DESIGN AND OVERVIEW OF THE DATA PROCESSING

J. Torra[1,3], F. Figueras[1,3], C. Jordi[1,3], X. Luri[1,3], E. Masana[1], C. Fabricius[2]

[1]Universitat de Barcelona, Diagonal, 647, 08032 Barcelona, Spain
[3]Institut d'Estudis Espacials de Catalunya, Gran Capità, 2, 08034 Barcelona, Spain
[2]Copenhagen University Observatory, Juliane Maries Vej, 30, 2100 Copenhagen, Denmark

ABSTRACT

The unprecedented observing capabilities of Gaia make it a unique instrument with an impressive impact on most of the research fields of astrophysics. Both the huge amount of data and the precision imposed by the scientific goals make the data storage and data reduction a critical issue of the mission. As such, the Gaia data access and analysis system has been object of study since the first steps of the mission. In the present paper we give an overview of the task performed until now in this area, leading to a first prototype of the Gaia Database and Data Reduction system on which the most critical algorithms have been implemented and tested. An update of the system including more realistic details and more complex algorithms that is now being implemented is also described. Some insights on the future plans will complete this paper.

Key words: Gaia; Data Reduction; Global Iterative Solution.

1. INTRODUCTION

The storage and data reduction of the Gaia data is one of the challenges of the mission. This task of data treatment is made difficult by three main reasons: the first one is the large amount of data to be processed, the second one is the requested precision and the third reason is the complexity inherent to the Gaia measurements. Regarding the amount of data to be processed, Gaia will download some hundreds of TB, its manipulation and the introduction of intermediate data will easily lead to a data base of the order of a Petabyte. The precision requested, the microarcsecond, demands a model of the observations and the knowledge of the mission operations up to this level; in addition, from a numerical point of view, one microarcsecond is 10^{-12} radians, so numerical instability at this level must be avoided. The third aspect is not independent of the other two. If we want to obtain microarcsecond precision all the effects that typically are of second order, become important: astrometric observation must be corrected for chromaticity, thus requiring photometric observations; radial velocity enters in the perspective effect and a relativistic model of the observation is needed, in turn that means that accurate enough ephemerides of the Solar System bodies and of the satellite are needed. Additional complexity is added when considering that the continuous scanning of the sky results in a mixing of data in time and space, and that to obtain absolute parallaxes all observations of a 5-year mission should be treated together. In the following paragraphs we will review the design and implementation of a prototype of the Gaia data base and its algorithmic framework. We start by introducing the concept of the data reduction for Gaia (Section 2) from which we extract requirements for the system (Section 3) which is discussed in Sections 4 and 5. Finally, some insights of the future are given in Section 6.

2. GAIA DATA REDUCTION CONCEPT

The concept of the data reduction is simple. It is based on the comparison between the calculated and observed field angles giving the direction to a given star as seen from the satellite in a given instant of time. This direction can be calculated if we have a model of the observed object, the instrument geometrical and photometrical calibrations, satellite orbital and attitude data and a relativistic model for the astrometric observations (see Figure 1).

The differences between observed and calculated field angles for each individual observation *i*, that is, a transit in a given CCD producing a sample (Bastian 2004), can be explained by a linear model in terms of the corrections to a given parameter, that is:

$$f_iobs - f_ical = \sum_j \frac{\partial f_i}{\partial p_j} \Delta a_j \qquad (1)$$

where p_j stands for a given parameter we want to update, and $\frac{\partial f_i}{\partial p_j}$ is the calculated partial derivative of the field angle with respect to p_j.

The p_j parameter may be any one of the parameters entering in the description of:

- the attitude of the satellite
- the astrometric parameters of a star

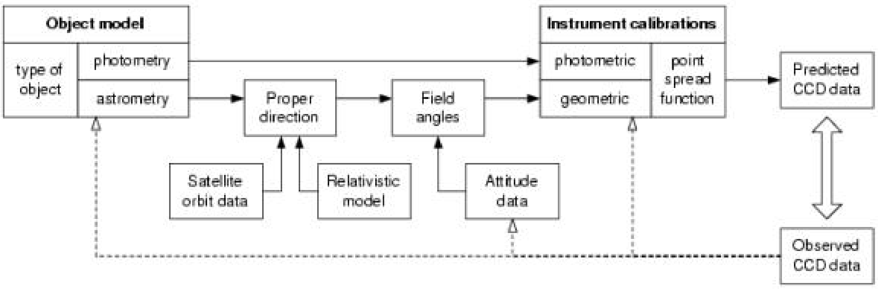

Figure 1. Gaia data reduction concept from ESA (2000).

