Gaia DPAC Project Office, Coordinating the Production of the Largest Stars Catalogue

Gracia-Abril, G.,[1,2] Teyssier,D.,[1,3] Portell,J.,[1,4] Brown,A.,[5] Vallenari,A.,[6] Jansen,F.,[7] and Lammers, U.[7]

[1]*Gaia DPAC Project Office, ESAC, Villanueva de la Cañada, Madrid, Spain*

[2]*Astronomisches Rechen-Institut, Zentrum für Astronomie, University of Heidelberg, Heidelberg, Baden-Wurtemberg, Germany*

[3]*Telespazio Vega UK Ltd for ESA / ESAC, Villanueva de la Cañada, Madrid, Spain*

[4]*Institut de Ciencies del Cosmos, Universitat de Barcelona (ICCUB/IEEC-UB), Barcelona, Spain*

[5]*Leiden Observatory, Leiden University, Leiden, The Netherlands*

[6]*INAF - Osservatorio astronomico di Padova, Padova, Italy*

[7]*Mission Operations Division, Operations Department, Directorate of Science,European Space Research and Technology Centre (ESTEC/ESA), Leiden, The Netherlands*

[8]*European Space Astronomy Centre (ESA/ESAC), Villanueva de la Cañada, Madrid, Spain*

Abstract. The European Space Agency mission Gaia is creating the most complete and accurate map of the Milky Way. It was launched in December 2013 to its orbit around the Lagrange point 2. It is now approaching the end of its nominal mission at the end of July 2019, although mission extension beyond that date has received an initial approval. Gaia final product is the Archive (`http://gea.esac.esa.int/archive/`), accessible from its central hub at ESAC (ESA) and many other data centres around the world. Two Data Releases have been done so far. The number of scientific papers based on Gaia Data Release 2 since it was published on April 25th, 2018, more than 400 papers, gives an idea of the impact Gaia is having in the astronomy community.

The DPAC (Data Processing and Analysis Consortium) is in charge of processing the raw data received daily from the spacecraft to produce data products usable by the astronomers. DPAC is a multinational consortium, with contributions from over 20 countries and near 100 institutions, mainly European. The DPAC Project Office (PO), together with the DPAC Executive, is responsible for the coordination of the consortium. Created in 2009 the PO is a key point to maintain the consortium tight and focus in the common goal of creating the best possible Gaia archive. A well balanced composition of the PO, including management, scientific and engineering expertise, is fundamental to accomplish its role in a very complex structure as DPAC, where common Industry management practices are not fully applicable and speaking a common language to the scientists and engineers is critical

1. Gaia and the Data Analysis and Processing Consortium (DPAC)

Gaia is continuously scanning the sky taking one dimensional images of all the stars crossing the field of view of the two Gaia telescopes. Gaia data, collected by its three instruments, astrometric field, photometer and spectrometer, are downloaded to Earth daily and processed immediately. The main goal of the daily processing is to assess the health of the payload and detect any issue affecting the quality of the scientific data and react quickly to minimize any data loss. Besides the satellite monitoring, the daily pipelines generate some initial calibrations and two dedicated pipelines raise alerts on events which may require immediate follow up from ground, photometric alerts (e.g. possible new supernovae) and detections of unknown asteroids. In both cases these alert are published and available to the community on line. The intermediate daily data products generated by the daily systems are stored in the Main Data Base.

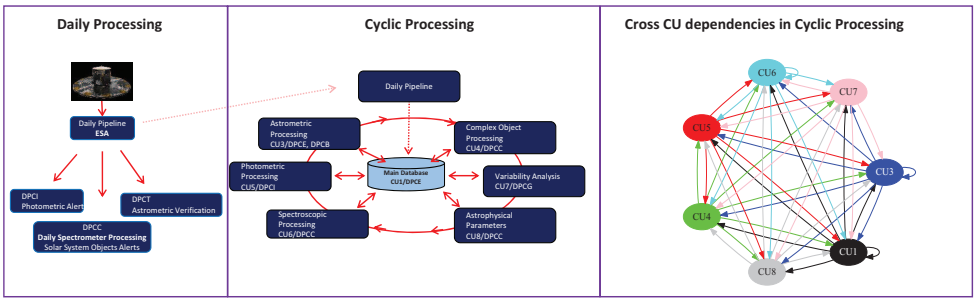

Figure 1. Schematic view of Gaia data processing flow

Data reduction after the daily processing is iterative: in each processing cycle the data generated in the previous cycle, and new data from the daily pipeline until a given date, is consistently processed by all pipelines. This consistency is required because of the dependencies between the different subsystems: photometry and spectroscopic processing needs astrometric results and downstream systems need the data from astrometry, photometry and spectroscopy processing. Likewise, these three upstream CUs need some results from the downstream systems which are only available to them in the following processing cycle.

2. The Data Processing and Analysis Consortium, DPAC

The Data Processing and Analysis Consortium (DPAC) is in charge of the Gaia data reduction and the production of the Gaia Data Releases. DPAC is a international consortium with more than 400 members distributed in 83 institutes in 21 countries plus the European Space Agency (ESA). Processing software development and the scientific validation of the data produced is the responsibility of 9 *Coordination Units (the CUs)*. Each CU, is in charge of a specific aspect of the data processing. Each CU is composed of several dozen members, spread around various (academic) institutes in various, mostly European countries.

