The Challenge of Data Reduction for Multiple Instruments on Stratospheric Observatory For Infrared Astronomy (SOFIA)

Miguel V. Charcos-Llorens[(1)]; Bob Krzaczek[(2)]; Ralph. Y. Shuping[(3)]; and Lan Lin[(1)]

(1) Universities Space Research Association, NASA Ames Research Center, Mail Stop N211-3, Moffett Field, CA 94035, USA

(2) Chester F. Carlson Center for Imaging Science, RIT, 54 Lomb Memorial Drive, Rochester NY 14623, USA

(3) Space Science Institute, 4750 Walnut Street, Boulder, Colorado 80301, USA

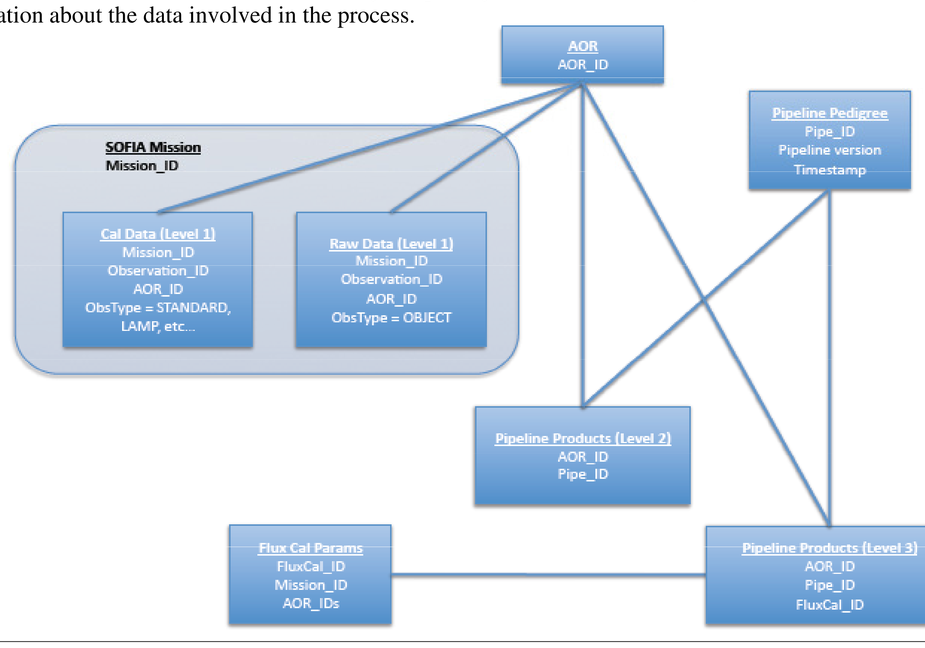
1- INTRODUCTION

SOFIA is an airborne telescope for infrared and sub-millimeter astronomy. Its design combines the challenges of ground and space observatories. Contrary to satellite observatories, SOFIA hosts a large variety of instruments observing in wavelength ranges from 0.3 to 600 microns which will be upgraded over time. This will produce a large diversity of data types which will likely increase as new generations of instruments are operated. The SOFIA Data Cycle System (DCS)[1] will support, for both facility and visitor instruments ,all aspects of data processing including archiving and pipelining of data of level 1, (raw), 2 (processed-instrumental and background removal), 3 (flux calibrated), and 4 (higher processing- mosaicing and source catalogs). The DCS will provide uniform, extensible and supportable framework for all aspects of the data cycle.

Data processing includes all steps required to obtain good quality flux calibrated data of spectroscopy, imaging, fast-acquisition, polarimetry, etc. Processing each data type requires a sequence of unique or common algorithms with specific parameters to be tuned. DCS will incorporate, improve and maintain these algorithms which are provided by the instrument teams and developed in a variety of environments. In addition, these algorithms may require user-interaction or fine tuning of input parameters in order to return good quality data.

2- CONCEPTS AND ASSOCIATIONS

The Astronomical Observation Request (AOR) is the link between scientific and calibration data of the same observation type. It defines the parameters necessary for the observation and post-processing. Therefore, it will identify the reduction pipeline and its parameters. For each level 2 product, the Pipeline Pedigree (PP) records the pipeline generating the data, the parameters, the processing date and the data involved in the process. AOR and PP concepts has been implemented and are operative in DCS. A similar concept will be necessary to track calibration activities. DCS will include a Flux Calibration Parameter (FCP) which will support the calibration process in order to document and reproduce the same results as needed. AOR, PP and FCP are characterized by unique key numbers that identify them as well as information about the data involved in the process.



