Communications of BAO, Vol. 67, Issue 2, 2020, pp. 323-328

Resurrecting DFBS into the Virtual Observatory

Markus Demleitner*[1], Aram Knyazyan[2], Daniel Baghdasaryan[2], Gor Mikayelyan[2], and Areg Mickaelian[2]

[1]Universität Heidelberg, Zentrum für Astronomie, Mönchhofstr. 12-14, 69221 Heidelberg, Germany
[2]Byurakan Astrophysical Observatory, Byurakan, Armenia

Abstract

The Digitised First Byurakan Survey has digitised and processed about 1900 photographic plates from the objective prism surveys conducted in Byurakan by Benjamin Markarian and collaborators in the 1960ies and 1970ies. After digitization, a custom web service was built and operated first in Rome, then in Trieste. However, as astronomical data systems and standards evolved, it became desirable to update the data and build standards-compliant, Virtual Observatory services from it.

This contribution reports on the challenges encountered during this migration, the solutions we chose, and the lessons to be learned. It also discusses the use of the resulting services.

Keywords: *Virtual Observatories, Spectroscopy, Surveys, Astronomy Databases*

1. Introduction

The first Byurakan survey was an objective-prism survey of the sky north of −15° and farther than 15° from the Galactic plane, performed between 1965 and 1980 by Benjamin Markarian and his team. This important resource was digitised in the early 2000s in the DFBS project (Mickaelian et al., 2007), which yielded a set of scanned plates and low-resolution spectra for about 20 million objects extracted from them. The spatial distribution of these data products is shown in Fig. 1.

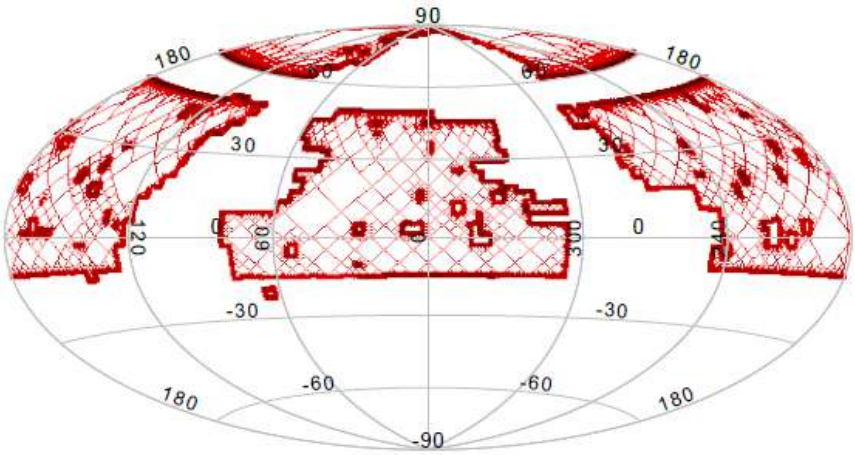

Figure 1. Spatial coverage of the scanned plates from the first Byurakan Survey; this figure was generated by plotting the result of `SELECT SUM(MOC(coverage)) FROM dfbs.plates` on the GAVO TAP service though a TOPCAT sky plot area control. The conspicuous gap around 50 degrees of declination is because the plates in that zone were used with the second Byurakan survey.

*msdemlei@ari.uni-heidelberg.de, Corresponding author

DOI: 10.52526/25792776-2020.67.2-323

When the programme reported on here was started in 2017, the results of this effort were available on an aging, unmaintained machine in Rome, accessible only through a browser interface, with spectra given in text form. It was clear that a migration of this valuable data towards standard discovery and usage patterns was highly desirable.

The reminder of this paper discusses what challenges had to be overcome in order to perform this migration in sect. 2 and goes on to describe the services now available through Virtual Observatory protocols in sect. 3. It concludes with some proposals on how the utility of the data could further be improved and a set of lessons that can be derived from our experience.

