Maximising data reach: bringing the Gaia dataset to the world


ESAC Science Data Centre (ESDC) - ESA
Why Open Archives?
Why Open Archives?

Gaia
Referred Gaia papers since launch

<table>
<thead>
<tr>
<th>Number of Papers:</th>
<th>Date Created:</th>
<th>Date Last Modified:</th>
</tr>
</thead>
<tbody>
<tr>
<td>2306</td>
<td>Apr 2 2019, 2:18pm</td>
<td>Sep 23 2019, 9:41pm</td>
</tr>
</tbody>
</table>

![Bar chart showing the number of refereed papers from 2014 to 2019.](chart.png)
Why Open Archives?

- How can open scientific archives increase the scientific output/outcome of a mission?
Why Open Archives?

- How can open scientific archives increase the scientific output/outcome of a mission?

- By **maximising data availability**: Retrieve and analyse data seamlessly -> increase in productivity
Why Open Archives?

- How can open scientific archives increase the scientific output/outcome of a mission?

  - By **maximising data availability**: Retrieve and analyse data seamlessly -> increase in productivity

  - Open data and software policies also enable a **more efficient usage of project resources** through **reuse and collaboration**
Why Open Archives?

- How can open scientific archives increase the scientific output/outcome of a mission?

  - By **maximising data availability**: Retrieve and analyse data seamlessly -> increase in productivity
  
  - Open data and software policies also enable a **more efficient usage of project resources** through **reuse and collaboration**
  
  - Improve Data **preservation** through Open Data Models
Openness in Scientific Archives

- The ESA Gaia Archive, ESA Gaia Mission and DPAC Consortium have made efforts in 4 areas:
  - Open Data
  - Open Software
  - Open Protocols
  - Open Data Models
Open Data

- Gaia Data License:
  - "The Gaia data are **open** and **free** to use, provided credit is given to 'ESA/Gaia/DPAC' [...]"

- Does not specifically state limitations to use and reuse. Enables distribution by third parties of the full dataset (redistribution)

- Attribution requirement – compatible with Open Data

- No proprietary exploitation period
  - Direct access to all Scientific community
Data redistribution

- DPAC Partner and Associate Data Centres:
  - Validated data delivered in advance
  - Eg. CDS, AIP, ARI, ASI SSDC...

- Positive effects in functionality
  - Local copies per service enable for quicker data correlation
  - Eg. crossmatch between catalogues

- Traffic is lower, more stable and more predictable due to aggregation of traffic
  - Less resources per datacentre needed, and more predictable
Open Software

OS Off the Shelf

- DB
  - PostgreSQL
  - Greenplum
- Content SV
  - Apache Tomcat
  - Red Hat

Components/FW

- GWT
- OpenJDK
- spring security
- jasig

CI/CC

- Maven
- Nexus
- sonarqube
- Jenkins

OS Libraries

- VOLLT
- TAPLib
- UWSLib
- ADQLLib

OS Communities

- Python
- GitHub
- Astropy Project
Database Software

• PostgreSQL
  • Ready to use: only requiring tuning to specific HW and general administration costs – efficient usage of resources
  • High uptimes (No DB software caused downtime in 3 years of public ops)
  • Remarkable performance
  • Cluster versions (Greenplum CE)

• Q3C
  • Geometrical indexing extension
  • S.Koposov, O. Bartunov.
Open Software TAP

- VOLLT: Java library implementing several VO protocols – G. Mantelelt
  - TAPLib, UWSLib, ADQLLib
- In-house development only of TAP+ adaptation (authentication, user spaces, etc.) – **reduced dev. effort**
Open Software TAP

- VOLLT: Java library implementing several VO protocols – G. Mantelet
  - TAPLib, UWSLib, ADQLLib
- In-house development only of TAP+ adaptation (authentication, user spaces, etc.) – **reduced dev. effort**
- Released TAP+ fork (2017)
  - GNU-LGPL
  - https://github.com/esdc-esac-esa-int/gaia
Open Source communities

- Data Centres should jump beyond the “service” level to provide data access libraries on Astronomical data analysis packages that are Open Source.

