Astronomical Data Analysis Software and Systems XXVIII
ASP Conference Series, Vol. 523
P.J. Teuben, M.W. Pound, B.A. Thomas, and E.M. Warner, eds.

TOPCAT and Gaia

M. B. Taylor

H. H. Wills Physics Laboratory, Tyndall Avenue, University of Bristol, UK;
m.b.taylor@bristol.ac.uk

Abstract. TOPCAT, and its command line counterpart STILTS, are powerful tools for working with large source catalogs. ESA's *Gaia* mission, most recently with its second data release, is producing source catalogs of unprecedented quality for more than a billion sources. This paper presents some examples of how TOPCAT and STILTS can be used for analysis of *Gaia* data.

1. Introduction

TOPCAT (Taylor 2005) and STILTS (Taylor 2006) are respectively GUI and command-line analysis packages for working with tabular data in astronomy, and as such offer many facilities for manipulation of data such as source catalogs. The recent second data release from ESA's *Gaia* mission (Gaia Collaboration et al. 2018) has produced a source catalog of exceptional quality, and TOPCAT/STILTS are well-placed to provide analysis capabilities for exploitation of this data set. This paper discusses some of the features of the software most relevant for working with the *Gaia* DR2 catalog. Some have been added specifically with *Gaia* data in mind, but in most cases they are general purpose capabilities that are also suitable for use with other datasets.

2. Data Access

The primary access to the *Gaia* catalog is via Virtual Observatory (VO) protocols, provided from the main archive service at ESAC and a number of other data centers.

The most capable of these access protocols is TAP, the Table Access Protocol, which allows execution in the archive database of user-supplied SQL-like queries and retrieval of the resulting tables. TAP, while allowing extremely powerful queries to be performed, is a complex protocol stack which presents some challenges for the client software and user alike. TOPCAT provides the user with a GUI client for interacting with TAP services that integrates functions such as metadata browsing, query validation, table upload and query submission to make the use of TAP as straightforward as possible for the user, without obscuring the flexibility it offers (Taylor 2017). For simpler queries, a Cone Search client is also provided for retrieving source lists based on sky position alone. The *Gaia* catalog additionally contains non-scalar data for some rows, exposed using the VO DataLink protocol. At DR2 this array data is limited to epoch photometry of a relatively small number of known variable sources, but much more, including spectrophotometry, will be provided in future data releases. TOP-

CAT's *Activation Action* toolbox has been overhauled in recent versions to work with this array data.

Since these services follow the VO standards, no *Gaia*-specific code is required or implemented in TOPCAT to provide data access. This means that the same clients can be used for working with copies of the *Gaia* catalog in the main archive and elsewhere, as well as with other VO-compliant services. This standardization has benefits for both the implementer and user of the software.

The only truly *Gaia*-specific code in TOPCAT is a reader for the GBIN format used internally by the analysis consortium. This is a specialized capability of no interest to the general astronomy user, but it has proved valuable for DPAC members working with the data prior to catalog publication.

3. Expression Language

TOPCAT provides a powerful language for evaluating algebraic expressions to define new columns, row selections and plot coordinates. As well as standard arithmetic, trigonometric and astronomical operations, the library contains a number of astrometric functions:

- Propagation of astrometric parameters to earlier/later epoch, with or without errors and correlations
- Conversion of positions and velocities from astrometric parameters to Cartesian coordinates in ICRS, galactic or ecliptic coordinates
- Bayesian estimation of distances and uncertainties from parallax, using the expressions from Astraatmadja & Bailer-Jones (2016)

These are not exactly specific to *Gaia*, but they have been added as they are likely to be often needed when working with *Gaia* data, and they are specified and documented in a way that makes them easy to use in that context. For instance the following expression calculates the (U, V, W) components of velocity in the Galactic coordinate system (without adjusting for local standard of rest, and assuming that parallax error is low):

```
icrsToGal(astromUVW(ra, dec, pmra, pmdec,
radial_velocity, 1000./parallax, false))
```

The variable names here are `gaia_source` catalog column names, and the units are as supplied in the catalog.