- the geometric and photometric calibration, or
- the global parameters in the astrometric model.

In the real mission this process has to be applied to a subset of some 100 million of ‘well behaved’ stars. For them the update of the four categories of parameters listed above has to be performed in a sequential iterative process improving in each step one set of parameters. Note that calibration, science data and attitude data are simultaneously obtained. This process is called the Global Iterative Solution (GIS).

An important characteristic of the GIS is that the data needed in each one of the steps are accessed using different criteria (see Figure 2). To improve the attitude in a given interval of time all the data of all the sources observed and all the calibration data in the time slot are required, so access by time to the data base should be optimized. To update the astrometric parameter of a source, all the data of the source plus all the calibration and attitude data corresponding to the source observations must be retrieved, demanding optimization of source access, while to perform the calibration of a given CCD all the data of all the sources observed in this given CCD are requested. Finally, if we want to improve a parameter of the astrometric model we need all the observations of all the observed sources. All these somehow conflicting requirements must be implemented in the data base and access system.

3. THE GAIA DATA BASE

The Gaia data base should be designed to store and retrieve all the data produced by the telemetry flow, both from the Astro and from the Spectro instruments, as well as all the (useful) intermediate data generated and final scientific results. Satellite data providing attitude and calibration of the instruments should be also stored, together with model parameters and auxiliary data needed to perform any one of the expected calculations.

On the other hand the data base must isolate data from algorithmics and preserve the raw data in order to make possible a future different approach to data reduction if needed.

The design of the data base must consider the tools to retrieve data according to the priorities set by the needs of the most demanding algorithms, that is: time, CCD, position, star, etc. It must be scalable, growing according to the increasing amount of data as the mission goes on, and must support the relationship between different types of data associated to a given source, at the same time the observations of a given source, obtained at different epochs as the satellite scans the sky, must be linked together.

In order to perform the large amount of operations required, the data base must be ready to work in a distributed environment.

4. THE GAIA DATA ACCESS AND ANALYSIS STUDY

The complexities involved in the Gaia data treatment were very soon identified (ESA 2000) and it was recognized that some Hipparcos-like approximation was not appropriate. Instead of using flat files, an object-oriented system was considered and some experiments based on Hipparcos data took place (O’Mullane 1999).

It was considered that a deep study aiming to establish the baseline concepts for the system on a realistic basis and aiming to prove the feasibility of the approach chosen for the reduction of the mission was needed.

The Announcement of Opportunities for the Gaia Data Access and Analysis Study (GDAAS) was issued by ESA in 2000 with the objective of defining an efficient, scalable, maintainable and usable system for populating the Gaia Data Base from the satellite stream allowing the data storage as well as the processing of the data. The contract was awarded to the consortium constituted by the software company GMV (Madrid, Spain), the University of Barcelona (UB) and the Super Computing Centre of Catalonia (CESCA).

4.1. The GDAAS-1 Prototype

During a first phase of GDAAS (GDAAS-1) that took place in the period 2000-02, a first prototype was built. Its deployment comprises:

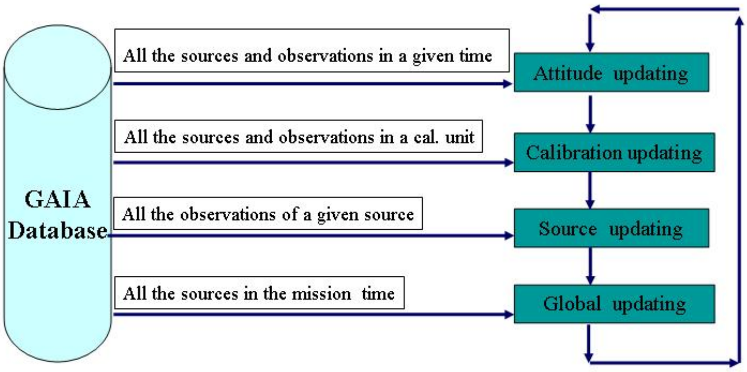
GAIA Database
All the sources and observations in a given time
All the sources and observations in a cal. unit
All the observations of a given source
All the sources in the mission time
Attitude updating
Calibration updating
Source updating
Global updating

Figure 2. Access to the Gaia Data Base of the GIS algorithms.

