ICFA Data Lifecycle Panel: Terms and definitions

ICFA Data LifeCycle Panel meeting- April 15, 2024

Discussion

Terms and definitions – 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.

Data lifecycle - perplexity.ai

A data lifecycle refers to the sequence of stages that a particular unit of data goes through, from its initial creation or collection to its eventual archiving or deletion. The key stages in the data lifecycle typically include:

1. Data Collection/Creation: This is the initial stage where data is generated, acquired, or captured from various sources, such as user interactions, sensors, manual entry, or external data providers.[1][3]

2. Data Storage: The collected data is stored and organized in databases, data warehouses, or other storage systems for further processing and analysis.[4]

3. Data Preparation/Processing: The stored data is cleaned, transformed, and prepared for analysis by removing errors, handling missing values, and converting it into a suitable format.[3]

4. Data Analysis/Usage: The prepared data is analyzed, interpreted, and used to derive insights, make decisions, or support business operations and processes.[2][3]

5. Data Sharing/Communication: The analyzed data or insights derived from it may be shared or communicated with relevant stakeholders, such as decision-makers, customers, or other systems.[5]

6. Data Archiving: When the data is no longer actively used or needed for current operations, it can be moved to long-term storage or an archive for future reference or compliance purposes.[1][4][5]

7. Data Deletion/Destruction: At the end of the lifecycle, data that is no longer needed or has reached the end of its retention period is permanently deleted or destroyed in a secure manner.[4][5]

The data lifecycle is an ongoing process, and different types of data may have different lifecycle durations or stages depending on their relevance, usage, and regulatory requirements.[4] Effective data lifecycle management involves implementing policies, tools, and processes to ensure data quality, security, and compliance throughout its lifecycle stages.[2][3]

Citations: [1] https://www.techtarget.com/whatis/definition/data-life-cycle
[2] https://ecampusontario.pressbooks.pub/drivingchangeinthehealthsectoranintegratedapproach/chapter/section-4-data-lifecycle/
[3] https://segment.com/blog/data-life-cycle/
[4] https://coresignal.com/blog/data-lifecycle/
[5] https://indigenousdatatoolkit.ca/data-management/the-data-lifecycle/

Data Lifecycle - from the discussion

- Our role is to generate some benefits through changes, evolutions or specific developments
- We need to consider the impact: in the chain of data lifecycle some parts have constraints e.g. due funding or infrastructure and that has an impact
- The sort of knowledge we are handling with is not necessarily widely known and it is our role to spread that knowledge

Open science and FAIR principles

https://www.unesco.org/en/open-science

https://www.go-fair.org/fair-principles/

FAIR practices for

- data
- research software
- computational workflows

in connection with “research integrity”:

- the ethical and professional standards that should guide the conduct of research.

Terms - Mission

global coordination

is:

international, infrastructure is global in HEP, need to cooperate with neighbouring fields (= fields with similar needs or challenges)

is not:

regional, or if yes: we should transfer knowledge from one region to the next

strategic questions

is:

ultimately related with money, steer funding agencies
long term planning
manpower

Tactics have to be included with milestones etc

is and is not:

tactics (how to move towards the strategic goals by doing work, having coherent milestones, deliverables)

systems competence

is:

Data lifecycle could be thought of as one big system that includes actors, data, infrastructure, collaborations that are interdependent and need to be looked at as a whole to understand the interactions

Different people have different view of the system

is not:

Terms and definitions - Mandate

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: topic 1 - terms

structured and integrated systems approach

is:

have subgroups in the panel that feed in the knowledge of different topics

some of the topics 1.1-1.6 are already being worked on, other need to be done (real work by this panel)

now discussing only terms and definitions, the actual actions will be discussed in the coming meetings

is not:

or ignore this term and consider points 1.1 - 1.6 that follow?

Terms and definitions - Mandate

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: topic 2 - terms

awareness for the importance of the data lifecycle

is:

reaching to all levels, to all stakeholders, not only individuals but also e.g. in the experiments to the medium and high management

not only training but making it possible for individual to work following FAIR practices and towards open science

audience is also the funding agencies

the examples are very much about training but important part is also the collaboration decisions e.g. requesting certain practices and making them mandatory

the mandate can be tuned up by the panel to explicitly mention these points

is not:

recognition is very important but it has its own entry in the mandate (point 5)

Terms and definitions - Mandate

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: topic 3 - terms

Big and distributed data management

examples:

Is this more about methods than tools?

In HEP:

WLCG

Open Science Data Federation (OSDF), a new program in the OSG ecosystem

In other fields:

HPC in general

Data management systems

examples:

If the first term is about methods, tools would be here.

in HEP:

rucio, frontier, XRootD, root

in other fields:

?

Data preservation systems

examples:

CERN Open Data portal?

HEPData?

Other data repositories?

Terms and definitions - Mandate

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.

Mandate: topic 4, 5 - terms

any terms to clarify?