3- ARCHITECTURE

The DCS provides the framework for a typical processing which requires switching between automatic pipelining and human activities. There are four main scenario defining data processing:

(1) EoF automatic pipelining producing immediate level2 products
(2) flux calibration producing level3 products
(3) manual processing and inspection (all levels)
(4) re-processing (all levels)

DCS will host automatic pipelining at EoF, user-initiated pipelining and user-interaction processing and analysis. In addition, the DCS ingestion tool allows archiving level 3 data that the Science Mission Operation (SMO) performs (outside DCS) from level 2 data previously processed within DCS. The diagram below illustrates how these scenarios relate to each other. The CORE is in charge of data processing within DCS (green). User can perform data processing outside DCS (red) and use DCS tools to extract and archive data. We show data flows as dashed arrows and process requests as plain arrows.

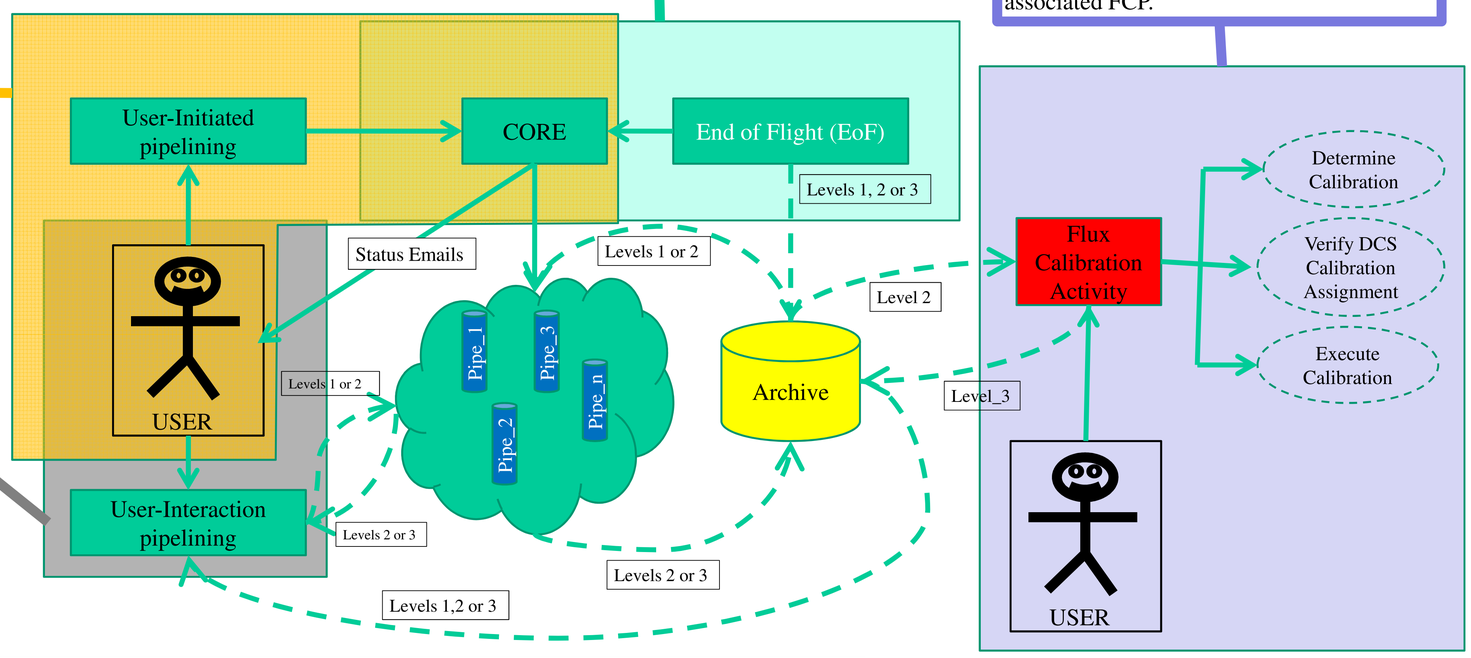
SCENARIO 1: AUTOMATIC PIPELINING
Flight data, as for example raw observations or flight-processed products, are ingested at EoF. DCS calls pipelines automatically after ingestion of data observed during flight operation. Products from data reduction, typically level 2 data, are automatically archived as the data are processed, making them quickly available for scientific analysis.

