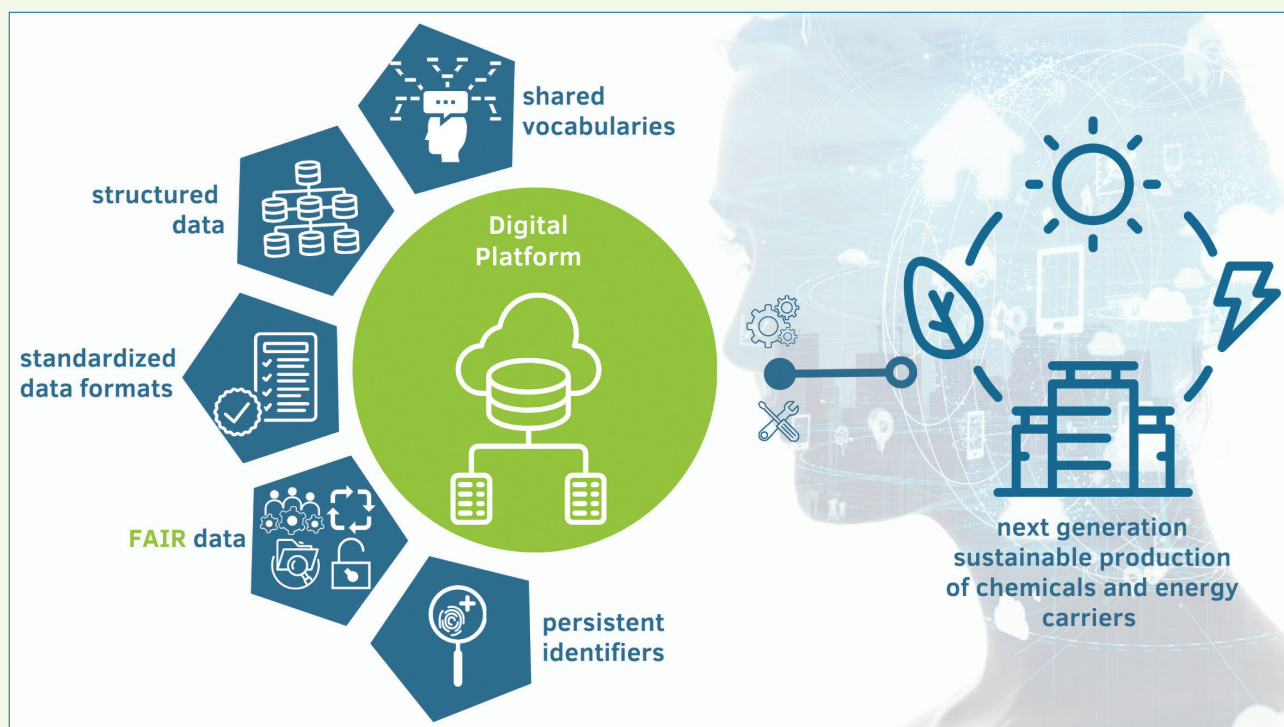**www.nfdi4cat.org**

Contents

1. About NFDI4Cat

Catalysis is a very complex and interdisciplinary scientific field that enables the efficient production of a wide range of products for different industries and at different production scales. Thus, catalysis is one of the key technologies for addressing major challenges such as a sustainable energy supply or climate change. Data management in catalysis is currently organized mainly at institutional or working group level and based on local conventions. Overarching repositories and standards on how data including metadata should be stored exist only in rudimentary form.

In order to generate the highest possible added value from catalysis research, a fundamental change in catalysis research and catalysis-related sciences such as chemical engineering and process engineering is required. Bringing together the different disciplines of catalysis in terms of data management is the central challenge. Out of this need, the NFDI4Cat consortium was founded as part of the *National Research Data Infrastructure* (NFDI) initiative.

Scientific data management should abide to the **FAIR** guiding principles. This means that data must be *Findable*, *Accessible*, *Interoperable* and *Reusable*. To this end, the development of vocabularies towards the formation of catalysis-specific ontologies is of outmost importance.

