

PID4nfdi



Example Use Cases and Services for NFDI

NFDITalk, February 19, 2024

Stephanie Hagemann-Wilholt  – Leibniz Information Centre for Science and Technology (TIB) 

Matthias Lange  – Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) 

Daniel Martini  – Kuratorium für Technik und Bauwesen in der Landwirtschaft (KTBL) 

Philipp Wieder  – Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen (GWDG) 

What is PID4nfdi?



Why PIDs?

- backbone of FAIR RDM
- persistence → not just a technical issue
- addressing whole lifecycle of research data
- support eco-system of PIDs



HELMHOLTZ
Open Science





No PID? Not FAIR!

To be Findable:

- **F1. (meta)data are assigned a globally unique and eternally persistent identifier.**
- **F2. data are described with rich metadata.**
- **F3. (meta)data are registered or indexed in a searchable resource.**
- **F4. metadata specify the data identifier.**

To be Accessible:

- **A1 (meta)data are retrievable by their identifier using a standardized communications protocol.**
- A1.1 the protocol is open, free, and universally implementable.
- A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
- **A2 metadata are accessible, even when the data are no longer available.**

To be Interoperable:

- **I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.**
- I2. (meta)data use vocabularies that follow FAIR principles.
- **I3. (meta)data include qualified references to other (meta)data.**

To be Re-usable:

- **R1. meta(data) have a plurality of accurate and relevant attributes.**
- **R1.1. (meta)data are released with a clear and accessible data usage license.**
- **R1.2. (meta)data are associated with their provenance.**
- **R1.3. (meta)data meet domain-relevant community standards.**



Benefits to the NFDI

Landscape: PID solutions for different use cases across different disciplines and stages of RD lifecycle

Interoperability: mature PID services operating with internationally established standards

Support: training to unleash full potential of PIDs and metadata
→ operationalize FAIR

Governance: requirements for persistence, cross-service usage

Networking: Connecting with national and international PID stakeholders



PID4nfdi: Phases and main objectives

Initialisation

- NFDI-wide PID strategy
- Blueprint for further development
- Covering various use cases and requirements of consortia
- Main outcome: fundamental information for implementing PID services concepts and solutions

Integration

- Support of technical, organizational and methodological implementations
- Identify further requirements
- Connected to existing sustainable PID infrastructure
- Policies & Governance

Ramping up

- Cross-service synchronisation
- NFDI PID Policy
- Central Support: Helpdesk
- Regular Training
- Metadata Quality Tools

Analyse PID Landscape and Requirements

Objectives

- Current use of PIDs in NFDI
- identify common features
- Evaluate optimization potentials
- identify agnostic solutions & community-specific requirements

contact:

pid4nfdi@lists.nfdi.de

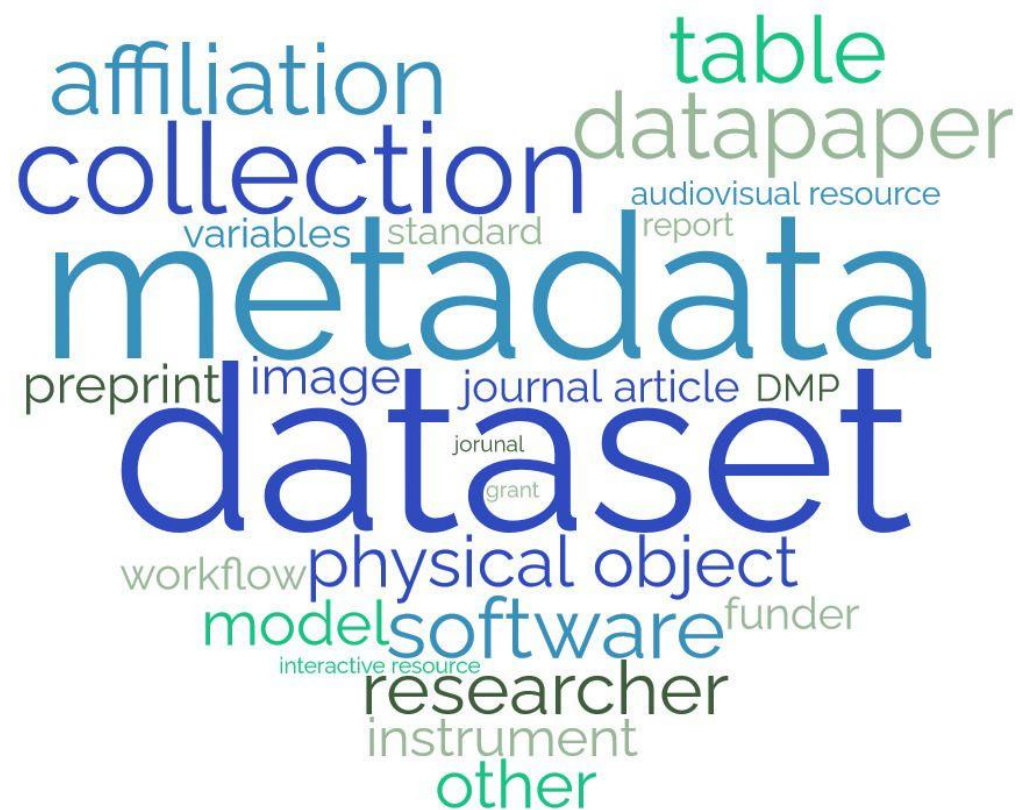
Activities

- In preparation: survey & interviews with stakeholders
- preliminary work 2023: survey & workshop





Preliminary findings



Most frequently mentioned resources with identifiers

- (planned to be) integrated from other sources for re-use in own infrastructure
- (planned to be) registered with own infrastructure

Interoperability

Objectives

- Develop concepts for technical integration and metadata interoperability
- Harmonize existing PID services and standards, consortia solutions and identified gaps

Activities

- Mapping use cases to existing PID services
- Evaluation of options for use of services, e.g. PID Graph, PID metaresolver
- Metadata quality assessment



