Data Access Working Group
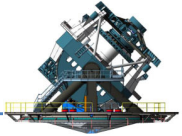
**Beth Willman/William O'Mullane**
LSST Deputy Director/DM Project Manager

Science Advisory Committee Meeting
21st February 2018

Background

- Only two years from the acquisition of commissioning data with ComCam. Many questions regarding data rights and data access remain unanswered, and pressure is growing to have consistent, and well-documented, responses.

- These include both technically oriented questions (How many Event Brokers will be supported?), policy oriented questions (What is a *derived data product*?), and those that fall in between (How will independent DACs be managed?).

- The Data Access Working Group was initiated in late 2017, to define a framework for answering these questions.
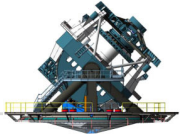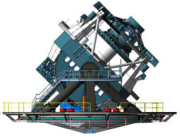
Working Group Scope

LPM-221 (the DAWG charge) was issued in Nov 2017

This working group will collect technical and policy questions related to data access policy for LSST. This short-lived working group will identify the appropriate timescales and people for taking decisions on these questions, will codify the answers to questions that can be answered in the near-term, and will provide a plan going forward on how to consistently answer the full suite of questions on data access. This group will start working early in November and will finish by February 2018.

Membership: Beth Willman, Wil O'Mullane, Melissa Graham, Don Petravick

Working Group Responsibilities

1. Take input from a broad range of interested parties on data access.

2. Catalog and categorize the inputs in terms of technical/policy or mixed.

3. Assign responsible parties where possible to follow those issues and write technical notes or policy documents to answer them.

4. Coordinate with the LSSTC, and other key data rights stakeholders, to ensure the value of data rights is not diluted.

5. Take into consideration off-boarding LSST members, not just adding new ones.
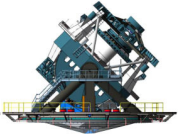
Working Group Tasks (original list)

1. Update Data Access white paper (Document-5373)

2. Draft a policy for Data Access decision making, to be a launching point for discussion with the key stakeholders (including the SAC)

3. Draft an Implementation Plan

   The implementation of any policies or standing up a committee to deal with them will be gradual. The most important aspect of the implementation plan will be the communication plan: When is it appropriate to begin communicating answers to specific data policy questions? What is the appropriate way to be answering questions for which a certain answer does not yet exist? The group should define a time line for implementation of the agreed and future policies. This may take the form of identifying when certain policy decisions are needed i.e. how far in advance of the Data Access center availability do we need to resolve issues around US community membership.

DAWG Meetings

- Nov 27 – Identify Doc 5373 (Access), 13380 (Policy) to be updated. Also discussed brief document for the LSSTC regarding Data Access costs and benefits

- Nov 30 - "Who is a scientist?", Some technical discussions on inCommons and self registration. Draft one pager for LSSTC. Start gathering list of data rights questions.

- Dec 21 (MLG and WOM) - Sorted list of questions. Decided to propose LSE-349 as new draft policy doc replacing Document-13380.

- Jan 4 – Membership and Authorization technical options

- Feb 1 (MLG, WOM) – Working meeting

Example Data Rights Questions

1. Definition of membership for groups of LSST users presently understood at the conceptual level (i.e., how to interpret "U.S. and Chilean Science Community")

2. In addition to the LSST alert stream and EPO quota (defined in LSE-131), are other data products public? (Engineering facility database? orbit catalog? etc.).

3. Initial resource allocation policies (e.g., bandwidth devoted to brokers vs. data downloads vs. EPO processing/data transfer).

4. We've said we'll make images (difference images and nightly coadds, see LSE-163) available in 24 hours, but haven't yet said how much any given person/entity will be able to download. Will full-blown real-time mirroring be supported?

5. How do we address the perception that project people have privileged access to the data and hence may publish ahead of others?

6. How do we deal with Zooniverse integration with the DAC, including metadata scrubbing for data use within citizen science projects – i.e. a data rights holder may set up a citizen science project and select cut outs of many objects to use – some may be in the proprietary data?

Working Group Approach

- Working on private Confluence pages
- Initial list of questions was added to and categorized at least into:
  - Technical
  - Policy
- We are considering how to:
  - Define members (i.e. who in the US has data access?)
  - Define qualifying institutions (to which most members belong – helps w/ implementation)
  - Define a framework for dealing with violations
- We need a future committee, with more complete stakeholder representation, to deal with access/rights questions and requests.

- – LSE so it is under Project control

- – Internal draft being worked on

  - • Will address some questions, and identify others for ongoing consideration

  - • Will include proposed definition for "members" and relevant committees

  - • Differentiates proposed policy statements from explanatory text

- – Not yet ready to discuss with the SAC

  - • Needs to be internally developed (including discussion w/ PST)

  - • Then discussed with Management Board/LSSTC/AMCL for Policy-related aspects

  - • Then with SAC – Hopefully by early summer.

