

# Development of an automatic text-based (inter)disciplinary classification of projects and publications using NLP

Research Information Platform with VIVO

@Berlin University Alliance

Florian Kotschka & Rolf Guescini

October 10<sup>th</sup>, 2023

# COMMUNITY



Network:  
BUA/DE/INT



[X/Twitter](#)

264 Follower



[VIVO Talks!](#)  
8 Events, Blog Articles



HU/CMS  
Communication



Project-[Webseite](#)  
& Platform



Conferences & Papers



Mailing list, [Newsletter](#)



VIVO [Membership](#)

## The Goal of BUA

„Scientists cooperate across disciplinary and institutional boundaries.“



The Research Information platform displays this cooperation digitally.



- Boost visibility & findability
- Link and structure research activities
- Facilitate external collaborations
- Connect internal projects

## OVERVIEW

As automated as possible and without much effort for the scientists.

## Welcome to BUA

Research Information Platform

All

Search form

Search



Using VIVO to connect research organizations, people, outputs, activities and things.

**Partners**

Charité – Universitätsmedizin Berlin  
 Freie Universität Berlin  
 Humboldt-Universität zu Berlin  
 Technische Universität Berlin

**Excellence Clusters**

Matters of Activity Cluster of Excellence  
 NeuroCure Cluster of Excellence  
 Science of Intelligence Cluster of Excellence

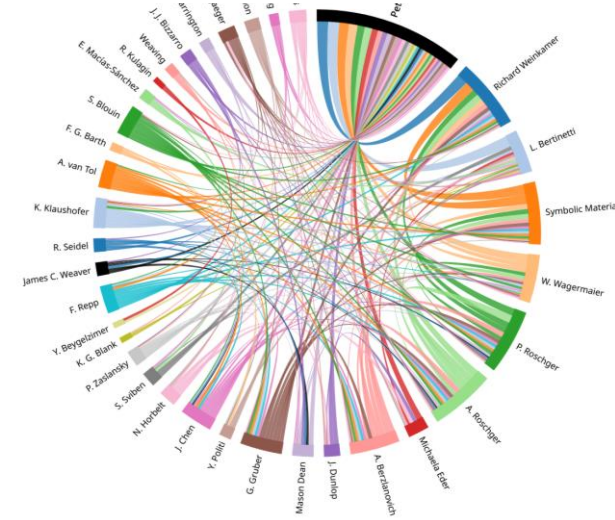
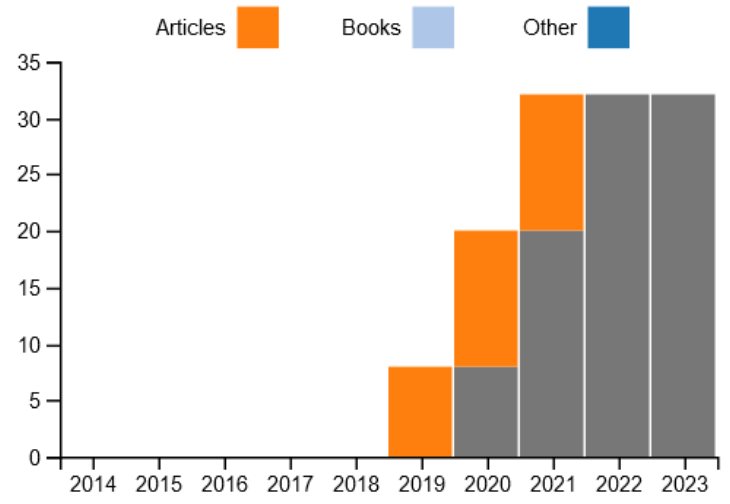
[View all ...](#)

**Projects**

Capabilities and consequences of recursive, hierarchical information processing in visual systems  
 Concept of intelligence  
 Intelligent kinematic problem solving  
 Multimodal Interaction and Communication

# VIVO

## @BERLIN UNIVERSITY ALLIANCE





Search

Materials Science Concept

Associated Departments

Matters of Activity Cluster of Excellence

Overview

knowledge area for

Department of Stability and Failure of Functionally Optimised Structures Subject domain  
 Experimental Physics / Material Sciences Subject domain

research area of

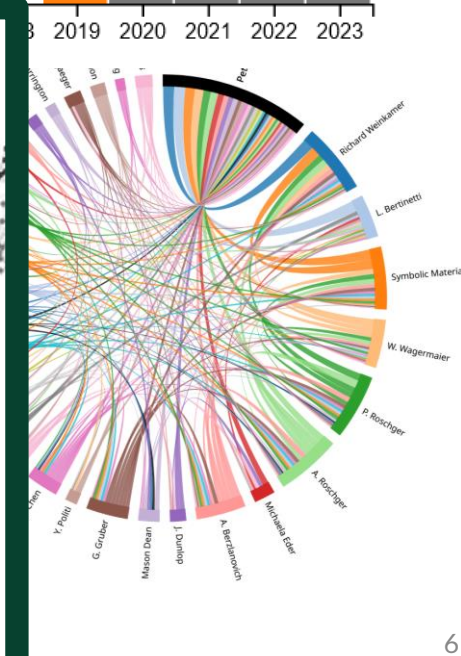
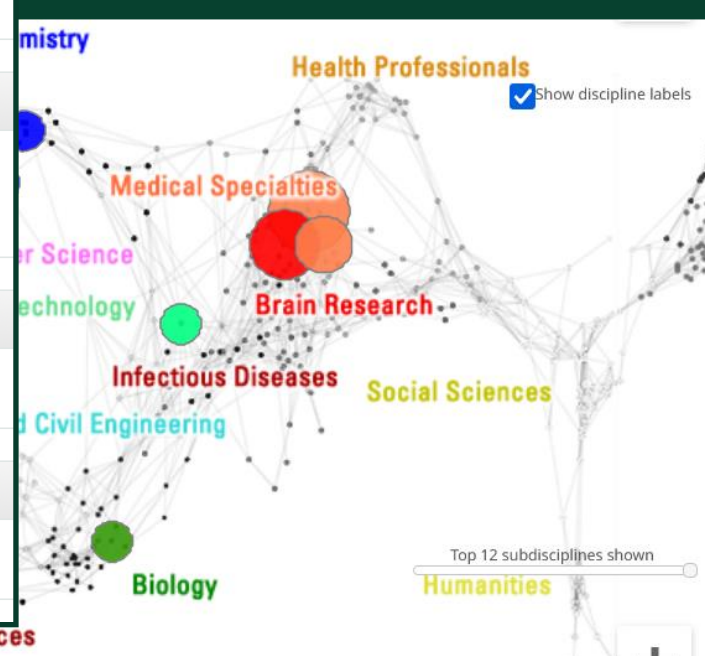
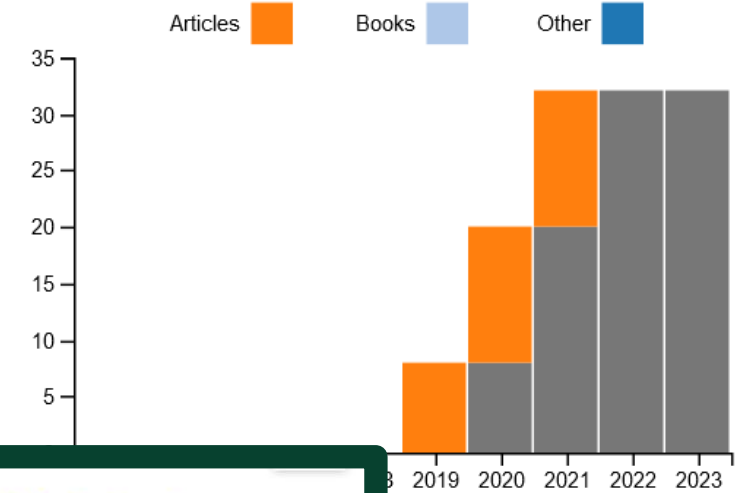
Peter Fratzl, Prof. Dr. Dr. h.c., Matters of Activity Cluster of Excellence  
 Richard Weinkamer, Member, Matters of Activity Cluster of Excellence

