SF2A 2009
M. Heydari-Malayeri, R. Reylé and R. Samadi (eds)

GAIA CATALOGUE AND ARCHIVE, PLANS AND STATUS

O'Mullane, W.[1]

Abstract. The Gaia Data Processing and Analysis Consortium (DPAC) has been, and continues to, work on the software for processing the Gaia data. The result will be an unprecedented celestial catalogue of about one thousand million objects with astrometric accuracies far exceeding anything currently available. But little has been heard about the catalogue itself, its availability, content etc. Some work has been done in this area and the status is explained in this short article.

1 Introduction

The Gaia Satellite and its capabilities are covered by de Bruijne (2009). ESA will build, under contract to EADS Astrium, the satellite and launch it from French Guiana aboard a Soyuz in 2012. The Data processing is a community task and the DPAC was officially put in place in 2006 to perform this task. Currently over 300 DPAC members (a DPAC member must work a minimum of 10% on Gaia) are working on the data processing systems. There is also some ESA involvement in DPAC with ESAC contributing to many tasks an leading some. In particular CU1 concerning the overall architecture of the system is led and heavily supported by ESAC. There will also be ESAC involvement in the archive.

When DPAC was made official via an Announcement of Opportunity from ESA in 2006 the catalogue/archive was explicitly excluded. DPAC as agreed in 2006 consists of nine coordination units, the ninth of which was to deal with the archive and to be activated at a later stage. This may entail a further announcement of opportunity but it is a matter for the DPAC Executive and Gaia Science Teams to agree with the Project Scientist.

Hence what we are really discussing here is Coordination Unit 9 (CU9). A first set of requirements (O'Mullane 2009) regarding the Archive has been agreed with DPACE and GST and are discussed below. Some tentative agreements on data releases are included in the document and will be outlined below.

2 Architecture of the archive

Within the DPAC architecture the Gaia Main Database (MDB) will contain all processed data from the satellite. This will be versioned at regular intervals and is the logical starting point for creating the Gaia catalogue as depicted in Figure 1. The catalogue is not however a simple copy of the MDB, it must be more refined and must present a single coherent dataset to the community (no longer divided by coordination unit). Furthermore the MDB is not designed for the type of arbitrary querying or data mining which will be required of the archive.

Figure 1 also shows multiple archives, ESA must have a copy of the archive as ESAC has been designated repository of all space science data for the agency. This does not preclude other copies for institutes who seriously want a copy. The system should be fairly portable.

In general CU9 will be an integral part of DPAC - it could function in no other way. Indeed, as for CU1, some CU9 members will have to come from the existing CUs, of course in many cases this will mean finding additional effort within the CU. For the purposes of discussion the archive is seen as comprising several components as depicted in Figure 2. Some of these are discussed further below. The components are :

- **Ingestor (ING)** to populate the archive.

[1] European Space Astronomy Centre, Madrid, Spain

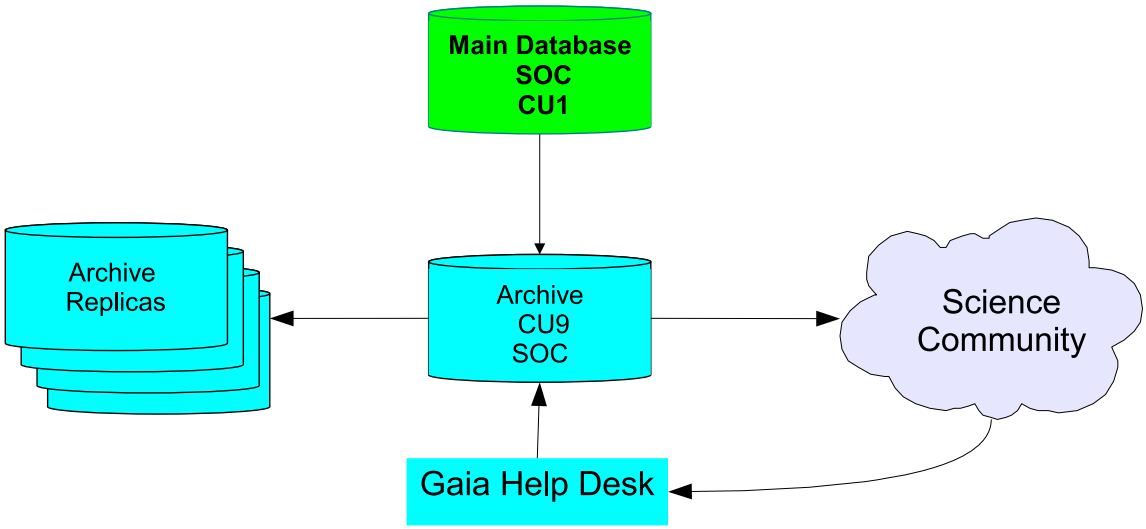

Fig. 1. Context of the Gaia archive. Derived from the Main Database the archive provides an interface to the general community. The possibility for multiple archive copies is also considered.

