European Organisation for Astronomical Research in the Southern Hemisphere
Organisation Européenne pour des Recherches Astronomiques dans l'Hémisphère Austral
Europäische Organisation für astronomische Forschung in der südlichen Hemisphäre

Data Products Department

Science Grade Data Products

prepared by Wolfram Freudling

September 17, 2008

Abstract The production of science grade data products requires an integrated approach to planning of the science and associated calibration observations, as well as software to produce the final product. Any effort to improve the science quality of data products needs therefore to look at the whole workflow from the OBs to the pipeline output. This document lists a short list of high-level requirements and considerations for each of the steps. It is meant as a guide to review the current state of existing instruments modes, and as input for the planning of the workflow for future instruments or instrument modes. In the last section, a survey of the current status of science products is described and the questionnaire of the survey it attached.

1 Introduction

1.1 Purpose

ESO supports the production of science grade data products for all of its VLT instruments. This requires an integrated approach to science and calibration observations as well as software to reduce and calibrate the raw data. The level to which this is currently achieves varies greatly from instrument mode to instrument mode. The first step to improve the overall science quality of data produced and delivered from ESO data is to survey the current status for all VLT instruments. The purpose of this document is to describe the top-level requirements to produce science ready data, and to provide a framework to survey the current status of VLT data products. The goal of the attached survey is to provide an overview of the current status of VLT data products. This overview is will be used to characterize data products in view of their publication in the archive, and to identify those data products which need further improvement.

1.2 Definitions

The degree to which a data product can directly be used to "do science" without further processing depends on both the quality of the data product and the particular application. Many terms are in common use to describe the level or reduction and calibration of data products, such as "science ready",

"science grade", "advanced" or "high level". To avoid confusion, we define in this section the terms as used in this document. These definitions are useful to distinguish between different quality levels for data products, but should be used as guideline rather than literally. For consistency, we suggest to use the same terminology throughout ESO.

Science grade data products (SGDPs) are data products which can be used as-is to extract scientific conclusions, or to carry out quantitative measurements. This implies that the instrument signature has been removed, the SGDP is calibrated in physical units, and the signal-to-noise ratio is close to the optimum which can be achieved. All SGDPs also include error estimates. Typical SGDPs are fully calibrated and mosaiced images which include noise maps, one or two-dimensional flux calibrated spectra with error bars, or three dimensional position-wavelength cubes. Any assumption used in the creation of SGDPs are independent of the science goals, such as assumptions on instrument properties, environmental conditions or noise properties. Assumptions on the scientific contents of the data, external knowledge which is not generally valid or depends on the targets, or scientific judgment related to the contents of the images is not used for the production of SGDPs. SGDPs are therefore general purpose data products and independent of specific targets. For example, photometric redshifts are not SGDPs since they depend on templates. Single-line redshifts (e.g. the results of Ly-α of Hα) searches are also not SGDPs since they depend on external judgment to identify the line. On the other hand, redshifts based on unambiguous sets of lines might be SGDPs.

Advanced data products (ADPs) are science products extracted from SGDPs which can directly be used for science analysis. In many cases, ADPs are tuned to a particular science application. In deriving ADPs, assumption might be used which are valid only under special circumstances or for some of the targets. A detailed description of the underlying assumption is essential for ADPs. While it is possible to use ADPs directly for scientific analysis, in many cases they will be re-derived by users who use them as a starting point for tuning processing parameters, or who use them to judge whether a dataset is useful for a particular purpose. ADPs are science grade in the sense that they are publication ready quality, but they are not necessarily optimized in a general sense. Typical ADPs are source catalogs extracted from images which include parameters such as shape parameters or redshifts, or line lists and/or classification of spectra. Another category of ADPs are images, spectra or data cubes produced by non-standard co-adding of the corresponding SGDPs. An example for a co-added image which qualifies as ADP as opposed to just a SGDP is a mosaic specifically created for the purpose of detecting weak lensing. The choices for distortion correction, weighting and PSF matching for such an application are different from the vast majority of uses of imaging data.

Metadata (MD) are auxiliary data which describe a data product. Every data product includes at least some MD. This might include single value parameters ("keywords"), tables, two or higher dimensional arrays, or human readable documentation. MDs are created during the observation and during the processing of the data. Some of the MD is necessary to exploit a data set, other MD are created to allow searches for specific data in the archive. Some of the requirements for SGDPs, such as the flux scale or error estimates, might be fulfilled by providing appropriate MDs. Specification of VO services (SIAP and SSAP) require mandatory keywords for VO-compliant data. This **VO metadata (VOMD)** are part of all SGDPs. A non-trivial example of MD are object catalogs meant to describe the data product. Such catalogs are not meant for science as are ADP catalogs, but are simply a description of the data to enable searches e.g. for non-crowded fields, or to search the archive for data on a particular object.

