Highlights of Spanish Astrophysics VIII, Proceedings of the XI Scientific Meeting of the Spanish Astronomical Society held on September 8–12, 2014, in Teruel, Spain. A. J. Cenarro, F. Figueras, C. Hernández-Monteagudo, J. Trujillo Bueno, and L. Valdivielso (eds.)

Daily processing of Gaia data

Jordi Portell[1], Claus Fabricius[1], Jordi Torra[1], Nora Garralda[1], Juanjo González[1] and Javier Castañeda[1]

[1] Dept. d'Astronomia i Meteorologia, Institut de Ciències del Cosmos, Universitat de Barcelona (IEEC-UB), c/Martí Franquès 1, E08028 Barcelona, Spain.

Abstract

Gaia, the space astrometry satellite of the European Space Agency, was successfully launched on 2013 December 19^{th}. A vast amount of data has already been received by the on-ground data processing systems running at the European Space Astronomy Centre near Madrid. In this paper we describe the mentioned systems, focusing on the Initial Data Treatment system which is coordinated by the Gaia group at the University of Barcelona. We present some of the results obtained during the first months of nominal operations.

1 Introduction

Gaia (4) will measure the positions, distances, motions and many physical characteristics of more than one billion stars in our Galaxy and beyond. The satellite is orbiting around the Sun-Earth L2 point at 1.5 million km from the Earth. During at least 5 years, Gaia will be spinning once every 6 hours, with its axis at 45° with respect to the Sun and precessing once every 70 days. This scanning law will allow observing the complete sky several times during the mission, with an average of about 80 times per star.

The payload is composed of an optical bench mounted on a highly stable SiC torus. Two telescopes separated by a *basic angle* of 106.5° project their images onto a big focal plane composed of 106 charge coupled devices (CCD), each with 8.8 megapixels (1). Part of the images are passed through a set of prisms in order to get spectroscopy and high-resolution photometry. The CCD cameras are arranged in seven rows, each controlled by a Video Processing Unit (VPU) that automatically detects and observes the stars entering the field of view. It allows acquiring only small *windows* of typically 12×12 pixels around each star, thus largely reducing the amount of data to be downlinked. Despite of this, about 7 Mbps are generated by the instruments on average, so the ground segment receives some 25 GB of compressed data every day — leading to about 60 GB with astrometry,

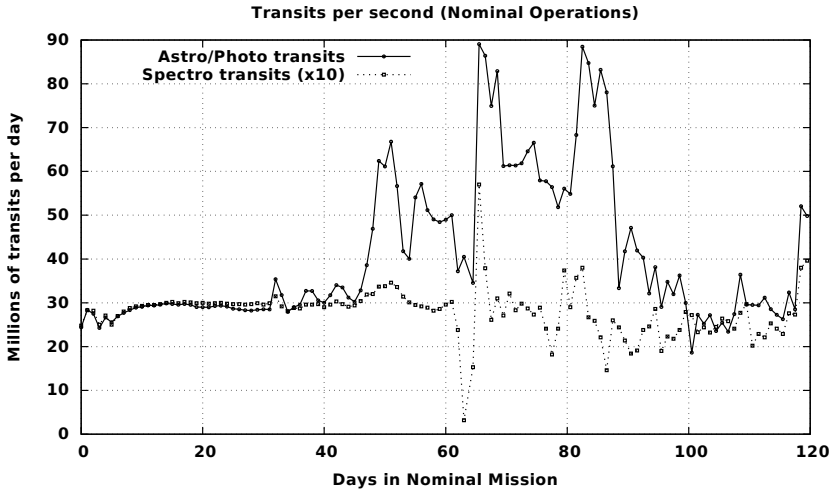

Figure 1: Daily transits received from Gaia during the first months of nominal operations. Variations are caused by changes in the scanning law and by the star densities being observed.

photometry and spectroscopy data. The ground stations available to Gaia are Cebreros (Spain), New Norcia (Australia) and Malargüe (Argentina). Data is then transferred to the Mission Operations Centre in ESOC (Darmstadt, Germany), and relayed to the Science Operations Centre in ESAC (near Madrid, Spain). The number of stars observed every day is not uniform throughout the mission, as shown in Fig. 1. In some cases it can significantly increase, such as when scanning the Galactic Plane almost tangentially — which will happen a few times during the mission. When needed, more than one ground station may be used to increase the downlink capability. In any case, data is downloaded following a specific priority scheme that depends on the features of each measurement, such as the brightness of the star. This scheme is designed in a way that enforces completeness for the most interesting data, and allows some on-board deletion of "bonus" data such as observations of the faintest stars.