Metadata Quality

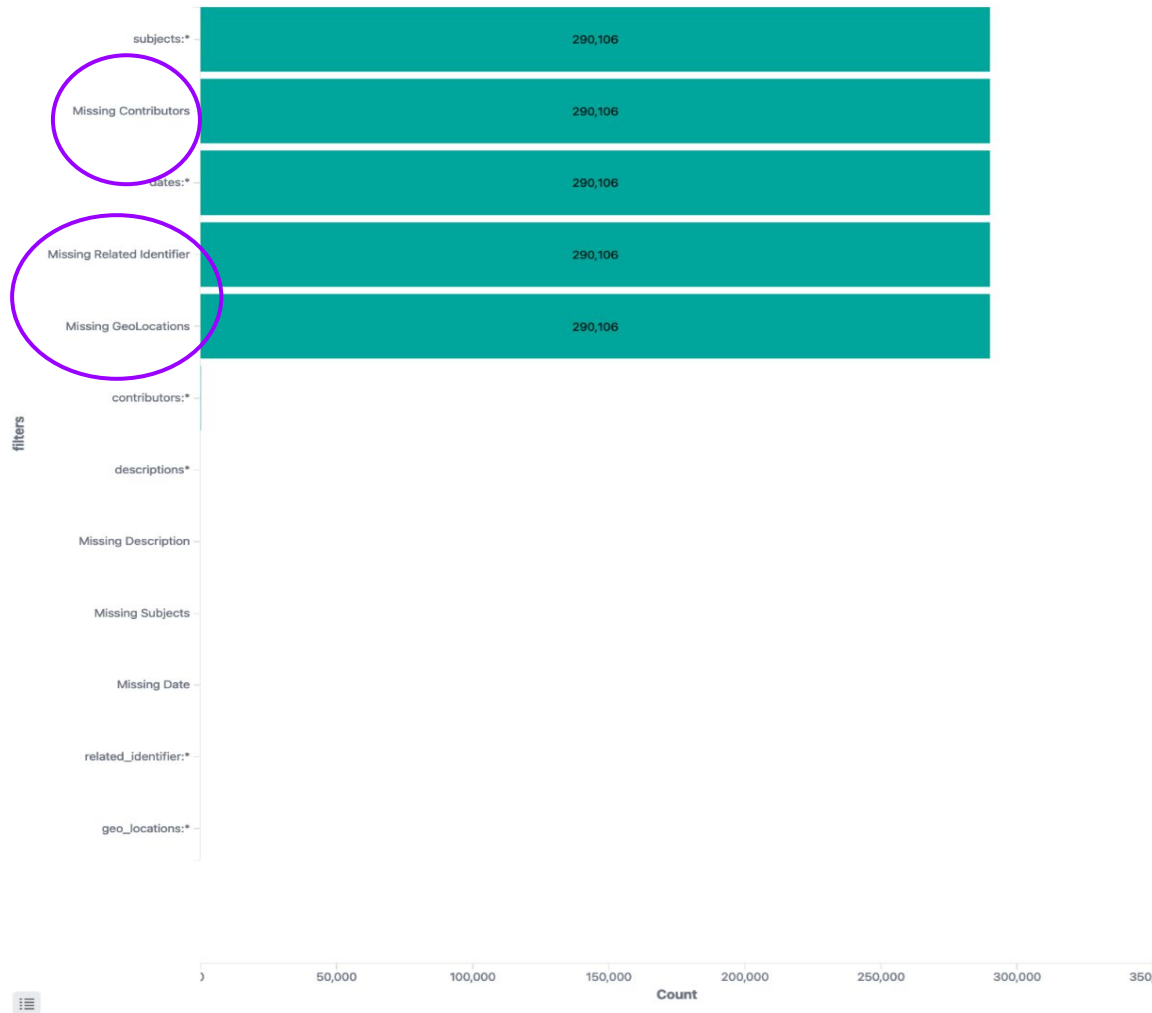


DFG Guidelines for Safeguarding Good Research Practice (Code of Conduct) ¹	Quality Criteria for Metadata ²	Operationalisation (Selection)
Methods and standards (G11)	Conformance to expectations Accuracy Completeness Logical consistency and coherence	General: accepted metadata standard scheme more: Field classification, Description of methods, Temporal data on collection/ publication/ change
Cross-phase quality insurance (G7) → Corrections of inconsistencies/errors	Accuracy of data and metadata	Versioning for additions and corrections Meta metadata
Cross-phase Quality Insurance (G7)→ origin of resources	Provenance Completeness	Authorship, curation information citation/linking information
Documentation (G12) + Providing public access (G13) → transparency	Accessibility Completeness	At least all metadata required to register a PID
Legal and ethical frameworks, usage rights (G10)	Accessibility Completeness	License information

1: Deutsche Forschungsgemeinschaft. (2022). Guidelines for Safeguarding Good Research Practice. Code of Conduct. <https://doi.org/10.5281/zenodo.6472827>.

2: Bruce/Hillmann (2004): The Continuum of Metadata Quality. Defining, Expressing, Exploiting, pp.5-10; see also: Dorothea Strecker (2021): Quantitative assessment of metadata collections of research data repositories. Berlin: HU Berlin (Berliner Handreichungen zur Bibliotheks- und Informationswissenschaft, 470).<https://doi.org/10.18452/22916>.

