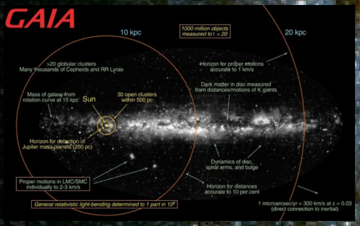
# Gaia Data Processing

## Status, Architecture, first results

Jose Hernández, John Hoar, Uwe Lammers,
Mark ter Linden, William O'Mullane, Paul Parsons,
Frederic Raison, Pilar de Teodoro
European Space Astronomy Centre
http://www.rssd.esa.int/gaia




**ABSTRACT**

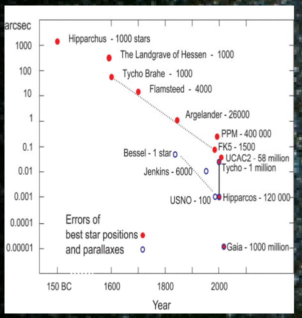
Gaia is ESA's ambitious space astrometry mission the main objective of which is to astrometrically and spectro-photometrically map $10^9$ celestial objects (mostly in our galaxy) with unprecedented accuracy. The satellite will downlink close to 100 TB of raw telemetry data over 5 years. To achieve its required accuracy of a few 10s of micro-arcsecond astrometry, intricate processing of this data is required. In addition to the main astrometric instrument Gaia will host a Radial Velocity instrument, two low-resolution dispersers for multi-colour photometry and two Star Mappers.

The basic principle of the astrometric data reduction is to determine all the variables of the system (source positions, attitude, calibration, photometry, spectroscopy...) in order to achieve an optimal match with the observations. The data reduction will be performed in several cycles, in each cycle new data from the satellite will be processed and the starting solution will be improved.

The Gaia Data Processing and Analysis Consortium (DPAC) has been formed recently and will answer ESA's announcement of opportunity for the data processing. All of Europe's expertise in astrometry will be needed in order to reduce Gaia data to the level of accuracy needed to meet the mission goals.

Keywords: Gaia, Astrometry, Survey

### Progress in astrometric accuracy




### The Challenge

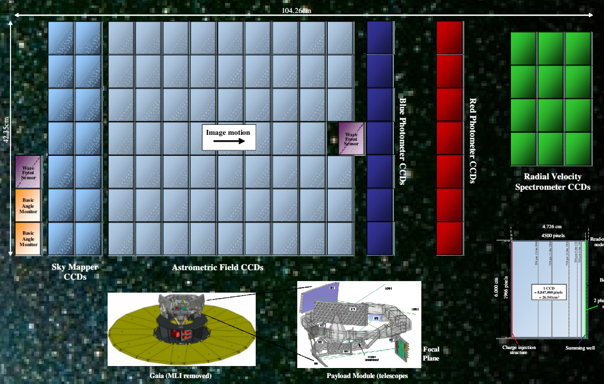
The data processing requirements for Gaia are amongst the most challenging of any scientific endeavour to date. Due to the immense volume of data that will be collected, for $10^9$ stars, it will be a major challenge, even by the standards of computational power in the next decade, to process, manage and extract the scientific results necessary to build a phase space map of our Galaxy, the Milky Way. Gaia is in some senses the astronomical equivalent of the Human Genome Project, and is a pioneering undertaking being led by ESA.

A total of some 100 Terabytes of science data will be collected during Gaia's lifetime. The total data archive may surpass one Petabyte, at current state-of-the-art disc access rates, this would require 40 days simply to read.

The required numerical processing is colossal, of order $10^{21}$ flops. Using the world's fastest computer (as of early 2006, IMB/DOE's BlueGene 136 TFlop system), would require 85 days of processor power. The size of the problem, and the requirements for rigorous testing and optimisation, can be emphasised by noting that just 1 second of CPU time devoted to any specific task, per object (e.g. variability classification), would require 30 years of CPU time for the entire data set, per task.

### Gaia Focal Plane

106 CCDs ≈ 938 million pixels ≈ 2800 cm²




### Astrometric Global Iterative Solution

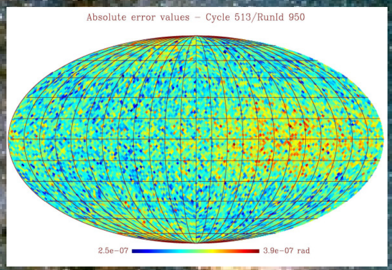
The core of the Gaia data reduction consists on the accurate determination of the source positions, spacecraft attitude and instrument calibrations. The Astrometric Global Iterative solution is the realization of the mathematical framework and algorithms developed by Lennart Lindegren (Lund Observatory).

The plausibility and feasibilty of the Global iterative solution has been a concern in the Gaia community for the past several years and has led to the development of several prototypes of increasing complexity.