2 Daily data processing needs

The daily data volume may not look very high for nowadays computing standards, but the complexity lies in the millions of star transits contained in such data. As seen in Fig. 1, some 50 million of transits are received every day, with peaks that can surpass 100 million transits. Each transit contains 10 astrometric measurements, 2 spectro-photometric measurements (for the red and blue bands), and a significant fraction also include 3 spectroscopic measurements. An adequate data processing architecture had to be designed to handle this challenging and precious information (3; 2). We may have opted for just accumulating all the data and process it at once towards the end of the mission, but that would have led to prohibitive data processing requirements for such a one-go processing to get the results in a reasonable amount of time. Most important, we must carefully and promptly examine the data generated by Gaia to diagnose the health of all its systems and instruments. Finally, progressively processing

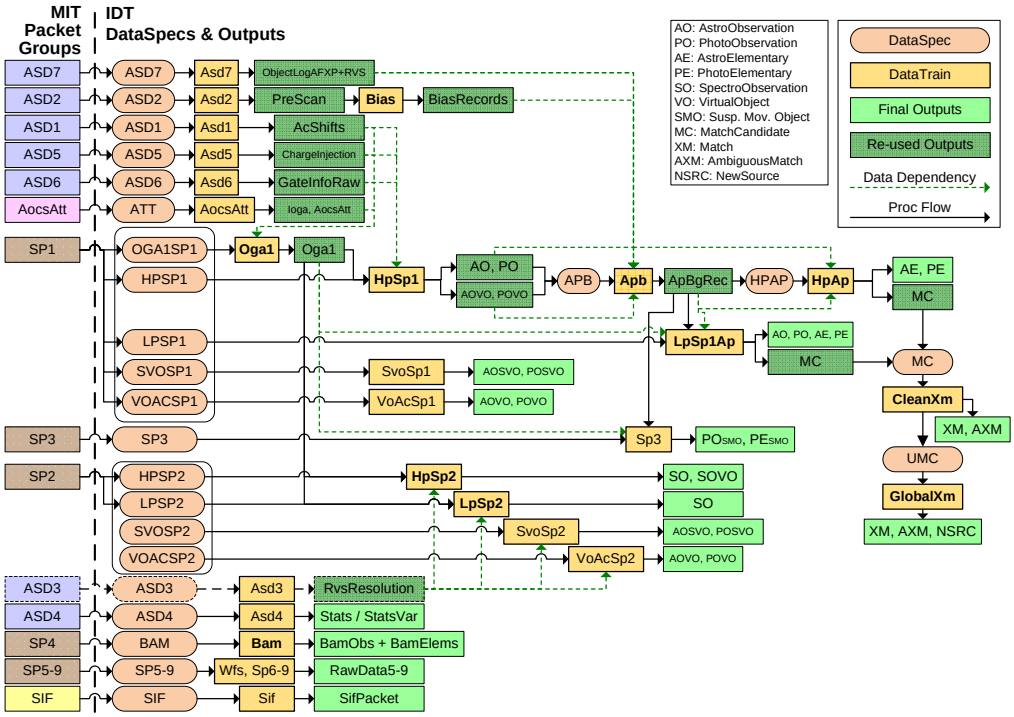

Figure 2: Overview of IDT, with its main processing modules (or Data Trains), flow control elements (or Data Specs), inputs (MIT Packet Groups), outputs and dependencies.

the data as it arrives allows to get early scientific results, including preliminary catalogue releases well before to the end of operations.

3 Initial Data Treatment

When data arrives at the Gaia SOC it is first reconstructed by composing measurement packets from the raw telemetry packets, decompressed and backed up. Then it enters the first of the scientific data processing systems, which is called IDT (Initial Data Treatment). Fig. 2 illustrates the overall operation of IDT, which has to detect and identify the incoming data and trigger the associated processing module — which, in turn, can trigger other modules. All this turns into data processing *jobs*, each processing a portion of data of a given type. A *whiteboard* helps managing this, where jobs are published by an IDT Coordinator and picked up by IDT *workers*. So far, 8 computing nodes (each with 32 processor cores) have proven to be more than enough — able to process about 3 times faster than the incoming rate. Storage is handled by Intersystems Caché®, a fast database engine. It is worth noting that we have followed the *data train* approach (3), that is, we first prepare each piece of data and then we pass it to the appropriate algorithms. It has proven to be much more efficient than the "classical" approach, that is, letting each algorithm pick up the necessary data from the database.

