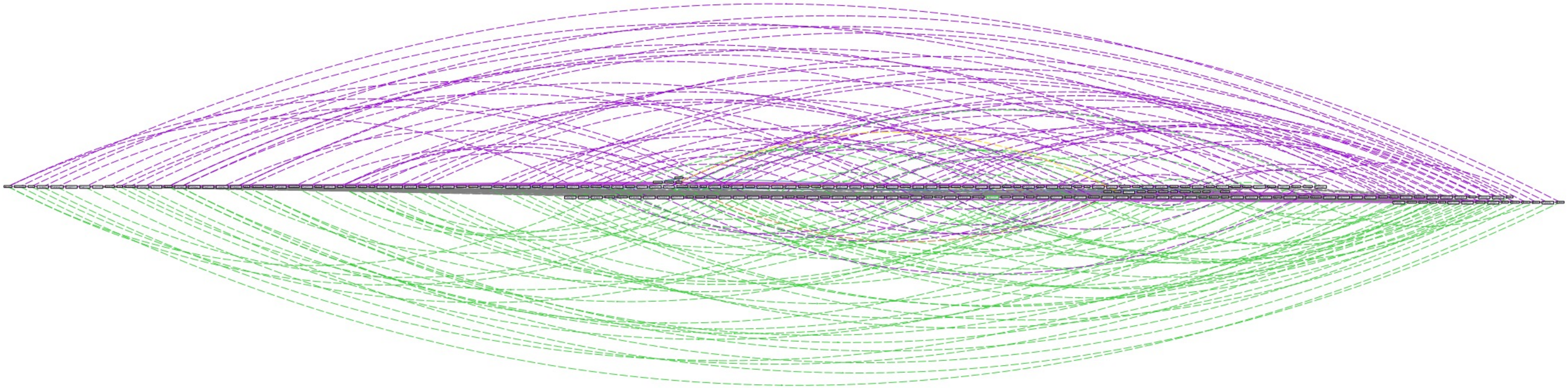
Data Collections Explorer

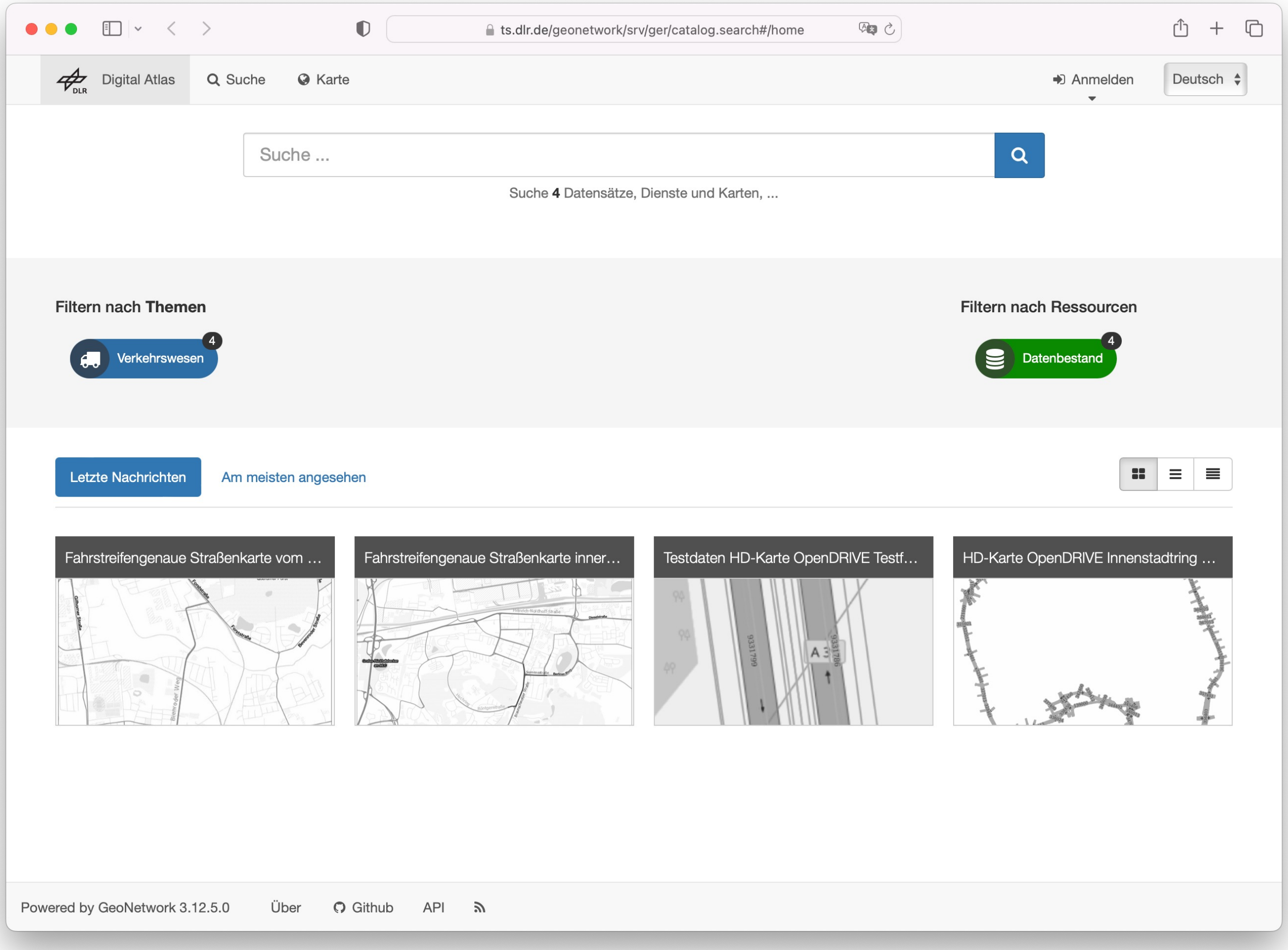
An easy-to-use tool for sharing and discovering research data