NFDI4Cat focuses on shaping the digital future of catalysis.

shared vocabularies

structured data

Digital Platform

standardized data formats

FAIR data

persistent identifiers

next generation sustainable production of chemicals and energy carriers
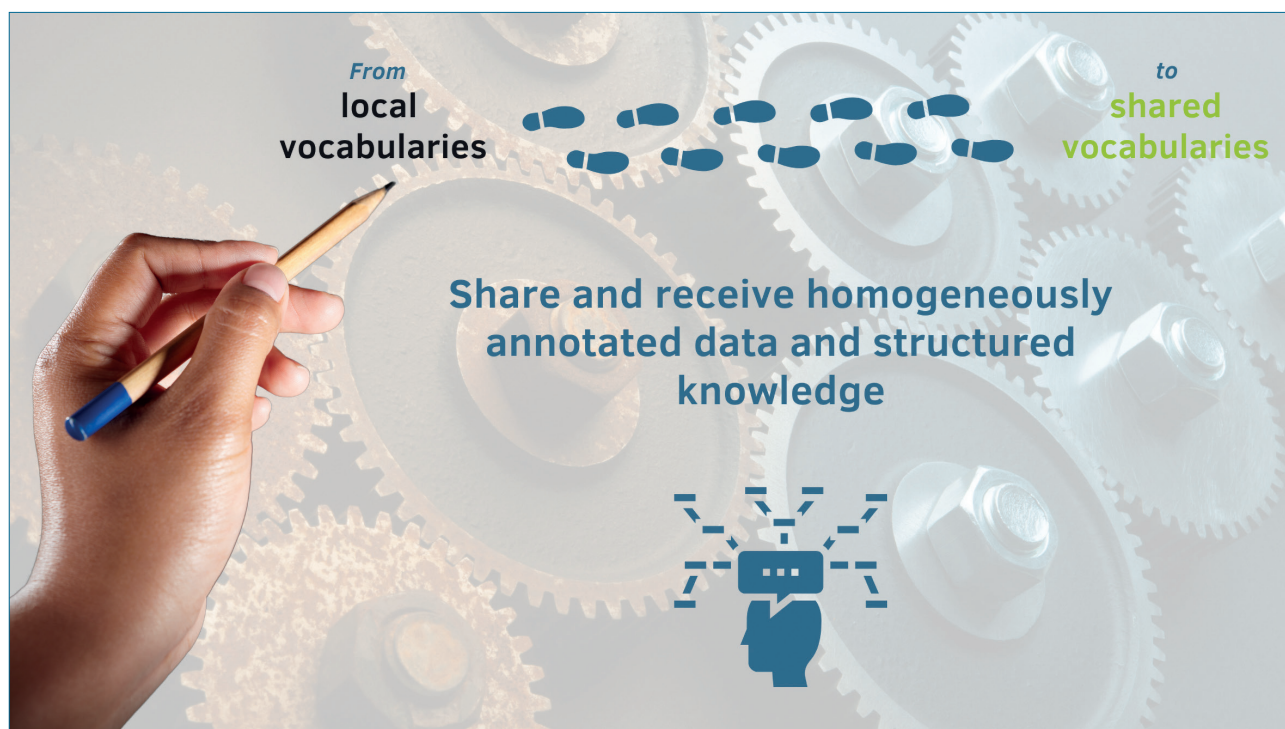
2. Project partners at a glance

NFDI4Cat consists of 16 experienced partners from all areas of catalysis. The disciplines of reaction and process engineering are also represented in the consortium. The catalysis and engineering competencies are complemented by expertise in the fields of data science, high-performance computing and machine learning.

3. About the VoG4Cat guidelines

The VoG4Cat guidelines have been developed as a blueprint towards suggesting, adding, and editing content to the vocabularies developed throughout NFDI4Cat. The aim of VoG4Cat is to provide guidelines to guarantee consistency and coherence on selection of concepts and terms between all catalysis-related vocabularies in NFDI4Cat. VoG4Cat, initially inspired by the AGROVOC editorial guidelines of FAO,[1] evolved over the course of *Task Area 1 (TA1): Ontology Development and Metadata Standards* of NFDI4Cat.

