Astronomical Data Analysis Software and Systems XVII **A.09**
ASP Conference Series, Vol. XXX, 2008
J. Lewis, R. Argyle, P. Bunclarck, D. Evans, and E. Gonzales-Solares, eds.

Automated data releases for the WFCAM Science Archive

Ross Collins, Nicholas Cross, Eckhard Sutorius, Mike Read, Nigel Hambly

Wide-Field Astronomy Unit, Institute for Astronomy, University of Edinburgh, UK

Abstract. The wide-field camera (WFCAM) on UKIRT is performing a variety of infrared surveys, from the large consortium-led UKIRT Infrared Deep Sky Surveys (UKIDSS) to the smaller private programmes that vary considerably in duration and nature. All of these data are released in the WFCAM Science Archive (WSA), initially only to the principal investigators and then publicly once a proprietary period has elapsed, prepared by the Wide-Field Astronomy Unit (WFAU) at the University of Edinburgh. Development of released data products has thus far focused on the requirements of UKIDSS, because the other WFCAM programmes are both too numerous and varied to support as fully. Here we describe our approach towards the automated preparation of a standard set of release products for all of these private programmes. This includes support for most of the features available in UKIDSS releases, such as tables of passband-merged, unique sources, as well as deep image stacking and synoptic data sets for surveys with repeated observations. The same approach can also now be adopted to private data in the forthcoming VISTA Science Archive.

1. Introduction to the data

All data taken by the infrared wide-field camera (WFCAM) on UKIRT are archived by the Wide-Field Astronomy Unit (WFAU) at the University of Edinburgh in the WFCAM Science Archive (WSA; see Collins et al. 2006, Hambly et al. 2008). Most data form part of the consortium of UKIRT Infrared Deep Sky Surveys (UKIDSS) that consists of 5 survey programmes for the ESO community (see Lawrence et al. 2007). Open time is also available for private programmes, typically averaging 50 in number per year. Each of these programmes may follow a wide variety of observing strategies, with many different pointing combinations and up to 10 different filters. For example, the programmes may choose fields that overlap or even possibly repeat observations of the same field in the same filter for greater depth or temporal coverage.

When providing an automated data preparation and release system one cannot rely on users, the programme principal investigator (PI) in this case, to always provide accurate details of the programme. Instead the data preparation and release is driven by the metadata of the observations that have been archived for the given programme. These metadata are initially provided by the summit software at UKIRT, which stores each observation in FITS files with header keywords providing metadata on programme, filter, time, position, grouping etc. Nightly pipeline-processing of the data is performed by the Cambridge Astronomical Survey Unit (CASU), which produces stacked images and extracts catalogues, all with FITS header metadata describing provenance etc. The metadata of these processed products as well as the complete detection catalogues are ingested into the WSA database once transferred to WFAU.

2. Automated data release software design

When the observation metadata are ingested into the WSA database, new programmes, identified by unique project keyword values in the FITS header, are automatically assigned a unique programme ID, which is registered in a database table named `Programme` (see figure 1). A table for that programme's detection catalogues is automatically created based on the standard schema template and into which the catalogues are ingested.

After the initial ingest of the pipeline-processed observations, a standardised sub-set of archive products are prepared for each programme depending on the type of observations taken by the programme. These are:

- For programmes with repeated observations of the same field with the same filter (the *deep* surveys):
    - Inter-night deep-stacked images with detection catalogues.
    - Tables identifying unique sources in each re-observation of the same field providing synoptic information (see Cross et al. 2009).
- For all programmes:
    - Tables of unique sources where detection catalogues from fields re-observed in different filters are merged to provide colour information.
    - Tables of neighbouring sources within a fixed radius.
    - Tables of sources cross-matched with different data sets, such as the 2MASS survey.

The preparation of a release database for the UKIDSS surveys is carefully designed around the needs of each of the 5 survey PIs. Every image is scrutinised during the quality control process and the detection catalogues go through several stages of quality control, both automatic and manual, before and after passband-merging. This level of control is impossible to achieve for the large number of private programmes hosted in the WSA database, without an equally large staffing effort to handle the logistics. Instead quality control of private programme data is only applied where it can be done so automatically, including:

- Exclusion of observations marked as rejected at summit.
- Rejection of observations with incomplete/invalid metadata.
- Rejection of stacked images without corresponding catalogues or invalid catalogue attributes.
- Flagging of the relative quality of source detections to identify suspected artefacts, false detections or just poor quality data.

The schema of the source and synoptic tables is individually tailored for each programme, based on its requirements, to provide appropriate columns for the available colour combinations etc. For the private programmes these requirements are automatically determined by inspecting the complete set of quality-controlled metadata available at the time of release.

