

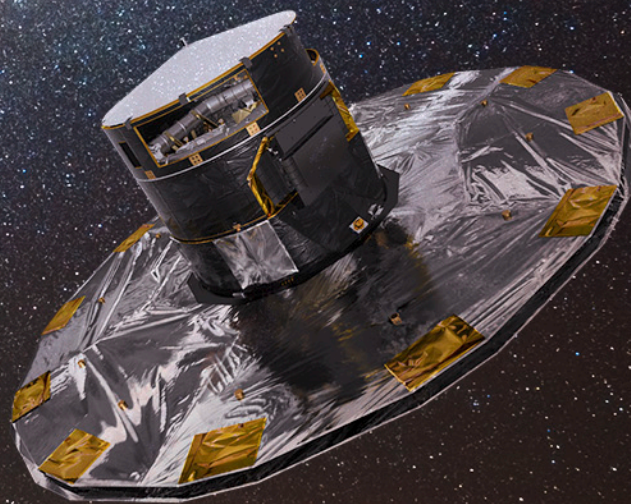
Open data Access / Provisioning



# Maximising data reach: bringing the Gaia dataset to the world

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

ESAC Science Data Centre (ESDC) - ESA



ESA UNCLASSIFIED - For Official Use



European Space Agency

# Why Open Archives?



# Why Open Archives?

## Gaia

Refereed Gaia papers since launch

Number of Papers:

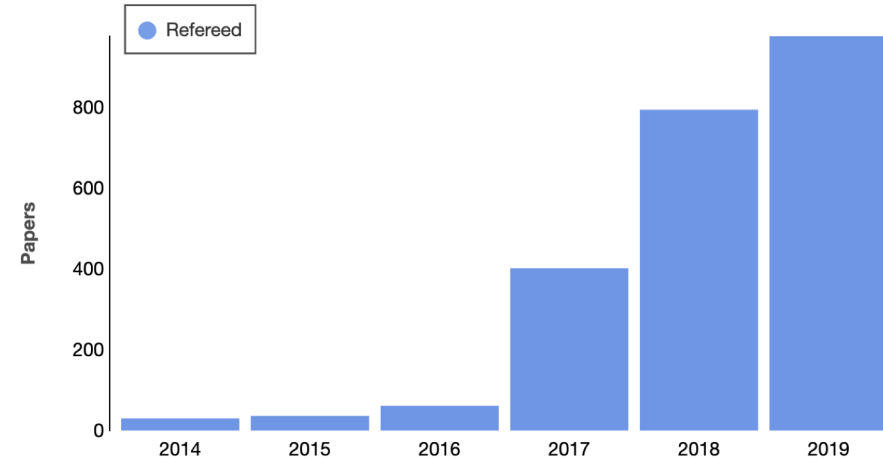
2306

Date Created:

Apr 2 2019, 2:18pm

Date Last Modified:

Sep 23 2019, 9:41pm



# Why Open Archives?

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

## Gaia

Refereed Gaia papers since launch

Number of Papers:

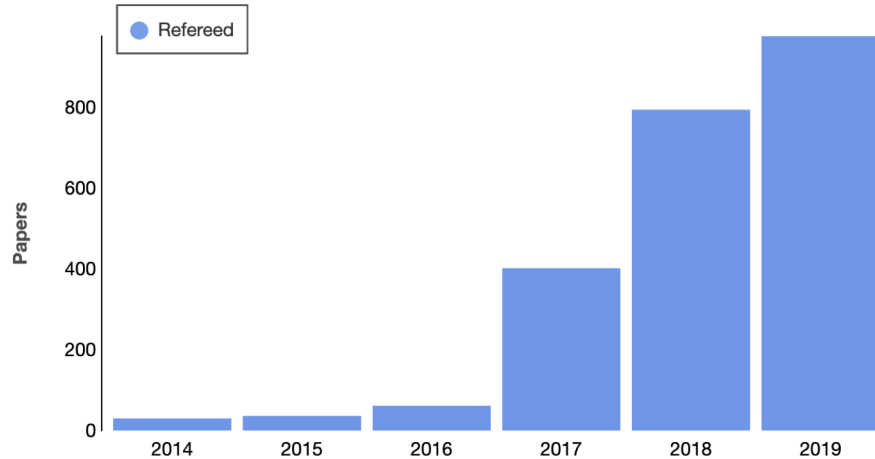
2306

Date Created:

Apr 2 2019, 2:18pm

Date Last Modified:

Sep 23 2019, 9:41pm





# Why Open Archives?

## Gaia

Refereed Gaia papers since launch

Number of Papers:

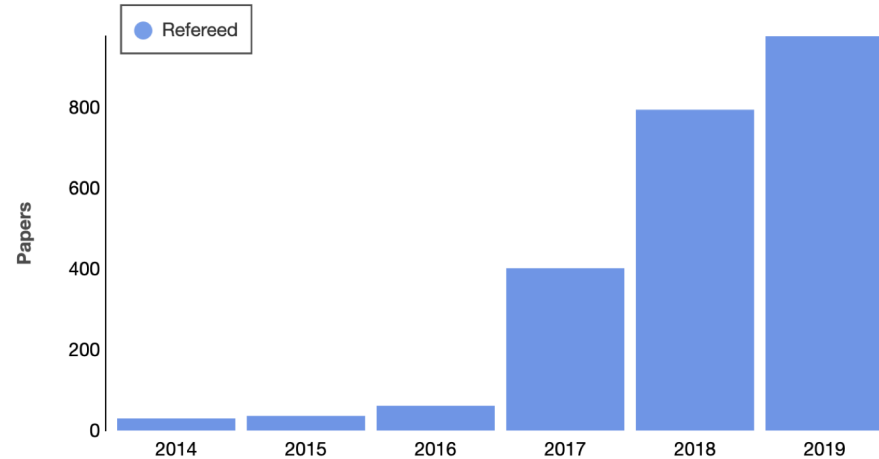
2306

Date Created:

Apr 2 2019, 2:18pm

Date Last Modified:

Sep 23 2019, 9:41pm



- 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?

