# VAMOS-GRIT-AGATA DMP

A.Matta, A. Lemasson, J. Dudouet

| | |
|---|---|
| Document Metadata Version | 1.0 |
| Title | VAMOS-GRIT-AGATA DMP |
| Filename | GRIT_DMP_V1.1.pdf |
| Authors | A.Matta, A. Lemasson, J. Dudouet |
| License | CC-BY 4.0 |
| Creation date | 2024-12-21 |
| Revision date | 2025-01-09 |
| Revision | 1.1 |
| Public | MB |
| Status | Drafting |
| Validation date | N/A |
| Validation body | N/A |
| DOI | N/A |

# 1 Introduction

## 1.1 SCOPE OF THE DOCUMENT

This document detail the data and software strategy of the VAMOS-GRIT-AGATA (VGA) campaign at GANIL. It aims at delivering a common working scheme between the different actors of the campaign to insure reproducibility and openness of the data produced during the campaign and its following analysis.

This document specifies the different policies implemented by the collaboration to pursue this goal. It describes both the technical requirements and the organisation that allows implementation of the policy. It defines the role of the Data Manager and Analysis Officer and their responsibilities to the different collaboration stakeholders.

# 2 Scientific data

## 2.1 EXPECTED DATA VOLUMETRY

The detector will start taking data from 2028 for an expected 5 year-period. A typical experiment will produce around 1 TB of raw, analysed and simulated data. One to five experiments will be performed every year, translating into a typical 5 TB of data produced per year.

Analysis and simulations will be performed by members of the collaboration at the storage location as well as at collaboration members local sites.

## 2.2 GENERAL POLICY

- Licensing
- metadata file naming
- metadata format
- data conservation
- reference document detailing the dataset structure and data content as well as update mechanism.

**Policy D1.1:** Scientific data are curated by the Data Manager (see Sec. 4.1).

**Policy D1.2:** All scientific data produced by the collaboration are placed under CC-BY 4.0 open data license.

**Policy D1.3:** All data files are associated to a separate metadata file.

**Policy D1.4:** All metadata files are machine readable and contain all necessary information on provenance and/or allowing reproducibility of the associated data file.

**Policy D1.5:** Data and metadata format, as well as dataset folder structure, are defined in a document produced by the Data Manager and the Analysis Officer and validated by the campaign management board.

| Level | Description |
|---|---|
| L0 | Raw data in DAQ normal or compressed native format |
| L1 | Raw data converted in root format |
| L2 | Calibration, Threshold and Reconstructed Data |
| L3 | Computation of physics observables |
| L4 | Histograms, tables and graphs |

Table 1: Data level description

## 2.3 DETECTOR DATA

**Definition D2.1:** Detector data are produced by the detector system during operation. This includes both test, calibration, and experiments. It also includes all derivative files produced during subsequent analysis.

**Policy D2.2:** Any person who participated in the detector construction or updates is de facto considered a co-author of any data produced by the detector system. In addition, any person directly involved in the data-taking process is a de facto co-author of the resulting data. Any person directly involved in the data processing process is a de facto co-author of the derived data produced.

## 2.4 SIMULATED DETECTOR DATA

**Definition D3.1:** Simulated detector data are produced through a simulation programme. Simulated detector data should adhere to the detector data file format and be fully compatible with any analysis software used for detector data.

**Policy D3.2:** Any person directly involved in running the simulations is a de facto co-author of the released simulated dataset. In addition, any person directly involved in developing new and released functionalities allowing to run these simulated data sets is a de facto co-author of the simulated dataset.

**Policy D3.3:** Simulated data and dataset quality are assessed by the Data Manager. Curation of the dataset is the responsibility of the Data Manager, including deletion of non-essential data files.

## 2.5 AUXILIARY DATA

**Policy D4.1:** All auxiliary data collected during the experimental process and allowing exploitation of the data are to be stored. These data include, but are not limited to, human and machine-generated logs.

## 2.6 DATA CATALOGUE

**Policy D5.1:** The Data Manager is responsible for implementing and maintaining a machine-actionable data catalogue. This catalogue should be accessible to members of the collaboration through a suitable Authentication Authorization Infrastructure (AAI) service interface. This catalogue includes the openness status of the data.

