Raw Diffraction data reuse: the good, the bad and the challenging

Managing and curating data flows at PETRA-IV

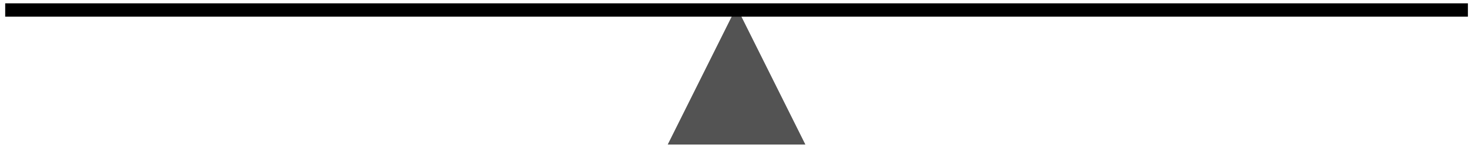
Don't become storage or compute limited
within a reasonable budget envelope

Anton Barty
DESY Photon Science
Scientific Computing

DESY.
HELMHOLTZ

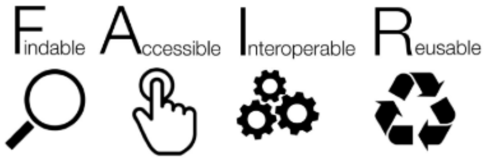
Keeping data has economic consequences

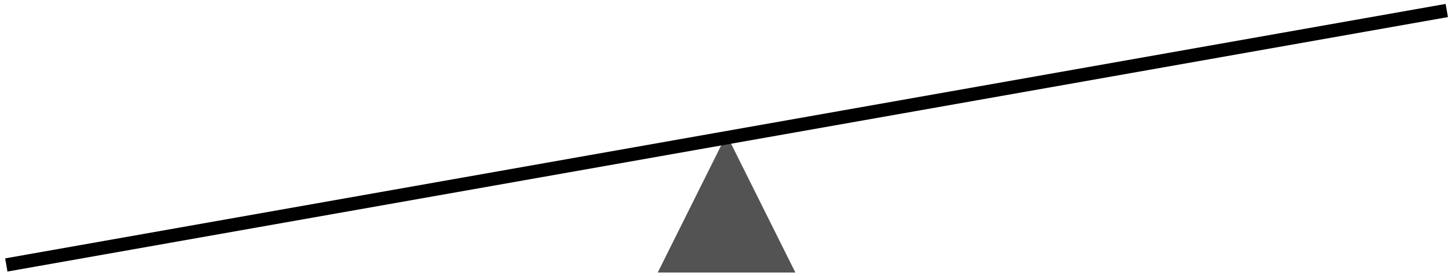
What is the economic value of data? Who pays?



Keeping data has economic consequences

What is the economic value of data? Who pays?

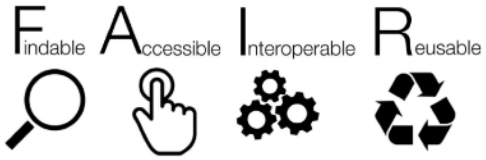


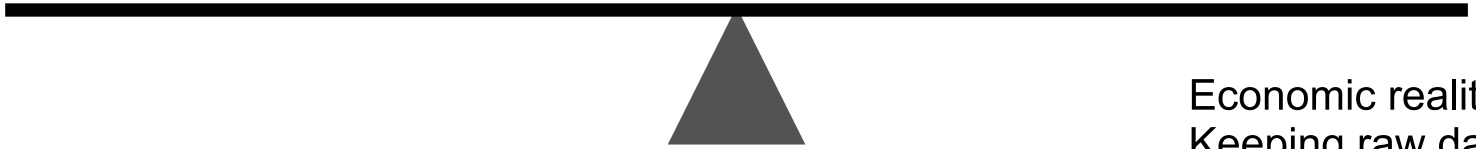


- Academic tradition
- ’Good scientific practice’
- Sometimes mandated by law (USA)?
- Typically archive all ‘raw’ data for 10 years
- Including data known to be ‘dud’
- A ‘nice to have’ or ‘must have’?

Keeping data has economic consequences

What is the economic value of data? Who pays?





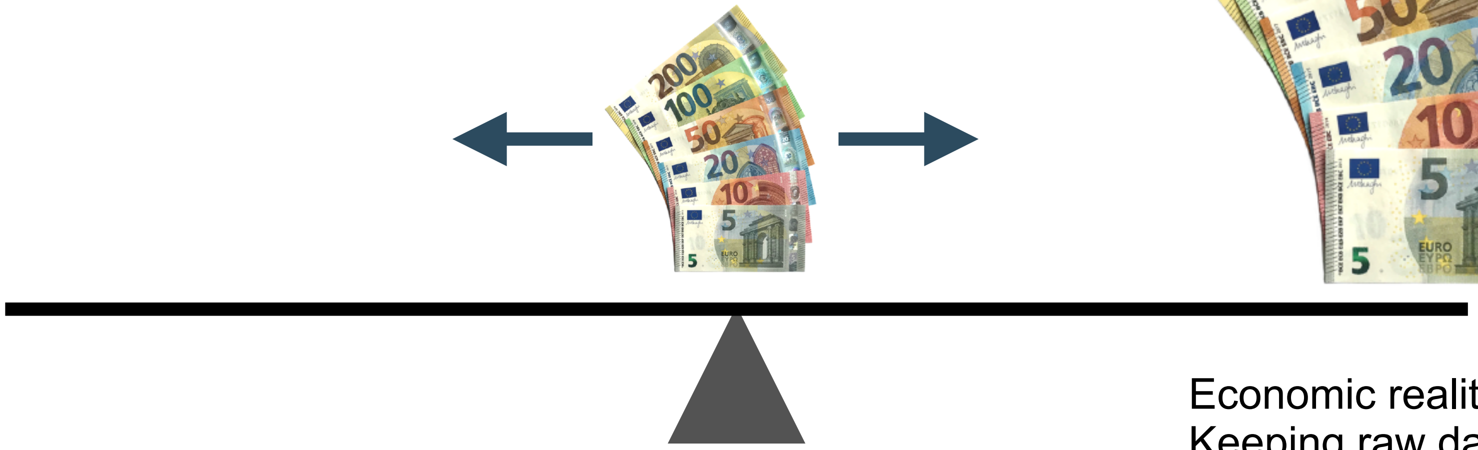
- Academic tradition
- ’Good scientific practice’
- Sometimes mandated by law (USA)?
- Typically archive all ‘raw’ data for 10 years
- Including data known to be ‘dud’
- A ‘nice to have’ or ‘must have’?

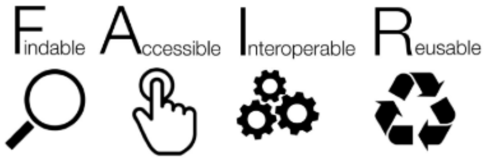
Economic reality:
Keeping raw data costs significant money (M€) and energy (MW)
Keeping all data for lots of experiments becomes expensive very quickly
Facility cost or user’s own cost?

Keeping data has economic consequences

What is the economic value of data? Who pays?

Most samples and experiments
are replaceable at some cost
in time and money





- Academic tradition
- ’Good scientific practice’
- Sometimes mandated by law (USA)?
- Typically archive all ‘raw’ data for 10 years
- Including data known to be ‘dud’
- A ‘nice to have’ or ‘must have’?

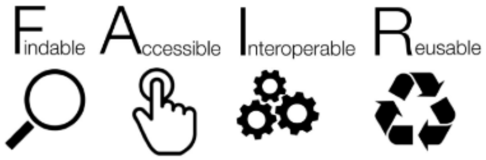
Economic reality:
Keeping raw data costs significant money (M€) and energy (MW)
Keeping all data for lots of experiments becomes expensive very quickly
Facility cost or user’s own cost?

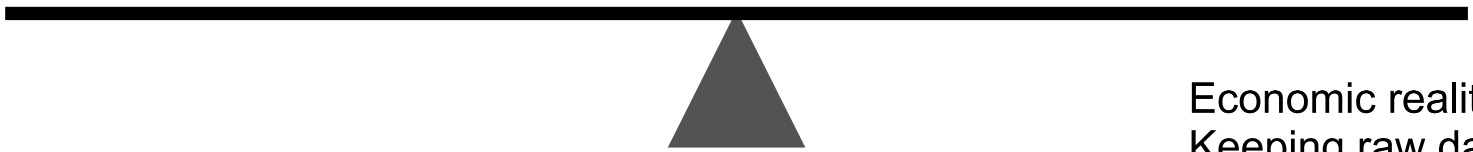
Keeping data has economic consequences

What is the economic value of data? Who pays?

How much are we willing to spend to retain data?
What data gives best value for money?
What are we keeping it for?

Most samples and experiments
are replaceable at some cost
in time and money





- Academic tradition
- 'Good scientific practice'
- Sometimes mandated by law (USA)?
- Typically archive all 'raw' data for 10 years
- Including data known to be 'dud'
- A 'nice to have' or 'must have'?

Economic reality:
Keeping raw data costs significant money (M€) and energy (MW)
Keeping all data for lots of experiments becomes expensive very quickly
Facility cost or user's own cost?

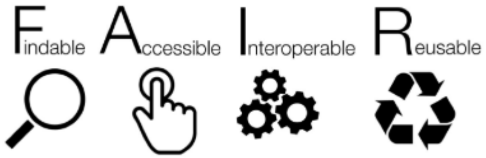
Keeping data has economic consequences

What is the economic value of data? Who pays?

How much are we willing to spend to retain data?
What data gives best value for money?
What are we keeping it for?

How much (limited) money do we spend on old data vs new outcomes?

Most samples and experiments are replaceable at some cost in time and money

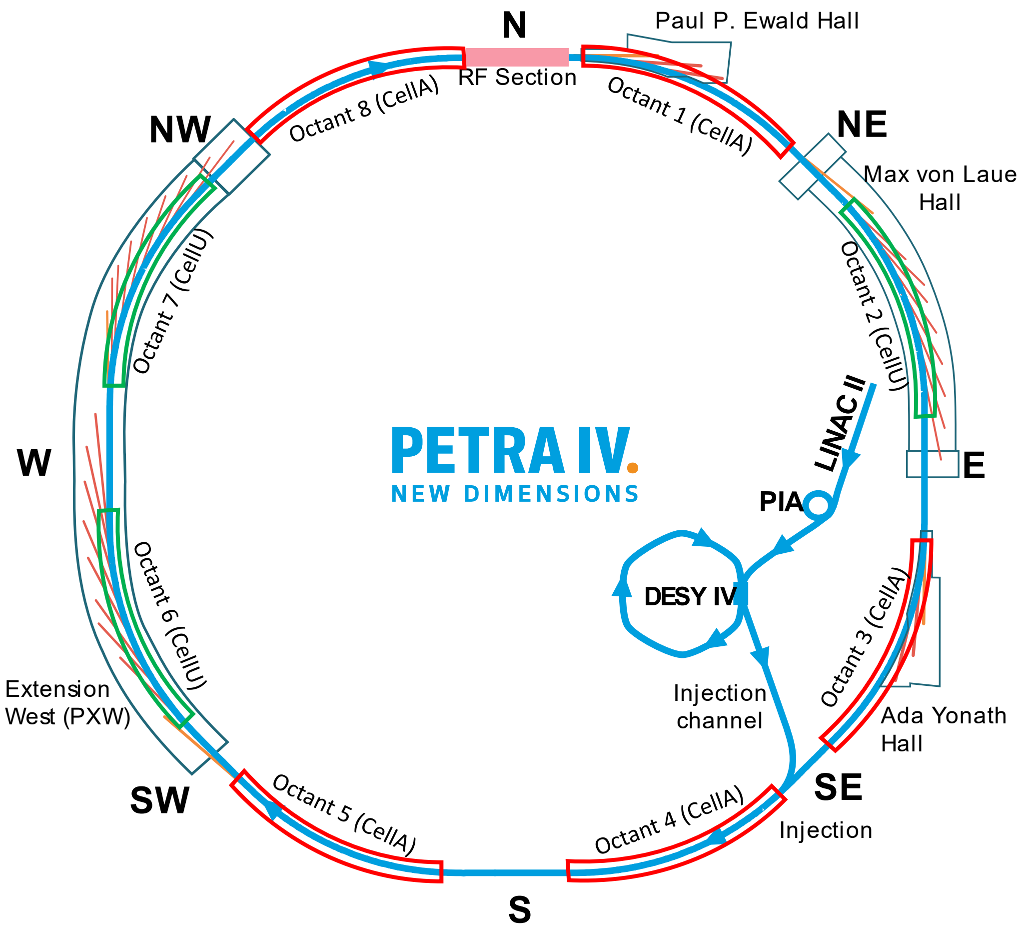


- Academic tradition
- ’Good scientific practice’
- Sometimes mandated by law (USA)?
- Typically archive all ‘raw’ data for 10 years
- Including data known to be ‘dud’
- A ‘nice to have’ or ‘must have’?

Economic reality:
Keeping raw data costs significant money (M€) and energy (MW)
Keeping all data for lots of experiments becomes expensive very quickly
Facility cost or user’s own cost?

