Building a master catalogue of unique UV sources

Juan C. Vallejo, Marta L. García, Raúl de la Fuente Marcos, Ana Inés Gómez de Castro
AEGORA research group - Universidad Complutense de Madrid

AEGORA COMPLUTENSE RESEARCH GROUP ON SPACE ASTRONOMY AND DATA MINING

JcUVa

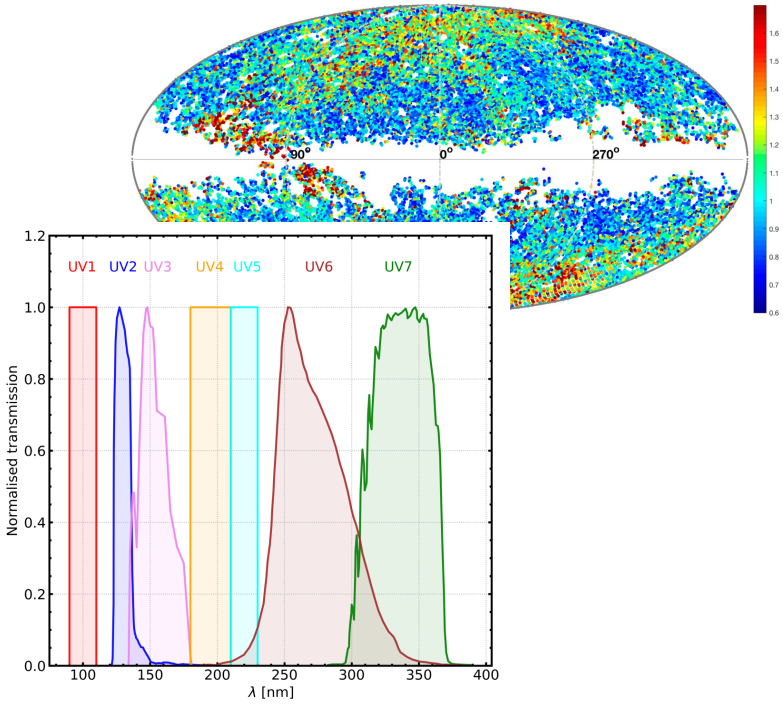
NUVA

Outline

- Introduction.
- The catalogue.
- The infrastructure.
- Testing through use cases and future work.

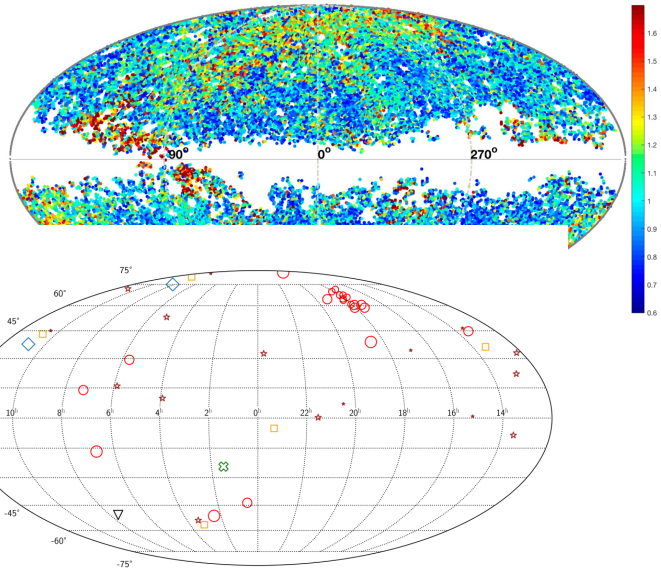
Introduction

- UV data is fundamental in many areas of astrophysical research, but there are some **obstacles**,
- The lack of UV standards of reference.
- Existing UV data is dispersed along a variety of photometric and spectral measurements coming from a variety of space missions.



Introduction

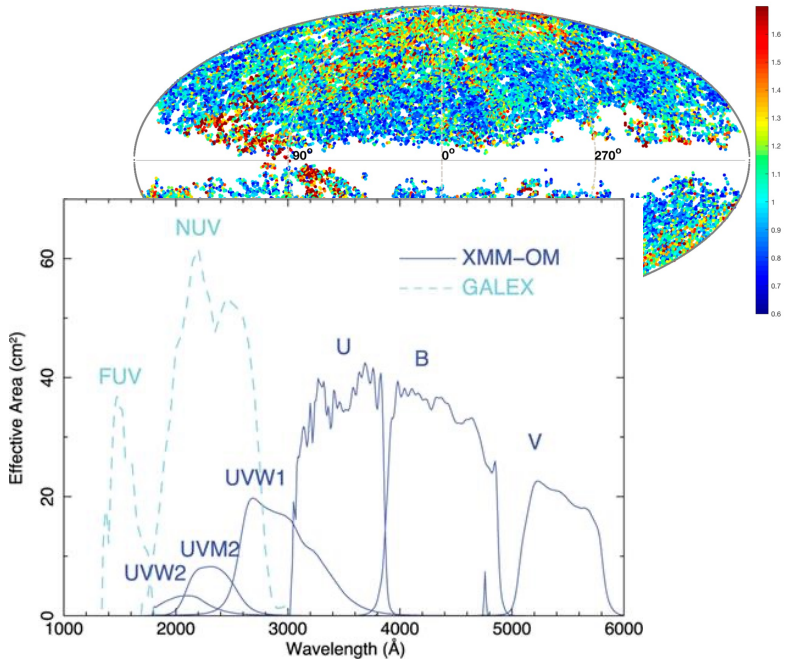
- UV data is fundamental in many areas of astrophysical research, but there are some **obstacles**,
- The lack of UV standards of reference.
- Existing UV data is dispersed along a variety of photometric and spectral measurements coming from a variety of space missions.



