ICFA Data Lifecycle Panel: Medium-term goals & Plan of action Summary from the May meeting

ICFA Data Lifecycle panel meeting - June 18, 2024

Kati Lassila-Perini
Helsinki Institute of Physics - Finland

Input so far

Suggested goals and actions grouped by themes

Survey questions

Goals:

“Suggest medium-terms goals for the panel, i.e. what would you like to see achieved in 1-2 years.”

Actions:

“Suggest actions through which we could reach those goals.”

Understanding the problems

Goals

Improve our common understanding of the mandate

Document the current approaches adopted with respect to data lifecycle by the different communities, starting from HEP experiments.

Actions

List the problems

Invite expert talks

Prioritize

Install sub-panels

Raise awareness of open science and the FAIR principles

Goals

Make sure that the importance of the Data Lifecycle is highlighted in the EU strategy for particle physics

Disseminate the importance of computing to all community members.

Facilitating more awareness about open data e.g. CMS & ATLAS

Practical guide for different stakeholders (researchers, group leaders, experiment management) on actions to achieve FAIRness of research software and workflows

Actions

representation in the EPPS process

Training

Goals

Raise the awareness for training and the recognition of people who engage in training.

Be an advocate for common software solutions

Come up with a set of training materials which are needed in software and computing for accelerator physics.

A well-defined, HEP-specific training curriculum on research software best practices based on the existing training resources (eventually in agreement with universities: ECTS credits)

Actions

Use synergies among the activities that have started or are running.

ICFA could be body to endorse a set of training materials useful for the community or even suggest a curriculum which can be proposed to schools and universities.

Tools

Goals

Consolidate current status of various computing resources, tools (e.g. file systems) and limitations (mostly from user perspective)

There is a surge of AI/ML tools but without much awareness on what are actually beneficial vs computing resources heavy.

Provide information and pointers to tools that make it easier to achieve FAIR data management for current and future experiments

Find commonalities in the current approaches adopted with respect to data lifecycle by the different communities and define a reference "implementation" (a set of guidelines that is known to work)

Understand what would be a reference implementation wrt the tools

Actions

A meta-repository of data sources (based on what exists, without duplicating)

Networking

Goals

(1) Beginning to deploy/integrate SENSE and related advanced network+Site services with the mainstream LHC data management tools. M1,M3

(2) Build a paragon network + data management system to show current capabilities (an order of magnitude beyond DC24 for example) - both scale and already existing functionalities and tools. This, and projections of technologies for 2026 and beyond are essential for setting real requirements for the HL LHC era; this will also enable more effective production and analysis workflows and their management. M1 M3

Actions

Engage with the GNA-G and its working groups.

Teach HEP about current server, network and interface technologies, and their projected evolution over the next 1-5 years

Learn from the SENSE network + site management teams

Learn about programmable network capabilities and methods also in the Global P4 Lab

Oversee and enable stronger ties and begin closer joint work between the above efforts and the LHC data management teams as well as the at-large HEP communities.

Recognition

Goals

Raise the awareness for training and the recognition of people who engage in training.

Disseminate the importance of computing to all community members.

Promote the work that goes in organizing training and learning, and value the time that is invested in following best practices

Enhance careers in HEP Data Lifecycle, including aspects of training, ethics, gender equity etc

Actions

Lobby for rewards for those who work on this topic

Community workshop (possible attached to one CHEP/ICHEP/EPS conference): issue a panel report with proposed measures.

Picked up for your

Cms Secretariat

Reply | Reply all | Forward | ...

To: **cms-members (All CMS Members)**

Mon 17/06/2024 14:57

Dear colleagues,

As part of the ECFA-ECR [1] community, we are conducting a study on the availability and quality of training programs in machine learning and software for young researchers in experimental and applied physics.

