ICFA Data Lifecycle Panel: Next steps

ICFA Data Lifecycle Panel meeting - October 8, 2024

Kati Lassila-Perini
Helsinki Institute of Physics - Finland

Recommendations

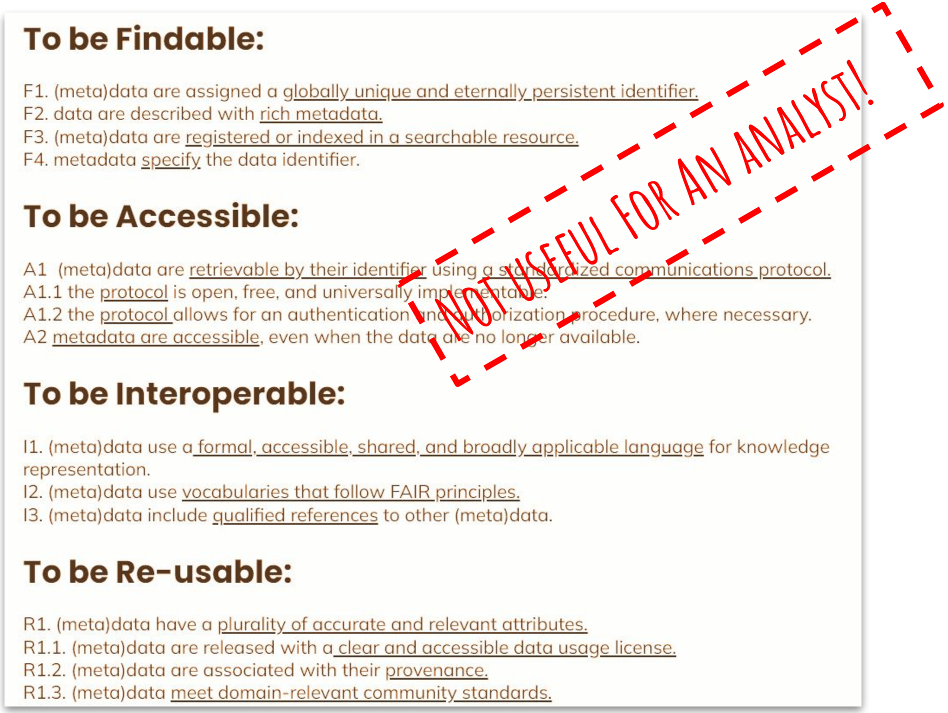
- It is in our mandate to develop recommendations for best practices to facilitate data preservation and data reuse
  - We want them to be concrete, specific and relevant to our domain.
  - We want them to be understandable to all stakeholders: from students and analysts to the experiment management.
- Therefore
  - Reach out to *enablers* in our domain to hear their view:
    - DPHEP – last week
    - HSF training organizers, trainers of AP skills in the experiments – CHEP
  - Follow the ongoing work for KPIs (Key Performance Indicators) for Open Science at CERN (and elsewhere?)
    - Recommendations and KPIs should match.

Learn from the past
Learn from the present

- The DP/AP community has a lots of expertise, but in small experiment-specific teams
- Pass the knowledge:
  - How to avoid the obstacles that the past experiments are facing?
  - How to learn from the best practices in other labs / experiments?
- Involve the community in writing
  - Inspired by the bottom-up process to write the CERN Open Data policy.

What to avoid?

- Repeating FAIR principles is not very useful
- Provide concrete suggestions that can be followed



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.