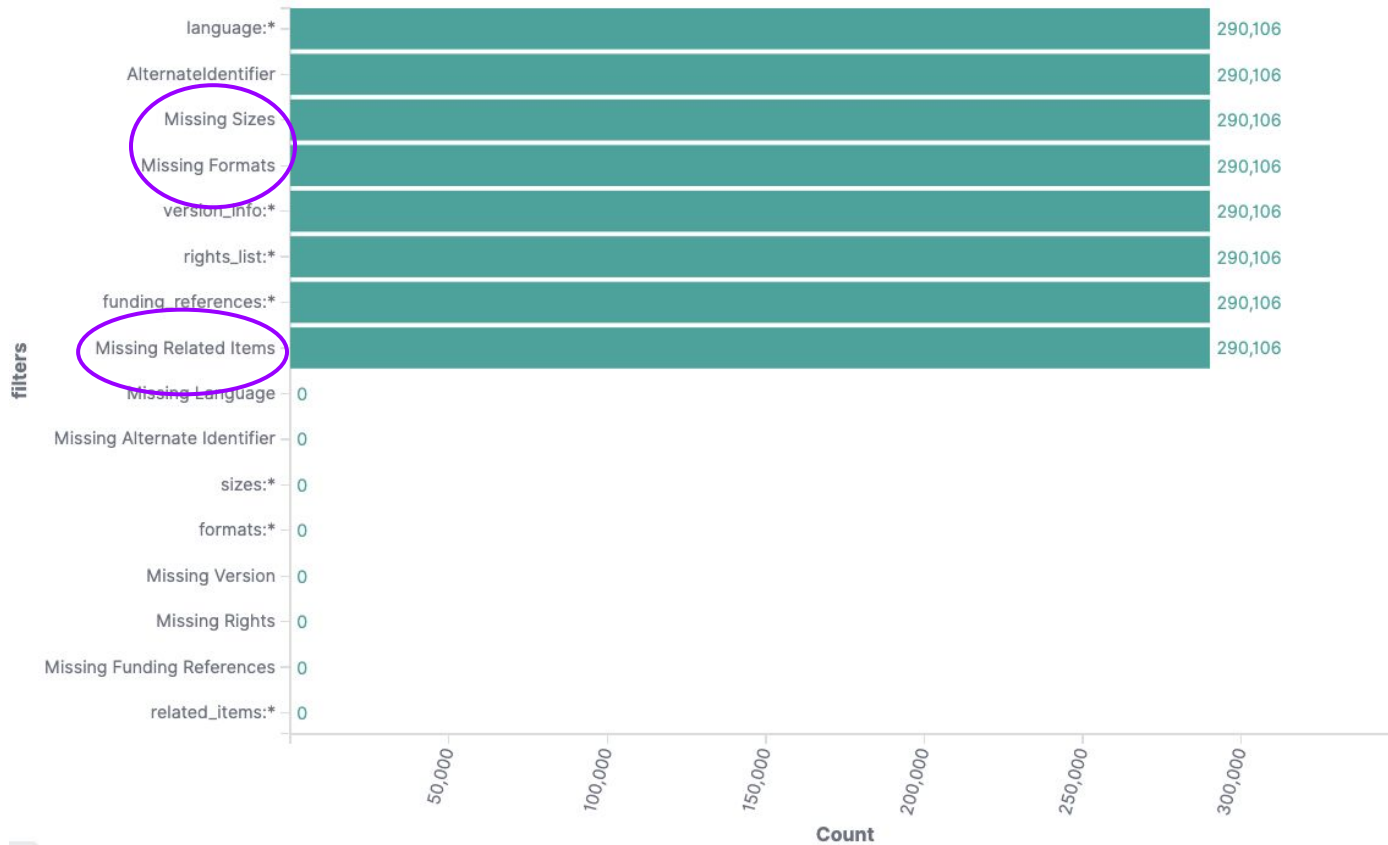
Metadata completeness



Missing (sub-)properties in **Recommended** fields:

- related Identifier → lack of related resources means it is not linking the resource to other outputs
- geolocation
- contributors

Metadata completeness



Analysis of metadata completeness in repository

Missing (sub-)properties in **Optional** fields:

- sizes
- formats
- related items

Support & Training

Objectives

- Raise awareness for added values of PIDs
- Support implementation and management of PIDs
- Support knowledge transfer into and out of NFDI

Activities

- Concept for trainings formats
- Cookbooks for use cases and generic PID registration in research workflows



Governance

Objectives

- Promoting concept of persistence management as process based on formal and informal commitment
- Identify opportunities and gaps to enable sustainably financed PID registration based on requirements

Activities

- Overview and evaluation of relevant business, governance and licence models and their modalities
- Concept for organisational integration of PID infrastructure



Outreach & Networking

Objectives

- Raise awareness for PIDs, need for interoperability and existing services
- Ensure link to international endeavours (esp. EOSC)

Activities

- Communication strategy
- Establish knowledge base on PIDs in NFDI
- Stakeholder Workshop





Example Use Case I – Agronomy

Use Cases for PIDs in agronomy: digital twins in plant genetic resources and soil samples

Matthias Lange¹, Daniel Martini²

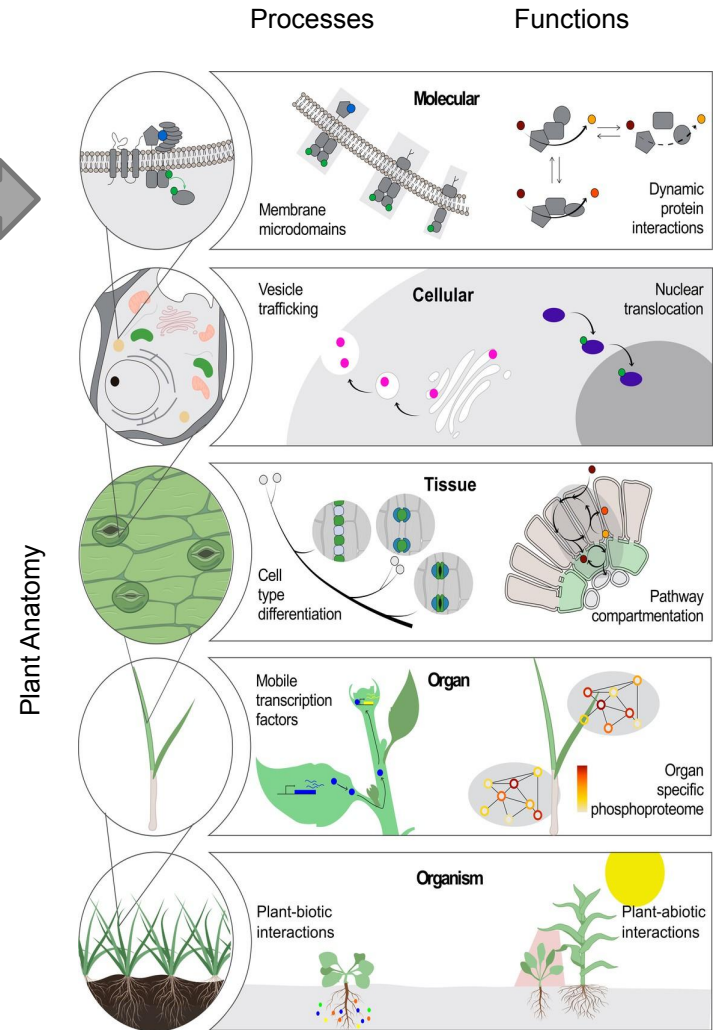
¹Leibniz Institute of Plant Genetics and Crop Plant Research

²Kuratorium für Technik und Bauwesen in der Landwirtschaft e.V.