To gain a better understanding of the current landscape and to provide valuable feedback to the organizers of these training programs, we invite young scientists to participate in our survey:

https://docs.google.com/forms/d/e/1FAIpQLSeJZcknKj0DHrCEZ4jbq6j7AecvjCEGxnyL3YIQbJ-kslixTQ/viewform

More details can be found in the survey description.

Thank you very much for your support. Your input is very important and can truly make a difference in shaping the future of our training programs.

Feel free to contact us at ecfa-ecr-sw-ml-instrumentation@cern.ch with any questions or feedback.

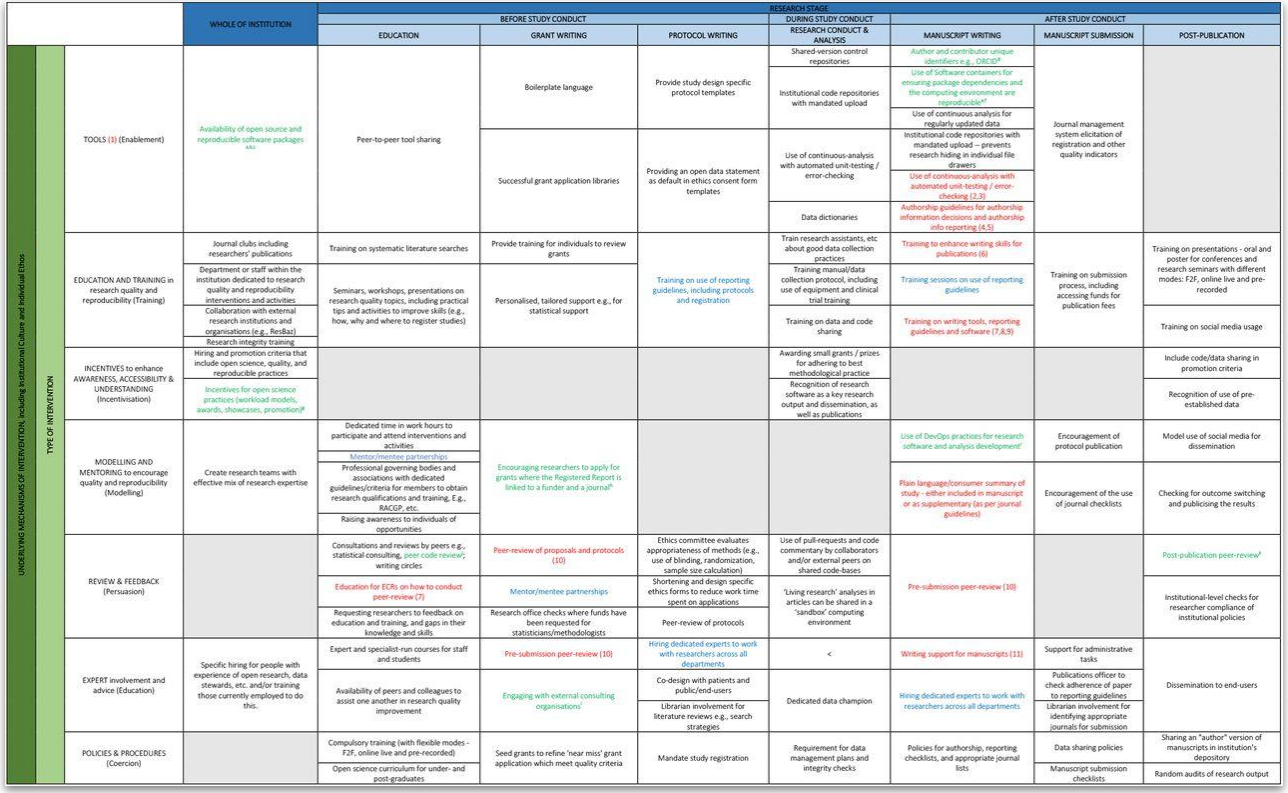
Best regards,
ECFA-ECR Software/Machine Learning applications for Future Colliders working group