From
local vocabularies

to
shared vocabularies

Share and receive homogeneously annotated data and structured knowledge

---

[1] FAO. 2022. The AGROVOC Editorial Guidelines 2020 – Second edition. Rome. (https://doi.org/10.4060/cb8640en)

4. Editorial group members

The editorial group of the first draft, comprises of the members of the NFDI4Cat project involved in the preparation and development of the vocabularies in

Task Area 1 (TA1): Ontology Development and Metadata Standards:

In alphabetical (based on surname) order:

Prefix	Name	Surname	Affiliation	Status
M. Sc.	Alexander	Behr	TUDO	Active
M. Sc.	Hendrik	Borgelt	TUDO	Active
Dr.	Vishwas	Chandraskekhar	LIKAT	Inactive
Dr.	Mark	Dörr	Uni Greifswald	Active
Dr.	Nadiia	Huskova	HLRS	Active
Dr.	Rachit	Khare	TUM	Active
	Julia	Köbl	FAU	Inactive
Dr.	David	Linke	LIKAT	Active
Dr.	Nikolaos	Moustakas	LIKAT	Active
Dr.	Taras	Petrenko	HLRS	Active
M. Sc.	Manal	Talab	HLRS	Inactive

TUDO: Technical University of Dortmund, LIKAT: Leibniz-Institute for Catalysis, HLRS: High-Performance Computing Center Stuttgart, TUM: Technical University of Munich, FAU: Friedrich-Alexander-Universität Erlangen-Nürnberg, Uni Greifswald: University of Greifswald.

The term ''Editorial group'' is used to express provenance of self-made definitions of concepts included in the vocabularies developed in TA1. Members of the editorial group also have the role of the curation of the developed vocabularies.

5. Vocabularies and ontologies

To create shareable and interoperable (meta) data in the field of catalysis research, the presence of common machine-readable identifiers is practical as it facilitates faster processing and enhances the findability and reusability of each (meta) data set.

In NFDI4Cat we created a vocabulary containing such unique identifiers (concepts). A vocabulary should be:

» Formal: unified regularized spelling coupled with an interna-tional unique identifier (IRI) allowing for machine interpretabil-ity

» Explicit: strict definitions allow for a common interpretation of synonymously used words and concepts

» Shared: approved by the community and actively supported and curated.

This vocabulary serves as the foundation for sharing data between available datasets and databases within NFDI4Cat. It shall form the basis of more logical stringent sets of terms, relations and constraints, called ''ontologies''. These ontologies have a graph-like structure and will allow logic reasoning and derivatization of knowledge by logical inference.

Ontologies represent knowledge explicitly in a machine-readable way by interconnecting data via its properties, relations and constraints using standardized concepts. The standardized concepts being machine-readable identifiers, can in turn be used for, e.g., data annotation. Such a graph-based interconnection of knowledge allows therefore for an increase of data FAIRness.