On behalf of the FAIRagro consortium



Plant Genetic Resources for Food and Agriculture

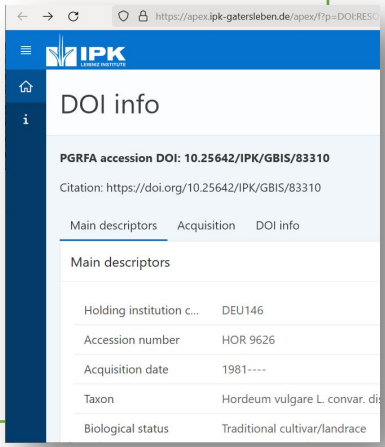


PGRFA Materials

- 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
 - multi crop passport descriptor



- Research Data Infrastructures**
- National
 - Europe
 - Global - GL
-

Digital Twins for PGRFA samples

- PUIDs**
- domain specific databases e.g. sequences: EMBL-ENA, BioSamples
 - proprietary global resolver <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

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

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

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, CO₂ concentration
Root environment: soil temperature, water content, nutrition levels, pH

Metadata

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

isatab
miappe

Sample flow

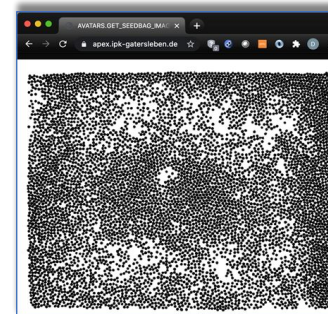
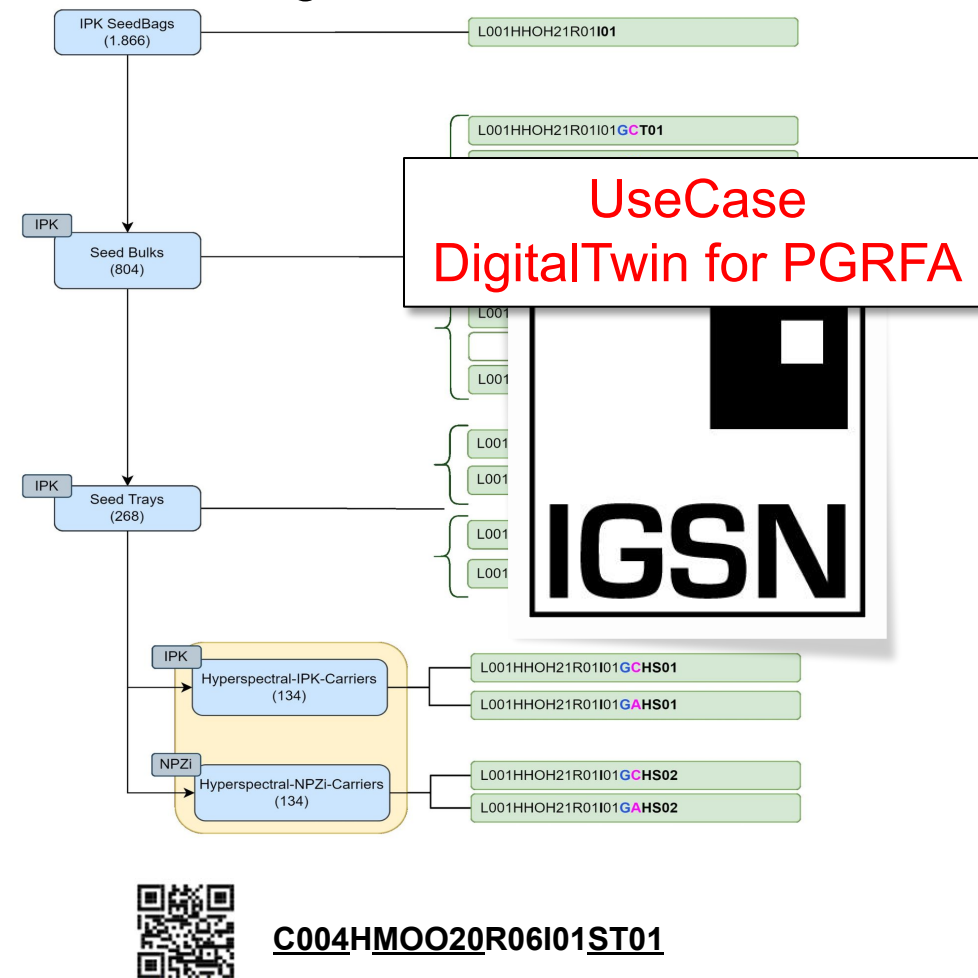


image source: [AVATARS project](#) IPK, NPZi

Digital Twins



genotype | field plot | season | physical object

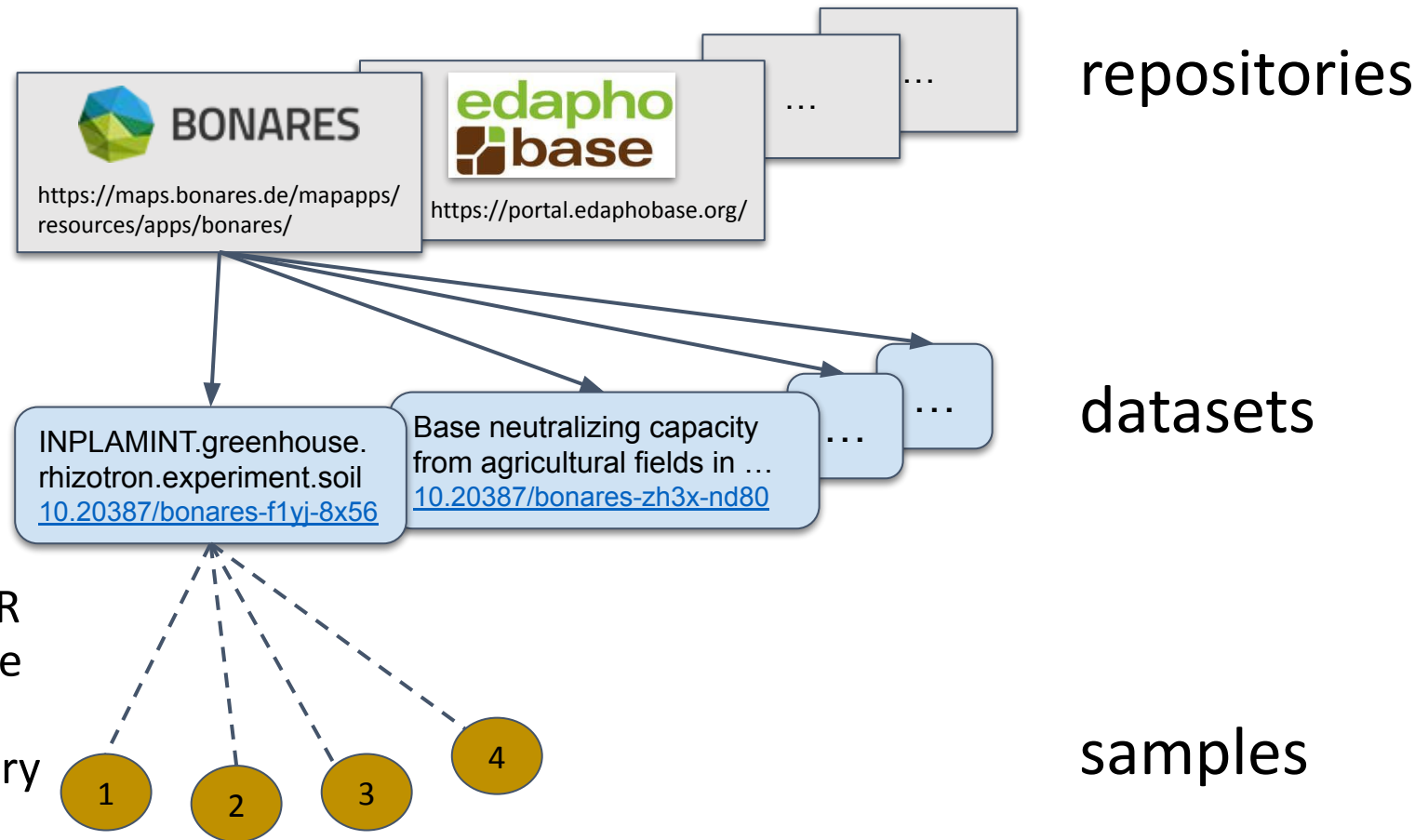
(Rey-Mazon, NPZi; Plant 2030 status seminar; 2023)