The definition of a standard photometric system in UV astronomy was a first step in overcoming these issues.

Introduction

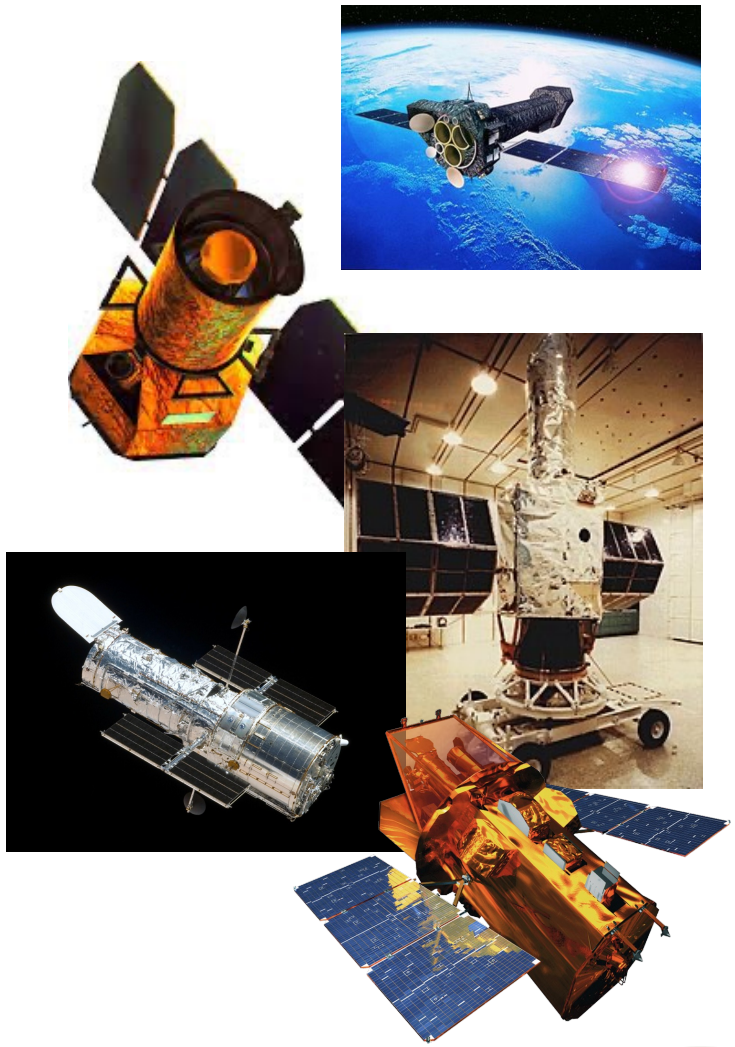
- UV data is fundamental in many areas of astrophysical research, but there are some **obstacles**,
- The lack of UV standards of reference.
- Existing UV data is dispersed along a variety of photometric and spectral measurements coming from a variety of space missions.



Next step…

Variety of sources

- **UV data** heavily rely on space missions. Their archives are the legacy of the past (and present) missions.
- Classic/old catalogues (UV?) were designed as **standalone archives**: isolation.
- Existing facilities (such as cross match from Vizier, MAST discovery portal, standard libraries for accessing modern archives) help to **mitigate this** isolation.



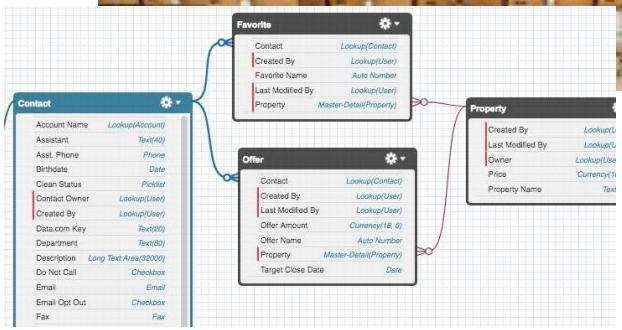
The catalogue

The catalogue

- Hence, AEGORA research group has started the process of creating a master list of **unique UV astronomical sources**.

- This list of unique sources is correlated with an UV-focused approach.

- **Matched metadata allow the extraction of information from parent catalogues.** This will enhance the usefulness of the list and re-enables the science value of old archives/catalogues.



