



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

VERY LARGE TELESCOPE

CalSelector: Use Cases and User Requirements

Doc. No.: GEN-SPE-ESO-33000-5192

Issue: 2

Date: 24.01.2014

Author:	Martino Romaniello	<u>24.01.14</u>	<u>Martino Romaniello</u>
Approver:	Martino Romaniello	<u>24.01.14</u>	<u>Martino Romaniello</u>
Releaser:	Michael Sterzik	<u>27. Jan. 2014</u>	<u>Michael Sterzik</u>
	Name	Date	Signature



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 2 of 37

CHANGE RECORD

ISSUE	DATE	SECTION/PARA. AFFECTED	REASON/INITIATION DOCUMENTS/REMARKS
0.1	18.01.2011	All	First draft, distributed to the CSARs, cc to SZA.
0.2	07.02.2011	All; added Use Case 8, section 7, Appendices B and C.	First round of comments from CSARs incorporated. Back to the CSARs for final comments.
0.3	11.02.2011	Minor changes and corrections of typos; MKG added to project Team.	Distributed to the Project Team, cc to the CSARs.
1	18.04.2011	Clarification of propagation of certification from masters to raw parents; typos. Comments from Comerón. First official release.	
2 draft 1	09.01.2014	Mainly the sections on Use Cases 2 and 3. Several minor changes throughout. Status and priority of all use cases updated and revised.	New Issue for implementation of CalSelector 2: Use Case 2 (raw-to-masters) and decommissioning of the static channel for Use Case 3 (raw-to-raw).
2 draft 2	14.01.2014	Use Case 2 (raw-to masters)	Added requirement to trigger raw-to-raw if association is empty or incomplete.
2 draft 3	15.01.2014	Use Cases 2 and 3.	Added requirement on "type" tag in xml file. Need to resolve calibrations of calibrations removed from Use Case 2.



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 3 of 37

2 draft 4	21.01.2014	Use Cases 2 and 3, Appendix B.	Feedback from Vincenzo Forchì and Christophe Moins. Appendix B updated w/ 2013 stats.
2	24.01.2014	Minor cosmetic changes.	Official release.



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 4 of 37

TABLE OF CONTENTS

1. Scope.....	6
2. Organization of the project.....	6
3. List of Abbreviations & Acronyms	7
4. List of Applicable and Referenced Documents	8
4.1. Applicable Documents	8
4.2. Referenced Documents	8
5. Preamble: science data, calibrations and associations	9
6. Use Cases and corresponding User Requirements	10
Use Case 1 : Proprietary PI Packs [Decommissioned]	10
Use Case 2 : From raw science to master calibrations [Priority 1]	11
Breakdown of the use case	11
User requirements	12
Miscellaneous considerations	15
Use Case 3 : From raw science to raw calibrations [Implemented]	16
Breakdown of the use case	16
User requirements	17
Miscellaneous considerations	20
Use Case 4 : From raw science to reduced science and raw and master calibrations [Priority 2]	22
Breakdown of the use case	22
User requirements	23
Miscellaneous considerations	24
Use Case 5 : Non-proprietary PI Packs [Obsolete]	25
Breakdown of the use case	25
User requirements	25
Miscellaneous considerations	26
Use Case 6 : From reduced science to raw science and associations [Priority 3]	26
Breakdown of the use case	26
User requirements	27
Miscellaneous considerations	28
Use Case 7 : From raw file to NightLog excerpts [Obsolete]	29
Breakdown of the use case	29
User requirements	29
Miscellaneous considerations	30
Use Case 8 : From raw or master calibrations to raw or reduced science [Priority 4]	30
Breakdown of the use case	30
User requirements	31
Miscellaneous considerations	32
7. Final miscellaneous considerations	33
Appendix A : Mind Map representation of the use cases.....	35



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 5 of 37

Appendix B : Statistics on CalSelector usage..... 36
Appendix C : Overview of currently available association information..... 37

	CalSelector: Use Cases and User Requirements	Doc: GEN-SPE-ESO-33000-5192 Issue 2 Date 24.01.2014 Page 6 of 37
--	---	---

1. Scope

Issue 2 of this document is in support of the development of version 2 of the software. The main changes with respect to Issue 1 [RD04] are in support of the implementation of Use Case 2 (raw science to master calibrations), which is now set to highest priority, and of the decommissioning of the static channel in the implemented Use Case 3 (raw science to raw calibrations). Other sections underwent minor editing, as well, but were not thoroughly checked for overall consistency. Care should be taken to do so before implementing the corresponding use cases.

This document details the use cases and resulting user requirements for the CalSelector: a set of tools that, given the result of a query by a user of ESO's Science Archive Facility, associate and return relevant files in addition to those directly selected with the query. The scope of CalSelector is focused on data from the Paranal and, possibly, La Silla observatories. For ALMA data the functionalities are provided as part of the corresponding archive¹ and for APEX data association is provided by the APEX staff as part of the end-to-end operational model².

For each use case there is a section on "miscellaneous considerations" with ideas and thoughts that do not strictly belong to a requirements document: they are intended to expand on and clarify some aspects of the use case and/or of the related requirements.

The use cases listed here have an associated priority ranking. In most cases all requirements within a use case are equally important. Occasionally, though, the individual requirements within a use case are prioritized as well, according to the following convention: A Critical, B Important and C Desired.

The specification of detailed requirements for the archive web interfaces that may be needed to support users to request and download associations of files is outside the scope of the present document.

Several early discussions with Stefano Zampieri and Reinhard Hanuschik were instrumental in shaping v0.1 document and are hereby gratefully acknowledged. The input of the CSARs and of the Project Team was fundamental to finalize the requirements.

Discussions with Vincenzo Forchì, Wolfram Freudling, Reinhard Hanuschik and Christophe Moins while preparing Issue 2 are gratefully acknowledged.

2. Organization of the project

¹ Paola Andreani, private communication.

² Carlos De Breuck, private communication.



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 7 of 37

The 2011 core development Project Team, which implemented Use Case 3, was formed by: Adam Dobrzycki (DMO/ACH), Vincenzo Forchì (DoE/DFI), Maurice Klein Gebbinck (DoE/SED), Reinhard Hanuschik (DMO/QCG), Alberto Micol (DMO/ASG), Christophe Moins (DoE/DFI), Martino Romaniello (Project Scientist, DMO/BOD), Ignacio Vera Sequeiros (DMO/ACH), Stefano Zampieri (Project Manager, DoE/DFI).
2014 Team XXX.

Scientific advice to the Project Team was provided in 2011 by the CalSelector Astronomy Reviewers (CSARs): Magda Arnaboldi (DMO/ASG), Reinhard Hanuschik (DMO/QCG), Vincenzo Mainieri (DMO/USD), Martino Romaniello (Chair, DMO/BOD), Jeremy Walsh (DMO).

3. List of Abbreviations & Acronyms

This document employs several abbreviations and acronyms to refer concisely to an item, after it has been introduced. The following list is aimed to help the reader in recalling the extended meaning of each short expression. Organizational units are in italics; acronyms as they appear in the ERP system are used for them. Instrument and telescope names are not reported here, even when they are acronyms.

AB	Association Block
<i>ACH</i>	Archive Content Handling group, belongs to BOD
<i>ASG</i>	Archive Science group, belongs to BOD
<i>BOD</i>	Back-end Operations Department, belongs to DMO
co-I	co-investigator
CSARs	CalSelector Astronomy Reviewers
<i>DFI</i>	Data Flow Infrastructure department, belongs to DoE
<i>DoE</i>	Directorate of Engineering
<i>DMO</i>	Data Management and Operations division, belongs to the Directorate of Operations
FITS	Flexible Image Transport System
gNLT	Garching NightLog Tool
NGRH	Next Generation Request Handler
OB	Observing Block
OCA	Organization, Classification and Association framework
PI	Principal Investigator

	CalSelector: Use Cases and User Requirements	Doc: GEN-SPE-ESO-33000-5192
		Issue 2
		Date 24.01.2014
		Page 8 of 37

PSO	Paranal Science Operations
QC	Quality Control
QCG	Data Processing and Quality Control group, belongs to BOD; it is usually referred to simply as QC, which generates some confusion with the process of the same name
Reflex	Recipe Flexible Execution workbench
RRM	Rapid Response Mode
SAF	Science Archive Facility (entry point at http://archive.eso.org)
SED	System Engineering Department, belongs to DoE
ToO	Target of Opportunity
UP	User Portal

4. List of Applicable and Referenced Documents

4.1. Applicable Documents

Ref	Title	Author	Doc number

4.2. Referenced Documents

Ref	Title	Author	Doc number
RD01	Data and metadata access control requirements	N. Delmotte	GEN-SPE-ESO-32000-0046, Issue 1
RD02	ESO Data Access Policy Implementation Guidelines	F. Comerón	dmo/memo/2009-027
RD03	Requirements for the request and access control for internal and external data products from the archive	N. Delmotte	dmo/memo/2010-079
RD04	Provenance "White Paper"	N. Delmotte	dmo/memo/2009-026
RD04	CalSelector: Use Cases and User Requirements	M. Romaniello	GEN-SPE-ESO-33000-5192, Issue 1



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 9 of 37

5. Preamble: science data, calibrations and associations

In this section we give a brief introduction to the nature of and motivation for the need of data association.