Ontologies exist for a variety of research fields, but so far there isn't one that can describe sufficiently catalysis-related data. NFDI4Cat and *Task Area 1 (TA1): Ontology Development and Metadata Standards* of NFDI4Cat aims to create the first catalysis-specific vocabulary and ontology, supported by the catalysis community members, and periodically curated by field experts.
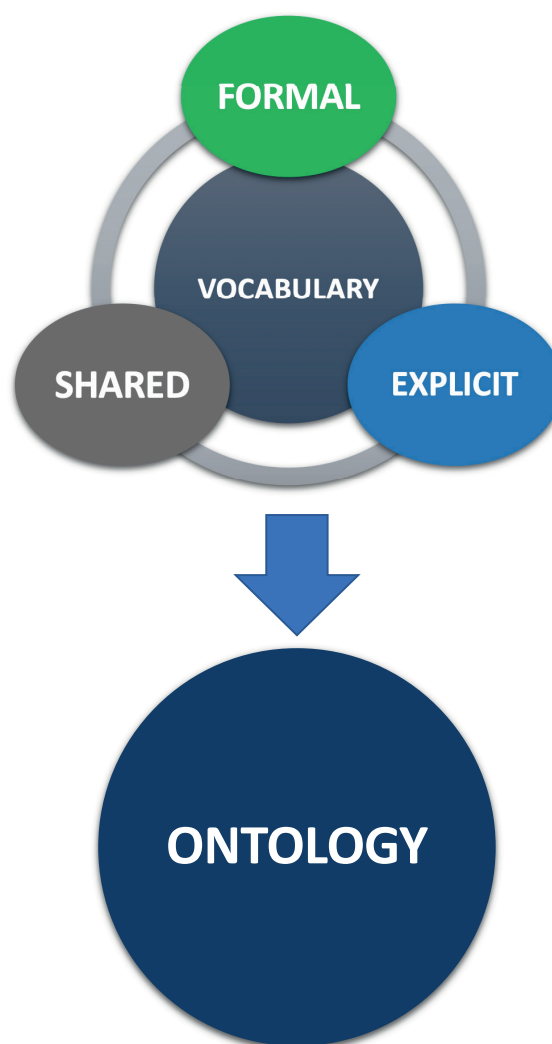
6. The VocExcel template

For the simplified collection and development of vocabularies for NFDI4Cat, a workflow, called the VocExcel, developed by an Australian company, (Surround Australia Pty Ltd.) was adapted and extended resulting in a more universal tool.

Introduction sheet of the VocExcel template

The developed vocabularies include a variety of concepts and their definitions describing the most often referenced terms in the scientific literature. The template allows for the selection of a Preferred Label (Best name for a concept), Alternate Labels (Other names for a concept), for Children – Parent relations and Related – Close – Exact – Narrower – Broader matches between concepts.
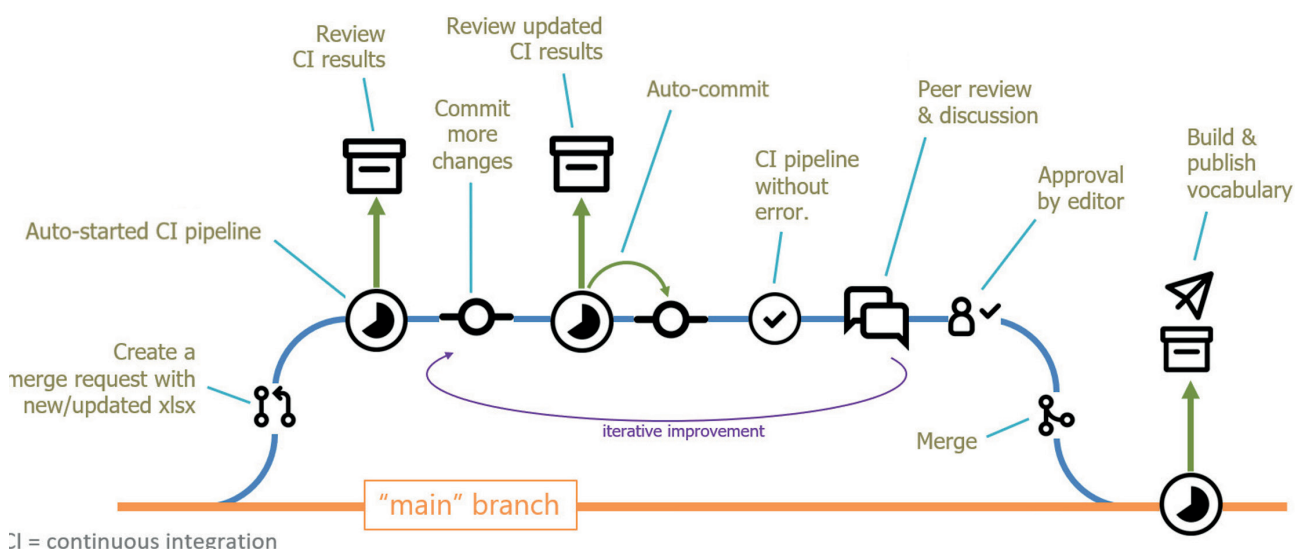
VocExcel can be accessed using the link below:
https://github.com/nfdi4cat/VocExcel

7. Editing NFDI4Cat vocabularies

A vocabulary contribution workflow has been developed that allows an easy integration of contribution from the catalysis community members.