- The definition of models for the instruments and its operation at an appropriate level of detail, with the requirement that the data downloaded allows the reconstruction of the observations.
- The design of the Gaia Data Base and the processing framework. A Data manipulation layer was implemented in order to insulate data and processes and a data model was designed according to the needs of telemetry and of the processes to be implemented.
- The data base and processing framework were implemented. Algorithms were coded and implemented in the system for the processes we will describe below.
- The full system was integrated and tested at CESCA premises.

The level of realism on the description of Gaia observations was enough to prove the main features of the system and of the data treatment and was a good first order approximation to the Gaia instrument and mission. Let us recall that Gaia was fully reviewed and redesigned in 2002. Nonetheless our design was frozen to the old one (Gaia-1). The main features of the design were:

- Only the astrometric instrument was considered.
- Optics: Two focal planes. LSF constant in the focal plane and along the time. No chromaticity effects were considered.
- Astrometry: The astrometric parameters of a star are $\alpha, \delta, \mu_\alpha, \mu_\delta$ and π, thus neglecting the effect of the radial velocity.
- Calibration: It is assumed that a CCD is characterized geometrically by two constants (position along and across scan) and photometrically by the average of the measurements performed in it. No variation from pixel-column to pixel-column was assumed. It was assumed that these parameters remain constant for a three-month period.
- Global parameter: The only parameter entering the astrometric model was the γ parameter. Only the Sun was considered as a light deflector.
- Satellite motion: It was assumed that the orbit is L2 around the Sun. The attitude is assumed to be nominal plus some noise in pointing. The scan law was assumed to be nominal.

4.2. Main Processes in GDAAS-1

In Figure 3 we show a schematic view of the GDAAS-1 prototype indicating the processes implemented which were those considered critical in the Gaia mission. We had:

- Telemetry Ingestion: Reading and decoding of data coming from the satellite. At the same time some Initial Data Treatment (IDT) is performed in order to create the ‘elementary observation’ associated to a transit of a star in a given CCD. It performs a centroiding of the observed samples and provides a rough estimation of its position. The determination of the elementary observation depends on the calibration parameters of the instrument, so it should be recalculated after any modification or improvement in these parameters. In GDAAS-1 it was considered that this effect was negligible.

  The system must be capable to cope with the daily observations of the satellite and with areas of the sky with different stellar density.

  All the telemetry files are produced by the GASS simulator described in a separate paper (Luri et al.

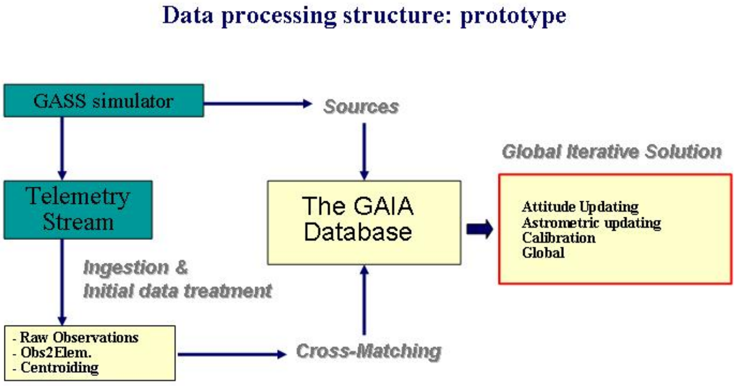

Figure 3. The GDAAS-1 prototype.

2005). Let us note that the assumptions for GDAAS must be coded in the simulator. It also provides all the values requested for the sources, values that can be stored in the DB as needed without passing through the ingestion and IDT process.

- Cross-Matching: As the satellite scans the sky, a given source is observed several times under different paths. In each crossing of the focal plane it is measured in a number of CCDs producing the corresponding elementary observations. Each elementary observation should be assigned to a source and all the elementary observations of a given source must be linked together. The Cross-Matching process, prepared by the group at the Torino Observatory, performs this task. Linking an incoming observation to an existing source in the data base or creating a new one if no cross-match has been possible.

- Core Processing: It includes all the processes needed to run the GIS. As described above this solution can be obtained by the iteration of four steps (Lindegren 2001) each one of them improving one set of parameters:
  - Attitude of the satellite: It is described by quaternions. Each one of them is described in terms of B-splines. The updating of the attitude is performed by improving the B-splines coefficients. To do that in a given time interval, all the observations performed in it are used, as well as the astrometric data of the stars considered and the calibration and model parameters defined for this same time interval. Some auxiliary data (like ephemerides of the satellite and of the Solar System are also needed).
  - Calibration: Geometrical and photometrical calibration parameters for each one of the CCDs on the focal plane are obtained for a given time interval (that is called a calibration unit), using attitude data, model global parameters and all the data of the stars observed in a given CCD.
  - Astrometric parameters: The astrometry of a star is obtained by using all its elementary observations as well as the attitude, astrometric model parameters and calibration data. Note that calibration data are associated to the CCD where the observation was performed.
  - Global parameters. It includes the updating of parameters entering the astrometric or payload model. The example is the γ parameter of the relativistic model for the astrometric observations. To perform it all the observations along the whole mission are necessary.

