Workshop on PIDs within NFDI

Report of the Working Group “Persistent Identifiers (PID)” of the Section *Common Infrastructures* of the NFDI

Imprint

Authors:

Daniel Arend, Janete Bach, Kirsten Elger, Sandra Göller, Stephanie Hagemann-Wilholt, Rolf Krahl, Matthias Lange, David Linke, Desiree Mayer, Peter Mutschke, Lorenz Reimer, Markus Scheidgen, Antonia C. Schrader, Michael Selzer, Philipp Wieder

As of

13.02.2023. Version. 1.0

Table of Content

Abstract

In order to gain an overview of the current state of the discussion on PIDs and for the identification of use cases for the initiation phase of a PID service within the NFDI basic services, the working group Persistent Identifier of the Section *Common Infrastructures* of the NFDI hosted an online workshop in January 2023. In the course of the workshop, members of nine different NFDI consortia presented the current application of PIDs in their consortia.

Introduction

In recent years, persistent identifiers (PIDs) have been widely accepted in the general scientific community to identify research assets like data objects, general research outputs, or the researchers themselves. Because of this, members of all NFDI consortia use different kinds of PIDs already in their everyday work. As of today the assignment of identifiers in all NFDI consortia is scattered and heterogeneous in terms of actors, services, scope, quality, and costs involved. Having the core function of PIDs in RDM in mind, an analysis of existing gaps and the development of joint solutions in order to serve the needs of the individual communities and the NFDI as a whole are needed.
For these reasons the working group Persistent Identifier within the Section *Common Infrastructures* of the NFDI, with representatives of existing PID services and infrastructures who are already involved in various domain-specific consortia, was initiated. The aim of the working group is to develop a common strategy for the implementation and extension of community-based and broadly used PID services that is closely aligned with the needs of NFDI consortia. This includes the necessary technical & organisational measures as well as the training and educational aspects particularly around standardised and complete metadata ensuring compliance with the FAIR principles.

In order to get a joint overview of the current state of the discussion on PIDs and, as examples, their current applications in the NFDI consortia, an online workshop with PID experts from the NFDI consortia and the working group members was conducted on January 27, 2023, from 09 to 11 a.m. The results of the workshop also served as a basis for the planned application for PID services in the NFDI and for the identification of use cases for the initiation phase of a PID service within the NFDI basic services.

After an introductory presentation on the motivation and objectives of the working group and the planned basic services, members of the following nine NFDI consortia presented the current application of PIDs in their consortia:

- Daphne4NFDI
- FAIRagro
- FAIRmat
- KonsortSWD
- NFDI4Cat
- NFDI4Earth

- NFDI4Ing
- NFDI4Microbiota
- NFDI4Culture & RADAR4Culture

The presentation slot was followed by an open question and discussion slot. Leading questions were:

- What are the developments in the different NFDI communities?
- How can we learn mutual?
- What do I want from the Persistent Identifier working group?
- What are my wishes for Base4NFDI?
- What are the next steps?

The present report documents key findings and discussion points.

Tab. 1: Program of the Workshop, 27.01.2023

Time	**Agenda**	**Speaker**
09:00-09:15	Welcoming and Introduction	Philipp Wieder (GWGD)
09:15-09:20	PIDs for data repository entries, datasets, and schemas (FAIRmat)	Markus Scheidgen (HU Berlin, NOMAD)
09:20-09:25	PIDs for Instruments (NFDI4Ing)	Michael Selzer (KIT)
09:25-09:30	PIDs for Samples (Daphne4NFDI)	Rolf Krahl (HZB)
09:30-09:35	IGSN (NFDI4Earth)	Kirsten Elger (GFZ)
09:35-09:40	Survey variables (KonsortSWD)	Janete Bach (GESIS) Peter Mutschke (GESIS)
09:40-09:45	*Break*	
09:45-09:55	PIDs for Catalysis (NFDI4Cat)	David Linke (Leibniz-Institut für Katalyse e.V.)

09:55-10:00	StrainInfo PIDs for microbial strains (NFDI4Microbiota)	Lorenz Reimer (DSMZ)
10:00-10:05	PIDs for Cultural Data (NFDI4Culture & RADAR4Culture)	Desiree Mayer (SLUB Dresden) Sandra Göller (FIZ Karlsruhe)
10:05-10:10	Digital twins of plant genetic resources (FAIRagro)	Matthias Lange (IPK) Daniel Arend (IPK)
10:10-10:15	Summary	Britta Dreyer (TIB)
10:15-11:00	Questions & Discussion	Moderated by Philipp Wieder

Presentations

PIDs for data repository entries, datasets, and schemas (FAIRmat)

We currently provide DOIs for datasets to our users on request. In the future, we want to add handle-based PIDs on all our entries (~10 million) automatically. We are also interested in PIDs for users, schemas and data-type definitions, samples, instruments, and more. However, we would prefer a singular service and API with abstract metadata over a pluralism of different PID types. The PID metadata should inform about the type/schema and leave the domain specific metadata to the identified (meta-)data.

PIDs for Instruments (NFDI4Ing)

In the context of NFDI4Ing the KIT plan to expand it's "Gerätepool" (means instrument data base) and to implement PIDs for uniquely identifying the scientific instruments included.

PIDs for Samples (Daphne4NFDI)

DAPHNE stands for DAta for PHoton and Neutron Experiments. The project focuses on research with photons and neutrons at large-scale research facilities. As distinguished from most other NFDI consortia, DAPHNE's scope is not a particular field of research, but rather a

class of experimental methods. The users of photon and neutron facilities come from almost all areas of research.

As a result, a huge variety of samples need to be considered. To illustrate this with at least a few examples: we have simple material samples, such as single crystals, crystal powders, or novel functional materials. But we also have complex structures, such as solar cells or batteries. An important case are macromolecular crystals, e.g. proteins. But we also have archeological artefacts, fossils, artwork, e.g. paintings, and even living plants. A special case are “ephemeral samples”: samples that are created, measured and destroyed during one single experiment, such as a liquid jet injected in the measurement chamber.

The accurate description of the sample being measured is an important part in the metadata for the resulting dataset. In many cases, the full history of sample preparation steps constitutes such a description. Assuming that each preparation and characterization step in turn generates a FAIR dataset, what we actually need is a sample PID that ties all these datasets together and allows us to track the sample's history.

Note that the lab that created the sample may be a different institution than the facility performing the measurement. Therefore, an internal sample id of the lab will not be enough. Furthermore the sample may be held in a collection outside the context of the experiment. So we need to be able to integrate existing external IDs.

Samples may be complex structures composed out of individual samples, each having their own history. On the other hand, a sample may be cut into pieces, each part being treated differently in subsequent steps. Therefore a parent / child relation is needed between sample PIDs.

How we should deal with ephemeral samples is still an open discussion.

IGSN (NFDI4Earth)

The International Generic Sample Number (IGSN) is a globally unique, resolvable persistent identifier for physical objects originally developed for earth sciences samples. IGSN identifier link to sample descriptions on the internet supporting sample discovery and enabling the connection between datasets, publications and the originating samples.

Since 2012, IGSNs have been registered as Handles (one namespace: 10273, internal organization managed by adding centrally-managed agent-specific namespaces). IGSN is an addition to the original sample name and shall not replace it. If cited in research articles, IGSNs have the possibility to “outsource” detailed sample descriptions from the manuscript texts. IGSN is already an established PID system with >10 M. registered samples and managed by IGSN e.V., a global, non-profit organization. Recent activities of IGSN e.V. reach out beyond the Geosciences domain (e.g. life sciences, archaeology, Daphne4NFDI).

In October 2021, DataCite and IGSN announced their strategic partnership and from January 2023, IGSNs are registered as DataCite DOIs. Consequently, each DataCite member can assign a DOI-IGSN. IGSN has modular and quite rich metadata with a “common kernel” ideally provided for all samples which can be further customized to serve more subdisciplines or disciplines. There are agreed recommendations to map IGSN metadata to the DataCite Schema, however, only using the DataCite schema (that has been developed for different purposes) would “only” represent a summary of the sample description, because the DataCite Schema is much more generic that IGSN metadata; we still need to find out a convenient model for not losing the richness of the IGSN metadata and sample descriptions in the new world of DataCite DOI-IGSNs.

Links and further information on IGSN:

- IGSN e.V: www.igsn.org
- IGSN GitHub at https://igsn.github.io
- IGSN-related project (2022/2023): FAIR WISH - FAIR Workflows to establish IGSN for Samples in the Helmholtz Association - (i) expanding domain-specific metadata for different geo-bio samples, (ii) development of a sample description template. All project output available in the FAIR WISH Zenodo Community: https://zenodo.org/communities/fair_wish
- Latest publication about IGSN: Klump, J., Lehnert, K., Ulbricht, D., Devaraju, A., Elger, K., Fleischer, D., Ramdeen, S. and Wyborn, L., 2021. Towards Globally Unique Identification of Physical Samples: Governance and Technical Implementation of the IGSN Global Sample Number. Data Science Journal, 20(1), p.33. DOI: http://doi.org/10.5334/dsj-2021-033

Survey variables (KonsortSWD)

Referencing research data and there inherit detailed entities supports FAIR usage. We enhance the state of the art of citing research data, by developing an infrastructure to reference detailed attributes, here initially variables, within such data. By assigning PIDs to these attributes, individual elements of the data files can be referenced and retrieved with the required metadata for machine-actionable and human access. The PIDs will not only enable citeability within scientific papers but also give access for processing the contained data itself, e.g., within script languages like R or python. We provide recommendations for assigning Persistent Identifiers (PIDs) below the study level. The recommendations rely on use case partners' data types and services. However, it would also benefit other institutions, such as data repositories that hold tabular data and want to take advantage of uniquely

identifying their data at a lower granularity level. To this end, the da|ra PID registration service was enlarged under the KonsortSWD Measure TA.5-M.1.

PIDs for Catalysis (NFDI4Cat)

In catalysis samples prepared by machines or in a lab play an important role. These samples are often difficult to reproduce even when they are prepared by the same protocol. Therefore, tracking the materials used and knowing for which sample instance catalytic or characterisation measurements were carried out is of huge importance in the community (as confirmed by a survey). Often the work on one sample/material is moreover distributed across labs and institutions which would also benefit from a wider use of PIDs.
Early 2022 NFDI4Cat created an internal working group on PIDs with the goal to make PIDs available to catalysis researchers from all institutions and possibly also to industrial partners which contribute materials to catalysis research. One goal is to make PIDs available “from the beginning”, that is for the synthesis in the lab. No relabelling or renaming should be necessary until publication in the NFDI4Cat RDM portal. We have been developing a handle-based service as part of our shared infrastructure that can be interfaced from local RDM tools like ELNs (electronic lab notebooks) and the central NFDI4Cat repository that is being built. Besides for samples the PIDs will also be useful for other resources applied in the research process such as materials, devices, instruments, models, schemas etc. A layered metadata schema is suggested similar to IGSN that also reflects confidentiality requirements.

StrainInfo PIDs for microbial strains (NFDI4Microbiota)

Within NFDI4Microbiota the database StrainInfo (https://straininfo.dsmz.de/) is developed. Straininfo is a database for collecting, matching and providing persistent identifiers for microbial strains. In the field of bacteria and archaea every newly cultivated and described species needs to be deposited as type strain in at least two international culture collections. Each collection is identifying strains with their own stable identifier that is recognizable. Through time, bacterial strains are assigned with multiple identifiers, which again are used in literature and databases to refer to used strains. To resolve if two identifiers from two collections point to the same strain can be cumbersome. Therefore the StrainInfo database collects and matches strain identifiers on a large scale, which is not limited to culture collection numbers but also includes sequence accession number and DOIs. In the long run a new central registry for microbial strains should be established, that offers persistent

identifiers independent from culture collections and thereby can be applied to any microbial strain.

PIDs for Cultural Data (NFDI4Culture & RADAR4Culture)

In NFDI4Culture, authority data IDs are the most widely used PIDs for research data. ORCID, ROR, geoNames and also wikidata are quite commonly used.
To meet these special needs in the 4Culture-Community, RADAR4Culture, the publication service for NFDI4Culture from FIZ Karlsruhe has implemented the possibility to enter GND-IDs for subject headings in the metadata-field “keywords”. This is an important step towards FAIR publications and the linkage between research data. Current discussions are about more controlled fields in metadata schemas of repositories that should require PIDs, the use of PIDs for different data types, e.g. PIDs for complex search queries (with timestamps), DOIs for annotations or specific elements of a publication, such as single entries in work catalogs, and the need of stable URIs for vocabularies and their provision to make them mappable to larger terminologies.

More information:

- About RADAR: https://radar.products.fiz-karlsruhe.de/en/radarabout/ueber-radar
- Event of NFDI4Culture: https://nfdi4culture.de/news-events/events/forum-data-publication-and-availability-5-persistent-identifiers.html

PIDs for digital twins of plant genetic resources (FAIRagro)

Agriculture is facing increasing challenges, e.g., growing demand for food with stagnant productivity, climate change, biodiversity loss, and natural resource degradation. One pillar is plant breeding, which can in principle make use of a big data ecosystem (Tian et al. 2021). Around collections and gene banks to preserve PGRFA, the volume of data potentially useful for plant breeding is rapidly increasing due to technological advancements in genome analysis, precision phenotyping, and digitization in the plant breeding value chain. Data is generated, stored, and analyzed by a variety of stakeholders, such as private sector companies, resort research, or scientific research institutions with different primary interests.
The types of acquired data generated range from temporally and spatially resolved data to analysis results on quantitative and qualitative traits or molecular characteristics, developed software, and records of field experiments, which may include scoring data on agronomic and breeding traits as well as information on fertilization, crop protection, field and soil conditions, geodata, and weather data. The data differ not only in their

object of study, but also in their type, format, and context of origin. The NFDI consortium FAIRAgro therefore forms in this domain important anchors for the use cases for the infrastructure being developed.

This proposed use case aims to implement a Digital Twin (DT) service to use IGSNs as PUIDs for material samples, complement DOIs as recommended PUIDs for Plant Genetic Resources for Food and Agriculture (PGRFA) (Manzella, Daniele 2016) by individual samples. The unique identification of all samples, materials, collected data, and annotated metadata, as well as their relation to each other, must performed throughout the lifecycle to enable their integration into a shareable data space.This is where the Digital Twin concept (DT) (Vachalek et al. 2017) comes in as an extension of the Fair Digital Object (FDO) (Wharton 2022). DTs are virtual, digital representations of products or services and consist of a physical part and a virtual representation as well as a bridging link between the two (Portela et al. 2020), thereby also mapping the dynamics of the object. Digital twins are about the **"as-is" state** of an entity, while an FDO is a **historical collection** of data that may affect many entities over a period of time. This concept, originally built on as approach to integrate complex industrial and digital processes, could be complementary to DOIs, as an identifier for PGRFA, towards a solution to identify samples, which based on a particular material, and could be the basis for FAIR and transparent documentation of the research data lifecycle for plant research and plant breeding data in field and lab.

Requests for the PID4NFDI proposal

During the discussion slot the participants were asked for their wishes and requests for the planned PID basic service (PID4NFDI). The participants highlighted that the development of cook books, guidance, best practices, and the support of open discussions in the NFDI community would be very helpful.

A participant also mentioned that PIDs and confidentiality is not a contradiction and emphasized that one doesn't need to expose any information when minting e.g. a handle.

Metadata richness for PIDs, which was seen closely connected with Interoperability, were also topics of the discussion. At the end of the workshop, a participant made the point to take the discussion further on if there is a need for a unified PID service to avoid the splitting up of the PID landscape.

Further reading

Bertelmann, R., Buys, M., Kett, J., Pampel, H., Pieper, D., Scholze, F., Sens, I., Burger, F., Dreyer, B., Glagla-Dietz, S., Hagemann-Wilholt, S., Hartmann, S., Schrader, A., Schirrwagen, J., Summann, F., & Vierkant, P. (2023). PID Network Deutschland. Netzwerk für die Förderung von persistenten Identifikatoren in Wissenschaft und Kultur. Helmholtz Open Science Office. https://doi.org/10.48440/os.helmholtz.059

Bingert, Sven, Brase, Jan, Burger, Felix, Dreyer, Britta, Hagemann-Wilholt, Stephanie, Vierkant, Paul, & Wieder, Philipp. (2022). Concept for Setting up the Persistent Identifier Services Working Group in the NFDI Section "Common Infrastructures" (1.0). Zenodo. https://doi.org/10.5281/zenodo.6507760

Klump, J., Lehnert, K., Ulbricht, D., Devaraju, A., Elger, K., Fleischer, D., Ramdeen, S., & Wyborn, L. (2021). Towards Globally Unique Identification of Physical Samples: Governance and Technical Implementation of the IGSN Global Sample Number. In Data Science Journal (Vol. 20). Ubiquity Press, Ltd. https://doi.org/10.5334/dsj-2021-033

Appendix

- PID und Basisdienste - Philipp Wieder
- PIDs for data repository entries, datasets, and schemas (FAIRmat) - Markus Scheidgen
- PIDs for Instruments (NFDI4Ing) - Michael Selzer
- PIDs for Samples (Daphne4NFDI) - Rolf Krahl
- IGSN (NFDI4Earth) - Kirsten Elger
- Survey variables (KonsortSWD) - Janete Bach, Claus-Peter Klas & Peter Mutschke
- PIDs for Catalysis (NFDI4Cat) - David Linke
- StrainInfo PIDs for microbial strains (NFDI4Microbiota) - Lorenz Reimer
- PIDs for Cultural Data (NFDI4Culture & RADAR4Culture) - Desiree Mayer & Sandra Göller
- Digital twins of plant genetic resources (FAIRagro) - Daniel Arend

PID und Basisdienste

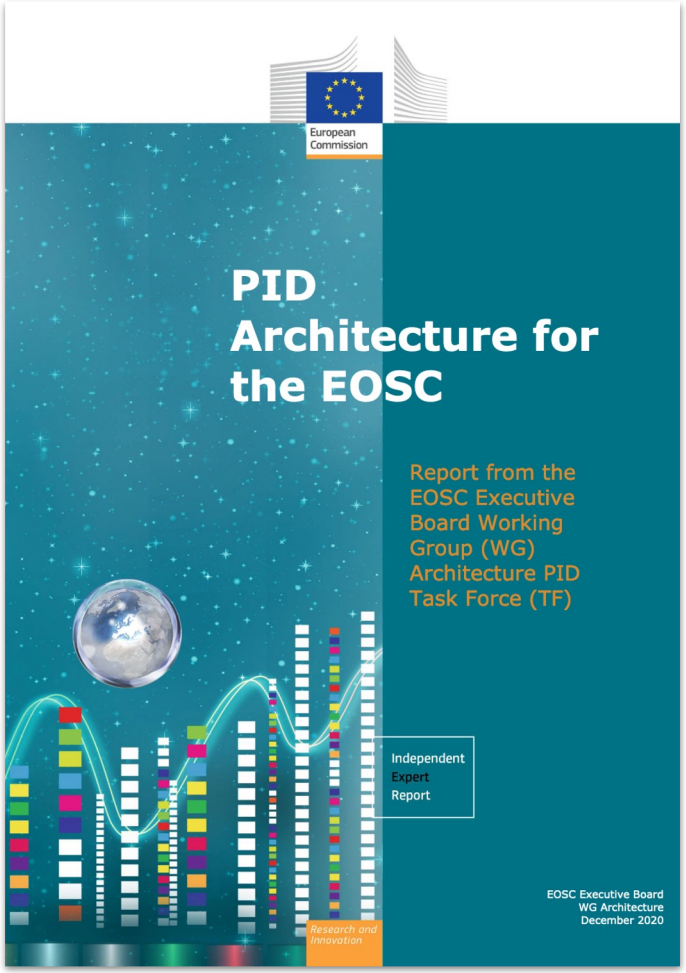
Ausgangslage für PID Entwicklungen

Vorarbeiten und aktuelle Entwicklungen

- EOSC
  - Publication: PID Architecture for the EOSC [1]
  - PID Policy and Implementation Task Force
- Verschiedene Gruppen der Research Data Alliance
- Projekten, die EOSC zuarbeiten
  - FREYA (beendet)
  - DICE
  - FAIRCORE4EOSC

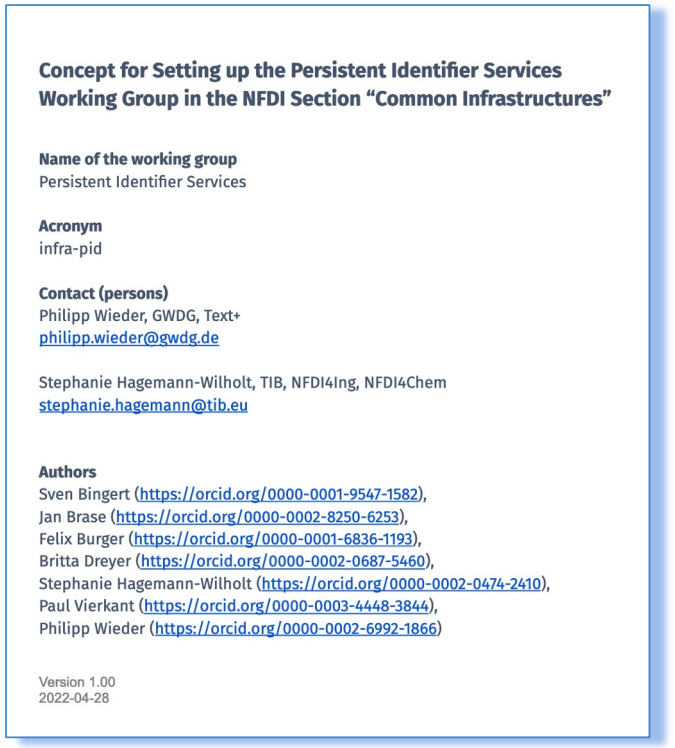
NFDI Persistent Identifier Services Working Group (PID WG)

- Wichtigkeit des Themas innerhalb der NFDI unbestritten
- Konzept ist publiziert [2]
- Aktuell 13 Einrichtungen aktiv



Ziele der PID Working Group

1. Erfassung der PID Use Cases und Anforderungsanalyse
   - Evaluierung der Disziplinen/Konsortien
   - Anforderungsanalyse
   - Abbildung der Anforderungen auf existierende Services
2. Entwicklung NFDI-weiter Konzepte für die Nutzung und Integration relevanter PID Systeme
   - Fokus auf einfacher Nutzung & Integration
   - Interoperabilität von Metadaten
   - Einbettung in “NFDI Service Architektur”



Concept for Setting up the Persistent Identifier Services Working Group in the NFDI Section “Common Infrastructures”

Name of the working group
Persistent Identifier Services

Acronym
infra-pid

Contact (persons)
Philipp Wieder, GWDG, Text+
philipp.wieder@gwdg.de

Stephanie Hagemann-Wilholt, TIB, NFDI4Ing, NFDI4Chem
stephanie.hagemann@tib.eu

Authors
Sven Bingert (https://orcid.org/0000-0001-9547-1582),
Jan Brase (https://orcid.org/0000-0002-8250-6253),
Felix Burger (https://orcid.org/0000-0001-6836-1193),
Britta Dreyer (https://orcid.org/0000-0002-0687-5460),
Stephanie Hagemann-Wilholt (https://orcid.org/0000-0002-0474-2410),
Paul Vierkant (https://orcid.org/0000-0003-4448-3844),
Philipp Wieder (https://orcid.org/0000-0002-6992-1866)

Version 1.00
2022-04-28

https://zenodo.org/record/6507760#.YxBrbvHP0-R

Ziele der PID Working Group

3. Wissenstransfer zur Unterstützung Nutzung und Entwicklung von PID Diensten
    - Cook Books und Best Practices
    - Weitergabe von Informationen an Projekte, Entwickler & weitere Stakeholder
4. Entwicklung von Governance- und Lizenzmodellen für die NFDI
    - Auswertung existierender Modelle
    - Definition & Umsetzung von spezifischen Erweiterungen für die NFDI
5. Zusammenarbeit mit deutschen und internationalen Institutionen bzgl. PIDs
    - Beratung von & Zusammenarbeit mit NFDI Konsortien
    - Fokus auf EOSC bzgl. generellen Dienstangebot
    - Kollaboration mit (inter)nationalen Playern

Deliverables

- Sammlung von konsolidierten Anforderungen der NFDI Konsortien
- Einheitlicher, einfacher Zugang zur Nutzung und Registrierung von PIDs
- Schnittstellen zur Nutzung und Integration von PID Services
- Technische (Weiter)entwicklung: z.B. neue PID Typen, einheitliche Metadaten, …
- Organisatorische Konzepte bzgl. Governance und Lizenzen
- Monitoring der Dienste
- Betriebsmodell und rechtlich geprüftes Rahmenwerk für NFDI-weiten Betrieb

Persistent Identifier Services als Teil von Base4NFDI

- **Antragsverfahren**: drei Phasen (initialisation, development, integration) können beantragt werden
  - Details: https://base4nfdi.de/
- Erste Einreichung am **15.2.2023** möglich
- Antragsdokument wird gerade von der PID Services AG vorbereitet

→ Ergebnisse des heutigen Workshops fließen noch in den Antrag ein und bilden, zusammen mit den Ergebnissen der aktuellen Umfrage, die Basis für die Arbeit des potentiellen Basisdienstes

Leitfragen für Heute

- Welche Entwicklungen gibt es in den verschiedenen NFDI Communities?
- Was kann ich von den anderen lernen?
- Welche Wünsche habe ich an Persistent Identifier AG?
- Welche Wünsche habe ich an Base4NFDI?
- Wie geht es weiter?

