PROCEEDINGS OF SCIENCE

Astrophysical Online Data Analysis powered by provenance data model

V. Savchenko[*]
University of Geneva, Department of Astronomy
E-mail: Volodymyr.Savchenko@unige.ch

I present the architecture of the backend of Online Data Analysis (ODA) platform, developed at Common Data Center Infrastructure of UniGE to facilitate reduction and analytics of astronomical data. The currently publicly available version allows to effortlessly explore and exploit observations of ESA's INternational Gamma-Ray Laboratory (INTEGRAL) as well as Polar GRB polarization detector. The platform can be accessed through a user-friendly on-line interface, suitable for researchers not familiar with INTEGRAL or Polar data reduction but interested in astrophysical phenomena. More involved configuration of the analysis can be achieved through detailed formulation of the requests to one of levels of the backend service infrastructure through an HTTP API. The backend relies on a provenance-driven storage and workflow management infrastructure, which ensures transparent, non-redundant, and extensible management of the scientific data and data analysis processes.

The New Era of Multi-Messenger Astrophysics - Asterics2019
25 - 29 March, 2019
Groningen, The Netherlands

[*]Speaker.

The Online Data Analysis [ODA, 15] platform developed at Common Data Center Infrasture of Department of Astronomy at University of Geneva is designed to address shared needs of the project-specific datacenters, hosted by the Department of Astronomy. It provides several interoperable interfaces for to access to the data analysis workflows of astrophysical observatories and detectors. The current public prototype incorporates the full power of the INTEGRAL/ISGRI [17] and a subset POLAR [16] data analysis.

In this paper I focus on the design of the backend of this platform, which relies on workflow-as-a-service infrastructure powered by provenance-backed storage to provide complete, efficient, non-redundant, and transpartent data and data analysis workflow management.

1. Provenance-indexed database

The design design of the ODA platform backend originated from the needs of organizing heterogenous archive of scientific data and data analysis results collected by the INTEGRAL gamma-ray space observatory [17]. While it is possible to construct a database of pre-cooked high-level data analysis products (images, light curves, spectra, etc), defining a useful selection of these products without an unreasonable loss of generality may be pose a considerable challenge. Which is why instead, I opted for developing a self-organizing living archive of data products at different levels of reduction. Both the raw data and the high-level products are preserved, with the possibility to add new products, as the demand becomes important.

The key feature of this system is to organize the storage according to the data lineage [e.g. 14], a specific case of data provenance [13]. The terms “data provenance” and “data lineage” are sometimes used interchangeably, depending on the domain. It is commonly accepted that while provenance describes everything that happened to data in the past (e.g. copy-provenance or how-provenance [14]), lineage addresses a specific aspect of provenance, the why-provenance. However for many purposes it is not necessary to distinguish between provenance and lineage. Tracing of *provenance* for the purposes of data identification and annotation alongside with the Not Only Structured Query Language (SQL), or sometimes Non SQL (NoSQL) data management became apparent with development of decentralized systems, which could not anymore easily satisfy Atomicity, Consistency, Isolation, Durability (ACID) principles of relational databases [10, 14].

In a narrow sense, data lineage metadata comprises the sequence (a collection with sufficient ordering) of the analysis steps (the analysis or workflow nodes) undertaken to produce the given data. The ordering is typically partial, since many steps follow independent paths, as can be represented by a directed acyclic graph (DAG). In other words, the metadata consists in a particular description of the associated scientific data analysis workflow (simply workflow, hereafter). The ontology of the workflow nodes prescribes specific metadata associated with each step, and induces the collection of metadata of the final product.

In order to define efficient lineage-indexed store, the workflow nodes must satisfy two properties:

1. each workflow node must be a pure function of its dependencies, i.e. it should not depend on any implicit state or properties of the analysis environment.

2. the workflow node metadata should describe accurately the transformation performed by the node, and only this transformation. Workflow node metadata are prescribed by the appropriate ontology.

If these two conditions are fulfilled, the lineage metadata contains all relevant properties of the data, and only the relevant data. Duplication of data is automatically avoided, and as new workflows are constructed by extending the ones already available, they are re-using the high-level (and well as mid-level, and any other level) products.

Such a workflow has been developed for the INTEGRAL instruments ISGRI and JEM-X [1]. The level of detail of the workflow expression is very high in the case of ISGRI (which required considerable effort from the instrument experts) and fairly low in the case of JEM-X (implemented with minimal effort). These two edge cases demonstrate the possibilities of a gradual adoption of the available instrumental knowledge.

Since many workflow nodes have more than one input, the workflows typically have a structure which can be represented by a Directed Acyclic Graph (DAG). It is possible to store this metadata in a relational database [11], for example, as a collection of edges. However, the graph structure of the indices favor a graph database solutions [5, 7], accessible with NoSQL queries, such as GraphQL[1].

Since workflow definitions are used to derive *data lineage*, they providing a *unique identity of the data* within a given namespace, with associated ontology. The namespace for the INTEGRAL products has been developed (see Figures 1, 2, 3 for a simple graphical representation). Complemented with the reference to the official *INTEGRAL product namespace*[2] these identities can be seen as *globally unique*, and may be assigned conventional citeable identifiers.

