# Administrative Details

**Project Name: Gaia**

**Project Website:** https://www.cosmos.esa.int/web/gaia, https://www.gaia.ac.uk and https://www.gaia.ac.uk/data/uk-gaia-data-mining-platform

**Applicable Grants:**

- Core processing via ST/X00158X/1. This is the current Cambridge grant as Lead institute. There are additionally associated grants at Edinburgh (ST/X001601/1), Leicester (ST/X001687/1) and UCL/MSSL (ST/X001636/1).
- Data Mining Platform via ST/S002103/1 (This is the Cambridge grant as Lead institute. There are additionally associated grants at Edinburgh and Bristol. The Edinburgh grant specifically resourcing the DMP is ST/S001948/1)

**Lead Contact: Nicholas Walton naw@ast.cam.ac.uk**

**Further Contacts:**

- **Core Processing: Patrick Burgess: pwb@ast.cam.ac.uk**
- **Data Platform: Nigel Hambly: nch@roe.ac.uk / Dave Morris: dmr@roe.ac.uk**

*Version: v2: 20230104*

# Science Programme to be supported by IRIS

## Overview

Gaia[1] is a key ESA Cornerstone mission, launched in Dec 2013, with its most recent Data Release DR3 in June 2022. The Gaia in flight operations period has recently been extended by ESA's SPC to Dec 2022, with a pre-approval until 2025 (this being the point at which cold gas propellant for Gaia will be exhausted and in-flight operations will end). The extension to Gaia's Multi-Lateral Agreement governing ESA operations and national agency support until the final data release in 2030 has been agreed at the Nov 2022 meeting of the ESA SPC and is now being signed off by all the funding agencies.

The main Gaia UK activity is in operation and development of the Photometric Analysis System pipelines for Gaia, with the Gaia Data Processing Centre for photometry (DPCI) and photometric science alerts being located in Cambridge. In addition the UK have some involvement in the design of the Gaia Archive.

This proposal indicates a request for use of IRIS resources in two main areas as described below. These are a) longer term provision of resources to support the core UK Gaia data processing (recognising that the current Gaia hardware cluster at DPCI will no longer be supported post mid 2023) and b) support of the development of an added value science

[1] See https://www.cosmos.esa.int/web/gaia for information on Gaia. The latest June 2022 data release is at https://www.cosmos.esa.int/web/gaia/data-release-3

platform for end user Gaia data analysis, this supporting the Gaia Data Mining Platform (DMP) development.

The cases of both strands of Gaia work are presented in this unified case. The Core processing activity is supported by a UKSA processing grant “Gaia Data Flow System: 2021-2024”, whilst the Science Platform activity is one of the activities supported by a current UKRI-STFC grant “UK Gaia CU9: Delivering Gaia to the Community: 2019-2023”.

Gaia is a core UKSA and STFC mission and is grant funded by both the UKSA and STFC. Both funding lines are expected to be continued through to the final data release in 2030. The STFC grant renewal for the period 2023-2026 is currently being review by STFC’s PPRP. The next UKSA grant for the period 2024-2027 will be submitted later in 2023.

## Workflow 1: Gaia Core Data Processing

DPCI has provided hardware and software infrastructure for the processing of the photometric and low-resolution spectroscopy data collected by the Gaia satellite since the start of operations and during the pre-launch preparation and testing activities. Funding for DPCI hardware and staff resources has been part of a larger programme funded by the UK Space Agency and covering the larger Gaia-UK contribution to this European project. Activities directly related to the scientific research aspects of Gaia science are supported by the Science and Technology Facilities Council (STFC).

DPCI operations currently run on dedicated hardware hosted at the West Cambridge Data Centre. The current hardware is under maintenance contract until Spring 2023. Gaia in flight operations will run until early 2025. This implies that DPCI will need to make provisions for an alternative hardware solution for the final reduction of the data which will take place at the end of daily operations and will lead to the final data release in 2030.

The two first Gaia data releases had a huge impact on the scientific community, with a few papers per day using Gaia data since DR2. The third data release (DR3) was released June 2022 with an early release of some data including updated photometric data occurred in December 2020 (EDR3). DR3 has enhanced the catalogue with the addition of the low-resolution spectroscopy. Photometry and low-resolution spectroscopy are the prime products of the core processing operated by DPCI. Work on the fourth data release (DR4) is already in progress.

With three cycles of operations already completed, DPCI has a very clear understanding of the processing requirements both in terms of CPU and storage capacity.

Gaia-UK plans to use IRIS as a replacement for the current data processing infrastructure. The core photometric data processing is carried out in the UK on a locally hosted cluster in Cambridge which reaches end-of-support in Spring 2023. Data processing resources are required beyond the current lifetime of the cluster in order to complete processing for DR4 and, further ahead, for the fifth and final Gaia data release.