4. Scalability

Gaia DR2 contains 1.7 billion sources, and investigating this dataset often requires working with large tables. TOPCAT is well-suited for interactive analysis, including flexible exploratory visualization, of tables (for instance selections from the full catalog) with the order of 10^6–10^7 rows. This regime is quite usable on modest hardware with no special data preparation, for instance data downloaded from external services, or loaded from local FITS or even CSV files. TOPCAT can be used with tables larger than this, but interactive performance may be poor or memory exhausted.

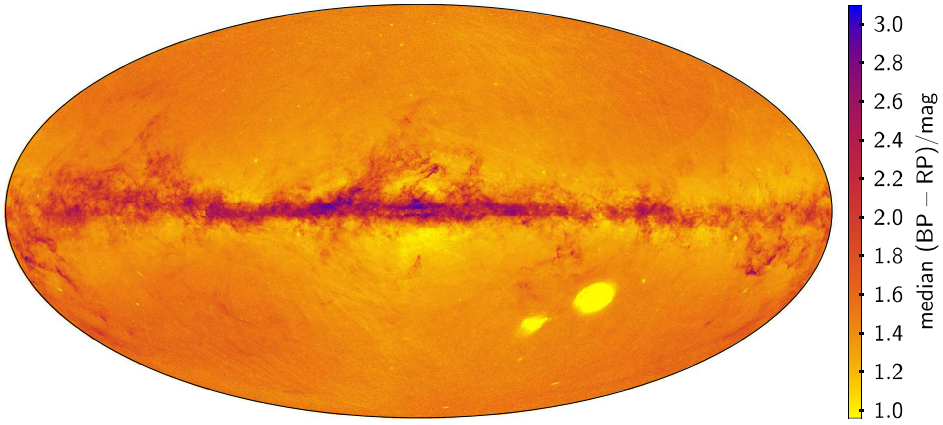

Figure 1. All-sky plot of $BP - RP$ color for 1.4 billion *Gaia* sources. The colors represent the median within each level 8 Halifax pixel.

STILTS on the other hand processes data for the most part in streaming mode, so can cope with arbitrarily large tables in fixed memory. This means that non-interactive calculations or preparation of graphics for the entire *Gaia* catalog is quite feasible. A set of all-sky weighted density maps using all 1.7 billion rows was prepared as follows:

1. download 61 000 (0.5 Tb) gzipped CSV files from ESAC (`wget`, 10 hours)
2. convert to 61 000 small FITS files (STILTS `tpipe`, 5 days)
3. convert to single 0.8 Tb column-oriented FITS file (STILTS `tpipe`, 12 hours)
4. aggregate into level-9 HEALPix map (STILTS `tskymap`, 45 minutes)
5. render plot to PDF or PNG (STILTS `plot2sky`, a few seconds)

In practice, steps 4 and 5 were iterated using TOPCAT visualization interactively on smaller datasets to archive the best results. An example is shown in Figure 1.

Note, this is not necessarily the best way to prepare such maps; executing the calculations near the data is in general more efficient (Taylor et al. 2016).

5. Visualization

TOPCAT has many visualization modes enabling highly configurable interactive exploration of high-dimensional data, and suitable for both large and small data sets. Special attention is given to providing comprehensible representations of large data sets — a simple scatter plot is not useful when there are many more points than pixels. Two plot types have been specifically introduced or enhanced for *Gaia* data: the *Sky Vector* plot displays proper motion vectors on the sky, and the *Sky/XY Correlation* plots show error ellipses based on the astrometric error and correlation quantities provided in the *Gaia* catalog; see Figure 2. The many non-*Gaia*-specific visualization options, too numerous to describe here, are however in most cases the core of TOPCAT's analysis capabilities for working with *Gaia* and non-*Gaia* data alike.