subject classification for

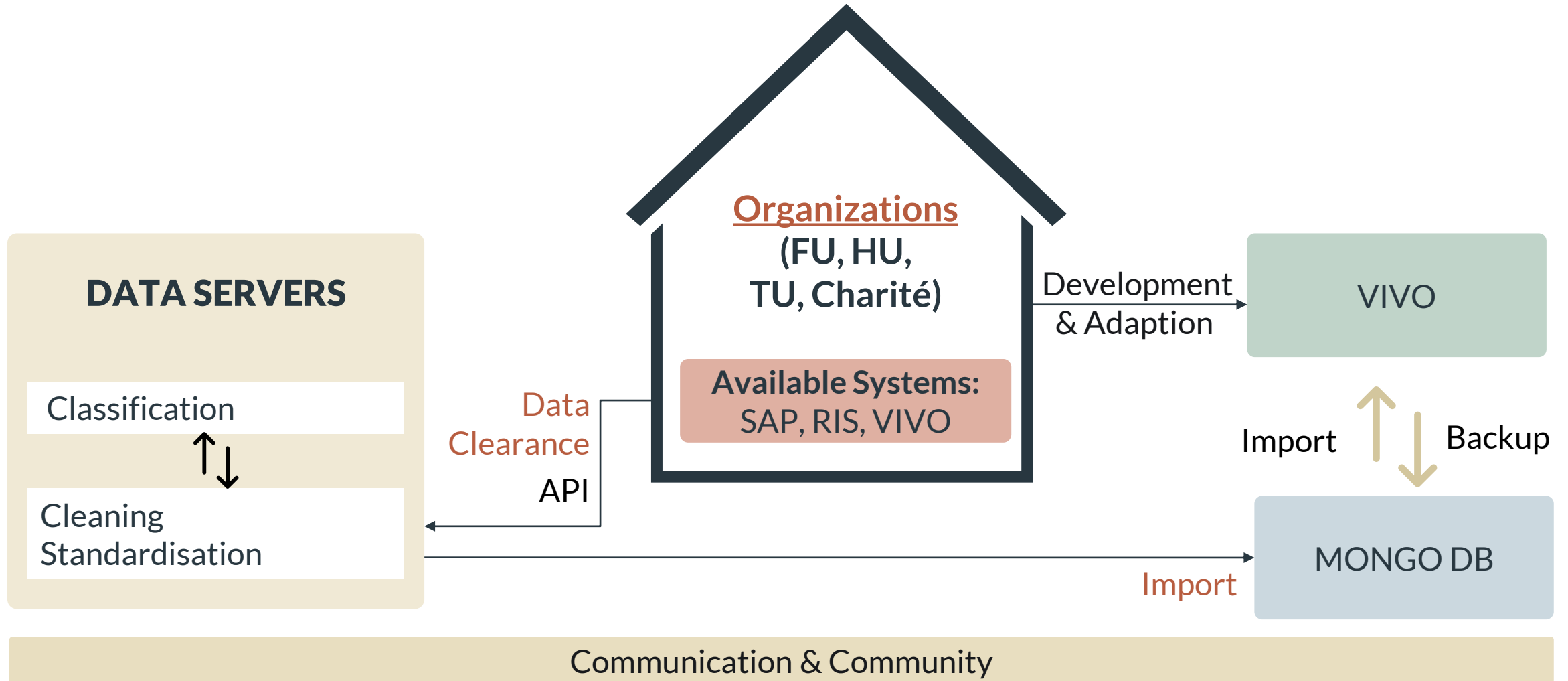
Matters of Activity Cluster of Excellence Cluster of Excellence

broader concept

Materials Science And Engineering Concept



# OVERVIEW





# CLASSIFICATION: CHALLENGE



Automatic text analysis to classify research entities using ML, LLM (Large Language Models)