Quellen

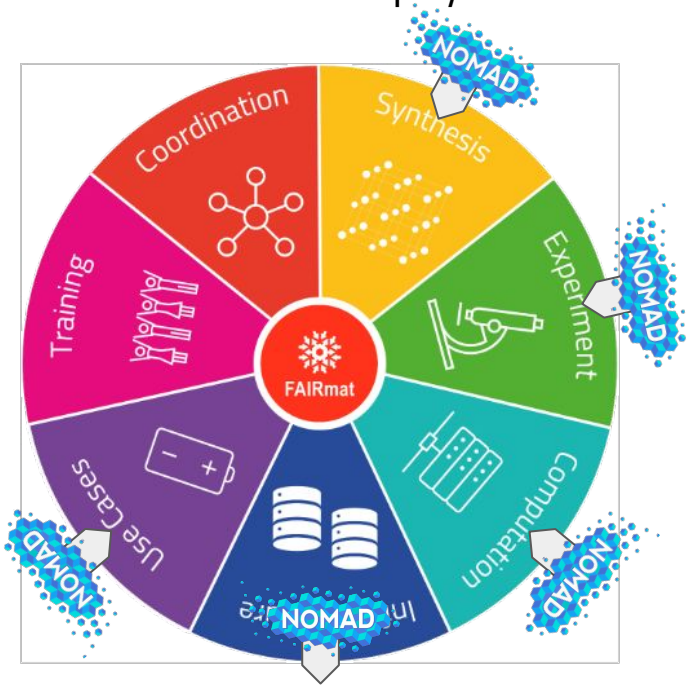
[1] PID architecture for the EOSC: https://s.gwdg.de/iWKgoE

[2] Concept for Setting up the Persistent Identifier Services Working Group in the NFDI Section “Common Infrastructures”: https://doi.org/10.5281/zenodo.6507759

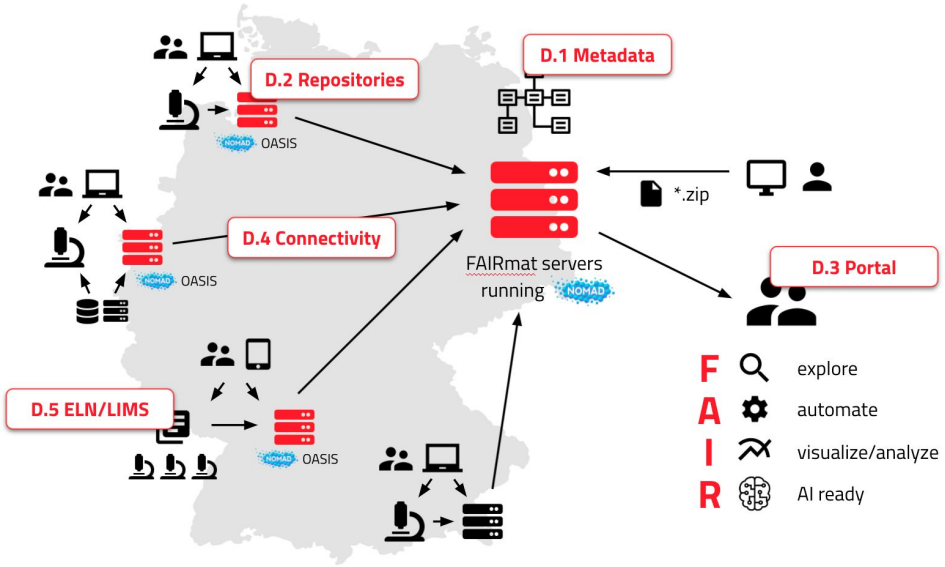
[3] Nationale Forschungsdateninfrastruktur: Förderung von NFDI-weiten Basisdiensten: https://s.gwdg.de/APBrFz

What is FAIRmat and NOMAD?

FAIRmat is the NDFI consortium to build a FAIR federated data infrastructure for solid state physics



NOMAD is a web-based software for FAIR research data management in materials science



What we need from PID and PID services

- a **singular** PID service to issue and resolve PIDs for data from all domains
- meaningful and flexible metadata schema for PID metadata
  - dublin-core-style metadata
  - metadata to describe the type of data (e.g. via schema-pid)
  - metadata to describe non-human programs and instruments as source/"author" (e.g.via source-pid)
  - metadata to describe the repository (e.g. via repository-pid)
- we want to use PIDs for
  - repository entries (simulations, samples, instruments, ELNs, measurement data, documents, …); ~10^7 of them
  - user curated datasets; ~10^3 of them
  - schemas and data-type definitions; ~10^3 of them
- ideally the service would be able to "promote" PIDs to become DOIs resolveable through the respective existing DOI services (datacite, etc.)
- we want to transition from issuing DOIs on demand to issuing PIDs automatically
- data and PIDs should go through a life-cycle together
  - might end in deletion (ephemeral PIDs) or long term archival
  - PID metadata and what it resolves to has to be changeable throughout the whole life-cycle

Karlsruhe Institute of Technology

Karlsruhe Institute of Technology

Instrument data base, meta data and related topics

A short introduction to the setup of a KIT Gerätepool and a brief outlook on further developments and possible fields of application

The KIT instrument data base („Gerätepool“)

- Background - Why do we need a Gerätepool?
- Who is responsible for the development of the Gerätepool?
- First bundle of instruments
- First round of questions to the instrument experts
- Where and how will the Gerätepool be realised?
- Future goals

The KIT instrument data base („Gerätepool“)
- First round of questions -

Karlsruhe Institute of Technology

Erfassung von Gerätedaten

AG Gerätepool
Ansprechpartner:
Dr. Axel Schumacher (BIFTM) axel.schumacher@kit.edu +49 721 608-24855
Prof. Dr. Christian Kübel (INT) christian.kuebel@kit.edu +49 721 608-28970
Dr. Matthias Mail (INT) matthias.mail@kit.edu +49 721 608-28909

Karlsruhe Institute of Technology

Anlage: xxxxx

Anlagenbezeichnung: _____

Gerätebezeichnung: _____
(vom Hersteller)

Hersteller: _____ Erstzugang am: _____

Gerätezustand: ☐ Orignalzustand ☐ Upgrade _____
Beschreibung

☐ Generalüberholt _____ ☐ Modofiziert _____
Datum Beschreibung

Software für das Gerät: _____

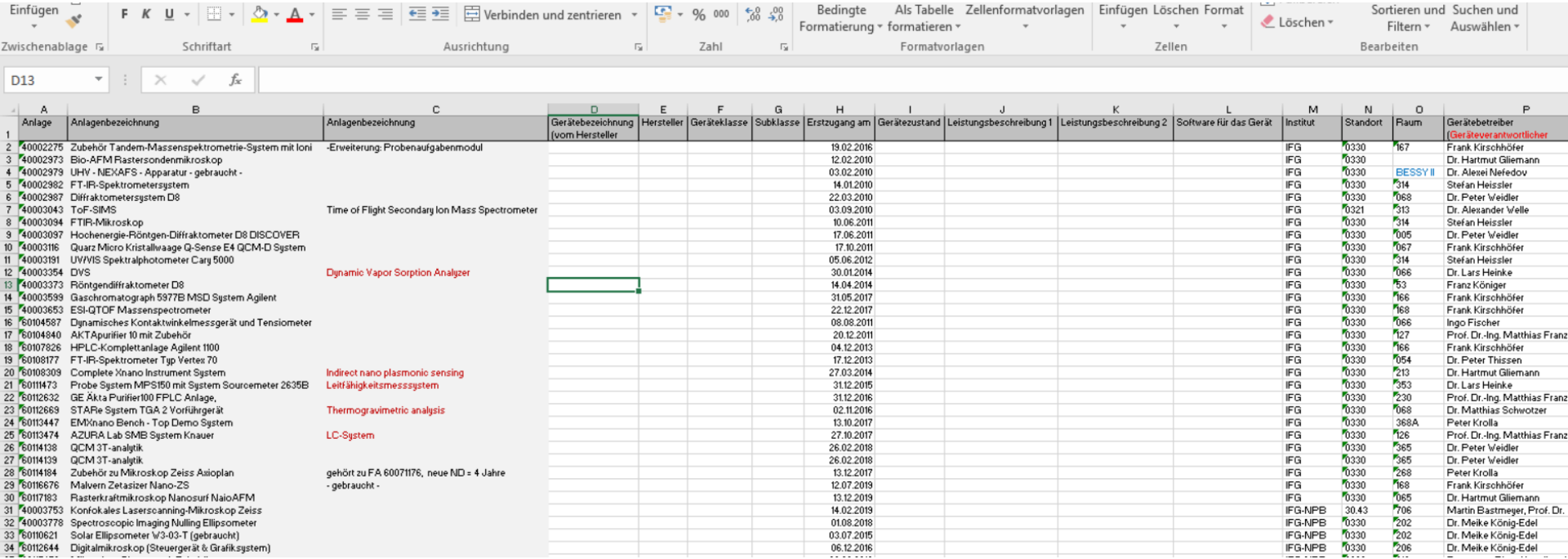
Institut: _____

Standort: _____

Raum: _____

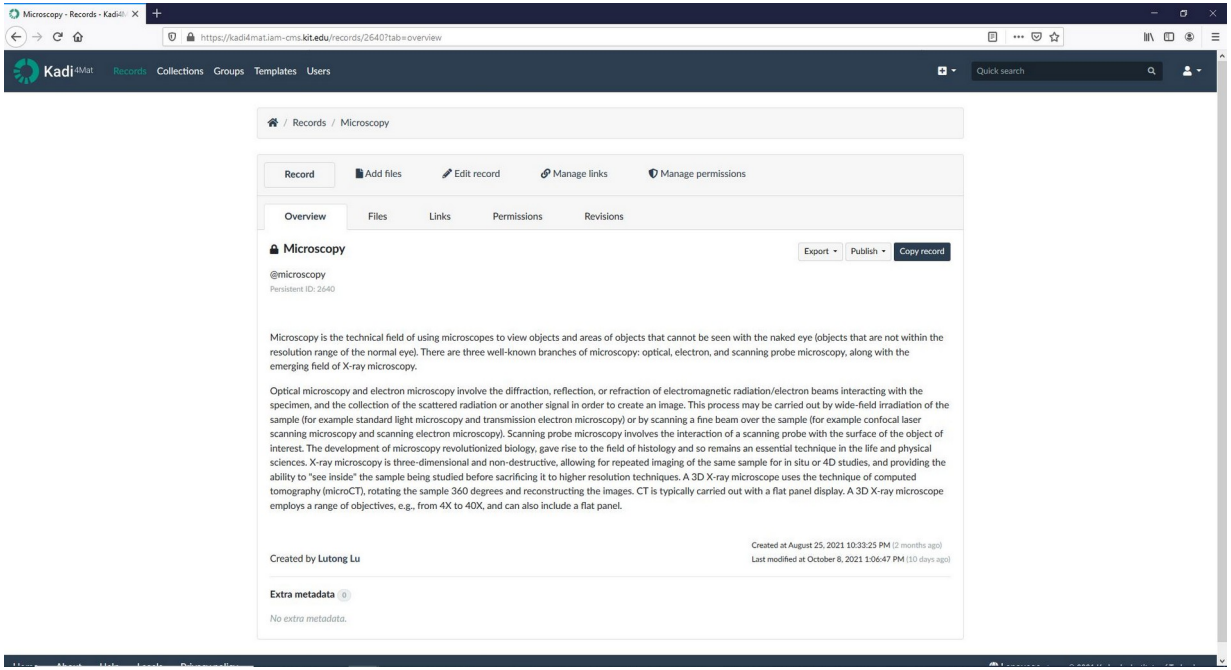
The KIT instrument data base („Gerätepool“)
- First bundle of instruments -

Karlsruhe Institute of Technology



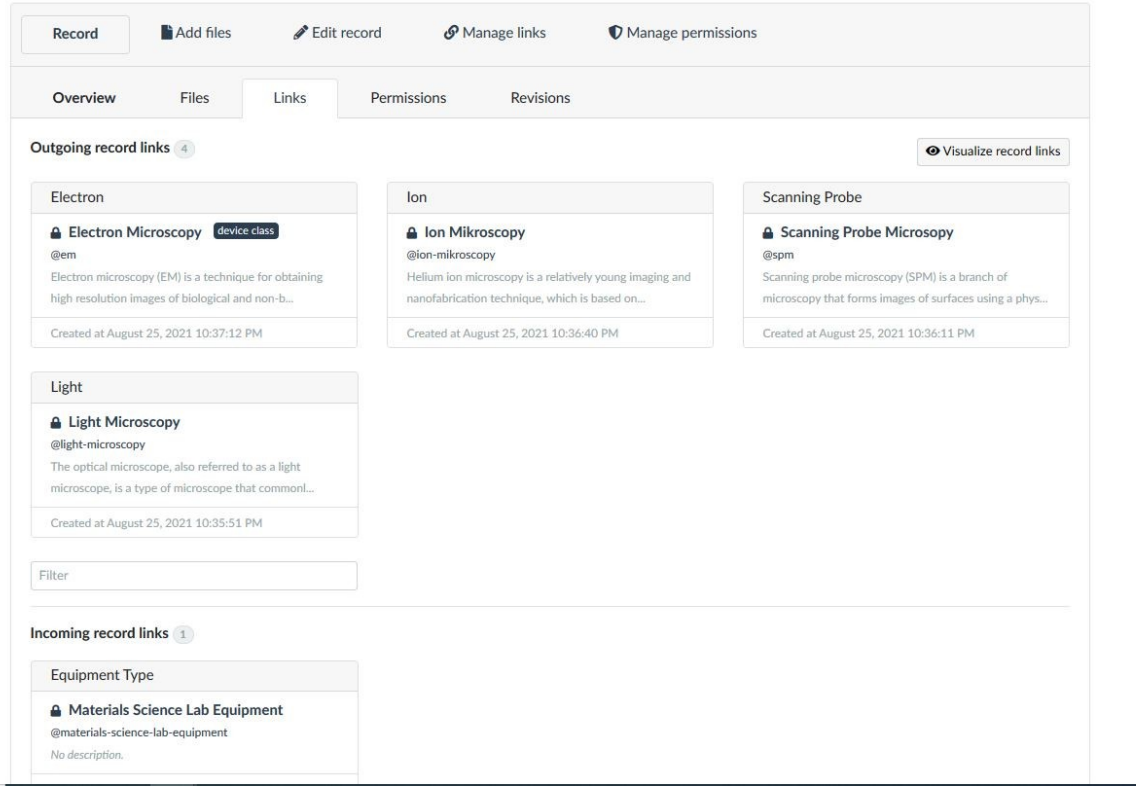
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Anlage	Anlagenbezeichnung	Anlagenbezeichnung	Gerätebezeichnung (vom Hersteller	Hersteller	Geräteklasse	Subklasse	Erstzugang am	Gerätezustand	Leistungsbeschreibung 1	Leistungsbeschreibung 2	Software für das Gerät	Institut	Standort	Raum	Gerätebetreiber (Geräteverantwortlicher
2	40002275	Zubehör Tandem-Massenspektrometrie-System mit Ioni	-Erweiterung: Probenaufgabenmodul					19.02.2016					IFG	0330	167	Frank Kirschhöfer
3	40002973	Bio-AFM Rastersondenmikroskop						12.02.2010					IFG	0330		Dr. Hartmut Gliemann
4	40002979	UHV - NEXAFS - Apparatur - gebraucht -						03.02.2010					IFG	0330	BESSY II	Dr. Alexei Nefedov
5	40002982	FT-IR-Spektrometersystem						14.01.2010					IFG	0330	314	Stefan Heissler
6	40002987	Diffraktometersystem D8						22.03.2010					IFG	0330	068	Dr. Peter Weidler
7	40003043	ToF-SIMS	Time of Flight Secondary Ion Mass Spectrometer					03.09.2010					IFG	0321	313	Dr. Alexander Welle
8	40003094	FTIR-Mikroskop						10.06.2011					IFG	0330	314	Stefan Heissler
9	40003097	Hochenergie-Röntgen-Diffraktometer D8 DISCOVER						17.06.2011					IFG	0330	005	Dr. Peter Weidler
10	40003116	Quarz Micro Kristallwaage Q-Sense E4 QCM-D System						17.10.2011					IFG	0330	067	Frank Kirschhöfer
11	40003191	UV/VIS Spektralphotometer Cary 5000						05.06.2012					IFG	0330	314	Stefan Heissler
12	40003354	DVS	Dynamic Vapor Sorption Analyzer					30.01.2014					IFG	0330	066	Dr. Lars Heinke
13	40003373	Röntgendiffraktometer D8						14.04.2014					IFG	0330	53	Franz Königer
14	40003599	Gaschromatograph 5977B MSD System Agilent						31.05.2017					IFG	0330	166	Frank Kirschhöfer
15	40003653	ESI-QTOF Massenspectrometer						22.12.2017					IFG	0330	168	Frank Kirschhöfer
16	60104587	Dynamisches Kontaktwinkelmessgerät und Tensiometer						08.08.2011					IFG	0330	066	Ingo Fischer
17	60104840	ÄKTApurifier 10 mit Zubehör						20.12.2011					IFG	0330	127	Prof. Dr.-Ing. Matthias Franz
18	60107826	HPLC-Komplettanlage Agilent 1100						04.12.2013					IFG	0330	166	Frank Kirschhöfer
19	60108177	FT-IR-Spektrometer Typ Vertex 70						17.12.2013					IFG	0330	054	Dr. Peter Thissen
20	60108309	Complete Xnano Instrument System	Indirect nano plasmonic sensing					27.03.2014					IFG	0330	213	Dr. Hartmut Gliemann
21	60111473	Probe System MPS150 mit System Sourcemeter 2635B	Leitfähigkeitsmesssystem					31.12.2015					IFG	0330	353	Dr. Lars Heinke
22	60112632	GE Äkta Purifier100 FPLC Anlage,						31.12.2016					IFG	0330	230	Prof. Dr.-Ing. Matthias Franz
23	60112669	STARe System TGA 2 Vorführgerät	Thermogravimetric analysis					02.11.2016					IFG	0330	068	Dr. Matthias Schwotzer
24	60113447	EMXnano Bench - Top Demo System						13.10.2017					IFG	0330	368A	Peter Krolla
25	60113474	AZURA Lab SMB System Knauer	LC-System					27.10.2017					IFG	0330	126	Prof. Dr.-Ing. Matthias Franz
26	60114138	QCM 3T-analytik						26.02.2018					IFG	0330	365	Dr. Peter Weidler
27	60114139	QCM 3T-analytik						26.02.2018					IFG	0330	365	Dr. Peter Weidler
28	60114184	Zubehör zu Mikroskop Zeiss Axioplan	gehört zu FA 60071176, neue ND = 4 Jahre					13.12.2017					IFG	0330	268	Peter Krolla
29	60116676	Malvern Zetasizer Nano-ZS	- gebraucht -					12.07.2019					IFG	0330	168	Frank Kirschhöfer
30	60117183	Rasterkraftmikroskop Nanosurf NaioAFM						13.12.2019					IFG	0330	065	Dr. Hartmut Gliemann
31	40003753	Konfokales Laserscanning-Mikroskop Zeiss						14.02.2019					IFG-NPB	30.43	706	Martin Bastmeyer, Prof. Dr.
32	40003778	Spectroscopic Imaging Nulling Ellipsometer						01.08.2018					IFG-NPB	0330	202	Dr. Meike König-Edel
33	60110621	Solar Ellipsometer W3-03-T (gebraucht)						03.07.2015					IFG-NPB	0330	202	Dr. Meike König-Edel
34	60112644	Digitalmikroskop (Steuergerät & Grafiksystem)						06.12.2016					IFG-NPB	0330	206	Dr. Meike König-Edel

The KIT instrument data base („Gerätepool“) - Inside Kadi4Mat



Kadi4Mat is the **Karlsruhe Data Infrastructure for Materials Science**, a software for managing research data with the aim of combining new concepts with established technologies and existing solutions.
https://kadi4mat.iam-cms.kit.edu/about

The KIT instrument data base („Gerätepool“) - Inside Kadi4Mat - Overview

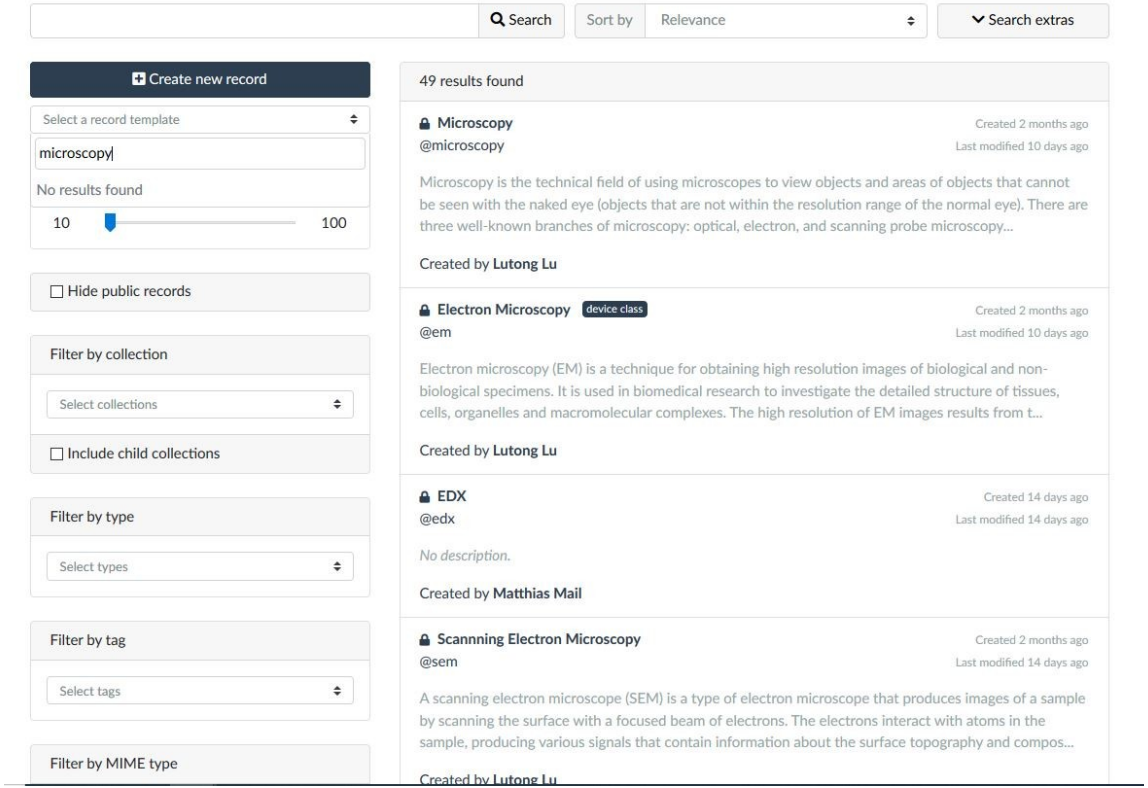


Karlsruhe Institute of Technology

Kadi4Mat

Kadi4Mat is the **Karlsruhe Data Infrastructure for Materials Science**, a software for managing research data with the aim of combining new concepts with established technologies and existing solutions.
https://kadi4mat.iam-cms.kit.edu/about