A **Science Data Reduction Package (SDRP)** is an integrated software package which can be used to convert raw data into SGDPs. A SDRP typically includes a number of modules which are run in sequence. These modules are called recipes. Ideally, SGDPs can be produced without any interactivity, but there are observing modes where interactive recipes are neccesaary to create SGDP. A

SDRP might therefore include both interactive and non-interactive recipes. A **pipeline** is a sequence of non-interactive recipes which can be run stand-alone.

Observations at ESO are specified in **Observing Blocks (OBs)**, which contain all the information necessary to obtain a "single" observation. These include the target position, the instrument and exposure setup parameters, and, in Service Mode, also the scheduling requirements, and time constraints. Such a single OB can contain in principle one or multiple exposures, or even multiple instrument configurations with multiple exposures.

2 Requirements for SGDPs

2.1 OBs

Science grade output from pipelines starts with the planning of the observations. Observations are specified in Observing Blocks (OBs). Information entered during the OB creation can be used by the pipeline. High level requirements relevant to the creations of SGDP are:

1. All necessary information to create SGDPs can be specified. For example, a SGDP might require multi-OB observations. In that case, a mechanism to associate observations already during Phase 2 preparation should be supported.

2. Calibration requirements can be specified in a systematic way, not just as comments. Calibration requirements to produce "science grade" output depend on the science objective of the observations. It is important that calibration requirements on top of those specified in the calibration plan can be specified in a systematic way so that observations can be carried out accordingly, and the information can be carried forward to the reduction pipelines. For example, photometric observations require multiple observations of photometric standards during a night. These calibrations can be used to compute more accurate photometric zero points for the night, as well as estimate the fluctuations in transparency of the atmosphere. To create photometrically calibrated SGDPs to a specified accuracy, there must be a way to specify the level of desired photometric calibration in OBs.

3. Observing constraints can be specified in a way that correlates directly with the quality of the science data. For example, specifying observing constraints in terms of physical units such as the maximum background level rather than moon phase will in some case improve the quality of the data products. Other examples for such specifications are the coherence time for VLTI, and the precipitable water vapour content for mid-IR observations.

2.2 Calibration Data

Routine calibrations are carried out according to the calibration plans. Calibration plans specify the routine calibrations carried out in regular intervals for each instruments. The purpose of a calibration plan is to to provide a basic level of calibration which is needed for any science application of the data, and to generate data for the monitoring of the instrument performance. Users are not charged for calibrations as specified in the calibration plan, but can request additional calibrations taken from their time allocation to improve the accuracy of the SGDPs. Requirements for the calibration plans are:

1. The main requirement for the calibration plan is that it specifies calibration data which are necessary and sufficient to create SGDPs, including realistic error estimates.

2. At the same time, our most precious resource, the night time observing time, should be efficiently used. This implies that calibrations use as little night time telescope time as possible, and that the level of calibration is the one needed for the vast majority of science applications.

3. The accuracy of the calibration data has to be well documented to allow estimates of systematic uncertainties.

4. The calibration plan should also outline a strategy users can employ to improve the calibration on top of the routine calibration carried out by the observatory.

In some cases, changes in the pipeline to improve the quality of the SGDP might require different calibration data, which not necessarily implies additional telescope time. For example, photometric calibration of FORS images requires second-order corrections to the flatfield. Changing the photometric standard fields and dithering of the photometric standard observations were needed to support the new module in the pipeline.

2.3 Pipelines

Pipelines aim to produce those "science-grade" data products which do not require interactivity. They need to be fed with both the right science and calibration data. The data products include metadata which completely describe the product. At a minimum, the final data products include information on flux, position, and wavelength with error estimates.

In most cases, steps carried out by science-grade pipelines will not be re-done by the user. Re-processing might lead to improved results, but such improvements are incremental. Re-processing of science grade data products is typically necessary if small improvements in accuracy matter, if the user wants to optimize the processing for a special use of the data, or if re-processing can achieve substantial improvements through interactive techniques which cannot be implemented in a pipeline.