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

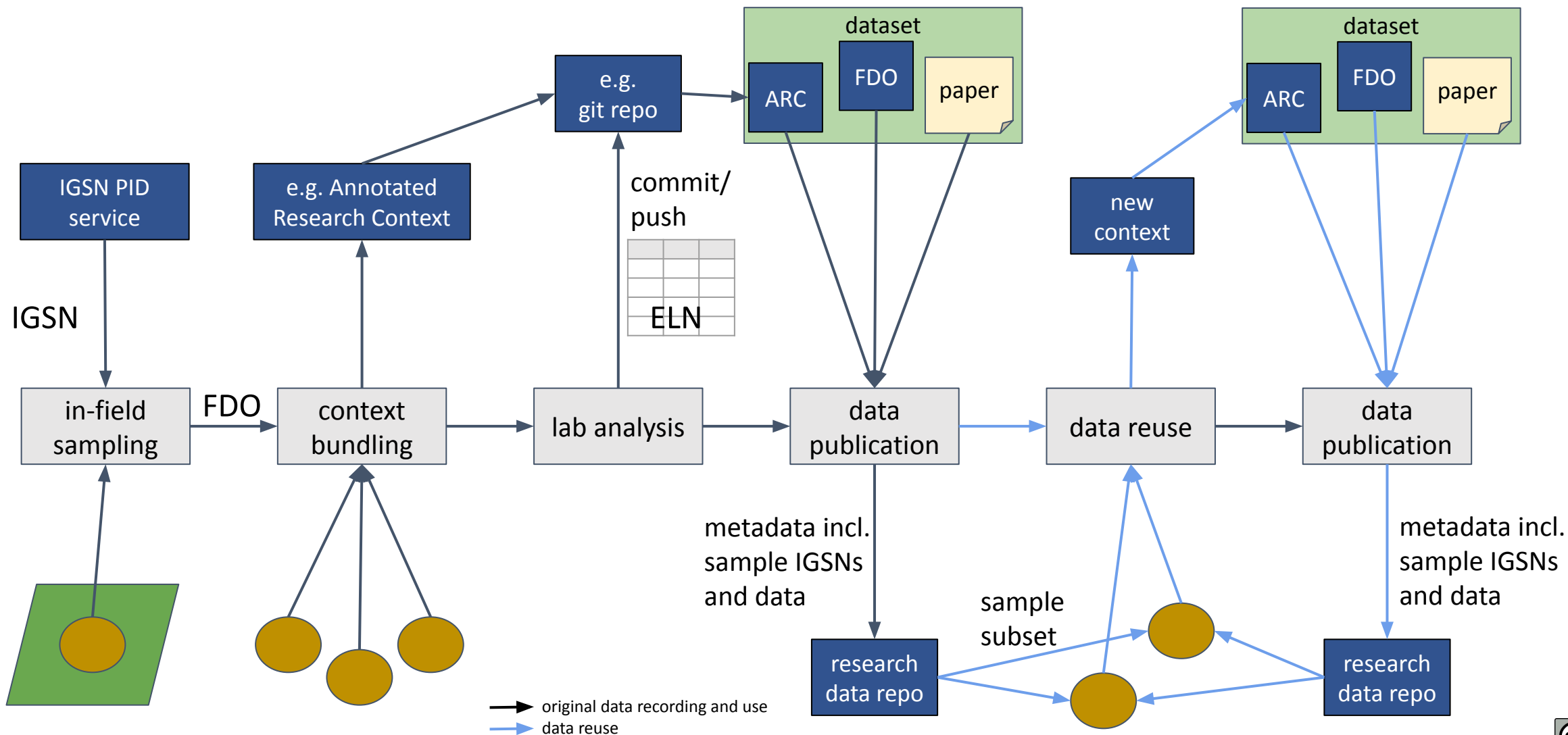


towards FAIR F1 in soil sample data: baseline situation

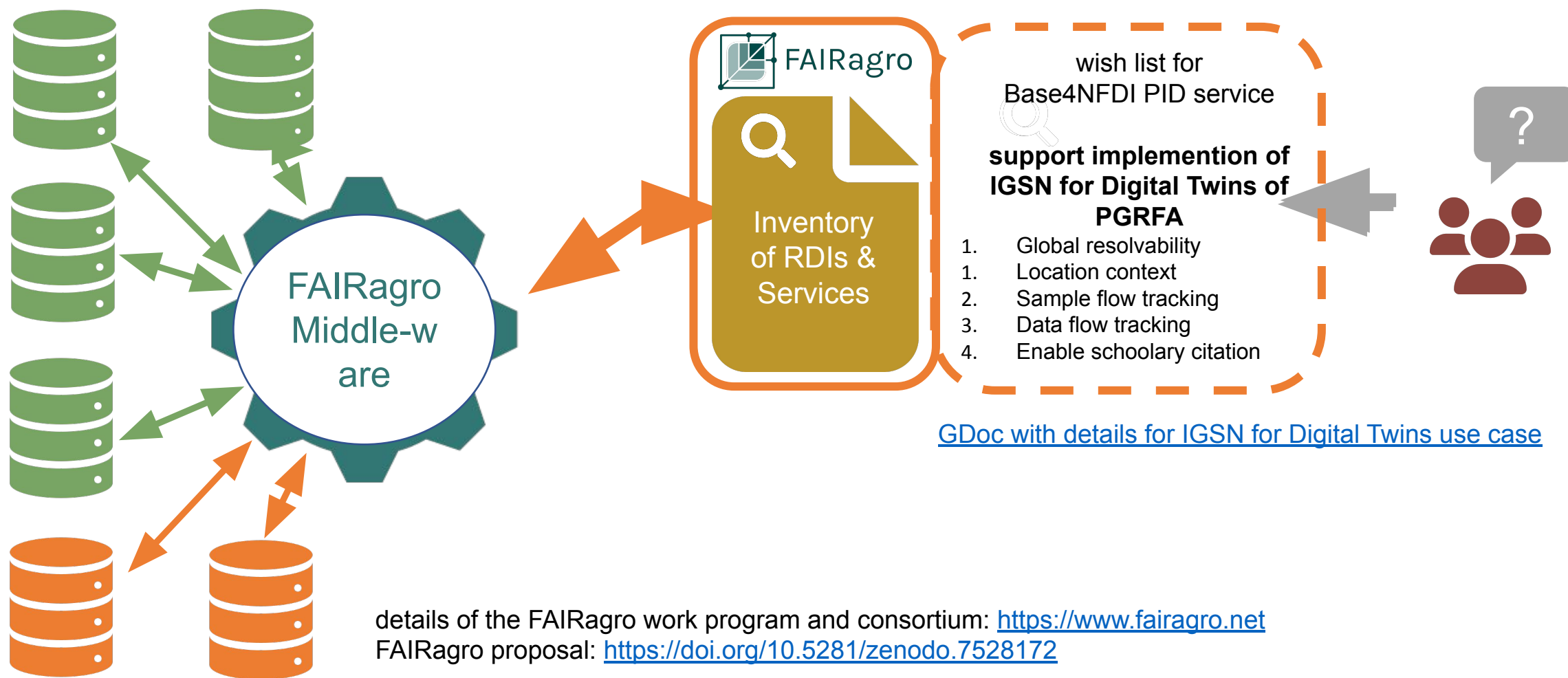
- soil sample relevance:
 - BonaRes repository:
 - soil science
 - Edaphobase:
 - soil organisms
 - ...
- datasets are well identified (DOIs)
- sample data are usually CSV files
 - identifier assignment up to the researcher