2. Migration Challenges

The first problem to be overcome was that the server that published the results of the digitisation project was running without maintenance, and the persons who had set it up had left the hosting institution. In particular, none of the project partners had any access to the machine's mass storage any more.

Hence, we started the migration with data obtained through web crawling; this concerned both the plate images and the spectra, which amounted to about 20 million little text files.

Expectably, the result was highly incomplete data, riddled with truncated files when the connections were interrupted during the transfer. Still, it was enough to set up a prototype service by writing two resource descriptors for the DaCHS publication package (Demleitner et al., 2014). The resulting resource directories are available under version control, one each for the plates[1] and for the spectra[2]. Of course, their current form is rather different from the early prototypes, but the basic structure has remained stable.

The important next step was to obtain access to the file system of the old publishing machine. This was eventually established through the help of Trieste Observatory, who managed to transfer the machine from Rome and save its mass storage to a modern virtual machine, which keeps running the legacy web interface for the time being; by today's standards, the raw data volumes are relatively modest (about 300 GB plate scans and 17 GB extracted spectra).

In this way, we could safely transmit the collected data through rsync.

The plates had expectably been stored in FITS format; however, the FITS headers were lacking important information, such as the observation date or the emulsion used. More seriously, their astrometric calibration was in the now severely outdated format of the Digitised Sky Survey (Lasker et al., 2008).

To remedy the first deficiency, we obtained the additional metadata from a TAP-accessible version of the Wide Field Plate Database[3] (Tsvetkova & Tsvetkov, 2006) and used this to add FITS headers in the Tuvikene convention for scanned photographic plates[4].

Repairing the second deficiency was harder. It appears that no analytic method of transforming from DSS calibration to modern WCS-SIP (Calabretta & Greisen, 2002, Shupe et al., 2005) that preserves the plate corrections has been worked out. Hence, we ended up creating the WCS-SIP headers numerically by computing a mesh of grid points (in practice, we used one per 200 pixels on each axis) and using astrometry.net's `fit-wcs` utility to compute the header (Lang et al., 2010). The implementation is found in the `compute_WCS` method in `addstandardheaders.py` in the DFBS RD repository mentioned above.

The extracted spectra were delivered to us in the form of an otherwise undocumented dump of a MySQL database that, for instance, still contained a large table of logs of the extraction procedures. The actual spectra turned out to be in per-plate tables using a custom array serialisation. That understood, it was straightforward to serialise all spectra into a single PostgreSQL table, with the flux values ending up in an array-valued column. Since the original extraction already has constant spectral bins, the spectral axis is added as a constant through a database view.

[1] https://svn.ari.uni-heidelberg.de/svn/gavo/hdinputs/dfbs
[2] https://svn.ari.uni-heidelberg.de/svn/gavo/hdinputs/dfbsspec
[3] ivo://org.gavo.dc/wfpdb/q/cone
[4] https://www.plate-archive.org/applause/project/fits-header-for-photoplates/.

In addition to the technical migration issues, there were also minor curational problems. One that deserves mentioning because it baffled us for a while although the original extraction team must already have noticed it is that the survey staff, presumably inadvertently, assigned some plate identifiers twice, namely 326, 449, and 966. This was a problem for us because spectra metadata like the emulsion used or the observation date is established through the plate identifier alone. Fortunately, in these three cases, the fields imaged are far apart, and so a certain identification of the source plate of a spectrum can be effected through a combination of plate number and position. It later turned out that of the affected plates, only the first one, FBS 0326, first epoch, were scanned, so that the duplication has not impacted data previously available.

3. Services After Migration

The result of our efforts is a set of standard services registered in the Virtual Observatory, available through both the data centers of the German Astrophysical Virtual Observatory GAVO and the Armenian Virtual Observatory ArVO. In this section, we will discuss these services and present usage scenarios.

Figure 2. DFBS spectra discovery in Aladin. See the text for how to reproduce the figure.