High-level requirements for science-grade pipelines are:

1. Pipelines support the general infrastructure for classification, grouping and association of data. They are modular and produces the intermediate products necessary to judge the success of the algorithm. Pipelines are required to support three different modes: 1) an on-line mode used for QC at the observatory, 2) an off-line mode used for quality control in Garching which can produce SGDPs, and 3) a user mode in which the reduction process can be stopped, restarted, and/or partially re-executed. There is substantial overlap between the modes, and in some cases they might be identical. The user mode of future ESO pipelines can be run within the Reflex environment.

2. Whenever possible, measured quantities are given in physical units, or the factors to convert to physical units is included in the metadata. The minimum set of information is flux scale, position and wavelength. This information might either be included as part of the actual data (e.g. the wavelength bins of a spectrum) or be part of the metadata (e.g. wavelength range for images).

3. Instrument signatures are removed where possible. Instrument signatures which cannot be removed are described in the metadata. This description includes the necessary details for subsequent attempts to improve the data product, or to evaluate the impact on science results.

4. The algorithms achieve close to optimal signal/noise, i.e. the measured noise is within 50% of what can be achieved. The sources of the measured noise in the data products are understood. The

contribution of computational noise is negligible. The noise added to the data by the processing is determined by the availability of calibration data. Noise contributed by calibration data which do not require night time telescope time (e.g. biases) is negligible.

5. All measured quantities are given with error estimates, including random and systematic errors. The random error estimates are based on error propagation from known sources of noise such as detector readout noise or Poisson noise. When possible, the errors are given for individual data points (e.g. pixels or wavelength bins). Estimates of the accuracy of the scale and systematic errors are given in the metadata separately from the random errors. For example, a systematic uncertainty of the flux scale is introduced by errors in the photometric zero point. This is given as an error estimate of the zero point in the meta data.

6. Atmospheric and environmental conditions are described in the metadata. The effect of the atmosphere is removed where possible. Random and systematic errors introduced by atmospheric and environmental conditions are included in the error estimates. For example, the extinction for a night determines the flux scale and therefore also the scale for the flux error estimate. The photometric quality of the night determines the accuracy of the photometric zero point and thereby the systematic errors.

7. Data products are in a form which is understood by commonly used analysis tools. Formats comply with applicable standards. Where possible, multiple standards are supported. For example, essential information is stored both in ESO Hierarchical keywords as well as standard FITS keywords.

Pipelines might also produce high level or advanced data products such as catalogs of objects. By definition, the production of high level data products requires decisions based on judgment and assumptions, and such decisions are not unique. Such high level data products are useful for example for archive searches, but they are not necessarily meant for science.

2.4 Interactive Tools

In some cases, science grade reduction might not be possible in a pipeline. Such interactive tools are different from analysis tools which are used to carry out measurements on reduced data. These cases should be clearly documented, and tools should be made available to accomplish science grade reductions of ESO data. Reflex workflows should include the option to use interactive tools.

3 Status of Science Products – Instructions to fill the form

Currently, SGDPs are routinely produced for only a small subset of instrument modes. Expanding the availability of SGDP to all VLT instrument modes is a huge task given the large number of instruments and modes. As a guide to direct our effort, and as a first step towards improving the VLT science output, we are carrying out a survey of the readiness of data products for all instruments. Attached to this memo is the survey form. The purpose of the form is to collect input on the status of the data products for a particular instrument mode.

For instrument modes which routinely produce SGDPs, the form should be used to describe the accuracy of these products. For instrument which need improvements to data products to make the science grade, the form should used to propose particular changes to the current workflow of the instrument. This might be a change to the pipeline, the calibration data or anything else in the

data flow. Note that the nature of changes to be proposed here is different from tickets submitted to SDF. Tickets should address a specific request for a software change, whereas in this form we ask for broader ideas to improve the science grade of the products. Each proposed change might translate to one or several tickets.

One form should be filled out for each proposed improvement, i.e. several forms for the same instrument should be used to propose more than one change. A reduced version of the form is shown on below. For the survey, the actual pdf form template should be used and filled out on a computer, for example with *Adobe Acrobat Professional* under Windows, or *acroread* under Linux. If possible, the form should be saved with the form data. Otherwise the form should be filled and printed to a file as postscript or pdf. Please use the following scheme for the filename: `instrument_mode_myfirstname_mylastname_n.pdf`, where `_n` is a running number in case several forms for the same instrument mode are submitted. This file should be mailed to sdp@eso.org.

The instructions for each section of the form are as follows:

In the form header, specify the instrument mode, your name and expertise. People with any level of expertise are welcome to fill this form.