Most science data as generated by telescopes and instruments are generally useless for scientific analysis until they are processed for the dual purpose of removing the signatures imprinted on the raw data by the telescope/instrument apparatus and by the earth's atmosphere and of converting the instrumental signal into physical units. Dedicated data ("calibration frames" or "calibrations") are taken for this purpose. Both the observational apparatus and the earth's atmosphere change in time, so calibrations need to be taken repeatedly to properly calibrate the science data. In addition, calibrations may also include static files for properties that do not change with time, e.g. the wavelengths of spectral lines in the comparison spectra ("arcs") used to calibrate in wavelength the science spectra. Some calibrations are taken during the night, either because direct observation of celestial objects are needed, e.g. flux standard stars, or to reproduce the exact system setup, e.g. the position of the instrument and telescope. Some others are acquired during the day by means of internal calibration units. Dedicated calibration plans are in place for all ESO instruments to generate calibrations appropriate both to process and calibrate science data, as well as for instrument monitoring. The use of calibrations to characterize the instrument performance and monitor its temporal evolution is outside of the scope of this document and will not be further considered here.

The process leading to calibrated science data, aka "data reduction", is a cascade where at each step raw calibrations are themselves calibrated until they can be meaningfully applied to the science frame(s). For example, in order to reduce an optical CCD image, a master bias frame needs to be created first from raw bias frames. Together with raw flat-field frames, the master bias then enters the creation of a master flat-field frame. Finally, both master bias and master flat are applied to the raw science frame.

Associating the data that need to be processed together, then, is the very first, obligatory step to data reduction. Conversely, starting from a calibrated science files, it is often useful to reconstruct the set of files used to generate it, e.g. to be able to check or even repeat the data reduction itself. Broadly speaking, these two are the main use cases for the CalSelector. Strictly speaking, then, CalSelector is actually a misnomer and "File Associator" would be a more appropriate name. However, for historical reasons, the former term is kept³.

The remainder of this document is devoted to describing the detailed use cases and corresponding user requirements for the CalSelector tool. At a high level, the workflow for most of the use cases is very similar. Archive users select a certain type

³ Plus, it turns out that "associator" is not an actual word in the English language. ;-)



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 10 of 37

of files from a query interface. Together with the results of the search they are, then, presented with the possibility of relevant files to be added to their original selection. Of course, the precise nature of these additional files varies among the different use cases. The user can, then, request and retrieve the entire dataset. A prime example of such an association is constituted by the (minimal) set of calibration files needed to reduce a set of raw science frames. This and the other use cases considered for implementation are detailed below.

6. Use Cases and corresponding User Requirements

In this section we list the use cases and the corresponding user requirements. The numbering does not reflect relative priorities among the use cases. It does, however, follow the one in the mind map in Appendix A, which makes the associations among the different file types apparent at a glance.

In an attempt to make the description of each use case self-contained and independent from the other ones, similar concepts are repeated several times throughout the document.

The priority ranking of the use cases is as follows:

- Use Case 1: : Proprietary PI Packs [Decommissioned]
- Use Case 2: From raw science to master calibrations [Priority 1]
- Use Case 3: From raw science to raw calibrations [Implemented]
- Use Case 4: From raw science to reduced science and raw and master calibrations [Priority 2]
- Use Case 5: Non-proprietary PI Packs [Obsolete]
- Use Case 6: From reduced science to raw science and associations [Priority 3]
- Use Case 7: From raw file to NightLog excerpts [Obsolete]. Relevant NightLog excerpts are now served directly by the NGRH.
- Use Case 8: From raw or master calibrations to raw or reduced science [Priority 4]

Use Case 1: Proprietary PI Packs [Decommissioned]

This use case allows PIs to access the data packs (aka PI Packs) containing raw data (science and associated calibrations), data products (for pipeline supported observing modes) and ancillary information (QC reports, NightLogs, etc.) as created incrementally by QCG as data were acquired and processed.



CalSelector: Use Cases and User Requirements

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 11 of 37

Use Case 1 was operational between June 29th, 2009 and September 2nd, 2013, when it was decommissioned. We still mention it here mainly for completeness. PI Packs were available from the archive for Service Mode observing programmes active as of April 1st, 2009 (ESO Observing Period 83). PI Packs for Visitor Mode observing programmes are available as of October 1st, 2009 (ESO Observing Period 84). Generation of new PI Packs was discontinued in November 2011 as the result of adopting a new operational scheme for QCG (aka QC XXLight). Access to historical PI Packs was discontinued on September 2nd, 2013.

By registering through the User Portal, PIs could access the data while still proprietary, which are packaged as tarballs before they are made available for the user to download. Users could retrieve temporal slices by specifying a time interval in which the data was acquired. Full details on the implementation can be found at (ESO internal access only):

<https://websqa.hq.eso.org/sdd/bin/view/BET/CalSelectorForPIPacks>

Use Case 2: From raw science to master calibrations [Priority 1]

This use case deals with associating to raw science files the quality certified master calibrations generated by QCG and ingested in the archive as part of the daily workflow⁴. In this context, a master calibration file is intended as any input file to the final science recipe in the calibration cascade other than the raw files that trigger the recipe itself. The use case is, thus, intended for users who want to take advantage of the added value of the certified master calibrations available in the archive and apply them to the raw science data for the final step of the data reduction. In this respect, it is worth noticing that master calibrations carry an imprint that is generally specific to the data reduction tools used to generate them and are likely to be useful only in that context. The fact that a given master calibration was certified for quality does not imply that all of the raw calibrations that went into it are also of certified quality (for example, a saturated frame will be excluded by a median filter provided that enough good ones are used). It does, however, imply that they can be processed into a certified master, which is an important aspect of the “added value” of the use case.

Breakdown of the use case

The breakdown of the use case is as follows:

UC 2.1 An archive user searches raw science observations from an archive query interface. According to general policy, logging into the User Portal is not required at this stage (browsing the archive content).

UC 2.2 Search results are returned in the result page.

⁴ Master calibrations were sparsely ingested in the archive for the years 2000-2003 and are complete from 2003.



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 12 of 37

- UC 2.3** The user selects a set of results and activates the option “Find Master Calibrations”.
- UC 2.4** Associations are formed and selected information on them is returned to the user (see UR 2.2 for the details).
- UC 2.5** The user selects the datasets to be retrieved, which will include the raw science files and the associated ones.
- UC 2.6** The user submits a request for those datasets. According to ESO policy, the user must be logged into the user Portal to be able to successfully complete the request.
- UC 2.7** The request is processed and the requested data are made available to the users subject to proprietary protection. In this respect, only associations linked to science frames the user is actually entitled to retrieve are made available to them.

User requirements

The corresponding user requirements are as follows:

- UR 2.1** Associations shall be formed dynamically by evaluating a set of OCA rules against the current content of the archive⁵.
- UR 2.2** The following items shall be delivered to the user:
- a. The raw science frames that resulted from the archive query in UC 2.1, grouped as per the criteria in the association rules⁶.
 - b. For each group of files in UR 2.2a, the complete minimum set of quality certified master calibrations and static calibrations needed to fully process the science frame.
 - c. For each group of files in UR 2.2a, the excerpts from the NightLog relevant to them. The content and format of the files currently served by the operational NGRH shall be used⁷. They include the overall QC0 grade, the individual Phase 2 constraints and whether or not each was met, text comments from the observer and the actual conditions under which the file was acquired (e.g. image quality, sky transparency, moon distance, etc.).
 - d. The relations among the associated files, including whether the association was complete, i.e. if enough calibrations were identified to process the raw

⁵ Experience with CalSelector version 1.x, which implemented both a dynamic as and a static channel, showed that this latter is not desirable in that it raised operational issues if changes need to be applied to cope, for example, with associated files that, a posteriori, turn out to be faulty.