- Bringing specific data access mechanisms in the languages/environments where data analysis is happening dramatically reduces the data access barrier, increases data usage and scientific productivity.

- Community contributions keep the overall development cost low to each data centre.
Open Source communities

- Eg. Astroquery
- GITHUB with Branch/Pull request mechanism enables to integrate minor contributions from many developers/institutions in an agile workflow
  - Branch (open)
  - Codify
  - Pull request (open)
  - Review
  - Merge
Open Source communities

- Gaia: Extensions of Astroquery contributed on specific packages, eg.
  - astroquery.gaia
  - astroquery.utils.tap (under int. with PyVo)
Open Source communities

- Gaia: Extensions of Astroquery contributed on specific packages, eg.
  - astroquery.gaia
  - astroquery.utils.tap (under int. with PyVo)

- ESDC has made several other contributions to OS communities:
  - astroquery.esasky
  - astroquery.hubble
  - astroquery.xmm (ongoing)
  - Proba2 (ongoing)
Open Protocols

• Long standing commitment by ESA with the IVOA

• IVOA fosters the creation of open standards
  • Freely available to use and re-use
  • Open definition process (transparency, broad consensus)

• Key dates
  • 2002 – ESA joined IVOA upon its foundation
  • 2005 – First ESA VO services (on-top layer)
    • ISO & XMM Image/Spectra (SIAP/SSAP) services
  • 2014 – First ESA VO Inside archive (Gaia)
  • 2016 – First ESA VO Inside multi-mission archive (ESASky)
Open Protocols: Early implementations

- ESA VO Registry of Resources
  - Released **June 2005**
  - First ESA service with **ADQL query**
  - Before ADQL became REC
  - Experimental **14 years ago**

- Early implementations help to drive standards development and also **build specific knowledge**

WHERE #description# LIKE '%esa%' AND #curation/publish
Open Protocols in the ESA Gaia Archive

• Many VO protocols are the **core** backbone of the Gaia Archive server side, not an on-top addition over tailored protocols
  • TAP -> catalogues
  • DataLink -> associated data products

• All APIs used by the Archive **are public and documented**

• When a VO protocol does not fully fit the purpose, it is **extended**, keeping compatibility. Eg. TAP+

http://archives.esac.esa.int/gaia
### Gaia Query sources

- Others can query your service!
  - Availability through other data access tools increases the data availability and hence, scientific productivity

- Usage of open protocols as internal data centre protocols reduces implementation costs and associated costs of external interfaces, for a more efficient development

<table>
<thead>
<tr>
<th>I/F</th>
<th>Tool</th>
<th>Origin</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAP</td>
<td>Python</td>
<td>astroquery.gaia</td>
<td>17.7M</td>
</tr>
<tr>
<td>ConeSearch</td>
<td>TOPCAT</td>
<td></td>
<td>9.1M</td>
</tr>
<tr>
<td>TAP</td>
<td>Python</td>
<td>other</td>
<td>184K</td>
</tr>
<tr>
<td>TAP</td>
<td>Web</td>
<td>archives.esac.esa.int/gaia</td>
<td>128K</td>
</tr>
<tr>
<td>TAP</td>
<td>TOPCAT</td>
<td></td>
<td>72K</td>
</tr>
<tr>
<td>TAP</td>
<td>Wget/curl</td>
<td></td>
<td>69K</td>
</tr>
<tr>
<td>TAP</td>
<td>Python</td>
<td>PyVo</td>
<td>26K</td>
</tr>
<tr>
<td>TAP</td>
<td>Browser</td>
<td></td>
<td>12K</td>
</tr>
<tr>
<td>TAP</td>
<td>Java</td>
<td>GACS Java lib</td>
<td>11K</td>
</tr>
<tr>
<td>TAP</td>
<td>Java</td>
<td>other</td>
<td>10K</td>
</tr>
<tr>
<td>ConeSearch</td>
<td>Java</td>
<td></td>
<td>5K</td>
</tr>
<tr>
<td>TAP</td>
<td>Python</td>
<td>urllib</td>
<td>1.2K</td>
</tr>
</tbody>
</table>