The catalogue (II)

- The catalogue is made up of two main tables,
- **Master**, that contains the list of unique UV sources processed from the other UV catalogues
- **Cross Match**, that collects the information from the master with the rest of the UV catalogues, and it is required as an intermediary to perform massive searches.

- …plus several mission specific tables.

The Master, currently contains 105,864,244 unique UV sources (Id, Ra, Dec, Common Name, Ancillary).
The cross match contains 20,311,662 entries.

The mission tables

CAT_(CATALOGUE_VERSION)

Catalogue	***Data*** (see ***cross_observation***)	***Type*** (see ***cross_observation_type***)
CAT_GALEX_V2	Sources	Photometry and Spectroscopic
CAT_HSC_V0	Sources	Photometry
CAT_HST_V0	Observations	Photometry and Spectroscopic
CAT_HST_V0_D1	Observations	Photometry and Spectroscopic
CAT_IUE_V6	Observations	Spectroscopic
CAT_OM_V2	Sources	Photometry
CAT_SWIFTOBS_V0	Observations	Photometry
CAT_SWIFTSC_V0	Sources	Photometry

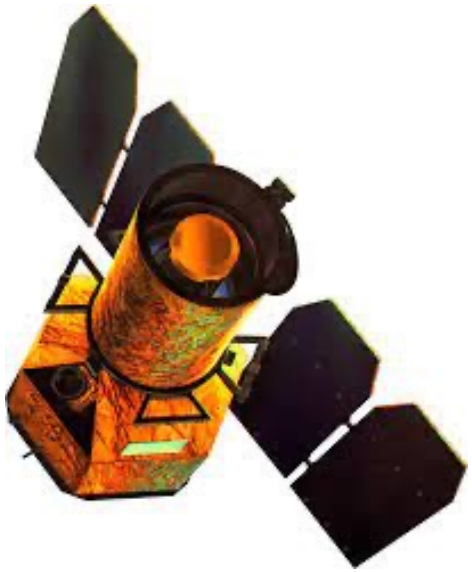
All these catalogues contain:

- all the columns corresponding to their original catalogue with a name of the form ***catalogue_version_originalcolumn***
- two added columns that allows to join those catalogues with the CAT_CROSS_MATCHING table:

- ***catalogue_version_id***: varchar identifier of the row in the catalogue. Corresponds to *cross_external_id*.
- ***catalogue_version_source_id***: Source varchar identifier in the catalogue. Corresponds to *cross_external_source_id*.

The GALEX revised catalogue

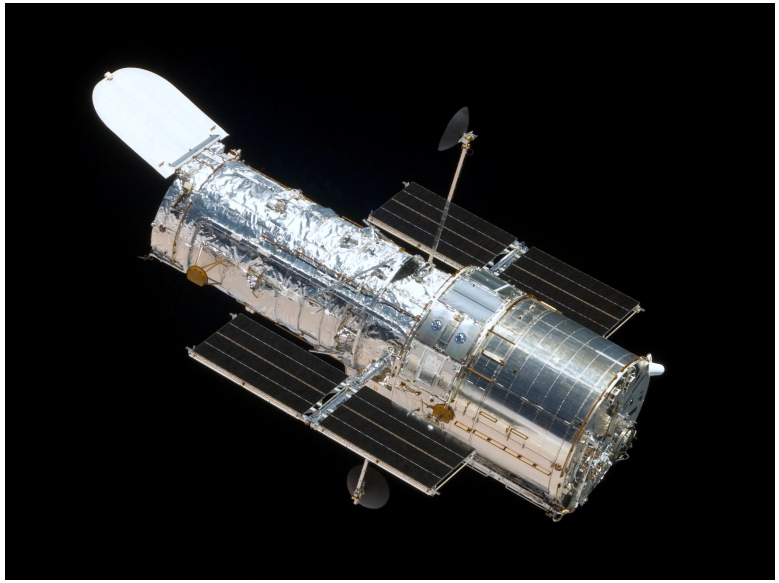
- The catalogue revised by Bianchi et al. (2017) includes **sources from observations** with both FUV and NUV detectors on.
- Within the central 0.55 radius of the field of view to avoid sources with poor photometry and astrometry near the edge of the field, and rim artifacts.
- With NUV magnitude errors ≤ 0.5 mag.
- Unique, i.e., duplicate measurements of the same source are identified and removed.



AEGORA JcUVa NUVA

The HST catalogues

- The **HST catalogue**, catalogue of photometric and spectrometric observations made with UV instrumentation of the Hubble Space Telescope (HST).
- The **HST source catalogue**, Whitmore et al. 2016, designed to optimise science by combining the tens of thousands of visit-based source lists in the Hubble Legacy Archive (HLA) into a single master catalogue.

