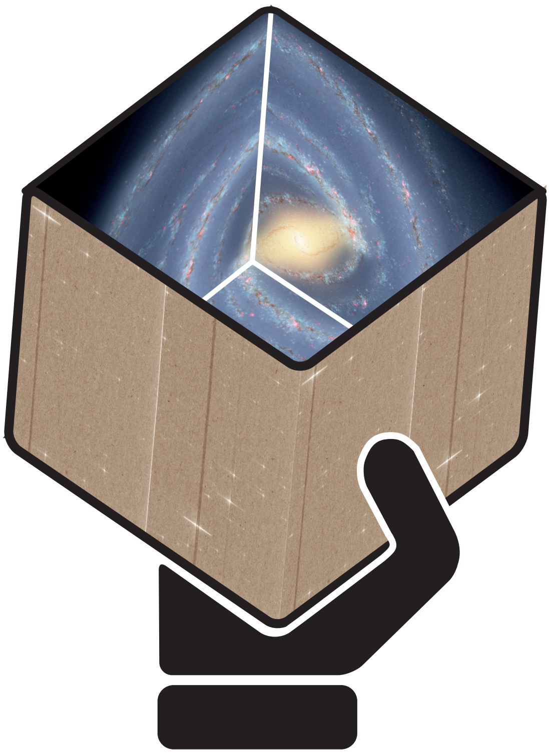


Delivering the promise of Gaia

Response to ESA's Announcement of Opportunity

Proposal for the Gaia Archive

date of issue	09 January 2013
status	Submitted to ESA

Compiled by:

J. Alves, F. Arenou, A.G.A. Brown, S. Els, N. Hambly, A. Helmi, S. Jordan,
A. Krone-Martins, F. van Leeuwen, X. Luri, E. Masana, P. Di Matteo, E. Mercier,
A. Moitinho, W. O'Mullane, P. Osuna, J. Salgado, D. Tapiador, N. Walton

Editors: X. Luri and W. O'Mullane

Cover art design by C. Brown. Image credits: R. Hurt (SSC)
NASA/JPL-Caltech and M. Davidson, University of Edinburgh

Contents

1 Executive Summary

ESA's Gaia mission is the next European breakthrough in astrophysics, a cornerstone mission scheduled for launch in the final quarter of 2013 aimed at producing the most accurate 3D map of the Milky Way to date (Section 2). The resulting stereoscopic census of our Galaxy will represent a giant leap in astrometric accuracy complemented by the only full sky homogeneous photometric survey with an angular resolution comparable to that of the Hubble Space Telescope, as well as the largest spectroscopic survey ever undertaken. The scientific bounty will be immense, not only unravelling the formation history and evolution of our Galaxy but also revealing and classifying thousands of extra-solar planetary systems, minor bodies within our solar system and millions of extragalactic objects, including some 500 000 quasars. Moreover, such a massive survey is bound to uncover many surprises that the universe still holds in store for us.

This document, answering ESA's *Announcement of Opportunity for the Gaia Data Processing Archive Access element* (AO), describes our proposal for the design, implementation, and operation of the Gaia archive in the context of the Gaia Data Processing and Analysis Consortium (DPAC). It has been prepared by a team of scientists and engineers structured around the Gaia Archive Preparation (GAP) group, an initiative of (but not limited to) DPAC to prepare for the starting of the task of the archive implementation.

The main goal (Section 3) is to provide a comprehensive repository of the rich data products to be generated by Gaia, and a range of access mechanisms and associated helper applications to maximize the scientific exploitation and public impact of the Gaia data. The Gaia data set will be large but, beyond its sheer size, will also be intricately interrelated. In this it will be unique, no other survey exists or is planned which delivers all sky photometry, astrometry and spectroscopy.

In this context, to support the main goal of unravelling the formation history of our galaxy, the archive must be able to answer complex questions involving the grouping and analysis on one billion or more objects (Section 4). Technically we have already demonstrated, using a simulated catalogue, technologies such as Hadoop (DTP-004) and Aladin (see Figure 6). But there is much more to do to make ingestion, storage and retrieval feasible and faster for the entire one petabyte final mission data archive.

We stress here that we do not consider the implementation and operation of the Gaia archive simply a technical challenge. The archive must be driven by science considerations. Consequently, science requirements have been gathered from the astronomical community and we will continue to request new science cases during the mission (Section 5). A first ranking and classification is provided in AB-026, a document we intend to refresh every few years or so during the mission. The cyclic development approach, already used by DPAC, is ideal to cope with the influx of new requirements which will result from this process. We expect a great many perceptions to change once the first catalogue is out and once the notion of doing science with a billion-source catalogue is understood.

As specified by the AO, the DPAC Coordination Unit resulting from the implementation of the proposal (CU9) will be embedded in DPAC, following the same rules as its other CUs. We have therefore devised the structure of the development effort following the DPAC guidelines, with some deviations from the standards to adapt to the specific needs of the archive development (Section 7). It is worth highlighting here, again following the AO specifications, that ESA will play a key role in CU9. The Satellite Archive Team at ESAC has the remit to archive and make available all European space science data, and as such their efforts form and integral part of this proposal so that we can together create an outstanding and invaluable resource for space science.

An additional and critical constraint to take into account is the proper handling of access to the data. CU9 is part of DPAC and is bound by the Science Management Plan (ESA/SPC(2006)45), which states that there will not be proprietary data rights for Gaia, and therefore special measures are described in our proposal (Section 8) which are aimed at ensuring that prior to official data releases the data is not used for scientific purposes. Hence within CU9 access to actual Gaia data will be restricted as much as possible before actual data releases. Instead simulated Gaia data will be extensively used for development, testing and validation of software. In the cases where real data needs to be used the members of CU9 will be bound by the DPAC publication policy (FM-039).

Based on the guidelines described above we have defined a development plan and schedule (Section 9) tied to the DPAC planning and to the expected data releases (the current baseline for these data releases is discussed in Prusti (TJP-011) and O'Mullane & van Leeuwen (WOM-066)). The CU9 effort itself has been structured into eight high-level work packages — described in Sections 10.1 to 10.8 — that cover management, documentation, archive architecture and development, data validation, operation and services, education and outreach, science enabling applications and visualisation. Let us remark here that we have made the production of outreach and educational resources from the Gaia archive one of our priorities, since we feel that the visibility of the mission is an essential asset to ensure its overall success, as well as our duty towards society. We also want to remark that we have included the visualisation of the Gaia data as another area where we are convinced we can make a significant contribution to better explore and exploit the Gaia data set. The ability to see and link objects in an easy manner across multiple data spaces within the entire archive will be challenging, but when achieved will provide powerful discovery tools.

With this structuring of its work the CU9 will deliver in a timely manner data access, data handling and manipulation capabilities, meeting the requirements of the community, to allow for a full scientific exploitation of the Gaia data releases. The key elements are:

- A thoroughly validated and well documented set of Gaia data, ready for scientific exploitation.
- A functional, single point access to all Gaia science data, including user support.
- A defined API to the science data repository to allow the development of high throughput access applications and visualization tools.
- A set of advanced applications to allow for defined manipulation and visualization of the data.
- A set of tools and content for outreach, tailored for different targets (media, academic and general public) to maximize the social impact of the mission.

Let us conclude by paraphrasing Sir Isaac Newton and stating that we feel we are standing on the shoulders of giants. Hipparcos paved the way for astrometry from space and has been a basic reference for astrometry in particular, and astronomy in general, for nearly two decades. We now look forward to ushering in the era of Gaia.

2 Gaia Mission Overview

The Gaia space mission, to be launched by the European Space Agency towards the end of 2013, will provide a stereoscopic census of our Galaxy through the measurement of high accuracy astrometry, radial velocities and multi-colour photometry. Over the course of its baseline five year mission Gaia will measure parallaxes and proper motions for every object in the sky brighter than (G) magnitude 20 — amounting to over 1 billion stars, galaxies, quasars and solar system objects. It will achieve an astrometric accuracy of 12–25 μas, depending on colour, at 15th magnitude and 100–300 μas at 20th magnitude. Multi-colour photometry will be obtained for all objects by means of low-resolution spectrophotometry which is realized through prisms dispersing the light entering the field of view. One disperser — called BP for Blue Photometer — operates in the wavelength range 330–680 nm; the other — called RP for Red Photometer — covers the wavelength range 640–1050 nm. In addition to the photometric instrument, Gaia features the so-called Radial Velocity Spectrometer (RVS). The RVS provides the third component of the space velocity of each star down to about 17th magnitude (V-band). The RVS instrument is a near-infrared (847–874 nm), medium-resolution ($\lambda/\Delta\lambda \sim 11\,000$), integral-field spectrograph dispersing all the light entering the field of view. With this instrument radial velocities with a precision of 1–15 km s^{-1} will be measured for all objects to $V \sim 16$–17, depending on colour, thus complementing the astrometry to provide full six-dimensional phase space information for the brighter sources. Gaia will achieve the complete all-sky survey to its limiting magnitude via real-time on board detection.

The primary scientific aim of the mission is to map the structure of our Galaxy and unravel its formation history and subsequent evolution. A number of current cosmological models envisage the formation of large galaxies through the merging of smaller structures. Deciphering the assembly history of our Galaxy requires a detailed mapping of the structure, dynamics, chemical composition, and age distribution of its stellar populations. Ideally one would like to ‘tag’ individual stars to each of the progenitor building blocks of the Galaxy. The Gaia mission is designed to provide the required fundamental data in the form of distances (through parallaxes), space velocities (through proper motions and radial velocities) and astrophysical characterisation (through multi-colour photometry) for massive numbers of stars throughout most of the Galaxy. It should be stressed however that Gaia is not simply a ‘Milky Way mission’ but is truly a multi-faceted *astrophysics mission* which will provide exciting scientific outcomes on the following (incomplete list of) topics: fundamental stellar data across the Hertzsprung-Russell diagram, the characterisation of tens of millions of binary stars, unique samples of variable stars of nearly all types (including key cosmological distance calibrators), detection and orbital classification of thousands of extra-solar planetary systems, a comprehensive survey of objects ranging from huge numbers of minor bodies in our Solar System, through galaxies in the nearby Universe, to some 500 000 distant quasars. Gaia will also provide a number of stringent tests of general relativity. Last but not least, a massive survey such as Gaia will surely uncover many surprises that the Universe still holds in store for us.

Looking ahead, the Gaia Data Processing and Analysis Consortium's (DPAC) processing of the rich and complex data from Gaia coupled with the efforts described in this proposal will lead to a catalogue and data archive of immense scientific utility.

The data products to be generated by the Gaia satellite and produced by the DPAC cover a wide range, containing the following. In Gaia's own broad-band magnitude G the number of stars in the catalogue is estimated to be $\sim 7 \times 10^5$ to $G = 10$, 48×10^6 to $G = 15$ and 1.1×10^9 to $G = 20$. About 60 million stars are expected to be seen as binary or multiple systems by Gaia, among which millions of eclipsing binaries. For each source observed by Gaia the following information is provided: positions, parallax, proper motions, the full covariance matrix of the astrometric parameters (standard errors and correlations) and astrometric solution quality indicators; broad band fluxes in the G, $G_{\rm BP}$, $G_{\rm RP}$ and $G_{\rm RVS}$ bands, as well as the prism spectra measured by the blue and red photometers. Variability indicators will be provided for all stars together with epoch photometry. Solution classifications for multiple stars are provided and, where relevant, orbital parameters together with covariance matrices and quality flags. The photometric, spectroscopic and parallax information will be combined to derive as much astrophysical information on each star as possible. The astrophysical parameters include $T_{\rm eff}$, A_V, $\log g$, [M/H], and $[\alpha/{\rm Fe}]$ where possible, and luminosities and ages will also be provided. For about 10^8 stars a variability analysis will be provided and estimates indicate that about 20×10^6 classical variables and $1\text{–}5 \times 10^6$ eclipsing binaries will be found, among which ~ 5000 Cepheids and 70 000 RR Lyrae. The spectroscopic data from the RVS instrument will lead to radial velocities for the $\sim 150 \times 10^6$ stars at $V \leq 17$; rotational velocities ($v \sin i$), atmospheric parameters, and interstellar reddening for the $\sim 5 \times 10^6$ stars at $V \leq 13$; abundances for the $\sim 2 \times 10^6$ stars at $V \leq 12$; accumulated spectra, allowing spectral analysis for all stars at $V \leq 13$. The spectroscopic data are also expected to contain about 10^6 spectroscopic binaries and about 10^5 eclipsing binaries. In addition the catalogue will contain astrometry and photometry for $\sim 3 \times 10^5$ solar systems bodies, $\sim 5 \times 10^5$ quasars, and some 10^6–10^7 galaxies.

The Gaia mission will thus result in an astronomical catalogue and data archive of unprecedented scope, accuracy and completeness, which will be *the* astronomical data archive for decades to come. As is clear from the contents summary above, the data set will be very large and rich in content, making it complex to deal with. Unlocking the immense scientific and outreach potential thus requires a matching effort which we propose to undertake. The data processing outputs from each of the DPAC coordination units (the organisational elements of the DPAC responsible for the analysis and generation of data products covering the areas of astrometry, photometry, spectroscopy, non single stars and solar system objects, variable stars, and higher level astrophysical parameters of Gaia objects) need to be *validated*, *documented*, and *integrated* into a single Gaia archive and catalogue which will be made available to the scientific end user community and the wider public. The *science services* provided to the astronomical community will form a layer on top of the catalogue and archive providing at least the following facilities: quick and easy basic access to the catalogue data; advanced science exploitation tools making use of both the catalogue and archive, including efficient inclusion of complementary ground- or space-based surveys; visualisation tools tailored to the high dimensionality and information context of the Gaia data products; tools for developing outreach activities to communicate Gaia results to the public at large as well as pre-packaged outreach activities. Because Gaia will undoubtedly stimulate interest in astronomy throughout society, the Gaia catalogue and archive will be open for use also by individuals and groups from the education and outreach communities as well as the public at large.

As mentioned earlier, Gaia will be launched at the end of 2013. Its baseline operational period is for 5 years, with a possible 1–2 year operational extension. The final Gaia post mission analysis will complete within 2–3 years after Gaia ceases to operate. It is anticipated that the first Gaia science data (that from the daily photometric science alerts) will start to be issued by the end of 2014. Full data releases will commence from late 2015 with a number of major releases foreseen at regular points thereafter - as outlined in the following sections, with the final full catalogues and data product releases in the 2021–2023 time frame (depending on possible mission extension). Thus, this proposal to develop and operate Gaia data access will be responsible for providing systems able to support access to early Gaia data in the near future, whilst developing more complex access systems able to support subsequent and correspondingly richer Gaia data releases later in the decade (see Prusti (TJP-011) and O'Mullane & van Leeuwen (WOM-066)).

Gaia will soon commence its operational phase, generating a rich and complex set of high quality science data which will enable a paradigm shift in our understanding of the Galaxy and the wider Universe. Gaia will transform astronomy in the coming decade and it will be a powerful example of European industrial and scientific expertise. The activities to be carried out by the DPAC developing the Gaia Data Access system as described in the following sections will support the effective exploitation of Gaia data and thus will maximise the science return from Gaia.

3 Mission goal for the archive

Gaia will provide an unprecedented census of our Galaxy in size, scope, and accuracy, encompassing astrometry, radial velocities and multi-colour photometry for over one billion objects in the sky. The primary scientific aim of the Gaia mission is to map the structure of the Galaxy and unravel its formation history. The main mission goal for the Gaia archive is to provide a comprehensive repository of the rich data products to be generated by Gaia, and a range of access mechanisms and associated helper applications to enable effective access to the Gaia data by the end user science community.

3.1 Complex data and complex questions

Today's astronomical data archives and catalogues are steadily becoming larger and richer, while the science questions being asked of the data are becoming more complex. Gaia will provide high quality science data which will be of fundamental importance in aiding research astronomers to answer a wide range of key astronomical problems.

One of the major science cases for Gaia concerns building up an improved understanding of the current structure of our Galaxy, and using this to gain an insight into the formation and evolutionary processes at work. During the lifetime of Gaia, it will generate, year on year, improved information on a large sample of up to one billion stars in the Milky Way. The high precision parallaxes determined for these stars will, in principle allow the distances of these to be determined, and thus, with the large numbers of objects, astronomers will be able to build up a three dimensional view of the Milky Way. However, from the perspective of the Gaia archive, the end user astronomers will not simply issue a bulk query requesting the distance for the complete set of Gaia stars. The analysis is a significantly more complicated process, and will necessarily involve the interplay between observational data and detailed models. Properties of the stars will need to be understood to allow for the selection of stellar types: intrinsically brighter stars – supergiants for example – will probe greater distances in the Galaxy. Interstellar extinction will need to be analysed in more detail from the Gaia data itself and external data sets. The space motions of the stars will enable them to be characterised as to where in the Galaxy they lie, and to which population they belong, for instance old halo stars or young disk stars. These inferences will be supported by astrophysical characterisation data for the Gaia objects, for instance chemical compositions and gravities, which again aid in allowing a fuller description and characterisation of them.

The structure of the Galaxy is intimately related to its formation history. Let us consider the following example. The concordance cosmological model tells us that the halo and perhaps also some fraction of the thick disk formed via mergers. Thus to establish the importance of assembly through mergers requires knowledge of the full phase space distribution of the more metal-poor stars. But how do we find the accreted stars? Brute force sifting through a catalogue that contains 10^9 objects, of which likely less than 1% are in the halo, is not the smart way. Also, it is important to establish how our selection/prejudices (in this case metallicity) affect our conclusions, so multiple exploratory paths are necessary to robustly assess the importance of the merger process.

Furthermore, estimates of the total mass of the Galaxy, its density profile, shape, etc, again require selection of tracers. The astronomer will need to have the tools to make different selections in an efficient way, and then will likely do the statistical analysis, or modelling offline, on his/her own desktop. To make this feasible an archive that is easy to query is needed, a low entry threshold will enable the community to maximize scientific exploitation.

All of the above imply complex queries to the Gaia archive, exposing the rich data attributes available for each Gaia object together with access to related information, either singularly or as an ensemble, from external ancillary data sources. These might include ground based survey data, for instance optical imagery from SDSS, or infrared photometric data from the ESO VISTA surveys.

In conclusion, although an archive or catalogue may be considered products as such, their full potential is only realised when the information content is efficiently accessible and widely used. This enables and fosters new scientific discoveries, improves our present understanding of Nature in significant ways, and leads to serendipitous results unthinkable at the time the mission/experiment was conceived.

3.2 Maximising the impact of Gaia

The driving design goal of the Gaia mission archive is to maximise the scientific and public impact of the Gaia catalogue. This will be accomplished by providing science-enabling services to the astronomical community, which will simplify the path from the complex catalogue and mission data to scientific discoveries. These services will be underpinned by a thorough validation of the archive and catalogue data and they will be accompanied by extensive documentation. Access to the catalogue data will be provided also to the general public, empowering them to better understand our Milky Way, our place in the Universe and our origins.

A science driven archive: The Gaia archive and associated data access systems will be developed ensuring that the needs of the scientific community are fully met. Thus, development will be driven by the implementation of the science use-cases, which have been already contributed by the greater astronomical community in the context of the Gaia Archive Preparation Working Group, and by providing examples of how to use the archive to accomplish certain common, and not so common, astronomical tasks.

Enabling the best science: In order to maximise the scientific impact of the Gaia products, the archive and the services built on top of it must be designed to aid users in the discovery process. Accessing and exploring the Gaia archive must not represent an extra burden. We thus intend to develop and provide a suite of tools and data interfaces that will allow users to explore the scientific products of the mission in the most productive ways: from small scale file exchanges, to high-throughput data streaming services. Our goal is to deliver fast and efficient access methods to query what is expected to become one of the most important and complex data structures in astronomy for the coming decades. As illustrated above, a use of the Gaia archive which is purely based on simple coordinate queries would severely limit the scientific impact of this unprecedented data set. However, more complex queries require a deeper understanding of the data set. The case of Gaia is especially challenging due to its unique data treatment. The archive design will take all this into consideration, and aims at minimising the overhead for the common user, and thus to make the effort needed for using the Gaia data as light as possible.

Science from quality assured data: The complicated data treatment leading to the Gaia archive and catalogue coupled with the rich information content and its high-dimensionality means that a serious effort has to be invested in validating the Gaia data products before releasing them to the community. We therefore have a dedicated work package focusing on the data validation effort. This effort will build on the science enabling tools which will be developed in parallel. To take one example, correlations between the parameters, due to systematics not detected in the data processing, low S/N etc, can lead to spurious features which may only be noticed once all of the information is combined. But in order to do this, we need clustering analysis and other appropriate automated statistical analysis tools to allow for a rapid search through the multi parameter data.

Documentation: Users of the Gaia archive will only truly profit from the rich data set if it is accompanied by sufficient documentation, describing and explaining both the data products and the services and tools provided with the archive. We thus consider documentation to be a central aspect of this endeavour and we will ensure that scientists outside the DPAC will be able to understand how to correctly and effectively use the Gaia data.

Reaching out to the general public: Gaia will provide a 'stereoscopic view' of the skies and of our home in the universe, the Milky Way. This is bound to inspire the general public and we intend to make outreach and education activities an integral part for the Gaia archive effort. The Gaia data products will not only be accessible to the professional astronomer but also to the interested lay person, educators, amateur astronomers, etc. Data access will be tailored to the needs of the end user, and this includes the general public. Outreach with Gaia will present its own challenges as Gaia does not produce the high impact imagery of missions such as the Hubble Space Telescope. Thus for certain communities, such as the school level, a certain degree of care and tailoring must be undertaken to provide effective usage of the archive. We intend to develop appropriate visualisations of the Gaia archive data, which especially for the Milky Way will have a powerful inspirational effect, triggering very effective outreach, as well as motivating new generations of scientists and space engineers, and so, invigorating the present and future of Space exploration and Astronomy.

3.3 Delivering the promise of Gaia

A carefully built archive paves the way to scientific discoveries, but its success requires more than providing only a basic data interface. The Gaia archive will empower the astronomical community as well as the general public to make full use of the Gaia products, by assuring that software tools and services of widespread use in astronomical research and outreach will be ready to deal with the Gaia data.

As a further motivation for embarking on the effort to deliver the Gaia archive we comment here on the remarkable success story of the archive and services provided by the Sloan Digital Sky Survey. By the time the SDSS archive was designed, an important computer scientist and winner of the Turing Award declared that “the average Wal-Mart manager has a better database than the average astronomer” (Thakar, 2008), which was certainly true at that time, and to a certain extent is still true today. But the SDSS scientific branch brought to astronomers the language of databases in an accessible way, and even motivated a private company to evolve its database product to the limits of reliability and speed through the addition of advanced data mining features. More importantly, it fostered a fruitful collaboration between Computer Science and Astronomy, that in a certain way paved the way for a new paradigm in Science, today known as “Data Intensive Science” or the “Fourth Paradigm” — and for what is today known in Astronomy as “Astroinformatics”.

At the same time, the SDSS archive brought the complex data of what was the largest survey of the time to the hands of the general public and amateur astronomers, and its byproducts enabled the general public to make scientific discoveries. Such a dedicated outreach vision is important not only for teaching and encouraging future scientists, but also to the actual construction of the scientific endeavour itself, bringing to people’s hands a chance to add small bricks, allowing everyone to contribute to the construction of humankind’s body of knowledge. Building on the SDSS experience we expect that the much larger and richer Gaia data set will have an even greater impact on science and society.

Apart from the very broad science that Gaia is expected to address, we can surely expect that many serendipitous discoveries will be made using this astronomical data archive, which will be the reference for the next decades. Moreover, the Gaia discoveries may encourage humankind to undertake new enterprises. Although we cannot be sure of the directions in which these discoveries will take us, a carefully built Gaia archive will be an indispensable guide.

4 Technical Challenges and Opportunities

The Gaia mission will pose several challenges for current data archiving technologies, mainly due to the unprecedented amount of data that will be produced. The final data delivery will not only include the catalogue of one billion sources but also the single epoch CCD transit data that was used in its computation (including spectra), making an estimated total data set of around 1 PB.