Philipp Ost, Yusra Shakeel, Philipp Tögel – 1st CoRDI, Karlsruhe, 12.–14. September 2023



Motivation – How to share datasets?

- DLR Digital Atlas
  - Hosted and curated by DLR
  - Easy sharing of data with collaboration partners



Suche **4** Datensätze, Dienste und Karten, ...

Filtern nach Themen

Filtern nach Ressourcen

Fahrstreifengenaue Straßenkarte vom ...

Fahrstreifengenaue Straßenkarte inner...

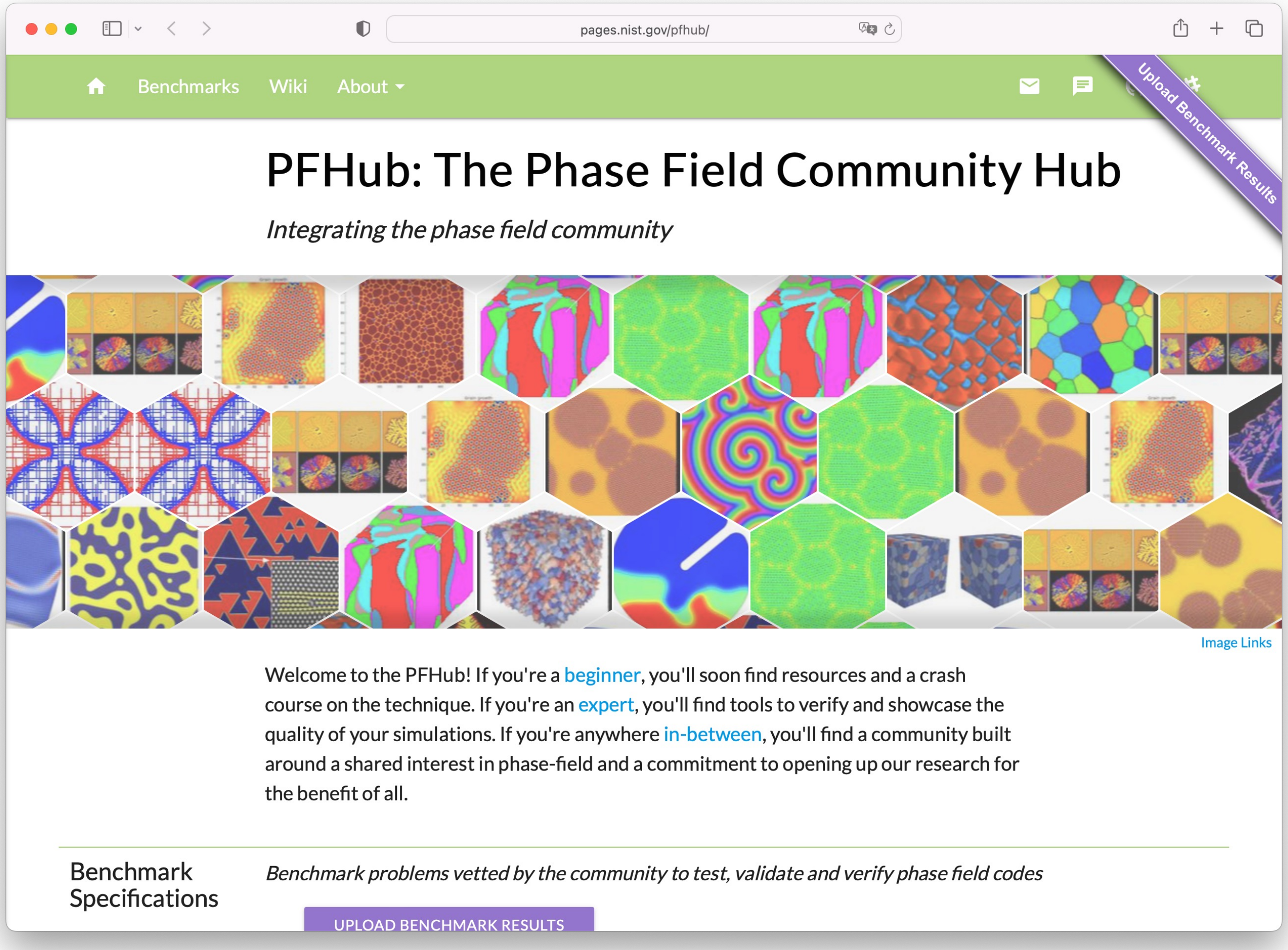
Testdaten HD-Karte OpenDRIVE Testf...

HD-Karte OpenDRIVE Innenstadtring ...

Powered by GeoNetwork 3.12.5.0

Motivation – How to serve a community?

- PFHub: The Phase Field Community Hub
  - Hosted by NIST
  - Curated by the phase field community



