

Brings together

- KFS and KFN
- Large-scale photon and neutron research facilities
- Universities
- Research institutions
- Wider community



Deutsche  
Forschungsgemeinschaft



Nationale  
Forschungsdaten  
Infrastruktur

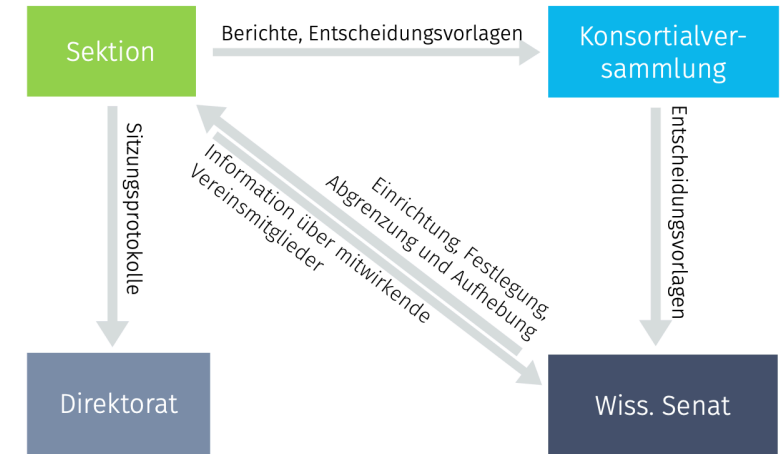
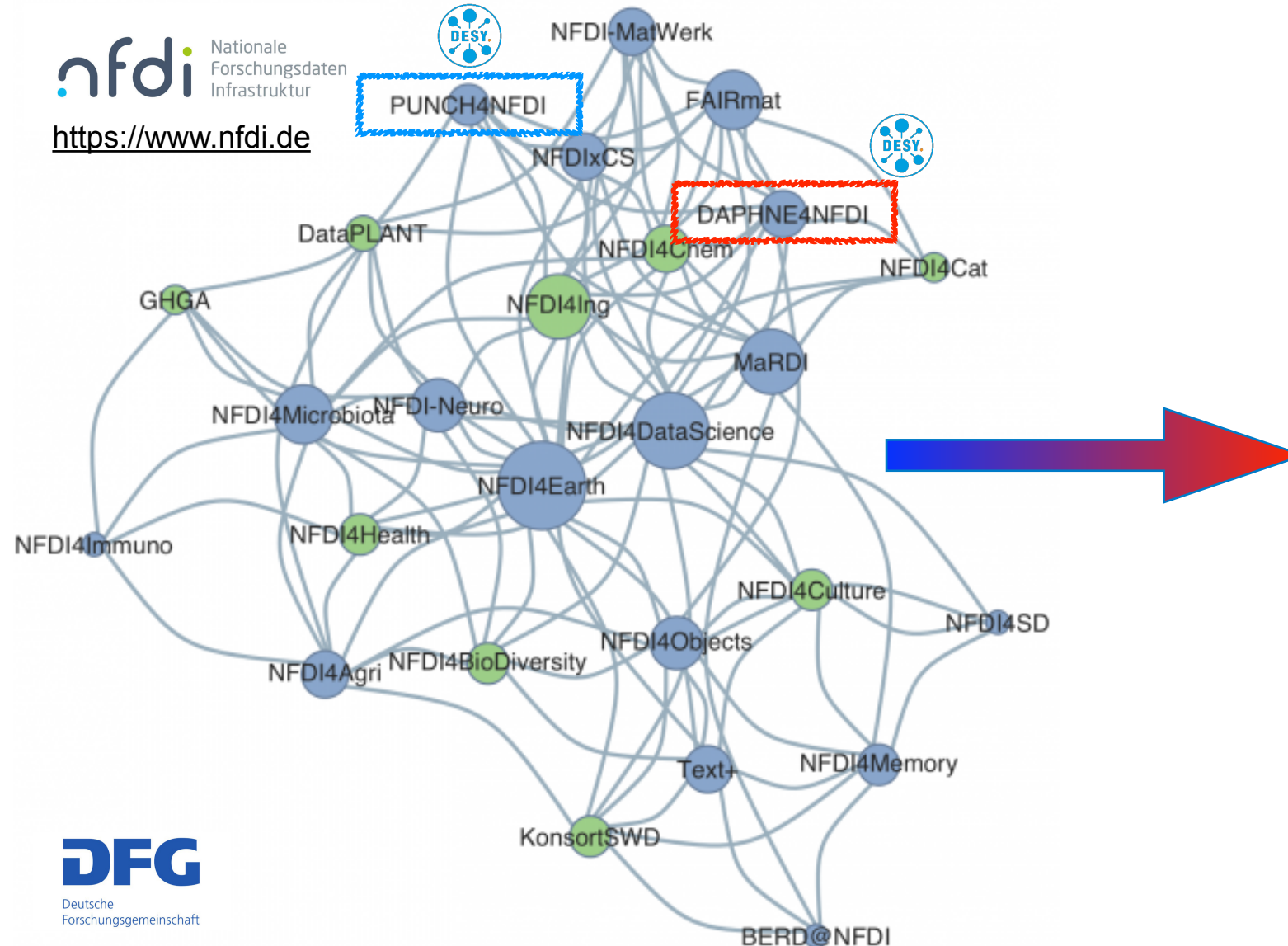
Research data infrastructure, without hardware



