Large and Long Program Processed Data Requirements for Archive Ingestion

Kathleen Labrie, Kenneth Anderson

Science User Support Department

V1.0 – 28 February 2018

Revision History

V1.0 – 28 February 2018 Kathleen Labrie

Document ID: GOA-USER-101_LPPProcDataReqs

Document Purpose

This document defines the minimum requirement for the ingestion of Large and Long program processed data into the archive. Here we focus on the technical and provenance requirements.

Intended Audience

This document is intended for the investigators of Gemini Large and Long Programs who are required to provide processed products to be ingested in the archive and made publicly available.

Table of Contents

1. Background

Gemini Large and Long Programs (LLP) teams are required to provide the Observatory with processed data products for ingestion to the Gemini Observatory Archive (GOA) within one year of the planned end date of the program. For proper ingestion and to retain the ability to search

the data, the data products must meet a certain number of criteria, as described in this document. Also included are requirements regarding the provenance and the processing of the data.

2. Applicability

The current version of this document, v1.0, applies to LLP programs accepted in the 2018 proposal cycle and beyond. Older LLP programs are also encouraged to submit reduced data to the archive, following the same requirements outlined here, but are not required to do so.

Table 1: Applicability of the rules by semester.

Proposal Cycle	Applicable version
2018B and beyond	V1.0
2018A and older	n/a

3. Types of Data Products

3.1 Imaging

Imaging data products will include frames corrected for the telescope, instrument, detector and atmospheric effects. For example, a detrended (dark subtracted and flat corrected), sky subtracted near-infrared image would be the minimum required data product. A better-value processed image would be a flux calibrated stack. If the data set is a time sequence, an image-time cube might be the natural product for the program. Preferably the data product will include error estimates and a data quality mask.

3.2 Spectroscopy

Spectroscopic data products can be in the form of a 2D spectrum, an extracted spectrum, or a spectroscopic cube. The minimum calibration for spectroscopy products must include wavelength calibration. Spectrophotometric flux calibration is considered added value. Preferably the data product should include error estimates and a data quality mask.

3.3 Other products

Advanced data products can include photometry tables or catalogs. At this time, we do not impose any restrictions as long as the product can be packaged into a Multi-extension FITS file. The metadata (headers) for such products need to follow the requirements defined in Section 6 to ensure that the files are associated with the correct program and data.

4. Filenames

The archive requires that filenames are not duplicated, since a file uploaded with a filename already existing in the archive will be recognized as a new version of the one already present, not as a different product. For this reason, **we require the following format for the data product filenames**:

```
GEMPRGID_LLPChosenName.fits
```

For example:

```
GN-2018B-LP-1_Object1stack.fits
```

This format prevents name collision across programs. It is the responsibility of the program to ensure that there are no filename collisions **within** their program's data products. Within a LLP program, target name, date of observation, filter, and such are examples of differentiating characteristics that could be added to the filename to ensure uniqueness.

5. File Format

The required file format is a Multi-Extension FITS file (MEF) that complies with the FITS standard (https://fits.gsfc.nasa.gov/fits_standard.html). The MEF can contain pixel data or binary table data.

6. Metadata (aka Headers)

The easiest way to ensure that all the required header keywords are in the data product MEF file is to simply propagate them through to the final products. The Gemini raw data have everything that is needed by the archive. If headers are carried along, the data product will have all that is required to make it searchable in the archive. Teams not using Gemini data reduction software need to make sure that the raw data headers are copied over to the final product.

6.1 Highly critical header keywords

The following header keywords, located in the PHU of the Gemini MEF data **must remain intact**:

- GEMPRGID
- INSTRUME
- TELESCOP
- OBSERVAT

There is no leeway here, those header keywords must be found, unchanged, in the PHU of the data product regardless of the type of data product.

6.2 Important header keywords for searchability

The following header keywords **must be present**, although the content can be modified to match the processing:

- DATE-OBS
- OBJECT
- RA and DEC

- Keywords associated with filter, disperser, central wavelength in the format used by the instrument.

DATE-OBS is a date-time string that allows an archive search in a date range. In the case of a stack, the investigators can decide the most meaningful value for that keyword. For example, one might decide to use the mid-point of a series of observation, the mid-point of the survey, etc.

In the case of a common, already known target, OBJECT should preferably match a SIMBAD alias to facilitate the search.

RA and DEC of the target should be included for searchability. These are added value if they are adjusted during processing to produce more accurate astrometry.

Each Gemini instrument uses various keywords to record the instrument configuration, and characteristics like the filter band, the disperser (grism, prism), pixel scale, spectral resolution, the

central wavelength and such. This information about the instrument configuration should be preserved in the final data products for increased searchability and increased science value.

6.3 Important header keywords for science value

The following header keywords **must be present** when available to ensure the scientific value of the product:

- AIRMASS
- EXPTIME
- OBSID

- WCS keywords
- Flux calibration information, e.g. zero point.
- Provenance information

In the case of a stack, the investigator can decide what value for AIRMASS is the most meaningful scientifically. The normal assumption would be that it refers to the average airmass of the individual observations.

The EXPTIME of a stack must reflect the total exposure time, or at least the exposure time that matches the zero point if the data has been flux calibrated.

OBSID is the observation ID of a sequence of exposures. If a stack is created from a single OBSID, then it should be included in the final product. In some cases, more than one observation sequence is used to create a final stack. In such case, OBSID has limited benefits. See the discussion on provenance below.

The World Coordinate System (WCS) information for the processed data is scientifically very valuable, sometimes critical. For an image, the WCS gives the RA and Dec of each pixel, and contains the pixel scale. For a 1D spectrum, the WCS is the wavelength calibration.

In the case of imaging, a flux calibrated image is added value. Header of flux calibrated images must contain the information necessary to convert the counts in the image to an accurate magnitude. This can be stored as a zeropoint keyword, for example.

The provenance relates to what specific data went into the production of the processed data and how they were used. At a minimum, we recommend including in the header a list of the observations (raw filenames, data labels, etc.) that went into a stack.

7. Documentation

7.1 Purpose and Content

The team benefits from the uploaded data products by having their work more easily available publicly and hence citable. However, the data products are uploaded to the archive for the benefits of “other” people too, people who are not on the LLP team. Therefore, it is important to provide as much information as possible about the data product and how it was created.

If the data product is associated with a published paper, the investigators should add the reference to the product as a header keyword.

Details on the reduction should be provided in documents alongside the data products. Such additional information can include the reduction scripts used to create the final products. The

accepted formats are ASCII text files or PDF files. It could be a description of some algorithms, or special software used. If the software is available publicly, include a URL for it and state the version used. Going back to the provenance issue discussed at the end of Section 6.3, maybe here is a good place to detail which datasets were used. A discussion of the systematics or remaining artifacts are other examples of scientifically valuable information for third-party users.

There are no specific rules yet, other than “Make the user of your data products confident about them.” The value-added is when the products can be used confidently for science.

7.2 Distribution

The documentation will be uploaded to the archive as a tar file. The Gemini program ID (GEMPRGID) will be associated with the tar file, allowing people to find the documentation alongside the data products.

8. Data Ingestion

At this time, there is no automated data transfer system set up. Instead, we request that the LLP investigators set up their own data transfer location and notify Gemini when a new group of data products is ready for ingestion. Gemini will then download the products and proceed to the archive upload.

9. Detailed Revision History

v1.0 28 February, 2018 Kathleen Labrie

Initial revision. Distributed with 2018B Call for Proposal.