PFHub: The Phase Field Community Hub

Integrating the phase field community

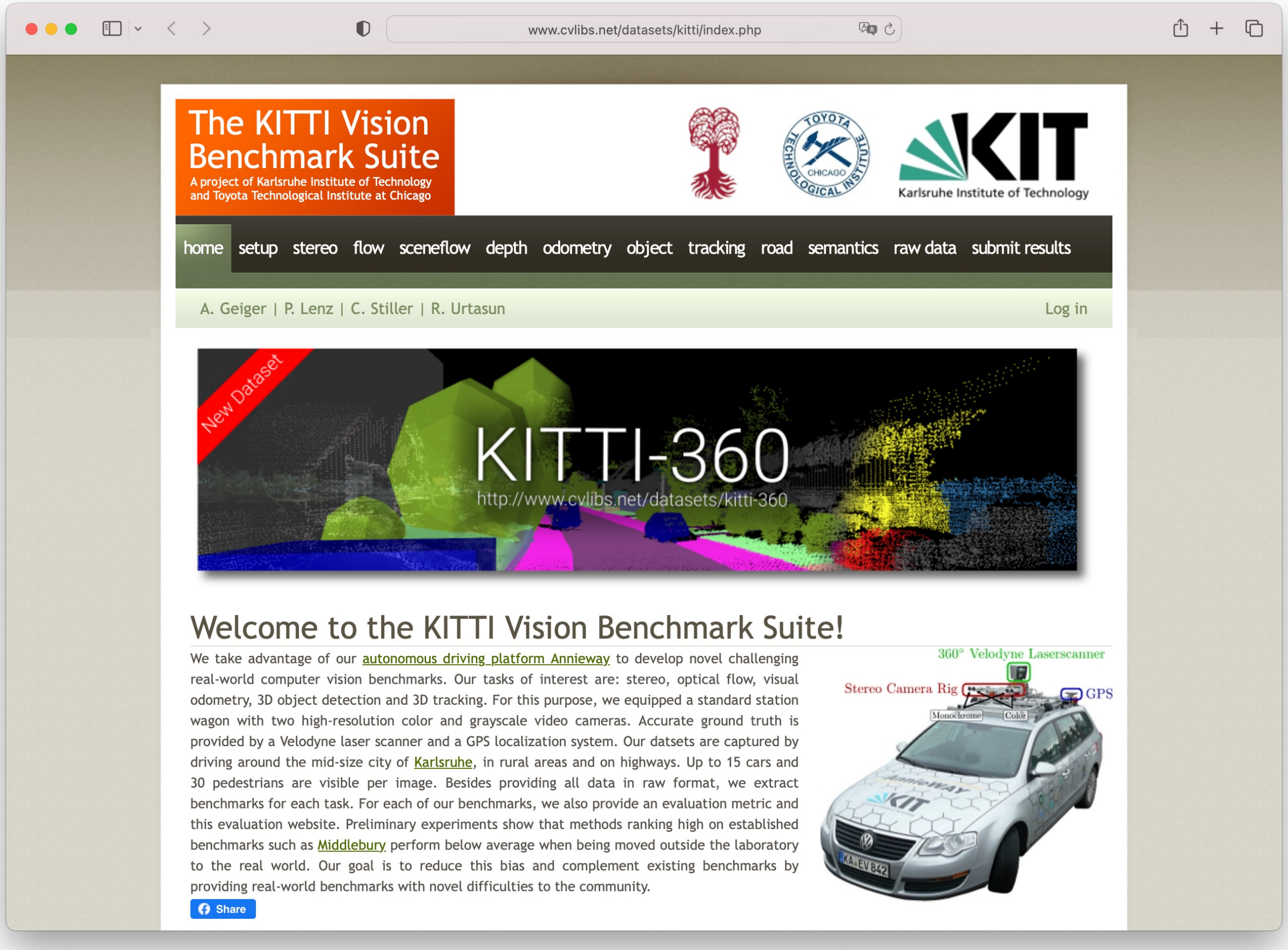
Welcome to the PFHub! If you're a beginner, you'll soon find resources and a crash course on the technique. If you're an expert, you'll find tools to verify and showcase the quality of your simulations. If you're anywhere in-between, you'll find a community built around a shared interest in phase-field and a commitment to opening up our research for the benefit of all.

Benchmark Specifications

Benchmark problems vetted by the community to test, validate and verify phase field codes

Motivation – How to share a dataset?

- KITTI Vision Benchmark Suite
  - Hosted by responsible research group
  - Data access requires an account
  - Usage restrictions apply


www.cvlibs.net/datasets/kitti/index.php
The KITTI Vision Benchmark Suite
A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago
Karlsruhe Institute of Technology
home
setup
stereo
flow
sceneflow
depth
odometry
object
tracking
road
semantics
raw data
submit results
A. Geiger | P. Lenz | C. Stiller | R. Urtasun
Log in
New Dataset
KITTI-360
http://www.cvlibs.net/datasets/kitti-360
360° Velodyne Laserscanner
Stereo Camera Rig
GPS
Monochrome
Color
Share

Welcome to the KITTI Vision Benchmark Suite!

We take advantage of our autonomous driving platform Annieway to develop novel challenging real-world computer vision benchmarks. Our tasks of interest are: stereo, optical flow, visual odometry, 3D object detection and 3D tracking. For this purpose, we equipped a standard station wagon with two high-resolution color and grayscale video cameras. Accurate ground truth is provided by a Velodyne laser scanner and a GPS localization system. Our datsets are captured by driving around the mid-size city of Karlsruhe, in rural areas and on highways. Up to 15 cars and 30 pedestrians are visible per image. Besides providing all data in raw format, we extract benchmarks for each task. For each of our benchmarks, we also provide an evaluation metric and this evaluation website. Preliminary experiments show that methods ranking high on established benchmarks such as Middlebury perform below average when being moved outside the laboratory to the real world. Our goal is to reduce this bias and complement existing benchmarks by providing real-world benchmarks with novel difficulties to the community.