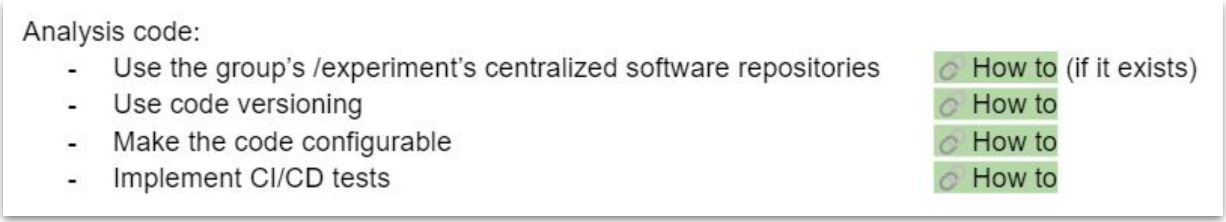
To be Re-usable:

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.

NOT USEFUL FOR AN ANALYST!

What’s new?

- How would this differ from the Open Data / Open Science policies and implementation plans?
  - Be concrete, practical and role specific- an example in the domain of analysis preservation for an analyst:

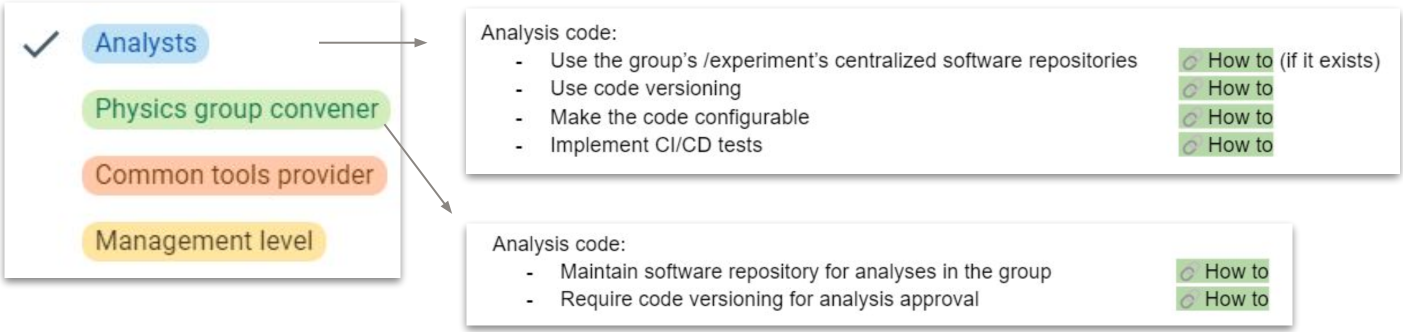


Analysis code:

- Use the group’s /experiment’s centralized software repositories — How to (if it exists)
- Use code versioning — How to
- Make the code configurable — How to
- Implement CI/CD tests — How to

What’s new?

- How to address different audiences with diverse tasks?
  - In an online document, define tabs (or similar) for different roles



Analysis code:

- Use the group’s /experiment’s centralized software repositories — How to (if it exists)
- Use code versioning — How to
- Make the code configurable — How to
- Implement CI/CD tests — How to

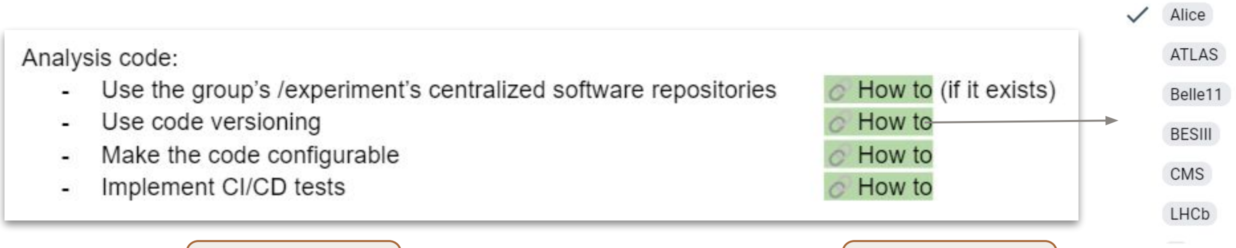
Analysis code:

- Maintain software repository for analyses in the group — How to
- Require code versioning for analysis approval — How to

  - Persistent choice over the document
  - These are quick examples just for illustration!