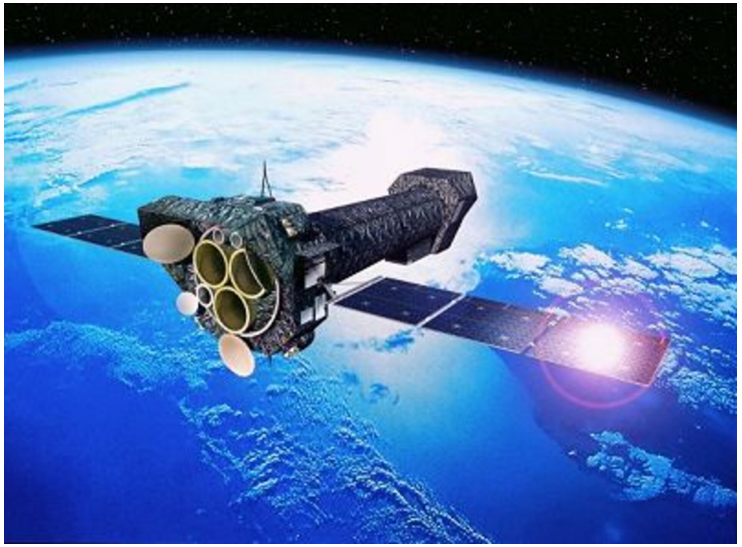


Final merged log of IUE observations

- The Final Merged Log of IUE Observations (NASA-ESA, 2000).
- Summary of all IUE Newly Extracted Spectra (INES) which resulted from a post-processing effort at Vilspa (nowadays ESAC).
- Constructed by using verified data from the IUE Final Archive Master Catalogue.

The XMM-Newton OM serendipitous source survey

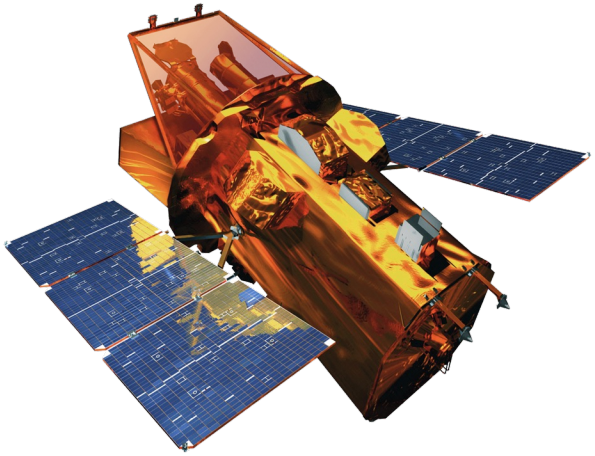
- The XMM-Newton Optical Monitor Serendipitous Source Survey Catalogue (Page et al. 2012)
- Ultraviolet (UV) **sources detected serendipitously by the XMM-OM**.
- UV-detected sources collected from 2417 XMM-OM observations in one to six broad-band UV and optical filters.
- The primary contents of the catalogue are source positions, magnitudes and fluxes in one to six passbands, profile diagnostics, variability statistics.



AEGORA JcUVa NUVA

The Swift catalogues

- The Ultraviolet Optical Telescope (UVOT) **observation catalogue**, list of all targets observed with UVOT onboard the Swift observatory.
- The Swift UVOT **Serendipitous Source Catalogue** (UVOTSSC), Page et al. 2015, produced by processing the data obtained from the Swift, that contains positions, photometry in three UV and three optical bands, morphological information and quality flags.



AEGORA

JcUVa

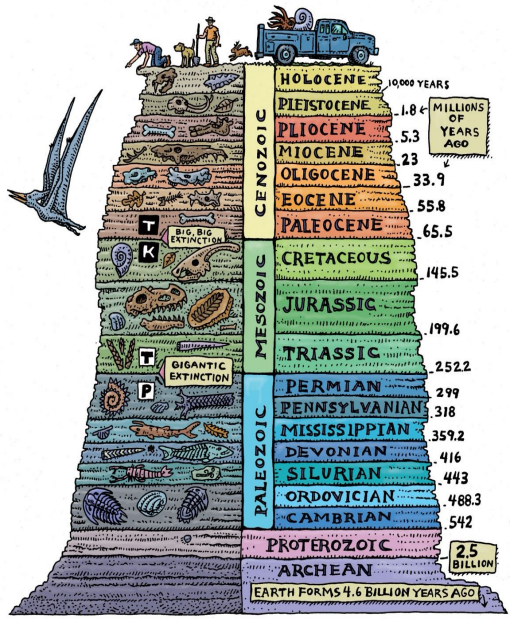
NUVA

Building the prototype

Timescales are important

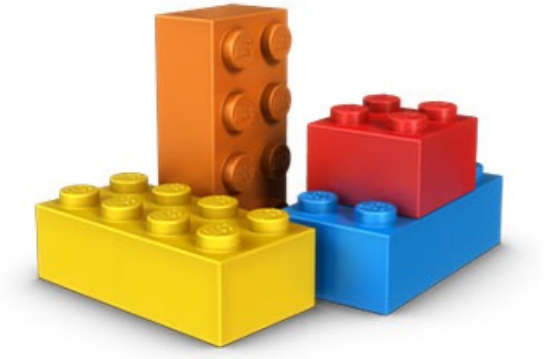
- Archives are “forever”.
- A period of one decade (or even shorter) in s/w and h/w is like a geological era: archives systems must cope with this.