⁶ The primary example of grouping among science files is that of files observed in a sequence to be processed together, e.g. an imaging jitter or a spectroscopic nodding sequence.

⁷ For reference, the current version of the NGRH is 1.6.5.



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 13 of 37

science frame. In particular, raw science frames for which an incomplete set of calibrations was found should be highlighted to the user⁸.

The content and format of the file encapsulating the relations among the files themselves shall be based on the (xml) files currently served by the operational CalSelector⁹. A new requirement is that the association shall be tagged at its root node with its type, i.e. in this case with “type=raw2master”.

UR 2.3 The list of associated files shall be purged from those exclusively linked to science frames the user is not actually entitled to retrieve.

UR 2.4 The list of associated files shall be purged from duplicates so that each file is only served once to the archive user.

- a. With reference to UR 2.6b, if a duplicate file belongs to different types of associations, i.e. raw-to-masters vs. raw-to-raw, and, hence, is to be saved in to two different subdirectories on the user’s disk¹⁰, then the one in the “raw2raw” directory shall be a symbolic link to the actual file in the “raw2master” directory.

In practice, the possibility for a file to belong to different types of associations is limited to static calibrations, which in general represent a small volume of a typical request. It is, then, also acceptable to simply transfer these files twice, one in the “raw2raw” and once in the “raw2master” directory.

UR 2.5 The items in UR 2.2a (raw science) and UR 2.2b (master and static calibrations) are FITS files.

- a. The connection between these files shall be solely based on matching combinations of FITS header keywords according to rules.
- b. These rules shall be configurable.
- c. All configurations shall be conveniently accessible and editable by the operational groups in charge¹¹.

⁸ The association will only return results for pipeline-supported modes, for which master calibrations could be generated by processing raw calibrations (see <http://www.eso.org/pipelines>). Otherwise, in principle all raw science frames stored in the archive should be calibratable, so incomplete calibration sets for pipeline-supported modes only occur in extreme cases, e.g.. if science data are graded 'C' calibrations need not be obtained (usually are, night-time calibrations usually are not); non-standard modes (e.g. 1x1 binning data taken with FORS receive no twilight flats).

⁹ For reference, the current version of the CalSelector is 1.2.3.

¹⁰ The rationale here being that all the files to be processed together are present in the same directory (this is how Reflex expects them, for example).

¹¹ For example, by providing the ability to graphically visualize the hierarchical structure intrinsic to



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 14 of 37

- d. It shall be possible to define break points in time when the definition of a complete set of calibration changes¹². The history of such break points shall be preserved so that the appropriate definition can be applied to files taken at different times in the life of an instrument¹³ (parameterization of association rules, with the parameters being the beginning and end of the rule validity period).

UR 2.6 For the datasets in the request for which the raw-to-masters association is empty or incomplete a raw-to-raw (Use Case 3) association shall be triggered, and, even if empty or incomplete, its results served to the user¹⁴.

The following items have a direct impact on the NGRH:

- a. The resulting association shall be graphically highlighted in the CalSelector result page by, e.g., a specific background colour to the corresponding dataset row(s). The choice of background colours shall avoid those commonly associated with a specific meaning, e.g. red for an error condition, or green for OK.
- b. The datasets resulting from the two flavours of associations, i.e. raw-to-master and raw-to raw, shall be saved in two distinct subdirectories of the (user-chosen) download directory¹⁵. The names of these two subdirectories do not need to be chosen by the user (they could, for example, be hardcoded to “raw2master” and “raw2raw”, respectively).

UR 2.7 Access to proprietary data is regulated by the provisions in RD01 and RD02.

UR 2.8 The performance of the “Find Master Calibrations” function (UC 2.3) shall allow for a responsive interactive user experience, as it is currently the case for the implemented raw-to-raw case (Use Case 3).

association rules.

¹² A break point is when the association rules change, not when the instrument itself changes due to, e.g. an earthquake or a hardware intervention, without affecting the rules themselves. Examples of break points include the addition or removal of a given type of calibrations and the addition or removal of header keywords.

¹³ Operationally, it may not be possible to reconstruct the complete history of such break points for each instrument mode. This is unfortunate, but acceptable, and should be documented on a case-by-case basis.

¹⁴ The rationale is twofold: privilege completeness over processing level (raw files are bound to be as complete or more than masters) and uniformity over processing level (a dataset including both raw calibrations and masters is less convenient to process than one only containing only one of the two flavors).

¹⁵ This provides an easy way for the user to have a permanent record of the different associations and provides a clean input to Reflex for actual processing.

	CalSelector: Use Cases and User Requirements	Doc: GEN-SPE-ESO-33000-5192 Issue 2 Date 24.01.2014 Page 15 of 37
--	---	--

UR 2.9 Authentication through the User Portal shall be possible as required by general policies (see also UC 2.1 and UC 2.6).

Miscellaneous considerations

A functionality related to Use Case 2 is the possibility to browse the archive directly for master calibrations, without involving science files. This does not imply any association among files and is, thus, not considered here.

Since QCG can only process data if a pipeline exists, master calibrations only exist for pipeline-supported modes. The list and status of the ESO pipelines can be found at: <http://www.eso.org/pipelines>.

In ground-based astronomy the process of calibrating the data is for the most part a calibration of the earth's atmosphere at the time when the observations were carried out. As such, and at variance with space-borne astronomy, calibrations do not usually improve with time and selecting the ones closest in time to the observations generally leads to the best results. It is, of course, possible that the choice of the best calibrations for a specific science file needs to be revised long after it was acquired, but these are by and large uncommon exceptions. All in all, calibrations are typically acquired within hours, days at most, from the science observations.

Use Case 2 is *de facto* a quasi-static problem, in that the connections among the different files can be established once, relatively close in time to the data acquisition, and subsequent updates are infrequent. From this point of view, then, there is no strict need to freshly form the required associations among files dynamically every time at submission of the archive request. Rather, it is conceivable that, once formed, the associations are stored in a suitable structure and read from there at the time of the submission of the archive request. This may be a particularly attractive option should performance issues linked to UR 2.8 arise. Having said this, experience with the implemented raw-to-raw case (Use Case 3) shows that performance is not really an issue (the bottleneck is, in fact, the interaction between the NGRH and CalSelector, not in putting together the actual association). Furthermore, the static channel as implemented for Use Case 3 has shown some operational issues if changes do need to be applied to cope, for example, with associated files that, a posteriori, turn out to be faulty. A purely dynamical implementation of CalSelector is, then, deemed desirable.

At the heart of Use Case 2 there is the matching of FITS header keywords (UR 2.5a). It is worth noticing here that ESO already has a very powerful mechanism to match FITS header keywords: the OCA (Organization, Classification and Association) framework¹⁶. It is used, e.g., to prepare data for processing, both in operational environments and in the offline environment Reflex^{17,18}, to set transfer priorities

¹⁶ More on OCA in the project pages at: <https://websqa.hq.eso.org/sdd/bin/view/BET/OCA>.

¹⁷ More on Reflex in the public pages at: <http://www.eso.org/reflex>.



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 16 of 37

based on file type in the Data Transfer System and is also part of the Phase 3 Infrastructure¹⁹ to validate the metadata items that describe the data itself. Reuse to the extent possible of the OCA framework within CalSelector would allow for considerable reuse of existing development and operational expertise.

As noted in the introduction to the use case, master calibrations are directly useful for further processing only if the same data reduction system is utilized. Of more general interest is the fact that they underwent certification in QCG, which extends also the raw files that went into generating them in the sense that it is proven that they can be processed into a meaningful master calibration file.

In order to facilitate user processing of the raw-to-master datasets defined here, dedicated Reflex workflows are being developed.

Use Case 3: From raw science to raw calibrations [Implemented]

This use case deals with associating raw calibrations to raw science files. It is intended for users who want to reduce the data themselves all the way from the very beginning, e.g. with Reflex.

Breakdown of the use case

The breakdown of the use case is as follows:

- UC 3.1** An archive user searches raw science observations from an archive query interface. According to general policy, logging into the User Portal is not required at this stage (browsing the archive content).
- UC 3.2** Search results are returned in the result page.
- UC 3.3** Optionally, the user uploads the set of rules to be used to generate the associations.
- UC 3.4** The user selects a set of results and activates the option “Find Raw Calibrations”.
- UC 3.5** Associations are formed and selected information on them is returned to the user (see UR 3.2 for the details).
- UC 3.6** The user selects the datasets to be retrieved, which will include the raw science files and the associated ones.