The Petra-IV upgrade project

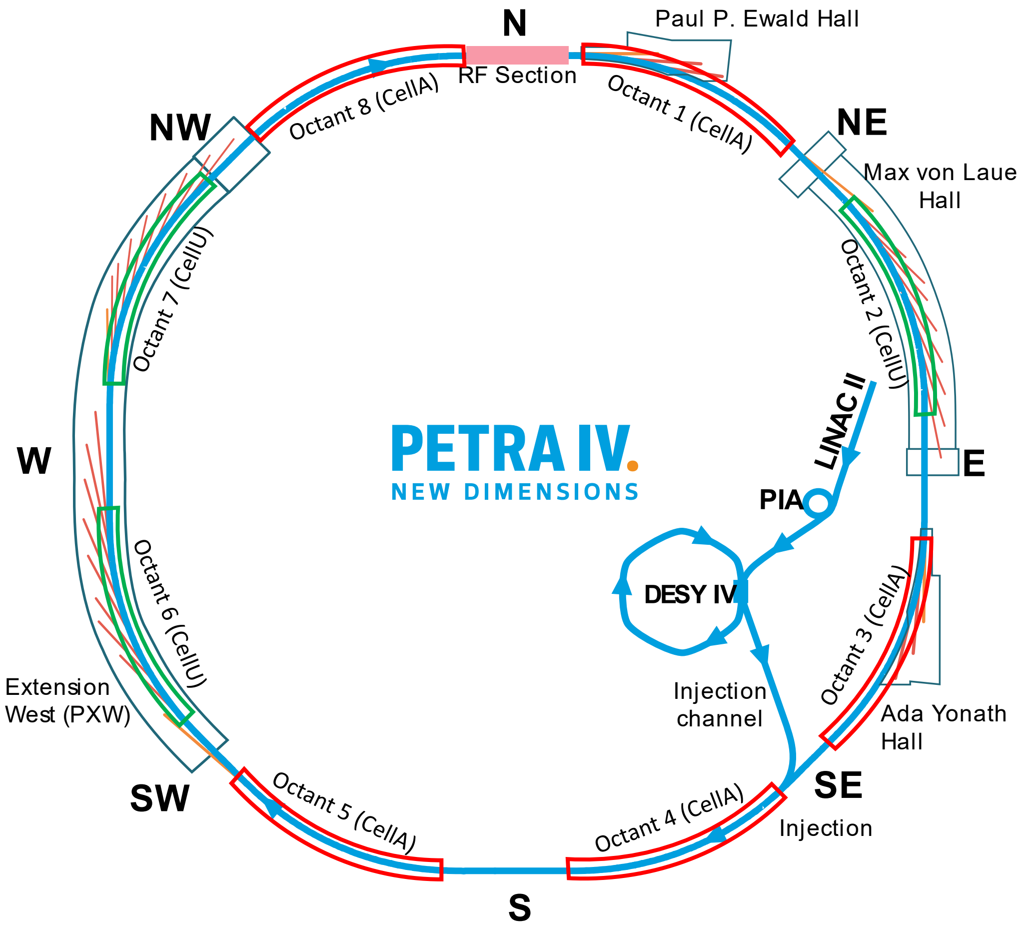
A new ring and an updated operation model serving as a national analytic facility

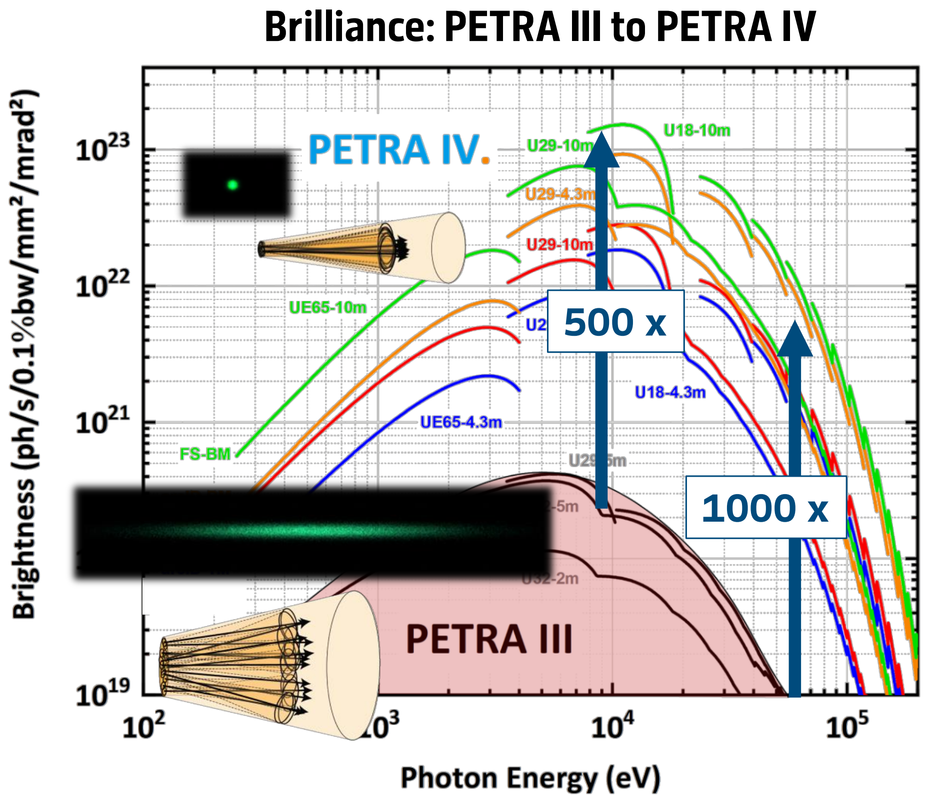


- A rebuilt low emittance ring
- 28-30 instruments
- A completely new hall to the west
- Ready for operation 2028 (or so)

The Petra-IV upgrade project

A new ring and an updated operation model serving as a national analytic facility

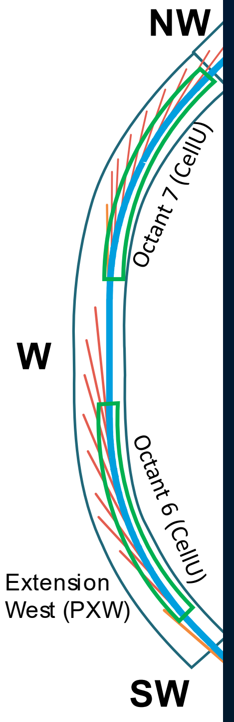


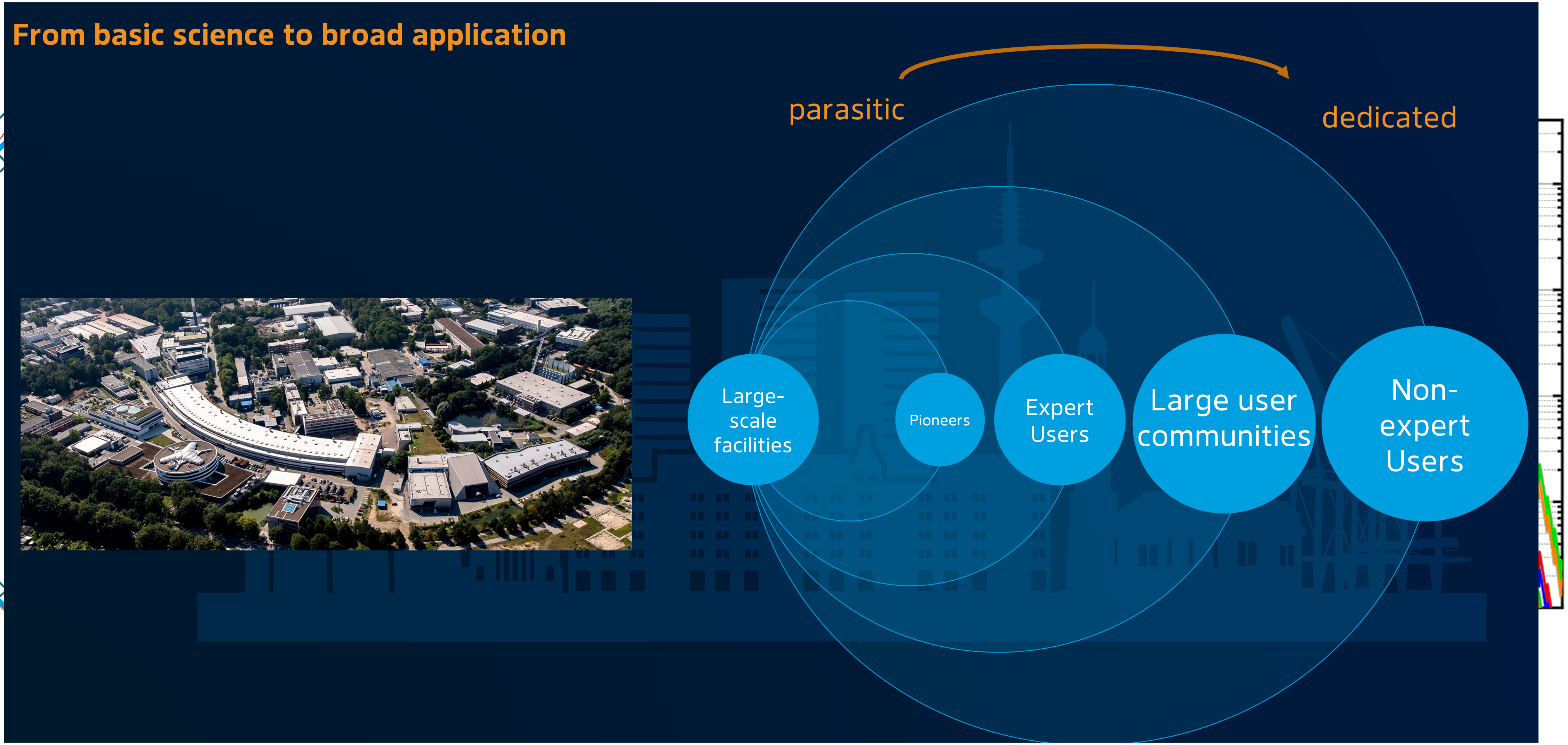


- A rebuilt low emittance ring
- 28-30 instruments
- A completely new hall to the west
- Ready for operation 2028 (or so)

The Petra-IV upgrade project

A new ring and an updated operation model serving as a national analytic facility





- A rebuilt low emittance ring
- 28-30 instruments
- A completely new hall to the west
- Ready for operation 2028 (or so)

- Support for non-expert users
- Faster turnaround from proposal to measurement
- Increased use of automation
- Deliver outcomes rather than data on disk

Data production and retention at PETRA-III today

A snapshot of the status quo

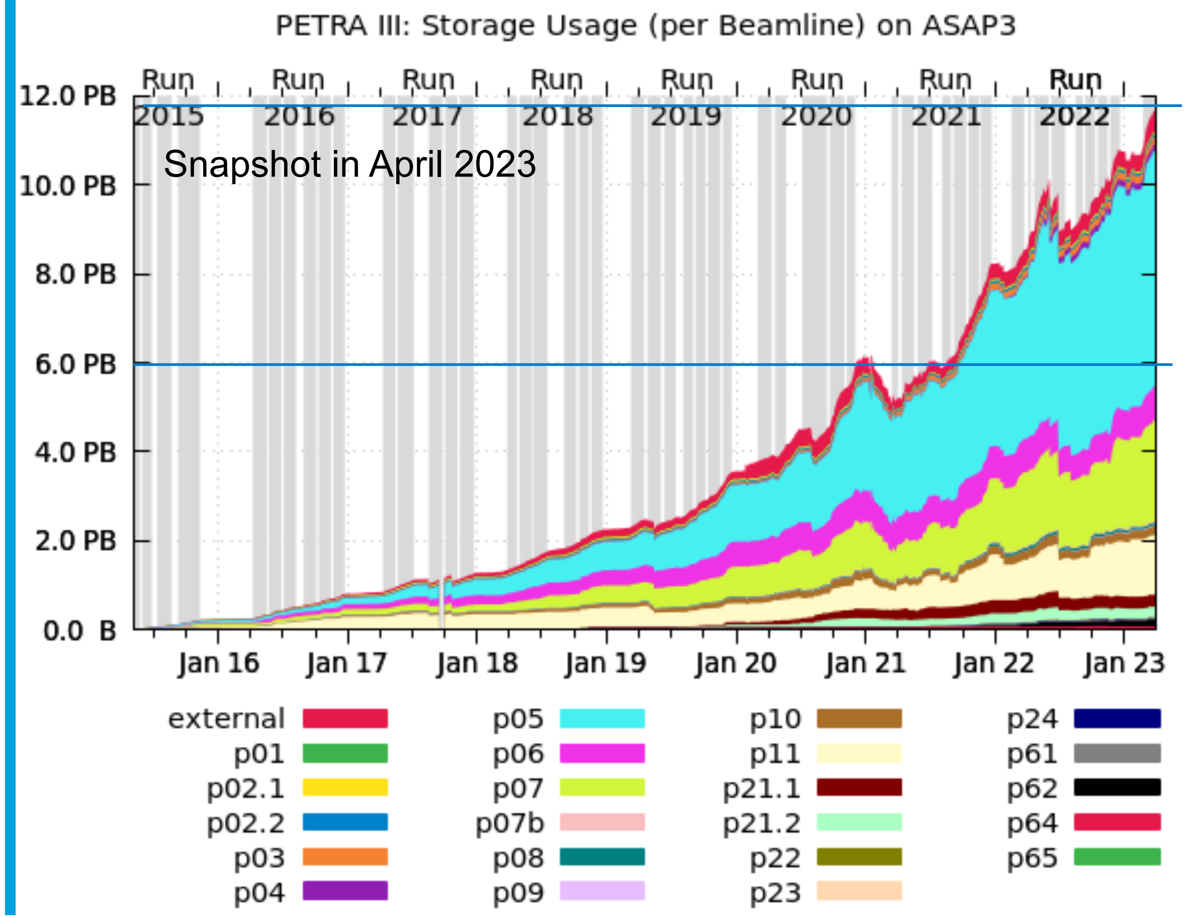
- Data policy
  - Data on disk for 180 days after measurement
    - (was: 180 days after last access)

  - Data migrated to tape after 180 days
    - retention on site (dCache), dual tape copy
    - 4.5 PB ingested to GPFS in past 12 months
    - 6 PB/year archived to tape
    - 12 PB tapes/yr with dual copy (€20K/PB/10YR)

- Usage highly variable between instruments

- Time to analyse data often limits publication rate
  - ~2 years from measurement to publication