New challenges will arise in two main areas: state of the art computing technologies will have to be applied for the Ingestion, Data Management and Storage processes of the OAIS model, and new methodologies and protocols will have to arise, based in current VO technology, to fulfil new requirements in the Access process. The combination of both processes will give the opportunity to deliver unprecedented levels of accessibility to perform high performance computation over large amounts of scientific data. Additionally, Gaia will as well present a great opportunity for the development of Visualisation techniques and tools.

4.1 Big Data Handling

Storage and Management of PB data volumes cannot be easily dealt with in the traditional monolithic approaches and some kind of parallelism will be needed to properly address its scientific exploitation. Data analysis will require making use of emerging architectures like those referred to as Cloud environments where users or scientists can upload the data analysis work flows so that they run in a low-latency environment accessing every single bit of information. In addition, raw data (re-)analysis is becoming an asset for scientific research as it opens up new possibilities to scientists that may lead to more accurate results, enlarging the scientific return of the mission. In the case of Gaia, this sometimes complex reanalysis of epoch-level data requires low-latency and Massively Parallel Processing (MPP) environments that can efficiently perform complex operations such as data mining algorithms like cluster analysis, pattern recognition, N-point correlation, etc.

4.1.1 Storage and Analysis beyond relational approaches

There are different ways to address these challenges which can be summarized into two slightly different storage approaches. The first one comprises the relational databases (RDBMS) and data warehouses that provide users with a powerful relational algebra that can be used to access the data in many different ways through a well-known declarative language (SQL). Another alternative is to use a new computing paradigm, the so-called NoSQL. This new architecture emphasizes the scalability and availability of the system over the structure of the information and the savings in storage hardware this may produce. It defines a simple programming model where the user just focuses on the algorithm implementation and the framework itself deals with the distribution and execution in the MPP environment taking care of failures and other contingencies that may arise in such a large distributed system. The former system (RDBMS) might allow the concept of a living archive to some extent while the latter (MapReduce) might be useful for adding archive reprocessing capabilities (a much needed feature in today's scientific archives), as well as a way to confront an archive model against real observed data or to validate and/or summarize a new data set for a better understanding of its contents. However, the MapReduce model may not be appropriate for all kinds of problems that will be faced in the exploitation of the Gaia mission data products. Other approaches will have to be studied as well.

4.2 Exposing Big Data to the Community - Access layer

Bringing in new SW services or data delivery patterns implies changes in the archives interface. Current VO technology will have to be extended to provide additional capabilities, and elements from other IT fields will have to be incorporated, in order to provide the most powerful Cloud services.

4.2.1 Dealing with Big Data in the VO and beyond: Enhancing Service-Oriented Architectures

Virtual Observatory has for a long time been the most innovative effort within the Astronomical Community to provide seamless discovery, access, combination and integrated analysis of astronomical data worldwide.

The computing architectures defined over the years, have been based on the emerging patterns widely used at the time of their definition, i.e. provision of Service Oriented Architectures (SOA) based on loosely coupled Web Services. VO Astronomy is performed nowadays through a whole set of inter-operating Web Services that communicate using well-standardized protocols, either for Resource Discovery (Registry Interfaces), data Access (S*AP, TAP), data Storage (VOSpace), etc.

While the VO has been quite successful, its service-centric conception sets certain challenges for scaling the services to the large data sets (e.g., Gaia, LSST) that will be available in the near future. This is not unique to VO technologies; in many IT fields the volume of data has become large enough to require bringing the software to the data instead of moving the data across services. As data volumes and user numbers increase the cost of moving software becomes negligible compared to the cost of moving the data. Volumes of data generated, as well as storage HW available, seem to grow much faster than network speeds, for which the performance is limited by the switching technologies available. Moreover, technologies currently under development do not appear to offer any improvement in this situation in the near future.

Hence the current trend is to bring the software to the data, and the VO has already started on the integration of Server-Side computing services. The Table Access Protocol (TAP) allows for query (and cross-processing) of astronomical catalogues, and Universal Worker Service (UWS). This provision of multi-purpose execution of server-side software provides the basic background for Software plus Services architectures. This will allow data to flow between data from providers and services providing analysis capabilities. However, this will not be enough in the future: software will have to move even closer to the data more like in Cloud Computing Software as a Service (SaaS) architectures.

Data providers in this new architecture have to provide hybrid capabilities over the data they expose. Provision of data access alone will no longer be sufficient - analysis and storage at the very least will be required also. Client tools will have to either evolve to work with these new services or be embedded, at least partially, as a server-side service.

There will be drawbacks to this integration of data providers and analysis software. For example data storage will often be done by a third party (know as “data escrow”). This exists partially already in the VO where some centres provide VOSpace and a process in some data centre kicked off by a user at home will be requested to place the output in the VOSpace of the user in a different centre. This will become more prevalent and raises all sorts of trust and security issues. Additionally analysis tools and data provision will become closer - the VO has benefited to date from the separation of these responsibilities. Closer cooperation and enforcement of standards will be a major concern for all parties involved.

4.2.2 Delivering Software as a Service: Additional Cloud technologies

Software as a Service (SaaS) is understood in the IT world as a software delivery model in which software and its associated data (both forming the archive) are hosted somewhere in the Internet (i.e., a private cloud in this case), and whose functionality is typically accessed by users with a thin client (i.e., a web browser) over the Internet. The idea is to keep software and data in the data centres allowing scientists to do research from anywhere with just an Internet connection and a browser. This would be a way to really bring the software to the data, in which all services provided are automatically orchestrated and scaled when needed, data sets are semantically linked with one another, etc. Furthermore, the SaaS model is somehow based on the idea of Service-Oriented Architecture (SOA) where services are composed dynamically, as needed, binding several lower-level ones. In addition, SaaS is becoming an increasingly prevalent delivery model as underlying technologies that support Web services (e.g., REST) and SOA mature, and new developmental approaches such as AJAX become popular.

From an architectural point of view, SaaS involves several aspects related to services that must be considered from the beginning. For instance, service description both from a technical (operations, argument types, etc) and semantic (what do the arguments mean? What are those operations for?) perspective, capabilities for service discovery (again both technical and semantic), negotiation (terms and conditions under which the service will be delivered), delivery (invocation, provision and monitoring), composition (automatically whenever possible to provide higher level services) and last but not least scalability (grow and shrink the amount of resources automatically to meet the service level requirements).

As described above, existing VO technology forms an excellent starting point due to its SOA architecture: well defined semantic information representing different magnitudes, a composition or orchestration of different relatively low level services for providing higher level and richer functionality, etc. The SaaS approach may thus rely upon these features to really allow scientists to work in environments where only (ideally) the final products (e.g., the graphs for a publication) are retrieved remotely. This will only be achieved if proper collaborative tools and environments are put in place (already being looked at in projects like *CyberSKA*) without leaving out unique features present in social networks like *Twitter* or *Facebook*. In addition, some kind of Astro applications store might be a good choice for sharing and ranking advanced applications that can be deployed with just one click. This allows customisation of user interfaces for different purposes depending on the user preferences. The original interfaces should be simple enough and focused on the general public (for outreach reasons) and can later on be seamlessly contextualized with more complex applications. Again in this case, the technology for dealing with these matters is already available to some extent.

Some of the emerging architectures presented have been studied to some extent with Gaia simulated data within the GAP working group to assess their suitability for an astronomical archive, to better understand the issues that will be faced during the development phase, to obtain some insights of the scientific community expectations, to benchmark some of the most common work flows by using state-of-the-art configurations, and to provide a proof of concept of the early devised technical architecture. A summary of those studies can be found in DTP-004 (“A Framework for Building Hypercubes using MapReduce", submitted to FGCS) where a well-known MPP relational database management system (RDBMS) and the MapReduce paradigm are tried out respectively.

4.3 Visualisation Tools and Techniques

Visualisation is yet another field that will pose many challenges and create many opportunities. It will probably also need a powerful data analysis architecture that can cover the highly demanding processing capabilities of real-time visualisation. The opportunities for the visualisation of the Gaia archive are not only immense, but also critical from a science return point of view. Given the dimensions of the archive, it is fair to speculate that major results from the Gaia mission are not included in any current Gaia science case. This has been the case for all major science missions in the past and will surely be the case for Gaia. Visualisation of the Gaia archive is perhaps one of the most pressing needs, not only for final data release in a decade from now, but also for the early data releases and as part of a follow-up development of a "living" Gaia archive.

However, discovery requires the ability to "see" and to "link" in an easy manner. In the case of Gaia, this means the ability to interact in an easy way with the entire archive, i.e., filter, select, display, link, and connect to other archives. This poses a major challenge. Although it is currently possible to interactively visualize 10 billion particles (Fraedrich et al., 2009), this cannot be achieved without expensive custom developed software and hardware, currently out of reach for most of the target audience of the Gaia mission, the astronomers of the world. The field of visualisation is a fast changing one, essentially capitalizing on the fast development of computer processing power, and we want to be ready to profit from these developments.

Innovative visualisation techniques associated with faster processors are not silver bullets. They enable new ground to be explored, but given the sheer size and dimensionality of the data, these techniques need to be honed and guided by scientific questions. The current Gaia science cases concentrate on particular objects, parameters, or regions, hence effectively reducing the size of the data visualisation and analysis problem. This data filtering makes these science cases viable for full exploration by an astronomer in a decade, assuming the expected development of computer processing power. But if one attempts a less conservative approach and is faced with the full brunt of the Gaia archive while attempting to tackle less restrictive science cases, even if technically possible, one is immediately "data stunned". Not in the sense of too much data, if the technical challenge is solved, but in the sense of too many relations between the data, necessarily aggravated by noise. One possible way around this "data stunning" is the development of powerful "data simplifiers" generated for the full archive, and consisting of 3D density surfaces/volumes of meaningful astrophysical quantities.

5 Meeting the Science Requirements

The development of CU9 will follow a highly science driven approach. Through the Gaia project, and the Gaia Research for European Astronomy Training (GREAT) network (http://www.great-esf.eu), a requirements gathering exercise was launched by the Gaia Archive Preparation group (GAP) at the GREAT Plenary Meeting PM4 in Brussels (21–23 June 2011).

The community was asked to provide usage scenarios describing how they might wish to access and interrogate the Gaia science data products. These scenarios did not imply technical solutions - but rather expressed the functionality required. The link http://great.ast.cam.ac.uk/Greatwiki/GaiaDataAccess has the complete set of science scenarios.

At the end of November 2011 some 120 scenarios had been submitted. These were then analysed by a working group of the GAP - where the scenarios were classified into a set of science and functional areas (GDAS = Gaia Data Access Scenario):

1. GDAS-BR: Browsing and qualitative exploration: including simple usages of first-time users
2. GDAS-SA: Science alerts
3. GDAS-ED: Early data access: thus access to data from the first releases
4. GDAS-EG: Extragalactic science
5. GDAS-GA: Galactic science
6. GDAS-ST: Stars and Stellar Physics science
7. GDAS-SO: Solar System science
8. GDAS-FP: Fundamental Physics science
9. GDAS-PR: Public outreach for the non-astronomer/non-scientist
10. GDAS-OA: Other and advanced usage scenarios

Each scenario was analysed and rated against a number of key attributes: Urgency, General/Specific, Science rank, Scale, Frequency, Rating, Related to, Inputs, Tasks, Roles, Information required for the roles.

The entire requirements collection process is described in the Gaia technical note: Brown et al., Gaia data access scenarios summary, 2012, (AB-026), which is publicly available at `http://www.rssd.esa.int/doc_fetch.php?id=3125400`.

In parallel to the access scenarios provided by the Gaia end user community, there has been a study of the set of science data products to be produced by the DPAC. The development of the access mechanisms required to provide baseline access to these products, as they are provided by the DPAC processing centres, is a central and core task of CU9 in ensuring that the science requirements of the Gaia mission are fulfilled.

The process by which the development of CU9 systems is matched against science requirements is assured through a careful adherence to the development principles outlined in this proposal. WP910 activities will track development against science requirements. WP940 will ensure the scientific validation of all data made accessible via CU9 interfaces and applications. The operational system (WP950) will take into account scientific demands, ensuring that end users are able to access in a timely manner their required scientific data, fully documented (WP920) to allow quantitative scientific analysis of the data.

Thus, during the course of the Gaia operations phase, CU9 will be developing and operating a set of access capabilities, carefully tuned to ensure that the full science potential of Gaia is reached.

6 Development Approach

CU9 is a fully integrated part of DPAC and will follow the overall DPAC guidelines in terms of development and monitoring. Here we explain the product oriented approach and the development scheme as applied to CU9.

6.1 Overall DPAC development approach applied to CU9

From its inception DPAC has followed a cyclical development approach for all of its software. This ‘eXtreme programming’ approach, detailed for a specific project in WOM-006 and adopted in a modified form for DPAC in general, will be applied to the CU9 products under development and during operations. A primary reason for the adoption of such an ‘agile’ approach is to adapt to change, mainly changing requirements. Inevitably when a product is constructed users will see that it was not as they expected or that what they wanted originally is not quite correct. Hence, CU9 must be able to cope easily with changes. This is especially important in light of the data release dates which are dependent on the progress of the DPAC data processing and may thus change. Any contingency during the operations could affect the duration of the planned data processing cycle as well as the data itself. CU9, being the most downstream CU within DPAC, should be ready to face those contingencies and possibly re-plan or adapt on a very short time scale.

Details on the cyclic development strategy can be found in the Gaia DPAC Development Plan (RD-010). The following points are specific to CU9 and in particular for the Work Packages within which products will be developed.

- The 6 month cycles adopted by DPAC have worked well and will be used in CU9. In addition we must take into account the intermediate data releases based on the scenario described in WOM-066. Each cycle will consist of common generic phases. The cyclic scheme allows CU product deliveries at anytime in the cycle and most CUs already provide an early, intermediate, and final release of their products within the cycle. This will be formalised for CU9 to include 'internal releases' meant for the early integration and testing of CU9 systems and the exercising of CU9 operations. These internal releases are not shown in the following diagrams.
- Toward the end of a cycle CU9 software should be stable and enter a validation phase which is described in Section 9.
- The development effort within the various work packages will be synchronized, however for some work packages the delivery of the products will occur slightly ahead of the others.
- The CU9 software is delivered to and accepted by the CU9 operations team (WP950).

6.1.1 Development and delivery of software products

The software development life cycle as recommended in WOM-012 and followed by all DPAC software will be applied to CU9 products. For more detail see RD-010. The DPAC Project Office Information Management Tool will be used to follow the level of completeness of CU9 software requirements.

6.1.2 Software Product Assurance, Configuration and Technical Support

CU9 will follow the DPAC standards for product assurance and configuration control. This means the DPAC Product Assurance Plan TL-001, which provides all the details about the centralized and automated PA process, will be applied, while the configuration of CU9 products will follow the approach described in the DPAC Configuration Plan WOM-012.

The DPAC Project Office Quality Assurance and Configuration managers will act as points of contact and support for the CU9 WP leaders for any PA and configuration related matters.

For technical support CU9 can rely internally on the efforts within WP930 while CU1 will provide the same technical support to CU9 as it provides to the rest of DPAC. In particular CU1 will help new DPAC people follow the development standards and make available the DPAC wide collaborative tools.

6.1.2.1 ESAC Science Archives specifics: The ESAC Science Archives Team (SAT) development approach is based on the former PSS-05-0 standards taking into account the peculiarities of Archival Software. This is embodied in the SAT Software Configuration and Management Plan (SATSCMP). This existing (and compatible) plan will be followed for the development of the Gaia Archive Core Systems. The code for all the Science Archives of the different missions is located in a SAT Subversion repository, protected with project specific rights, that can be, in this case, referenced from the Gaia subversion repository.

Code dependencies on external or internal libraries are already managed using a Dependency Manager (Apache Ivy) hosted on the SAT Configuration Management Server. Also, integration of code developed by the different Gaia Archive Core Systems developers is done using a Continuous Integration System (e.g., Jenkins[1]) on the same server.

Issues and requirements are managed using a Project Management Tool (Redmine) and the tracking of SPRs and SCREWs (Software Change Requirements and Extra Wishes) are tracked and included in the development phase using the *Scrum*[2] development paradigm.

Gaia Archive Core Systems releases will follow the SATSCMP Software Releases Procedure, in line with the rest of the ESA Science Archives.

6.1.3 Planning and scheduling

The planning and scheduling of CU9 effort will be done by the CU9 leader together with the WP managers. The DPAC Project Office, through the Project Office Scheduler, will support this process and maintain the CU9 milestones as described in the DPAC Milestones Approach (GGA-002).

[1] `http://jenkins-ci.org/`

[2] One of the “Agile” development approaches.

6.1.4 Internal meetings and associated documentation

As part of the monitoring of its activities and to strengthen internal communications regular meetings (face to face or telecons) will be held within CU9. In particular at the start of development cycles a meeting will take place to coordinate the goals and the planning for the various work packages. The CU9 documentation (for instance the Software Development Plan) will then be updated accordingly.

Following a data release a meeting will take place to review the activities performed within the various work packages and to learn the lessons from the problems encountered. If needed the development plans up to the next release will be adapted.

Plenary face-to-face meetings will be organized by the CU9 leader on a regular basis. These will be the occasions to foster the collaboration between the teams carrying out the various work packages. In addition participants from the other DPAC CUs will be invited to attend in order to ensure a tight connection between CU9 and the rest of DPAC.

6.2 Detailed Development Approach

CU9 is structured around a set of work packages. Each work package corresponds to at least one CU9 product. Those products may be used internally to CU9, to support the archive development, or directly delivered to the community. To understand the detailed development strategy of CU9 we need to introduce the different work packages, their products and the sequence of interactions between the work packages in preparation for a data release.

6.2.1 Work Packages

CU9 is composed of 8 work packages listed in Table 1 along with their managers, and short CVs for these managers are included in the appendices document. Two work packages are considered as ‘horizontal’ meaning that they benefit or oversee all the activities of CU9. The remaining six work packages are considered as ‘vertical’ and deliver their products in a specific area. Figure 1 shows the work packages, their role (horizontal, vertical) and the objective of CU9 which is the production of an Archive and the associated services. The work packages are detailed in Section 10 which describes the work breakdown structure.

Num	Title	Manager & deputy manager
WP910	Management	X. Luri & W. O'Mullane
WP920	Documentation	F. van Leeuwen & A.G.A. Brown
WP930	Architecture, Design and Development	N. Hambly & J. Salgado
WP940	Validation	F. Arenou & P. di Matteo
WP950	Operations and Services	E. Mercier & J. Hernandez
WP960	Education and Outreach	S. Jordan & E. Masana
WP970	Science Enabling Applications	X. Luri & S. Jordan
WP980	Visualisation	A. Moitinho & J. Alves

Table 1: CU9 main work packages

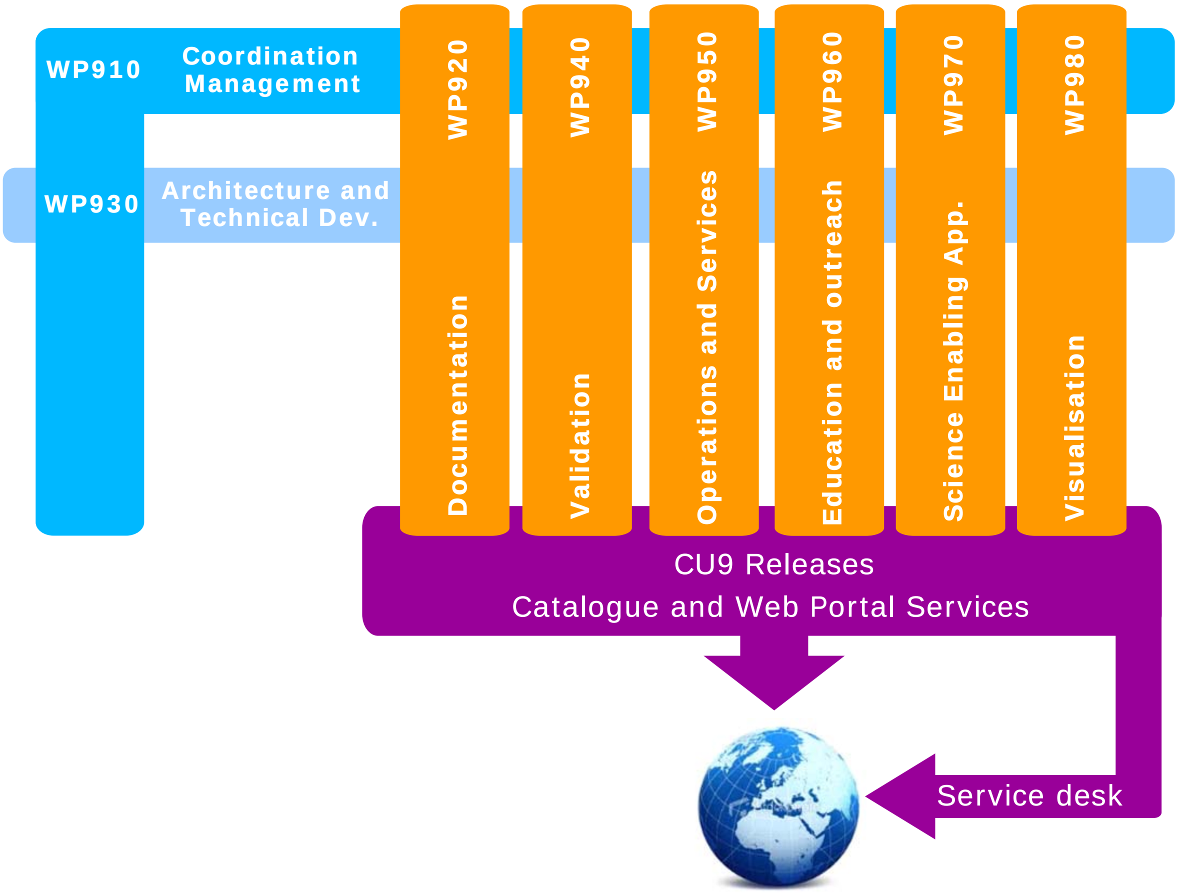


Figure 1: CU9 Work Packages.

6.2.2 Products

As described in WOM-066, CU9 is expected to produce several data releases. A data release is then a product delivered by CU9 to the astronomical community and to the public at large[3]. A data release consists of two sub-products, the core data and a set of services which facilitate the unlocking of the scientific potential of the data. The core data is made of three sub-products: the archive, the catalogue, and the documentation describing both. The services consist of different products such as the science enabling and visualisation tools, auxiliary data sets, web content, documentation, etc. Figure 2 presents the products composing a CU9 data release.

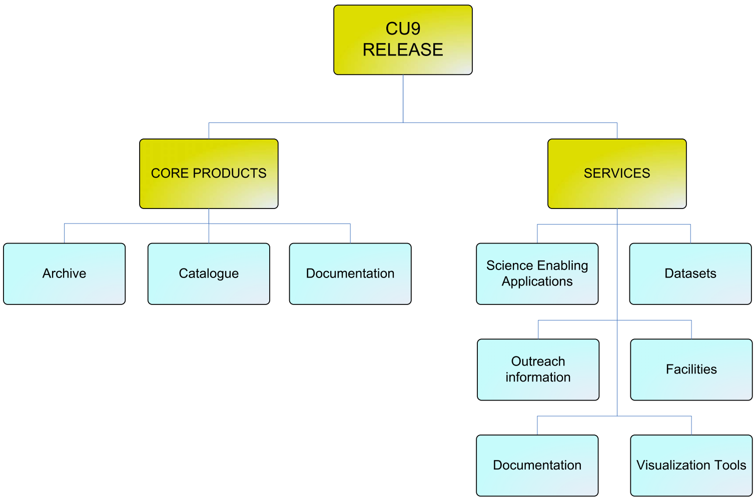


Figure 2: CU9 Products.

We stress that at the time of writing this proposal an exhaustive identification of all CU9 products has not been performed. We expect that the Gaia archive access effort will evolve, for example in response to requests from the astronomical community, which may lead to new products. We are however confident that the top level products are all identified.

