HSF Training and ICFA Data Lifecycle Panel

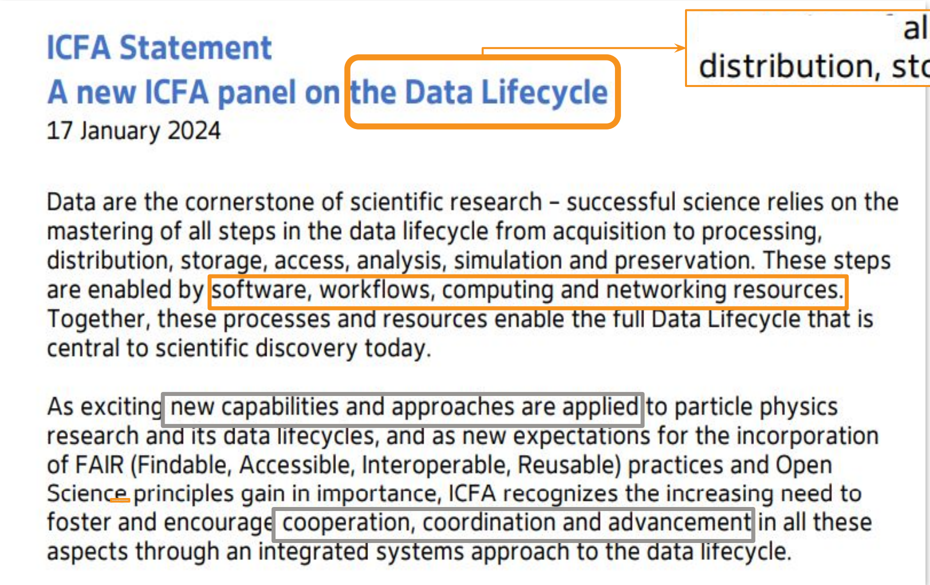
Pre-CHEP HSF training workshop - October 20, 2024

Kati Lassila-Perini
Helsinki Institute of Physics - Finland

0

ICFA?

International Committee for Future Accelerators: https://icfa.hep.net/



ICFA Statement
A new ICFA panel on the Data Lifecycle

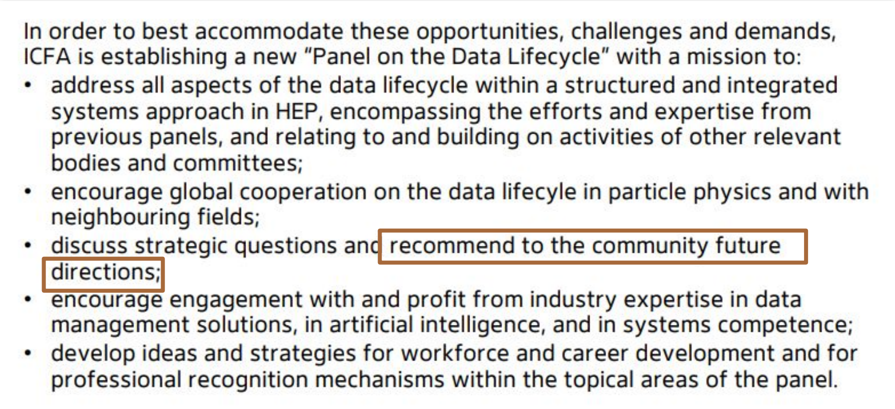
17 January 2024

Data are the cornerstone of scientific research – successful science relies on the mastering of all steps in the data lifecycle from acquisition to processing, distribution, storage, access, analysis, simulation and preservation. These steps are enabled by software, workflows, computing and networking resources. Together, these processes and resources enable the full Data Lifecycle that is central to scientific discovery today.

As exciting new capabilities and approaches are applied to particle physics research and its data lifecycles, and as new expectations for the incorporation of FAIR (Findable, Accessible, Interoperable, Reusable) practices and Open Science principles gain in importance, ICFA recognizes the increasing need to foster and encourage cooperation, coordination and advancement in all these aspects through an integrated systems approach to the data lifecycle.

all steps in the data lifecycle from acquisition to processing, distribution, storage, access, analysis, simulation and preservation.

Software, workflows, computing, networking

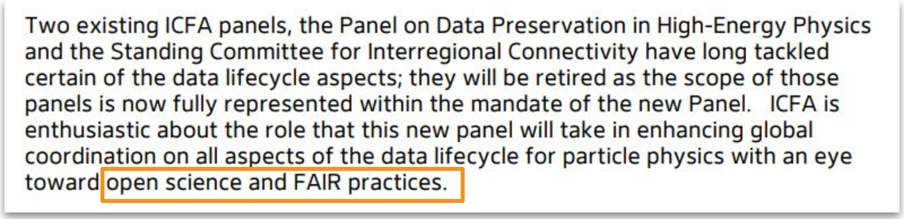


In order to best accommodate these opportunities, challenges and demands, ICFA is establishing a new "Panel on the Data Lifecycle" with a mission to:

- address all aspects of the data lifecycle within a structured and integrated systems approach in HEP, encompassing the efforts and expertise from previous panels, and relating to and building on activities of other relevant bodies and committees;
- encourage global cooperation on the data lifecyle in particle physics and with neighbouring fields;
- discuss strategic questions and recommend to the community future directions;
- encourage engagement with and profit from industry expertise in data management solutions, in artificial intelligence, and in systems competence;
- develop ideas and strategies for workforce and career development and for professional recognition mechanisms within the topical areas of the panel.

FAIR practices
Open Science

Cooperation,
coordination
advancement

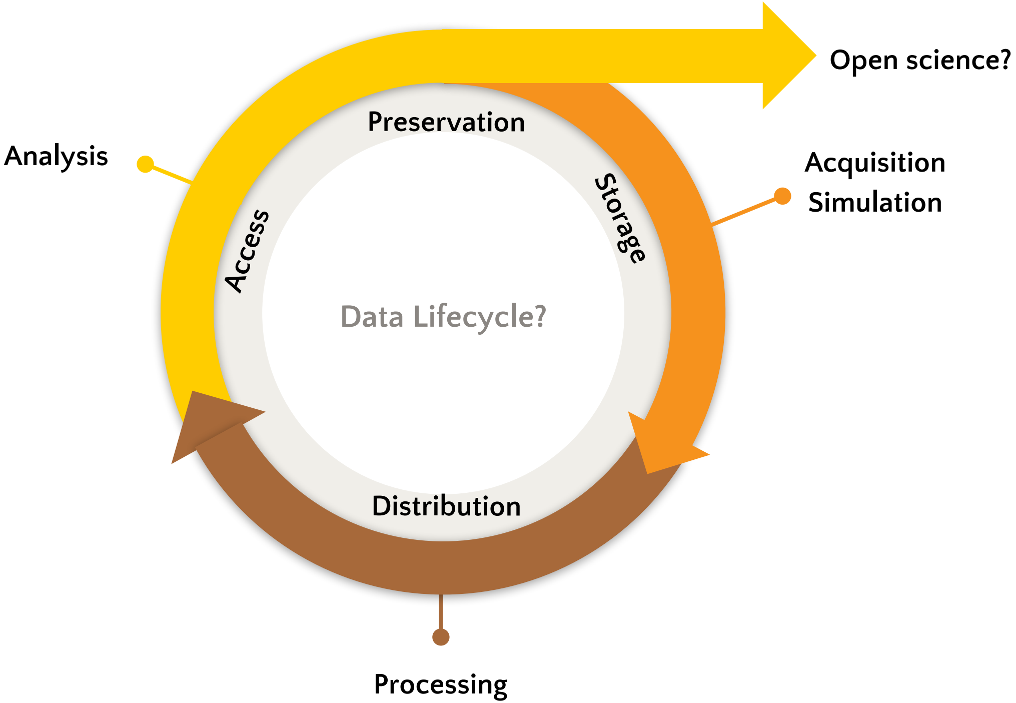


Two existing ICFA panels, the Panel on Data Preservation in High-Energy Physics and the Standing Committee for Interregional Connectivity have long tackled certain of the data lifecycle aspects; they will be retired as the scope of those panels is now fully represented within the mandate of the new Panel. ICFA is enthusiastic about the role that this new panel will take in enhancing global coordination on all aspects of the data lifecycle for particle physics with an eye toward open science and FAIR practices.