¹⁸ Specifically, the DataOrganizer actor in Reflex implements raw-to-raw associations, just like Use Case 3. The most obvious difference is that the DataOrganizer in Reflex works on files stored on the local disk, whereas the CalSelector would have to work against the whole content of the archive.

¹⁹ More on the Phase 3 Infrastructure in the project pages at: <https://nagios-vip/twiki/bin/view/EDP/Phase3Project>.



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 17 of 37

UC 3.7 The user submits a request for those datasets. According to ESO policy, the user must be logged into the user Portal to be able to successfully complete the request.

UC 3.8 The request is processed and the requested data are made available to the users subject to proprietary protection. In this respect, only associations linked to science frames the user is actually entitled to retrieve are made available to them.

User requirements

The corresponding user requirements are as follows:

UR 3.1 Associations shall be formed dynamically by evaluating a set of OCA rules against the current content of the archive²⁰.

UR 3.2 The following items shall be associated and delivered to the user:

- a. The raw science frames that resulted from the archive query in UC 3.1, grouped as per the criteria in the association rules²¹.
- b. For each group of files in UR 3.2a, the complete set of raw and static calibrations needed to fully process it. Identifying this complete set may require an iterative process to resolve all the dependencies among files (e.g. identify the “calibrations of calibrations”, i.e. calibration files need to process intermediate calibrations, but not the science itself²²).

The introduction of a database quality flag at the level of individual raw files merits a specific (short) discussion here²³. In short, it can have two values: “quality unknown” and “quality certified”. It defaults to “quality unknown” and is set to “quality certified” as part of the QCG daily workflow for files that pass the certification²⁴. Faulty frames are hidden via a separate

²⁰ Experience with CalSelector version 1.x, which implemented both a dynamic as well as a static channel, showed that this latter is not desirable in that it raised operational issues if changes need to be applied to cope, for example, with associated files that, a posteriori, turn out to be faulty.

²¹ The primary example of grouping among science files is that of files observed in a sequence to be processed together, e.g. an imaging jitter or a spectroscopic nodding sequence.

²² Examples include flux standards, which may require different calibration files than the science frames because of different instrumental setups, or master dark frames, which may be very distant in time from the science frames, and thus need to be calibrated with their own specific set of raw biases.

²³ At the time of writing, the presence of this flag is actually not yet not implemented. It is scheduled to be introduced with the upgraded version of the cdbingest tool, which then becomes a prerequisite for implementing this requirement.

²⁴ Since certification is not done directly on the individual files, but on the master resulting from their processing, extending it to the individual raw files has to be understood in the following way: individual raw frames can potentially be faulty, but result in a meaningful product when processed together. Example: an individual flat exposure is saturated, but still produces a meaningful master flat when median filtered with other non saturated ones.

	CalSelector: Use Cases and User Requirements	Doc: GEN-SPE-ESO-33000-5192 Issue 2 Date 24.01.2014 Page 18 of 37
--	---	--

mechanism, so that no archive service, including CalSelector, even considers them.

When dealing with this quality flag, the requirement on CalSelector is as follows. The tool shall first attempt to form a complete association by only using quality certified files, as identified by the corresponding database flag. If this results in a complete association, then the association itself shall be used. If not, then also non-certified files shall be considered (but never the hidden files). Ideally one would want to use non-certified files only for those blocks in which no certified files could be found, or, in other words, a certified association is preferable to a non certified one even if the latter is closer in time to the file it calibrates. In practice the difference is limited to rare exotic cases, so, in this second attempt, it is acceptable to directly consider all files, independently of their certified flag. The association formed by considering all values of the certification flag shall be served to users even if the association itself is incomplete.

- c. For each group of files in UR 3.2a and UR 3.2b, the excerpts from the NightLog relevant to them. The content and format of the files currently served by the operational NGRH shall be used²⁵. They include the overall QC0 grade, the individual Phase 2 constraints and whether or not each was met, text comments from the observer and the actual conditions under which the file was acquired (e.g. image quality, sky transparency, moon distance, etc.).
- d. The relations among the associated files, including whether the association was complete, i.e. if enough calibrations were identified to process the raw science frame. In particular, raw science frames for which an incomplete set of calibrations was found should be highlighted to the user²⁶.

The content and format of the file encapsulating the relations among the files themselves shall be based on the (xml) files currently served by the operational CalSelector²⁷. New requirements on the content of the xml file are:

- Each block in the association cascade shall now be flagged as containing certified or uncertified files, as opposed to the present case in which only the root one is (in the current CalSelector certification information is not available at the individual file level). The assumption here is that all files in the same block share the

²⁵ For reference, the current version of the NGRH is 1.6.5.

²⁶ In principle, all raw science frames stored in the archive should be calibratable, so incomplete calibration sets should be a rare occurrence.

²⁷ For reference, the current version of the CalSelector is 1.2.3.



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 19 of 37

same certification status. Departures from this imply an inconsistency between the OCA rules used to process the data and those used by CalSelector and it should be addressed at that level.

- The association shall be tagged at its root node with its type, i.e. in this case with “type=raw2raw”.

UR 3.3 The resulting associations shall be graphically highlighted in the CalSelector result page by, e.g., a specific background color to the corresponding dataset row(s), in accordance with UR 2.6a.

UR 3.4 If not satisfied with the default ones, users shall be allowed to upload the set of rules that describe the association. [Priority C]

- a. Users shall be allowed to download the default set of rules (to edit them at their leisure). [Priority C]

UR 3.5 The list of associated files in UR 3.1 shall be purged from those exclusively linked to science frames the user is not actually entitled to retrieve.

UR 3.6 The list of items in UR 3.1 shall be purged from duplicates so that each one is served only once to the archive user.

- a. If the raw-to-raw association was triggered because of an empty or incomplete raw-to-masters one (see UR 2.6), handling of duplicates shall comply with UR 2.4a.

UR 3.7 The items in UR 3.2a and UR 3.2b are FITS files.

- a. The connection between the files shall be solely based on matching combinations of FITS header keywords according to rules.
- b. These rules shall be configurable.
- c. All configurations shall be conveniently accessible and editable by the operational groups in charge²⁸.
- d. It shall be possible to define break points in time when the definition of a complete set of calibration changes²⁹. The history of such break points shall be preserved so that the appropriate definition can be applied to files taken at different times in the life of an instrument³⁰ (parameterization of

²⁸ For example, by providing the ability to graphically visualize the hierarchical structure intrinsic to association rules.

²⁹ A break point is when the association rules change, not when the instrument itself changes due to, e.g. an earthquake or a hardware intervention, without affecting the rules themselves. Examples of break points include the addition or removal of a given type of calibrations and the addition or removal of header keywords.

³⁰ Operationally, it may not be possible to reconstruct the complete history of such break points for



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 20 of 37

association rules, with the parameters being the beginning and end of the rule validity period).

UR 3.8 The performance of the “Find Raw Calibrations” (UC 3.4) function shall allow for a responsive interactive user experience, as it is currently the case with the operational version (1.2.3).

UR 3.9 Access to proprietary data is regulated by the provisions in RD01 and RD02.

UR 3.10 Authentication through the User Portal shall be possible as required by general policies (see also UC 3.1 and UC 3.7).

Miscellaneous considerations

UR 3.3 and its associated UC 3.3 were assigned a low priority. This is because of two considerations. Firstly, they would imply to expose the OCA syntax to the users for them to edit, which is likely to result in a niche usage. Secondly, users who need more or different calibrations than the standard set can conceivably assemble them separately, e.g. through the instrument specific query forms (<http://archive.eso.org/cms/eso-data/instrument-specific-query-forms>).

In ground-based astronomy the process of calibrating the data is for the most part a calibration of the earth’s atmosphere at the time when the observations were carried out. As such, and at variance with space-borne astronomy, calibrations do not usually improve with time and selecting the ones closest in time to the observations generally leads to the best results. It is, of course, possible that the choice of the best calibrations for a specific science file needs to be revised long after it was acquired, but these are by and large uncommon exceptions. All in all, calibrations are typically acquired within hours, days at most, from the science observations.

If UC 3.3 and UR 3.3 are neglected³¹, Use Case 3 is *de facto* a quasi-static problem, in that the connections among the different files can be established once, relatively close in time to the data acquisition, and subsequent updates are infrequent. From this point of view, then, there is no strict need to freshly form the required associations among files dynamically every time at submission of the archive request. Rather, it is conceivable that, once formed, the associations are stored in a suitable structure and read from there at the time of the submission of the archive request. This may be a particularly attractive options should performance issues linked to UR 3.8 arise.

each instrument mode. This is unfortunate, but acceptable, and should be documented on a case-by-case basis.