June-September 2019 data
Gaia Query sources

<table>
<thead>
<tr>
<th>I/F</th>
<th>Tool</th>
<th>Origin</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAP</td>
<td>Python</td>
<td>astroquery.gaia</td>
<td>17.7M</td>
</tr>
<tr>
<td>ConeSearch</td>
<td>TOPCAT</td>
<td></td>
<td>9.1M</td>
</tr>
<tr>
<td>TAP</td>
<td>Python</td>
<td>other</td>
<td>184K</td>
</tr>
<tr>
<td>TAP</td>
<td>Web</td>
<td>archives.esac.esa.int/gaia</td>
<td>128K</td>
</tr>
<tr>
<td>TAP</td>
<td>TOPCAT</td>
<td></td>
<td>72K</td>
</tr>
<tr>
<td>TAP</td>
<td>Wget/curl</td>
<td></td>
<td>69K</td>
</tr>
<tr>
<td>TAP</td>
<td>Python</td>
<td>PyVo</td>
<td>26K</td>
</tr>
<tr>
<td>TAP</td>
<td>Browser</td>
<td></td>
<td>12K</td>
</tr>
<tr>
<td>TAP</td>
<td>Java</td>
<td>GACS Java lib</td>
<td>11K</td>
</tr>
<tr>
<td>TAP</td>
<td>Java</td>
<td>other</td>
<td>10K</td>
</tr>
<tr>
<td>ConeSearch</td>
<td>Java</td>
<td></td>
<td>5K</td>
</tr>
<tr>
<td>TAP</td>
<td>Python</td>
<td>urllib</td>
<td>1.2K</td>
</tr>
</tbody>
</table>

June-September 2019 data

- Correspondence between # of queries and knowledge extracted is by no means direct, but...
Correspondence between # of queries and knowledge extracted is by no means direct, but…

- **65%** of the query traffic corresponds to Open Source projects

### Gaia Query sources

<table>
<thead>
<tr>
<th>I/F</th>
<th>Tool</th>
<th>Origin</th>
<th>Traffic</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAP</td>
<td>Python</td>
<td>astroquery.gaia</td>
<td>17.7M</td>
</tr>
<tr>
<td>ConeSearch</td>
<td>TOPCAT</td>
<td></td>
<td>9.1M</td>
</tr>
<tr>
<td>TAP</td>
<td>Python</td>
<td>other</td>
<td>184K</td>
</tr>
<tr>
<td>TAP</td>
<td>Web</td>
<td>archives.esac.esa.int/gaia</td>
<td>128K</td>
</tr>
<tr>
<td>TAP</td>
<td>TOPCAT</td>
<td></td>
<td>72K</td>
</tr>
<tr>
<td>TAP</td>
<td>Wget/curl</td>
<td></td>
<td>69K</td>
</tr>
<tr>
<td>TAP</td>
<td>Python</td>
<td>PyVo</td>
<td>26K</td>
</tr>
<tr>
<td>TAP</td>
<td>Browser</td>
<td></td>
<td>12K</td>
</tr>
<tr>
<td>TAP</td>
<td>Java</td>
<td>GACS Java lib</td>
<td>11K</td>
</tr>
<tr>
<td>TAP</td>
<td>Java</td>
<td>other</td>
<td>10K</td>
</tr>
<tr>
<td>ConeSearch</td>
<td>Java</td>
<td></td>
<td>5K</td>
</tr>
<tr>
<td>TAP</td>
<td>Python</td>
<td>urllib</td>
<td>1.2K</td>
</tr>
</tbody>
</table>

June-September 2019 data
Correspondence between # of queries and knowledge extracted is by no means direct, but...