1

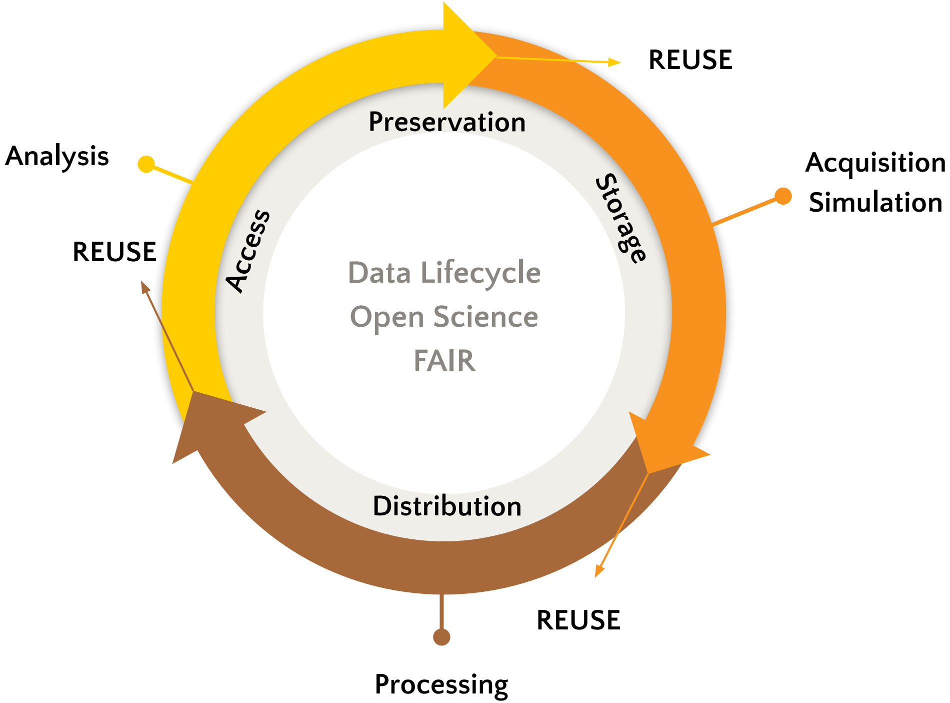
Data Lifecycle?

Does Open Science come out of the last step of the data lifecycle?



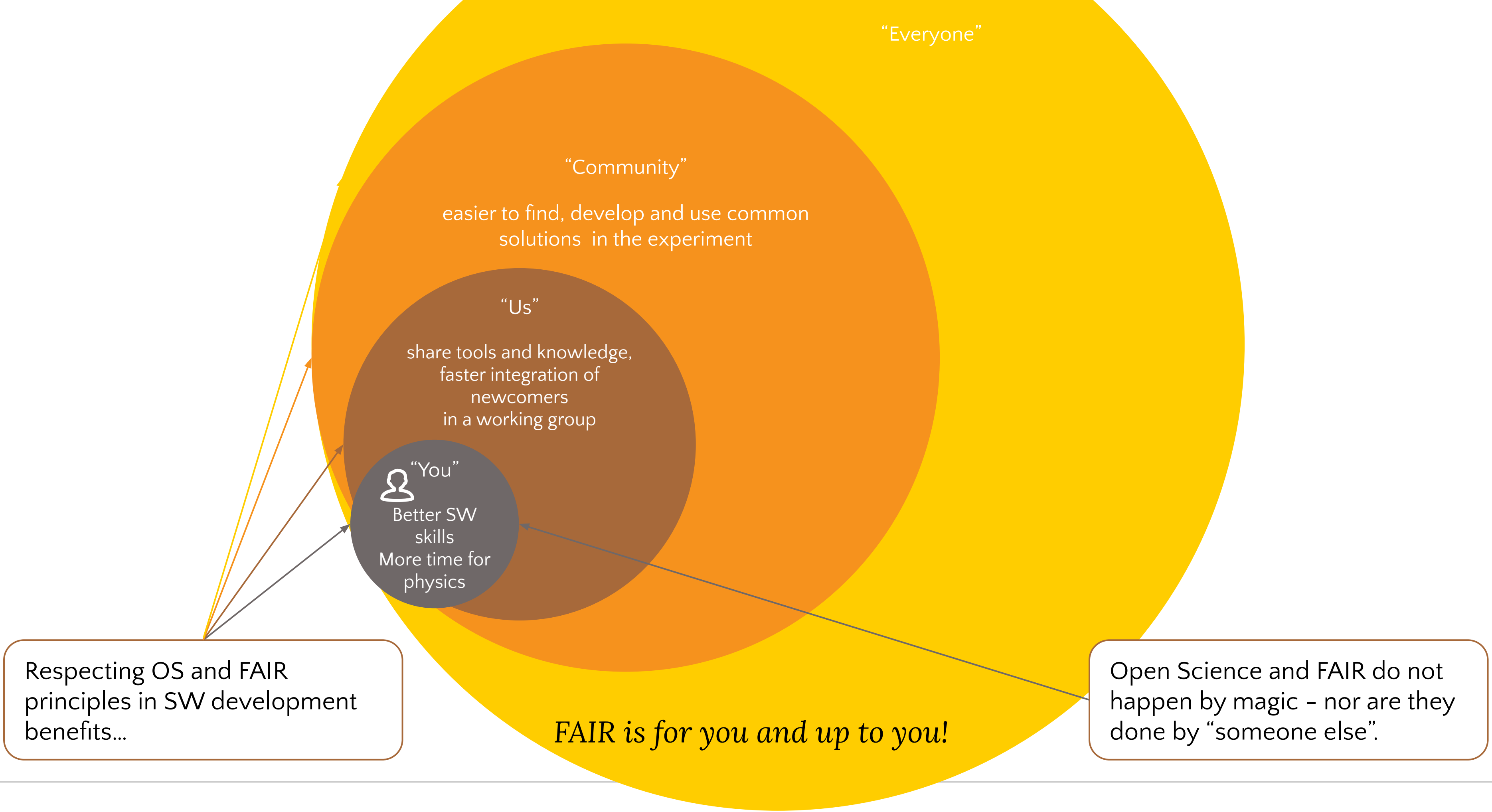
Preserved ripple marks from 2 Gy ago
Noitatunturi – Finland

No, the key is REUSE.



Survey, DPHEP workshop:

Lack of preserved analysis code was mentioned as the biggest challenge for the usability of preserved data.