6.2.3 Work flow for a Gaia data release

In this section we describe the sequence of steps involved in the production of a Gaia data release, showing at the same time the role of the various work packages. Figure 3 visualizes the work flow leading up to a data release. The numbers besides the arrows showing the work flow will be referred to in the description that follows.

[3] Data releases are not the only products to be delivered by CU9 - WP960 Outreach for instance will deliver some products to the users independently from a data release.

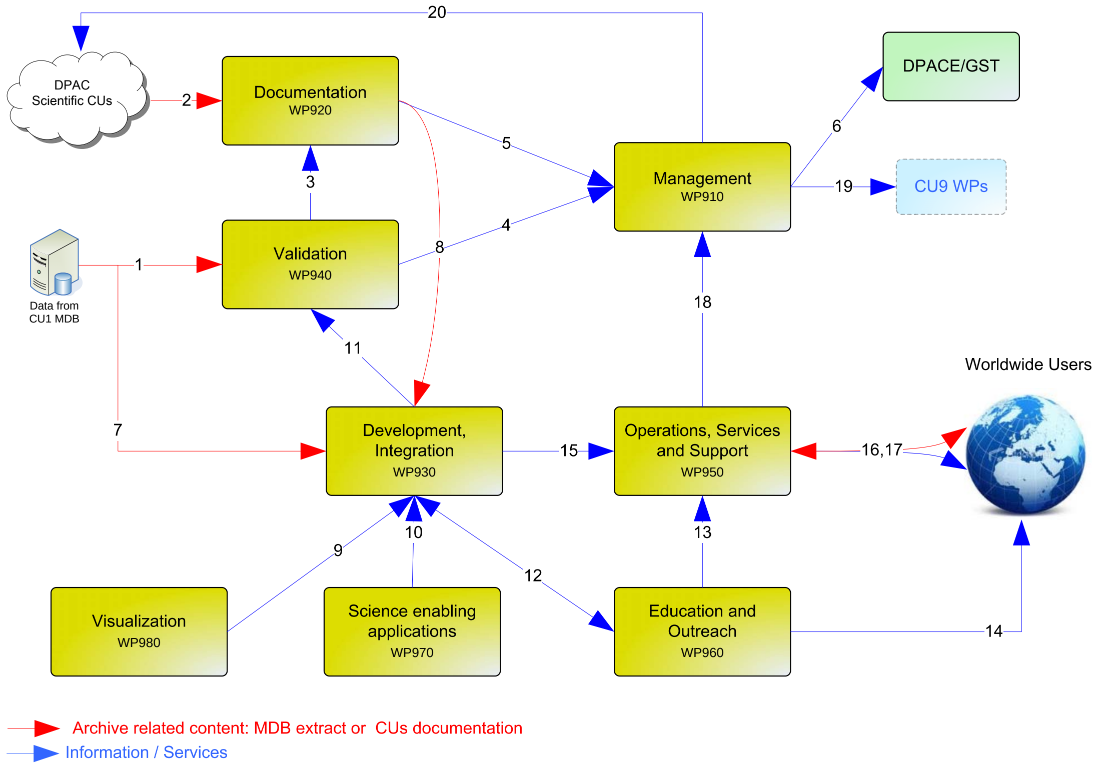


Figure 3: The work flow leading up to a Gaia data release.

The first step (1) in the work flow prior to a data release is that DPAC provides an extract of its Main Database (MDB) to the data validation team (WP940). In parallel (2) the documentation produced by the various CUs is delivered to the CU9 documentation team (WP920). This documentation describes the data products being validated within WP940. During the validation process some issues may be discovered, which in the worst case may prevent the release but which will typically lead to additional documentation (3). The efforts within WP920 and WP940 are rounded off with summary reports which are sent to the CU9 management (4, 5) as inputs to the data release acceptance process. This process will result in a recommendation to the Gaia Science Team (GST) and DPACE (6). The latter two bodies together decide on whether or not a data release goes ahead.

After GST/DPACE approval of the data release the MDB extract will be made available (7) to the team responsible for the archive development (WP930) in order to integrate the MDB results into the Gaia archive and catalogue. Note that in practice steps (1) and (7) will take place in parallel, both to enable the validation team to make use of the archive facilities and to expedite the data release process as a whole. Similarly the documentation is made available (8) to the archive team for incorporation into the Gaia archive.

The services on top of the archive will be developed as follows. The science enabling and visualisation teams (WP970, WP980) will deliver the latest versions of their tools to the archive team (9, 10). The latter will integrate these tools and make them available as services for use by the validation and outreach (WP960) teams (11, 12). The outreach team will provide content (13) to be integrated into the archive and services. In addition the outreach team may provide products directly (14) to the astronomical community or the general public (for example events, press kits, etc).

Once the archive integration team rounds off its activities the operations team (WP950) will take over (15) and assume the responsibility for providing (16) the archive and the associated services to the astronomical community and the wider public. This will take place through the Gaia portal and some supporting web sites[4] where the contents and services will be installed.

The users of the Gaia archive will be requested to provide feedback which will be collected (17) by the operations team who will assist in synthesizing the feedback into a report to be provided to the management of CU9 (18). This report serves as one of the inputs in the process of evaluating the Gaia archive performance and in planning its future evolution. The results are provided to the CU9 WP leaders (19) in order to guide their developments and planning for the next data release. In addition the leaders from the other DPAC CUs will get feedback (20) on the documentation they provided.

7 CU9 in DPAC and relationship to ESA

The data processing for Gaia is undertaken by the scientific community in Europe which has organized itself into the Gaia Data Processing and Analysis Consortium (DPAC). This consortium has been in place since 2006 and has the task to develop the data processing *algorithms*, the corresponding *software*, and the *IT infrastructure*, as well as to execute the algorithms during the mission in order to turn the raw telemetry from Gaia into the final scientific data products that will be released to the scientific community. The data processing activities are structured around (currently) eight 'coordination units' (CUs) and six data processing centres. Each CU is responsible for delivering a specific part of the overall data processing system for Gaia. The task of delivering the resulting Gaia catalogue and archive as well as the corresponding facilities is not currently part of DPAC. It is the aim of the efforts described in this proposal to fill that gap and bring into being the missing Coordination Unit CU9.

The announcement of opportunity clearly spells out that the Gaia archive access effort should be embedded in the ongoing DPAC efforts. We strongly believe that this is the only viable approach. DPAC is the only existing community that can support the deliv-

[4] for instance the CDS, where Aladin and Simbad will be adapted for Gaia, see Section 10.7.

ery of the Gaia archive as envisioned in Section 3. DPAC has a deep understanding of the Gaia data processing and is thus in a position of enabling the broad astronomical community to most efficiently use the Gaia archive. The actual data exploitation must be performed with query systems that are efficient and optimised for the nature of the data structure and the queries that will be performed, and such that these systems are minimally intrusive to the discovery process. This can only be archived in an archive system that is built in very close cooperation with the groups that are producing the data — a process that is even more effective if the groups are partners in the archive construction itself, as is the case for the present proposal. Finally, complementing the DPAC expertise with a large effort to develop science enabling applications and sophisticated visualisation tools is an essential task that we intend to fulfil with the proposed efforts. Leaving these tasks to individual groups would hamper a quick and efficient scientific exploitation of the Gaia data and would most likely lead to only a superficial exploitation of the data set. In addition, as explained above, these tools will be employed in the data validation effort.

7.1 CU9 in DPAC

CU9 will have the same Scientific and Technical duality outlined in WOM-001 and will follow the existing standards such as TL-001 and WOM-012. Hence the new DPAC organisational chart would look like Figure 4.

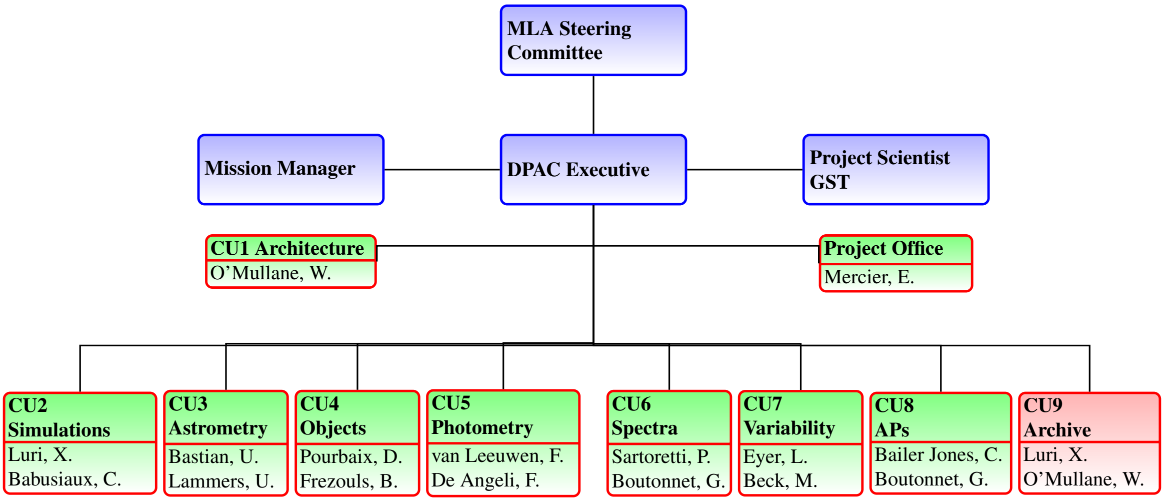


Figure 4: DPAC organisation including CU9 (in red). Data processing centres have been left out for simplicity.

This proposal arises from the Gaia Access Preparation Working Group and as such it has already been working within the normal DPAC framework. Many members are already contributors to DPAC while new members have already been exposed to the DPAC standards and processes. Thus we are confident that our proposed programme of work for CU9 will fit neatly into DPAC.

CU9 fits in DPAC but it is clear that some restrictions agreed in DPAC should not necessarily apply to CU9. In particular the single language constraint (JH-001) is not considered appropriate. The best technologies for data exploitation should be used and demonstrated by CU9, not one single set of tools in one language.

7.2 CU9 relationship to ESA

There are two main points of Contact between CU9 and ESA. For the public outreach and web contents we will keep in close contact with the ESTEC science communications office. Within the Gaia Archive Preparation (GAP) working group we have already had extensive interactions with the Science Archives and VO Team (SAT) in ESAC.

ESA's European Space Astronomy Centre (ESAC), near Madrid, Spain, hosts most of the ESA space based missions' scientific archives; in planetary exploration (Mars Express, Venus Express, Rosetta, Huygens, Giotto, Smart-1), in astronomy (XMM-Newton, Herschel, ISO, Integral, Exosat, Planck) and in solar physics (Soho). All these science archives are operated by a dedicated Science Archives and Virtual Observatory Team (SAT) at ESAC, enabling common and efficient design, development, operations and maintenance of the archive software systems. This also ensures long term preservation and availability of such science archives, as a sustainable service to the science community. Virtual Observatory activities are also fully part of the ESA archiving strategy and ESA is a very active partner in the VO initiatives in Europe and worldwide through the IVOA (International Virtual Observatory Alliance) and the IPDA (International Planetary Data Alliance).

The ESA Science Archives and VO Team is organized in a mixed structure, divided by projects for optimal vertical support and with a Core Development Team to ensure horizontal support activities. The project–specific development group, the SAT Gaia Team in this case, is devoted to developing and maintaining the Gaia Archive Core Systems. Apart from the specific project dependent members, there is a SAT Core Development Team composed of a selected group of experts that guarantee the uniformity of the archives, code reuse, horizontal support whenever applicable, legacy maintenance and some level of coordination of general SAT technical activities, ensuring synergy and preventing duplication of work. SAT follows a *Scrum* approach which is analogous to and compatible with the DPAC development approach (Section 6).

One of the early identified key areas for the CU9–ESA relationship is the correct integration of the different products and tools built throughout CU9. In this respect, a deep involvement of the stakeholders will be needed as the technical architecture will be developed by ESA (SAT), the so-called Gaia Archive Core Systems (as is shown in Figure 5), and that infrastructure will be used in many different ways for many different purposes by the rest of the CU9 participants.

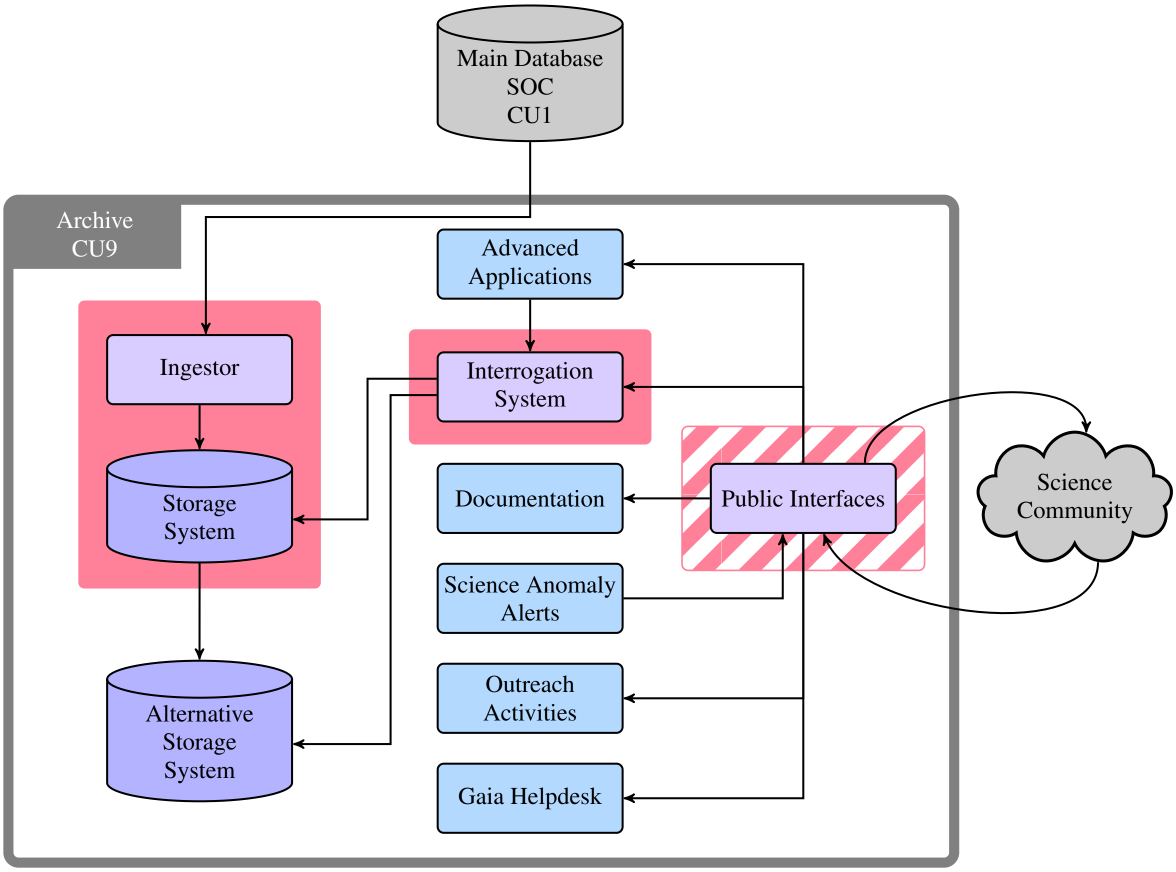


Figure 5: Gaia Archive Core Systems (highlighted in magenta) to be developed by ESA (SAT). It is important to remark that there may be other public interfaces as well.

The development approach must comprise a smooth and steady collaboration among all parties concerned, holding regular status meetings where the tasks and actions are specified and prioritized depending on the demands of the different groups.

Furthermore, system integration exercises will be carried out regularly to test the correct functioning of subsystems altogether, and internal workshops and conferences will be held in order to discuss and approach new technologies in a coordinated manner. It is also of great importance that proper quality assurance procedures are followed from the very beginning and that the software developed is as uniform as possible, considering that there will be some code reuse from SAT common system infrastructure as well as DPAC.

This approach will ensure a proper software development process, although agreements made throughout any of the interactions will be enforced by the creation of a relevant ICD defining the responsibilities of each group.

In order to ensure the proper ingestion of data and the integration of external code components, applications and libraries (which is going to be a key point for the Gaia Science Archive), SAT will contribute and encourage the definition of standard ICDs (Interface Control Documents). These ICDs documents will be crucial e.g. in the ingestion of data from other CUs (like CU1- Data Extraction and CU2- Simulations, CU3- Core Processing if applicable) and other packages within CU9 (e.g. Visualisation Tools).

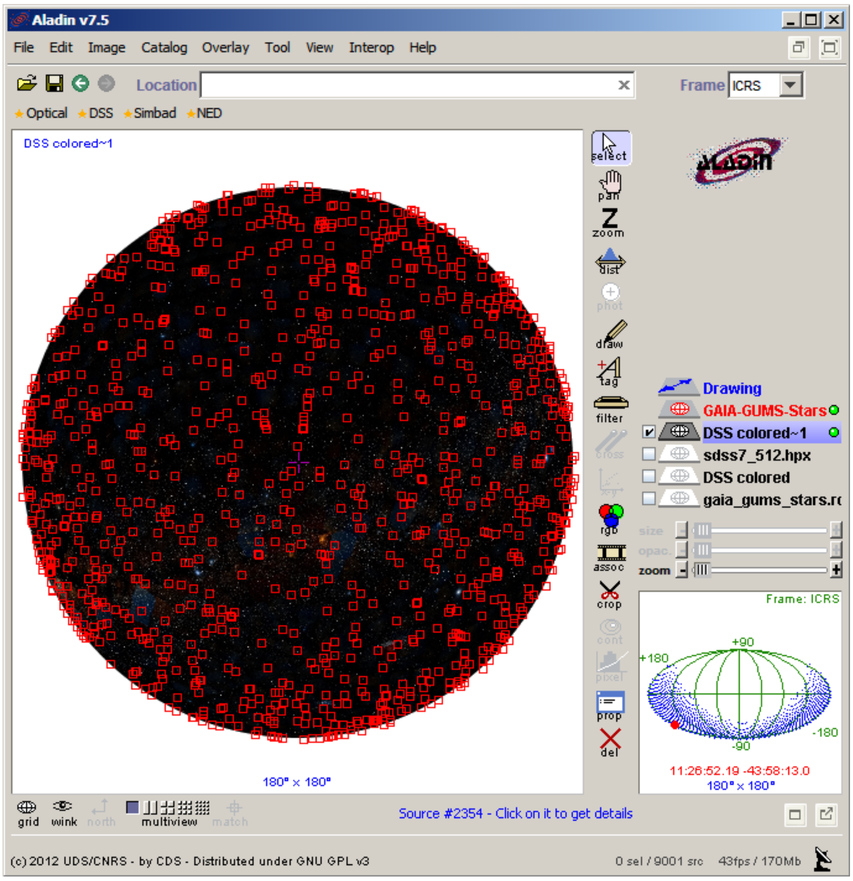
These guidelines are described in detail in the Science Archives and VO Team (SAT) Management Plan document (SATMP).

8 Data Rights in CU9

CU9 is part of DPAC and is bound by the Science Management Plan (ESA/SPC(2006)45) which states that there will not be proprietary data rights for Gaia. Therefore special care has to be taken to ensure that, before the official releases, the data is not used for scientific purposes. This is further reinforced by DPAC's own data access policy (FM-044) which is also binding for CU9, and will be especially delicate in this CU, which has the responsibility of making the processed data available.

Hence in the work within CU9 access to actual Gaia data will be restricted as much as possible before actual data releases. Instead simulated Gaia data generated in coordination with CU2 (see Section 10.5) will be extensively used for development, testing and validation of software. In the cases where real data needs to be used the members of CU9 will be bound by the publication policy (FM-039).

We have already been working with this principle for over a year within the Gaia Archive Preparation (GAP) working group. Indeed we have made simulations available to the world and demonstrated that CDS tools such as Aladin (see Figure 6) can handle billion star catalogues.



Source #2354 - Click on it to get details

(c) 2012 UDS/CNRS - by CDS - Distributed under GNU GPL v3 0 sel / 9001 src 43fps / 170Mb

Figure 6: CDS Aladin (Bonnarel et al., 2000) showing an all–sky view of the Gaia Universe Model Catalogue (Robin et al., 2012). This figure shows the progressive loading feature thus the all–sky view only shows *close* objects. It is a nice demonstration of collaboration in CU9, use of simulated Gaia data and availability of visualisation tools for billion star catalogues.

9 Schedule and Project Phases

This section details the CU9 phases leading to the generation of the catalogue releases and associated milestones.

9.1 Inputs

The main inputs to the CU9 schedule are the Gaia mission planning and the Gaia Science Ground Segment Operations Plan (JSH-033). The Gaia mission planning is based on the mission phases defined by ESA in ECSS-M-30B. With the current launch date Gaia mission phases are:

Time frame	ESA Phase	ESA Name
2013-2014	D2	Commissioning
2014-2016	E1a	Early mission
2016-2018	E1b	Late mission
2018-2020	E2	Extended mission
2020-2023	F	Post mission

The commissioning phase starts right after launch, which is currently planned to take place in the last quarter of 2013. CU9 activities will start no later than at launch. The Gaia mission duration will, obviously, have a direct impact on the production of the final Gaia Catalogue and the CU9 schedule.

9.1.1 Gaia DPAC processing plan and CU9 release scenario

At the end of the commissioning phase DPAC will start its data processing. Over the 5 years of the mission DPAC foresees at least the data processing cycles listed below, as described in detail in JSH-033:

Name	Duration	Time Frame
Data Processing Cycle 00	$\sim$ 10m	2014
Data Processing Cycle 01	$\sim$ 6m	2014-2015
Data Processing Cycle 02	$\sim$ 6m	2015
Data Processing Cycle 03	$\sim$ 12m	2015-2016
Data Processing Cycle 04	$\sim$ 12m	2016-2017
Data Processing Cycle 05	$\sim$ 15m	2017-2019

At the end of each data processing cycle, DPAC will produce a new version of its integrated main database (MDB), which incorporates all the data products delivered by the DPAC coordination units. These data products are derived from the raw satellite data.

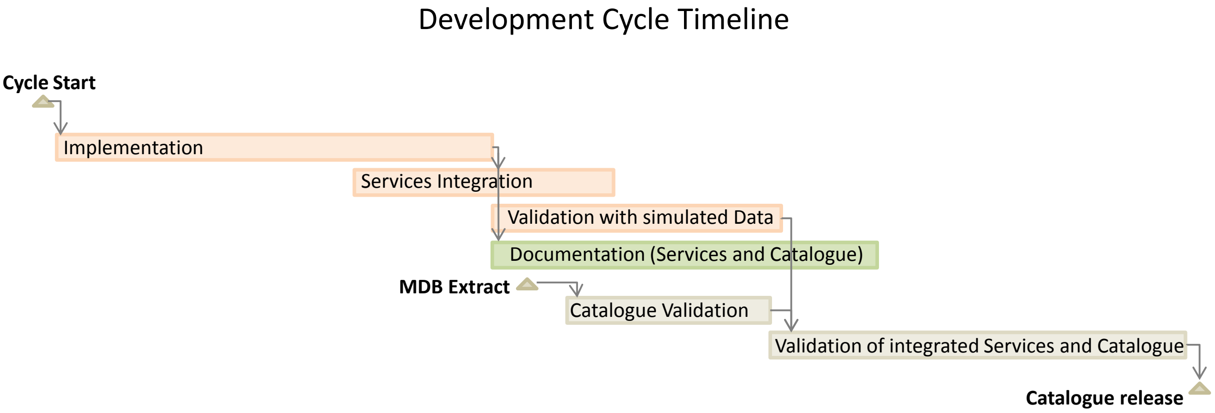


Figure 7: A typical CU9 development cycle.

An extract of this database *plus* the related CU data documentation will constitute the main delivery from the DPAC processing CUs to CU9 and will mark the start of the data release process. Changes in the DPAC processing schedule will impact the high level CU9 milestones and the releases of the archive and catalogue.

The current CU9 data releases baseline, outlined in the document 'Release scenarios for the Gaia archive' (WOM-066) assumes a release between 3 and 4 months after the end of the first four processing cycles while the last release would be up to three years after the end of the mission.