1. Select the current overall status of this instrument mode. If the status is "science grade, no further improvement", proceed to question 7.

   Select "science grade, archive data should be reprocessed" if data products are science grade and recently improved enough to warrant re-processing of archive data. This could be due to an upgrade of the pipeline, or because new calibration data available which are applicable for archive data.

   If the data products are already useful for science, i.e. science grade or almost science grade, select the third option.

   The last two options imply that ESO currently does not produce science grade products for this mode.

2. If an improvement in the current pipeline is proposed, give this proposal a one line title. The title should emphasize the resulting improvement. For example, "Improve Astrometry" is a better title than "More digits in WCS". The priority should rank the importance in case that several changes are proposed for the same instrument mode.

   If this proposal corresponds to one or several tickets with SDF, check the appropriate box. If it is unknown whether there is a ticket or not, do not check the box. In addition, if possible the ticket numbers should be given.

3. Check any category which needs work to implement the proposed improvement. One or several categories can be checked. For example, if the proposal is to create mosaics from several OBs, it might be necessary to put some information on associations into the OBs, change the calibration plan to get better astrometry, change the pipelines to read read and interpret this information and add a new module to the pipeline to create the mosaics.

4. List the science parameter which will improve, e.g. "magnitudes" if the proposal is to improve the flatfielding of images.

5. Categorize the nature of the improvement. Again, think about the impact of the change rather than what has to be done.

6. Any proposal will require work to be implemented, which will be estimated later. However, here the question is the impact *after* implementation. The important distinction is between changes which are a one-time effort with little or no impact later, and those which require continued effort.

7. Describe the proposal and/or the status of the current pipeline. If the status given in question 1 was given as "routinely produces science grade output", describe the data products, in what sense they are science grade, and if possible give quantitative estimates the accuracy achieved (e.g. "wavelength calibration routinely better than 0.1 pixels"). If an improvement of the data quality is proposed, give the details here and again where possible include quantitative estimates for improvements, e.g. "error estimates can be improved by factor of 2".

8. Any suggestions about improvement to the pipelines, calibration plan or OBs which are *not* aimed at improving the science quality of the data products should be described here. For example, missing quality control parameters a pipeline should produce from the science frames should be described here.

9. Put in any comments, including comments on the usefulness of this form or the contents of the document.

Status of Science Data Products

Instrument:

Mode:

filled out by:

name

email

ESO division/group

level of expertise:

- ◯ I observed with this instrument mode for my research
- ◯ I used archive data from this instrument mode
- ◯ I am responsible for this instrument mode for my work
- ◯ other

1. Status of data products as routinely produced by current pipeline and calibration plan:

- ◯ science grade, no further improvement required
- ◯ science grade, archive data should be reprocessed
- ◯ useful for science but can be improved
- ◯ not useful for science
- ◯ mode requires interactive tools to produce science grade output
- ◯ there is no science data reduction package supported by ESO
- ◯ other

2. What do we need to change to produce science grade data products for this mode?

one line description of proposed change

Priority: (lower number = higher priority)

☐ proposed change is already fully or partially covered by ticket(s) with SDF.

ticket number(s) if known:

3. Proposed Change:

Pipeline:	◯ no change	◯ minor changes	◯ major additions / rewrite
Calibration plan:	◯ same as now	◯ change in procedure	◯ new / additional calibration
OBs:	◯ same as now	◯ change in procedure	◯ new parameters

4. Parameter to be improved (e.g. flux, wavelength, position):

5. Nature of improvement:

- ◯ compute physical scale (e.g. give mJy instead of counts)
- ◯ improved accuracy of calibration
- ◯ reduced noise
- ◯ more accurate error estimates
- ◯ additional description of data (e.g. shape of PSF)
- ◯ reduce bad data points (e.g. cosmic rays)
- ◯ new data products (e.g. catalogs, co-additions)
- ◯ improved format of product (e.g. 3D data cube)
- ◯ other

6. Impact of the proposed change once it has been implemented and tested:

quality control:	◯ unknown	◯ negligible	◯ some	◯ major / new resource
software maintenance:	◯ unknown	◯ negligible	◯ some	◯ major / new resource
calibration time:	◯ unknown	◯ negligible	◯ some	◯ major / new resource
telescope time:	◯ unknown	◯ negligible	◯ some	◯ major / new resource

7. Details on science readiness of data products

8. other suggestions for pipeline / calibration plan / OBs

9. General comments