The actual data processing is done in 6 data processing centres (DPCs) located in ESAC (DPCE, Madrid), BSC Barcelona (DPCB), CNES Toulouse (DPCC), Institute

of Astronomy (IoA) in Cambridge (DPCI), Geneva (DPCG), and Turin (DPCT). Each DPC receives and operates the software modules from one, or several, CUs, in such a way that each processing system only runs in one of the DPCs. At the end of each processing cycle data products generated in every DPC is transferred to the Main Data Base (MDB) located at DPCE. All the data types are combined and redistributed back to all DPCs to serve as input for the next processing cycle.

CUs basic description and their relation with the DPCs is provided in table 2

Table 1. Coordination Units and Data Processing Centres in DPAC

Coordination Unit	Task	Procesing Centre	Main countries
CU1-Architecture	Overall architecture, Main Data Base, common tools, data transfers, etc.	DPCE	ES, FR, IT, UK, ESA
CU2-Simulations	Generation of simulated data (mainly active before launch)	DPCB	ES, FR
CU3-Core Processing	Astrometry	DPCE, DPCB, DPCT	DE, ES, SW, IT, USA, ESA
CU4-Object Processing	Multiple systems, extended objects, solar system objects	DPCC	BE, FR, IT, PT, GR, FI, CH, ESA
CU5-Photometry	Fluxes and magnitudes in the three bands and photometer spectra	DPCI	UK, ES, IT, NL, PL
CU6-Spectroscopy	Radial velocities and spectrometer processing	DPCC	FR, UK, BE, DE, SL
CU7-Variability	Variable sources classification and characterisation	DPCG	CH, IT, IL, PT, AT, UK, ESA
CU8-Astrophysical Parameters	Object classification and determination of physical properties	DPCC	DE, FR, IT, BE, ES, SW, GR
CU9 Validation and archive	Final cross CU validation and archive publication	ESAC	ES, FR, UK, IT, PT, DE, NL, BE, SW, FI, AT, ESA

3. Gaia-DPAC Project Office

The DPAC Gaia project office was created in 2009. Until the Gaia launch, the PO worked was strongly linked with CU1, the coordination unit in charge of designing the general architecture of the consortium. The CU1 leader was also member of the PO. Among other tasks some of the activities the PO did before launch were:

- Definition of the cross CU interfaces writing the the DPAC Operations Interface Control Document.
- Coordination of the development of the different sub-systems which are part of the daily pipeline. Those systems had to be ready from the very first moment after launch and its development to be able to process the data during commissioning.
- Preparation of operations rehearsals, tests aimed to simulate daily normal operation after launch in realistic conditions, and end to end tests of the cyclic processing pipelines.

- Preparation, together with the CUs and DPCs, of the documentations required for the ESA reviews
- Cross CU coordination of QA activities and DPAC level risk management

During the nominal mission the PO activities have evolved to focus on four main types of activities:

- Coordination of DPAC operations: creation and maintenance of the operational schedule. The operational flow must take into account all the dependencies between the processing pipelines, including deadlines for deliveries between the different groups. Deliveries of preliminary validation data are also needed in the development and tests phases. The top level DPAC priorities must always be taken into account when planning the individual DPCs operations.
- Technical support to the CUs and DPCs (together with CU1: provide technical support and expertise to the different groups and advise in the use of common tools. The PO also fosters technical synergies between the groups to avoid, when possible, duplication of work trying to apply similar solutions to equivalent problems.
- Scientific interfaces:The dependencies between processing pipelines, with all systems using data produced by upstream pipelines require a good understanding of data caveats and clear flagging of outliers and special cases. DPAC is processing the data, but also in continuous development to improve the results and reduce the systematic errors. Data producers are very well aware of the remaining weaknesses in their software but data consumers tend to assume they will get data of a given quality. Despite the technical interface is defined communicating concrete data issues between the units is critical.
- Data release preparation: the publication of the data in the Gaia Data Releases require new communications channels. The data in the MDB is optimised for internal use, while the data model used in the archive should serve a different purpose, usability by the end users of the archive avoiding internal slang. To go from the MDB to the archive requires data conversions and data filtering. Both, conversions and filters, must be communicated clearly to CU9 by the rest of DPAC. Thanks to its overall knowledge of the operations and the data produced, the PO has played a very active role in the definition.

The composition of the PO has evolved since it was created, it is necessary to have a well balanced mixture of technical, management and coordination, and scientific expertise. All members of the PO have at least experience in two of these aspects, either scientist with coordination or management experience, or engineers with experience in science. This mixed profiles is critical, communication is the key word in all the PO activities. The PO does not have any real executive power and its tasks can only be successful if the different groups have developed enough confidence in the PO, this is not easy and can only be achieved if the PO members are able to speak the language of the scientist and engineers bringing some added value which at the end contributes to the general DPAC goal.