"Everyone"

"Community"

easier to find, develop and use common solutions in the experiment

"Us"

share tools and knowledge, faster integration of newcomers in a working group

"You"

Better SW skills More time for physics

Respecting OS and FAIR principles in SW development benefits...

Open Science and FAIR do not happen by magic – nor are they done by "someone else".

FAIR is for you and up to you!

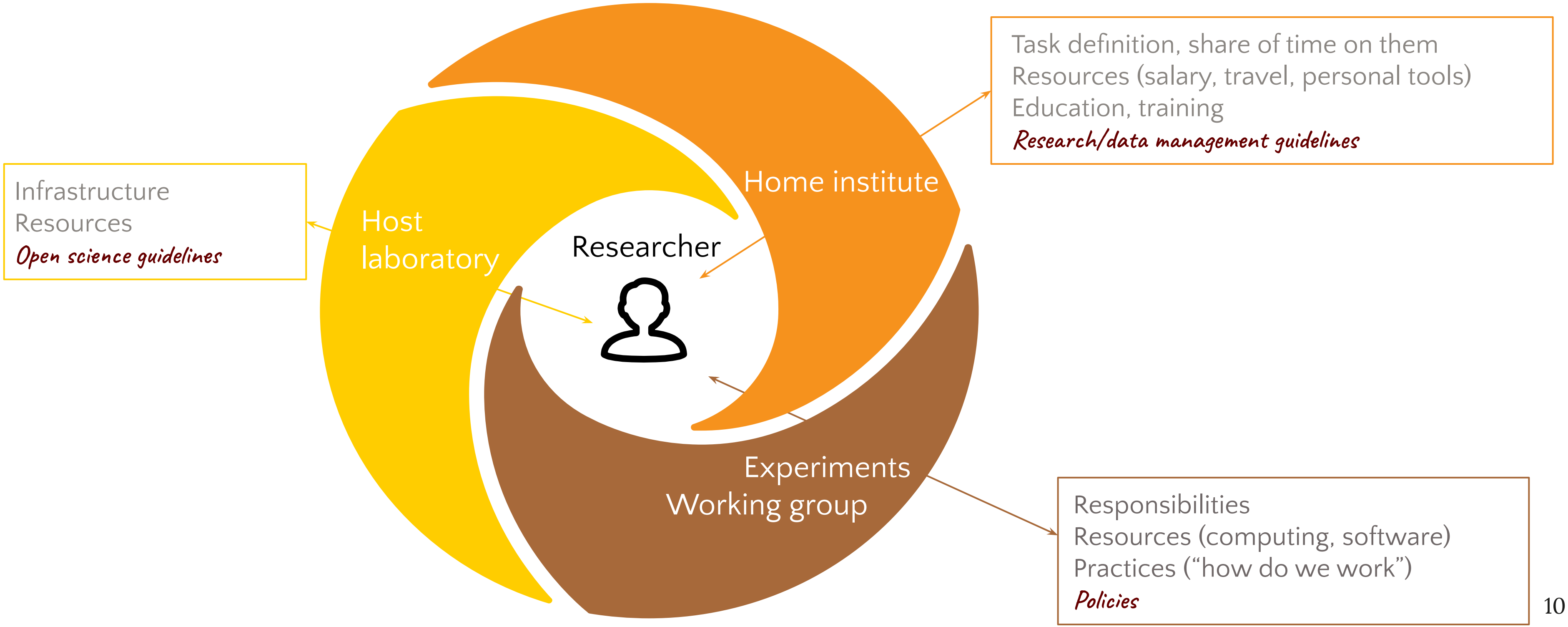
2

Stakeholders

Individuals researchers, within the collaborations, carry out the work.

The surrounding stakeholders may either empower or restrict the researchers' ability to adopt best practices for the full “data lifecycle".

Stakeholders

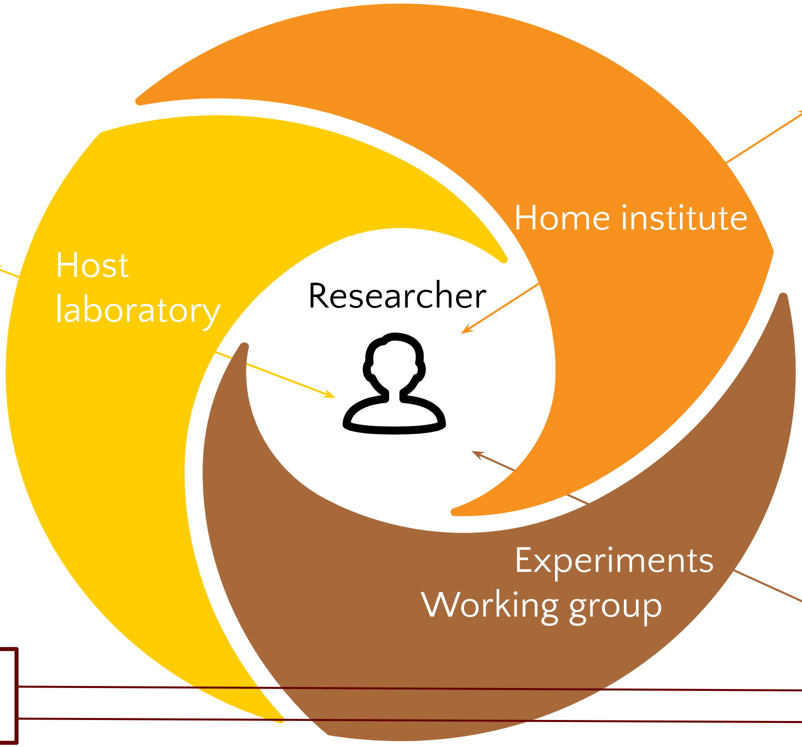


Infrastructure
Resources
Open science guidelines

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

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

Stakeholders - words



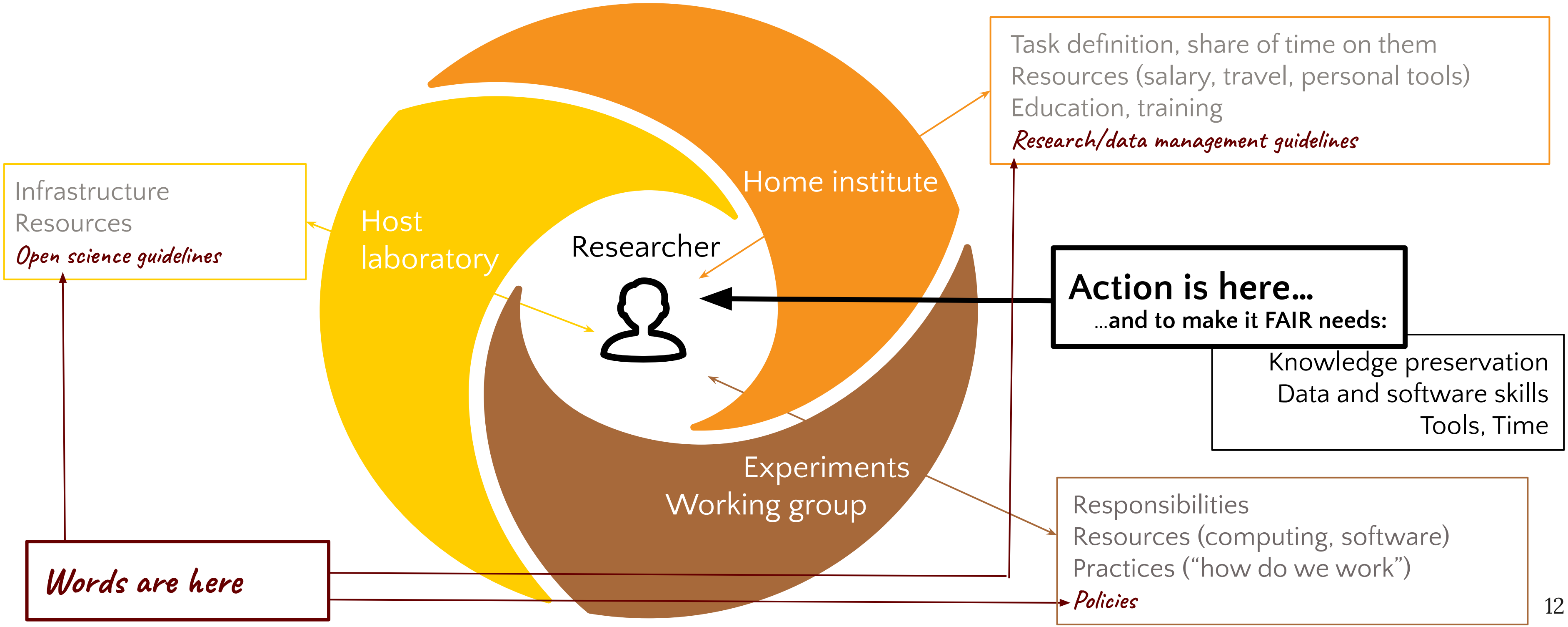
Infrastructure
Resources
Open science guidelines

