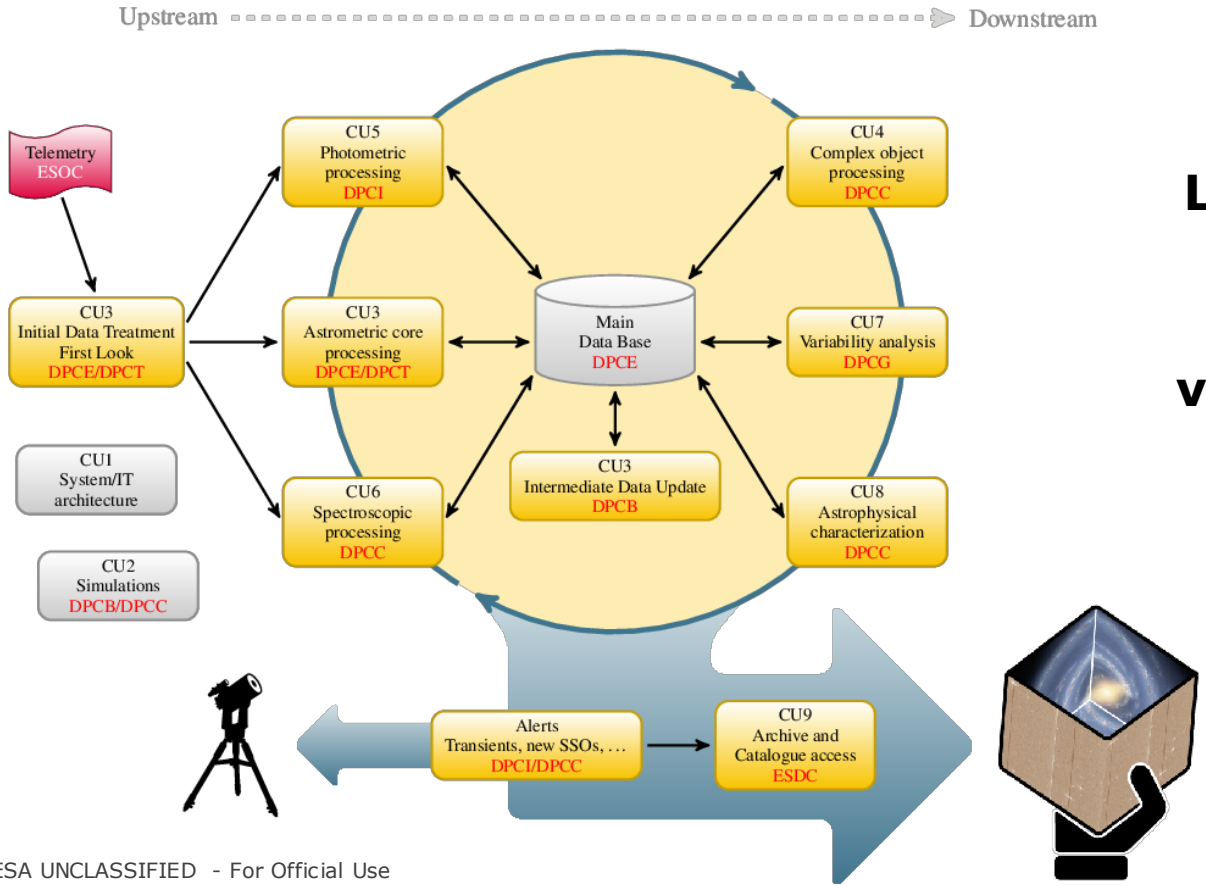
Preparation of Gaia Data Releases

Enrique Utrilla

IN2P3 meeting, 6/02/2019, ESAC

How do we get the data to start with…



Big, but not huge
LSST data is much larger

Nevertheless,
very complex processing

Interdependencies of data
Self-calibration
Time variability
…

Over 400 people in the DPAC consortium

Data Volume in Gaia Releases

Release	Data Volume
DR1	192 GB
DR2	581 GB (main tables) 241 GB (crossmatches)
DR3	Estimated a few tens of TB
DR4	Estimated a few hundreds of TB