3.1. Spectra for Simple Clients

In the VO, the family of "S-protocols" (where the "S" stands for "Simple") gives straightforward, HTTP-based interfaces for the most common discovery tasks. In the case of spectra, this is the Simple Spectral Access Protocol SSAP (Tody et al., 2012).

We therefore publish the extracted spectra through SSAP[5]. A simple usage scenario employs the Aladin client (Bonnarel et al., 2000), in which one would point it to a location in the DFBS' coverage (M51, say), then look for "Byurakan" in Aladin's discovery tree and select one of the two DFBS services. In the resulting pop-up dialogue, check "in view" and then then click the Load button. Figure 2 illustrates the density of spectra in DFBS.

The spectra discovered in this way can be sent to TOPCAT, Splat, or some other table-processing application via the SAMP desktop messaging protocol for visualisation or analysis.

3.2. Datalink for Spectra

In a data collection like the DFBS extracted spectra, being able to inspect the provenance – in this case, the appearance of the objective prism spectra on the plates – is particularly important, because unusal spectral features may simply be due to plate defects or, in particular in crowded regions, spectral blends. Our services offer access to this provenance using the IVOA's Datalink protocol

[5] `ivo://org.gavo.dc/dfbsspec/q/getssa` and `ivo://arvo/dfbsspec/q/getssa`.

(Dowler et al., 2015). Datalink lets operators associate service calls to rows in query results, with a well-defined format of the results of these service calls.

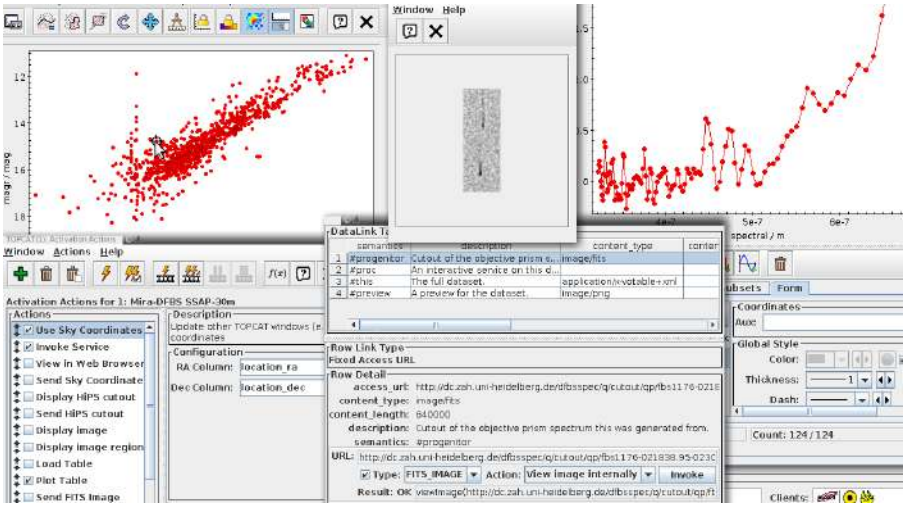

Figure 3. Accessing a cutout for a DFBS spectrum through Datalink with TOPCAT

To see how this works, start the TOPCAT VO client (Taylor, 2005). In this case, SSAP access is through the "Spectral Query" item in the VO menu. In the resulting dialogue, again look for "Byurakan" and choose one of the DFBS services. To reproduce Fig. 3, run a query on a 0.5 degrees vicinity of Mira. For this example, you could plot the result by, say, the magb versus the magr columns.

To view the spectra, in Views/Activation Action, check "Plot Table"; thanks to the metadata associated with the SSAP response, the activation action is already properly pre-configured. TOPCAT will now open a plot of the corresponding spectrum if you click on a point in the mag/mag plot.

If you additionally check "Invoke Service", another window will come up after selecting a point in the plot; this shows the extra Datalinks, among which there is one with a semantics of "#progenitor". Clicking "Invoke" will show the image; it could again be sent and displayed in Aladin using SAMP.

3.3. Bulk Spectra Processing Using TAP

For low-resolution spectra, the information content of a single dataset is limited. Their true power comes from bulk analysis of many spectra at a time.