The KIT instrument data base („Gerätepool“) - Inside Kadi4Mat - Searching

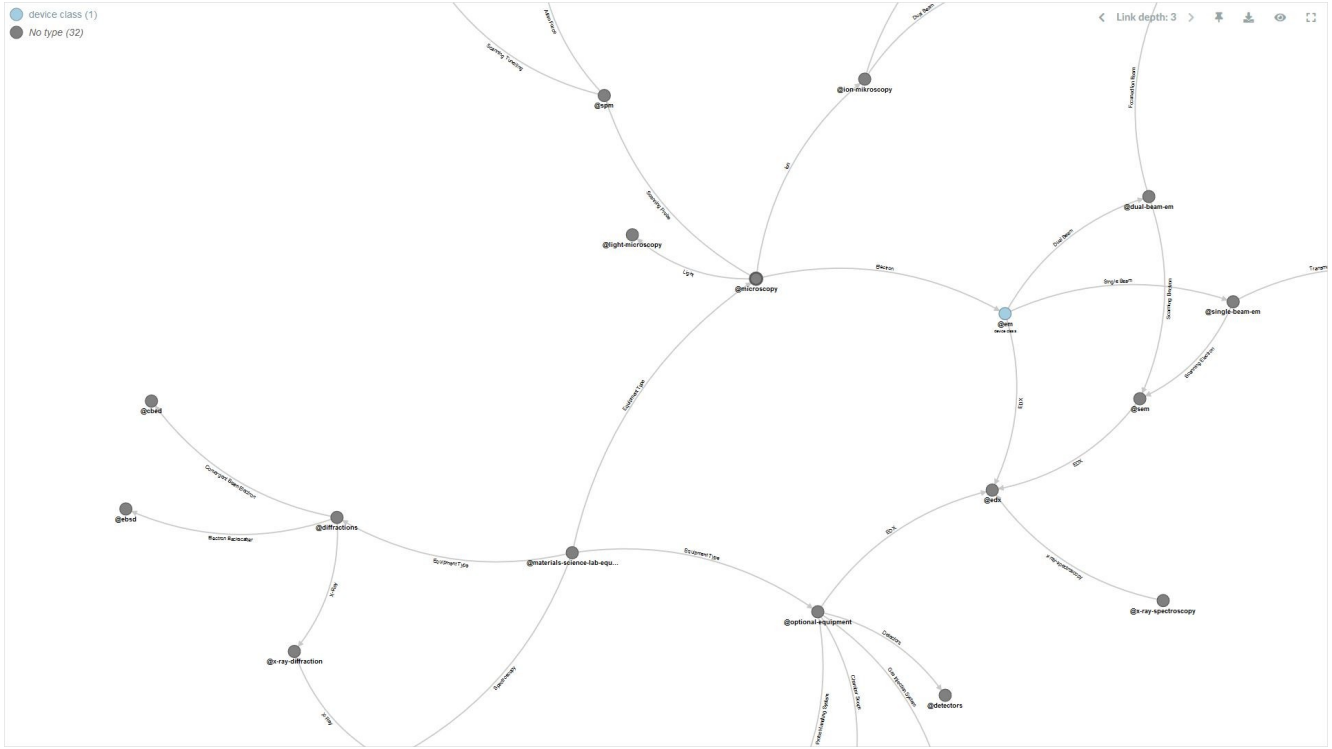


Karlsruhe Institute of Technology

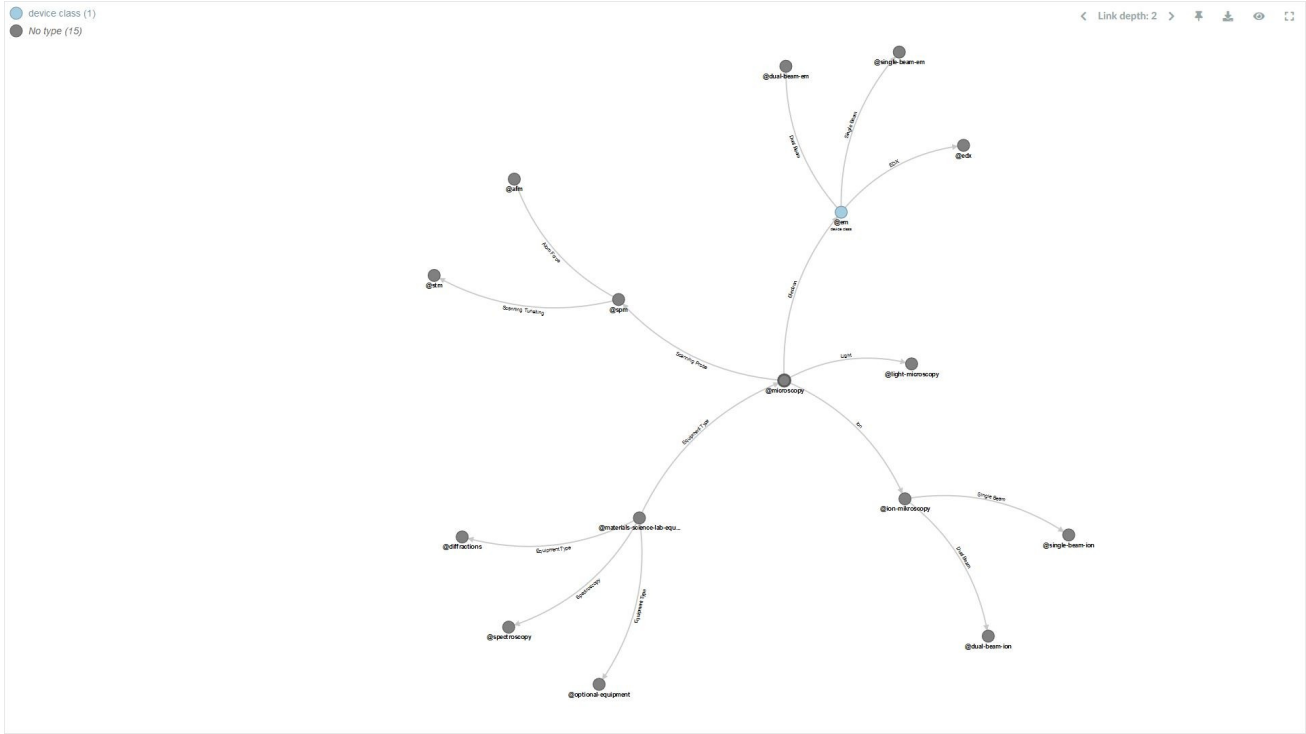
Kadi4Mat

Kadi4Mat is the **Karlsruhe Data Infrastructure for Materials Science**, a software for managing research data with the aim of combining new concepts with established technologies and existing solutions.
https://kadi4mat.iam-cms.kit.edu/about

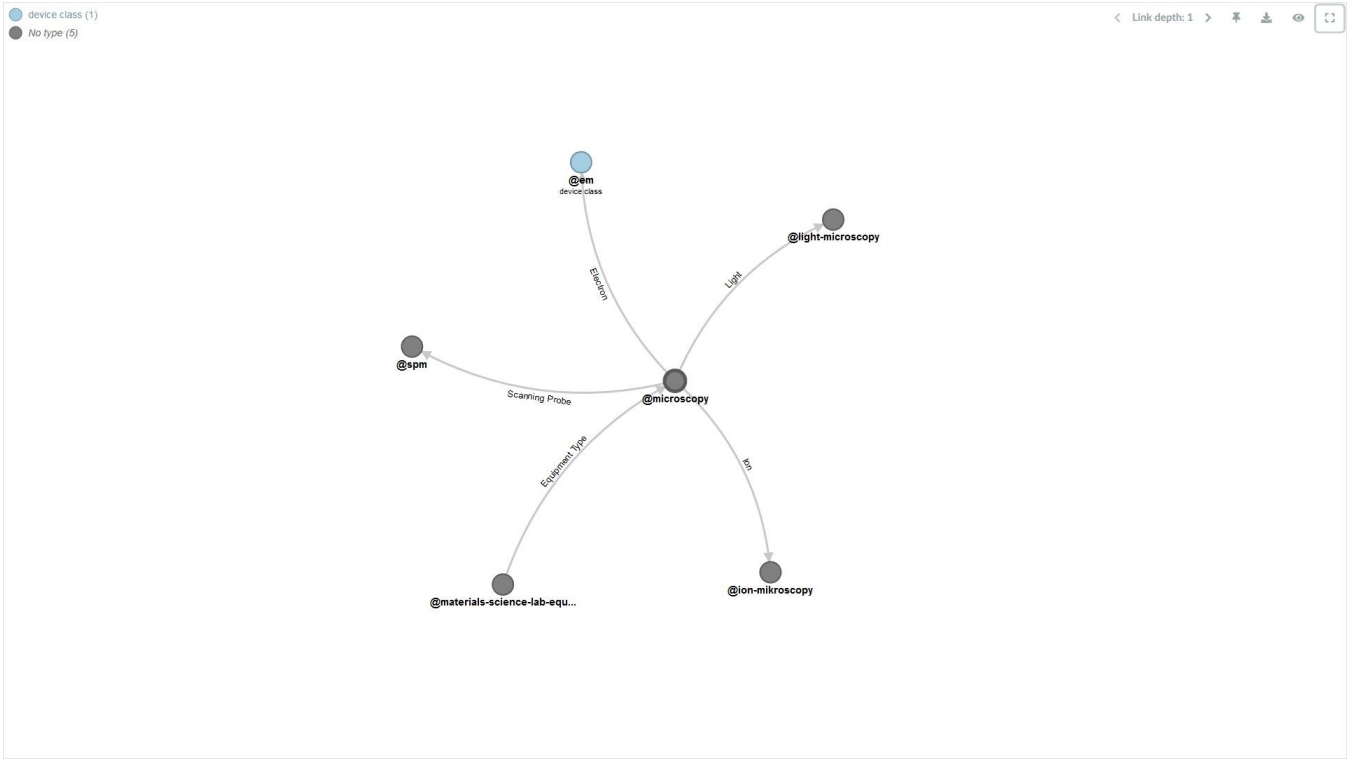
The KIT instrument data base („Gerätepool“) - Inside Kadi4Mat - Searching



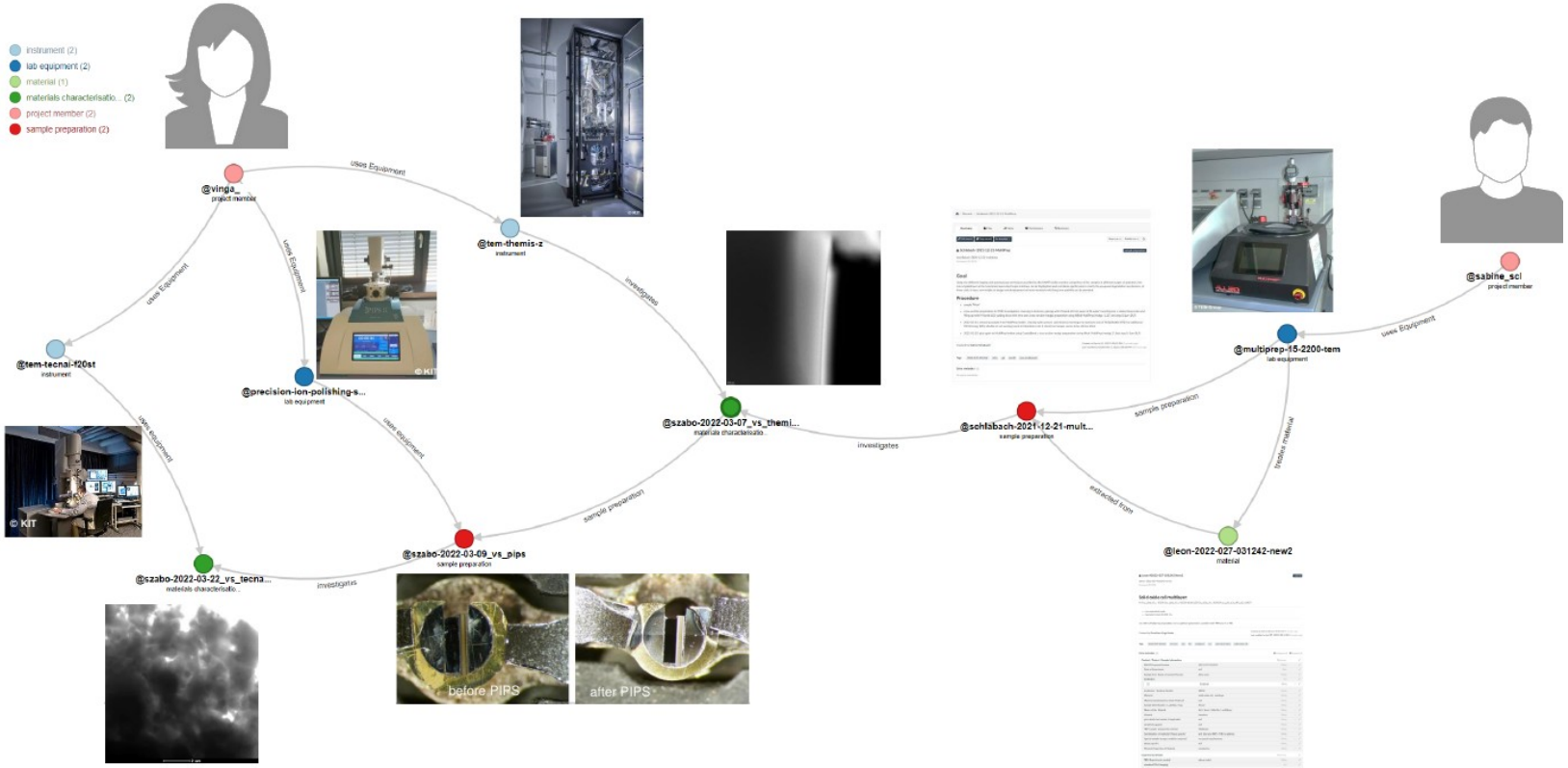
The KIT instrument data base („Gerätepool“) - Inside Kadi4Mat - Searching



The KIT instrument data base („Gerätepool“) - Inside Kadi4Mat - Searching



Usage of Instrument descriptions in the research process



The KIT instrument data base („Gerätepool“)
- Future goals -

- Including more instruments
  - Older
  - Smaller/Cheaper ones
  - etc.…
- Full description of all instruments
- Using the database for other topics, e.g.
  - Ontology development
  - Correlative approaches
  - Electronic Lab Notebooks
  - etc.

- Thank you for listening
- Any questions?

HZB Helmholtz Zentrum Berlin

PIDs for Sample in DAPHNE4NFDI

Rolf Krahl

Workshop on PIDs within NFDI, 27 January 2023

DAPHNE4NFDI

- DAta from PHoton and Neutron Experiments (DAPHNE4NFDI): photons and neutrons at large-scale research facilities.
- Scope of DAPHNE is a class of experimental methods, not a particular field of research. We have research from all areas.
- Accurate description of the sample being measured is an important part in the metadata for the resulting dataset.
- Huge diversity of samples.

Note

Discussion on sample PIDs in DAPHNE is at the very beginning. No conclusive concept yet. I can only present my own personal view on the matter here.

Samples at Photon and Neutron Facilities

We have all sorts of different samples. Some examples:

- single crystals, crystal powder, novel functional materials
- complex structures: solar cells, batteries
- macromolecular crystals, e.g. proteins
- archeological artefacts, fossils
- artwork, e.g. paintings
- living plants

Samples at Photon and Neutron Facilities

Diversity of live cycles:

- Most samples are specifically created for the particular experiment, often in the user's home lab.
- Samples may have a complex history of preparation steps.
- Samples may be split in parts, each part being treated differently in subsequent steps.
- Samples may be held in a collection outside the context of the experiment.
- Operando measurements: samples may be changing during the experiment.
- Ephemeral samples: samples may be created, measured and destroyed during one experiment, e.g. the sample may be a liquid jet injested in the measurement chamber.

Sample PIDs

Why do we need sample PIDs, what are the requirements:

- The full history of sample preparation steps constitutes the sample description.
- If each preparation and characterization step generates a dataset, if all these datasets are FAIR and if they are all linked together by referencing the sample PID, we get the full picture.
- If the same sample is measured in more than one independent investigations, we want to be able to combine the results.
- We need a parent / child relation for sample PIDs.
- We need to be able to integrate existing external ID, e.g. from collections holding objects.
- Open discussion: ephemeral samples.

IGSN
International Generic Sample Numbers
uniquely identifying your samples

Kirsten Elger

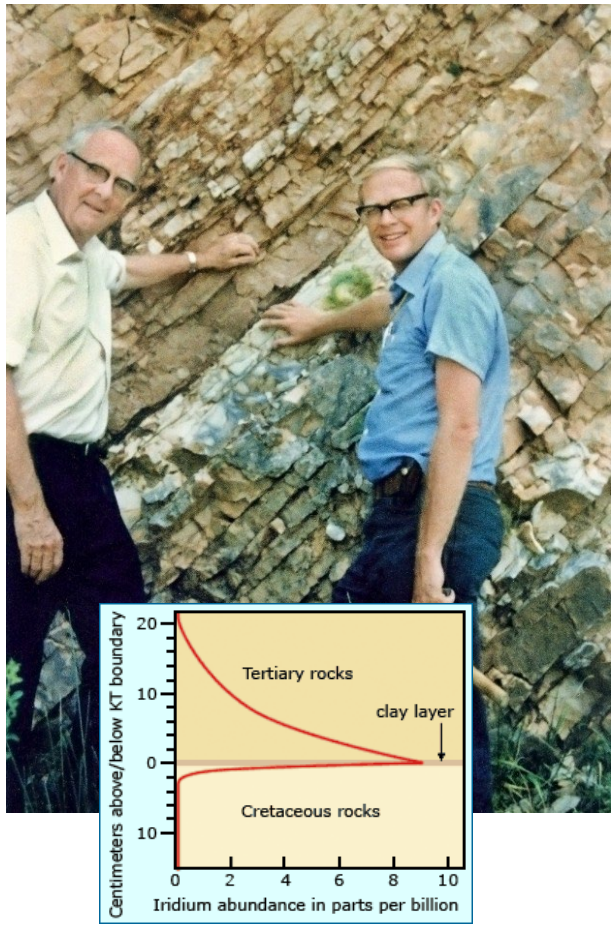
GFZ German Research Centre for Geosciences, Library and Information Services
Vice president of the IGSN e.V.

GFZ Helmholtz Centre Potsdam

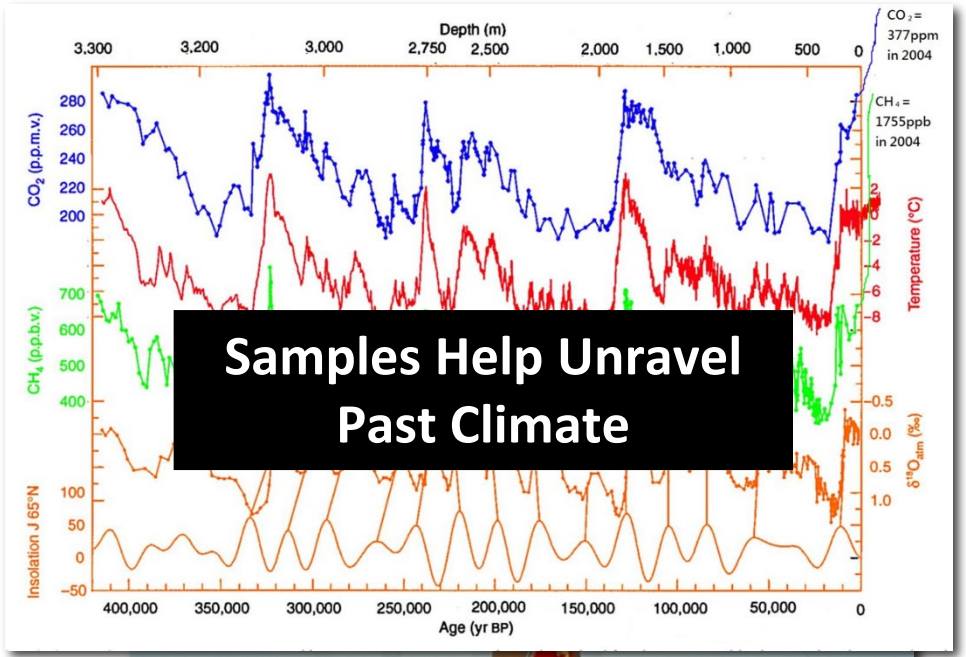
GFZ Data Services
Geosciences Data Publisher

HELMHOLTZ

Samples & Science



Samples Record Unique Events in History



Samples Help Unravel Past Climate

Archaean zircons in Miocene oceanic hotspot rocks establish ancient continental crust beneath Mauritius

Lewis D. Ashwal, Michael Wiedenbeck & Trond H. Torsvik

Nature Communications **8,** Received: 20 September 2016

Article

doi:10. 2017

Samples Provide Access to the Inaccessible

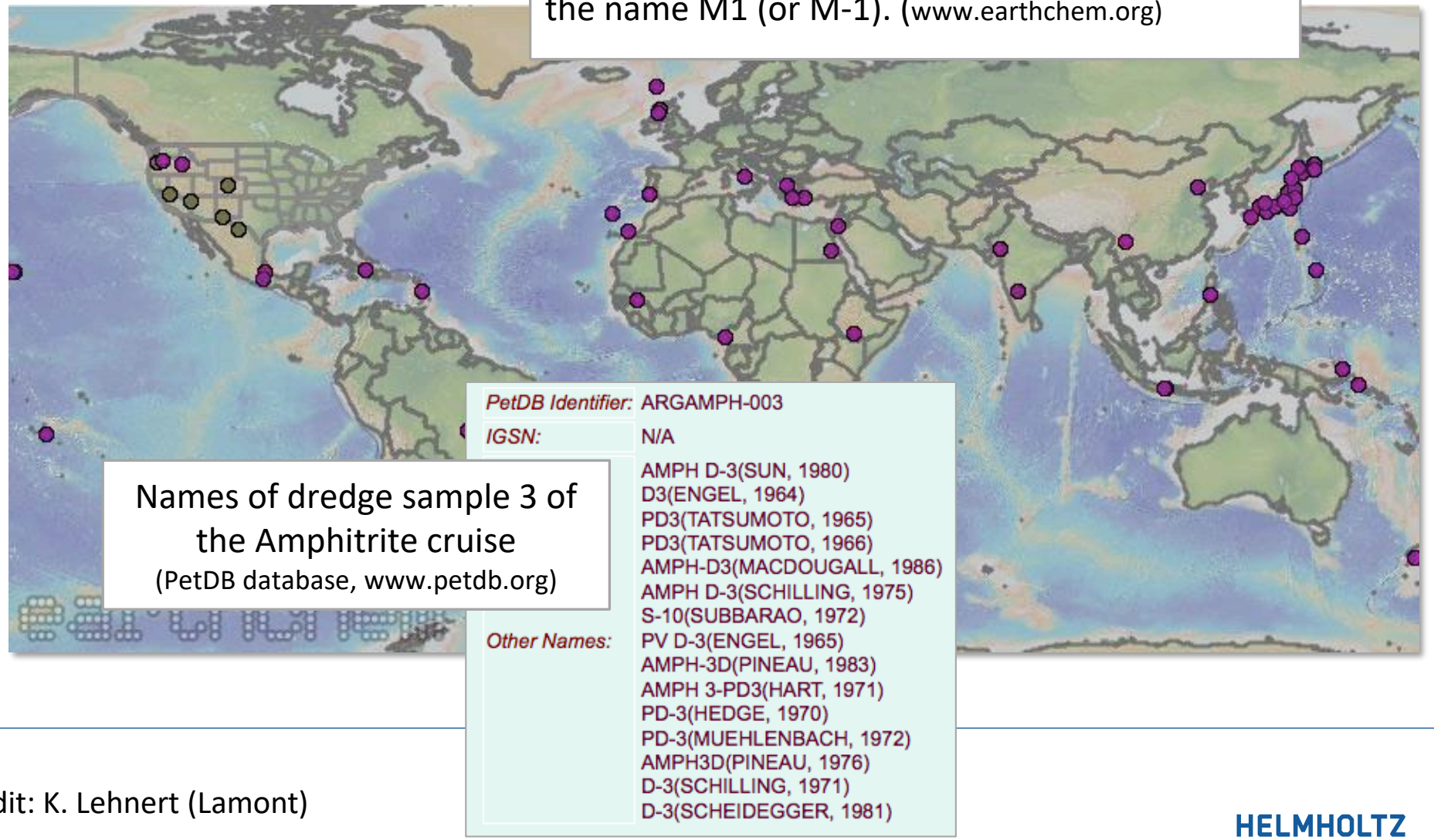
Credit: K. Lehnert (Lamont)

Why IGSN? Unambiguous names

EarthChem Portal

- One-stop shop for geochemical data (at LDEO)
- gives users the ability to search various federated databases simultaneously
- https://www.earthchem.org/

The EarthChem Portal includes 75 samples with the name M1 (or M-1). (www.earthchem.org)



Names of dredge sample 3 of the Amphitrite cruise
(PetDB database, www.petdb.org)

Credit: K. Lehnert (Lamont)

IGSN International Generic Sample Number

A globally unique and persistent identifier (PID) for physical objects (originally derived from the Earth Sciences)

- guaranteed to be unique via a centralized control mechanism (unique name spaces, Handle identifier)
- resolves to virtual sample representations (sample metadata profiles) managed at federated IGSN Allocating Agents.