In Figure 2 a schematic view of the core process is presented. The algorithms for the core process were provided by Lindegren (2001a) and adapted by our group (Luri et al. 2002; Figueras et al. 2002a; Torra et al. 2002; Figueras et al. 2002b) to the different steps. We should mention here what has been an important problem in the GDAAS-1 implementation: All the code has been written in Fortran while the system is an object oriented one and Java is the native language. To solve these problems complex wrappers linking Java and Fortran via C++ have been written and tested.

4.3. GDAAS-1 conclusions

Up until now a quite large number of tests of the different processes have been run in GDAAS-1. The most important results proving the feasibility of the GIS are given in a separate presentation by (Figueras et al. 2005) and have been carefully described in several technical notes (see Figueras et al. 2004a,b, and references therein). The ingestion and cross-matching were tested under very different circunstances, low and high stellar densities, very crowded areas such as Baade's window, and different scan conditions. The results on time to cope with the ingestion and cross-matching process, size of the data base, processors involved and so on can de found in González et al. (2002). On average one single processor needs one hour and a half to ingest and cross-match one day of observations. The extrapolation of the size of the data base to the real mission estimates that its size would be of some 500 TB not including Spectro data.

From the point of view of software considerations GDAAS-1 showed that the approach and the design chosen were correct. Object oriented techniques and UML (Unified Modelling Language) tools demonstrated their advantages in the implementation of a complex system as required by Gaia. The Java language has proven to be ideal for the problems posed by the system, although the use of wrappers added much more complexity to the system and should be avoided as much as possible.

The choice of the data base management system was a key element. The Objectivity system was unable to cope with the framework requested and made the running of the tests difficult. It was decided to move to Oracle even if that required some re-design of the system.

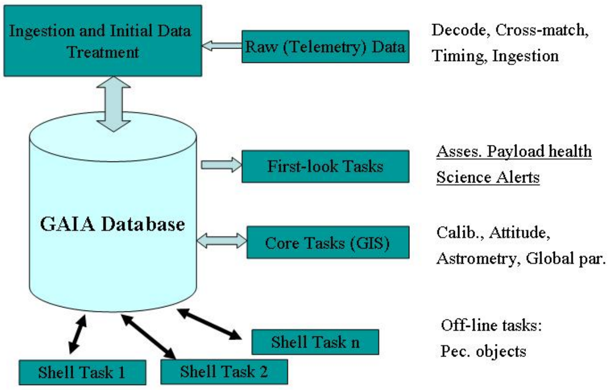
Ingestion and Initial Data Treatment
Raw (Telemetry) Data
Decode, Cross-match, Timing, Ingestion
GAIA Database
First-look Tasks
Asses. Payload health
Science Alerts
Core Tasks (GIS)
Calib., Attitude, Astrometry, Global par.
Shell Task n
Shell Task 1
Shell Task 2
Off-line tasks: Pec. objects

Figure 4. GDAAS-2 main processes.

5. GDAAS-2: THE SECOND PHASE OF GDAAS

The objective of the second phase of the GDAAS is to provide complete confidence in the overall Gaia data processing approach, identifying interfaces with all data reduction steps, implementing and testing an agreed package of algorithms provided by the scientific community and demonstrating the scalability to a final processing system.

During this second phase the Gaia-2 design has been implemented, although in order to obtain as much information as possible, a set of tests on the former design and system are being performed in parallel.

From the point of view of the modelling of the observations and instrument a step forward has been done. The implementation of new and more demanding algorithms is also reflected in the gain of realism of the system, we have:

- Optics: Only Astro instrument. LSF is now a function of the position in the focal plane and time and chromaticity is taken into account
- Astrometry: The astrometric parameters of a star are $\alpha, \delta, \mu_\alpha, \mu_\delta$ and π, thus neglecting the effect of the radial velocity.
- Calibration: In addition to the CCD units, pixel columns are calibrated. Geometric and photometric calibrations will be performed at large and short temporal scales (see Lindegren 2002, for the details). LSF and Chromaticity are calibrated in GIS.
- Global parameter: The only parameter entering the astrometric model is the γ parameter. Sun and planets are considered. A full relativistic model (Klioner 2003) has been implemented in the simulator and in the GDAAS.
- Satellite motion: The orbit is Lissajous around L2. The attitude is assumed to be nominal plus noise. The scan law is assumed to be nominal.