³¹ Neglecting UR 3.3 is, *de facto*, what the DataOrganizer in Reflex does. While it is theoretically possible for a user to edit the appropriate configuration files to change the default number of associated files of different types, this is rather cumbersome and very few people, if any, are expected to do so.

	CalSelector: Use Cases and User Requirements	Doc: GEN-SPE-ESO-33000-5192 Issue 2 Date 24.01.2014 Page 21 of 37
--	---	--

At the heart of Use Case 3 there is the matching of FITS header keywords (UR 3.7a). It is worth noticing here that ESO already has a very powerful mechanism to match FITS header keywords: the OCA (Organization, Classification and Association) framework³². It is used, e.g., to prepare data for processing, both in operational environments and in the offline environment Reflex^{33,34}, to set transfer priorities based on file type in the Data Transfer System and is also part of the Phase 3 Infrastructure³⁵ to validate the metadata items that describe the data itself. Reuse to the extent possible of the OCA framework within CalSelector would allow for considerable reuse of existing development and operational expertise.

The delivery to PIs of ToO and RRM data can be considered as a special case of Use Case 3. With data arriving into the archive in Garching in a matter of minutes after acquisition, the content of what is currently delivered to PIs by PSO in an ad-hoc, semi-manual way, namely the raw science frames and the raw calibrations needed to process them, would be naturally available to them by querying the archive and using the “Find Raw Calibrations” function. Two caveats are worth mentioning here. Firstly, the data delegation function in the User Portal needs to be operational so that PIs can have co-Is access the data if they so wish and RRM data can be shared among competing teams, as required by ESO policy. Secondly, it is essential for the scientific exploitation of ToOs and RRMs that the data is made available to the PIs in a matter of hours after acquisition.

Finally, it should be noted that QCG routinely generates associations among data as part of the daily workflow (see Appendix C). The connection between the raw science frames and the raw calibrations needed to process them that serves Use Case 3 is established in two steps: raw calibrations are processed into masters, which are then certified for quality and applied to raw science files to generate science data products. If this information is used for CalSelector, then from the point of view of CalSelector itself the associations are based on processing provenance information in the FITS headers because the rule-based connections in UR 3.7 above were already established by QCG.

Using the associations established by QCG to actually implement Use Case 3 presents some advantages, as well as limitations that are listed here for consideration. On the plus side, only raw calibrations that lead to quality certified masters are used, which would ensure that only raw calibrations that can be processed into meaningful masters are distributed; the break points of UR 3.7d would

³² More on OCA in the project pages at: <https://websqa.hq.eso.org/sdd/bin/view/BET/OCA>.

³³ More on Reflex in the public pages at: <http://www.eso.org/reflex>.

³⁴ Specifically, the DataOrganizer actor in Reflex implements raw-to-raw associations, just like Use Case 3. The most obvious difference is that the DataOrganizer in Reflex works on files stored on the local disk, whereas the CalSelector would have to work against the whole content of the archive.

³⁵ More on the Phase 3 Infrastructure in the project pages at: <https://nagios-vip/twiki/bin/view/EDP/Phase3Project>.



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 22 of 37

be taken care of automatically, since QCG always uses the appropriate rules at any given point in time; since CalSelector would not need access to the rules to create the associations (see UR 3.7) they would only have to exist, hence be maintained, only once, i.e. in the QCG environment. Current limitations include: the association is performed only for pipeline-supported modes³⁶, because it is driven by the mandate to actually process the data³⁷; the associations may not necessarily include the best possible calibrations, if these were not available at the time of processing; tying the distribution of ToO/RRM data to the QCG workflow would introduce a significant delay, because human intervention is needed to certify the master calibrations before the connection is made to the raw science frames. Also, at the time of writing the QCG-generated associations are not complete, especially for La Silla instruments (see Appendix C). They would initially need to be extended as far back as possible. From then on, the association formed daily would be used.

Use Case 4: From raw science to reduced science and raw and master calibrations [Priority 2]

This use case deals with associating calibrations and science data products to raw science files. It is intended for users who do not necessarily want to reduce the data themselves, hence they request the data products, but request to have the raw files at hand, e.g. to double check some aspects of the reduction.

Breakdown of the use case

The breakdown of the use case is as follows:

- UC 4.1** An archive user searches raw science observations from an archive query interface. According to general policy, logging into the User Portal is not required at this stage (browsing the archive content).
- UC 4.2** Search results are returned in the result page.
- UC 4.3** The user selects a set of results and activates the option “Find Raw Calibrations and Data Products”.
- UC 4.4** Associations are formed and selected information on them is returned to the user (see UR 4.6 for the details).
- UC 4.5** The user selects the datasets to be retrieved, which will include the raw science files and the associated ones.

³⁶ The list and status of the ESO pipelines can be found at: <http://www.eso.org/pipelines>.

³⁷ At variance with Use Case 3, the fact that the associations are available only for pipeline-supported modes is not a limitation for Use Case 2, since these are the only modes for which the masters exist in the first place



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 23 of 37

UC 4.6 The user submits a request for those datasets. According to ESO policy, the user must be logged into the user Portal to be able to successfully complete the request.

UC 4.7 The request is processed and the requested data are made available to the users subject to proprietary protection. In this respect, only associations linked to science frames the user is actually entitled to retrieve are made available to them.

User requirements

The corresponding user requirements are as follows:

UR 4.1 The following items shall be delivered to the user:

- a. The science raw frames that resulted from the archive query in UC 4.1.
- b. For each file in UR 4.1a, the data products obtained by reducing it.
- c. The complete set of master and static calibrations needed to generate the data products in UR 4.1b. Identifying this complete set may require an iterative process to resolve all the dependencies among files (e.g. identify the “calibrations of calibrations”, i.e. calibration files need to process intermediate calibrations, but not the science itself³⁸).
- d. The complete set of raw and static calibrations needed to generate the master calibrations in UR 4.1c.
- e. The complete set of ancillary files generated during the data reduction (e.g. weight maps, error files, etc.).
- f. For each raw file in UR 4.1a and UR 4.1d, the excerpts from the NightLog relevant to them: these include the overall QC0 grade, the individual Phase 2 constraints and whether or not each was met, text comments from the observer.
- g. If applicable, for each file in UR 4.1a and UR 4.1d, the actual conditions under which the file was acquired (e.g. image quality, sky transparency, moon distance, etc.).
- h. If available, QC1 parameters and related information (QC plots, etc.).
- i. The relations among the associated files. If applicable, the format shall be suitable to be imported into the Reflex DataOrganizer actor³⁹.

³⁸ Examples include flux standards, which may require different calibration files than the science frames because of different instrumental setups, or master dark frames, which may be very distant in time from the science frames, and thus need to be calibrated with their own specific set of raw biases.

³⁹ At the time of writing, the functionality of the Reflex DataOrganizer to import data associations is not implemented.



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 24 of 37

- UR 4.2** The list of associated files shall be purged from those exclusively linked to science frames the user is not actually entitled to retrieve.
- UR 4.3** The list of associated files shall be purged from duplicates so that each file is only served once to the archive user.
- UR 4.4** The items in UR 4.1a, UR 4.1b, UR 4.1c, UR 4.1d and UR 4.1e are FITS files.
- The connection among the raw science frames (UR 4.1a) and the science products (UR 4.1b) and, from those, to the master and raw calibration files (UR 4.1c and UR 4.1d, respectively) shall be based on processing provenance information (cf. RD04) as stored in their headers⁴⁰.
 - The connection among the science products (UR 4.1b) and the ancillary files (UR 4.1e) shall be based on organizational provenance information (cf. RD04) as stored in their headers.
- UR 4.5** The connection between the raw nighttime files (UR 4.1a and UR 4.1d) and the excerpts from the NightLogs (UR 4.1f) shall be based on the OB ID and on the time at which the OB was executed⁴¹. Both are FITS header keywords.
- UR 4.6** With respect to UC 4.4, after running the association the following information shall be provided to the archive user:
- The list of associated frames can potentially contain several hundred entries, which makes it likely confusing and, in fact, useless to list them all. Aggregate information (e.g. total number of frames, total size, etc.) should be displayed instead.
- UR 4.7** Access to proprietary data is regulated by the provisions in RD01, RD02 and RD03.
- UR 4.8** The performance of the “Find Raw Calibrations and Data Products” function shall allow for a responsive interactive user experience.
- UR 4.9** Authentication through the User Portal shall be possible as required by general policies (see also UC 4.1 and UC 4.6).