IGSN: GRO000188

IGSN:	GRO000188
Sample Name:	12PTX04
Other Name(s):	Rhodes at Pawtuxet
Sample Type:	Individual Sample
Parent IGSN:	Not Provided

Description

Material:	Liquid>aqueous
Classification:	Not Provided
Field Name:	Not Provided
Description:	Not Provided

IGSN: GMY00007W

IGSN:	GMY00007W
Sample Name:	TN182_47_002
Other Name(s):	
Sample Type:	Individual Sample
Parent IGSN:	GMY00001B

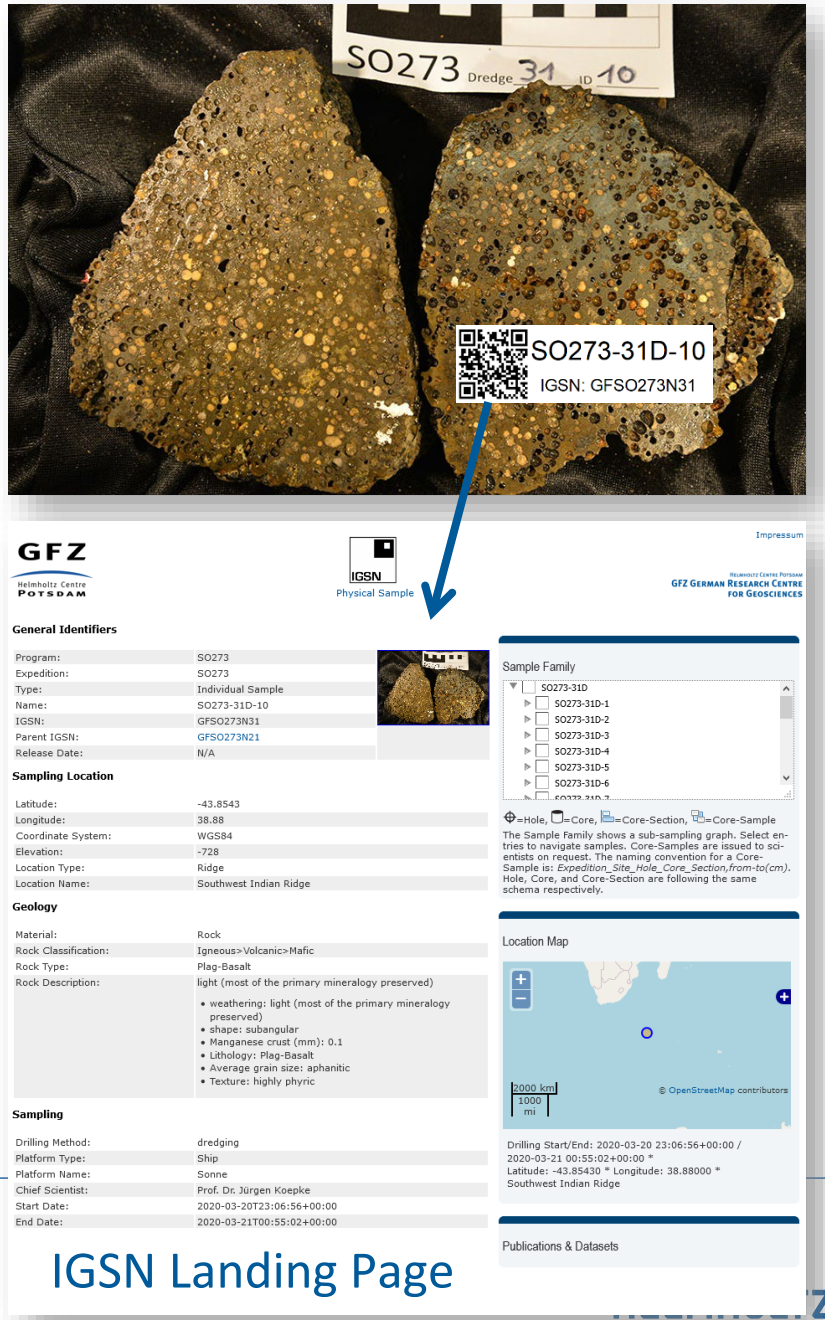
Description

Material:	Rock
Classification:	Igneous>Plutonic>Mafic
Field Name:	gabbro, hornblende gabbro
Description:	mafic plutonic rock

Credit: K. Lehnert (Lamont)

Benefits of IGSN

- unambiguously cite and track physical samples on a global scale:
  - allows previously impossible **linking of samples to data and publications**,
  - allows previously impossible **linking** and integration of **sample-based observations across data systems**, and
  - paves the road towards advanced data mining of sample-based data.
- persistently link to online digital representations of samples (landing pages)
  - builds a federated global sample catalog.
  - are citable in scholarly literature

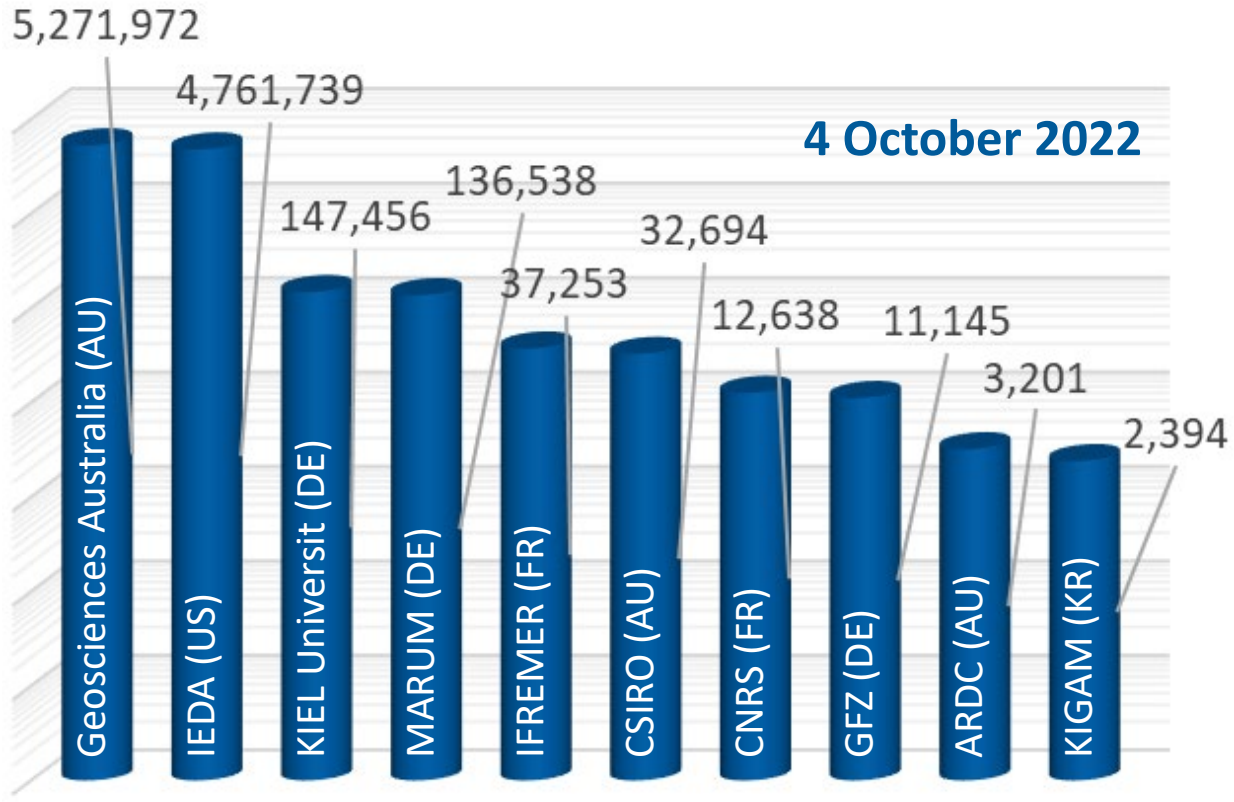


https://www.igsn.org/

IGSN e.V.

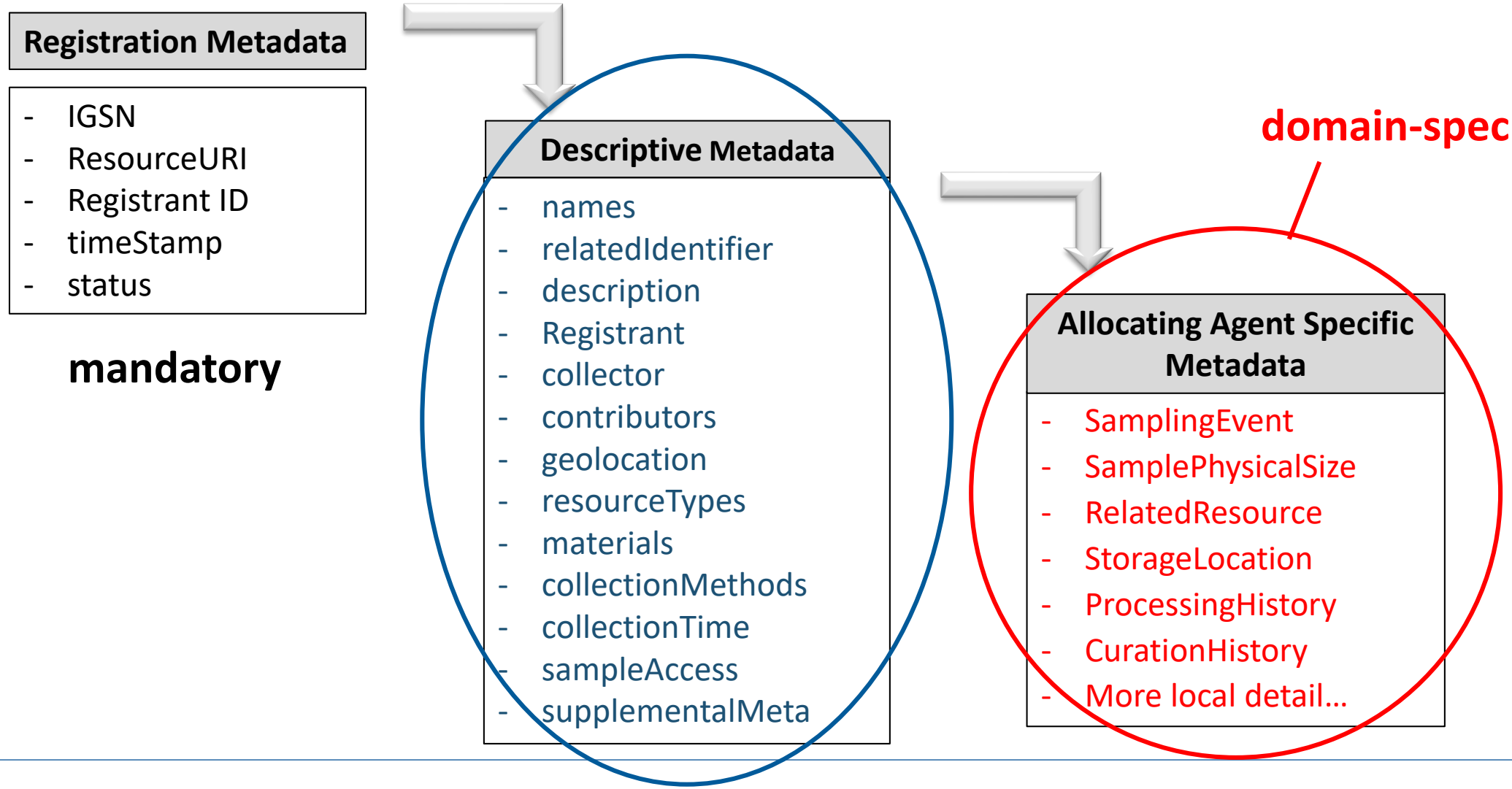
- established in 2011, registered as non-profit organization in Germany
- Members are organizations that want to provide registration and catalogue services (Allocating Agents)
  - currently 16 members,
  - 6 affiliate members and
  - 10 active Allocating Agents

Number of IGSN by Allocating Agent



Total = 10,417,030 registered IGSNs

IGSN Metadata Levels



Common Kernel: recommended for data discovery

IGSN metadata schema (XML)

```
<?xml version="1.0" encoding="UTF-8"?>
- <resource type="Sample" xsi:schemaLocation="http://pmd.gfz-potsdam.de/igsn/schemas/description/1.3 http://pmd.gfz-potsdam.de/igsn/schemas/description/1.3/resource.xsd" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://pmd.gfz-potsdam.de/igsn/schemas/description/1.3">
    <identifier type="IGSN">SSDPRR01ECV2101</identifier>
    <name>RR01_1_A_3</name>
    <parentIdentifier type="IGSN">SSDPRR01EH40001</parentIdentifier>
  - <registrant>
      <name>GFZ Data Services</name>
    - <affiliation>
        <identifier type="URL">https://ror.org/04z8jg394</identifier>
        <name>GFZ Potsdam</name>
      </affiliation>
    </registrant>
  - <collector>
      <identifier type="ORCID">0000-0003-2776-0846</identifier>
      <name>Christopher Juhlin</name>
    - <affiliation>
        <identifier type="URL">https://www.ror.org/048a87296</identifier>
        <name>Uppsala University</name>
      </affiliation>
    </collector>
  - <contributors>
    - <contributor type="ProjectLeader">
        <identifier type="ORCID">0000-0003-2776-0846</identifier>
        <name>Christopher Juhlin</name>
      </contributor>
    - <contributor type="Other">
        <name>Riksriggen, Engineering Geology</name>
      </contributor>
    </contributors>
  - <geoLocations>
    - <geoLocation>
        <geometry type="Point" sridType="4326">15.255724 59.224578</ge
      </geoLocation>
    </geoLocations>
  - <resourceTypes>
      <resourceType>http://vocabulary.odm2.org/samplingfeaturetype/spec
    </resourceTypes>
  - <materials>
      <material>http://vocabulary.odm2.org/medium/rock/</material>
    </materials>
  - <collectionMethods>
      <collectionMethod>Corer:Rock</collectionMethod>
    </collectionMethods>
```

IGSN Description Metadata is intended to describe the core elements of a specimen. The set of attributes is seen as the "birth certificate" of a specimens and should not contain stateful attributes, where possible. The base document for the development are the notes from the IGSN Metadata Kernel Workshop held in Los Angeles in September 2015.

IGSN Descriptive Metadata Elements

ID	Element	A/C	Occ	Definition	Description and instructions
	resource	Root element			IGSN "birth certificate" for a physical sample, associated feature, or collection
1	identifier	C	1	string	The Identifier is a unique string that identifies a resource. IGSN (International GeoSample Number) registered by an IGSN member. Format should be: "10273/foo"
1.1	identifierType	A	1	include/identifierType.xsd	currently only type=IGSN is supported
2	name		1...1, not nillable	string	Text string for people to understand what is identified. What would typically be presented in a user interface. Collector's or contributor's local/field name used to name the specimen; not globally unique but typically unique within a set of specimens.
3	alternateIdentifiers		0...1		Other formal identifiers for this resource, in addition to the IGSN.
3.1	alternateIdentifier	C	0...n	string	An identifier or identifiers other than the primary Identifier applied to the resource being registered. This may be any

IGSN technical (2012-2022)

- Central registry using the handle.net system
- Handle namespace: 10273
- One central Handle server operated at GFZ
- Uniqueness of IGSNs organized via different, centrally registered, namespaces (for allocating agents) and individual sub-namespaces
- IGSN resolve via https://igsn.org/<IGSN-number> or https://hdl.handle.net/10273/<IGSN-number>“ to IGSN Landing Pages

The future of IGSN
(from 1.1.2023 on)

- IGSNs will be registered as DataCite DOIs (resource type „physical object“)
- IGSN metadata schema will be mapped to DataCite schema (Guidelines available)
- Each DataCite Member or DataCite Consortium Member can assign DataCite „DOI IGSNs“
- DataCite recommends to use individual DOI prefixes for IGSN DOIs

DataCite Blog

October 2021

Partnership between IGSN and DataCite

October 28, 2021 Matt Buys and Kerstin Lehnert

https://doi.org/10.5438/7z70-1155

Earlier this year, DataCite and IGSN announced their roadmap towards a partnership to support the global adoption, implementation, and use of physical sample identifiers. Today, we are pleased to share the announcement of the partnership agreement.

DataCite is a community-led organisation with a vision to connect research and identify knowledge. We have been providing the means to create, find, cite, connect, and use research across 48 countries globally since 2009. In addition to DataCite core services, we support the scaling efforts of several identifier communities through governance, sustainability, insurance, and technical implementation.

IGSN e.V. is an international, non-profit organization with more than 20 members, which has operated a central registration system for IGSNs since 2011. The IGSN is a globally unique and persistent identifier for physical samples. The core purpose of IGSN is to enable transparent and traceable connections between research activities and objects, including samples, collections, instruments, grants, data, publications, people and organizations.

IGSN and DataCite have a common purpose, and a close relationship in the future will provide mutual benefit to our shared vision of connecting research and identifying knowledge. The partnership brings years of experience across our organizations and communities to scale sample community engagement, develop sample identifier practice standards, and increase adoption globally.

KonsortSWD

Consortium for the
Social, Behavioural, Educational
and Economic Sciences

KonsortSWD PID registration Service

Workshop on PIDs within NFDI

NFDI4 PID Working Group 2023

Presentation
27 January 2023

Janete Saldanha Bach
Claus-Peter Klas
Peter Mutschke

GESIS – Leibniz Institute for the Social Sciences

nfdi

gesis
Leibniz Institute
for the Social Sciences

Janete Bach

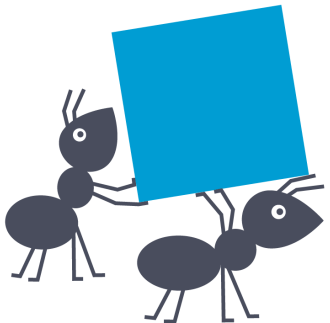
Janete Saldanha Bach is a Researcher at GESIS – Leibniz Institute for the Social Sciences, based in the Knowledge Technologies (KTS) Department, team FAIR Data and Human Information Interaction, working in the consortia KonsortSWD Project of the National Research Data Infrastructure (NFDI). She holds a Ph.D. and a Master's degree in Science and Technology Studies (STS) and a bachelor's degree in Information Science. Her research expertise is in Open Science, especially in research data management and data reuse in the Social Sciences. She is currently involved in consortium KonsortSWD, Task Area 5 Measure 1 - developing the conceptual framework for the PID registration service at a variable level and Task Area 5 Measure 2 Enhancing data findability.

Claus-Peter Klas

Claus-Peter Klas is lead of the Data & Service Engineering team in the department Knowledge Technologies for the Social Sciences of GESIS. He received his PhD in computer science at the University of Duisburg-Essen and was a postdoctoral researcher in the Department of Multimedia and Internet Applications, Faculty of Mathematics and Computer Science, University of Hagen, Germany. His research focuses on information retrieval, interactive information retrieval, information systems, databases, digital libraries, preservation and grid and cloud architectures. He developed the software Daffodil founded on a nation research project and worked in national and European research projects such as The European Film Gateway, SHAMAN (Sustaining Heritage Access through Multivalent ArchiviNg) and Smart Vortex (Scalable Semantic Product Data Stream Management for Collaboration and Decision Making in Engineering). He is currently responsible for several infrastructure projects within GESIS, such as da|ra, SowiDataNet or Missy, all concerned with providing information and data for social scientists. In addition, he lead the measure PID Services in the national research infrastructure project NFDI. In his team, they are developing an open source DDI suite to support getting DDI into operation.

Peter Mutschke

Peter Mutschke is deputy head of the department “Knowledge Technologies for the Social Sciences (KTS)" and leader of the team "FAIR Data and Human Information Interaction" of KTS. His research interests include Information Retrieval, Network Analysis and Open Science. He worked in a number of national and international research projects, such as the DFG projects DAFFODIL and IRM and the EU projects WeGov, SENSE4US, OpenMinTeD and MOVING. Peter served as a member of the management committee of the Leibniz research alliance “Science 2.0/Open Science” from 2013-2021. He founded and coordinates the GO FAIR Implementation Network "Cross-Domain Interoperability of Heterogeneous Research Data (Go Inter)", and he is member of the steering committee of the FAIR Digital Objects Forum (fairdo.org) where he also co-chairs a working group on semantics. He is currently involved in consortia KonsortSWD, NFDI4DataScience and BERD@NFDI of the National Research Data Infrastructure (NFDI). ORCID: https://orcid.org/0000-0003-3517-8071.