What’s new?

- Isn’t it challenging to make it generic?
  - As an online document, make it configurable for labs /experiments
  - Keep the concept generic, but optionally link to specific instructions



Analysis code:

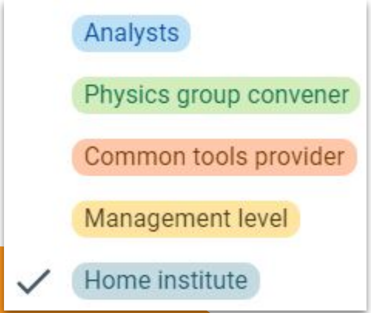
- Use the group’s /experiment’s centralized software repositories — How to (if it exists)
- Use code versioning — How to
- Make the code configurable — How to
- Implement CI/CD tests — How to

Generic

Specific

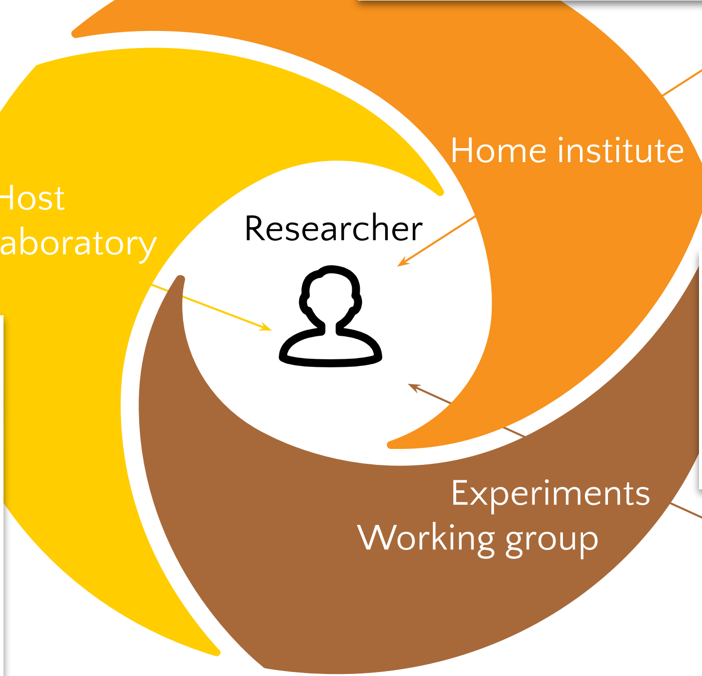
  - Experiments assessing if the instructions exist will be part of the process – expect this to take quite some effort…

Other audiences?
To be decided

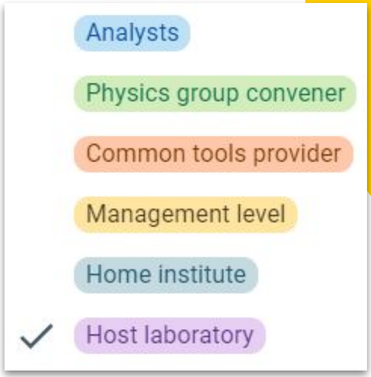


- Analysts
- Physics group convener
- Common tools provider
- Management level
- ✓ Home institute

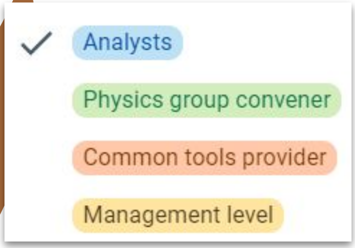
Task definition, share of time on them
Resources (salary, travel, personal tools)
Education, training
Research/data management guidelines