The workflow definitions are expressed in a abstract way, focusing on functional relations between different operations instead of details of the implementation. They can be naturally expressed in other formats suitable for capturing links between different entities, such as RDF, and queried with SPARQL. At the stage of the execution, the workflows are either formatted in CWL [3][9] (for example on *Reana*[4] k8s runner) or evaluated directly on a custom engine.

2. Ontology

Expressions of the workflows and associated provenance rely on a collection of well-defined related concepts - the ontology. Semantic relations between workflows are induced by their properties, such as:

1. functional signature (types of input)
2. used software, methods, and other implicit dependencies
3. inheritance relations (if any)
4. history, such as commit history

[1] https://github.com/volodymyrss/data-analysis, https://github.com/volodymyrss/dda-ddosa

[2] All the output products of INTEGRAL OSA analysis are described in the user manuals available at ISDC website. The ontology of these products will be part of the INTEGRAL legacy archive.

[3] https://www.commonwl.org/

[4] https://www.reanahub.io/

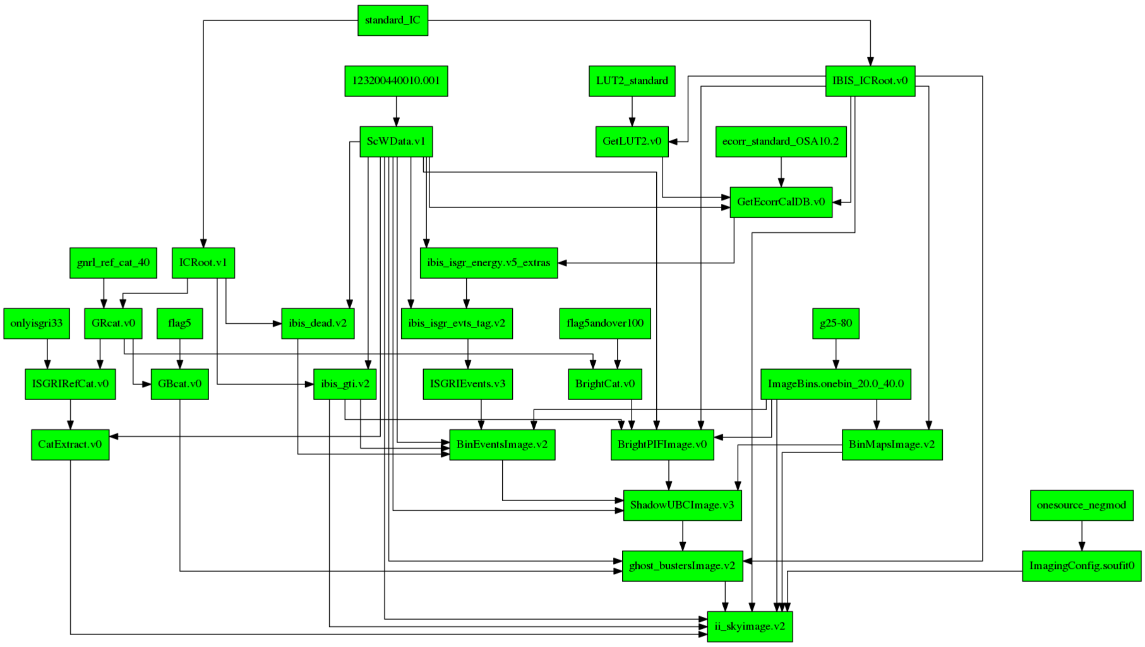

Figure 1: Example of simplified provenance information of an sky image derived for a single INTEGRAL ISGRI ScW, in a single energy range. Each node is represented by the name of the function (lineage operator) unique in the given namespace. Complete metadata of a node contains the information about the properties of the functions, and additional provenance information associated with the development of the code itself: commit history and class inheritance (see explanations in the text).

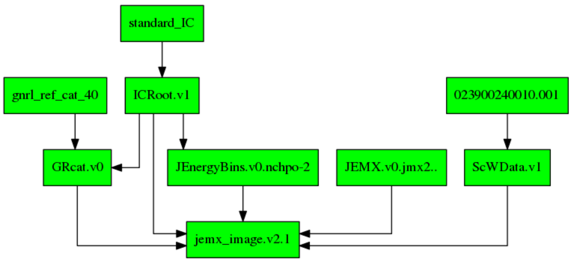

Figure 2: The same as Figure 1, but for a sky image derived for a single INTEGRAL JEM-X2 ScW, in a single energy range. Note the comparing to the lineage illustrated in Figure 1 for ISGRI, the level of detail in the JEM-X workflow is considerably smaller. It also is possible to expose further detail of this workflow while preserving validity of the available high-level JEM-X lineage by expanding some of the composed (compound) functions.

5. embedding in scientific context

This workflow ontology can be associated with existing ontology of CWL workflows. Although, to the best of our knowledge, no semnatic workflow model satisfyting implementing the aforementioned properties is currently available.

