TA5 Metadata document

Inter-TA-Meeting

26/04/2023

M. Kramer, A. Redelbach

TA5 Metadata document

Status, overview

- Metadata document sent to MB on December 12
- Updated version on Indico since April 20
- Link to document: https://www.overleaf.com/4394671859tvxrcknqksxr

→ **Goal: Coordination of metadata document with other TAs**

Contents

1 Introduction **2**

2 Concepts **3**

3 Data irreversibility and metadata **3**
3.1 Short overview of work in TA5 . . . 3
3.2 Data reduction and the challenges for metadata . . . 5
3.3 Hierarchical dynamic metadata . . . 6
3.4 Recursive metadata . . . 9

4 Use cases **9**
4.1 Data from tracking in high-energy physics . . . 9
4.2 Data from the ground-based air-shower observations . . . 11
4.3 Metadata in Pulsar searches . . . 12
4.4 Concepts for related data from simulations . . . 13

5 Previous approaches and frameworks **14**
5.1 Data provenance . . . 14
5.2 Frameworks for Big Data . . . 14
5.3 PUNCH4NFDI . . . 15
5.4 Data Processing Levels in NASA/EOSDIS . . . 16
5.5 CERN open data and preservation . . . 16
5.6 Data preservation for the HERA experiment . . . 17

6 Requirements for metadata in PUNCH **17**
6.1 WP 1 - Discovery potential and reproducibility . . . 18
6.2 WP 2 - Dynamic Filtering . . . 20
6.3 WP 3 - Dynamic Archiving . . . 21
6.4 WP 4 - Scalability . . . 22
6.5 WP 5 - Evaluation and validation of instrument response & characteristics . . . 23
6.6 Metadata and workflows in the dynamic life-cycle . . . 25
6.7 Extra requirements from anomaly detection workflows . . . 26
6.8 Metadata storage size . . . 27

7 Towards the dynamical data life-cycle **27**

TA5 Metadata document

Update: Preample and Introduction

Preamble added after discussions in the CollabTools meeting last week

→ Sharpening of context, relations to other TAs and scope

Preamble

The curation of data and the concept of the associated metadata are relevant for all TAs in PUNCH4NFDI and, obviously, also very much relevant beyond our own consortium for the whole of NFDI. A number of specific challenges arrive with the focus on TA5, caused by the huge data streams and the needs for heavy on-line processing. Solutions to address these challenges must not, however, be designed in isolation of TA5 but must find the applicability also in other TAs, if not now then certainly in the future. Vice-versa, concepts and implementations in other TAs must be flexible enough to accommodate TA5 requirements in the future. The aim of this document is therefore *not* to provide a general and complete description of metadata in all fields of PUNCH sciences, but to start a discussion of the relevant topics by highlighting some of the specific TA5 challenges. Consequently, the document is naturally biased towards TA5 needs to convey our *current* thinking. That thinking will evolve with time as part of a process including ongoing and future TA5 work and discussions with other TAs. This document is a snapshot of this process.

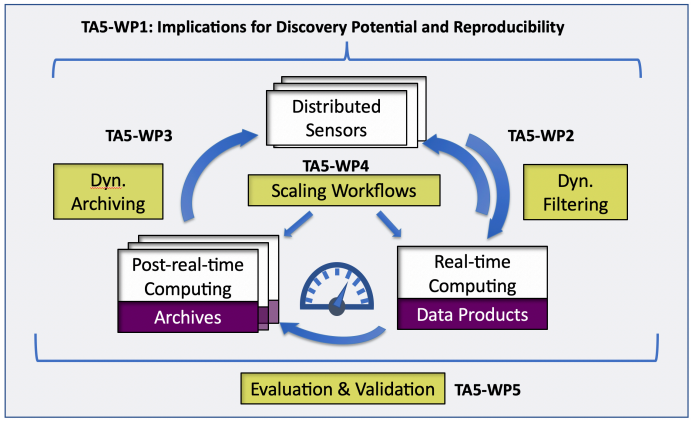
Introduction

- Metadata: “data that provide information about other data”
- Metadata describing a measurement or experiment should ideally not only describe a data set and its relevant parameters, but they should also contain information about the experiment itself, environmental conditions and
- In particular, any relevant information about how and why certain information was selected and, ideally, why other were not.
- Increasingly, metadata by themselves can become very large as a consequence.