9.2 Development Cycle Phases

The overall CU9 schedule will start at launch and will finish with the final Gaia catalogue release. We note that the final data release and the services associated with it will last, in principle, indefinitely, but this extended period is not within the scope of this proposal.

As described in Section 6 the CU9 development will follow an iterative approach. Some CU9 cycles will contain catalogue releases. Between releases, several iterations are planned for integration and internal validation purposes. The following activities will be done during a CU9 cycle. The first 4 steps will be done in each cycle, while 5 to 8 are only performed in those cycles leading to a catalogue release.

Figure 7 shows schematically the activities which will be carried out during each development cycle.

1. Implementation of the services
2. Integration of the services

3. Validation of the integrated services with simulated data and/or earlier MDB extract
4. Testing and integration of the services into the Gaia portal and supporting web sites
5. MDB extraction
6. Catalogue validation
7. Catalogue and integrated services documentation
8. Validation of integrated catalogue and services

Once the catalogue is released it will remain available indefinitely. The services and science enabling applications are considered to be in operation until the next release is available. Typically CU9 will handle activities of two releases in parallel. It will operate a version of the Catalogue/Archive and associated documentation while preparing the next release activities.

A CU9 release is composed of the catalogue and archive and its associated science enabling applications and services plus documentation and visualisation tools. The end of the integration of the catalogue and services into the Gaia portal and supporting web sites leads to the CU9 Release *actual* milestone. The MDB data extraction, integration and validation of the catalogue and the services will be on the critical path for every CU9 release.

10 Work Breakdown

The CU9 effort is divided into several 'work packages'. Each major work package is described in the sub sections that follow and details at the sub workpackage level are provided in a separate document, Appendix C (although as an additional content and not part of the formal response to the CU9 AO).

Table 2: Summary of effort (in Staff Years) per year to 2017 and cumulative for 2018-2022 for top level Work Packages

WP	**Description**	2013	2014	2015	2016	2017	2018- 2022	**Total (SY)**
910	Management	1.8	1.8	1.8	1.8	1.8	9.0	**18.0**
920	Documentation	1.5	1.4	1.1	1.4	2.5	3.5	**11.4**
930	Architecture Design/Development	7.35	7.85	7.6	4.9	4.3	19.5	**51.5**
940	Validation	6.5	8.4	7.9	8.4	10.4	29.7	**71.3**
950	Operations Services and Support	2.8	3.9	3.9	3.9	3.4	13.3	**31.2**
960	Education and Outreach	2.2	2.3	3.5	6.5	3.0	11.9	**29.4**
970	Science Enabling Applications	7.75	9.85	9.95	10.15	8.95	38.55	**85.2**
980	Visualisation	5.5	5.7	5.7	5.7	5.4	21.7	**49.7**
Total								**347.7**

Table 2 provides an overview of the effort estimates for these packages while details at the sub workpackage level are given in Table 4. In addition Table 3 provides a breakdown per institute in the proposal. A **type flag** has been included in these tables in order to target in a more flexible way the effort required at this still provisional stage, pending funding confirmation by the national agencies to the groups participating in this proposal:

Type 1: mandatory work packages, strictly needed for the working of CU9 and that should be kept in all cases

Type 2: in case of real necessity some of them can be dropped at the price of only partially fulfilling some of the CU9 goals

Table 3: Work package effort estimates per institute in Staff Years for the entire project 2013-2022.

Institute	Type 1	Type 2	**Effort (SY)**
Univ. Barcelona	25.4	24.7	50.1
ESAC	32.1	13.1	45.2
IT(ASDC INAF Univ)	16.1	27.0	43.1
ARI	6.3	21.8	28.1
ObsPM	22.7	3.5	26.2
MPIA		17.0	17.0
Lisbon Univ.	14.1		14.1
IoA Cambridge	7.4	3.7	11.1
Univ. Coruna	3.0	7.5	10.5
Vienna	2.0	8.0	10.0
Geneva	7.0	2.7	9.7
OCA	8.0		8.0
INTA	1.0	6.5	7.5
AIP	4.0	3.0	7.0
CDS	4.8	2.0	6.8
Groningen	3.0	3.4	6.4
Innsbruck		5.0	5.0
ARI/HdA		4.9	4.9
IfA Edinburgh	2.2	1.8	4.0
Besançon	4.0		4.0
KU Leuven		3.0	3.0
Lund	3.0		3.0
UNINOVA	3.0		3.0
PO	2.2		2.2
Royal Obs Belgium	1.0	1.1	2.1
ESTEC	2.0		2.0
Groningen/Leiden		2.0	2.0
DPAC-PO			2.0
DPAC PO	1.0	0.9	1.9
Univ. Coruna/Vigo		1.6	1.6
INAF-OABo		1.4	1.4
Laboratoire d'Astrophysique de Bordeaux			1.2
Bristol	1.2		1.2
Leiden	1.0		1.0
INAF Bologna Obs	0.9		0.9
Milan		0.5	0.5
TOTAL	178.4	166.1	347.7

As defined for DPAC in (WOM-001) estimates in work package descriptions are given in Staff Months (SM). A Full Time Equivalent (FTE) is considered to provide 10 months of effort in one year. Hence to convert estimates to FTEs one should divide by 10. The tables are provided in Staff Years for convenience (and are labeled as such).

10.1 WP910 – Management

10.1.1 DPAC framework

As described in Section 7 the CU9, as required in the AO, will be part of DPAC like other CUs. It will have the same Scientific and Technical duality outlined in WOM-001 and will follow the existing standards such as TL-001 and WOM-012.

First of all, following TL-001, the CU9 has to identify three key persons to lead the CU9 (CU-L, CU-T & CU-Q). The proposal for the start of CU9 is:

CU leader: in charge of the overall management of CU9. X. Luri, editor of this document.

CU technical leader: in charge of the overall technical coordination of CU9. W.J. O'Mullane, co-editor of this document.

CU9 quality assurance: in charge of QA for CU9. G. Gracia, acting as project management support and QA manager, sharing time with DPAC PO tasks.

As in the rest of DPAC CUs, the leadership of CU9 can be changed by internal agreement between its members and posterior notification to DPACE.

Again following TL-001, Development Unit leaders have to be identified for each major work package. The proposed initial DU-Ls are listed in Table 1. Furthermore, tentative managers for lower level work packages are listed in the appendix document (Appendix C), although not as part of the formal proposal.

In addition, to ensure effective coordination with the external science community, and reflecting the science driven nature of CU9, the coordinator of the GREAT consortium (N. Walton) will attend CU9 management meetings, providing an interface with the wider Gaia GREAT science community.

The CU9 leaders will monitor the work and ensure that the development of CU9 follows the schedule defined in Section 9. The development will be tracked through:

Formal documentation: CU9 will follow the adapted ECSS standard used by DPAC. The development can be tracked at a formal level through the set of standard documents defined there. A documentalist will be named (following the WOM-012 requirement) to handle CU9 documentation. Final versions of documents will be approved by the CU9 CCB (see below) and put in the DPAC Livelink system.

Tracking of milestones and schedule: as described in Section 6.1.3.

Regular meetings: several levels of meetings will be held regularly as described in Section 6.1.4. These meetings will allow the CU9 management to stay fully updated on the development status of CU9. In addition, ad-hoc teleconferences for specific topics will be held on an as-needed basis.

This tracking will be carried out in close coordination with the DPAC Project Office, as it is currently done with the already existing CUs. Furthermore, the CU9-L will also report to the DPACE on CU9 progress at each DPACE meeting and the DPAC wiki will be used for collaboration and dissemination of information within CU9.

At an internal level, the software development will be coordinated, also following the DPAC practices, by a Configuration and Control Board (CCB, as described in WOM-012) in charge of issue management:

- Keep track of the software issues, that will be reported and tracked through the DPAC Mantis system.
- Approve software change requests (but not bug fixes) to CU9 products
- Check and approve all software release notes.
- Approve CU9 ECSS documents, including and specially the SRS.

10.1.2 External reviewing

The previous section has covered the aspects of the management of CU9 as part of the DPAC. However, considering the special role of CU9 as the deliverer of the Gaia data to the overall scientific community we consider it necessary to add an advisory body, the **External Advisory Board** in order to provide independent advice from experts to avoid the risk of becoming a development community that is too insular. The role of the advisory board will be to carry out an independent (outside DPAC) overview of the CU9 progress and it will be formed by leading experts in astrophysics and engineering. Specifically, we aim to include senior scientists involved in the preparation and delivery of the Hipparcos Catalogue — in order to benefit from their experience in tasks closely related to the ones in hand for CU9 — as well as experts involved in other scientific projects with massive data deliveries, in order to benefit from their experience in large-scale repositories of science data.

The External Advisory Board will receive the relevant CU9 documentation allowing its members to track the project and will participate in the plenary CU9 meetings. It should produce a short report with a critical review of the CU9 status and advice for the continuation of the development.

10.1.3 User-driven development

The mission of CU9 is not only to bring the Gaia data to the scientific community but to do so in a way that facilitates its use and maximizes the scientific exploitation of the archive. To achieve these goals the development of the CU9 systems has to be user-driven rather than technology driven, and therefore the requirements to build these systems have to be captured from the actual users. The process to capture requirements is described in Section 5 but in addition to this, somewhat limited, interaction with the user community we also intend to involve the latter in the work of CU9 through:

- Selecting some Gaia archive users to act as our "beta testing team", so that the systems are thoroughly tested by actual users and not only by personnel already involved in its development.
- Making part of the CU9 plenary meetings open to the community and including special presentations of the CU9 systems, sometimes including hands-on exercises on the archive.
- Additionaly, workshops or schools dedicated to the efficient use of the Gaia Archive Tools can be organised to improve the skills of the community in the use of the archive.

10.2 WP920 – Documentation

The documentation will cover the following aspects:

1. A description of the various data types, including contents and an overview of statistical properties
2. A description of the access facilities to those data types

3. A description of how the data types have been derived, including a detailed description of the processing.

These different aspects will each develop over the various releases as needed. In other words, every release will (correct and) extend the previous release of documentation, to cover new data types and updated statistical properties.

The documentation activities require a very sound knowledge of the data, data reduction methods, or in case of point 2 mentioned above, of the data release facilities. For that reason, WP920 will rely heavily on contributions from experts in the various CUs, and staff effort assignments need to be agreed between CU9 and the data processing-related CUs. The current estimate is that over the mission this will amount to one to two staff years per CU. There is in addition the need for 2 to 3 staff years of effort from CU9 directly for coordination and editorial activities.

The documentation needs to be fully accessible in electronic form, and as such linked in with the data access facilities. It is therefore important to coordinate these two activities from the start.

Preliminary staff effort requirements summary for a period of 6 years, spanning from 2016 to 2022, are:

- Seconded from other CUs (including WG activities such as GBOG and RTF) 9 staff years;
- CU9 specific man power: 3 staff years (0.5 FTE), preferably spread over two persons.

Details of sub workpackage effort levels are given in Table 4.

10.3 WP930 – Archive architecture design and development

As shown in Figure 3 of Section 6.2.3, WP930 plays a core role in the product flow of CU9. In particular, this work package will involve:

- setting up the overall system architecture of CU9;
- defining the interfaces between the Gaia core systems at ESAC and the services developed by DPAC institutes;
- developing and integrating the core components of the Gaia Archive at ESAC;

- integrating the services developed by other WPs and making them ready for Operations (WP950).

As described in Section 6.2.2 a CU9 Release will be made of the core systems developed by the SAT Team at ESAC and the different services from WPs 970 and 980 as well as the product developed by WP960. The Gaia core system at ESAC will be composed of a database and its related query engine as detailed in Section 10.3.2. The interfaces between this core system and the numerous CU9 services are presented in Sections 10.3.3, 10.3.4 and 10.3.3. Furthermore, the overall system must be Virtual Observatory compliant and therefore will include a VO layer and support for the relevant metadata to drive application functionality.

The technology choices and the design of the systems should be carefully based on the real user needs, as explored and defined in WP910. The activities of WP930 will follow the overall CU9 strategy based on a phased delivery of systems presenting increasingly rich functionality over the incremental Gaia data releases (see Section 6).

10.3.1 WP931: Management

The purpose of this sub–workpackage is to ensure that the work undertaken is well coordinated, and that the requirements and goals are met within the context of the project as a whole.

10.3.2 WP932: Gaia archive core systems

Terabyte–scale, billion–row survey catalogues of the current generation are served to the community using conventional relational technology, e.g. Szalay et al. (2001)[5] and Hambly et al. (2008) and references therein. This brings many advantages for applications design, and, more importantly, users have become accustomed to the power provided by Structured Query Language (SQL) access to the full relational schema for a complex database, rather than being restricted to certain access patterns and query types implemented as webforms. The success of these systems suggests that a similar conventional RDBMS could be considered as the baseline for Gaia, but several newer developments in database technology are particularly relevant and should be exploited. Firstly, in recent years there has been a growing realisation that many scientific databases – which typically feature wide tables, within which a small number of columns are very popular – are well suited to column-oriented databases, which minimise the data I/O needed for queries on small numbers of columns that are typical in scientific database workloads. Secondly, the Gaia project at ESAC is using a high–

[5] http://research.microsoft.com/apps/pubs/default.aspx?id=69896

performance object–oriented database, Intersystems Caché[6], for the persistence of Java objects in the core processing systems and, since it also supports an SQL interface, it would be advisable to see whether this same system can support the user scenarios developed through WP910, to assess whether a single DBMS solution can be used for both processing and end–user access.

Instances of various DBMSs will be implemented on identical hardware and loaded with a testbed data set of sufficient size that a realistic query workload can be executed upon it. Different DBMSs are likely to perform best for different types of query: as noted in the Science Requirements Specification for the *Interrogator* component of the Gaia archive (Tapiador, 2011) *'It is assumed . . . that there will not be a single technology that will out–perform the others in all the different use cases foreseen . . . there will probably be more than one solution in operation . . . '*, and that *'. . . there might be cases where two or more back–ends are put in place . . . so that different queries might be redirected to the most appropriate back–end in terms of performance or other service.'* The outcome of this work will be a detailed understanding of which technology performs best in each part of the expected Gaia archive workload. In addition to performance against the benchmark workload, it is important to assess how easy it will be to load data from the Gaia Main Database (MDB) into each candidate for the archive DBMS. In the language of the 'Gaia Catalogue and Archive Software Requirements and Specification' (WOM-033), we need to prototype the implementation of an *Ingestor* for each system – see for example DTP-002 – and then benchmark each.

There are several areas of work (for further details, see WP932 in the appendix document) that can be summarized as those related to data management: Gaia catalogue and metadata access, ingestion from the SOC, archival storage for data products and Cloud services.

10.3.3 WP933: Consortium/ESAC–SAT interface control

WP932 describes the 'in–house' work that ESAC must undertake to deliver generic archive core systems for all ESA missions, including Gaia. It is anticipated that certain specialised server–side software subsystems for Gaia will be required that need contributions from CU9 (see the following sections). It is clearly important to establish early on an effective coordination policy to avoid interface mismatches, duplication of effort and other problems. This sub-WP aims to do just that, via an agreed 'interface' (in the broadest sense) coordination document. For further details, see WP933 in the Appendix document.

[6] http://www.intersystems.com/cache/

10.3.4 WP934: Database collaboration

ESA now routinely publishes data from its astronomy missions through archives located at ESAC, and Gaia will be no exception. The Science Archive Team at ESAC are developing a common archive architecture, which is being employed for mission archives from ISO, through XMM–Newton and Integral to Herschel, as well as a number of planetary science missions. These are all observatory missions, comprising separate pointed observations, so that the functionality that must be delivered centres on the selection and extraction of standard data products from pointed observations. Since Gaia is a survey mission its requirements (WOM-033) are quite different. While there will be some data product extraction activities (e.g. downloading RVS spectra), many of the most important analyses of Gaia data will involve sophisticated queries on source parameters in the catalogue (e.g. selecting sources in a particular region of the sky with certain colours and kinematic properties) or will require specialised tools running on those parameters (e.g. variability measurement and characterisation), and many will also require the use of data from other archives (e.g. photometry in different passbands), accessed through the Virtual Observatory (VO). All of this, together with the fundamental fact that the natural organisation of the data in the Gaia Catalogue is by celestial source, rather than by individual observation, makes it clear that the Gaia Archive must differ greatly from a 'standard' archive architecture, and that the bulk of the additional Gaia–specific design and development work must be contributed by DPAC through CU9: the SRS notes that CU9 can assume that technical support from ESAC will be included in the planning. This support will be in the back end data store, so the main CU9 contribution must be in the provision of infrastructural support for large–scale survey science, which is an area where the non–ESAC groups contributing to WP930 have considerable expertise. For further details of the proposed contributions, see WP934 in the Appendix document.

10.3.5 WP935: Database interface design

The Gaia archive has to cater for a variety of demands from the community, and thus needs specialised interfaces customized to these demands. It will be necessary to design these interfaces along the requirements and use cases already gathered, and fit these to the capabilities of the archive systems. Then, the implementation of these interfaces has to be done, and the feedback of the community will require iterations of the requirements over implementation cycles.

With the Table Access Protocol (TAP[7]) the VO provides a standard means of querying tabular data sets, and with the advent of the TAP factory (Hume et al., 2012) it has become possible to execute multiple, distributed TAP queries. In a traditional IVOA TAP scenario, single TAP endpoints provide the means for VO clients to present the user with a data resource schema and then to service an ADQL query on that resource, but it is

[7] http://www.ivoa.net/Documents/TAP/

then up to further, separate client–end manipulations to join data for multiwavelength science. TAP Factory takes this further by combining TAP with the Open Grid Service Architecture Data Access Infrastructure (OGSA–DAI) middleware to provide a means of creating TAP end-points on–the–fly, and, thereby, facilitating the cross-querying of distributed resources by TAP clients.

Such a system supports one of the fundamental usage scenarios for the VO. A user can select a set of data resources published using TAP on which to execute a distributed query. From the metadata exposed by the individual TAP services, TAP Factory is able to create a new TAP end-point on–the–fly for the distributed query and present the user with the metadata of the virtual data federation thus generated. The user can then pose a query against this virtual federation as if querying a single TAP service, and, when coupled with the MyDB–like personal database, it enables users to create sophisticated sets of cross–catalogue queries, as required for the full exploitation of Gaia data. The key point here is that a data resource can be incorporated into a virtual federation without requiring any action on the part of the staff of the data centre that curate it; so, in the case of Gaia, it is possible for higher level services like these to be developed and deployed, without requiring any action from (or placing any obligations on) the staff at ESAC. For further details, see WP935 in the appendix document.

10.3.6 WP936: Correlation functions

It goes without saying that all the basic source parameters in the Gaia archive – and in particular the astrometric parameters – will be provided with their associated standard errors, or other measures of uncertainty, as calculated in the normal data processing. This is important in many investigations, where the significance of a result may depend critically on the estimated uncertainties of the derived quantities.

When the calculated quantity depends on several astrometric parameters, it is in general necessary, when performing the error propagation, to take into account not only the standard errors of the parameters but also their statistical correlations. This is especially important when combining data for many sources, where even a rather small positive correlation ρ effectively prevents the expected $N^{-1/2}$ improvement with the number of sources beyond some moderate number $N \simeq \rho^{-1}$ (Holl et al., 2010). The functionality implemented with this work package thus caters for a data access scenario such as *'I want to average astrometric measurements over groups of sources and account correctly for the star-to-star correlations'*. Ideally the Gaia archive should include the complete covariances of all the data but this is clearly infeasible from a data management viewpoint and also undesirable because only a small subset will be needed in typical cases, and then only certain functions (quadratic forms) of the subset. The proposed strategy is therefore to compute the covariance of the desired quantities (e.g., the mean parallax and mean proper motion components of a group of sources) directly as needed for a particular application.

This work package belongs to the Archive architecture design and development because it will likely require some very specific and potentially complex access patterns, involving many more sources than the ones for which the correlations are requested. The physical origin of the star-to-star correlations is mainly the attitude errors in the core astrometric solution, which affect all the sources observed at a given instant of time in a similar manner. However, these attitude errors in turn depend on the astrometric errors of all the (primary) sources that contribute to the attitude determination at that instant. To estimate the covariances for just a few target sources it is thus necessary to consider at least all the primary sources that were observed roughly simultaneously with the target sources, and possibly a much larger set must be considered recursively, depending on the required accuracy of the estimate. A possible algorithm is described in Holl & Lindegren (2012) and Holl et al. (2012).

In the simplest case the user interface for this task would ask for a list of the astrometric parameters to be combined together with the Jacobian matrix for the desired transformation, and would return the estimated covariance matrix for the transformed variables. However, a more useful and elegant approach would be for the user to supply only the transformation itself, in some high-level code, whereupon the archive system carries out not only the transformation itself but also the associated error propagation, using a built-in algebraic manipulator to calculate the Jacobian.

10.4 WP940 – Data validation

'*To contribute usefully to the scientific progress, one must sometimes not disdain from undertaking simple verifications*' (Foucault, 1847). Indeed, despite the precautions taken when building the data processing, completely avoiding errors in a one billion source catalogue, with many intricate data for each object, is an impossible task. Before the publication of the Gaia archive, an in-depth validation of its contents should thus be undertaken. This will rely on methods and tools developed in this or in the other CU9 Work Packages, and which can also be used, with little or no adaptation, by the scientific community for the analysis of the catalogue.

Not only the formal responsibility but also common sense dictates that Gaia data validation should occur within DPAC, not outside: establishing evidence that the output Catalogue meets the Gaia mission requirements necessitates knowledge of various aspects of the mission; understanding possible problems, in particular those coming from instrumental effects, requires an expertise more likely present within DPAC.

Also for the Gaia predecessor, Hipparcos, were substantial efforts dedicated to data validation: not less than 4 chapters are devoted to the validation aspects in the printed

Catalogue (Perryman & ESA, 1997). Several people within the Gaia community participated in this validation (see e.g. Arenou, 1993; Lindegren, 1995; Brown et al., 1997; van Leeuwen, 2007). This experience has shown that users easily misinterpret the data content, which is statistical by nature. Validation also has the indirect effect of assessing both astrophysical and statistical properties of the Catalogue.

10.4.1 Principles and boundaries

The validation will start where the verifications stop. All DPAC Coordination Units have implemented their own internal tests to verify their intermediate (DU) or final outputs. Validation can be compared to an integration test while each CU verification can be compared to a unit test, both being necessary and complementary. While the verifications within CUs will primarily use their own data, about which the CUs have the best expertise, the CU9 validation perspective should be transverse, benefiting both from external and *cross-CU* data in order to check that DPAC produces a consistent Catalogue.

The validation should stop where the science exploitation starts. Validation requires an in-depth, though mostly statistical, view of the Catalogue in order to check that there are no remaining artifacts from the data reduction process. Still, it is likely that outlying characteristics are uncovered, which can be attributed either to some interesting scientific property, or to some processing artifact. If this latter explanation can be ruled out, no science exploitation should be pursued as this is prohibited by the Gaia data right policy. As requested by the CU9 Announcement of Opportunity (Prusti, TJP-013), participation in Gaia data validation, like any other early access to the reduced data, is an explicit acceptance of this policy.

Obligation of means vs obligation of results. Without any doubt, there *will* remain errors in the Catalogue, and it is also very true that the validation process will be unable to spot all these errors. Still, the validation process has an obligation of devoting enough effort and putting priorities on what is the most important to validate, and this must also be phased with the content of each successive data release, thus broadening with time. However, even when validations are performed, their conclusion will not always be unambiguous. This is illustrated with the obvious example of the correlation at small angles of Hipparcos measurements: although anticipated in 1988, documented and accounted for in the scientific exploitation in 1998, their effect is still being disputed since (and the outcome will have to wait for Gaia). The last word is always the one from the scientific community and this word can take some time to be articulated.

10.4.2 Work areas

The definition of what a systematic error is will lead to checks for suspect spatial (or epoch, colour, magnitude, ...) variations suggesting that the data under scrutiny may have processing artifacts. The validation will address the statistical distribution of the astrometric parameters, typically the unbiasedness of the parallaxes and proper motions, as one of the main outcome of the mission, and which may be subject to many perturbations. It will also scrutinize the photometry and spectroscopy data, assisted by the important classification task performed in Gaia, and, when available, external data. The unique point of view of CU9 will allow use and validation of all the Gaia scientific products simultaneously.