- **Storage System** Physical disk, machines and DBMS.
- **Interrogation System (ITG)** to effectively query the archive.
- **Advanced Applications (ADV)** for value added access.
- **Documentation (DOC)** to allow users understand the archive.
- **Science Alerts (SAA)** Anomaly based Science Alerting.
- **Public Outreach (PBO)** to engage general public.
- **Help desk (HLP)** to answer users' questions.
- **Community Interface (CIF)** to provide a consolidated portal.

2.1 Community Interface

The general idea is that the archive should have a public oriented feel - no special "professional" site should exist. Rather all come to the same starting point but the professional obviously will dig deeper to a more extensive set of tools and information. Some form of no-login access should be provided with advanced features available with fairly easy self-registration. Many tools should be included in the archive for example; some sort of sky browser - perhaps a customisation of the SDSS Sky Browser (Szalay 2002) or Google Sky. Availability, of at least higher level information, in multiple language of course goes without saying.

2.2 Documentation

Documenting the catalogue will be CU9's most important task. Again much information exists in the document repository of DPAC and the MDB Dictionary Tool but it will need to be made presentable to a general audience and brought up to date. Knowledge of the algorithms will need to be extracted from the various CUs. There should be extensive pre-calculated statistics e.g. source and observation maps, histograms of sources per magnitude bin, spectral type, etc . Additionally some informational plots should be provided e.g. Hertzsprung-Russell diagrams, Hess diagrams, galactic-kinematics diagrams etc. A project history should be provided.

Although undoubtedly too large to print in its entirety some printed volume or volumes would be interesting. Perhaps the source catalogue, without transits and spectra, could be provided on some form of media.

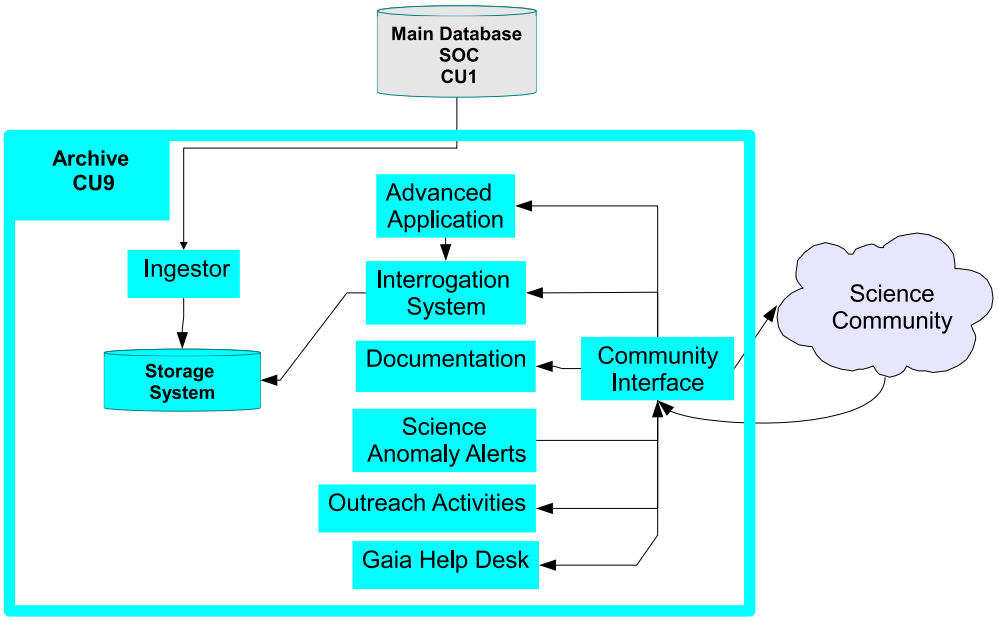

Fig. 2. Gaia archive components. This decomposition was done only to facilitate thinking about the archive, it does not mean CU9 will end up with this exact set of components.