On the contrary, provenance ontology is developed in the context of W3C (and derived Virtual Observatory efforts), such as PROV-O. The abstract high-level description provenance derived from the workflow definition, can be transcribed in existing provenance standards. However, such as

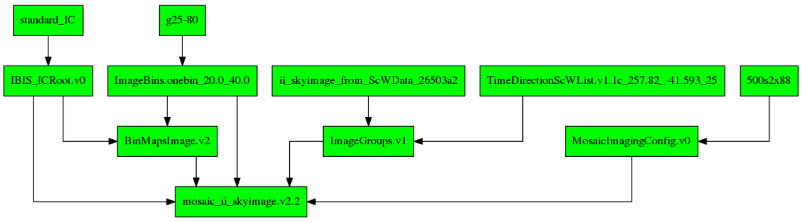

Figure 3: Lineage of a INTEGRAL ISGRI mosaic in a selection of the ScWs from a particular sky direction and time interval. Note that the mosaic is produced from individual ScWs by repeatedly applying to the TimeDirectionScWList the composed function deriving ISGRI image from a single pointing (see Figure 1 for a lineage resulting from applying this function to a particular pointing).

transformation often results in a massive increase of the description size, principally because of lack of compound provenance operators and higher-order operators (such as Map or Factorize).

3. Conclusion

Key component of the UniGe CDCI's ODA backend design is the data management infrastructure, which is using data provenance not just to annotate the data, but to index, partition, and replicate it. Exploiting data lineage, apart from purely technical advantages, favors reuse of the data and code. By building upon the acquired experience, it increases the visibility and impact of the scientific data. Holistic approach to Data Center activities, pursued by CDCI and ODA, requires careful balance of project-specific efforts vs more universal developments. I stress that increase in the role cross-project and cross-domain research and developement requires dedicated management strategies supported by broad outlook and perspective awareness.

At this point, ODA is the primary tool used by the official INTEGRAL collaboration for follow-up of multi-messenger transients, demostrating capacity to integrate collaborative contributions at diffent levels of software and domain-specific expertise, fostering broad collaborations.

Lineage provides numerous additional benefits. It allows to *credit (or blame)* the entities involved in the production of the given data products. Preserving the Offline Scientific Analysis (OSA) pipeline in a suitable fashion will allow to *expose the scientific workflow* and enable *knowledge transfer to future experiments, observatories, and missions*, while maintaining appropriate credits.

Ongoing development involves adding new instruments to the ODA infrastructure, progressively publicly releasing, documenting, and exposing, backend design, aligning it as much as possible with existing standards.

References

[1] GraphQL | A query language for your API.

[2] IVOA Recommendation - IVOA Simple Image Access.

[3] IVOA.net.

[4] JupyterLab.

[5] Ontotext GraphDB^{TM} - a Semantic Graph Database Free Download.

[6] RenkuLab.

[7] The Neo4j Graph Platform – The #1 Platform for Connected Data.

[8] The Swan Service | SWAN.

[9] Peter Amstutz, Michael R. Crusoe, Nebojša Tijanić, Brad Chapman, John Chilton, Michael Heuer, Andrey Kartashov, Dan Leehr, Hervé Ménager, Maya Nedeljkovich, Matt Scales, Stian Soiland-Reyes, and Luka Stojanovic. Common Workflow Language, v1.0. Specification, Common Workflow Language working group. 2016.

[10] James Cheney, Laura Chiticariu, and Wang-Chiew Tan. Provenance in Databases: Why, How, and Where. *Foundations and Trends in Databases*, 1(4):379–474, 2007.

[11] E. F. Codd and E. F. A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377–387, jun 1970.

[12] K.~M. Górski, E Hivon, A.~J. Banday, B.~D. Wandelt, F.~K. Hansen, M Reinecke, and M Bartelmann. HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere. ApJ, 622:759–771, apr 2005.

[13] Amarnath Gupta. *Data Provenance*, page 608. Springer US, Boston, MA, 2009.

[14] Robert Ikeda and Jennifer Widom. Data Lineage: A Survey. 2009.

[15] Andrii Neronov. An online data analysis system of INTEGRAL telescope. *in preparation*, 2019.

[16] N Produit, F Barao, S Deluit, W Hajdas, C Leluc, M Pohl, D Rapin, J P Vialle, R Walter, and C Wigger. POLAR, a compact detector for gamma-ray bursts photon polarization measurements. *NIMPR A*, 550:616–625, sep 2005.

[17] C Winkler, T J.-L. Courvoisier, G Di Cocco, N Gehrels, A Gim*nez, S Grebenev, W Hermsen, J M Mas-Hesse, F Lebrun, N Lund, G G C Palumbo, J Paul, J.-P. Roques, H Schnopper, V Sch*nfelder, R Sunyaev, B Teegarden, P Ubertini, G Vedrenne, and A J Dean. The INTEGRAL mission. A&A, 411(1):L1, nov 2003.