

TA5 Metadata document

Some questions for discussion

10.11.2022

A. Redelbach

Metadata document – Table of contents

1	Concepts	2
2	Data irreversibility and metadata	4
2.1	Short overview of work in TA5	4
2.2	Data reduction and the challenges for metadata	5
2.3	Hierarchical dynamic metadata	6
2.4	Recursive meta-data	7
3	Use cases	7
3.1	Data from tracking in high energy physics	7
3.2	Data from the ground-based air-shower observations	8
3.3	Concepts for related data from simulations	9
4	Previous approaches	9
4.0.1	CERN open data and preservation	11
4.0.2	Data preservation for the HERA experiment	11
5	Requirements for meta-data in PUNCH	12
5.1	WP 1 - Discovery potential and reproducibility	12
5.2	WP 2 - Dynamic Filtering	12
5.3	WP 3 - Dynamic Archiving	13
5.4	WP 4 - Scalability	13
5.5	WP 5 - Evaluation and validation of instrument response & characteristics	14
5.6	Meta-meta data and workflows in the dynamic life-cycle	16
5.7	Extra requirements from anomaly detection workflows	17
5.8	Metadata storage size	18
6	Towards the dynamical data life-cycle	18

Section 1 – Questions

Concepts:

Flexibility as central requirement to be highlighted

Metadata correspond to **annotations**?

Table: **Simulation step** related to level 1?

Table: We need **explanations** in form of an use case to better understand the context/workflows.

Can we add a **similar table for HEP** also?

For discussion (from Hans):

- 1) There is a special form of metadata, level 0, which could be regarded as part of the data in the hierarchical definition. If level 0 metadata is lost, it would make the data less informative or even useless. For most purposes, **level 0 metadata can be considered part of the data**, and it is not necessary to distinguish it explicitly from data.
- 2) I define the hierarchy of metadata based on the way how metadata is constructed. This implies that **for different experiments and processing pipelines, level 2 metadata may be different things**.
- 3) For example, in collider experiments, level 1 metadata is a track and for the Pierre Auger Observatory, level 1 is a shower. Level 2 for a collider experiment then is a reconstructed decay candidate, but I think there is no such operational level for the Pierre Auger Observatory.

Section 2 – Questions

Base metadata and meta metadata → obsolete with new hierarchical dynamic metadata

Base metadata → **hierarchical metadata**

Meta metadata → **dynamic workflow metadata**: feedback loops and complex decisions,
e.g. anomaly detection, related to **dynamic data lifecycle**

Metadata are assigned only to digital data that are permanently stored (tbd)

Section 3 – Questions

Data levels (0,1,2,...) should be added in these use cases to illustrate the scheme

Simulated data as extra data and/or metadata (tbd)

Section 4 – Questions

We need some kind of **ordering of the projects** (chronologically, HEP, Astro, metadata, frameworks).

Reference to PUNCH publication

Survey of Open Data Concepts Within Fundamental Physics: An Initiative of the PUNCH4NFDI Consortium

<https://doi.org/10.1007/s41781-022-00081-7>

→to be added

Section 5 – Questions

Base metadata and meta metadata are still used → needs update (see section 2):

When will the **WPs „finalize“ their subsections?**

Anomaly detection workflows could be related to tasks of WPs and dynamic life cycle of data

Meta data storage size: Huge data volumes might be drastically reduced if **recursive metadata** concept can be realized → still to be investigated!

Key requirement: Metadata should allow **reproducibility** of data processing

Section 6 – Questions

What are the key messages?

Which next steps to be mentioned?