The data processing for Gaia involves hundreds of people in different institutions across Europe. Many different teams rely on timely delivery of processed data according to schedules agreed sometimes years in advance in order to deliver a coherent and consistent data release. Reliability and predictability of the infrastructure is critical. Delays in delivering allocated hardware resources or unplanned long periods of downtime could cause critical problems across the Europe-wide consortium and in the worst case lead to compromised data quality or delays to publish data release milestones.

## Workflow 2: Gaia Data Mining Platform

This resource request element is for compute resources to enable further development and operation of an end-user data-mining platform for public Gaia data releases. Ultimately we envision IRIS hosting a data-intensive science exploitation facility. The Gaia mission is currently in progress, and periodic static data releases are made to the world scientific community at appropriate stages in the data processing schedule to enable timely scientific exploitation. At the time of writing, there have been four releases from ESA/DPAC: DR1

(September 2016), DR2 (April 2018), EDR3 (December 2020) and "full" DR3 (June 2022). The next major release will be DR4, no earlier than the end of Q4 in 2025. The full DR3 data set comprises nearly 10 TB of science-ready data; DR4 is anticipated to be several hundreds of TB.

Up to EDR3 Gaia data releases consisted of billion-row, mainly flat, tabular data sets that map well to normal relational form. The data are hosted centrally at the European Space Astronomy Centre (ESAC) Science Data Centre and by several partner data centres around Europe plus affiliates around the world. Those centres present the data through user interfaces that expose a structured query language interface (using the IVOA standard Astronomy Data Query Language, an extension of standard SQL) that allows limited server-side manipulation of row subsets. However, from full DR3 onwards the scale and structure of the released data is expanding significantly. Time-resolved measurements and spectra are adding new array dimensionality and multiplying row counts by 10s to 100s. Advanced modelling of derived data products by the DPAC will produce statistical information (covariance, parameter probability density functions, etc.) that will add further to the volume and structures. Advanced usage modes for these data require high performance in IO and computation and these are not well served given the constraints imposed by SQL, relational design and the relational storage engine. Hence the Gaia project, in common with many large-scale astronomy survey missions, has prototyped a code-to-data platform-as-a-service for deployment on scale-out computing infrastructure in order to facilitate exploitation of the rapidly expanding survey mission data products. We have adapted the prototype and deployed on IRIS infrastructure to create the UK Gaia Data Mining Platform for end user science exploitation.

## Science Goals and Data Intensive Science

General science usage scenarios gathered in order to inform the design process for Gaia archive access have been well documented (Brown et al. 2012). While they are far too numerous to list here, concrete examples of science exploitation anticipated are well illustrated at recent workshops and "sprints": for example the 2019 Oxford "sprint"; "The Gaia universe: 53rd ESLAB symposium"; and the 2019 Santa Barbara "sprint". Many of these usage scenarios involve data-intensive workflows and data-mining techniques. They cover a multitude of science areas ranging from solar system science, through the solar neighbourhood and Galactic locale to astronomy on the largest Galactic scales. They concern the formation and evolution of planetary systems, stars and the Galaxy and impact nearly all areas of astrophysics at some level.

As stated above, the relational paradigm for data serving simply cannot cope with the above scale-out usages. This is why we have developed a science platform to address the inevitable "code-to-data" requirements of data-intensive workflows. The great opportunity afforded by the IRIS initiative is in addressing the devilish detail in deploying a practical solution to the problems encountered in making this a reality. While it is relatively easy to prototype a static bare-metal system based on third-party and COTS software stacks such as the Apache Spark ecosystem (see below) it is much harder to design and maintain a system that can cope with highly variable demand, scale as the data volume grows considerably, and is sufficiently flexible to be redeployed as infrastructure develops and changes. In particular it is now clear that virtualization and cloud infrastructure will play a big

role. We are demonstrating that the IRIS collaboration provides the means to address these fundamental development issues as well as to host a platform for UK data science exploitation in and beyond the medium term.

A typical workflow consists of a research astronomer writing bespoke application code that trawls/mines the bulk data: the input to a workflow is then the data set corresponding to a particular data release. Existing facilities do not allow this kind of open-ended, bulk data analysis - a science platform is required to provide the performance in terms of throughput and computation in a time frame amenable to exploration of the data and experimentation with alternative algorithms or ML tuning for example. Outputs can be anything from statistical summaries, selection of astronomical sources matching a certain pattern, candidate rare objects, improved models of time evolution in position (e.g. solar system objects, exoplanetary systems, …) or brightness (transiting exoplanets, microlensing events…). The scale-out platform available via IRIS provides the means to service these more ambitious usage modes.

## Science Director Approval

(Please leave blank for Science Director)