# The NFDI is a network of consortia across many areas

A network of common cross-consortia sections and topics spanning diverse research topics

**nfdi** Nationale Forschungsdaten Infrastruktur  
<https://www.nfdi.de>



## Section Common Basic Infrastructure:

- Long term archiving
- Identity management
- Federated cloud
- Research Software Engineering
- Data integration
- AI / ML

## Section Metadata:

- Persistent identifiers
- Semantic interoperability and terminology services
- Ontology harmonisation and mapping
- Provenance verification

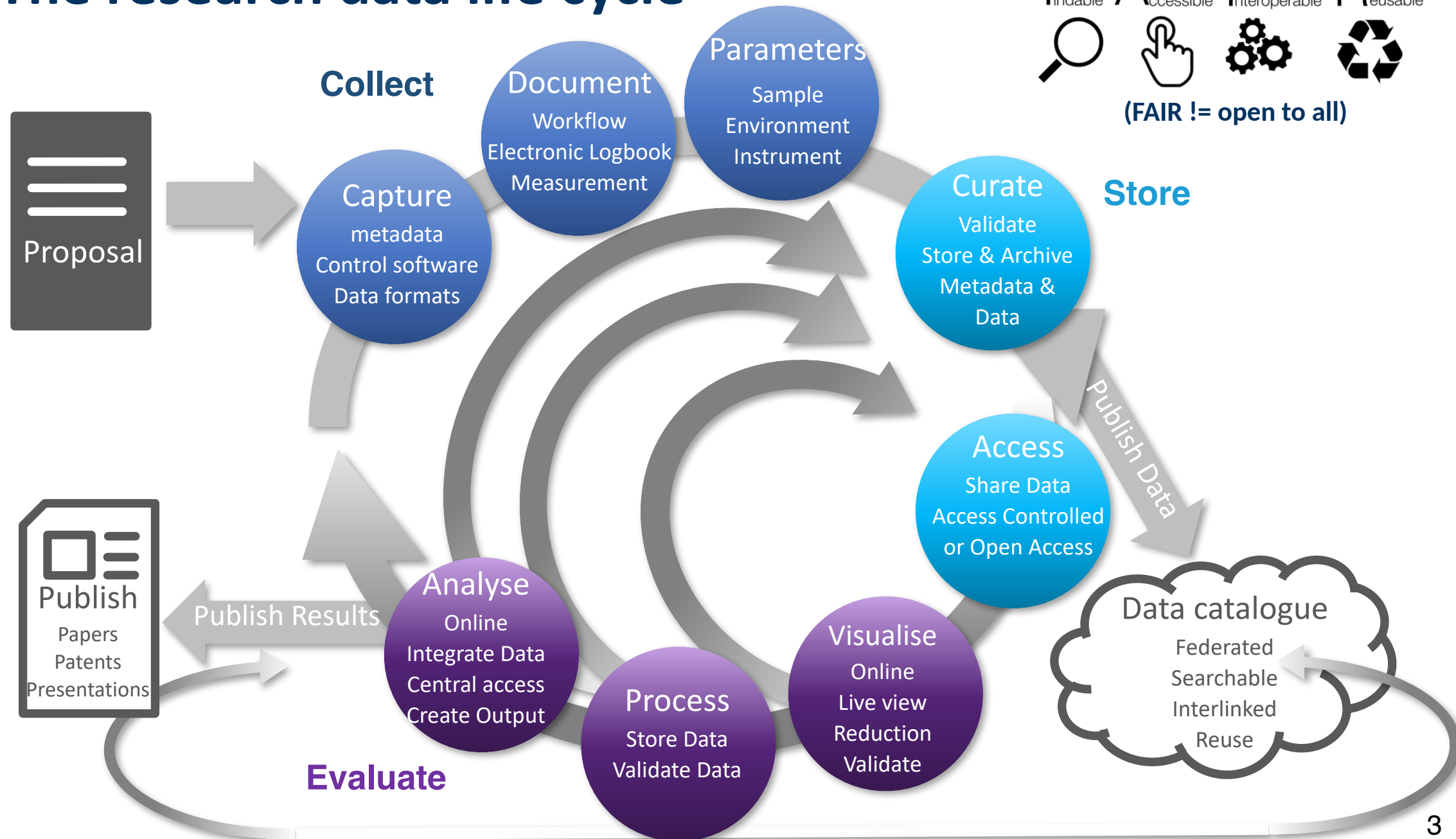
## Section Ethical and legal issues...

