Review of HSF Metadata Paper

John De Stefano (BNL, IT)
Elizabeth Gallas (Oxford, ATLAS, databases)
Giacomo Govi (INFN Padova, CMS)
Thomas Kuhr (LMU Munich, Belle II)
Igor Mandrichenko (FNAL, IT)
Tibor Simko (CERN, IT, reusable analyses)

23.05.2022

The Paper

Constraints on future analysis metadata systems in High Energy Physics

T. J. Khoo[5], A. Reinsvold Hall[10], N. Skidmore[16], S. Alderweireldt[15], J. Anders[13], C. Burr[3], W. Buttinger[9], P. David[11], L. Gouskos[3], L. Gray[4], S. Hageböck[3], A. Krasznahorkay[3], P. Laycock[1], A. Lister[14], Z. Marshall[6], A. B. Meyer[2], T. Novak[2], S. Rappoccio[12], M. Ritter[7], E. Rodrigues[8], J. Rumsevicius[3], L. Sexton-Kennedy[4], N. Smith[4], G. A. Stewart[3], and S. Wertz[11]

- arXiv:2203.00463
- Conditionally accepted for publication by Computing and Software for Big Science

Structure

Abstract

In High Energy Physics (HEP), analysis metadata comes in many forms – from theoretical cross-sections, to calibration corrections, to details about file processing. Correctly applying metadata is a crucial and often time-consuming step in an analysis, but designing analysis metadata systems has historically received little direct attention. Among other considerations, an ideal metadata tool should be easy to use by new analysers, should scale to large data volumes and diverse processing paradigms, and should enable future analysis reinterpretation. This document, which is the product of community discussions organised by the HEP Software Foundation, categorises types of metadata by scope and format and gives examples of current metadata solutions. Important design considerations for metadata systems, including sociological factors, analysis preservation efforts, and technical factors, are discussed. A list of best practices and technical requirements for future analysis metadata systems is presented. These best practices could guide the development of a future cross-experimental effort for analysis metadata tools.

1. Introduction
   types of metadata, metadata scopes
2. Motivations
   sociological factors, analysis preservation, book-keeping
3. Technological considerations
   metadata formats, repository structure, access interface
4. Technical specification
   Technical requirements, desired features and other considerations
5. Summary

Review Guidelines

1) Are the requirements / specifications well stated from a technical viewpoint?

2) Are the requirements / specifications complete? Are there missing use cases / requirements / specifications?

3) Is it worth identifying the “systems” in the paper title, and should the requirements / specifications be factorised and associated with those systems?

General Comments

- The reviewers much appreciate that the analysis metadata topic is addressed, for the first time, by a community wide effort.
- The paper addresses many aspects that resonate well with the reviewers.
- The paper seems to be more a summary of a discussion and an overview of selected current implementation approaches than a formal determination of requirements. It is a first step towards and provides valuable input to a more formal requirements engineering process that should be tackled next.

Suggestions for Next Step

- ➔ Provide a detailed discussion of use cases
- ➔ Describe what problems should be addressed, not what the solution is
- ➔ Avoid mixing the discussion of the design of a system and how it is used
- ➔ Sharpen metadata scopes definitions
- ➔ Derive requirements from use cases, assign them to metadata scopes
- ➔ Discuss arguments for or against common solutions (across metadata scopes and experiments)

Metadata Scopes

- Analysis metadata (including examples 4 and 7 above) – describes features of an analysis, such as lists of required datasets and how they are used, versions of calibration metadata used to produce final results, and so on. "Datasets" here refer to samples of events as they are organised in persistent storage, usually according to some useful common criteria, e.g., data recorded during the same run with the same triggers or data simulated with common parameters
- Dataset metadata (includes examples 1, 6, and 7, arguably 4 and 5) – describes either features of datasets, or information about how to analyse datasets.
- Time-dependent metadata (includes examples 3, 4, and 5) – describes information that varies over the course of data collection, typically by being tied to timestamps on the detector data, defining "intervals of validity" (IOVs). In the case of calibration data derived through analysis of simulated or recorded data, IOVs may be as wide as "one specified year" or "one multi-year run", and handled similarly to dataset metadata.
- File-dependent metadata (includes examples 1, 2, and 3) – information about a single file, therefore typically related to the mechanics of file processing. Note that this is not the same as metadata stored in the file, which may in fact be dataset metadata or time-dependent metadata.

➔ Some examples appear in multiple scopes

Metadata Scopes

- Analysis metadata and analysis-oriented dataset metadata are two different things
- Information about features of a dataset and how it is used are two different things
- Dependency may not be on time only, but e.g. run/event (and/or calibration version, etc.)
  - Categorize as data needed for event processing, but not part of event data?
  - Interpolatable or not?

Metadata Scopes Recommendations

- Sharpen definition and general applicability of metadata scopes
- Take a more conceptual/abstract view on metadata scopes
  - Metadata is data about data. Which data?
  - On what does the metadata depend?
  - Is the metadata known at the time of data production?
  - How is the metadata produced and by whom?
  - How is the metadata used and by whom?
  - Who is the owner of the metadata?
  - What relations between different levels of metadata exist and how are they handled?

Further Questions

- Is relocatability and the distribution of metadata to multiple sites really a requirement or a feature imposed to a specific technology choice?
- How important is it to have correct metadata for partial datasets?
- How is the evolution of metadata schemata, API versions, software tools handled?
- What are advantages and disadvantages of storing the metadata together with/inside the data vs cross-referencing several metadata objects?
- Are MC datasets fundamentally different from measurement datasets?
- What are the requirements and solutions of experiments other than the LHC experiments and Belle II?
- ...

Recommendations

- We strongly encourage the community to continue the systematic and coordinated effort of addressing the metadata structure, tools, access, preservation, and reuse challenge.
- The term metadata is used for many things. The valuable discussion of metadata scopes in the paper should be continued to reach an even clearer definition and common understanding of metadata scopes.
- A more formal collection of use cases and determination of requirements would be the desirable next step.
- We very much hope that HSF will take care of the problem ownership and make sure the next steps are taken for the benefit of the HEP community.