- Not all data can be stored – including not all metadata! (”persistent” and “transient”) data
- What data do we need to keep to capture an experiment and understand its results?
- Note that in (time-domain) astrophysics, one cannot reproduce an experiment

- An archive is constantly modified, but an experiment perhaps also
- This modification may be motivated by results from the archive

- Dynamical life-cycle of data



Dynamic Life Cycle Model

TA5 Metadata document

Update: Figures illustrating workflows/data/metadata

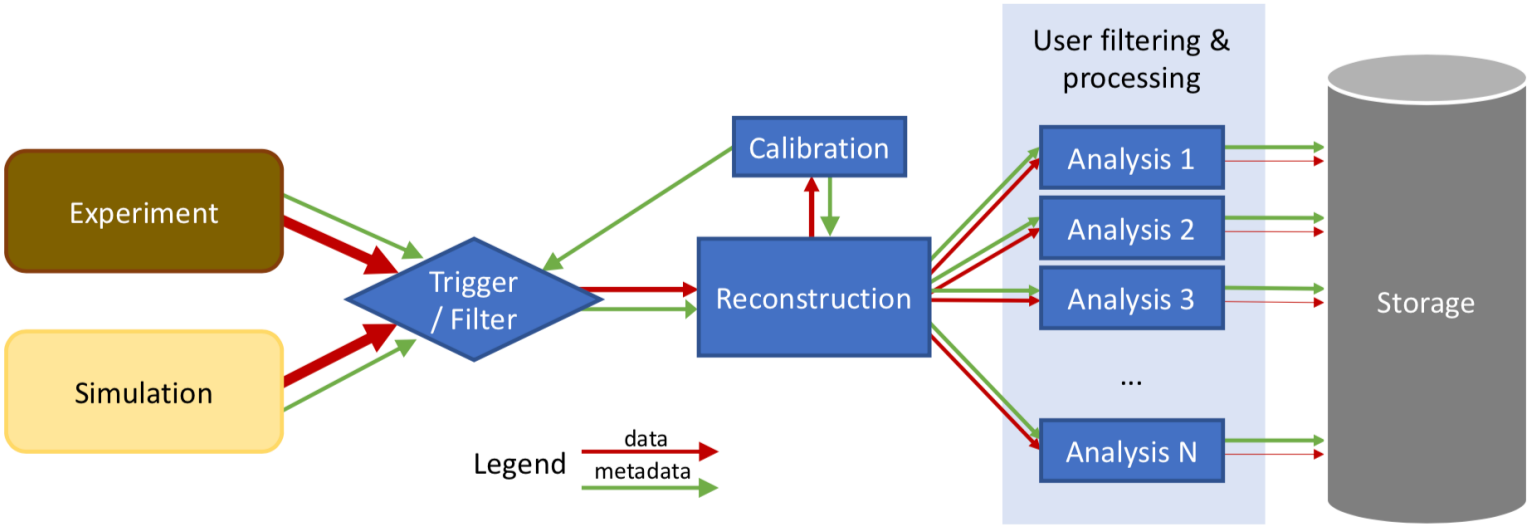


Figure 2: General data processing graph for particle and astroparticle experiments. Variations of the data flow and triggering scheme are possible. The arrow width qualitatively indicates the data rate.