Miscellaneous considerations

This use case is in a way similar to Use Case 1 and Use Case 5, in that the content of the packages are identical in structure: raw science files to which the corresponding calibrations (raw and masters) and data products are associated. The

⁴⁰ For internal data products the processing provenance is encoded in the PRO.REC keywords stored in each product (as of the year 2003).

⁴¹ An OB can be executed multiple times until the observing constraints are met, hence the OB ID itself is not enough to uniquely identify the relevant NightLog.



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 25 of 37

differences arise in the way the raw science files are selected. In Use Case 1 and Use Case 5, all of the raw science files belong to the same run ID and the content of the package is static, as provided by the data provider (e.g. QCG for internal data products, an external user for external data products, etc.). For Use Case 4, on the other hand, the raw science frames are the result of an archive query by the archive user.

When it comes to the observing conditions, the difference between UR 4.1f and UR 4.1g is that the former is concerned with whether or not the Phase 2 constraints were met. It does not explicitly return the actual atmospheric conditions, which is the scope of the latter.

Use Case 5: Non-proprietary PI Packs [Obsolete]

The use case is now obsolete, because PI Packs are not generated any more as of November 1st, 2011. The following text is kept here just for the record.

This use case extends Use Case 1 to allow access to PI Packs to all archive users once the proprietary protections have duly expired.

Breakdown of the use case

The breakdown of the use case is as follows:

- UC 5.1** An archive user browses the run IDs for which the PI Packs are publicly available or searches for specific run IDs, with the possibility of filtering by, e.g. instrument, start and end date of the observations, etc. According to general policy, logging into the User Portal is not required at this stage (browsing the archive content).
- UC 5.2** When selecting a run ID, the package prepared for the original PI, or a slice thereof, is returned. According to ESO policy, the user must be logged into the User Portal to be able to successfully complete the request.

User requirements

The corresponding user requirements are as follows:

- UR 5.1** The content of each package is the same as in Use Case 1 above.
- UR 5.2** With respect to the query interface:
- a. It shall be possible to filter by instrument the list of publicly available run IDs.
 - b. It shall be possible to enter a list of run IDs. The use of wildcards shall be supported.
 - c. It shall be possible for the user to select multiple run IDs at once for data retrieval.



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 26 of 37

- d. It shall be possible to select a slice of a complete package by specifying a time interval for when the data was acquired (already implemented for Use Case 1).

UR 5.3 PI Packs generally contain a mixture of raw data and data products. Access to proprietary data is regulated by the provisions in RD01, RD02 and RD03.

UR 5.4 Authentication through the User Portal shall be possible as required by general policies (see also UC 5.1 and UC 5.2).

Miscellaneous considerations

As noted earlier, this use case is an extension of Use Case 1, which is already operational. What is essentially missing to fulfil Use Case 5 is to implement access control (UR 5.3) and a suitable query interface (UR 5.2).

Use Case 6: From reduced science to raw science and associations [Priority 3]

Here the archive user selects a set of science data products and requests the files that were used to generate them (raw science and calibrations, master calibrations, static calibrations, etc.), the ancillary files generated in the data processing (e.g. weight maps) and associate information (excerpts from NightLogs, etc.).

Breakdown of the use case

The breakdown of the use case is as follows:

UC 6.1 An archive user searches science data products (reduced files) from an archive query interface. According to general policy, logging into the User Portal is not required at this stage (browsing the archive content).

UC 6.2 Search results are returned in the result page.

UC 6.3 The user selects a set of results and activates the option “Find Associated Files”.

UC 6.4 Associations are formed and selected information on them is returned to the user (see UR 6.6 for the details).

UC 6.5 The user selects the datasets to be retrieved, which will include the science data products and the associated ones.

UC 6.6 The user submits a request for those datasets. According to ESO policy, the user must be logged into the user Portal to be able to successfully complete the request.

UC 6.7 The request is processed and the requested data are made available to the users subject to proprietary protection. In this respect, only associations linked to science frames the user is actually entitled to retrieve are made available to them.



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 27 of 37

User requirements

The corresponding user requirements are as follows:

UR 6.1 The following items shall be delivered to the user:

- a. The science data products that resulted from the archive query in UC 6.1.
- b. For each file in UR 6.1a, the complete set of raw science files needed to generate it.
- c. For each file in UR 6.1a, the master and static calibrations that were used to generate it together with the raw science files (UR 6.1b). Identifying this complete set may require an iterative process to resolve all the dependencies among files (e.g. identify the “calibrations of calibrations”, i.e. calibration files need to process intermediate calibrations, but not the science itself⁴²).
- d. The complete set of raw calibrations used to generate the master calibrations in UR 6.1c.
- e. The complete set of ancillary files generated during the data reduction (e.g. weight maps, error files, etc.).
- f. For each science frame in UR 6.1b and raw calibration in UR 6.1d, the excerpts from the NightLog relevant to them: these include the overall QC0 grade, the individual Phase 2 constraints and whether or not each was met, text comments from the observer.
- g. If applicable, for each science frame in UR 6.1b and raw calibration in UR 6.1d, the actual conditions under which the file was acquired (e.g. image quality, sky transparency, moon distance, etc.).
- h. If available, QC1 parameters and related information (QC plots, etc.).
- i. The relations among the associated files. If applicable, the format shall be suitable to be imported into the Reflex DataOrganizer actor⁴³.

UR 6.2 The list of associated files shall be purged from those exclusively linked to science frames the user is not actually entitled to retrieve.

UR 6.3 The list of associated files shall be purged from duplicates so that each file is only served once to the archive user.

⁴² Examples include flux standards, which may require different calibration files than the science frames because of different instrumental setups, or master dark frames, which may be very distant in time from the science frames, and thus need to be calibrated with their own specific set of raw biases.

⁴³ At the time of writing, the functionality of the Reflex DataOrganizer to import data associations is not implemented.



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 28 of 37

- UR 6.4** The items in UR 6.1a, UR 6.1b, UR 6.1c, UR 6.1d and UR 6.1e are FITS files.
- a. The connections between the science data products (UR 6.1a), raw science files (UR 6.1b), master and static calibrations (UR 6.1c) and raw calibrations (UR 6.1d) shall be based on processing provenance⁴⁴ (cf. RD04).
 - b. The connections between the science data products (UR 6.1a) and the ancillary files (UR 6.1e) shall be based on organizational provenance (cf. RD04).
- UR 6.5** The connection between the raw nighttime files and the excerpts from the NightLogs (UR 6.1f) shall be based on the OB ID and on the time at which the OB was executed⁴⁵. Both are FITS header keywords.
- UR 6.6** With respect to UC 6.4, after running the association the following information shall be provided to the archive user:
- a. The list of associated frames can potentially contain several hundred entries, which makes it likely confusing and, in fact, useless to list them all. Aggregate information (e.g. total number of frames, total size, etc.) should be displayed instead.
- UR 6.7** Access to proprietary data is regulated by the provisions in RD01, RD02 and RD03.
- UR 6.8** The performance of the “Find Associated Files” function shall allow for a responsive interactive user experience.
- UR 6.9** Authentication through the User Portal shall be possible as required by general policies (see also UC 6.1 and UC 6.6).

Miscellaneous considerations

As detailed in UR 6.4, all associations among FITS files related to this use case are pre-defined and immutable and are solely based on processing and organizational provenance (cf. RD04).

When it comes to the observing conditions, the difference between UR 6.1f and UR 6.1g is that the former is concerned with whether or not the Phase 2 constraints were met. It does not explicitly return the actual atmospheric conditions, which is the scope of the latter.

⁴⁴ For internal data products the processing provenance is encoded in the PRO.REC keywords stored in each product (as of the year 2003).

⁴⁵ An OB can be executed multiple times until the observing constraints are met, hence the OB ID itself is not enough to uniquely identify the relevant NightLog.



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 29 of 37

Use Case 7: From raw file to NightLog excerpts [Obsolete]

This use case is rendered obsolete by the fact that relevant NightLog excerpts as extracted from the Garching NightLog Tool (gNLT) database are now handled directly by the NGHR.

This functionality is to associate to raw science frames the relevant NightLog excerpts.