Although the Catalogue can be accessed in many different ways, it will be of obvious interest to use the different tools developed within CU9 to check the data content and thus apply the tools in an operational context. The validation task can thus not only validate the Catalogue but also indirectly contribute to validate the Services.

10.4.2.1 WP942: Internal consistency and more complex scenarios Gaia is a complete observatory in orbit, combining astrometric, photometric and spectroscopic information, and this implies some redundancy which can be exploited for validation purposes; for example, photometry should be consistent with spectroscopy. Other intrinsic correlations between parameters can be used to build those tests, such as e.g. the dependence of proper motion on distance.

More elaborate scenarios are also needed with the perspective of investigating what kind of problems could occur and what consequences (systematic errors) this would have on observed parameters, typically instrumental problems, calibration errors, data processing shortcuts or approximate models.

10.4.2.2 WP943: Comparing models with data. Models have been used in the DPAC CU2 (the 'Universe Model') for the training of the data reduction algorithms. They can also be used for Gaia data validation, provided that large enough margins are accounted for. Once projected in the observable domain, confidence intervals, correlations and other statistics for the parameters can be computed to which the actual parameters observed with Gaia can be compared. Working with truncated, censored or correlated data, in some limited magnitude range, with a relative precision censorship, will however require adapted statistical tools. Not going into details, rather checking whether the large, expected structures are present, can then help to uncover unexpected residual effects. And, again, the goal is only to detect possible artifacts and no further scientific exploitation may be done.

10.4.2.3 WP944: Confrontation with external archives. Thanks to the use of external and potentially more precise data, it may be possible to search for possible biases within the Gaia data. This applies e.g. to stellar photometry or special objects like binaries or solar system objects. The caveat is that validation is supposed to be the validation of Gaia data, *not* that of the external archives. Robustness is thus mandatory in front of the lack of data, the lack of precision, and the high level of systematics that may be present in external archives and which could wrongly be interpreted as problems within Gaia data.

10.4.2.4 WP945: Statistical tools: data mining and data demining. Outliers being by definition objects which deviate from an assumed model, it would be surprising that a mission such as Gaia planned for deciphering the complex structure of the Galaxy would exhibit no outliers departing from our current knowledge. Data mining tools using artificial intelligence techniques will be used to identify outliers and check whether they originate from artifacts. Not only outliers but also unexpected correlations between astrophysical parameters can also unveil the presence of systematic errors. In this way the first order risk associated with the presence of problems or systematic errors in the Catalogue will have been handled, yet another issue is an incorrect interpretation of data features. For example, although the community was warned (Brown et al., 1997) about the precautions to be taken with the analysis of the Hipparcos data, this did not prevent incorrect exploitation of the astrometric data. In this respect, being able to show that objects or substructures are not outlying is perhaps as important and should be present in the documentation.

10.4.2.5 WP946: Time series and variability While a large fraction of the sources are expected to be intrinsically variable in flux, some assumed constancy of many other sources is also what permits the principles of the data reduction. Conversely, an unexpected variability can also be the signature of an acquisition or data reduction problem. As variability is transversal to the validation process, this work package will validate the cross-CU astrometric, photometric, spectro-photometric and spectroscopic reduction from the point of view of time series and variability, e.g. to detect if constant sources, or small amplitude variables have some residual effects coming from the satellite or perturbations from that data acquisition mode, or reduction method.

10.5 WP950 – Operations and Services

WP950 Operations and Services is underpinning contact between the Gaia mission and the worldwide community of users. WP950 Operations will sustain availability of the Gaia facilities that consist of the data archive, its associated services and science enabling applications. The archive related activities will be performed by the ESAC SAT team while the development and operations of most services and applications will be provided by a set of European institutes involved in CU9.

In addition to the core task of ensuring the operation of the systems, WP950 also includes the task of provision of simulated data for development, testing and validation purposes (Section 8). These data will be generated in coordination with DPAC CU2, using mainly the GOG data generator developed by this unit.

The operations of CU9 are driven by several principles:

1. availability through a single access point of the CU9 services
2. reliability of the CU9 services
3. user orientation of the CU9 services
4. improvement of the CU9 services

Single Access Point: The CU9 goal is to provide a *single* facility allowing access to the archive and the CU9 services. This single facility should be a public interface - typically an internet portal - from where the users will use online tools and applications, download some of them for offline use as well as retrieve the data from the archive. In a more advanced scenario there may be mirrors of the catalogue (and metadata) which could be utilised by smart tools; if possible this should be transparent to users.

The development of a useful facility will require close collaboration between ESA and the community - we include a work package specifically for this. Hence the operations of CU9 imply the operations of the archive itself and any partial mirrors, as well as the operations of the CU9 services. WP950 will have to perform integration activities for the CU9 services and the facility itself. The data archive and some core CU9 services will be integrated at ESAC by WP930 prior to a delivery to WP950.

Implementation of a single access facility implies several features:

1. the access to the data and services should remain free for all users. This includes:
   (a) the overall data archive as stored at ESAC
   (b) the Advanced data access tools as on line services or downloadable for an offline use
   (c) the Data Mining tools as online services or downloadable for an offline use
   (d) the cross–match with external catalogues as online services or downloadable for offline use
   (e) the visualisation tools as offline or online tools
   (f) any outreach content
   (g) a documented API to allow programmatic access to Gaia data resources
   (h) any Newsflash including for instance some Gaia science alerts information

2. the operations shall try to ensure that any interruption shall be fixed within 3 hours of the interruption. If this is not possible, perhaps due to specific circumstances outside of the normal control of the Gaia archive, then users will be advised as to the actions being taken to restore service.
3. the architecture of a portal will conform to current internet standards and the content shall include some dynamically generated information which shall be updated on a weekly basis at the longest
4. the design plus look–and–feel of any services or facility shall relay a clear and recognizable Gaia identity
5. interfaces shall be designed such that the professional astronomers as well as the general public will find the requested services
6. all agencies contributing to CU9 shall be given visibility in a contributors area with display of logo, names and links etc.
7. the possibility of providing the release content information in more than one language shall be considered

Reliability: The services provided by CU9 shall be reliable. WP950 will act as a DPAC Data Processing Centre (WOM-017) and will have the responsibility to approve the incoming products (archive, tools, content, ...) through a quality assurance acceptance process. Only after this acceptance review should the products be put into operations. Any static content will also be under configuration control. At each new release, a new version of the facility will be issued (v.1.0 for the first release) embedding all the new services and data. Any major update made to the facility in between two releases will increment the revision number. For instance, the first update after the first release will be version v.1.1.

The following prerequisites will aid in a reliable system:

1. the inputs of WP950 shall be reliable: any products to be operated by WP950 will be subject to an operations acceptance by the WP950 leader and the CU9 QA officer
2. the services as operated by WP950 shall be closely monitored: any technical problems will be tracked through Mantis and their closure followed up an reported to WP950-L
3. any user reports concerning an error shall be traced, acknowledged and properly resolved

4. the DPAC configuration approach (WOM-012) will be applied to the WP950 products

Responding to changing user requirements: WP950 is a user oriented work package - there will be permanent contact with the Gaia users to improve the overall quality of the CU9 services. We have some user scenarios now but understand these will look dated in a few years - we must not be restricted to this limited view of a possible archive.

Feedback will be collected via the web but also through some dedicated workshops planned around the releases dates. WP950 will issue regular user feedback reports including recommendations on some possible improvements to be performed. The implementation of those changes will be steered by CU9-M.

Improvement of services: In order to improve its services, CU9 Operations will:

1. implement the community approach initiated by the management of CU9
2. provide the user with some registration facilities
3. provide the users with some information on the community of users
4. allow the users to access any scientific documentation related to Gaia
5. maintain a help desk for technical and scientific matters
6. maintain a FAQ section
7. provide the users with a forum to exchange on some preselected topics
8. acknowledge the reception of any user request for technical or scientific matters
9. organize user feedback workshops in order to collect their feedback on a face to face basis

WP950 is structured in the sub-work packages listed in Table 4.

10.6 WP960 – Education and Outreach

The CU9 WP960 will provide the “face” of the Gaia project and its products not only to the interested science community, but also to the interested sections of the general public. The popularity of astronomy and space sciences can be seen easily by the large number of space-related articles in the media, the significant number of amateur astronomers and the huge interest amongst the general public. In this context, the Gaia mission offers a unique opportunity to bring those disciplines closer to the public. Outreach in Gaia is a particular challenge since Gaia does not directly offer pretty pictures like other space observatories but on the other hand Gaia is related to almost all topics in astronomy – from the solar system to cosmology – in a direct or indirect way.

The primary goal of Gaia outreach is to make available the Gaia achievements to the general public by collecting and monitoring the Gaia science and technological impact. In addition there are several scientific and technological aspects of the Gaia mission that can be used as an excellent introduction to more general problems. The access and use of huge databases and the study of our Galaxy are two examples. We would like to bring basic concepts in Astronomy, such as the determination of distances to the celestial bodies, to students and the general public. One of the main aims of the outreach task is to engage young people in scientific and technological careers, in particular related with Space Sciences.

We have identified four work areas: General Resources, Academic Outreach, Educational and Public Outreach and Outreach for the media. With them we want to reach a wide variety of people, covering different ages and education levels: from children in primary school to graduate students, from general public to advanced amateur astronomers. Also, the public interfaces have to support accessibility for persons who need special support e.g. by using larger fonts and control elements.

We will use a large variety of different media: comics, booklets, web sites, educational texts, multimedia material, and tools to access and visualize Gaia data. The explanations should be available as far as possible in the languages spoken in the countries involved in the mission.

Finally we will offer support to the media, preparing material for press and media and serving as contact between the media and Gaia team specialists in specific subjects, in close cooperation with the press offices of the DPAC and ESA to avoid duplication of work.

10.6.1 WP961: Management

The Education and Outreach task will to be coordinated by making a very detailed plan for all sub-workpackages with detailed milestones and delivery dates, by regularly having video conferences with the WP leaders and additional persons involved, by feeding the DPAC CU9 Wiki with up-to-date information, by exchanging information via subversion and Email and by using the infrastructure of web portals. Besides the interaction between the members of WP960 it is also necessary to coordinate the task with other workpackages which have relevance for the education and outreach (e.g. Visualisation).

10.6.2 WP962: General Resources

- One of the goals will be to collect outreach resources already available on the Internet and to adapt them to the special needs of Gaia. Good examples of astronomical and non-astronomical outreach resources from other projects should be investigated in order to inspire the Gaia Outreach Program.
- The public material shall be localisable. It should be easy to add new language versions of pages. This could be done by using a suitable content management system.
- The workpackage shall offer editorial support by offering material in a standardised way.
- Early technical and science demonstration cases shall be supported.
- The Gaia science and the technology impact of Gaia shall be documented and monitored.
- Information for the Gaia web portal and supporting web sites shall be collected and consolidated.
- The Visualisation task WP980 should be supported with the focus on visualisation for public outreach.

10.6.3 WP963: Academic Outreach

CU9 WP963 is expected to provide the interested scientific and academic community with the tools and resources for a best possible exploitation of Gaia products. This workpackage extends the documentation defined in WP920 in the sense of enhancing its user friendliness.

Our goal is

- to provide users with a friendly environment, in general terms, to access Gaia data, tools, resources and science capabilities.
- to make a best possible use of current technological facilities (Internet, Web, social networks) to keep the interested academic community permanently up to date on new developments, improvements or facilities.

- to organize a very useful tutorial system (written material plus videos), with step-by-step showcases, on scientific and academic exploitation of data.
- to provide an efficient and fast help desk to answer technical and scientific questions on Gaia data and applications.
- to establish a feedback system for improvement of all the above.

10.6.4 WP964: General Public Outreach

- The public-outreach component shall have a “questions to the experts” section (interactive, with a real person answering), a “frequently asked questions” section (archival only), and a kids- and youth-oriented section.
- Some general resources should be developed in the framework of the most common social network environments (Facebook, Twitter, blogs, forums, etc), that will constitute a direct interface with the public.
- Citizen science projects “Gaia@home”, in the manner of SDSS “Galaxy Zoo” classification project (http://www.galaxyzoo.org) or the Aerogel stardust trail search project (http://stardustathome.ssl.berkeley.edu) can be defined.
- The public-outreach component shall include a discussion forum for direct contact between the general-public users of the archive. This forum will not necessarily provide professional interaction/advice/replies to questions. The forum will, however, be regularly monitored by a CU9 member for non-topical, offensive, commercial, . . . entries or discussion threads.
- CU9 shall elaborate educational material such as a set of exercises and guidelines for teachers to illustrate astronomical and astrophysical concepts/problems and their investigation using Gaia data.
- Presentation material in the form of exhibition posters, a traveling show, scale models (in paper or with future 3d printing technology) and demos can be prepared, being partially available through the Internet in different languages.

10.6.5 WP965: Outreach for the media

- This workpackage shall provide material for the press and media, e.g. short articles, review articles, graphical and audiovisual material.
- Requests for information, material and interviews shall be coordinated.
- This task shall serve as a contact between the media and Gaia team specialists in technical/scientific specific subjects.

10.7 WP970 – Science Enabling Applications

In order to unlock the full potential of the Gaia data and thus fully exploit the billion object data set, efficient and powerful data access and exploration tools are needed. This work package is dedicated to the adaptation of existing tools and the development of new tools to enable the community to work with the archive.

The developments in this work package are strongly tied to the ones in WP930, Archive architecture design and development. WP930 provides the server–side infrastructure of the archive, while the tools adapted/developed in WP970 deal with the client side. Furthermore, these tools will put requirements on the archive framework and vice versa and require a close tailoring to the archive framework to ensure their performance and usability. The design of the science enabling tools will build on the Virtual Observatory protocols and standards implemented by the archive system – in particular the TAP interface – but also supports other protocols needed by tools used in the community to provide the best possible service.

WP970, on the other hand, is also tied to WP980, Visualisation, since the representation of the data by powerful visualisation tools (remotely or on site) is a key element of working with the data, allowing the results of the science enabling applications to be visualized.

10.7.1 WP972: Advanced Data Access Tools

The category of advanced data access tools to be developed in WP970 requires optimised methods to access and process the massive amounts of data contained in the Gaia archive. The collection of use cases from the GREAT workshops and resulting publication is large. The advanced data access tools identify and initially draw their requirements from these use cases.

The tools should be provided along with basic SQL/ADQL assistance for querying basic statistical properties and binning of subsets of the data by many properties. Within the scientific workflow, the first phase consists of the identification of the subsets of the archive to be mined, and on their transformation into a representation space suitable for these tasks. Then, the second stage models or analyses the resulting data sets. Providing access to the data for the second phase needs temporary computing and storage co-located with the data. This might not be achievable for all problems, but some limited userspace for modelling and analysis should be provided, for example, interactive fitting of spectra, or timeseries analysis.

Finally, the Gaia catalogues (various releases) will be integrated in the widely used CDS tools (Strasbourg astronomical Data Center[8]) in order to make them available to the astronomical community, notably in VizieR and Aladin. On the other hand, the use of suitable VO data models (e.g. Photometry Data Model), and VO protocols such as SSAP and TAP will be employed while tools like VOSA, VOSED will be adapted to the needs of the Gaia archive.

10.7.2 WP973: Data Mining

This sub-work package covers the development of data mining tools and infrastructure adapted to the characteristics of the archive (both to its contents and the archive system), allowing the users to perform data mining tasks and extract new knowledge.

The Gaia archive will represent an unmatched opportunity to apply data mining techniques and algorithms as discovery tools in a domain where there is no alternative to automated methods based on statistical learning. Since human exploration is certainly not feasible except for very limited subsets of data, the application of data mining algorithms is essential for a full scientific exploitation of the Gaia data. The benefits of data mining astronomical data archives in general (see e.g. SDSS) and specifically with Gaia data has been discussed extensively in the scientific community (e.g. Astrostatistics and Datamining, Sarro et al. (2012)) and roads to possible implementations were explored (WOM-057).

The application of data mining tools is expected to reveal patterns and relationships within the astronomical data that can lead to the detection of new (sub)types of objects or isolated, exotic objects that represent rapid stages of stellar evolution and/or new astrophysical scenarios. Thus, the capability of automated dimensionality reduction (feature extraction, feature selection) and the development of key learning algorithms (clustering, outlier analysis, swarm intelligence,...) are important and need to be optimised for parallel processing. For these algorithms to work efficiently, queries on the Gaia archive have to be processed swiftly and the appropriate indexes such as k-d trees[9] or HEALPix indexing[10] are required. This requires tight interactions with the Gaia archive framework development on how these specific data structures are accessed by the data mining tools.

From the architecture point of view, the data mining module will have to scale to the entire Gaia data set and allow for a flexible use of underlying infrastructure such as Cloud Computing, High Performance computing (HPC), GRID computing, and other emerging technologies. The approach currently envisaged is an architecture where the mining algorithms are built on the paradigm of Software as a Service (SaaS, see Section 4.2.2)

[8] http://cdsweb.u-strasbg.fr/about

[9] http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.115.8996

[10] http://www.g-vo.org/pmwiki/Products/HEALPixIndexing

over a service oriented architecture. However, the actual implementation will have to be matched to the archive architecture defined in WP930 and shall be based on already ongoing data mining developments.

10.7.3 WP974: Auxiliary data

The Gaia astrometric, spectrophotometric and radial velocity data promise to enable excellent science in a wide range of research fields. However, many scientific questions will be best addressed through the combination of Gaia data with existing and forthcoming surveys such as 2MASS, VISTA, and LSST. Although the DPAC processing does not take such external data into account for the production of the Gaia catalogue it does use external data – for the calibration of Gaia data and the training of some of its algorithms – that should be preserved also. The ”Auxilary Data” WP will therefore take up the following two tasks:

1. The gathering, preserving, homogenizing, and preparing of all data that were used for Gaia data calibration and DPAC algorithm training, so that every Gaia data release is accompanied, by its set of calibration data. The challenge comes from the heterogeneity of the data and of the scientific communities preparing them, ranging from solar system objects to distant galaxies, some already organized with their own databases, web portals and dissemination tools, some self-organized on local machines, some with their own well established standards.
2. The cross-matching of the Gaia Archive with external catalogues. The challenge here is twofold. On one hand, the matching of large surveys (usually in the optical range) involves processing a huge amount of data, and the accompanying performance issues will therefore also be addressed by WP930 (Archive Design). On the other hand, the cross-matching of data sets from largely different wavelength ranges involves dealing with very heterogeneous data. This problem was also experienced by different VO partners, and this WP will therefore benefit from their experience and developments. In order to further facilitate the usage of the Gaia data, the cross matching results will be integrated into existing services, such as the well-known CDS Xmatch service - allowing to cross-match Gaia catalogue with any other of the 10,000 catalogues available.

10.7.4 WP975: Science Alerts Access

The Gaia flux-based science alert stream will be issued to the community through the science alert processing carried out at the Cambridge Photometric Data Processing Centre (DPCI). The science alerts processing will issue basic information for each flux alert via the VOEvent system to the community in a timely fashion (with alerts being produced 12 days after observation by Gaia). The alert packet will contain basic characterisation information for each event, including parameters such as estimated alert object type, and more advanced classification for certain objects such as supernovae (SNe). For these, inherent Gaia photometric data will be used to provide additional information concerning SNe alerts including class, epoch, redshift, reddening.

The ”Science Alerts Access” WP will therefore take up the following two key tasks to ensure that science alerts can be accessed and visualised utilising standard tools such as Aladin or the WorldWideTelescope:

1. The Alerts Access work to be carried out in this WP will develop the interfaces required to connect the real time science alerts classification processing to the main Gaia data products. Thus, as the mission evolves, and more knowledge is accumulated about objects measured by Gaia as it successively scans the sky, there will be opportunity to cross reference new alerts against previous knowledge of that sky point as well as previous alerts against new information. Thus for instance, irregular outburst events may show up multiple times during the Gaia mission. Identification will be improved through correlation with earlier Gaia knowledge.

2. The Alerts Access will in addition provide linkages to external data resources provided through CU9, in particular via interfaces to the archive development through WP930. Finally the alerts access system will plug into the main CU9 data portal developed in WP932.

10.8 WP980 – Visualisation

The goal of the WP980 (Visualisation) is to procure, construct and integrate advanced visualisation tools that will enable a global exploration of the Gaia archive, something that is not possible out of the box with currently available software. Data visualisation is deeply linked to the scientific process. It is also part of the infrastructure of several of the WPs in this proposal, noticeably WP940 (Validation), WP960 (Education and Outreach) and WP970 (Science Enabling tools). WP980 will provide tools for making the exploration of the Gaia archive a global and interactive process. Interactive exploratory data visualisation can give far more scientific insight than an approach where blind data processing and statistical analysis are followed, rather than accompanied, by visualisation (e.g. Goodman, 2012). Most of the times this visual insight is the starting point and even the guiding reference for scientific thought.

10.8.1 Visualisation of the Gaia archive

Scientific discovery requires the ability to “see” and to “link” in an easy manner. In the case of Gaia, this means the ability to interact in an easy way with the entire archive, i.e., the ability to select, cross-link (e.g. identify a certain object or region in another visual representation), change the view in real time (e.g. navigation, panning, zooming), and link to external archives. Interaction with visual representations of the data allows selecting complex structure that cannot be specified easily in an analytical way. Take the example of spiral arms in a galaxy, which although well perceived visually are hard to isolate in the data. Spiral arms are extended irregular clumpy structures. Even fairly sophisticated spatial models fail to capture the range of stellar densities, non-trivial distribution of stars and isolate inter-arm connections. They can be better separated in certain ranges in velocity-position diagrams (as in HI radio maps), but this only becomes clear after visual inspection. Only after such inspection can one start to devise adequate automated selection and analysis algorithms.

Although the Gaia archive will be multi-dimensional, with attributes such as coordinates, distances, magnitudes and proper motions (to name a few), current day visual exploration is mostly done using 2D representations. But reduced dimensionality has a price: it easily hides features and relations in the data and, for large data sets, produces cluttered views. Multiple 2D panels are often used to improve this problem, but the reduction from 3D to 2D is quantised (as opposed to continuous), and the link between data in different panels is frequently not clear. Curiously, 3D visualisation is not widespread in space sciences, where most of the data are individual entities (stars, galaxies, asteroids). It is almost exclusively used in simulations of astrophysical fluids and fields, which are extended bodies. One of the reasons is a lack of good tools for 3D selection and interaction with point clouds. We struggle to perform even simple tasks, such as selecting a region of 3D space, using mice and 2D screens (or even 2D paper), and tend to avoid 3D representations of data as much as possible. This is also seen outside space sciences, where 3D interactive visualisation is almost exclusively used in analysis of extended bodies, such as in medical imaging, fluid dynamics, and CAD applications. Considering the specific needs of the astronomical community, such as Virtual Observatory capabilities and the ability to visualise interactively using common workstations and operating systems, one finds that no adequate 3D visualisation solutions are available today for analysis of such large point data sets.

The ability to handle big data sets in real-time constitutes a challenge we will need to face. Although it is possible at the present to interactively visualise 0.6TB of point source data at 7 frames/s, (Hassan et al., 2012), this requires extremely expensive GPU clusters - not the kind of hardware available to most users of the Gaia archive. Even though it is to be expected that by 2022, the available hardware will be able to handle sets comparable to the Gaia catalogue, the volume of external data produced by other astronomical surveys will grow more quickly than computing power. This will render impossible a brute force approach to the visualisation of Gaia products combined with those of other surveys, with a resulting loss in new groundbreaking science. In this perspective, offloading strategies must be developed that will bring Gaia and its combination with other data sets to the desktop (or even tablet). One possible solution to be studied would be the pre-computation and storage of some common data set visualisations or simplified versions of the data set. For example, if a certain kind of data is usually visualised as iso-surfaces by many users, the system would pre-compute and store these iso-surfaces as independent objects – but will keep them linked to the original data and codes used to produce them, through some data-lineage methodology. Also, in order to make sense of a huge amount of data points ($> 10^9$), together with their uncertainties, machine learning techniques will probably be needed similar to those found on Amazon, Google, Bing, etc., to suggest as well as prioritise what is more important on a given context. This will likely involve the adoption of non-deterministic algorithms, and will need to involve some kind of human (researcher) behaviour prediction. This methodology could, for example, analyse how an ensemble of past users of the system have been using it, and would try to predict and possibly suggest how a certain user working the data set could better visualise it, suggesting colormaps, smoothing or density estimation schemes, error visualisation, etc.