- Analyze
- Categorize
- Link



Projects  
Publications

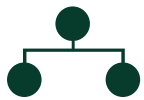




# CLASSIFICATION: VOCABULARIES



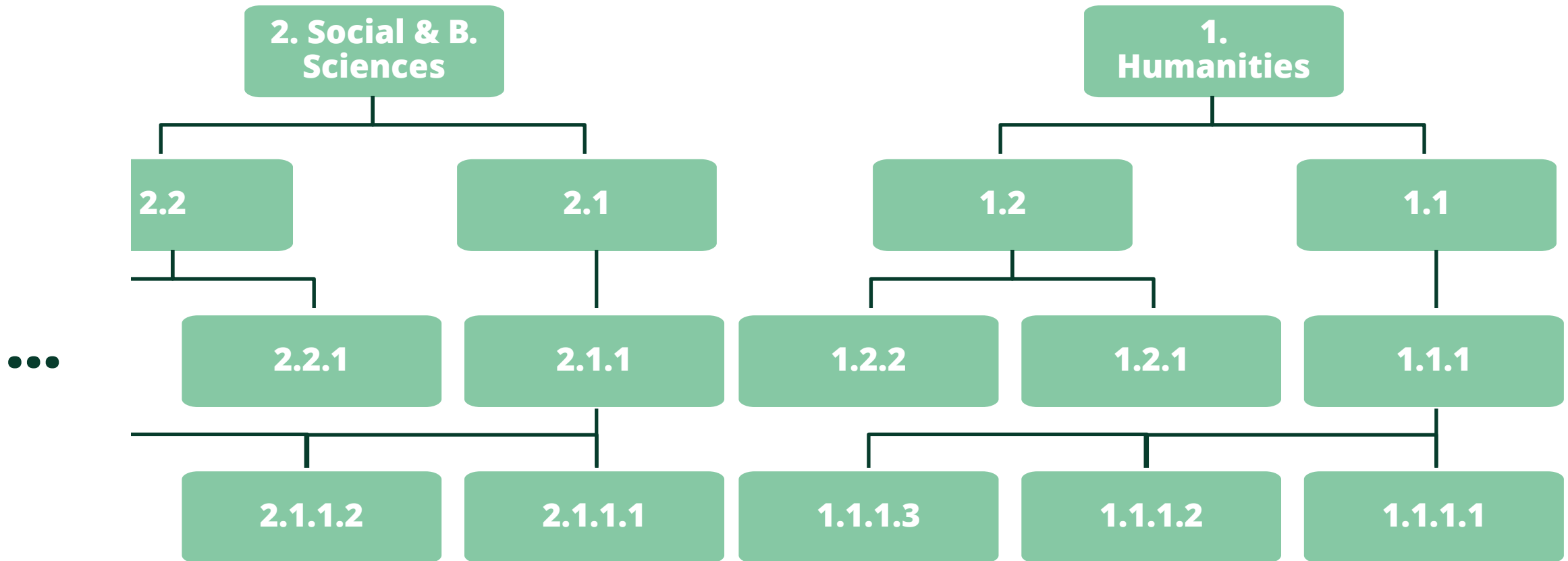
Disciplinary vocabulary  
Extend to interdisciplinary vocabulary



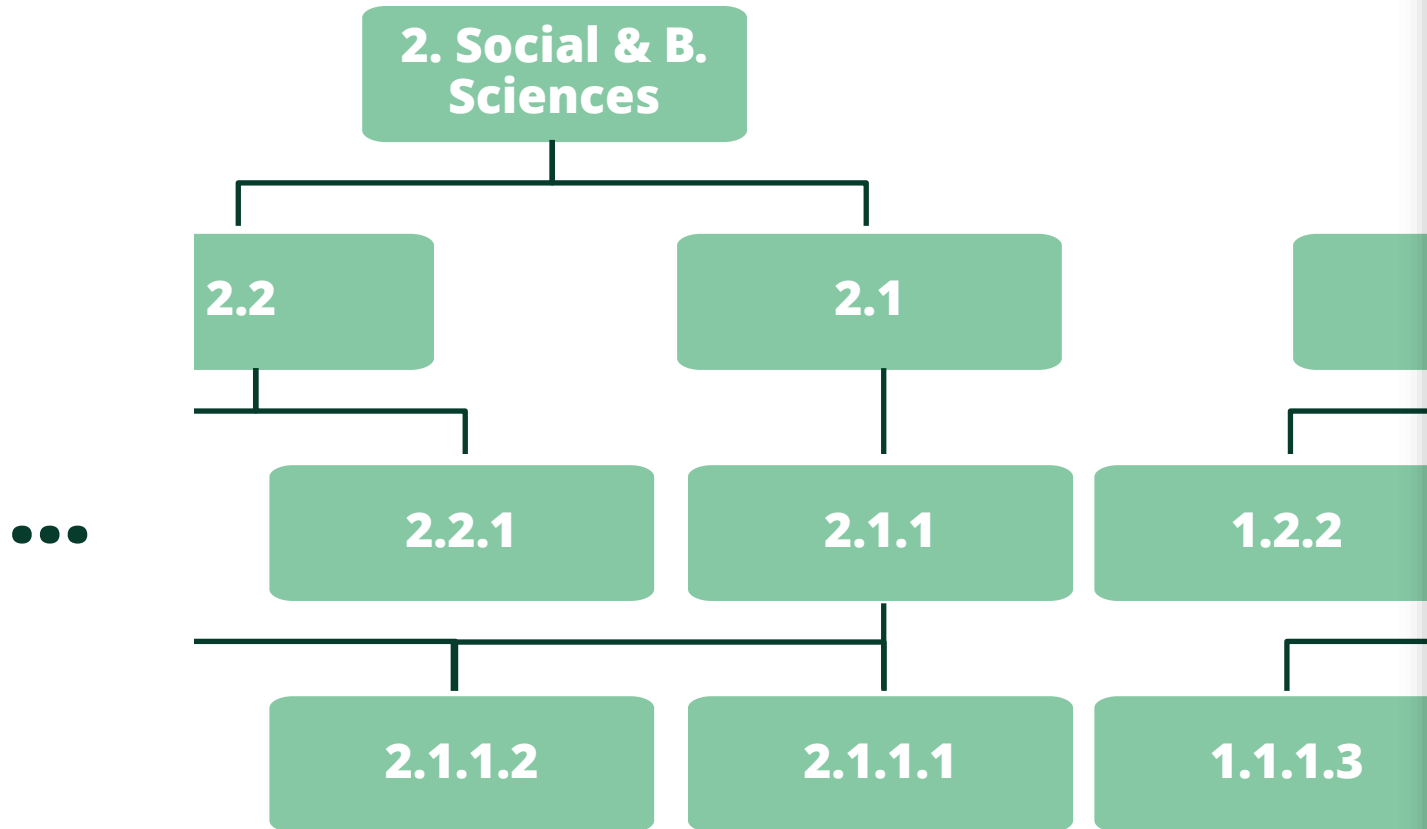
Disciplines [EUDAT-B2FIND](#)  
Interdisciplinary research fields [KDSE](#)