## Section training and education...

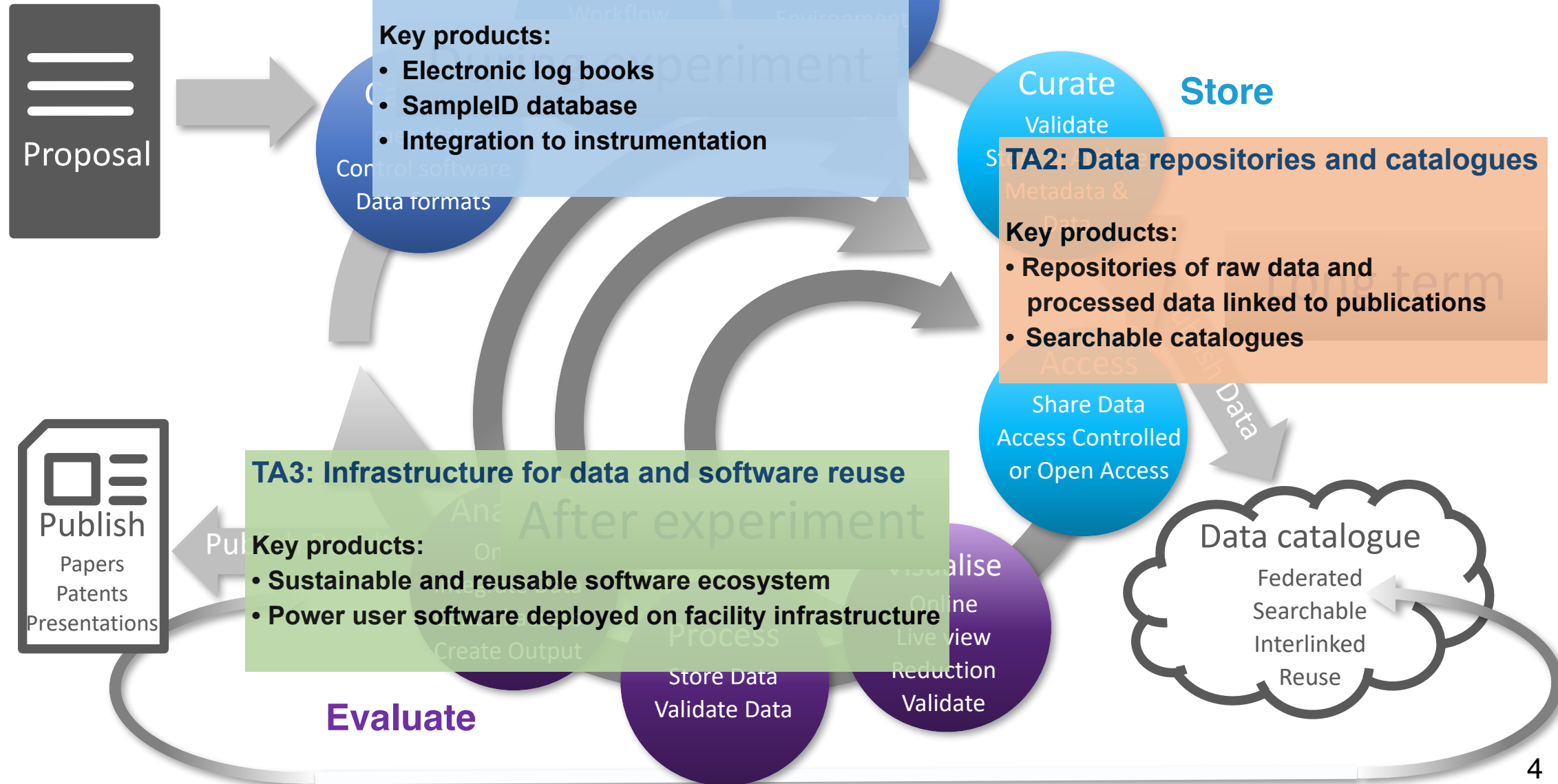
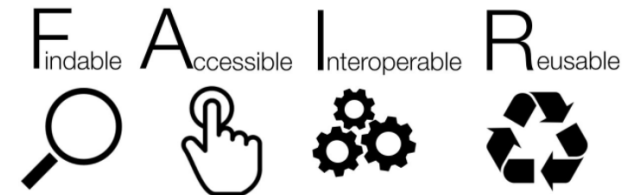