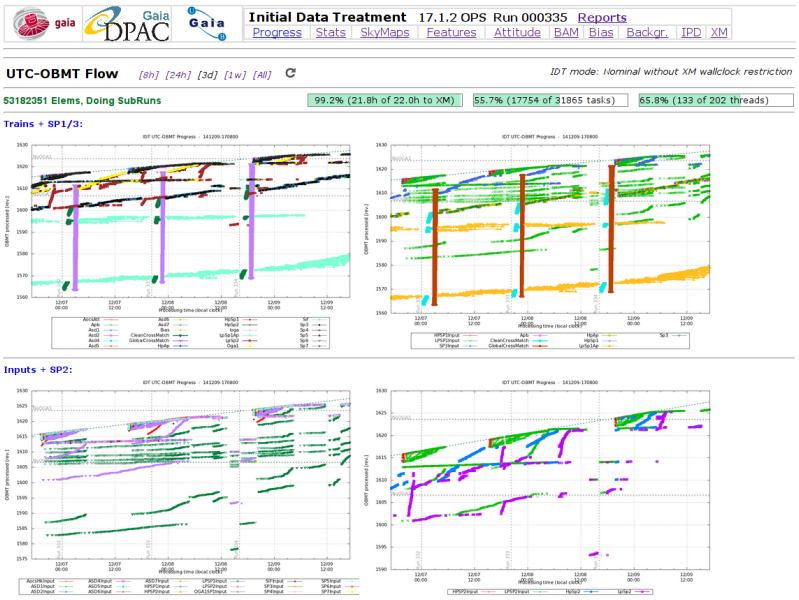

Figure 3: Web-based near-realtime monitoring and diagnostics for IDT.

The most important operations executed by IDT are the following. Raw observations are reconstructed by combining raw science packets (SP) with ancillary science data (ASD) and instrument configuration data, leading to self-contained *observation* records for astrometric, photometric and spectroscopic measurements. Science packets from the basic angle monitoring device (BAM), allow determining variations of just a few microarcseconds (μas) in such angle. IDT also refines the raw attitude quaternions provided by Gaia by combining them with the brightest astrometric observations, cross-matching them against a reference catalogue, and finally applying a Kalman filter on the field angle differences, leading to the first on-ground attitude results (OGA1). Electronic bias of the image samples is determined from some ASD packets. The on-board VPUs acquire some *virtual objects* (or empty windows) which are then used by IDT to determine the astrophysical background (APB). The instrumental background is determined as well, including the memory effects caused by charge transfer inefficiency in the CCDs.

The core modules of IDT are related to astro-photometric data (SP1). A preliminary centroid is determined for each of the astrometric images using a Tukey Biweight algorithm, also estimating the star flux. OGA1 allows propagating the centroids to the photometric instrument to estimate the location of a reference wavelength on the dispersed image. Multi-band photometric features of the star are then determined, namely, blue, red and spectroscopic integrated fluxes, 8 narrow bands, and an effective wavelength which is correlated with the star temperature. We then retrieve the most appropriate model of the point spread function (PSF) for that star, which depends on such photometric features. Finally, we fit the PSF of the star on the raw samples by means of a maximum likelihood algorithm. The image parameters obtained include more accurate centroids and star fluxes, goodness-of-fit values, formal errors and ancillary flags. Finally, for each transit IDT finds all the possible matching

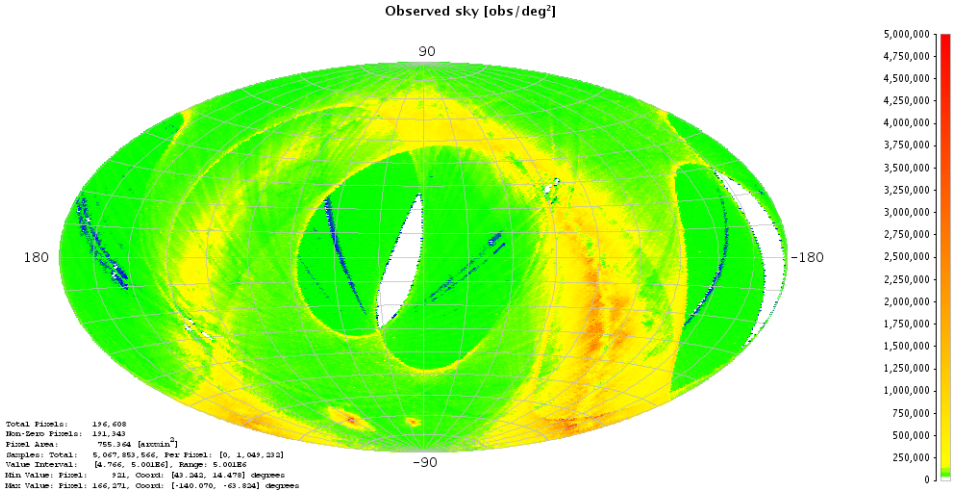

Figure 4: Sky map (equatorial coordinates) with the observations density after 3 months.