An emerging aspect to take into consideration is that of collaboration. Visually oriented frameworks provide an unmatched support for collaborative development of ideas. Technologies for multi-user marking or highlighting objects or regions, adding comments, web links and other meta-data, are currently used in social networks and widely adopted by the younger generations, those that will constitute the main fraction of the future users of the Gaia archive.

10.8.2 WP980 Work areas

WP980 is designed to address the visualisation challenges posed by the Gaia archive. It has been decomposed into seven sub-WPs which together address the specific aspects to be covered by the visualisation architecture, while grouping the relevant expertise within the involved teams:

10.8.2.1 WP981 Management. The goal of this WP is to assure that the objectives and schedule of WP980 are met.

10.8.2.2 WP982 Data and services interface. The main goal of this sub-workpackage is to assure a robust connection between the visualisation infrastructure and the Gaia archive. It will also provide a user command line interface with scripting capabilities (python or R foreseen) to the visualisation framework.

10.8.2.3 WP983 Visualisation infrastructure. This sub-workpackage will provide an interactive visualisation infrastructure. Off loading strategies will be studied and implemented in order to bring Gaia visualisation to the common workstation. Off-the-shelf tools that are adequate, or almost adequate, for certain aspects of the visualisation of the Gaia archive (e.g. 2D plotting and image overlay) will be procured. However, some improvements to these tools are foreseen. This is the case, for instance, of endowing Topcat with the ability of keeping its state for later resuming a task. Interactive 3D solutions, for which the state of the art applications are still lacking the desired interactivity (especially for point clouds - which is the case of the Gaia catalogue) and ability to run on common workstations, will be developed in house. The infrastructure will integrate these applications in a comprehensive environment providing linked views of the data between different panels. It is currently foreseen that this integration will be presented to the user in a web browser. Collaborative visualisation functionalities will also be assessed during the design of the infrastructure.

10.8.2.4 WP984 Volume/isosurface rendering for extended structures. The main goal of this sub-workpackage is the development of robust algorithms to provide astrophysically meaningful surfaces (e.g., A_V, metallicity, etc.) making use of the Gaia database. These surfaces will be implemented in the visualisation infrastructure.

10.8.2.5 WP985 Clustering and advanced data selection for multi-D visualisation. The focus of this sub-workpackage is to tackle the exploitation of the multi-dimensional character of the Gaia database through implementation of dimensionality reduction methods. Data clustering algorithms will be implemented for assisting multi-dimensional data selection in 2D and 3D projections.

10.8.2.6 WP986 Time-domain visualisation. The goal of this sub-workpackage is to tackle the visualisation of the time-domain aspect of the Gaia database (applied to individual objects).

Each sub-WP is described in the appendix document.

The development timeline has been planned in the perspective of the contents and dates of each data release as described in the CU9 AO Information Package:

2015/16 - First data release: 2D positions and G magnitudes for around 1 billion sources. Hundred Thousand Proper Motions (HTPM) catalogue.
Tools: 2D browsing and plotting tool with overlay support and VO layer. Single panel 3D navigator with linked views for small (HTPM) 3D data sets. Solutions exist at the

present. Mostly an integration effort.
Intermediate data releases: Increased number of parameters for all Gaia detected sources
Tools: gradual improvements until final release.
2021/22 - Final data release: Full Gaia Catalogue.
Tools: Capable of addressing the challenges described in this section, including Multi-panel 3D visualisation with linked views. Web browser (or 2021 equivalent) integration. Capable of dealing with large data sets directly and with off-loading strategies. Collaborative visualisation.

10.9 Effort estimates

Table 4: Detailed Work Package Effort Estimates in Staff Years

WP	Description	T	2013	2014	2015	2016	2017	2018- 2022	Total (SY)
910	**Management**								
911	General management	1	1.2	1.2	1.2	1.2	1.2	6.0	**12.0**
912	Technical management	1	0.4	0.4	0.4	0.4	0.4	2.0	**4.0**
913	Science scenarios and requirements mangement	1	0.2	0.2	0.2	0.2	0.2	1.0	**2.0**
920	**Documentation**								
921	Management	1	0.4	0.2	0.2	0.2	0.2	1.0	**2.2**
922	Documentation Editorial Team	1	1.0	1.0	0.8	1.0	1.7	2.5	**8.0**
923	Collect CU documentation	1	0.1	0.2	0.1	0.2	0.6	0	**1.2**
930	**Architecture Design/Development**								
931	Management	1	0.2	0.2	0.2	0.2	0.2	1.0	**2.0**
932	Gaia Archive Core Systems		2.9	3.4	2.9	2.4	3.2	14.0	**28.8**
933	Consortium / ESAC-SAT Interface Control	1	0.1	0.1	0.1	0	0	0	**0.3**
934	DB collaboration	2	0.95	0.95	1.2	0.4	0.4	2.0	**5.9**
935	DB Interface Design (TAP+Server side)	1	2.2	2.2	2.2	1.9	0.5	2.5	**11.5**
936	Correlation Function	1	1.0	1.0	1.0	0	0	0	**3.0**
940	**Validation**								
941	Management	1	0.2	0.2	0.2	0.2	0.2	1.0	**2.0**
942	Scenarios	1	1.4	1.4	1.5	1.5	1.5	1.2	**8.5**
943	Models	1	1.4	2.5	1.5	1.6	3.6	5.0	**15.6**
944	External archives	1	1.6	1.8	1.8	2.2	2.2	10.0	**19.6**
945	Statistical tools	1	1.3	1.9	2.3	2.3	2.3	9.5	**19.6**
946	Time series and variability	1	0.6	0.6	0.6	0.6	0.6	3.0	**6.0**
950	**Operations Services and Support**								
951	Management	1	0.1	0.1	0.1	0.1	0.1	0.5	**1.0**
952	Archive maintenance	1	1.5	2.25	2.25	2.25	1.75	5.6	**15.6**
953	User Support and HelpDesk	1	0.4	0.65	0.65	0.65	0.65	3.0	**6.0**
954	Service Monitoring and feedback	1	0.5	0.5	0.5	0.5	0.5	1.5	**4.0**
955	Preservation	1	0.0	0.0	0.0	0.0	0.0	1.6	**1.6**
956	Cloud/workflow	2	0.1	0.1	0.1	0.1	0.1	0.5	**1.0**
957	Auxiliary data	1	0.1	0.1	0.1	0.1	0.1	0.4	**0.9**
958	Provision of simulations (GOG)	1	0.1	0.2	0.2	0.2	0.2	0.2	**1.1**
960	**Education and Outreach**								
961	Management	2	0.2	0.2	0.2	0.2	0.2	1.0	**2.0**
962	General resources	2	0.4	0.4	1.3	1.9	1.5	5.0	**10.5**
963	Academic outreach	2	0.6	0.6	0.6	3.0	0	0	**4.8**
964	Education and general public outreach	2	0.6	0.7	0.9	0.9	0.8	3.8	**7.7**
965	Media	2	0.4	0.4	0.5	0.5	0.5	2.1	**4.4**
970	**Science Enabling Applications**								
971	Management	2	0.2	0.2	0.2	0.2	0.2	1.0	**2.0**
972	Advanced data access tools	2	2.7	3.2	3.1	3.3	2.6	11.4	**26.3**
973	Data Mining	2	1.95	2.45	2.65	2.65	2.15	9.75	**21.6**
974	Auxiliary and external data	2	2.6	3.7	3.7	3.7	3.7	15.2	**32.6**
975	Science Alerts	2	0.3	0.3	0.3	0.3	0.3	1.2	**2.7**
980	**Visualisation**								
981	Management	1	0.4	0.3	0.3	0.3	0.3	1.5	**3.1**
982	Data and services interface	1	0.3	0.3	0.3	0.3	0.3	1.5	**3.0**
983	Visualisation infrastructure	1	2.2	2.2	2.2	2.2	2.2	9.0	**20.0**
984	Volume/isosurface rendering for extended structures	2	1.4	1.4	1.4	1.4	1.4	6.5	**13.5**
985	Clustering and advanced data selection for multi-D visualisation	2	0.9	1.2	1.2	1.2	0.9	2.0	**7.4**
986	Time-domain visualisation	2	0.3	0.3	0.3	0.3	0.3	1.2	**2.7**
Total									**347.7**

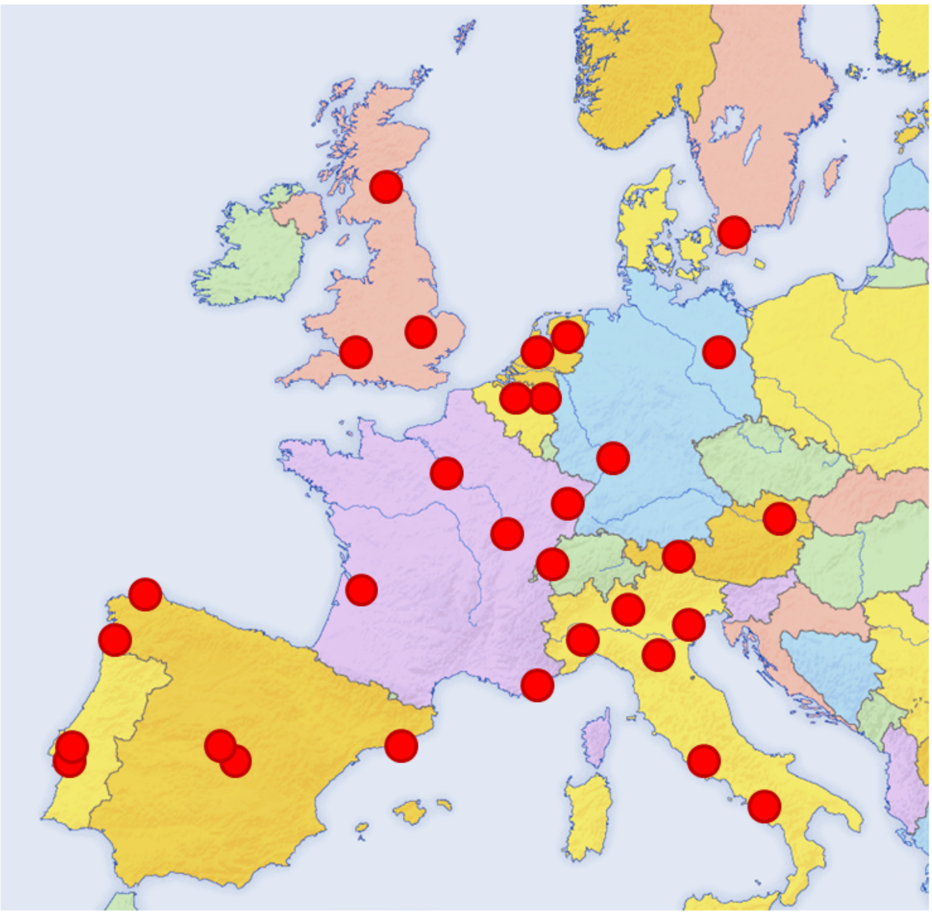


Figure 8: Geographical distribution of the institutes participating in the response (background map from Wikimedia Commons, `http://commons.wikimedia.org/wiki/File:Europe_countries_map_2.png`)

11 References

Arenou, F., 1993, *Contribution à la validation statistique des données d'Hipparcos: Catalogue d'Entrée et données préliminaires*, Ph.D. thesis, Observatoire de Paris, CNRS

Bonnarel, F., Fernique, P., Bienaymé, O., et al., 2000, A&AS, 143, 33, ADS Link

[AB-026], Brown, A., Arenou, F., Hambly, N., et al., 2012, *Gaia data access scenarios summary*,
GAIA-C9-TN-LEI-AB-026,
URL http://www.rssd.esa.int/llink/livelink/open/3125400

Brown, A.G.A., Arenou, F., van Leeuwen, F., Lindegren, L., Luri, X., 1997, In: IAU Joint Discussion, vol. 14 of IAU Joint Discussion, ADS Link

[ECSS-M-30B], ESA Publications Division, 2003, *Project Phasing and Planning*,
ECSS-M-30B

Foucault, L., 1847, Journal des Débats politiques et littéraires, 29 septembre 1847, p. 1,
URL http://gallica.bnf.fr/ark:/12148/bpt6k4477844

Fraedrich, R., Schneider, J., Westermann, R., 2009, IEEE Transactions on Visualization and Computer Graphics (Proceedings Visualization / Information Visualization 2009), 15, to appear

[ESA/SPC(2006)45], Gaia Project Scientist, 2006, *Revised Gaia Science Management Plan (SMP)*,
ESA/SPC(2006)45,
URL http://www.rssd.esa.int/llink/livelink/open/2720576

Goodman, A., 2012, Astronomsche Nachrichten, 333, 505

[GGA-002], Gracia, G., 2010, *DPAC Milestones Technical Approach*,
GAIA-PO-TN-ESAC-GGA-002,
URL http://www.rssd.esa.int/llink/livelink/open/2907536

Hambly, N.C., Collins, R.S., Cross, N.J.G., et al., 2008, MNRAS, 384, 637, ADS Link

Hassan, A.H., Fluke, C.J., Barnes, D.G., Kilborn, V.A., 2012, ArXiv e-prints, ADS Link

[JH-001], Hernandez, J., 2012, *Main Database Interface Control Document*,
GAIA-C1-SP-ESAC-JH-001,
URL http://www.rssd.esa.int/llink/livelink/open/2786145

Holl, B., Lindegren, L., 2012, A&A, 543, A14, ADS Link

Holl, B., Hobbs, D., Lindegren, L., 2010, In: S. A. Klioner, P. K. Seidelmann, & M. H. Soffel (ed.) IAU Symposium, vol. 261 of IAU Symposium, 320–324, ADS Link

Holl, B., Lindegren, L., Hobbs, D., 2012, A&A, 543, A15, ADS Link

Hume, A.C., Krause, A., Holliman, M., et al., 2012, In: Ballester, P., Egret, D., Lorente, N.P.F. (eds.) Astronomical Data Analysis Software and Systems XXI, vol. 461 of Astronomical Society of the Pacific Conference Series, 359, ADS Link

[SATSCMP], Leon, I., 2009, *SCIENCE ARCHIVE - SOFTWARE CONFIGURATION MANAGEMENT PLAN (SCMP)*, SAT_GEN_PL_2.0_01_SCMP_05112009, URL http://www.rssd.esa.int/llink/livelink/fetch/-415780/495310/1051419/Sw_Conf_Mng_Plan_v2-0.pdf?nodeid=2942288&vernum=-2

[TL-001], Levoir, T., Damery, J., Hoar, J., et al., 2012, *DPAC Product Assurance Plan*, GAIA-C1-PL-CNES-TL-001, URL http://www.rssd.esa.int/llink/livelink/open/2439085

Lindegren, L., 1995, Astronomy & Astrophysics, 304, 61, ADS Link

[JSH-033], Mercier, E., Hoar, J., 2012, *Gaia Science Ground Segment Operations Plan*, GAIA-C1-PL-ESAC-JSH-033, URL http://www.rssd.esa.int/llink/livelink/open/2892601

[RD-010], Mercier, E., Drimmel, R., O'Mullane, W., et al., 2011, *DPAC Project Development Plan*, GAIA-CD-PL-INAF-RD-010, URL http://www.rssd.esa.int/llink/livelink/open/2786669

[FM-039], Mignard, F., 2009, *DPAC Publication Policy*, GAIA-CD-PL-OCA-FM-039, URL http://www.rssd.esa.int/llink/livelink/open/2960957

[FM-044], Mignard., F., 2011, *Data access policy during processing*, GAIA-CD-PL-OCA-FM-044, URL http://www.rssd.esa.int/llink/livelink/open/3066329

[WOM-033], O'Mullane, W., 2009, *Gaia Catalogue and Archive Software Requirements and Specification*, GAIA-C9-SD-ESAC-WOM-033, URL http://www.rssd.esa.int/llink/livelink/open/2907710

[WOM-057], O'Mullane, W., 2011, *Blue skies and clouds, archives of the future*, GAIA-C9-TN-ESAC-WOM-057, URL http://www.rssd.esa.int/llink/livelink/open/3072045

[WOM-001], O'Mullane, W., Lammers, U., 2007, *Work breakdown structures for DPAC*,
GAIA-C1-TN-ESAC-WOM-001,
URL http://www.rssd.esa.int/llink/livelink/open/497865

[WOM-066], O'Mullane, W., van Leeuwen, F., 2012, *Release scenarios for the Gaia archive*,
GAIA-C9-TN-ESAC-WOM-066,
URL http://www.rssd.esa.int/llink/livelink/open/3111396

[WOM-006], O'Mullane, W., Hernandez, J., Hoar, J., et al., 2006, *Tackling AGIS development with Agile programming*,
GAIA-C1-TN-ESAC-WOM-006,
URL http://www.rssd.esa.int/llink/livelink/open/535099

[WOM-017], O'Mullane, W., Drimmel, R., Mignard, F., et al., 2008, *Project Implementation Plan (PIP)*,
GAIA-CD-PL-ESAC-WOM-017,
URL http://www.rssd.esa.int/llink/livelink/open/2812481

[WOM-012], O'Mullane, W., Nguyen, A.T., Hoar, J., et al., 2009, *DPAC Configuration Management Plan*,
GAIA-C1-PL-ESAC-WOM-012,
URL http://www.rssd.esa.int/llink/livelink/open/2760363

[SATMP], Osuna, P., 2011, *Science Archives and VO Team (SAT) Management Plan*,
SAT_GEN_PL_3.0_06_MP_30_May_2011,
URL http://www.rssd.esa.int/llink/livelink/fetch/-415780/2741092/SAT_GEN_PL_3.0_06_MP_30May2011.pdf?nodeid=3120171&vernum=-2

Perryman, M.A.C., ESA (eds.), 1997, *The HIPPARCOS and TYCHO catalogues. Astrometric and photometric star catalogues derived from the ESA HIPPARCOS Space Astrometry Mission*, vol. 1200 of ESA Special Publication, ADS Link

[TJP-011], Prusti, T., 2012, *Gaia Intermediate Data Release Scenario*,
GAIA-CG-PL-ESA-TJP-011,
URL http://www.rssd.esa.int/llink/livelink/open/3145458

[TJP-013], Prusti, T., 2012, *Announcement of Opportunity for the Gaia Data Processing Archive Access Co-Ordination Unit*,
GAIA-CG-PL-ESA-TJP-013,
URL http://sci.esa.int/science-e/www/object/index.cfm?fobjectid=51116

Robin, A.C., Luri, X., Reylé, C., et al., 2012, ArXiv e-prints, ADS Link

Sarro, L.M., Eyer, L., O'Mullane, W., De Ridder, J., 2012, *Astrostatistics and Data Mining*

Szalay, A., Gray, J., Thakar, A., et al., 2001, eprint arXiv:cs/0111015, ADS Link

[DTP-002], Tapiador, D., 2011, *Deployment of Intersystems Cache with GUMS on Amazon EC2*,
`GAIA-C9-TN-ESAC-DTP-002`,
URL `http://www.rssd.esa.int/llink/livelink/open/3104896`

Tapiador, D., 2011, *Interrogator Software Requirements and Specification*, Tech. Rep. GAIA-C9-SP-ESAC-DTP-001, ESAC

[DTP-004], Tapiador, D., 2012, *Deployment of Greenplum parallel DBMS on Amazon EC2 with GUMS*,
`GAIA-C9-TN-ESAC-DTP-004`,
URL `http://www.rssd.esa.int/llink/livelink/open/3133836`

Thakar, A.R., 2008, Computing in Science and Engineering, 10, 9, ADS Link

van Leeuwen, F., 2007, Astronomy & Astrophysics, 474, 653, ADS Link

12 Acronyms

The following table has been generated from the on-line Gaia acronym list:

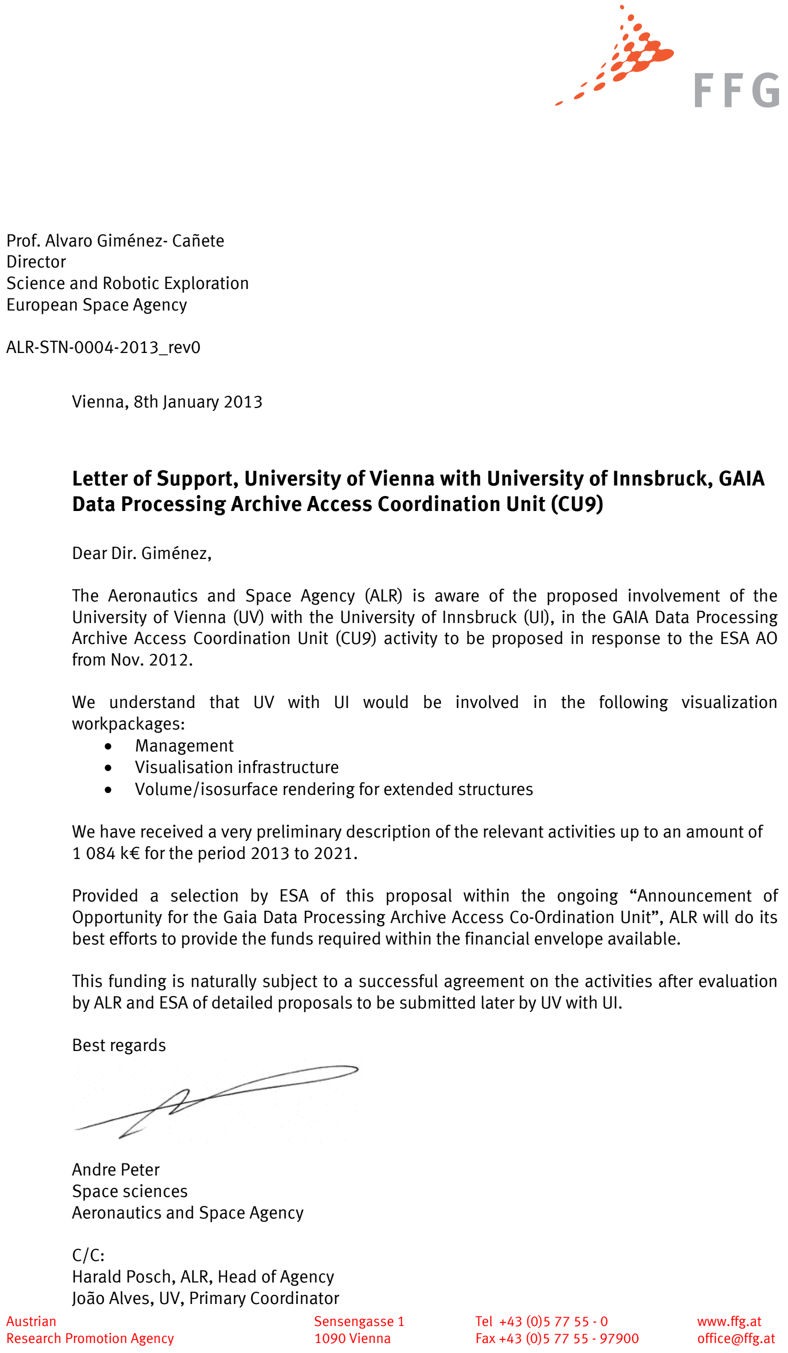
Acronym	Description
2MASS	Two-Micron All Sky Survey
ADQL	Astronomical Data Query Language
AIP	Astrophysikalisches Institut Potsdam
AO	Announcement of Opportunity
ARI	Astronomisches Rechen-Institut (Heidelberg; part of ZAH)
ASDC	ASI Science Data Centre
AT	Average Tilt
BE	Big Endian
BP	Blue Photometer
CADC	Canadian Astronomical Data Centre
CANFAR	Canadian Advanced Network for Astronomical Research
CDS	Centre de Donnes Strasbourg
CESCA	Centre de Serveis Científics i Acadèmics de Catalunya
CH	Switserland
CU	Coordination Unit (in DPAC)
CV	Curriculum Vitae
DB	DataBase
DBMS	DataBase Management System
DE	Developmental Ephemerides (from JPL/NASA)
DPAC	Data Processing and Analysis Consortium
DPACE	Data Processing and Analysis Consortium Executive
DPCB	Data Processing Centre Barcelona
DPCI	Data Processing Centre (Institute of Astronomy) Cambridge
DS9	Deep Space 9 (specific astronomical data visualisation application; SAOImage)
DSS	Digital Sun Sensor
ECSS	European Cooperation for Space Standardisation
EGI	European Grid Infrastructure
ES	España (Spain)
ESA	European Space Agency
ESAC	European Space Astronomy Centre (VilSpa)
ESTEC	European Space research and TEchnology Centre (ESA)
FAQ	Frequently Asked Questions
FP7	Seventh Research Framework Programme
FR	Final Review