---
[1] European Committee for Future Accelerators - Early-Career Researchers panel https://ecfa.web.cern.ch/ecfa-early-career-researchers-panel

ECFA-ECR survey on training programs

On increasing reproducibility

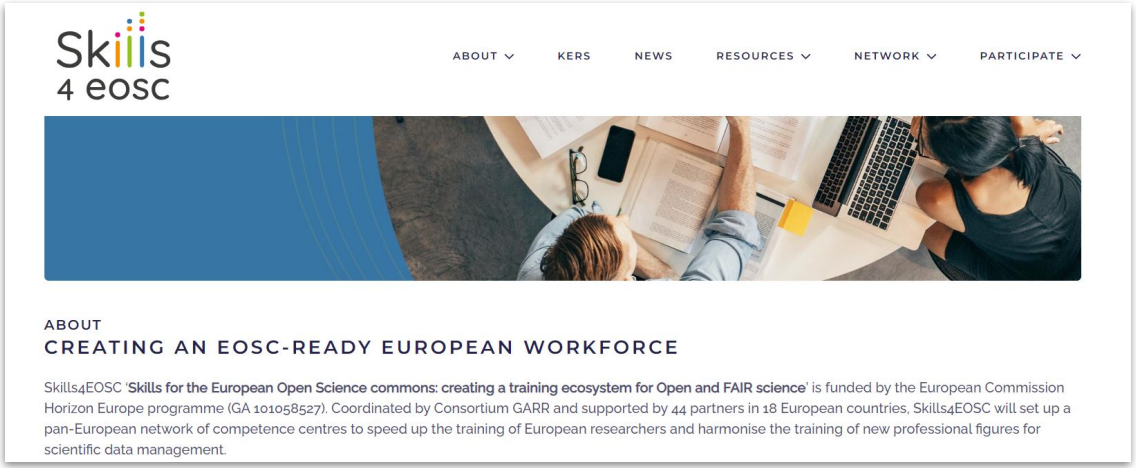
https://www.knowledge-exchange.info/news/articles/26-3-24