Let us insist on the fact that any modification in the design should be considered in the simulator at the same level of approximation.

From the point of view of the computation some important factors are being considered namely:

- The elementary observations are recalculated each time that a new calibration is available, that means that IDT runs not only during the ingestion but inside the GIS (see Figure 5) and requires that raw data be stored and accessed from different processes.
- As the wrappers were a source of problems in the first phase, some routines of GIS have been translated to Java thus modifying the structure of the processing framework accordingly.

5.1. New Algorithms

The system must perform all the tasks already done by GDAAS-1, using new or adapted algorithms, plus some new tasks. A list of about thirty algorithms was proposed for implementation on GDAAS-2.

According to its impact on the data base, by its requirements for access and computing needs, the algorithms to be implemented are classified as core or shell (see Figure 6). Core tasks are the most demanding ones, and are crucial for the mission. The GIS processes are the core processes. Other tasks that can be run offline are labeled shell. They carry the treatment of a particular type of object or search for the solution of a particular problem. Other tasks, such as the First-look task, are devoted to assessing the quality of the data without going to run a full GIS were considered. Major improvements and additions are:

- Ingestion and Initial Data Treatment: IDT has been largely improved according to the new design and new requirements. That implies a major revision of the Data Model. LSF and chromaticity are included, and an improved version of the cross-matching will be implemented.
- Core tasks: As mentioned above a new design of GIS including the repeated use of raw data is being implemented.
  - The calibration task is being recoded as it now includes the calibration of LSF, chromaticity and geometric calibration and large scale (CCD) and short scale (pixel column) photometric calibration.
  - Source astrometry: An approximation taking into account the relativistic model has been implemented and new code is ready for the astrometry update. Ephemerides for planets are available from the data base.
  - Source photometry: A new code is being implemented.
  - Attitude and global updates are using the same model although its implementation is different. In order to simplify as much as possible the wrapping, the routines providing the field angles corresponding to one observation and which were present in all the GIS steps have been recoded in Java and can be operated individually.

  Other tasks such as Data Base Initialization have been designed and implemented in order to improve the efficiency of the system.

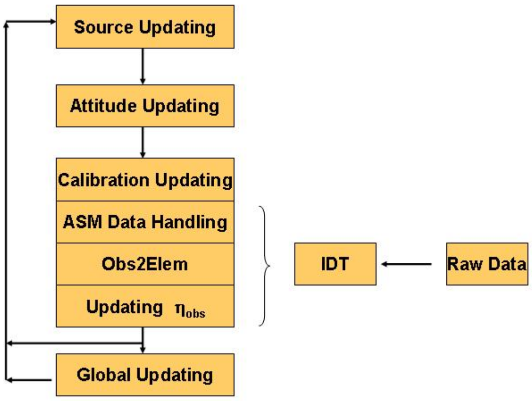

Figure 5. GIS structure in GDAAS-2.

- Shell tasks: Only two of the proposed shell algorithms will be implemented in the present phase:
  - Astrometric binary stars analysis
  - Variability analysis

5.2. The Hardware

Both systems GDAAS-1 and GDAAS-2 have been implemented in CESCA's computers. The hardware has undergone several modfications thus proving the portability of the system. First runs were in a IBM SP2 machine, which was soon substituted by HP V2500, HP N4000 and is at present running in a HPC Compaq320. It offers 32 processors giving some 50 Gflop s^{-1}. The computer is shared with several users.

Other experiments with different platforms have been performed in GDAAS. The system was succesfully run in a beowulf computer in the Universidad Complutense de Madrid. More recently several test combining nodes in HPC320 plus a beowulf system at CESCA plus some PCs at CESCA and UB have been performed, trying to compare configurations and identify the best suited for GIS.

It has been proposed, and it is now being tested, to run the shell algorithms in a GRID distributed system (Ansari 2005).

5.3. Status

It is expected to have implemented at the end of this year all the core processes. The goal is to have a full test of GIS in the GDAAS-2 frame at mid-2005. Up until now the experiments performed with the GDAAS-1-GIS test have allowed us to understand lots of problems that at the beginning were even not considered at all. The lesson learned is that the Gaia Data Access and Analysis System is much more complex than expected.