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

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

Words are here

Stakeholders - words vs action



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

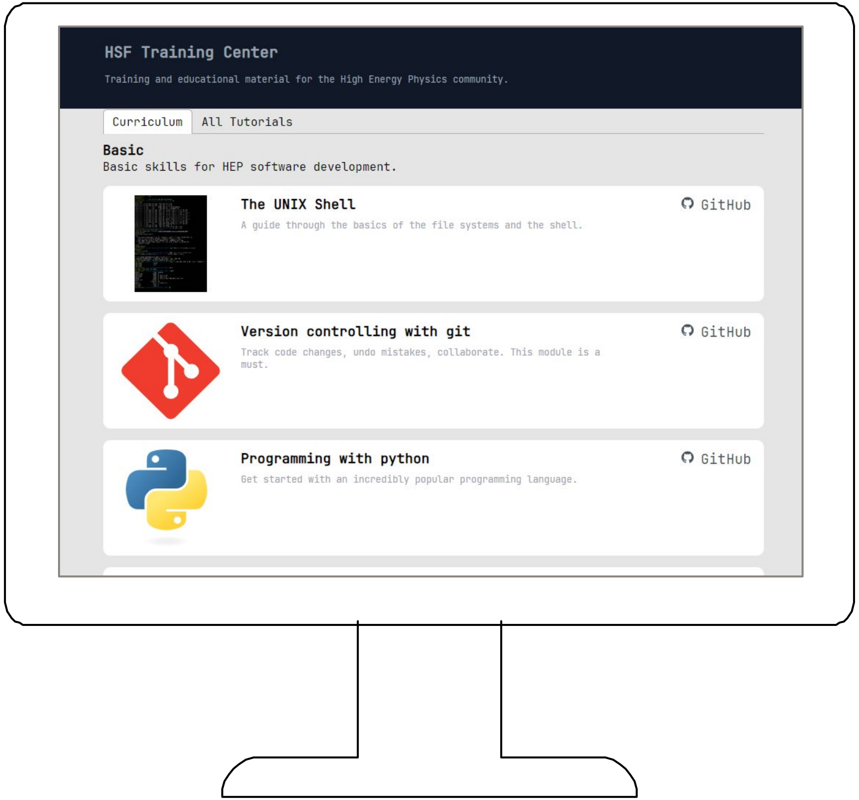
Infrastructure
Resources
Open science guidelines

Action is here…
…and to make it FAIR needs:

Knowledge preservation
Data and software skills
Tools, Time

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

Words are here



HSF Training Center

Training and educational material for the High Energy Physics community.

Basic

Basic skills for HEP software development.

The UNIX Shell

A guide through the basics of the file systems and the shell.

Version controlling with git

Track code changes, undo mistakes, collaborate. This module is a must.

Programming with python

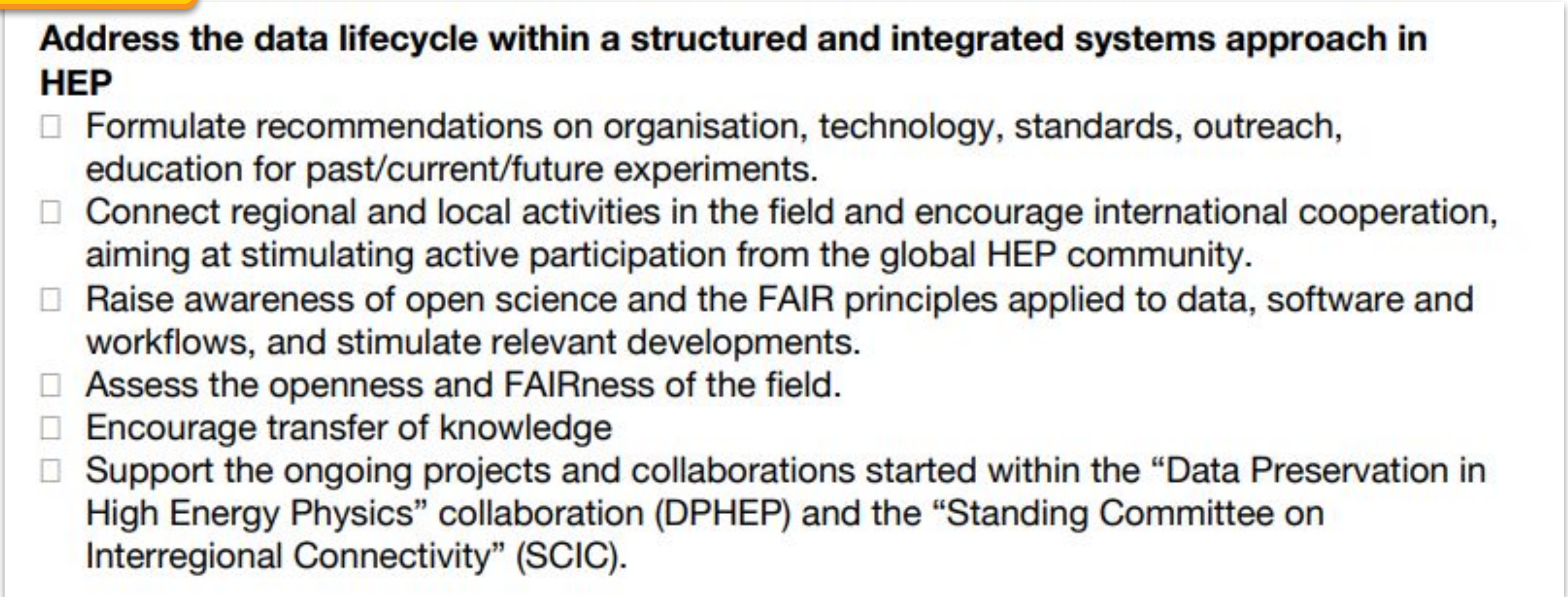
Get started with an incredibly popular programming language.