# The research data life cycle

F<sub>indable</sub> A<sub>ccessible</sub> I<sub>nteroperable</sub> R<sub>eusable</sub>  
     
(FAIR != open to all)

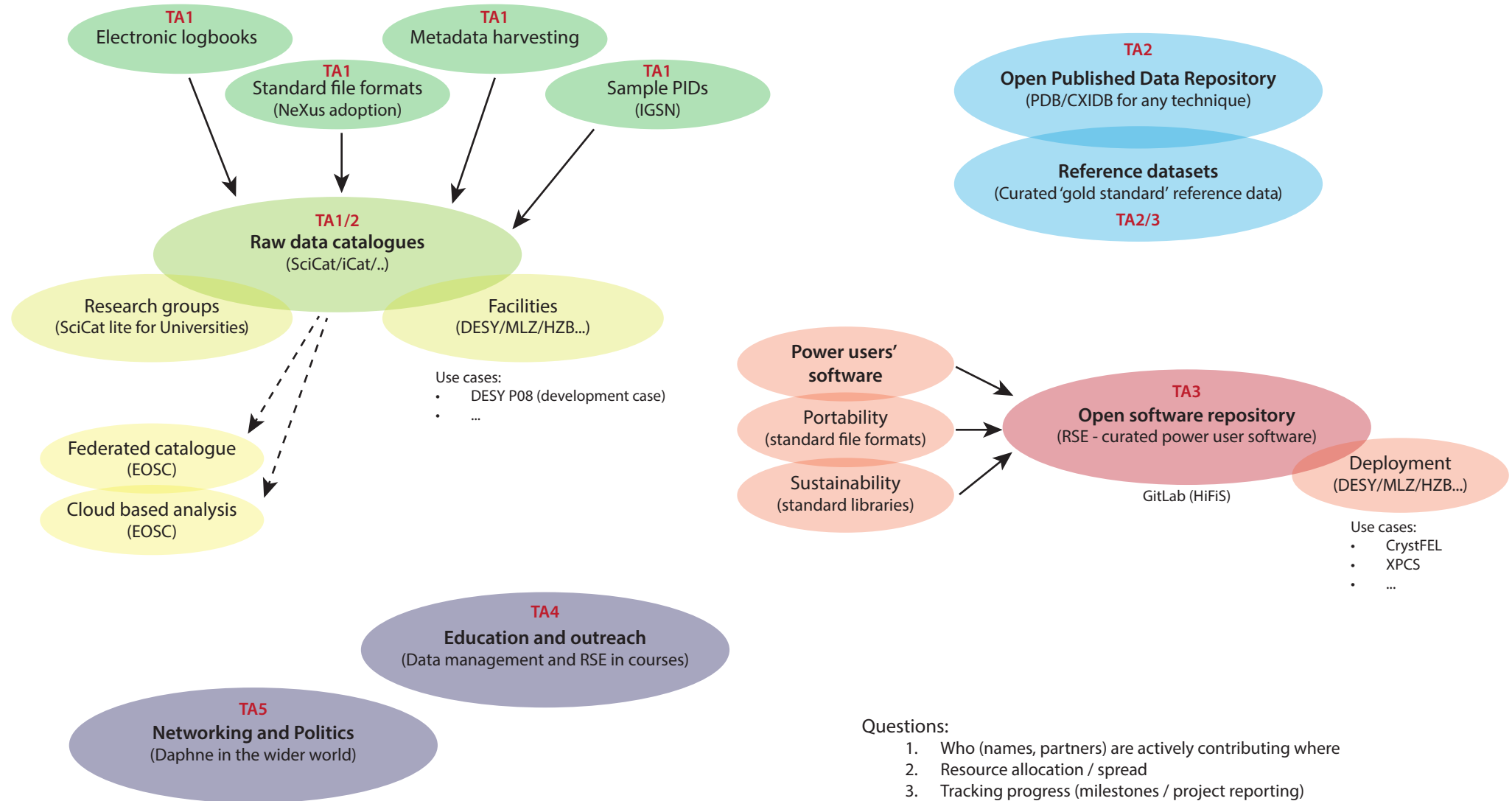


# The research data life cycle








# Daphne work program







# 1. SciCat database as a catalogue foundation




Discover data via WebUI




Findable Accessible Interoperable Reusable



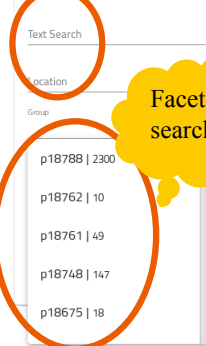
User specific data



Archive Interface





Facet search



Name	Source Folder	Size	Start Time	Type	Proposal ID	Group	Data Status
029_estaillades1_q01_fw085_ss	...1_fw085_ss	1 TB	2020-12-23 Wed 00:05	derived	p17614		retrievable
020_estaillades1_q01_fw085_us	...1_fw085_us	729 GB	2020-12-23 Wed 00:05	derived	p17614		retrievable
019_estaillades1_q01_fw085_us	...1_fw085_us	376 GB	2020-12-23 Wed 00:05	derived	p17614		retrievable
018_estaillades1_q01_fw085_us	...1_fw085_us	376 GB	2020-12-23 Wed 00:05	derived	p17614		retrievable
031_estaillades1_q01_fw085_ss	...1_fw085_ss	4 TB	2020-12-22 Tue 22:02	derived	p17614		retrievable
20201214_ANAXAM/11_360_	...AM/11_360_	47 GB	2020-12-14 Mon 20:59	raw	unknown	p17896	archivable