The system GDAAS-1, a testbed for GDAAS-2, has proven to be robust enough to permit test running with algorithmic and parameters modification.

At the time of writing the IDT of GDASS-2 has been already tested and validated. New tests with the old configuration (GDAAS-1) are running. Work is in progress.

6. GDAAS-3: THE NEXT STEP

In a next phase the system has to be made closer to the operational one. In addition to refinements in the models, instrumentation and algorithms already implemented the main goals will be:

- To prove the GIS approach at a level sufficient to extrapolate and validate to the real mission.

- To implement the shell algorithms at a level sufficient to extrapolate and validate to the real mission.
- To implement the Medium Band Photometry (MBP) and Radial Velocity measurements (RVS) in the data base and in the reduction scheme, making a first approach to a 'unique' instrument integrating all the measurements.
- To perform large scale tests considering the full five years mission and some millions of stars.

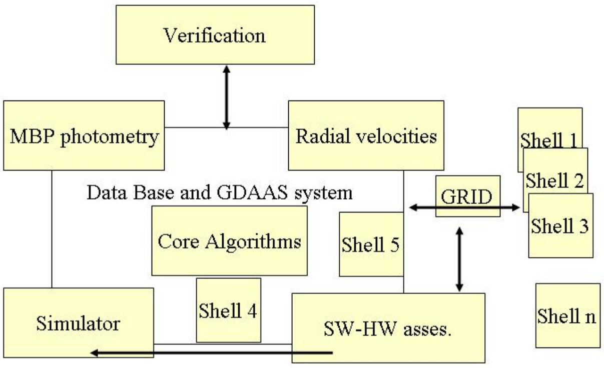

Figure 6. A proposal for GDAAS-3.

The structure of GDAAS-3 can be as shown in Figure 6. A central core and data base with a direct link to some new boxes devoted to MBP and RVS. Special links must be established with the simulator tasks as well as with the tasks of hardware and software assessment and implementation. A new task of verification of the full GDAAS task performed by an independent team can be considered.

As proposed, a number of shell algorithms run via GRID techniques while some others perform their specific tasks *in situ*.

ACKNOWLEDGEMENTS

The GDAAS tasks have largely benefited from several contributors providing algorithms, ideas and criticisms. We warmly thank all of them.

The GDAAS related tasks have been supported by ESA under contracts 3691/00/NL/GS and 16439/02/NL/VD, and by the Spanish Ministerio de Ciencia y Tecnología under contract PNE2003-04352.

REFERENCES

Ansari, S., 2005, ESA SP-576, this volume

Bastian, U., 2004, Gaia technical report GAIA-ARI-BAS-003

ESA, 2000, *GAIA: Composition, Formation and Evolution of the Galaxy*, ESA-SCI(2000)4

Figueras, F., Torra, J., Luri, X., Jordi, C., Masana, E., 2002a, Gaia technical report UB-GDAAS-TN011

Figueras, F., Torra, J., Luri, X., Jordi, C., Masana, E., 2002b, Gaia technical report UB-GDAAS-TN013

Figueras, F., Torra, J., Serraller, I., Masana, E., Luri, X., Jordi, C., Llimona, P., Pérez, P., 2004a, Gaia technical report UB-GDAAS2-TR01

Figueras, F., López, B., Fabricius, C., Torra, J., Jordi, C., Llimona, Masana, E., Luri, X., 2004b, Gaia technical report UB-GDAAS2-TN020

Figueras, F., López, B., Fabricius, C., et al., 2005, ESA SP-576, this volume.

González, L.M. Serraller, I., Torra, J., et al.,2002, Technical Report GMV-GDAAS-RP-001

Klioner, S., 2003, As. J., 125, 1580

Lindegren, L., 2001a, Gaia technical report Gaia-LL-34

Lindegren, L., 2001b, Gaia technical report Gaia-LL-37

Lindegren, L., 2002, Gaia technical report Gaia-LL-44

Luri, X., Figueras, F., Torra, J., Jordi, C., Masana, E., 2002, Gaia technical report UB-GDAAS-TN010 v2.1

Luri, X., Babusiaux, C., Masana, E., 2005, ESA SP-576, this volume

O'Mullane, W. Lindegren, L., 1999, Baltic Astron. 8, 57

Torra, J., Figueras, F., Luri, X., Jordi, C., Masana, E., 2002, Gaia technical report UB-GDAAS-TN012 v2.2