 - missing linkability (in the sense of FAIR F1/I3) from sample to dataset and vice versa
 - missing cross-dataset / cross-repository linkability
 - sample provenance/history difficult to track



towards FAIR F1 in soil sample data: idealized workflow and data life cycle



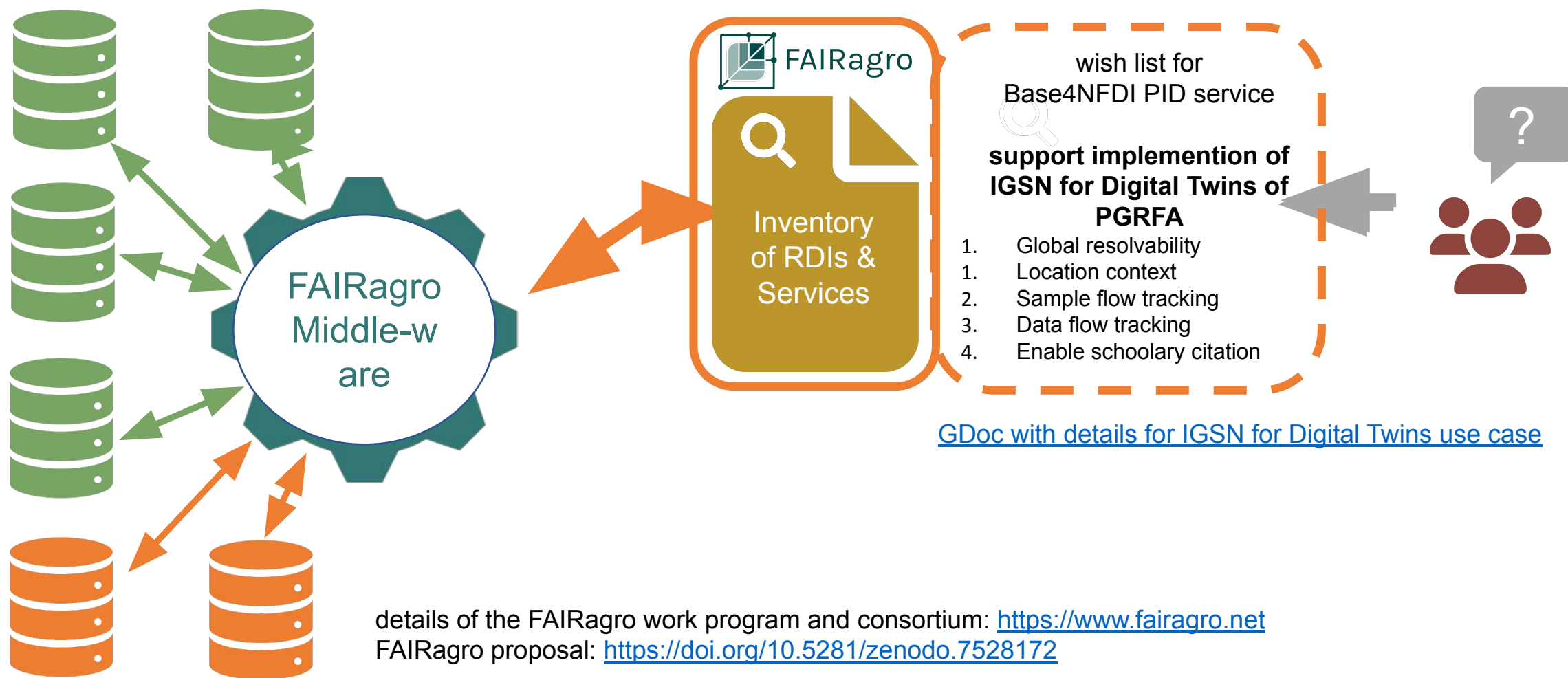
Integration into FAIRragro – Inventory and Search Portal