# DISCIPLINES [B2FIND, 336 CATEGORIES]



# DISCIPLINES [B2FIND, 336 CATEGORIES]



- - Engineering Science
    - + Computer Science, Electrical and System Engineering
    - + Construction Engineering and Architecture
    - + Materials Science and Engineering
    - + Mechanical and industrial Engineering
    - + Thermal Engineering/Process Engineering
  - - Humanities
    - + Ancient Cultures
    - + Fine Arts, Music, Theatre and Media Studies
    - + History
    - + Linguistics
    - + Literary Studies
    - + Philosophy
    - + Social and Cultural Anthropology
    - + Theology and Religion Studies
  - - Life Sciences
    - + Agriculture, Forestry, Horticulture and Veterinary Medicine
    - + Biology
    - + Medicine
  - - Natural Sciences
    - + Chemistry
    - + Mathematics
    - + Physics
  - - Social and Behavioural Sciences
    - + Economics
    - + Education Sciences
    - + Jurisprudence
    - + Psychology
    - + Social Sciences
- 
- + Work and Economy
  - + Earth and cosmos
  - + Globalization and sustainability
  - + Industry
  - + Information technology
  - + Infrastructure
  - + Cognition and knowledge
  - + Culture
  - + Life and wellbeing
  - + Materials
  - + People and Society
  - + Nature and environment
  - + Technology
  - + Science

# CLASSIFICATION: TRAINING DATA



20000 documents

HU [edoc-server](#)

- Unbalanced Corpus: some categories are amply represented, others not
- Heterogeneous, multilingual, voluminous, wide range of topics
- No reliable training corpus available



# APPROACH

## Pre-Classification



## Training and Testing



# KEYWORDS CREATION: CHATGTP

```
{  
  "id": "b2find:3.2.1",  
  "name_en": "Microbiology, Virology and Immunologie",  
  "name_de": "Mikrobiologie, Virologie Und Immunologie",  
  "keywords_de": [  
    {  
      "keyword": "Mikroorganismen",  
      "Wahrscheinlichkeit": 0.8  
    },  
    {  
      "keyword": "Viren",  
      "Wahrscheinlichkeit": 0.85  
    },  
    {  
      "keyword": "Immunsystem",  
      "Wahrscheinlichkeit": 0.83  
    }  
  ]  
}
```



Wir suchen keywords mit denen sich Texte kategorisieren lassen. Diese keywords sollen möglichst einzigartig pro kategorie sein, aber gleichzeitig allgemein genug, als das sie in möglichst vielen Texten einer kategorie vorkommen. verstehst du?



Ja, ich verstehe Ihre Anforderungen. Sie suchen nach Keywords, die Texte in bestimmten Kategorien eindeutig identifizieren können, aber gleichzeitig auch allgemein genug sind, um in vielen Texten derselben Kategorie vorkommen zu können.



gut.  
Bist du in der lage mir pro kategorie und keywords eine wahrscheinlichkeit zu geben, mit der das keyword auf die kategorie hindeutet?



Als Sprachmodell kann ich statistische Wahrscheinlichkeiten berechnen, aber ich benötige eine spezifische Datengrundlage, um genaue Prognosen zu machen. Wenn Sie mir eine große Menge an Texten in jeder Kategorie zur Verfügung stellen, kann ich diese analysieren und eine Wahrscheinlichkeit berechnen, mit der jedes Keyword auf die jeweilige Kategorie hindeutet.

