Who is In, and Who is Not? Determining the Gaia Survey Selection Function - GaiaUnlimited Initial Data Management Plan

This project is funded by the European Union's Horizon 2020 research and innovation program under grant agreement No 101004110.

1. Data summary

Purpose of the data collection/generation

The Gaia survey selection function cannot be specified solely in terms of analytical or numerical functions derived from basic principles. A detailed knowledge and modelling of the Gaia sky survey strategy, the on-board data collection (observation) process, and the subsequent steps taken in the data processing on ground are needed to generate a selection function. This represents data in the form of for example:

- The pointing of Gaia as a function of time (scanning law)
- A table of time intervals during which data collection was interrupted or data was lost
- Tables containing the filtering on data quality that was done during the processing
- Gaia and other survey data needed as input to construct selection functions

The project will also generate data in the form of numerical tables that can be used in conjunction with the selection function software tools to incorporate the Gaia survey selection function into science applications. Subsets of surveys will also be generated to serve as test data sets for the selection function tools.

Relation to the objectives of the project

The objectives of the project are to provide a detailed description of the Gaia survey selection function in the form of numerical tables and open source computer applications which can be applied to astronomical inference problems. The data listed above thus form an integral part of the project.

Data description

Data collected	**Type**	**Format**	**Volume**
Gaia catalogue data	Numerical	FITS/VoTable/CSV	GB to TB level
Data from other astronomical surveys	Numerical	FITS/VoTable/CSV	GB to TB level
Gaia scanning law	Numerical	FITS/VoTable/CSV	GB level
Gaia event logs	Mixed numeric/string	Spreadsheet	MB level
Data filtering logs	Mixed numeric/string	Spreadsheet	MB level
Data generated	**Type**	**Format**	**Volume**
Survey subsets	Numerical	FITS/VoTable/CSV	GB to 100 GB

Selection function tables	Numerical	FITS/VoTable/CSV	GB to 100 GB
TBD other generated data	Numerical	FITS/VoTable/CSV	TBD

- The collected survey data is all publicly available (and hence is re-used).
- Data on the Gaia scanning law, event logs, and filtering logs is available only within the Gaia Data Processing and Analysis Consortium, but will be turned into publicly available versions.

Data utility

- The collected data will be used within the GaiaUnlimited project to construct the survey selection function and develop the corresponding tools.
- The generated data will be used by scientists interested in including the Gaia survey selection function in their projects.

2. FAIR DATA

2.1 MAKING DATA FINDABLE, INCLUDING PROVISIONS FOR METADATA:

Discoverability

The GaiaUnlimited web portal (https://gaia-unlimited.org) will be the primary gateway to accessing the data generated in this project. The page dedicated to the data will contain the links to the generated data sets. The latter will be hosted at zenodo.org and at the ESA Gaia archive (https://gea.esac.esa.int/archive/).

Identifiability

The data sets will be given a descriptive name, a DOI, and will be tagged with keywords.

Naming conventions

The precise naming conventions remain to be defined. We will strive for descriptive names which include time stamps and versioning information.

Search keywords

Data sets will be tagged with standard astronomical keywords, as well as specific terms such as "Gaia", "selection function", etc.

Versioning

The approach to versioning remains to be defined. For software tools the versioning will be done through Github. For the data files we will very likely follow the Gaia Data Processing and Analysis Consortium standards and practices.

Metadata standards

Standards commonly employed in astronomy will be used:

- FITS
- IVOA (VoTable)
- CSV

The choice of standard will depend on the use case and size of the data set.

2.2 Making data openly accessible:

Which data

All data generated in the context of the GaiaUnlimited project, and which is needed to use the survey selection function tools, will be made openly available.

Remarks:

- The astronomical data from public surveys such as Gaia are already openly available and fall outside the scope of this document.
- As noted above, DPAC internal data will not be made openly available in direct form, but will be turned into openly available versions that can be used with the selection function tools.

How and where

The data and associated metadata will be made available through the ESA Gaia archives and any other astronomical data centres interested in hosting these data. In addition the data and metadata will be deposited with Zenodo.

Data access methods and tools

- The data can be queried (through ADQL) from the ESA Gaia archives or directly downloaded and then used with any tools preferred by the user.
- However, typically the user will use the software tools developed by GaiaUnlimited. These tools will be made available open source through the GaiaUnlimited website which will link to the Github repository.
- Documentation will be included with the software tools, for example through the facilities provided by https://readthedocs.org/.

2.3 Making data interoperable

- When accessing the data through the ESA Gaia archives or through another astronomical data center, the standards of the data center will be followed. These comply to the International Virtual Observatory Alliance (IVOA) standards.
- The data provided in directly downloadable format will be provided as FITS, VoTable, CSV, eCSV files.
- The vocabulary will follow as much as possible standard astronomical usage.

2.4 Increase data re-use (through clarifying licenses)

License

As the main repository for the data will be the ESA Gaia Archives, the applicable license is the Gaia Data License. It is quoted here for clarity:

The Gaia data are open and free to use, provided credit is given to 'ESA/Gaia/DPAC'. In general, access to, and use of, ESA's Gaia Archive (hereafter called 'the website') constitutes acceptance of the following general terms and conditions. Neither ESA nor any other party involved in creating, producing, or delivering the website shall be liable for any direct, incidental, consequential, indirect, or punitive damages arising out of user access to, or use of, the website. The website does not guarantee the accuracy of information provided by external sources and accepts no responsibility or liability for any consequences arising from the use of such data.

Restrictions

Once the data generated by GaiaUnlimited has been made openly available no re-use restrictions apply, including after the end of the project.

Quality assurance

GaiaUnlimited will follow the standard Gaia Data Processing and Analysis Consortium QA practices. This includes:

- A well designed and documented data model.
- Versioning of the data ingested into the ESA Gaia archives.
- Quality control at the moment the data is included in the archive (e.g., numerical range checking against the data model).
- Before making the data openly available it will be extensively validated through use within the GaiaUnlimited project and by users participating in the community workshops.

Life cycle

The data will remain re-usable forever.

3. Allocation of resources

Responsibilities

The data management for GaiaUnlimited will be the responsibility of A. Brown, the project PI and leader of the management work package (WP1). He will liaise with the relevant persons in the Gaia Data Processing and Analysis consortium to ensure the GaiaUnlimited data is included in the ESA Gaia Archive and he will oversee the deposition of data elsewhere, such as at Zenodo.

Value

For as long as the Gaia data releases will be used by scientists around the world (and this is expected to be the case for decades to a century) the selection function data and tools will remain indispensable.

Costs

The data generated by GaiaUnlimited is expected to be sufficiently small in volume that no significant storage costs will be incurred. The data can reside on the storage available at the institutes participating in GaiaUnlimited and the archiving of the open available data products will be done at the ESA Gaia Archives, or at Zenodo, both publicly funded.

4. Data security

- The openly available data will be hosted at the ESA Gaia Archive and Zenodo, both of which have the facilities in place for secure storage and data recovery.
- The data collected and generated in the course of the GaiaUnlimited project will be stored at Leiden University where the university IT team will take care of securing the storage and backups.
- No use of sensitive data is foreseen.

5. Ethical aspects

Ethical issues are not foreseen.

6. OTHER

N/A