## Gaia

Refereed Gaia papers since launch

Number of Papers:

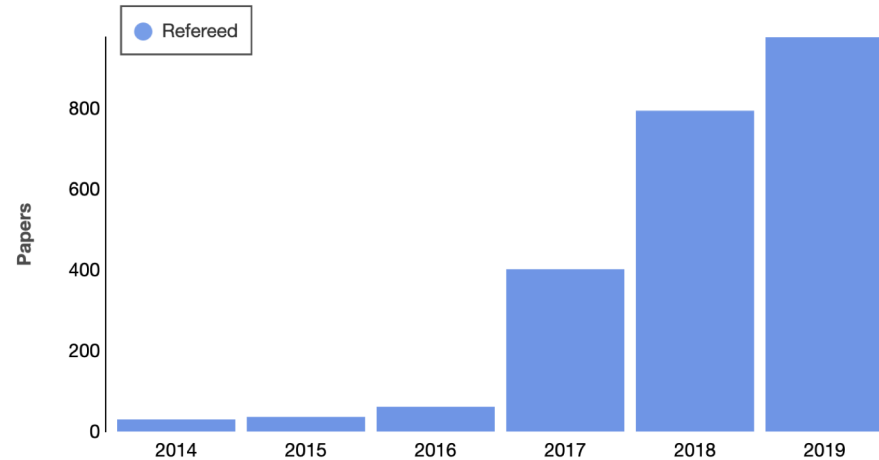
2306

Date Created:

Apr 2 2019, 2:18pm

Date Last Modified:

Sep 23 2019, 9:41pm



- 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?

## Gaia

Refereed Gaia papers since launch

Number of Papers:

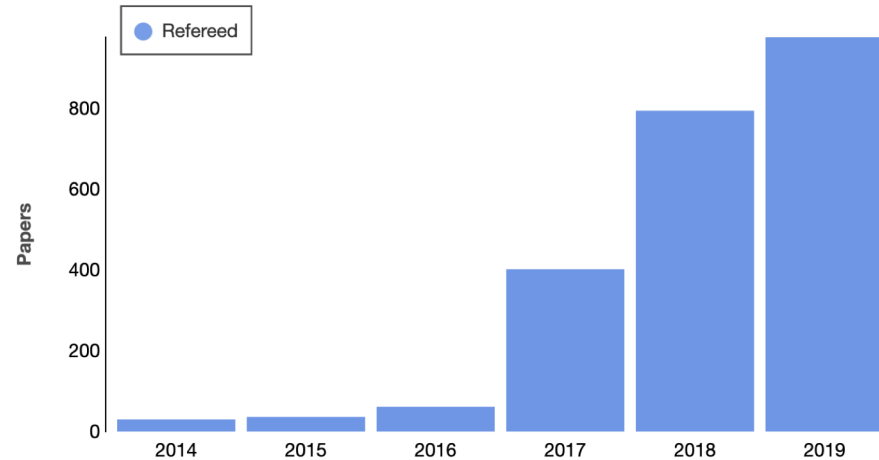
2306

Date Created:

Apr 2 2019, 2:18pm

Date Last Modified:

Sep 23 2019, 9:41pm



- 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

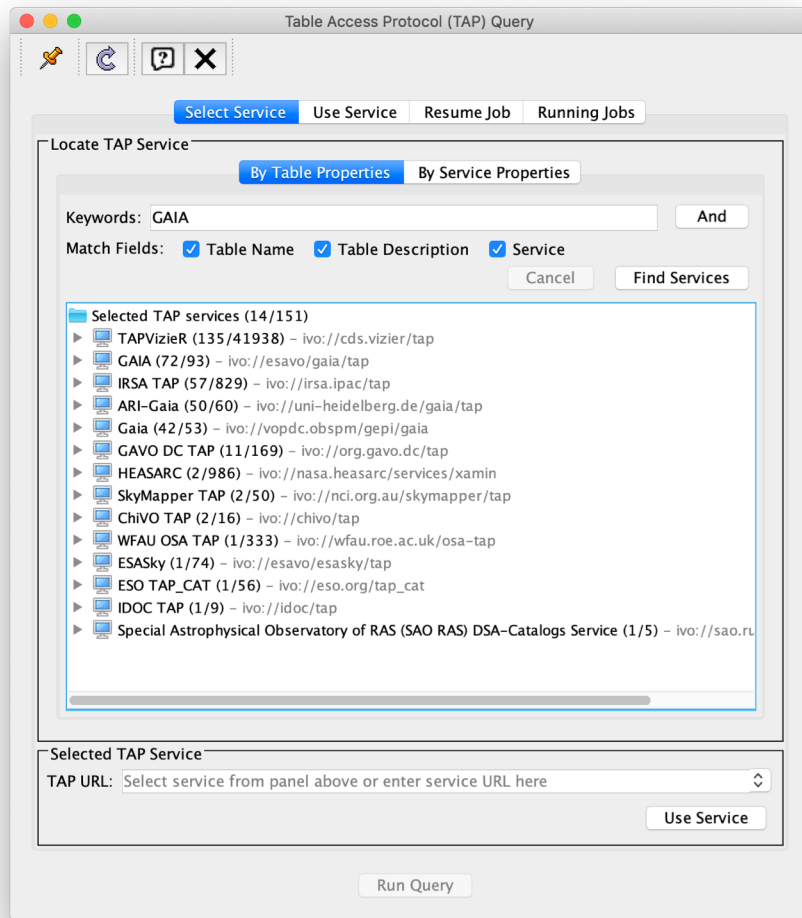




# Open Data

- 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





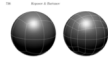
# Data redistribution

- 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



### Content SV



Apache Tomcat



### Components/FW



GWT



OpenJDK



### CI/CC



## OS Libraries

- VOLLT
  - TAPLib
  - UWSLib
  - ADQLLib

## OS Communities



python™

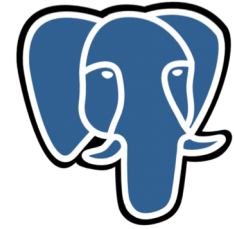


GitHub



# 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.



PostgreSQL

736

Koposov & Bartunov

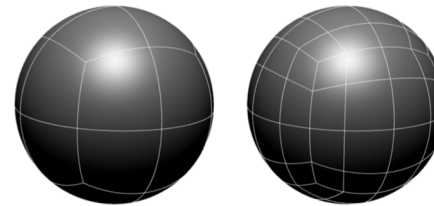
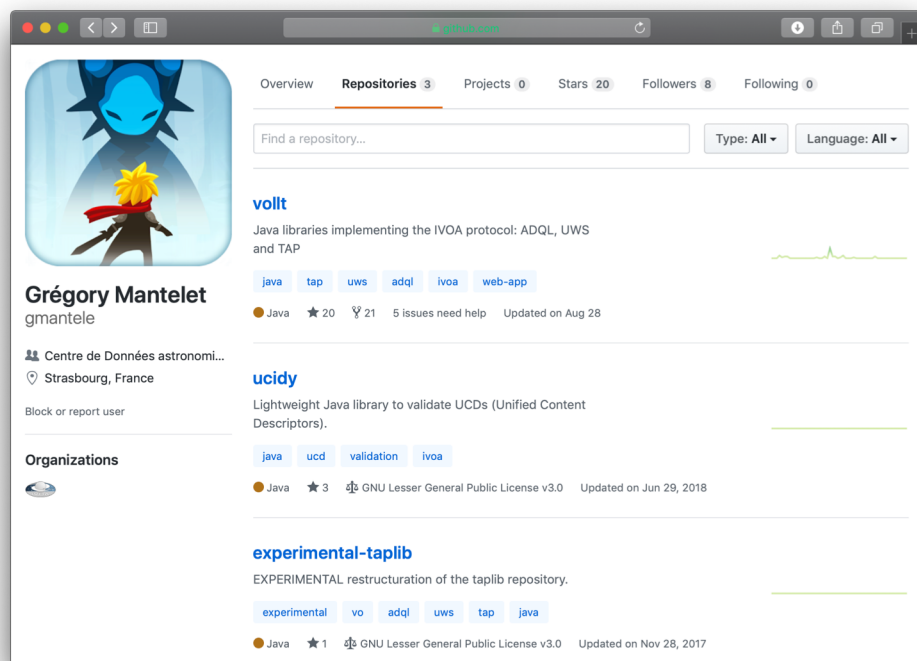


Figure 1. The sphere segmentation in Q3C.



