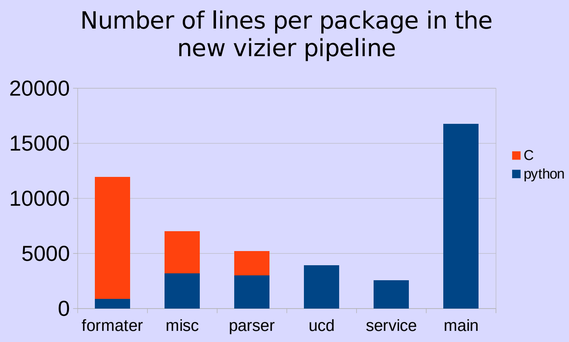
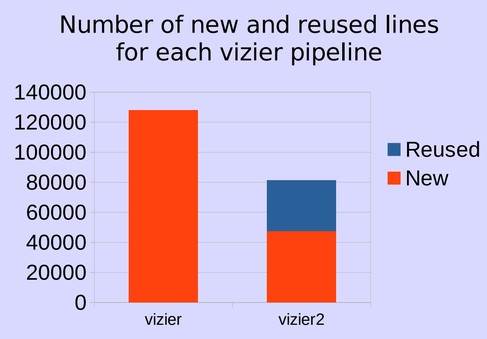
Re-writing the VizieR catalogue ingestion pipeline

VizieR is a curated library of astronomical catalogues that provides enriched, verified data. Users have multiple ways of accessing the data, be it through the html pages, TAP queries, or interoperating with Topcat or astropy. On the internal side, Vizier works with a PostgreSQL database to store the data, as well as multiple files that describe and enrich catalogues, adding descriptions or links to other resources for instance. The internal workflow by which catalogues are ingested into VizieR is a semi-automated workflow : the description of catalogues and their enrichment with metadata are done by the documentalists at CDS, with the help of tools that streamline the process (for instance, the UCD builder, that gives proposition of UCDs for columns in the catalogue, that documentalists can verify and edit). The first part of the workflow builds the metadata around the catalogue, and relies heavily on the standardised description of the catalogue, while the second part, populating the database, needs no human intervention.

Program length comparison



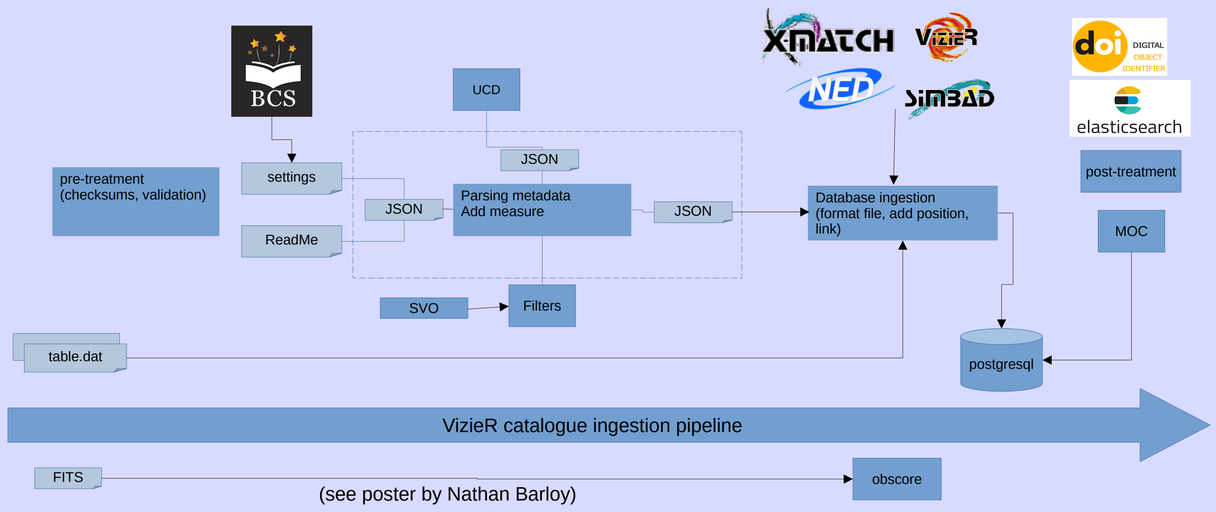


The original ingestion main program is written in C and is 17000 lines long. As such, maintaining the program and adding new features gets harder and harder. That is why we re-wrote the program : compartmentalising different parts of the workflow in different packages, which allowed us to reduce the size of the main program, making, in turn, the reading and understanding of the architecture easier for future developments. The new pipeline is in no way a simple transformation from C to python.