Infrastructure
Resources
Open science guidelines

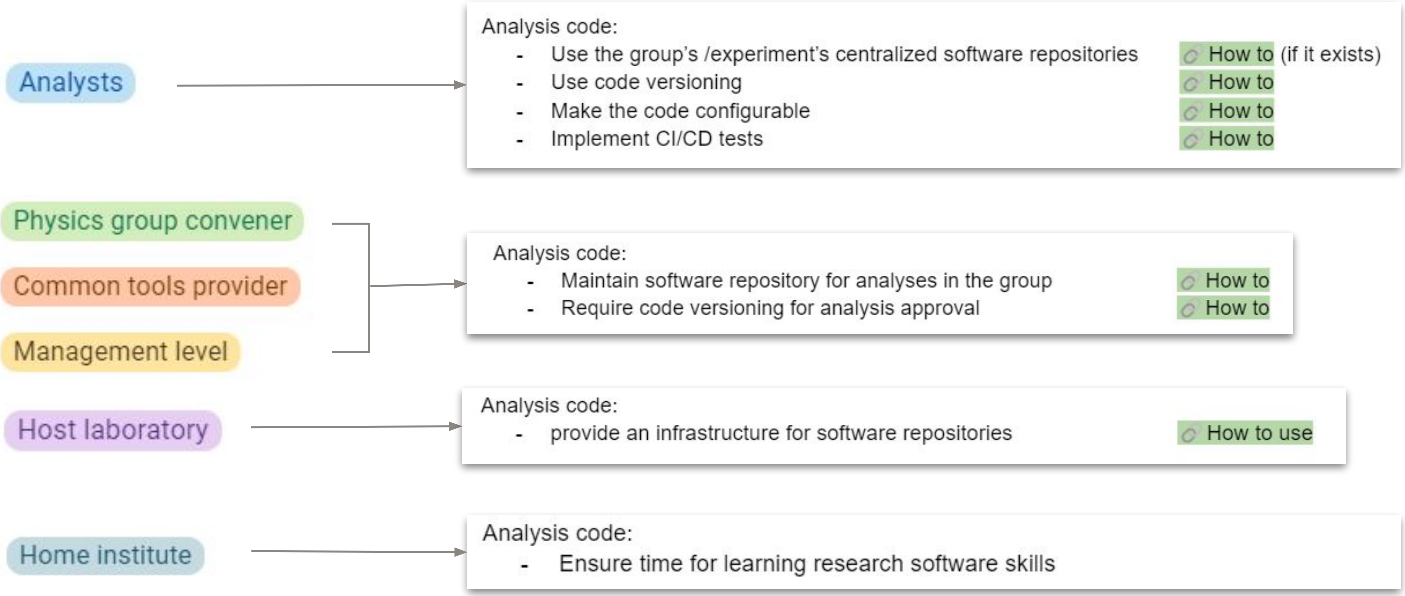


- Analysts
- Physics group convener
- Common tools provider
- Management level
- Home institute
- ✓ Host laboratory



- ✓ Analysts
- Physics group convener
- Common tools provider
- Management level

Responsibilities
Resources (computing, software)
Practices (“how do we work”)
Policies



Analysts

Analysis code:

- Use the group's /experiment's centralized software repositories — How to (if it exists)
- Use code versioning — How to
- Make the code configurable — How to
- Implement CI/CD tests — How to

Physics group convener

Common tools provider

Management level

Analysis code:

- Maintain software repository for analyses in the group — How to
- Require code versioning for analysis approval — How to

Host laboratory

Analysis code:

- provide an infrastructure for software repositories — How to use

Home institute

Analysis code:

- Ensure time for learning research software skills

Proposal for a program of work:

Nov 24

- Analyse the input from the surveys
  - Define the topics to be covered
    - Match to the mandate of the DLC panel

Nov 24
Mar 25

- First draft by a working group with volunteers from
  - this panel
  - people involved in DP/AP/docs in the past and present experiments
  - people involved in the OS/FAIR initiatives (HSF / EVERSE / FAIROS-HEP)
  - people involved in the OS KPI (Key Performance Indicator) definition

1st half 25

- Organize a workshop / a retreat to work on details
  - FAIROS-HEP can help in funding.
- Circulate for a wider feedback.