candidates available in a star catalogue (initialised from ground-based catalogues and later updated with our own outputs), by means of a nearest-neighbour algorithm with a search radius of 2 arcseconds. Candidates are consolidated by removing inconsistent matches, such as transits close in time linked to the same star. We also disregard invalid transits, such as spurious detections onboard along the PSF wings of some bright stars. The transits without any catalogue candidate are consolidated, leading to new entries in the star catalogue.

As we can see, IDT must perform quite a lot of complex operations on a large number of measurements every day. An adequate monitoring and diagnostic system is necessary to assess its correct operation, both technically and scientifically. We have implemented a web-based system, illustrated in Fig. 3, which allows browsing most of the near-realtime diagnostics. Lower left panel shows data entering IDT, with colors telling the data types and axes telling the mission time versus wallclock time. Other panels show processes launched, the processing progress or resources occupation. Detailed diagnostics are accessible through links on the top of the page. PDF documents are also automatically generated on a daily basis with most of these diagnostics, thus providing a very rich scientific and technical feedback.

4 Results obtained on real data

Fig. 4 illustrates the sky areas that have already been observed by Gaia, received and processed by IDT, and transferred to the rest of data processing systems and centres. We can easily see the higher star densities along the Galactic Plane, as well as some effects caused by the scanning law of the satellite. These will obviously disappear in the final catalogue. Note that almost all the sky has already been observed in just 100 days.

Table 1 summarizes some of the most relevant results obtained by IDT on real Gaia data during these first months of nominal operations. These excellent accuracies will allow a smooth ramp up of the rest of data processing systems throughout DPAC, the Data Processing

and Analysis Consortium built for Gaia (3). The technical numbers are also excellent: in the first 4 months of nominal mission, IDT has processed over 6 billion star transits without major problems, meaning an average of about 46 million transits per day. More than 3.6 TB of raw data (26 GB/day) have been received from the satellite, whereas the main database has grown beyond 30 TB (compressed).

Table 1: Some scientific results obtained by IDT on real Gaia data.

Module	Results
Attitude (OGA1)	$\sim$ 50 mas accuracy
Basic Angle	$\sim$ 10 μas accuracy
CCD readout noise	4-5 e^- typ.
Astrophys. background	Typ. <10 e^-/pix/s
Centroid formal error	<250 μas (<15^{mag}), <6 mas (<20^{mag})
Photometry formal error	<5 mmag (<15^{mag}), <0.03 mag (<20^{mag})
Star motions (equatorial)	$\sim$1 mas/sec typ. accuracy
On-board detection coordinates	$\sim$200 mas typ. absolute error

5 Conclusions

We have presented and briefly described IDT, the core element in the daily processing of Gaia data. This system has proven to correctly process the large and complex data set received from the satellite, offering outstanding scientific results despite of being just the very first stage of the Gaia data reduction pipeline.

Acknowledgments

This work was supported by the MINECO (Spanish Ministry of Economy) – FEDER through grant AYA2012-39551-C02-01 and ESP2013-48318-C2-1-R.

References

[1] de Bruijne, J., Kohley, R., & Prusti, T. 2010, in , 77311C–77311C–15

[2] de Teodoro, P., Hutton, A., Frezouls, B., et al. 2012, in Springer Series in Astrostatistics, Vol. 2, Astrostatistics and Data Mining, ed. L. M. Sarro, L. Eyer, W. O'Mullane, & J. De Ridder (Springer New York), 107–115

[3] O'Mullane, W., Lammers, U., Bailer-Jones, C., et al. 2007, in Astronomical Society of the Pacific Conference Series, Vol. 376, Astronomical Data Analysis Software and Systems XVI, ed. R. A. Shaw, F. Hill, & D. J. Bell, 99

[4] Perryman, M. A. C., de Boer, K. S., Gilmore, G., et al. 2001, A&A, 369, 339