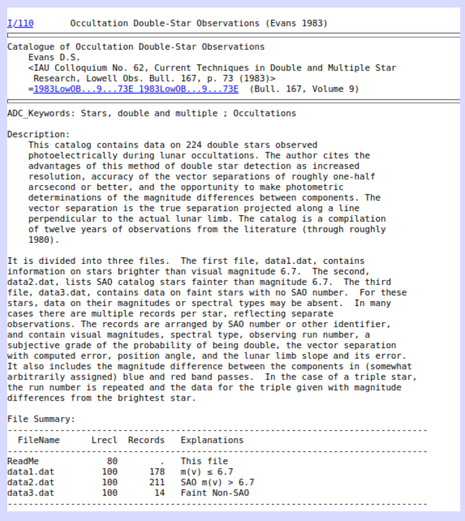
Original pipeline

New pipeline

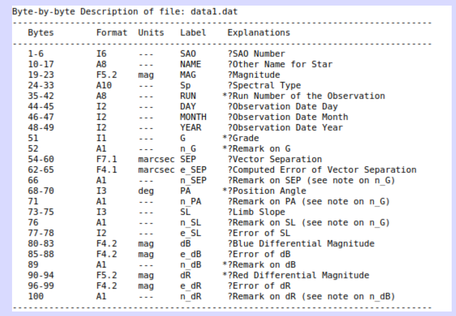
Vizier catalogue ingestion pipeline



The new ingestion pipeline is written in C and Python, and reuses most of the ideas behind the original pipeline. The new pipeline is subdivided in different packages for different tools. During the building of the metadata, the standardised ReadMe file is read and combined with a settings file. They are transformed into a json file containing all metadata relevant to the catalogue. This file is then used to transform the data files : adding links to other data, adding columns such as position columns, merging files... Afterwards, data are ingested in the database.



```
I/110       Occultation Double-Star Observations (Evans 1983)

Catalogue of Occultation Double-Star Observations
    Evans D.S.
    <IAU Colloquium No. 62, Current Techniques in Double and Multiple Star
    Research, Lowell Obs. Bull. 167, p. 73 (1983)>
    =1983LowOB...9...73E 1983LowOB...9....73E  (Bull. 167, Volume 9)

ADC_Keywords: Stars, double and multiple ; Occultations

Description:
    This catalog contains data on 224 double stars observed
    photoelectrically during lunar occultations. The author cites the
    advantages of this method of double star detection as increased
    resolution, accuracy of the vector separations of roughly one-half
    arcsecond or better, and the opportunity to make photometric
    determinations of the magnitude differences between components. The
    vector separation is the true separation projected along a line
    perpendicular to the actual lunar limb. The catalog is a compilation
    of twelve years of observations from the literature (through roughly
    1980).

It is divided into three files.  The first file, data1.dat, contains
information on stars brighter than visual magnitude 6.7.  The second,
data2.dat, lists SAO catalog stars fainter than magnitude 6.7.  The third
file, data3.dat, contains data on faint stars with no SAO number.  For these
stars, data on their magnitudes or spectral types may be absent.  In many
cases there are multiple records per star, reflecting separate
observations. The records are arranged by SAO number or other identifier,
and contain visual magnitudes, spectral type, observing run number, a
subjective grade of the probability of being double, the vector separation
with computed error, position angle, and the lunar limb slope and its error.
It also includes the magnitude difference between the components in (somewhat
arbitrarily assigned) blue and red band passes.  In the case of a triple star,
the run number is repeated and the data for the triple given with magnitude
differences from the brightest star.

File Summary:
--------------------------------------------------------------------------------
  FileName      Lrecl  Records   Explanations
--------------------------------------------------------------------------------
ReadMe            80        .   This file
data1.dat        100      178   m(v) ≤ 6.7
data2.dat        100      211   SAO m(v) > 6.7
data3.dat        100       14   Faint Non-SAO
--------------------------------------------------------------------------------
```

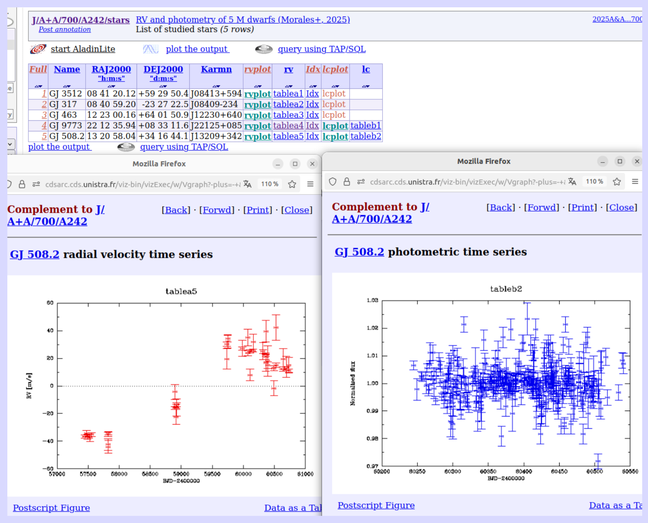


```
Byte-by-byte Description of file: data1.dat
--------------------------------------------------------------------------------
  Bytes      Format  Units   Label     Explanations
--------------------------------------------------------------------------------
  1-6          I6    ---     SAO       ?SAO Number
 10-17         A8    ---     NAME      ?Other Name for Star
 19-23         F5.2  mag     MAG       ?Magnitude
 24-33         A10   ---     Sp        ?Spectral Type
 35-42         A8    ---     RUN     *?Run Number of the Observation
 44-45         I2    ---     DAY       ?Observation Date Day
 46-47         I2    ---     MONTH     ?Observation Date Month
 48-49         I2    ---     YEAR      ?Observation Date Year
 51            I1    ---     G       *?Grade
 52            A1    ---     n_G     *?Remark on G
 54-60         F7.1  marcsec SEP       ?Vector Separation
 62-65         F4.1  marcsec e_SEP     ?Computed Error of Vector Separation
 66            A1    ---     n_SEP     ?Remark on SEP (see note on n_G)
 68-70         I3    deg     PA      *?Position Angle
 71            A1    ---     n_PA      ?Remark on PA (see note on n_G)
 73-75         I3    ---     SL        ?Limb Slope
 76            A1    ---     n_SL      ?Remark on SL (see note on n_G)
 77-78         I2    ---     e_SL      ?Error of SL
 80-83         F4.2  mag     dB        ?Blue Differential Magnitude
 85-88         F4.2  mag     e_dB      ?Error of dB
 89            A1    ---     n_dB    *?Remark on dB
 90-94         F5.2  mag     dR      *?Red Differential Magnitude
 96-99         F4.2  mag     e_dR      ?Error of dR
 100           A1    ---     n_dR      ?Remark on dR (see note on n_dB)
--------------------------------------------------------------------------------
```