TA5 Metadata document

Update: Figures illustrating workflows/data/metadata

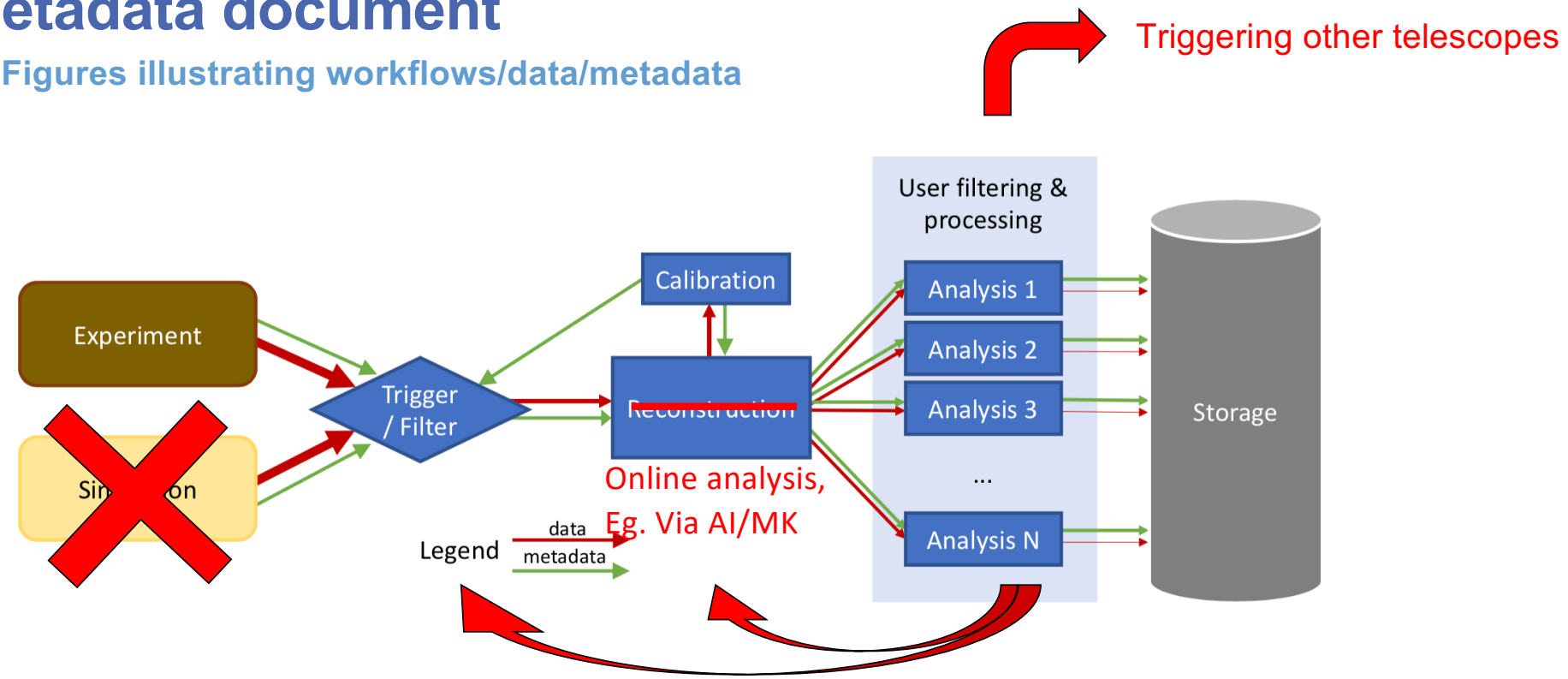


Figure 2: General data processing graph for particle and astroparticle experiments. Variations of the data flow and triggering scheme are possible. The arrow width qualitatively indicates the data rate.