Research Data Infrastructures (RDIs) for PGRFA



Integration into FAIRragro – Inventory and Search Portal



Research Data Infrastructures (RDIs) for PGRFA



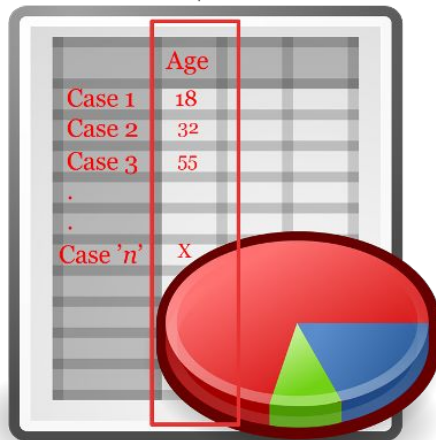
Example Use Case II – Social Sciences

Research Data Granularity & PID Assignment Practice



- In the Social Sciences, PIDs are usually assigned only at study and dataset level

- PIDs for dataset elements (e.g., survey variables) are missing



	Age
Case 1	18
Case 2	32
Case 3	55
...	
Case 'n'	X

PIDs missing

**More general
(Increase ambiguity)**

PID assigned

Study Level

PID assigned

Dataset

**TA5.M1 - PID
data extension
service**

Dataset
elements

More specific

Hurdles of Data Citation Practice in Social Sciences



- Often, researchers do not use the **entire dataset**, but a **set of dataset elements**, e.g. survey variables
- However, **due to the lack of PIDs for dataset elements**, researchers use **textual descriptions of the data** they've been using which are often **incomplete and semantically ambiguous**

Religiosity. General religiosity was measured through the ISSP 2008 item: "Would you describe yourself as. . . ?" (responses ranged from 1 = extremely religious to 7 = extremely non-religious). For the analyses, scores were reversed. Religious practice was measured through three ISSP 2008 items assessing frequency of prayer, religious attendance, and visitation to holy places (responses ranged from 1 = never to 11 = once a day; $\alpha = .61$; α across samples: .43-.64).¹

not clear which variables were actually used

participation rather than opinions and beliefs. The key variables concern attendance of religious services and several demographic and socioeconomic characteristics, such as age, work status, and income.

Several variables used below deserve a more precise definition. First, two levels of attendance are distinguished in the analysis based on the question: "How often do you attend religious services?" Weekly attendance means that a respondent claims to attend a religious service at least once a week; yearly attendance signifies participation at least once a year. Second, employment

paraphrasing the question text as a substitute for citing the variables used

Goal & Status KonsortSWD (TA5-M1)



PID service for dataset elements to facilitate clear and reliable data citation on a more fine-grained level:

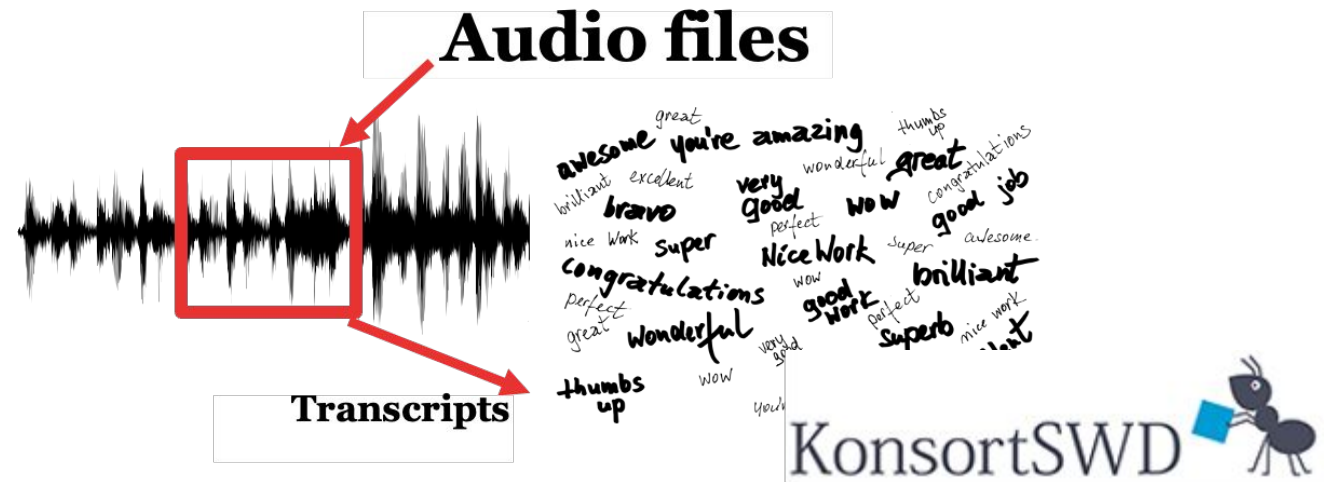
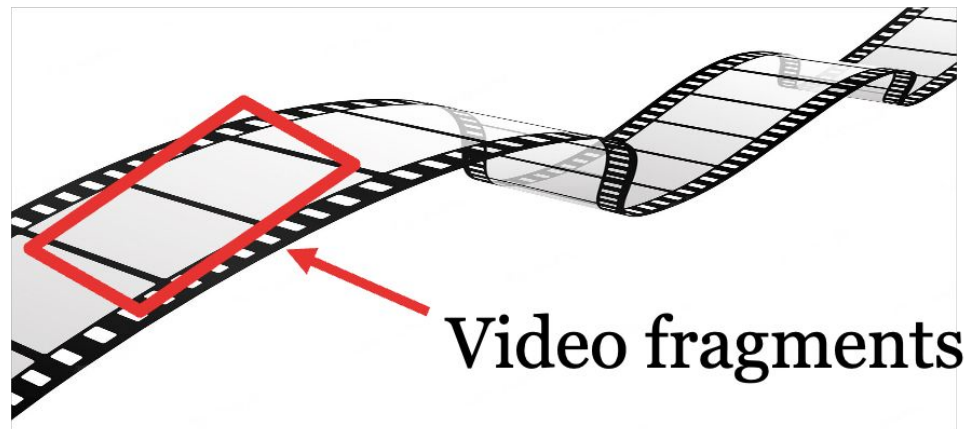
- make research data easier to find
- to boost subsequent citation,
- to enable direct (meta)data access
- to promote data reuse