Example of a Readme file (see 10.1051/aas:2000169)

New feature

Using the new basis for the handling of the metadata in the VizieR pipeline, we are able to integrate new standards in the service. A long-standing issue for VizeR has been the accessibilty of associated data (images, spectra...) of catalogues. Indeed, apart from the visual html interface, it is impossible to access linked resources. To address this problem, we are implementing the Datalink standard (https://www.ivoa.net/documents/DataLink/20231215/REC-DataLink-1.1.html). We can add a column to the result of the user query, containing links to datalink instances, which themselves link back to the associated resources, with a short description for each.



Time series for object GJ 508.2 of catalogue J/A+A/700/A242

```
<VOTABLE version="1.4">
 <RESOURCE type="results">
  <TABLE>
   <FIELD name="ID" datatype="char" arraysize="*" ucd="meta.id;meta.main"/>
   <FIELD name="access_url" datatype="char" arraysize="*" ucd="meta.ref.url"/>
   <FIELD name="service_def" datatype="char" arraysize="*" ucd="meta.ref"/>
   <FIELD name="error_message" datatype="char" arraysize="*" ucd="meta.code.error"/>
   <FIELD name="semantics" datatype="char" arraysize="*" ucd="meta.code"/>
   <FIELD name="description" datatype="char" arraysize="*" ucd="meta.note"/>
   <FIELD name="content_type" datatype="char" arraysize="*" ucd="meta.code.mime"/>
   <FIELD name="content_qualifier" datatype="char" arraysize="*" ucd="meta.code.mime"/>
  <DATA>
   <TABLEDATA>
    <TR>
     <TD>ivo://CDS.VizieR/J/A+A/700/A242?stars___recno=4</TD>
     <TD>
      http://cdsarc.cds.unistra.fr/viz-bin/vizExec/w/Vgraph?-plus=-+\&J/A%2BA/700/A242/tablea4
     </TD>
     <TD/>
     <TD/>
     <TD>auxiliary</TD>
     <TD>
      Show the Radial Velocity curves (timeserie) from J/A+A/700/A242
     </TD>
     <TD>text/html</TD>
     <TD>timeserie</TD>
    </TR>
    <TR>
     <TD>ivo://CDS.VizieR/J/A+A/700/A242?stars___recno=4</TD>
     <TD>
      http://cdsarc.cds.unistra.fr/viz-bin/vizExec/w/Vgraph?-plus=-+\&J/A%2BA/700/A242/tableb1
     </TD>
     <TD/>
     <TD/>
     <TD>auxiliary</TD>
     <TD>
      Show the light curves (timeserie) from J/A+A/700/A242
     </TD>
     <TD>text/html</TD>
     <TD>timeserie</TD>
    </TR>
   </TABLEDATA>
  </DATA>
 </TABLE>
</RESOURCE>
</VOTABLE>
```

Datalink for time series for object GJ 508.2 of catalogue J/A+A/700/A242

Retrocompatibility

A large part of the work behind writing the new VizieR pipeline has been making sure that it is backwards compatible with the ~30 thousand catalogues saved in our database. Although it mainly means abiding by pre-exisiting rules, it also means scouring the database to find exceptions, outliers, and even unspoken rules.

Links

The VizieR website

The VizieR database of astronomical catalogues article

Lastest Datalink documentation

Ivan Brossard
i.brossard@astro.unistra.fr
With the help of Gilles Landais
gilles.landais@astro.unistra.fr

ADASS 2025 – Görlitz
November 9–13 2025 - CDS – Observatoire astronomique de Strasbourg