		WHOLE OF INSTITUTION	RESEARCH STAGE: BEFORE STUDY CONDUCT – EDUCATION	BEFORE STUDY CONDUCT – GRANT WRITING	BEFORE STUDY CONDUCT – PROTOCOL WRITING	DURING STUDY CONDUCT – RESEARCH CONDUCT & ANALYSIS	AFTER STUDY CONDUCT – MANUSCRIPT WRITING	AFTER STUDY CONDUCT – MANUSCRIPT SUBMISSION	AFTER STUDY CONDUCT – POST-PUBLICATION
UNDERLYING MECHANISMS OF INTERVENTION, including Institutional Culture and Individual Ethos	TYPE OF INTERVENTION								
	TOOLS (1) (Enablement)	Availability of open source and reproducible software packages [4,5]	Peer-to-peer tool sharing	Boilerplate language	Provide study design specific protocol templates	Shared-version control repositories	Author and contributor unique identifiers e.g., ORCID[4]	Journal management system elicitation of registration and other quality indicators	
						Institutional code repositories with mandated upload	Use of Software containers for ensuring package dependencies and the computing environment are reproducible[4]		
							Use of continuous analysis for regularly updated data		
				Successful grant application libraries	Providing an open data statement as default in ethics consent form templates	Use of continuous-analysis with automated unit-testing / error-checking	Institutional code repositories with mandated upload – prevents research hiding in individual file drawers		
							Use of continuous-analysis with automated unit-testing / error-checking (2,3)		
						Data dictionaries	Authorship guidelines for authorship information decisions and authorship info reporting (4,5)		
	EDUCATION AND TRAINING in research quality and reproducibility (Training)	Journal clubs including researchers' publications	Training on systematic literature searches	Provide training for individuals to review grants	Training on use of reporting guidelines, including protocols and registration	Train research assistants, etc about good data collection practices	Training to enhance writing skills for publications (6)	Training on submission process, including accessing funds for publication fees	Training on presentations - oral and poster for conferences and research seminars with different modes: F2F, online live and pre-recorded
		Department or staff within the institution dedicated to research quality and reproducibility interventions and activities	Seminars, workshops, presentations on research quality topics, including practical tips and activities to improve skills (e.g., how, why and where to register studies)	Personalised, tailored support e.g., for statistical support		Training manual/data collection protocol, including use of equipment and clinical trial training	Training sessions on use of reporting guidelines		
		Collaboration with external research institutions and organisations (e.g., ResBaz)				Training on data and code sharing	Training on writing tools, reporting guidelines and software (7,8,9)		Training on social media usage
		Research integrity training							
	INCENTIVES to enhance AWARENESS, ACCESSIBILITY & UNDERSTANDING (Incentivisation)	Hiring and promotion criteria that include open science, quality, and reproducible practices				Awarding small grants / prizes for adhering to best methodological practice			Include code/data sharing in promotion criteria
		Incentives for open science practices (workload models, awards, showcases, promotion)[4]				Recognition of research software as a key research output and dissemination, as well as publications			Recognition of use of pre-established data
	MODELLING AND MENTORING to encourage quality and reproducibility (Modelling)	Create research teams with effective mix of research expertise	Dedicated time in work hours to participate and attend interventions and activities	Encouraging researchers to apply for grants where the Registered Report is linked to a funder and a journal[4]			Use of DevOps practices for research software and analysis development[4]	Encouragement of protocol publication	Model use of social media for dissemination
			Mentor/mentee partnerships						
			Professional governing bodies and associations with dedicated guidelines/criteria for members to obtain research qualifications and training, E.g., RACGP, etc.				Plain language/consumer summary of study - either included in manuscript or as supplementary (as per journal guidelines)	Encouragement of the use of journal checklists	Checking for outcome switching and publicising the results
			Raising awareness to individuals of opportunities						
	REVIEW & FEEDBACK (Persuasion)		Consultations and reviews by peers e.g., statistical consulting, peer code review[4]; writing circles	Peer-review of proposals and protocols (10)	Ethics committee evaluates appropriateness of methods (e.g., use of blinding, randomization, sample size calculation)	Use of pull-requests and code commentary by collaborators and/or external peers on shared code-bases	Pre-submission peer-review (10)		Post-publication peer-review[4]
			Education for ECRs on how to conduct peer-review (7)	Mentor/mentee partnerships	Shortening and design specific ethics forms to reduce work time spent on applications	'Living research' analyses in articles can be shared in a 'sandbox' computing environment			Institutional-level checks for researcher compliance of institutional policies
			Requesting researchers to feedback on education and training, and gaps in their knowledge and skills	Research office checks where funds have been requested for statisticians/methodologists	Peer-review of protocols				
	EXPERT involvement and advice (Education)	Specific hiring for people with experience of open research, data stewards, etc. and/or training those currently employed to do this.	Expert and specialist-run courses for staff and students	Pre-submission peer-review (10)	Hiring dedicated experts to work with researchers across all departments	<	Writing support for manuscripts (11)	Support for administrative tasks	Dissemination to end-users
			Availability of peers and colleagues to assist one another in research quality improvement	Engaging with external consulting organisations[4]	Co-design with patients and public/end-users	Dedicated data champion	Hiring dedicated experts to work with researchers across all departments	Publications officer to check adherence of paper to reporting guidelines	
					Librarian involvement for literature reviews e.g., search strategies			Librarian involvement for identifying appropriate journals for submission	
	POLICIES & PROCEDURES (Coercion)		Compulsory training (with flexible modes - F2F, online live and pre-recorded)	Seed grants to refine 'near miss' grant application which meet quality criteria	Mandate study registration	Requirement for data management plans and integrity checks	Policies for authorship, reporting checklists, and appropriate journal lists	Data sharing policies	Sharing an "author" version of manuscripts in institution's depository
			Open science curriculum for under- and post-graduates					Manuscript submission checklists	Random audits of research output

Examples of the interventions identified to improve research quality

Table 1 in https://www.biorxiv.org/content/10.1101/2022.12.08.519666v1.full



conducting in collaboration with other institutions (TU Delft, KIT, etc.) as part of the Skills4EOSC project. The study aims to stimulate the uptake of the FAIR principles for improving Research Data Management in the ML/AI domain.