GAIA	Global Astrometric Interferometer for Astrophysics (obsolete; now spelled as Gaia)
GALEX	GALaxy Evolution eXplorer
GAP	Gaia Archive Preparations (DPAC WG)
GBIN	Gaia binary format
GBOG	Ground-Based Observations for Gaia (DPAC)
GOG	Gaia Object Generator
GPU	Gaia Processing Unit
GREAT	Gaia Research for European Astronomy Training
GST	Gaia Science Team
HEALPix	Hierarchical Equal-Area iso-Latitude Pixelisation
HPC	High-Performance Computing
HTML	HyperText Markup Language
HW	Hardware (also denoted H/W)
ICD	Interface Control Document
IFA	Institute for Astronomy, Edinburgh
INAF	Instituto Nazionale di Astrofisica (Italy)
INTA	Laboratorio de Astrofísica Espacial y Física Fundamental (LAEFF-INTA)
IPDA	International Planetary Data Alliance
ISO	International Organisation for Standardisation (Geneva, Switzerland)
IT	Information Technology
IVOA	International Virtual Observatory Alliance
IfA	Institute for Astronomy (Edinburgh)
IoA	Institute of Astronomy (Cambridge)
LAMOST	Large Sky Area Multi-Object Fibre Spectroscopic Telescope (Guoshoujing Telescope)
LSST	Large-aperture Synoptic Survey Telescope
MDB	Main Database
MPIA	Max Planck Institute für Astronomy (Heidelberg)
MSR	Microsoft Research
NL	NetherLands
OCA	Observatoire de la Côte d'Azur (Nice)
ObsGE	Observatory of GEneva (Switzerland)
PA	Product Assurance
PCA	Principal Component Analysis
PDF	Portable Document Format
PO	(DPAC) Project Office
PT	Portugal
REST	REpresentational State Transfer
RP	Red Photometer

RVS	Radial Velocity Spectrometer
SAMP	Simple Application Messaging Protocol
SAT	Satellite Archive Team
SCREW	Software Change Requirements and Extra Wishes
SDD	Software Design Document
SDP	Software Development Plan
SDSS	Sloan Digital Sky Survey
SE	Sweden
SIAP	Simple Image Access Protocol
SOAP	Simple Object Access Protocol
SOC	Science Operations Centre
SOM	Spacecraft Operations Manager
SPLAT	Spectral Analysis Tool (VO)
SPR	Software Problem Report
SQL	Structured Query Language
SRS	Scanning Reference System
SSAP	Simple Spectral Access Protocol
SSC	Spitzer Science Centre (NASA)
SSO	Solar System Object
STILTS	Starlink Tables Infrastructure Library Tool Set
STR	Software Test Report
STS	System Test Specification
SY	Staff Year
SaaS	Software as a Service
TAP	Table Access Protocol
TN	Technical Note
TOPCAT	Tool for OPerations on Catalogues And Tables
UK	United Kingdom
UKIDSS	UKIRT Infrared Deep Sky Survey
UNINOVA	Institute for the Development of New Tecnologies (University of Lisbon, Portugal)
UV	UltraViolet
VISTA	Visible and Infrared Survey Telescope for Astronomy
VO	Virtual Observatory
VOSA	VOSA
VOSED	VO SED Building Tool
WBS	Work Breakdown Structure
WP	Work Package
WWT	World Wide Telescope

13 Support letters from funding agencies

13.1 Austria



FFG

Prof. Alvaro Giménez- Cañete
Director
Science and Robotic Exploration
European Space Agency

ALR-STN-0004-2013_rev0

Vienna, 8th January 2013

Letter of Support, University of Vienna with University of Innsbruck, GAIA Data Processing Archive Access Coordination Unit (CU9)

Dear Dir. Giménez,

The Aeronautics and Space Agency (ALR) is aware of the proposed involvement of the University of Vienna (UV) with the University of Innsbruck (UI), in the GAIA Data Processing Archive Access Coordination Unit (CU9) activity to be proposed in response to the ESA AO from Nov. 2012.

We understand that UV with UI would be involved in the following visualization workpackages:

- Management
- Visualisation infrastructure
- Volume/isosurface rendering for extended structures

We have received a very preliminary description of the relevant activities up to an amount of 1 084 k€ for the period 2013 to 2021.

Provided a selection by ESA of this proposal within the ongoing “Announcement of Opportunity for the Gaia Data Processing Archive Access Co-Ordination Unit”, ALR will do its best efforts to provide the funds required within the financial envelope available.

This funding is naturally subject to a successful agreement on the activities after evaluation by ALR and ESA of detailed proposals to be submitted later by UV with UI.

Best regards

Andre Peter
Space sciences
Aeronautics and Space Agency

C/C:
Harald Posch, ALR, Head of Agency
João Alves, UV, Primary Coordinator

Austrian
Research Promotion Agency

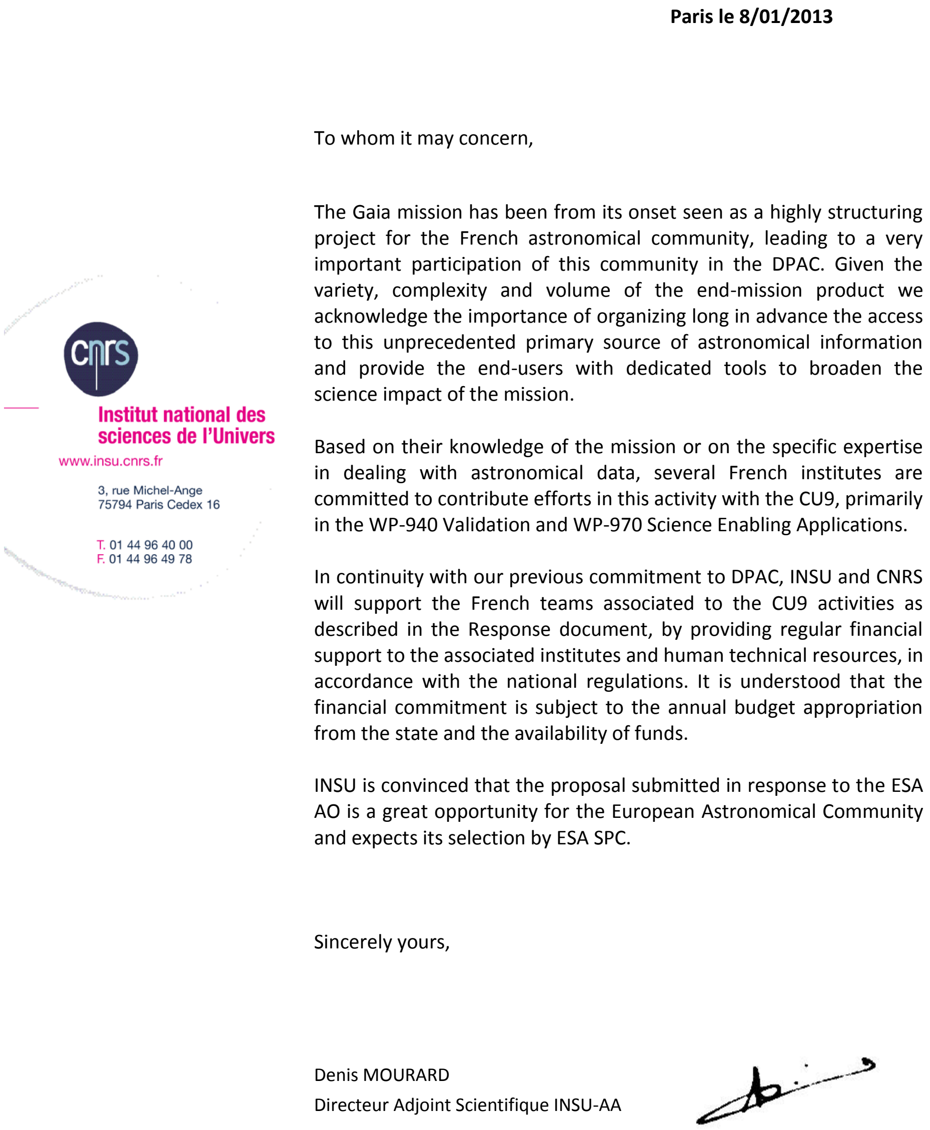
Sensengasse 1
1090 Vienna

Tel +43 (0)5 77 55 - 0
Fax +43 (0)5 77 55 - 97900

www.ffg.at
office@ffg.at

13.2 France

13.2.1 INSU



Paris le 8/01/2013

Institut national des sciences de l'Univers
www.insu.cnrs.fr

3, rue Michel-Ange
75794 Paris Cedex 16

T. 01 44 96 40 00
F. 01 44 96 49 78

To whom it may concern,

The Gaia mission has been from its onset seen as a highly structuring project for the French astronomical community, leading to a very important participation of this community in the DPAC. Given the variety, complexity and volume of the end-mission product we acknowledge the importance of organizing long in advance the access to this unprecedented primary source of astronomical information and provide the end-users with dedicated tools to broaden the science impact of the mission.

Based on their knowledge of the mission or on the specific expertise in dealing with astronomical data, several French institutes are committed to contribute efforts in this activity with the CU9, primarily in the WP-940 Validation and WP-970 Science Enabling Applications.

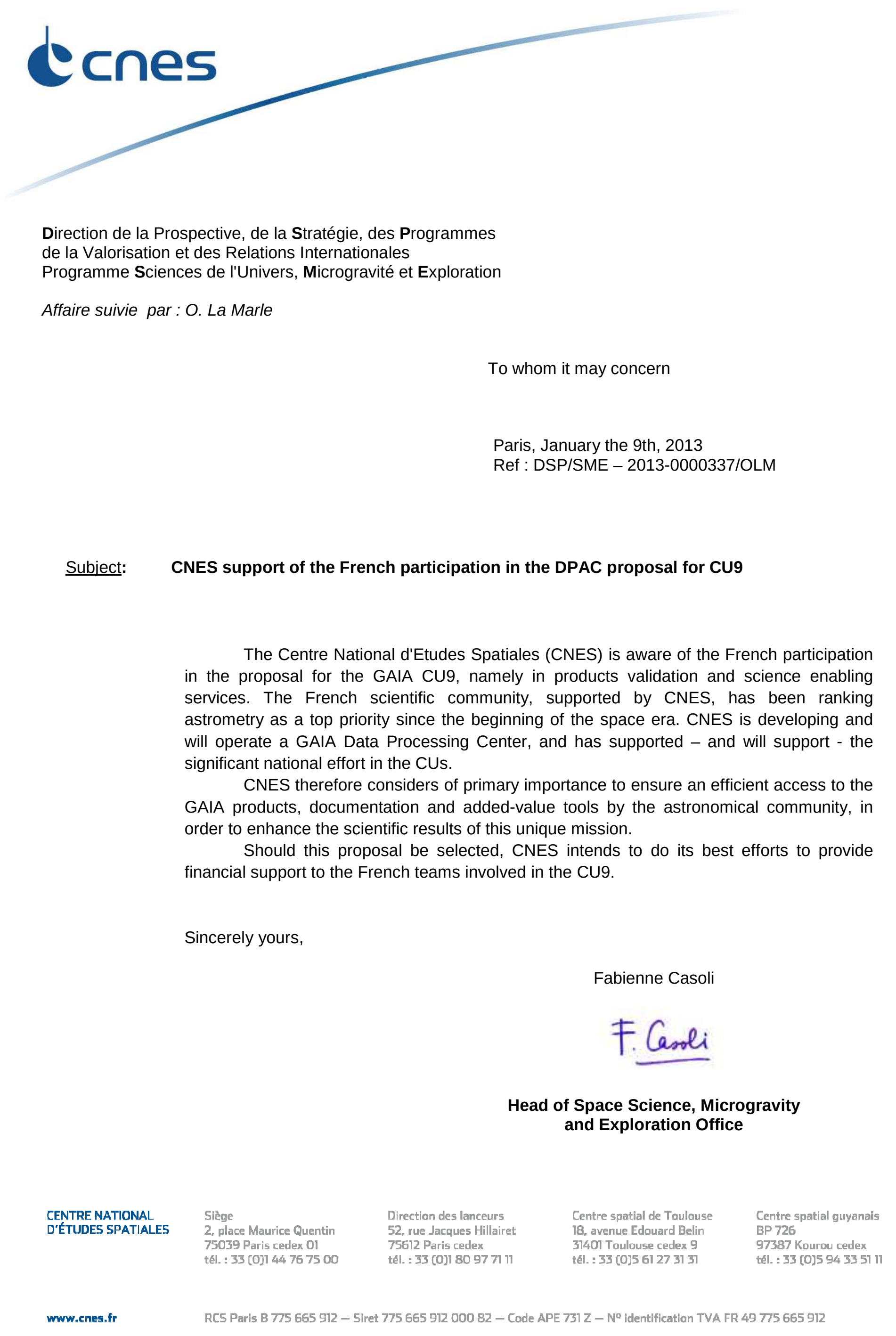
In continuity with our previous commitment to DPAC, INSU and CNRS will support the French teams associated to the CU9 activities as described in the Response document, by providing regular financial support to the associated institutes and human technical resources, in accordance with the national regulations. It is understood that the financial commitment is subject to the annual budget appropriation from the state and the availability of funds.

INSU is convinced that the proposal submitted in response to the ESA AO is a great opportunity for the European Astronomical Community and expects its selection by ESA SPC.

Sincerely yours,

Denis MOURARD
Directeur Adjoint Scientifique INSU-AA

13.2.2 CNES



Direction de la Prospective, de la Stratégie, des Programmes
de la Valorisation et des Relations Internationales
Programme Sciences de l'Univers, Microgravité et Exploration

Affaire suivie par : O. La Marle

To whom it may concern

Paris, January the 9th, 2013
Ref : DSP/SME – 2013-0000337/OLM

Subject: **CNES support of the French participation in the DPAC proposal for CU9**

The Centre National d'Etudes Spatiales (CNES) is aware of the French participation in the proposal for the GAIA CU9, namely in products validation and science enabling services. The French scientific community, supported by CNES, has been ranking astrometry as a top priority since the beginning of the space era. CNES is developing and will operate a GAIA Data Processing Center, and has supported – and will support - the significant national effort in the CUs.

CNES therefore considers of primary importance to ensure an efficient access to the GAIA products, documentation and added-value tools by the astronomical community, in order to enhance the scientific results of this unique mission.

Should this proposal be selected, CNES intends to do its best efforts to provide financial support to the French teams involved in the CU9.

Sincerely yours,

Fabienne Casoli

F. Casoli

Head of Space Science, Microgravity and Exploration Office

CENTRE NATIONAL
D'ÉTUDES SPATIALES

Siège
2, place Maurice Quentin
75039 Paris cedex 01
tél. : 33 (0)1 44 76 75 00

Direction des lanceurs
52, rue Jacques Hillairet
75612 Paris cedex
tél. : 33 (0)1 80 97 71 11

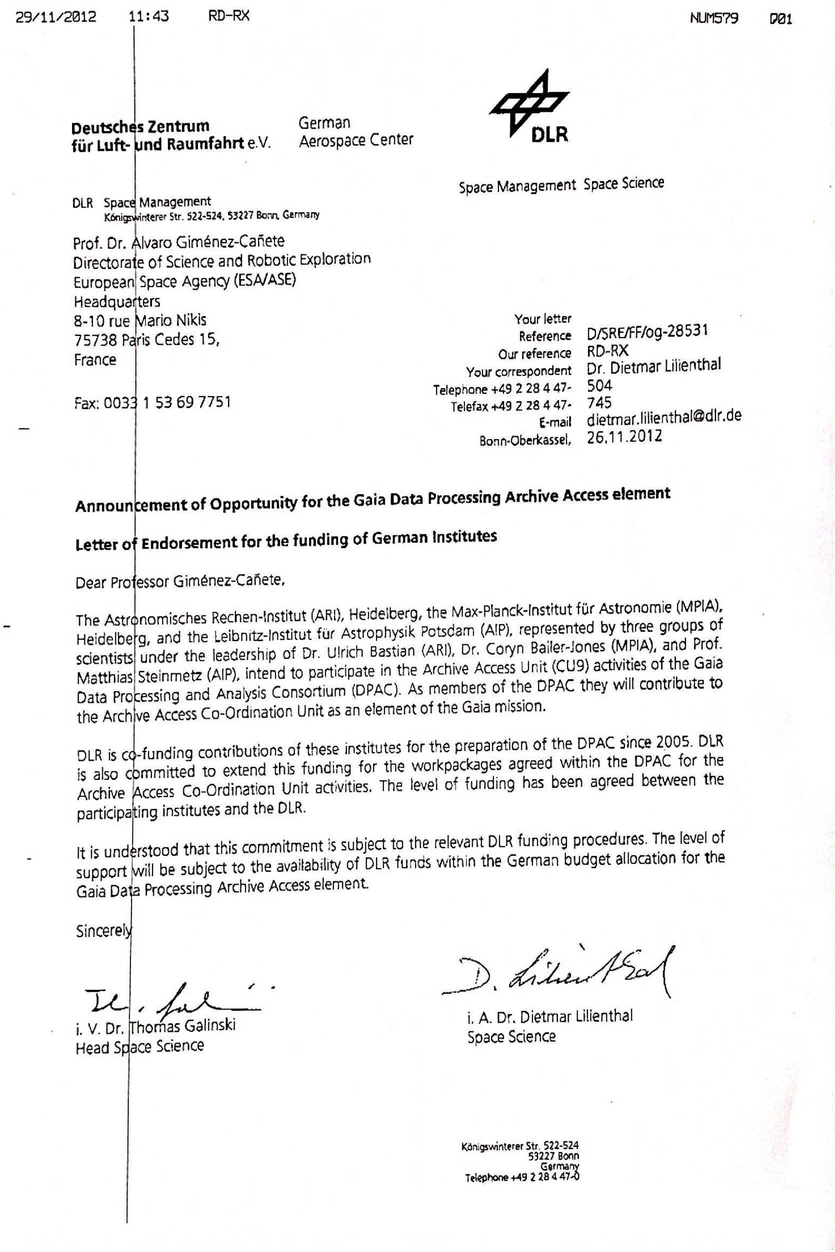
Centre spatial de Toulouse
18, avenue Edouard Belin
31401 Toulouse cedex 9
tél. : 33 (0)5 61 27 31 31

Centre spatial guyanais
BP 726
97387 Kourou cedex
tél. : 33 (0)5 94 33 51 11

www.cnes.fr

RCS Paris B 775 665 912 – Siret 775 665 912 000 82 – Code APE 731 Z – Nº identification TVA FR 49 775 665 912

13.3 Germany



29/11/2012 11:43 RD-RX NUM579 001

Deutsches Zentrum für Luft- und Raumfahrt e.V.
German Aerospace Center

DLR

Space Management Space Science

DLR Space Management
Königswinterer Str. 522-524, 53227 Bonn, Germany

Prof. Dr. Alvaro Giménez-Cañete
Directorate of Science and Robotic Exploration
European Space Agency (ESA/ASE)
Headquarters
8-10 rue Mario Nikis
75738 Paris Cedes 15,
France

Fax: 0033 1 53 69 7751

Your letter	
Reference	D/SRE/FF/og-28531
Our reference	RD-RX
Your correspondent	Dr. Dietmar Lilienthal
Telephone +49 2 28 4 47-	504
Telefax +49 2 28 4 47-	745
E-mail	dietmar.lilienthal@dlr.de
Bonn-Oberkassel,	26.11.2012

Announcement of Opportunity for the Gaia Data Processing Archive Access element

Letter of Endorsement for the funding of German Institutes

Dear Professor Giménez-Cañete,

The Astronomisches Rechen-Institut (ARI), Heidelberg, the Max-Planck-Institut für Astronomie (MPIA), Heidelberg, and the Leibnitz-Institut für Astrophysik Potsdam (AIP), represented by three groups of scientists under the leadership of Dr. Ulrich Bastian (ARI), Dr. Coryn Bailer-Jones (MPIA), and Prof. Matthias Steinmetz (AIP), intend to participate in the Archive Access Unit (CU9) activities of the Gaia Data Processing and Analysis Consortium (DPAC). As members of the DPAC they will contribute to the Archive Access Co-Ordination Unit as an element of the Gaia mission.

DLR is co-funding contributions of these institutes for the preparation of the DPAC since 2005. DLR is also committed to extend this funding for the workpackages agreed within the DPAC for the Archive Access Co-Ordination Unit activities. The level of funding has been agreed between the participating institutes and the DLR.

It is understood that this commitment is subject to the relevant DLR funding procedures. The level of support will be subject to the availability of DLR funds within the German budget allocation for the Gaia Data Processing Archive Access element.