Breakdown of the use case

The breakdown of the use case is as follows:

- UC 7.1** An archive user searches raw science observations from an archive query interface. According to general policy, logging into the User Portal is not required at this stage (browsing the archive content).
- UC 7.2** Search results are returned in the result page.
- UC 7.3** The user selects a set of results and activates the option “Find excerpts from the NightLog”.
 - a. Alternatively, if providing these items of information is the default behavior, activating the option is not required.
- UC 7.4** Associations are formed.
- UC 7.5** The user selects the datasets to be retrieved, which will include the raw science files and the excerpts from the relevant NightLogs.
- UC 7.6** The user submits a request for those datasets. According to ESO policy, the user must be logged into the user Portal to be able to successfully complete the request.
- UC 7.7** The request is processed and the requested data are made available to the users subject to proprietary protection. In this respect, only associations linked to science frames the user is actually entitled to retrieve are made available to them.

User requirements

The associated user requirements are as follows:

- UR 7.1** The following items shall be associated to each raw science frame and delivered to the user:
 - a. The science raw frames that resulted from the archive query in UC 7.1.
 - b. For each file in UR 7.1a, the excerpts from the NightLog relevant to it.
- UR 7.2** The list of associated files shall be purged from those exclusively linked to science frames the user is not actually entitled to retrieve.



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 30 of 37

- UR 7.3** The list of associated files shall be purged from duplicates so that each file is only served once to the archive user.
- UR 7.4** The connection between the raw nighttime files and the excerpts from the NightLogs (UR 7.1b) shall be based on the OB ID and on the time at which the OB was executed⁴⁶. Both are FITS header keywords.
- UR 7.5** Access to proprietary data is regulated by the provisions in RD01 and RD02.
- UR 7.6** The performance of the “Find excerpts from the NightLog” function shall allow for a responsive interactive user experience.
- UR 7.7** Authentication through the User Portal shall be possible as required by general policies (see also UC 7.1 and UC 7.6).

Miscellaneous considerations

This is essentially like Use Case 3, minus the raw and static calibrations.

Providing this type of very basic ancillary information should be the default behaviour when raw data are queried.

Use Case 8: From raw or master calibrations to raw or reduced science
[Priority 4]

This use case is best illustrated with an example. Case: User finds out his line ratios are not as expected and wonders if the calibration data are at fault. He looks for other reduced or raw science using the same calibrations to see if they exhibit the same effect. He would download other reduced and or raw data to determine if the effect was in the calibration files used.

Breakdown of the use case

The breakdown of the use case is as follows:

- UC 8.1** An archive user searches raw or master calibrations from an archive query interface. According to general policy, logging into the User Portal is not required at this stage (browsing the archive content).
- UC 8.2** Search results are returned in the result page.
- UC 8.3** The user selects a set of results and activates the option “Find Associated Science”.
- UC 8.4** Associations are formed and selected information on them is returned to the user (see UR 8.6 for the details).

⁴⁶ An OB can be executed multiple times until the observing constraints are met, hence the OB ID itself is not enough to uniquely identify the relevant NightLog.



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 31 of 37

- UC 8.5** The user selects the datasets to be retrieved, which will include the calibrations and the files associated to them.
- UC 8.6** The user submits a request for those datasets. According to ESO policy, the user must be logged into the user Portal to be able to successfully complete the request.
- UC 8.7** The request is processed and the requested data are made available to the users subject to proprietary protection. In this respect, only associations linked to science frames the user is actually entitled to retrieve are made available to them.

User requirements

The corresponding user requirements are as follows:

- UR 8.1** The following items shall be associated and delivered to the user:
- a. The raw or master calibrations that resulted from the archive query in UC 8.1.
 - b. For each file in UR 8.1a, the complete set of science data products that were calibrated with them and the complete set of raw science data that would be calibrated with them. Identifying this complete set may require an iterative process to resolve all the dependencies among files.
 - c. For each raw file in UR 8.1a and UR 8.1b, the excerpts from the NightLog relevant to them: these include the overall QC0 grade, the individual Phase 2 constraints and whether or not each was met, text comments from the observer.
 - d. If applicable, for each raw file in UR 8.1a and UR 8.1b, the actual conditions under which the file was acquired (e.g. image quality, sky transparency, moon distance, etc.).
 - e. If available, QC1 parameters and related information (QC plots, etc.).
 - f. The relations among the associated files. If applicable, the format shall be suitable to be imported into the Reflex DataOrganizer actor⁴⁷.
- UR 8.2** The list of associated files in UR 8.1 shall be purged from those exclusively linked to science frames the user is not actually entitled to retrieve.
- UR 8.3** The list of items in UR 8.1 shall be purged from duplicates so that each one is served only once to the archive user.
- UR 8.4** The items in UR 8.1a and UR 8.1b are FITS files.

⁴⁷ At the time of writing, the functionality of the Reflex DataOrganizer to import data associations is not implemented.



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 32 of 37

- a. The connection between the files shall be solely based on matching combinations of FITS header keywords according to rules.
- b. These rules shall be configurable.
- c. All configurations shall be conveniently accessible and editable by the operational groups in charge⁴⁸.
- d. It shall be possible to define break points in time when the definition of a complete set of calibration changes⁴⁹. The history of such break points shall be preserved so that the appropriate definition can be applied to files taken at different times in the life of an instrument⁵⁰ (parameterization of association rules, with the parameters being the beginning and end of the rule validity period).

UR 8.5 The connection between the raw files and the excerpts from the NightLogs (UR 8.1c) shall be based on the OB ID and on the time at which the OB was executed⁵¹. Both are FITS header keywords.

UR 8.6 With respect to UC 8.4, after running the association the following information shall be provided to the archive user:

- a. The list of associated frames can potentially contain several hundred entries, which makes it likely confusing and, in fact, useless to list them all. Aggregate information (e.g. total number of frames, total size, etc.) should be displayed instead.

UR 8.7 The performance of the “Find Associated Science” (UC 8.3) function shall allow for a responsive interactive user experience.

UR 8.8 Access to proprietary data is regulated by the provisions in RD01, RD02 and RD03.

UR 8.9 Authentication through the User Portal shall be possible as required by general policies (see also UC 8.1 and UC 8.6).

Miscellaneous considerations

⁴⁸ For example, by providing the ability to graphically visualize the hierarchical structure intrinsic to association rules.

⁴⁹ A break point is when the association rules change, not when the instrument itself changes due to, e.g. an earthquake or a hardware intervention, without affecting the rules themselves. Examples of break points include the addition or removal of a given type of calibrations and the addition or removal of header keywords.

⁵⁰ Operationally, it may not be possible to reconstruct the complete history of such break points for each instrument mode. This is unfortunate, but acceptable, and should be documented on a case-by-case basis.

⁵¹ An OB can be executed multiple times until the observing constraints are met, hence the OB ID itself is not enough to uniquely identify the relevant NightLog.



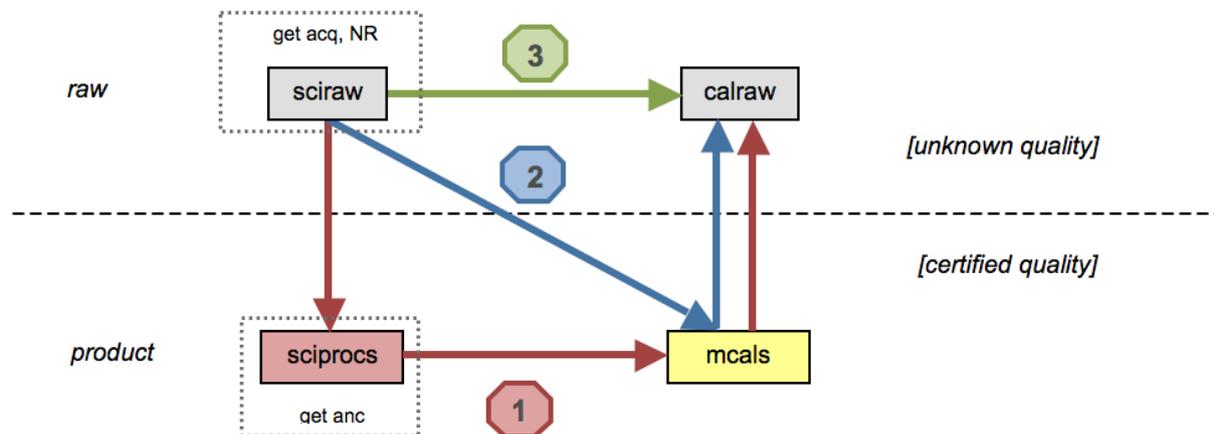
CalSelector: Use Cases and User Requirements

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 33 of 37

This use case is a niche one and has the lowest overall priority ranking. Its implementation is, then, not urgent at this stage and it is listed here for completeness.

