Upcoming Data challenges in Gaia and the VO

Juan González-Núñez, A. Mora, J. Salgado, R. Gutiérrez-Sánchez, JC. Segovia, J. Duran, E. Racero, J. Osinde, M. Marcos, J. Bakker, D. Baines, B. Merín, C. Arviset

ESAC Science Data Centre (ESDC)
European Space Agency

Where we stand: DR1 and DR2

- DR1: mainly catalogues
  - gaia_source: positions and magnitudes
  - tgas_source: 5-p astr. solution XM w/Tycho2

- DR2: catalogues and first multidimensional data
  - gaia_source, variables, asteroids, cross-matches
  - ~550,000 light curves. BP, G and RP bands

Gaia DPAC Data Processing & Analysis Consortium

gaia

DR3, EDR3 and DR4: contents

- EDR3: Q3 2020
  - catalogues: astrometry, int. Photometry, ext. Sources
- DR3: H2 2021
  - Based on same 34 months of input data
  - catalogues: source classification, astrophysical parameters, radial velocities, phot. variability
  - Mean BP/RP/RVS Spectra *
  - Light curves
- DR4: TBD
  - All epoch and transit data for all sources *

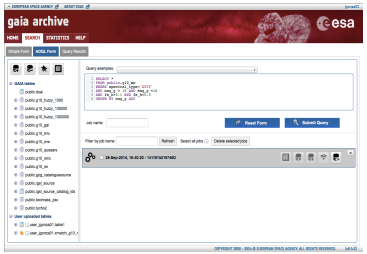
DR3, EDR3 and DR4: **sizes**

- EDR3: equivalent to DR2
- DR3:
  - Photometry: ~100M variables x 250 points ~ 25B epochs
  - Spectroscopy: ~10-100M spectra x 100-1000 el. ~ 10B el.
- DR4
  - Epoch astrometry: ~2B objects ~1000 epoch ~ 2T epochs
  - Epoch spectra: ~200B spectra x ~100 res. el. ~ 20T res. el.
  - Auxiliary data? TBD

Gaia DPAC Data Processing & Analysis Consortium

gaia

The ESA Gaia Archive: A VO Inside Architecture

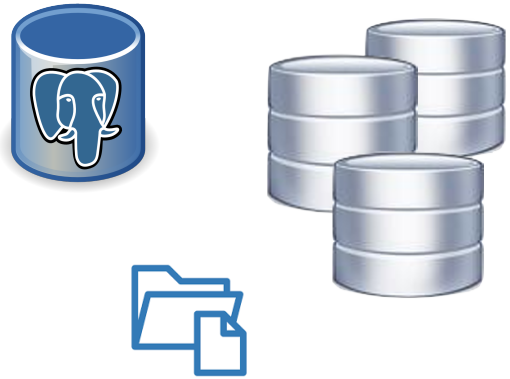


Programmatic: Python, Java, command Line

VO Enabled Apps

Open APIs

- TAP+
- DataLink
- VOSpace
- CSearch



Public area

- Publicly released data

Restricted area

- Dataduring validation

User Space

- User-uploaded data

Scaling up: Serving Multidimensional Data Products

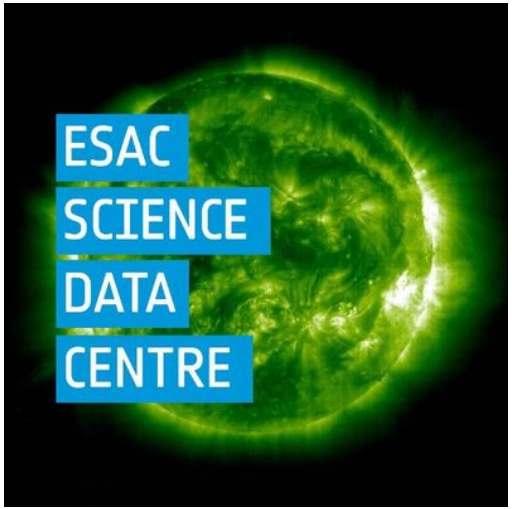
- **TAP+**
  - Catalogues, source clasification, SSOs.
  - Efficiently “indexable” data
  - Benefits from storage in RDBMS

- **DataLink**
  - Associated data products (Spectra, Light Curves).
  - DataLink allows for efficient DataModel-agnostic search over large datasets based on product level metadata; perfect fit for Gaia Spectra or Light curves
  - Mechanisms for linking TAP searches with associated data products
  - Scales to DR3/4 data volumes

Scaling up: Bringing code to the data

- **Gaia** Python modules:
  - **astroquery.gaia**: Gaia TAP+, **astroquery.utils.tap**: generic TAP
  - **DataLink** access (in dev.)
- **2** usage scenarios:
  - Advanced data **selection, filtering, reprocessing through archive services and analysis** in the client side
    - TAP, DataLink, SODA, etc. + astroquery libraries
    - Best fit for highly selective usage scenarios
  - Direct data analysis in the server side through **Notebook** services
    - ESA **SEPP Plattform** development ongoing with ESA data access libraries, including massive parallel processing infrastructures
    - Best fit for data access intensive applications

Gaia Archive and the VO



- ESDC a long time partner of the IVOA
  - Implementation of VO protocols in mission archives
  - Development of VO tools and Registries
  - Contribution to standards development
- Gaia DPAC includes major VO experts that provide guidance
- VO awareness
  - Interoperable data access protocols
  - Interoperable data models and formats, with VO awareness fed back to the data production chain

esa VO Virtual Observatory

Gaia DPAC Data Processing & Analysis Consortium

Gaia Archive and the VO – Next Steps

- Evolution of current standards
  - ADQL 2.1
  - TAP 1.1
  - DataLink 1.1

- Server Side data processing
  - Eg. SODA 1.0?

- Data Model definition and implementation
  - Spectral DM
  - Time Series DM/ Serialisation draft
  - Source DM draft