20201214_ANAXAM/10_360_	...AM/10_360_	47 GB	2020-12-14 Mon 20:37	raw	unknown	p17896	archivable
09_360/09_360_S13_	...9_360_S13_	47 GB	2020-12-14 Mon 20:09	raw	unknown	p17896	archivable
09_360/09_360_S12_	...9_360_S12_	47 GB	2020-12-14 Mon 20:03	raw	unknown	p17896	archivable
09_360/09_360_S11_	...9_360_S11_	47 GB	2020-12-14 Mon 19:57	raw	unknown	p17896	archivable
09_360/09_360_S10_	...9_360_S10_	47 GB	2020-12-14 Mon 19:52	raw	unknown	p17896	archivable
09_360/09_360_S09_	...9_360_S09_	47 GB	2020-12-14 Mon 19:46	raw	unknown	p17896	archivable
09_360/09_360_S08_	...9_360_S08_	47 GB	2020-12-14 Mon 19:40	raw	unknown	p17896	archivable
09_360/09_360_S07_	...9_360_S07_	47 GB	2020-12-14 Mon 19:35	raw	unknown	p17896	archivable
09_360/09_360_S06_	...9_360_S06_	47 GB	2020-12-14 Mon 19:29	raw	unknown	p17896	archivable

PAUL SCHERRER INSTITUT



EUROPEAN SPALLATION SOURCE

## Some features:

- Data browsing
- Data search
- Data download
- Metadata collection
- Online logbooks
- Online chat session
- DataDOI generation
- Archiving interface
- Analysis previews
- 'Data lake' for
  - simulations
  - LK-II data
  - reference datasets