Data and software skills: training!

3

Data Lifecycle panel

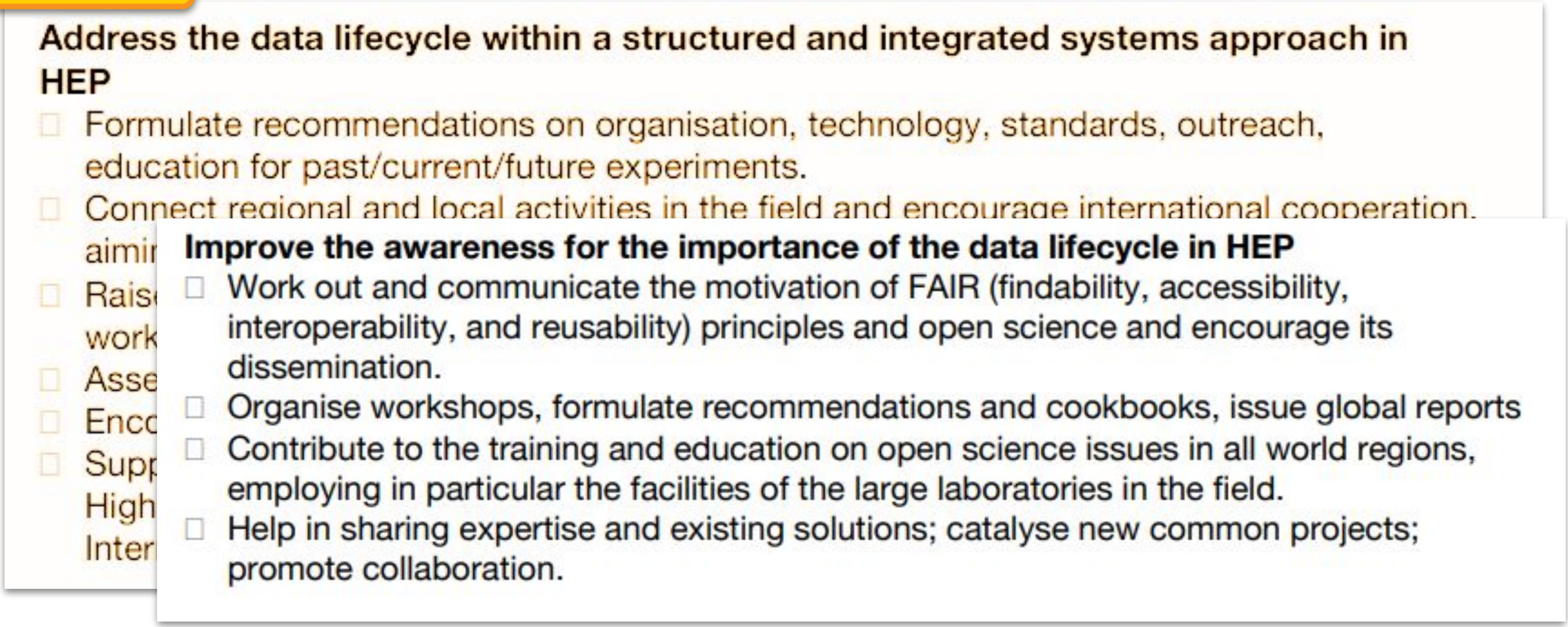
From the mandate



Address the data lifecycle within a structured and integrated systems approach in HEP

- □ Formulate recommendations on organisation, technology, standards, outreach, education for past/current/future experiments.
- □ Connect regional and local activities in the field and encourage international cooperation, aiming at stimulating active participation from the global HEP community.
- □ Raise awareness of open science and the FAIR principles applied to data, software and workflows, and stimulate relevant developments.
- □ Assess the openness and FAIRness of the field.
- □ Encourage transfer of knowledge
- □ Support the ongoing projects and collaborations started within the “Data Preservation in High Energy Physics” collaboration (DPHEP) and the “Standing Committee on Interregional Connectivity” (SCIC).

From the mandate



Address the data lifecycle within a structured and integrated systems approach in HEP

□ Formulate recommendations on organisation, technology, standards, outreach, education for past/current/future experiments.

□ Connect regional and local activities in the field and encourage international cooperation, aimi

□ Rais work

□ Asse

□ Enc

□ Sup High Inter

Improve the awareness for the importance of the data lifecycle in HEP

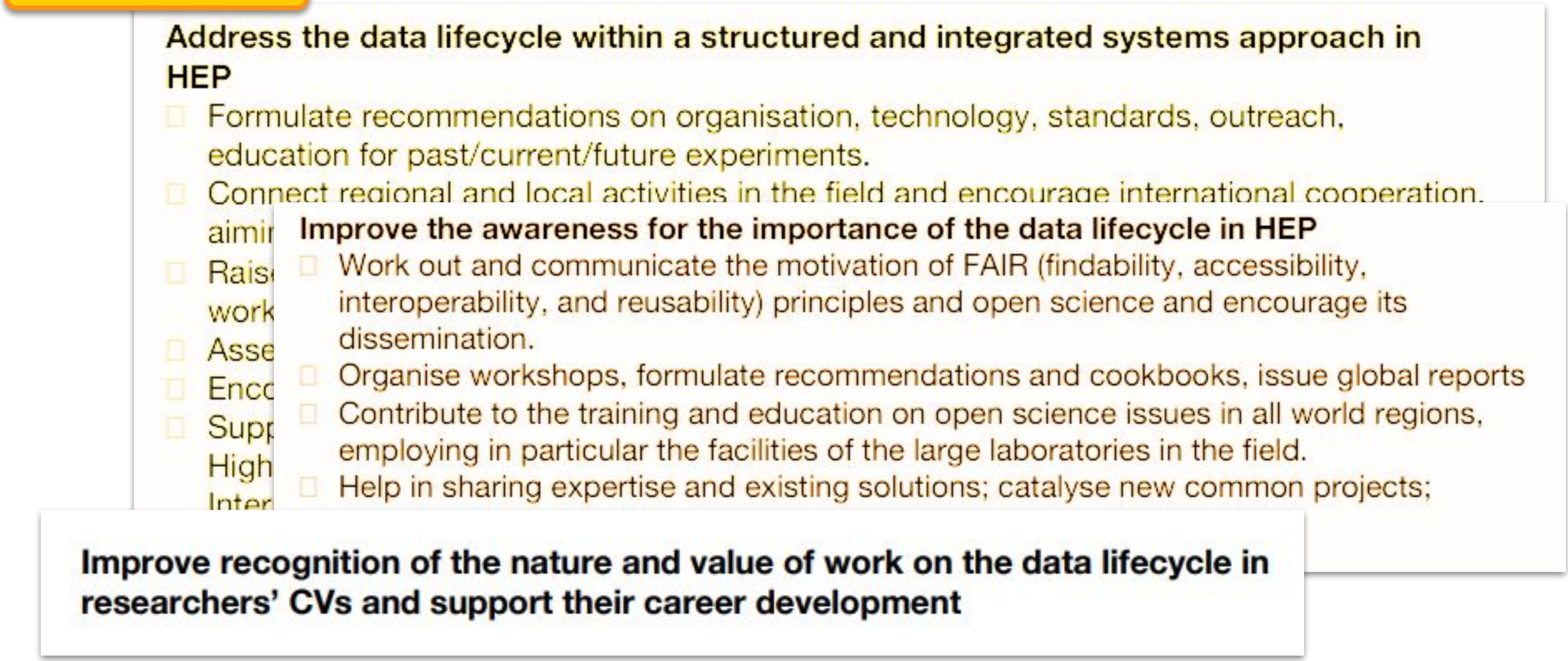
□ Work out and communicate the motivation of FAIR (findability, accessibility, interoperability, and reusability) principles and open science and encourage its dissemination.