Vocabulary Contribution Workflow

To contribute concrete new vocabularies or change existing ones, a community member may either submit a contribution as an Excel-file (recommended to use the VocExcel template) or (as an expert) as new/change turtle file. The steps for submitting updates in a turtle format are not different from contributing source code to other github repositories. Before submitting a change or an addition of a concept, it is advised that the community member checks if the concept already exists, as duplicates should be avoided. In addition, if a proposed concept can be broken down to simpler/individual terms, it is advised to do so.

To contribute your changes, use the link below:
https://github.com/nfdi4cat/voc4cat

The steps for submitting changes in an Excel form are:

1. Download the Excel-vocabulary file.
   a. For an existing vocabulary, download the latest version from NEED TO ADD A LINK
   b. For a new vocabulary, use the template from the tem-plates-folder NEED TO ADD A LINK

2. Add the proposed changes to the Excel file.

3. Create a merge request with the updated Excel-file.
   a. The name of the Excel-file must match the vocabulary that you want to update (e.g., catvoc.xlsx to update a vocabu-lary named ''catvoc'')
   b. New vocabularies will be named like the excel-file (minus the .xlsx extension).

4. The merge request will be processed automatically by the CI/CD pipeline.

5. Artifacts/logs generated by the CI pipeline are reviewed.

6. If no errors occur, the member's contribution will either:
   a. be directly merged by the maintenance – curation team,
   b. require more changes to be committed,
   c. or be rejected.

7. An updated version of the vocabulary will be published.

8. Vocabulary guidelines

The following paragraphs contain the guidelines to be followed when creating a new vocabulary, or when suggesting changes in the already developed vocabularies of NFDI4Cat. Adhering to these guidelines will ensure a more uniform vocabulary and a faster and more accurate curation of the vocabularies.

8.1 Duplicates

Duplicates of existing concepts must be avoided: synonyms, spelling variants, singular VS plural and hyphenation must be taken into consideration.

8.2 Spelling variants

British English spelling will be used as standard in the development of the vocabularies. Alternative spellings can be added as an *Alternate Label* in a concept.

Spelling rules for English:

» words ending in -re/-er. Prefer -re: litre, metre;

» words ending in -our/-or. Prefer -our: colour;

» words ending in -ence/-ense. Prefer -ence (for nouns): licence;

» words ending in -l and followed by a suffix. Prefer -ll (not -l): modelling;

» words ending in -ize/-ise. Prefer -ize: immobilize, categorize;

» words ending in -yse/-yze. Prefer -yse: analyse, catalyse

8.3 Hyphenation

The non-hyphenated version of a concept is always preferred, unless the hyphenated version is always the one used in literature. Both versions of a concept (hyphenated and non-hyphenated) must be included (one of them as an alternate label). Examples: photoexcitation VS. photo-excitation, pretreatment VS. pre-treatment, counter electrode VS. counter-electrode, transition metal VS. transition-metal.

8.4 Slashes

» The slash (/) must be avoided (e.g., oxidation / reduction, forward / backward reaction, working / counter electrodes) wherever possible.

» Two separate concepts must be used if the terms differ in meaning, or one of them should be set as an alternate label if the meaning is the same.

» Antonyms should be generated into a separate concept.

8.5 One- VS. two-word forms

» If terms are seen in both two-word and one-word forms (e.g., ''thermal conductivity'' VS. ''thermoconductivity''), the preferred term should be the one-word form. The two-word form can be included as an alternate label.