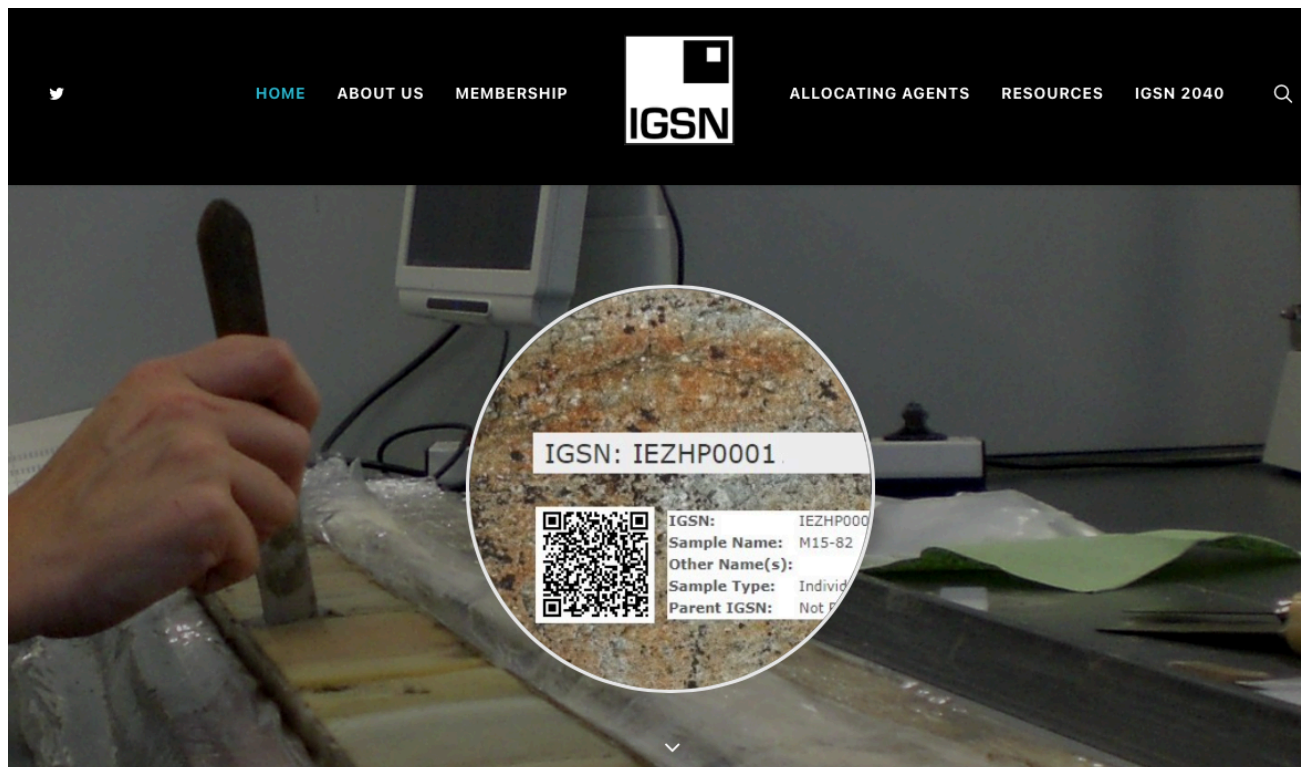


## 2. IGSN as a unique sample identifier

### Challenge:

- Uniquely identify samples so that they can be tracked through logbooks and datasets
- The identifier itself should be unique and persistent even though samples themselves are not always persistent
- Should be simple, cheap and easy to use (and not a minefield of paperwork)