To facilitate that without having to download a large number of spectra to local storage, the

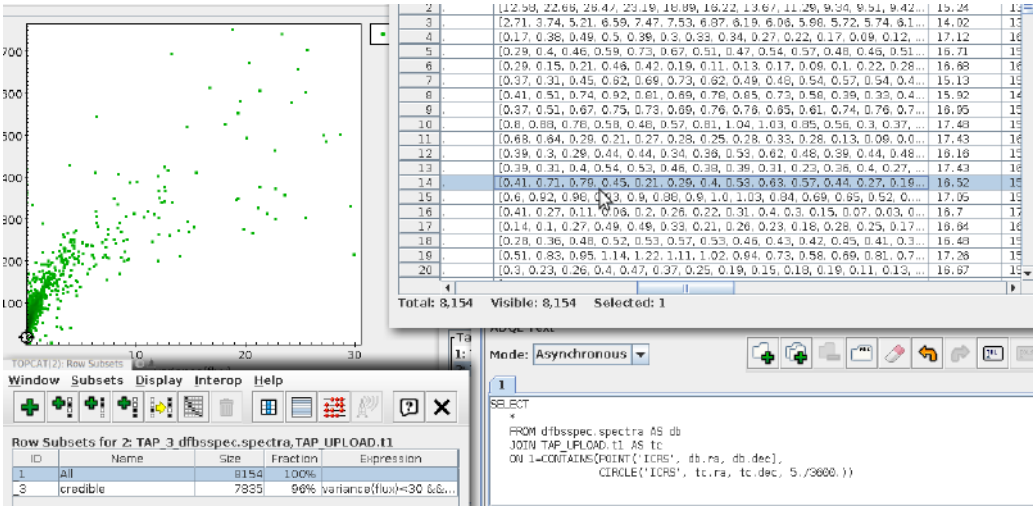

Figure 4. A sample bulk analysis of spectra of Seyfert galaxies in TOPCAT.

Byurakan spectra are also available through TAP services[6]. TAP, the IVOA's Table Access Protocol (Dowler et al., 2015), is a powerful means of performing selection and (to some extent) computation tasks in relational databases under the control of remote users. A full scenario involving outlier analysis has been written in the context of the project reported on here and is available as Demleitner et al. (2019).

A quick illustration of the power of the approach could be an overview of the spectral properties of known Seyfert galaxies as in Fig. 4.

To reproduce this, in TOPCAT, open the TAP window. To obtain a list of Seyfert galaxies, look for Simbad's TAP service and click "Use Service". On Simbad, run a query like

```
SELECT main_id, ra, dec
FROM basic
WHERE otype='Sy2'
  AND dec>-15
```

(the constraint on the declination reflects the DFBS' coverage).

This table can now be matched against the DFBS; to do that, in the TAP window, go back to the service selection tab and look for Byurakan. Select either of the ArVO or GAVO services, click "Use Service" again and locate the DFBS spectra table in the table browser. Select the table dfbsspec.spectra.

With this preparation, TOPCAT can generate an ADQL query performing a join between the local Simbad result and the remote spectra table by clicking on the "Examples" button and choosing Upload/Upload Join. The resulting query could be further edited, but it will work as generated by TOPCAT. Since the upload is relatively large, choose "async" mode before sending off the query.

The result of the query has the spectra DFBS has for the positions given by Simbad for the Seyfert galaxies as arrays in the flux column. A natural analysis might plot `sum(flux)` against `variance(flux)` (these expressions can be entered in the plot fields).

A few very noisy spectra will disturb the plot; to filter them out, create a subset through Views/Row Subsets, using `variance(flux)<30 && sum(flux)>0`. Objects standing out in the resulting plot can then be investigated using the techniques discussed above.

4. Further Work

The data as published could be made even more useful by addressing some open issues.