- Hardware typically has a 5 year lifetime
  - Budget for regular replacement



Projection for PETRA-IV operation in 2028

PETRA-IV science output should not be storage or compute limited

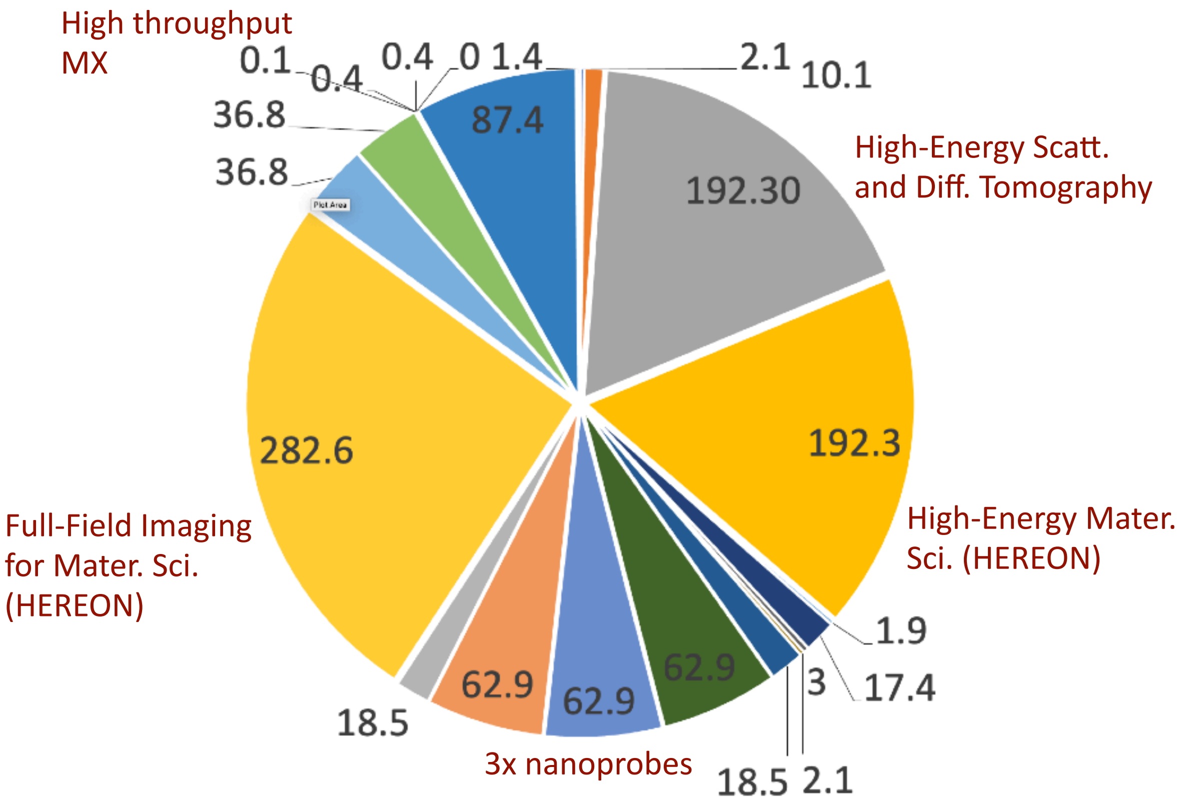
Peak total daily data generation **will exceed 1PB per day** based on actual peak 2021 GPFS usage

- Operation of any one instrument should not jeopardise operation of other instruments

By 2028, detectors will be larger and faster:

- Planned 130 kHz detector with a frame size of 10 MP and dynamical range of 2 Bytes, would produce **2.5 TB/s**
- Some **individual instruments will produce >1PB per day**
  - Luckily, not at all instruments are data volcanoes
- Increase inevitable **almost regardless of PETRA-IV project**

Peak daily data generated (in 2021)



Numbers are the **actual peak TB generated in 24 hours** by the comparable PETRA-III instrument in 2021

Projection for PETRA-IV operation in 2028

PETRA-IV science output should not be storage or compute limited

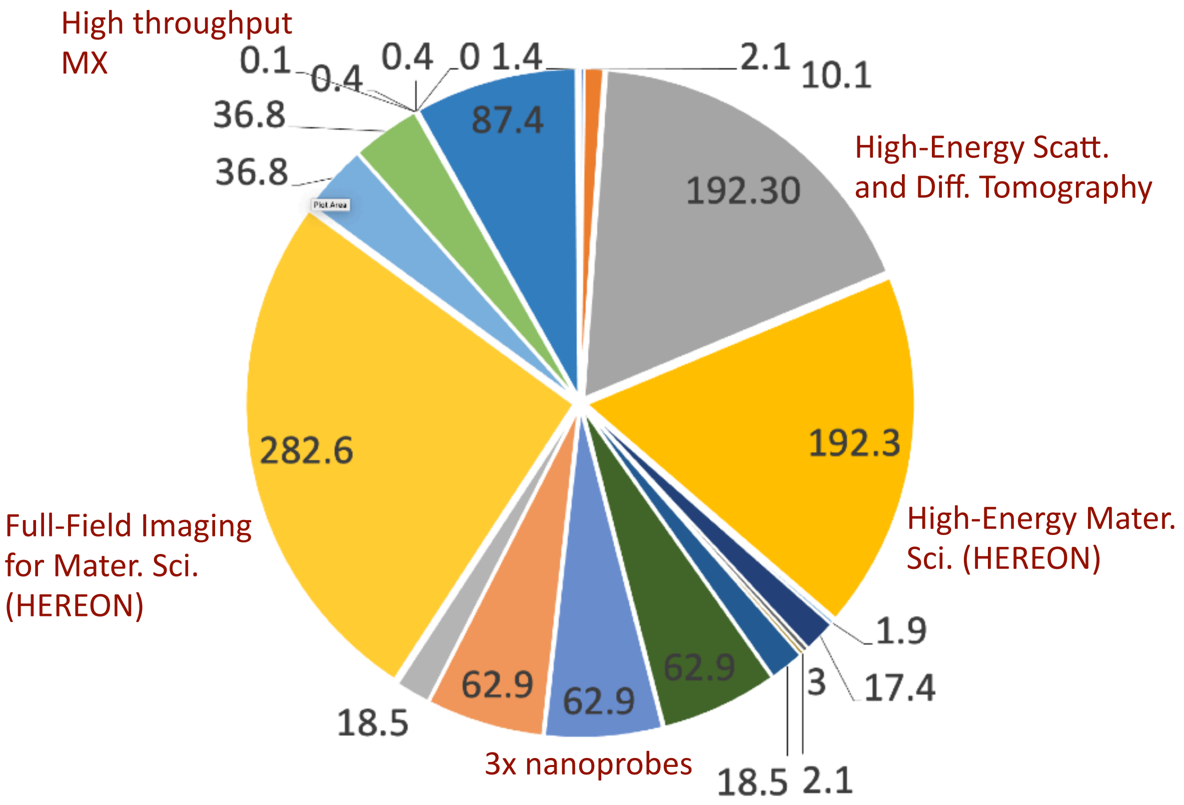
Peak total daily data generation **will exceed 1PB per day** based on actual peak 2021 GPFS usage

- Operation of any one instrument should not jeopardise operation of other instruments

By 2028, detectors will be larger and faster:

- Planned 130 kHz detector with a frame size of 10 MP and dynamical range of 2 Bytes, would produce **2.5 TB/s**
- Some **individual instruments will produce >1PB per day**
  - Luckily, not at all instruments are data volcanoes
- Increase inevitable **almost regardless of PETRA-IV project**

Peak daily data generated (in 2021)



Numbers are the **actual peak TB generated in 24 hours** by the comparable PETRA-III instrument in 2021

Reality check:

- Some instruments at ESRF already produce 1 PB per day
- In 2022, EuXFEL operating only 3 instruments simultaneously has produced 7 PB in a week (=364 PB/yr)
- 1 PB/day * 5 big data instruments * 180 days = 900 PB

In the future, retaining all data for 10 years is unaffordable

This problem will exist regardless of PETRA-IV

Continuing "business as usual" will:

- Over 500PB of disk space to keep data for 180 days, and up to 1EB of tape storage per year
- Cost > €150M for disks, plus > €50M per year for consumables and upkeep
- Consume between 1-2 MW of power and exceed the current data centre space
- Swamp users with complicated data further increasing time on disk and slowing science output
  - Performance metric is publications and citations (re-use) not PB on disk

In the future, retaining all data for 10 years is unaffordable

This problem will exist regardless of PETRA-IV

Continuing "business as usual" will:
- Over 500PB of disk space to keep data for 180 days, and up to 1EB of tape storage per year
- Cost > €150M for disks, plus > €50M per year for consumables and upkeep
- Consume between 1-2 MW of power and exceed the current data centre space
- Swamp users with complicated data further increasing time on disk and slowing science output
  - Performance metric is publications and citations (re-use) not PB on disk

Reality check:
- Some instruments at ESRF already produce 1 PB per day
- In 2022, EuXFEL operating only 3 instruments simultaneously has produced 7 PB in a week (=364 PB/yr)
- 1 PB/day * 5 instruments * 180 days = 900 PB

In the future, retaining all data for 10 years is unaffordable

This problem will exist regardless of PETRA-IV

Continuing “business as usual” will:

- Over 500PB of disk space to keep data for 180 days, and up to 1EB of tape storage per year
- Cost > €150M for disks, plus > €50M per year for consumables and upkeep
- Consume between 1-2 MW of power and exceed the current data centre space
- Swamp users with complicated data further increasing time on disk and slowing science output
    - Performance metric is publications and citations (re-use) not PB on disk

Cost driver

- PB of data generated (per instrument per day)
- PB of data saved to disk (reduction, veto or quotas)
- Number of high data rate instruments
- Number of days of operation
- Efficiency of data collection (and automation)

- Data reduction/compression ratios
- All data on GPFS for 180 days
- Efficiency of analysis (time on disk)
- Archive all data for 10 years

Reality check:

- Some instruments at ESRF already produce 1 PB per day
- In 2022, EuXFEL operating only 3 instruments simultaneously has produced 7 PB in a week (=364 PB/yr)
- 1 PB/day * 5 instruments * 180 days = 900 PB

In the future, retaining all data for 10 years is unaffordable

This problem will exist regardless of PETRA-IV

Continuing “business as usual” will:

- Over 500PB of disk space to keep data for 180 days, and up to 1EB of tape storage per year
- Cost > €150M for disks, plus > €50M per year for consumables and upkeep
- Consume between 1-2 MW of power and exceed the current data centre space
- Swamp users with complicated data further increasing time on disk and slowing science output
    - Performance metric is publications and citations (re-use) not PB on disk

Cost driver

- PB of data generated (per instrument per day)
- PB of data saved to disk (reduction, veto or quotas)
- Number of high data rate instruments
- Number of days of operation
- Efficiency of data collection (and automation)

- Data reduction/compression ratios
- All data on GPFS for 180 days
- Efficiency of analysis (time on disk)
- Archive all data for 10 years

Remedial action

- Smaller, slower detectors; shorter measurement time
- Enforce limited data quotas (eg: 1PB/week)
- High data instruments are mostly the flagships
- Reduce user hours (ie: measurement time)
- Operate inefficiently (eg: manual alignment…)

Reality check:

- Some instruments at ESRF already produce 1 PB per day
- In 2022, EuXFEL operating only 3 instruments simultaneously has produced 7 PB in a week (=364 PB/yr)
- 1 PB/day * 5 instruments * 180 days = 900 PB

In the future, retaining all data for 10 years is unaffordable

This problem will exist regardless of PETRA-IV

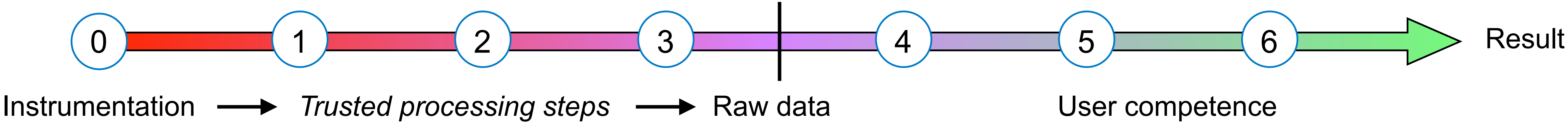
Continuing "business as usual" will:
- Over 500PB of disk space to keep data for 180 days, and up to 1EB of tape storage per year
- Cost > €150M for disks, plus > €50M per year for consumables and upkeep
- Consume between 1-2 MW of power and exceed the current data centre space
- Swamp users with complicated data further increasing time on disk and slowing science output
  - Performance metric is publications and citations (re-use) not PB on disk

Cost driver	**Remedial action**
PB of data generated (per instrument per day)	Smaller, slower detectors; shorter measurement time
PB of data saved to disk (reduction, veto or quotas)	Enforce limited data quotas (eg: 1PB/week)
Number of high data rate instruments	High data instruments are mostly the flagships
Number of days of operation	Reduce user hours (ie: measurement time)
Efficiency of data collection (and automation)	Operate inefficiently (eg: manual alignment…)
Data reduction/compression ratios	Will do this anyway (lossless ~4x, lossy varies a lot)
All data on GPFS for 180 days	Data stays on GPFS for ~30 days (6x reduction)
Efficiency of analysis (time on disk)	Optimised pipelines, particularly for measurements
Archive all data for 10 years	Redefine what gets kept after 30 days

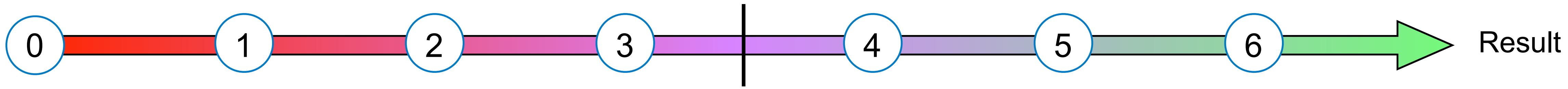
Reality check:
- Some instruments at ESRF already produce 1 PB per day
- In 2022, EuXFEL operating only 3 instruments simultaneously has produced 7 PB in a week (=364 PB/yr)
- 1 PB/day * 5 instruments * 180 days = 900 PB

The definition of raw data depends on your starting point

The definition of raw data depends on your starting point



The definition of raw data depends on your starting point



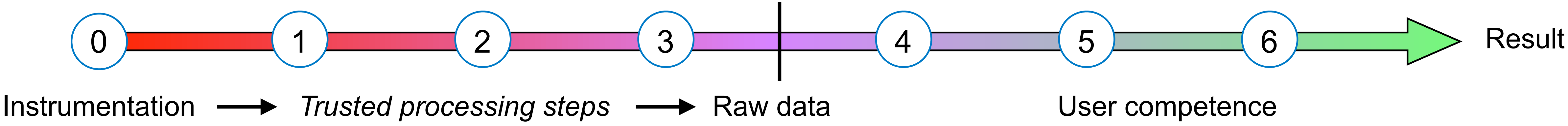
Instrumentation → *Trusted processing steps* → Raw data   User competence

I need the raw detector data

Here it is in ADU

Can you give me the photons per pixel?