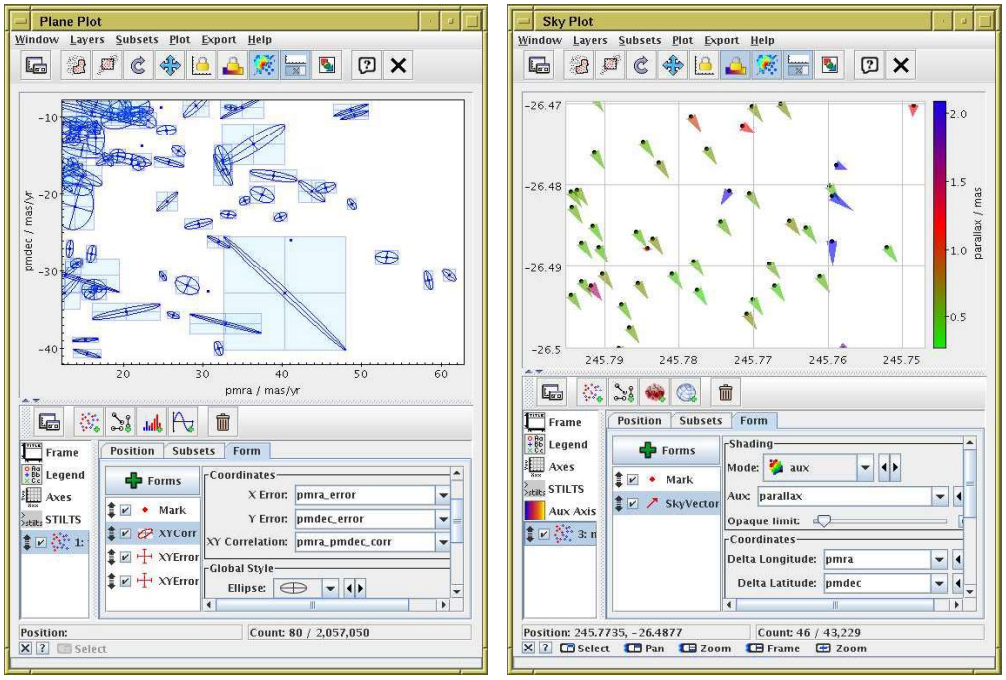

Figure 2. Interactive visualization of high-dimensional data in TOPCAT. The left hand figure displays proper motions with error ellipses derived from the Gaia `pmra_pmdec_corr` column, which show much more information than the simple `pmra_error`/`pmdec_error` error boxes. The right hand figure shows proper motion vectors by shape, and parallaxes by color. In each case, five dimensions are visualized.

Acknowledgments. This work has been primarily funded by the UK's Science and Technology Facilities Council. It has made use of data from the European Space Agency (ESA) mission *Gaia* (`https://www.cosmos.esa.int/gaia`), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, `https://www.cosmos.esa.int/web/gaia/dpac/consortium`). Special thanks to the EU Horizon 2020 project ASTERICS for funding presentation of this work at ADASS 2018.

References

Astraatmadja, T. L., & Bailer-Jones, C. A. L. 2016, ApJ, 833, 119. `1609.07369`

Gaia Collaboration, Brown, A. G. A., Vallenari, A., Prusti, T., de Bruijne, J. H. J., & al. 2018, A&A, 616, A1. `1804.09365`

Taylor, M. B. 2005, in ADASS XIV, edited by P. Shopbell, M. Britton, & R. Ebert, vol. 347 of ASP Conf. Ser., 29

— 2006, in ADASS XV, edited by C. Gabriel, C. Arviset, D. Ponz, & S. Enrique, vol. 351 of ASP Conf. Ser., 666

— 2017, in ADASS XXV, edited by N. P. F. Lorente, K. Shortridge, & R. Wayth, vol. 512 of ASP Conf. Ser., 589

Taylor, M. B., Mantelet, G., & Demleitner, M. 2016, in ADASS XXVI, ASP Conf. Ser. `1611.09190`