For one, the DFBS used a number of emulsions, which obviously will influence the shape of the spectra in ways relevant for, for instance, classification tasks. There are 31 different emulsions – most of which are rather similar, though – given in the database. Characterising them better would make global analysis more robust. Taking this endeavour further, a reliable flux calibration might be attempted, potentially using Gaia's RP/BP spectra when these become available; conversely, such a flux calibration might be used to characterise the emulsions, which would be a useful resource with a view to working with scanned plates in general.

For robust classification, estimates of the flux errors would be highly desirable. There are several conceivable paths towards providing such estimates, but all of them require resources beyond the means of our rather modest project. A side benefit of such an endeavour would be that the negative or excessively large fluxes that the extraction pipeline occasionally produced – they are what required the subsetting in the TAP example above – might be remedied.

5. Conclusions

Conclusions projects planning on larger-scale data publications could draw from our experiences include:

[6] `ivo://org.gavo.dc/tap`, `ivo://arvo/tap`

- Avoid proprietary, stand-alone proprietary solutions: They do not age well. Yes, this includes custom-made web pages.
- If you really cannot avoid proprietary code, try to co-locate it with established data centers that will (hopefully) maintain and migrate it as technology evolves.
- Make sure you keep contact with staff operating the machines that publish the data physically; you should always retain an option to access the file system the published data resides on.
- As people leave, make sure there is documentation on the files, tables, and procedures and, probably even more importantly, that someone staying knows where this documentation is.
- Plan upgrades to your data products' metadata and data formats as the relevant standards and practices progress.

Acknowledgements

This work was supported by a joint German-Armenian programme, BMBF FKZ 01DK17055, and the e-inf-astro project, BMBF FKZ 05A17VH2. This work would not have been possible without the help of Marco Molinaro of Trieste Observatory. We thank François Bonnarel for helping us out with the DSS polynomials, which otherwise were hard to locate.

References

Bonnarel F., et al., 2000, Astronomy and Astrophysics Supplement, 143, 33

Calabretta M. R., Greisen E. W., 2002, A&A, 395, 1077

Demleitner M., Neves M. C., Rothmaier F., Wambsganss J., 2014, Astronomy and Computing, 7, 27

Demleitner M. aand Mickaelian A., Knyazyan A., Baghdasaryan D., Mikayelyan G., 2019, Outlier Analysis in Low-Resolution Spectra: DFBS and Beyond, VO Tutorial, doi:10.21938/MxUvKseqU˙gxCwyGocudlg, http://www.g-vo.org/tutorials/dfbs.pdf

Dowler P., Bonnarel F., Michel L., Demleitner M., 2015, IVOA DataLink Version 1.0, IVOA Recommendation 17 June 2015 (arXiv:1509.06152), doi:10.5479/ADS/bib/2015ivoa.spec.0617D

Lang D., Hogg D. W., Mierle K., Blanton M., Roweis S., 2010, AJ, 139, 1782

Lasker B. M., et al., 2008, AJ, 136, 735

Mickaelian A. M., et al., 2007, A&A, 464, 1177

Shupe D. L., Moshir M., Li J., Makovoz D., Narron R., Hook R. N., 2005, in Shopbell P., Britton M., Ebert R., eds, Astronomical Society of the Pacific Conference Series Vol. 347, Astronomical Data Analysis Software and Systems XIV. p. 491

Taylor M. B., 2005, in Shopbell P., Britton M., Ebert R., eds, Astronomical Society of the Pacific Conference Series Vol. 347, Astronomical Data Analysis Software and Systems XIV. p. 29

Tody D., et al., 2012, Simple Spectral Access Protocol Version 1.1, IVOA Recommendation 10 February 2012 (arXiv:1203.5725), doi:10.5479/ADS/bib/2012ivoa.spec.0210T

Tsvetkova K. P., Tsvetkov M. K., 2006, in Tsvetkov M., Golev V., Murtagh F., Molina R., eds, Virtual Observatory: Plate Content Digitization, Archive Mining and Image Sequence Processing. pp 45–53