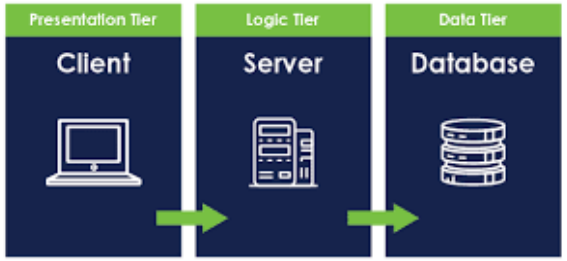
- A pluggable blocks approach allows future enhancements and major changes.
- Allows future paradigms which at the moment are not mature enough.
- Clean/Clear interfaces are a key issue.



Modular approach

- **Modularity** isolates technologies and allows future replacements or migrations.
- Need of tight control of interface layers.
- Reuse of “old” modules possible, reuse of proven tools reduces cost and resources.
- Might increase heterogeneity.





- Well proven approach (old fashioned?) for starting: three tier architecture, which **separates database from application from GUI**.
- We must also tackle mission catalogues evolution, aka **versioning**.

Catalogue versioning

CAT_(CATALOGUE_VERSION)

Catalogue	***Data*** (see ***cross_observation***)	***Type*** (see ***cross_observation_type***)
CAT_GALEX_V2	Sources	Photometry and Spectroscopic
CAT_HSC_V0	Sources	Photometry
CAT_HST_V0	Observations	Photometry and Spectroscopic
CAT_HST_V0_D1	Observations	Photometry and Spectroscopic
CAT_IUE_V6	Observations	Spectroscopic
CAT_OM_V2	Sources	Photometry
CAT_SWIFTOBS_V0	Observations	Photometry
CAT_SWIFTSC_V0	Sources	Photometry

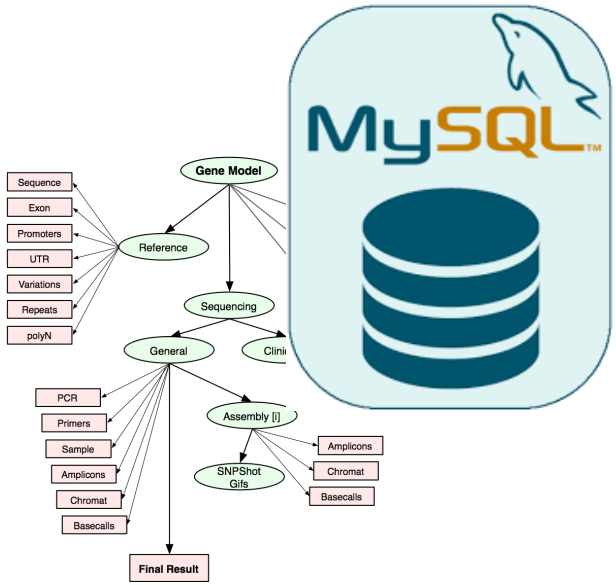
All these catalogues contain:

- all the columns corresponding to their original catalogue with a name of the form ***catalogue_version_originalcolumn***
- two added columns that allows to join those catalogues with the CAT_CROSS_MATCHING table:

- ***catalogue_version_id***: varchar identifier of the row in the catalogue. Corresponds to *cross_external_id*.
- ***catalogue_version_source_id***: Source varchar identifier in the catalogue. Corresponds to *cross_external_source_id*.

Data model, data formats

- **Free storage**: plainfiles, metainformation is encoded in the filename and any hierarchical relationship relies on the file system folder structure.
- **Structured storage**: data is kept in a RDBMS. Each type of cohesive data is placed in a separate table, maintaining explicit relations among tables that contain related data.
- **Mixed storage**: Metadata in RDBMS, with pointers to Data in Files.

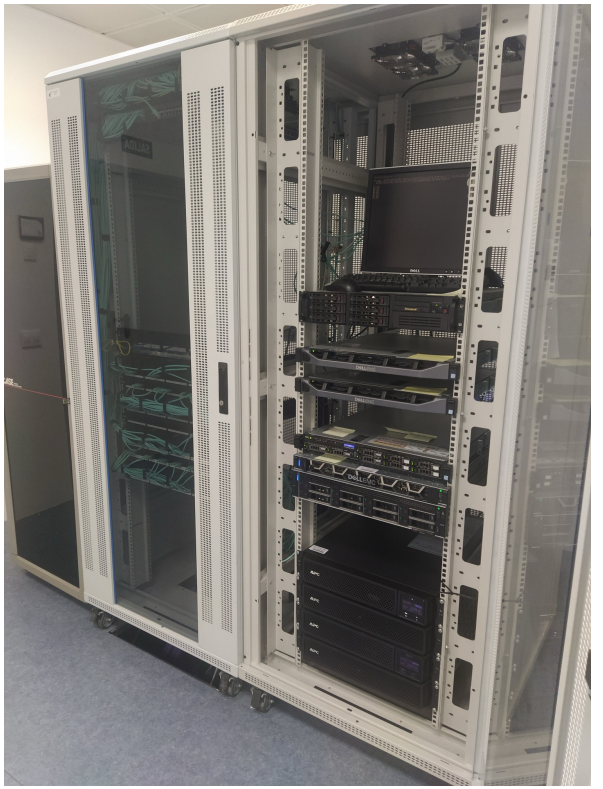


Here, pointers can link to existing external facilities, And use external services.