The definition of raw data depends on your starting point



I need the raw detector data

Here it is in ADU

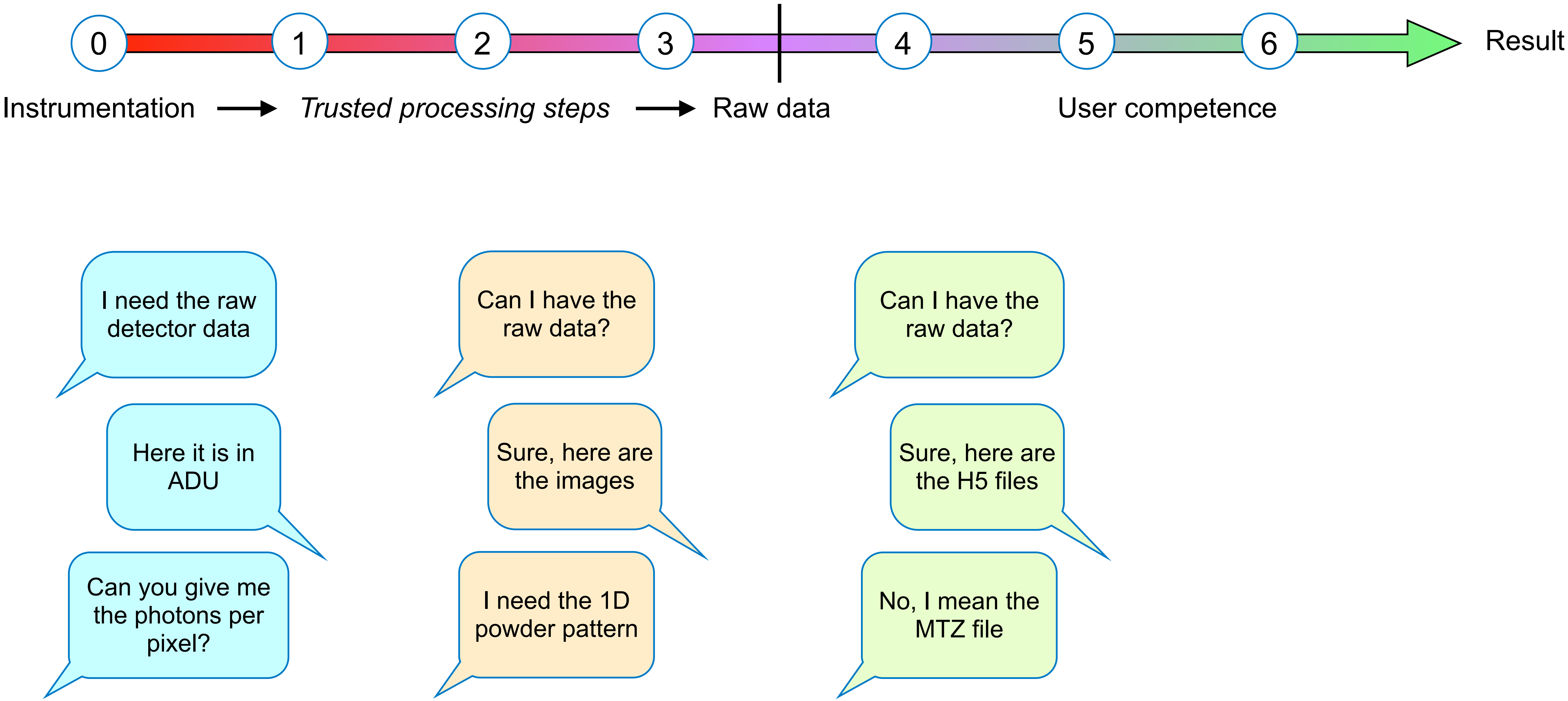
Can you give me the photons per pixel?

Can I have the raw data?

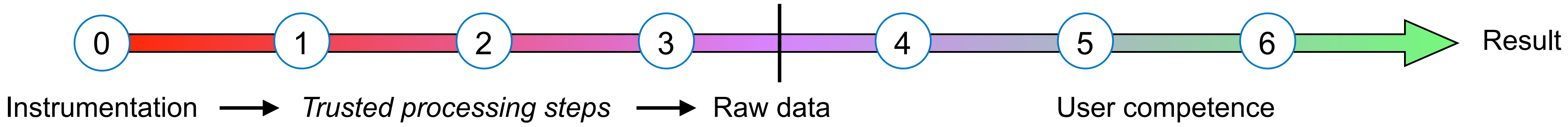
Sure, here are the images

I need the 1D powder pattern

The definition of raw data depends on your starting point



The definition of raw data depends on your starting point



I need the raw detector data

Here it is in ADU

Can you give me the photons per pixel?

Can I have the raw data?

Sure, here are the images

I need the 1D powder pattern

Can I have the raw data?

Sure, here are the H5 files

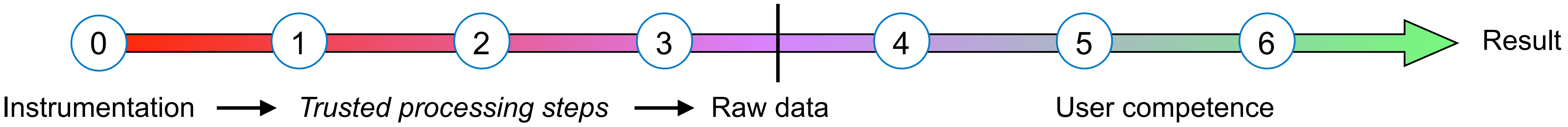
No, I mean the MTZ file

Can I have the raw fluorescence data?

The event stream from the detector is here

Can I have the chemical concentration?

The definition of raw data depends on your starting point



'Raw data' is typically the input to the user's own data analysis pipeline or the limit of their expertise

I need the raw detector data

Here it is in ADU

Can you give me the photons per pixel?

Can I have the raw data?

Sure, here are the images

I need the 1D powder pattern

Can I have the raw data?

Sure, here are the H5 files

No, I mean the MTZ file

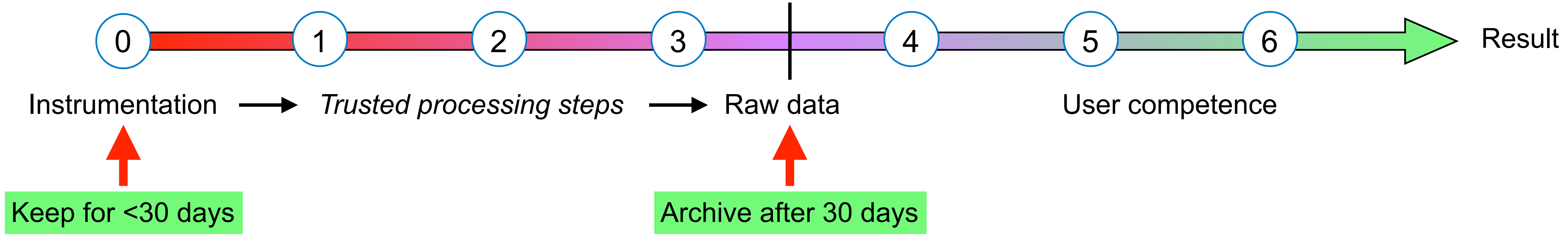
Can I have the raw fluorescence data?

The event stream from the detector is here

Can I have the chemical concentration?

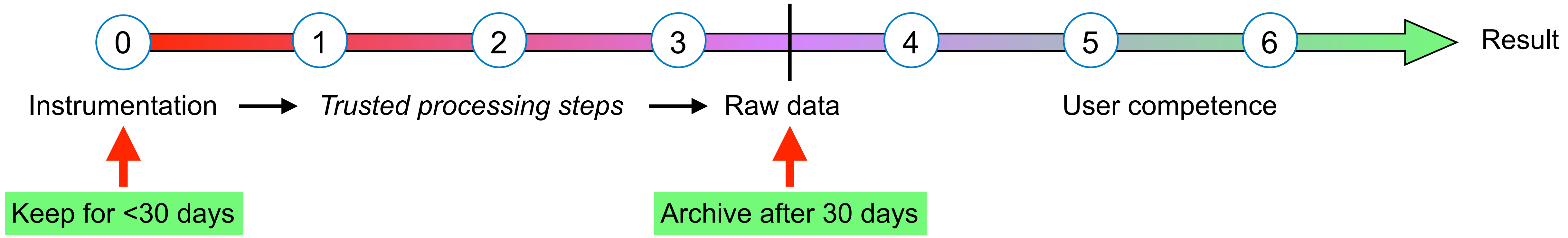
PETRA-IV requires a different approach to data retention

Provide results to users and retain low-level data for only 30 days



PETRA-IV requires a different approach to data retention

Provide results to users and retain low-level data for only 30 days

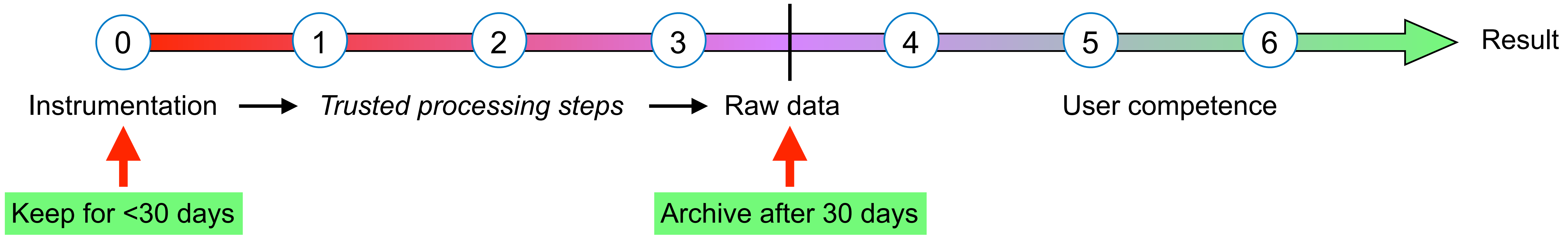


Aim to **maximise science content of stored data:**

1) Develop trusted and validated processing pipelines to efficiently deliver results to users
2) Processing output is the product we give to the users,
3) Keep instrument data for 30 days during which time processing problems can be corrected
4) Develop a data weeding strategy (policy for discarding data), including (maybe) deleting raw data

PETRA-IV requires a different approach to data retention

Provide results to users and retain low-level data for only 30 days

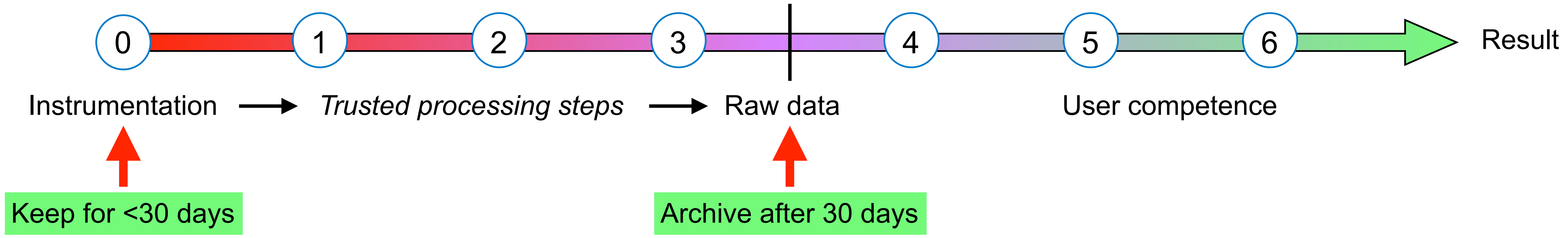


Aim to **maximise science content of stored data:**

1) Develop trusted and validated processing pipelines to efficiently deliver results to users
2) Processing output is the product we give to the users,
3) Keep instrument data for 30 days during which time processing problems can be corrected
4) Develop a data weeding strategy (policy for discarding data), including (maybe) deleting raw data

PETRA-IV requires a different approach to data retention

Provide results to users and retain low-level data for only 30 days



Aim to **maximise science content of stored data:**

1) Develop trusted and validated processing pipelines to efficiently deliver results to users
2) Processing output is the product we give to the users,
3) Keep instrument data for 30 days during which time processing problems can be corrected
4) Develop a data weeding strategy (policy for discarding data), including (maybe) deleting raw data

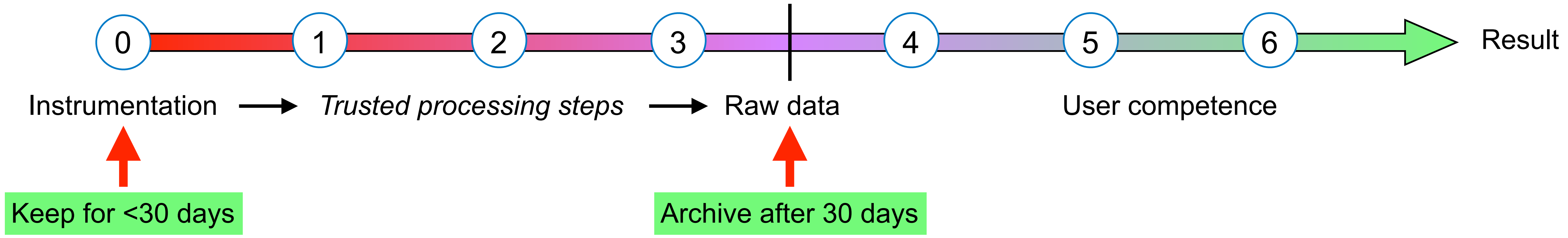
Data processing vs analysis:

- **Data analysis** is where researchers turn processed data into scientific knowledge
  - Domain specialists interpret processed data to answer their science question(s)