Easy Sharing of Research Data

- What do these examples have in common?
  - Allow relatively easy sharing of research data
  - Allow working with/on the data as part of a community
  - None of them are registered in re3data (as of 30.08.’23)
    - How does one find them?
    - There is a certain amount of “insider knowledge” required to know where to look

- How to share
  - … interim results before publication, possibly with a wider audience?
  - … data and related results with project partners?
  - … data jointly produced/curated/… by a group of researchers?

Current State of the Art

- re3data – Repository of Research Data Repositories
  - "Gold standard" for research data repositories
  - Admits only quality controlled repositories

- Community specific initiative: Data Repository Finder
  - Geared towards the Life Sciences
  - Overview of repositories
  - "…helps researchers find data repositories where they can share data…"[1]

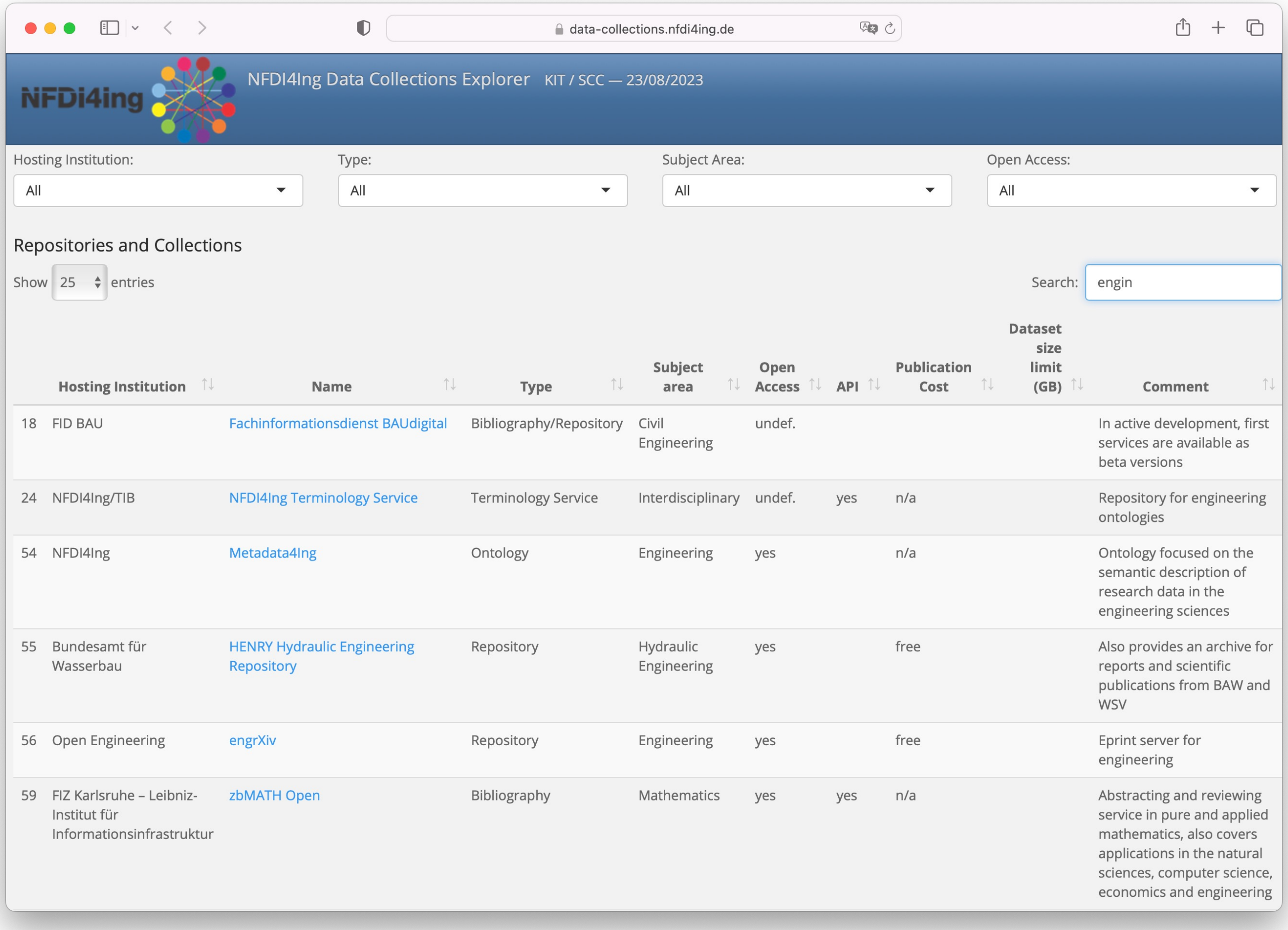
[1] https://data-repository-finder.ll.mit.edu