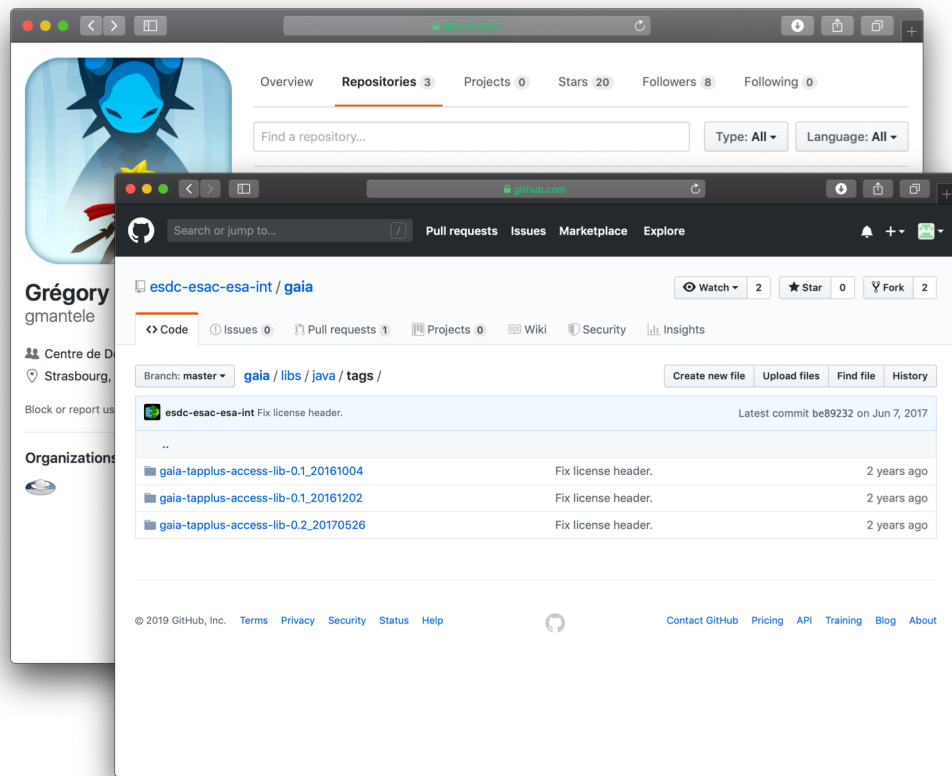


# Open Software TAP



- 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**

# Open Software TAP



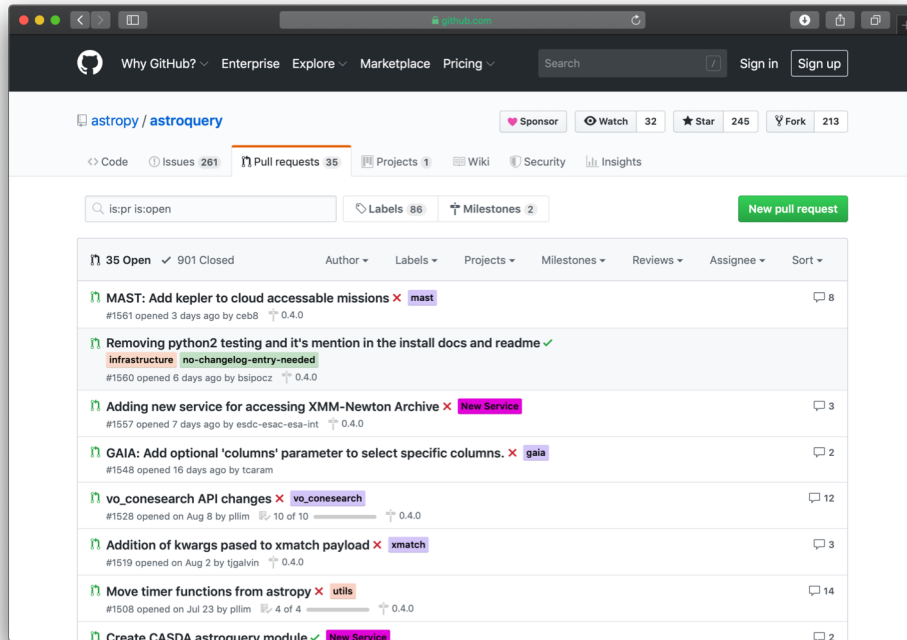
- VOLLT: Java library implementing several VO protocols – G. Mantelet
  - TAPLib, UWSLib, ADQLLib
- In-house development only of TAP+ adaptation (authentication, user spaces, etc.) – **reduced dev. effort**
- Released TAP+ fork (2017)
  - GNU-LGPL
  - <https://github.com/esdc-esac-esa-int/gaia>