Data Server

- In its **initial** shape, the catalogue is just a RDBMS in a server plus a web-based app (application+GUI).
  - RDBMS MySQL, Dell PowerEdge R550 (4Tb disk)
  - Reuse of JCUVA Web infrastructure.

- In its **final** shape, the catalogue will be a "storage service“, exposing the operations that let client applications to access heterogeneous resources in an homogeneous manner.



Pre-processing of mission data

- Data selection and preprocessing activities are required **before the actual ingestion**.
- Because of data size, files grouped by region were downloaded and merged into a single file.
- Headers are deleted, ….
- Final csv files corresponding to every catalogue are produced and ingested into the RDBMS engine, by adding two new columns/fields.

The cross-match (I)

- Identifier corresponding to the match.
- Id corresponding to the master source.
- There will be several sources sharing this field, since sources are found in more than one secondary catalogue.
- External id: link to the original secondary catalogue, corresponding to the added field.
- External source id: source identifier in the original secondary catalogue, corresponding to the added field.

The cross-match (II)

- Cone criteria: If the source has a counterpart in master, this field contains the distance at which the match was successful.
- If there was not any matching sources and the object was added to master, the value of this field is -1.

- The matching in all cases was done using a criteria of 2 arcsec[(1)].
- Observation/Source flag: If the source comes from a secondary catalogue of observations or a (point) source catalogue.
- Phot/Spec flag.

[(1)] See Gómez de Castro et al, Astrop.Space Sci, 335, 97,(2011)

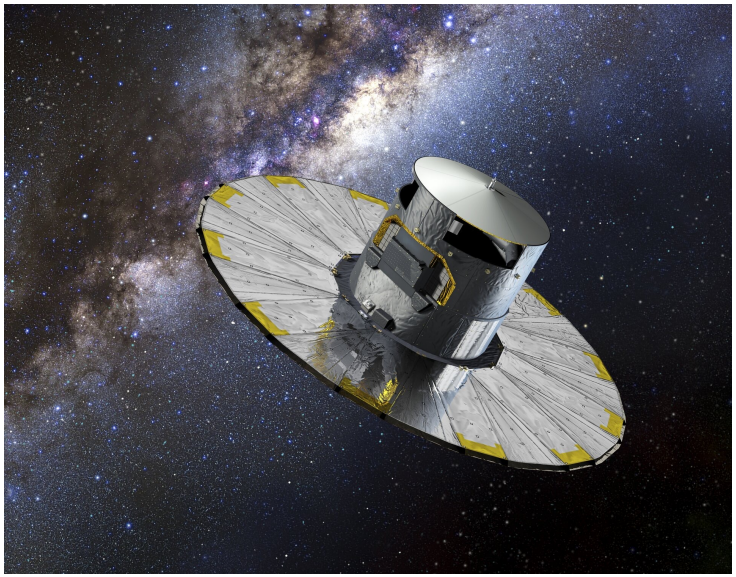
The Master, currently contains 105,864,244 unique UV sources.
The cross match contains 20,311,662 entries.

Preliminar use cases, Next steps

First use case: GAIA cross match

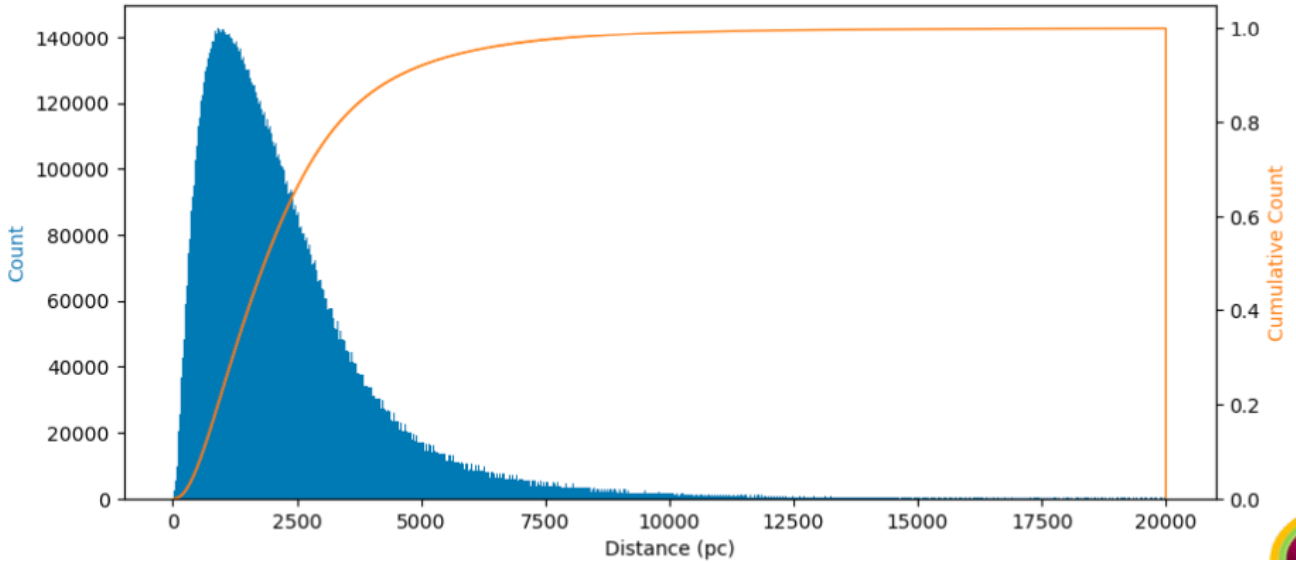
- **Science** Use Case: to obtain three-dimensional coordinates of the UV sources and make a 3D map of the sky.