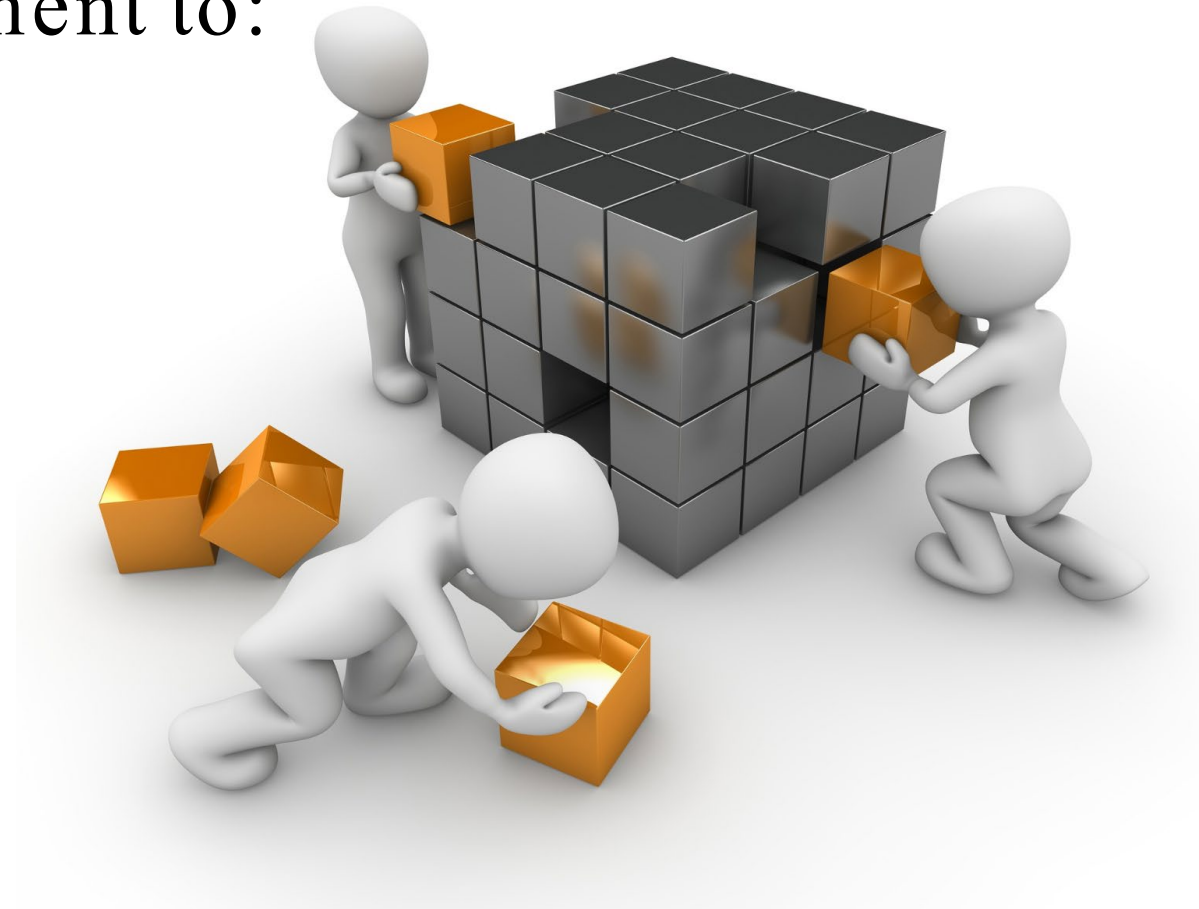


Agenda

- The PID Registration service for variables
  - Goal
  - The Research data granularity levels
  - Data formats: initial approach and future use
  - PID registration service provider
  - Use cases

Goal

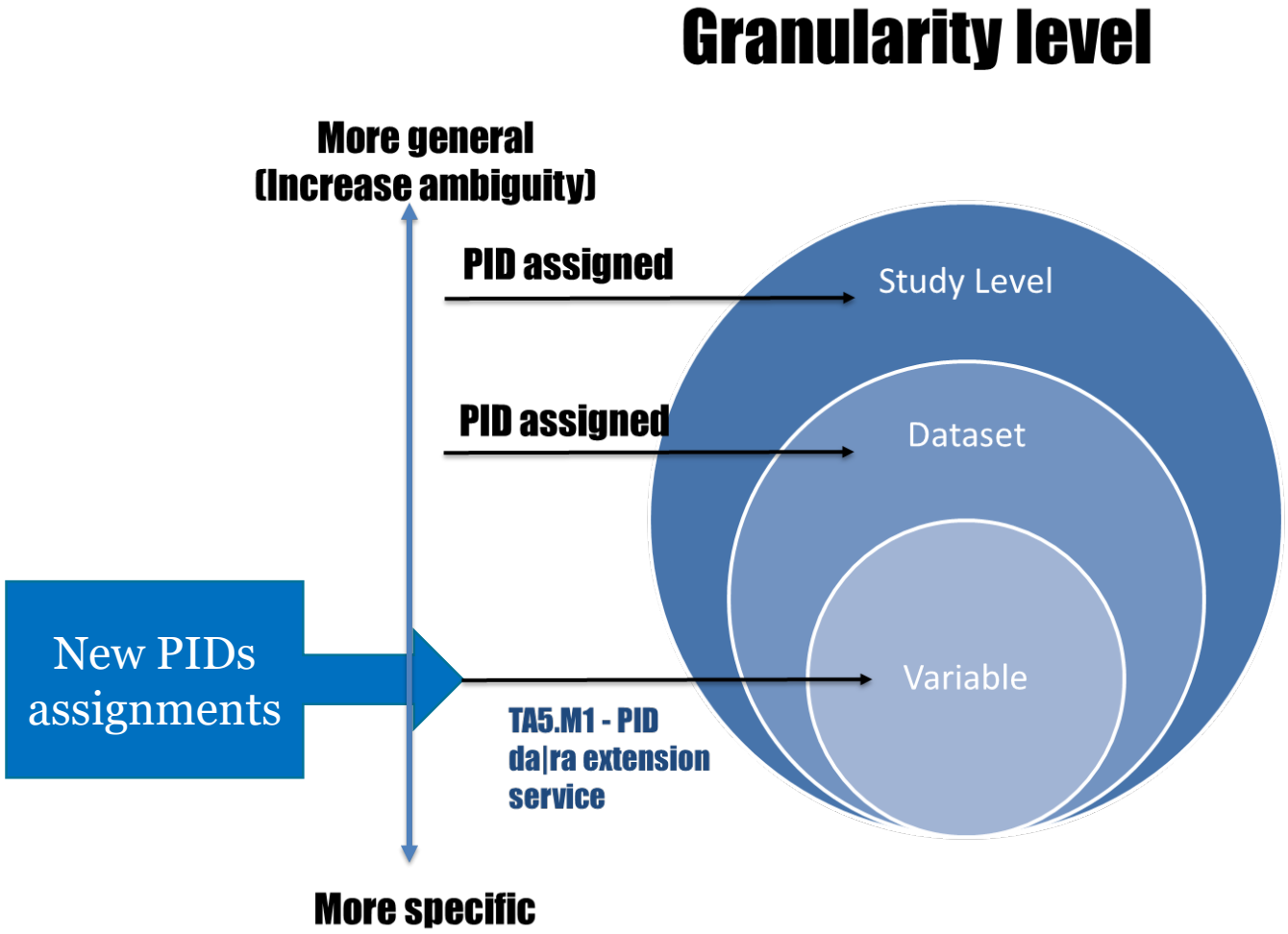
- Identify survey variables, **using one identifier**- the PID - will simplify FAIR data management to:
  - to boost subsequent citation,
  - get direct (meta)- data access, and
  - promote data reuse.



FAIR: Findability, Accessibility, Interoperability, and Reusability

Image by Peggy und Marco Lachmann-Anke from Pixabay

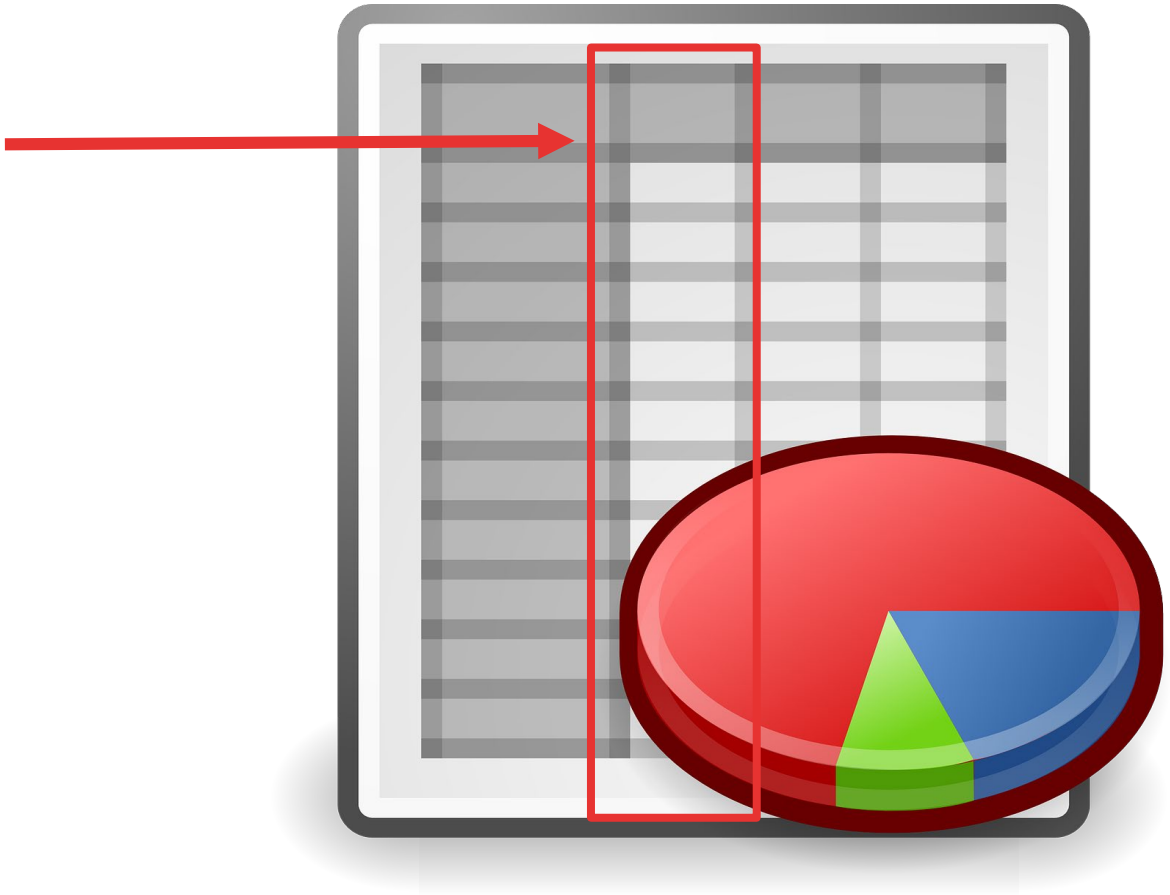
The Research data granularity levels



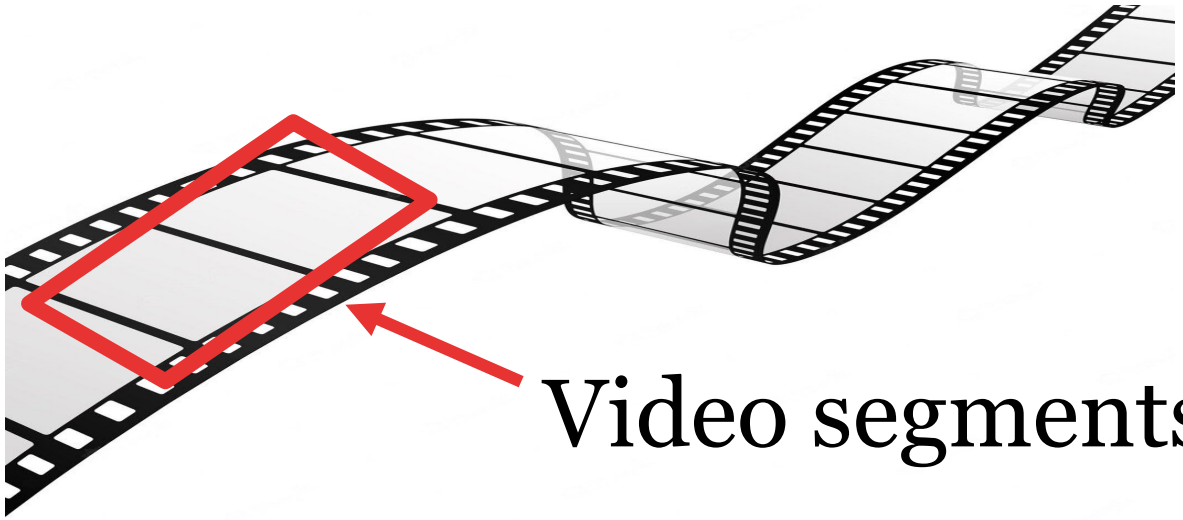
Klas, Claus-Peter, Zloch, Matthäus, Bach, Janete Saldanha, Baran, Erdal, & Mutschke, Peter. (2022). KonsortSWD Measure 5.1: PID service for variables report (1.0). Zenodo. https://doi.org/10.5281/zenodo.6397367

Data formats: initial approach

Variables in tabular data format

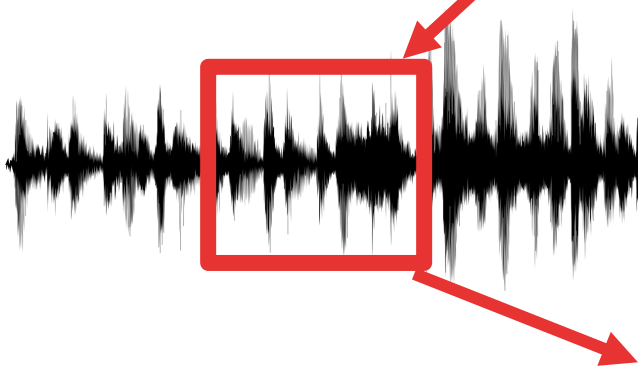


Data formats possible in the future



Video segments

Audio files



Transcripts

PID registration service provider

- This is a **da|ra** service **widening** and assigns a PID with **Handle** standard (ePIC);

- The service will be upgraded to **handle PIDs on variable** level;

da|ra

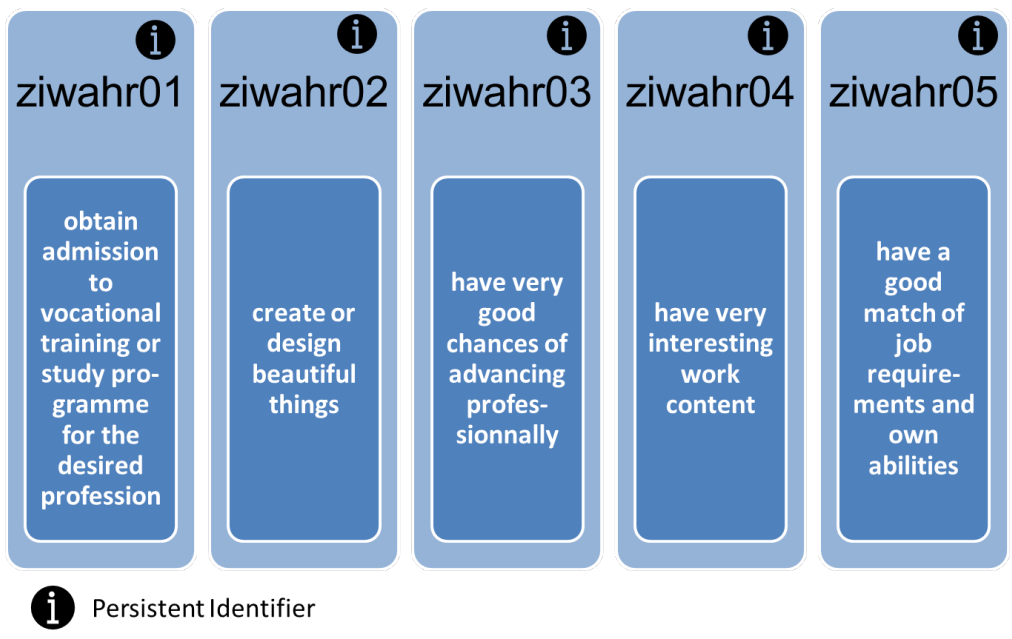
ePIC
Persistent Identifiers for eResearch

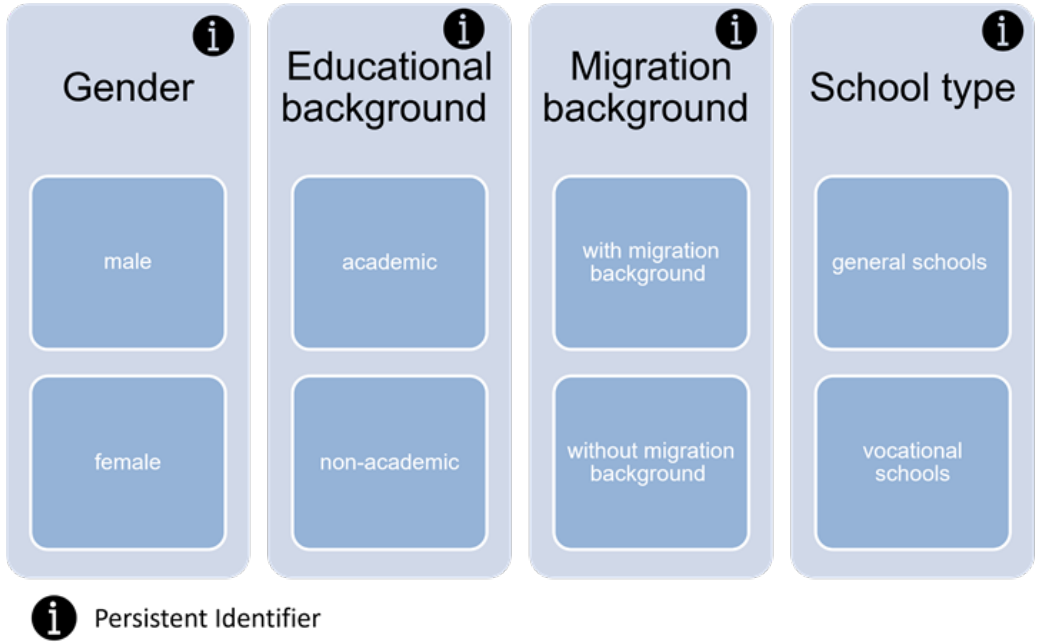
Use Case

DZHW
Deutsches Zentrum für Hochschul- und Wissenschaftsforschung

Higher Education Analytical Data System (HEADS) project at the DZHW needs a standard of data citation is to make its results widely usable and citable, particularly the entire information packages that comprise a central reporting variable (“indicator”) and the related multivariate analyses conducted in HEADS.

- PIDs for each variable, i.e., for each indicator or differentiation variables





Use Case

DZHW
Deutsches Zentrum für
Hochschul- und Wissenschaftsforschung

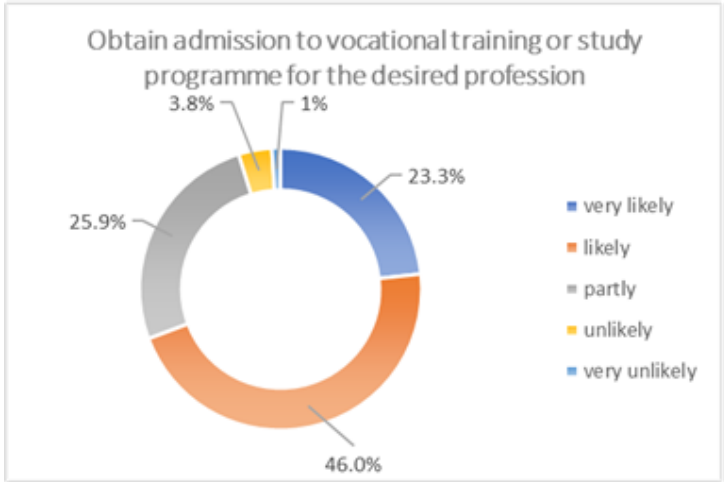
- PIDs for information packages

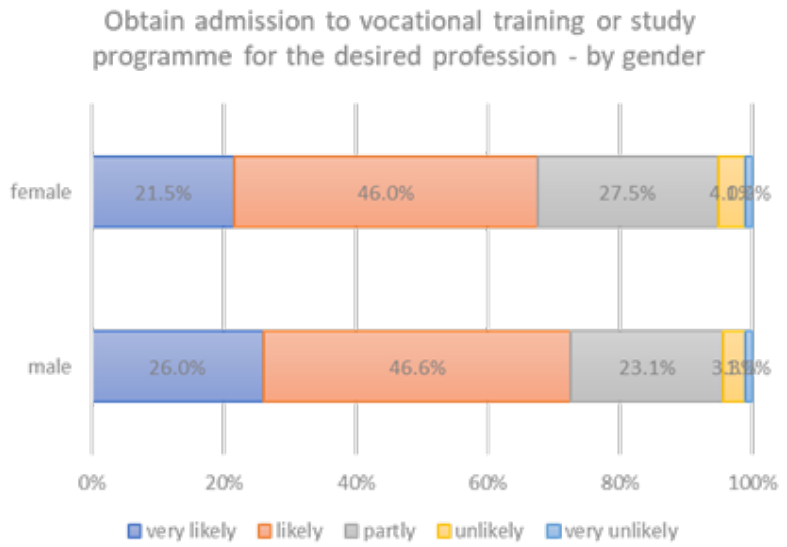
Students with university entrance qualifications in 2015 half a year before graduation

Indicator: Career goals - probability of realization
Question text: Everyone has goals. How likely do you think it is to achieve the following career goals?

Statement
obtain admission to vocational training or study programme for the desired profession (total value)

Feature
Gender





Options for differentiation variables
Gender
Educational background
Migration background
School type

Persistent Identifier

Use Case

gesis
Leibniz Institute
for the Social Sciences

- PIDs for more than 500,000 variables from 6,500 national and international studies covering various topics in the Social Sciences, Economics, and Behaviour Sciences.

Use Case

ONBound
Old And New Boundaries

HaSpaD

- PIDs for variables from harmonization tools and services
- Automatic access to variable data using scripts:
  - researchers are responsible for getting access to the datasets directly from the data providers;
  - With unique identifiers assigned, the data could be automatically accessed;
  - it makes it easier to use dozens of harmonised variables of the same topic from numerous diverse instruments.