□ Organise workshops, formulate recommendations and cookbooks, issue global reports

□ Contribute to the training and education on open science issues in all world regions, employing in particular the facilities of the large laboratories in the field.

□ Help in sharing expertise and existing solutions; catalyse new common projects; promote collaboration.

From the mandate



Address the data lifecycle within a structured and integrated systems approach in HEP

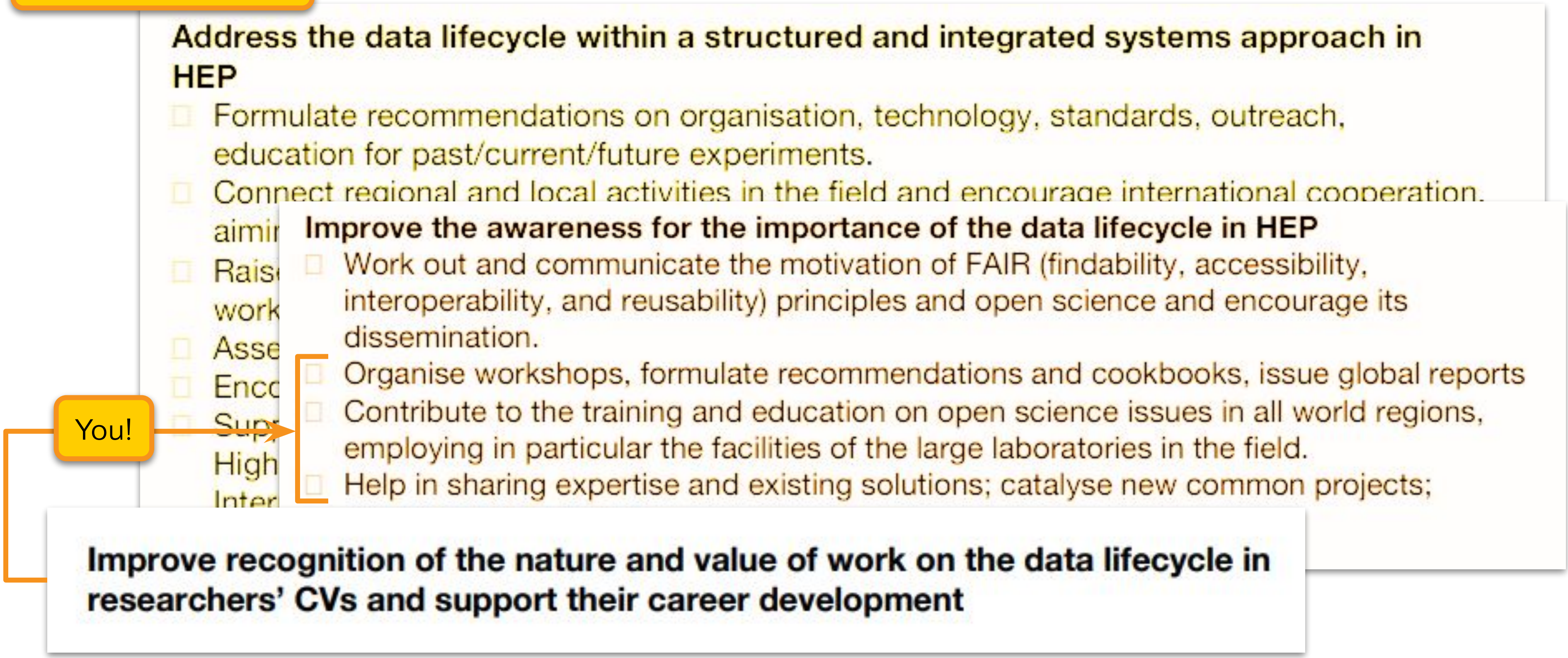
- □ Formulate recommendations on organisation, technology, standards, outreach, education for past/current/future experiments.
- □ Connect regional and local activities in the field and encourage international cooperation, aimi
- □ Rais work
- □ Asse
- □ Enc
- □ Supp High Inter

Improve the awareness for the importance of the data lifecycle in HEP

- □ Work out and communicate the motivation of FAIR (findability, accessibility, interoperability, and reusability) principles and open science and encourage its dissemination.
- □ Organise workshops, formulate recommendations and cookbooks, issue global reports
- □ Contribute to the training and education on open science issues in all world regions, employing in particular the facilities of the large laboratories in the field.
- □ Help in sharing expertise and existing solutions; catalyse new common projects;

Improve recognition of the nature and value of work on the data lifecycle in researchers' CVs and support their career development

From the mandate



Address the data lifecycle within a structured and integrated systems approach in HEP

□ Formulate recommendations on organisation, technology, standards, outreach, education for past/current/future experiments.

□ Connect regional and local activities in the field and encourage international cooperation, aimi

□ Raise work

□ Asse

□ Enco

□ Sup High Inter

Improve the awareness for the importance of the data lifecycle in HEP

□ Work out and communicate the motivation of FAIR (findability, accessibility, interoperability, and reusability) principles and open science and encourage its dissemination.

□ Organise workshops, formulate recommendations and cookbooks, issue global reports

□ Contribute to the training and education on open science issues in all world regions, employing in particular the facilities of the large laboratories in the field.

□ Help in sharing expertise and existing solutions; catalyse new common projects;

Improve recognition of the nature and value of work on the data lifecycle in researchers' CVs and support their career development

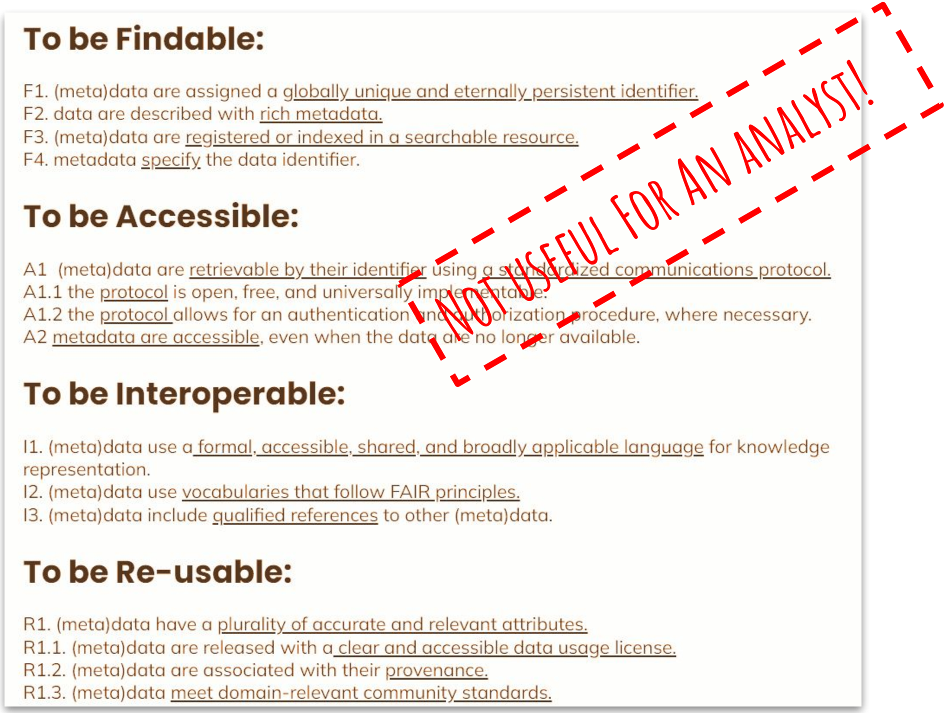
You!