7. Final miscellaneous considerations

Use Case 2, Use Case 3 and Use Case 4 share the same starting and end points, i.e. the raw science files and the raw calibrations, respectively⁵². What differentiate these three use cases are the paths connecting the starting and end points, which result in different data types that are included in the associations along the way. This is illustrated in the sketch below (courtesy of Reinhard Hanuschik):



The different use cases can in principle be implemented following different paths, with user options not to serve the files that are not directly relevant to the use case:

- Use Case 2: **Path 2** (direct path) or **Path 1** (if processed science data are available)
- Use Case 3: **Path 3** (direct path) or **Path 1** (if processed science data and master calibrations are available) or **Path 2** (if master calibrations are available)
- Use Case 4: **Path 1** (direct path). In this case the direct path is the only possible one. In this sense, this use case encompasses the other two.

The possibility of actually following the different paths depends on the actual availability of the different data types, which is summarized in Appendix C.

⁵² This would also apply for Use Case 1 and Use Case 5. However, let us leave them aside for the moment because the former is already implemented and the latter is simply about extending access to data packs of the former.



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 34 of 37

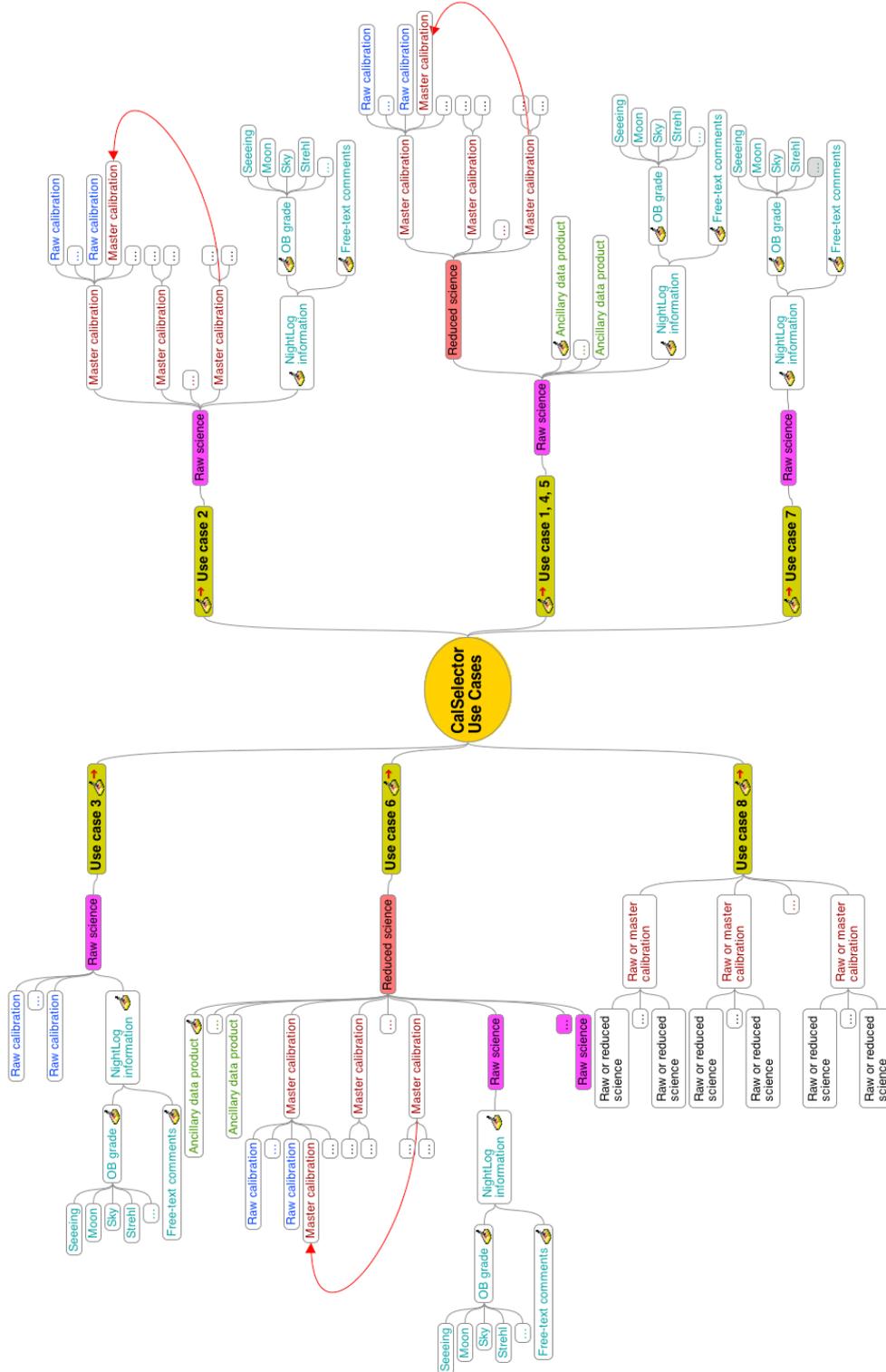
Not taking the most direct path to implement a given use case involves more highly processed data than strictly necessary, which may actually be advantageous, in that it adds useful additional information and quality assurance. For example, and perhaps most notably, implementing Use Case 3 through [Path 2](#), i.e. taking a detour through master calibrations and stored ABs, as opposed to the more direct [Path 3](#), has the advantage of taking automatically into account break points (UR 3.7d) and of only delivering raw calibrations that are certified to yield meaningful master calibrations.



CalSelector: Use Cases and User Requirements

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 35 of 37

Appendix A: Mind Map representation of the use cases





**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 36 of 37

Appendix B: Statistics on CalSelector usage

In order to give a feeling for the load generated by CalSelector requests, we report here aggregate statistics on its usage in the year 2013 (courtesy of Nathalie Fourniol). All numbers refer to raw-to-raw associations (Use Case 3), the only one operational in the reporting period.

- Number of requests: 8,886.
- Number of files: 6,991,294.
- Data volume: 91.5 TB.
- Median execution time of a request: 14 seconds.



**CalSelector:
Use Cases and User Requirements**

Doc: GEN-SPE-ESO-33000-5192
Issue 2
Date 24.01.2014
Page 37 of 37

Appendix C: Overview of currently available association information

The following table contains an overview of the current availability of associations generated by QCG as part of the daily workflow (courtesy of Reinhard Hanuschik). The corresponding Association Blocks (ABs) are accessible as text files from <http://safweb1.hq.eso.org/~qc/<instr>/logs>. In addition, as of the year 2003, the processing provenance of Internal Data Products is recorded in each file's FITS header via the PRO REC FITS keywords.

Instrument	ABs available since*	Science data since*	Status
AMBER	2005-02-01	2005-10-01	SM complete
CRIFES	2006-12-28	2007-04	SM complete
FORS1	2005-10-01	1999-04	1999-2001 missing
FORS2	2005-10-01	2000-04	2000-2002 missing
GIRAFFE	2003-02-01	2003-04	Complete**
HAWKI	2008-11-17	2008-04	2008 missing
ISAAC	2008-02-06	1999-04	1999-2008 missing
MIDI	2008-06-19	2004-07	2004-2006 missing
NACO	2008-12-27	2002-10	2002-2008 missing
SINFONI	2008-01-13	2005-04	2005-2007 missing
UVES	2003-01-04	2000-04	Complete***
VIMOS	2002-12-01	2001-04	2001-2002 missing
VIRCAM	2008-10-01	2009-10	Complete
VISIR	2009-04-05	2005-04	2005-2009 missing
XSHOOTER	2009-04-30	2009-10	Complete

* Most dates refer to SM only; SM+VM information since 2009-06

** Reprocessing: SM complete; VM Medusa complete

*** Reprocessing: SM+VM complete for ECHELLE mode

The mandate of QCG on data processing and quality control does not extend to data from La Silla, so the associations listed in the table above only apply to Paranal data. Nonetheless, between the years 2007 and 2009 the group was tasked to pack Service Mode data from La Silla to further distribute them to the respective PIs. As a result, the corresponding Association Blocks were created and are now available on dfo21 machine. Since this was for SM only, the population of dates is sometimes sparse and strictly limited to the time interval between June 15th, 2007 and June 7th, 2009. The feasibility of creating association rules for La Silla needs to be assessed.