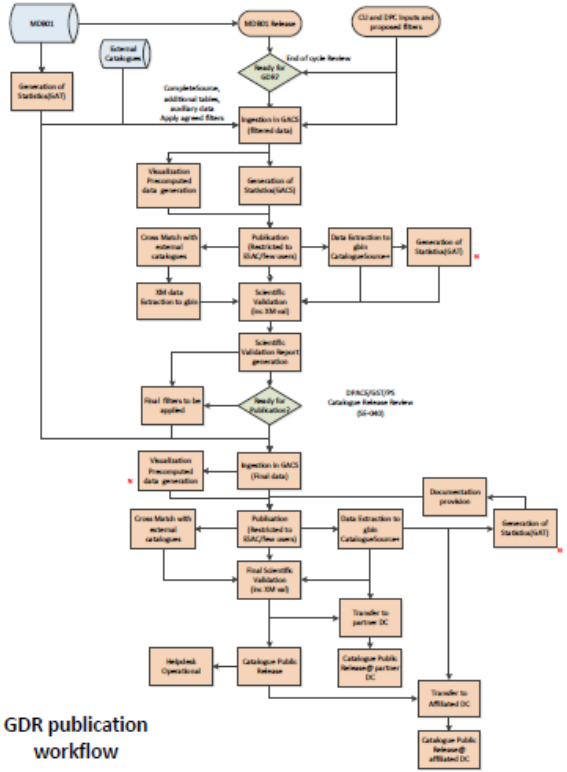
Figures for Gaia-specific compressed binary format: GBIN

What does a release involve?

- Functional aspects
  - What tasks have to be performed
- Integration and testing
  - Make sure that we do those tasks properly
- HW/SW environments
  - Where can we perform those tasks
- Configuration control
  - For both Software and Data
- Documentation & tools
  - Making data public only when intended
- The actual release
  - Making data public only when intended

Functional aspects

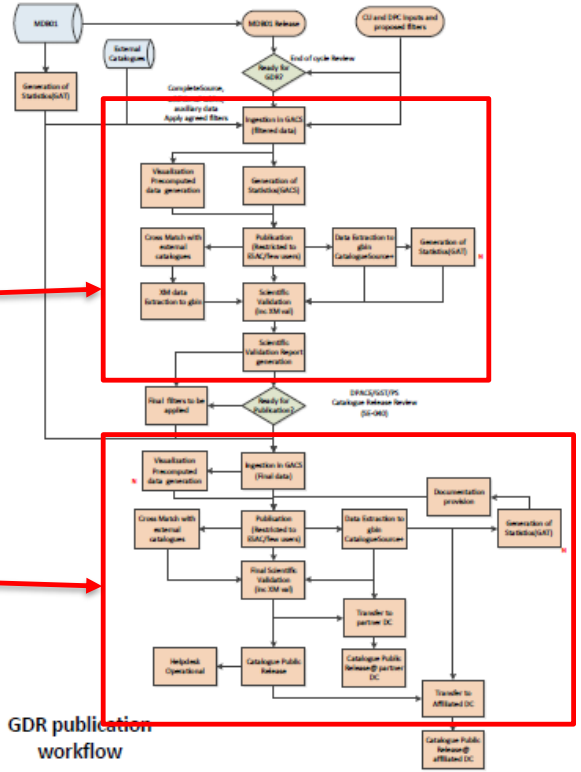
- Data curation & conversion
  - Filtering out data that doesn't meet quality requirements
  - Unit conversion, flattening, calculation of derived quantities, …
- Generation of precomputed crossmatches with other catalogues
  - 7 in DR1, 11 in DR2
  - AllWISE, GSC 2.3, PPMXL, SDSS, 2MASS, …
- Database ingestion in Archive (GACS)
- Data validation
  - Consistency checks
  - Extraction of statistics, comparison with other catalogues