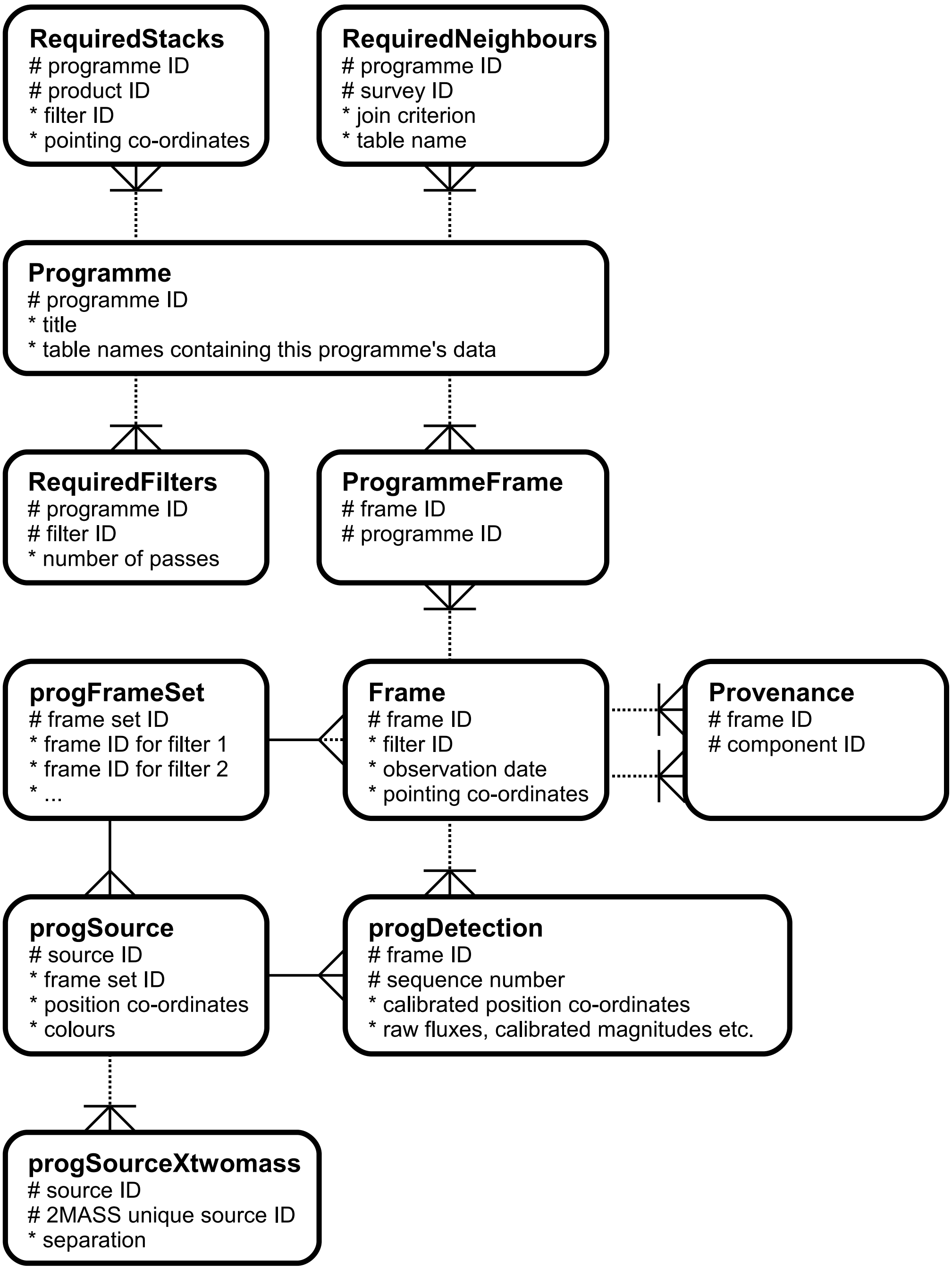

Figure 1. A simplified view of some of the relevant tables in the WSA database design (see Hambly et al. 2008 for the complete database design and see Cross et al. 2009 for the synoptic design). Each table with the *prog* prefix in its name represents a single table for an individual programme with its schema tailored to that programme's requirements.

From the metadata the set of filters used for the programme's observations are automatically determined and entered into a database table named `RequiredFilters` (see figure 1). If there are repeated pointings identified for a given filter then entries are inserted into a `RequiredStacks` table that defines the parameters for producing deep stacked images of these pointings. Additionally the synoptic parameters of the observations are defined in the `RequiredSynoptic` table (see Cross et al. 2009) and the cross-matches required are listed in a `RequiredNeighbours` table. From the parameters listed in these tables the schema for the release database tables can be automatically created to produce passband-merged source tables containing the correct colours, cross-match and synoptic tables where required.

With the programme requirements and database schema defined, the software developed for curating UKIDSS survey data can be automatically run for each private programme. This software, written in Python[1], was designed to be database/schema-driven, so that it can be applied to any programme using the parameters defined in the database to produce the standard set of products listed above. If required, deep image stacks are created, from which catalogues are extracted and ingested into the database. Tables of unique sources, merged across the available passbands, are then created along with the tables listing neighbouring sources, cross-matches with other survey databases as well as tables of synoptic data where appropriate. In the final step of the automated curation process, individual private release databases are created for each programme by copying out that programme's fully-prepared data from the archive's load database.

Once a semester's worth of pipeline-processed observations have been ingested into the WSA, a release of all private programme data is made. A script is executed that identifies newly-registered, unreleased private programmes and automatically performs each stage of curation as necessary for the data up to the point where a private release database has been created for each programme. A typical semester of data, containing 25 new programmes, with an approximately even split between *deep* and *shallow* surveys, will take approximately 3 days of processing to prepare the release databases, with the time dominated by the preparation of the deep stacked products.

References

Collins, R. S. et al. 2006, in ASP Conf. Ser. 351, ADASS XV, ed. C. Gabriel, C. Arviset, D. Ponz, & E. Solano (San Francisco: ASP)

Cross, N. J. G., Collins, R. S., Hambly, N. C., Read, M. A., & Sutorius, E. T. W. S. 2009, in ASP Conf. Ser. XXX, ADASS XVIII, ed. D.A. Bohlender, D. Durand & P. Dowler (San Francisco: ASP), [A.10]

Hambly, N. C. et al. 2008, MNRAS, 384, 637

Lawrence, A. et al. 2007, MNRAS, 379, 1599

[1]See http://www.roe.ac.uk/~rsc/wsa for the software documentation