# 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



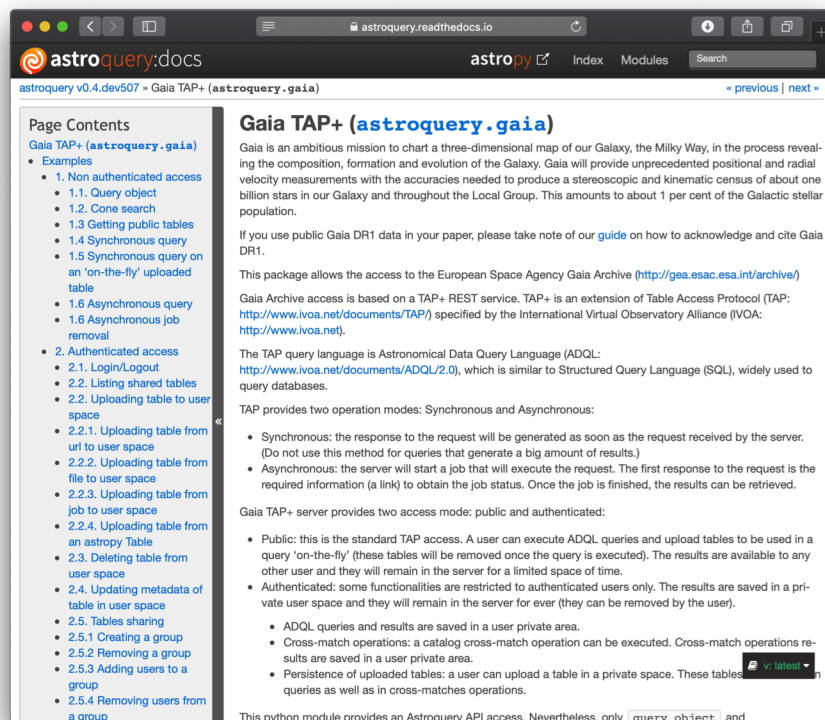
# Open Source communities



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

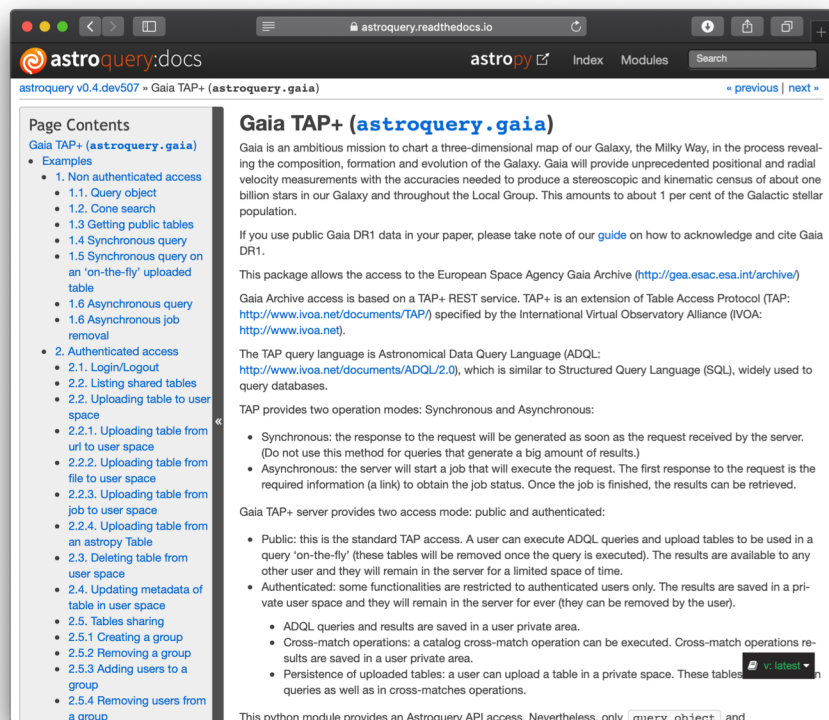


# Open Source communities



- 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)

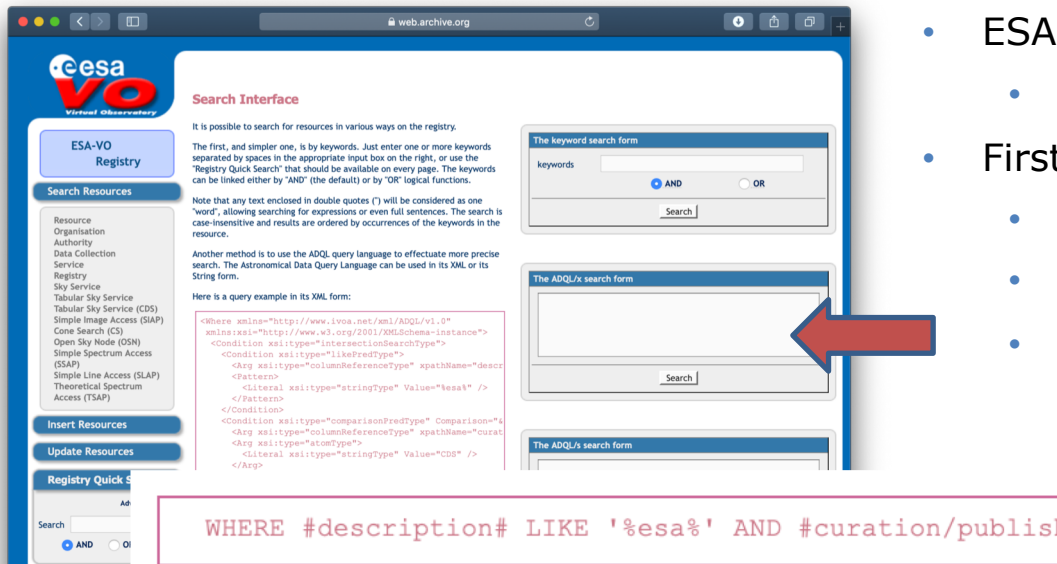


# Open Protocols

- 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



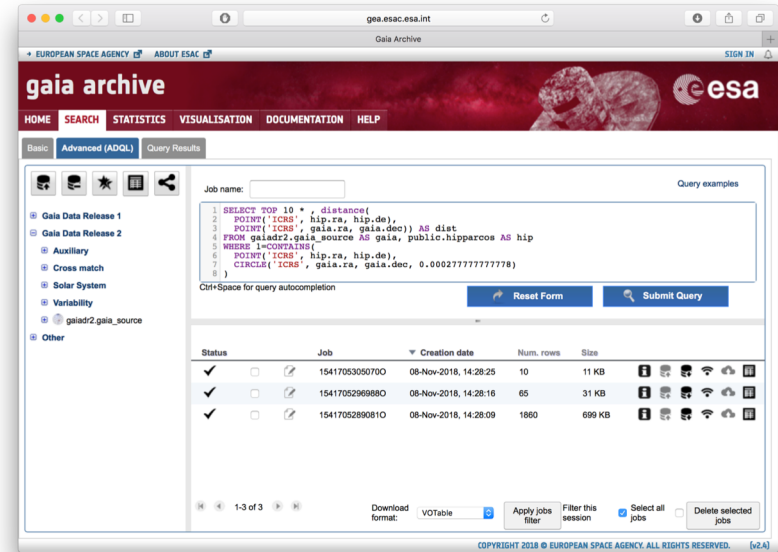
Note that the elements to search should be specified using the `xpathName` attribute for the XML form and enclosed in sharp characters (#) for the String form. The ADQL versions currently supported are v0.7.4, v0.8, v0.9 and v1.0 ; it must be specified by using the proper XML namespace declaration.