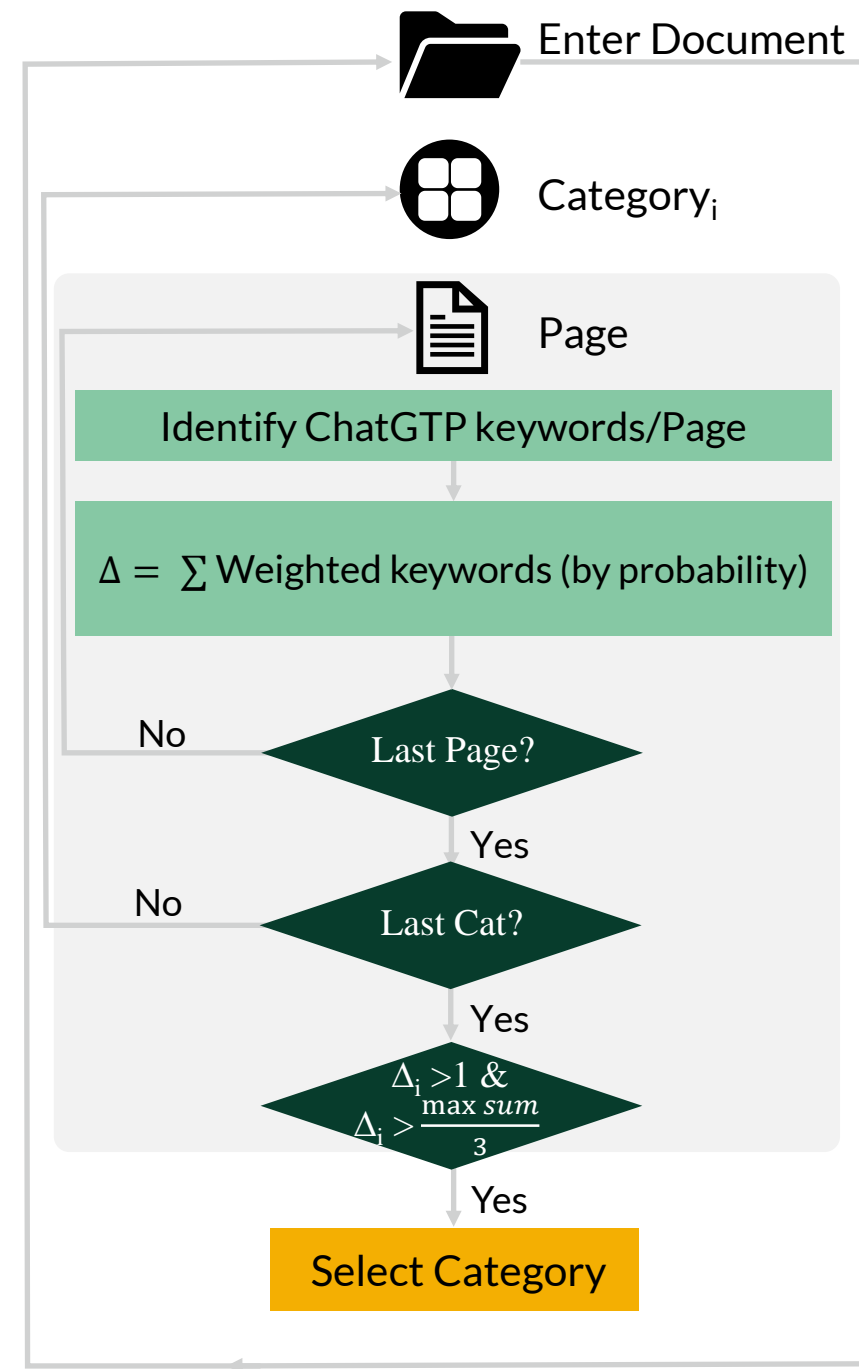
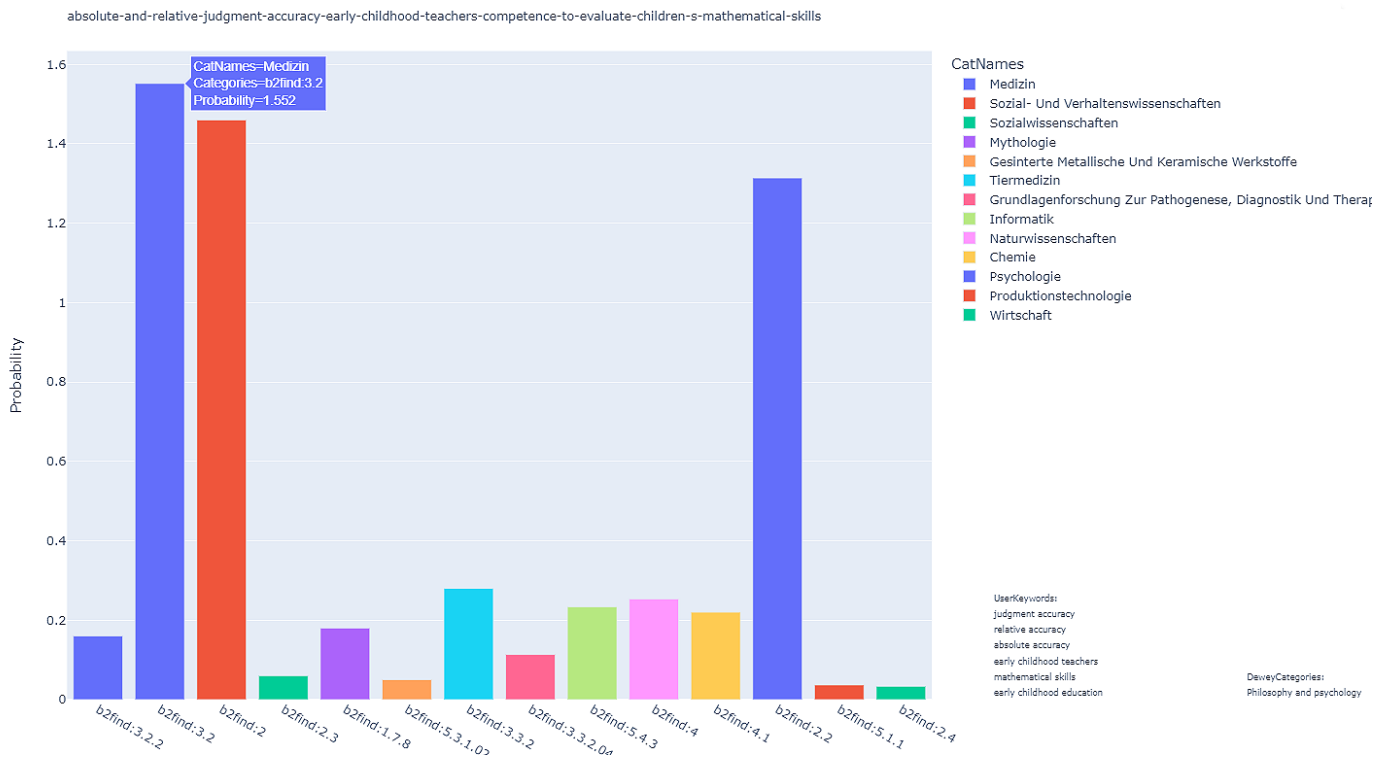


als sprachmodell benutzt du ähnlichkeitsvektoren. Kannst du diese vektoren als berechnungsgrundlage verwenden?



Ja, als Sprachmodell basiere ich auf der Verwendung von ähnlichkeitsbasierten Vektoren, die auch als "Word Embeddings" bezeichnet werden. Diese Vektoren können verwendet werden, um die semantische Ähnlichkeit zwischen Wörtern und Konzepten zu quantifizieren und können als Grundlage für die Berechnung von Wahrscheinlichkeiten dienen.