GDR publication workflow

Functional aspects

- Final release requires going through this process at least twice
  - First run:
    - Preparation of preliminary catalogue using predefined quality filters (e.g. Number of Observations > 5)
    - Find specific outliers and problematic areas
    - Assessment and proposal of additional filters
  - Second run:
    - Preparation of release candidate catalogue using revised quality filters
    - Revalidation, check that problems are gone
    - Go/no go decision for release



GDR publication workflow

Integration and testing

- Mutual dependency:
  - Our software validates the data from the rest of DPAC
  - But must use the DPAC data to be validated itself!
- For each data release we have had to repeat the whole process 5-6 times
  - At least 2 iterations for the final release
  - Several more times with preliminary data for software testing
    - as data becomes available,
    - and time allows
- CU9 is at the very end of the DPAC chain
  - Usually no time to go back, data must be fixed here.
  - Not much time to react, the earlier problems are found, the better

HW/SW Environments

- Data conversion & Validation use dedicated hardware
  - Usually not needed to have separate environments (except for configuration control)
  - Different tests use different resources/inputs: gbin files, DB queries, …
  - Increase in the volume of data to process is forcing us to move towards other High Performance processing environments:
    - Grid computing already used in DR2,
    - Spark cluster is being evaluated for DR3
- GACS is used also as a tool for validation and data inspection
  - Need to have at the very least **three (or four)** HW environments
    - GACS development
    - Integration, for full scale testing and internal data validation (≅ Operational)
    - Operational, to keep on servicing previous releases. Duplicated for redundancy.

Configuration control

- Several teams involved, sequences of steps repeated a number of times…

- Configuration control and traceability is a must:
  - Datamodel
  - Inputs from Gaia Main DB
  - Processing software
  - Intermediate data products
  - Publishable outputs (in different formats)
  - Release documentation
  - Auxiliary data
  - …

Documentation and tools

- Each data release must have a set of documentation for the users:
  - How the catalogue was produced
  - Which limitations, caveats and known problems it has
  - Description of the published data model
  - How to use the data (best practices, recipes, …)

- Additionally, there might be different tools, each with its own help pages / docs
  - GACS
  - Visualization
  - TOPCAT
  - GaiaDataLibs (library for partner data centres to read gbins)
  - …

The actual release

- Constraints:
  - Data must be visible for general users from a set date and time, and not before
    - GACS / TAP
    - Bulk download repository
  - Bulk data must be accessible for some Partner Data Centres some time in advance
  - Bulk data must be accessible for Affiliated Data Centres only from the public release
  - Data from previous releases must be accessible as much as possible during the preparation of the release

The actual release

- Once the dataset has been curated and validated (Integration environment):
    1. A ‘public’ dataset is generated by removing some columns for internal use only from each table
    2. This dataset is generated in several formats (GBIN, compressed csv, …)
    3. A checksum is calculated in each column of the public datasets to check data integrity
    4. These public datasets are put under configuration control
    5. The gbin dataset is published in a server for Partner Data Centres to download
    6. The compressed csv format is uploaded to a commercial CDN (but not released)
    7. The Operational GACS server is taken offline for the general public, so it is only accessible from ESAC, in order to ingest the public data of the new release.
    8. The new documentation sets, visualization tools, GACS, etc. are uploaded to the operational server

The actual release

- At a predefined hour, everything is released. This involves:
  - Changes in the network configuration at ESAC to put back online the OPS server
  - Changing permissions in the local content server so that Affiliated Data Centres can download the bulk data
  - Open up the CDN for the bulk download for the general public
  - Coordination with other Partners so that they perform their own releases

- Things to monitor:
  - Number of accesses to the DB, documentation, downloads, etc.
  - DB, network, HW statistics: expect high volume of traffic

- Be ready for hiccups and remember Murphy's law

Any questions?