- **Processed data** is in a form suitable for the non-expert user to continue with their analysis
  - Processing is often generic and deterministic; optimised pipelines can be provided
  - Calibration, geometry and masking procedures must be standardised, minimise parameter tweaking
  - Often highly reduced data volumes (at worst, avoid data duplication)

PETRA-IV requires a different approach to data retention

Provide results to users and retain low-level data for only 30 days



Aim to **maximise science content of stored data:**

1) Develop trusted and validated processing pipelines to efficiently deliver results to users
2) Processing output is the product we give to the users,
3) Keep instrument data for 30 days during which time processing problems can be corrected
4) Develop a data weeding strategy (policy for discarding data), including (maybe) deleting raw data

Data processing vs analysis:

- **Data analysis** is where researchers turn processed data into scientific knowledge
  - Domain specialists interpret processed data to answer their science question(s)

- **Processed data** is in a form suitable for the non-expert user to continue with their analysis
  - Processing is often generic and deterministic; optimised pipelines can be provided
  - Calibration, geometry and masking procedures must be standardised, minimise parameter tweaking
  - Often highly reduced data volumes (at worst, avoid data duplication)

Not all instruments or samples are equal - policies must be implemented sensibly

The human factor: trust is a big issue

Confidence that initial processing is correct must be ensured

The human factor: trust is a big issue

Confidence that initial processing is correct must be ensured

Online analysis may be 80% good enough,
but I need 95%-100% for my paper

The human factor: trust is a big issue

Confidence that initial processing is correct must be ensured

Online analysis may be 80% good enough, but I need 95%-100% for my paper

Detector location needs to be refined from the data (experiment geometry)

The detector response is not properly calibrated

Beam centre may move, I need to refine it from the data

Can’t trust the metadata (detector distance, wavelength)

ROI adjustment, eg: unexpected bad regions on the detectors (shadows)

The human factor: trust is a big issue

Confidence that initial processing is correct must be ensured

Online analysis may be 80% good enough, but I need 95%-100% for my paper

Detector location needs to be refined from the data (experiment geometry)

The detector response is not properly calibrated

Beam centre may move, I need to refine it from the data

Can't trust the metadata (detector distance, wavelength)

ROI adjustment, eg: unexpected bad regions on the detectors (shadows)

I need to convert file formats

PhD student needs to do the analysis / write the code

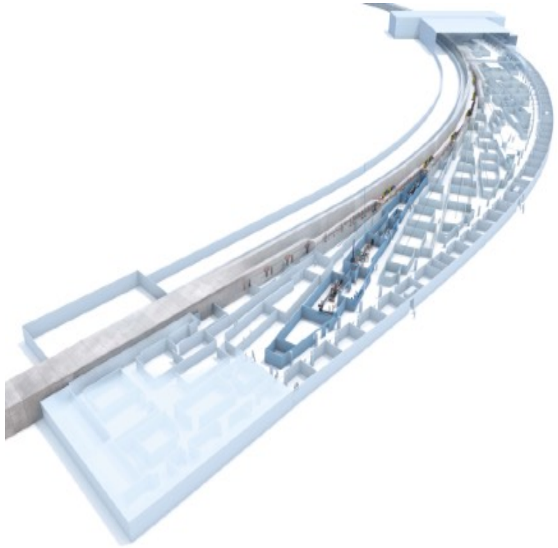
My analysis only works using files

There are parameters/ thresholds to tweak

Just in case… I need to check it's right myself

We move data to the central data centre as soon as possible

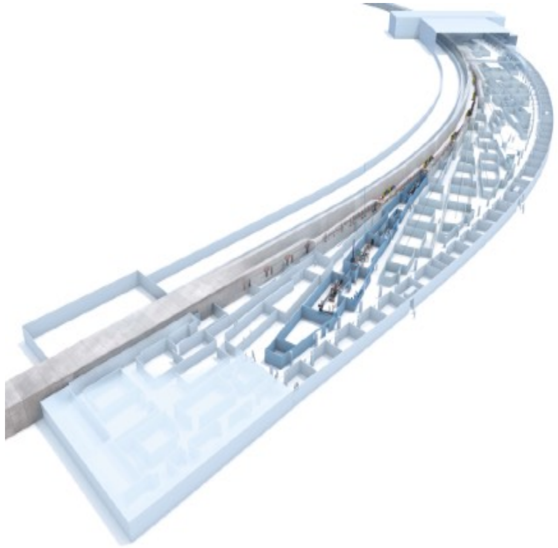
Exploit large scale shared infrastructure



Dedicated 96x 400GE fibre
per instrument

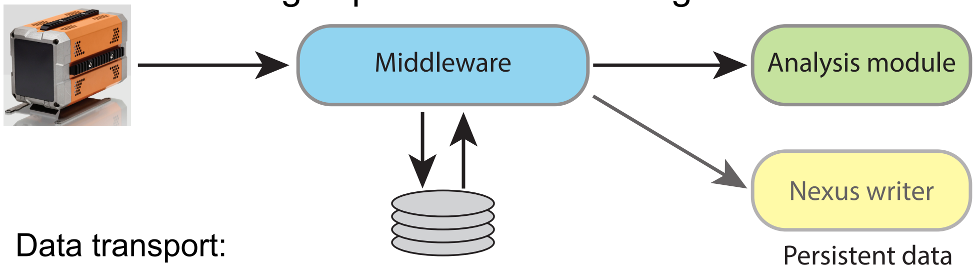
We move data to the central data centre as soon as possible

Exploit large scale shared infrastructure



Dedicated 96x 400GE fibre
per instrument

ASAP::O
High speed data streaming



Data transport:

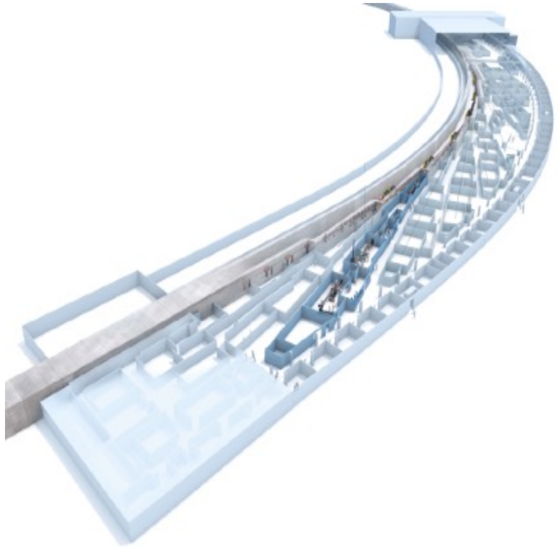
- Infiniband
- RDMA
- TCP…

Configurable data cache
• Zero size = purely live
• Memory size = buffered
• A few minutes = 'replay' of recent data
• Whole beamtime for Nexus conversion

Persistent data
"The keep"

We move data to the central data centre as soon as possible

Exploit large scale shared infrastructure



Dedicated 96x 400GE fibre
per instrument

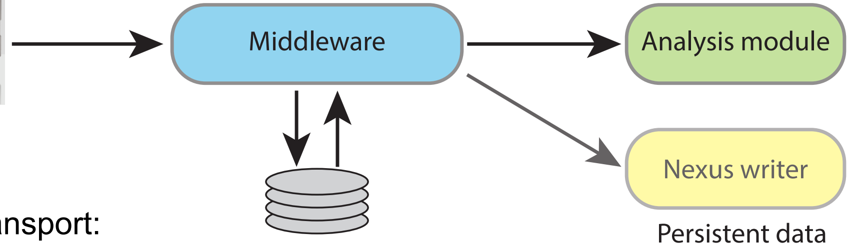
Tier-0

On site

Initial storage:
1. In-memory data access
2. Fast SSD burst cache
3. High performance GPFS

Online computing
(CPU + GPU)

ASAP::O
High speed data streaming



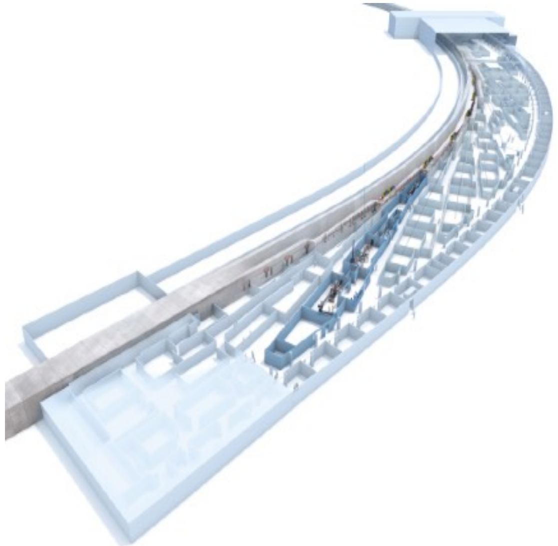
Data transport:
- Infiniband
- RDMA
- TCP…

Configurable data cache
- Zero size = purely live
- Memory size = buffered
- A few minutes = 'replay' of recent data
- Whole beamtime for Nexus conversion

Persistent data
"The keep"

We move data to the central data centre as soon as possible

Exploit large scale shared infrastructure



Dedicated 96x 400GE fibre
per instrument

Tier-0

On site

Initial storage:
1. In-memory data access
2. Fast SSD burst cache
3. High performance GPFS

Online computing
(CPU + GPU)

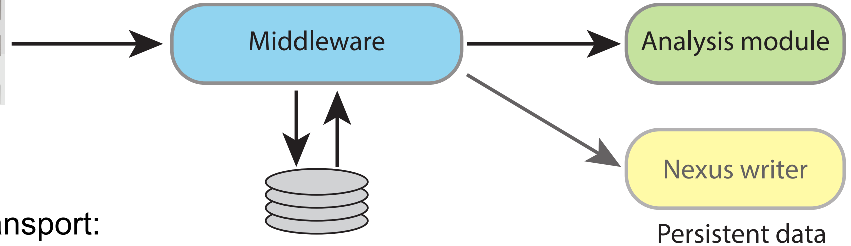
Tier-1

On site or federated

Longer term storage:
1. Commodity dCache
2. Tape archive

Offline computing
(CPU + GPU)

ASAP::O
High speed data streaming



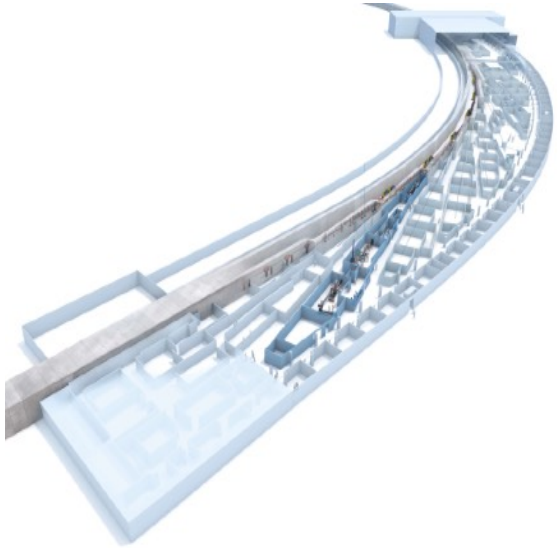
Data transport:
- Infiniband
- RDMA
- TCP…

Configurable data cache
- Zero size = purely live
- Memory size = buffered
- A few minutes = 'replay' of recent data
- Whole beamtime for Nexus conversion

Persistent data
"The keep"

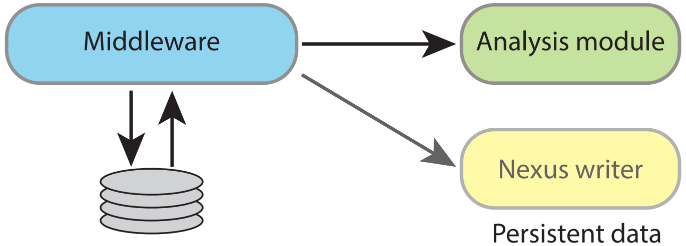
We move data to the central data centre as soon as possible

Exploit large scale shared infrastructure



Dedicated 96x 400GE fibre
per instrument

ASAP::O
High speed data streaming



Data transport:
- Infiniband
- RDMA
- TCP…

Configurable data cache
- Zero size = purely live
- Memory size = buffered
- A few minutes = 'replay' of recent data
- Whole beamtime for Nexus conversion

Persistent data
"The keep"

Tier-0

On site

Initial storage:
1. In-memory data access
2. Fast SSD burst cache
3. High performance GPFS

Online computing
(CPU + GPU)

On-site load balance

Tier-1

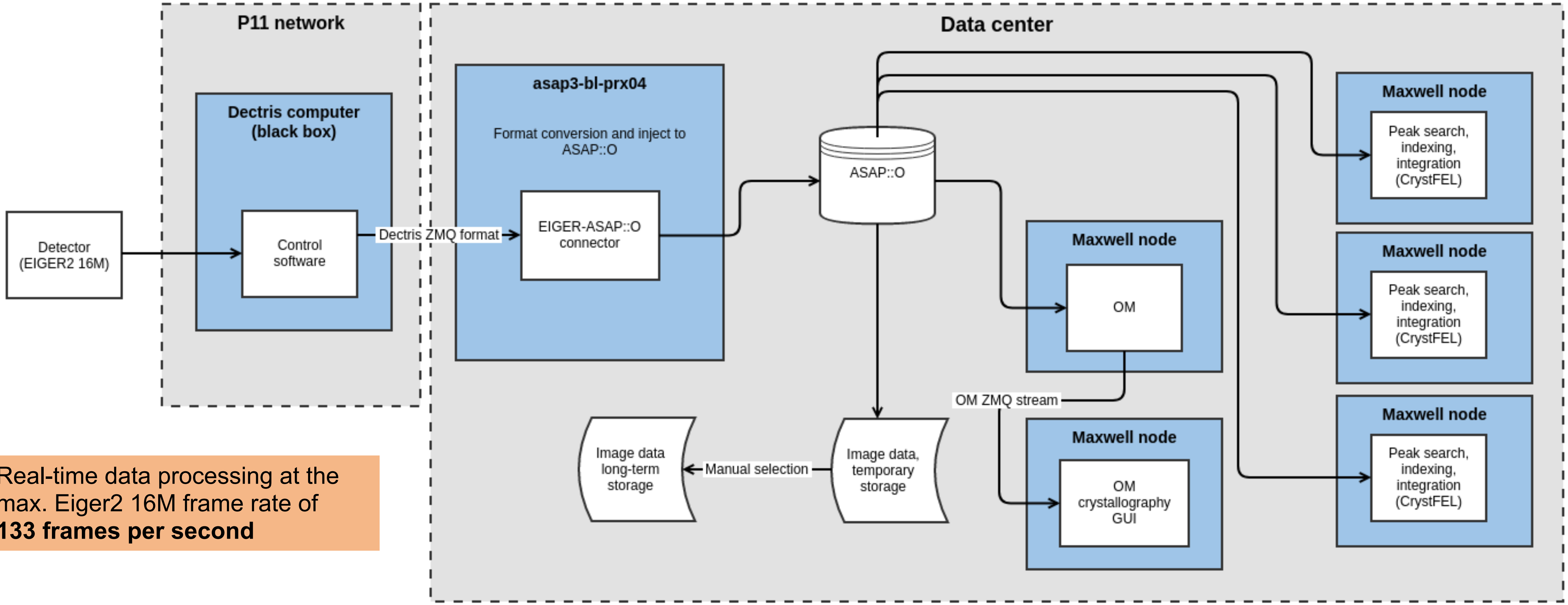
On site or federated

Longer term storage:
1. Commodity dCache
2. Tape archive

Offline computing
(CPU + GPU)