Data Collections Explorer – Use Cases

- Typical use cases considered during design and implementation:
  - Scientists looking for repositories to publish their research data
    - Example: PhD students create data sets as part of their research
      - Where to publish them?
      - Are there size limits or costs involved?
  - Scientists searching for data sets
    - Example: An engineer is interested in material properties
      - Are there other materials that might fit the requirements?

Data Collections Explorer – Current Version

- Successful proof-of-concept
  - Started out with 38 entries
  - Continuously updated
    - Currently at 87 entries
- Access it at data-collections.nfdi4ing.de



NFDI4Ing Data Collections Explorer KIT / SCC — 23/08/2023

Hosting Institution: All

Type: All

Subject Area: All

Open Access: All

Repositories and Collections

Show 25 entries

Search: engin

	Hosting Institution	Name	Type	Subject area	Open Access	API	Publication Cost	Dataset size limit (GB)	Comment
18	FID BAU	Fachinformationsdienst BAUdigital	Bibliography/Repository	Civil Engineering	undef.				In active development, first services are available as beta versions
24	NFDI4Ing/TIB	NFDI4Ing Terminology Service	Terminology Service	Interdisciplinary	undef.	yes	n/a		Repository for engineering ontologies
54	NFDI4Ing	Metadata4Ing	Ontology	Engineering	yes		n/a		Ontology focused on the semantic description of research data in the engineering sciences
55	Bundesamt für Wasserbau	HENRY Hydraulic Engineering Repository	Repository	Hydraulic Engineering	yes		free		Also provides an archive for reports and scientific publications from BAW and WSV
56	Open Engineering	engrXiv	Repository	Engineering	yes		free		Eprint server for engineering
59	FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur	zbMATH Open	Bibliography	Mathematics	yes	yes	n/a		Abstracting and reviewing service in pure and applied mathematics, also covers applications in the natural sciences, computer science, economics and engineering