# PRECLASSIFICATION





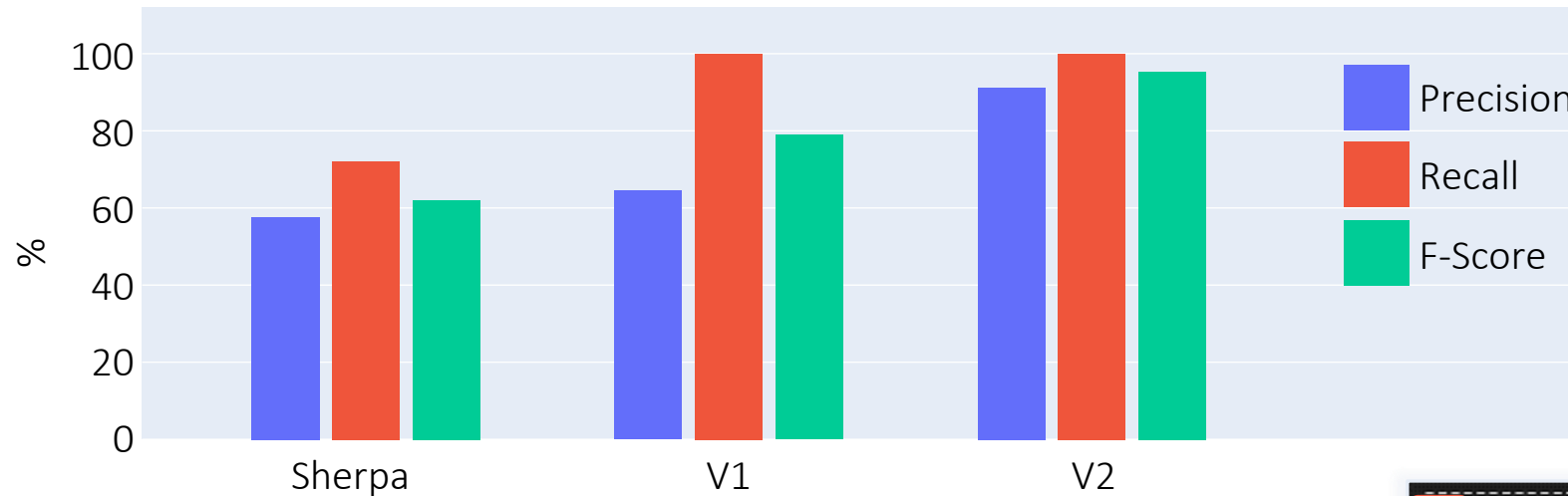
# CLASSIFICATION: RESULTS

EDOC with 9k Documents



# CLASSIFICATION: RESULTS

## EDOC with 19k Documents



```
-----  
V1: Exact Categories, no Hierarchies involved  
-----  
Precision: 0.651755164518536  
Recall: 0.9983792544570502  
F-Score: 0.7886616117546423  
-----  
V2: Hierarchy Version, upper Categorie Hits are counted as correct  
-----  
Precision: 0.9039978390059431  
Recall: 0.9983792544570502  
F-Score: 0.9488473043949578  
-----  
Sample Size: 617
```

# CLASSIFICATION: RESULTS - LMU DATASET

Ästhetisches Erkenntnisinteresse und literarische Artikulationsform.

	keywords	dewey	APIclass
0	Erkenntnis	100 Philosophie und Psychologie	Geisteswissenschaften
1	Interesse	100 Philosophie\n\n400 Sprache	Antike Kulturen
2	Ästhetik	400 Sprache\n\n700 Künste und Unterhaltung	Alte Geschichte
3	literarische Form	700 Künste\n\n800 Literatur	Geschichte
4	Jürgen Habermas	800 Literatur, Rhetorik, Literaturwissenschaft	Literaturwissenschaft
5	None	None	Literaturtheorie Und Vergleichende Literaturwi...
6	None	None	Philosophie
7	None	None	Geschichte Der Philosophie
8	None	None	Theoretische Philosophie

Re-imagining cities as ecosystems: environmental subject formation in Auckland and Mexico City.

	keywords	dewey	APIclass
0	Governmentality	300 Sozialwissenschaften	Biowissenschaften
1	environmentality	300 Sozialwissenschaft, Soziologie\n\n900 Ges...	Tierökologie, Biodiversität Und Ökosystemforsc...
2	neoliberal urbanization	910 Geografie, Reisen	Land- Und Forstwirtschaft, Gartenbau Und Veter...
3	urban governance	None	Forstwirtschaft
4	ecosystems	None	Tiermedizin
5	None	None	Tiernahrung Und Ernährungsphysiologie
6	None	None	Naturwissenschaften
7	None	None	Geowissenschaften
8	None	None	Biosphärenwissenschaften
9	None	None	Ökologie
10	None	None	Umweltforschung

Immune Landscape of Invasive Ductal Carcinoma Tumor Microenvironment Identifies a Prognostic

	keywords	dewey	APIclass
0	Immune landscape	500 Naturwissenschaften und Mathematik	Biowissenschaften
1	immune signature	540 Chemie\n\n500 Naturwissenschaften und Mat...	Biologie
2	survival	500 Naturwissenschaften	Medizin
3	invasive ductal carcinoma	None	Mikrobiologie, Virologie Und Immunolog
4	immune checkpoint inhibitor	None	Medizin

# CLASSIFICATION: RESULTS - LMU DATASET

Ästhetisches Erkenntnisinteresse und literarische Artikulationsform.

	keywords	dewey	APIclass
0	Erkenntnis	100 Philosophie und Psychologie	Geisteswissenschaften
1	Interesse	100 Philosophie\n\n400 Sprache	Antike Kulturen
2	Ästhetik	400 Sprache\n\n700 Künste und Unterhaltung	Alte Geschichte
3	literarische Form	700 Künste\n\n800 Literatur	Geschichte
4	Jürgen Habermas	800 Literatur, Rhetorik, Literaturwissenschaft	Literaturwissenschaft
5	None	None	Literaturtheorie Und Vergleichende Literaturwi...
6	None	None	Philosophie
7	None	None	Geschichte Der Philosophie
8	None	None	Theoretische Philosophie

Re-imagining cities as ecosystems: environmental subject formation in Auckland and Mexico City.

	keywords	dewey	APIclass
0	Governmentality	300 Sozialwissenschaften	Biowissenschaften
1	environmentality	300 Sozialwissenschaft, Soziologie\n\n900 Ges...	Tierökologie, Biodiversität Und Ökosystemforsc...
2	neoliberal urbanization	910 Geografie, Reisen	Land- Und Forstwirtschaft, Gartenbau Und Veter...
3	urban governance	None	Forstwirtschaft
4	ecosystems	None	Tiermedizin
5	None	None	Tiernahrung Und Ernährungsphysiologie
6	None	None	Naturwissenschaften
7	None	None	Geowissenschaften
8	None	None	Biosphärenwissenschaften
9	None	None	Ökologie
10	None	None	Umweltforschung

Immune Landscape of Invasive Ductal Carcinoma Tumor Microenvironment Identifi...

	keywords	dewey
0	Immune landscape	500 Naturwissenschaften und Mathematik
1	immune signature	540 Chemie\n\n500 Naturwissenschaften und Mat...
2	survival	500 Naturwissenschaften
3	invasive ductal carcinoma	None
4	immune checkpoint inhibitor	None

## Abstract

The constitution of environmental subjects by governments and civil society organizations around the world has taken place within a framework of neoliberal urbanization. This entails promoting an individual sense of responsibility over urban environments among city dwellers. The approach used is not so much governmentality as environmentality, because of its focus on environmental mat- ters. We claim that the tools used in this process are designed to generate among urban dwellers an imaginary of the cities they inhabit as ecosystems. Using qualitative methods, we examine cases in Auckland and Mexico City regarding water management.