## 2.7 DATA CURATION AND PRESERVATION

**Policy D6.1:** All scientific data are preserved at, and distributed from CC-IN2P3 through the existing AGATA data infrastructure under the supervision of the Data Manager. Members of the collaboration are required to provide to the Data Manager all produced derived data as a structured dataset for safekeeping at CC-IN2P3 following the procedure described in section 4.

**Policy D6.2:** Overall dataset lifetime is set to 10 years after the final experimental data taking. Lifetime of the different data inside the dataset is dependent on the Publication date of a research article. L0 data are to be kept for at least 2 years after initial research publication. L1 and L2 data are to be kept for at least 10 years after first publication. L3 data are to be published in a long term open repository of the collaboration choice before publication of the research article. *At the end of a dataset lifetime, L2 data as well as all metadata are to be published in a preservation repository of the Data Manager choice.(A affiné)* Simulated data are to be kept for at least 2 years after publication. The lifetime of data could be extended by a specific and motivated request from any of the co-authors to the Data Manager.

**Policy D6.3:** The Data Manager can propose to the management board the re-qualification of data files as junk. Junk files could be deleted as long as their metadata file are preserved. The Data Manager is responsible for proposing data deletion to the Management Board and Steering Committee. Once both MB and SC agreed, the Data Manager can proceed with data deletion. *In any case before deletion of any L0 data, the Data Manager is responsible for overseeing production of decimated data file keeping a random 10% of the global statistics in each of them* This decimated data file is to be kept for the duration of the dataset lifetime.

# 3 Software

This section describes the software strategy adopted for the VGA campaign, including reproducibility requirements, licensing, version control, metadata production, and long-term preservation.

## 3.1 REPRODUCIBLE ANALYSIS AND SIMULATION

**Policy S1.1:** All analysis and simulation activities performed within the VGA campaign must be reproducible throughout the entire dataset lifetime by any member of the collaboration, given a reasonable amount of effort. This implies that all software source codes are openly available to the collaboration, properly documented, versioned, and that all analysis and simulation steps are fully machine-actionable.

**Policy S1.2:** Each analysis or simulation associated with a given dataset must be delivered as a single, self-contained software project. This project shall include:

- the complete source code together with its associated license and list of authors and contributors,
- a complete description of software dependencies and, when relevant, compilation or build instructions,
- all configuration, calibration, and auxiliary files required to run the analysis or simulation,
- a documented description of how to execute the software and of the nature of the expected outputs,
- metadata files describing all produced data outputs.

**Policy S1.3:** All analysis and simulation software used within the VGA campaign must be released under a recognized open-source license compatible with scientific reuse and long-term preservation. Permissive open-source licenses such as Apache License 2.0 or MIT are recommended, as they facilitate broad reuse and integration. Apache License 2.0 may be preferred when explicit patent protection is required, while MIT is well suited when maximal accessibility and minimal constraints for future users are desired.

**Policy S1.4:** All software source codes must be managed through the official version control infrastructure of the VGA campaign. Any software version used to produce a finalized dataset (e.g. calibration releases, reconstructed data, or final analysis outputs) must be identified by a corresponding version tag in the version control system. These tags shall be referenced in the metadata files of the produced datasets.

**Policy S1.5:** Detailed user and developer documentation describing installation, configuration, execution, and expected outputs must be provided within each software repository.

**Policy S1.6:** In addition to source code distribution, each tagged version must be associated with a container image allowing reprocessing of the data in a controlled and reproducible environment.

**Policy S1.7:** Each analysis must provide a clear and reproducible description of the workflow used to produce the results, including the sequence of processing steps and how to execute them. This description must be sufficiently detailed to allow an independent user to rerun the analysis. Workflow management systems are encouraged (e.g. Snakemake).

## 3.2 COMMON SOFTWARE STRATEGY FOR THE VGA CAMPAIGN

**Policy S2.1:** Each sub-system of the VGA campaign (VAMOS, GRIT, and AGATA) maintains its own dedicated software suites for detector control, calibration, simulation, and physics analysis. While these software stacks may differ in implementation and internal workflows, they must follow a common software and metadata policy as defined in this document.