- 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**



# 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

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

June-September 2019 data

- 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

# Gaia Query sources

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

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

June-September 2019 data

# Gaia Query sources

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

- Correspondence between # of queries and knowledge extracted is by no means direct, but...
- **65%** of the query traffic corresponds to Open Source projects

June-September 2019 data

# Gaia Query sources

Query origin:

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

- 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

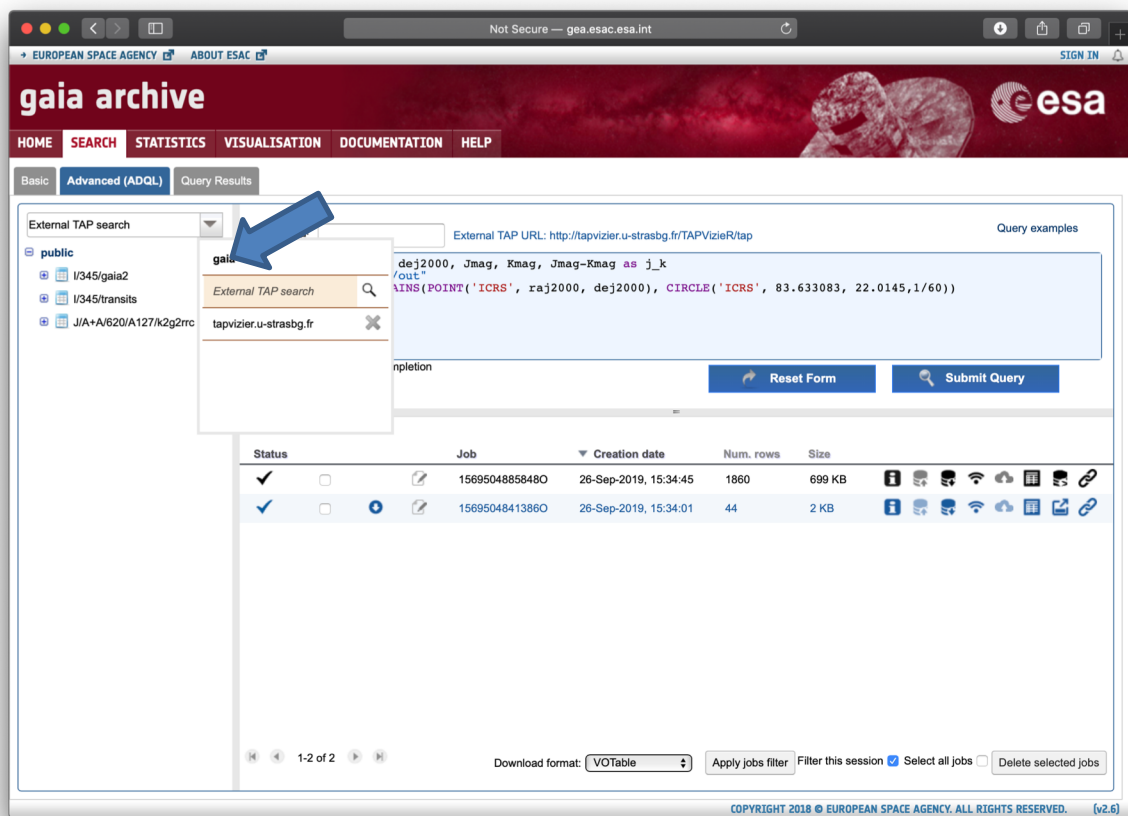
June-September 2019 data

# 98%

Of all the query traffic would not exist without

- Open Source projects contributions
- Open VO protocols compatibility

# External queries in the Gaia archive



The screenshot shows the Gaia archive web interface. The 'External TAP search' dropdown menu is open, showing a list of search options. A blue arrow points to the 'gaia' option. The 'Query examples' section displays a SQL query for a TAP query. The 'Query results' table shows two jobs with their status, creation date, number of rows, and size.

External TAP search:

Query examples

```
dej2000, Jmag, Kmag, Jmag-Kmag as j_k  
/out+  
AINS(POINT('ICRS', raj2000, dej2000), CIRCLE('ICRS', 83.633083, 22.0145, 1/60))
```

Reset Form Submit Query

Status	Job	Creation date	Num. rows	Size
✓	15695048858480	26-Sep-2019, 15:34:45	1860	699 KB
✓	15695048413860	26-Sep-2019, 15:34:01	44	2 KB

Download format: VOTable Apply jobs filter Filter this session Select all jobs Delete selected jobs

COPYRIGHT 2018 © EUROPEAN SPACE AGENCY. ALL RIGHTS RESERVED. (v2.6)

- 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

gaia archive

HOME SEARCH STATISTICS

Basic Advanced (ADQL) Query

External TAP search

public

- I/345/gaia2
- I/345/transits
- J/A+A/620/A127/k2g2rrc

External TAP search

Search keywords:  Submit Query

First 200 results

External TAP: <http://tapvizier.u-strasbg.fr/TAPVizieR/tap>

- ☐ I/345/allwise  
Allwise AGN Gaia DR2 cross-identification (aux\_allwise\_agn\_gdr2\_cross\_id) (Gaia collaboration)
- ☒ I/345/gaia2  
GaiaSource DR2 data (Gaia collaboration)
- ☐ I/345/numtrans  
Calibrated FoV transit photometry from CU5, consolidated and provided by CU7 for variable stars in Gaia DR2 (epoch\_photometry, part 1) (Gaia collaboration)
- ☐ I/345/transits  
Calibrated FoV transit photometry for CU5, consolidated and provided by CU7 for variable stars in Gaia DR2 (epoch\_photometry, part 2) (Gaia collaboration)
- ☐ J/A+A/611/A11/tableb0  
\*Bolometric correction for Johnson-Cousins, 2MASS, SDSS, and Gaia systems (updated with Gaia DR2 bolometric corrections) (Chiavassa A., Casagrande L., Collet R., Magic Z., Bigot L., ...)