# CLASSIFICATION: FINE TUNING

anthropogenic-disturbance-of-silt-mudflats-a-study-of-effects-on-populations-of-co...

- C Animal Ecology, Biodiversity and Ecosystem R...
- C Biogeography
- C Zoology

## Abstract

Anthropogenic disturbance is known to be highly influential on ecosystems. Sensitive systems ratio of specialized Fauna and Flora are often more affected than those inhabited with generalists. A sensitive ecosystem is the mudflats located off List on the island of Sylt. To investigate the disturbance of the mudflats, we examined populations of *Peringia ulvae* and *Corophium volutator*. In this article we show that already a one-time disturbance of the mudflats leads to a significant reduction of *P. ulvae*. No significant difference could be observed between the area disturbed multiple times and the control zone that was disturbed only once, indicating that a single disturbance has drastic effects on the population. Total numbers of *C. volutator* increased in zones that were disturbed compared to the control zone.

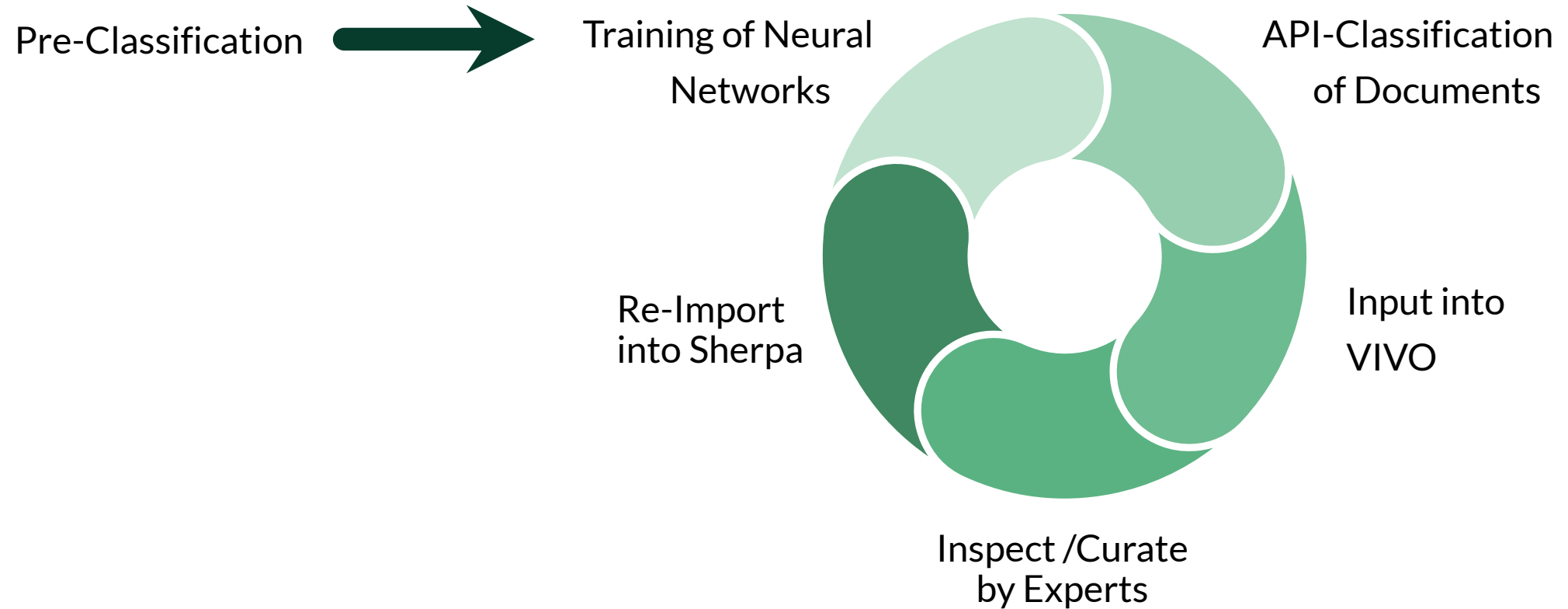
- Ionospheric physics
- Food Safety
- Basic Biolo. and Medic. Resea.
- Systems Engineering
- Theater
- Volcanology
- Cardiothoracic Surgery
- Ancient Cultures
- Roman Catholic Theology
- Theology and Religion Studies
- Geology