We already process and reduce data before it is saved to disk

Real time serial crystallography at P11 using central compute resources



Real-time data processing at the max. Eiger2 16M frame rate of **133 frames per second**

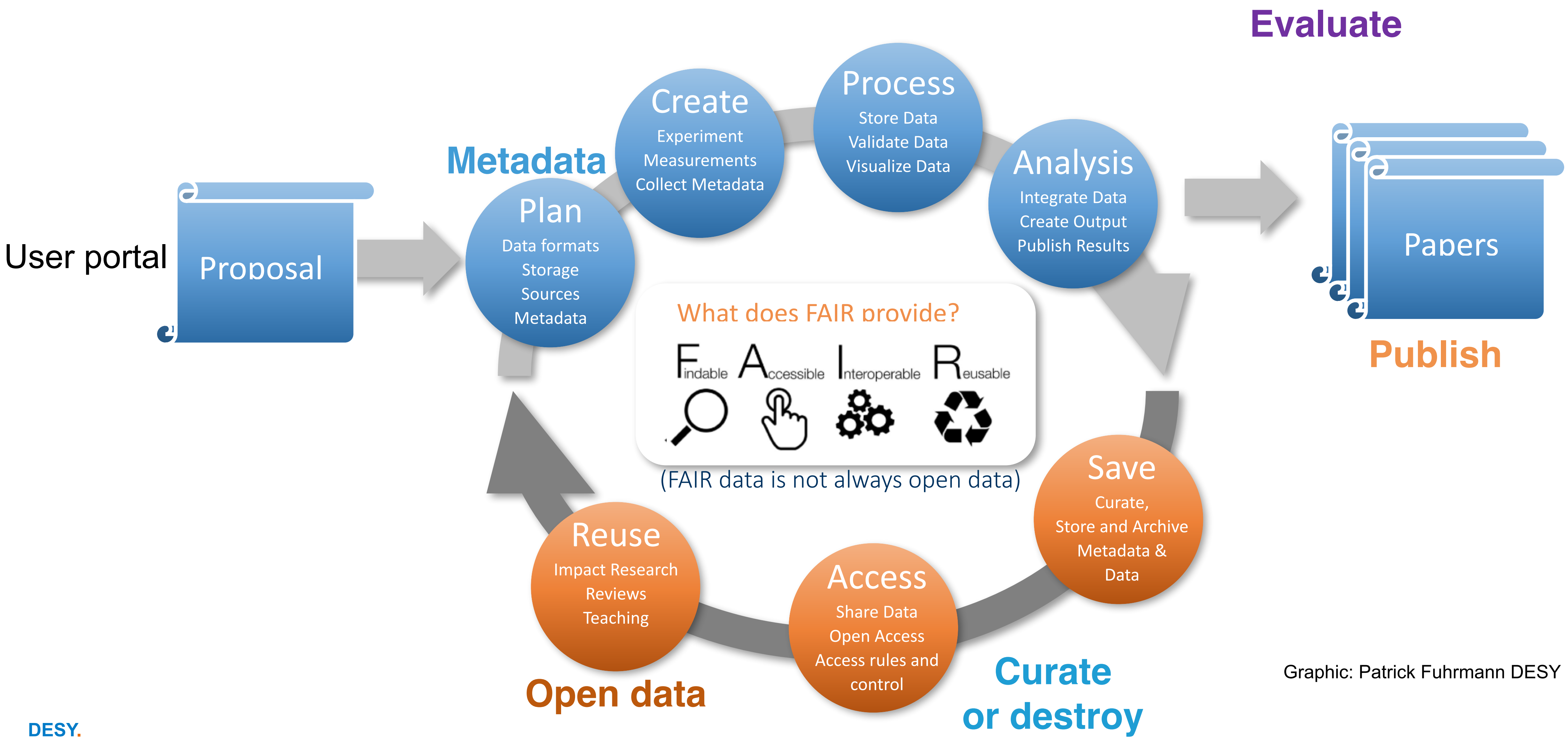
- Average 200 ms per frame (5 frames per second per CPU) when working on the full 16 megapixel frames from Eiger (16 bits per pixel).
- Uses two dedicated computers (2x 192 CPUs) running CrystFEL plus other parts of the pipeline (NeXus writer, OM, OM GUI, binning worker)
- ASAP::O handles high speed data transfer, bookkeeping, etc; always performed after a similar experiment for which the calibrations exist

More details in Alexandra's talk later today

Tom White, et.al. DESY

PETRA-IV will offer services for the complete data life cycle

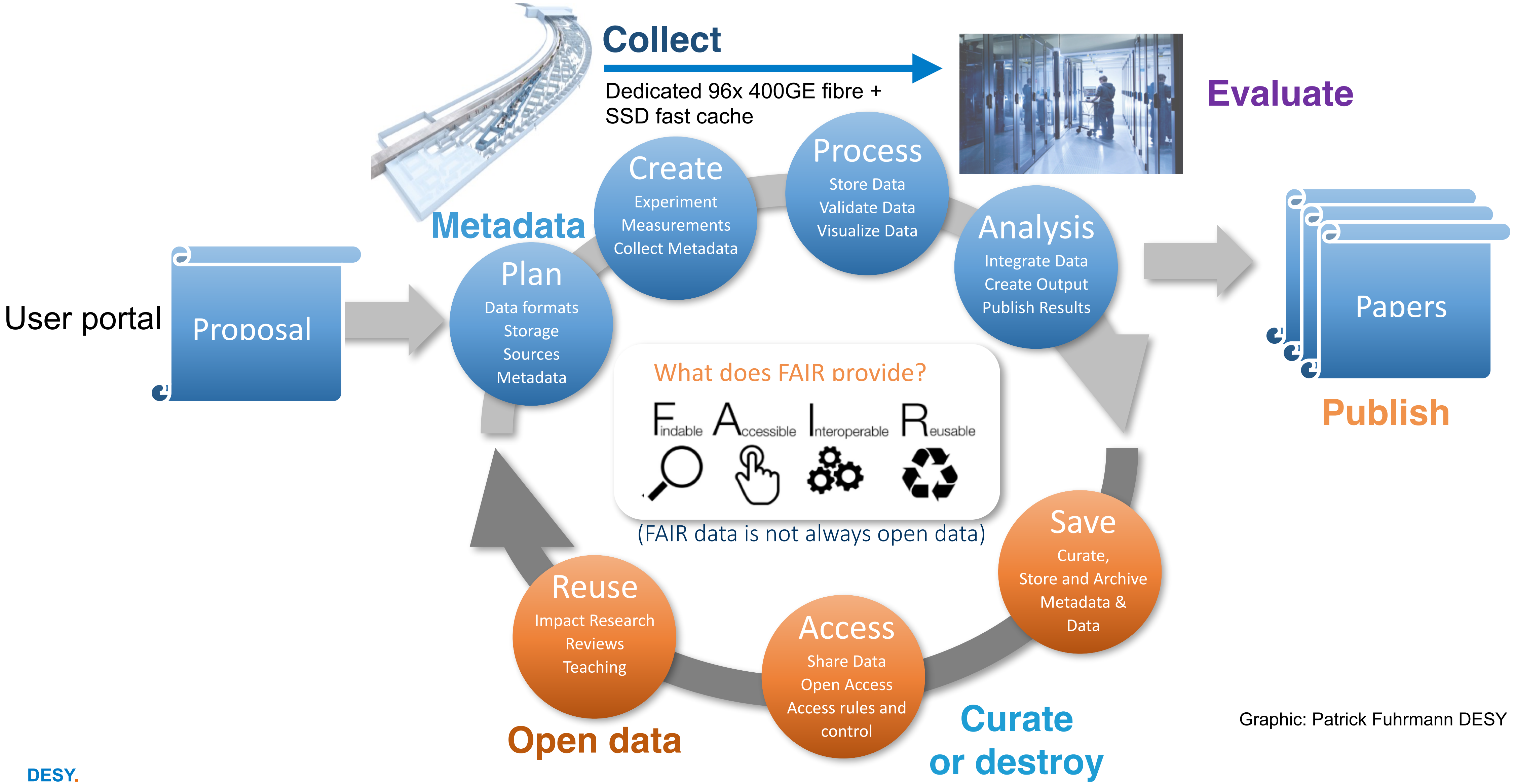
Data management by the facility for the a non expert user community



Graphic: Patrick Fuhrmann DESY

PETRA-IV will offer services for the complete data life cycle

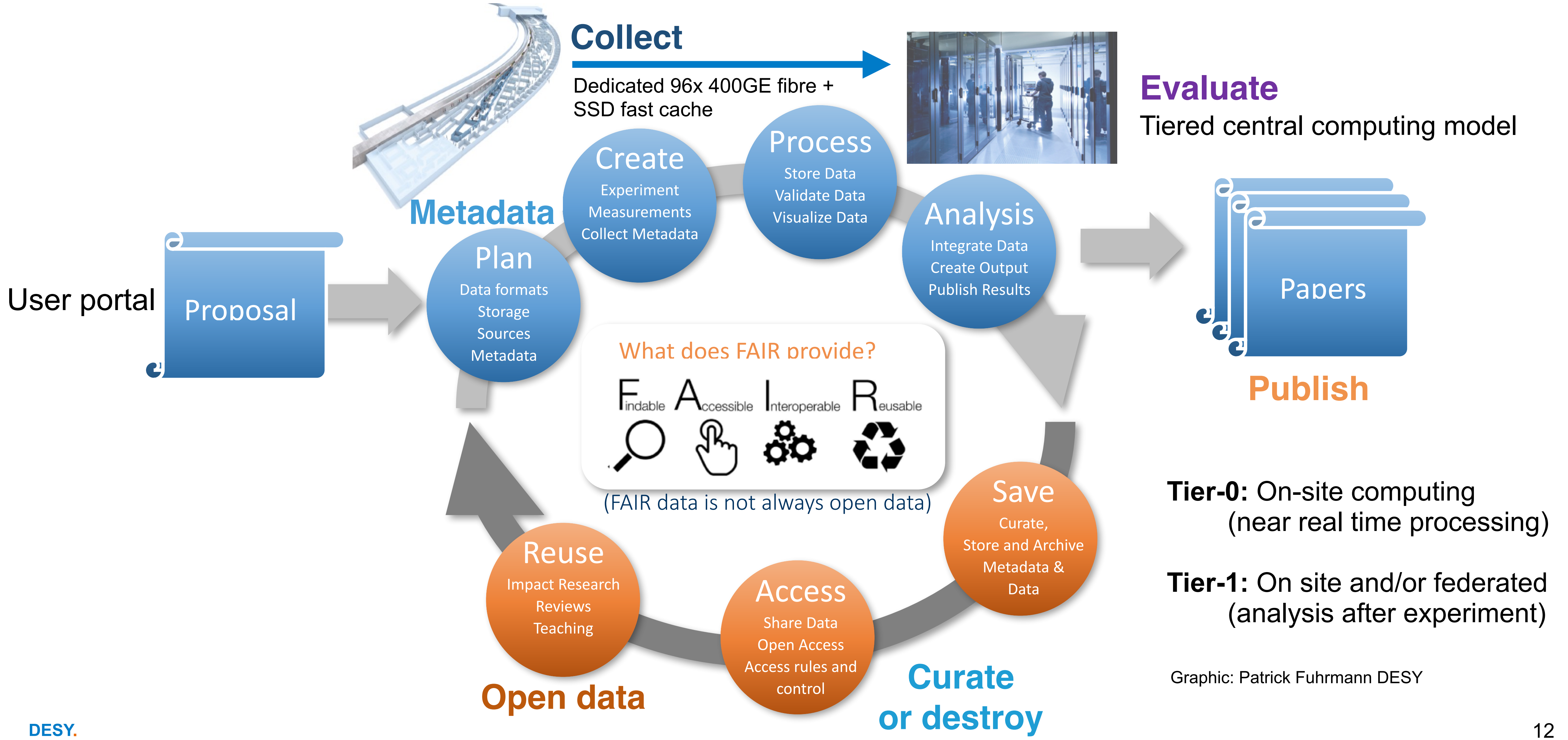
Data management by the facility for the a non expert user community