2.3 Interrogator

There will be both graphical and programmatic interfaces to the archive. Virtual observatory protocols such as TAP and SIAP shall be implemented. Users will have access to a SQL equivalent and/or ADQL. All of these interfaces require a fast engine to answer queries. Simply putting the data in a Database will not work - it will need to be tuned and will need to make use of spatial indexing techniques such as HTM and HEALPix (O'Mullane 2001). Whatever it is, it needs to be fast !

2.4 Science Alerts

From early in the mission, flux based alerts will come from CU5 - these should be VOEvent based. Additionally CU4 are in direct contact with IMCCE to provide NEO orbits as they become available. The archive should contain at least a record of the alerts and possibly a publishing mechanism for alerts coming from other CUs. Currently DPAC is concentrating on data processing - possibly other alerts will arise later. In any case first results coming from Gaia should be available not too long after the nominal mission starts.

3 Schedule

A definitive release schedule has not yet been finalised between GST, DPACE and ESA and depends very much on the data quality from the satellite and the processing systems. The intention is certainly not to wait until 2020 and the *final* catalogue before getting data to the community. It should be possible to provide some form of astrometry and possibly photometry about two years in to the mission. This could be followed in the fifth year by an update, with the final release after eight years as planned. Only the final date is definitive of course but there is agreement to do several other intermediate releases, the Science Management Plan (Gaia Project Scientist 2006) calls for at least one intermediate catalogue release.

Mignard also considers a special release after possibly six months of nominal mission. With the first full sky from Gaia it would be possible to pick out the hundred thousand Hipparcos stars and provide new proper motions for them spanning 21 years and with accuracies of 50 to 100 $\mu as/year$. This is generally seen as desirable.

Technically it is not feasible to have a data release until about eighteen months of nominal mission data have been collected. Allowing for processing time it would be about two years of nominal mission before a reasonable

release could be made. One must remember that most of Gaia's calibration involves the data taken by Gaia itself during the mission. Hence CU9 could start after launch which would reduce risk for CU9 of launch delays. It is acknowledged that some work does need to start before that time. GST and DPACE are considering how best to achieve this, (O'Mullane 2009) is a start.

4 Open Areas

There are several open areas which require far more investigation in CU9. The final archive is so far in the future that we should not be bound by our current thinking as to what constitutes an archive.

Brown suggests an attempt at a *living archive* in which new ground based observations could be coupled with Gaia data to improve the source catalogue. This, for example, would allow improved solutions for binaries. The question of allowing additions to the released catalogue is quite tricky, there are issues of quality, security and maintenance. But since a complete printed catalogue is impossible why not a completely new type of archive?

Modelling is also very popular and sophisticated these days. Binney asks how we will be able to compare a model to the Gaia catalogue. Should such a facility be provided ? How would it work? Hogg (Hogg 2008) goes further and suggests archives should be encoded in a model to answer other questions, not just to compare models.

Although virtual observatory protocols will be implemented and the VO provides dynamic cross matching it is felt that perhaps a few major catalogues should be matched. The VO can do no more than give the lowest common denominator - a focused match could provide better results. It may merit including some match tables in the archive.

Szalay has said for many years that with the new surveys we need to bring the processing to the data not the data to the processing. Virtualisation could be an excellent way to do this - one could make a range of virtual machines available *in the archive* for users to install and run their programs on. Then just let them download the result. The local storage provided by CasJobs and VO Space partially does this but the complexity of code which can be sent is limited i.e. for CasJobs one can only send SQL programs. Virtualisation with appropriate access libraries would allow almost any code to be run. This may be the only way to bring fully general models to the data. This obviously has very serious security implications.

5 Conclusion

Coordination Unit 9 (CU9) will be tasked with setting up the Gaia archive. Since most data will not be suitable for public consumption before about two years into the mission, CU9 could start in earnest after launch. Initial ground work would need to be laid in the coming years however. The Gaia archive will be an excellent resource challenging us to rethink our concept of an archive. With Science alerts indeed the first Gaia data will be available before any catalogue is released.

The Gaia Data Processing Consortium Executive and the Science Team Members have provided valuable insights in the production of (O'Mullane 2009) upon which this short article is based.

References

de Bruijne, J., 2009, These Proceedings

Gaia Project Scientist, (ESA/SPC(2006)45).

Hogg, D. W., Lang D., 2008, arXiv:0810.3851v1. Classification and Discovery in Large Astronomical Surveys proceedings.

O'Mullane, W., DPACE, 2009, DPAC Document WOM-033

O'Mullane, W., Banday, A. J., Górski, K. M., Kunszt, P., Szalay, A. S, 2001, Mining the Sky Proceedings

Szalay, A., et al., 2002, ACM SIGMOD 2002 proceedings.