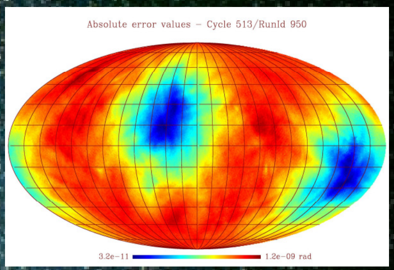
The ESAC AGIS system consists of the three astrometric blocks: Source, Attitude and Calibration. This system is data driven, each observation is read from disk precisely once per iteration. The calibration and attitude are loaded once per iteration per processor, where every attempt has been made to reduce the disk accesses due to possibility of multiple terabytes of data being handled.

The system can process five years worth of simulated observations of one million objects in under four hours. AGIS requires around 25 iterations to converge on the sub microarcsecond error limit when the initial data contains errors of the order of 100 miliarcseconds. It runs on a cluster of 18 dual-3.6GHz Xeon processors giving a overall performance of 140 GFLOPs. Up until Gaia's launch (end of 2011) the system will be scaled and exercised using up to 100 million sources and more complex data.

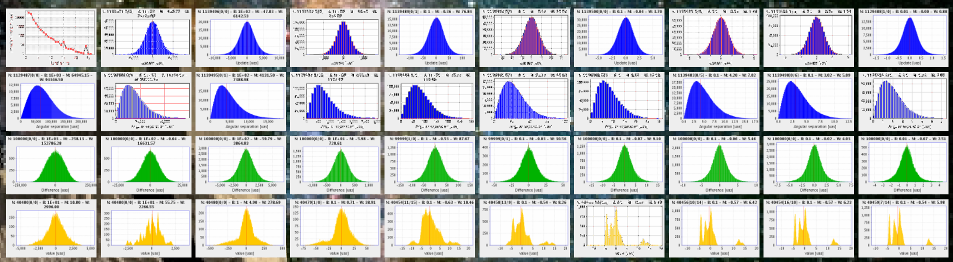
**Initial Errors In Source Positions**

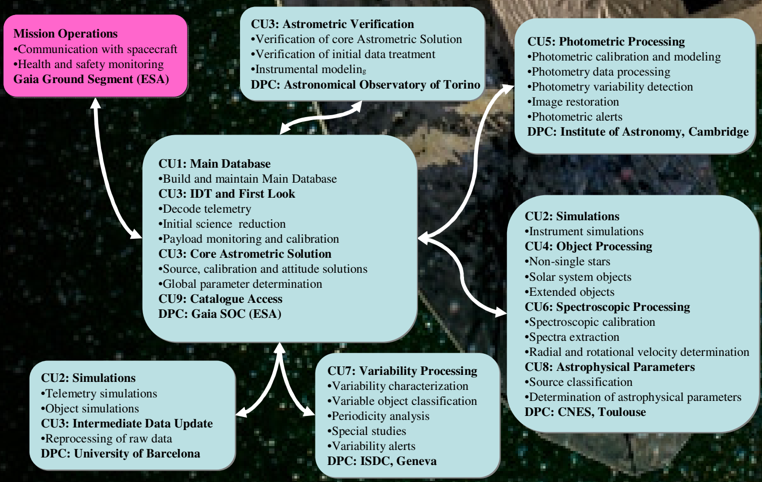



**Errors In Source Positions after AGIS convergence**




**Source parallax, Source position, Attitude and Calibration updates on successive AGIS Iterations**






### Gaia Data Processing

The AGIS system will produce the core astrometric results from the Gaia mission. The Gaia mission will also produce an enormous range of astrophysical information from the combined astrometric, photometric and spectroscopic data produced by the Gaia instruments.

The diverse range of processing tasks required to extract the true potential of the Gaia data is too great for any one organization to undertake alone. Over 250 scientists and engineers from across Europe will form the Gaia Data Processing and Analysis Consortium (DPAC) responsible for the production, validation and operation of the processing systems needed.

The DPAC is logically divided into Coordination Units (CUs) according to scientific nature of the processing required and are tasked with the production of the processing systems. The CUs are supported by a set of Data Processing Centres (DPCs) responsible for the operation of the processing systems. An executive committee oversees the activities of the DPAC.

The large scope and cutting-edge nature of many of the data reduction activities implies a high risk. In order to manage these risks an iterative, incremental development process will be used throughout the DPAC. Six-month synchronized development cycles are used to partition development activities, with frequent integration and validation of systems across the entire DPAC. We aim to ensure that unknown or poorly-specified requirements do not lead to 'late design breakage'.

The diagram on the left presents the overall distribution of relationship between the CUs and DPCs and some of the scientific and technical tasks tackled by each CU with the overall data flow indicated.

**For more information on Gaia, visit http://www.rssd.esa.int/gaia**