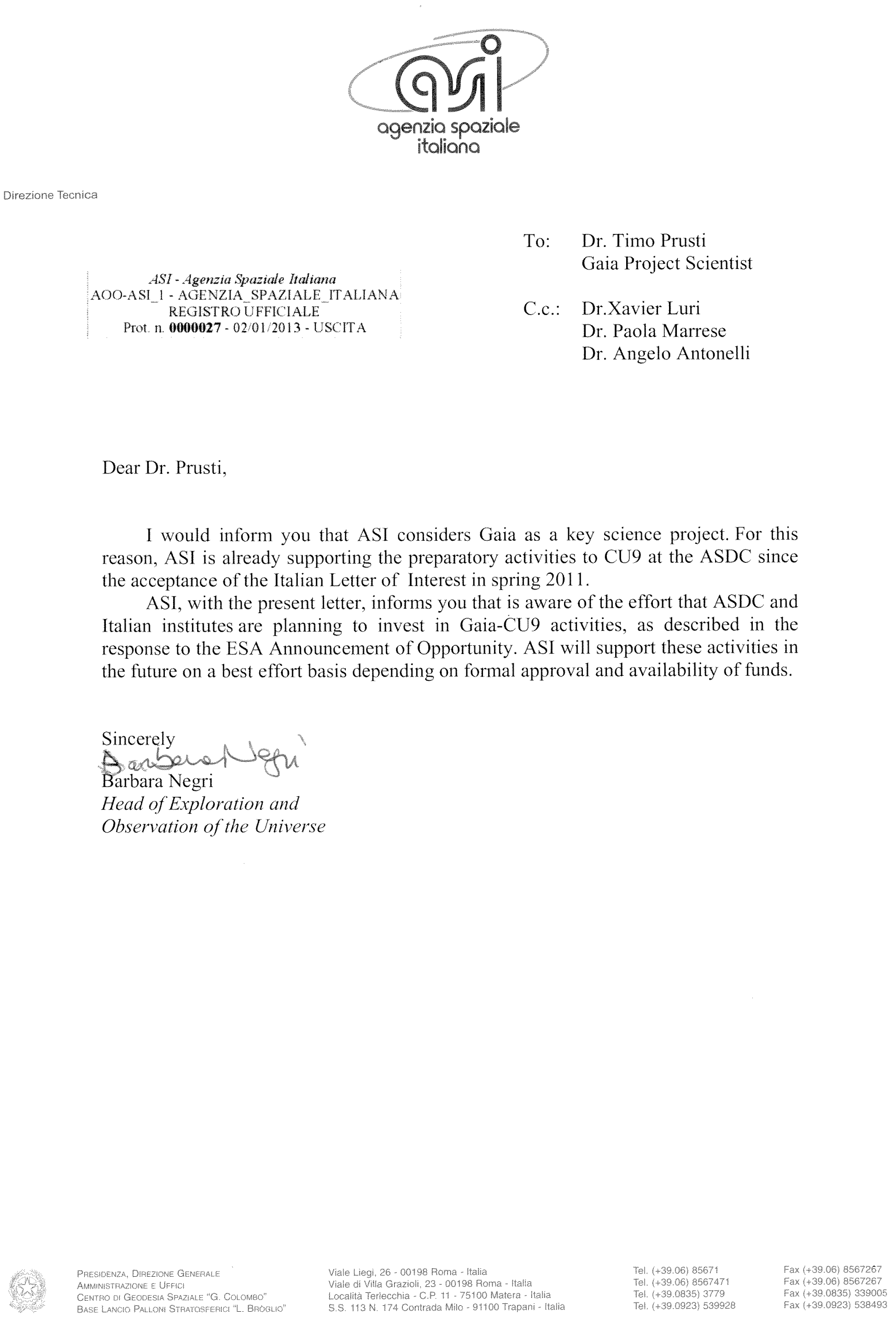
Sincerely

i. V. Dr. Thomas Galinski
Head Space Science

D. Lilienthal
i. A. Dr. Dietmar Lilienthal
Space Science

Königswinterer Str. 522-524
53227 Bonn
Germany
Telephone +49 2 28 4 47-0

13.4 Italy



agenzia spaziale italiana

Direzione Tecnica

ASI - Agenzia Spaziale Italiana
AOO-ASI_1 - AGENZIA_SPAZIALE_ITALIANA
REGISTRO UFFICIALE
Prot. n. **0000027** - 02/01/2013 - USCITA

To: Dr. Timo Prusti
Gaia Project Scientist

C.c.: Dr.Xavier Luri
Dr. Paola Marrese
Dr. Angelo Antonelli

Dear Dr. Prusti,

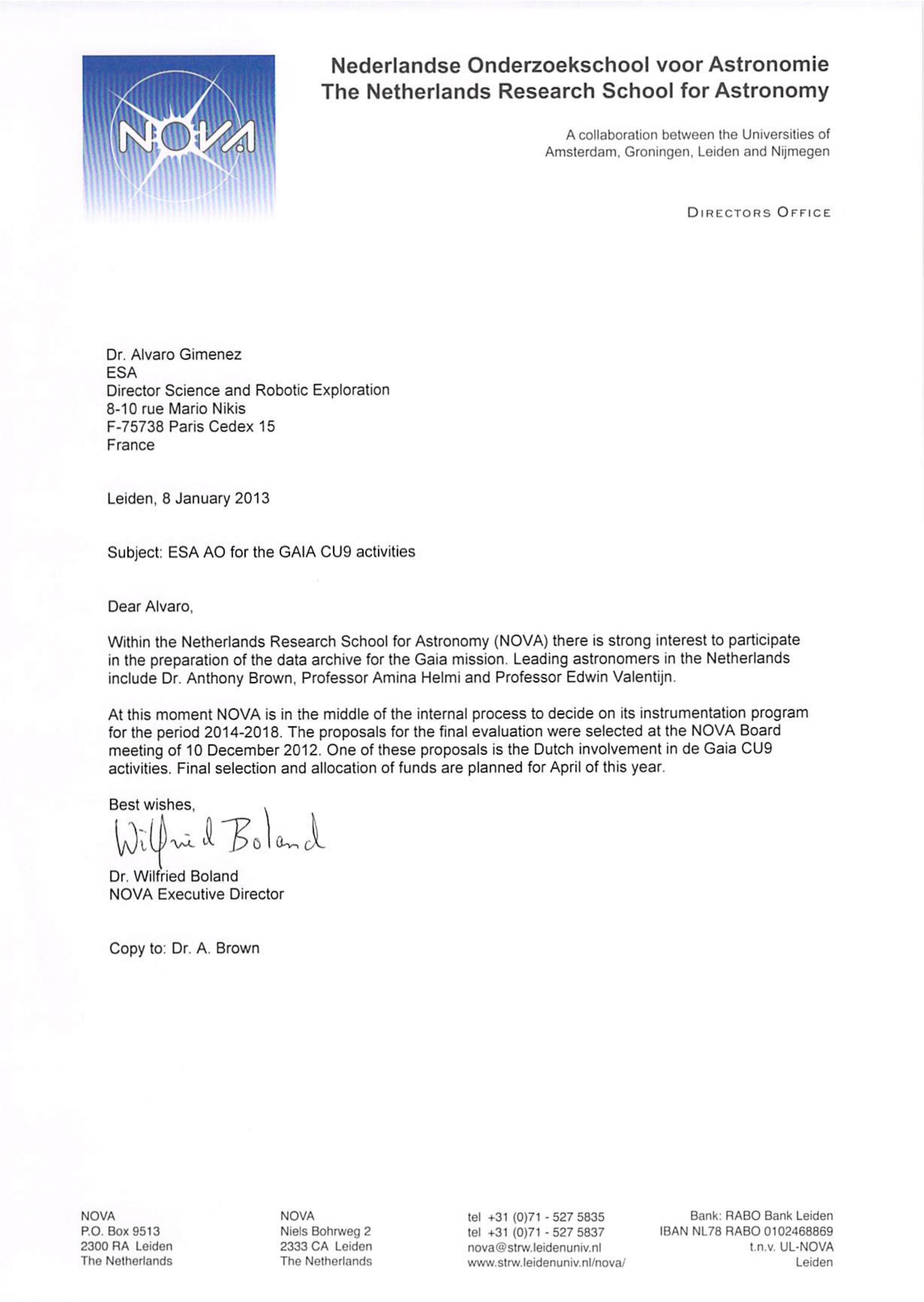
I would inform you that ASI considers Gaia as a key science project. For this reason, ASI is already supporting the preparatory activities to CU9 at the ASDC since the acceptance of the Italian Letter of Interest in spring 2011.

ASI, with the present letter, informs you that is aware of the effort that ASDC and Italian institutes are planning to invest in Gaia-CU9 activities, as described in the response to the ESA Announcement of Opportunity. ASI will support these activities in the future on a best effort basis depending on formal approval and availability of funds.

Sincerely
Barbara Negri
Barbara Negri
Head of Exploration and
Observation of the Universe

Presidenza, Direzione Generale	Viale Liegi, 26 - 00198 Roma - Italia	Tel. (+39.06) 85671	Fax (+39.06) 8567267
Amministrazione e Uffici	Viale di Villa Grazioli, 23 - 00198 Roma - Italia	Tel. (+39.06) 8567471	Fax (+39.06) 8567267
Centro di Geodesia Spaziale "G. Colombo"	Località Terlecchia - C.P. 11 - 75100 Matera - Italia	Tel. (+39.0835) 3779	Fax (+39.0835) 339005
Base Lancio Palloni Stratosferici "L. Broglio"	S.S. 113 N. 174 Contrada Milo - 91100 Trapani - Italia	Tel. (+39.0923) 539928	Fax (+39.0923) 538493

13.5 Netherlands



Nederlandse Onderzoekschool voor Astronomie
The Netherlands Research School for Astronomy

A collaboration between the Universities of
Amsterdam, Groningen, Leiden and Nijmegen

DIRECTORS OFFICE

Dr. Alvaro Gimenez
ESA
Director Science and Robotic Exploration
8-10 rue Mario Nikis
F-75738 Paris Cedex 15
France

Leiden, 8 January 2013

Subject: ESA AO for the GAIA CU9 activities

Dear Alvaro,

Within the Netherlands Research School for Astronomy (NOVA) there is strong interest to participate in the preparation of the data archive for the Gaia mission. Leading astronomers in the Netherlands include Dr. Anthony Brown, Professor Amina Helmi and Professor Edwin Valentijn.

At this moment NOVA is in the middle of the internal process to decide on its instrumentation program for the period 2014-2018. The proposals for the final evaluation were selected at the NOVA Board meeting of 10 December 2012. One of these proposals is the Dutch involvement in de Gaia CU9 activities. Final selection and allocation of funds are planned for April of this year.

Best wishes,

Wilfried Boland

Dr. Wilfried Boland
NOVA Executive Director

Copy to: Dr. A. Brown

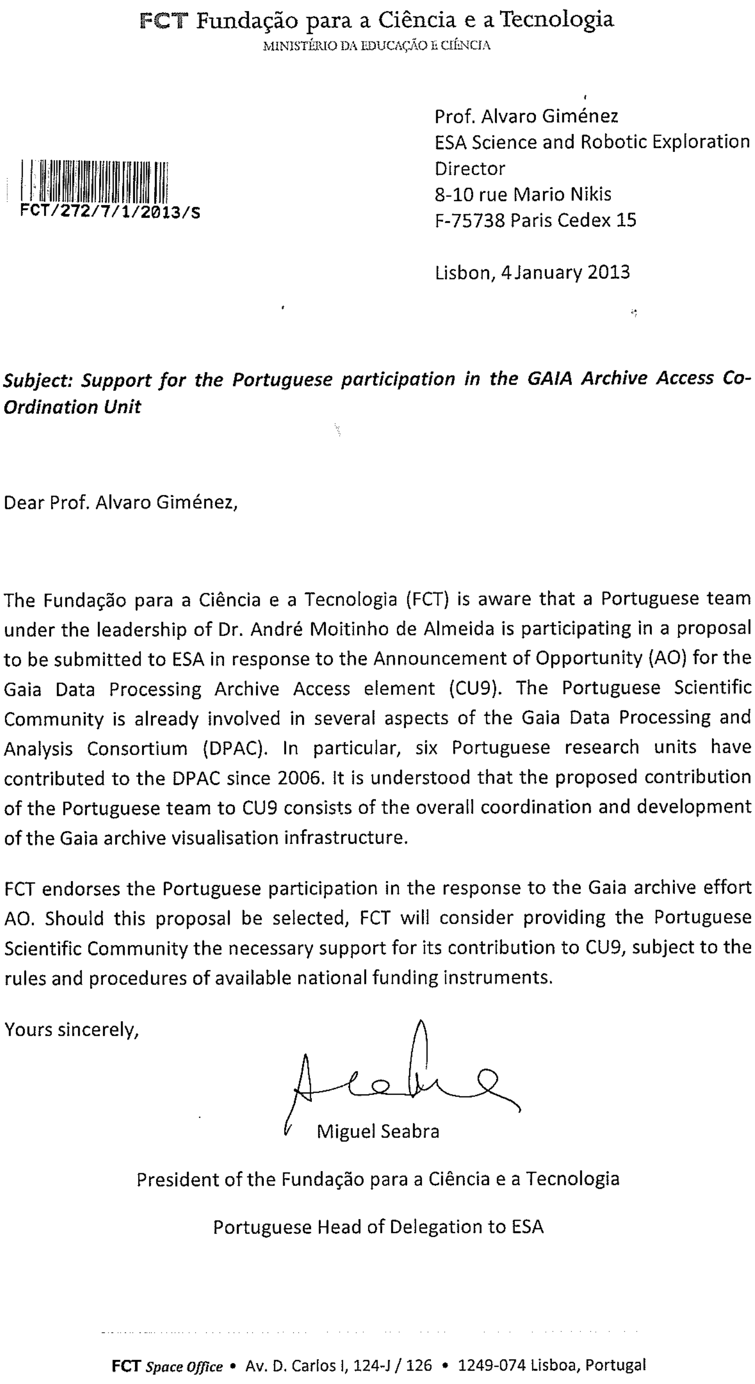
NOVA
P.O. Box 9513
2300 RA Leiden
The Netherlands

NOVA
Niels Bohrweg 2
2333 CA Leiden
The Netherlands

tel +31 (0)71 - 527 5835
tel +31 (0)71 - 527 5837
nova@strw.leidenuniv.nl
www.strw.leidenuniv.nl/nova/

Bank: RABO Bank Leiden
IBAN NL78 RABO 0102468869
t.n.v. UL-NOVA
Leiden

13.6 Portugal



FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA

FCT/272/7/1/2013/S

Prof. Alvaro Giménez
ESA Science and Robotic Exploration
Director
8-10 rue Mario Nikis
F-75738 Paris Cedex 15

Lisbon, 4 January 2013

Subject: Support for the Portuguese participation in the GAIA Archive Access Co-Ordination Unit

Dear Prof. Alvaro Giménez,

The Fundação para a Ciência e a Tecnologia (FCT) is aware that a Portuguese team under the leadership of Dr. André Moitinho de Almeida is participating in a proposal to be submitted to ESA in response to the Announcement of Opportunity (AO) for the Gaia Data Processing Archive Access element (CU9). The Portuguese Scientific Community is already involved in several aspects of the Gaia Data Processing and Analysis Consortium (DPAC). In particular, six Portuguese research units have contributed to the DPAC since 2006. It is understood that the proposed contribution of the Portuguese team to CU9 consists of the overall coordination and development of the Gaia archive visualisation infrastructure.

FCT endorses the Portuguese participation in the response to the Gaia archive effort AO. Should this proposal be selected, FCT will consider providing the Portuguese Scientific Community the necessary support for its contribution to CU9, subject to the rules and procedures of available national funding instruments.

Yours sincerely,

Miguel Seabra

President of the Fundação para a Ciência e a Tecnologia

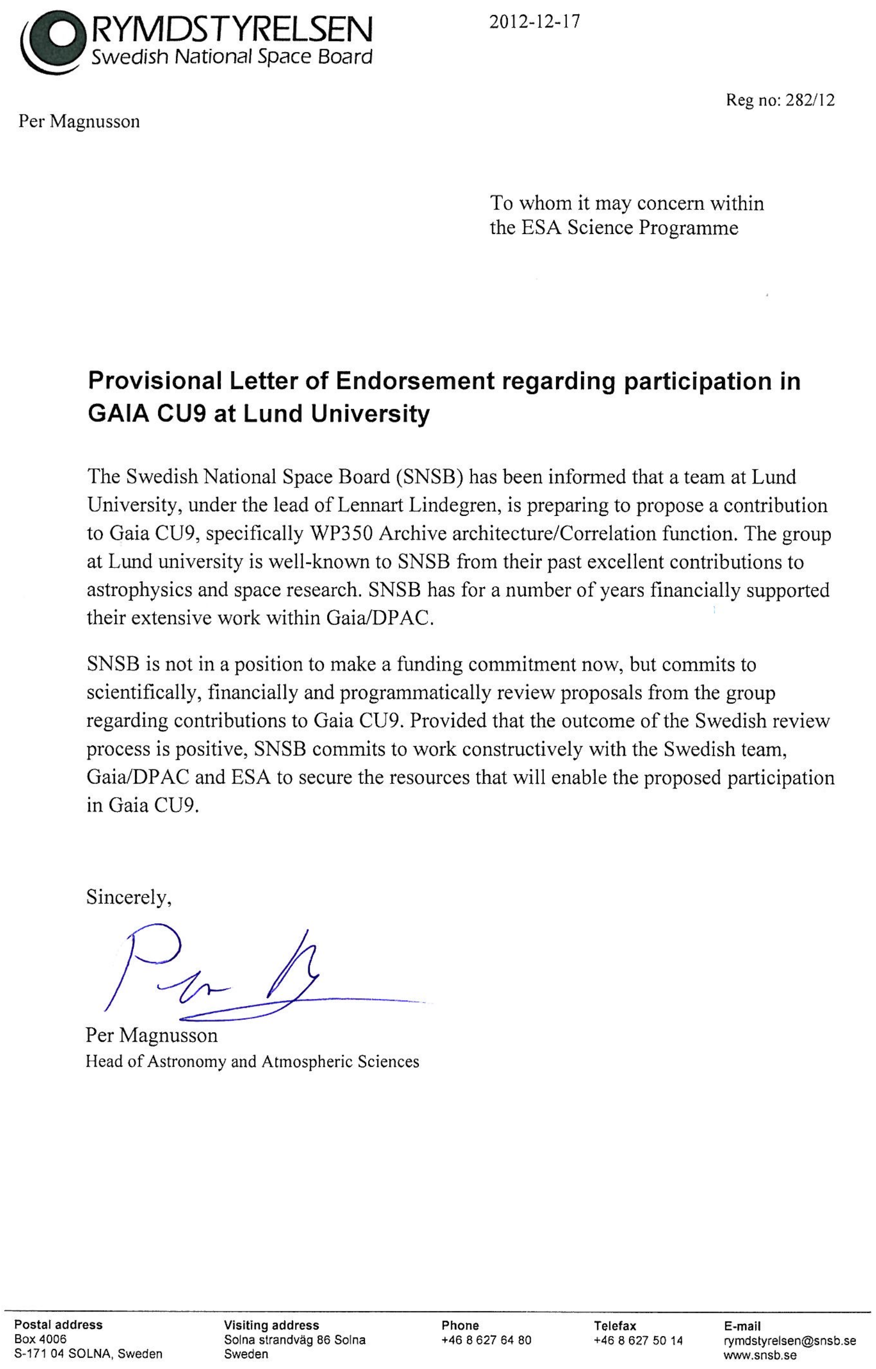
Portuguese Head of Delegation to ESA

FCT *Space Office* • Av. D. Carlos I, 124-J / 126 • 1249-074 Lisboa, Portugal

13.7 Spain

The support letter from the Spanish funding agency is confirmed but was not received at the time of submission of this document, and will be delivered separately when received.

13.8 Sweden



RYMDSTYRELSEN
Swedish National Space Board

2012-12-17

Reg no: 282/12

Per Magnusson

To whom it may concern within
the ESA Science Programme

Provisional Letter of Endorsement regarding participation in GAIA CU9 at Lund University

The Swedish National Space Board (SNSB) has been informed that a team at Lund University, under the lead of Lennart Lindegren, is preparing to propose a contribution to Gaia CU9, specifically WP350 Archive architecture/Correlation function. The group at Lund university is well-known to SNSB from their past excellent contributions to astrophysics and space research. SNSB has for a number of years financially supported their extensive work within Gaia/DPAC.

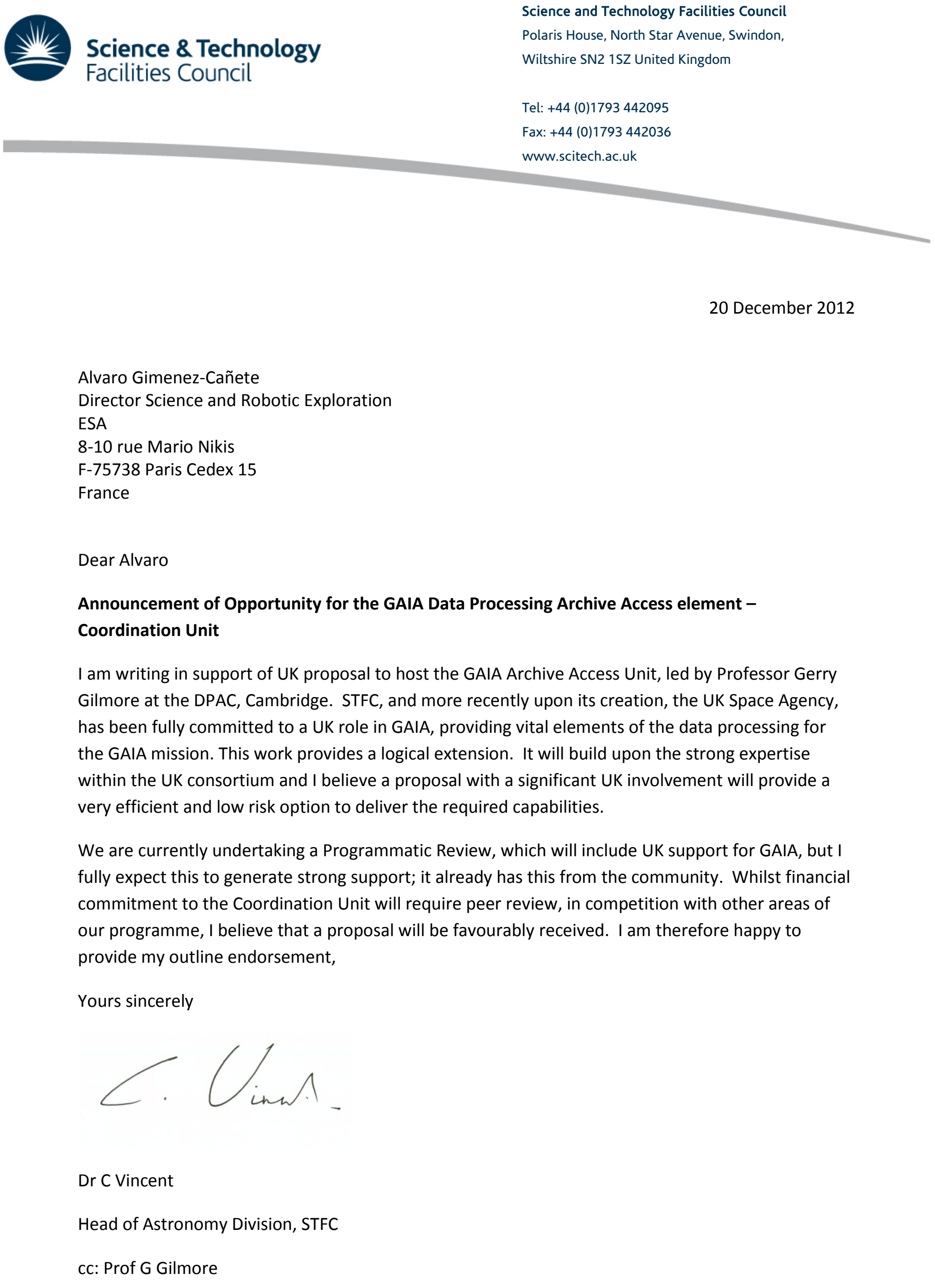
SNSB is not in a position to make a funding commitment now, but commits to scientifically, financially and programmatically review proposals from the group regarding contributions to Gaia CU9. Provided that the outcome of the Swedish review process is positive, SNSB commits to work constructively with the Swedish team, Gaia/DPAC and ESA to secure the resources that will enable the proposed participation in Gaia CU9.

Sincerely,

Per Magnusson
Head of Astronomy and Atmospheric Sciences

Postal address	Visiting address	Phone	Telefax	E-mail
Box 4006 S-171 04 SOLNA, Sweden	Solna strandväg 86 Solna Sweden	+46 8 627 64 80	+46 8 627 50 14	rymdstyrelsen@snsb.se www.snsb.se

13.9 UK



Science & Technology Facilities Council

Science and Technology Facilities Council
Polaris House, North Star Avenue, Swindon,
Wiltshire SN2 1SZ United Kingdom

Tel: +44 (0)1793 442095
Fax: +44 (0)1793 442036
www.scitech.ac.uk

20 December 2012

Alvaro Gimenez-Cañete
Director Science and Robotic Exploration
ESA
8-10 rue Mario Nikis
F-75738 Paris Cedex 15
France

Dear Alvaro

Announcement of Opportunity for the GAIA Data Processing Archive Access element – Coordination Unit

I am writing in support of UK proposal to host the GAIA Archive Access Unit, led by Professor Gerry Gilmore at the DPAC, Cambridge. STFC, and more recently upon its creation, the UK Space Agency, has been fully committed to a UK role in GAIA, providing vital elements of the data processing for the GAIA mission. This work provides a logical extension. It will build upon the strong expertise within the UK consortium and I believe a proposal with a significant UK involvement will provide a very efficient and low risk option to deliver the required capabilities.

We are currently undertaking a Programmatic Review, which will include UK support for GAIA, but I fully expect this to generate strong support; it already has this from the community. Whilst financial commitment to the Coordination Unit will require peer review, in competition with other areas of our programme, I believe that a proposal will be favourably received. I am therefore happy to provide my outline endorsement,

Yours sincerely

Dr C Vincent

Head of Astronomy Division, STFC

cc: Prof G Gilmore