Technical Issues

- Credential Authorization and Security
  - Is 7500 users correct?
  - Identifying via association with a qualifying institution might ease the administrative burden at the DAC.
    - InCommon lists over 600 higher-ed participating institutions – identifies faculty etc. (Many missing - could not be sole source)
  - For independent DACs - What is the security situation?
- Resource Capabilities and Costs
  - Can members "pool" together their quotas to access?

Technical implementation of Policies

- Document- 5373 "Scientific Access to LSST Data and Computing" mixes technical implementation and policy.

- This should probably be replaced by a new LPM (probably) which speaks of technically implementing the policy in LSE-349.

- Such a technical document is likely not the work of the working group, but could instead be a direct action to Data Management and/or the LSST Data Facility (NCSA).

WG Charge – Summary Status

1. Take input from a broad range of interested parties on data access.

   We have a lot already – We will take more in the commenting round

2. Catalog (perhaps in Jira) and categorize the inputs in terms of technical/policy or mixed.

   Currently on Confluence

3. Assign responsible parties where possible to follow those issues and write technical notes or policy documents to answer them.
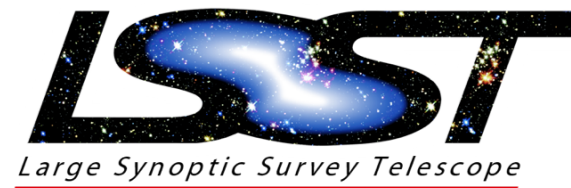
   Partially done on Confluence and LSE-349

4. Coordinate with the LSSTC, and other key data rights stakeholders, to ensure the value of data rights is not diluted.
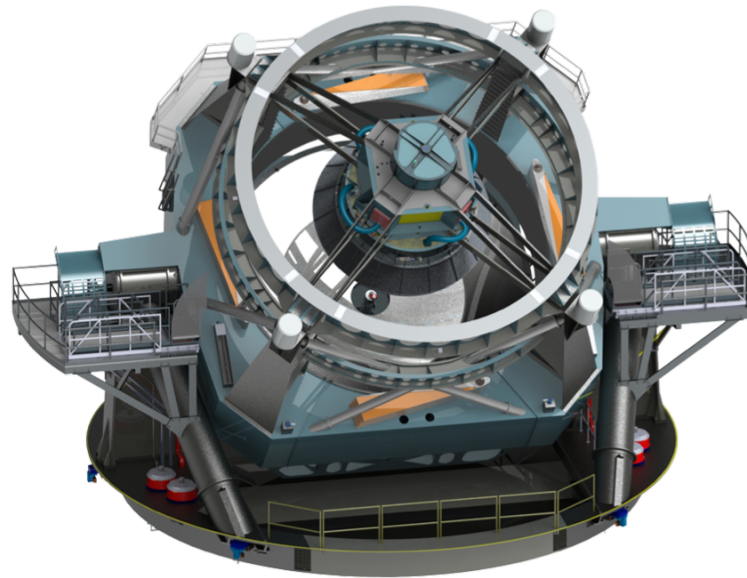
   Next step

5. Take into consideration off-boarding LSST members, not just adding new ones. -

   In the technical list

The End

- 1. Definition of membership for groups of LSST users presently understood at the conceptual level (i.e., how to interpret "U.S. and Chilean Science Community").

- 2. Definition of access rights to specific LSST data products (e.g., the Solar System Object catalogs).

- 3. Definition of access rights to specific LSST U.S. and Chile DAC services (e.g., access to the LSST alert filtering service, or added value catalogs built by users but served from LSST's DACs).

- 4. Will Chilean scientists be able to access the US DAC and vice versa?

- 5. Codification of data and data access rights during Commissioning.

- 6. Initial resource allocation policies (e.g., bandwidth devoted to brokers vs. data downloads vs. EPO processing/data transfer).

- 7. Who is a "scientist" in terms of LSST agreements?

- 8. In addition to the LSST alert stream and EPO quota (defined in LSE-131), are other data products public? (Engineering facility database? orbit catalog? etc.).

- 9. It is currently possible to have data rights but that does not give DAC access rights. What services, if any, does having the data rights entail?

- 10. If a Science Collaboration (SC) decides to build and offer to its members a common data product hosted in the DAC, will their non-DAC-rights members be able to access it?

- 11. We've said we'll make images (difference images and nightly coadds, see LSE-163) available in 24 hours, but haven't yet said how much any given person/entity will be able to download. Will full-blown real-time mirroring be supported?

- 12. Will International Contributors w/o DAC access have access to the LSST software helpdesk?

- 13. How do we deal with unauthorized access? Ensure editors on journals know our policies and who they should ask if they are suspicious?

- 14. How do we address the perception that project people have privileged access to the data and hence may publish ahead of others?

- 15. How do we deal with Zooniverse integration with the DAC, including metadata scrubbing for data use within citizen science projects – i.e. a data rights holder may set up a citizen science project and select cut outs of many objects to use – some may be in the proprietary data?

- 16. What are the data rights on EPO animated images? The exact form of these "movies" is not yet known but we should consider what the access policy is on them.