- The JCUVA UV Catalogue was cross-matched with the Gaia DR3 catalogue.



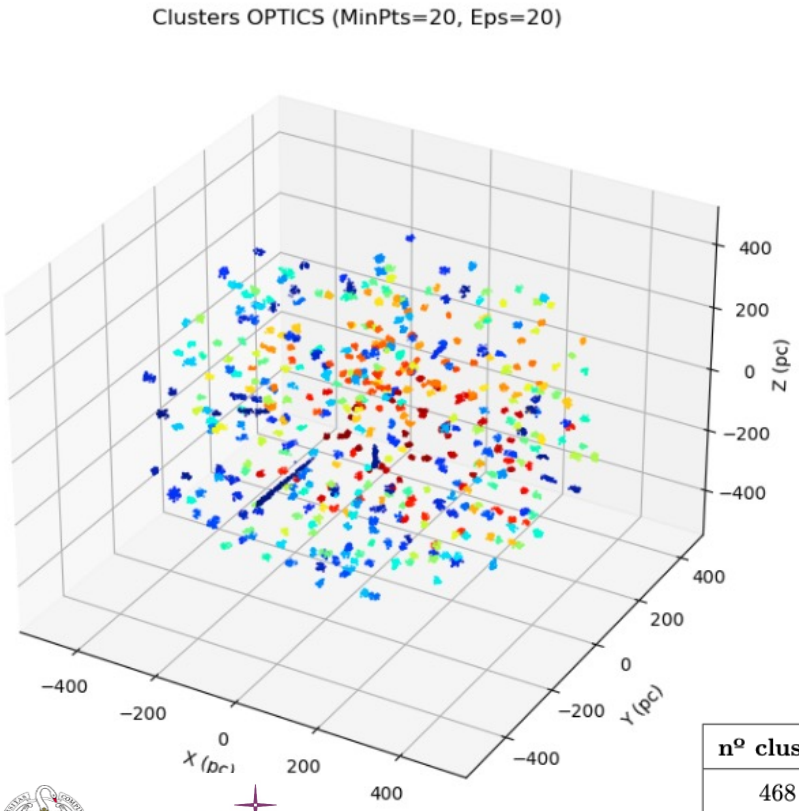
Performance test bed

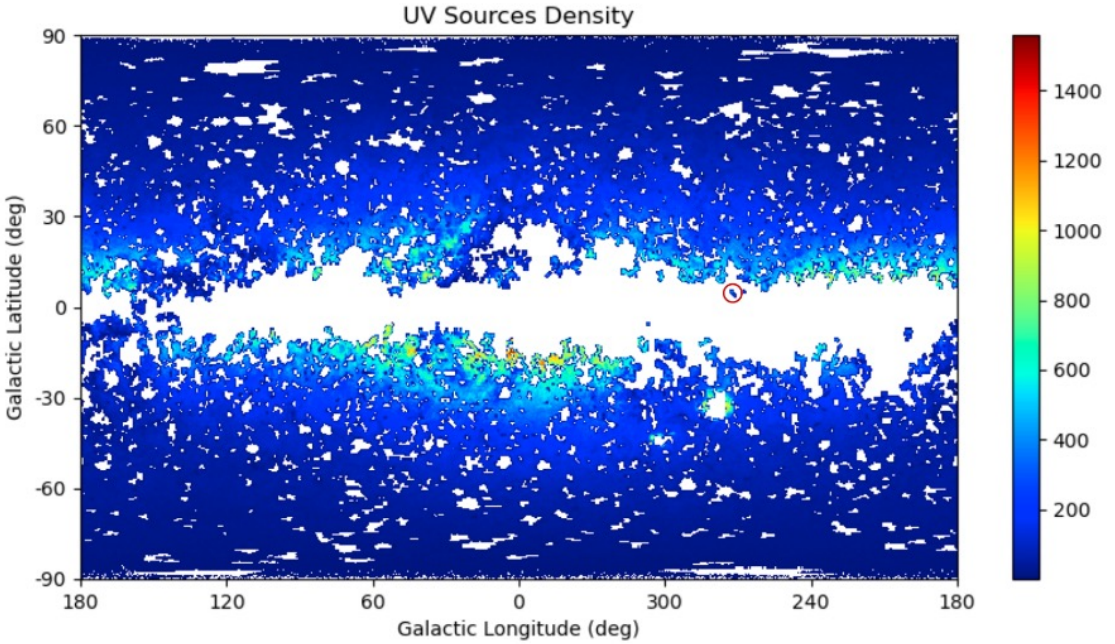
- A good testing of performance in addition to scientific uses.
- A good testing for 2arc sec criteria, additional catalogues needed for GAIA.
- 25,8 millions UV entries. Once proper ruwe index and positive parallaxes are taken into account, 22,45 millions entries.
- Then one takes into account positive G and NUV, plus ”good” values of FUV (blanking flag from GALEX): 1,1 millions.