Analyzed Document

Delete unfitting categories

Manually add categories if required

# CLASSIFICATION: UPDATE CONCEPT



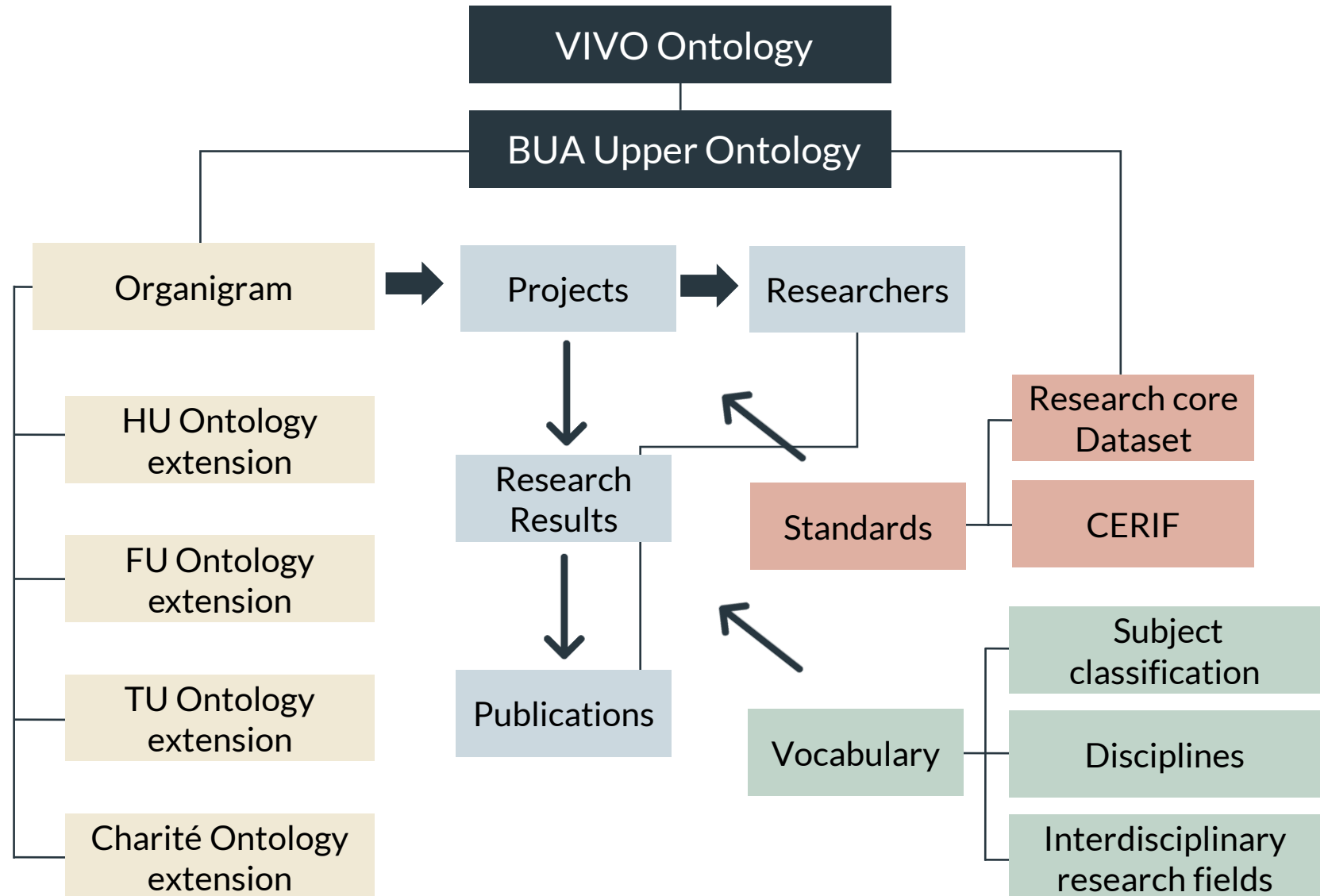
# ONTOLOGIES

Formal representation of concepts within a specific domain

Semantically and structurally relating research entities

## Vocabularies

Disambiguate ontology entities to ensure consistent understanding across different Organizational structures





# ONTOLOGIES AND VOCABULARIES

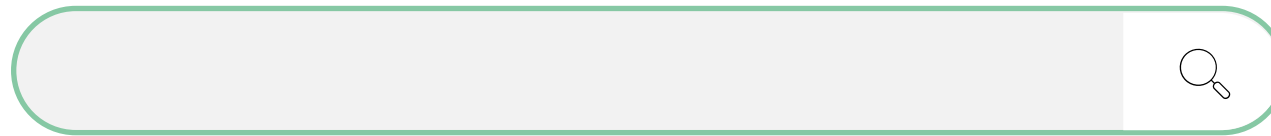
- BUA-Organigrams as Ontologies
  - Published on [GitHub](#)
- VIVO Ontology extensions for the German academic domain
  - Published on [GitHub](#)
- Vocabulary for the classification
  - Research Disciplines
    - [EUDAT-B2FIND](#), on [GitHub](#)
  - Academic Subjects
    - [DESTATIS](#), auf [GitHub](#)
  - Interdisciplinary research fields
    - [KDSF](#), on [GitHub](#)

- - Engineering Science
  - + Computer Science, Electrical and System Engineering
  - + Construction Engineering and Architecture
  - + Materials Science and Engineering
  - + Mechanical and industrial Engineering
  - + Thermal Engineering/Process Engineering
- - Humanities
  - + Ancient Cultures
  - + Fine Arts, Music, Theatre and Media Studies
  - + History
  - + Linguistics
  - + Literary Studies
  - + Philosophy
  - + Social and Cultural Anthropology
  - + Theology and Religion Studies
- - Life Sciences
  - + Agriculture, Forestry, Horticulture and Veterinary Medicine
  - + Biology
  - + Medicine
- - Natural Sciences
  - + Chemistry
  - + Mathematics
  - + Physics
- - Social and Behavioural Sciences
  - + Economics
  - + Education Sciences
  - + Jurisprudence
  - + Psychology
  - + Social Sciences
- + Work and Economy
- + Earth and cosmos
- + Globalization and sustainability
- + Industry
- + Information technology
- + Infrastructure
- + Cognition and knowledge
- + Culture
- + Life and wellbeing
- + Materials
- + People and Society
- + Nature and environment
- + Technology
- + Science

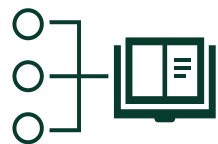


# ONTOLOGIES: CONTEXT

Query expansion and refinement  
Expanding contextual understanding  
Faceted search and navigation



For a given institutional unit, find research projects of other units, which have addressed the same research problems



Based on the categorization of my research results, find matching research results from other institutions, which are categorized within the same research field(s)



For a given funder, find other funded projects, coordinating institution and project objectives, which are categorized under interdisciplinary research

# CLASSIFICATION: SUMMARY

**“Every Technology has its limits...”**

- Vocabularies are made for Humans
  - Overlapping Categories
  - Bundled Categories
- Accuracy, Validity and Relevance of Pre-Classification
  - GPT as a black hole
- Sufficient documents for each category
  - Limiting Datasets

**“... and at the limit awaits the human”**

# ACKNOWLEDGEMENT



**THANK YOU!**