- **65%** of the query traffic corresponds to Open Source projects
- **98%** of the query traffic corresponds to Open Source projects + VO tools

<table>
<thead>
<tr>
<th>I/F</th>
<th>Tool</th>
<th>Origin</th>
<th>Traffic</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAP</td>
<td>Python</td>
<td>astroquery.gaia</td>
<td>17.7M</td>
</tr>
<tr>
<td>ConeSearch</td>
<td>TOPCAT</td>
<td></td>
<td>9.1M</td>
</tr>
<tr>
<td>TAP</td>
<td>Python</td>
<td>other</td>
<td>184K</td>
</tr>
<tr>
<td>TAP</td>
<td>Web</td>
<td>archives.esac.esa.int/gaia</td>
<td>128K</td>
</tr>
<tr>
<td>TAP</td>
<td>TOPCAT</td>
<td></td>
<td>72K</td>
</tr>
<tr>
<td>TAP</td>
<td>Wget/curl</td>
<td></td>
<td>69K</td>
</tr>
<tr>
<td>TAP</td>
<td>Python</td>
<td>PyVo</td>
<td>26K</td>
</tr>
<tr>
<td>TAP</td>
<td>Browser</td>
<td></td>
<td>12K</td>
</tr>
<tr>
<td>TAP</td>
<td>Java</td>
<td>GACS Java lib</td>
<td>11K</td>
</tr>
<tr>
<td>TAP</td>
<td>Java</td>
<td>other</td>
<td>10K</td>
</tr>
<tr>
<td>ConeSearch</td>
<td>Java</td>
<td></td>
<td>5K</td>
</tr>
<tr>
<td>TAP</td>
<td>Python</td>
<td>urllib</td>
<td>1.2K</td>
</tr>
</tbody>
</table>

Query origin:

June-September 2019 data
98%

Of all the query traffic would not exist without

- Open Source projects contributions
- Open VO protocols compatibility
External queries in the Gaia archive

• Since V2.6 You can also query other TAPs!
  • Search on GloTS: table level metadata search for any public table
  • Almost 50K catalogues (and growing!)
External queries in the Gaia archive

- Since V2.6 You can also query other TAPs!
- Search on GloTS: table level metadata search for any public table
- Almost 50K catalogues (and growing!)
• External query results (in blue) are stored in the Gaia TAP service
• Possibility to upload them to user DB spaces or cross reference in subsequent queries
Open Protocols

- Many other service implementors rely on open protocols: a quickly increasing toolbox
  
- IVOA protocols require 2 reference implementations to become REC
  - Mainly Open Source
  - Great efficiency by using these implementations and spending project resouces in extending them

- TAP, ADQL, UWS, VOSpace, DataLink, etc.
Open Data Models

- Requires to go from Mission level thinking to global, long-term metadata
- Long term data access is guaranteed by
  - Adaptation of internal mission DM with view in **standard DMs**
  - **On the fly** serialisation through DataLink (adaptable) for data products

- Gaia DM

- VO DM serialisations
  - Spectrum, TimeSeries, etc.
Is it working?

- It is the overall usage of Gaia mission data that matters
Is it working?

- It is the overall usage of Gaia mission data that matters
- Still, ESA Gaia service stats provide some positive trends
- Publications with references to the Archive increasing in line with total Gaia publications
Is it working?

- It is the overall usage of Gaia mission data that matters
- Still, ESA Gaia service stats provide some positive trends
  - Publications with references to the Archive increasing in line with total Gaia publications
  - 5-6K users retrieving science level data from the archive monthly
Is it working?

- It is the overall usage of Gaia mission data that matters
- Still, ESA Gaia service stats provide some positive trends
  - Publications with references to the Archive increasing in line with total Gaia publications
  - 5-6K users retrieving science level data from the archive monthly
  - 2K registered archive users
Thanks for your attention

Questions?