SCENARIO 2: FLUX CALIBRATIONS
Outstanding scientific results can be obtained only with calibrated data. Flux calibration are complicated processes that are hardly automated. The difficulty of defining a metric of the data quality makes necessary intervention of experienced scientists. Final level 3 products can be archived in SOFIA database as well as their associated FCP.

SCENARIO 4: USER INITIATED PIPELINING
Pipelining can also be manually initiated by SMO scientists. For example, they may re-pipeline the data with modified parameters which would improve the quality of the final results.

Iterative interactions

SCENARIO 3: MANUAL PROCESSING
Likely, human intervention is often needed to verify results at any of the data levels. DCS will provide an interaction interface to extract data from the archive and run locally the same algorithms used during automatic pipelining. This allows the user to analyze the data at any step of the process, eliminate undesirable data and fine-tune parameters of the reduction. This step will result on the data validation or the appropriate parameters required to re-pipeline data in order to improve the quality of the final product.



4- PIPELINNING APPROACHES

A pipeline is a collection of algorithms which are run in a particular order. The DCS will store pipelines coded in IDL, Python and C++ which are delivered by the instrument teams in association with a XML file describing the pipeline recipes. With the appropriate pipeline specification, DCS can currently run pipelines in any language with no modification of the code as soon as the pipeline is delivered as an executable, likely the same that runs in SMO machines outside DCS. Because DCS does not have a knowledge of the details of the pipeline execution after it is called, we name this pipeline "blackbox". Level 2 blackboxes are applied based on the specifications of the AOR - which is detailed before the flight as part of the observation planning process – and the details of the process are recorded on the PP – which is created after pipeling – as explained in section 2. This approach is currently implemented in the DCS and embraces automatic pipelining at EoF and manually pipelining initiation within the same framework. This answers the need for re-pipelining with the goal of improving the quality of the Level 2 data by fine tuning pipeline parameters after manual inspection or applying an improved version of the pipeline. Although, this approach represents an enormous cost saving on the implementation and maintenance of the pipelines it lacks of the advanced functionalities that the DCS could offer including parallel execution of processes of a single pipeline, status report , and intermediate user intervention.

We plan to complement the current functionality with another approach allowing human interaction. User interaction is required for step-by-step data processing and intermediate data analysis. These will be performed using a DCS graphical interface tool which runs user-interaction data process and analysis tools locally (outside DCS) after downloading updated algorithms from DCS. As a long term goal, DCS will integrate user-interaction pipelining within the same framework as automatic pipelining. For that purpose, pipelines will be delivered as a collection of functions (modules) performing a portion of the pipeline and XML files describing them. The pipeline recipe (another XML file) will describe how modules are executed, the order of execution and how data is transfer between modules. Technically, the pipeline manager objects (pipe_man) are in charge of executing specific modules (module->process method) or the whole pipeline (pipe_man->run method). This new approach fits in the actual black box structure by calling run method as the pipeline executable. When implemented within DCS, pipe_man will be able to process modules in parallel, control their execution, and allow user data analysis. In addition, pipe_man will manage modules in different computer languages for the same pipeline thus, reducing the number of algorithms in the system. Instrument teams will be encouraged to use existing algorithms when developing their pipelines, resulting in a common library of algorithms which will decrease the efforts of the instrument teams for developing pipelines and of the DCS team for maintaining and upgrading them.

5- CONCLUSION

Combining automatic pipelining and user interaction of processing algorithms which are developed in various languages presents an important challenge to the DCS. Especially, when trying to minimize efforts required for long-term maintenance and upgrade of the code. We divide the problem in four distinct cases of interaction with the data. These scenario can be developed independently but are based on a common architecture. The case of automatic pipelining, either at EoF or user-initiated, is already implemented and tested with FLITECAM data. User-interaction pipelining is in its design phase but we show its feasibility using a prototype design in IDL. Flux calibration activity is out of the scope of the current DCS version but we provide the require tools for the user to ingest human validated data.

(1) http://dcs.sofia.usra.edu/

For more information about SOFIA: http://www.sofia.usra.edu/