Data Collections Explorer

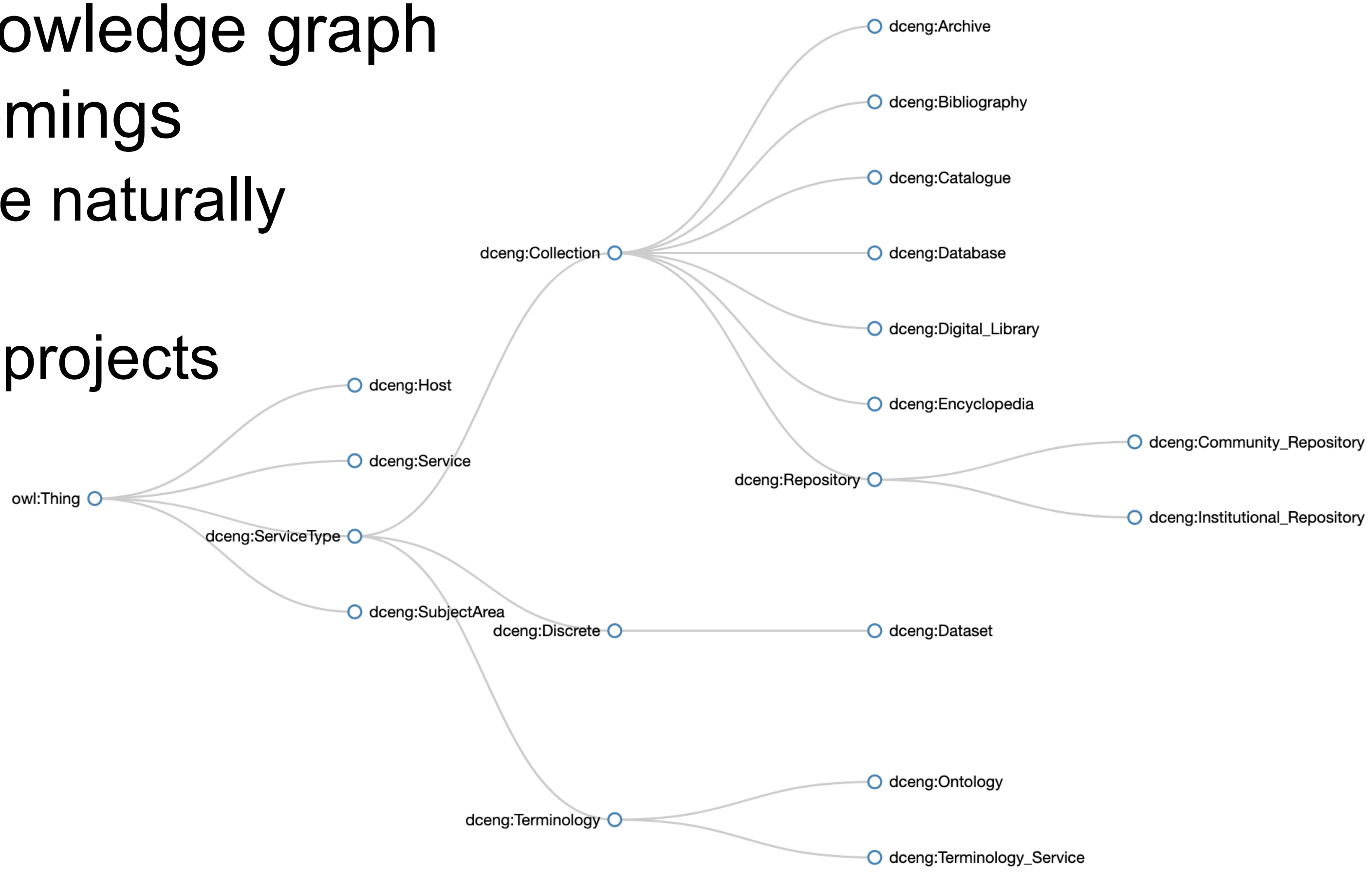
- Basic implementation completed and available
  - CSV table to store entries
  - Served as HTML table using R Markdown and shiny
- Code is available on GitHub: github.com/kit-data-manager/Data-Collections-Explorer
  - Adapt it, use it for your own community
- Constant updates
  - Mostly triggered by input from scientists
- Nonetheless, some caveats apply…

Collecting Scientists’ Feedback

- Feedback collected from all NFDI4Ing Task Areas
  - Questions of interest:
    - What are scientists requirements and wishes of scientists regarding the Data Collections Explorer?
    - How would they like to work with the Data Collections Explorer?
    - Which repositories or data sets are important for their daily work?
  - Aggregated results:
    - Easy-to-use interfaces, including submission of new entries
    - API access
    - Flexible search
    - Use of vocabularies
- Discussions take time, but provide valuable input

Data Collections Explorer – Knowledge Graph

- Replace CSV table with a knowledge graph
- Addresses almost all shortcomings
  - One-to-many mappings come naturally
  - Access via SPARQL
  - Easier integration with other projects
  - More flexibility
  - Easy machine accessibility



Conclusions

- The Data Collections Explorer is an easy-to-use flexible tool
  - Share and find data focussed on a dedicated community
  - No big entry hurdles compared to re3data
  - Built for NFDI4Ing, easily adaptable for other communities

- New graph based version is work in progress; what to look forward to:
  - API access
  - New user interface
  - Vocabulary integration
  - And more… ☺