**Policy S2.2:** A reference software repository for the VGA campaign shall be established and maintained on the official version control infrastructure of the VGA campaign. The reference repository provides a common software framework shared across the VAMOS, GRIT, and AGATA sub-systems,

including shared utilities, metadata templates, workflow examples, and documentation of agreed software interfaces.

**Policy S2.3:** As part of the reference VGA software repository, a base analysis package shall be provided and maintained. It includes common tools for combined VGA analyses, in particular an event-merger producing unified datasets in which data from VAMOS, GRIT, and AGATA are correlated in time on an event-by-event basis. This package may be forked by users as a starting point for their own analysis workflows.

## 3.3 METADATA PRODUCTION

**Policy S3.1:** Software used within the VGA campaign must automatically generate a metadata file for each data file produced. These metadata are an integral part of the dataset and are required to ensure traceability, reproducibility, and long-term reuse of the data.

**Policy S3.2:** Metadata files must be machine-readable and contain sufficient information to unambiguously reproduce the associated data product. This includes the identification of the software components used, their versions, the workflow or execution context, the full set of input data and configuration files, as well as the command-line parameters or equivalent execution instructions.

**Policy S3.3:** Each metadata file shall follow a common structure agreed upon at the campaign level and maintained by the Data Manager and the Analysis Officer. Metadata files must include both campaign-level information describing the dataset and software-level information describing the analysis or simulation process that produced the data.

**Policy S3.4:** Metadata files must be versioned consistently with the associated software and datasets and preserved together with the associated data products. Any change to the data that impacts its provenance or reproducibility must result in a corresponding update of the metadata.

## 3.4 SOFTWARE PUBLICATION

**Policy S4.1:** Software used to produce results published in scientific articles must be archived in a long-term preservation repository at the time of publication and appropriately cited in the corresponding research output.

**Policy S4.2:** The preferred preservation repository for software developed within the VGA campaign is the ESCAPE Open-Source Software Repository (OSSR), or any equivalent repository approved by the campaign management.

# 4 Organisation

## 4.1 DATA MANAGER

The Data Manager ensures compliance with the VGA collaboration's data management policies, overseeing the entire lifecycle of datasets, access control, best practices, and policy updates. This role requires close coordination with the AGATA, GRIT, and VAMOS data managers, as well as the Analysis Officer, Dataset Analysis Spokespersons, and Campaign Steering Committee.

The Data Manager is responsible to ensure that the data management policies described in the present document are followed by the different stake-holders of the collaboration.

**Policy O1.1:** Appointment : The Data Manager is appointed by the Campaign Steering Committee to ensure alignment with the collaboration's scientific and operational objectives.

**Policy O1.2:** The Data Manager is responsible for:

- Coordinating the curation of all datasets, ensuring consistency in metadata, formatting, and documentation.
- Supervising the full lifecycle of datasets, from collection and processing to long-term storage and archiving.
- Collaborating with AGATA, GRIT, and VAMOS data managers to harmonize standards and resolve interdependencies.

**Policy O1.3:** The Data Manager must:

- Ensure the availability and reliability of data storage infrastructure, monitoring capacity and performance.
- Enforce access authorization procedures, including:
  - Granting or revoking access based on roles and project needs.
  - Implementing role-based permissions to protect sensitive data.
  - Conducting regular audits of access logs to ensure compliance with security protocols.

**Policy O1.4:** The Data Manager is responsible for:

- Updating data management policies in coordination with the Analysis Officer.
- Submitting updates for validation to the Management Board and Steering Committee.
- Co-organizing the annual analysis workshop, including logistics and post-workshop reporting.

**Policy O1.5:** The Data Manager must:

- Compile and distribute an annual report summarizing:
  - A complete inventory of datasets, including names, IDs, sizes, and locations.
  - An assessment of data quality, including completeness, accuracy, and metadata.
  - In Coordination with the Analysis officer, The status of data analysis, such as ongoing projects, publications, and pending tasks.
  - Access statistics, including usage metrics and user feedback.
- Submit the report to the Campaign Steering Committee for validation before distribution.