Status

- PID infrastructure ready to go, based on the ePIC API (Handle system)
- Applied to some KonsortSWD use cases (survey variables so far)



... to be applied to further Entity Types such as





Example Use Case III – Humanities



Research Data in the Humanities

- Metadata, bibliographical data, finding aids
- Digital and/or digitized data and/or digital representation of analog data
- Digital objects
- Full text, transcripts
- Enriched full text
- Images, movies, music & notes
- Authority files, controlled vocabularies, ontologies
- And many more ...



Example: Epigraphical Database



Photography



Symbols

Name: David (genannt Hammerschlag) ben Natan [26.10.1686] Seitenanfang

Edition und Übersetzung Text/Zeilenansicht Textansicht Seitenanfang

<p>ה[ג] הויסם ונסמן : איש נאמן ודוד עלה בפעלה רום חביון : וחפץ בחיים לישב בסתר עליון : היה האלוף המרוםם כהריר דוד בן ה והקצין מיו כהריר נתן זיל מהילוסים מן המרשלאק ויקרבו ימי דוד למות הלך לעולמו יום ג' חמי חשוון ש' תמיו לפיק ושבק חיים לכל חי הנבבדי</p>	<p>Hier errichtet (ein Mal) und ward geborgen ein getreuer Mann, »und David stieg hinauf« hoch »ins Versteck, »der, der Leben begehrt«, »wird sitzen im Schutze des Höchsten«, es ist der Vornehme, der Erhabene, der geehrte Meister, Herr David, Sohn des Vornehmen und Einflußreichen, Vorstehers und Leiters, des geehrten Meisters, Herrn Natan, sein Andenken zum Segen, aus Hildesheim, genannt Hammerschlag. »Und es nahten Davids Tage dem Tod« und »er ging hin in seine Welt« (am) Tag 3, 18. Cheschan des Jahres 447 der kleinen Zählung, »und er ließ das Leben wie alles Lebende«. Seine Seele sei eingebunden in das Bündel des Lebens</p>
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Zl 3: 2Sam 15,30 Zl 4: Ps 34,13 Zl 4f: Ps 91,1 Zl 9: 1Kön 2,1 Zl 9f: Koh 12,5 Zl 11f: nach bBer 61b

Scholarly Edition – synoptical view inscriptions

	P	O	N	M	L	K	I	H	G	F	E	D	C	B	A
A			NA	MA	LA										
B			NB	MB	LB	KB	IB								
C			NC	MC	LC	KC	IC	HC	GC						
D			OD	MD	LD	KD	ID	HD	GD	FD	ED	DD			
E	PE	OE	NE	ME	LE	KE	IE	HE	GE	FE	DE	EE			
F	PF	OF	NF	MF	LF	KF	IF	HF	GF	FF	EF	DF	BF	AF	
G	PG	OG	NG	MG	LG	KG	IG	HG	GG	FG	DG	DG	BG	AG	
H	PH	OH	NH	MH	LH	KH	IH	GH	FH	EH	DH	CH	BH	AH	
I								HI	GI	FI	EI	DI	CI	AI	
K								HK	GK	FK	EK	DK	CK	BK	
L								HL	GL	FL	EL	DL			
M								HM	GM	FM	EM	DM	CM	BM	
N								HN	GN	FN	EN	DN	CN	BN	
O								HO	GO	FO	EO				
P								HP	GP	FP	EP				
Q								HQ	GQ	FQ	EQ				
R								HR	GR	FR					

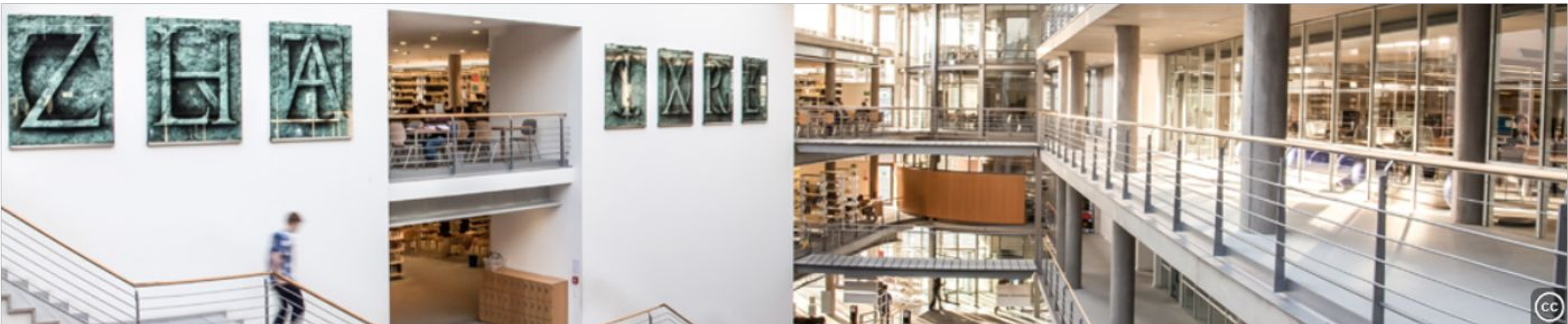
Burial site



Text+ and DARIAH



- DOI assignment integral part of the DARIAH repository
 - Part of the Text+ service portfolio
- SUB Göttingen leads a DataCite consortium with approx. 60 partners to act as a central DOI helpdesk for the Humanities in Germany





Wishlist for PID4NFDI

Overall: PID registration should be free of charge for users and easy to obtain within the NFDI

- Central overview and analysis of PID assignments through all NFDI
- Share experiences from the different disciplines
- Find a sustainable business model for PID assignment in the NFDI
- Central support, held-desk and training



More information

<https://base4nfdi.de/projects/pid4nfdi>

Contact

pid4nfdi@lists.nfdi.de



CC BY 4.0 Deed Attribution 4.0 International
<https://creativecommons.org/licenses/by/4.0/>



In cooperation with



Funded by