» Multi-word terms should express a single concept or unit of thought. Complex compound concepts are discouraged. It is recommended (if possible) to split compound words into simpler concepts. Example: "reduction in hydrogen atmosphere" can be generalised to "gas treatment" which may be in hydrogen atmosphere.

8.6 Singular VS plural

Singular will always be used when naming concepts.

8.7 Lower case VS Capital letters

» Common nouns and noun phrases use lower case.

» The first letter of proper names must be a capital, when the proper name occurs withing longer terms, such in combination with ''method'', ''theory'' or ''process'' (e.g., 'Brønsted acid site, Lewis acid site, Röntgen crystallography, Langmuir-Hinshelwood mechanism).

8.8 Punctuation, diacritics and special characters

Appropriate punctuation, diacritics and other special characters of an individual language should be used (e.g., ü, ö, ß, ô, ñ). Example: Ångström, Brønsted acid sites, Röntgen crystallography.

8.9 Empty spaces and commas

(Extra) empty spaces must be avoided before, within and after a term. Unnecessary empty spaces could compromise search results. Examples (_ = empty space):

» _ semiconductor

» semiconductor_

» Charge _ _ transfer

Commas should be avoided unless in chemical substances.

8.10 Verbs and adjectives

» Verbs should be nominalized (transformed into nouns).

» The use of adjectives (without a noun), verbs and initial articles (e.g., ''A'', ''An'' or ''The'') should be avoided. Example: drying in an oven vs drying in oven (prefer drying in oven).

» Adjectives should only be used in compound terms, such as ''continuous flow'', but never alone.

8.11 Prepositional phrases

» Prepositional phrases should be avoided (e.g., alcohol oxidation rather than oxidation of alcohol, catalyst pretreatment rather than pretreatment of catalyst).

8.12 Abbreviations and acronyms

» The full form (written out) of a word / term should be generally selected as the preferred term. The abbreviated form or acronym should be included as the non-preferred label (e.g., gas chromatography VS. GC, density functional theory vs DFT).

8.13 Chemical compounds and elements

» Chemical compounds and elements in their full form, are written in lower case, such as ''carbon'', while the chemical symbol(s) is(are) always capitalized, such as ''C''.

» The full form is always the preferred term, whereas the symbol is the non-preferred term.

» Chemical compounds and chemical elements written out are non-countable nouns and should be explained in the singular.

» The definition should contain a unique identifier for chemical substances, such as the CAS number.

» Use the English spelling e.g, prefer aluminium (EN) over aluminum (USA).

8.14 Gender-neutral language

Gender-neutral language must be used whenever possible.

8.15 Other recommendations

» The following should be avoided:

- Trademarked names (e.g., brand names and commercial names). Example: Teflon VS. polytetrafluoroethylene.

- Names of individuals (unless as a part of the name of a method or technique)

- Names of specific software or models of devices.

- Names of policies, programs, or initiatives.

- Neologisms, slang and jargon.

8.16 Definitions

» A definition is a statement or formal explanation of the meaning of a concept. The following considerations should be followed when adding definitions:

- Definitions consist of a narrative and its source, which are both mandatory.

- A good definition explains what a concept is, not what it does.

- Only one definition per language is recommended.

- A definition is a sentence that requires orthographical rules, such as starting a sentence with a capital letter and ending a sentence with a full stop (period). The definitions need to be concise (ideally one sentence) and clear.

- A definition many be expressed in one or more languages, but the default choice should be English.

- A definition does not have to be a translation of an existing definition, but all definitions for a concept should have the same conceptual meaning.

- A definition should implicitly explain the difference from closely related concepts.

- Circular, imprecise or negative definitions should be avoided.

- A trusted and stable source for the definition, such as an external glossary or thesaurus of a well-known institution, is required. Credit must be given to the creator. When copyright is restrictive, it must be respected. The identify the source, a URL or a descriptive source (with details of the publication) can be used.

NFDI4(at

NFDI for Catalysis-Related Sciences