## 4.2 ANALYSIS OFFICER

The Analysis Officer is a central leadership role within the VGA collaboration, responsible for ensuring the application of the analysis policies outlined in this document. As the primary coordinator for data

analysis activities, the Analysis Officer oversees technical and organizational alignment, promotes best practices in software and methodology, and facilitates collaboration among all stakeholders.

The Analysis Officer acts as a bridge between the Campaign Steering Committee, the analysis work-package leaders of AGATA, GRIT, and VAMOS, and the Dataset Analysis Spokespersons, ensuring that scientific objectives and operational excellence are met.

**Policy O2.1:** Appointment : The Analysis Officer is appointed by the Campaign Steering Committee to ensure alignment with the collaboration's scientific goals and operational standards.

**Policy O2.2:** Coordination of Analysis Activities : The Analysis Officer is responsible for:

- Coordinating the analysis of datasets across the collaboration, ensuring consistency and quality in results.
- Facilitating collaboration with the analysis work-package leaders of AGATA, GRIT, and VAMOS to align methodologies and tools.
- Monitoring progress on analysis tasks, including timelines, resource allocation, and deliverables.
- Resolving technical or organizational bottlenecks that may arise during analysis activities.

**Policy O2.3:**

- Promoting best practices, such as FAIR principles, version control, and reproducibility.
- Proposing standardized workflows for data processing and analysis.
- Organizing training sessions to educate collaboration members on these practices.

Promotion of Best Practices The Analysis Officer advocates for and implements best practices in software, tools, and methodologies. He provides collaboration members with the necessary training, documentation, and resources to adhere to this policy.

**Policy O2.4:** The Analysis Officer is responsible for

- Updating data management policies in coordination with the Analysis Officer.
- Submitting updates for validation to the Management Board and Steering Committee.
- Co-organizing the annual analysis workshop, including logistics and post-workshop reporting.

## 4.3 DATASET ANALYSIS SPOKESPERSON

The Dataset Analysis Spokesperson is a designated role within the VGA collaboration, responsible for the stewardship, analysis, and publication of a specific dataset. This role ensures that datasets are analyzed rigorously, results are published promptly, and collaboration policies are followed. The Spokesperson acts as the primary point of contact for all matters related to their assigned dataset, coordinating with the Analysis Officer, Data Manager, and other stakeholders to meet scientific and operational objectives.

**Policy O3.1:** Appointment and Accountability

- Each dataset is attributed to a Dataset Analysis Spokesperson.
- By default, the spokesperson of an experiment serves as the Dataset Analysis Spokesperson, provided they comply with the policies outlined in this document.
- If the default spokesperson is found in breach of their responsibilities (e.g., failure to deliver results, non-compliance with policies), the Campaign Steering Committee may appoint a new spokesperson to ensure proper stewardship of the dataset.
- The Campaign Steering Committee will notify all relevant parties of the change through official collaboration channels.

**Policy O3.2:** Oversight of Analysis and Publication

The Dataset Analysis Spokesperson is responsible for:

- Overseeing the analysis of the dataset, ensuring methodological rigor and adherence to collaboration standards.
- Coordinating publication efforts, including:
  - Drafting manuscripts, figures, and supplementary materials.
  - Ensuring timely submission to journals or conferences.
  - Managing author lists and acknowledgments in accordance with collaboration authorship policies.

**Policy O3.3:** Delivery of Datasets and Analysis Environments

The Dataset Analysis Spokesperson must:

- Deliver L2 (processed) and L3 (final analysis-ready) datasets to the collaboration, including:
  - Processed data files (e.g., ROOT files, histograms).
  - Metadata and documentation (e.g., data format, variables, units).
- Provide associated analysis environments
- Ensure the effective implementation of collaboration policies for their dataset, including:
  - Compliance with data management policies (e.g., metadata standards, access controls).
  - Adherence to analysis best practices.

**Policy O3.4:** Reporting

The Dataset Analysis Spokesperson will report on the analysis status when requested by the Steering Committee, Data Manager or Analysis Officer

## 4.4 ANALYSIS WORKSHOP

**Policy O4.1:** The Data Manager and Analysis Officer will co-organise a yearly analysis workshop. Attendance to this workshop of at least one member of the analysis team appointed by the corresponding Data Analysis Spokesperson is mandatory.