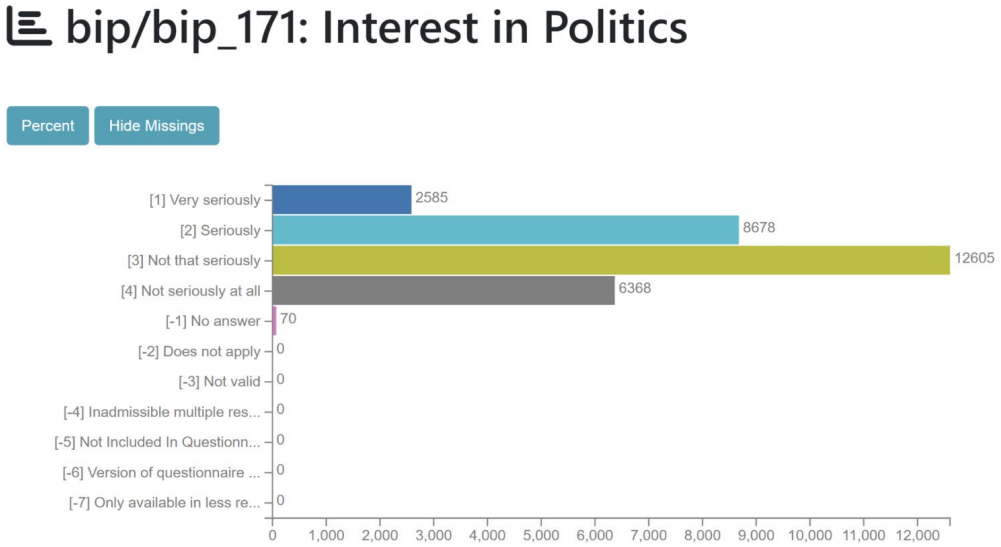
Use Case

SOEP The Socio-Economic Panel

Leben in Deutschland SOEP

Provide information on all household members: Germans living in the former eastern and western German states, foreign citizens, and immigrants residing in Germany. Some topics include household composition, occupational biographies, employment, earnings, health and satisfaction indicators.

- PIDs for 101.574 variables
  - available from 560 datasets,
  - distributed in 21.280 questions, and
  - 309 instruments



Variable graph: *bip/bip_171* Interest in Politics

https://paneldata.org/soep-core/datasets/bip/bip_171

Use Case

Qualiservice consists primarily of qualitative interview transcripts and context data in text, videos, and description data.

- PIDs are assigned at the file level for disambiguating similar data types and file naming;
- Provides a direct way of citing, identifying, and getting the target file.

Citation

Janete Saldanha Bach, Claus-Peter Klas and Peter Mutschke. 2023. KonsortSWD PID registration Service. In Workshop on PIDs within NFDI. *NFDI4 PID Working Group*, 27 Jan. 2023. 16 slides.

Acknowledgments

PID Service report
https://doi.org/10.5281/zenodo.6397367

Service demo application
https://youtu.be/fm8T-hlhsXg

The service is part of KonsortSWD project deliverable, NFDI funding number 442494171

nfdi

KonsortSWD

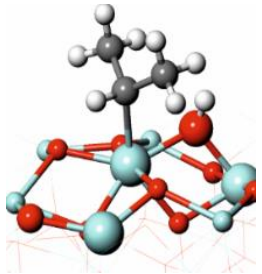
gesis
Leibniz Institute
for the Social Sciences

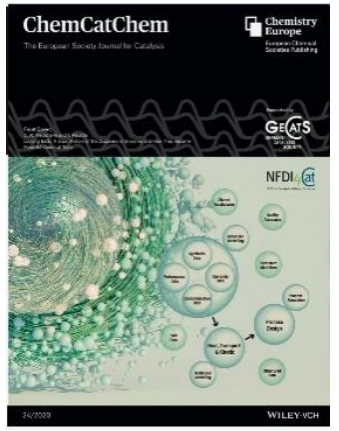
PIDs for Catalysis (NFDI4Cat)

2023-01-27, David Linke (LIKAT)

NFDI4Cat´s goal: A Data Value Chain in Catalysis

Syngas Plant
Linde Engineering





C. Wulf, M. Beller, T. Boenisch, O. Deutschmann, S. Hanf, N. Kockmann, R. Kraehnert, M. Oezaslan, S. Palkovits, S. Schimmler, S.A. Schunk, K. Wagemann, D. Linke*
A Unified Research Data Infrastructure for Catalysis Research - Challenges and Concepts, ChemCatChem 2021. DOI: 10.1002/cctc.202001974R2.

Why persistent identifier (PID)?

- PIDs are an important part of FAIR principles
  - *F1. (Meta)data are assigned a globally unique and **persistent identifier***
  - *A1. (Meta)data are **retrievable by** their **identifier** using a standardized communication protocol*
  - *I3. (Meta)data include **qualified references** to other (meta)data*
  - *R1. (Meta)data are richly described with a plurality of accurate and relevant attributes*
    - *R1.2. (Meta)data are associated with detailed **provenance***

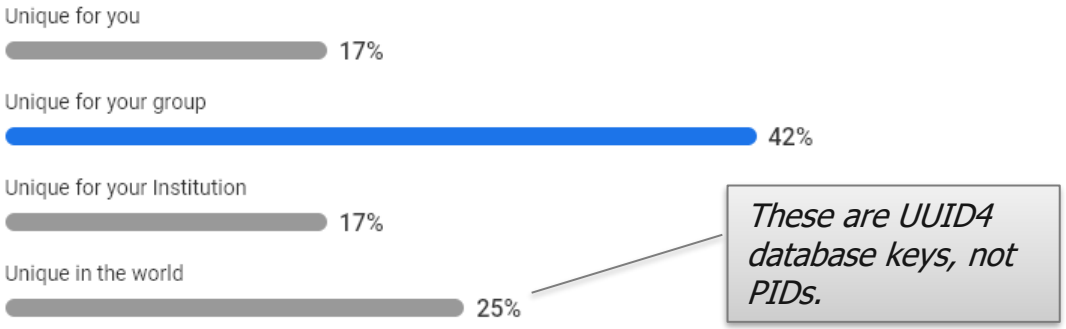
- Stable links to resources help to addresses the “reproducibility crisis” in science
- Links between data and other resources get more and more important to advance science.
- Long term preservation (findability) of the links to scientific resources must be assured.
  - Identifying resources only by its url in a storage system is too fragile.
  - Storage systems (“repositories”) will evolve/change over time.

- **Stable references (PIDs) are needed that are resolvable via a commonly agreed process.**

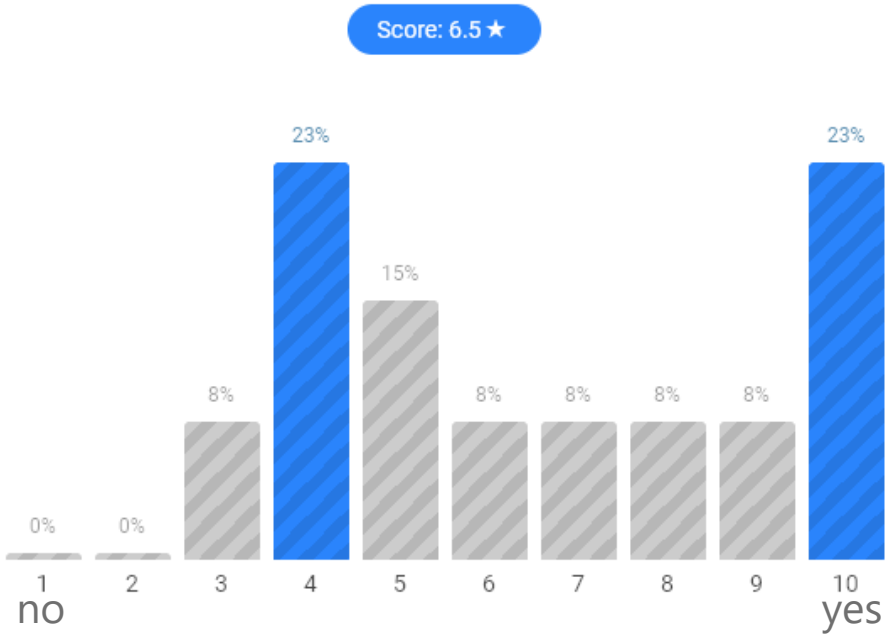
Online-Survey on PIDs from NFDI4Cat Consortium Meeting

Survey from Feb-2022

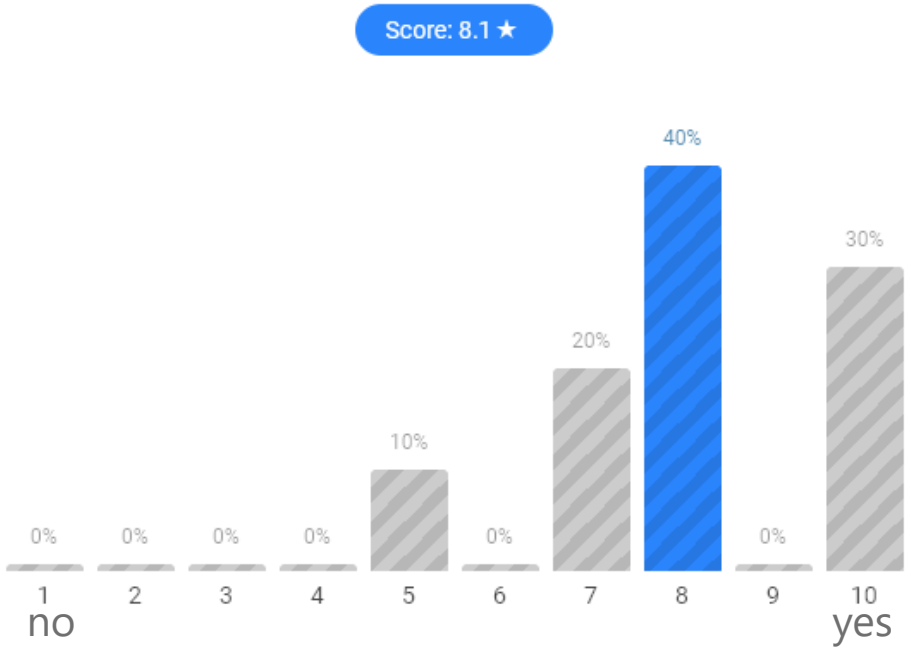
How unique are the IDs you use?



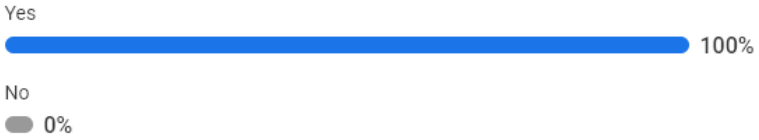
Will the sample IDs and all information associated with the IDs remain accessible and findable after you leave?



Would you switch to using Catalyst Ressource IDs if you can create and access them easily?



Are you interested in being part of a working group that promotes this topic?



=> As a result, a working group on PIDs was formed in March 2022.

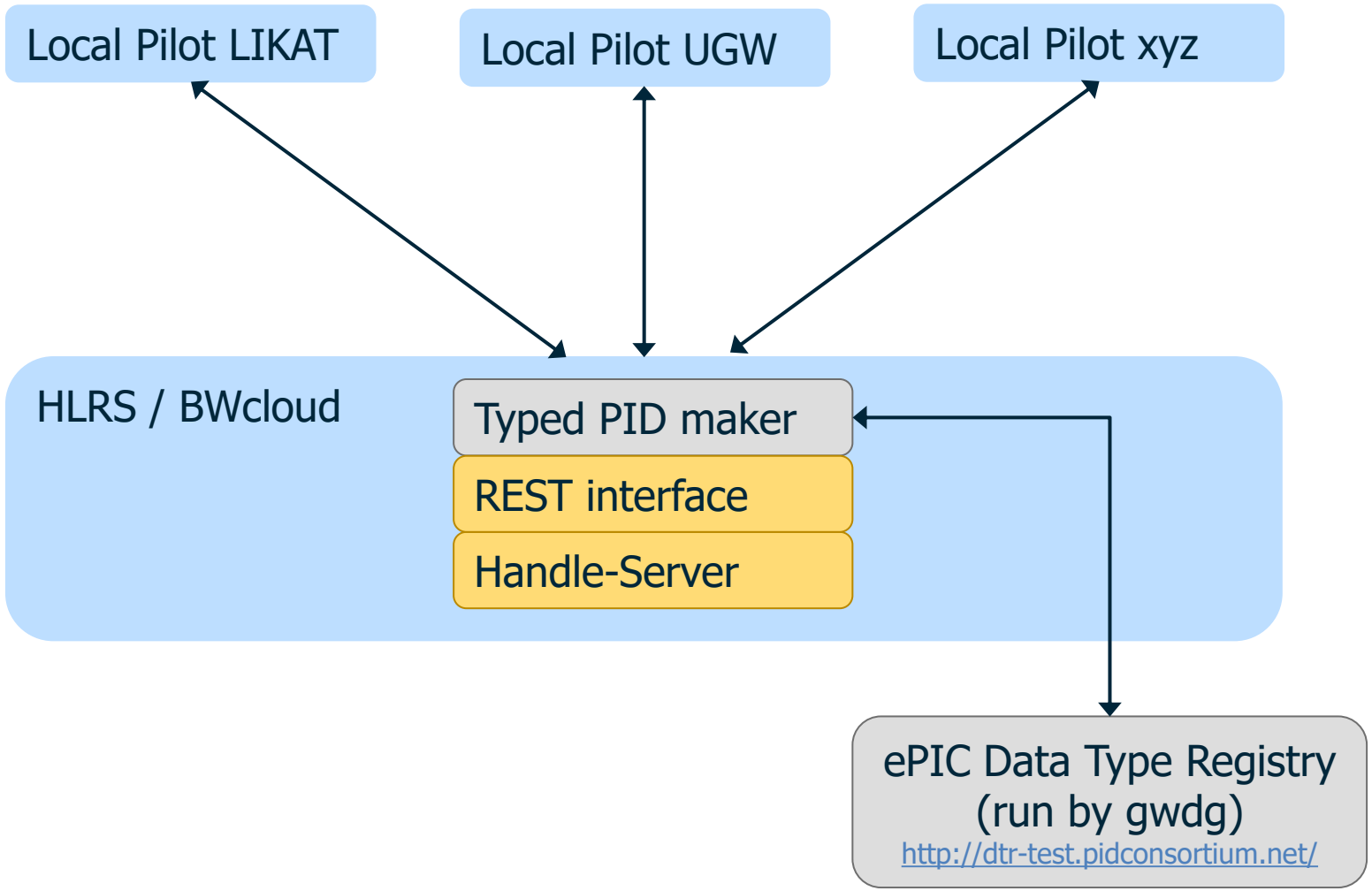
What do we have? What do we still need?

- There are already well established systems that should be used:
  - **DOI**s for publications, software, data collections
  - **ORCID**s for identifying researchers
  - **ROR**s for identifying institutions (e.g. https://ror.org/029hg0311)

- **There is no established PID system for other resources like samples, materials, devices, instruments, models etc.** which
  - supports **assigning PIDs in the early research phase** and can later be cited,
  - can be adjusted (by NFDI4Cat) to the metadata requirements for catalysis resources,
  - can cope with metadata requirements that differ between the resources,
  - has a low barrier of entry,
  - and low costs per PID.

- **IGSN & RRID cannot cover catalysis needs** but are excellent examples to learn from.

Proposed Solution: handle-system-based PIDs



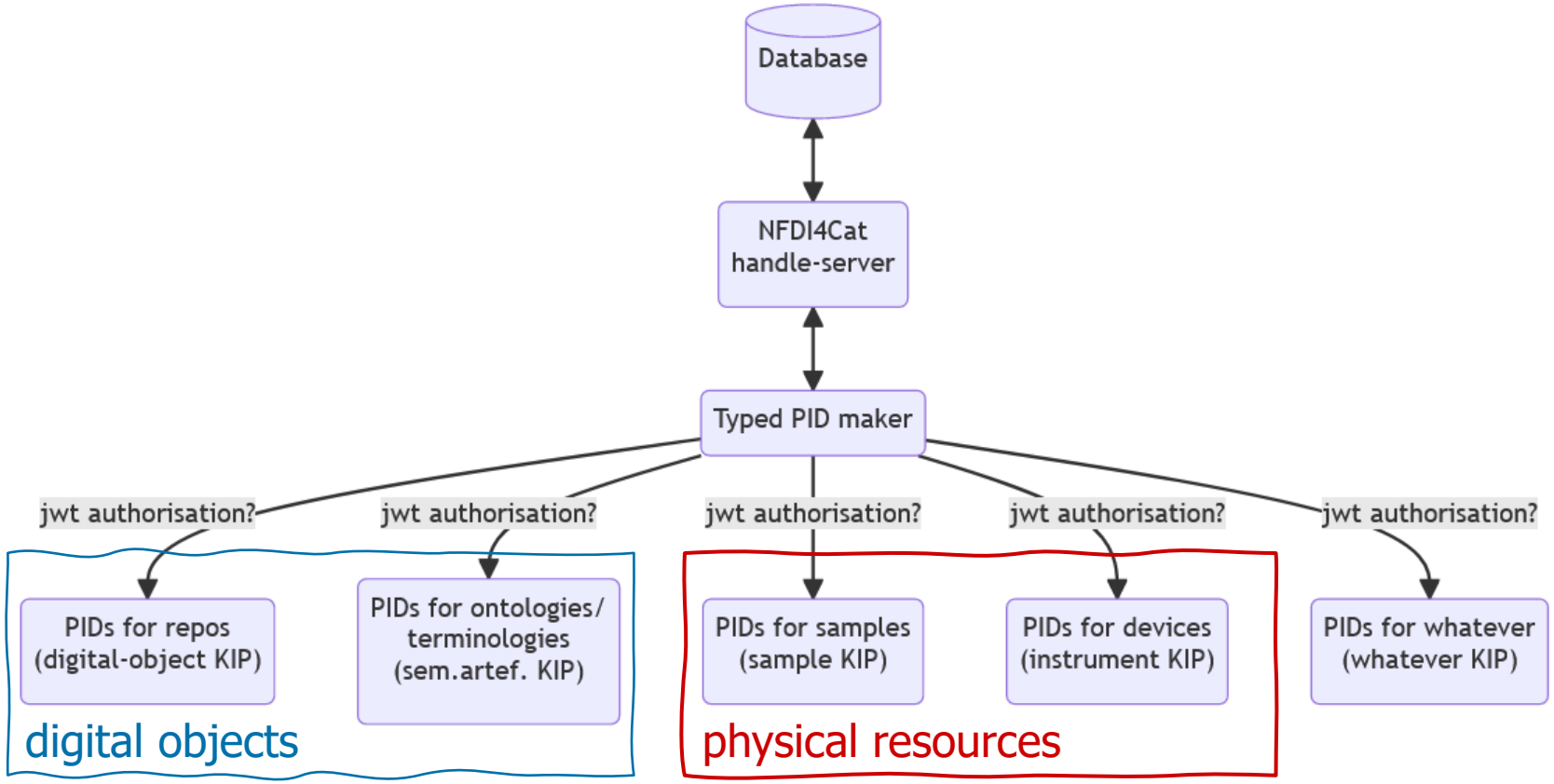
Several different institutional RDM systems (e.g. ELNs) will get PIDs from our central handle server.
Researchers can use these PIDs from the first lab book entry to final publication.

We will test **Typed PID Maker** from KIT/SCC for

- PID schema validation
- Authorization

<u>Fall-back alternative:</u>
Add fastapi REST interface to B2Handle/pyHandle (from ePIC)

Different PID (types) for different use cases in NFDI4Cat



Open: How many different types of PIDs do we want/need?

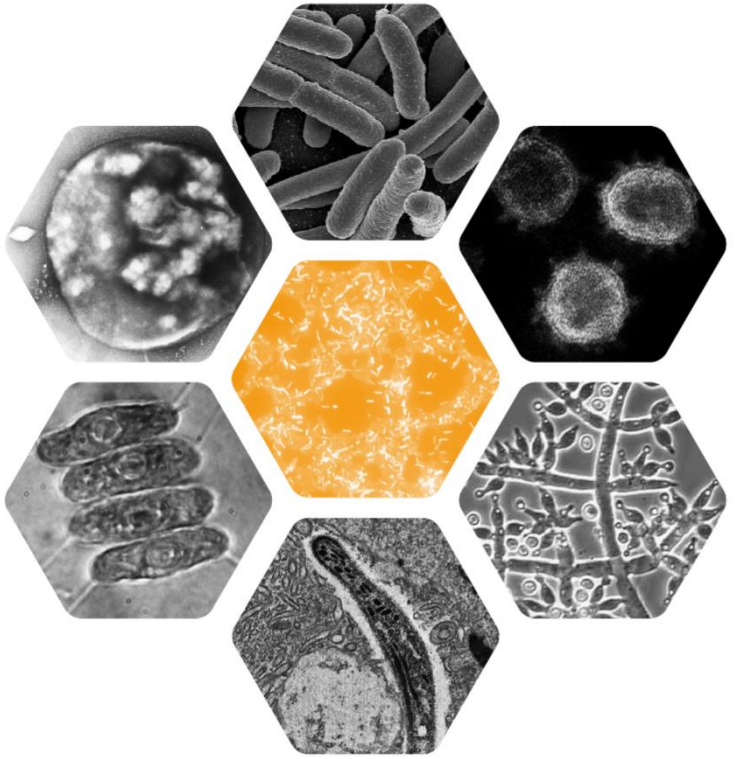
It is probably sensible to support at least two: one for physical resources & one for digital objects.

Stable URIs (PIDs?) for Concepts in Terminologies & Ontologies

- In principle PIDs could be used instead of urls.
  - » Idea was discussed with some participants in FAIR-impact event & members of RDA Fair-Digital Objects group
    - Idea is not new (implemented in a French biomedical project).
    - Most(?) experts consider this as out-of-scope use of PIDs.

- Current plan: use **w3id.org** as provider for permanent urls.
  - » https://w3id.org/nfdi4cat/ has been registered.
  - » We are going to use IDs as suffix-parts of the url.
    - e.g. for the photo-catalysis vocabulary https://w3id.org/nfdi4cat/**voc/photocatalysis/concept12345**

StrainInfo:
PIDs for microbial strains



Version 2023-01-27

Lorenz Reimer
DSMZ

NFDI4 MICROBIOTA

NCBI

ENA European Nucleotide Archive

JGI

DDBJ DNA Data Bank of Japan

BacDive

ATCC

INSTITUT PASTEUR

JCM

CCUG

DSMZ

NFDI4 MICROBIOTA

Find a microbial strain and information about it?