Deselect all Add selected

External TAP URL:  Add

Close

Query examples

Submit Query

Select all jobs ☐ Delete selected jobs

ESA

EUROPEAN SPACE AGENCY

ALL RIGHTS RESERVED. (v2.6)

- 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!)



Not Secure — gea.esac.esa.int

EUROPEAN SPACE AGENCY ABOUT ESAC SIGN IN

# gaia archive

HOME SEARCH STATISTICS VISUALISATION DOCUMENTATION HELP

Basic Advanced (ADQL) Query Results

tapvizier.u-strasbg.fr

Job name: External TAP URL: <http://tapvizier.u-strasbg.fr/TAPVizieR/tap> Query examples

```

1 SELECT raj2000, dej2000, Jmag, Kmag, Jmag-Kmag as j_k
2 FROM "II/246/out"
3 WHERE 1=CONTAINS(POINT('ICRS', raj2000, dej2000), CIRCLE('ICRS', 83.633083, 22.0145, 1/60))

```

Ctrl+Space for query autocompletion

Reset Form Submit Query

Status	Job	Creation date	Num. rows	Size
✓	15695048858480	26-Sep-2019, 15:34:45	1860	699 KB
✓	15695048413860	26-Sep-2019, 15:34:01	44	2 KB

1-2 of 2

Download format: VOTable

Apply jobs filter Filter this session Select all jobs Delete selected jobs

COPYRIGHT 2018 © EUROPEAN SPACE AGENCY. ALL RIGHTS RESERVED. (v2.6)

- 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

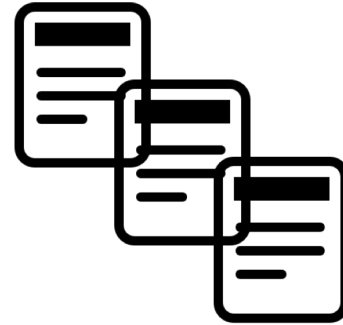
- Many other service implementors rely on open protocols: a quickly increasing toolbox
- IVOA protocols require 2 reference implementations to become REC
  - Mainly Open Source
  - Great efficiency by **using** these implementations and spending project resources in **extending** them
- TAP, ADQL, UWS,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

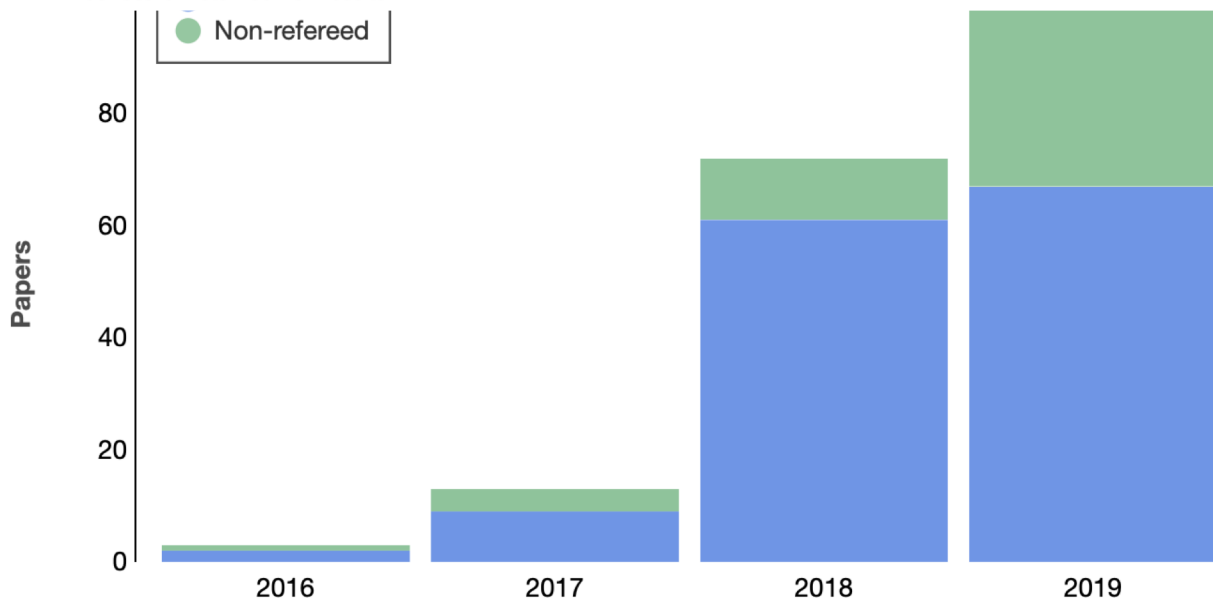
- It is the overall usage of Gaia mission data that matters