### Solution?



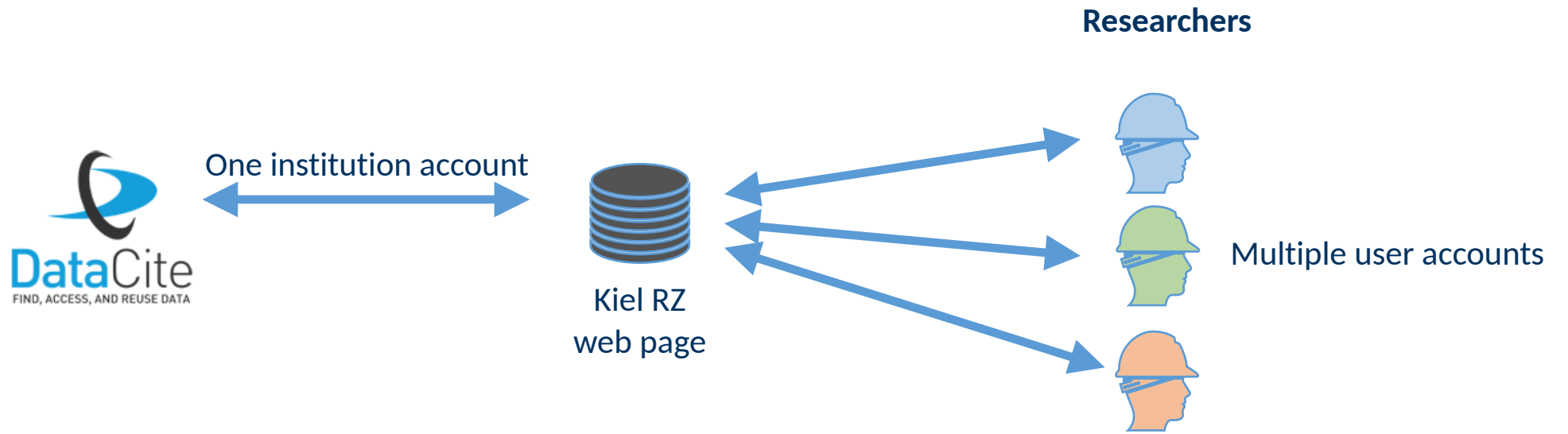
<https://www.igsn.org/>

<https://ardc.edu.au/services>



In September 2021, IGSN e.V. and DataCite entered a partnership under which DataCite will provide the IGSN ID registration services and supporting technology to enable the ongoing sustainability of the IGSN PID infrastructure.

## 2. IGSN as a unique sample identifier



Lightweight to issue unique IDs  
ID resolves to page at RZ with user info and 'light' metadata  
Registered in DataCite only if requested

(Philipp - Kiel?)



# 3. Standardising ontologies and vocabularies

## Standard file formats

- Nexus adoption is a starting point
- What about downstream data?

## Standard metadata

- Community languages
- Essential for interoperable catalogues

## Interoperability validators and libraries

- eg: PDB check

*Daphne brings communities and facilities together*

# 4: Community open data repository (or repositories)

A place to find published data - and in some cases the ability to reprocess data

ht

http

<https://www.eosc-portal.eu>

**DataCite**  
FIND, ACCESS, AND REUSE DATA  
<https://datacite.org>

**DataDOI tracks usage of open data**

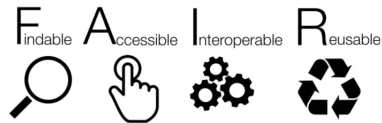


# 5. Managing overwhelming data volumes in the FAIR data era

Archive all data for 10 years:

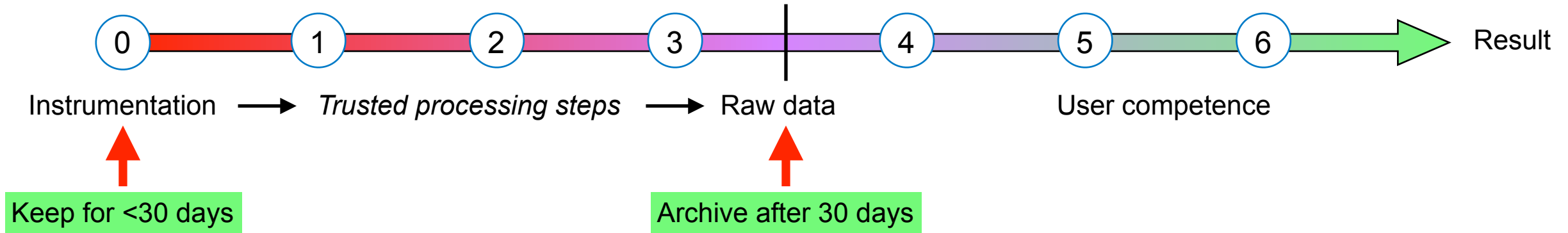
- Current data policy
- Tradition and 'good practice'

Raw data kept on GPFS for 180 days



Economic reality:  
Keeping raw data costs significant  
money (M€) and energy (MW)

## 5. Managing overwhelming data volumes in the FAIR data era



### What is raw data anyway?

1. Develop trusted and validated analysis pipelines to efficiently deliver results to users.
2. Keep all data for a 'safety period' of only 30 days during which low lying problems can be corrected
3. Standard pipeline output becomes the raw data we keep (and the product we give to the users).
4. After 30 days the low level data is deleted (policy, exceptions are possible)
5. Focus on the high data rate instruments