Science: clustering analyses

- OPTICS algorithm.

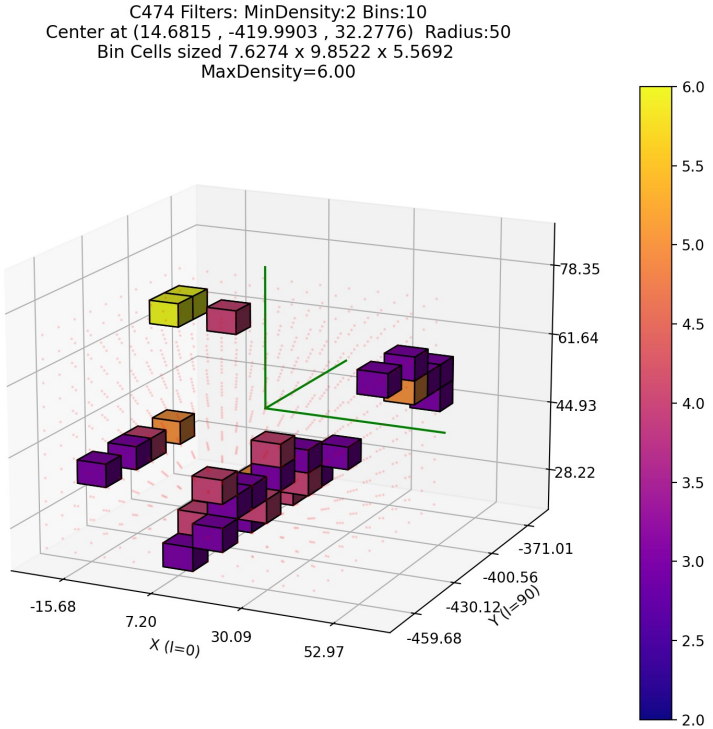


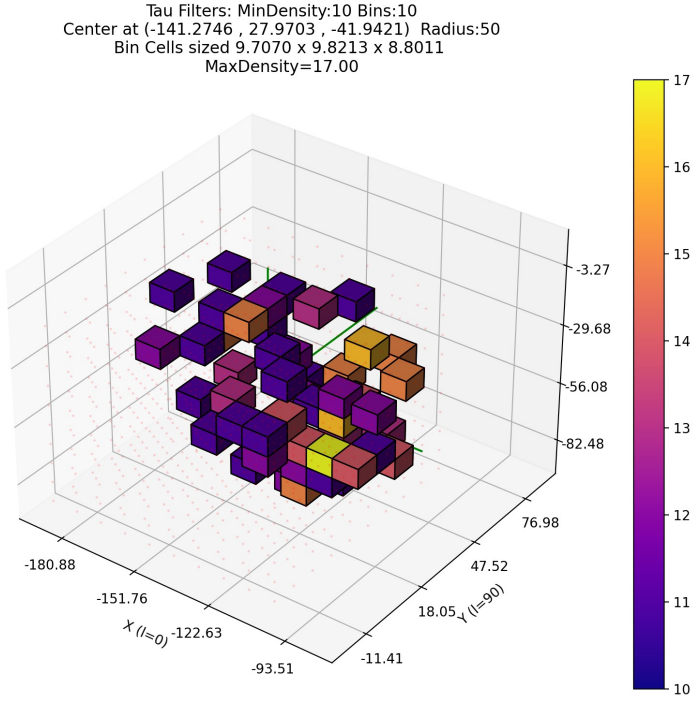


nº cluster	nº sources	X	Y	Z	l	b	fuv	nuv	g
468	148	14,68	-419,99	32,28	272	4,39	19,16	18,42	13,52

Science: density analyses

- Density algorithm.





Future work

- Future steps will aim to develop a user interface.
- Generating a database of UV photometry in the IAU system.
- Multi-epoch variability studies

Thank you for your attention…

- Work in progress, ideas/comments/… are appreciated.
- Questions…?