Recommendations

- Panel's mandate: to formulate recommendations for best practices to achieve Open Science and follow FAIR principles
  - 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 and home institutes.
- Therefore
  - Reach out to *enablers* in our domain to hear their view:
    - DPHEP workshop– 2-3 Obtober
    - HSF training workshop – right now!
  - Follow the ongoing work for
    - KPIs (Key Performance Indicators) for Open Science at CERN (and else?)
      - Recommendations and KPIs should match
    - Everse et al. (see Stefan Roiser's slides)
    - process of defining a strategy for Open Data Management in France

What to avoid?

- Repeating FAIR principles is not very useful.
- Provide concrete suggestions at the level in which the reader can take action.



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:

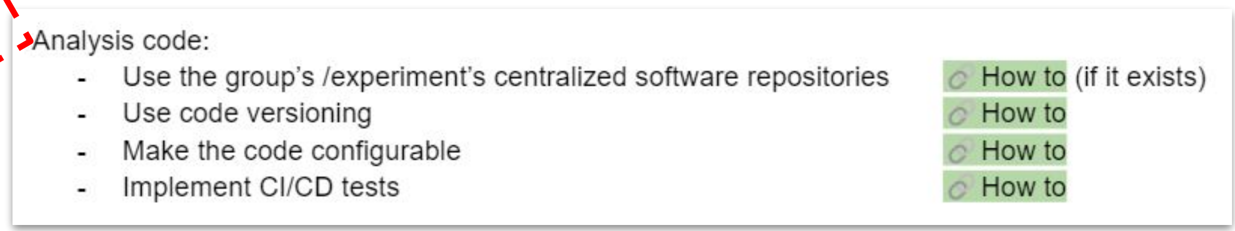
R1. (meta)data have a plurality of accurate and relevant attributes.
R1.1. (meta)data are released with a clear and accessible data usage license.
R1.2. (meta)data are associated with their provenance.
R1.3. (meta)data meet domain-relevant community standards.

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:

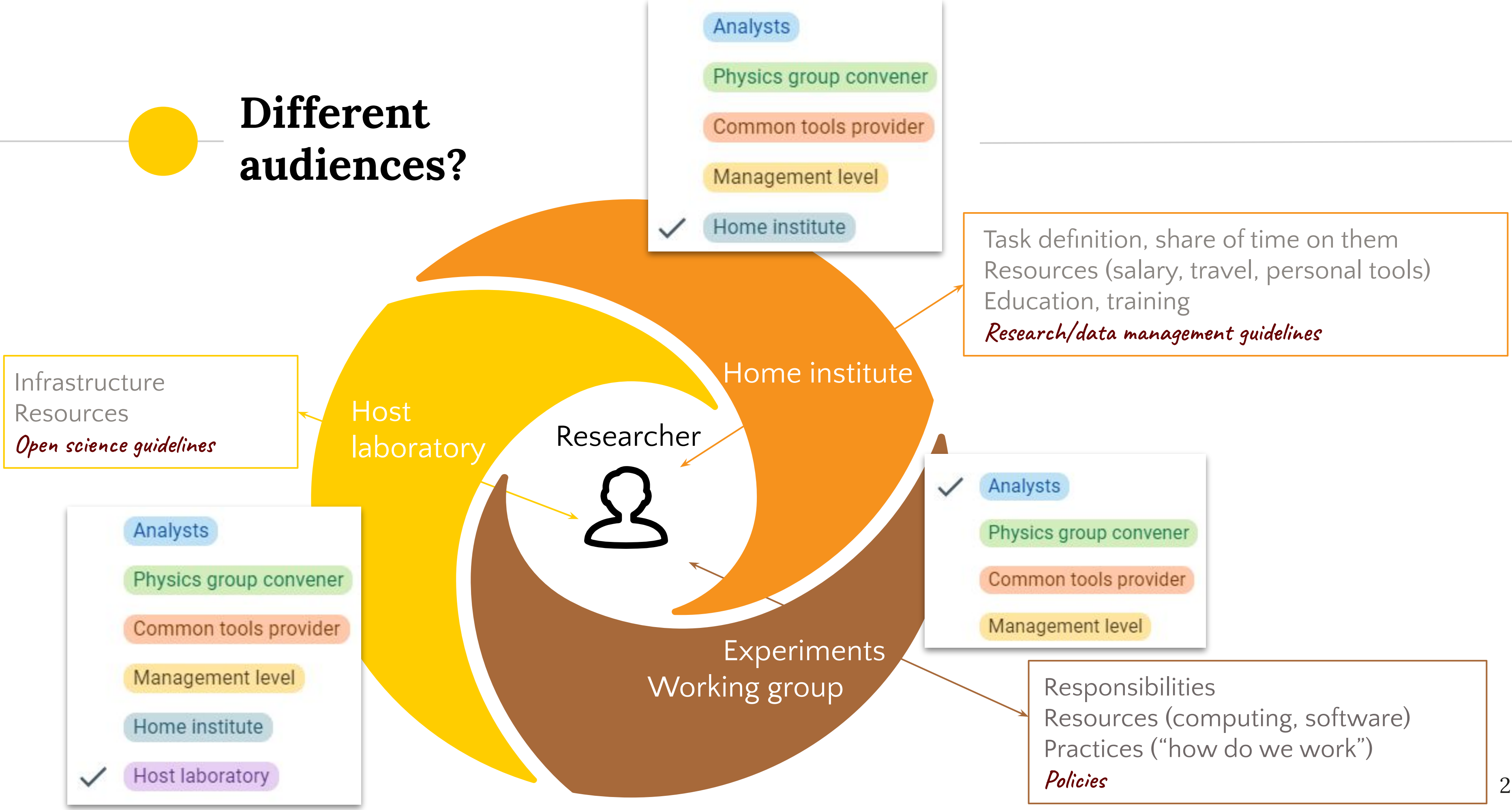
JUST FOR ILLUSTRATION!



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

Different audiences?