Graphic: Patrick Fuhrmann DESY

PETRA-IV will offer services for the complete data life cycle

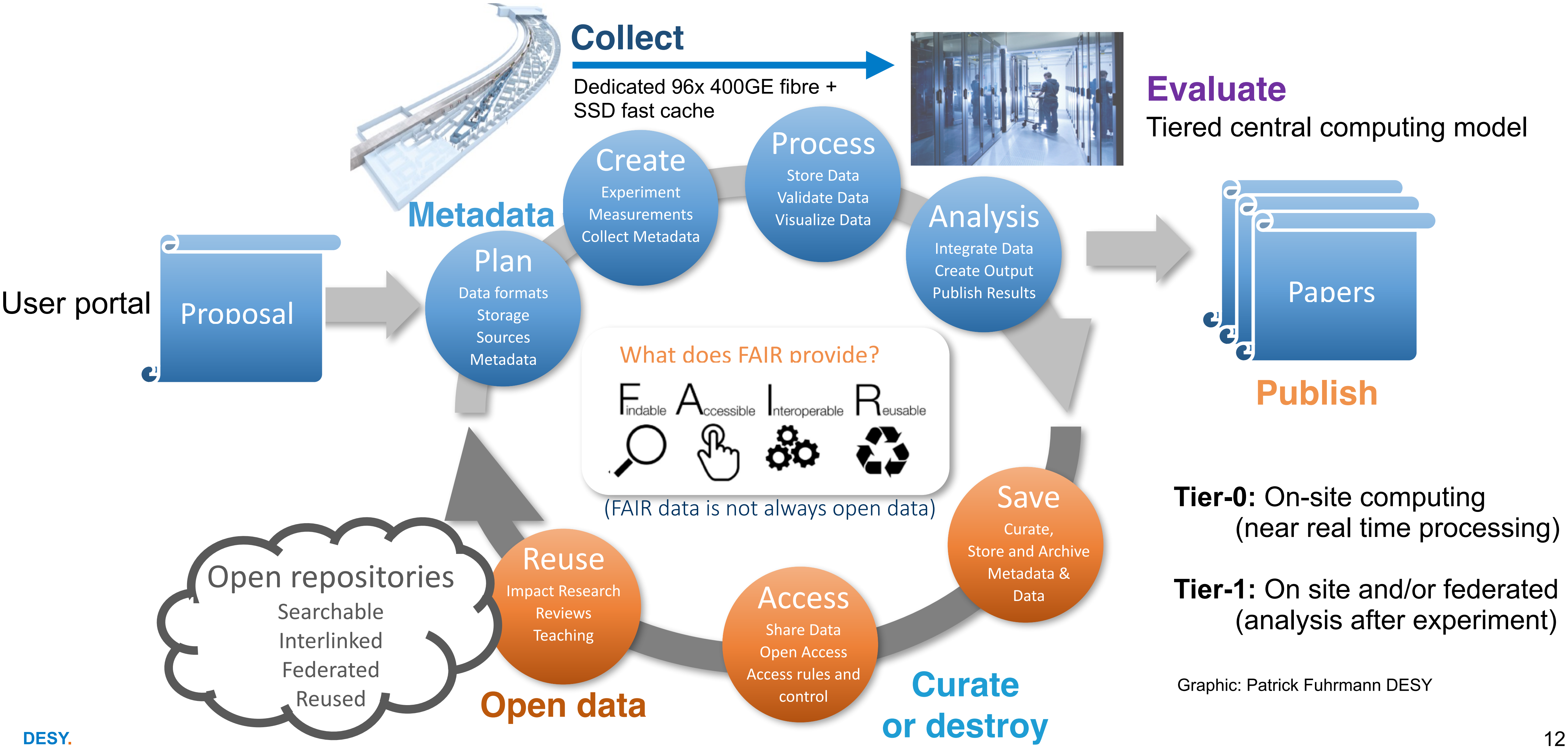
Data management by the facility for the a non expert user community



Graphic: Patrick Fuhrmann DESY

PETRA-IV will offer services for the complete data life cycle

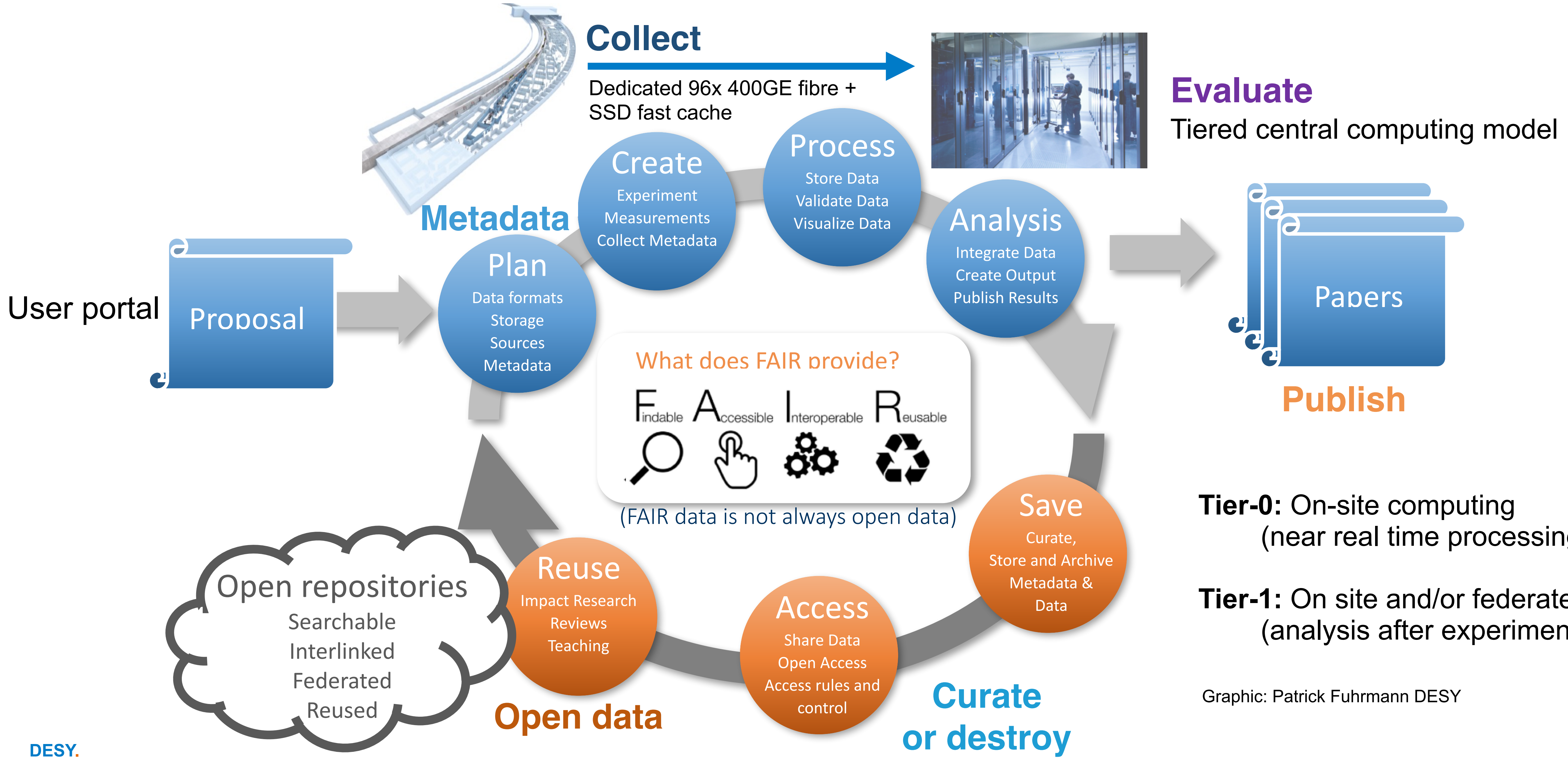
Data management by the facility for the a non expert user community



Graphic: Patrick Fuhrmann DESY

PETRA-IV will offer complete data life cycle services

Data management for the non expert user community



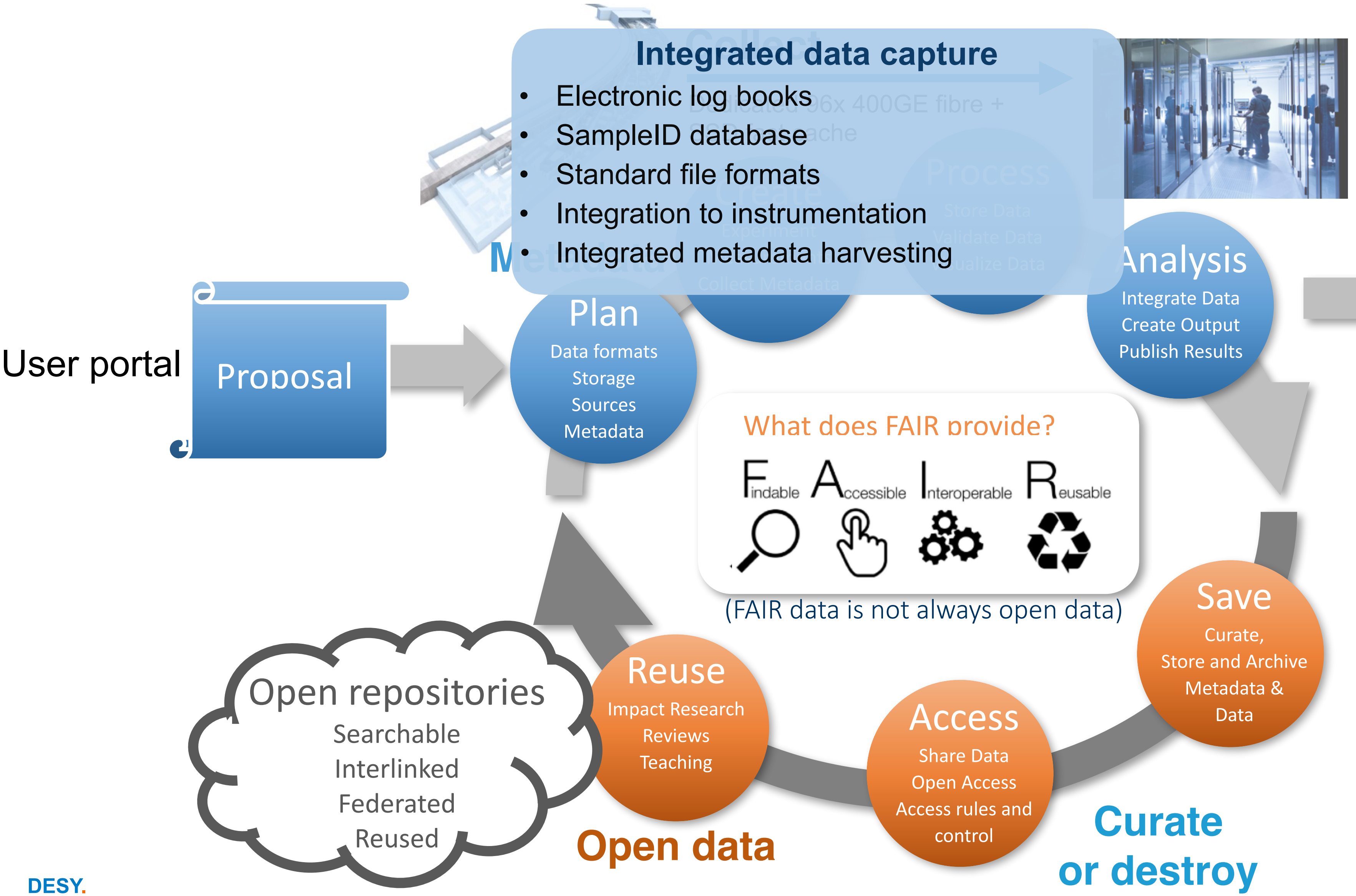
Tier-0: On-site computing
(near real time processing)

Tier-1: On site and/or federated
(analysis after experiment)

Graphic: Patrick Fuhrmann DESY

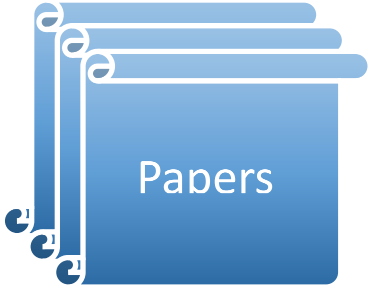
PETRA-IV will offer complete data life cycle services

Data management for the non expert user community



Evaluate

Tiered central computing model



Publish

Tier-0: On-site computing
(near real time processing)

Tier-1: On site and/or federated
(analysis after experiment)

Graphic: Patrick Fuhrmann DESY

PETRA-IV will offer complete data life cycle services

Data management for the non expert user community

Integrated data capture

- Electronic log books
- SampleID database
- Standard file formats
- Integration to instrumentation
- Integrated metadata harvesting

Evaluate

Tiered central computing model

Infrastructure for data evaluation

- Central HPC resources
- Sustainable and reusable software ecosystem
- Power user software deployed for all users
- Remote data evaluation
- Containerisation

Tier-0: On-site computing
(near real time processing)

Tier-1: On site and/or federated
(analysis after experiment)

User portal

Proposal

Plan
Data formats
Storage
Sources
Metadata

Analysis
Integrate Data

What does FAIR provide?

Findable Accessible Interoperable Reusable

(FAIR data is not always open data)

Save
Curate,
Store and Archive
Metadata &
Data

Access
Share Data
Open Access
Access rules and
control

Reuse
Impact Research
Reviews
Teaching

Open repositories
Searchable
Interlinked
Federated
Reused

Open data

Curate or destroy

Graphic: Patrick Fuhrmann DESY

PETRA-IV will offer complete data life cycle services

Data management for the non expert user community

Integrated data capture

- Electronic log books
- SampleID database
- Standard file formats
- Integration to instrumentation
- Integrated metadata harvesting

Evaluate

Tiered central computing model

User portal

Proposal

Plan

Data formats
Storage
Sources
Metadata

Analysis

Integrate Data

Infrastructure for data evaluation

- Central HPC resources
- Sustainable and reusable software ecosystem
- Power user software deployed for all users
- Remote data evaluation
- Containerisation

Tier-0: On-site computing
(near real time processing)

Tier-1: On site and/or federated
(analysis after experiment)

Graphic: Patrick Fuhrmann DESY

Open repositories

Searchable
Interlinked
Federated
Reused

(Open) data repositories and catalogues

- Raw data can be made open
- Searchable federated catalogues (with access control)
- Common user IDs (AAI)
- DOI minting
- Place to put open results with processed data

or destroy

PETRA-IV will offer complete data life cycle services

Data management for the non expert user community

Integrated data capture

- Electronic log books
- SampleID database
- Standard file formats
- Integration to instrumentation
- Integrated metadata harvesting

IGSN

Evaluate

Tiered central computing model

User portal

Proposal

Plan
Data formats
Storage
Sources
Metadata

Analysis
Integrate Data

What does FAIR provide?