Input from the DPHEP Workshop

2 –3 October, CERN and hybrid

So far: Facilitating factors

- Institutional support
- Technical solutions:
  - Reduced data formats
  - Open-source software
  - Software containers
  - Decoupling from specific environments
  - Data migration to more accessible storage
- Policy and collaboration:
  - LHC Open Data Policy
  - Best effort agreements with IT
  - Collaboration with CERN IT open data team

- Ongoing research needs:
  - Continued data analyses beyond main funding period
  - Regular publications from datasets
- Documentation and accessibility:
  - Full data provenance information
  - Dedicated tools for open data access
  - Clear instructions and easier processing

- Community factors:
  - Increased appreciation of preservation efforts
  - Positive feedback from the community
  - Small group of committed individuals
- Standardization:
  - Use of common packages and standard techniques
  - Central storage of experiment-specific software and documentation

So far: Obstacles...

- Resource constraints:
  - Limited funding
  - Lack of dedicated person-power
  - Time constraints, especially at the end of analysis processes
- Policy and understanding issues:
  - Restrictive data access policies
  - Misunderstanding of data preservation vs. open data
  - Lack of awareness about preservation policies within experiments
- Technical challenges:
  - Proliferation of analysis frameworks
  - Complexity of analysis preservation
  - Software maintenance over long periods
  - Adapting to changes in computing infrastructure and OS support
- Documentation and standardization:
  - Sparse or fragmented documentation
  - Non-standardized recording of analysis information
  - Use of non-open formats for documentation

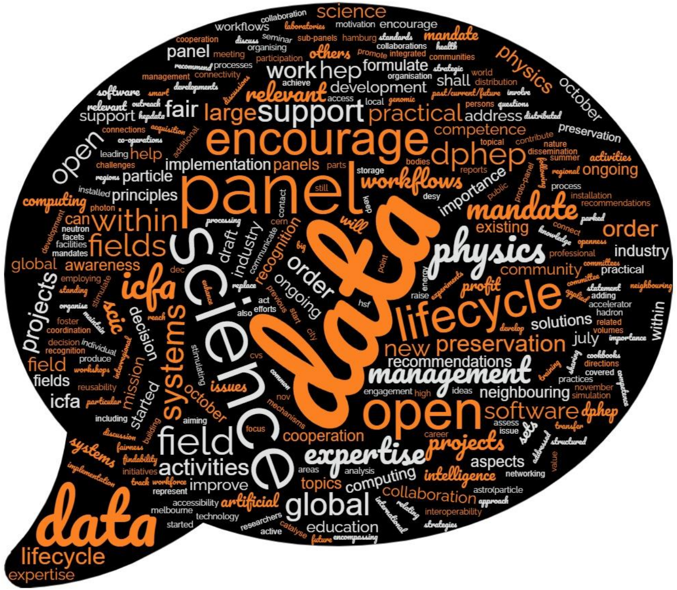
So far: ...Obstacles...

- Continuity and knowledge transfer:
  - Loss of human knowledge over time
  - Lack of continuity when individuals move on
  - Information stored in personal directories that may be deleted
- Commitment and coordination:
  - Reliance on individual initiatives
  - Difficulty in uniting around a common vision
  - Weak language in policies leading to ambiguity
- Cultural factors:
  - Perception of low return on investment for preservation efforts
  - Pushback from various parties within experiments
  - Difficulty in openly discussing challenges across experiments

Outlook

Working towards recommendations for best practices data preservation and FAIR principles

This will require quite some effort, looking forward to getting started!



Thank you!

Questions?

And thanks to SlidesCarnival for this free presentation template

ICFA statement on the Data Lifecycle Panel

Mandate of the Data Lifecycle Panel