# Is it working?

QUICK FIELD: [Author](#) [First Author](#) [Abstract](#) [Year](#) [Fulltext](#) [All Search Terms](#) ▾

full:"gea.esac.esa.int" OR full:"archives.esac.esa.int/gaia" ✕ 🔍

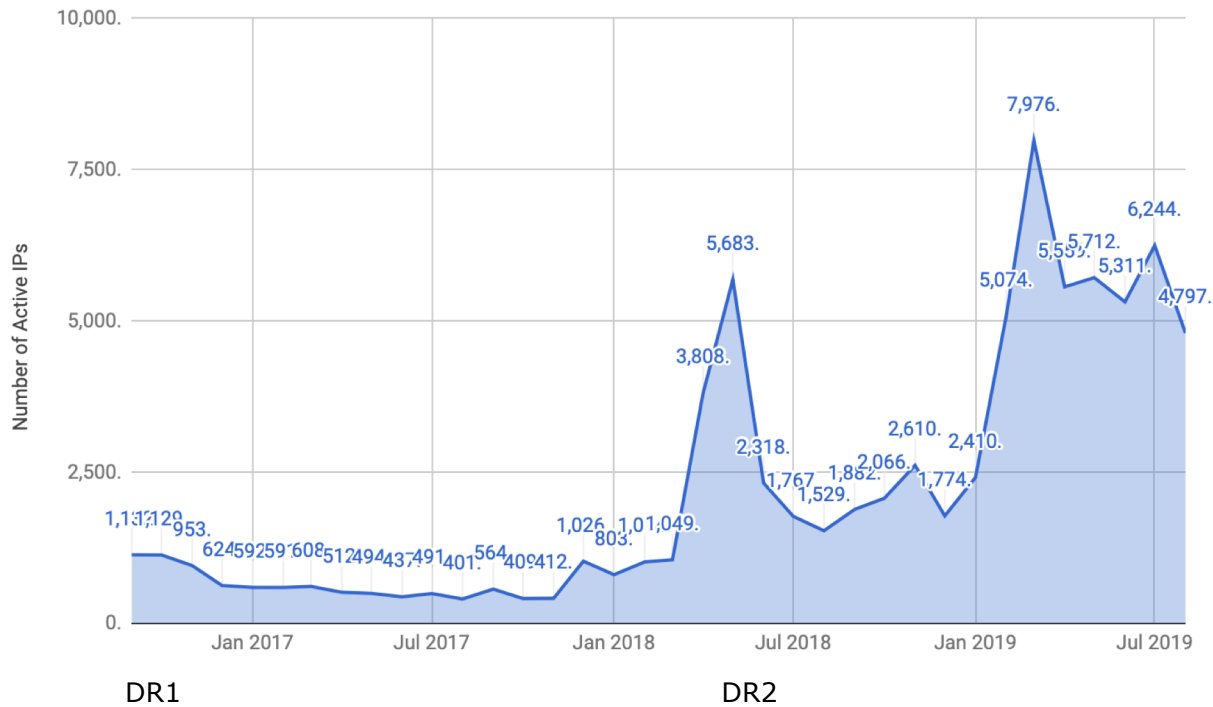
Your search returned **187** results



- 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?

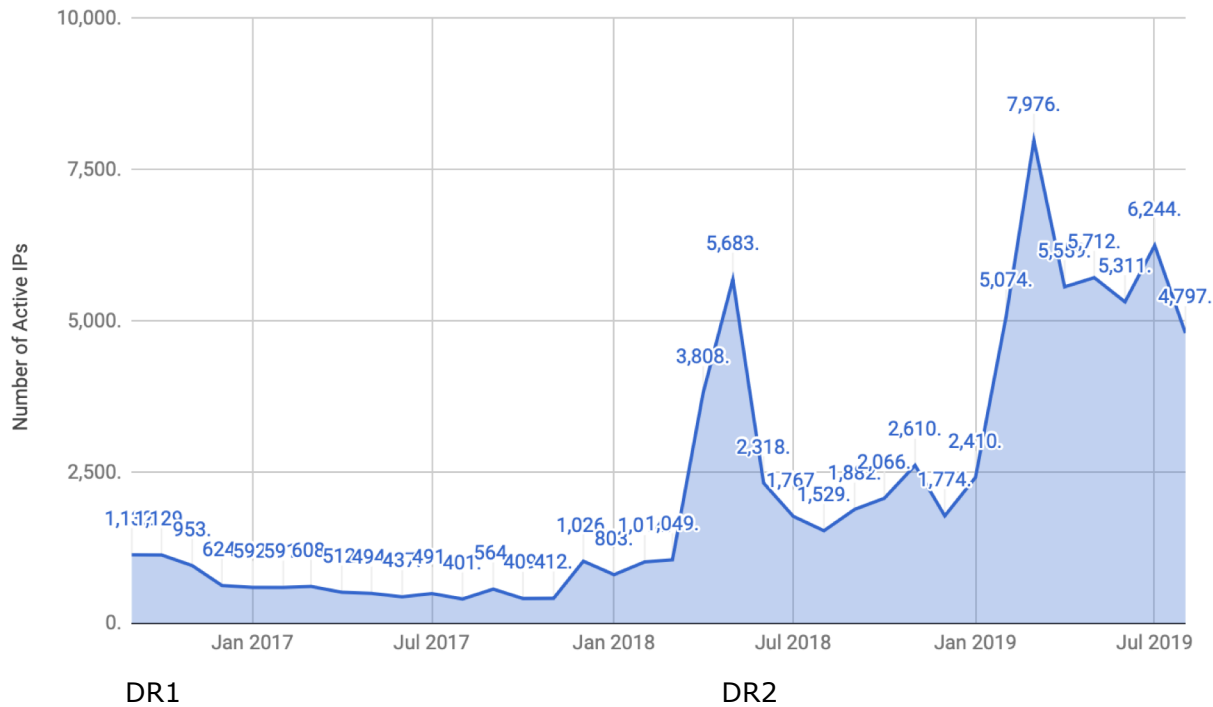
Archive Users (different IP Addresses) downloading data per month



- 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?

Archive Users (different IP Addresses) downloading data per month



- 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?