Findable Accessible Interoperable Reusable

Infrastructure for data evaluation

- Central HPC resources
- Sustainable and reusable software ecosystem
- Power user software deployed for all users
- Remote data evaluation
- Containerisation

Tier-0: On-site computing
(near real time processing)

Tier-1: On site and/or federated
(analysis after experiment)

Open repositories
Searchable
Interlinked
Federated
Reused

(Open) data repositories and catalogues

- Raw data can be made open
- Searchable federated catalogues (with access control)
- Common user IDs (AAI)
- DOI minting
- Place to put open results with processed data

or destroy

Graphic: Patrick Fuhrmann DESY

PETRA-IV will offer complete data life cycle services

Data management for the non expert user community

User portal

Proposal

Plan
Data formats
Storage
Sources
Metadata

Integrated data capture
- Electronic log books
- SampleID database
- Standard file formats
- Integration to instrumentation
- Integrated metadata harvesting

Evaluate
Tiered central computing model

Analysis
Integrate Data

Infrastructure for data evaluation
- Central HPC resources
- Sustainable and reusable software ecosystem
- Power user software deployed for all users
- Remote data evaluation
- Containerisation

Tier-0: On-site computing
(near real time processing)

Tier-1: On site and/or federated
(analysis after experiment)

Open repositories
Searchable
Interlinked
Federated
Reused

(Open) data repositories and catalogues
- Raw data can be made open
- Searchable federated catalogues (with access control)
- Common user IDs (AAI)
- DOI minting
- Place to put open results with processed data

or destroy

Graphic: Patrick Fuhrmann DESY

PETRA-IV will offer complete data life cycle services

Data management for the non expert user community

DAPHNE 4NFDI

nfdi Nationale Forschungsdaten Infrastruktur

User portal

Proposal

Plan
Data formats
Storage
Sources
Metadata

Integrated data capture

- Electronic log books
- SampleID database
- Standard file formats
- Integration to instrumentation
- Integrated metadata harvesting

IGSN

Evaluate

Tiered central computing model

Analysis
Integrate Data

Infrastructure for data evaluation

- Central HPC resources
- Sustainable and reusable software ecosystem
- Power user software deployed for all users
- Remote data evaluation
- Containerisation

What does FAIR provide?

Findable Accessible Interoperable Reusable

Tier-0: On-site computing
(near real time processing)

Tier-1: On site and/or federated
(analysis after experiment)

(Open) data repositories and catalogues

- Raw data can be made open
- Searchable federated catalogues (with access control)
- Common user IDs (AAI)
- DOI minting
- Place to put open results with processed data

SciCat

Open repositories
Searchable
Interlinked
Federated
Reused

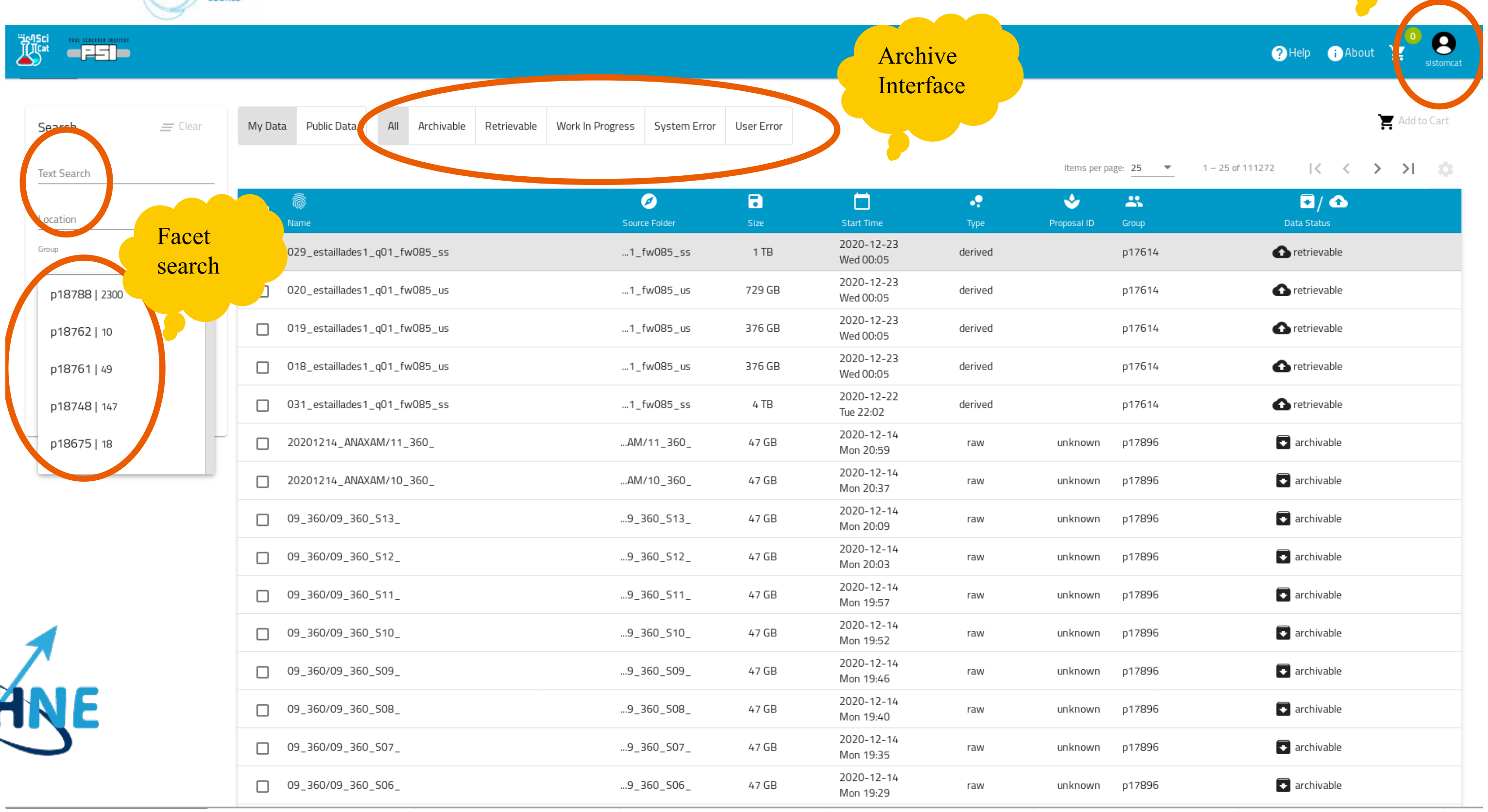
or destroy

Graphic: Patrick Fuhrmann DESY

SciCat as a catalogue foundation

We are in the process of deploying and developing SciCat as our data catalogue

Discover data via WebUI



Some features:

- Data browsing
- Data search
- Data download
- Access control
- Federated login
- Metadata management
- Online logbooks
- Online chat session
- DataDOI generation
- Archive interface
- Catalogue harvesting
- Data previews
- 'Data lake' for
  - reference datasets
  - simulations
  - research group data

Findable Accessible Interoperable Reusable

Initial development by

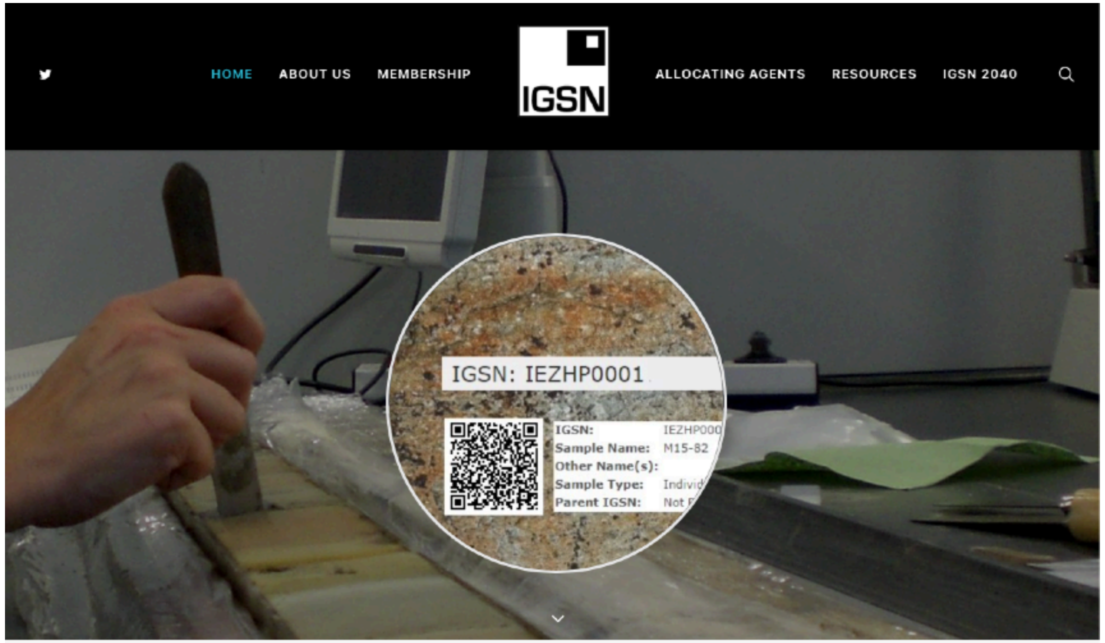
Unique sample identifiers

Tracking samples from creation through to data and publication

- Uniquely identify samples so that they can be tracked through logbooks and datasets
- Identifier should be unique and persistent - even though samples themselves may not always persistent
- Must be simple, easy to use, minimal paperwork overhead

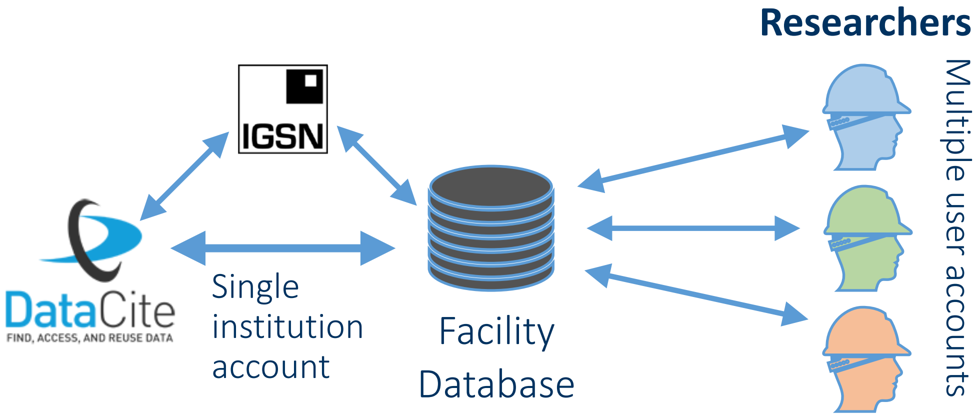
The IGSN* system has been developed for other disciplines

IGSN is a globally unique and persistent identifier for material samples.



https://www.igsn.org/ https://ardc.edu.au/services

* International ~~Geo~~ Generic Sample Number



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

IGSN now works with DataCite

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.

Acknowledgments

Special thanks for contributions, ideas and inspiration from

FS-SC: Anton Barty, Tom White, Christoph Rosemann, Alexandra Tolstikova, Tim Schoof, Vijay Kartik, Mads Jakobsen, Sam Flewett, Gudrun Lotse, Diana Rueda, Marc-Olivier Andrez, Michael Größler, Lisa Amelung, Nicola Baark, Igor Khokhriakov, Olga Merkulova, Mikhail Karnevinsky, Abdullah Malik, Silvan Schön

Daphne4NFDI: Anton Barty, Bridget Murphy, Lisa Amelung, Christian Gutt, Astrid Schneidewind, Wiebke Lohstroh, Sebastien Busch, Frank Schreiber, Tobias Unruh, Jan-Dierk Grunwaldt and at least 50 others

Other DESY: Volker Gülzow, Martin Gasthuber, Patrick Fuhrmann, Paul Millar, Sophie Servan, Mikhail Karnevinsky, Kars Ohrenberg (Central IT), Thorsten Kracht, Linus Pithan (FS-EC), Harald Reichert, Kai Bagschik, Stephan Klumpp (PETRA-IV), and many others

Others: Valerio Mariani (SLAC)

Final thoughts

Final thoughts

- Keeping all detector output is feasible but can get very expensive
  - Who pays?
  - Who are we keeping it for?
  - Is it worth the cost?

Final thoughts

- Keeping all detector output is feasible but can get very expensive
  - Who pays?
  - Who are we keeping it for?
  - Is it worth the cost?

Final thoughts

- Keeping all detector output is feasible but can get very expensive
  - Who pays?
  - Who are we keeping it for?
  - Is it worth the cost?

- What is the ‘raw data’ we aim to keep?
  - Photons not ADU, 1D powder …
  - For what purpose is the data being kept?

Final thoughts

- Keeping all detector output is feasible but can get very expensive
  - Who pays?
  - Who are we keeping it for?
  - Is it worth the cost?
- What is the ‘raw data’ we aim to keep?
  - Photons not ADU, 1D powder …
  - For what purpose is the data being kept?
- Clarity on when to discard data is needed

Final thoughts

- Keeping all detector output is feasible but can get very expensive
  - Who pays?
  - Who are we keeping it for?
  - Is it worth the cost?
- What is the ‘raw data’ we aim to keep?
  - Photons not ADU, 1D powder …
  - For what purpose is the data being kept?
- Clarity on when to discard data is needed
- Temptation is to invest in new outcomes rather than old data
  - Money is limited and may come from the same (limited) budget

Final thoughts

- Keeping all detector output is feasible but can get very expensive
  - Who pays?
  - Who are we keeping it for?
  - Is it worth the cost?
- What is the ‘raw data’ we aim to keep?
  - Photons not ADU, 1D powder …
  - For what purpose is the data being kept?
- Clarity on when to discard data is needed
- Temptation is to invest in new outcomes rather than old data
  - Money is limited and may come from the same (limited) budget
- Persistent availability of data requires persistent funding
  - What happens at the end of a 5 year project?

End