In line with this aim, we are currently in the process of defining a list of **'Top 10 best practices for FAIR implementation in AI/ML'**. To establish a community consensus on these best practices, we are conducting a three-round modified **Delphi study**. The first two rounds of the study will be conducted through an online survey using **EUSurvey**.

During the **first round**, participants will be asked to vote on a list of 20 recommended practices and suggest any additional ones if desired. The list of 20 practices has been generated by the project team based on desk research and discussions with researchers working in the field. The first round is scheduled for the **last week of June 2024.**
In the **second round**, participants will re-vote the FAIR practices that didn't reach consensus, as well as vote on any additional practices that were suggested by participants during the first round.
The third round will be organized as an online consensus meeting, involving a selected group of participants to facilitate more in-depth discussions and reach consensus among them.

https://www.skills4eosc.eu/

Discussion

Outlook

Next meetings on July 16th

Follow up on the list of problems w.r.t panel's mandate topics
Hear from networking (Harvey)
Prepare a presentation for the ICFA meeting (July 20th)

Thank you!

Questions?

And thanks to SlidesCarnival for this free presentation template

Mission

Mission

The mission of the panel is to enhance global coordination on all aspects of the data lifecycle including acquisition, processing, distribution, storage, access, analysis, simulation, preservation, management, software, workflows, computing and networking in particle physics, with a focus on open science and FAIR practices.

In order to achieve this, the panel will

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

Mandate 1

Mandate

1. Address the data lifecycle within a structured and integrated systems approach in HEP
    - 1.1. Formulate recommendations on organisation, technology, standards, outreach, education for past/current/future experiments.
    - 1.2. Connect regional and local activities in the field and encourage international cooperation, aiming at stimulating active participation from the global HEP community.
    - 1.3. Raise awareness of open science and the FAIR principles applied to data, software and workflows, and stimulate relevant developments.
    - 1.4. Assess the openness and FAIRness of the field.
    - 1.5. Encourage transfer of knowledge
    - 1.6. 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).

Mandate 2

Mandate (cont)

2. Improve the awareness for the importance of the data lifecycle in HEP
   - 2.1. Work out and communicate the motivation of FAIR (findability, accessibility, interoperability, and reusability) principles and open science and encourage its dissemination.
   - 2.2. Organise workshops, formulate recommendations and cookbooks, issue global reports
   - 2.3. 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.
   - 2.4. Help in sharing expertise and existing solutions; catalyse new common projects; promote collaboration.

Mandate 3

Mandate (cont)

3. Encourage and foster connections to other fields of science, to industry and to open science initiatives in order to profit from their expertise and competence in the following fields:
   - 3.1. Big and distributed data management.
   - 3.2. Data management systems.
   - 3.3. Artificial intelligence.
   - 3.4. Open science processes.
   - 3.5. Data preservation systems.
   - 3.6. Reach out to neighbouring fields such as astro(particle) physics, hadron physics, and accelerator science, but also to the communities of photon and neutron science and others with large data volumes and related data challenges (genomic, public health, smart city, ...)

Mandate 4 & 5

Mandate (cont)

4. Help in organising practical support and act as point of contact for practical issues in the field of data, software, workflows and computing
    4.1. Support the ongoing projects and co-operations started within DPHEP in order to maintain data sets that (can) still produce science, keep track on parked data sets
5. Improve recognition of the nature and value of work on the data lifecycle in researchers' CVs and support their career development.