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

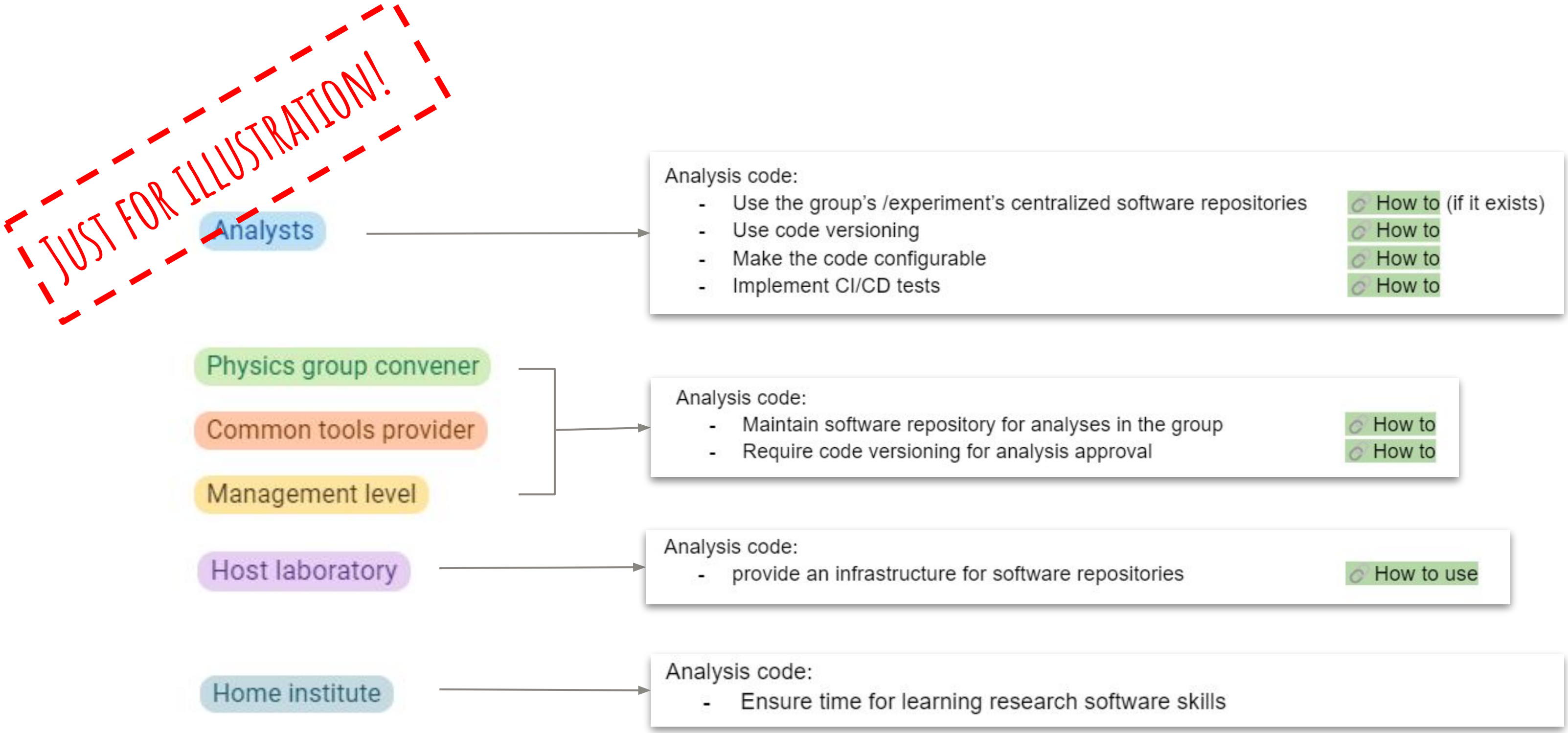
✓ Analysts

Physics group convener

Common tools provider

Management level

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



JUST FOR ILLUSTRATION!

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

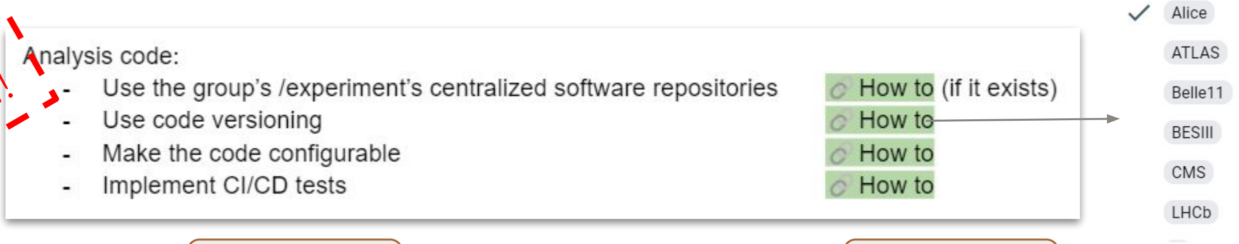
Analysis code:

- Ensure time for learning research software skills

What’s new?

- Isn’t it challenging to make it generic?
  - Online document: configurable for labs / experiments / audiences
  - Concept: generic – Instructions (link): specific

JUST FOR ILLUSTRATION!



Analysis code:

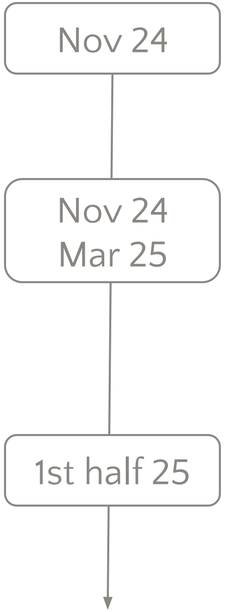
- Use the group’s /experiment’s centralized software repositories — How to (if it exists)
- Use code versioning — How to
- Make the code configurable — How to
- Implement CI/CD tests — How to

Generic

Specific

  - Agree on the generic concepts.
  - Experiments assessing if the instructions exist will be part of the process – expect this to take quite some effort…
  - Work together with the ongoing initiatives!

Proposal for a program of work:



Nov 24

Nov 24
Mar 25

1st half 25

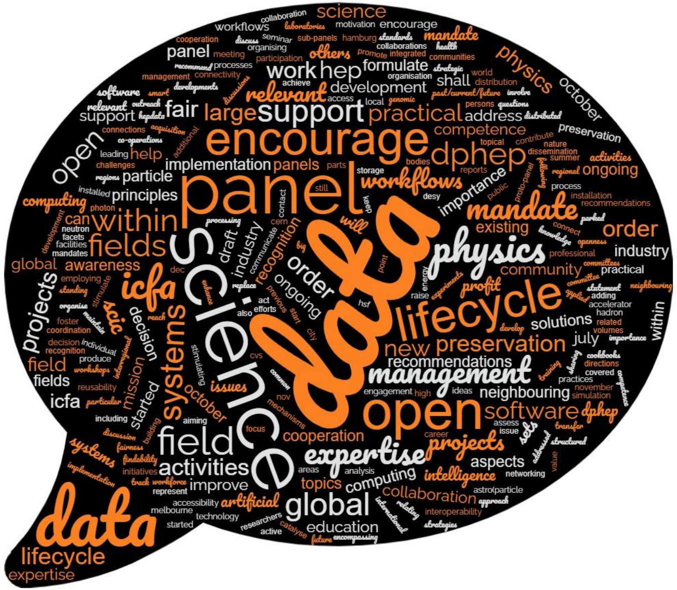
- Analyse the input from the surveys
  - Define the topics and concepts to be covered
- First draft by a working group with volunteers from
  - this audience
  - people involved in other OS/FAIR initiatives (EVERSE / FAIROS-HEP)
  - people involved in DP/AP/docs in the past and present experiments
  - people involved in the OS KPI (Key Performance Indicator) definition
- Organize a workshop / a retreat to work on details
- Circulate for a wider feedback.

Outlook

Your input counts!

Bottom-up approach, find out the pain points.

Goal: actionable recommendations that are concrete, specific and relevant to our domain.



Thank you!

Questions?

And thanks to SlidesCarnival for this free presentation template

Survey time!

ICFA statement on the Data Lifecycle Panel

Mandate of the Data Lifecycle Panel