TA5 Metadata document

Update: Figures illustrating workflows/data/metadata

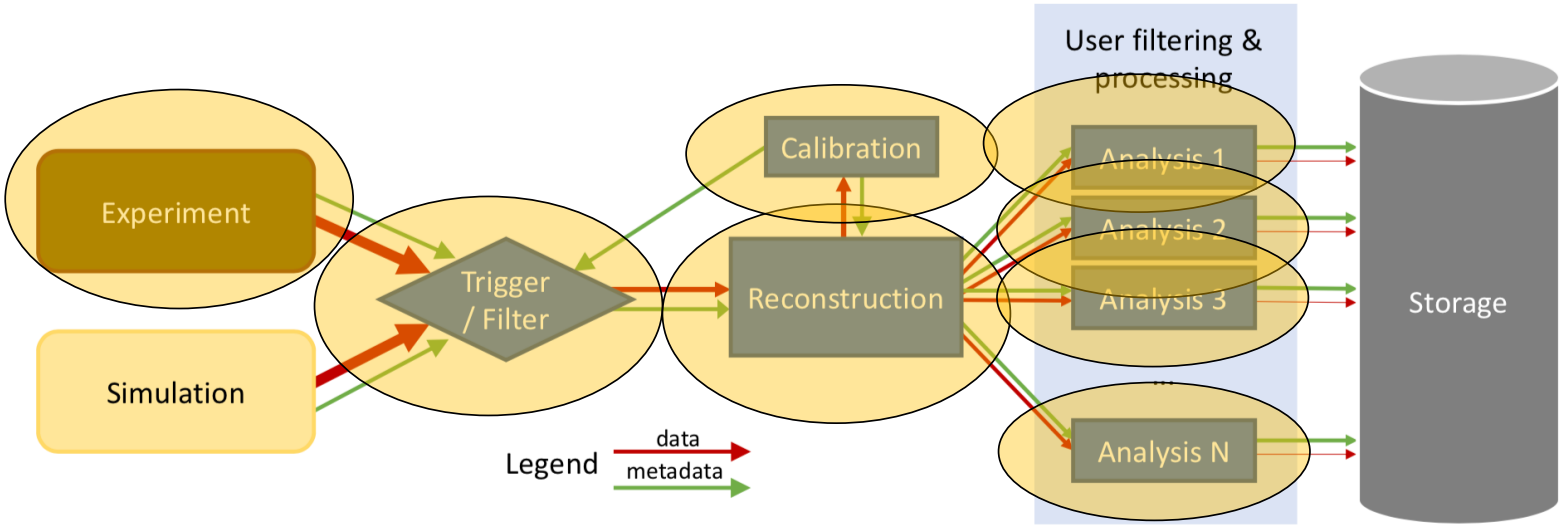


Figure 2: General data processing graph for particle and astroparticle experiments. Variations of the data flow and triggering scheme are possible. The arrow width qualitatively indicates the data rate.

TA5 Metadata document

Hierachical dynamic metadata

- Raw data with basic ”level-0” metadata
- Higher-level data and metadata are built, which form a natural data hierarchies
- Metadata is of a higher level if its construction depends on metadata of lower level or if it describes data of a lower level, otherwise it is of the same level
- With the level of metadata, the abstraction level increases and the distinction between data and metadata can become blurred as high level data directly depends on lower level metadata.
- Low-level metadata is often automatically created and centrally processed to higher levels, and most analyses operate entirely on higher level metadata.
- ”Data” and ”metadata” evolve into levels containing data of increasing complex-ity, and dependency on other sources of information
- Information at each level requires lower level to have existed in order to obtain meaning
- Data hierarchies can also branch, both in the sense of higher levels combining lower level data from multiple sources (sensors) as well as different higher level processes using the same lower level data.
- Common that complementary experiments are performed simultaneously during the same data-taking process.
- Data irreversibility emerges when some parts of this hierarchy is not available for subsequent analysis.
- Need decisions, how to handle different branches that have the same Level-0 or Level-1 origin. (Inadequate to keep copies of the same low-level items, but in order to avoid duplication lower levels in our metadata structure may be simply a reference, pointing to a single physical location of those items)
- There is a need for recursive metadata