NCBI

ENA European Nucleotide Archive

JGI

DDBJ DNA Data Bank of Japan

BacDive

ATCC

INSTITUT PASTEUR

JCM

CCUG

DSMZ

NFDI4 MICROBIOTA

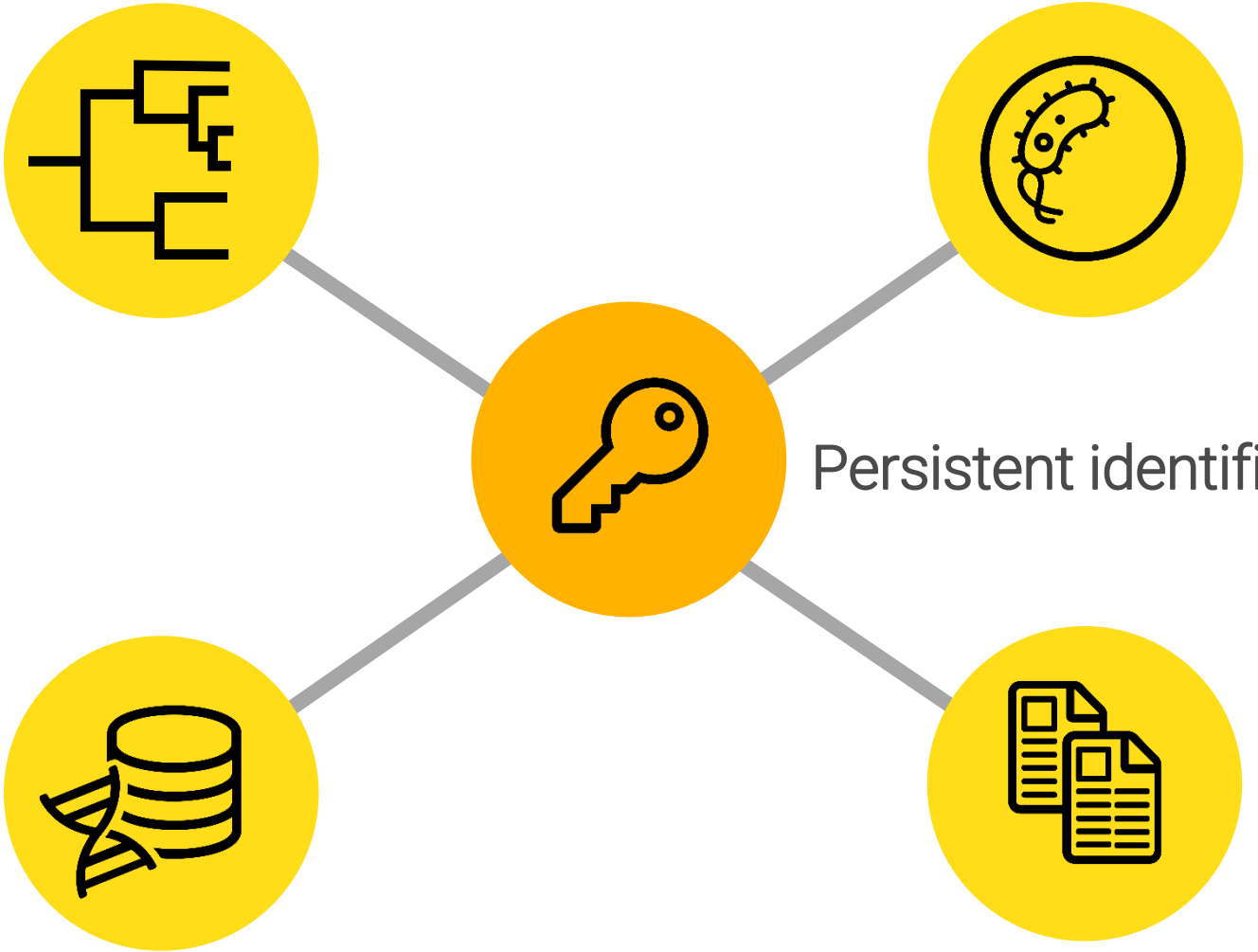
LPSN .dsmz.de

NCBI

Streptomyces hiroshimensis

- *Actinomyces biverticillatus*
- *Streptomyces fervens*
- *Streptomyces spitsbergensis*
- *Streptoverticillium salmonicida*
- *Streptomyces fervens subsp. melrosporus*

How to connect distributed data?



Home › Microbe Products › Bacteriology and Archaea › Anaerobes › 15703

Bifidobacterium adolescentis Reuter

15703™

DOWNLOAD GENOME

LEARN ABOUT OUR ENHANCED AUTHENTICATION INITIATIVE >

Bifidobacterium adolescentis strain E194a (Variant a) is a whole-genome sequenced type strain that was isolated from the intestine of an adult. It has applications in bioinformatics and is a microbiome standard component.

96/100 Bioz Stars 229 Product Citations

Product category	Bacteria
Product type	Anaerobe
Strain designation	E194a (Variant a)
Type strain	Yes
Genome sequenced strain	Yes
Isolation source	Intestine of adult
Applications	Bioinformatics
Product format	Freeze-dried
Storage conditions	2°C to 8°C

SEE ADDITIONAL PRODUCT INFORMATION >

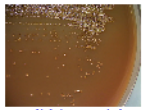
https://www.atcc.org/products/15703

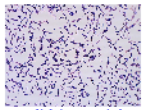
JCM Catalogue

Bifidobacterium adolescentis Reuter 1963

Taxonomy in NCBI database: *Bacteria*; *Actinobacteria*; *Actinomycetia*; *Bifidobacteriales*; *Bifidobacteriaceae*.

1275^{T} <-- ATCC 15703 <-- G. Reuter E194a (Variant a).
Accessioned in 1982.
=ATCC 15703 =BCRC 14606 =CCM 4987 =CCUG 17359 =CCUG 18363 =CCUG 45213 =CECT 5781 =CGMCC 1.2190 =CIP 64.59 =DSM 20083 =KCTC 3216 =LMG 10502 =NCTC 11814 =VTT E-981074.
Type strain [596].
Medium: 13; Temperature: 37°C; Anaerobic; Rehydration fluid: 663.


open link in new window


open link in new window

Source: Intestine of adult [110].
Biochemistry/Physiology: [3093,3126,4424].
Cell wall: Lys(Orn)-Asp [010,3126].
Enzyme electrophoretic profile: [3126,4425].
G+C (mol%): 58.9 (Tm) [181].
DNA-DNA relatedness: [023,3126].
Phylogeny: 16S rRNA (M58729), 16S rRNA gene (AB437354, LC071806), *hsp60* (AF210319) [5374].
DNA typing: RELP [4423], rRNA gene restriction pattern [4597], genome-wide tiling array [9039].
Other taxonomic data: Protein profile [1865].
Genome sequence: AP009256, UAQG00000000.
More information: Species-specific primer [4521].
NCBI Taxonomy ID: 1680.
Genomic DNA is available from RIKEN BRC-DNA Bank: JGD 07423.

Publication(s) using this strain [A04015, A04018, A04079, A04152, A04168, A04196, A05011, A05036, A05048, A06033, A07171, B07183, A08019, A08201, A08244, A08278, A08304, A09019, A09026, A09198, A09201, A10090, A10101, A10294, A11008, A11225, A11236, A11237, A11252, A11297, A12287, A12408, A12422, A12441, A13116, A13170, A13553, A14435, A15097, A15541, A16131, A17103, A17549, A18008, A18010, A18152, A18169, A18418, A18569, A19161, A19175, A19522, A19531, A20144, A20166, A20446, A20562, A21209, A21447, A21524, A21549].

Patent publication(s) using this strain [WO2007/020884, WO2007/114378, WO2007/099997, WO2007/023588, JP2007-112805A, JP2008-074768A, WO2008/062555, JP2009-125055A, JP2009-084215A, JP2009-242430A, WO2009/028253, WO2011/027875, WO2012/005240, JP2012-180288A, JP2012-180373A, JP2013-042749A, JP2014-055194A, 2015-198638, 2015-96555, 2016-028595, WO2016/103699, CN105671120, 2017-216979, 2019-94307, 2019-92469, 2020-132560, 2020-92704].

https://www.jcm.riken.jp/cgi-bin/jcm/jcm_number?JCM=1275

NFDI4 MICROBIOTA

Springer Link

Open Access | Published: 21 April 2017

Bifidobacterium adolescentis (DSM 20083) and *Lactobacillus casei* (Lafti L26-DSL): Probiotics Able to Block the In Vitro Adherence of Rotavirus in MA104 Cells

Karem Prunella Fernandez-Duarte, Nury Nathalia Olaya-Galán, Sandra Patricia Salas-Cárdenas, Jazmin Lopez-Rozo & I

Probiotics and A

2806 Accesses

Abstract

Rotavirus is tl
Even though t
strategies for
are constantly
order to demc
rotavirus infe
tested in MA1
treatment was
bacteria show

Springer Link

Original Paper | Published: 22 January 2004

Physico-chemical and transglucosylation properties of recombinant sucrose phosphorylase from *Bifidobacterium adolescentis* DSM20083

L. A. M. van den Broek, E. L. van Boxtel, R. P. Kievit, R. Verhoef, G. Beldman & A. G. J. Voragen

Applied Microbiology and Biotechnology **65**, 219–227 (2004) | Cite this article

1013 Accesses | **60** Citations | **6** Altmetric | Metrics

Abstract

Clones of a genomic library of *Bifidobacterium adolescentis* were grown in minimal medium with sucrose as sole carbon source. An enzymatic fructose dehydrogenase assay was used to identify sucrose-degrading enzymes. Plasmids were isolated from the positive colonies and sequence analysis revealed that two types of insert were present, which only differed with respect to their orientation in the plasmid. An open reading frame of 1,515 nucleotides with high homology for sucrose phosphorylases was detected on these inserts. The gene was designated *SucP* and encoded a protein of 56,189 Da. *SucP* was heterologously expressed in *Escherichia coli*, purified, and characterized. The molecular mass of SucP was 58 kDa, as estimated by SDS-PAGE, while 129 kDa was found with gel permeation, suggesting that the native enzyme was a dimer. The enzyme showed high activity towards sucrose and a lower extent towards α-glucose-1-phosphate. The transglucosylation properties were investigated

https://link.springer.com/article/10.1007/s12602-017-9277-7

https://link.springer.com/article/10.1007/s00253-003-1534-x

ORIGINAL RESEARCH article

Front. Microbiol., 08 April 2022
Sec. Food Microbiology
https://doi.org/10.3389/fmicb.2022.860014

Study on the Biochemical Characterization and Selectivity of Three β-Glucosidases From *Bifidobacterium adolescentis* ATCC15703

Yanbo Hu[1*], Liy

[1] School of Food Science and Engin
[2] School of Life Sciences, Changchu

Three β-glucosidases fr
were overexpressed in *E*
affinity chromatography
fold and specific activity
β-glucopyranoside (*p*NP
suggested that BaBgl1A
β-1,3-glucosidase and β
enzymes was further stu
chromatography and hi
highest bioconversion a
into the rare ginsenosid
where it hydrolyzed bot
was not active on Rb1 a

https://www.frontiersin.org/articles/10.3389/fmicb.2022.860014/full

Food Science & Nutrition Open Access

ORIGINAL RESEARCH | Open Access

Combining of transcriptome and metabolome analyses for understanding the utilization and metabolic pathways of Xylo-oligosaccharide in *Bifidobacterium adolescentis* ATCC 15703

Jian Yang, Qilong Tang, Lei Xu, Zhijiang Li, Yongqiang Ma, Di Yao

First published: 30 September 2019 | https://doi.org/10.1002/fsn3.1194 | Citations: 9

SECTIONS PDF TOOLS SHARE

Abstract

A combination of transcriptome and metabolome analyses was applied to understand the utilization and metabolism of Xylo-oligosaccharide (XOS) in *Bifidobacterium adolescentis* 15703 as well as identifying the key regulatory-related genes and metabolites. Samples of cultures grown on either XOS or xylose were collected. The transcript and metabolite profiles were obtained from high-throughput RNA-sequencing data analysis and UHPLC system. Compared with xylose, XOS highly promoted the

https://onlinelibrary.wiley.com/doi/10.1002/fsn3.1194

NFDI4 MICROBIOTA

Datasets Taxonomy **Genome** Gene Command-line tools Documentation

Datasets / Genome / ASM1042v1

Genome assembly ASM1042v1

Download datasets curl

Reference sequence	RefSeq GCF_000010425.1
Submitted sequence	GenBank GCA_000010425.1
Taxon	*Bifidobacterium adolescentis* ATCC 15703
Strain	ATCC 15703
Submitter	Gifu University, Life Science Research Center, Japan
Date	Dec 5, 2006

View the legacy Assembly page

Assembly statistics

These statistics describe the RefSeq genome sequence GCF_000010425.1

Genome size	2.1 Mb
Number of chromosomes	1
Number of scaffolds	1
GC percent	59
Assembly level	Complete Genome

https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000010425.1/

Datasets Taxonomy **Genome** Gene Command-line tools Documentation

Datasets / Genome / 49964_F01

Genome assembly 49964_F01

Download datasets curl

Reference sequence	RefSeq GCF_900445615.1
Submitted sequence	GenBank GCA_900445615.1
Taxon	*Bifidobacterium adolescentis*
Strain	NCTC11814
WGS project	UAQG01
Submitter	SC
Date	Jun 28, 2018

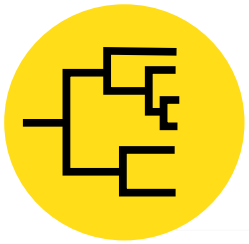
View the legacy Assembly page

Assembly statistics

These statistics describe the RefSeq genome sequence GCF_900445615.1

Genome size	2.2 Mb
Number of contigs	57
Contig N50	238.5 kb
Contig L50	3
GC percent	59
Assembly level	Contig

https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_900445615.1/



NFDI4 MICROBIOTA

LPSN .dsmz.de

Browse by rank

Advanced search
Subscribe

Main
Introduction
Navigation
Nomenclature
Etymology
Collections
Acknowledgements
Copyright
FAQ

Submit
Contact
Statistics
Downloads

Species *Bifidobacterium adolescentis*

ⓘ **Name:** *Bifidobacterium adolescentis* Reuter 1963 (Approved Lists 1980)

ⓘ **Category:** Species

ⓘ **Proposed as:** sp. nov.

ⓘ **Etymology:** ad.o.le.scen.tis. L. n. *adolescens -tis*, adolescent; L. gen. masc

ⓘ **Gender:** neuter

ⓘ **Type strain:** ATCC 15703; BCRC 14606; CCRC 14606; CCUG 17359; CCU

ⓘ See detailed strain information at **BacDive**

ⓘ Conduct genome-based taxonomy at TYGS

ⓘ **16S rRNA gene:** M58729 Analyse FASTA ENA N

ⓘ **Effective publication:** Reuter G. Vergleichenden Untersuchung über die Bi *Infektionskrankheiten und Hygiene. Abteilung I* 1963; **191**:486-507.

ⓘ **IJSEM list:** Skerman VBD, McGowan V, Sneath PHA. Approved lists of bac

ⓘ **Nomenclatural status:** validly published under the ICNP

ⓘ **Taxonomic status:** correct name

ⓘ **Risk group:** 1

ⓘ ▼ **Synonyms:**

Name	Kind
Bifidobacterium stercoris Kim *et al.* 2010	heterotypic synonym, validly

https://lpsn.dsmz.de/species/bifidobacterium-adolescentis

NCBI Taxonomy Browser

Entrez PubMed Nucleotide Protein Genome Structure

Search for as complete name lock Go Clear

Display 3 levels using filter: none

Bifidobacterium adolescentis ATCC 15703

Taxonomy ID: 367928 (for references in articles please use NCBI:txid367928)

current name

Bifidobacterium adolescentis ATCC 15703

equivalent: **Bifidobacterium adolescentis str. ATCC 15703**

NCBI BLAST name: **high G+C Gram-positive bacteria**
Rank: **strain**
Genetic code: Translation table 11 (Bacterial, Archaeal and Plant Plastid)

Lineage (*full*)

cellular organisms; Bacteria; Terrabacteria group; Actinobacteria; Actinomycetia; Bifidobacteriales; Bifidobacteriaceae; Bifidobacterium; Bifidobacterium adolescentis

Comments and References:

genome sequence
Determination of the DNA genome sequence of this strain has been or is being determined either in whole or in part.

https://www.ncbi.nlm.nih.gov/data-hub/taxonomy/1680/

How to connect distributed data?

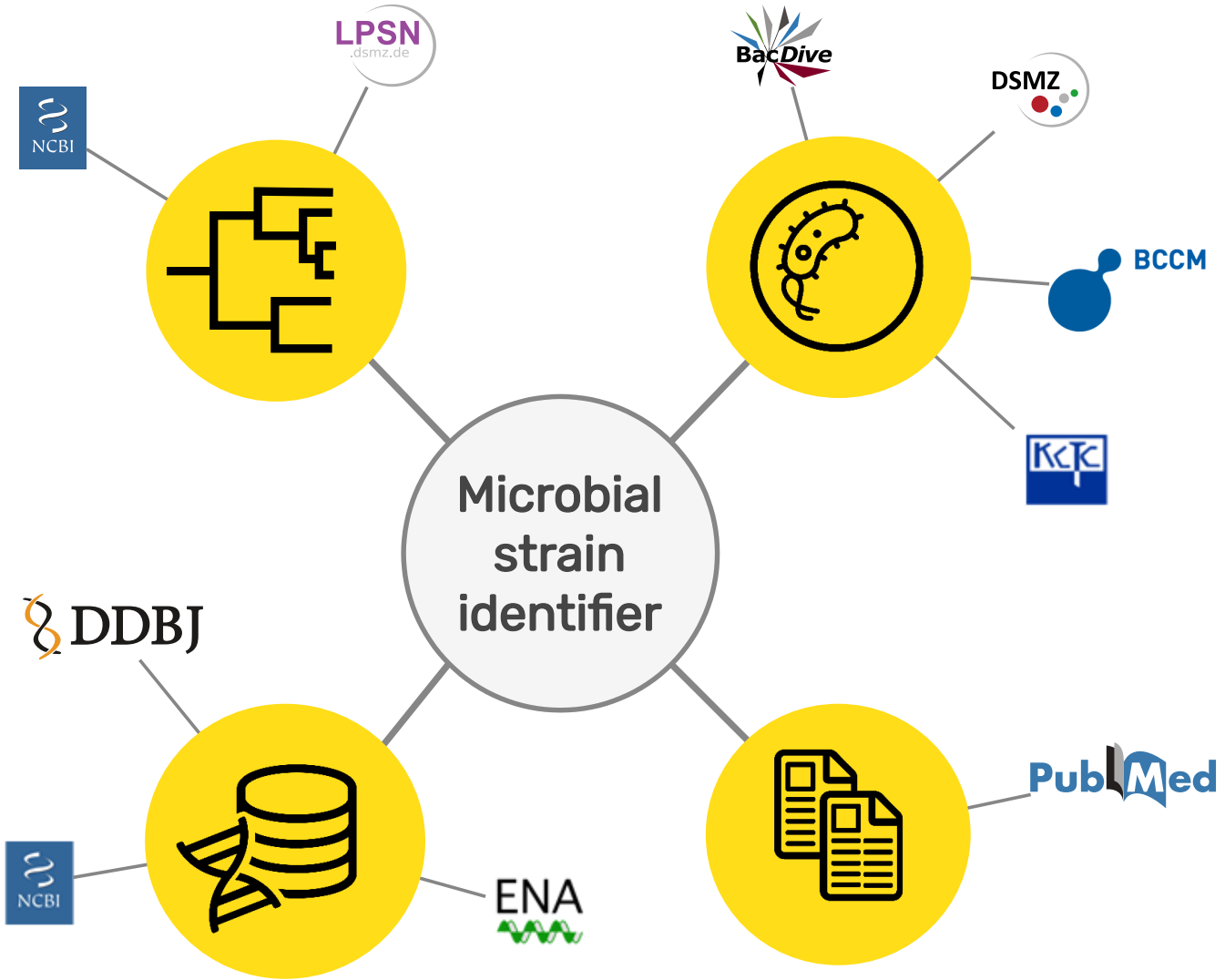
Persistent identifier:

- Culture collection identifier
- INSDC sequence accession no.
- DOIs (mostly in literature)

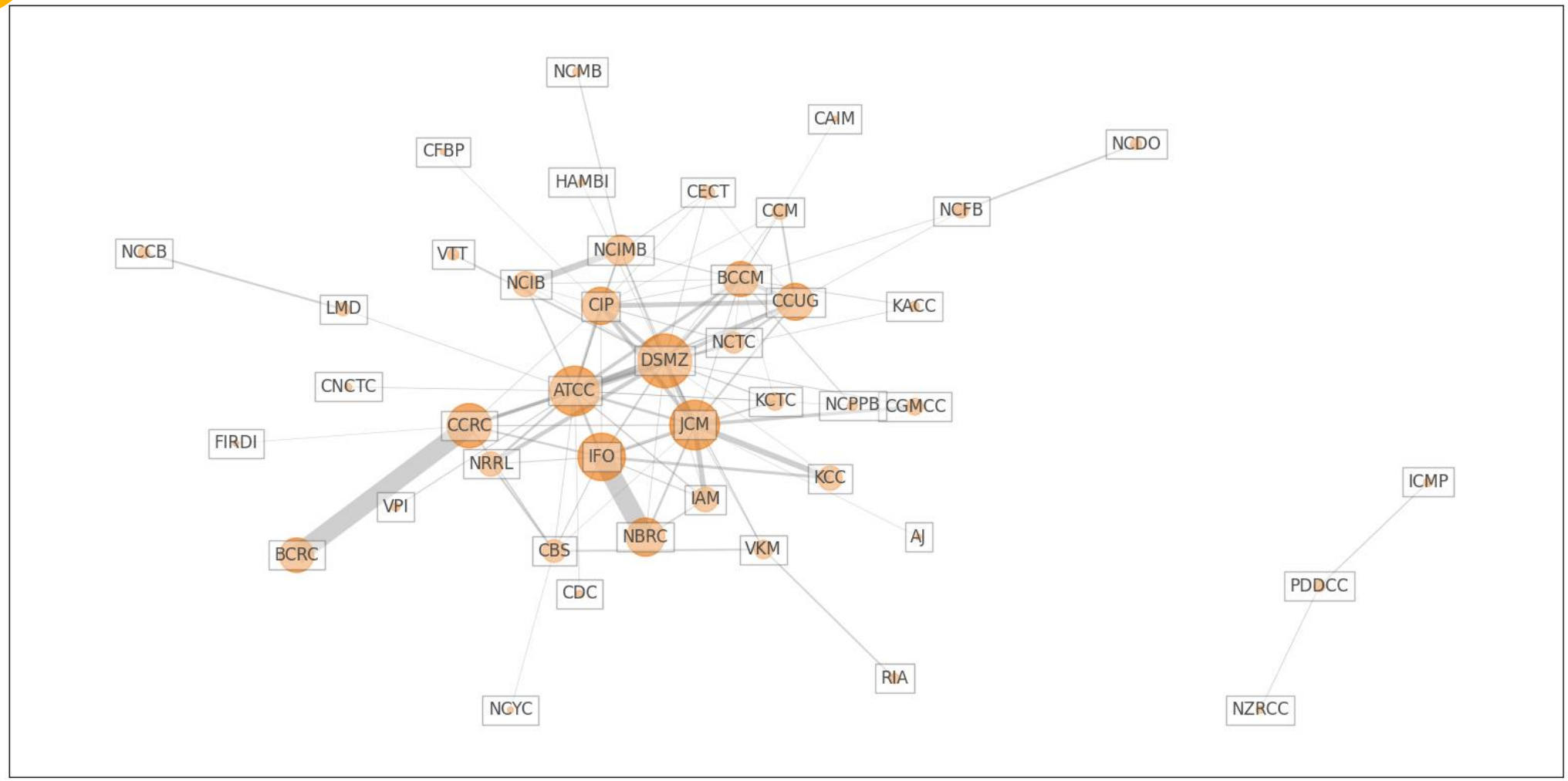
Problem to solve:

- Collect and connect them!
- Build a central database/registry

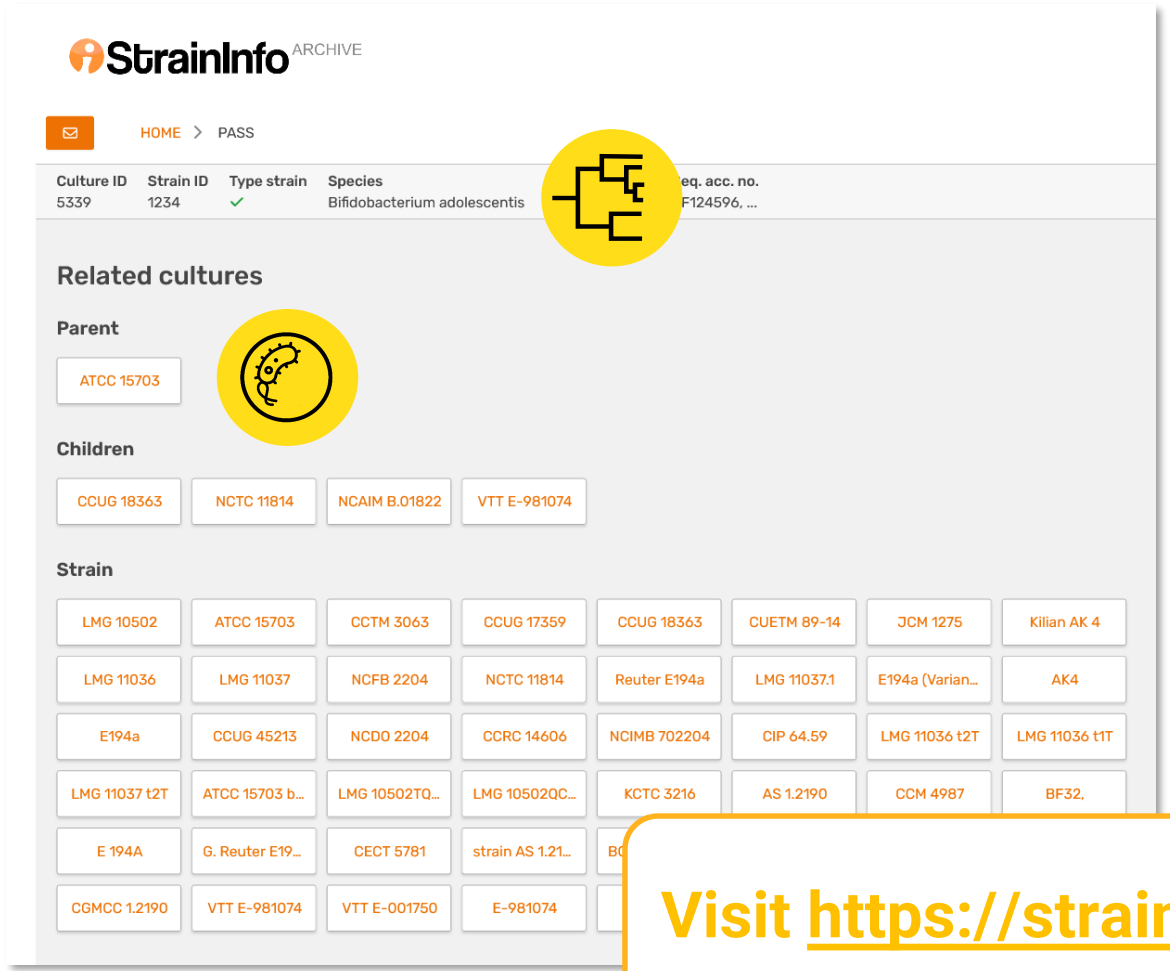
Use relations to connect data

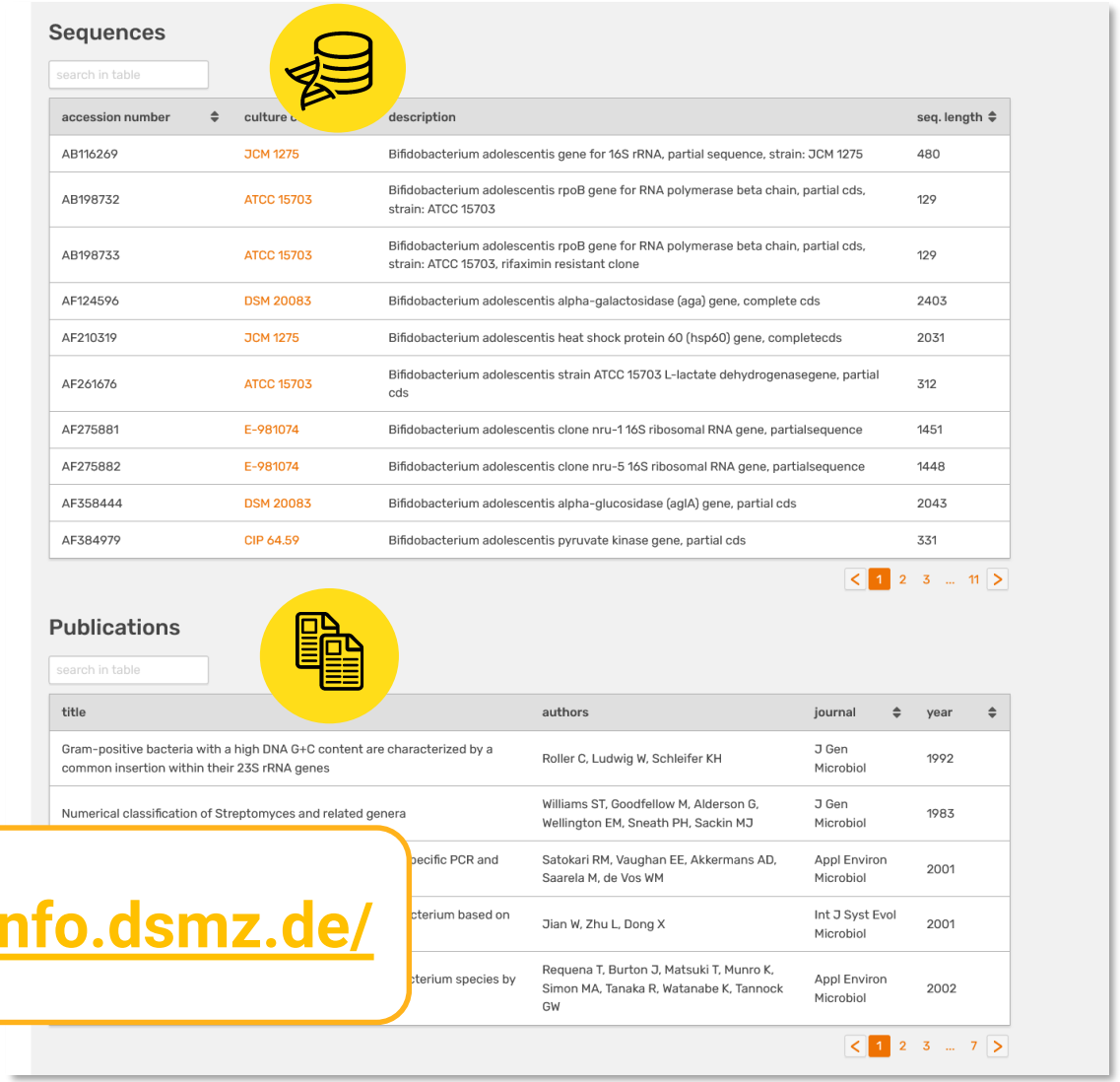


Reveal the relations



Provide a Webinterface & an API





Visit https://straininfo.dsmz.de/

Thank you!

Artur Lissin
artur.lissin@dsmz.de

NFDI4 MICROBIOTA

Lorenz Reimer
lorenz.reimer@dsmz.de

StrainInfo ARCHIVE

https://straininfo.dsmz.de/

www.nfdi4microbiota.de
@nfdi4microbiota

nationale
Forschungsdaten
Infrastruktur
for CULTURE

27th January 2023 | Workshop on PIDs within NFDI

PIDs in NFDI4Culture

Sandra Göller | FIZ Karlsruhe
Desiree Mayer | SLUB Dresden

Deutsche Forschungsgemeinschaft

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - 441958017

Overview:

- Which are the most important PIDs?
- Authority Data as PIDs
- PIDs in RADAR4Culture
- Status quo in NFDI4Culture/Desiderata/Discussions

Most important PIDs in NFDI4Culture

Highly sustainable and reliable
Controlled Standards from National Libraries

Authority Data, e.g.:

LIBRARY OF CONGRESS · BnF · VIAF · GND

Highly sustainable and reliable
Open to public / community driven

Structured Data, e.g.:

GeoNames · WIKIDATA · ORCID · ROR

metadata

(research)-data

PIDs in Repositories:

- make relations between publications and research data clear, controlled fields in the meta data schema are necessary.
- PIDs in controlled fields also enable search within a repository and from outside

Usage of PIDs within NFDI4Culture Repositories – An Example with RADAR4Culture

RADAR4Culture

- publishing service for research data on tangible and intangible cultural assets
- provided by FIZ Karlsruhe - Leibniz Institute for Information Infrastructure

RADAR Metadata Schema

- based on the DataCite metadata schema kernel 4.4
- Metadata Schema v9.1: 10 mandatory metadata fields (including 6 mandatory fields for the **DOI** registration), 13 optional metadata fields
- combination of free text fields and options for standardised or normed entries:
  - individuals (**ORCID**)
  - organisations (**ROR**)
  - funding organisations (**Crossref Funder Registry**)
  - keywords (Gemeinsame Normdatei, **GND**)

In Discussion

- GND-IDs as additional alternative for ORCID
- further authority data / vocabulary IDs for keywords
- vocabulary IDs for subject areas

Dataset: Example

RADAR Metadata | Content | Statistics | Technical Metadata

Creator/Author:	Göller, Sandra https://orcid.org/0000-0003-4553-3671 [FIZ Karlsruhe – Leibniz Institute for Information Infrastructure]
Title:	Example
Keywords:	Forschungsdaten
Publishers:	FIZ Karlsruhe – Leibniz Institute for Information Infrastructure
Production year:	2023
Subject areas:	Information Technology
Resource type:	Dataset
Publication year:	2023
Rights holders:	FIZ Karlsruhe – Leibniz Institute for Information Infrastructure
Funding:	Deutsche Forschungsgemeinschaft

Status quo NFDI4Culture:

- DOIs for publications; URN for digitizations (individual pages with permalinks)

Desiderata

- More controlled fields in metadata schemas of repositories should require PIDs
- More terms in subject headings or improved access to create subject headings in authority data (e.g. in GND)

Recent Discussions:

- use of multiple PIDs for individual objects
- permanent referencing of artworks and their components
- Transparency about possible PIDs in repositories
- handling of changes in metadata due to new research results
- Use of PIDs for different data types, e.g. PIDs for complex search queries (with timestamps)
- DOIs for Annotations or specific elements/ single entries in workcatalogues
- Stable URIs for vocabularies are required for mappings

nationale
Forschungsdaten
Infrastruktur
for CULTURE

27th January 2023 | Workshop on PIDs within NFDI

Thank you!

Sandra.Goeller@fiz-Karlsruhe.de
Desiree.Mayer@slub-dresden.de

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - 441958017

UseCase

“PID for digital twins in plant genetic resources”

within NFDI FAIRagro

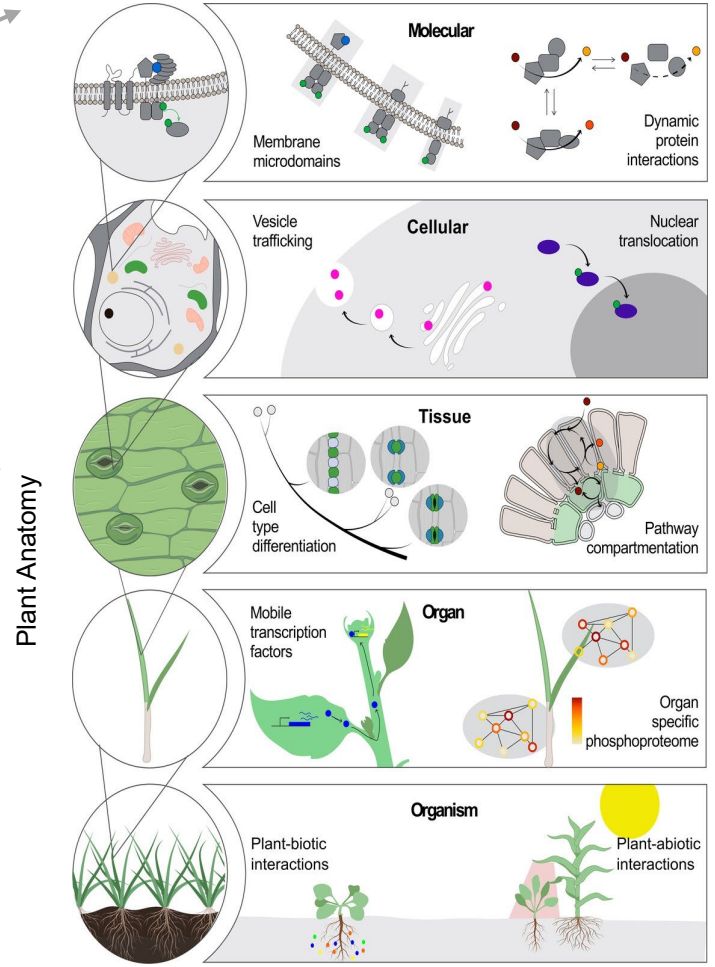
Daniel Arend, Matthias Lange
Leibniz Institute of Plant Genetics and Crop Plant Research

(Ewert et al. FAIRagro FAIR Data Infrastructure for Agrosystems - Proposal 2021, Zenodo, DOI: 10.5281/zenodo.7528172)

Plant Genetic Resources for Food and Agriculture (PGRFA)

1:n

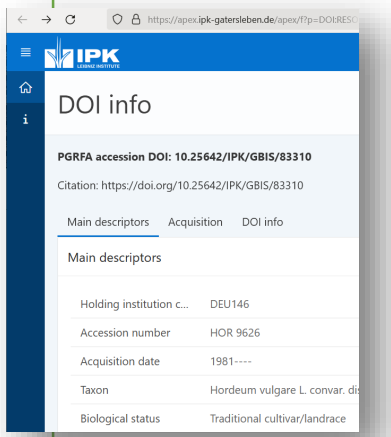
n:m



(Plant Cell Atlas Consortium et al., Cell Biology, 2021)

PGRFA Materials

- PUIDs
  - Genbanks specific: accession numbers
    e.g. HOR 9626 / IPK Gatersleben
  - Global PUIDs: DOIs
    e.g. 10.25642/IPK/GBIS/8331
- meta data
  - darwin core/data cite
  - MCPD - multi crop passport descriptor
- research data Infrastructures
  - National => GLIS – Global Information System
  - Germany -> in NFDI services ecosystem (FAIRAgro, DataPlant, 4BioDiversity)



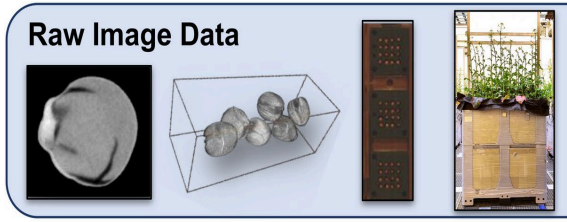
Digital Twins for PGRFA

- PUIDs for samples
  - samples
  - domain specific databases
    e.g. sequences: EMBL-ENA, e!DAL-PGP
  - proprietary global resolver: Identifier.org
    https://identifiers.org/ena.embl:PRJEB40589
- meta data - standard: MIAPPE+
  - sample property
  - environment
  - management
- research data infrastructures
  - institutional repos e.g. LIMS, DBs, file stores
    -> FAIR activation by FAIRAgro

https://www.ipk-gatersleben.de/forschung/genbank/genbankdokumentation

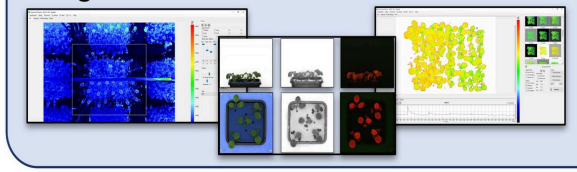
Digital Twins in plant research data ecosystem

Raw Image Data



Spectrum: images taken at visible light, static fluorescence, near-infrared wavelengths, NMR images, CT images
Angles: top, several side views

Image-Derived Traits

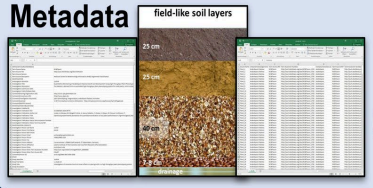


Architecture: plant height, projected leaf area, leaf angles, growth rate
Color: average leaf hue, green to brown ratio, variance in leaf color
Intensity: static fluorescence, near-infrared emitted radiation

Environmental Data

Shoot environment: air temperature, humidity, light intensity, CO_2 concentration
Root environment: soil temperature, water content, nutrition levels, pH

Metadata



isatab

miappe

Plant: species, genotype, seed origin
Conditions: soil and container type, watering regime, experiment location
Measurements: observation units, measurement methods, sensor types

(Arend et al. 2022 The Plant Journal; DOI: 10.1111/tpj.15804)

Sample flow

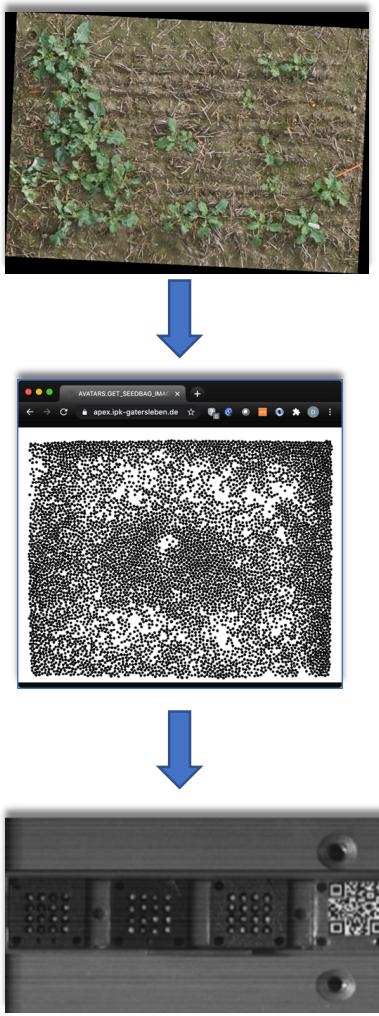
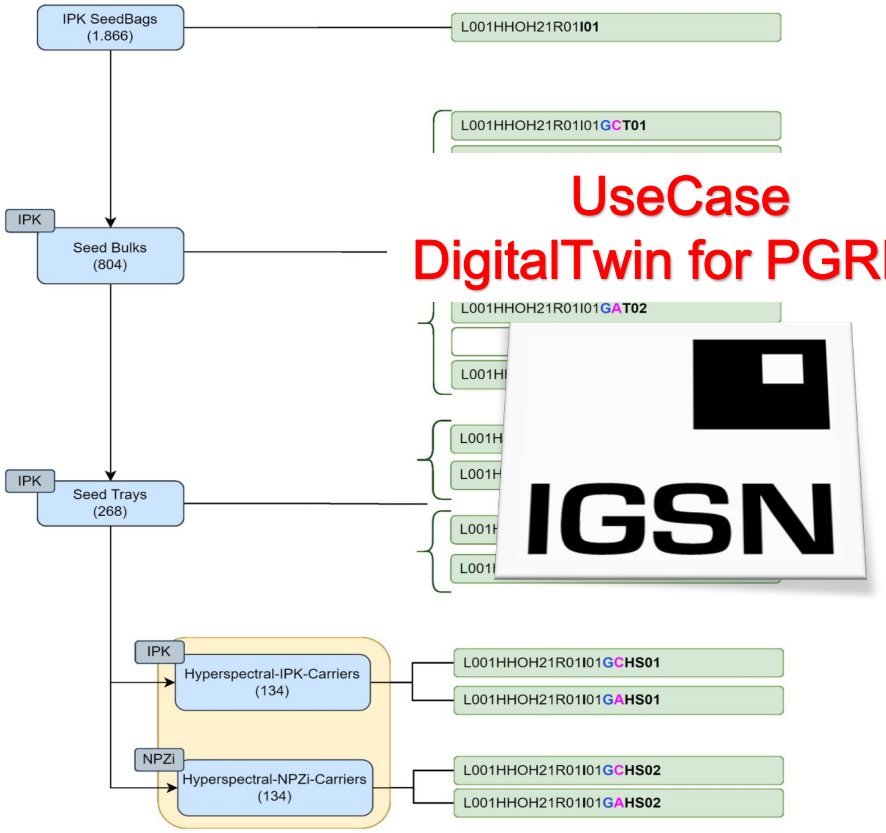


image source: AVATARS project IPK, NPZi

Digital Twins



UseCase DigitalTwin for PGRFA

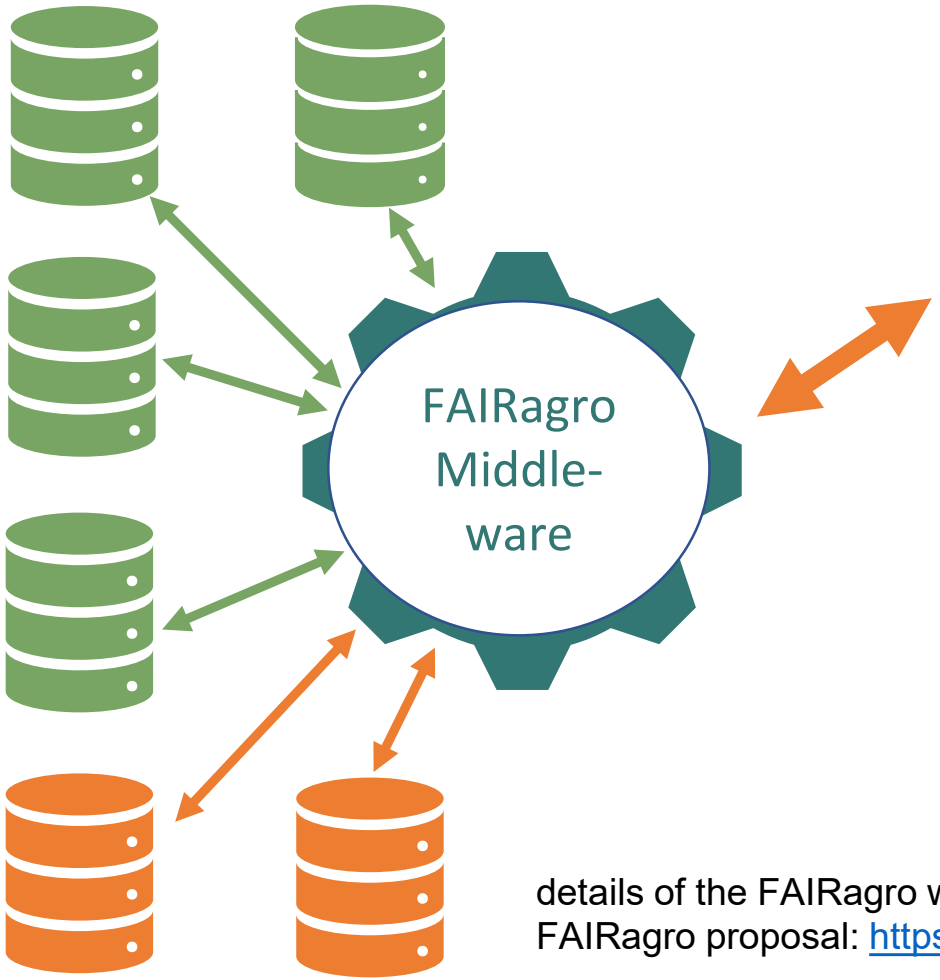
IGSN

C004HMOO20R06I01ST01

genotype* | *field plot* | *season* | *physical object

(Rey-Mazon, NPZi; Plant 2030 status seminar; 2022)

Integration into FAIRAgro – Inventory and Search Portal in Task Area 4



wish list for
Base4NFDI PID service

support implemention of IGSN for Digital Twins of PGRFA

1. Global resolvability
1. Location context
2. Sample flow tracking
3. Data flow tracking
4. Enable schoolary citation

[GDoc with details for IGSN for Digital Twins use case]

details of the FAIRagro work program and consortium: https://www.fairagro.net
FAIRagro proposal: https://doi.org/10.5281/zenodo.7528172

Research Data Infrastructures (RDIs) for PGRFA