TA5 Metadata document

Table: Illustrating levels of data/metadata in astrophysics

Data level	Content	Sample content
L0	A sensor measurement.	Count level in CCD detector.
L1	Annotations referencing L0	CCD temperature and readout gain.
L2	Operation on Level 0/1.	Photon flux reaching telescope.
L3	L2 annotations ("meta-data").	Reference to calibration algorithm / ancillary data.
L4	Operation on L2/3.	Brightness of astronomical source.
L5	L4 annotations.	Reference to astronomical catalog.
L4	Operation on L4/5.	Source classification.
L5	L4 annotations.	Reference to astronomical catalog.
L6	Analysis output (Operation on L4).	Real-time follow-up announcement.
L7	L6 annotations.	Publication reference.

Table 1: A description of data levels suitable to a sample analysis of astronomical data. Here, data and metadata levels are interleaved to emphasize that higher level data can depend on lower level metadata.

TA5 Metadata document

Section 4 Use cases

Description of representative use cases focusing on types of data/metadata arising in processing/workflows:

- Data from tracking in high-energy physics
  - Reconstruction and data reduction in high energy particle and nuclear physics
  - Natural hierarchy for more abstract levels of reconstruction
- Data from ground-based air-shower observations
  - Reconstruction of primary incoming particles and their properties
  - Mixture of data and metadata for modern approaches like machine learning
- Metadata in pulsar searches
  - Initial Fast Fourier Transformation for information of frequency and time
  - Subsequent searches for periodic or transient astrophysical signals
  - Goal: Optimising sensitivity for FRBs and pulsars
- Concepts for related data from simulations
  - Publishing of simulation data is rare in many sub-fields (of astrophysics)
  - Providing a comprehensive list of astrophysical and numerical parameters is quintessential for reproducibility
  - Metadata should also contain information about the computer used and the technical setup

→ **For discussion: Connecting these use cases to existing (originally defined) use cases**

TA5 Metadata document

Update: Figures illustrating workflows (example LHCb)

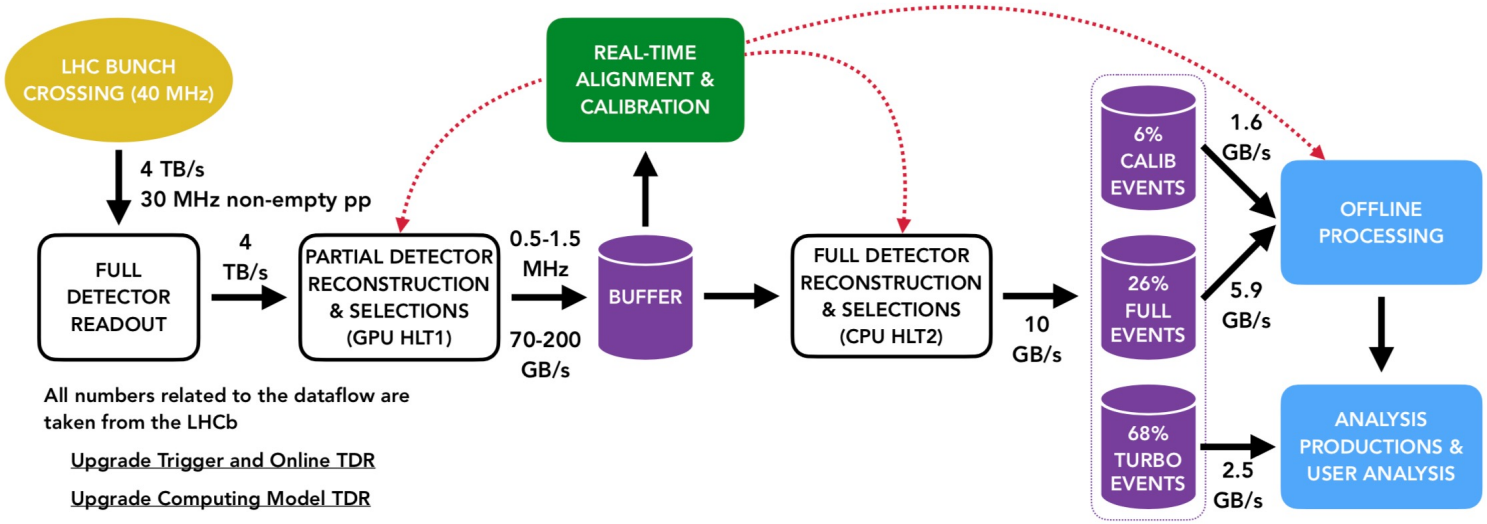


Figure 3: Current data processing pipeline of the LHCb experiment for proton-proton collisions [6, 7]. Arrows indicate data flow, which are annotated with event and data rates.

TA5 Metadata document

Section 5 Previous approaches and frameworks

Overview of existing approaches and solutions related to processing metadata as a basis for future developments:

- Data provenance: DataCite's Metadata schema, VAMPIRA project, Virtual Observatory with Table Access Protocol, FITS (Flexible Image Transport System) format, International Lattice Data Grid, ATLAS Metadata Interface (AMI) with Metadata Querying Language
- Frameworks for Big Data: Rucio and others
- PUNCH4NFDI (see next slide)
- Data Processing Levels in NASA/EOSDIS: Hierarchical levels (0 to 4)
- CERN open data and preservation: Hierarchical scheme (1 to 4, other direction for abstraction)

  → Availability of Level 4 data critical: Covering basic raw data with accompanying reconstruction and simulation software, allowing the production of new simulated signals or even re-reconstruction of collision and simulated data
- Data preservation for the HERA experiment

TA5 Metadata document – references

References to other PUNCH or NFDI projects/papers

Punch4nfdi consortium proposal

Victoria Tokareva. Metadata curation use cases in astroparticle physics

Thomas Schörner-Sadenius, Harry Enke, Andreas Haungs, Kilian Schwarz, Markus Demleitner, Achim Geiser, Lukas Heinrich, Michael Kramer, Gernot Maier, Dominik Schwarz, Hendrik Seitz-Moskaliuk, Hubert Simma, Michael Sterzik, and Stefan Typel.
Survey of Open Data Concepts Within Fundamental Physics

Sektionskonzept Meta(daten), Terminologien und Provenienz zur Einrichtung einer Sektion im Verein Nationale Forschungsdateninfrastruktur (NFDI) e.V.

→ To be extended or updated

TA5 Metadata document

Section 6 Requirements for metadata in PUNCH

Collection of requirements for future workflows and reproducibility

- WP 1 - Discovery potential and reproducibility: General strategies, e. g. tension between strong filtering and sufficient data/metadata for unexpected discoveries
- WP 2 - Dynamic Filtering:
  - Online calibration and alignment at LHCb during reconstruction
  - Dynamic variation of filtering due to varying radio frequency interference
- WP 3 - Dynamic Archiving: Availibility of metadata to re-investigate archival data after having obtained additional or new knowledge
- 6.4 WP 4 - Scalability: Flexible data models for increasing complexity & volumes of metadata
- 6.5 WP 5 - Evaluation and validation of instrument response & characteristics:
  E. g. continuous determination of trigger efficiencies in HEP
- Metadata and workflows in the dynamic life-cycle: Questions related to reproducibility in case of partially missing (meta)data and standardisation
- Extra requirements from anomaly detection workflows: Extra validations and data streams
- Metadata storage size: Options for data compression and “intelligent“ representation of metadata

→ **For discussion: These requirements essentially correspond to TA5 only (re-naming of section?)**

TA5 Metadata – input needed

For discussion

General feedback?

What is missing? Are all relevant areas covered?

How to cross-reference requirements for other types of PUNCH metadata

Interface to offline computing / SDP / other TAs ?

Illustrate another specific example or use case?

Where do we go from here?

Formal steps for possible publication?